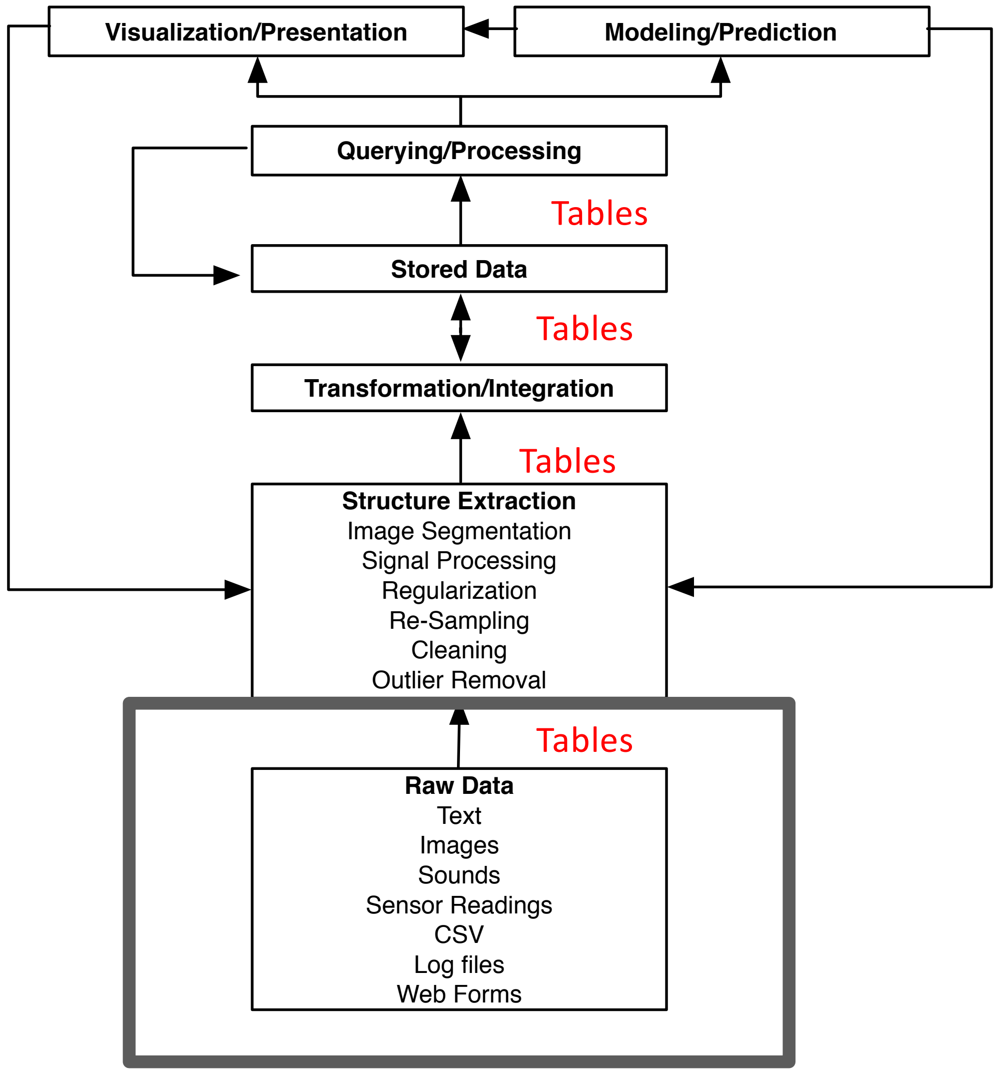
- Project Proposals (March 4)

6.S079 Lec 6

Data Cleaning & Entity Resolution

DATA SCIENCE PIPELINE



RECAP

Last time:

- Text manipulation tools:
  - grep, sed, awk
- Text similarity:
  - Jaccard similarity
  - Cosine distance
  - TF/IDF
  - Embeddings

`^x?$|^(xx+?)\1+$`

Generates a string of length n, to test if n is prime (match = not prime)

^x?$ — *base case*: 0 and 1 are not prime

(? matches preceding character 0 or 1 times)

| — *or*

^(xx+?) — *two or more xs*

(? makes + match smallest substring)

Without ?:

xxxxxx — *No match*

xxxxxx — *No match*

xxxxxx — *No match*

xxxxxx — *Match! ➔ Prime*

With ?:

xxxxxx — *Match!*

? does not affect correctness; any match indicates non-prime

Search algorithm is to look for smallest (w ?, largest) match; if none found, backtrack and repeated with one larger (smaller) subsequence

PERFORMANCE

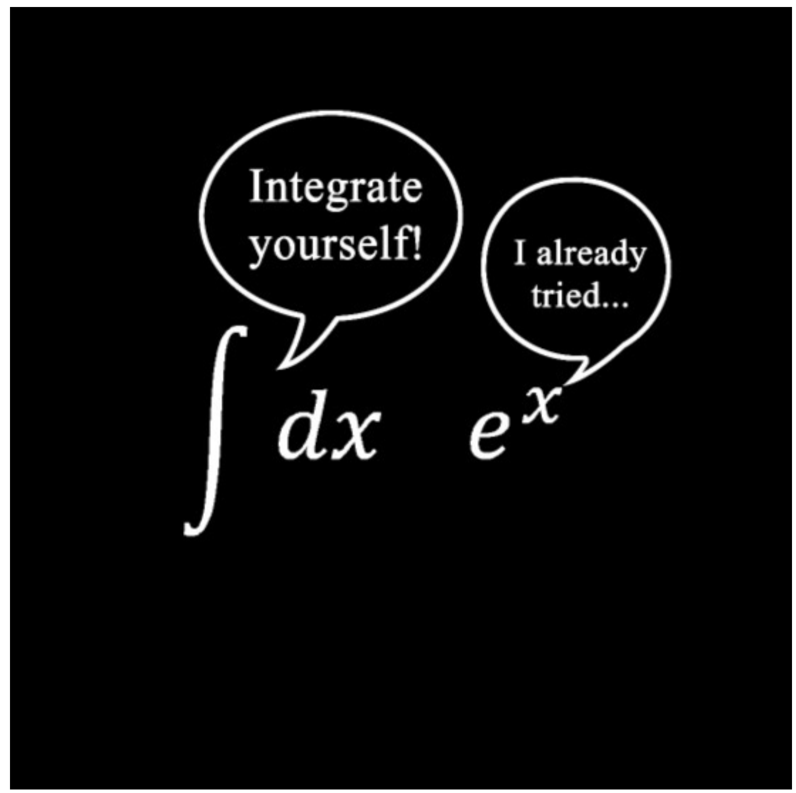
```
import re
import time
def prime(n):
    s = "x" * n
    return re.match("^x?$|^(xx+?)\\1+$", s)

def prime_largest(n):
    s = "x" * n
    return re.match("^x?$|^(xx+)\\1+$", s)


for n in [10000, 100000, 99991, 99999, 100000]:
    print(f"N = {n}")
    start = time.time()
    r1 = prime(n)
    end = time.time()
    print(f"\tsmallest first: {end - start:.2}")
    start = time.time()
    r2 = prime_largest(n)
    end = time.time()
    print(f"\tlargest first: {end - start:.2}")
```

N = 10000
    smallest first: 0.00021
    largest first: 0.0085
N = 100000
    smallest first: 0.0013
    largest first: 0.79
N = 99991
    smallest first: 3.2
    largest first: 3.2
N = 99999
    smallest first: 0.0026
    largest first: 1.4
N = 100000
    smallest first: 0.0015
    largest first: 0.79

Clearly, matching smallest first will perform better, since largest first always has to try at least first N/2 before it finds a match

THIS TIME

- Data Integration and Cleaning
  - Dealing with tabular data with errors
  - Combining tabular datasets
  - Handling missing data



EXAMPLE TASK

How many people work in the US IT industry?
What is the avg revenue per employee in the tech industry?

EXAMPLE TASK

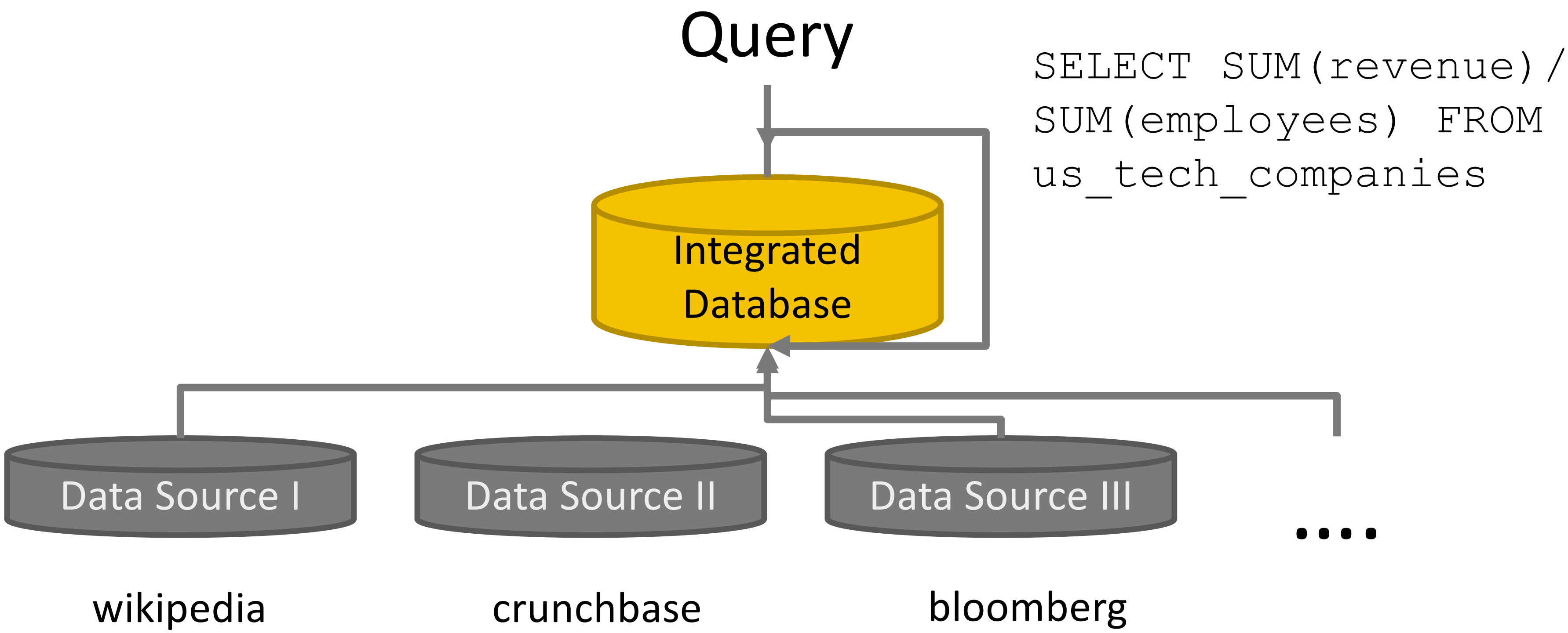
Rank[1]		Company	Fiscal Year Ending	Revenue ($B) USD	Employees	Headquarters
1		Apple Inc.	30 September 2017[2]	$229.2[1][3]	123,000[3]	Cupertino, CA, US
2		Samsung Electronics	31 December 2017[4]	$211.9[1][5][6]	320,670[7][8]	Suwon, South Korea
3		Amazon	31 December 2017[9][10]	$177.9[1][10]	613,300[11]	Seattle, WA, US
4		Foxconn	31 December 2017[12][13]	$154.7–158[1][13][14]	803,126[15]	New Taipei City, Taiwan
5		Alphabet Inc.	31 December 2017[16][17]	$110.8[1][17]	80,110[18]	Mountain View, CA, US
6		Microsoft	30 June 2017[19]	$90.0[1]	124,000[19]	Redmond, WA, US
7		Huawei	31 December 2017[20][21]	$89.3–92.5[1][21]	180,000	Shenzhen, China
8		Hitachi	31 March 2018[22]	$84.6[1]	307,275	Tokyo, Japan
9		IBM	31 December 2017[23][24]	$79.1[1]	397,800	Armonk, NY, US
10		Dell Technologies	31 January 2018[25][26]	$78.7[1][26]	145,000[25]	Round Rock, TX, US
11		Sony	31 March 2018[27]	$77.1[1][28]	117,300[27]	Tokyo, Japan
12		Panasonic	31 March 2018[29]	$72.0[1]	274,143	Osaka, Japan
13		Intel	31 December 2017[30]	$62.8[1]	102,700	Santa Clara, CA, US
14		LG Electronics	31 December 2017[31]	$54.3[1]	74,000	Seoul, South Korea
15		JD.com	31 December 2017[32]	$54.0[1]	157,831	Beijing, China
16		HP Inc.	31 October 2017[33]	$52.0[1]	49,000	Palo Alto, CA, US

Private and semipublic companies with the most employees in the world			
Rank	Employer	Country	Employees
1	Walmart	United States	2,200,000
2	China National Petroleum	China	1,382,401
3	China Post Group	China	935,191
4	State Grid	China	917,717
5	Hon Hai Precision Industry (Foxconn)	Taiwan	667,680
6	Volkswagen	Germany	664,496
7	Amazon	United States	647,500
8	Sinopec Group	China	619,151
9	Compass Group	United Kingdom	595,841
10	United States Postal Service	United States	565,802

United States Largest Private Employers (as of 2017)[1][2][3][4] [hide]			
Rank	Employer	Global number of Employees	Median annual pay
1	Walmart	2,300,000	$19,177
2	Amazon	469,690	$36,969
	Deutsche Post DHL	499,018	
3	United Parcel Service	456,415	$53,443
4	Yum! Brands	450,000	$9,111
5	Kroger	449,000	$21,075
6	Home Depot	413,000	$20,095
7	Berkshire Hathaway	377,000	$53,510 (BH directly employs c. 30 people. All the others are employed by the companies BH purchases.)
8	International Business Machines	366,000	$55,088
9	FedEx	357,000	$50,017
10	Target Corporation	345,000	$20,581
11	General Electric	313,000	$57,211
12	Walgreens Boots Alliance	290,000	$31,132
13	Starbucks	277,000	$12,754
14	Albertsons	273,000	
15	PepsiCo	263,000	$47,801
16	Wells Fargo	262,700	$60,466
17	Cognizant Technology Solutions	260,000	$31,998
18	UnitedHealth Group	260,000	$58,378
19	Lowe's	240,000	$23,905
20	AT&T	268,540	$95,814

```
,name,domain,year founded,industry,size range,locality,country,linkedin url,current employee
estimate,total employee estimate
5872184,ibm,ibm.com,1911,information technology and services,10001+,"new york, new york, united
states",united states,linkedin.com/company/ibm,274047,716906
4425416,tata consultancy services,tcs.com,1968,information technology and
services,10001+,"bombay, maharashtra, india",india,linkedin.com/company/tata-consultancy-
services,190771,341369
21074,accenture,accenture.com,1989,information technology and services,10001+,"dublin, dublin,
ireland",ireland,linkedin.com/company/accenture,190689,455768
2309813,us army,goarmy.com,1800,military,10001+,"alexandria, virginia, united states",united
states,linkedin.com/company/us-army,162163,445958
1558607,ey,ey.com,1989,accounting,10001+,"london, greater london, united kingdom",united
kingdom,linkedin.com/company/ernstandyoung,158363,428960
3844889,hewlett-packard,hpe.com,1939,information technology and services,10001+,"palo alto,
california, united states",united states,linkedin.com/company/hewlett-packard-
enterprise,127952,412952
2959148,cognizant technology solutions,cognizant.com,1994,information technology and
services,10001+,"teaneck, new jersey, united states",united
states,linkedin.com/company/cognizant,122031,210020
5944912,walmart,walmartcareers.com,1962,retail,10001+,"withee, wisconsin, united states",united
states,linkedin.com/company/walmart,120753,272827
3727010,microsoft,microsoft.com,1975,computer software,10001+,"redmond, washington, united
states",united states,linkedin.com/company/microsoft,116196,276983
3300741,at&t,att.com,1876,telecommunications,10001+,"dallas, texas, united states",united
states,linkedin.com/company/at&t,115188,269659
5412257,united states air force,airforce.com,1947,defense & space,10001+,"randolph, texas,
united states",united states,linkedin.com/company/united-states-air-force,113997,316549
2780814,pwc,pwc.com,1998,accounting,10001+,"new york, new york, united states",united
states,linkedin.com/company/pwc,111372,379447
3972223,wells fargo,wellsfargo.com,,financial services,10001+,"san francisco, california,
united states",united states,linkedin.com/company/wellsfargo,109532,264101
1454663,infosys,infosys.com,1981,information technology and services,10001+,"bangalore,
karnataka, india",india,linkedin.com/company/infosys,104752,215718
3221953,deloitte,deloitte.com,1900,management consulting,10001+,"new york, new york, united
```

EXAMPLE TASK



On average, what is the revenue per employee in the tech sector in the US?

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

What are some errors you see here?

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

Duplicate Entities
(Entity Resolution)

Pattern Violation

Outdated data / wrong data

Spelling mistakes / abbreviations

Encoding Error
(nb in thousands)

Rule Violations

Missing values
(known unknowns)

MORE?

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

Known Unknowns

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

Unknown Unknowns

OUTLINE

Data Integration

- Schema matching
- Entity resolution
- Blocking, etc

Data Cleaning

- **Missing values** → Value imputation
- **Missing records** → Species estimation

OUTLINE

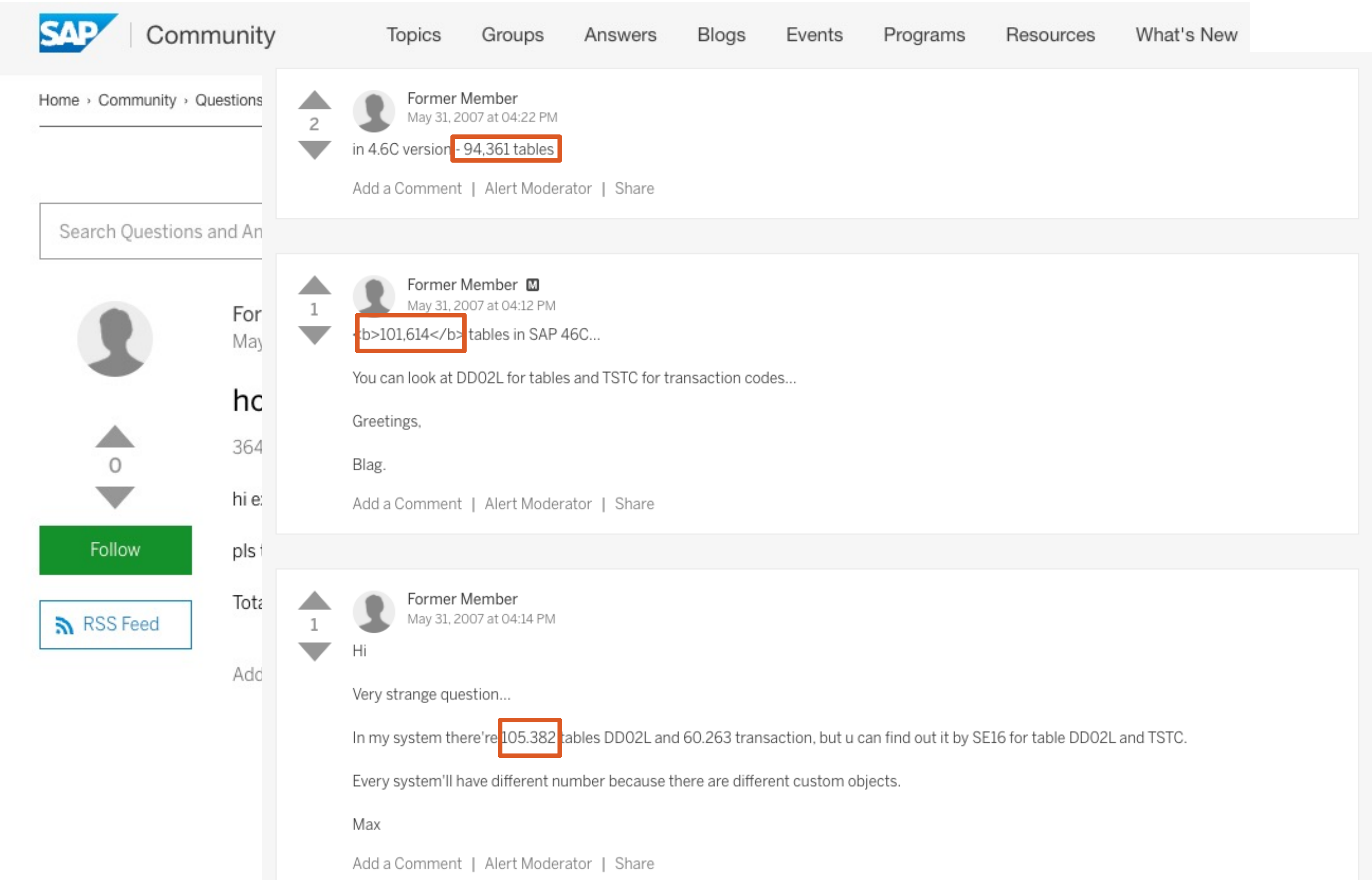
Data Integration

- Schema matching
- Entity resolution
- Blocking, etc

Data Cleaning

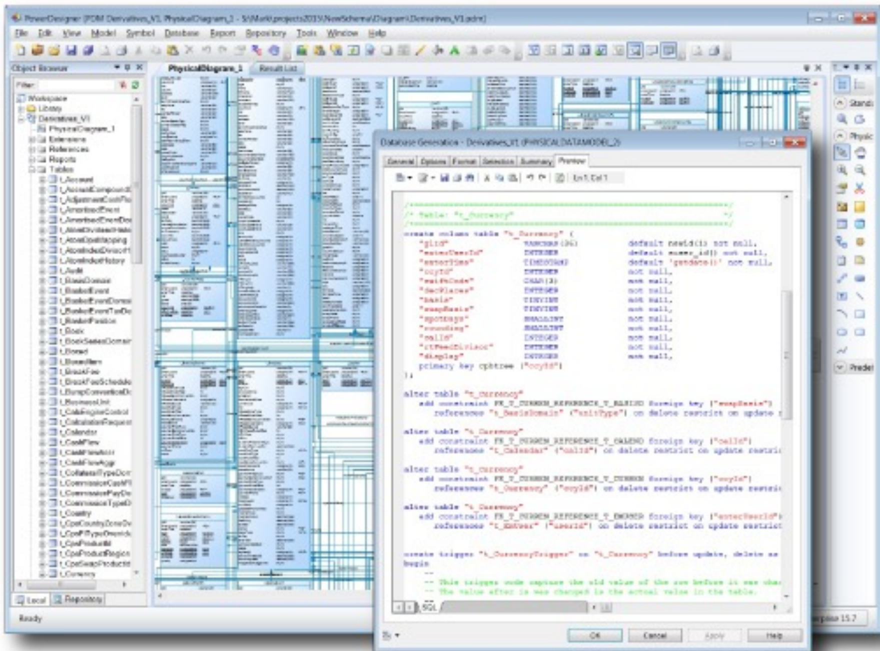
- **Missing values** → Value imputation
- **Missing records** → Species estimation

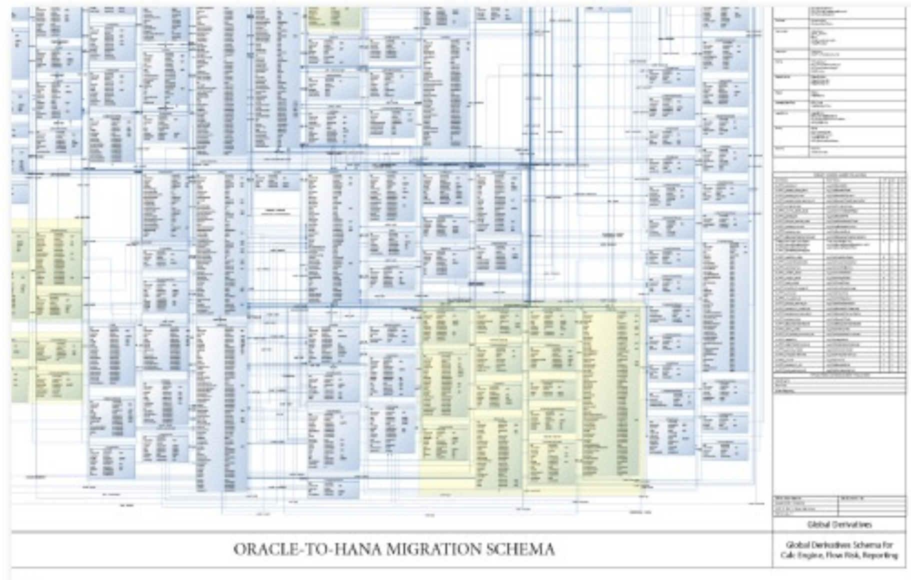
WHY IS SCHEMA MATCHING HARD



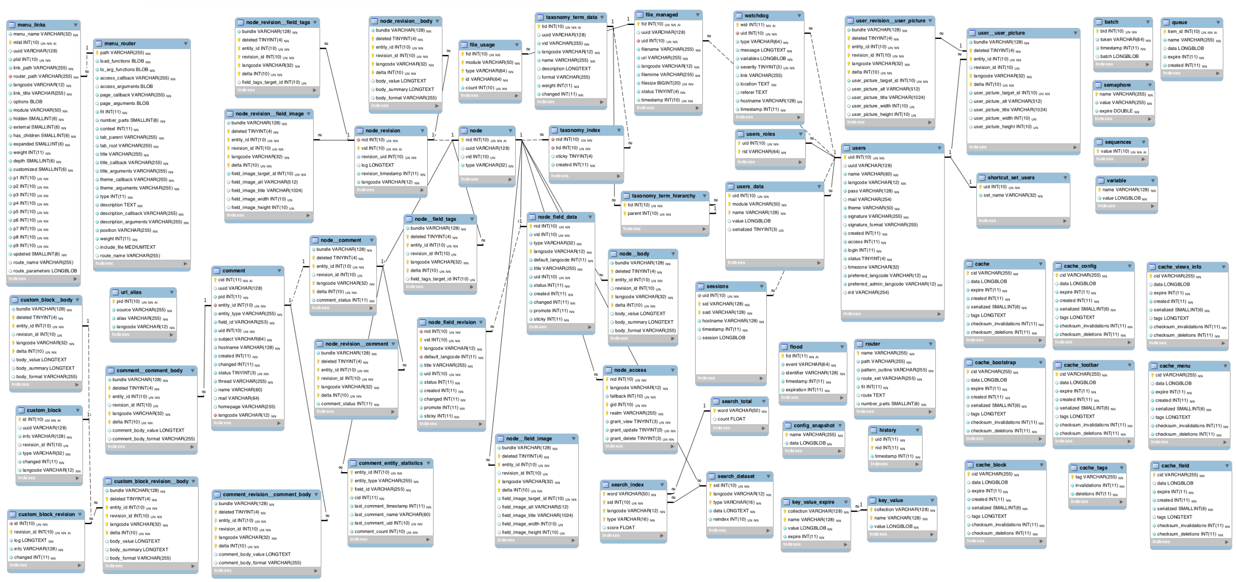
SCHEMAS CAN BE REALLY COMPLICATED

SAP (very small fraction)





Drupal 8



SCHEMA MATCHING

Goal is to match columns from two tables, to produce a single table with the same schema

Complicated because people use different names, types, #s of columns for attributes

E.g., name vs firstName, lastName

addr vs addrNo, addrSt, addrCty, addrState…

Typical approach: find columns with a similar name, the same data type, and high overlap in values

DATA OFTEN HAS MANY CONSTRAINTS

Key, uniqueness, functional dependencies, foreign keys

What do these terms mean?

Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

Takes_Course

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Courses

DATA OFTEN HAS MANY CONSTRAINTS

Value range, format, etc.

Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

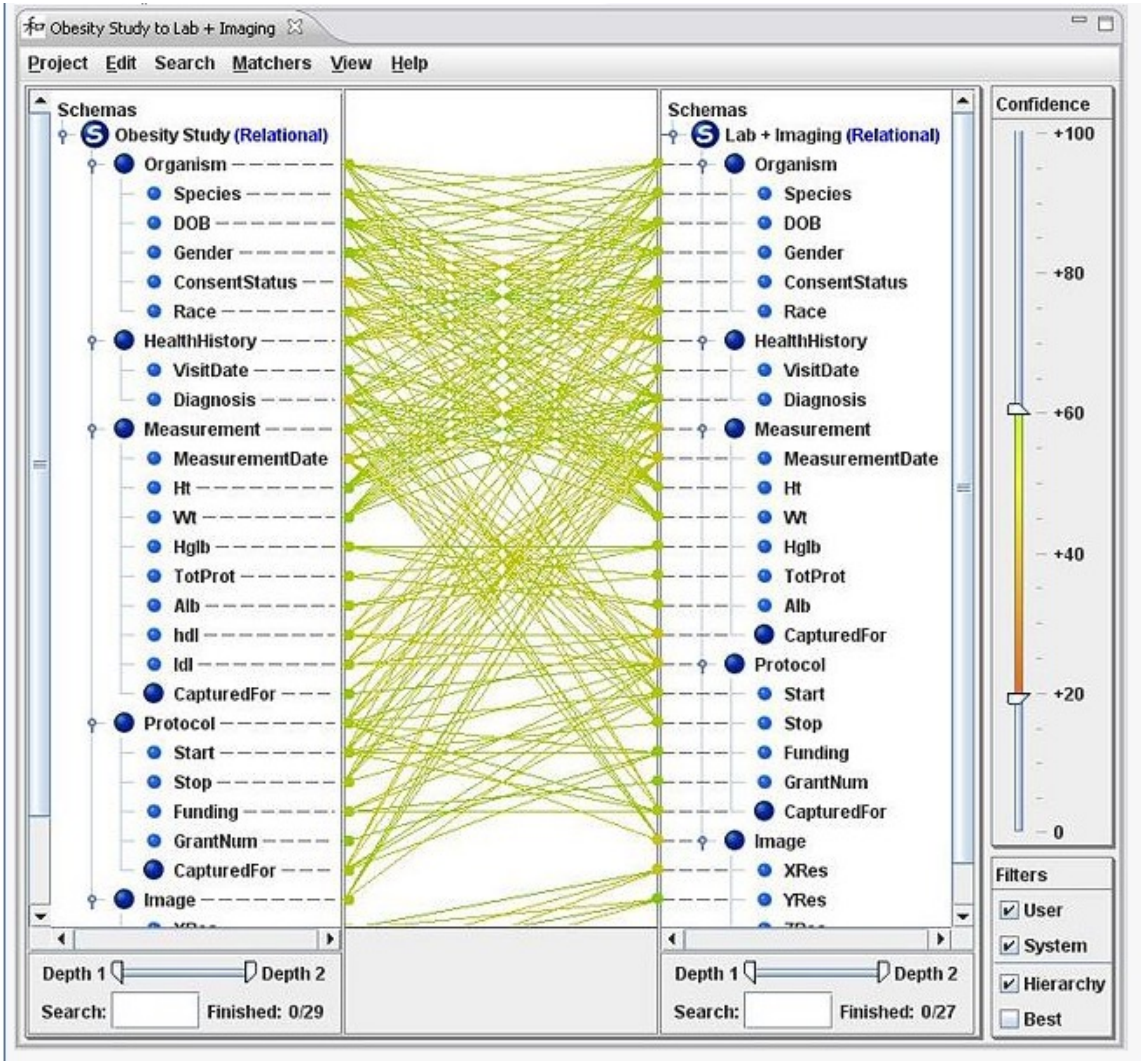
Takes_Course

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Courses

HARMONY

http://openii.sourceforge.net/



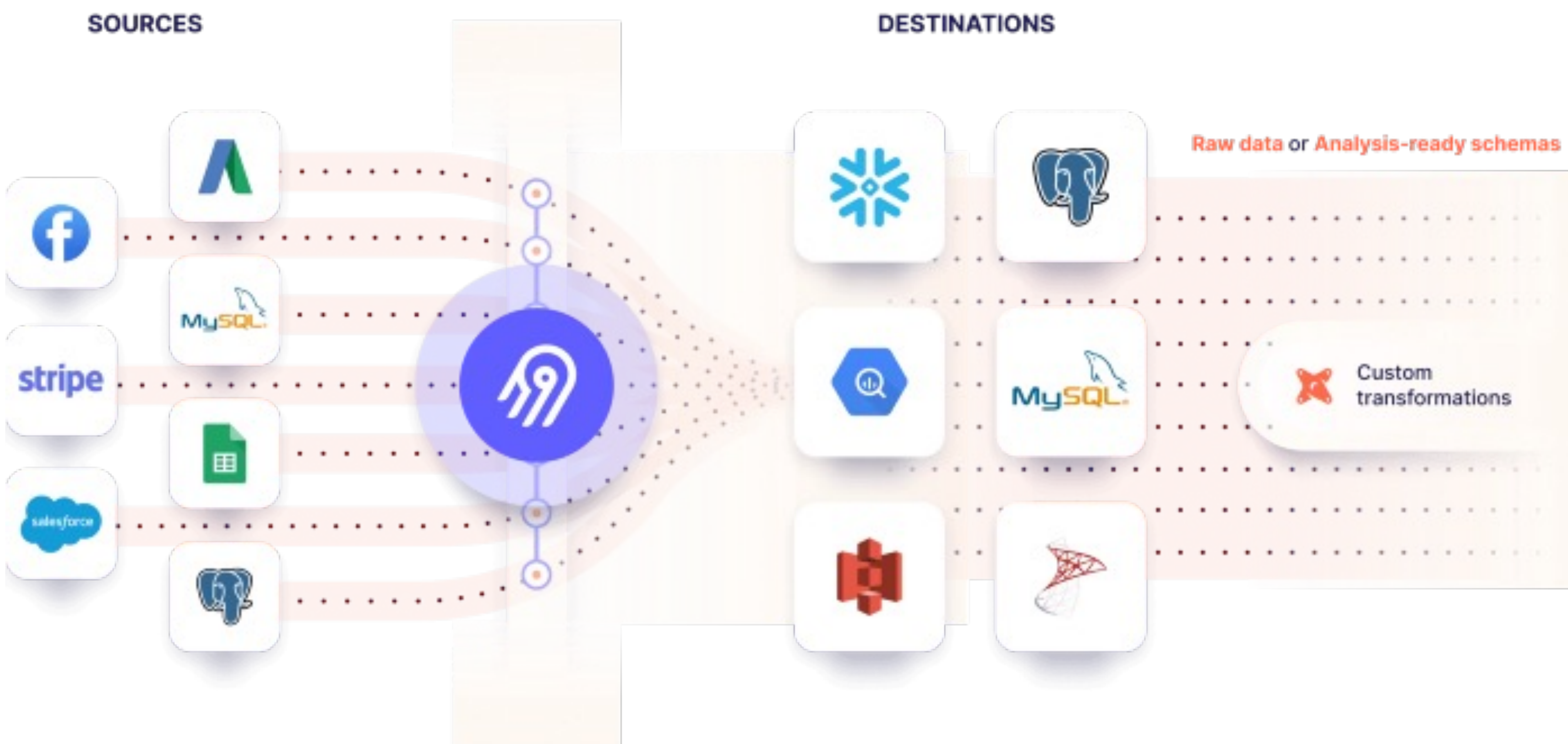
EVERY COMPANY HAS TO DEAL WITH IT

- salesforce
- Google Ads
- amazon advertising
- cart.com
- SAP
- HubSpot
- Marketo
- okta
- stripe
- ....

→

- amazon REDSHIFT
- snowflake
- Google Big Query
- Azure Synapse Analytics
- ....

DATA INTEGRATION OPEN-SOURCE/STARTUPS



Airbyte

Fivetran

DATA LAKES TO THE RESCUE?

OUTLINE

Data Integration

- Schema matching
- Entity resolution
- Blocking, etc

Data Cleaning

- **Missing values** → Value imputation
- **Missing records** → Species estimation

ENTITY RESOLUTION

"[The] problem of identifying and linking/grouping different manifestations of the same real world object."

Challenges

- Fundamental ambiguity
- Diversity in representations (format, truncation, ambiguity)
- Errors
- Missing data
- Records from different times
- Relationships in addition to equality

TEXT SIMILARITY

Customer

Id	Name	Street	City	State	P-Code	Age
1	J Smith	123 University Ave	Seattle	Washington	98106	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle	WA	98101	19
6	James Smith	123 Univ Ave	Seatle	WA	NULL	41
7	J Widom	123 University Ave	Palo Alto	CA	94305	NULL
…	…	…	…	…	…	…

TEXT SIMILARITY

String Similarity function:

- $Sim(string, string) \rightarrow numeric\ value$

A “good” similarity function:

- Strings representing the same concept ⇒ high similarity
- Strings representing different concepts ⇒ low similarity

EDIT DISTANCE

EditDistance(s1, s2):

- Minimum number of edits to transform s1 to s2

Edit:

- Insert a character
- Delete a character
- Substitute a character

Note: EditDistance(s1, s2) = EditDstance(s2, s1)

EDIT DISTANCE

EditDistance ("Provdince", "Providence") = 2

Provdince $\longrightarrow$ Providence $\longrightarrow$ Providence

EditDistance("Seattle", "Redmond") = 6

Seattle $\longrightarrow$ Reattle $\longrightarrow$ Redttle

Redmtle $\longrightarrow$ Redmole $\longrightarrow$ Redmone

$\longrightarrow$ Redmond

EDIT DISTANCE PROBLEMS

115th Waterman St., Providence, RI

EditDistance = 1

110th Waterman St., Providence, RI

Waterman Street, Providence, RI

EditDistance = 4

Waterman St, Providence, RI

Character Level vs. Word Level Similarity?

EDIT DISTANCE PROBLEMS

148th Ave NE, Redmond, WA

EditDist = 0

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

EditDist = 4

NE 148th Ave, Redmond, WA

Order sensitive Similarity?

JACCARD SIMILARITY

- **Saw last time**
- **Statistical measure**
- **Originally defined over sets**
- **String = set of words**

$$Jaccard(s1, s2) = \frac{|s1 \cap s2|}{|s1 \cup s2|}$$

- **Range of values = [0,1]**

OTHER SIMILARITY FUNCTIONS

- Embedding Distance (BERT, etc)
- Affine edit distance
- Cosine similarity
- Hamming distance
- Generalized edit distance
- Jaro distance
- Monge-Elkan distance
- Q-gram
- Smith-Warerman distance
- Soundex distance
- TF/IDF
- …many more

- No universally good similarity function
- Choice of similarity function depends on domains of interest, data instances, etc.

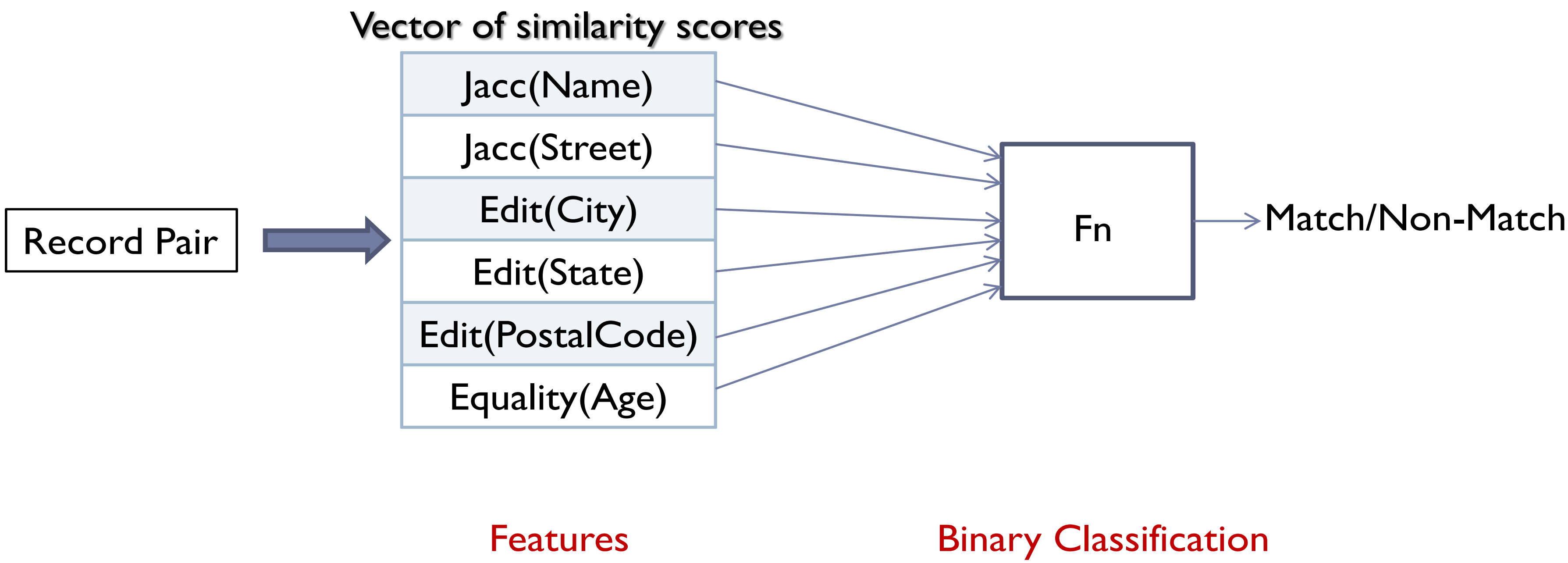
RECORD MATCHING PROBLEMS

Customer

Id	Name	Street	City	State	P-Code	Age
1	J Smith	123 University Ave	Seattle	Washington	98106	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle	WA	98101	19
6	James Smith	123 Univ Ave	Seatle	WA	NULL	41
7	J Widom	123 University Ave	Palo Alto	CA	94305	NULL
…	…	…	…	…	…	…

$WtJaccard =$ 0.57 0.91 1.0 0.0 1.0 1.0

COMBINING SIMILARITY FUNCTIONS



LEARNING-BASED APPROACH

Bob Wilson	345 Broadway	Seattle	Washington	98101	19
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Match

B Wilson	123 Broadway	Boise	Idaho	83712	19
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Non-Match

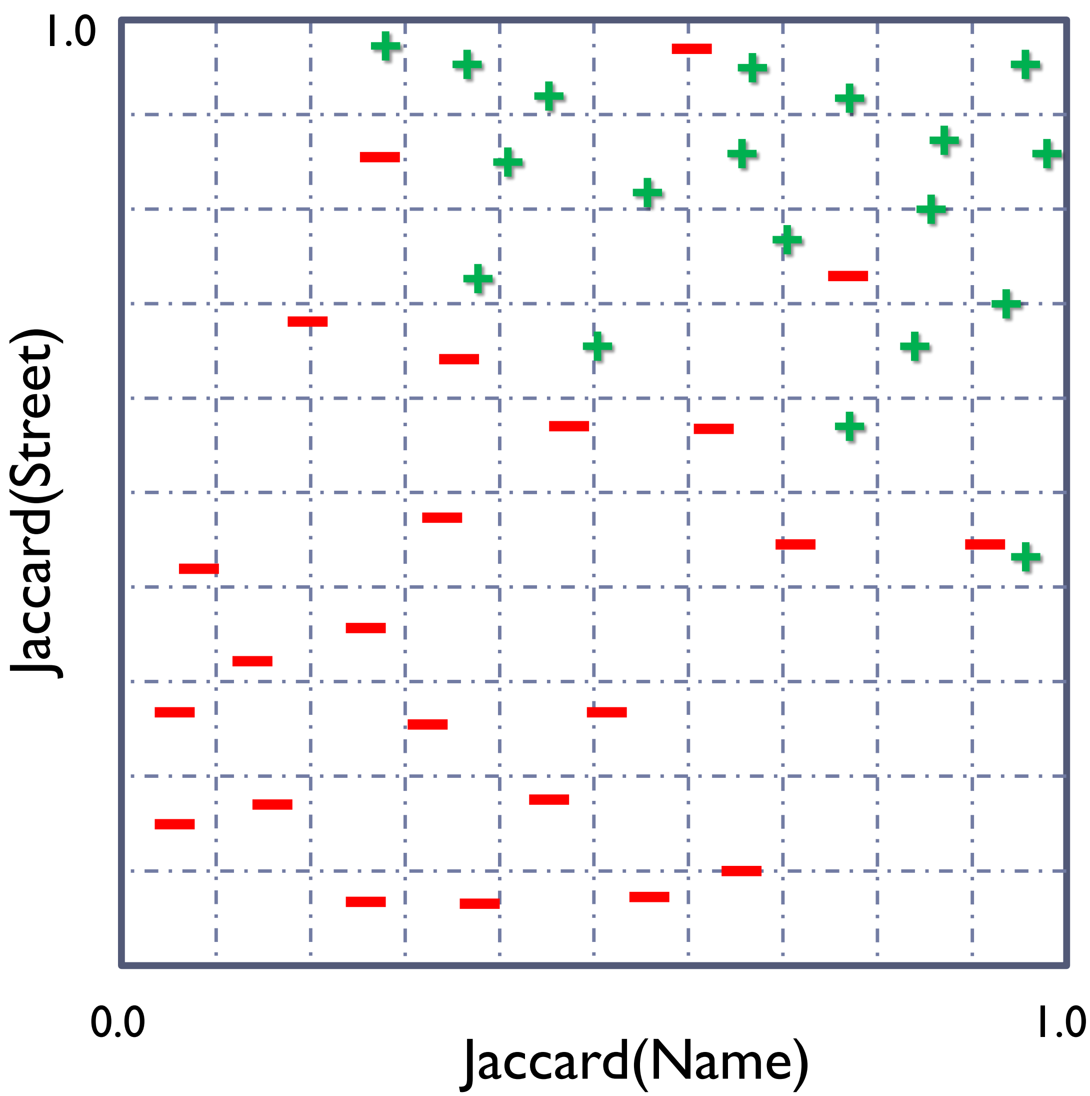
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
M Jones	245 Third Street	Redmond	NULL	98052	299

Match

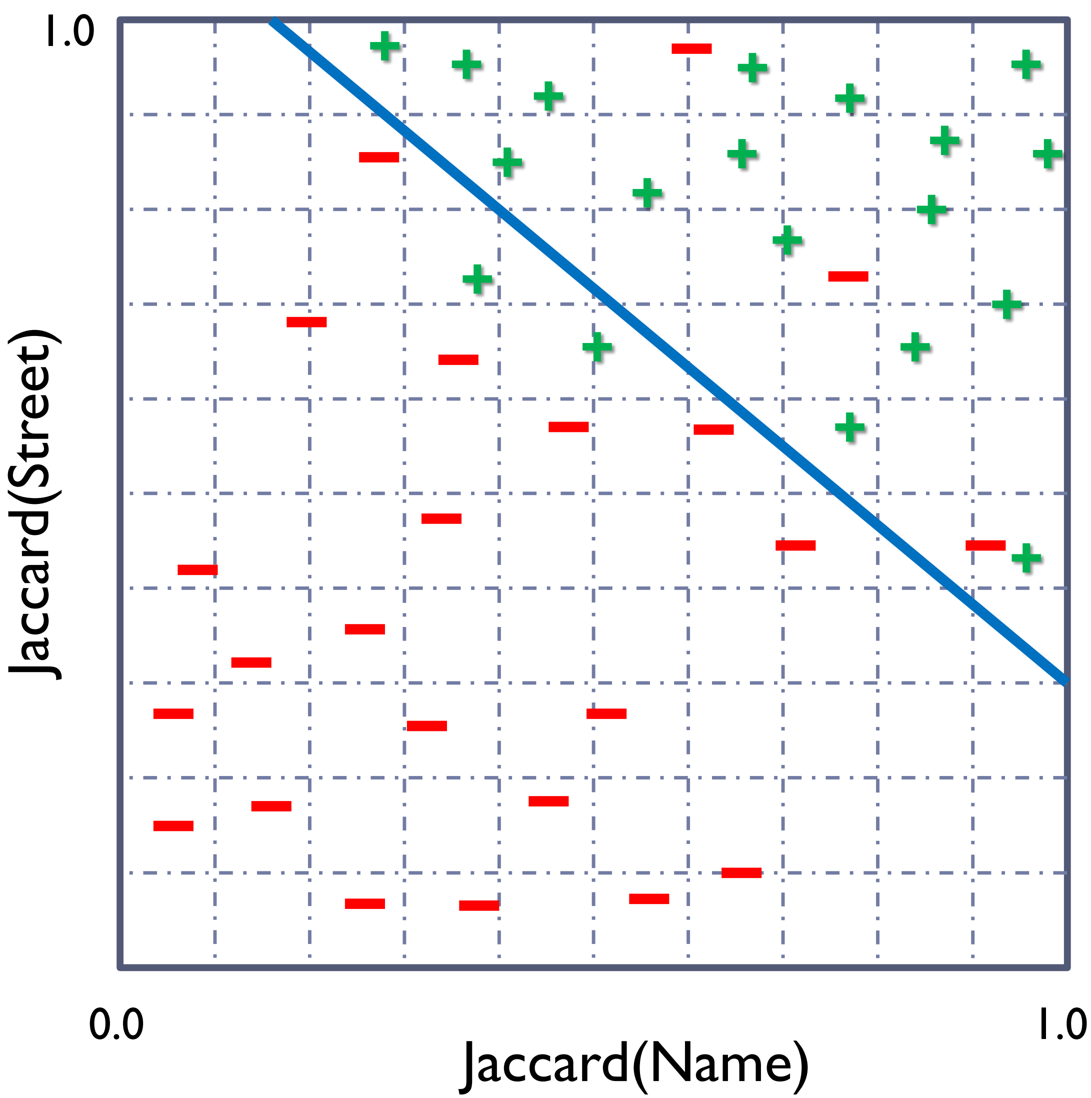
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Non-Match

LEARNING BASED APPROACH



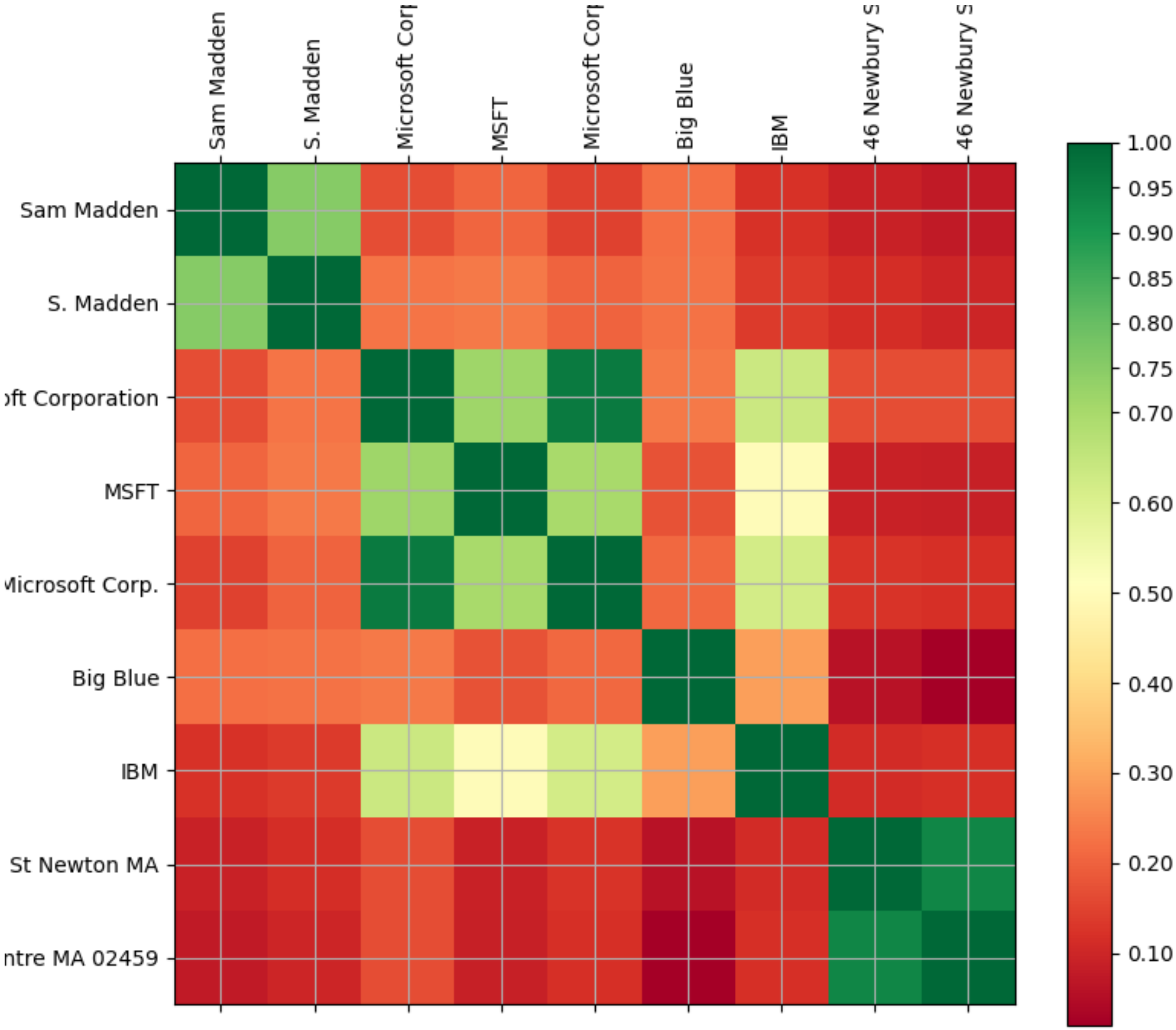
LEARNING BASED APPROACH



$$0.73\,Jacc(Name) + 0.89\,Jacc(Street) \geq 1$$

EMBEDDINGS TO THE RESCUE?

```python
def do_bert():
    model = SentenceTransformer('all-mpnet-base-v2')
    sen_embeddings = model.encode(sen)
    from sklearn.metrics.pairwise import cosine_similarity

    cos_sim = cosine_similarity(sen_embeddings)
    plot_sim_matrix(cos_sim, sen)

sen = [
    "Sam Madden",
    "S. Madden",
    "Microsoft Corporation",
    "MSFT",
    "Microsoft Corp.",
    "Big Blue",
    "IBM",
    "46 Newbury St Newton MA",
    "46 Newbury Street Newton Centre MA 02459"
]
```



Sam Madden
S. Madden
Microsoft Corp
MSFT
Microsoft Corp
Big Blue
IBM
46 Newbury S
46 Newbury S
Sam Madden
S. Madden
oft Corporation
MSFT
Microsoft Corp.
Big Blue
IBM
St Newton MA
ntre MA 02459
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.60
0.50
0.40
0.30
0.20
0.10

OUTLINE

Data Integration

- Schema matching
- Entity resolution
- Blocking, etc

Data Cleaning

- **Missing values** → Value imputation
- **Missing records** → Species estimation

SCALING CHALLENGE: BLOCKING

Matching is a quadratic process

Naively, have to compare every record in dataset A to every record in B

Idea: only compare similar records, i.e., by splitting records based on some attribute, either manually (e.g., using intuition) or automatically (e.g., using clustering)

Dataset 1

Name	Address	Dept
Sam	1st St	EECS
Mike	2nd Ave	ME
Mary	1st St	Physics
Yuan	2nd Ave	Math

Dataset 2

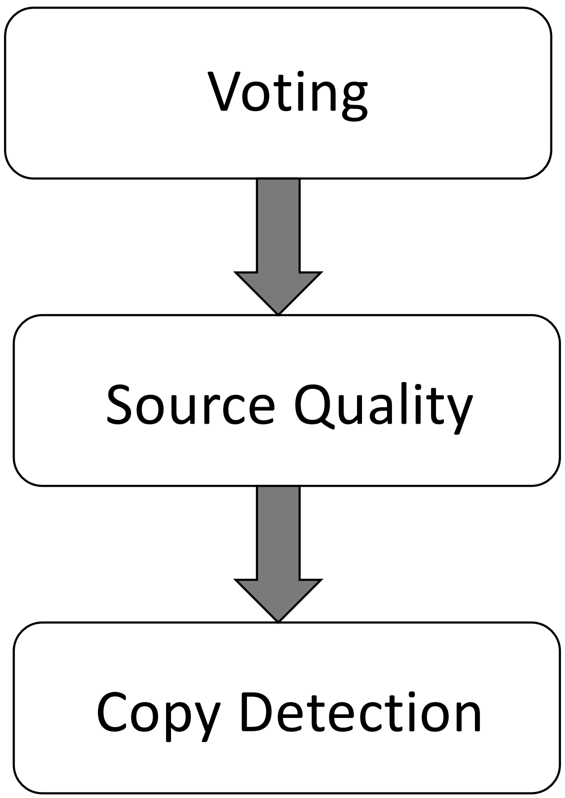
Name	Addr	Income
Samuel	123 1st	50k
M. Jones	348 1st	80k
Mikey	246 2nd	30k
Yuan Yuan	444 2nd	75k

Yields a set of blocks; only compare records in the same block

DATA FUSION: MULTI-SOURCE INTEGRATION

Voting + source quality + copy detection

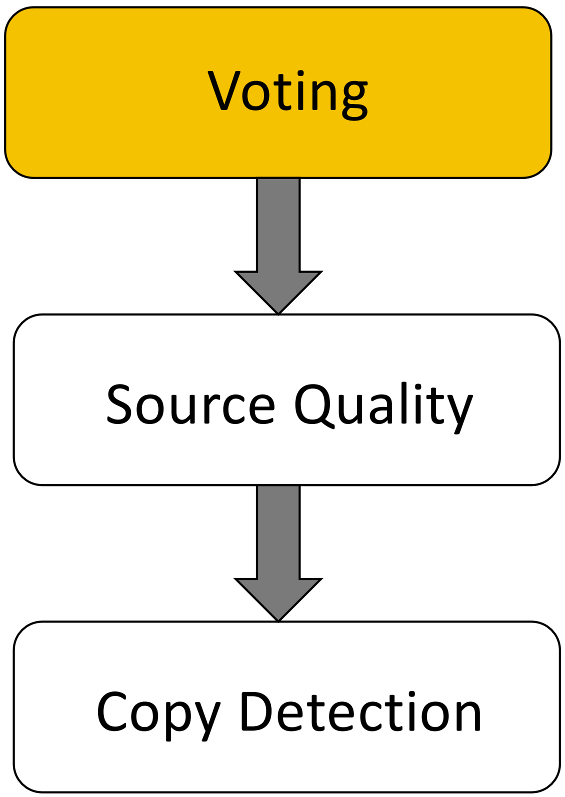
- Resolves inconsistency across diversity of sources



	S1	S2	S3	S4	S5
Jagadish	UM	ATT	UM	UM	UI
Dewitt	MSR	MSR	UW	UW	UW
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	ATT	BEA	BEA	BEA
Franklin	UCB	UCB	UMD	UMD	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

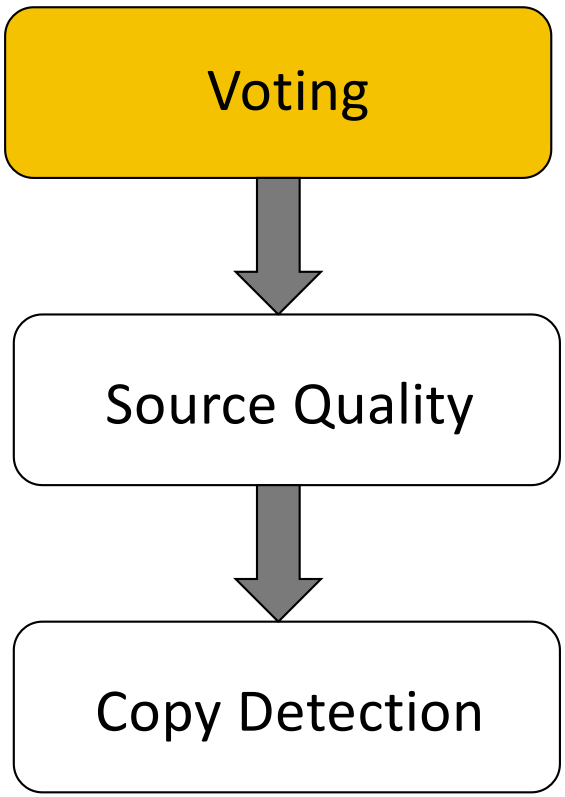


	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

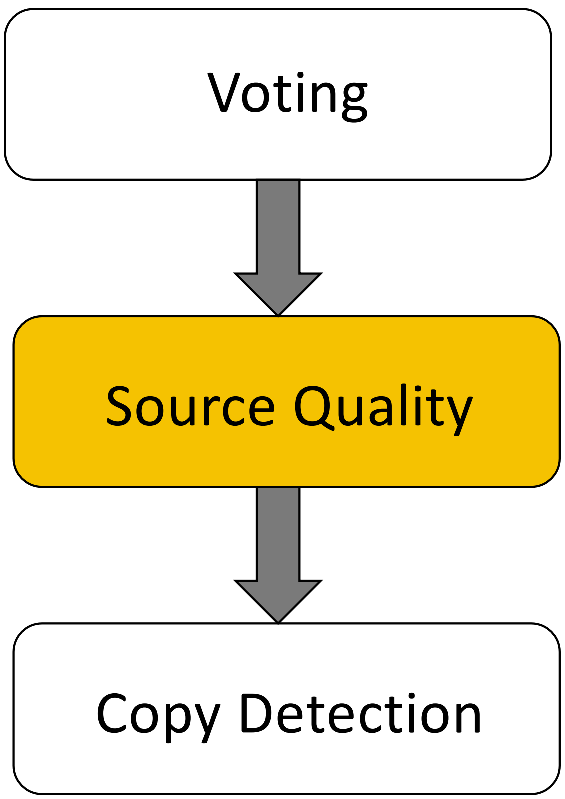
- Supports difference of opinion



	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

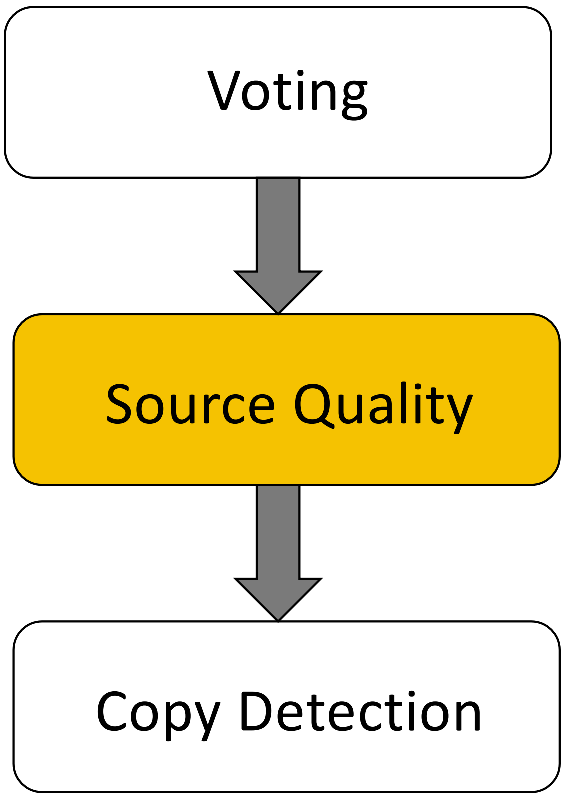


	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

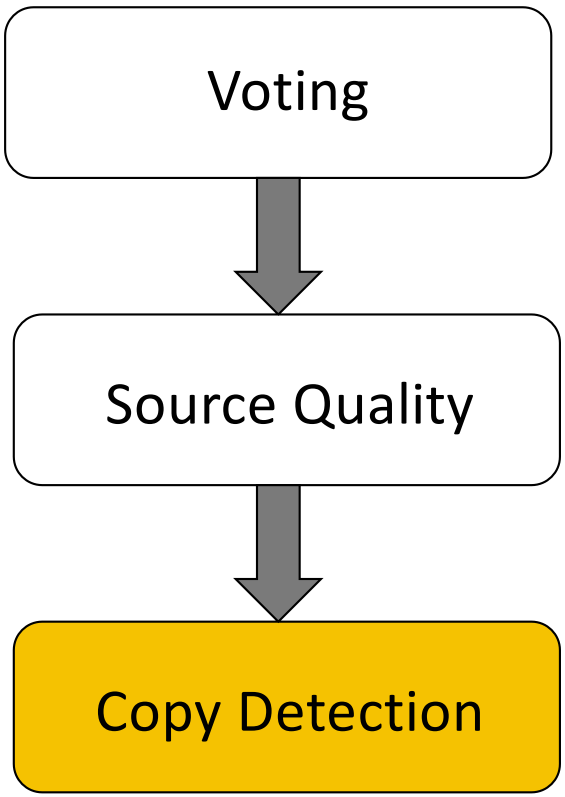
- Gives more weight to knowledgeable sources



	S1	S2	S3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	**UCI**	ATT	BEA
Franklin	UCB	UCB	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

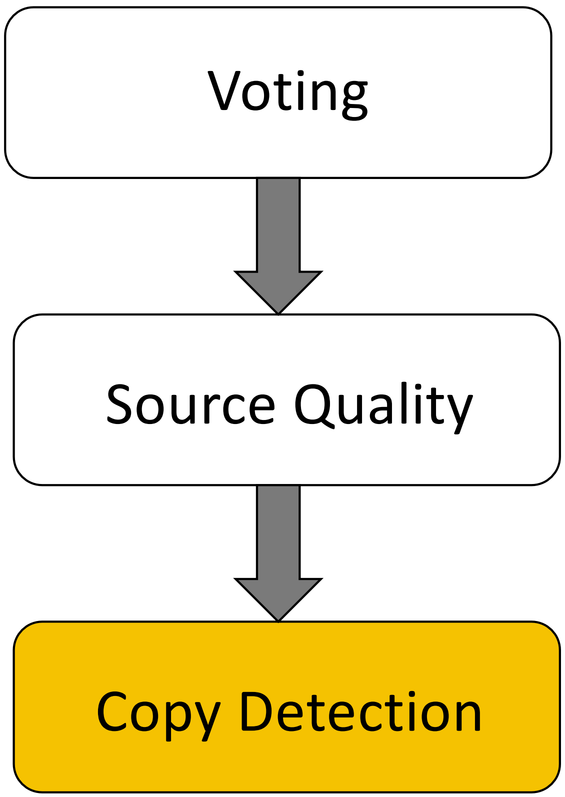


	S1	S2	S3	S4	S5
Jagadish	UM	ATT	UM	UM	UI
Dewitt	MSR	MSR	UW	UW	UW
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	ATT	BEA	BEA	BEA
Franklin	UCB	UCB	UMD	UMD	UMD

DATA FUSION

Data fusion: voting + source quality + copy detection

- Reduces weight of copied sources

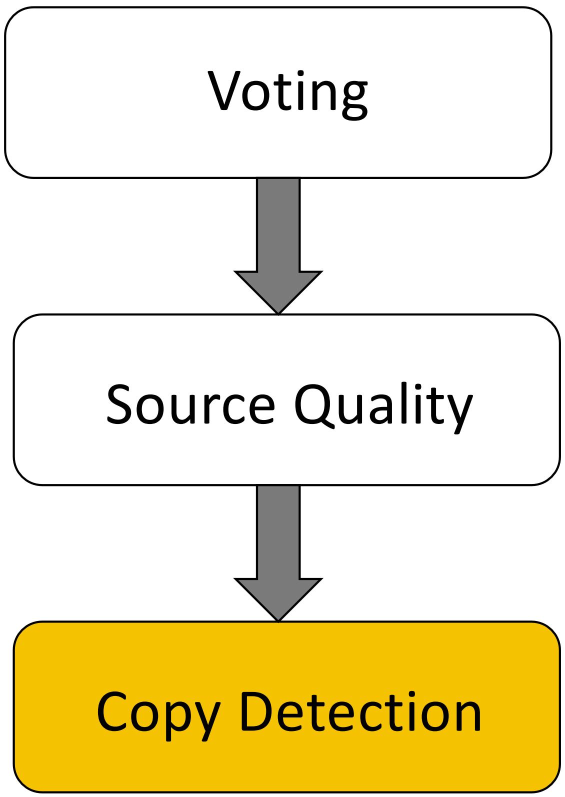


	S1	S2	S3	~~S4~~	~~S5~~
Jagadish	UM	ATT	UM	~~UM~~	~~UI~~
Dewitt	MSR	MSR	UW	~~UW~~	~~UW~~
Bernstein	MSR	MSR	MSR	~~MSR~~	~~MSR~~
Carey	UCI	ATT	BEA	~~BEA~~	~~BEA~~
Franklin	UCB	UCB	UMD	~~UMD~~	~~UMD~~

DATA FUSION

Data fusion: voting + source quality + copy detection

- Reduces weight of copied sources



	S1	S2	S3	~~S4~~	~~S5~~
Jagadish	UM	ATT	UM	~~UM~~	~~UI~~
Dewitt	MSR	MSR	UW	~~UW~~	~~UW~~
Bernstein	MSR	MSR	MSR	~~MSR~~	~~MSR~~
Carey	**UCI**	ATT	BEA	~~BEA~~	~~BEA~~
Franklin	UCB	UCB	UMD	~~UMD~~	~~UMD~~

OUTLINE

Data Integration

- **Different schemas** →Schema matching
- **Duplicates** → Entity resolution
- **Scale**→ Blocking, etc

Data Cleaning

- **Missing values** → Value imputation
- **Missing records** → Species estimation

TYPES OF MISSING VALUES

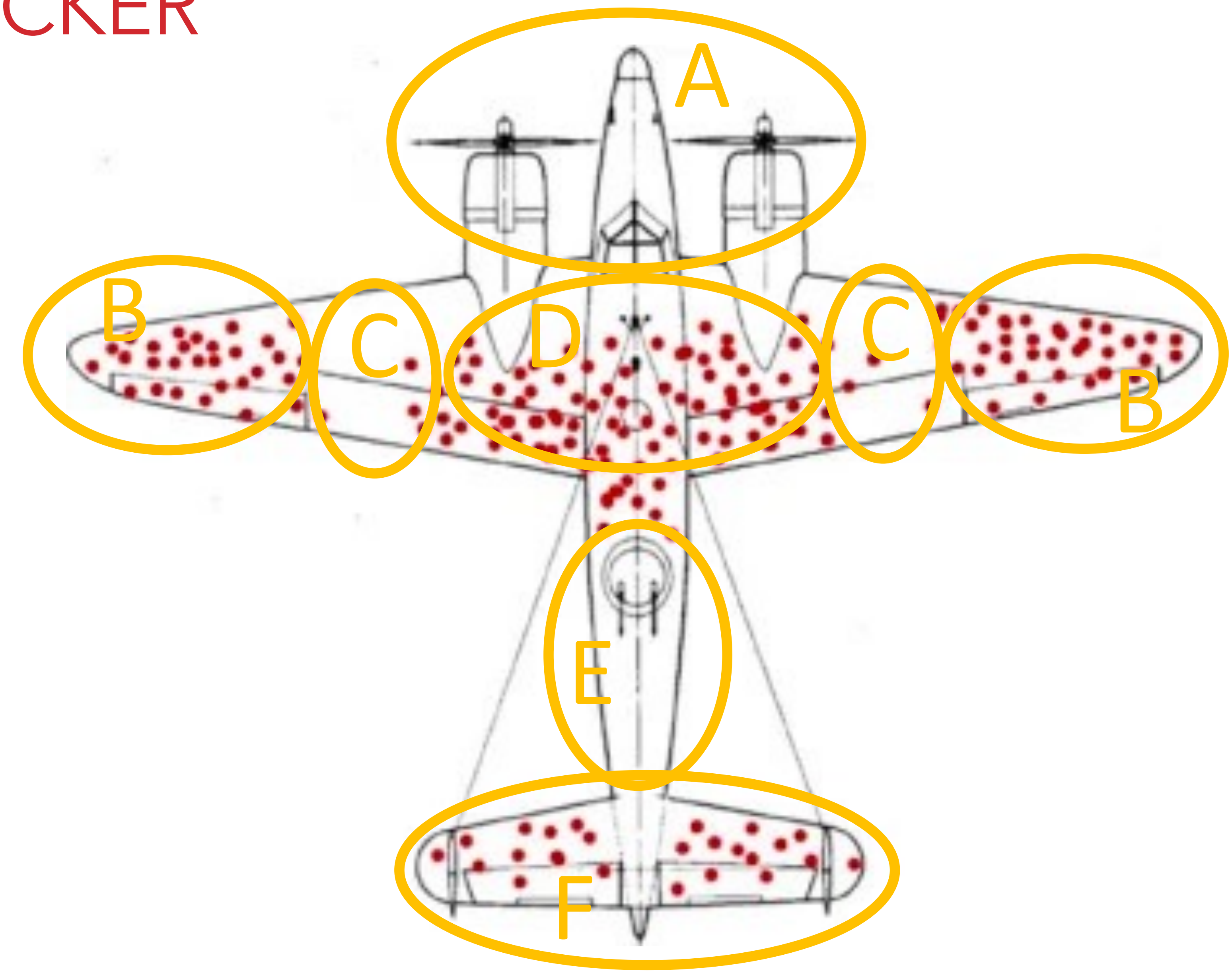
- **Missing Completely at Random (MCAR)**
  - Includes missing by design. For example: Survey randomly selects questions to reduce load

- **Missing at Random (MAR)**
  - Better name: Missing Conditionally at Random
  - Systematic relationship between the propensity of missing values and the *observed* data, but *not* the missing data.
    --> if we can control for this conditional variable, we can get a random subset.
  - Example: older people more likely to respond to telephone survey, thus more data missing from older people

- **Missing Not at Random, MNAR**
  - Relationship between the propensity of a value to be missing and its values.
  - Lowest education are missing on education or the sickest people are most likely to drop out of the study.
  - MNAR is called "non-ignorable" because the missing data mechanism itself has to be modeled as you deal with the missing data.

Note: null values are often encoded in various ways. Be aware of it!
Null, "null", n/a, "", 0, "empty", 99999, 200.

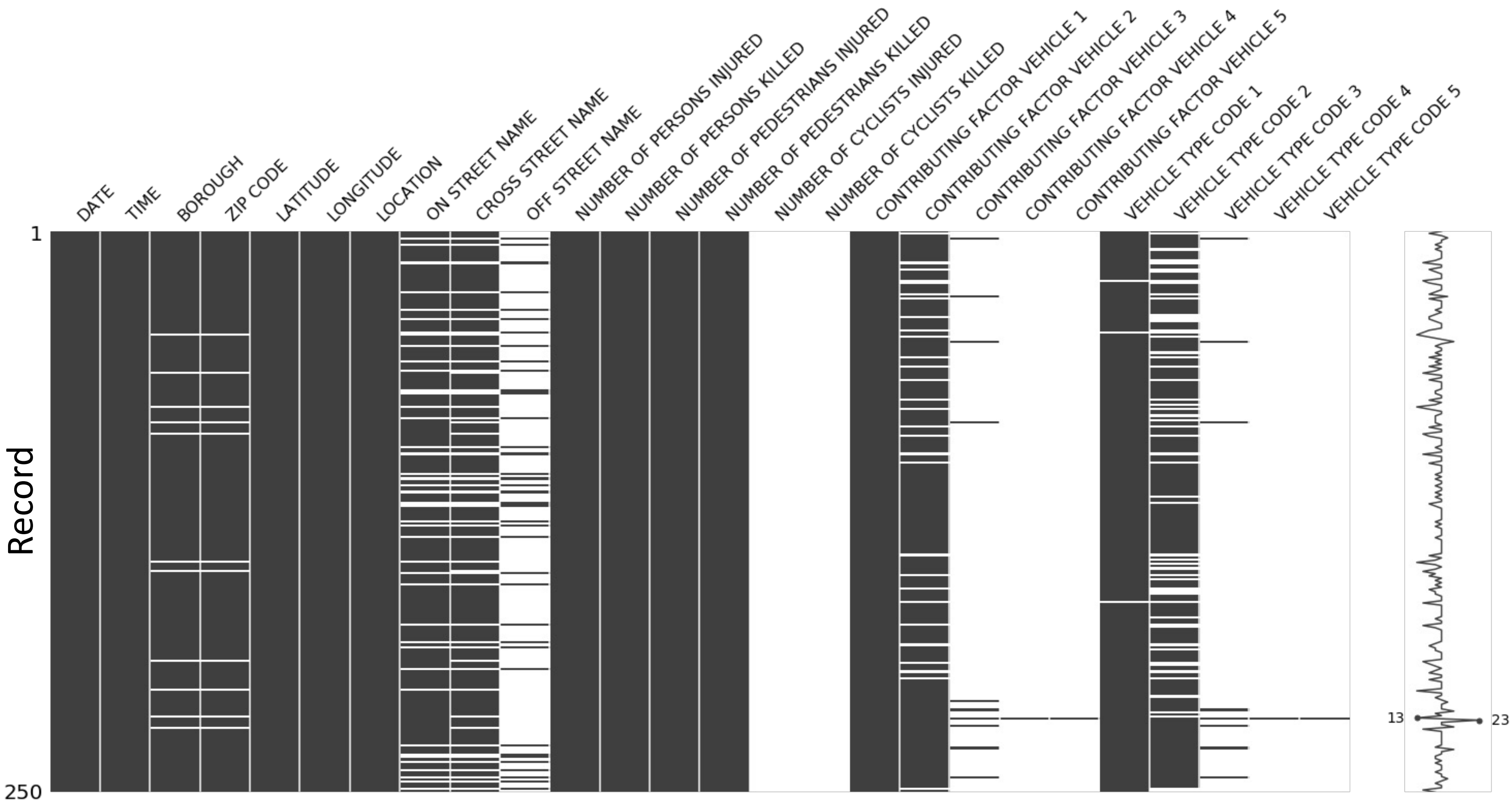
CLICKER



Where would you reinforce the plane?

HOW DO YOU START ADDRESSING MISSING VALUES?

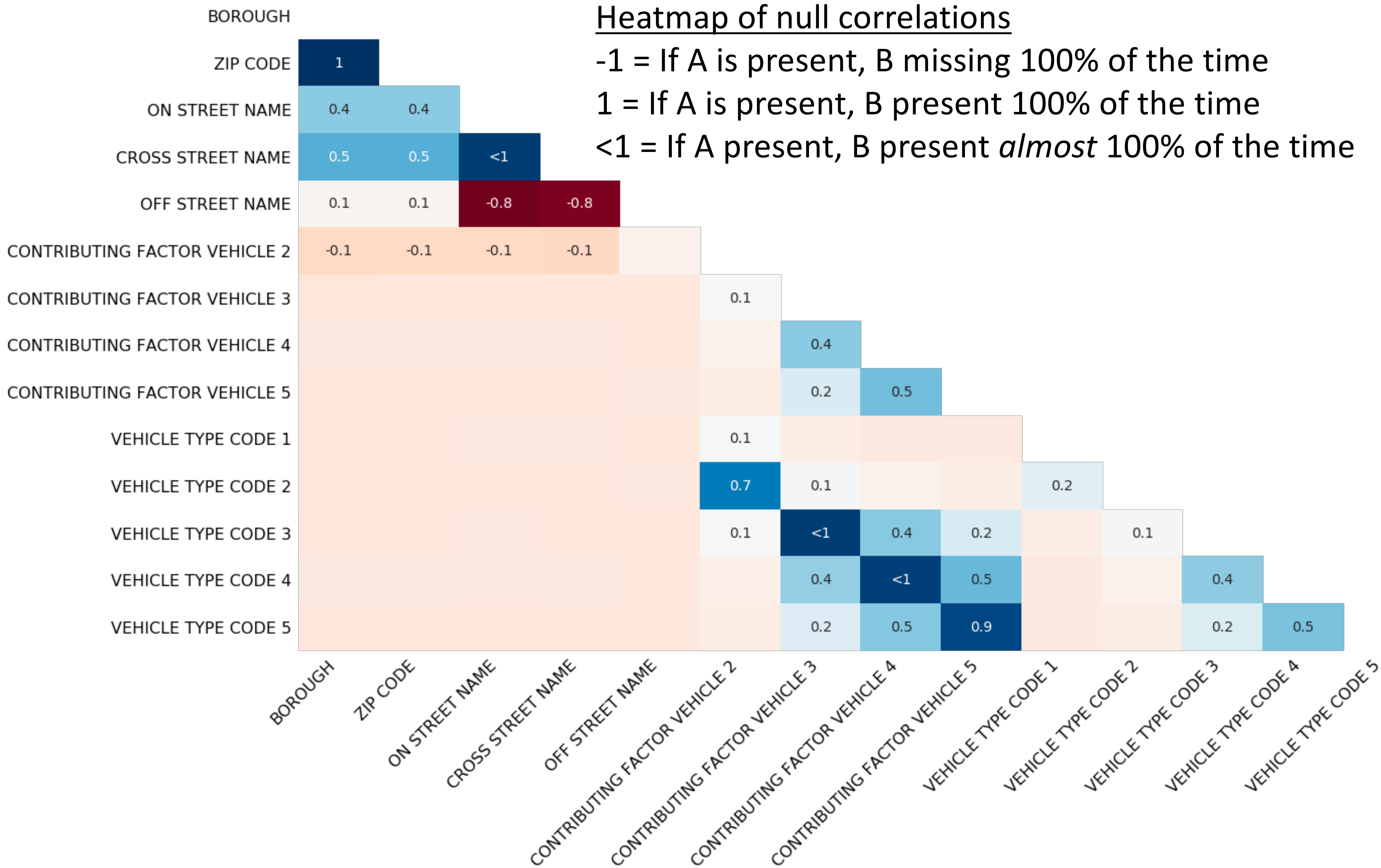
VISUALIZATIONS TO DETECT BIAS IN MISSING DATA



White = missing; black = present

A lot of tips here: https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS IN MISSING DATA



<u>Heatmap of null correlations</u>

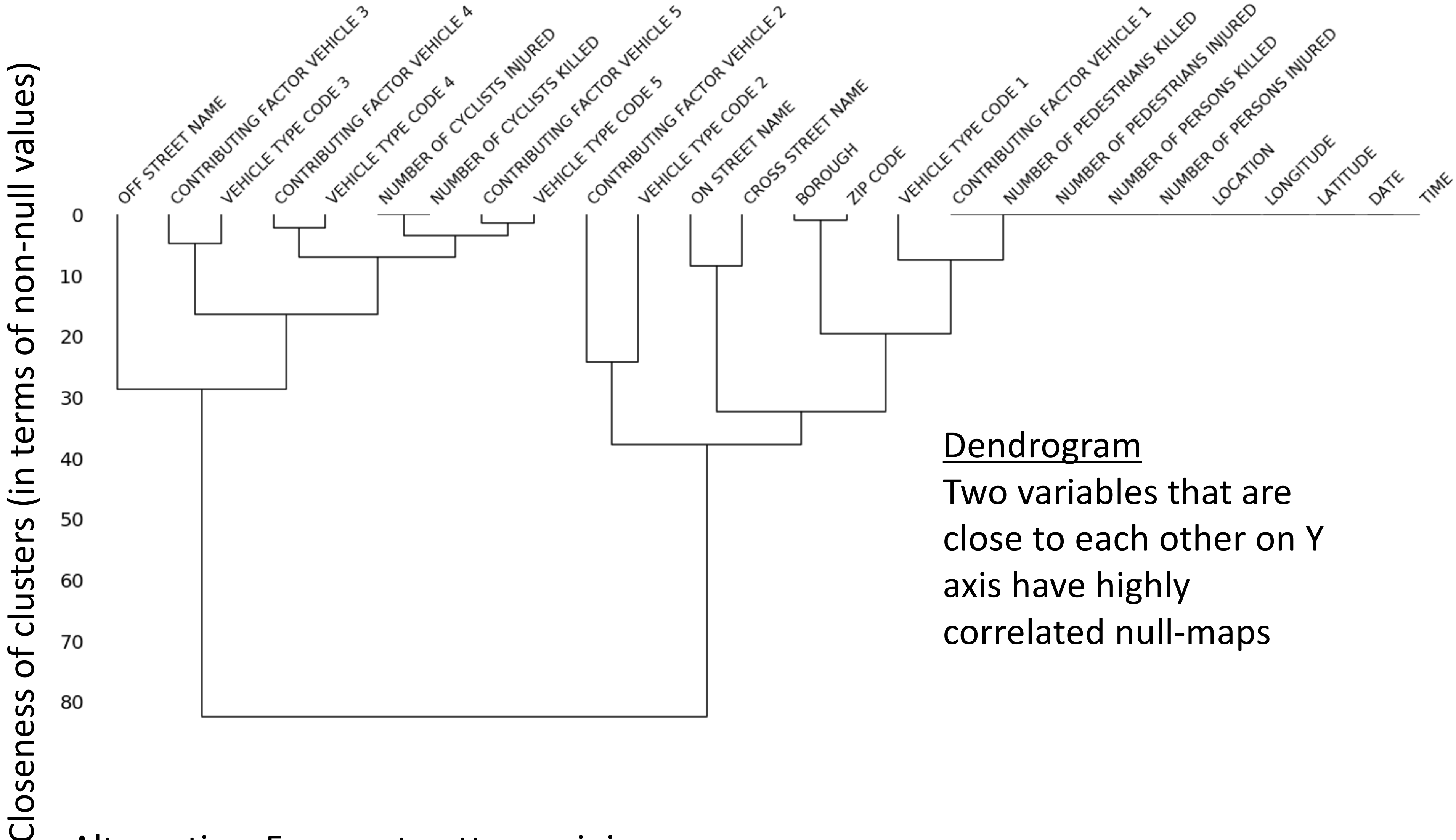
-1 = If A is present, B missing 100% of the time

1 = If A is present, B present 100% of the time

<1 = If A present, B present *almost* 100% of the time

https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS IN MISSING DATA



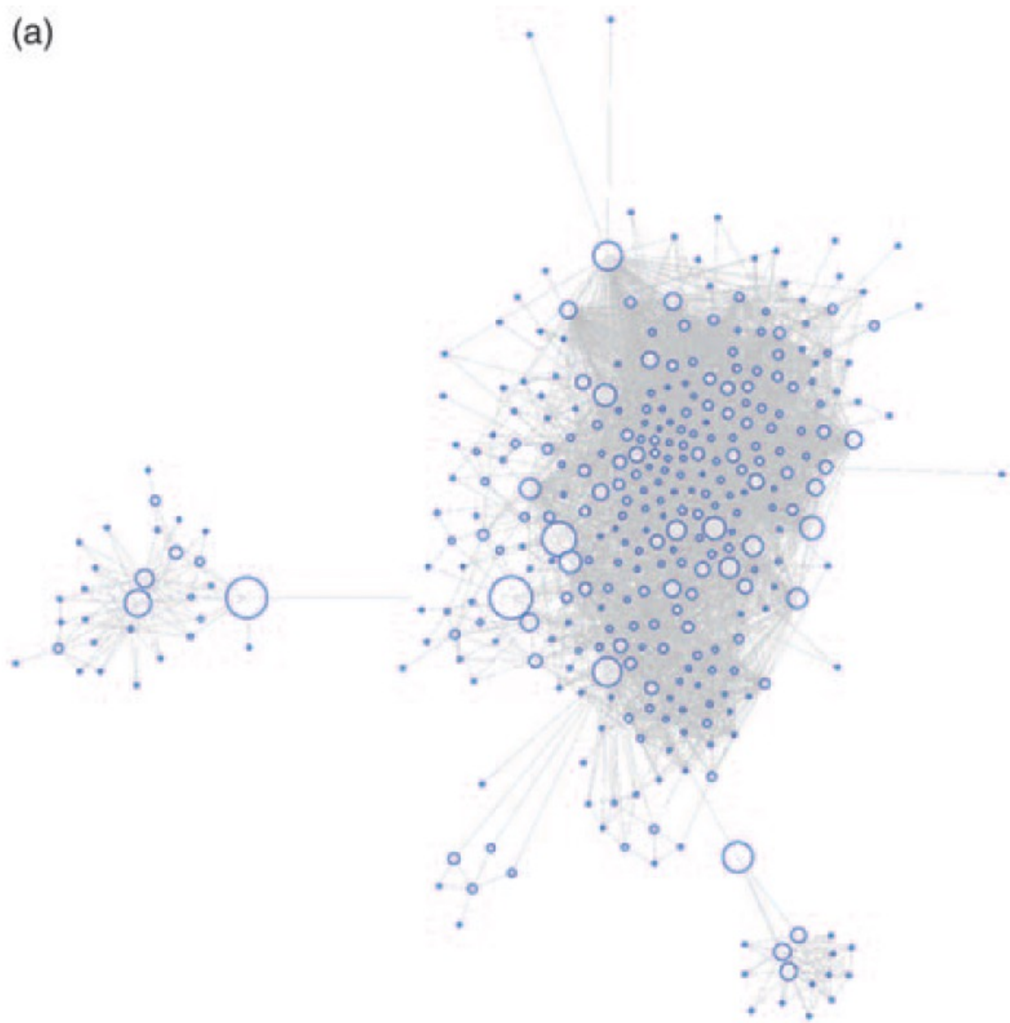
Dendrogram

Two variables that are close to each other on Y axis have highly correlated null-maps

Alternative: Frequent pattern mining

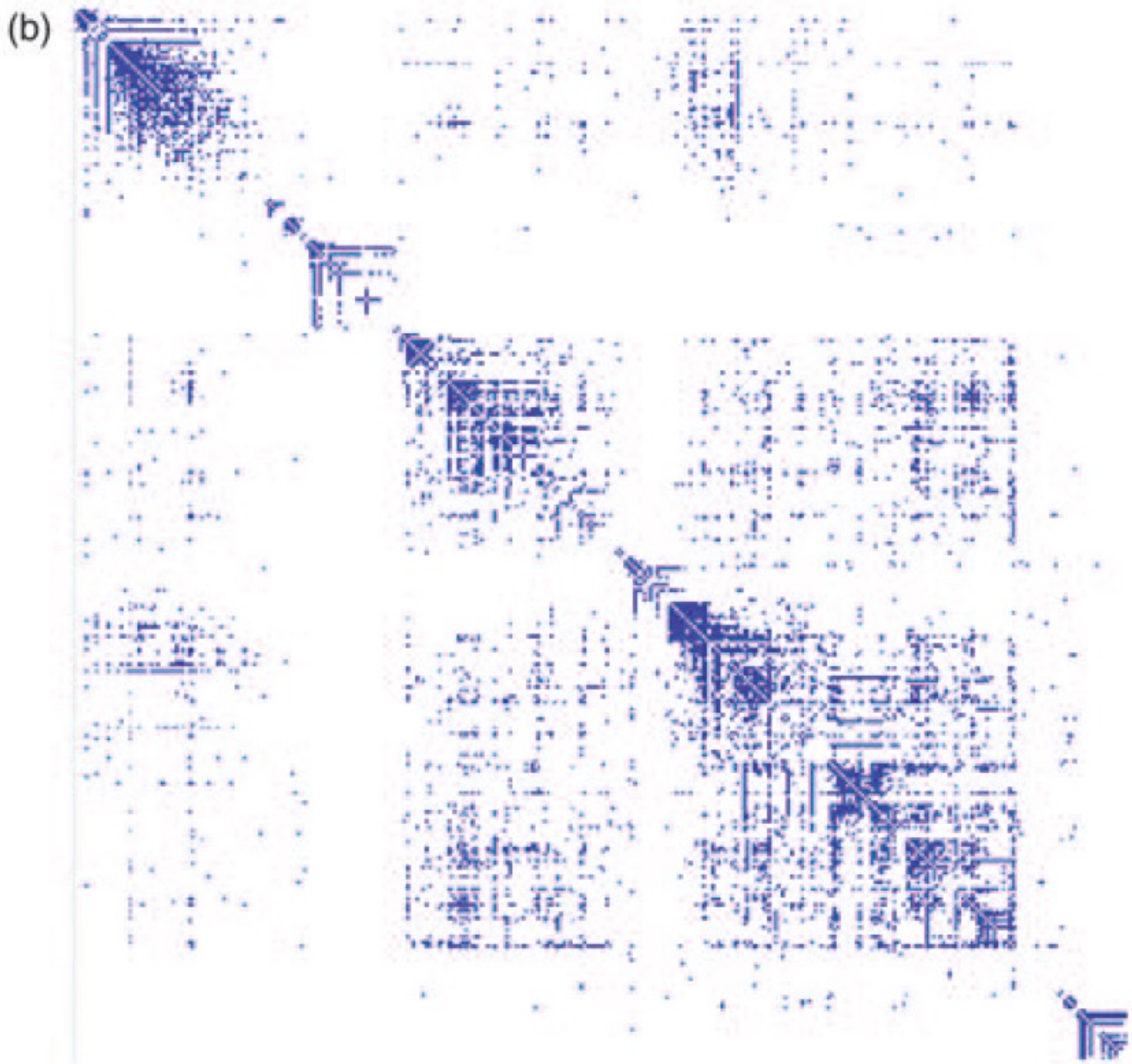
https://github.com/ResidentMario/missingno

FACEBOOK SOCIAL GRAPH: VISUALIZATION THE NODE-LINK DIAGRAM



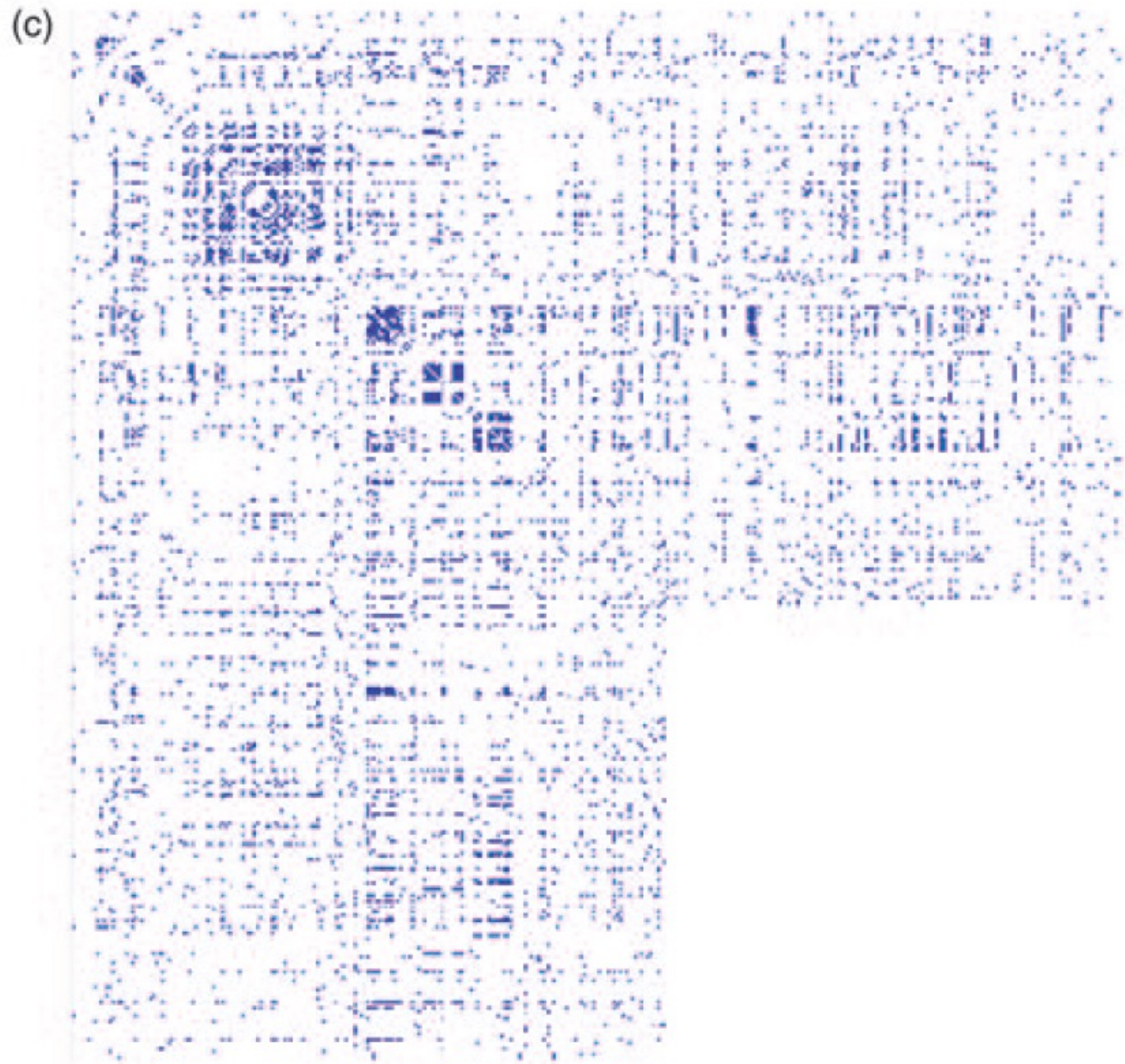
[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

FACEBOOK SOCIAL GRAPH: VISUALIZATION THE NODE-LINK DIAGRAM



[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

FACEBOOK SOCIAL GRAPH: SORTING BY RAW DATA



[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

CLASS TASK:

COME UP WITH AT LEAST 5 TECHNIQUES TO DEAL WITH MISSING VALUES

TECHNIQUES TO DEAL WITH MISSING VALUES (ONLY FOR MCAR / MAR)

Missing Completely at Random

- **Two broad choices**: Drop or Impute
- Drop Methods
  - Pairwise deletion
  - Listwise deletion

- Imputation Methods
  - Mean Substitution
  - Regression Methods
  - Random sample from existing values/ reasonable distribution
  - Multiple Imputation

PAIRWISE AND LISTWISE DELETION

```
SELECT SUM(revenue)/
SUM(employees) FROM
us_tech_companies
```

Pairwise Deletion

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

PAIRWISE AND LISTWISE DELETION

```
SELECT SUM(revenue)/
SUM(employees) FROM
us_tech_companies
```

Pairwise Deletion

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

Listwise Deletion

Name	Address	#Employees	Revenue	Profit
~~Google~~	~~1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA~~	~~60k~~	~~$89B~~	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

PAIRWISE AND LISTWISE DELETION

Pairwise Deletion

- Only cases relating to each pair of variables with missing data involved in an analysis are deleted.
- Advantage: keeps as many cases as possible for each analysis, uses all information possible with each analysis
- Disadvantage: cannot compare analyses because sample is different each time, sample size vary for each parameter estimation, can obtain nonsense results

Listwise Deletion

- Only analyze cases with available data on each variable
- Advantage: simplicity and comparability across analyses
- Disadvantage: reduces statistical power (reduced sample size), some information unused, estimates may be biased if data not MCAR

INITIAL CLEANING

Look for fields with very high percentage of missing fields

- It may be necessary to exclude field and use an alternative

Look for records with a high percentage of missing fields

- Consider excluding these
- For example, someone who has started inputting a survey and given up after two questions!

Document deletions!

UNIVARIATE SINGLE IMPUTATION
MEAN SUBSTITUTION

Mean Substitution

- Replace missing value with the sample mean or mode. Then, run analyses as if all complete cases

UNIVARIATE SINGLE IMPUTATION
MEAN SUBSTITUTION

Mean Substitution (do not use)

- Replace missing value with the sample mean or mode. Then, run analyses as if data is complete
- Advantage: Simple, no missing data
- Disadvantage: Reduces variability, weakens correlations, biases data
- Unless the proportion of missing data is low, do not use this method.
- Inappropriate for categorical variables.

https://scikit-learn.org/stable/modules/impute.html

SIMPLE STOCHASTIC IMPUTATION

Randomly sample from existing values:

- Randomly generate an integer from 1 to num. non-missing

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States		$5M	$8M

SIMPLE STOCHASTIC IMPUTATION

Randomly sample from existing values:

- Randomly generate an integer from 1 to num. non-missing
- E.g., Randomly generate number between 1 and 4: Say 2 → Set Tableau employees to Apple Employees (66k)

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	66k	$5M	$8M

SIMPLE STOCHASTIC IMPUTATION

Randomly sample from existing values:

- Randomly generate an integer from 1 to num. non-missing
- E.g., Randomly generate number between 1 and 4: Say 2 → Set Tableau employees to Apple Employees (66k)

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	66k	$5M	$8M

Disadvantage: May be very wrong for certain values

Hot-deck approach: draws are made from units with complete data that are 'similar' to the one with missing values (donors).

MULTIVARIATE IMPUTATION

Regression imputation

- Replace missing values with predicted score from regression equation. Use complete cases to regress the variable with incomplete data on the other complete variables.

https://scikit-learn.org/stable/modules/impute.html

MULTIVARIATE IMPUTATION

Regression imputation

- Replace missing values with predicted score from regression equation. Use complete cases to regress the variable with incomplete data on the other complete variables.
- Uses information from the observed data, gives better results than previous ones
- Emphasizes correlations present in the available data

Other models, e.g., maximum likelihood estimation, are possible (but we won't cover them)

https://scikit-learn.org/stable/modules/impute.html

DEMO

OTHER METHODS

Nearest-neighbors imputation

- KNN defines for each sample or individual a set of K-nearest neighbors and then replaces the missing data for a given variable by averaging (non-missing) values of its neighbors
- Advantage: Simple, uses information from the observed data, experimentally shows good performance
- Disadvantage: not statistically grounded, might over-estimates model fit and correlation

EM (Expectation Maximization)

Fuzzy K-means Clustering

Bayesian Principal Component Analysis

Deep Learning-based approaches

….

https://scikit-learn.org/stable/modules/impute.html

MULTIPLE IMPUTATION (MI)

Multiple imputation (MI) is a common method for general- purpose handling of missing data in multivariate analysis.

1. Impute missing values using an appropriate model that incorporates random variation.
2. Do this M times producing M "complete" data sets.
3. Perform the desired analysis on each data set using standard complete-data methods.
4. Average the values of the parameter estimates across the M samples to produce a single point estimate.
5. Calculate the standard errors by (a) averaging the squared standard errors of the M estimates (b) calculating the variance of the M parameter estimates across samples, and (c) combining the two quantities using a simple formula

OUTLINE

Data Integration

- **Different schemas** → Schema matching
- **Duplicates** → Entity resolution
- **Scale** → Blocking, etc

Data Cleaning

- **Missing values** → Value imputation
- **Missing records** → Species estimation

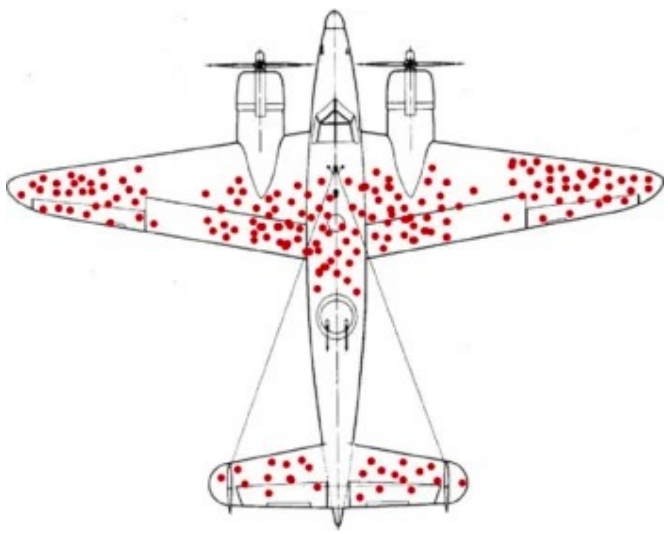
UNKNOWN UNKNOWNS

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway,	60k	$89B	null
Apple				45B
IBM				12B
Internationa Business M				12B
Microsoft				35B
Tableau				1B
Tamr	02138, United States			-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also **unknown unknowns**—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones."

Donald Rumsfeld (Defense Secretary, US, 2001-2006)

IF YOU CAN ESTIMATE THEM DEPENDS ON THE SAMPLING SCENARIO



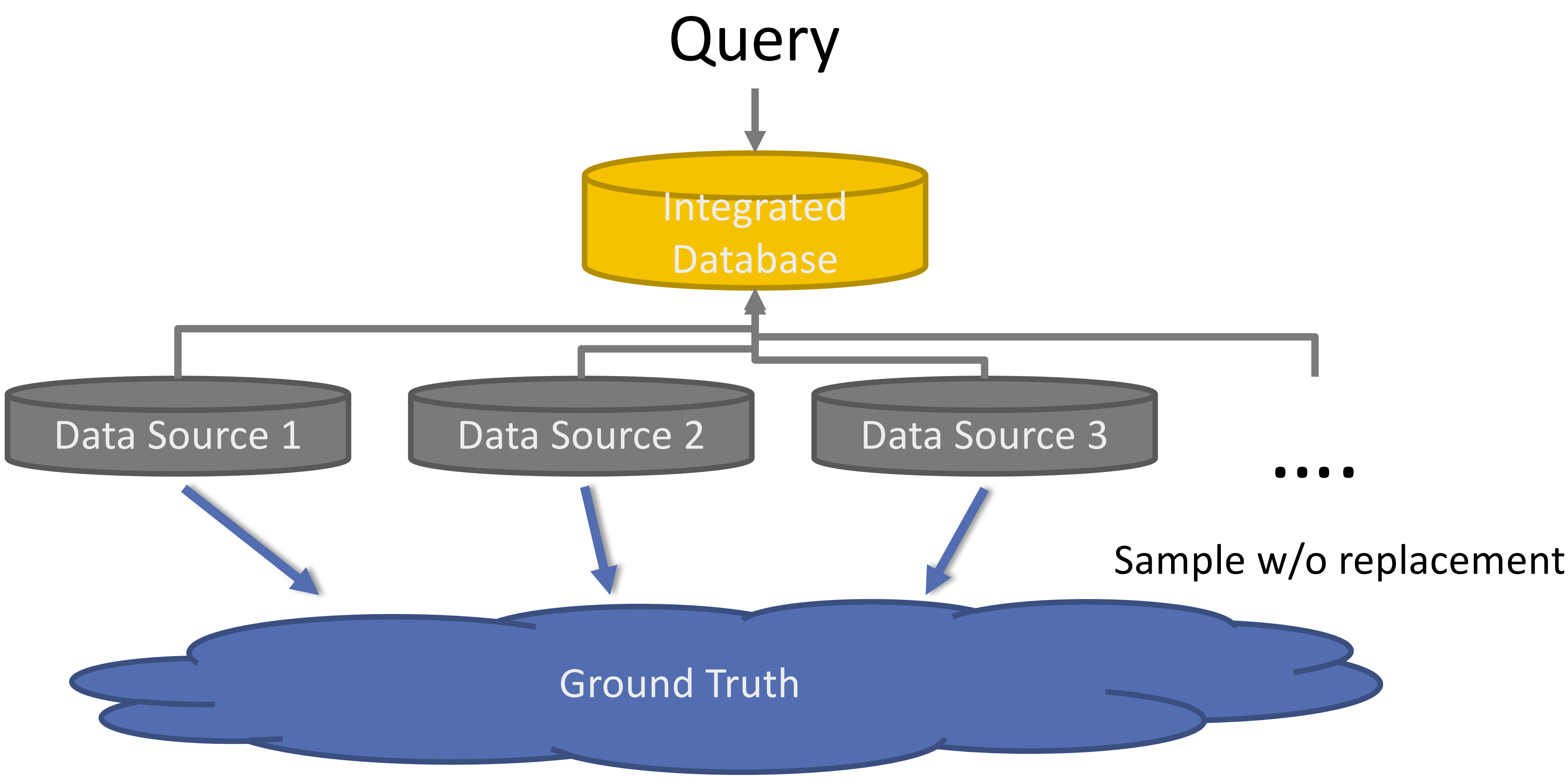
VS

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	1 New Orchard Rd; 10504	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

THE IMPACT OF THE **UNKNOWN UNKNOWNS** ON QUERY RESULTS

How many people work in the US IT industry

```
SELECT SUM(employees)
FROM us_tech_companies
```



Assumption: Enough data sources , Data sources are (semi-) independent

Sampling - Statistic

$$\sum$$

Name	Address	#Employees	Revenue	Profit	Frequency
Google	Address I	60k	$89B	$10B	3
Apple	Address II	66k	$215B	$45B	4
IBM	Address II	380k	$80B	$12B	2
Microsoft	Address	120k	$85B	$85B	2
Tableau	Address	3.2k	$500	$8M	1
Tamr	Address	20	$-X	$-Y	1

Frequency (i.e., f-statistic):

f_1: 2

f_2: 2

f_4: 1

f_3: 1

Singletons (items which were exactly observed once)

c = 6 unique companies

N = 3 + 4 + 2 + 2 + 1 + 1 = 13 observations

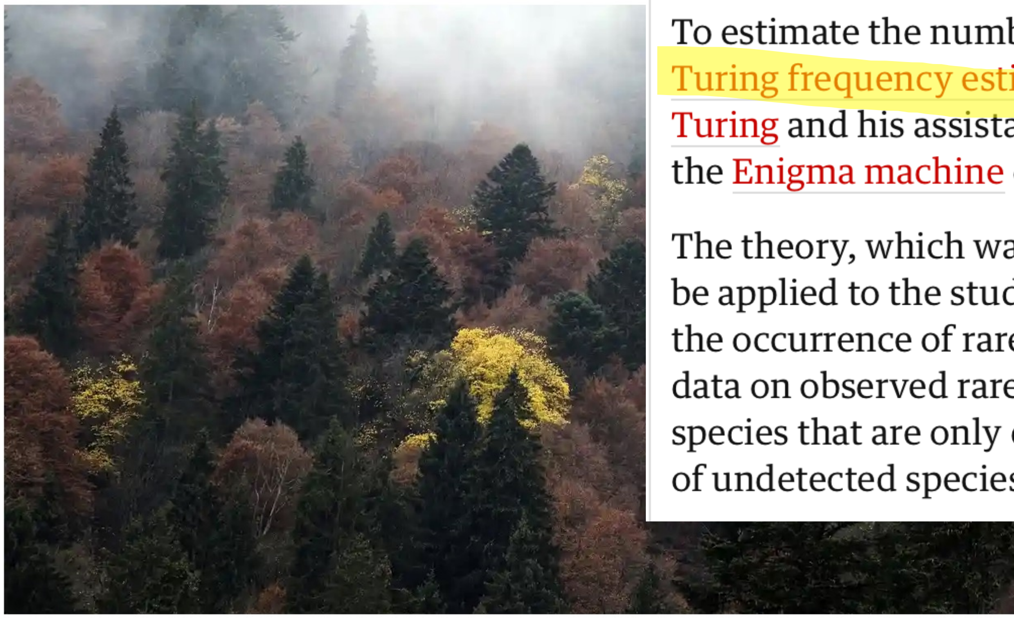
MANY WAYS TO ESTIMATE THE NUMBER OF MISSING ITEMS

- Good-Turing Estimate / Chao84
- Chao92
- Pattern Maximum Likelihood
- Linear programming-based solutions (see Valiant brothers)
- …

ESTIMATING THE NUMBER OF DISTINCT BUTTERFLY SPECIES

Global count estimates Earth has 73,000 tree species – 14% more than reported

Second world war codebreaking calculations used at Bletchley Park find 9,000 of those species are yet to be discovered



Researchers collected information on 38m trees in 90 countries as part of the global count. Photograph: Global Forest Biodiversity Initiative)

To estimate the number of unknown species, scientists used the Good-Turing frequency estimation, which was created by the codebreaker Alan Turing and his assistant Irving Good when trying to crack German codes for the Enigma machine during the second world war.

The theory, which was developed by the Taiwanese statistician Anne Chao to be applied to the study of undetected species, helped researchers work out the occurrence of rare events - in this case unknown species of trees - using data on observed rare species. Essentially, the code uses information on species that are only detected once or twice in data to estimate the number of undetected species.

There are an estimated 73,300 species of tree on Earth, 9,000 of which have yet to be discovered, according to a global count of tree species by thousands of researchers who used second world war codebreaking techniques created at Bletchley Park to evaluate the number of unknown species.

https://www.theguardian.com/environment/2022/jan/31/global-count-estimates-earth-has-73000-tree-species-bletchley-park-good-turing-frequency-estimation

17500 **species** known in the world

GOOD-TURING / CHAO84 ESTIMATE

Unique Items

$$\hat{N} = \frac{\boxed{c}}{\left(1 - \boxed{{f_1}/{n}}\right)}$$

Missing mass

Number of Unknown Unknowns:

$$M = \hat{N} - c$$

*Note, we usually prefer **Chao92**: A. Chao and S. Lee, "**Estimating the Number of Classes via Sample Coverage**," Journal of the American Statistical Association, vol. 87, no. 417, pp. 210–217, 1992*
*over **Chao84**: A. Chao, "**Nonparametric Estimation of the Number of Classes in a Population**," SJS, vol. 11, no. 4, 1984*

A NAÏVE ESTIMATOR FOR THE IMPACT OF THE UNKNOWN UNKNOWNS

```
SELECT SUM(employees)
FROM us_tech_companies
```

$$\sum employees\ ,\ \Delta(employees, fingerprint)$$

$$\Delta_{Naive} = M \cdot \emptyset$$

Estimate of Unknown Unknowns Count

Average Value of Knowns

(aka mean substitution)

A NAÏVE ESTIMATOR FOR THE IMPACT OF THE UNKNOWN UNKNOWNS

Number of unique records i.e., count(*)

Value sum over all unique items

$$\Delta_{Naive} = \frac{c}{\left(1 - {f_1}/{n}\right)} \cdot \frac{\sum_{\{c\}} v}{c}$$

Estimated number of missing records

Mean value

EXAMPLE

n = 13

c = 6

f_1=2

$$\widehat{N} = \frac{c}{\left(1 - {f_1}/{n}\right)} = 6/(1-2/13) = 7.09$$

$$\Delta_{Naive} = \frac{c}{\left(1 - {f_1}/{n}\right)} \cdot \frac{\sum_{\{c\}} v}{c}$$

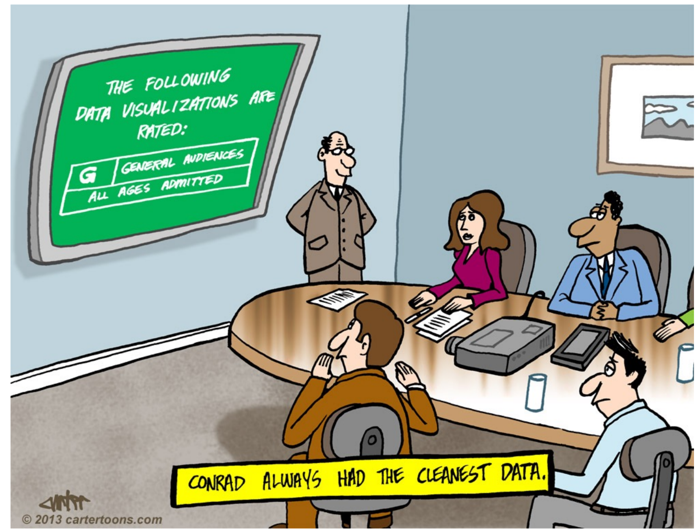
Naïve estimated revenue = 7.09 * $469B/6 = $554B

Name	Address	#Employees	Revenue	Profit	Frequency
Google	Address I	60k	$89B	$10B	3
Apple	Address II	66k	$215B	$45B	4
IBM	Address II	380k	$80B	$12B	2
Microsoft	Address	120k	$85B	$85B	2
Tableau	Address	3.2k	$500	$8M	1
Tamr	Address	20	$-X	$-Y	1

SUMMARY

Quick survey of data cleaning techniques

- Schema Matching
- Entity Resolution
- Dealing with missing values
  - Imputation
  - Species Estimation



Things we haven't touched on

- Detecting & repairing violations
- Outlier detection
- Data evolution and temporal linkage (i.e., data changes)