
CS 541: Advanced Data Management

Dong Deng

Self Introduction



Dong Deng

- Assistant Professor at Rutgers University
- Senior Scientist at Inception Institute of AI
- Postdoc from MIT Database Group
- Ph.D. from Tsinghua University

10+ years of research
experience in the **data
management** field

Course Object

- Introducing students the cutting-edge research on data management and data curation (preparation)
- Training students to master basic skills for being a researcher

Course Setup

- Grading
 - Paper Review: 25% (write review for 1 paper each week)
 - Paper Presentation: 25% (35mins + 15mins QA, 1 or 2 presentations depends on enrollment)
 - Paper Discussion: 10% (ask at least 10 questions during paper discussions)
 - Programming Assignments: $2 \times 10\% = 20\%$
 - Exam: 20% (based on contents in the lectures)
 - TA
 - Yanshi Luo yanshi.luo@rutgers.edu
-

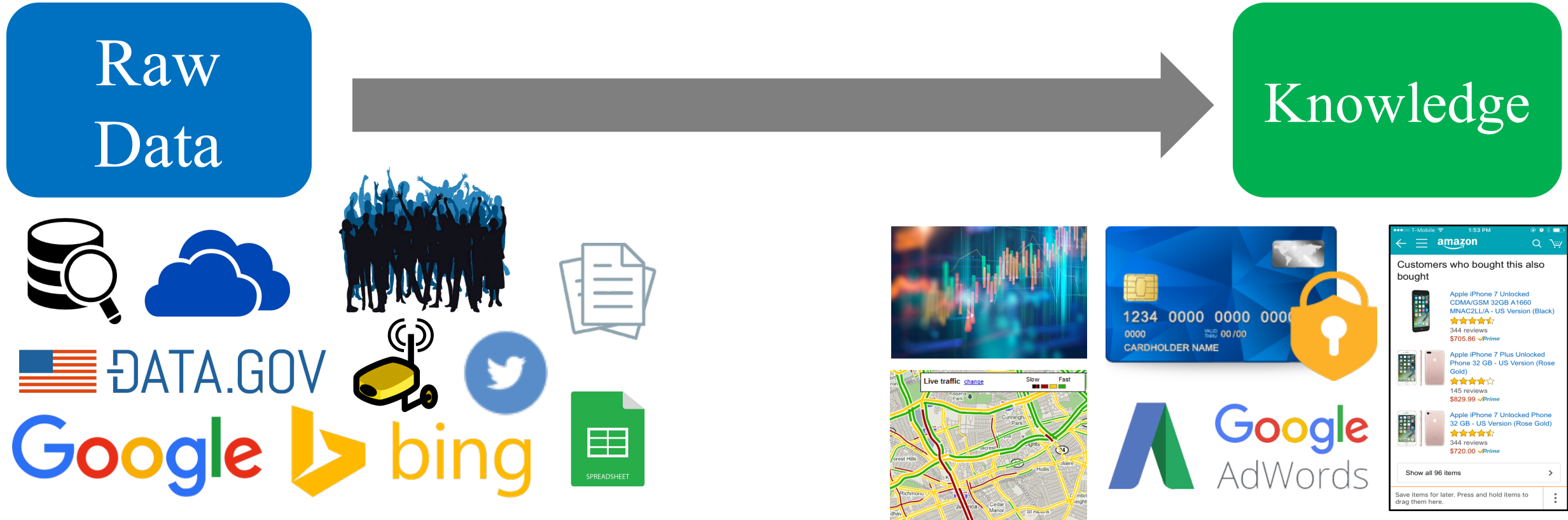
Policy

- Don't be Late
 - Everyone has a budget of 2 days to be used on assignments
 - Once it is used up, 20% penalty per day for each late day
- Don't Cheat
 - We will do plagiarism check
 - If you got caught, you will fail this course

If you are struggling, let us know!

Course Overview

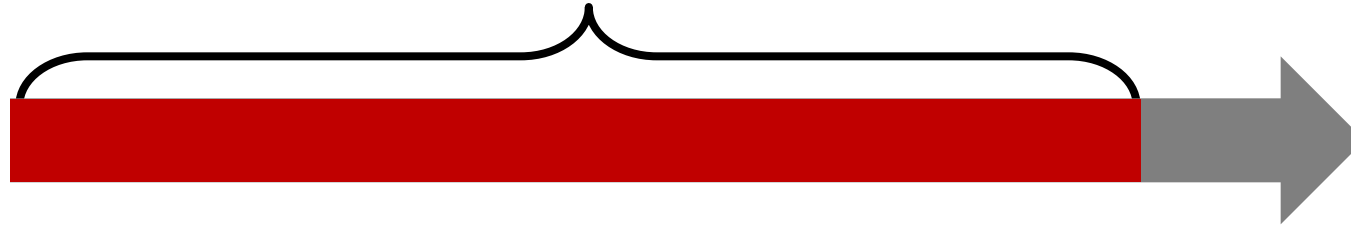
Data Science



Data Curation (Preparation)

Raw
Data

~ 80%



Knowledge

Fort

“Data pre]
about 80
scientists”

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Email

Share

Tweet

Save

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

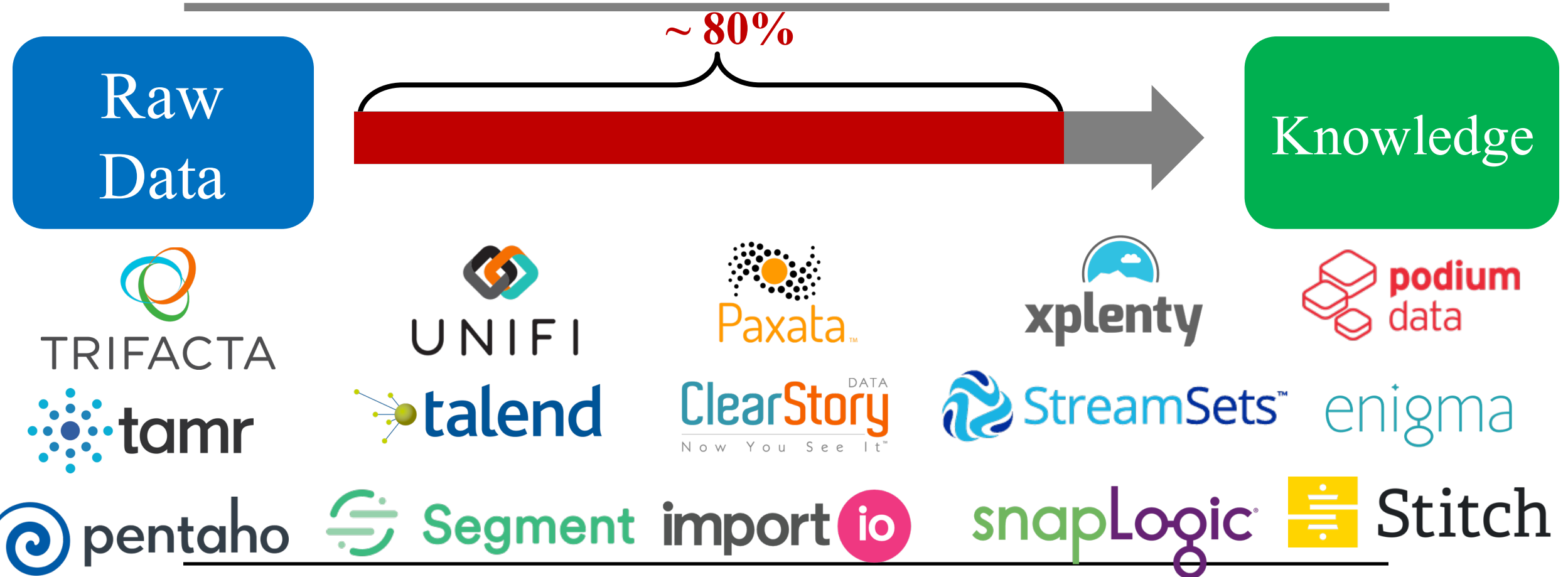
The field known as “big data” offers a contemporary case study. The catchphrase



imes

0% of their
unruly digital

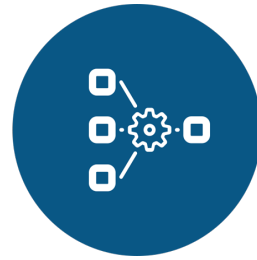
Data Curation (Preparation)



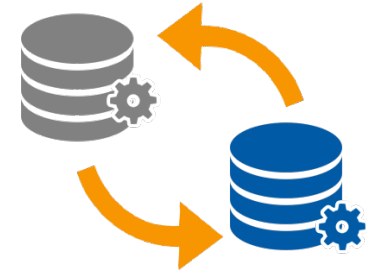
The Landscape of Data Curation



Data Cleaning



Data Integration



Data Transformation



Data Discovery



Data Wrangling
and many more ...



Data Standardization

Data Curation Example

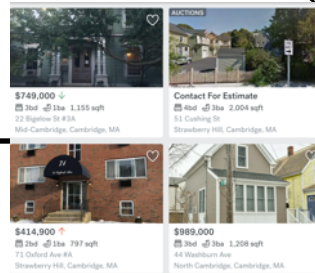
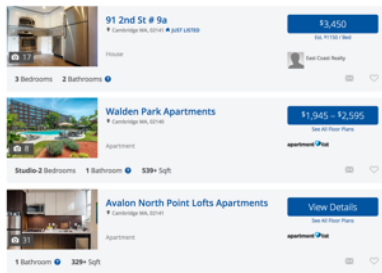


new faculty member

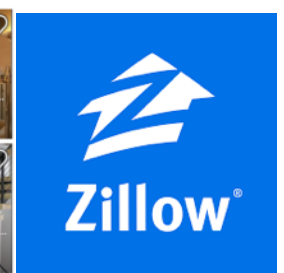
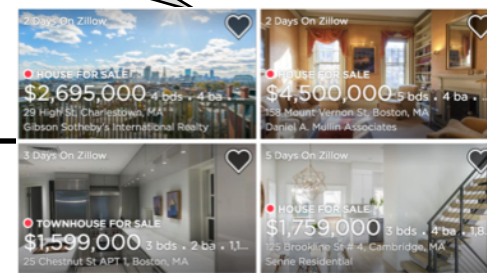


I'd like to go through all listings to make a good buy

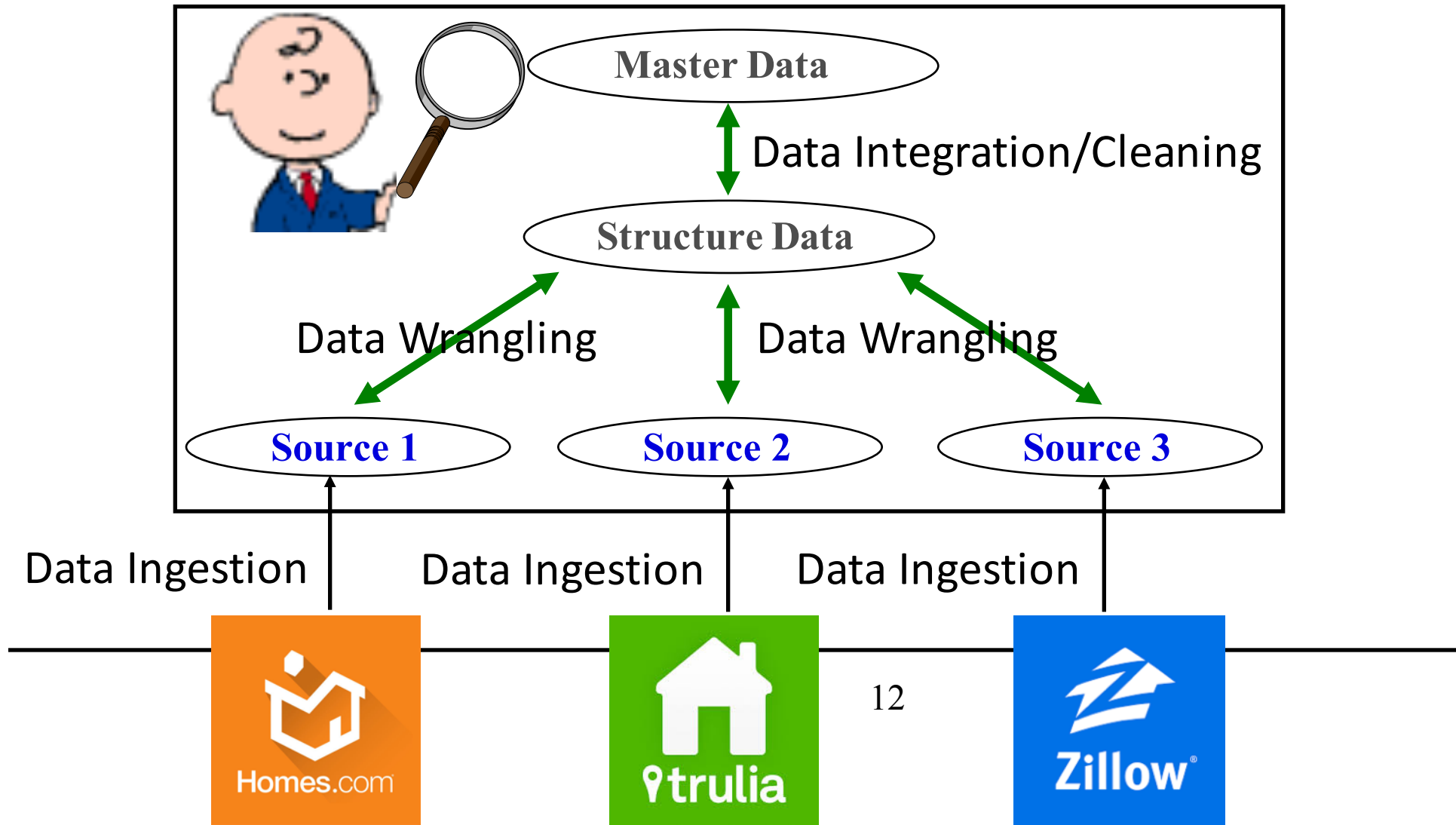
data discovery



11

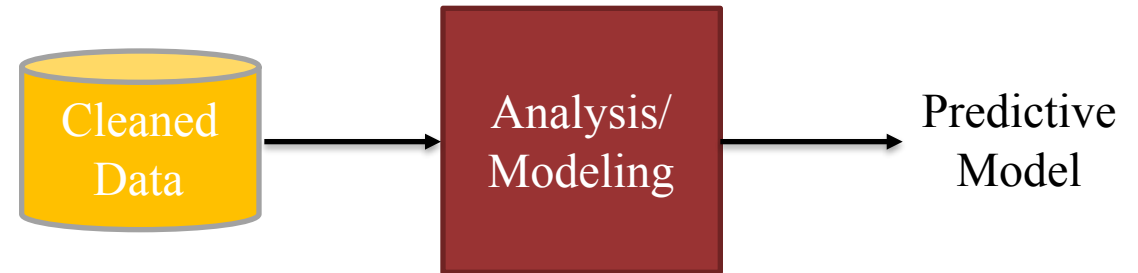


Data Curation Example

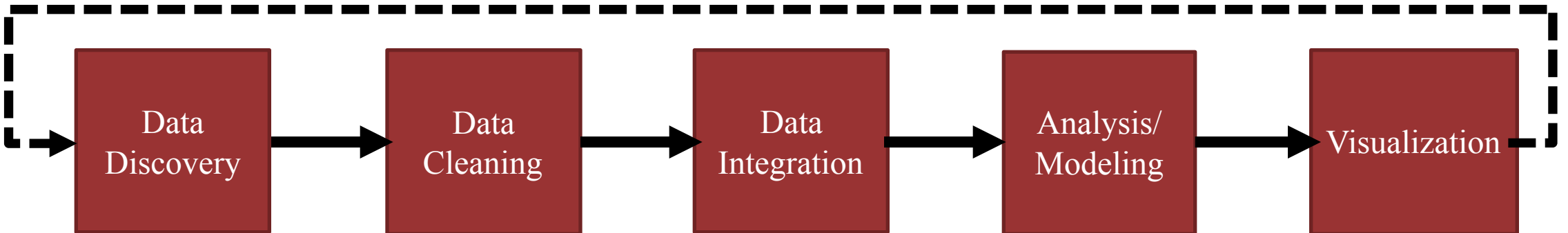


Data Processing Pipeline

- What you think you do?



- What you really do?



Course Topics

- Data Discovery
- Data Wrangling
- Data Transformation
- Data Standardization
- Data Cleaning
- Data Integration
- Data Visualization

They are all interrelated

Data Discovery

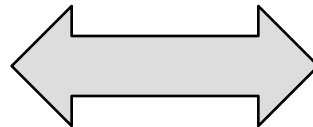
- Merck has approximated 4000 Oracle databases, a large data lake, and uncountable individual files. Data scientists spend >90% of their time finding datasets relevant to their task at hand.
- Many data lakes:
 - US Government open data: www.data.gov
 - Climate Data Library: <http://iridl.ldeo.columbia.edu/index.html>

Data Wrangling

- The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Bureau of I.A.	
Regional Director	Numbers
Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

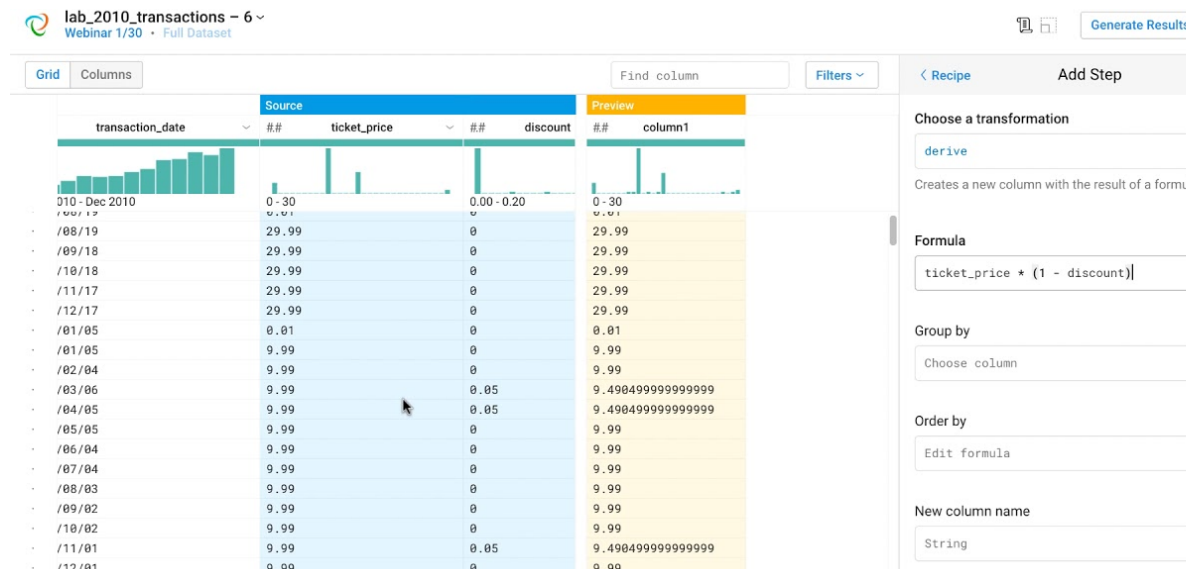
...



	Tel	Fax
Niles C.	(800)645-8397	(907)586-7252
Jean H.	(918)781-4600	(918)781-4604
Frank K.	(615)564-6500	(615)564-6701

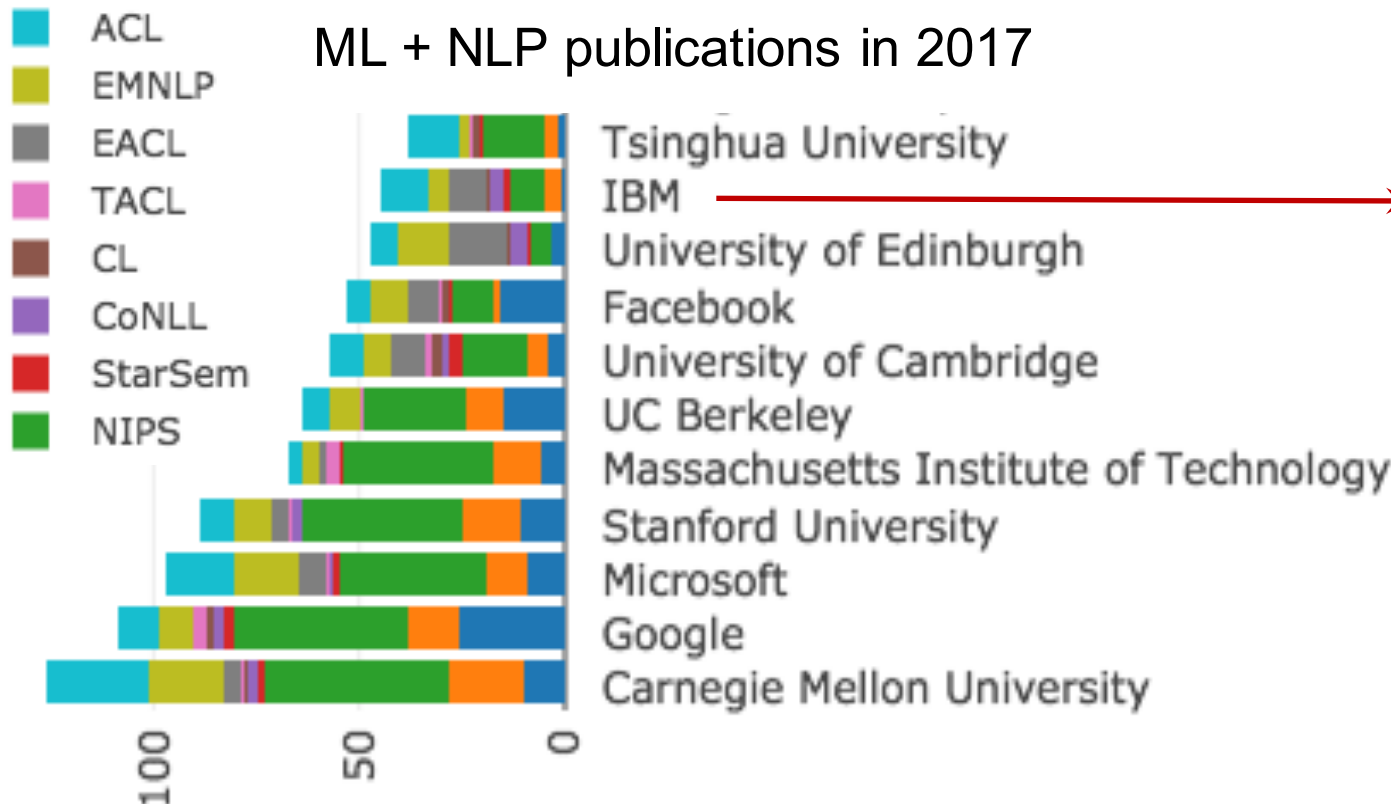
Data Wrangling

- The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.



Trifacta

Data Standardization



IBM Research
IBM Research, NY
IBM Research, USA
IBM T. J. Watson Research Center
IBM T. J. Watson Research
IBM T.J Watson Research Center
IBM T.J. Watson Research Center
IBM T.J. Watson Research Center
IBM Thomas J. Watson Research Center
IBM TJ Watson Research Center
... ..

Data Standardization

“It strikes me as a little comical that it required such a kludgy effort to pull these numbers together...”

“I wrote a bunch of manual patterns to map names to canonical versions (“UCL” to “University College London” and “Google Inc” to “Google”), although it is likely that I still missed some cases...”

“I’ve created a number of rules to map together alternative organization names and misspellings...”

IBM Research
IBM Research, NY
IBM Research, USA
IBM T. J. Watson Research Center
IBM T. J. Watson Research
IBM T.J Watson Research Center
IBM T.J. Watson Research Center
IBM T.J.Watson Research Center
IBM Thomas J. Watson Research Center
IBM TJ Watson Research Center
... ..

Data Cleaning

Inaccurate



The image shows a product listing for an HP ZBook 17 G2 Mobile Workstation. On the left is a large image of the laptop. To its right, the product name 'HP ZBook 17 G2 Mobile Workstation' is displayed, followed by a five-star rating and a link to 'Read all 1 reviews'. Below this, the price is shown as 'Was £2,378.30' and '£1.58', with 'VAT incl.' written underneath. A red oval highlights the '£1.58' price. To the right of the price is a purple 'SAVE' badge showing '£2,376.72'. Below the price, there is a checkbox for 'HP 5 year Next Business Day Onsite Hardware S...'. A quantity selector shows '1' with a dropdown arrow, and an orange 'ADD TO BASKET' button is next to it. At the bottom, a green checkmark indicates 'Delivered in 5-10 Working days'. A row of four small thumbnail images is at the bottom left, with the first one selected.

HP ZBook 17 G2 Mobile Workstation

★★★★★ [Read all 1 reviews](#)

Was £2,378.30
£1.58
VAT incl.

SAVE
£2,376.72

☐ HP 5 year Next Business Day Onsite Hardware S...

1 ▼ **ADD TO BASKET**

✓ Delivered in 5-10 Working days

Data Cleaning

Inconsistent

FlightView	FlightAware	Orbitz
American Airlines Flight Number 119 (AA119)	AAL119 (Track inbound flight) (web app) (API) (Status) American Airlines "American"	American Airlines # 119
FLIGHT TRACKER	Aircraft Boeing 737-800 (twin-jet) (B738/Q - track or photos) Origin Terminal A / Gate 32 / Newark Liberty Intl (KEWR - track) Destination Terminal 4 / Gate 42B / Los Angeles Intl (KLAX - track) Other flights between these airports	Leg 1: In Transit Departs: Newark (EWR) View real-time airport status Gate: 32
Departure Airport: Scheduled Time: 6:15 PM, Dec 08 Takeoff Time: 6:53 PM, Dec 08 Terminal - Gate: Terminal A - 32	Route ZIMMER Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J Date 2011年 12月 08日 (Thursday) Duration 5 hours 43 minutes 20 minutes left 5 hours 23 minutes Progress	Scheduled Estimated Actual 6:22p - 6:32p Dec 8 Dec 8
ArrivalStatus: In Air Airport: Scheduled Time: 9:40 PM, Dec 08 9:42 PM, Dec 08 Estimated Time: Track This Flight Live!	Status En Route (2,284 sm down; 168 sm to go) Distance Direct: 2,451 sm Planned: 2,458 Fare \$51.99 to \$3,561.11; average: \$241.96 (airline insight) Cabin First: Dinner / Economy: Food for sale Scheduled 7-day Average Actual/Estimated	Arrives: Los Angeles (LAX) View real-time airport status Gate: 42B
Time Remaining: 25 min Terminal - Gate: Terminal 4 - 42B Baggage Claim: 4	Departure 06:15PM EST 07:08PM EST 06:53PM EST Arrival 08:33PM PST 09:17PM PST 09:36PM PST	Scheduled Estimated Actual 9:54p 9:47p Dec 8 Dec 8

Data Cleaning

Incomplete

Country	UN R/P 10% ^[4]	UN R/P 20% ^[5]	World Bank Gini (%) ^[6]	WB Gini (year)	CIA R/P 10% ^[7]	Year	CIA Gini (%) ^[8]	CIA Gini (year)	GPI Gini (%) ^[9]
 Seychelles			65.8	2007					
 Comoros			64.3	2004					
 Namibia	106.6	56.1	63.9	2004	129.0	2003	59.7	2010	
 South Africa	33.1	17.9	63.1	2009	31.9	2000	65.0	2005	
 Botswana	43.0	20.4	61.0	1994			63	1993	
 Haiti	54.4	26.6	59.2	2001	68.1	2001	59.2	2001	
 Angola			58.6	2000					62.0
 Honduras	59.4	17.2	57.0	2009	35.2	2003	57.7	2007	

Data Cleaning



- Address errors caused **6.8 billion** undelivered mails in 2013
- Estimated **\$1.5 billion** spent on processing
- At least **\$3.4 billion** wasted postage



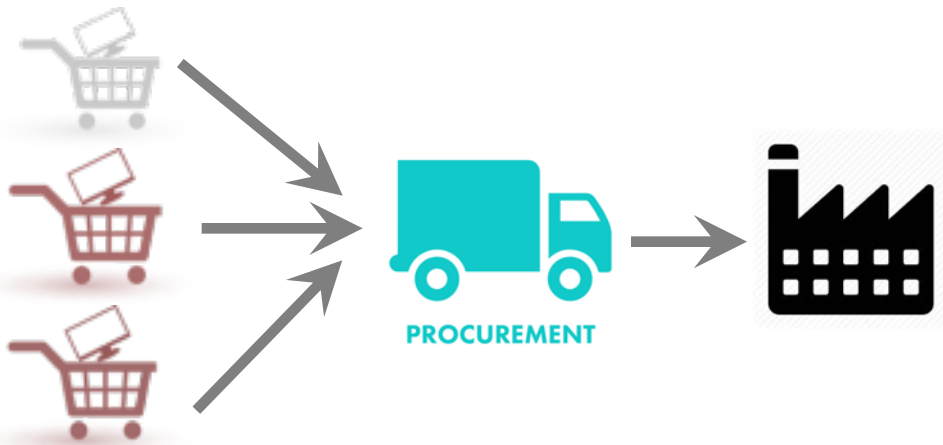
DATA

Bad Data Costs the U.S. \$3 Trillion Per Year

by [Thomas C. Redman](#)

SEPTEMBER 22, 2016

Data Integration



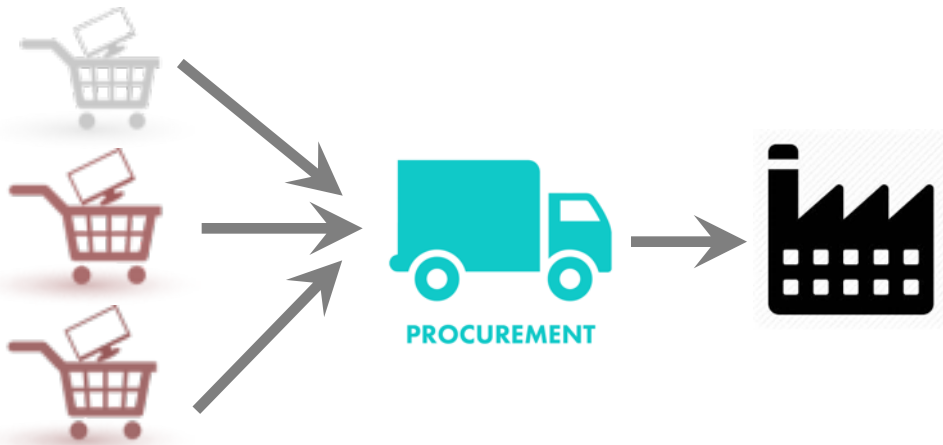
“GE estimates they would save
\$100M/year by integrating orders
for better pricing” **FORTUNE**



Comparison Shopping

Supplier Mastering,
Customer Mastering,
Collaborative Research etc.

Data Integration



“GE estimates they would save \$100M/year by integrating orders for better pricing”
FORTUNE



Comparison Shopping

Will be increasingly more important

- data sharing becomes pervasive
- translation of legacy data

Data Integration is Critical

Google Rachel Abrams

About 942,000 results (0.35 seconds)

Rachel Abrams - The New York Times
www.nytimes.com/by/rachel-abrams
Rachel Abrams joined The New York Times as a business reporter in 2013. Before that, she had written about the economics of the film industry for Variety magazine. ... Thomas Barrack's Colony Capital Ends Bid for Weinstein Studio.

Google Thinks I'm Dead - The New York Times
<https://www.nytimes.com/2017/12/16/business/google-thinks-im-dead.html>
Dec 16, 2017 - For much of the last week, I have been trying to persuade the world's most powerful search engine to remove my photo from biographical details that belong to someone else. A search for "Rachel Abrams" revealed that Google had mashed my picture from The New York Times's website with the Wikipedia ...

Rachel C. Abrams @RachelAbramsNY
How I spent a week trying to tell someone at Google that I wasn't dead yet

Rachel C. Abrams (@RachelAbramsNY) | Twitter
<https://twitter.com/rachelabramsny?lang=en>
3619 tweets • 249 photos/videos • 4894 followers. Check out the latest Tweets from Rachel C. Abrams (@RachelAbramsNY)

Rachel Abrams | Professional Profile - LinkedIn
<https://www.linkedin.com/in/rachel-abrams-64539783>
New York, New York - reporter - The New York Times

Rachel Abrams
American writer

Rachel Abrams was an American writer, editor, and artist. She was the wife of Elliott Abrams. [Wikipedia](#)

Born: January 2, 1951
Died: June 7, 2013
Spouse: Elliott Abrams (m. 1980–2013)
Parents: Midge Decter
Children: Sarah Abrams, Jacob Abrams, Joseph Abrams

People also search for

Elliott Abrams (Spouse)
Midge Decter (Mother)
John Podhoretz
Norman Podhoretz
Ricky Silberman

Could be me...?
Me
Not me
Definitely not me

Who has More Data Sources?

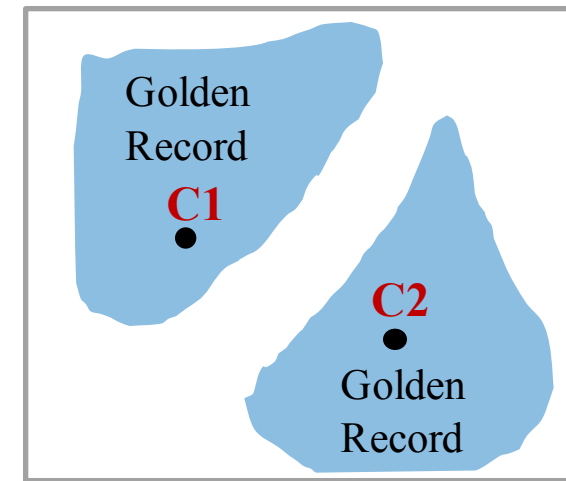
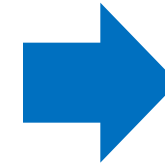
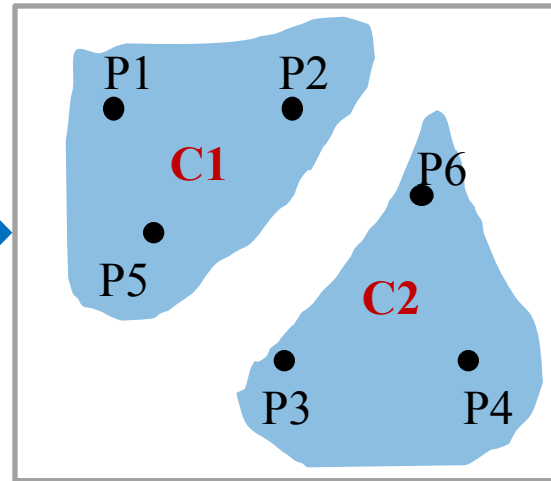
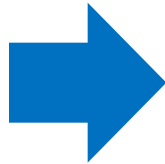
- Large manufacturing enterprise
 - Has 325 procurement systems
 - Estimates they would save \$100M/year by “most favored nation status”
 - Large drug company
 - Has 10,000 bench scientists
 - Wants to integrate their “electronic lab notebooks”
 - Large auto company
 - Wants to integrate customer databases In Europe — In 40 languages
-

A Closer Look at Data Integration

Entity Resolution:
Find Duplicate Records

Entity Consolidation:
Merge Duplicate Records

ID	Name	Address	Telephone
P1	Mary Lee	9 St, Wisconsin	(718) 453-0681
P2	M. Lee	9th St, WI	7184530681
P3	James Smith	3rd E Ave, CA	212-213-2888x264
P4	J. Smith	3 E Avenue, CA	(212) 213-2888
P5	Lee, Mary	9 Street, WI	+1-718-453-0681
P6	Smith, James	5th Street, WA	+1-212-213-2888



Schema Mapping



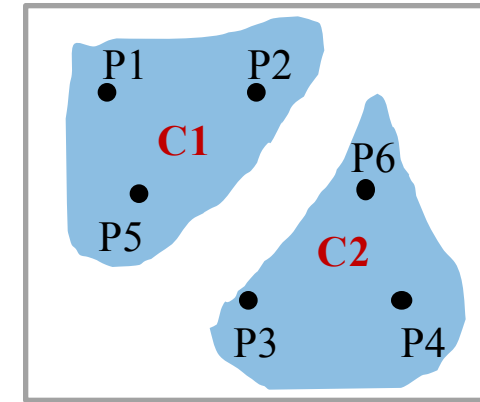
Entity Resolution

ID	Name	Address	Telephone
P1	Mary Lee	9th Street, WI	(718) 453-0681
P2	M. Lee	9th St, WI	718-453-0681
P3	James Smith	3rd E Ave, CA	212-213-2888x264
P4	J. Smith	3 E Avenue, CA	(212) 213-2888
P5	Lee, Mary	9 Street, WI	+1-718-453-0681
P6	Smith, James	5th Street, WA	+1-212-213-2888

Abstract as Sequences

$EditSimilarity = 0.86$

$EditSimilarity = 0.21$

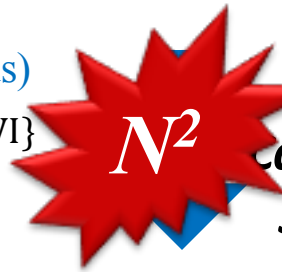


Abstract as Sets (Bag-of-Words)

$P1[Address] \rightarrow A = \{9th, Street, WI\}$

$P2[Address] \rightarrow B = \{9th, St, WI\}$

$Overlap(A, B) = |A \cap B| = 2$



Similarity Join:

Calculating All-Pair
Similarity Scores

machine-learning-based
&
pruning non-duplicates

All Pairs	Overlap of Address	Edit Similarity of Tel	...
(P5, P6)	1	0.43	...
(P4, P5)	0 less similar	0.21 less similar	
(P2, P5)	1	0.71	...
(P1, P2)	2 more similar	0.86 more similar	
...	...	...	...

Label



Label



Similarity Partial Order

Inference

29 Inference

Inference

Inference

$P4 \neq P5, P5 = P2 \rightarrow P4 \neq P2$

$P1 = P2, P2 = P5 \rightarrow P1 = P5$

Transitivity



Similarity Measures

	Unnormalized	Normalized
Sequence	Edit Distance	Edit Similarity
Set	Overlap Size	Jaccard/Cosine Similarity

INPUT

OUTPUT

Set Similarity Measures

	Unnormalized Score	Normalized Score
Sequence	Edit Distance	Edit Similarity
Set	Overlap Size	Jaccard/Cosine Similarity

$Overlap(A, B) = A \cap B $	3	A 	
$Jaccard(A, B) = \frac{ A \cap B }{ A \cup B }$	0.6		
$Cosine(A, B) = \frac{ A \cap B }{\sqrt{ A B }}$	0.75	B 	

Set Similarity Measures

	Unnormalized	Normalized
Sequence	Edit Distance	Edit Similarity
Set	Overlap Size	Jaccard/Cosine Similarity

$$Overlap(A, B) = |A \cap B| \quad 2$$

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad 0.66$$

David Smith → $A = \{\text{David, Smith}\}$

Smith R. David → $B = \{\text{David, R., Smith}\}$

$$Cosine(A, B) = \frac{|A \cap B|}{\sqrt{|A||B|}} \quad 0.82$$