Data and Society

Resources and Dangers and Opportunities

Kenneth W. Regan

(Includes material from Kenny A. Joseph and some other past CSE199 units.)

CSE199, Fall 2024

Main Problem...

Any Questions?

(Brain scan source, 1987 PSA source)

...And Problems

...And Problems

1. How has the advent of the Internet altered—

...And Problems

1. How has the advent of the Internet altered—
    - —our ecology of personhood?

...And Problems

1. How has the advent of the Internet altered—
    - —our ecology of personhood?
    - —our communal relationships?

...And Problems

1. How has the advent of the Internet altered—
    - —our ecology of personhood?
    - —our communal relationships?
    - —opportunity and equity in society?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?
2. In an Ocean of Data, will we develop "gills"?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?
2. In an Ocean of Data, will we develop “gills”?
3. How much Greater than Gutenberg?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?
2. In an Ocean of Data, will we develop “gills”?
3. How much Greater than Gutenberg?
   - The *Time–Life Top 100 Events of the Last Millennium* placed Gutenberg’s circa-1450 invention of the printing press at #1.

...And Problems

1. How has the advent of the Internet altered—
    - —our ecology of personhood?
    - —our communal relationships?
    - —opportunity and equity in society?
    - —our cognitive functions?
    - —our organization of life experiences?
2. In an Ocean of Data, will we develop “gills”?
3. How much Greater than Gutenberg?
    - The *Time–Life Top 100 Events of the Last Millennium* placed Gutenberg’s circa-1450 invention of the printing press at #1.
4. What ingredients and tools have enabled erecting all of this in only the past 30+ years?

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?
2. In an Ocean of Data, will we develop "gills"?
3. How much Greater than Gutenberg?
   - The *Time–Life Top 100 Events of the Last Millennium* placed Gutenberg's circa-1450 invention of the printing press at #1.
4. What ingredients and tools have enabled erecting all of this in only the past 30+ years?
5. **What tools enable us to understand it?**

...And Problems

1. How has the advent of the Internet altered—
   - —our ecology of personhood?
   - —our communal relationships?
   - —opportunity and equity in society?
   - —our cognitive functions?
   - —our organization of life experiences?
2. In an Ocean of Data, will we develop "gills"?
3. How much Greater than Gutenberg?
   - The *Time–Life Top 100 Events of the Last Millennium* placed Gutenberg's circa-1450 invention of the printing press at #1.
4. What ingredients and tools have enabled erecting all of this in only the past 30+ years?
5. **What tools enable us to understand it?** We will cover some: probabilistic modeling, regression, simulation, preference aggregation, causal graphs, other data analytics...

Picking Up the Gutenberg Theme

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books “Brain Extenders”?

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books “Brain Extenders”?
- As opposed to **Cognition Extenders** as we have today...

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books “Brain Extenders”?
- As opposed to **Cognition Extenders** as we have today...
- Midway: **Imagination Extenders**.

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books "Brain Extenders"?
- As opposed to **Cognition Extenders** as we have today...
- Midway: **Imagination Extenders**. (The writing of *Don Quixote* circa 1605 is #96 on the Time–Life list.)

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books "Brain Extenders"?
- As opposed to **Cognition Extenders** as we have today...
- Midway: **Imagination Extenders**. (The writing of *Don Quixote* circa 1605 is #96 on the Time–Life list.)
- One major impact of Gutenberg's mass democritization of affordable books was spreading political and cultural ideas in waves.

Picking Up the Gutenberg Theme

- **Books** existed long before the printing press.
- The scroll form dominated until the codex was invented around the time of Julius Caesar.
- The Herculaneum scrolls were the private library of the Roman poet/philosopher Philodemus and heirs before Mt. Vesuvius carbonized them in 79 CE.
- In what senses were those books "Brain Extenders"?
- As opposed to **Cognition Extenders** as we have today...
- Midway: **Imagination Extenders**. (The writing of *Don Quixote* circa 1605 is #96 on the Time–Life list.)
- One major impact of Gutenberg's mass democritization of affordable books was spreading political and cultural ideas in waves.
- How does that compare (in speed and mass) to "Memes" and viral content today?

Brain Extenders

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss "8 Hours Without Internet" essays.]**

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss "8 Hours Without Internet" essays.]**
- I [KWR] deal with a special kind of "brain extension": catching those cheat at human chess games by illicitly accessing computer input on which next move to make.

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss "8 Hours Without Internet" essays.]**
- I [KWR] deal with a special kind of "brain extension": catching those cheat at human chess games by illicitly accessing computer input on which next move to make.
- Since Deep Blue defeated Garry Kasparov in 1997, computers have grown to be far better than us at finding the *best next moves*.

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss “8 Hours Without Internet” essays.]**
- I [KWR] deal with a special kind of “brain extension”: catching those cheat at human chess games by illicitly accessing computer input on which next move to make.
- Since Deep Blue defeated Garry Kasparov in 1997, computers have grown to be far better than us at finding the *best next moves*.
- Large Language Models such as **ChatGPT** operate by finding the *best next words*.

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss "8 Hours Without Internet" essays.]**
- I [KWR] deal with a special kind of "brain extension": catching those cheat at human chess games by illicitly accessing computer input on which next move to make.
- Since Deep Blue defeated Garry Kasparov in 1997, computers have grown to be far better than us at finding the *best next moves*.
- Large Language Models such as **ChatGPT** operate by finding the *best next words*.
- Will they—and other forms of **AI** in general—soon supersede us?

Brain Extenders

- Not just Facts and Ideas and Data but also **Computation**.
- Compare using GPS to using a physical map...
- **[Discuss "8 Hours Without Internet" essays.]**
- I [KWR] deal with a special kind of "brain extension": catching those cheat at human chess games by illicitly accessing computer input on which next move to make.
- Since Deep Blue defeated Garry Kasparov in 1997, computers have grown to be far better than us at finding the *best next moves*.
- Large Language Models such as **ChatGPT** operate by finding the *best next words*.
- Will they—and other forms of **AI** in general—soon supersede us?
- Even nearer term: Elon Musk's **Neuralink** brain implant as used to play chess.

The Global Brain

The Global Brain

- E.M. Forster, 1909 short story "The Machine Stops."

The Global Brain

- E.M. Forster, 1909 short story "The Machine Stops."
- Arguably a critique of H.G. Wells's 1905 novel *A Modern Utopia*.

The Global Brain

- E.M. Forster, 1909 short story "The Machine Stops."
- Arguably a critique of H.G. Wells's 1905 novel *A Modern Utopia*.
- **Dystopian sci-fi**: humanity forced to rely on a giant machine regulating an underground biosphere and all aspects of life.

The Global Brain

- E.M. Forster, 1909 short story "The Machine Stops."
- Arguably a critique of H.G. Wells's 1905 novel *A Modern Utopia*.
- **Dystopian sci-fi**: humanity forced to rely on a giant machine regulating an underground biosphere and all aspects of life.
- **Actual reality**: the July 19, 2024 CrowdStrike Crash.

Low-Level Foundations

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.
- Our existing code base is code-based anyway.

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.
- Our existing code base is code-based anyway.
- Analogy: Venice was founded on about 10 million tree logs that were pile-driven into Adriatic Sea shallows.
  - The engineers of 1,100 years ago knew the logs wouldn't rot in that water.

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.
- Our existing code base is code-based anyway.
- Analogy: Venice was founded on about 10 million tree logs that were pile-driven into Adriatic Sea shallows.
  - The engineers of 1,100 years ago knew the logs wouldn't rot in that water.
- Does Code Rot?

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.
- Our existing code base is code-based anyway.
- Analogy: Venice was founded on about 10 million tree logs that were pile-driven into Adriatic Sea shallows.
  - The engineers of 1,100 years ago knew the logs wouldn't rot in that water.
- Does Code Rot? Does it slowly sink?

Low-Level Foundations

- The root cause of the Crowdstrike crash was **an attempted read from a null pointer** in C++ code.
- We will see other low-level bugs that caused famous breaches.
- "No Code" Software Development is not-here-yet and limited.
- Our existing code base is code-based anyway.
- Analogy: Venice was founded on about 10 million tree logs that were pile-driven into Adriatic Sea shallows.
    - The engineers of 1,100 years ago knew the logs wouldn't rot in that water.
- Does Code Rot? Does it slowly sink?
- Your further CS education will show how to build systems from the ground up.

High-Level Issues

High-Level Issues

- Increasingly more of our lives is governed by “Algorithms.”

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm."

High-Level Issues

- Increasingly more of our lives is governed by “Algorithms.”
- Not quite what our CS courses mean by “algorithm.” Often it’s the operation of a **predictive model.**

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications

High-Level Issues

- Increasingly more of our lives is governed by “Algorithms.”
- Not quite what our CS courses mean by “algorithm.” Often it’s the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions

High-Level Issues

- Increasingly more of our lives is governed by “Algorithms.”
- Not quite what our CS courses mean by “algorithm.” Often it’s the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions

High-Level Issues

- Increasingly more of our lives is governed by “Algorithms.”
- Not quite what our CS courses mean by “algorithm.” Often it’s the operation of a **predictive model.**
- Some examples:
    - bank loan applications
    - medical treatment decisions
    - credit scoring
    - college admissions
    - parole decisions

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions
  - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions
  - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions
  - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.
- **The model can be buggy.**

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
    - bank loan applications
    - medical treatment decisions
    - credit scoring
    - college admissions
    - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.
- **The model can be buggy.** (Some people think mine is.)

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions
  - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.
- **The model can be buggy.** (Some people think mine is.)
- **The data can be buggy.**

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
  - bank loan applications
  - medical treatment decisions
  - credit scoring
  - college admissions
  - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.
- **The model can be buggy.** (Some people think mine is.)
- **The data can be buggy.** (Covid greatly skewed **chess ratings**.)

High-Level Issues

- Increasingly more of our lives is governed by "Algorithms."
- Not quite what our CS courses mean by "algorithm." Often it's the operation of a **predictive model.**
- Some examples:
    - bank loan applications
    - medical treatment decisions
    - credit scoring
    - college admissions
    - parole decisions
- The *key ingredient* is the **data** on which the models are **trained**.
- I've built a predictive model trained on high level chess games.
- **The model can be buggy.** (Some people think mine is.)
- **The data can be buggy.** (Covid greatly skewed **chess ratings**.)
- **Datasets from the past have large racial and socioeconomic biases.**

The Ocean of ~~Language~~ ~~Information~~ Data

Before we can talk about **Misinformation**, we must note how **Claude Shannon** in 1947 essentially defined *information* merely as *data*.

> The information $I(x)$ in a datum x equals the minimum length of a program that **generates** x.

This *opposes* our human idea of information because:

The Ocean of ~~Language~~ ~~Information~~ Data

Before we can talk about **Misinformation**, we must note how **Claude Shannon** in 1947 essentially defined *information* merely as *data*.

> The information $I(x)$ in a datum x equals the minimum length of a program that **generates** x.

This *opposes* our human idea of information because:

- Anything with lots of **structure** is defined by a relatively short set of rules that generate it, hence has *low* information.

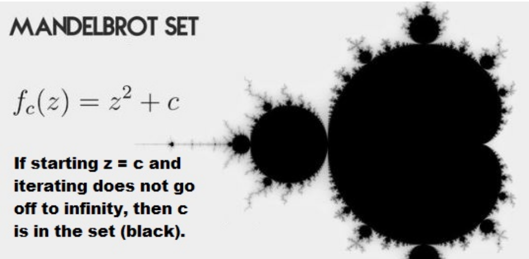
The Ocean of ~~Language~~ ~~Information~~ Data

Before we can talk about **Misinformation**, we must note how **Claude Shannon** in 1947 essentially defined *information* merely as *data*.

> The information $I(x)$ in a datum x equals the minimum length of a program that **generates** x.

This *opposes* our human idea of information because:

- Anything with lots of **structure** is defined by a relatively short set of rules that generate it, hence has *low* information. Example:



Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

> In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?**

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?** If P=NP, then *no*.

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?** If P=NP, then *no*.
- How about using GPT4 to generate lots of code from your problem spec?

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

> In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?** If P=NP, then *no*.
- How about using GPT4 to generate lots of code from your problem spec? (This leverages the **huge** but **fixed** background data that was used to train GPT4.)

Upshot: Any notion of *information* beyond (size-of-)*data* must involve extra criteria specific to its *sender* and *receiver*.

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?** If P=NP, then *no*.
- How about using GPT4 to generate lots of code from your problem spec? (This leverages the **huge** but **fixed** background data that was used to train GPT4.)

Upshot: Any notion of *information* beyond (size-of-)*data* must involve extra criteria specific to its *sender* and *receiver*. **Subjective?**

Data Versus Information—continued

The digits of π are another low-info example. *Whereas:*

- Completely random data has no rules, so no way to abbreviate, which means *high* (but useless!–?) information.

In over 75 years since Shannon, no one has pinned down what "Structured Information" should mean.

- Key impasse in my main professional field of **Computational Complexity**, including the infamous **P Versus NP** question.
- Also the #2 question in my field: **Are pseudorandom generators secure?** If P=NP, then *no*.
- How about using GPT4 to generate lots of code from your problem spec? (This leverages the **huge** but **fixed** background data that was used to train GPT4.)

Upshot: Any notion of *information* beyond (size-of-)*data* must involve extra criteria specific to its *sender* and *receiver*. **Subjective? Biased?**

Gleaning Information From (Your) Data

Gleaning Information From (Your) Data

- Many “Apps”—and what you call your “Algorithms”—are mainly ways of **querying** data stored in The Cloud.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.

Gleaning Information From (Your) Data

- Many “Apps”—and what you call your “Algorithms”—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as “Sequel”) allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**

Gleaning Information From (Your) Data

- Many “Apps”—and what you call your “Algorithms”—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as “Sequel”) allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**—as opposed to the **dark web**.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**—as opposed to the **dark web**.
    - (I maintain gigabytes of deep-web textual data...

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**—as opposed to the **dark web**.
  - (I maintain gigabytes of deep-web textual data... tracking chess tournaments for possible cheating.)

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**—as opposed to the **dark web**.
  - (I maintain gigabytes of deep-web textual data... tracking chess tournaments for possible cheating.)
- A step further is apps that make *inferences* from data. This is where we begin to speak of **Machine Learning**.

Gleaning Information From (Your) Data

- Many "Apps"—and what you call your "Algorithms"—are mainly ways of **querying** data stored in The Cloud.
- GPS is an example of mostly passive information.
- Apps built atop the **Structured Query Language** (SQL, pronounced that way or as "Sequel") allow interactive queries.
- Queries are formulated using Boolean logic, numerics, and other built-in or user-created predicates.
- Queries are addressed to a particular database.
- Internet **search**, on the other hand, can address the whole **searchable web**—as opposed to the **dark web**.
    - (I maintain gigabytes of deep-web textual data... tracking chess tournaments for possible cheating.)
- A step further is apps that make *inferences* from data. This is where we begin to speak of **Machine Learning**.
- Whether the info and inferences are **true** is secondary!

Outline For Remaining Lectures

1. Some further remarks about Data as time allows in this lecture.
2. Our Global Data Village
3. Data Analytics, Search, and AI
4. AI, continued—Project Ideas
5. Societal Computing and Fairness
6. Synthesis.

How Much Data Is There?

How Much Data Is There?

- That is, How Big Is the Internet?

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
    - One **terabyte** = 1,000 **gigabytes**.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
    - One **terabyte** = 1,000 **gigabytes**.
    - One **petabyte** = 1,000 **terabytes**. "**Big Data**"

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
  - One **terabyte** = 1,000 **gigabytes**.
  - One **petabyte** = 1,000 **terabytes**. "**Big Data**"
  - One **exabyte** = 1,000 **petabytes**.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
  - One **terabyte** = 1,000 **gigabytes**.
  - One **petabyte** = 1,000 **terabytes**. "**Big Data**"
  - One **exabyte** = 1,000 **petabytes**.
  - One **zettabyte** = 1,000 **exabytes**.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
  - One **terabyte** = 1,000 **gigabytes**.
  - One **petabyte** = 1,000 **terabytes**. "**Big Data**"
  - One **exabyte** = 1,000 **petabytes**.
  - One **zettabyte** = 1,000 **exabytes**.
  - Next level is called **yottabyte**.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
  - One **terabyte** = 1,000 **gigabytes**.
  - One **petabyte** = 1,000 **terabytes**. “**Big Data**”
  - One **exabyte** = 1,000 **petabytes**.
  - One **zettabyte** = 1,000 **exabytes**.
  - Next level is called **yottabyte**.
- Google now holds about 15 exabytes.

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
    - One **terabyte** = 1,000 **gigabytes**.
    - One **petabyte** = 1,000 **terabytes**. "**Big Data**"
    - One **exabyte** = 1,000 **petabytes**.
    - One **zettabyte** = 1,000 **exabytes**.
    - Next level is called **yottabyte**.
- Google now holds about 15 exabytes. Oops—10?

How Much Data Is There?

- That is, How Big Is the Internet?
- World Wide Web Size.
  - One **terabyte** = 1,000 **gigabytes**.
  - One **petabyte** = 1,000 **terabytes**. "**Big Data**"
  - One **exabyte** = 1,000 **petabytes**.
  - One **zettabyte** = 1,000 **exabytes**.
  - Next level is called **yottabyte**.
- Google now holds about 15 exabytes. Oops—10? OOPS—just 5??

Growth Rate of the Internet

Growth Rate of the Internet

- How much data is being added per minute?

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.
  - One vast category partly weaves through the graphic, but is largely off it.

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.
    - One vast category partly weaves through the graphic, but is largely off it.
    - Once estimated here as comprising **30%** of all Internet *traffic*.

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.
  - One vast category partly weaves through the graphic, but is largely off it.
  - Once estimated here as comprising **30%** of all Internet *traffic*.
  - The musical "Avenue Q" says the Internet was made for it...

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.
  - One vast category partly weaves through the graphic, but is largely off it.
  - Once estimated here as comprising **30%** of all Internet *traffic*.
  - The musical “Avenue Q” says the Internet was made for it...
  - Is it Data? OK, not for the rest of these lectures. . .

Growth Rate of the Internet

- How much data is being added per minute?
- This widget quickly counts up 1TB added data.
- This graphic shows how all the burgeoning data divides into categories.
  - One vast category partly weaves through the graphic, but is largely off it.
  - Once estimated here as comprising **30%** of all Internet *traffic*.
  - The musical "Avenue Q" says the Internet was made for it...
  - Is it Data? OK, not for the rest of these lectures...
- How can the Net's architecture absorb this expansion?

Where Data Lives

Where Data Lives

- Data physically resides on “hard media” in computer systems.

Where Data Lives

- Data physically resides on “hard media” in computer systems.
- **Data Centers**

Where Data Lives

- Data physically resides on “hard media” in computer systems.
- **Data Centers**
  - Often service governments—hopefully with redundancy.

Where Data Lives

- Data physically resides on “hard media” in computer systems.
- **Data Centers**
  - Often service governments—hopefully with redundancy.
  - Service multiple agencies and companies...

Where Data Lives

- Data physically resides on “hard media” in computer systems.
- **Data Centers**
  - Often service governments—hopefully with redundancy.
  - Service multiple agencies and companies...
  - ...as opposed to a data warehouse organized by one company or partnership.

Where Data Lives

- Data physically resides on “hard media” in computer systems.
- **Data Centers**
    - Often service governments—hopefully with redundancy.
    - Service multiple agencies and companies...
    - ...as opposed to a data warehouse organized by one company or partnership.
- Largest floor space is China Telecom–Inner Mongolia. Over 10M sq. ft., bigger than the Pentagon. (Note what first paragraph says about expectation of Google search.)

Where Data Lives

- Data physically resides on "hard media" in computer systems.
- **Data Centers**
  - Often service governments—hopefully with redundancy.
  - Service multiple agencies and companies...
  - ...as opposed to a data warehouse organized by one company or partnership.
- Largest floor space is China Telecom–Inner Mongolia. Over 10M sq. ft., bigger than the Pentagon. (Note what first paragraph says about expectation of Google search.)
- Nevada SuperNAP Reno: 6.2M sq. ft.

Where Data Lives

- Data physically resides on "hard media" in computer systems.
- **Data Centers**
  - Often service governments—hopefully with redundancy.
  - Service multiple agencies and companies...
  - ...as opposed to a data warehouse organized by one company or partnership.
- Largest floor space is China Telecom–Inner Mongolia. Over 10M sq. ft., bigger than the Pentagon. (Note what first paragraph says about expectation of Google search.)
- Nevada SuperNAP Reno: 6.2M sq. ft.
- Chicago Lakeside Technology Center, former champ at 1.1M sq. ft.

But for many users, where it lives virtually is in the Cloud.

Data Management and the Cloud

Data Management and the Cloud

- The Cloud fits under the heading of data management services.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
    - physical maintenance of data;

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
    - physical maintenance of data;
    - recoverability in event of mutation or loss;

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access...

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access...
  - ...**and also improper usage**;

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access...
  - ...**and also improper usage**;
  - compatibility and interoperability;

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access...
  - ...**and also improper usage**;
  - compatibility and interoperability;
  - algorithmic services.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access…
  - …**and also improper usage**;
  - compatibility and interoperability;
  - algorithmic services.
- Many data centers are augmented with **server farms** to do the processing.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access...
  - ...**and also improper usage**;
  - compatibility and interoperability;
  - algorithmic services.
- Many data centers are augmented with **server farms** to do the processing. Could even be for users training their own AI models.

Data Management and the Cloud

- The Cloud fits under the heading of data management services.
- Can be called an internetwork with common structures.
- Services are contracted to subscribers of all kinds: from individuals to huge consortia.
- Responsible for:
  - physical maintenance of data;
  - recoverability in event of mutation or loss;
  - governing access to data;
  - security mechanisms against unauthorized access. . .
  - . . . **and also improper usage**;
  - compatibility and interoperability;
  - algorithmic services.
- Many data centers are augmented with **server farms** to do the processing. Could even be for users training their own AI models.
- **Nontrivial portion of world energy consumption**. (Segue to next unit.)

Part II: A Global Data Village

Part II: A Global Data Village

- “No Man is an Island…” wrote John Donne in 1624.

Part II: A Global Data Village

- "No Man is an Island. . . " wrote John Donne in 1624.
- Then it was a "Meditation."

Part II: A Global Data Village

- "No Man is an Island..." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.

Part II: A Global Data Village

- "No Man is an Island. . ." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).

Part II: A Global Data Village

- "No Man is an Island..." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:

Part II: A Global Data Village

- "No Man is an Island..." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.

Part II: A Global Data Village

- “No Man is an Island…” wrote John Donne in 1624.
- Then it was a “Meditation.” Now pretty much a statement of fact.
- Article, “What Facebook Knows” (old, 2012, but valid).
- Even more along Donne’s lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.
- Some stories from Hurricane Helene now. Starlink story.

Part II: A Global Data Village

- "No Man is an Island. . . " wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.
- Some stories from Hurricane Helene now. Starlink story.

- We will first explore "the interconnectedness of humanity" mathematically.

Part II: A Global Data Village

- "No Man is an Island. . ." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.
- Some stories from Hurricane Helene now. Starlink story.

- We will first explore "the interconnectedness of humanity" mathematically.
- "Six Degrees of Separation" before Internet, now more all-to-all.

Part II: A Global Data Village

- "No Man is an Island..." wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.
- Some stories from Hurricane Helene now. Starlink story.

- We will first explore "the interconnectedness of humanity" mathematically.
- "Six Degrees of Separation" before Internet, now more all-to-all.
- We will relate **graphs** and **games** and Internet search.

Part II: A Global Data Village

- "No Man is an Island…" wrote John Donne in 1624.
- Then it was a "Meditation." Now pretty much a statement of fact.
- Article, "What Facebook Knows" (old, 2012, but valid).
- Even more along Donne's lines, a Floridian during Hurricane Irma was rescued by someone reading her Tweets in California:
- Oct. 2022 Gulf of Mexico rescue via text message story.
- Some stories from Hurricane Helene now. Starlink story.

- We will first explore "the interconnectedness of humanity" mathematically.
- "Six Degrees of Separation" before Internet, now more all-to-all.
- We will relate **graphs** and **games** and Internet search.
- Graphs can be **directed** with arrows or **undirected**.

Games, Payoff Matrices, and Graphs

- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.

Games, Payoff Matrices, and Graphs

- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.
- Chess counts even though it has drawn outcomes. Players have **perfect information** about the current state of the game *in principle*, but even computers find it too *complex* to play perfectly.

Games, Payoff Matrices, and Graphs

- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.
- Chess counts even though it has drawn outcomes. Players have **perfect information** about the current state of the game *in principle*, but even computers find it too *complex* to play perfectly.
- Poker is a zero-sum game of **imperfect information**—you don't know what cards others have.

Games, Payoff Matrices, and Graphs

- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.
- Chess counts even though it has drawn outcomes. Players have **perfect information** about the current state of the game *in principle*, but even computers find it too *complex* to play perfectly.
- Poker is a zero-sum game of **imperfect information**—you don't know what cards others have.
- Rock-Paper-Scissors is a simpler example with *simultaneous play*.

Games, Payoff Matrices, and Graphs

- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.
- Chess counts even though it has drawn outcomes. Players have **perfect information** about the current state of the game *in principle*, but even computers find it too *complex* to play perfectly.
- Poker is a zero-sum game of **imperfect information**—you don't know what cards others have.
- Rock-Paper-Scissors is a simpler example with *simultaneous play*.
- Describable as a **single-matrix game** like so:

You\Oppt.	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

Some Strategizing

Some Strategizing

- If you always pick Rock, Oppt. may **learn** to always pick Paper.

Some Strategizing

- If you always pick Rock, Oppt. may **learn** to always pick Paper.
- If Oppt. picked Rock last turn, you might reason: "ey won't play Rock again. So choose Scissors next..."

Some Strategizing

- If you always pick Rock, Oppt. may **learn** to always pick Paper.
- If Oppt. picked Rock last turn, you might reason: "ey won't play Rock again. So choose Scissors next..."
- Any completely rule-based (buzzword: *deterministic*) strategy can be beaten by someone *who knows your playbook*.

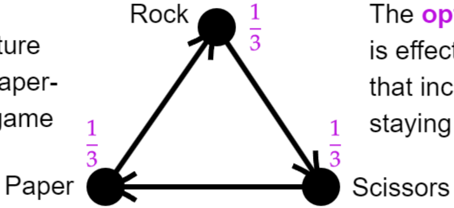
Some Strategizing

- If you always pick Rock, Oppt. may **learn** to always pick Paper.
- If Oppt. picked Rock last turn, you might reason: "ey won't play Rock again. So choose Scissors next..."
- Any completely rule-based (buzzword: *deterministic*) strategy can be beaten by someone *who knows your playbook*.
- Only foolproof way: a **completely random** strategy. Here: roll a die and play Rock on 1 or 2, Paper on 3 or 4, and Scissors on 5 or 6.

Some Strategizing

- If you always pick Rock, Oppt. may **learn** to always pick Paper.
- If Oppt. picked Rock last turn, you might reason: "ey won't play Rock again. So choose Scissors next..."
- Any completely rule-based (buzzword: *deterministic*) strategy can be beaten by someone *who knows your playbook*.
- Only foolproof way: a **completely random** strategy. Here: roll a die and play Rock on 1 or 2, Paper on 3 or 4, and Scissors on 5 or 6.
- But since this is a **fair game**, you can't expect to win either.

Graph picture of Rock-Paper-Scissors game



The **optimal random strategy** is effected by a **random walk** that includes the option of staying on your **current node**.

The walk is on the corresponding **undirected** graph.

Non-Zero-Sum and Multiplayer Games

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**. Mathematically detailed example.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**. Mathematically detailed example.
- Upshot is that **random walk** on G is often (near-)optimal strategy.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**. Mathematically detailed example.
- Upshot is that **random walk** on G is often (near-)optimal strategy.
- Can have $N > 2$ players or as *solitaire*—

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**. Mathematically detailed example.
- Upshot is that **random walk** on G is often (near-)optimal strategy.
- Can have $N > 2$ players or as *solitaire*—like vs. house at blackjack.

Non-Zero-Sum and Multiplayer Games

- Could play with rule that Paper+Scissors wins for *both* players.
- **Non-zero-sum** games use separate **payoff matrix** for each player.
- Could **pre-arrange** choices—but then the Paper player could **defect** by choosing Rock.
- Similar games can be played on any graph G, not just the triangle.
- Each player secretly chooses a node, win (or lose) if they're linked.
- If they choose the same node, can say *zero* or whatever disincentive.
- Then playing nodes with lots of links is attractive—but not **dominant**. Mathematically detailed example.
- Upshot is that **random walk** on G is often (near-)optimal strategy.
- Can have $N > 2$ players or as *solitaire*—like vs. house at blackjack.
- **Internet Search** is a solitaire game where the payoff to you is the *non-quantified* usefulness of the found pages to you.

The Internet as a Graph

The Internet as a Graph

- Webpages form a big graph with pages as nodes and links as edges.

The Internet as a Graph

- Webpages form a big graph with pages as nodes and links as edges.
- Web crawlers enable finding the entire accessible Web (not the dark web).

The Internet as a Graph

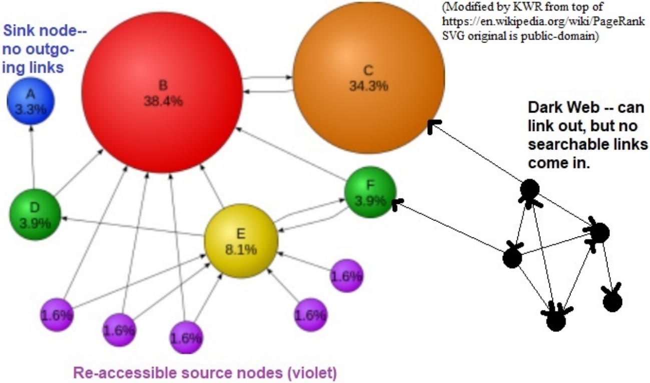
- Webpages form a big graph with pages as nodes and links as edges.
- Web crawlers enable finding the entire accessible Web (not the dark web). Need only store node URL and all its outgoing link URLs.

The Internet as a Graph

- Webpages form a big graph with pages as nodes and links as edges.
- Web crawlers enable finding the entire accessible Web (not the dark web). Need only store node URL and all its outgoing link URLs.
- Can remember links in reverse, so as to treat graph as undirected.

The Internet as a Graph

- Webpages form a big graph with pages as nodes and links as edges.
- Web crawlers enable finding the entire accessible Web (not the dark web). Need only store node URL and all its outgoing link URLs.
- Can remember links in reverse, so as to treat graph as undirected.
- You must cite Web pages used for HW and presentations.



Google PageRank's Graph-Structure Insight

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.
- If many pages link in to P, then many "vote" P as important.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.
- If many pages link in to P, then many "vote" P as important.
- A good **proxy** for the unknown—and unobservable—user value.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.
- If many pages link in to P, then many "vote" P as important.
- A good **proxy** for the unknown—and unobservable—user value.
- So this is a **solitaire** form of a game on a graph—focusing on the portion G_Q filtered by relevance to the query Q.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.
- If many pages link in to P, then many "vote" P as important.
- A good **proxy** for the unknown—and unobservable—user value.
- So this is a **solitaire** form of a game on a graph—focusing on the portion G_Q filtered by relevance to the query Q. **The insight:**

A random walk on G_Q is a good strategy in this game.

Computing the walk probabilities gives weights $w_Q(P)$ on pages in G_Q. Return them in that order.

Google PageRank's Graph-Structure Insight

- Early search engines only computed relevance scores $r(P, Q)$ of pages P to a query Q.
- Not always $\approx$ real user value. But useful as an initial stage.
- Google's founders recognized that links are user votes of value.
- If many pages link in to P, then many "vote" P as important.
- A good **proxy** for the unknown—and unobservable—user value.
- So this is a **solitaire** form of a game on a graph—focusing on the portion G_Q filtered by relevance to the query Q. **The insight:**

$$\boxed{\text{A random walk on } G_Q \text{ is a good strategy in this game.}}$$

Computing the walk probabilities gives weights $w_Q(P)$ on pages in G_Q. Return them in that order. (Alternative: in order of $w'_Q(P) \cdot r(P, Q)$.)

Societal Boons and Banes

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.
- Promotes **backscratching**.

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.
- Promotes **backscratching**. **Clique-ishness.** **Echo chambers...**

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.
- Promotes **backscratching**. **Clique-ishness. Echo chambers...**
- Google's $w(P)$ may be as purely **democratic** as possible—and neutral by design—but in practice leans Democratic.

Societal Boons and Banes

- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.
- Promotes **backscratching**. **Clique-ishness.** **Echo chambers...**
- Google's $w(P)$ may be as purely **democratic** as possible—and neutral by design—but in practice leans Democratic.
- **Fair** to "let chips fall where they may"?

Societal Boons and Banes

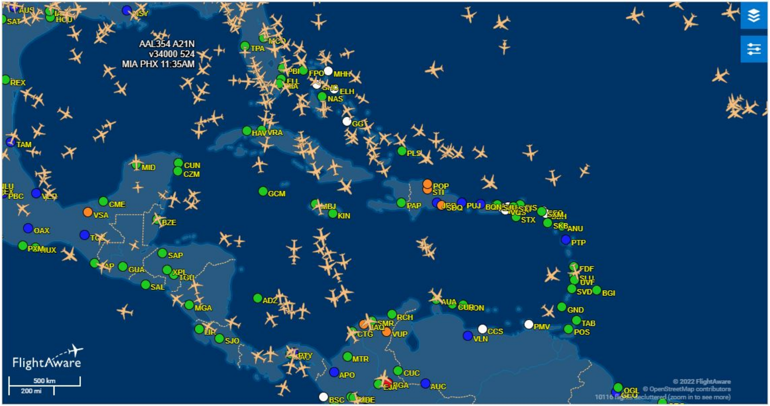
- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> It concentrates power according to those who create many well-linked webpages.

- Linking out to webpages Q that link in to your P raises $w(P)$...
- ...maybe even when you create lots of those pages Q yourself.
- Promotes **backscratching**. **Clique-ishness. Echo chambers...**
- Google's $w(P)$ may be as purely **democratic** as possible—and neutral by design—but in practice leans Democratic.
- **Fair** to "let chips fall where they may"? Or is there real collusion?

A “Semi-Structured” Example (of Inferencing)

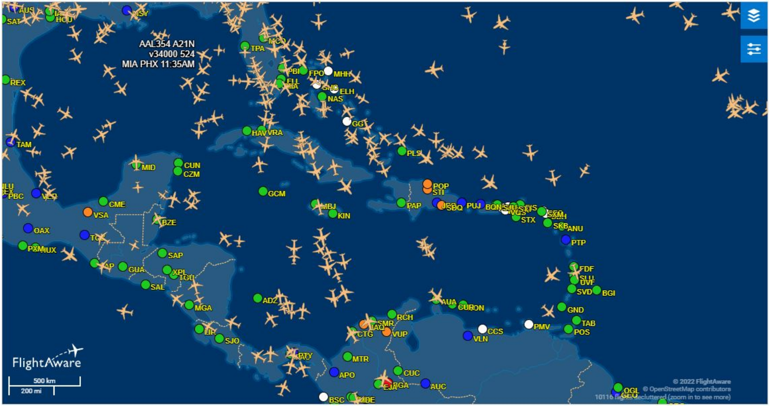
FlightAware Live Tracker, Monday 9/19/22, about 11am:



Why almost no planes over Puerto Rico and the Dominican Republic + Haiti?

A “Semi-Structured” Example (of Inferencing)

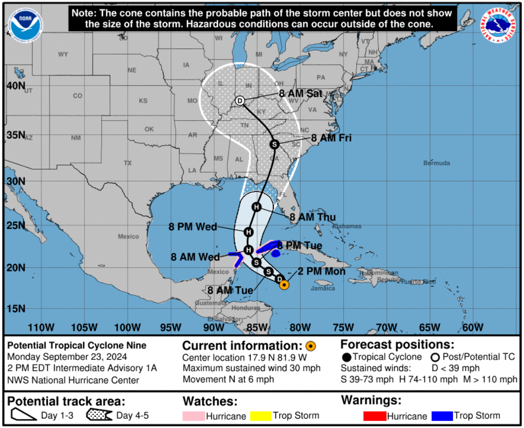
FlightAware Live Tracker, Monday 9/19/22, about 11am:



Why almost no planes over Puerto Rico and the Dominican Republic + Haiti? Compared to right now...

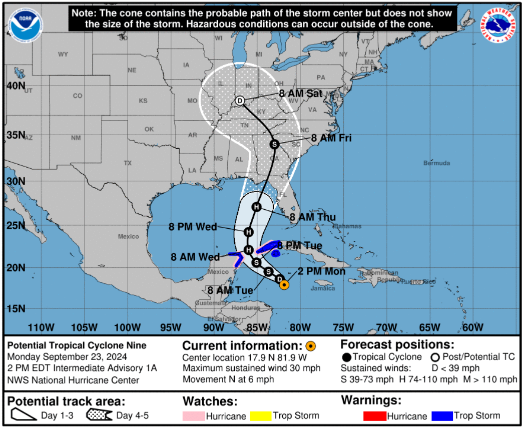
And what about north of the Black Sea?

Hurricane Tracking—Helene By NOAA



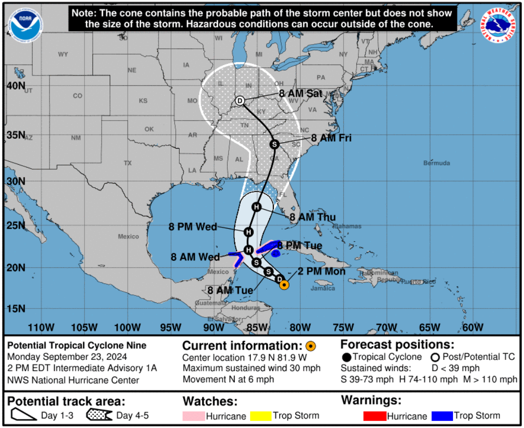
Note **error bars** around the forecasted track. Was spot-on.

Hurricane Tracking—Helene By NOAA



Note **error bars** around the forecasted track. Was spot-on. (But, Otis 2023 was a forecasting failure.)

Hurricane Tracking—Helene By NOAA



Note **error bars** around the forecasted track. Was spot-on. (But, Otis 2023 was a forecasting failure.) Still remnants. Track of power outages.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in “8 Hours” essays.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in “8 Hours” essays.
- Fire Information for Resource Management System (FIRMS).

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
    - Originally built by NASA for fighting wildfires.
    - But now used for war tracking.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.
- Also JominiW. Note the 5-to-10 kilometer scale in this example.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.
- Also JominiW. Note the 5-to-10 kilometer scale in this example.
- UA Weapons Tracker confirmed equipment losses until 10/23/2023.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in “8 Hours” essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.
- Also JominiW. Note the 5-to-10 kilometer scale in this example.
- UA Weapons Tracker confirmed equipment losses until 10/23/2023.
- For our purposes, we can say they can map the front within a mile or two of accuracy, also based on reports (when confirmed).

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.
- Also JominiW. Note the 5-to-10 kilometer scale in this example.
- UA Weapons Tracker confirmed equipment losses until 10/23/2023.
- For our purposes, we can say they can map the front within a mile or two of accuracy, also based on reports (when confirmed).
- Real-time location of individual units and large equipment is dicier...if *we* could do it, the other army could.

Mapping and Geolocation

- *Google Maps*, GPS, and similar—lack noted in "8 Hours" essays.
- Fire Information for Resource Management System (FIRMS).
  - Originally built by NASA for fighting wildfires.
  - But now used for war tracking.
- Daily mapping of the front in Ukraine: ISW, LiveUAMap, DefMon3.
- Also JominiW. Note the 5-to-10 kilometer scale in this example.
- UA Weapons Tracker confirmed equipment losses until 10/23/2023.
- For our purposes, we can say they can map the front within a mile or two of accuracy, also based on reports (when confirmed).
- Real-time location of individual units and large equipment is dicier...if *we* could do it, the other army could.
- Part of **OSINT**: Open-Source Intelligence.

OSINT on Home Soil?

Can we tell whether cats and dogs are being eaten in Springfield, Ohio?

OSINT on Home Soil?

Can we tell whether cats and dogs are being eaten in Springfield, Ohio?

- Definitely August 16 arrest of US citizen for eating cat in Canton.

OSINT on Home Soil?

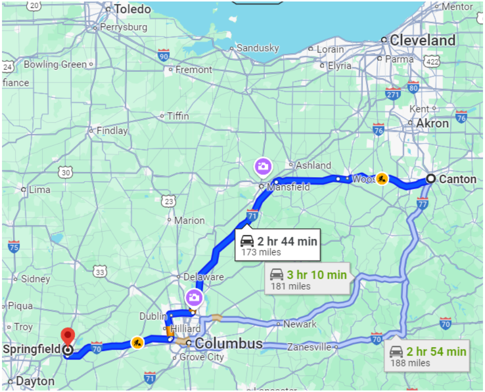
Can we tell whether cats and dogs are being eaten in Springfield, Ohio?

- Definitely August 16 arrest of US citizen for eating cat in Canton.
- Springfield is over 170 miles from Canton.

OSINT on Home Soil?

Can we tell whether cats and dogs are being eaten in Springfield, Ohio?

- Definitely August 16 arrest of US citizen for eating cat in Canton.
- Springfield is over 170 miles from Canton.
- Reports of lost pets in Springfield coming now—more than usual?



Other Internet "Truther"-to-Truthiness-to-Truth

Other Internet “Truther”-to-Truthiness-to-Truth

- **Trump Assassination Attempt.**

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.

Other Internet “Truther”-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC “debate whistleblower” complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.

Other Internet “Truther”-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC “debate whistleblower” complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?
- This example is better-**controlled** than other considered ones:

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?
- This example is better-**controlled** than other considered ones:
  - Can trace at least one referenced authority.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?
- This example is better-**controlled** than other considered ones:
  - Can trace at least one referenced authority.
  - Properties of Google's "Algorithm" arguably come into play.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?
- This example is better-**controlled** than other considered ones:
  - Can trace at least one referenced authority.
  - Properties of Google's "Algorithm" arguably come into play.
  - Dates to 2012 but still well-preserved on the Net.

Other Internet "Truther"-to-Truthiness-to-Truth

- **Trump Assassination Attempt**. IMO bullet visible in photo (alt), and people died, so end-of-story: not staged.
- Hurricane Helene Relief: Example. Not misinfo.
- Faked ABC "debate whistleblower" complaint.
- Vote Count Misinfo. 2020 article I wrote.
- Some cases clear, others unclear. *No magic truth bullet*.
- But tracing the *flow* of stories and assertions is very much a data task.
- **HW Example:** What % of the Internet is Porn? Is it **30%**?
- This example is better-**controlled** than other considered ones:
  - Can trace at least one referenced authority.
  - Properties of Google's "Algorithm" arguably come into play.
  - Dates to 2012 but still well-preserved on the Net.
  - In 2017 it passed my filters and those of some organizations that have since taken it down.

Scientific Data

Scientific Data

- Example: NIH Gene Expression Omnibus.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.
- Posting data makes this possible by 3rd-parties.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.
- Posting data makes this possible by 3rd-parties.
- Center For Open Science—emphasizes rigor and replication in social, medical, and environmental studies.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.
- Posting data makes this possible by 3rd-parties.
- Center For Open Science—emphasizes rigor and replication in social, medical, and environmental studies.
- Impetus to be public—except mainly for *privacy* concerns.

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.
- Posting data makes this possible by 3rd-parties.
- Center For Open Science—emphasizes rigor and replication in social, medical, and environmental studies.
- Impetus to be public—except mainly for *privacy* concerns.
- Tension over *proprietary* aspects, especially for NSF grants, public universities. . .

Scientific Data

- Example: NIH Gene Expression Omnibus.
- Accepts submissions from Excel, XML, even plaintext but formatted like this.
- NASA Exoplanet Archive
- Key concern is **Reproducibility**.
- For example, someone else analyzing the raw exoplanet data should reach closely similar conclusions.
- Posting data makes this possible by 3rd-parties.
- Center For Open Science—emphasizes rigor and replication in social, medical, and environmental studies.
- Impetus to be public—except mainly for *privacy* concerns.
- Tension over *proprietary* aspects, especially for NSF grants, public universities. . .
- Look at all these public datasets!

Business Data

Business Data

- Impetus to be *proprietary*.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.
- Same concerns apply to government agencies.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
    - Data contracted to be used by clients.
    - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*. . .

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
    - Data contracted to be used by clients.
    - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*...
- ...with less responsibility than scientists to establish *causation*.

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*...
- ...with less responsibility than scientists to establish *causation*.
- Example: "Binge-Watching TV Is Killing Us."

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*...
- ...with less responsibility than scientists to establish *causation*.
- Example: "Binge-Watching TV Is Killing Us."
- Or do already sick and less-active people watch more TV?

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*...
- ...with less responsibility than scientists to establish *causation*.
- Example: "Binge-Watching TV Is Killing Us."
- Or do already sick and less-active people watch more TV?
- Either way, can insert targeted ads...

Business Data

- Impetus to be *proprietary*.
- Profit$ replace reproducibility as regards validation.
- Two layers of privacy concerns:
  - Data contracted to be used by clients.
  - Data gathered on customers and competitors.
- Same concerns apply to government agencies.
- Can build *models* based on past record and *correlations*...
- ...with less responsibility than scientists to establish *causation*.
- Example: "Binge-Watching TV Is Killing Us."
- Or do already sick and less-active people watch more TV?
- Either way, can insert targeted ads...
- (Silly new example of correlation-versus-causation: do the KC Chiefs lose when Taylor Swift isn't at the game? Madden '24)

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

> Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title “Haupia (Coconut Pudding)” is public.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title "Haupia (Coconut Pudding)" is public.
- Has 13 ingredients and the recipe takes 17 steps, 3 unnecessary.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title "Haupia (Coconut Pudding)" is public.
- Has 13 ingredients and the recipe takes 17 steps, 3 unnecessary.

Does not give away the ingredients or their amounts or the instructions.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title “Haupia (Coconut Pudding)” is public.
- Has 13 ingredients and the recipe takes 17 steps, 3 unnecessary.

Does not give away the ingredients or their amounts or the instructions.

Metadata may be admissible in court when private content isn’t.

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title “Haupia (Coconut Pudding)” is public.
- Has 13 ingredients and the recipe takes 17 steps, 3 unnecessary.

Does not give away the ingredients or their amounts or the instructions.

Metadata may be admissible in court when private content isn’t.
E.g. time and duration (and recipient??) of cell phone calls.
[Discuss 2010 French chess cheating case and civil vs. criminal law.]

Data, Metadata, and Privacy

A rough working definition of **metadata** is:

Data in XML headers and in `<tag ATTR=...>` attributes

Recipe example (paywalled but *hommaged*) metadata includes:

- It is a dessert.
- Serves 6 people and takes 10 minutes to prepare.
- *Maybe* the title "Haupia (Coconut Pudding)" is public.
- Has 13 ingredients and the recipe takes 17 steps, 3 unnecessary.

Does not give away the ingredients or their amounts or the instructions.

Metadata may be admissible in court when private content isn't.
E.g. time and duration (and recipient??) of cell phone calls.
[Discuss 2010 French chess cheating case and civil vs. criminal law.]

• Major controversy over gathering metadata by law enforcement and intelligence.

Privacy Via Slightly Fake News

Privacy Via Slightly Fake News

- Many databases allow public access to “aggregates” such as mean, median, max, min, “90th percentile” values.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.
- Do the math: they scored only about 50 between them—they bombed it!

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.
- Do the math: they scored only about 50 between them—they bombed it!
- **Differential Privacy** says to fuzz up aggregate values by $\pm\epsilon$.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.
- Do the math: they scored only about 50 between them—they bombed it!
- **Differential Privacy** says to fuzz up aggregate values by $\pm\epsilon$.
- Say $\epsilon = 1\%$.

Privacy Via Slightly Fake News

- Many databases allow public access to “aggregates” such as mean, median, max, min, “90th percentile” values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.
- Do the math: they scored only about 50 between them—they bombed it!
- **Differential Privacy** says to fuzz up aggregate values by $\pm\epsilon$.
- Say $\epsilon = 1\%$. Then 75.0 vs. 74.0 could easily have been “random variation.” We don’t really know.

Privacy Via Slightly Fake News

- Many databases allow public access to "aggregates" such as mean, median, max, min, "90th percentile" values.
- Typified by allowing students to see the class average on UBLearns.
- Say 98 students average 75.1 on a test, then 2 in Band make it up.
- Say class average slips to 74.1.
- Do the math: they scored only about 50 between them—they bombed it!
- **Differential Privacy** says to fuzz up aggregate values by $\pm\epsilon$.
- Say $\epsilon = 1\%$. Then 75.0 vs. 74.0 could easily have been "random variation." We don't really know.
- Has been a special research topic at UB CSE.

Hacks, Crime, Legal Contours, and the Net

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause.

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today…

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
    - GLL blog post, "Security Via Surrender"
    - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.
- What does "copyright" mean when copying is so seamless? Article.

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today…
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.
- What does "copyright" mean when copying is so seamless? Article.
- Programming language meanings such as *read-only*, *local copy*, *temporary* are shaping legal contours.

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.
- What does "copyright" mean when copying is so seamless? Article.
- Programming language meanings such as *read-only*, *local copy*, *temporary* are shaping legal contours.
- After a "hack," who bears responsibility—and how much?

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifiying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.
- What does "copyright" mean when copying is so seamless? Article.
- Programming language meanings such as *read-only*, *local copy*, *temporary* are shaping legal contours.
- After a "hack," who bears responsibility—and how much?
- 1998 DMCA: Internet providers not responsible.

Hacks, Crime, Legal Contours, and the Net

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Cloudbleed and simple 2017 cause. Too many examples today...
- Systems may cope by *verifying* data and changing data's *nature*:
  - GLL blog post, "Security Via Surrender"
  - GLL blog post, "Making Public Information Secret"
- Even with authorized access, **fair use** of public data is an issue.
- What does "copyright" mean when copying is so seamless? Article.
- Programming language meanings such as *read-only*, *local copy*, *temporary* are shaping legal contours.
- After a "hack," who bears responsibility—and how much?
- 1998 DMCA: Internet providers not responsible.
- For misuse of Bram Cohen's BitTorrent—not so clear. Cut deal in 2005 with Motion Picture Association of America to follow DMCA.

Important CSE199 Items

- Regrade requests are due 11/22 for grades before 11/16
  - All materials except Cohort Survey 2 have been graded.
  - If you think something is wrong, request a regrade.
  - See the course announcements for instructions
  - Course Instructor Email addresses are in the syllabus!
  - Any regrade request that does not follow the instructions will be ignored
- No class next Monday (11/25)
- But otherwise, come to class!
  - Every 3 missed lectures/recitations is a half-letter drop in your grade!
- Nominate yourself to be a 199 TA next fall
  - Fill in the nomination form you'll get as a course announcement
- Reminder: All of this is in a course announcement on UBLearns that you got this morning.

Part III: Data Analytics

We will cover the following tools and some of their societal implications:

1. Linear Regression: $Y = a + bX$, $Z = a + bX + cY$, and so on.
2. Causal Inference, Graphs, and Caveats.
3. Probabilistic Modeling.
4. Predictive Modeling.
5. Preference Aggregation:
   - Voting.
   - Ranking and Rating.
   - Polling and Poll Aggregation.
6. Internet Search. (already covered last week)

Many topics are left uncovered. Search will reappear in Wednesday's coverage of machine learning, sentiment analysis, and AI.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run\_Likelihood} = a + b \cdot (\text{Pitch\_Count})$.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run\_Likelihood} = a + b{\cdot}(\text{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\mathrm{Run\_Likelihood} = a + b \cdot (\mathrm{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run\_Likelihood} = a + b \cdot (\text{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run\_Likelihood} = a + b \cdot (\text{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.
- Strength of **correlation** is measured by R—or its square, R^2.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run\_Likelihood} = a + b \cdot (\text{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.
- Strength of **correlation** is measured by R—or its square, R^2.
- But whether this amounts to **causation** may remain problematic.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.
- The **average error** per move, measured as **ACPL** for "average centipawn loss."

Chess Example

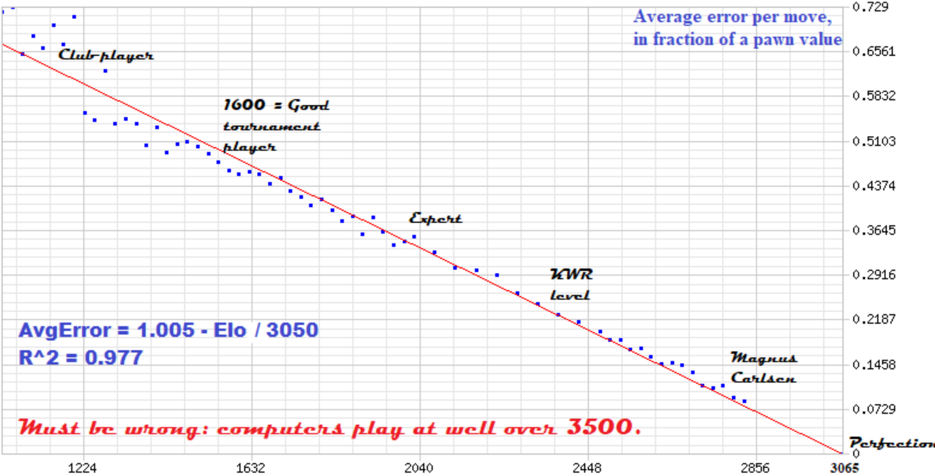
My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level.

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

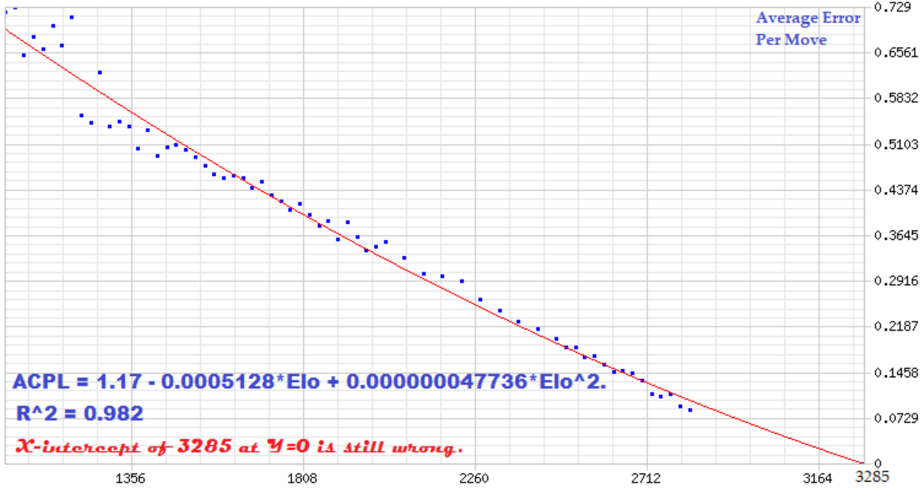
I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.
- The **average error** per move, measured as **ACPL** for "average centipawn loss." My **ASD** variant *scales* loss according to the overall advantage.

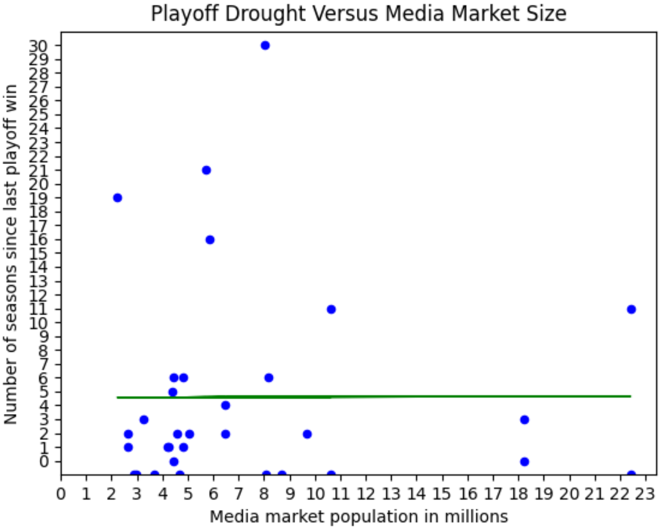
Linear Model: ACPL = $a + b \cdot$Elo Rating



Quadratic Fit—Only Marginally Better



A Desired Null Result? (data from a year ago)



Issues With Linear Models

Issues With Linear Models

- Do we have all the relevant input/independent factors?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
    - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
    - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
    - Hence must be more to chess **skill**—ACPL is at most **accuracy**.

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
    - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
    - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
    - If **one player's** performance is way above the line, = cheating?
    - What if ey had really easy positions?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
    - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
    - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
    - If **one player's** performance is way above the line, = cheating?
    - What if ey had really easy positions? I.e., does inference *scale down?*

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
    - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
    - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
    - If **one player's** performance is way above the line, = cheating?
    - What if ey had really easy positions? I.e., does inference *scale down?*
    - Can you use the regression line to **predict** on (i.e., *bet on*) a given game turn whether P will play the move the computer likes?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors—which might be **confounding** or **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?
  - What if ey had really easy positions? I.e., does inference *scale down?*
  - Can you use the regression line to **predict** on (i.e., *bet on*) a given game turn whether P will play the move the computer likes?
- We need a stronger **probabilistic model** that individuates game positions.

2. Causal Inference and Causal Graphs

- Causal graphs help map potential causes in a complex system.

2. Causal Inference and Causal Graphs

- Causal graphs help map potential causes in a complex system.
- Python Causality Handbook Ch. 4 (best source I've found)

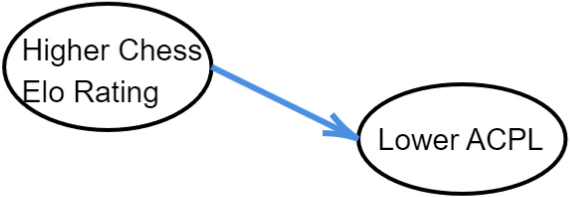
2. Causal Inference and Causal Graphs

- Causal graphs help map potential causes in a complex system.
- Python Causality Handbook Ch. 4 (best source I've found)
- Causal graphs have nodes—black or red for conditioning—and arrows.

2. Causal Inference and Causal Graphs

- Causal graphs help map potential causes in a complex system.
- Python Causality Handbook Ch. 4 (best source I've found)
- Causal graphs have nodes—black or red for conditioning—and arrows.
- They can help ascertain
  - which are genuine causes—as opposed to mere correlations or null effects,and
  - which variables in the system can helpfully be **regressed** or **conditioned on**.

Let's start with a simple example and see how considerations can mushroom:



Transitive and Confounding Causes

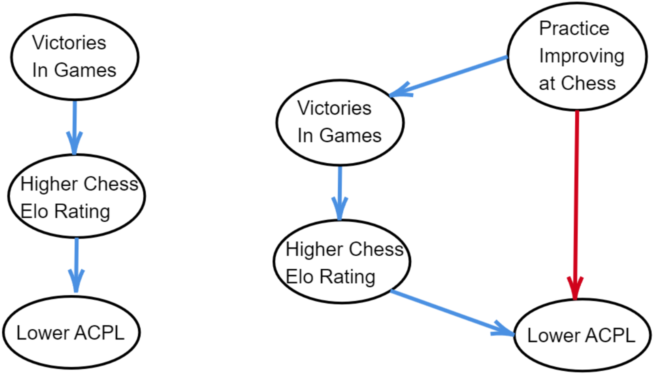
- **Transitive**: if A causes B and B causes C, then A causes C.

Transitive and Confounding Causes

- **Transitive**: if A causes B and B causes C, then A causes C.
- But if we have a lurking *common cause* D of both our *source* and *intended target*, then it can **confound** the smaller-scale analysis.

Transitive and Confounding Causes

- **Transitive**: if A causes B and B causes C, then A causes C.
- But if we have a lurking *common cause* D of both our *source* and *intended target*, then it can **confound** the smaller-scale analysis.
- I faced this when the pandemic caused official chess ratings to lag true skill. Case of wrongly accused player.



Conditioning

- Suppose we study players in separate segments of similar Elo ratings.

Conditioning

- Suppose we study players in separate segments of similar Elo ratings.
- Called a **cross-sectional** or **latitudinal** study.

Conditioning

- Suppose we study players in separate segments of similar Elo ratings.
- Called a **cross-sectional** or **latitudinal** study.
- The factor defining each segment is **conditioned on** and shows in **red**.

Conditioning on the middle node of a causal chain can sever the "A causes C" inference. A and C may even show as **independent** in the conditioned slices---here, because lower error (higher accuracy) might not imply more wins when players of the same rating are in action. Some players may even win more *via* higher ACPL if it tempts their opponents into playing wildly.

(We will do more causal graph examples next week.)

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
    - There is a hidden variable here: the **aptitudes** or the students.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
  - Much like chess ratings R—but student aptitudes aren't published.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
    - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
    - Much like chess ratings R—but student aptitudes aren't published.
    - Nevertheless, judging the difficulty/likelihoods of questions is needed to estimate the grading curve and so design a **fair** exam.

3. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
  - Much like chess ratings R—but student aptitudes aren't published.
  - Nevertheless, judging the difficulty/likelihoods of questions is needed to estimate the grading curve and so design a **fair** exam.
  - It helps to be *confident* that the class won't just bomb your exam.

Grading Probabilistic Forecasters

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?
- From "hedge-fund perspective," it was B: $14/18 = 77.8\%$.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?
- From "hedge-fund perspective," it was B: $14/18 = 77.8\%$.
- My chess model's probability forecasts are similarly accurate within $\sim 5\%$.

4. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$

4. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

4. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

In my model, the m_j are possible moves in chess positions.

Chess and Tests—With Partial Credits (Or LLMs?)

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

(a) vigor . . corroborated

(b) feebleness . . dashed

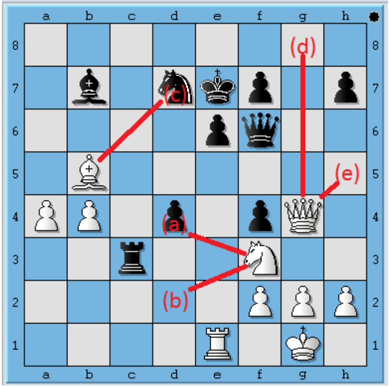
(c) proliferation . . blighted

(d) destruction . . disputed

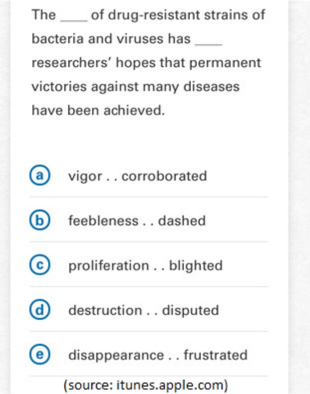
(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Chess and Tests—With Partial Credits (Or LLMs?)



The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

(a) vigor . . corroborated

(b) feebleness . . dashed

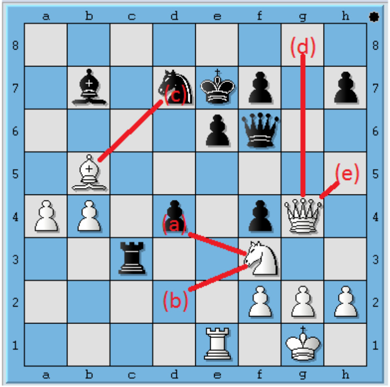
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

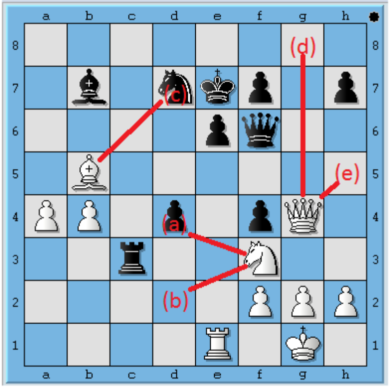
Chess and Tests—With Partial Credits (Or LLMs?)

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

- (a) vigor . . corroborated
- (b) feebleness . . dashed
- (c) proliferation . . blighted
- (d) destruction . . disputed
- (e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

How My Model Works

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.
- Use u_i's and s, c, h to generate probability p_i for each m_i.

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.
- Use u_i's and s, c, h to generate probability p_i for each m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.
- Use u_i's and s, c, h to generate probability p_i for each m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.
- Use u_i's and s, c, h to generate probability p_i for each m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.
- (Correct after-the-fact for chess decisions not being independent.)

Statistical application then follows by math known since Jacob Bernoulli and Carl Gauss over 200 years ago.

Validate the model on millions of randomized trials involving "Frankenstein Players" to ensure conformance to the standard bell curve at all rating levels.

How My Model Works

- Take parameters s, c, h from a player's rating (or "profile").
- Strong chess programs give *utility values* u_i for each legal move m_i.
- Use u_i's and s, c, h to generate probability p_i for each m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.
- (Correct after-the-fact for chess decisions not being independent.)

Statistical application then follows by math known since Jacob Bernoulli and Carl Gauss over 200 years ago.

Validate the model on millions of randomized trials involving "Frankenstein Players" to ensure conformance to the standard bell curve at all rating levels. This is also an example of **Simulation.**

Gaussian math yields confidence intervals that can enable **rejecting the null hypothesis** *of fair play* with high confidence.

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.
- A Likert scale converts strong/weak preferences to grainy numbers.

5. Preference Aggregation: Voting, Ranking, Ratings)

Thorny Question:

How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.
- A Likert scale converts strong/weak preferences to grainy numbers.
- Rating numbers convert into **rankings**. E.g. chess players by Elo.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random.

Polling

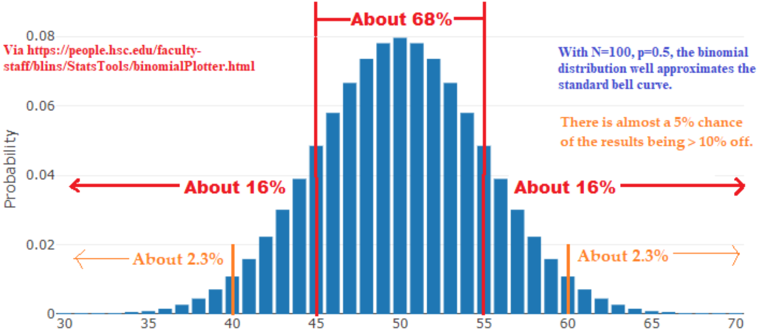
- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.

Polling

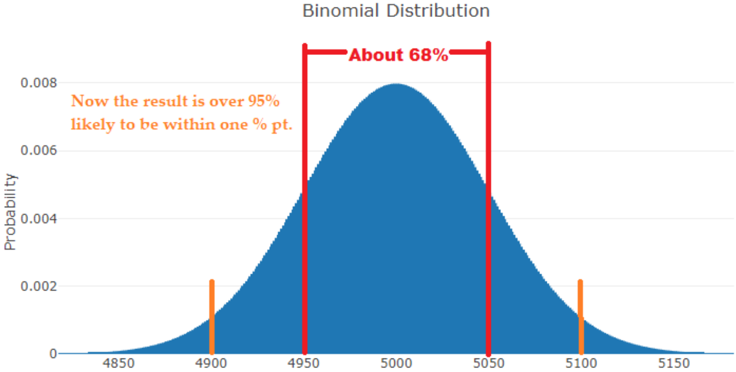
- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.
- **Mean** $\mu = 50$ and **std. deviation** $\sigma = \sqrt{Np(1-p)} = \sqrt{25} = 5$.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.
- **Mean** $\mu = 50$ and **std. deviation** $\sigma = \sqrt{Np(1-p)} = \sqrt{25} = 5$.
- Almost **one-third** chance poll results will be $<$ **45%** or $>$ **55%**.

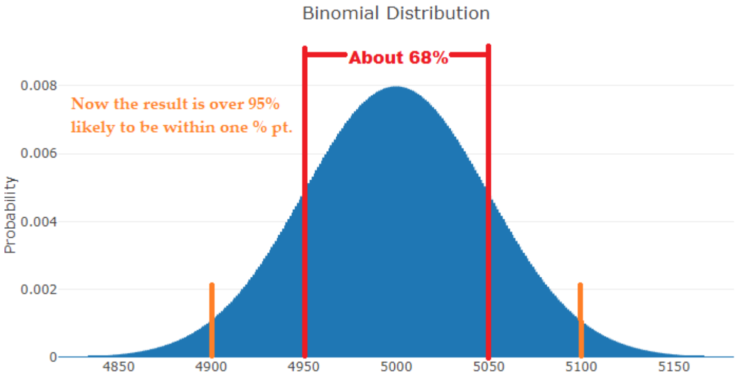


Try Polling 10,000 People Instead



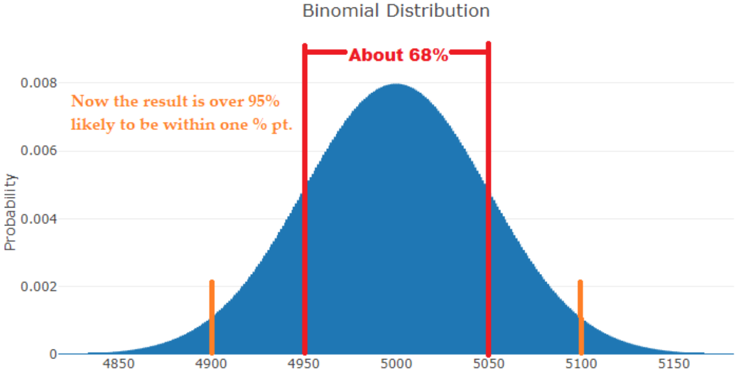
- Better—but polling 10,000 *es caro*!

Try Polling 10,000 People Instead



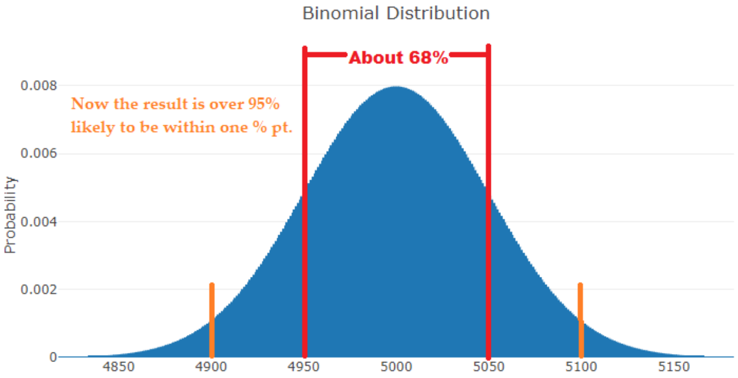
- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.

Try Polling 10,000 People Instead



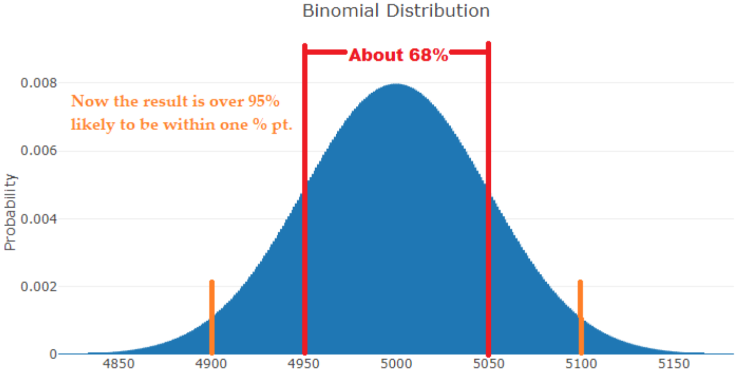
- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.

Try Polling 10,000 People Instead



- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.
- Gives $\pm 2\sigma/1024 = \pm\frac{1}{32} \approx \pm 3.1\%$ margin of "95%" confidence.

Try Polling 10,000 People Instead



- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.
- Gives $\pm 2\sigma/1024 = \pm\frac{1}{32} \approx \pm 3.1\%$ margin of "95%" confidence.
- So results 47%-to-53% count as "statistically tied" (yuck).

Precision, Accuracy, and a "Murphy's Law"

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.

Precision, Accuracy, and a “Murphy’s Law”

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.

Precision, Accuracy, and a “Murphy’s Law”

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using 100× more people brought only 10× more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using $100\times$ more people brought only $10\times$ more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate. Your $10\times$ narrower confidence intervals would give you misplaced confidence in a wrong result.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using 100× more people brought only 10× more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate. Your 10× narrower confidence intervals would give you misplaced confidence in a wrong result. **(Your HW will emphasize detecting possible sources of bias/inaccuracy and how to manage them.)**

Interpretation and Poll Aggregation

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a “95%” confidence interval of about 49% to 55%.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.
- Silver was non-renewed after ABC bought 538 and has his own site.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.
- Silver was non-renewed after ABC bought 538 and has his own site.
- I "poll' chess tournaments where proposition $A =$ player underperforms eir Elo rating, $B =$ the player overperforms it.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.
- Silver was non-renewed after ABC bought 538 and has his own site.
- I "poll' chess tournaments where proposition $A =$ player underperforms eir Elo rating, $B =$ the player overperforms it.
- Can both detect and rule out large-scale cheating.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.
- Silver was non-renewed after ABC bought 538 and has his own site.
- I "poll' chess tournaments where proposition A = player underperforms eir Elo rating, B = the player overperforms it.
- Can both detect and rule out large-scale cheating.
- Aggregating tournaments checks my formulas for accuracy and bias.

Optimization and Simulation and General Metrics

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.
- Simulations are especially useful for systems that **don't** have express models.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.
- Simulations are especially useful for systems that **don't** have express models.
- Search Engine Optimization (SEO) is a big area.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.
- Simulations are especially useful for systems that **don't** have express models.
- Search Engine Optimization (SEO) is a big area.
- **Can both simulate and solve the random walk on the Net Graph.**

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.
- Simulations are especially useful for systems that **don't** have express models.
- Search Engine Optimization (SEO) is a big area.
- **Can both simulate and solve the random walk on the Net Graph.**
- Any Q&A about that? (Mention Sentiment Analysis if time allows.)

Part VI: AI and Machine Learning

Alan Turing: Besides his WWII work on the Enigma machine (featured in the movie *The Imitation Game*) and **Turing Machine** theory of computation in his 1936-38 PhD thesis under Alonzo Church, he is considered the **founder** of Artificial Intelligence.

Part VI: AI and Machine Learning

Alan Turing: Besides his WWII work on the Enigma machine (featured in the movie *The Imitation Game*) and **Turing Machine** theory of computation in his 1936-38 PhD thesis under Alonzo Church, he is considered the **founder** of Artificial Intelligence.

The **Church-Turing Thesis** is primarily stated in terms of the class of *computable functions*, but here is Turing's angle:

> **Anything that human beings can consistently deduce or classify can also be achieved by computers acting alone.**

The **Turing Test** involves computers trying to be indistinguishable from humans in ordinary life communications and transactions.

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us.

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet
- Too breakable?

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet
- Too breakable? RECAPTCHA uses a different principle.

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet
- Too breakable? RECAPTCHA uses a different principle.

Also defies the logical **contrapositive** of Turing's Principle:

If it is really hard for computers then it should be hard for humans.

What we fear when worrying that AI will take away our jobs is:

Stuff that is hard for humans but easy for computers.

The logical **converse** of Turing's Principle acts as a brake, however:

If X is hard for humans—insofar as we can't consistently agree on answers—then X is hard for computers too.

Machine Learning (ML)

> The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called "learning."

Machine Learning (ML)

The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called ”learning.” **Some examples:**

Machine Learning (ML)

The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called "learning." **Some examples:**

- Building a **classifier**: able to distinguish pictures of cats, for instance.

Machine Learning (ML)

The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called ”learning.” **Some examples:**

- Building a **classifier**: able to distinguish pictures of cats, for instance.
- **Clustering analysis:** infer associations from nearness in “parameter space.”

Machine Learning (ML)

The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called ”learning.” **Some examples:**

- Building a **classifier**: able to distinguish pictures of cats, for instance.
- **Clustering analysis:** infer associations from nearness in “parameter space.”
- These and similar learning tasks can be automated using **neural nets**.

Machine Learning (ML)

The act of modifying a system or algorithm A via interactions with examples and other data so that A can emulate (and/or predict) the interactions without any more data.

Your Brain is Included—then this is just called "learning." **Some examples:**

- Building a **classifier**: able to distinguish pictures of cats, for instance.
- **Clustering analysis:** infer associations from nearness in "parameter space."
- These and similar learning tasks can be automated using **neural nets**.
- Using multiple **layers** of neural nets gives **deep learning**.

Some Hard Data Challenges (based on the converse principle)

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com
- Scene analysis in greater generality.

Some Hard Data Challenges (based on the converse principle)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com
- Scene analysis in greater generality.
- General anomaly alert systems.

Sentiment Analysis

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, “Vehicle defect discovery from social media.”

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, "Vehicle defect discovery from social media."
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, “Vehicle defect discovery from social media.”
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, "Vehicle defect discovery from social media."
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- (show Python 3 Trinkets web app and activity code.)

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, "Vehicle defect discovery from social media."
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- (show Python 3 Trinkets web app and activity code.)
- Simplistic idea: if the electorate is (un-)happy that's (bad) good news for the incumbent.

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, “Vehicle defect discovery from social media.”
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- (show Python 3 Trinkets web app and activity code.)
- Simplistic idea: if the electorate is (un-)happy that’s (bad) good news for the incumbent.
- “Joy” is an express term of the Harris-Walz campaign. Does it show?

(Chat)GPT, DALL-E, LaMDA, Etc.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications. Claimed by one engineer to evoke human-level *sentience* in conversations.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications. Claimed by one engineer to evoke human-level *sentience* in conversations.
- A big step up from 1960s "ELIZA." **New (11/28/22)**: ChatGPT.

(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications. Claimed by one engineer to evoke human-level *sentience* in conversations.
- A big step up from 1960s "ELIZA." **New (11/28/22)**: ChatGPT.
- Main paradigm of their operation is "find the next word" or "best next visual element."

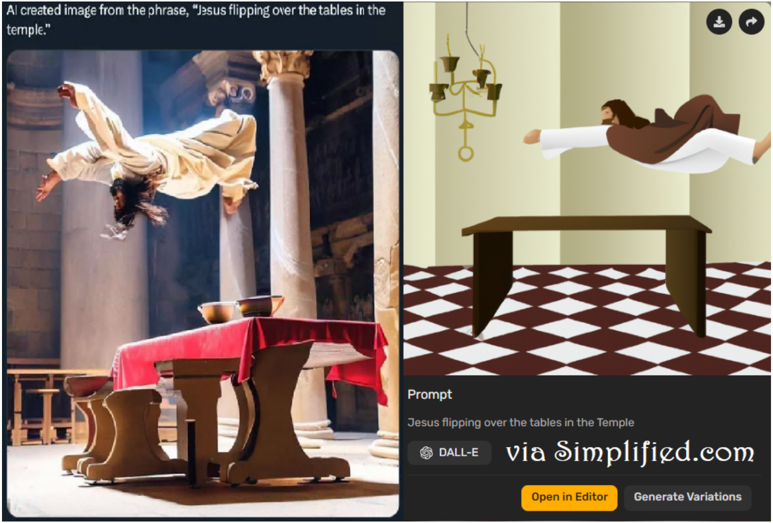
(Chat)GPT, DALL-E, LaMDA, Etc.

- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications. Claimed by one engineer to evoke human-level *sentience* in conversations.
- A big step up from 1960s "ELIZA." **New (11/28/22)**: ChatGPT.
- Main paradigm of their operation is "find the next word" or "best next visual element."
- But subject to **hallucinations** and other foibles—some shown by me here and here and here.

AI Art Adventure

"Jesus flipping over the tables in the Temple." From the movie *Jesus Christ Superstar*—then try it on Cutout or NightCafe or Simplified:

Two Results—one famous, one mine



ChatGPT Is Made of Us ("Pogo" Quote)

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]
- Does ChatGPT know the inner experience of writing poetry (in Latvian), or is it only shuffling symbols that imitate how poetry (in Latvian) has been written in the past?

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]
- Does ChatGPT know the inner experience of writing poetry (in Latvian), or is it only shuffling symbols that imitate how poetry (in Latvian) has been written in the past?
- This updates and focuses the "Chinese Room" Argument.

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]
- Does ChatGPT know the inner experience of writing poetry (in Latvian), or is it only shuffling symbols that imitate how poetry (in Latvian) has been written in the past?
- This updates and focuses the "Chinese Room" Argument.
- Given that ChatGPT has already processed the data and rules to write grammatical and cogent Latvian, a minimal threshold on the way to *sentience*, IMHO, is that a non-Latvian speaker like myself, giving examples of high-quality limericks in English and with no further Latvian data of any kind, should be able to get it to write superb limericks in Latvian.

ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]
- Does ChatGPT know the inner experience of writing poetry (in Latvian), or is it only shuffling symbols that imitate how poetry (in Latvian) has been written in the past?
- This updates and focuses the "Chinese Room" Argument.
- Given that ChatGPT has already processed the data and rules to write grammatical and cogent Latvian, a minimal threshold on the way to *sentience*, IMHO, is that a non-Latvian speaker like myself, giving examples of high-quality limericks in English and with no further Latvian data of any kind, should be able to get it to write superb limericks in Latvian.
- (But possibly I already pushed it to the limits of its current data.)

Another Example / AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:

Another Example / AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:

- By any chance was my blog horse picture "scraped" to contribute to this?

Another Example / AI Rights and Privacy Issues

“Cowboy closes barn door after the horse has left” via OpenAI API:

- By any chance was my blog horse picture “scraped” to contribute to this? ...without paying John Lund $35?

Another Example / AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:

- By any chance was my blog horse picture "scraped" to contribute to this? ...without paying John Lund $35?
- Goes even more for scraping copyrighted articles and books. Lawsuit.

Another Example / AI Rights and Privacy Issues

"Cowboy closes barn door after the horse has left" via OpenAI API:

- By any chance was my blog horse picture "scraped" to contribute to this? ...without paying John Lund $35?
- Goes even more for scraping copyrighted articles and books. Lawsuit. Worse stuff.

Will Extinction Be Academic?

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word."

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word."
- Two good articles by Stephen Wolfram on how ChatGPT works.

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word."
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."

Will Extinction Be Academic?

- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word."
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."
- Analogy to "find the best next move" in chess.
- Indeed, the architecture has affinity to AlphaZero.

Will Extinction Be Academic?

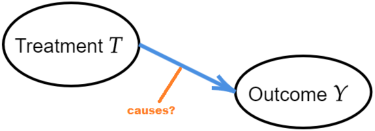
- https://rjlipton.wpcomstaging.com/2023/06/08/human-extinction/
- Note that Hava Siegelmann began by asking, can neural nets solve the Halting Problem? (covered in the "Abstraction" unit).
- Besides human demise, will AI cheating make academia extinct?
- Scott Aaronson article including his work on **watermark** cheating detection. Quote: "[ChatGPT's] only goal or intention in the world: to predict the next word."
- Two good articles by Stephen Wolfram on how ChatGPT works.
- Tag line at the top: "It's Just Adding One Word at a Time."
- Analogy to "find the best next move" in chess.
- Indeed, the architecture has affinity to AlphaZero.
- This *may* foster adapting my chess model for a "simple frequentist" kind of cheating detection.

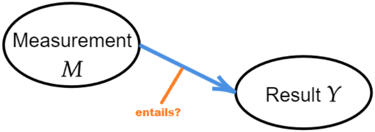
How AI Extends Search

[show Tonito.]

Part V: Societal Computing and Fairness

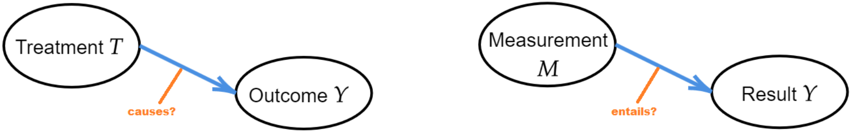
Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:





Part V: Societal Computing and Fairness

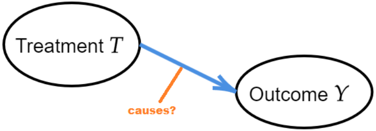
Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:

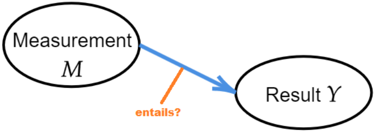


- Often Y is a binary choice: does a desired outcome happen? does the result go one way or the other way?

Part V: Societal Computing and Fairness

Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:

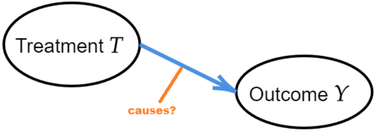


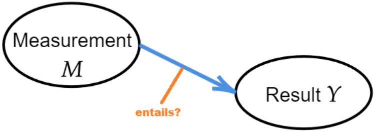


- Often Y is a binary choice: does a desired outcome happen? does the result go one way or the other way?
- The math for determining whether there is a significant causal relationship then resembles a simple poll.

Part V: Societal Computing and Fairness

Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:

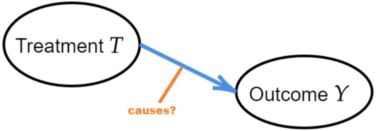


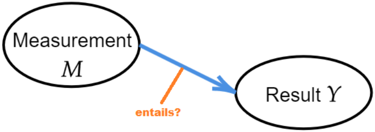


- Often Y is a binary choice: does a desired outcome happen? does the result go one way or the other way?
- The math for determining whether there is a significant causal relationship then resembles a simple poll.
- For a targeted value Y, the study's findings can be phrased as whether Y is significantly ahead.

Part V: Societal Computing and Fairness

Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:





- Often Y is a binary choice: does a desired outcome happen? does the result go one way or the other way?
- The math for determining whether there is a significant causal relationship then resembles a simple poll.
- For a targeted value Y, the study's findings can be phrased as whether Y is significantly ahead.
- I.e., is Y beyond the *margin of error* for the **null hypothesis** of no causation?

The Replication Crisis

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
    - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study...

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
    - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
    - More than 50 of the world's major institutions launch a study... privately.
    - The ground truth is "no effect"—analogous to our 50-50 election.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
    - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
    - More than 50 of the world's major institutions launch a study... privately.
    - The ground truth is "no effect"—analogous to our 50-50 election.
    - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
    - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
    - More than 50 of the world's major institutions launch a study... privately.
    - The ground truth is "no effect"—analogous to our 50-50 election.
    - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
    - This is the academic threshold to publish, so they do.
    - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.
- When others try to replicate the study, the ground truth proves out.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
    - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
    - More than 50 of the world's major institutions launch a study... privately.
    - The ground truth is "no effect"—analogous to our 50-50 election.
    - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
    - This is the academic threshold to publish, so they do.
    - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.
- When others try to replicate the study, the ground truth proves out.
- Can happen with 50 different big ideas, too (see this).

Study Size Matters

- Bookending this is that human-subject studies tend to be small.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.
  - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).

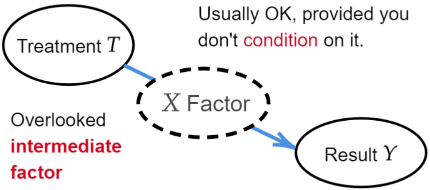
Study Size Matters

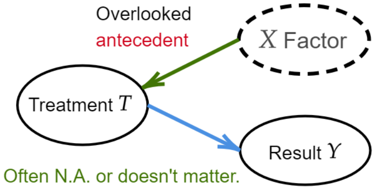
- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
    - Get higher N from less-targeted situations.
    - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).
- Mining social media is a major example.

Study Size Matters

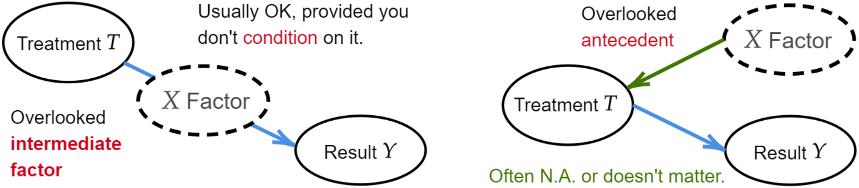
- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.
  - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).
- Mining social media is a major example.
- Can we make a tight enough relation between our measurements M and the results Y we are trying to capture?

Missing Factors in Studies—When Benign and...



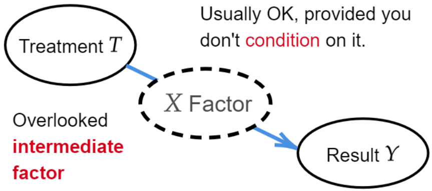


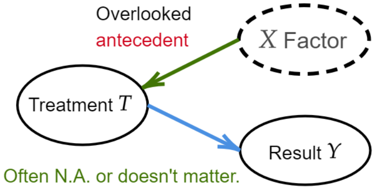
Missing Factors in Studies—When Benign and...



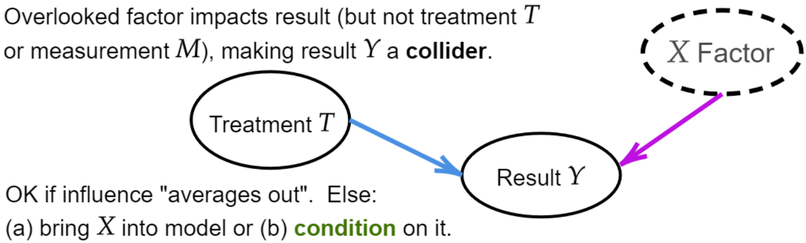
Top two are usually already taken care of. But **colliding** is a real issue:

Missing Factors in Studies—When Benign and...

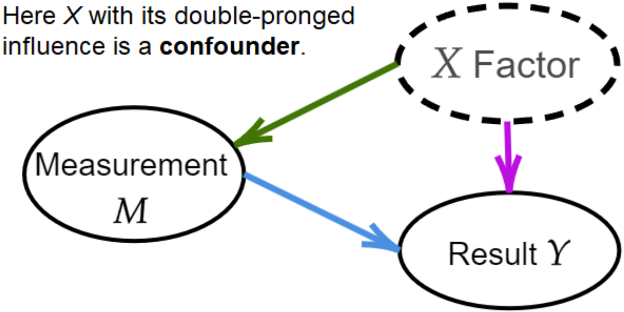




Top two are usually already taken care of. But **colliding** is a real issue:

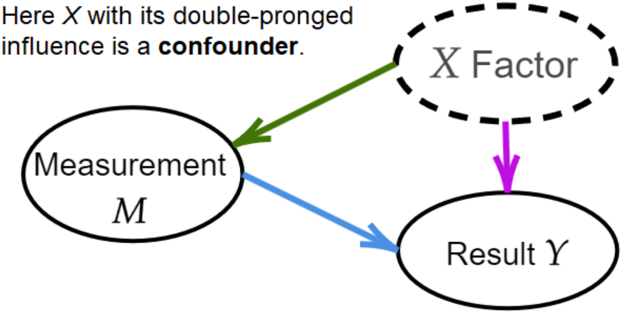


...When Not: I. Confounding Factors



The overlooked factor can **distort** the causal relationship of M to Y.

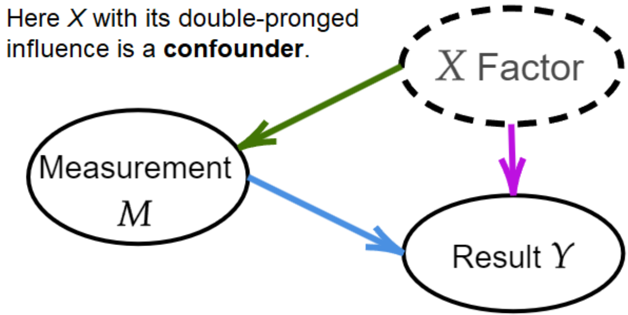
...When Not: I. Confounding Factors



The overlooked factor can **distort** the causal relationship of M to Y.

- Possible example: $X =$ a scandal, such as in North Carolina.

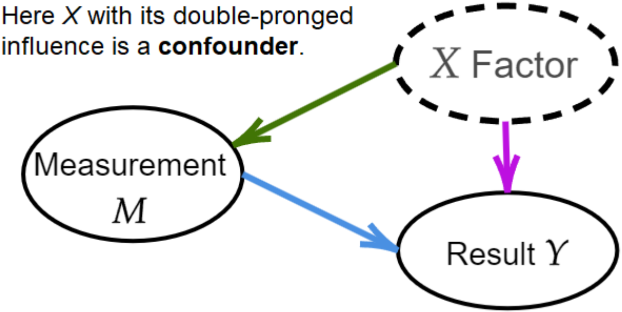
...When Not: I. Confounding Factors



The overlooked factor can **distort** the causal relationship of M to Y.

- Possible example: X = a scandal, such as in North Carolina.
- Can both *stimulate* M (such as "heat") while *inhibiting* Y (such as "Challenger Wins").

...When Not: I. Confounding Factors

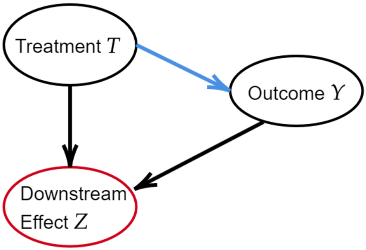


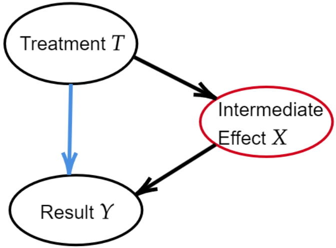
The overlooked factor can **distort** the causal relationship of M to Y.

- Possible example: X = a scandal, such as in North Carolina.
- Can both *stimulate* M (such as "heat") while *inhibiting* Y (such as "Challenger Wins").
- Even if impact is positive on both M and Y, X can dominate, drown out, or otherwise skew the effect we are trying to analyze.

II. Selection Bias From Conditioning

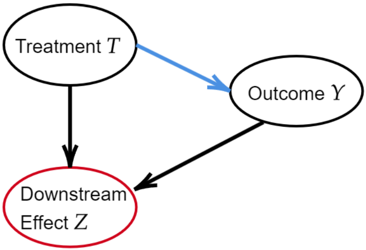
Conditioning on Other Effects

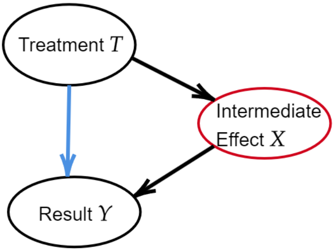




II. Selection Bias From Conditioning

Conditioning on Other Effects

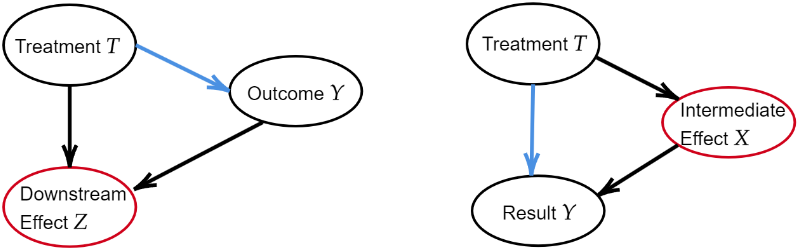




- Chess examples: (IIa) T = chess training, Y = more wins, Z = lower ACPL.

II. Selection Bias From Conditioning

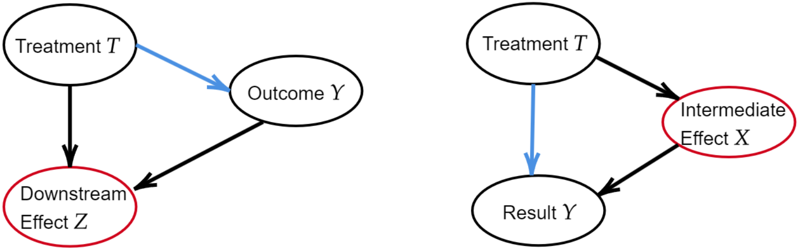
Conditioning on Other Effects



- Chess examples: (IIa) T = chess training, Y = more wins, Z = lower ACPL.
- (IIb): T = chess training, X = higher rating, Y = lower ACPL.

II. Selection Bias From Conditioning

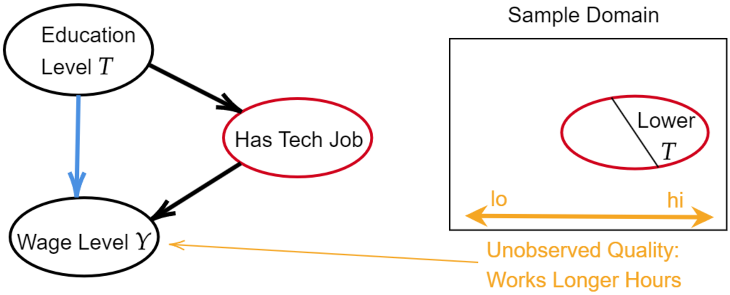
Conditioning on Other Effects



- Chess examples: (IIa) T = chess training, Y = more wins, Z = lower ACPL.
- (IIb): T = chess training, X = higher rating, Y = lower ACPL.
- Each way, conditioning on Z or X **selects** a **subsample** that may be skewed relative to the whole domain.

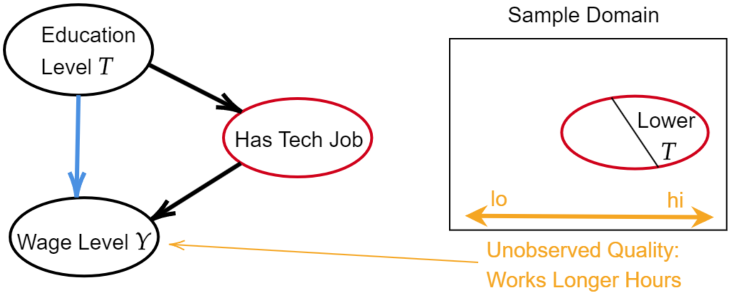
Non-Chess Example (adapted from here)

Suppose we are doing a large-scale study of the effect of education on wages, but decide to condition on people having tech jobs:



Non-Chess Example (adapted from here)

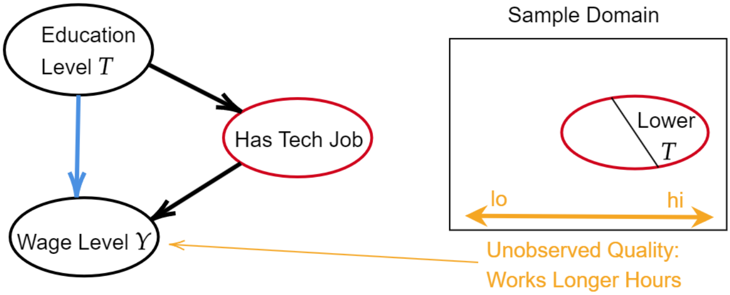
Suppose we are doing a large-scale study of the effect of education on wages, but decide to condition on people having tech jobs:



- Subsample from conditioned variable is skewed.

Non-Chess Example (adapted from here)

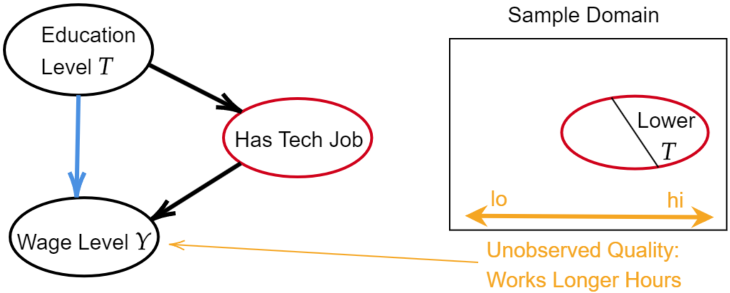
Suppose we are doing a large-scale study of the effect of education on wages, but decide to condition on people having tech jobs:



- Subsample from conditioned variable is skewed.
- (Source says "white-collar jobs" rather than "tech jobs.")

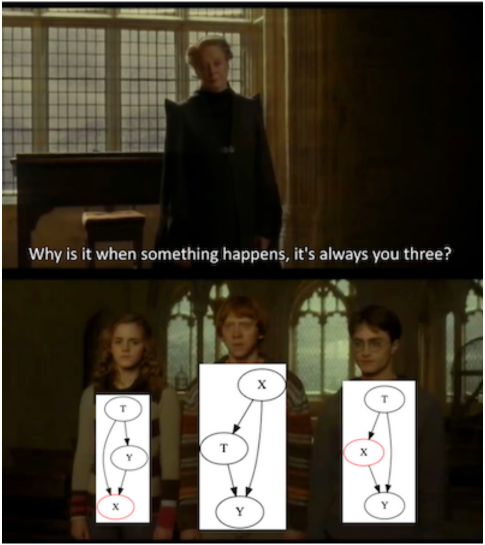
Non-Chess Example (adapted from here)

Suppose we are doing a large-scale study of the effect of education on wages, but decide to condition on people having tech jobs:



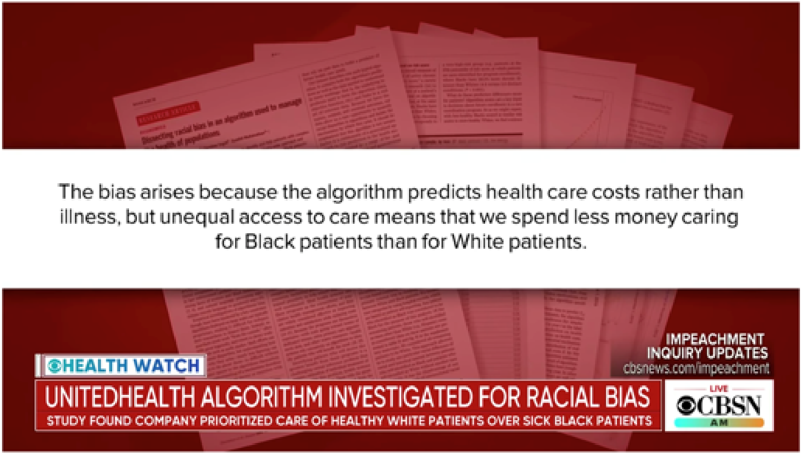
- Subsample from conditioned variable is skewed.
- (Source says “white-collar jobs” rather than “tech jobs.”)
- Can also happen from choices of unrepresentative proxy variables.

Harry Potter Meme (also from here)



Example of Bias From Proxy Variable (K. Joseph)

Here the variable Y' = health care costs used for Y = level of illness did implicit conditioning. Video.



Some Large-Scale Data Successes

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz. Thesis: Formal survey responses are inconsistent with opinions from the same populations mined on social media.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz. Thesis: Formal survey responses are inconsistent with opinions from the same populations mined on social media.
- Book *Weapons of Math Destruction*, by Cathy O'Neill.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz. Thesis: Formal survey responses are inconsistent with opinions from the same populations mined on social media.
- Book *Weapons of Math Destruction*, by Cathy O'Neill. Thesis: Mathematical models fossilize biases in data from remote history and skewed prior sources.

Some Large-Scale Data Successes

- Netflix and other recommender systems (also Spotify for music).
- Retailer recommendations (but also see this article).
- Product research via sentiment analysis.
- Google search and other web mining projects, e.g. Trends, N-grams.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz. Thesis: Formal survey responses are inconsistent with opinions from the same populations mined on social media.
- Book *Weapons of Math Destruction*, by Cathy O'Neill. Thesis: Mathematical models fossilize biases in data from remote history and skewed prior sources.
- Insofar as we are the training data for the Internet, the latter has baked in tangible amounts of racism and sexism.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.
- If time allows, show this GLL article about fairness in sampling and ranking.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.
- If time allows, show this GLL article about fairness in sampling and ranking.
- Main contentious issues come from AI Fairness.

Bias in Datasets

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)
- Another "Murphy's Law" situation: Minorities have **higher variance** in trained diagnoses simply because they comprise a minority in the datasets.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)
- Another "Murphy's Law" situation: Minorities have **higher variance** in trained diagnoses simply because they comprise a minority in the datasets.
- Fix only by having **ample** data from all groups.

Bias-Removal Dilemma

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*
- Doing so would make my model more predictive...

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit example of whether Google search should be affirmatively de-biased.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*
- Doing so would make my model more predictive...but less neutral.

Societal Impact of Tech, Again

Societal Impact of Tech, Again

- Stephen Brams, Game Theory and the Humanities and related papers and books.

Societal Impact of Tech, Again

- Stephen Brams, Game Theory and the Humanities and related papers and books.
- Liv Boeree, video: The Dark Side of Competition in AI.

Societal Impact of Tech, Again

- Stephen Brams, Game Theory and the Humanities and related papers and books.
- Liv Boeree, video: The Dark Side of Competition in AI.
- Liv Boeree, long video, Poker, Game Theory, AI, Simulation, Aliens, and Existential Risk.

Societal Impact of Tech, Again

- Stephen Brams, Game Theory and the Humanities and related papers and books.
- Liv Boeree, video: The Dark Side of Competition in AI.
- Liv Boeree, long video, Poker, Game Theory, AI, Simulation, Aliens, and Existential Risk.
- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.

Societal Impact of Tech, Again

- Stephen Brams, Game Theory and the Humanities and related papers and books.
- Liv Boeree, video: The Dark Side of Competition in AI.
- Liv Boeree, long video, Poker, Game Theory, AI, Simulation, Aliens, and Existential Risk.
- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.