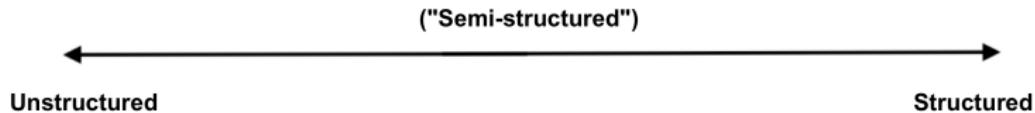


PLSC 197 / SODA 197N: Social Data, Technology, and Artificial Intelligence

Christopher Zorn

January 27, 2026

Data Structure



Key: The Data Model:

“...an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.”

Structured data (e.g., Ademola-Shanu 2025) has:

- A predefined *schema* (e.g., a “codebook”)
- Consistent *formatting*
- Ease of *storage*
- High *searchability*

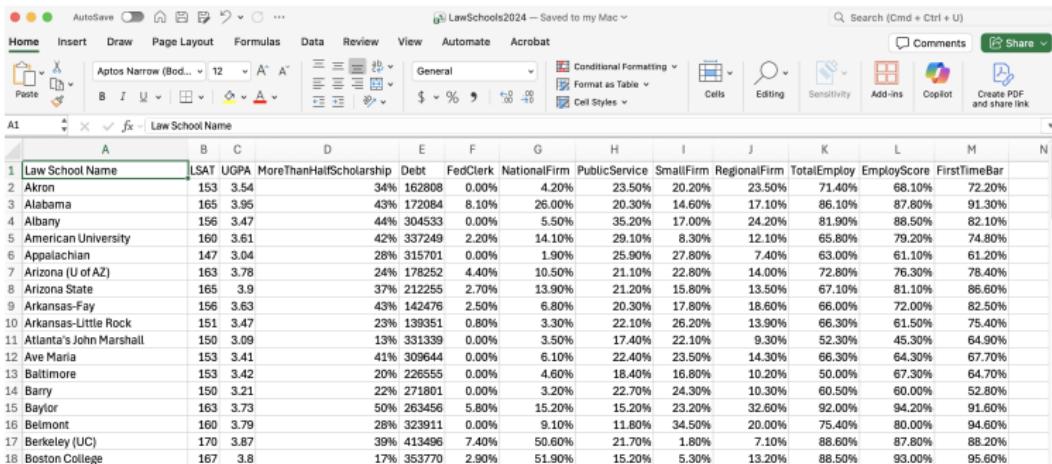
Characteristics:

- Intuitive / easy to use + understand
- Myriad tools for managing + analysis
- Central for machine learning

Structured Data: Examples

Rectangular / “linear” data:

- Rows are *cases* / *observations* (units / “things”)
- Columns are *variables* / *features* (characteristics of units)

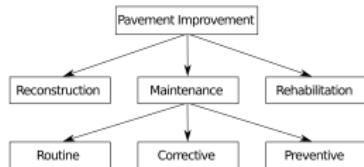


The screenshot shows a Microsoft Excel spreadsheet titled "LawSchools2024". The ribbon menu includes Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Automate, and Acrobat. The formula bar shows "A1 Law School Name". The data starts at A1 and consists of 18 rows of law school information. The columns are labeled from A to M and include: Law School Name, LSAT, UGPA, MoreThanHalfScholarship, Debt, FedClerk, NationalFirm, PublicService, SmallFirm, RegionalFirm, TotalEmploy, EmployScore, and FirstTimeBar.

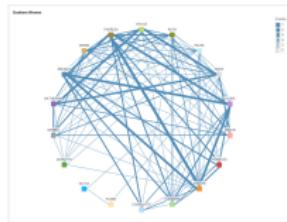
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Law School Name	LSAT	UGPA	MoreThanHalfScholarship	Debt	FedClerk	NationalFirm	PublicService	SmallFirm	RegionalFirm	TotalEmploy	EmployScore	FirstTimeBar	
2	Akron	153	3.54		34%	162808	0.00%	4.20%	23.50%	20.20%	23.50%	71.40%	68.10%	72.20%
3	Alabama	165	3.95		43%	172084	8.10%	26.00%	20.30%	14.60%	17.10%	86.10%	87.80%	91.30%
4	Albany	156	3.47		44%	304533	0.00%	5.50%	35.20%	17.00%	24.20%	81.90%	88.50%	82.10%
5	American University	160	3.61		42%	337249	2.20%	14.10%	29.10%	8.30%	12.10%	65.80%	79.20%	74.80%
6	Appalachian	147	3.04		28%	315701	0.00%	1.90%	25.90%	27.80%	7.40%	63.00%	61.10%	61.20%
7	Arizona (U of AZ)	163	3.78		24%	178252	4.40%	10.50%	21.10%	22.80%	14.00%	72.80%	76.30%	78.40%
8	Arizona State	165	3.9		37%	212255	2.70%	13.90%	21.20%	15.80%	13.50%	67.10%	81.10%	86.60%
9	Arkansas-Fay	156	3.63		43%	142476	2.50%	6.80%	20.30%	17.80%	18.60%	66.00%	72.00%	82.50%
10	Arkansas-Little Rock	151	3.47		23%	139351	0.80%	3.30%	22.10%	26.20%	13.90%	66.30%	61.50%	75.40%
11	Atlanta's John Marshall	150	3.09		13%	331339	0.00%	3.50%	17.40%	22.10%	9.30%	52.30%	45.30%	64.90%
12	Ave Maria	153	3.41		41%	309644	0.00%	6.10%	22.40%	23.50%	14.30%	66.30%	64.30%	67.70%
13	Baltimore	153	3.42		20%	266555	0.00%	4.60%	18.40%	16.80%	10.20%	50.00%	67.30%	64.70%
14	Barry	150	3.21		22%	271801	0.00%	3.20%	22.70%	24.30%	10.30%	60.50%	60.00%	52.80%
15	Baylor	163	3.73		50%	263456	5.80%	15.20%	15.20%	23.20%	32.60%	92.00%	94.20%	91.60%
16	Belmont	160	3.79		28%	323911	0.00%	9.10%	11.80%	34.50%	20.00%	75.40%	80.00%	94.60%
17	Berkeley (UC)	170	3.87		39%	413496	7.40%	50.60%	21.70%	1.80%	7.10%	88.60%	87.80%	88.20%
18	Boston College	167	3.8		17%	353770	2.90%	51.90%	15.20%	5.30%	13.20%	88.50%	93.00%	95.60%

Structured Data: Other Examples

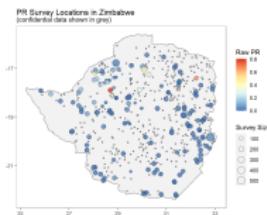
Hierarchical Model



- Hierarchical Data (nested)



- Network Data (nodes + edges)



- Spatial Data (geolocated)

Structured Data: Formats

Rectangular formats:

- .csv (Comma-separated values)
- .xlsx (Microsoft Excel)
- .parquet (Parquet)
- SQL tables (see below)
- Others (.tsv, .txt, etc.)

Hierarchical Data: Excel

The screenshot shows a Microsoft Excel spreadsheet with the following details:

- Header Row:** The first row contains column headers: A, B, G, K, BC, BD, BE, BF, BG, BH, and BI.
- Data Rows:** There are 12 data rows, each representing a case entry. The columns correspond to the headers: caseld, docketId, usCite, term, justiceName, vote, opinion, direction, majority, firstAgreement, and secondAgreement.
- Formatting:** The table has a light gray background. The header row is bolded. Column K (BC) is merged across the first two data rows. Columns BD, BE, BF, BG, BH, and BI are merged across the last six data rows.
- Clipboard:** The clipboard icon in the ribbon indicates that the last action was a copy operation.
- Font:** The font is set to "Aptos Narrow (Bodoni)" at size 12.
- Number Format:** The number format is set to "General".
- Cells:** Cell L1 is selected, and the formula bar shows "naturalCourt".

	A	B	G	K	BC	BD	BE	BF	BG	BH	BI
1	caseld	docketId	usCite	term	justiceName	vote	opinion	direction	majority	firstAgreement	secondAgreement
2	1946-001	1946-001-01	329 U.S. 1	1946	HHBurton	2	1	1	1		
3	1946-001	1946-001-01	329 U.S. 1	1946	RHJackson	1	1	2	2		
4	1946-001	1946-001-01	329 U.S. 1	1946	WODouglas	1	1	2	2		
5	1946-001	1946-001-01	329 U.S. 1	1946	FFrankfurter	4	2	2	2		
6	1946-001	1946-001-01	329 U.S. 1	1946	SReed	1	1	2	2		
7	1946-001	1946-001-01	329 U.S. 1	1946	HBlack	1	2	2	2		
8	1946-001	1946-001-01	329 U.S. 1	1946	WB Rutledge	1	1	2	2		
9	1946-001	1946-001-01	329 U.S. 1	1946	FMurphy	1	1	2	2		
10	1946-001	1946-001-01	329 U.S. 1	1946	FMVinson	1	1	2	2		
11	1946-002	1946-002-01	329 U.S. 14	1946	HHBurton	1	1	1	2		
12	1946-002	1946-002-01	329 U.S. 14	1946	RHJackson	2	3	2	1		

Network Data: Excel

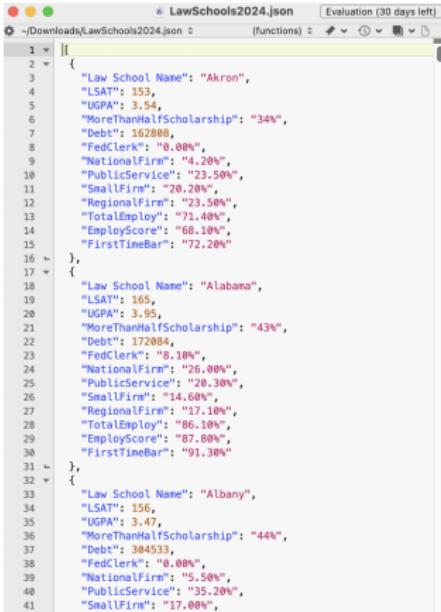
The screenshot shows a Microsoft Excel spreadsheet titled "N16". The ribbon menu is visible at the top, showing tabs for Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Automate, and Acrobat. The Home tab is selected. The formula bar shows the formula `Aptos Narrow (Bod... 16 A^ A^`. The main area displays a 10x10 grid of data representing a network adjacency matrix. The columns and rows are labeled with the names of Supreme Court Justices: Rehnquist, Stevens, OConnor, Scalia, Kennedy, Souter, Thomas, Ginsburg, and Breyer. The matrix entries are binary values (0 or 1), indicating connections between justices. For example, Rehnquist has connections to Stevens (0), OConnor (1), Scalia (1), Kennedy (1), Souter (0), Thomas (1), and Breyer (0). Stevens has connections to OConnor (0), Scalia (0), Kennedy (0), Souter (1), Thomas (0), and Breyer (1). The diagonal elements are all 1s, representing self-loops.

	A	B	C	D	E	F	G	H	I	J	K
1		Rehnquist	Stevens	OConnor	Scalia	Kennedy	Souter	Thomas	Ginsburg	Breyer	
2	Rehnquist	0	0	1	1	1	0	1	0	0	
3	Stevens	0	0	0	0	0	1	0	1	1	
4	OConnor	1	0	0	0	1	1	0	1	1	
5	Scalia	1	0	0	0	1	0	1	0	0	
6	Kennedy	1	0	1	1	0	1	1	1	1	
7	Souter	0	1	1	0	1	0	0	1	1	
8	Thomas	1	0	0	1	1	0	0	0	0	
9	Ginsburg	0	1	1	0	1	1	0	0	1	
10	Breyer	0	1	1	0	1	1	0	1	0	

Spatial Data: Excel

	A	B	C	F	L	M	O	Q	R	S
1	FID	ORGANIZA	FILENAME	CAUSE	STARTDATED	CONTRDATE	STATE	LATITUDE	LONGITUDE	TOTALACRES
2	1603	FWS	BIG BERTHA	Human	3/26/1988 0:00	3/27/1988 0:00	Arizona	31.58333	-111.55	1500
3	1605	FWS	MORMON	Human	5/15/1986 0:00	5/19/1986 0:00	Arizona	32.5	-111.51667	10390
4	1608	FWS	NORTH	Human	6/27/1986 0:00	6/28/1986 0:00	Montana	47.5	-111.43333	1400
5	1647	FWS	YELLOW	Human	2/28/2002 0:00	3/2/2002 0:00	Arizona	31.7	-111.483	1035
6	1668	FWS	GUS	Human	4/9/2000 0:00	4/10/2000 0:00	Arizona	31.516	-111.517	5700
7	1673	FWS	LANE	Human	5/14/2000 0:00	5/18/2000 0:00	Arizona	31.649	-111.491	2750
8	1675	FWS	CITY HALL	Human	5/14/2002 0:00	5/15/2002 0:00	Arizona	31.766	-111.483	5312
9	1677	FWS	CITYHALL2	Human	6/2/2000 0:00	6/4/2000 0:00	Arizona	31.791	-111.458	5200
10	1680	FWS	CITY HALL	Human	5/16/1991 0:00	5/18/1991 0:00	Arizona	31.75	-111.45	6530
11	1682	FWS	CUMERO	Natural	7/26/1991 0:00	7/30/1991 0:00	Arizona	31.46667	-111.43333	2500
12	1707	FWS	HIPPY	Human	7/6/1994 0:00	7/7/1994 0:00	Arizona	31.716	-111.433	2500
13	1708	FWS	SASABE	Human	7/4/1994 0:00	7/4/1994 0:00	Arizona	31.5	-111.55	1200
14	1760	FWS	WALTERS	Human	3/18/1992 0:00	3/20/1992 0:00	California	33.28333	-114.71667	1800
15	1775	FWS	CIBOLA	Natural	7/16/2006 0:00	7/23/2006 0:00	Arizona	33.32199	-114.705	4600
16	1777	FWS	CAMINO	Human	4/18/2005 0:00	4/21/2005 0:00	Arizona	32.25056	-113.63028	1025
17	1778	FWS	GROWLER PEAK	Human	5/13/2005 0:00	5/18/2005 0:00	Arizona	32.3888	-113.28639	7500
18	1803	FWS	SOUTH DIKE	Human	7/5/1998 0:00	9/2/1998 0:00	Arizona	34.7653	-114.533	2200
19	1846	FWS	FERGUSONLKL	Natural	5/7/1987 0:00	5/8/1987 0:00	California	33	-114.5	2500
20	1851	FWS	FERGUSON	Human	5/28/1989 0:00	6/1/1989 0:00	California	33.01667	-114.51667	3080
21	1875	FWS	KING VALLY	Human	10/1/2005 0:00	10/6/2005 0:00	Arizona	33.17889	-114.02139	26000
22	1892	FWS	SANTIAGO	Natural	6/19/1988 0:00	6/22/1988 0:00	California	34.9298	-119.3114	1900
23	1901	FWS	CLEAR	Natural	7/3/2001 0:00	7/5/2001 0:00	California	41.86667	-121.125	4317
24	1964	FWS	PIRU	Human	#####	10/24/1998 0:00	California	34.418	-118.794	1257
25	1966	FWS	HOPPER	Human	8/5/1997 0:00	8/11/1997 0:00	California	34.466	-118.867	24800
26	1968	FWS	PIRU	Human	#####	11/10/2003 0:00	California	34.48722	-118.75972	63719
27	1990	FWS	REFUGE	Human	8/4/1998 0:00	9/18/1998 0:00	California	41.95	-121.73333	1500
28	2106	FWS	TURKEYTR	Human	6/12/1992 0:00	6/12/1992 0:00	California	35.91667	-119.41667	1200
29	2114	FWS	HANFORD FA	Human	6/14/1993 0:00	6/14/1993 0:00	California	35.92333	-119.35	1560

Structured Data: JSON



A screenshot of a code editor window titled "LawSchools2024.json". The file contains an array of objects representing law schools. Each object has properties like "Law School Name", "LSAT", "UGPA", "MoreThanHalfScholarship", "Debt", "FedClerk", and various employment statistics. The code editor shows syntax highlighting and line numbers from 1 to 41.

```
1  [
2    {
3  "Law School Name": "Akron",
4  "LSAT": 153,
5  "UGPA": 3.54,
6  "MoreThanHalfScholarship": "34%",
7  "Debt": 162888,
8  "FedClerk": "0.00%",
9  "NationalFirm": "4.20%",
10 "PublicService": "23.50%",
11 "SmallFirm": "20.20%",
12 "RegionalFirm": "23.50%",
13 "TotalEmploy": "21.40%",
14 "EmployScore": "68.10%",
15 "FirstTimeBar": "72.20%"
16 },
17   {
18  "Law School Name": "Alabama",
19  "LSAT": 165,
20  "UGPA": 3.95,
21  "MoreThanHalfScholarship": "43%",
22  "Debt": 172084,
23  "FedClerk": "8.10%",
24  "NationalFirm": "26.00%",
25  "PublicService": "20.30%",
26  "SmallFirm": "14.60%",
27  "RegionalFirm": "17.10%",
28  "TotalEmploy": "86.10%",
29  "EmployScore": "87.80%",
30  "FirstTimeBar": "91.30%"
31 },
32   {
33  "Law School Name": "Albany",
34  "LSAT": 156,
35  "UGPA": 3.47,
36  "MoreThanHalfScholarship": "44%",
37  "Debt": 304533,
38  "FedClerk": "0.00%",
39  "NationalFirm": "5.50%",
40  "PublicService": "35.20%",
41  "SmallFirm": "17.00%"
```

JavaScript Object Notation (JSON) is:

- Lightweight (little extra formatting)
- Text-based
- Human-readable
- Language-independent
- Simple / fast
- Widely used for web/cloud-based applications

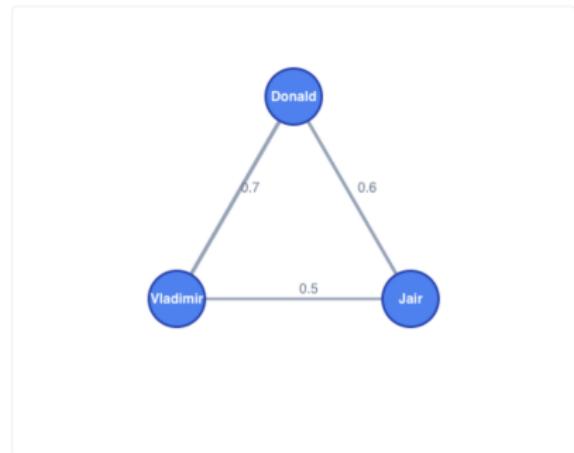
Hierarchical Data: JSON

A screenshot of a code editor window displaying hierarchical JSON data. The interface includes a toolbar at the top with font selection (Menlo, Regular, 18), color palette, and standard text editing icons. Below the toolbar is a horizontal ruler with numerical markers from 0 to 6. The JSON code itself is written in red and black text. It starts with a root object 'CEO' containing three nested objects: 'Manager A', 'Manager B', and 'Manager C'. Each manager object contains two employee objects: 'Employee A1' and 'Employee A2' for Manager A; 'Employee B1' and 'Employee B2' for Manager B; and 'Employee C1' and 'Employee C2' for Manager C. Each employee object maps to a single department name: Accounting, Distribution, Design, Engineering, Marketing, or Marketing.

```
{'CEO': { 'Manager A': { 'Employee A1': {Accounting}, 'Employee A2': {Distribution}}, 'Manager B': { 'Employee B1': {Design}, 'Employee B2': {Engineering}}, 'Manager C': { 'Employee C1': {Marketing}, 'Employee C2': {Marketing}}}}
```

Network Data: JSON

```
{  
  "nodes": [  
    {  
      "id": "1",  
      "label": "Donald"  
    },  
    {  
      "id": "2",  
      "label": "Vladimir"  
    },  
    {  
      "id": "3",  
      "label": "Jair"  
    }  
  ],  
  "edges": [  
    {  
      "source": "1",  
      "target": "2",  
      "value": 0.7  
    },  
    {  
      "source": "2",  
      "target": "3",  
      "value": 0.5  
    },  
    {  
      "source": "1",  
      "target": "3",  
      "value": 0.6  
    }  
  ]  
}
```



Databases are collections of structured data:

- Managed / analyzed using a “database management system (DBMS)”
- Original “navigational;” later primarily “relational” (SQL)
- More recently: “NewSQL” (relational 2.0)

Relational Databases

The primary form of structured database in use in AI is the *relational database*:

- Typically *rectangular tables* ("flat files") with identifying *keys* in each row
- Each table has a unique unit / item type that it stores
- Tables are connected by *relationships*
- Changes to one table leave others intact
- Interface: Structured Query Language (SQL)

Example: Comparative Legislators Database

Comparative Legislators Database

</>

The Comparative Legislators Database (CLD) is a one-stop shop for rich, diverse and integrated individual-level data on national political representatives.

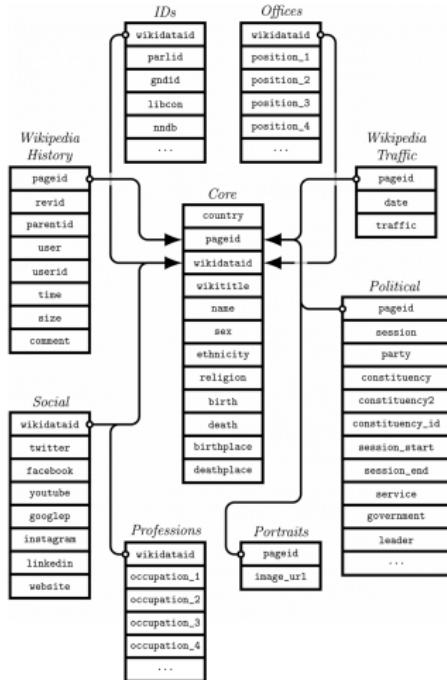
The database contains information for over 67,000 contemporary and historical legislators from 16 countries. It unites collaborative micro-data collection efforts and brings these together through the integration of data from Wikipedia, Wikidata, and other sources.

Use the menu above to learn more about the database, to receive an overview of the data, to check out the accompanying R package `legislatoR`, and to learn how to contribute.



- Over 67,000 contemporary and historical legislators
- 16 countries
- (As far back as) 1789 - 2022
- Includes sociodemographic, political, historical, occupational, and social network data, plus Wikipedia traffic
- Also: `LegislatoR` R package
- Website:
<https://complegdatabase.com/>

Comparative Legislators Database



- Each legislature = 9 tables
- Each table is a “flat file”
(think: Excel sheet)
- Each legislator’s characteristics are connected to data on their country, occupations, web presence, Wikipedia page(s), etc.

Unstructured Data

Data that do not fit the above-referenced structure(s).

Most data is unstructured...

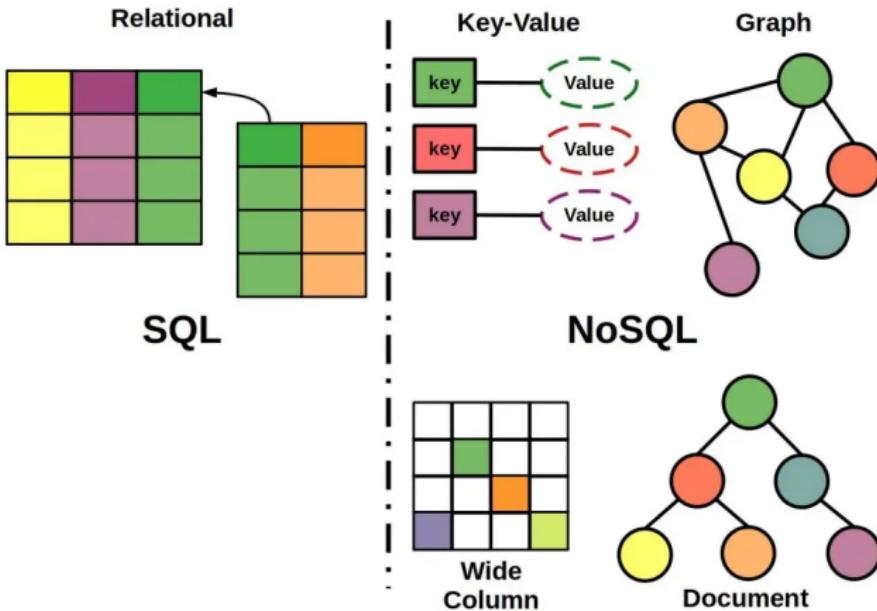
- Much text (social media, documents, etc.)
- Images, audio, video files
- Web content (blends of the above)

Semi-Structured Data

Semi-structured data...

- ...doesn't have a structured schema, but
- ...contains tags/markers that enforce types and hierarchies across objects.
- Data and schema are *integrated* ("self-describing")
- Formats: XML, JSON, YAML, etc.
- Interface: NoSQL

SQL vs. NoSQL



(Source)

Semi- vs. Structured Data

Characteristic	Structured/SQL	Semi-Structured/NoSQL
Data Model	Relational	Varies*
Schema	Predefined	Dynamic / Schema-on-read
Scales	Vertically	Horizontally
ACID**	Yes	No

* Relational, network, etc.

** Atomicity, Consistency, Isolation, Durability

Digression: ACID

ACID means:

1. **Atomicity:** All transactions must succeed or fail completely and cannot be left partially complete, even in the case of system failure.
2. **Consistency:** The database must follow rules that validate and prevent corruption at every step.
3. **Isolation:** Concurrent transactions cannot affect each other.
4. **Durability:** Transactions are final, and even system failure cannot “roll back” a complete transaction.

Schöch (2013): Data in the Humanities

Key points:

- “Big data” (volume, velocity, variety) vs. “smart data”
 - Which does he seem to prefer, and why?
 - Is it an accurate / fair description?
- Challenge: Make big data smarter, or smart data bigger... how?
- Compare to physical / natural sciences?
- Social / behavioral sciences?