PLSC 197 / SODA 197N: Social Data, Technology, and Artificial Intelligence

Christopher Zorn

January 20, 2026

Wherefore "data"?

Linguistics:

- "Data" in common usage: *mass noun* (like "rice," "rain," "music," etc.) → singular
- "Data" in scientific usage: plural of *datum* → plural

Usage:

- If: "data" $\equiv$ "information": *singular*
- If:
  - · "data" $\equiv$ "facts": *plural*
  - · "data" $\equiv$ "data points": *plural*

What is "data"?

"*Data*: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing."

- Borgman (2008)

Data isn't...(only) numbers

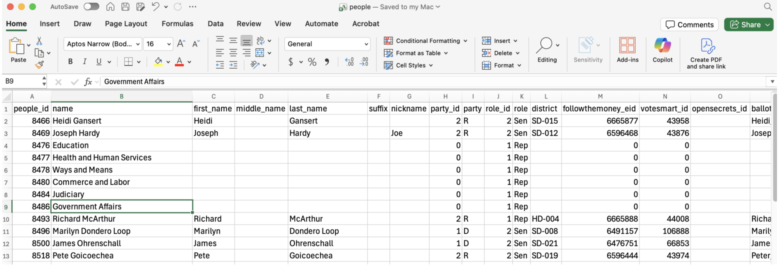
Basic data types:

- Numeric data:
  - Integers
  - Real numbers
  - Floating points, etc.
- Characters
- Booleans (true / false)
- Void / Null (empty)

Derived/user-defined data types:

- Functions
- Strings
- Lists
- Dates
- etc.

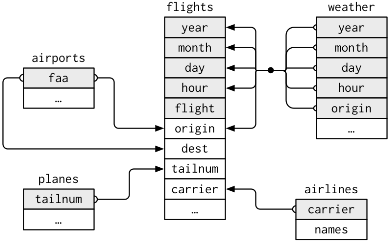
Data isn't… rectangular



B9 Government Affairs

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	people_id	name	first_name	middle_name	last_name	suffix	nickname	party_id	party	role_id	role	district	followthemoney_eid	votesmart_id	opensecrets_id	ballot
2	8466	Heidi Gansert	Heidi		Gansert			2	R	2	Sen	SD-015	6665877	43958		Heidi
3	8469	Joseph Hardy	Joseph		Hardy		Joe	2	R	2	Sen	SD-012	6596468	43876		Josep
4	8476	Education						0		1	Rep		0	0		
5	8477	Health and Human Services						0		1	Rep		0	0		
6	8478	Ways and Means						0		1	Rep		0	0		
7	8480	Commerce and Labor						0		1	Rep		0	0		
8	8484	Judiciary						0		1	Rep		0	0		
9	8486	Government Affairs						0		1	Rep		0	0		
10	8493	Richard McArthur	Richard		McArthur			2	R	1	Rep	HD-004	6665888	44008		Richa
11	8496	Marilyn Dondero Loop	Marilyn		Dondero Loop			1	D	2	Sen	SD-008	6491157	106888		Marily
12	8500	James Ohrenschall	James		Ohrenschall			1	D	2	Sen	SD-021	6476751	66853		Jame
13	8518	Pete Goicoechea	Pete		Goicoechea			2	R	2	Sen	SD-019	6596444	43974		Peter

```
{
    "type": "message",
    "channel": "C2147483705",
    "user": "U2147483697",
    "text": "Hello world",
    "ts": "1355517523.000005",
    "is_starred": true,
    "pinned_to": ["C024BE7LT", ...],
    "reactions": [
        {
            "name": "astonished",
            "count": 3,
            "users": [ "U1", "U2", "U3" ]
        },
        {
            "name": "facepalm",
            "count": 1034,
            "users": [ "U1", "U2", "U3", "U4", "U5"
]
        }
```



Data isn't...(only) information

{celery, enchant, borscht, symmetry, jovial, daybreak}

{daisy, refuge, imagine, armadillo, futon, frolic}

"Data":

- ...requires *context*
- ...requires *purpose*
- ...implies *value*

Data isn't...*natural*

"Data is the new oil."

- mathematician Clive Humby, 2006

Data:

- ...requires refining to be useful, *but*
- ...it's also renewable, non-rivalrous, non-excludable, non-fungible, etc. Moreover,
- ...data is a *specific asset*, which makes its value highly variable. And finally,
- ...it *doesn't exist at all without human intervention*

Data Across (Academic) Disciplines

Adapted from Borgman (2008):

	P&NS*	S&BS**	A&H***
Centrality	Highest	High	Moderate
Diversity	High	Higher	Highest
Structure†	High	Varies	Varies/Low
Data Generation	Common	Often	Rarely
Infrastructure	Highly Developed	Developed	Developing

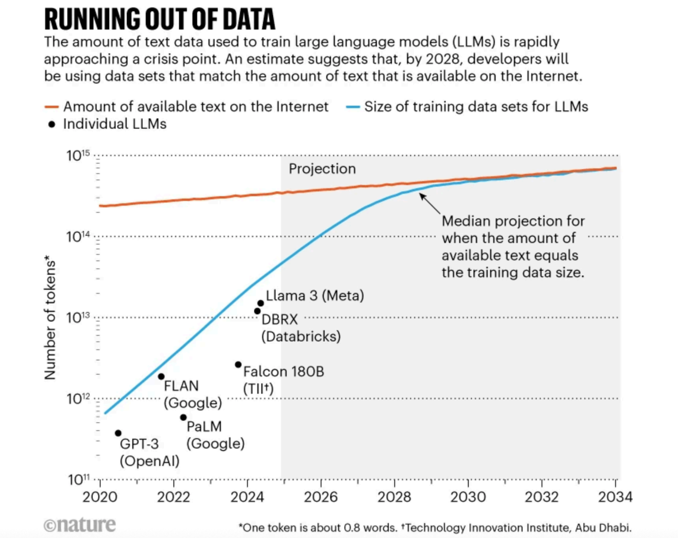
* Physical and natural sciences.

** Social and behavioral sciences.

*** Arts and humanities.

† More on this next week...

Jones (2024): “AI is running out of data...”



RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly approaching a crisis point. An estimate suggests that, by 2028, developers will be using data sets that match the amount of text that is available on the Internet.

*One token is about 0.8 words. †Technology Innovation Institute, Abu Dhabi.