Data and AI and Society

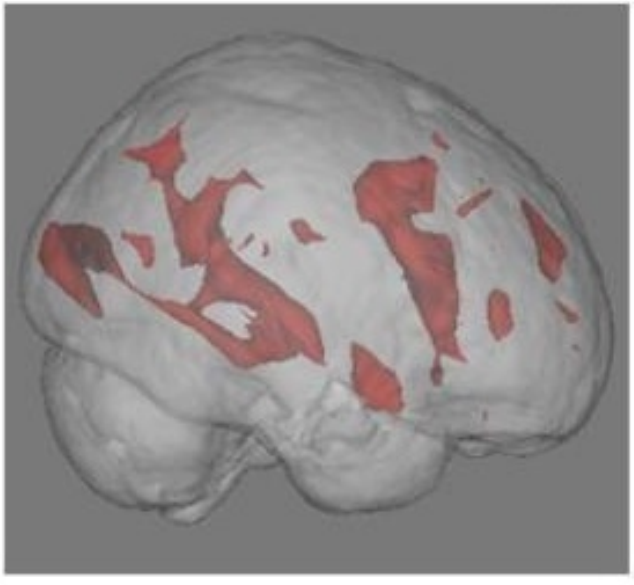
Resources and Dangers and Opportunities

Kenneth W. Regan

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

CSE199, Fall 2025

Main Problem...



THIS IS YOUR
BRAIN

THIS IS YOUR BRAIN
ON THE INTERNET

Any Questions?

(Brain scan source, 1987 PSA source)

...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

- **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

- 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*.
- Do we already invest them with **personhood**? Management too?
- Will they—and other forms of **AI** in general—soon supersede us?
- Now: Elon Musk's **Neuralink** brain implant as used to play chess.

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

- 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

- 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.**

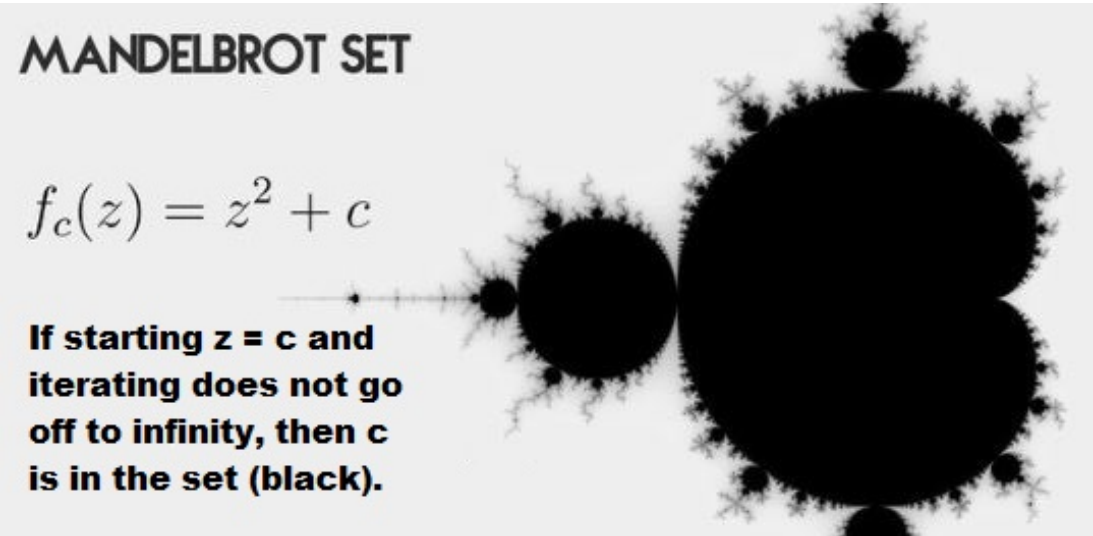
The Ocean of ~~Language~~ ~~Information~~ Data

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:

- 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:*

- 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

- 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?

- 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

- 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...
  - There is a virtual handout (not assigned HW this year, was so previously) to read "before or after" the next lecture.
- How can the Net's architecture absorb this expansion?

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

- 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.)