Internet and Data

Resources and Risks and Power

Kenneth W. Regan

CSE199, Fall 2023

Outline Week 1 of 2: Data and the Internet

- What is *data* exactly? How much is there? How is it growing?
- Where data resides—real and virtual. The Cloud. The Farm.
- How data may be accessed. Importance of **structure** and **markup**.
- Structures that help algorithms "crunch" data.
- Formats and protocols for *enabling* access to data.
- Protocols for *controlling* access and changes to data.
- **SQL**: `Select.` `Insert.` `Update.` `Delete.` `Create.` **Drop.**
- Dangers to privacy.
- Dangers of crime.
- (Dis-)Advantages of online data.
- [Week 1 Activity: Trying some SQL queries.]

What Exactly Is "Data"?

Several different aspects and definitions:

1. The entire track record of (your) online activity.
   - Note that any "real data" put online was part of online usage. Exception could be burning CD/DVDs and other hard media onto a server, but nowadays dwarfed by uploads. So this is the most inclusive and expansive definition.
   - Certainly what your carrier means by "data"—if you re-upload a file, it counts twice.
2. Structured information for a particular context or purpose.
   - What most people mean by "data."
   - Data repositories often specify the context and form.
   - Structure embodied in *formats* and *access protocols*.
3. In-between is what's commonly called "Unstructured Information"
   - Puts the *M* in *Data Mining*.
   - Hottest focus of consent, rights, and privacy issues. 2023 Example

How Much Data Is There?

- That is, How Big Is the Internet?
  - Searchable Web
  - Dark Web
  - (I maintain several gigabytes of deep-web textual data. . . tracking chess tournaments for possible cheating. Only tournament staff know the link—for their event only.)
- 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??
- Whichever, the Internet is in the "Zettabyte Epoch."

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? (Other lectures)
- *Access* to data: *who* and *how*, is key.

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.

Access to Data

- Some data you own—and you (or your group) have sole access to it.
- Other data you own but wish to share outside your group, even publicly.
- Access to **read**. . .
- Access to **modify**. . .
- Not just permission, but ease of interpreting data is paramount.
- Owner and/or provider are responsible for *structuring* data.

Prime Directive: Eliminate—or at least minimize—the one-off work a client needs to do to interface with your data.

Some Structural and Algorithmic Constraints

A Basic Dilemma—which will echo early on in your courses:

1. Expect certain data points at preappointed positions, or
2. Search open-endedly for **tags** identifying the data points.

- Array lookup `arr[i]` is an example of the first.
- Associative lookup `$table{key}` exemplifies the second.
- But if the tag could be "anywhere" in a mound of data, much waste of time.
- The Cloud cannot serve *Random Access* on a large scale.
- The "Three Rules" of Real Estate (on the Net):
  - **Locality.**
  - **Locality.**
  - **Locality.**
- Whole Net system architectures (MapReduce/Hadoop/Google File System, Amazon Elastic Compute Cloud...) are designed to ensure that data is *Stream-Friendly*.

Data File Formats

- Positional formats typified by CSV, BMP
- Whereas TIFF tags images, XLSX adds markup to XLS...

Markup

- Long predates the Internet.
- Publisher markup for editing and typesetting (and interpretation).
- Jerome Saltzer, 1964: RUNOFF, which led to ROFF.
- Later: TeX, LaTeX... (As opposed to WYSIWYG)
- William Tunnicliff, 1967: "Generic Coding."
- Charles Goldfarb, 1969: organize legal documents.
- Led to IBM's Generalized Markup Language (**GML**), 1973.
- Standard Generalized Markup Language (**SGML**), ISO 1986.
- Extensible Markup Language (**XML**) started as a simpler SGML.
- Hypertext Markup Language (**HTML**) imitated SGML.
  - Introduced by Tim Berners-Lee in a 1991 forum post which linked to a document titled "HTML Tags."
- JavaScript Object Notation (**JSON**), Douglas Crockford, 2001.
- Now main alternative to XML, especially for *object serialization*.

Markup Example: SGML (source, alt)

```
<recipe type="dessert" servings="6" preptime="10">  <!--Ten what?-->
<title>Haupia (Coconut Pudding)</title>
<ingredient-list>
<ingredient>
12 ounces coconut milk
</ingredient>      <!--Parser could allow omitting item close tag-->
<ingredient>
4 to 6 tablespoons sugar
...
</ingredient-list>
<instruction-list>
<step necessary="no">
Thoroughly wash and dry the pot you will use.
</step>
...
</instruction-list>
</recipe>
```

Example: The First HTML Doc (lightly altered)

```
<TITLE>Tags used in HTML</TITLE>
<NEXTID 22>
<H1>HTML Tags</H1>This is a list of tags used in the
<A NAME=0 HREF=MarkUp.html#4>HTML</A> language.
Each tag starts with a tag opener (a less than sign) and ends
with a tag closer (a greater than sign).
Many tags have corresponding closing tags which
identical except for a slash after the tag opener.
(For example, the <A NAME=3 HREF=#2>TITLE</A> tag).<P>
Some tags take parameters, called attributes.
...
Opening list tags are:
<DL>
...
</DL>
the closing tag must obviously match the opening tag.
```

Did not yet have `HEAD` and `BODY` structure. (Yes, word “are” is missing)

Example: XML and JSON Compared

From https://www.w3schools.com/js/js_json_xml.asp, XML first:

```
<employees>
    <employee>
        <firstName>John</firstName> <lastName>Doe</lastName>
    </employee>
    <employee>
        <firstName>Anna</firstName> <lastName>Smith</lastName>
    </employee>
    <employee>
        <firstName>Peter</firstName> <lastName>Jones</lastName>
    </employee>
</employees>
```

```
{"employees":[
    { "firstName":"John", "lastName":"Doe" },
    { "firstName":"Anna", "lastName":"Smith" },
    { "firstName":"Peter", "lastName":"Jones" }
]}
```

My Own Format Extending Chess “PGN” Standard

```
[GID "De Castellvi;Vinoles;Valencia;Valencia ESP;1475.??.??;?;1-0"]
[EID "Komodo-8-32bit"]
[Turn "6-w"]
[MovePlayed "h3"]
[EngineMove "Ne5"]
[Eval "+160"]
[Depth "12"]
...
       1    2    3    4    5    6    7    8    9   10   11   12
-----------------------------------------------------------------
Ne5  n.a. n.a. n.a. n.a. n.a. +142 +142 +140 +132 +147 +146 +160
d3   +110 NREC NREC NREC +053 +095 NREC NREC NREC NREC NREC NREC
Bxf7 n.a. n.a. n.a. n.a. +107 +079 NREC NREC NREC NREC NREC NREC
...
```

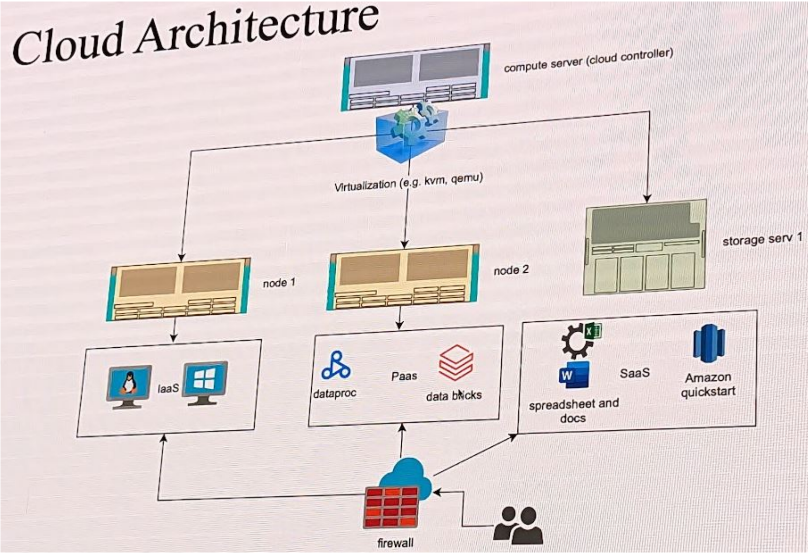
Mixes position-based and tagged elements. One [...] encloses tag and value.

Non-hierarchical structure.

Week 1 Homework Q&A and Activity Instructions

- No “bad porn” issues two weeks ago. Was my demo clear?
- **Activity Curveball 1:** Some Macbooks don’t show my `NFLTeams.xml` in Safari. **Try Chrome** (you may need to download and install it). Or Firefox.
- **Activity Curveball 2:** When trying to save the file from the browser, the system “helpfully” suggests a **mangled** filename that includes the URL. **You must save it with the exact name NFLTeams.xml** (case sensitive)...
- ...That’s because **Activity Curveball 3:** SQLizer names the SQL table after the filename rather than the outer XML tag name. :-(
- So in brief: (I) save from my site to your machine as `NFLTeams.xml`
- (II) Upload that to SQLizer and click Convert.
- (III) Select-all and mouse-copy the result into the left pane of DB-Fiddle. Click the “Run” triangle to verify.

Three Extra Slides on the Cloud, by Asif Imran



Cloud Architecture

compute server (cloud controller)

Virtualization (e.g. kvm, qemu)

node 1

node 2

storage serv 1

IaaS

dataproc

Paas

data bricks

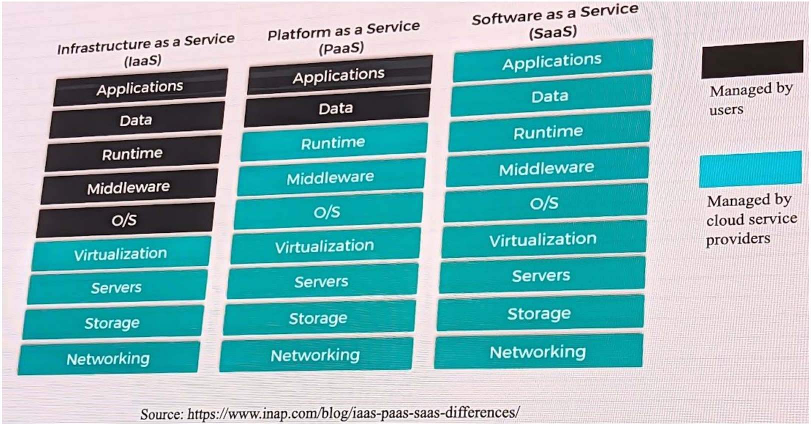
spreadsheet and docs

SaaS

Amazon quickstart

firewall

Levels of Using the Cloud



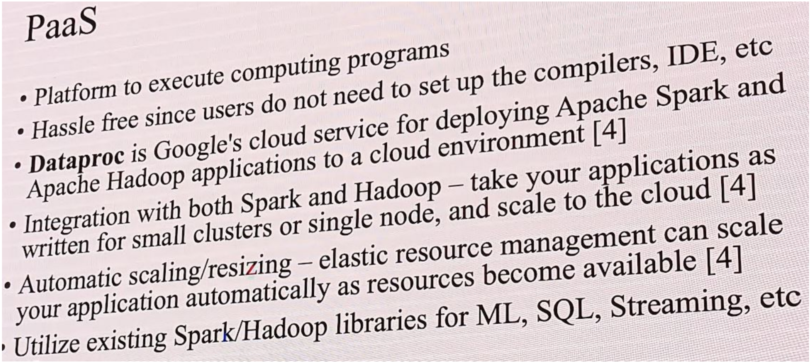
Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
⬛ Applications	⬛ Applications	🟦 Applications
⬛ Data	⬛ Data	🟦 Data
⬛ Runtime	🟦 Runtime	🟦 Runtime
⬛ Middleware	🟦 Middleware	🟦 Middleware
⬛ O/S	🟦 O/S	🟦 O/S
🟦 Virtualization	🟦 Virtualization	🟦 Virtualization
🟦 Servers	🟦 Servers	🟦 Servers
🟦 Storage	🟦 Storage	🟦 Storage
🟦 Networking	🟦 Networking	🟦 Networking

⬛ Managed by users

🟦 Managed by cloud service providers

Source: https://www.inap.com/blog/iaas-paas-saas-differences/

Cloud Platforms



PaaS

- Platform to execute computing programs
- Hassle free since users do not need to set up the compilers, IDE, etc
- **Dataproc** is Google's cloud service for deploying Apache Spark and Apache Hadoop applications to a cloud environment [4]
- Integration with both Spark and Hadoop – take your applications as written for small clusters or single node, and scale to the cloud [4]
- Automatic scaling/resizing – elastic resource management can scale your application automatically as resources become available [4]
- Utilize existing Spark/Hadoop libraries for ML, SQL, Streaming, etc

What Does ‘X’tensible Mean? (∼ Part II begin)

- Can tailor (to) data structures and interfaces.
- Can *define* a common user/program interface.
- E.g., common webpage display and protocol for new users of a web-deployed system.
- Close analogy to **Cascading Style Sheets** (CSS).
- Indeed, CSS interfaces with XML as a display and UI front end.
- **Document Type Definition** (DTD) specifies new SGML/XML elements and their syntax rules, which can allow “nesting.”
- End-user apps need to be tailored to the DTD but this can be automated, e.g. by an *XML parser generator*.
- **Meta**-level: can extend the language to produce whole hierarchies of DTDs and meta-rules for specifying them.
- (PGN and my AIF have no formal DTD, are minimally extensible.)

Three Functions With Data—All Handled By SQL

1. Data Definition/Creation
2. Data Manipulation (read-only access included in this heading)
3. Data Control.

The **Structured Query Language** (SQL) handles all three.

- Donald Chamberlain, Raymond Boyce, IBM, early 1970s.
- Originally **S**tructured **E**nglish **QUE**ry **L**anguage, but "SEQUEL" trademark was taken. Still often pronounced that way.
- **Oracle Corp.** both extended and "front-ended" SQL.

Largely embodies Edgar F. Codd's **Relational Model** (RM).

Relational not positional. *Declarative* in that users are responsible only for data and queries, not algorithms or code. RM governs how database is built. Queries are built from logic and numerical predicates.

Some SQL Commands

CREATE. Note that it creates a structure before you input data.

```
CREATE TABLE Games (
    gid             VARCHAR(128)       PRIMARY KEY,
    white_name      VARCHAR(50)        not null,
    black_name      VARCHAR(50)        not null,
    result          VARCHAR(7)         not null,
    white_rating    INTEGER
    black_rating    INTEGER
);
```

Here TABLE is a built-in SQL type, or rather template for the user-defined type Games. To kill it and all data you give both names:

```
DROP TABLE Games;
```

TRUNCATE TABLE Games; would destroy the entries but not the definition.

Inserting, Updating, and Removing Data

```
INSERT INTO Games (white_name, black_name, result)
   VALUES ('DeCastellvi', 'Vinoles', '1-0');
```

```
UPDATE Games SET gid = generate_game_id();
```

SQL allows user-defined functions, here to generate the game ID.

Since players didn't have ratings back in 1475, those fields can be left with a default `null` value. We could define a default of 0 but shouldn't—it would throw off `AVG` calculations. The `gid` field had a default which must be immediately changed, else the next insert will violate the `PRIMARY KEY` uniqueness constraint.

`DELETE FROM Games WHERE gid =` followed by the unique key removes just that game.

Can build by generating commands from data in XML/JSON/etc...

Selection and Logic in SQL

Suppose I want just the games where the lower-rated player won. A user-defined predicate `underdog_wins()` could have body:

```
(white_rating < black_rating AND result = ’1-0’)
   OR (white_rating > black_rating AND result = ’0-1’)
```

As with a *method* in OOP, the table object is implicit. Then

```
SELECT * FROM Games WHERE underdog_wins() = 1;
```

temporarily makes a table from just those games where the underdog won. In place of we could have listed just some fields to return.

User-defined functions can return whole tables. Tables can be `JOIN`ed together (in various ways) on common field(s).

(Yes, basic SQL needs that ‘`= 1`’)

Converting Data to SQL Entry (simplified)

```
<NFLTeams>
<Team code="ARI" teamName="Cardinals" region="Arizona"
      pop="4438000" lastPlayoffWin="2015"/>
<Team code="ATL" teamName="Falcons" region="Atlanta"
      pop="6462000" lastPlayoffWin="2016"/>
...
</NFLTeams>

CREATE TABLE NFLTeams (
   _code VARCHAR(3),
   _teamName VARCHAR(50),
   _region VARCHAR(50),
   _pop INT,
   _lastPlayoffWin INT
);
```

```
INSERT INTO NFLTeams VALUES (’ARI’, ’Cardinals’, ’Arizona’, 4438000, 2015);
INSERT INTO NFLTeams VALUES (’ATL’, ’Falcons’, ’Atlanta’, 6462000, 2016);
...
```

SQL Permissions

- The SQL standard finally includes a whole Data Control Language (DCL).
- Maintains a list of user IDs.
- Mostly done by two commands, `GRANT` and `REVOKE`.
- Rather than read-write-execute (`rwx`) permissions, it grants or withdraws allowed SQL commands. E.g.:
- `GRANT UPDATE ON Games TO garry_kasparov;`
- `REVOKE EXECUTE ON Games FROM PUBLIC;`
- Permissions can also be system-wide.
- Permissions can be grouped into *role* specifiers.
- Can build a management system on top of the SQL DCL.
- Permissions can be granted to not just people!
- Your "Al-Go-Rith-Ms" carry lots of SQL commands to submit...
- When "everything is data," those commands are data...and data is commands...

So Is This Data Heaven?

- **Structure**, **Extensibility**, and sheer computing power have built a brave new world.
- "Power Corrupts" is a **theorem** in CS.
- Microsoft Technet article on SQL serving:

> "Security is an exercise in creating enough barriers to the system such that the effort involved to attack a system exceeds the benefit derived from the data."

- It does *not* say, "Security is an exercise in making systems secure."
- Speedy execution cuts corners on safety.
- SQL by itself has several vulnerabilities.
- **Injection:** Trick a system into executing SQL privilege commands embedded in data.
- Show XKCD comic https://xkcd.com/327/

Other Potential Weaknesses

- Although SQL polices its own user-defined functions, it allows functions written in other languages.
- These can possibly import "unsafe code."
- Might exploit details/weaknesses in how the SQL system was implemented.
- Even within SQL, what happens if you give 100 chars to a VARCHAR(50)?
- Implementations "should" either (a) refuse or (b) truncate your string, but (a) can block a whole upload and (b) may cause constraint violations.
- For speed and simplicity too, systems might (c) take your whole string and overflow into another memory region.
- Such "buffer overflows" have bit from the 1988 Internet Worm to 2017's Cloudflare bug.
- I wrote a joint article on the latter.
- More about security in other weeks of this course...

Other Ways to Game a Database

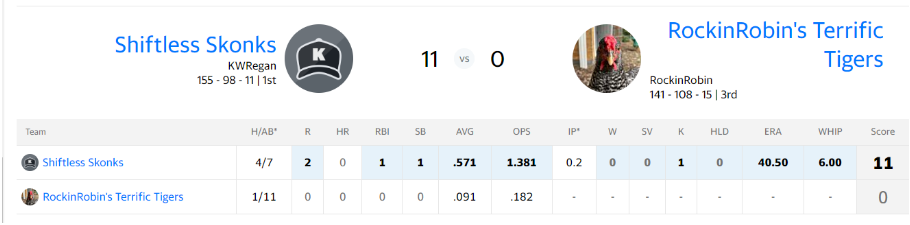
- Even if a database is completely sound, the combination of incautious programming and unseen defaults can leave loopholes.
- Suppose no one under 12 can ride a roller coaster, so they wrote:

```
SELECT * FROM Riders WHERE NOT(age < 12);
```

- And suppose Bart Simpson can upload or finagle his record not to have an `age` field.
- Even if default is `null` or something producing a non-number "`nan`" value, the `age < 12` comparison may fail "gracefully."
- Then `NOT(age < 12)` will *succeed*—and Bart gets to ride!
- Yes, `...WHERE age >= 12` would have averted the problem.
- Database can be vulnerable in-between restoring constraints after upload.
- Point is: we can't escape attention to low-level details.

Fresh Example of Edge-Case Issues

My fantasy baseball league final matchup status after 4 MLB games on Monday 9/25/23:



Shiftless Skonks
KWRegan
155 - 98 - 11 | 1st

11 vs 0

RockinRobin's Terrific Tigers
RockinRobin
141 - 108 - 15 | 3rd

Team	H/AB*	R	HR	RBI	SB	AVG	OPS	IP*	W	SV	K	HLD	ERA	WHIP	Score
Shiftless Skonks	4/7	**2**	0	**1**	**1**	**.571**	**1.381**	0.2	**0**	**0**	**1**	**0**	**40.50**	**6.00**	**11**
RockinRobin's Terrific Tigers	1/11	0	0	0	0	.091	.182	-	-	-	-	-	-	-	0

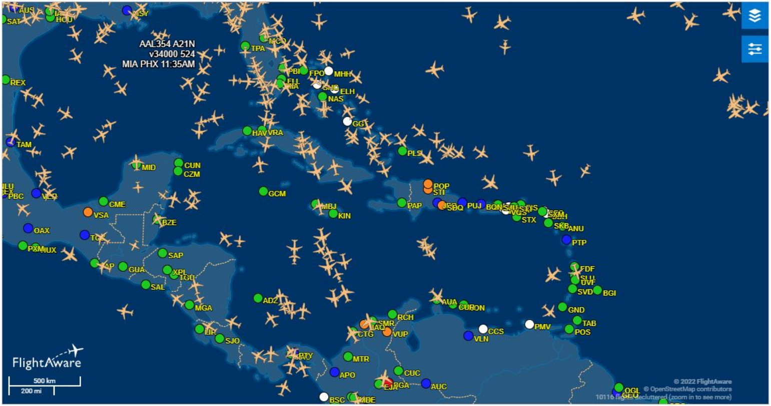
Because my opponent had no pitchers in those 4 games, his pitching scores were *null* not zero, and I got "credit" for 0 > *null*. (Never mind that 40.50 is a horrible ERA value—it still is considered to beat *null*.) This policy may nevertheless be correct on the simplest level.

Part III: A Global Data Village

- "No Man is an Island..." wrote John Donne in 1624.
- Then it was a "Meditation." Now it is 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.
- Some data is intended to be out there—other data not...
- **Structured Data** has a pre-defined format and (hence) purpose. If it wasn't meant to be out there, a breach has occurred.
- **Unstructured Data** may not have been originally intended as data.

A “Semi-Structred” Example (of Inferencing)

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

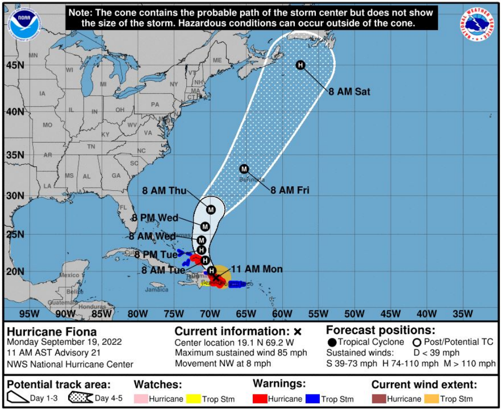


Why almost no planes over Puerto Rico and the Dominican Republic + Haiti? Compared to right now...

And what about north of the Black Sea?

Hurricane Tracking

NOAA (picture of Hurricane Fiona a year ago)



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

Abstraction in Modeling: This is done at only $1km^2$ resolution.

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. (Could discuss new Gaza examples.)

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.
- 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. (Revisit XKCD link.)
- 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.

Some Notorious Inferences and Model Decisions

- **Target**ing ads at a pregnant teen: article.
- Amazon often recommends to me the book *Quantum Algorithms Via Linear Algebra*. Problem is—I co-wrote it. Nice to hear…
- Bond and CDO (Collateralized Debt Obligation) ratings before the 2008 crash.
- Book *Weapons of Math Destruction*, by Cathy O'Neill. Thesis: Mathematical models fossilize biases in data from remote history and skewed prior sources.
- 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.
- Insofar as we are the training data for the Internet, the latter has baked in tangible amounts of racism and sexism.

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)

Part IV: Data and AI

What is **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.
- Simple Example: Building a Model.
- Simple Linear Regression Model: $Y = a + bX$.
- E.g. $\text{Walk\_Likelihood} = a + b \cdot (\text{Pitch\_Count})$.
- [show graphs from FanGraphs article, bottom of page.]
- Point is that the model can emulate/project the results of pitches by itself—when its projections go bad, the manager takes the actual flesh-and-blood pitcher out of the game. Like Blake Snell on 9/19/23!

Week 2 HW—due tonight, what to (not) look for...

- `NFLTeams.xml` data fed into a Simple Linear Model.
- The Python code grabs the XML file from my website, no download hassles... **But does your machine display Python code in the browser window?**
- Model is $Y = a + bX$ or $Y = a' + b'Z$ or both: $Y = a + b_1X + b_2Z$, where
    - Y = number of years without winning a playoff game
    - X = number of coaches since 2010
    - Z = media market size.
- How strong are the correlations? Look for R^2 in the printed lines.
- Technically, "R" from the correlation game is the square root of R^2. In this case you can kind-of ignore the difference.
- XYZ model "drought versus both" does **not** print a picture (would be 3D) but does give output. **Describe what you see and get.**
- **Main Qs:** Are the correlations strong? Would the NFL want them to be strong?

Week 2 Activity: Sentiment Analysis

Please read the activity sheet before your recitation.

- Simple idea: take words on a webpage and tell how “hot” they are.
- Also usable on Python 3 Trinket
- Can paste in my `heatindex.py` file into their `main.py`
- Use “+” button to create second file, **must name it heatlib.py** ! Paste in its code from my site.
- *Might* have to click the blue “Learn More” at top to get true Python 3. If you do so, first click the `inflammation-01.csv` tab you see, click the tab’s little gear icon, and trash the file. Then use “+” to create `heatlib.py` as above.
- Click the black triangle (familiar “play” icon) to run.
- Paste a URL from a webpage. **Twitter, wsj.com, and some other pages break...or wimp out...**
- You will probably need to restart by clicking the triangle every time, including to try the ‘`mul`’ feature.

Styles of Machine Learning

- **Supervised Learning:** examples are structured and desired responses are labeled.
  - Regression usually falls into this category.
  - *Classification* according to predetermined criteria. Knowledge Bases.
  - *Training* new Apple iPhone X on labeled datasets of faces.
- **Unsupervised Learning:** responses not labeled, data often unstructured. (Hallmark of "Big Data")
  - *Cluster analysis* is a typical example.
  - Finding patterns in data that are not pre-defined.
  - *Principal Component Analysis* (PCA)—used for face recognition too.
- **Semi-Supervised Learning:** Mix and match these approaches...
- **Reinforcement Learning:** Algorithms A act as autonomous *agents*, receive "rewards" and "demerits," and modify their parameters according to what gives increasing rewards.
- **Deep Learning:** Build layers on successful modeling...

Google and Twitter and Facebook and More

- Google Trends is not only browsable but allows interaction with staff and downloads. Used extensively by *Everybody Lies*.
- Using Twitter to predict (postdict) elections: Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- Paper, “Vehicle defect discovery from social media.”
- Consumer sentiment analysis: paper.
- Mining Facebook behavior by MasterCard: news article.
- Textbook: *Mining the Social Web*.
- Google Ngram Viewer tracks historical usage of terms and phrases.
- Automatic Text Summarization—as opposed to user-provided markup or abstracting.
- **New 11/14/23**: Hurricane forecasting by AI, incl. Lee and Otis.

Turing's Principle

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"

- 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.

Some Hard Data Challenges (based on the converse principle)

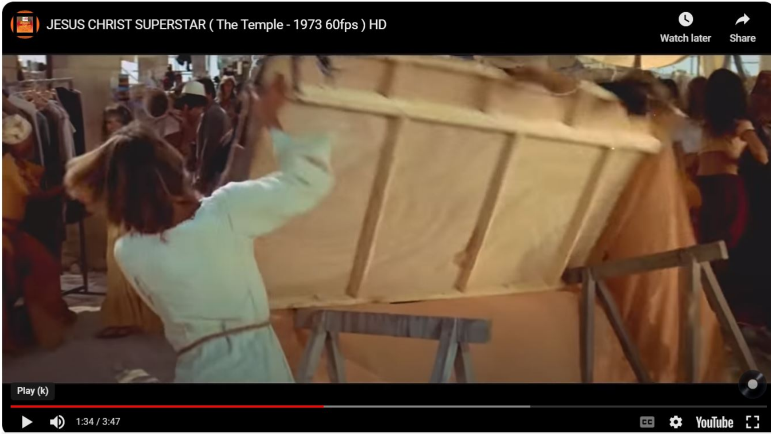
- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding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

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

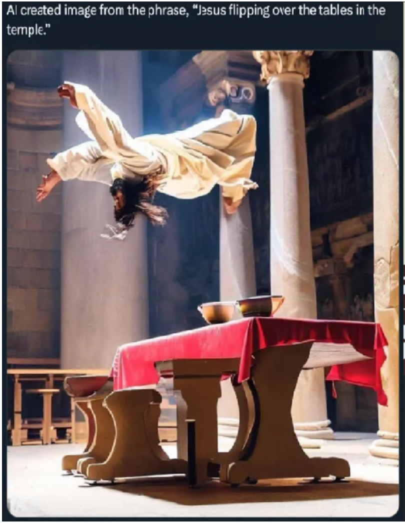
- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for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

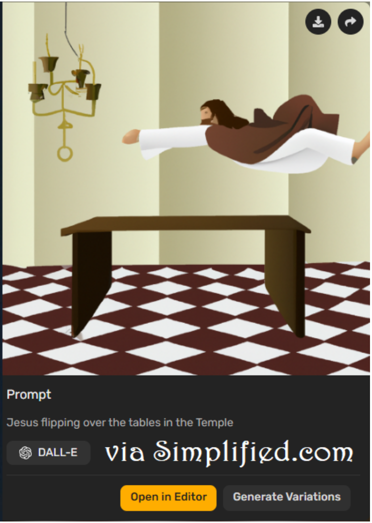
“Jesus flipping over the tables in the Temple.” From the movie *Jesus Christ Superstar*—then try it on Cutout or NightCafe or Simplified:



Two Results—one famous, one mine



AI created image from the phrase, “Jesus flipping over the tables in the temple.”



Prompt

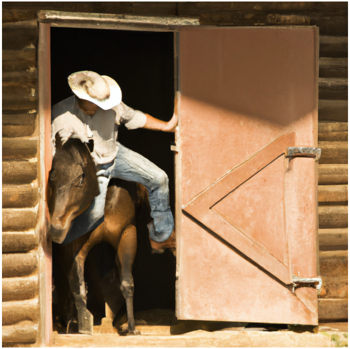
Jesus flipping over the tables in the Temple

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:



- By any chance did 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?

- 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. (END)