Data and Society

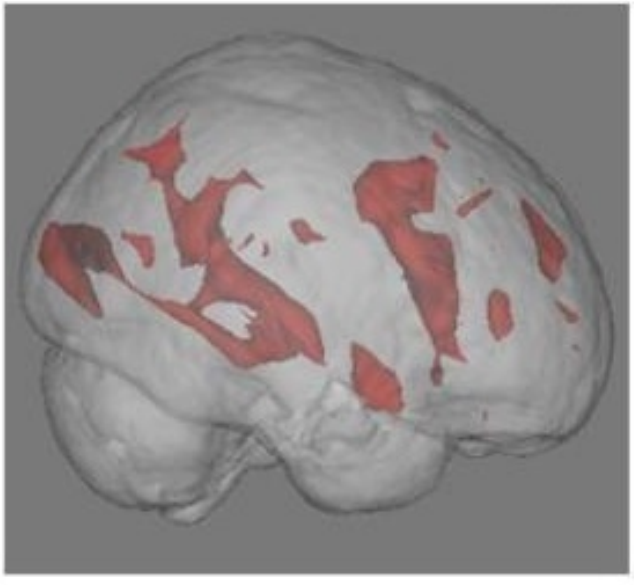
Resources and Dangers and Opportunities

Kenneth W. Regan

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

CSE199, Fall 2024

Main Problem...



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*.
- 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

- 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.**

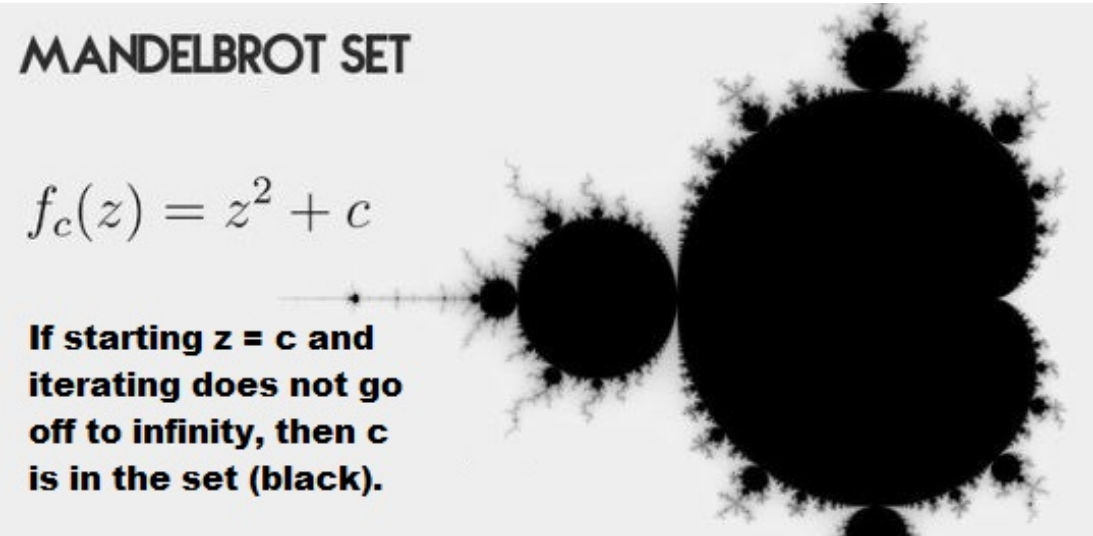
The Ocean of ~~Language~~ ~~Information~~ Data

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??
- How much data is being added per minute?
- This graphic shows how all the burgeoning data divides into categories.
- The Internet Archive Wayback Machine has indexed over **866 billion** webpages.

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 larger heading of data management services.
- Can be called an internetwork of data management services with common structures.
- Services are contracted to subscribers of all kinds: 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.

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: http://www.cnn.com/2017/09/11/us/social-media-irma-rescue-trnd/index.html
- Oct. 2022 Gulf of Mexico rescue via text message story.

- **Change for 2024: Begin with small-scale competition in two-player games—then scale ideas up to societal impacts.**
- **Doing so front-loads material for both this week's activity and next week's homework.**

Competition and Cooperation

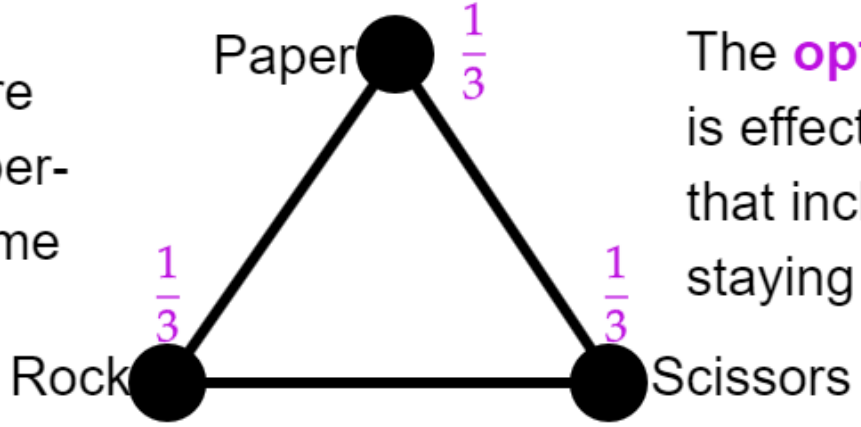
- Most familiar games are **zero-sum**, meaning that whenever and whatever one party wins, the others lose.
- Chess counts even though it has drawn outcomes. Players have **perfect information** about the current state of the game *in principle*, but even computers find it too *complex* to play perfectly.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

- 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



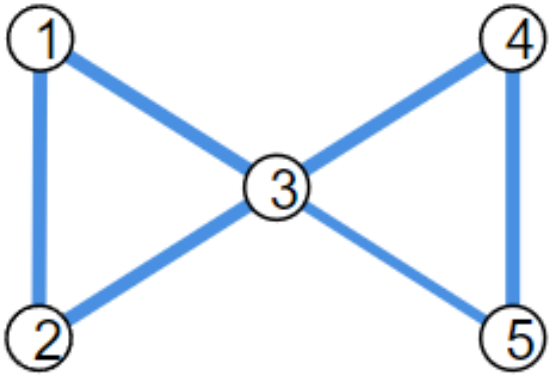
The **optimal random strategy** is effected by a **random walk** that includes the option of staying on your **current node**.

Another Single-Matrix Game

Imagine hunting a polar bear on ice floes in Arctic fog. When fog lifts:

- If hunter and bear are on adjacent floes, hunter shoots bear: $\rightarrow +\mathbf{1}$.
- If the bear is 2 or more floe-jumps away, the hunter misses: $\rightarrow \mathbf{0}$.
- If they find themselves on the same floe, $\rightarrow$ **?**.

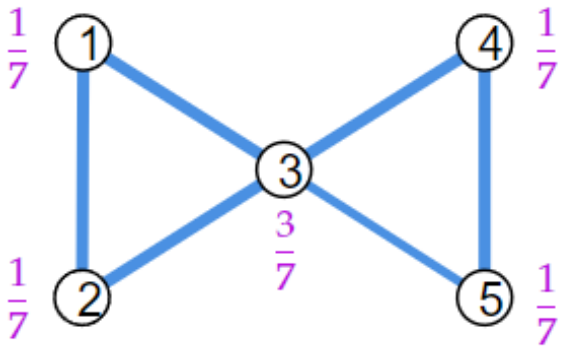
The network of adjacent floes can be represented as both a discrete graph and a matrix. Here is a picture of the game when five floes are arranged in a "bowtie" pattern:



You \ *Bear*	1	2	3	4	5
1	?	1	1	0	0
2	1	?	1	0	0
3	1	1	?	1	1
4	0	0	1	?	1
5	0	0	1	1	?

Bowtie Graph Game—Continued

If **?** $=$ **0** then the hunter achieves **expected value** $v = \frac{4}{7}$ by adopting the randomized strategy shown.



	1	2	3	4	5
1	**0**	1	1	0	0
2	1	**0**	1	0	0
3	1	1	**0**	1	1
4	0	0	1	**0**	1
5	0	0	1	1	**0**

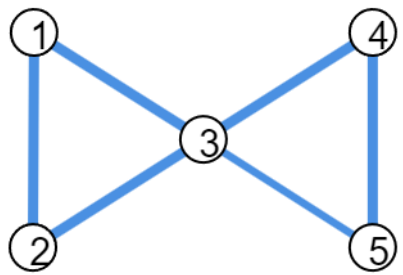
The "same" strategy by the bear assures losing no worse than $v = \frac{4}{7}$.

- Note that *both* choose the central floe (3) less than half the time.
- If **?** $=$ **+1** then the hunter **dominates** by always choosing (3).
- If **?** is negative the bear sometimes wins. What negative value makes the game *fair*—that is, both have expected value 0?
- Weird answer: $3 - \frac{16}{7-\sqrt{17}} = -2.56155...$
- If **?** $=$ **-1** then $v = \frac{1}{3}$ and both hunter and bear play (3) one-third of the time—same frequency as in a **random walk** of the graph.

Two-Matrix Games: Not Zero-Sum

Change the same-floe case to be: bear knocks the gun away but raids the hunter's lunch for **+3** value rather than kill em. Meanwhile the hunter videos the bear, for **+0.5** value. And in the two-floes-away case, let's penalize both of them **-0.5**, for missing and being inadvisably close. Now we need a separate **payoff matrix** for each:

H	1	2	3	4	5
1	0.5	1	1	−0.5	−0.5
2	1	0.5	1	−0.5	−0.5
3	1	1	0.5	1	1
4	−0.5	−0.5	1	0.5	1
5	−0.5	−0.5	1	1	0.5



B	1	2	3	4	5
1	3	−1	−1	−0.5	−0.5
2	−1	3	−1	−0.5	−0.5
3	−1	−1	3	−1	−1
4	−0.5	−0.5	−1	3	−1
5	−0.5	−0.5	−1	−1	3

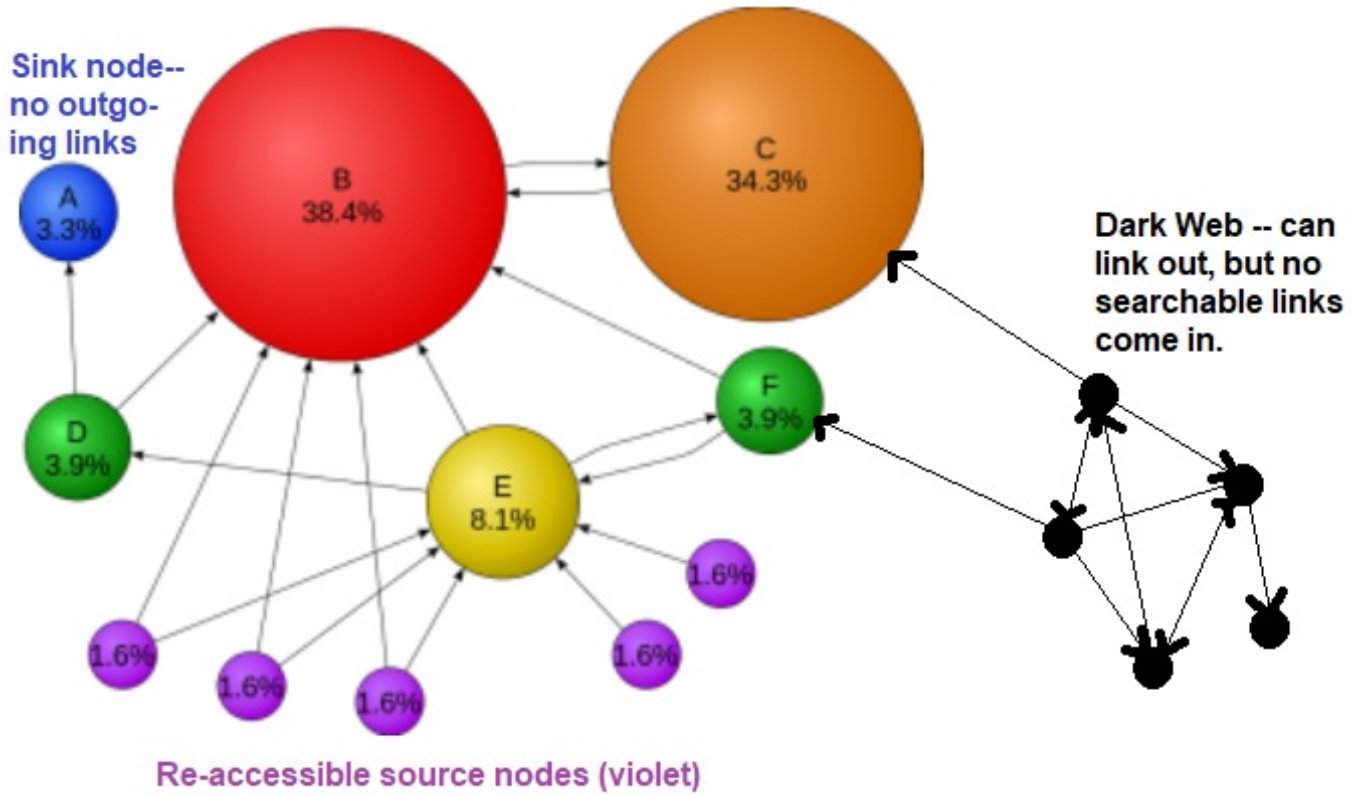
- If H and B agree on same floe, *both win*. No longer zero-sum!
- But H could gain by switching to an adjacent floe and shooting...
- Analysis becomes complicated! Semi-solved by **John Nash** in 1950.
- You will play a simpler(?) example game in recitations.

Multi-Player and Solitaire Games

- Games with $N > 2$ players are more complex, but many features of 2-player games apply.
- Nash Equilibrium: N strategies such that no player can improve by emself.
- Nash proved such an equilibrium always exists, but *finding* one is not known to be in **P**—and whether more than one equilibrium exists is **NP-complete**.
- A Nash equilibrium need not be optimal for all players. (Call it a "GNash equilibrium.")
- If all others' strategies are fixed (whether optimal or equilibrium or not, and whether you know the strategies or not), then the game becomes **solitaire** for you. Like playing the 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

- Webpages form a big graph with pages as nodes and links as edges.
- Graph is **directed** (one-way arrows). but many links are two-way.
- 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.



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. (Alternative: in order of $w'_Q(P) \cdot r(P, Q)$.)

Societal Boons and Banes

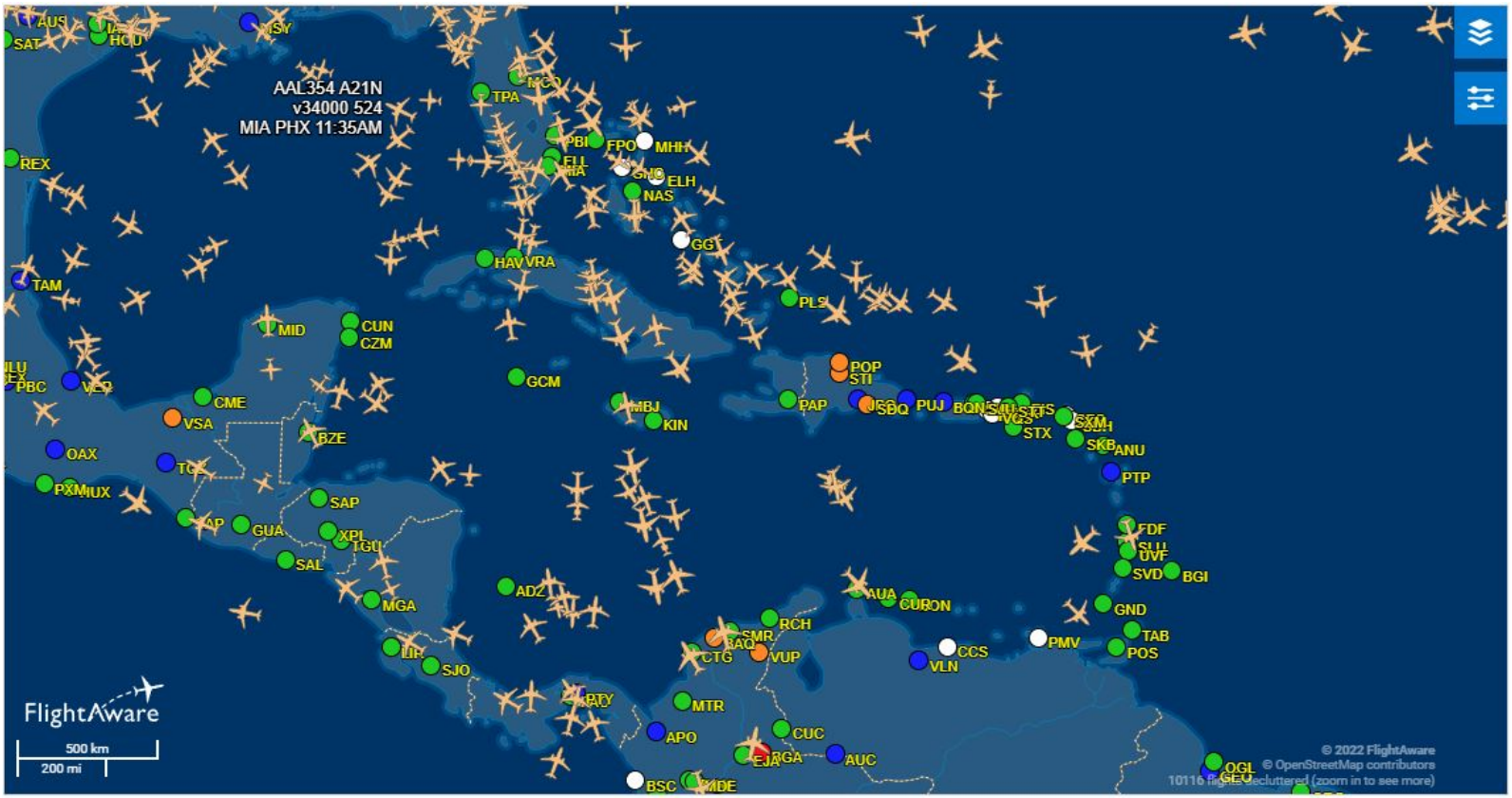
- The correspondence between high $w(P)$ and real usefulness of a page P was so great that Google slayed all its peer search engines.
- The graph principle still largely rules now. It is *organic*. **But:**

> 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-Structred" Example (of Inferencing)

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

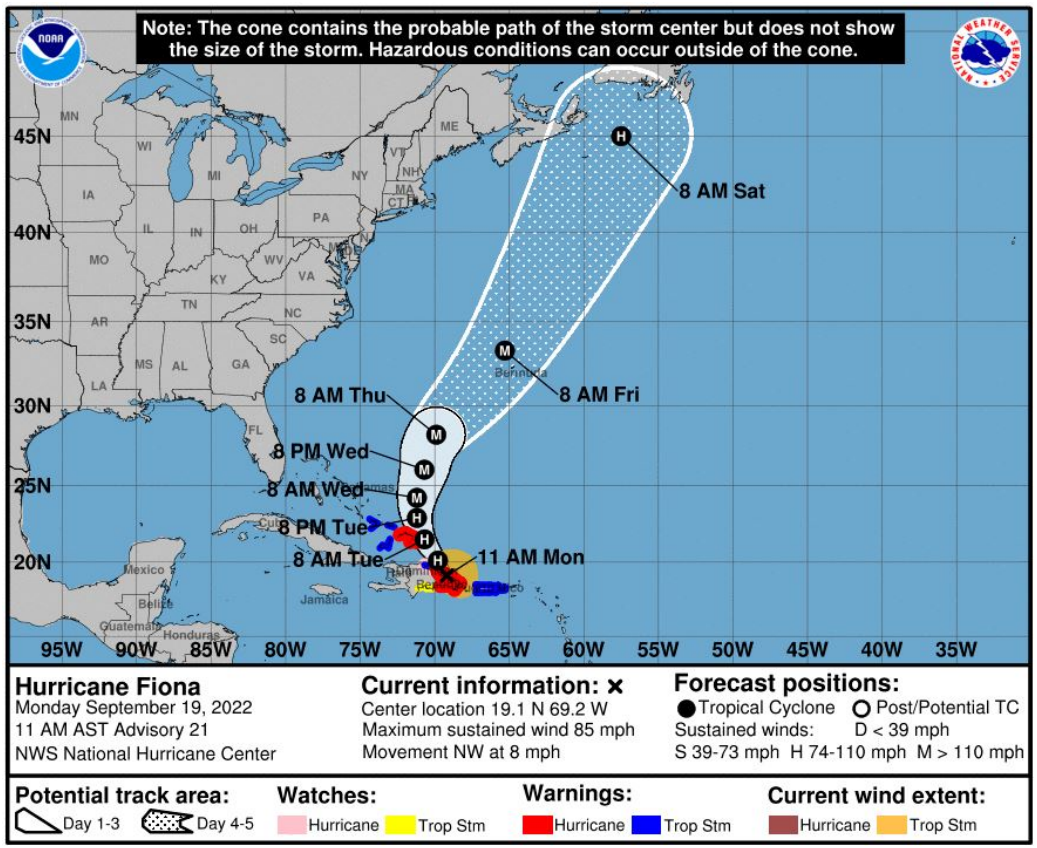


Why almost no planes over Puerto Rico and the Dominican Republic + Haiti? Compared to right now...

And what about north of the Black Sea?

Hurricane Tracking

NOAA (picture of Hurricane Fiona in 2022)



Note the error bars around the forecasted track. Trace of Hurricane Lee (But, Otis on Oct. 24, 2023 was a big forecasting failure.)
Hurricane Francine today.

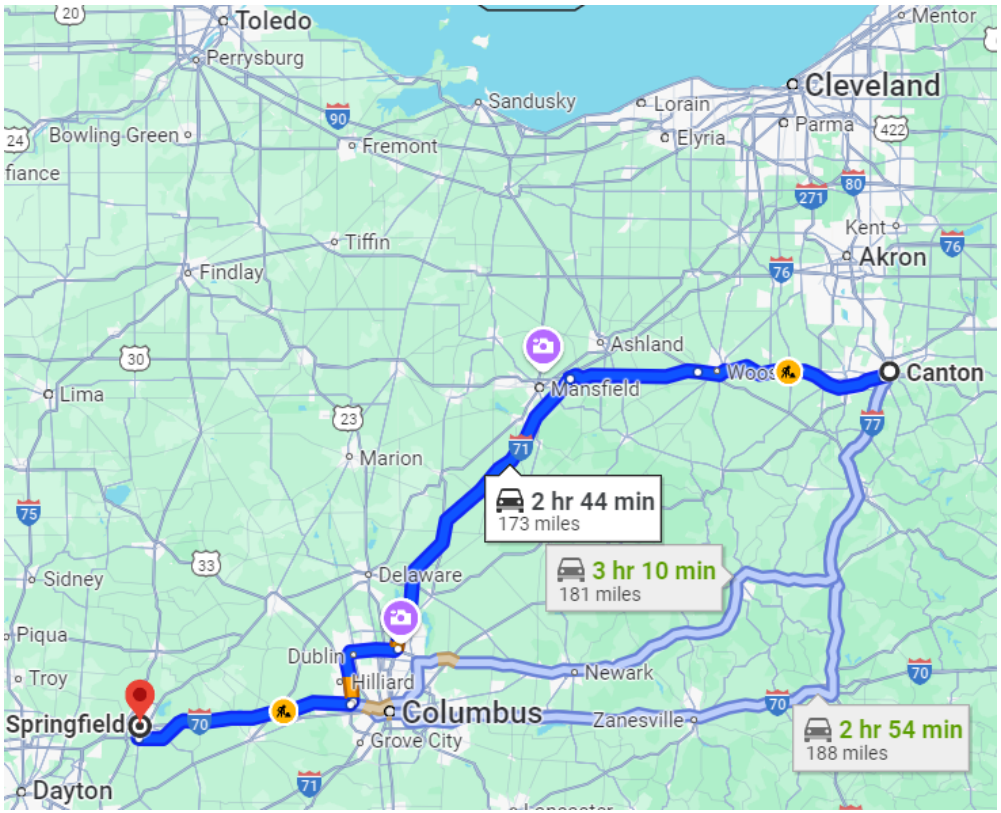
Mapping and Geolocation

- *Google Maps* and similar services—lack of them was cited often in the "8 Hours Without Internet" 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, less often (not since spring) but more detailed.
- Ukraine Weapons Tracker and others do visual confirmation of equipment losses.
- 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—New Example

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

- **Trump Assassination Attempt**. IMO bullet visible in photo, and people died, so end-of-story: not staged.
- **Other Fact-Checking Cases**: Can be hard to trace the record...
- ...and even harder to agree on language—AI examples next week.
- 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 others that were considered:
  - 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

- 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

- 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

In our previous `<recipe>` example this would include:

- 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

- 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

- First(?) Major Data Breach to Public: 2006 AOL "*Valdez*" (user search data, ID-ed by number but persons exposed).
- Too Many Examples today, clear thru to Equifax...
- Systems trying to cope by altering *verification* of data and *nature* of data:
  - 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?
- 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.