Advanced Data Management (CSCI 490/680)

Introduction

Dr. David Koop

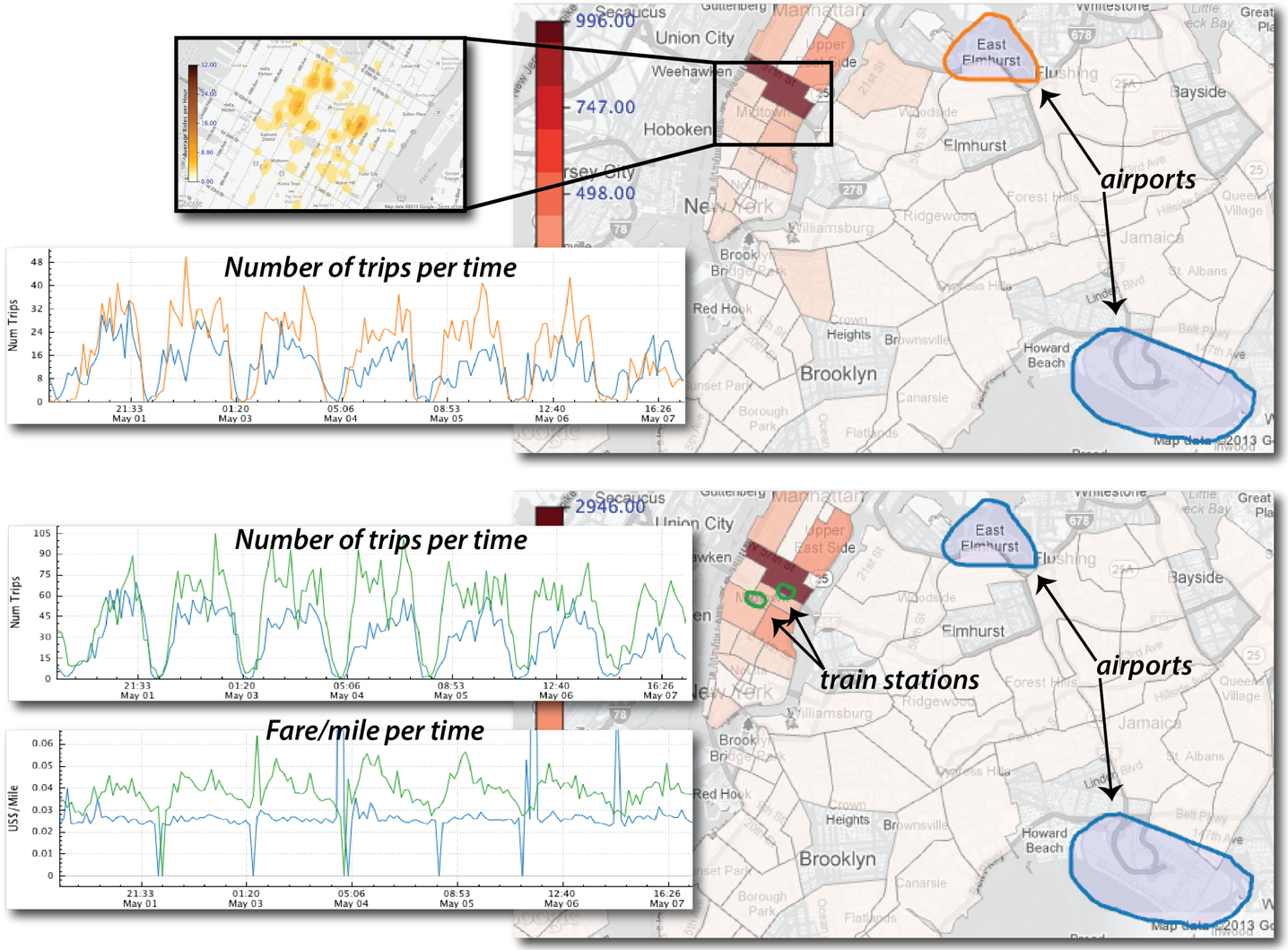
NYC Taxi Data

[Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance, T. W. Schneider]

NYC Taxi Data: Day analysis

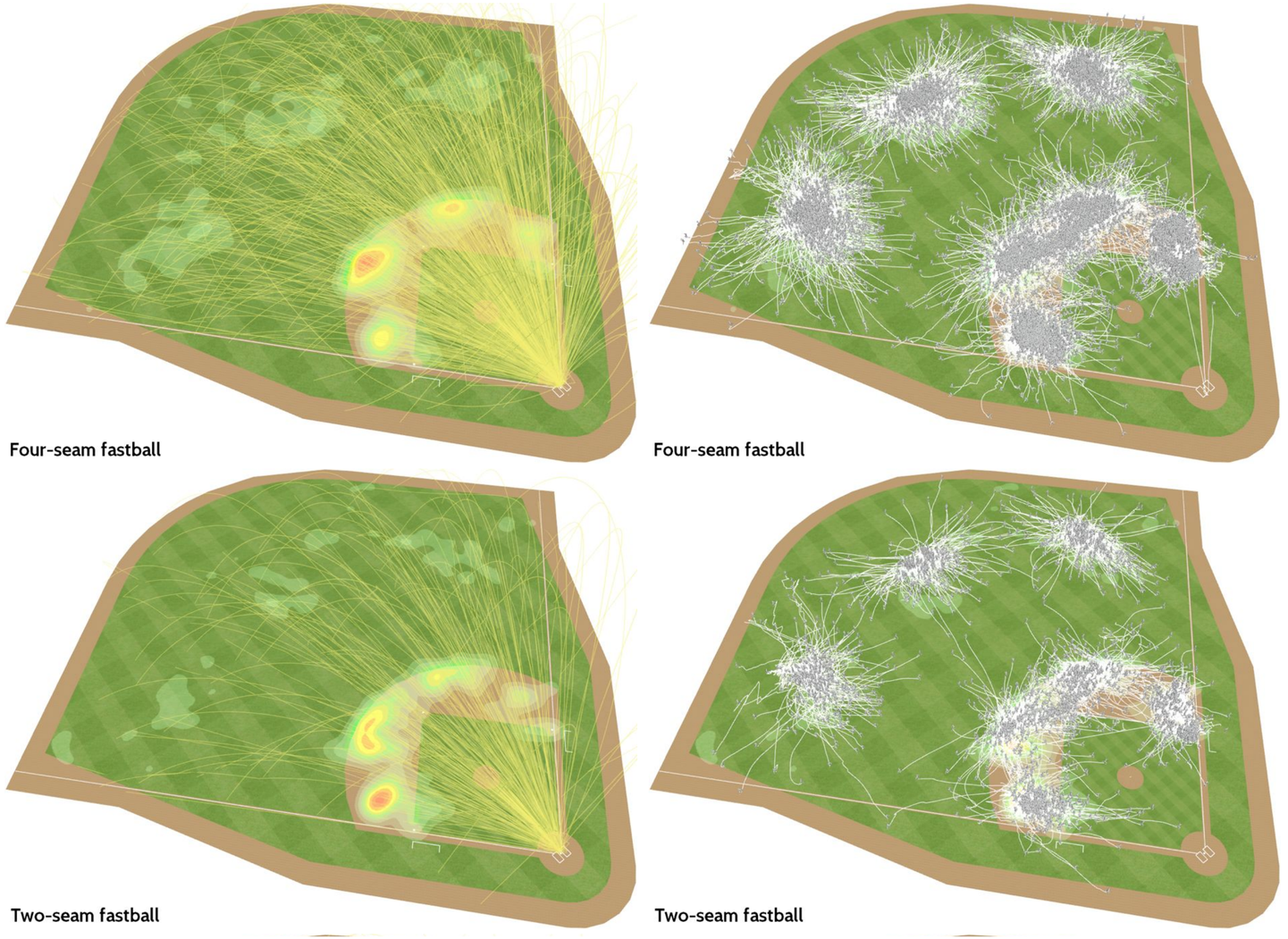
[Ferreira et al., 2013]

NYC Taxi Data: Region analysis



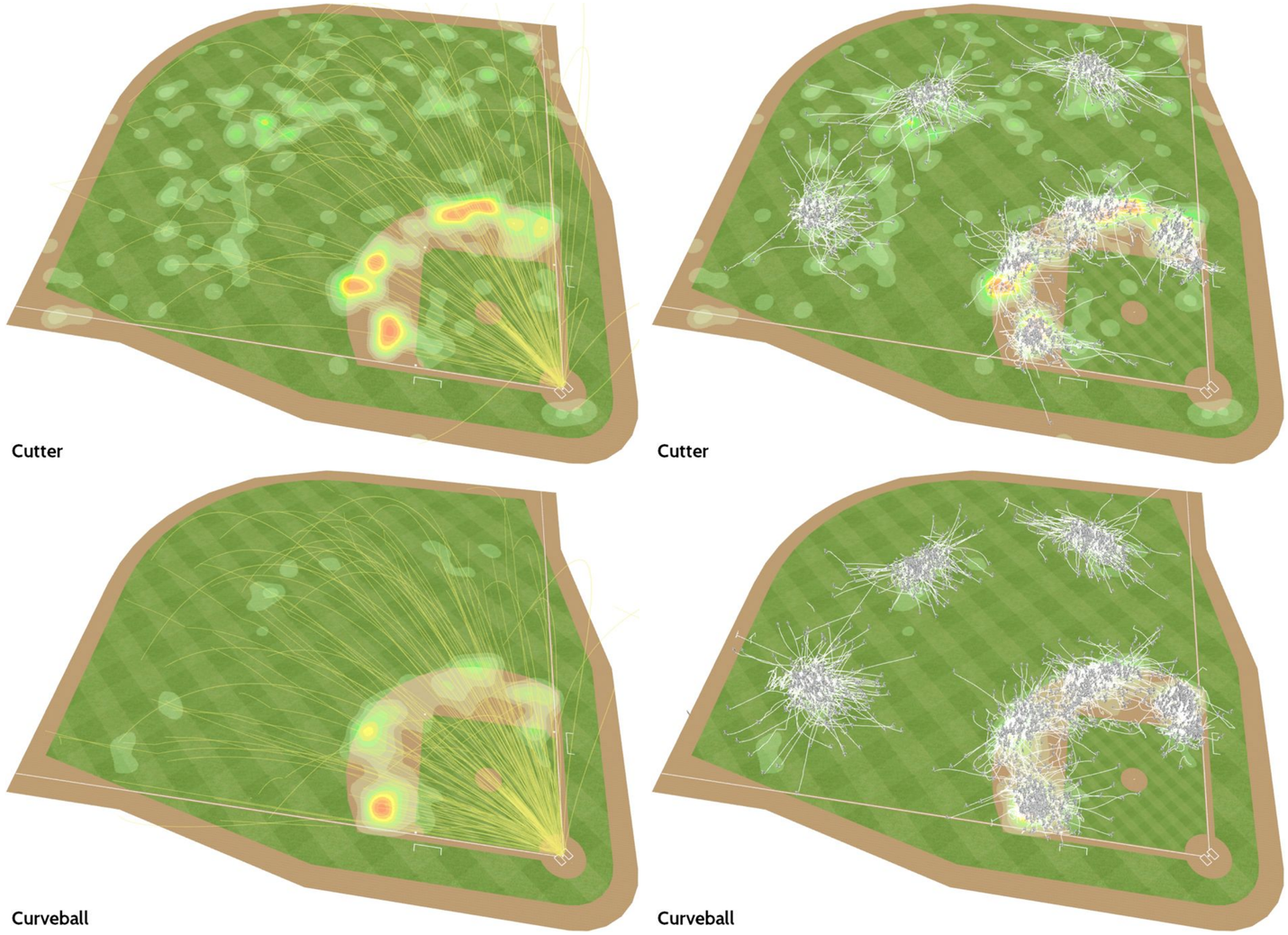
[Ferreira et al., 2013]

Baseball Data



[Deitrich et al., 2014]

Baseball Data



[Deitrich et al., 2014]

Real-time Analysis

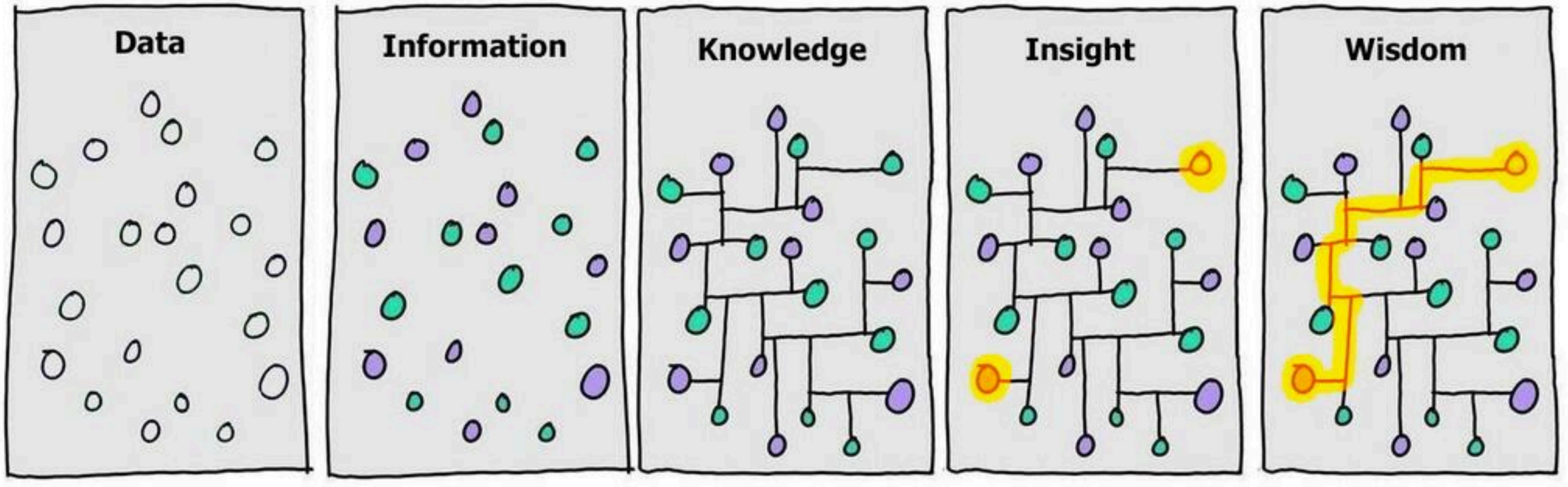
- Want to have results now
- How?
  - Faster machines
  - Clusters
  - Progressive techniques

What's involved in dealing with data?

Data Acquisition	Data Analysis	Data Curation	Data Storage	Data Usage
• Structured data • Unstructured data • Event processing • Sensor networks • Protocols • Real-time • Data streams • Multimodality	• Stream mining • Semantic analysis • Machine learning • Information extraction • Linked Data • Data discovery • 'Whole world' semantics • Ecosystems • Community data analysis • Cross-sectorial data analysis	• Data Quality • Trust / Provenance • Annotation • Data validation • Human-Data Interaction • Top-down/Bottom-up • Community / Crowd • Human Computation • Curation at scale • Incentivisation • Automation • Interoperability	• In-Memory DBs • NoSQL DBs • NewSQL DBs • Cloud storage • Query Interfaces • Scalability and Performance • Data Models • Consistency, Availability, Partition-tolerance • Security and Privacy • Standardization	• Decision support • Prediction • In-use analytics • Simulation • Exploration • Visualisation • Modeling • Control • Domain-specific usage

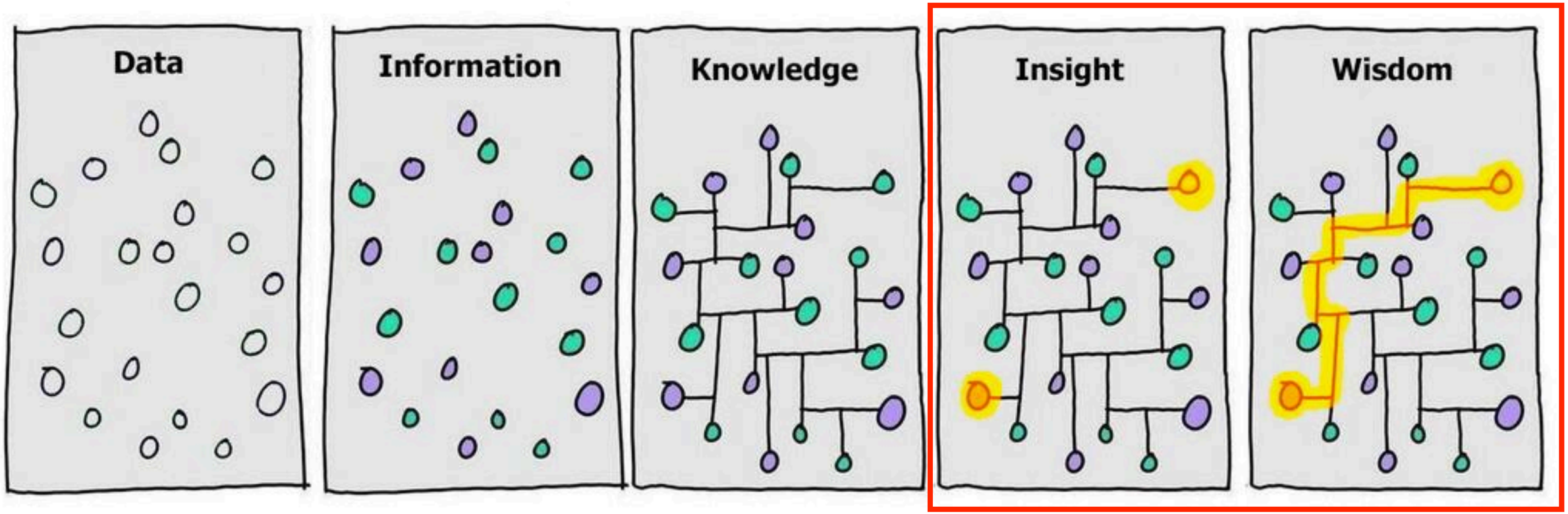
[Big Data Value Chain, Curry et al., 2014]

Data to Knowledge



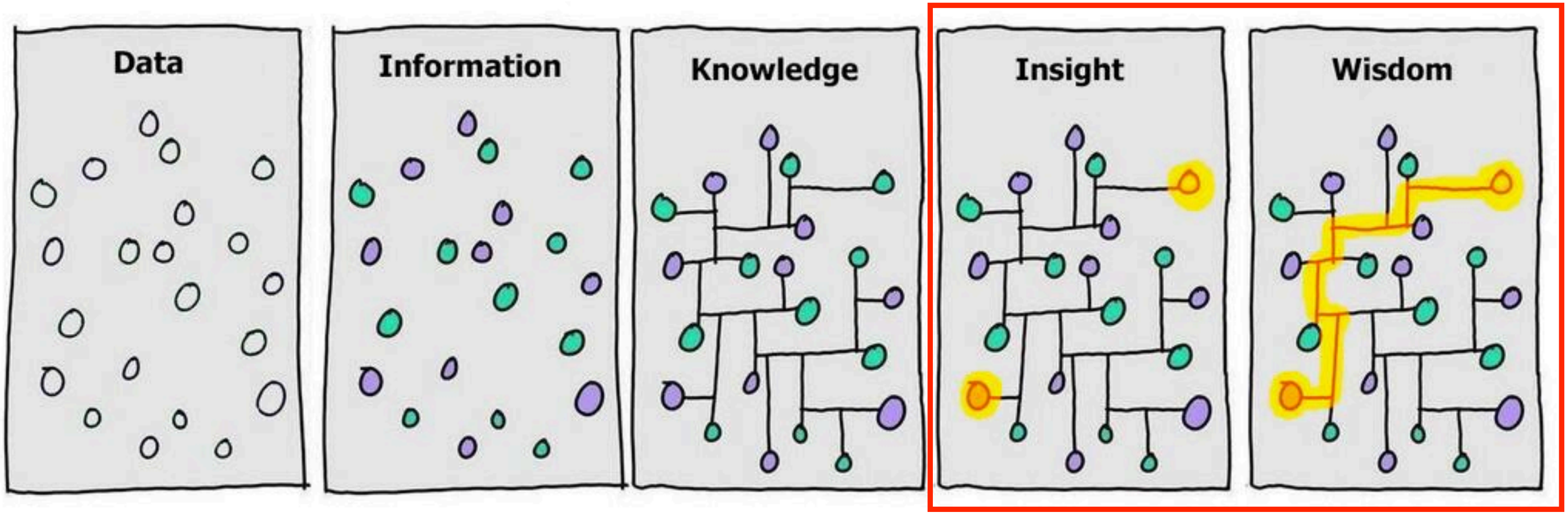
[D. Somerville, based on H. McLeod's original]

Data to Knowledge



[D. Somerville, based on H. McLeod's original]

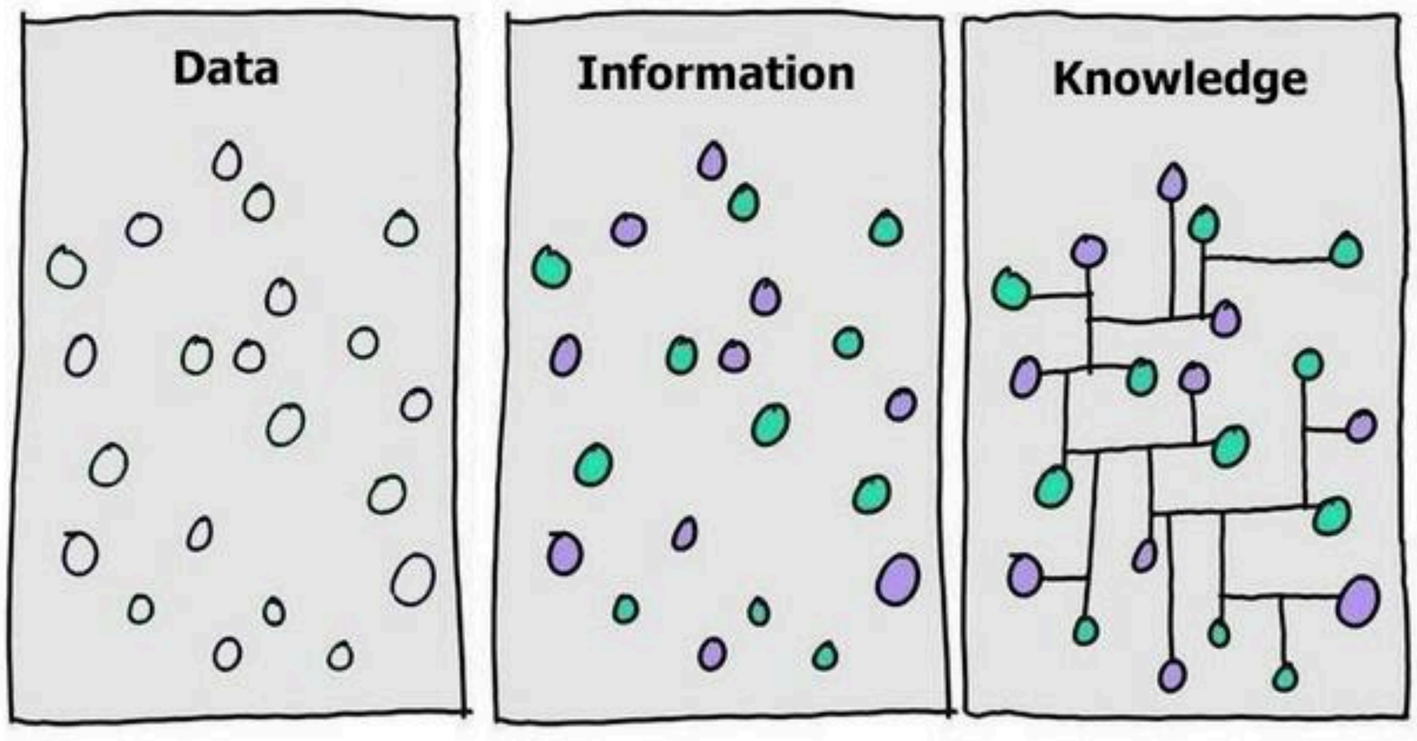
Data to Knowledge



Require People

[D. Somerville, based on H. McLeod's original]

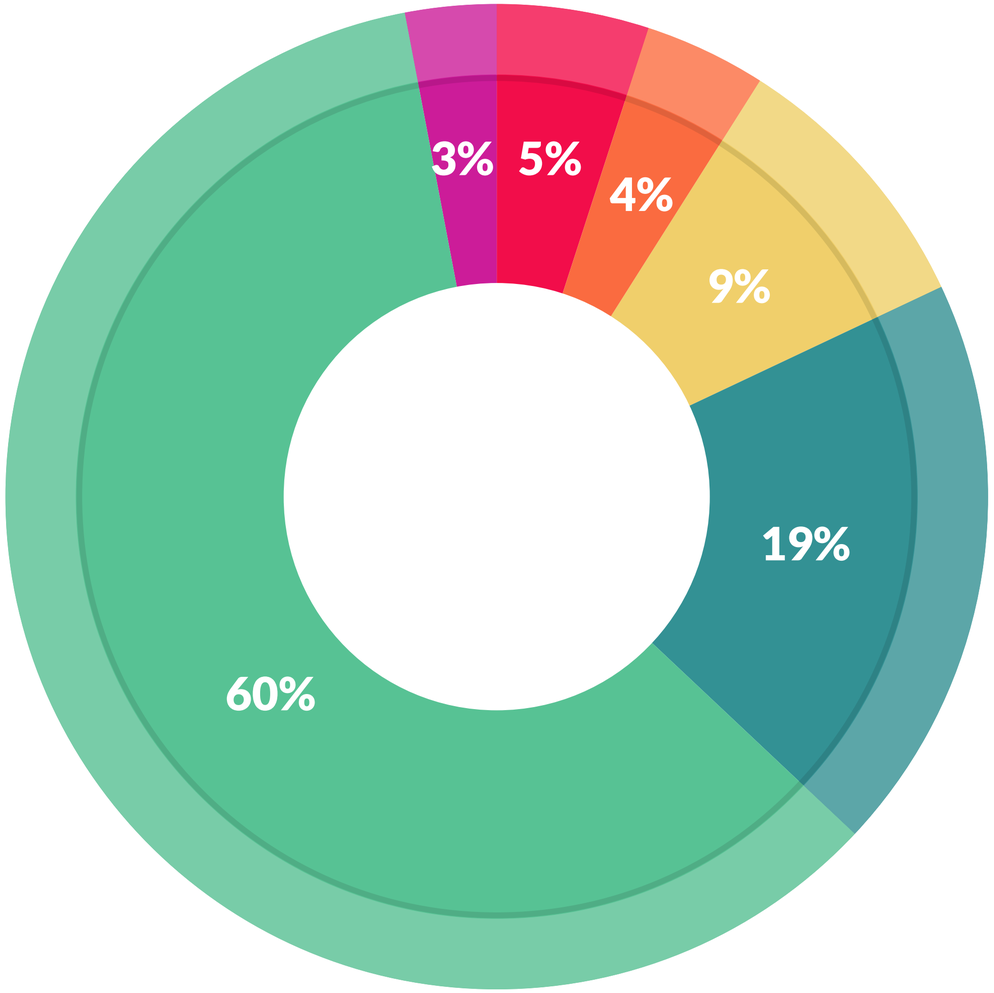
Data to Knowledge



Can computers do this for us?

[D. Somerville, based on H. McLeod's original]

How do data scientists spend their time?

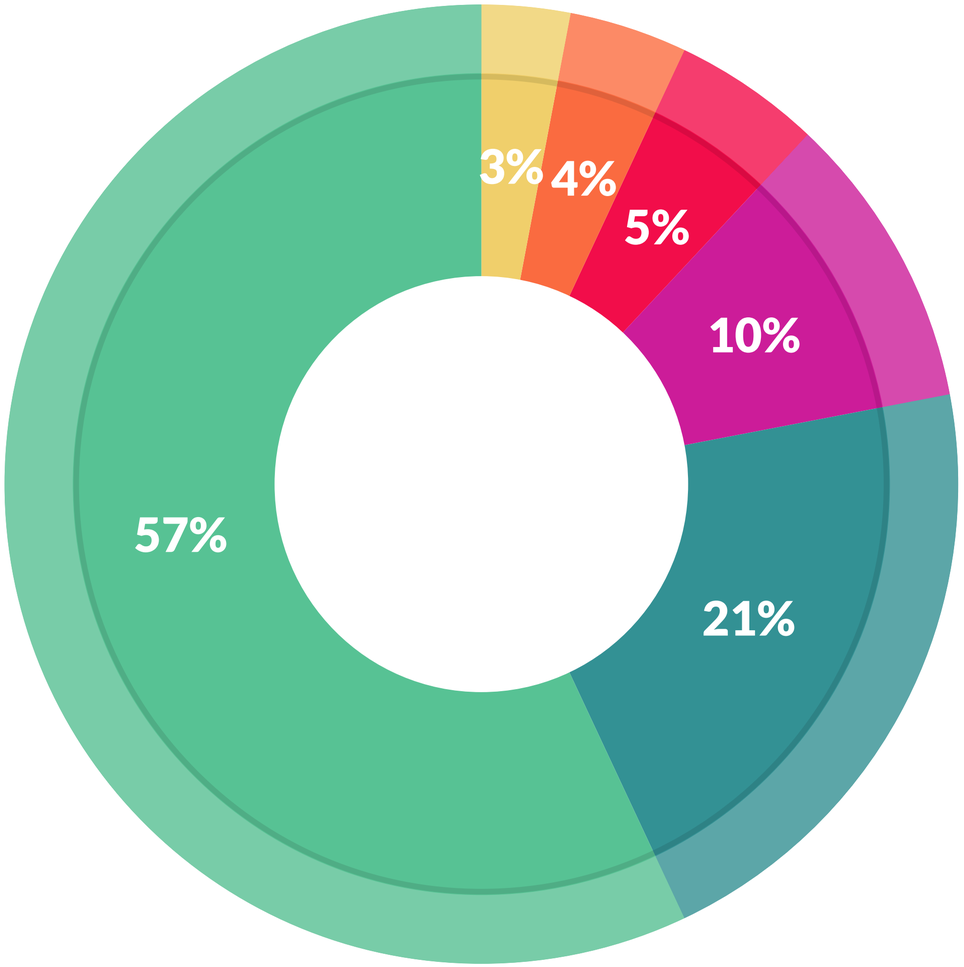


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

[CrowdFlower Data Science Report, 2016]

What do they like doing?

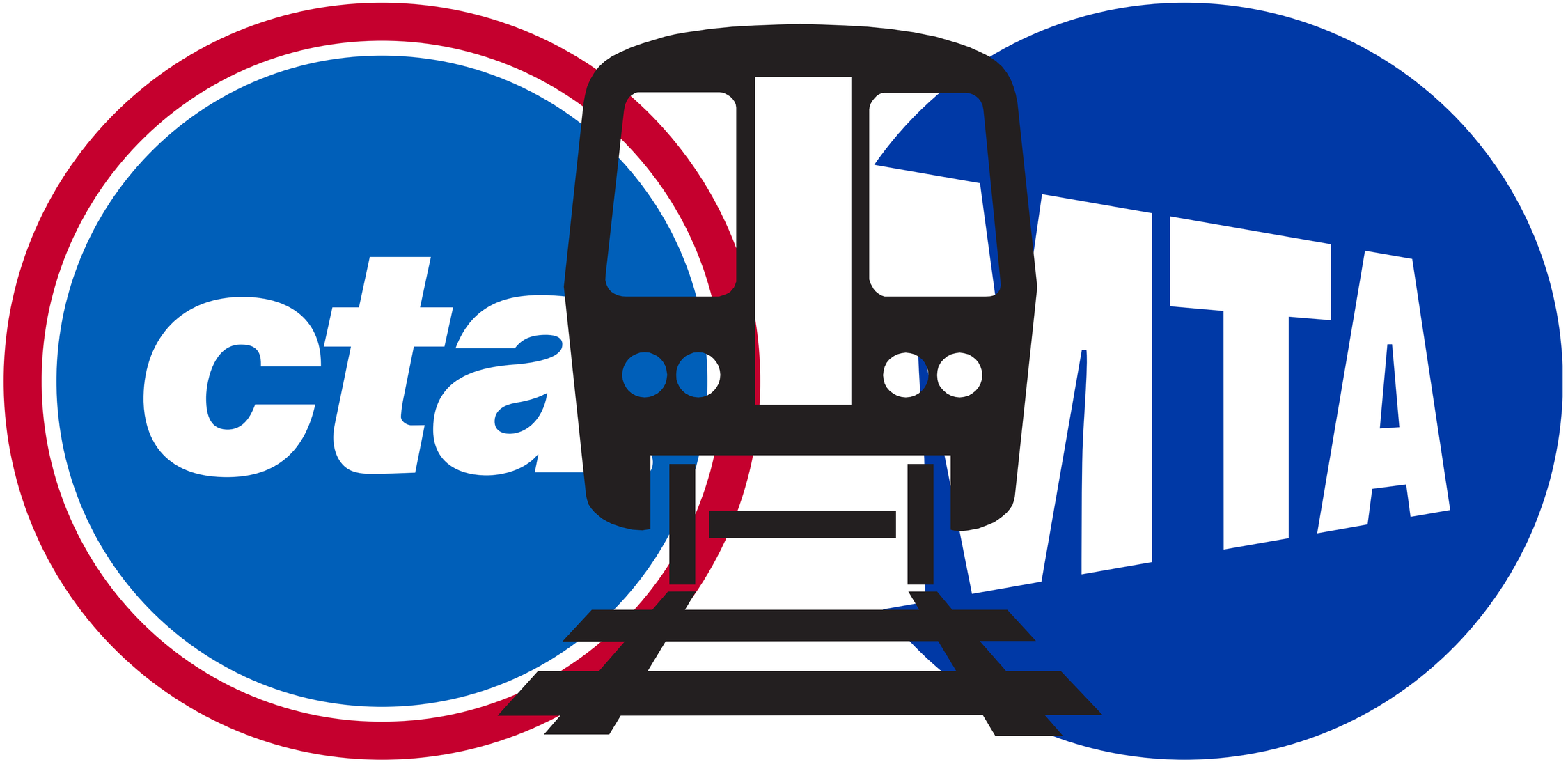


What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

[CrowdFlower Data Science Report, 2016]

Example: Compare public transit in Chicago and NYC



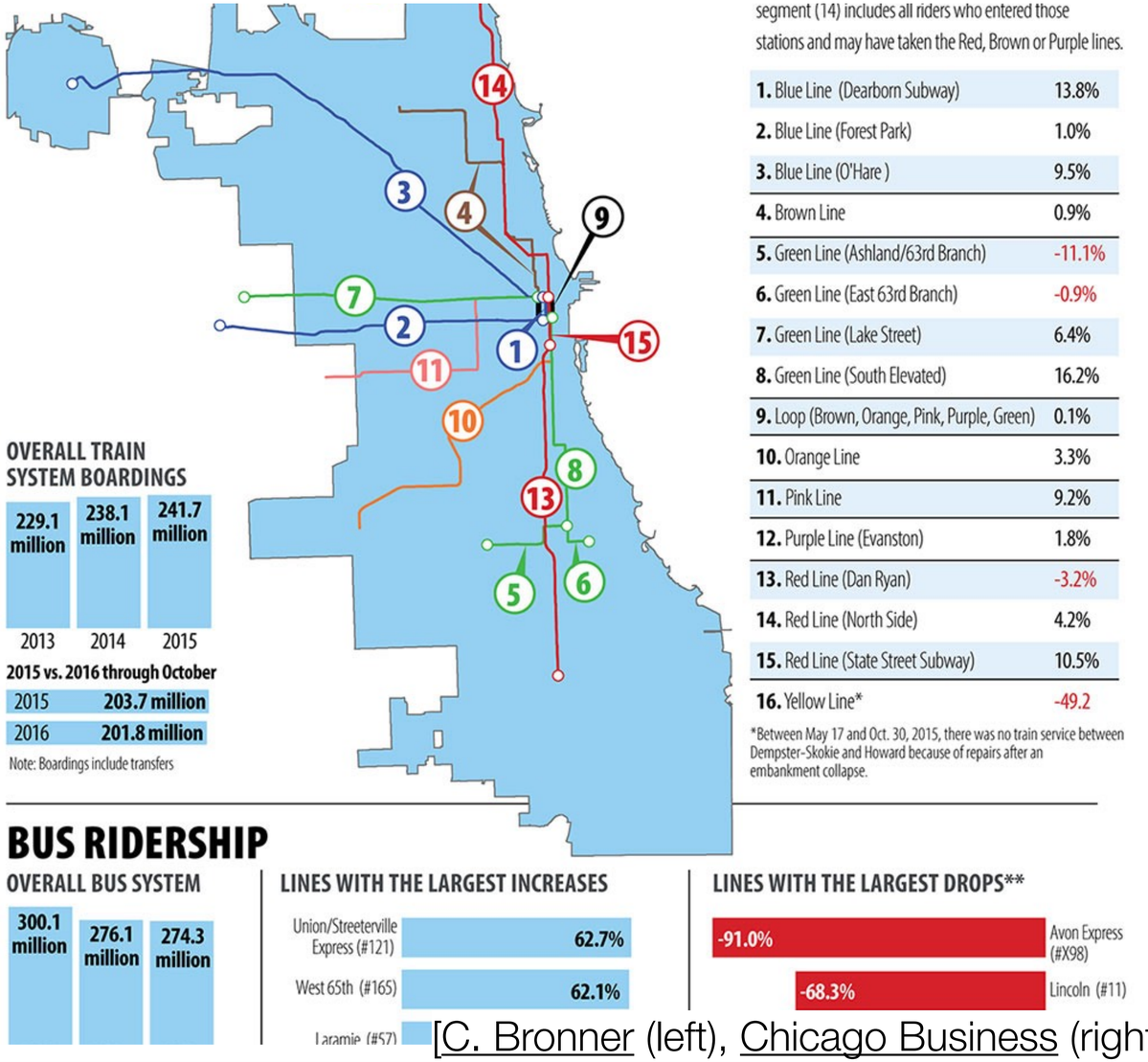
Public Transit Ridership Data

station_id	stationname	date	daytype	rides
40350	UIC-Halsted	01/01/2001	U	273
41130	Halsted-Orange	01/01/2001	U	306
40760	Granville	01/01/2001	U	1,059
40070	Jackson/Dearborn	01/01/2001	U	649
40090	Damen-Brown	01/01/2001	U	411
40590	Damen/Milwaukee	01/01/2001	U	870
40720	East 63rd-Cottage Grove	01/01/2001	U	391
41260	Austin-Lake	01/01/2001	U	399
40230	Cumberland	01/01/2001	U	788
41120	35-Bronzeville-IIT	01/01/2001	U	448
40810	Medical Center	01/01/2001	U	479
40330	Grand/State	01/01/2001	U	2,542
41050	Linden	01/01/2001	U	176
40140	Skokie	01/01/2001	U	0
40450	95th/Dan Ryan	01/01/2001	U	3,948
40400	Noyes	01/01/2001	U	72
40150	Pulaski-Cermak	01/01/2001	U	0
40690	Dempster	01/01/2001	U	177
40460	Merchandise Mart	01/01/2001	U	185
40840	South Boulevard	01/01/2001	U	202
41280	Jefferson Park	01/01/2001	U	1,302
40130	51st	01/01/2001	U	364
40870	Francisco	01/01/2001	U	196
40710	Chicago/Franklin	01/01/2001	U	384
40740	Western-Cermak	01/01/2001	U	0
40550	Irving Park-O'Hare	01/01/2001	U	731

< Previous Next > Showing Rows 1-100 out of 962,546

```
C/A,UNIT,SCP,STATION,LINENAME,DIVISION,DATE,TIME,DESC,ENTRIES,EXITS
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,03:00:00,REGULAR,0007331213,0002484849
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,07:00:00,REGULAR,0007331224,0002484861
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,11:00:00,REGULAR,0007331281,0002484936
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,15:00:00,REGULAR,0007331454,0002485014
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,19:00:00,REGULAR,0007331759,0002485106
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,23:00:00,REGULAR,0007331951,0002485166
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,03:00:00,REGULAR,0007331997,0002485182
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,07:00:00,REGULAR,0007332007,0002485190
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,11:00:00,REGULAR,0007332052,0002485249
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,15:00:00,REGULAR,0007332197,0002485308
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,19:00:00,REGULAR,0007332405,0002485369
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,23:00:00,REGULAR,0007332543,0002485396
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,03:00:00,REGULAR,0007332566,0002485402
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,07:00:00,REGULAR,0007332574,0002485431
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,11:00:00,REGULAR,0007332705,0002485725
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,15:00:00,REGULAR,0007332892,0002485801
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,19:00:00,REGULAR,0007333645,0002485891
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,23:00:00,REGULAR,0007333879,0002485925
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,03:00:00,REGULAR,0007333906,0002485935
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,07:00:00,REGULAR,0007333921,0002485986
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,11:00:00,REGULAR,0007334052,0002486261
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,15:00:00,REGULAR,0007334252,0002486319
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,19:00:00,REGULAR,0007335008,0002486391
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,23:00:00,REGULAR,0007335258,0002486432
A002,R051,02-00-00,59
```

Cool Machine Learning Model & Pretty Visualizations



[C. Bronner (left), Chicago Business (right)]

Wait… how do we actually get those results?

station_id	stationname	date	daytype	rides
40350	UIC-Halsted	01/01/2001	U	273
41130	Halsted-Orange	01/01/2001	U	306
40760	Granville	01/01/2001	U	1,059
40070	Jackson/Dearborn	01/01/2001	U	649
40090	Damen-Brown	01/01/2001	U	411
40590	Damen/Milwaukee	01/01/2001	U	870
40720	East 63rd-Cottage Grove	01/01/2001	U	391
41260	Austin-Lake	01/01/2001	U	399
40230	Cumberland	01/01/2001	U	788
41120	35-Bronzeville-IIT	01/01/2001	U	448
40810	Medical Center	01/01/2001	U	479
40330	Grand/State	01/01/2001	U	2,542
41050	Linden	01/01/2001	U	176
40140	Skokie	01/01/2001	U	0
40450	95th/Dan Ryan	01/01/2001	U	3,948
40400	Noyes	01/01/2001	U	72
40150	Pulaski-Cermak	01/01/2001	U	0
40690	Dempster	01/01/2001	U	177
40460	Merchandise Mart	01/01/2001	U	185
40840	South Boulevard	01/01/2001	U	202
41280	Jefferson Park	01/01/2001	U	1,302
40130	51st	01/01/2001	U	364
40870	Francisco	01/01/2001	U	196
40710	Chicago/Franklin	01/01/2001	U	384
40740	Western-Cermak	01/01/2001	U	0
40550	Irving Park-O'Hare	01/01/2001	U	731

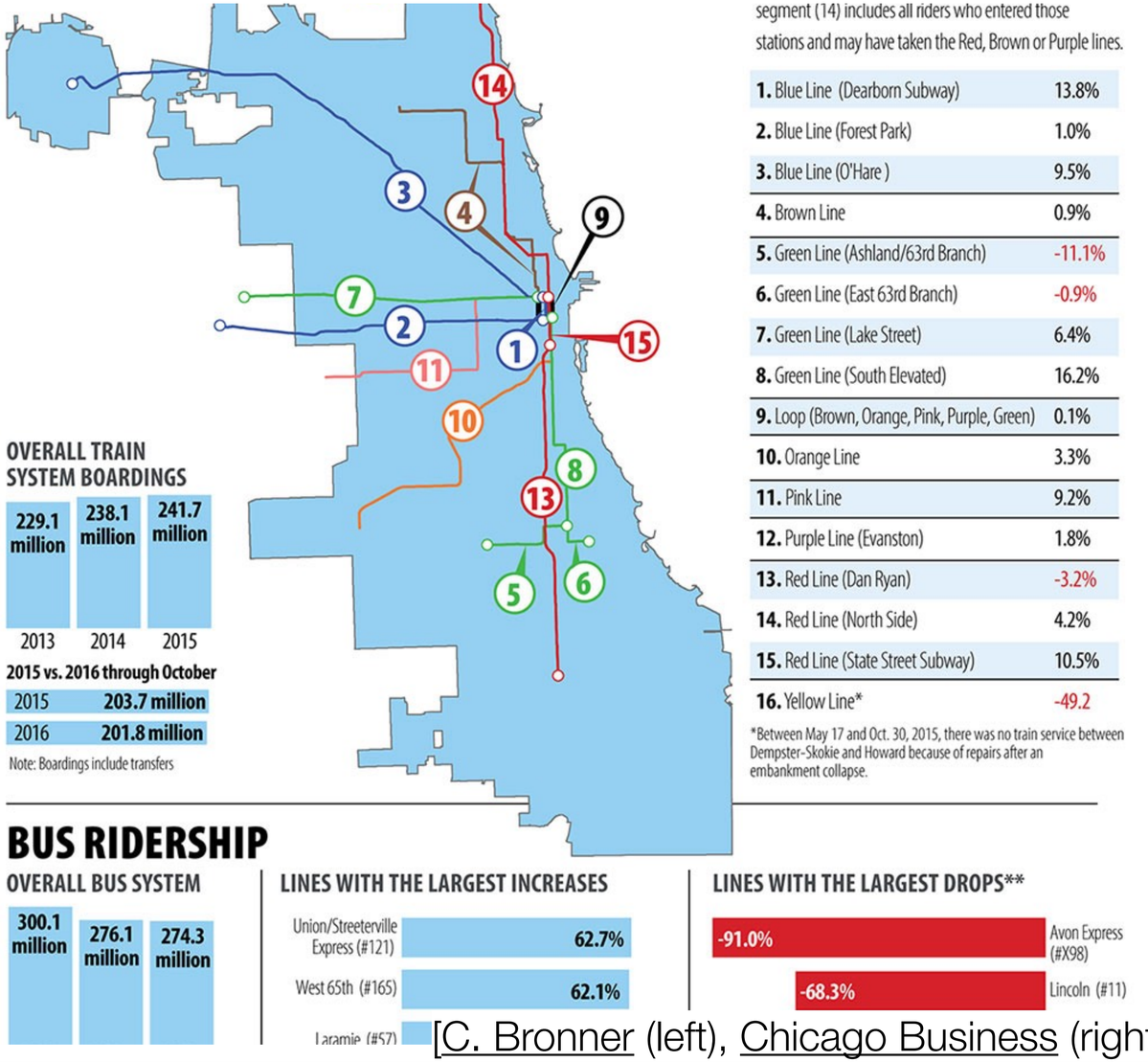
< Previous Next > Showing Rows 1-100 out of 962,546

```
C/A,UNIT,SCP,STATION,LINENAME,DIVISION,DATE,TIME,DESC,ENTRIES,EXITS
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,03:00:00,REGULAR,0007331213,0002484849
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,07:00:00,REGULAR,0007331224,0002484861
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,11:00:00,REGULAR,0007331281,0002484936
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,15:00:00,REGULAR,0007331454,0002485014
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,19:00:00,REGULAR,0007331759,0002485106
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,23:00:00,REGULAR,0007331951,0002485166
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,03:00:00,REGULAR,0007331997,0002485182
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,07:00:00,REGULAR,0007332007,0002485190
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,11:00:00,REGULAR,0007332052,0002485249
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,15:00:00,REGULAR,0007332197,0002485308
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,19:00:00,REGULAR,0007332405,0002485369
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/05/2020,23:00:00,REGULAR,0007332543,0002485396
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,03:00:00,REGULAR,0007332566,0002485402
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,07:00:00,REGULAR,0007332574,0002485431
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,11:00:00,REGULAR,0007332705,0002485725
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,15:00:00,REGULAR,0007332892,0002485801
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,19:00:00,REGULAR,0007333645,0002485891
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/06/2020,23:00:00,REGULAR,0007333879,0002485925
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,03:00:00,REGULAR,0007333906,0002485935
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,07:00:00,REGULAR,0007333921,0002485986
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,11:00:00,REGULAR,0007334052,0002486261
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,15:00:00,REGULAR,0007334252,0002486319
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,19:00:00,REGULAR,0007335008,0002486391
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/07/2020,23:00:00,REGULAR,0007335258,0002486432
A002,R051,02-00-00,59
```

Processing the data

- Data Ingestion
  - Need to understand format of the data
  - Need to understand what the data is (types and semantics)
- Data Wrangling
  - Get the data into a meaningful state
  - Check for errors in the data
  - Check for missing data and deal with it
- Data Integration
  - Make it so we can actually compare the data
  - Put the datasets together

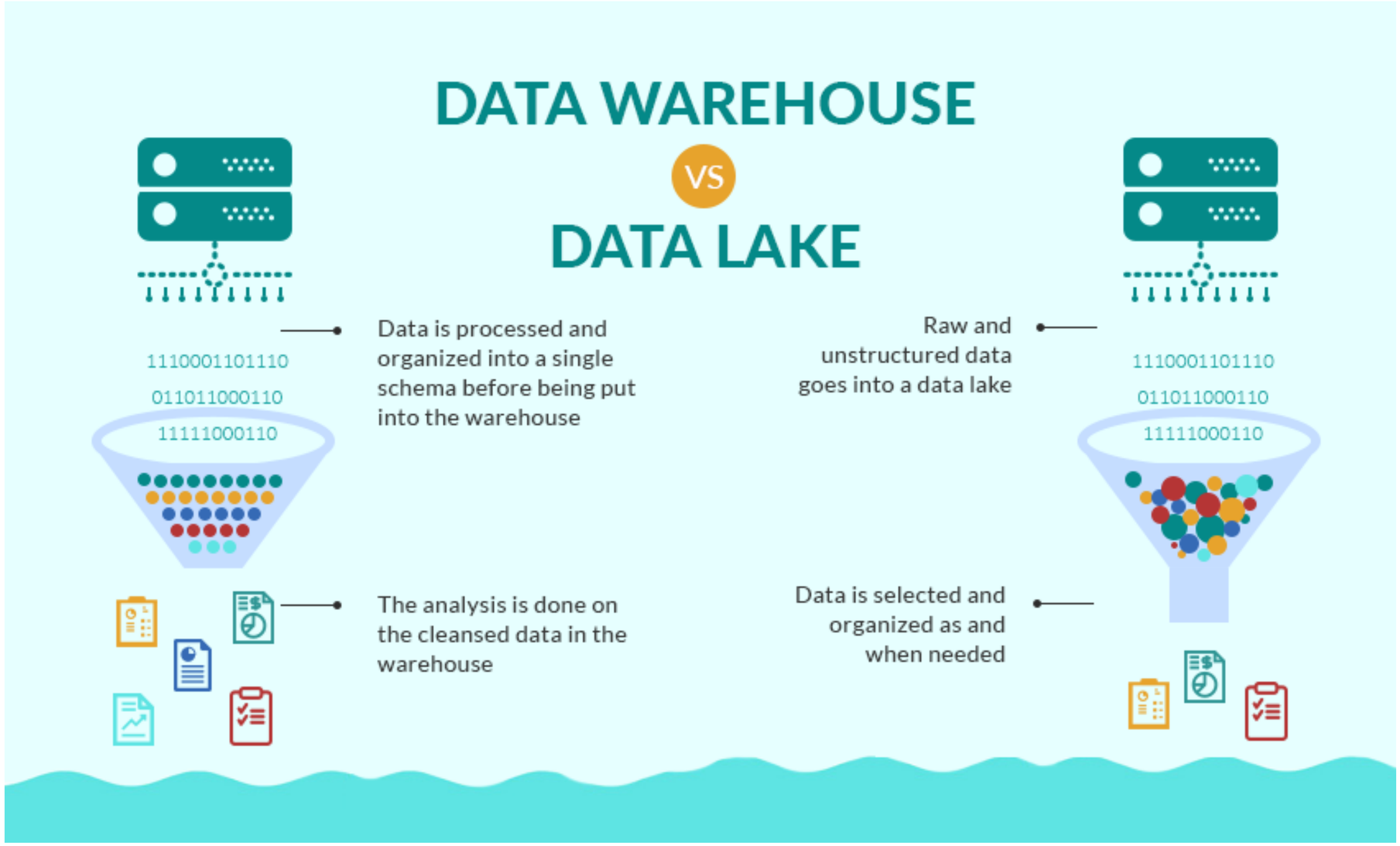
Cool Machine Learning Model & Pretty Visualizations



[C. Bronner (left), Chicago Business (right)]

Lots of topics related to this

Finding & Discovering Data (even data you already have!)



[S. Dewan]

Data Wrangling

	A	B	C	D
1	**Transaction Date**	**Customer Name**	**Phone Numbers**	**Address**
2	Wed, 12 Jan 2011	John K. Doe Jr.	(609)-993-3001	2196 184th Ave. NE, Redmond, 98052
3	Thu, 15 Sep 2011	Mr. Doe, John	609.993.3001 ext 2001	4297 148th Avenue NE, Bellevue, 98007
4	Mon, 17 Sep 2012	Jane A. Smith	+1-4250013981	2720 N Mesa St, El Paso, 79902, USA
5	2010-Nov-30 11:10:41	MS. Jane Smith	425 001 3981	3524 W Shore Rd APT 1002, Warwick
6	2011-Jan-11 02:27:21	Smith, Jane	tel: 4250013981	4740 N 132nd St Apt 417, Omaha, 68164
7	2011-Jan-12	Anthony R Von Fange II	650-384-9911	10508 Prairie Ln, Oklahoma City
8	2010-Dec-24	Mr. Peter Tyson	(405)123-3981	525 1st St, Marysville, WA 95901
9	9/22/2011	Dan E. Williams	1-650-1234183	211 W Ridge Dr, Waukon,52172
10	7/11/2012	James Davis Sr.	+1-425-736-9999	13120 Five Mile Rd, Brainerd
11	2/12/2012	Mr. James J. Davis	425.736.9999 x 9	602 Highland Ave, Shinnston, 26431
12	3/31/2013	Donald Edward Miller	(206) 309-8381	840 W Star St, Greenville, 27834
13	6/1/2009 12:01	Miller, Donald	206 309 8381	25571 Elba, Redford, 48239
14	2/26/2007 18:37	Rajesh Krishnan	206 456 8500 extension 1	539 Co Hwy 48, Sikeston, USA
15	1/4/2011 14:33	Daniel Chen	425 960 3566	1008 Whitlock Ave NW, Marietta, 30064

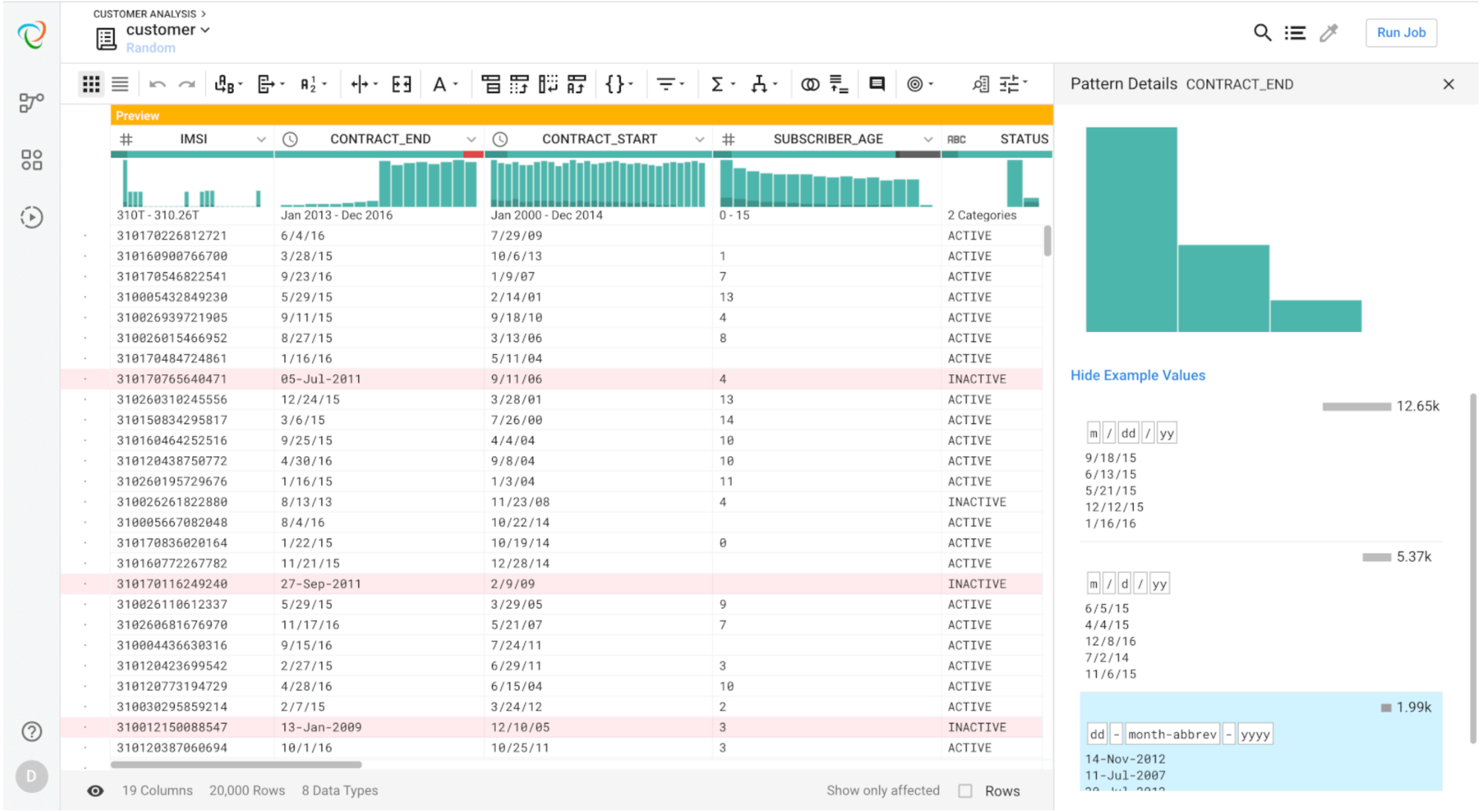
C	D
Transaction Date	**output**
Wed, 12 Jan 2011	2011-01-12-Wednesday
Thu, 15 Sep 2011	2011-09-15-Thursday
Mon, 17 Sep 2012	2012-09-17-Monday
2010-Nov-30 11:10:41	2010-11-30-Tuesday
2011-Jan-11 02:27:21	2011-01-11-Tuesday
2011-Jan-12	2011-01-12-Wednesday
2010-Dec-24	2010-12-24-Friday
9/22/2011	2011-09-22-Thursday
7/11/2012	2012-07-11-Wednesday
2/12/2012	2012-02-12-Sunday

C	D
Customer Name	**Output**
John K. **Doe** Jr.	Doe, John
Mr. **Doe**, **John**	Doe, John
Jane A. **Smith**	Smith, Jane
MS. **Jane Smith**	Smith, Jane
Smith, **Jane**	Smith, Jane
Dr **Anthony** R **Von Fange** III	Von Fange, Anthony
Peter Tyson	Tyson, Peter
Dan E. **Williams**	Williams, Dan
James Davis Sr.	Davis, James
James J. **Davis**	Davis, James
Mr. **Donald** Edward **Miller**	Miller, Donald

C	D
Address	**Output**
2196 184th Ave. NE Apt 417, **Redmond, 98052**	Redmond, WA, 98052
4297 148th Avenue NE L105, **Bellevue, WA 98007**	Bellevue, WA, 98007
2720 N Mesa St, **El Paso, 79902, USA**	El Paso, TX, 79902
3524 W Shore Rd APT 1002, **Warwick,02886**	Warwick, RI, 02886
4740 N 132nd St, **Omaha, 68164**	Omaha, NE, 68164
10508 Prairie Ln, **Oklahoma City**	Oklahoma City, OK, 73162
525 1st St, **Marysville, WA 95901**	Marysville, CA, 95901
211 W Ridge Dr, **Waukon,52172**	Waukon, IA, 52172
602 Highland Ave, **Shinnston, 26431**	Shinnston, WV, 26431
840 W Star St, **Greenville, 27834**	Greenville, NC, 27834

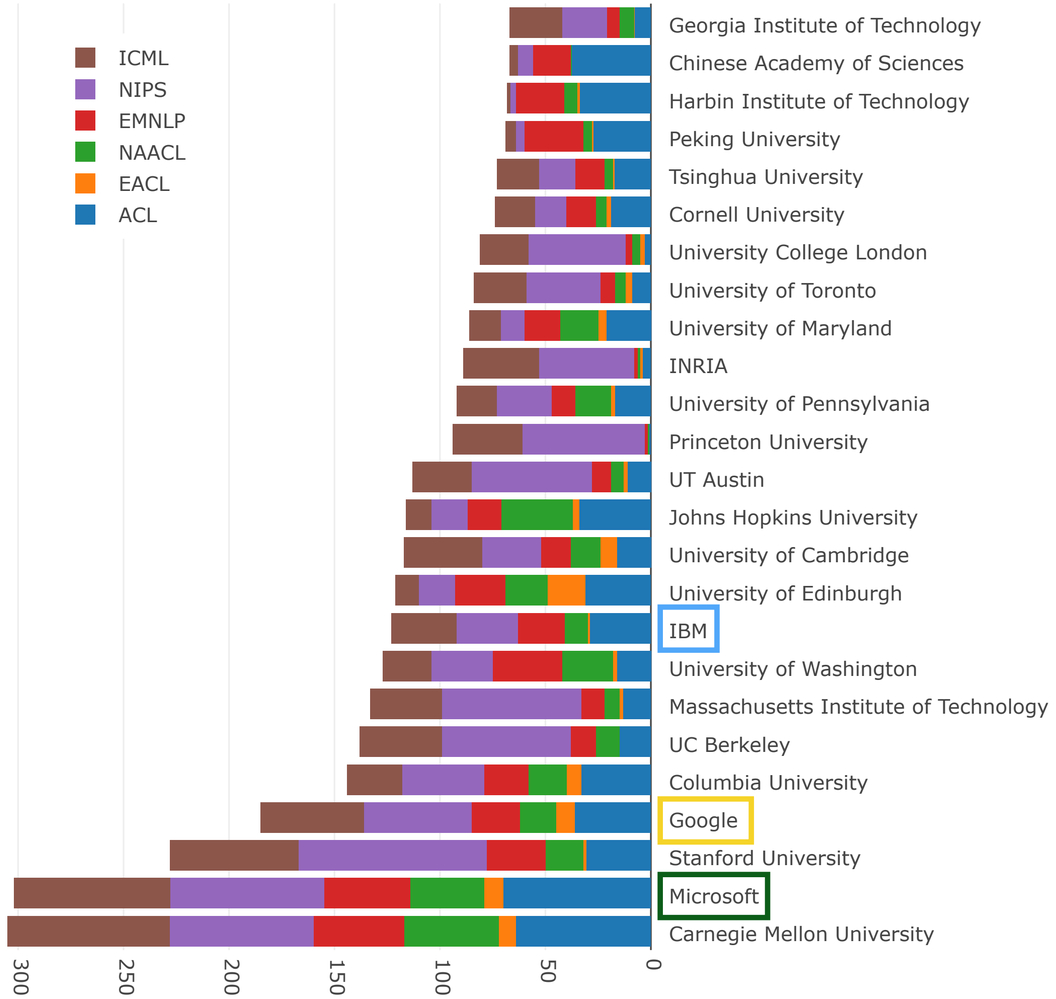
[Y. He et al., 2018]

Data Wrangling



[Trifacta]

Data Cleaning/Standardization (Aliases)




```
'google brain resident': 'google',
'google brain': 'google',
'google inc': 'google',
'google inc.':'google',
'google research nyc': 'google',
'google research': 'google',
'google, inc.': 'google',
'deepmind @ google': 'deepmind',
'deepmind technologies': 'deepmind',
'google deepmind': 'deepmind',
```

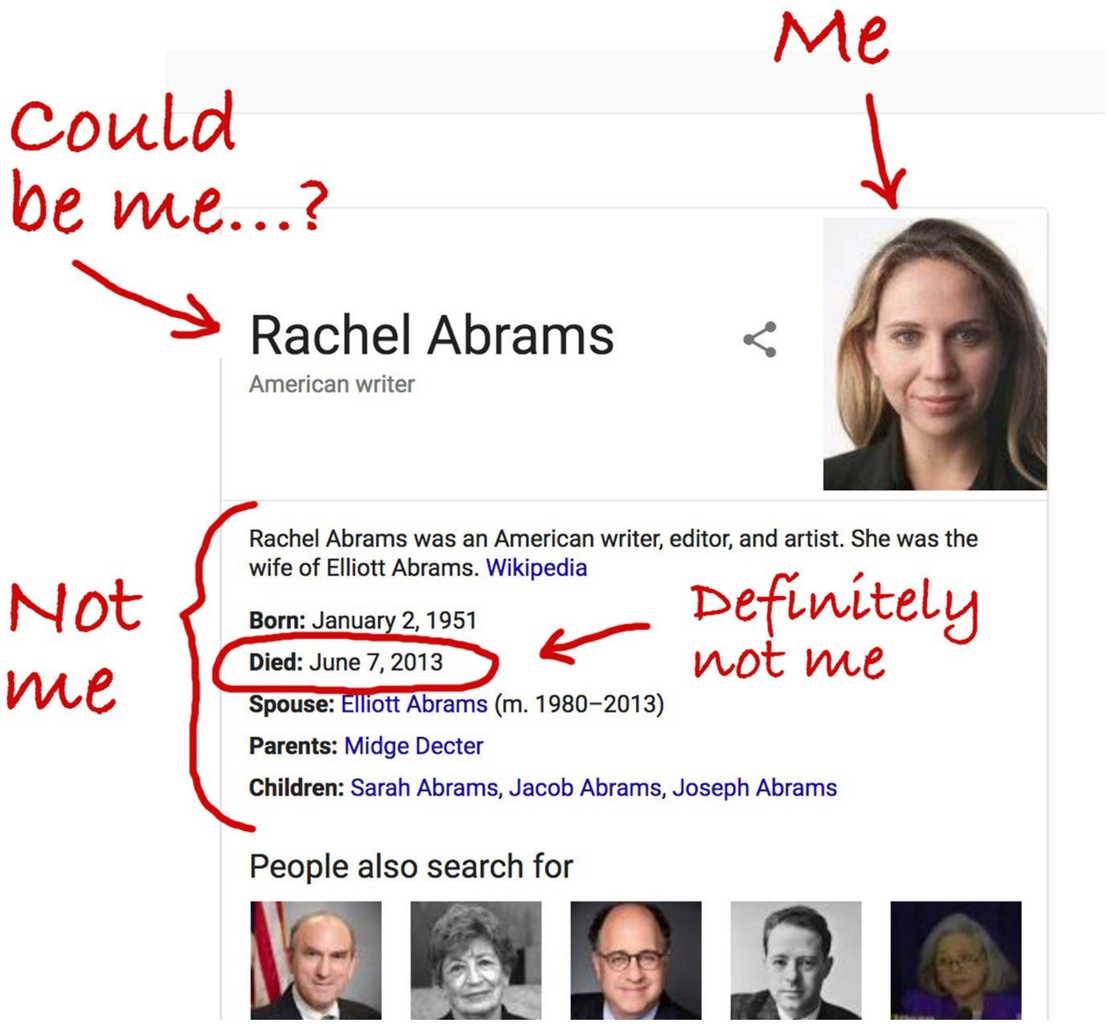
```
'ibm research - china':'ibm',
'ibm research':'ibm',
'ibm research, ny':'ibm',
'ibm research, usa':'ibm',
'ibm t. j. watson research center':'ibm',
'ibm t. j. watson research':'ibm',
'ibm t.j watson research center':'ibm',
'ibm t.j. watson research center':'ibm',
'ibm t.j.watson research center':'ibm',
'ibm thomas j. watson research center':'ibm',
'ibm tj watson research center':'ibm',
```

```
'microsoft research cambridge':'microsoft',
'microsoft research india':'microsoft',
'microsoft research maluuba':'microsoft',
'microsoft research new england':'microsoft',
'microsoft research':'microsoft',
'microsoft research, redmond, w':'microsoft',
'microsoft research, redmond, wa':'microsoft',
'miicrosoft research':'microsoft',
```

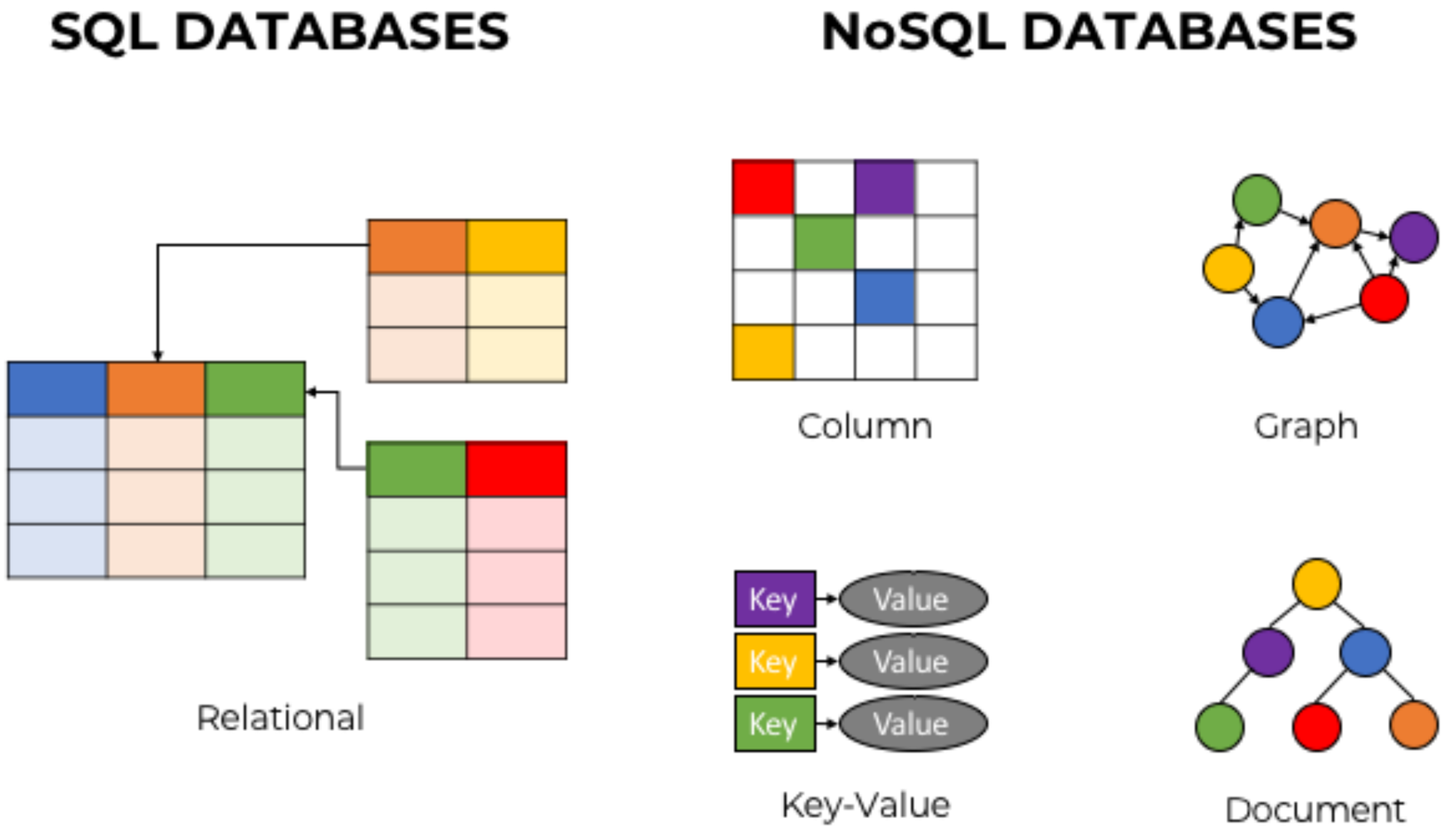
[NLP Publishing Stats, M. Rei & R. Allen]

Data Integration

- Google Thinks I'm Dead (I know otherwise.) [R. Abrams, NYTimes, 2017]
- Not only Google, but also Alexa:
  - "Alexa replies that Rachel Abrams is a sprinter from the Northern Mariana Islands (which is true of someone else)."
  - "He asks if Rachel Abrams is deceased, and Alexa responds yes, citing information in the Knowledge Graph panel."

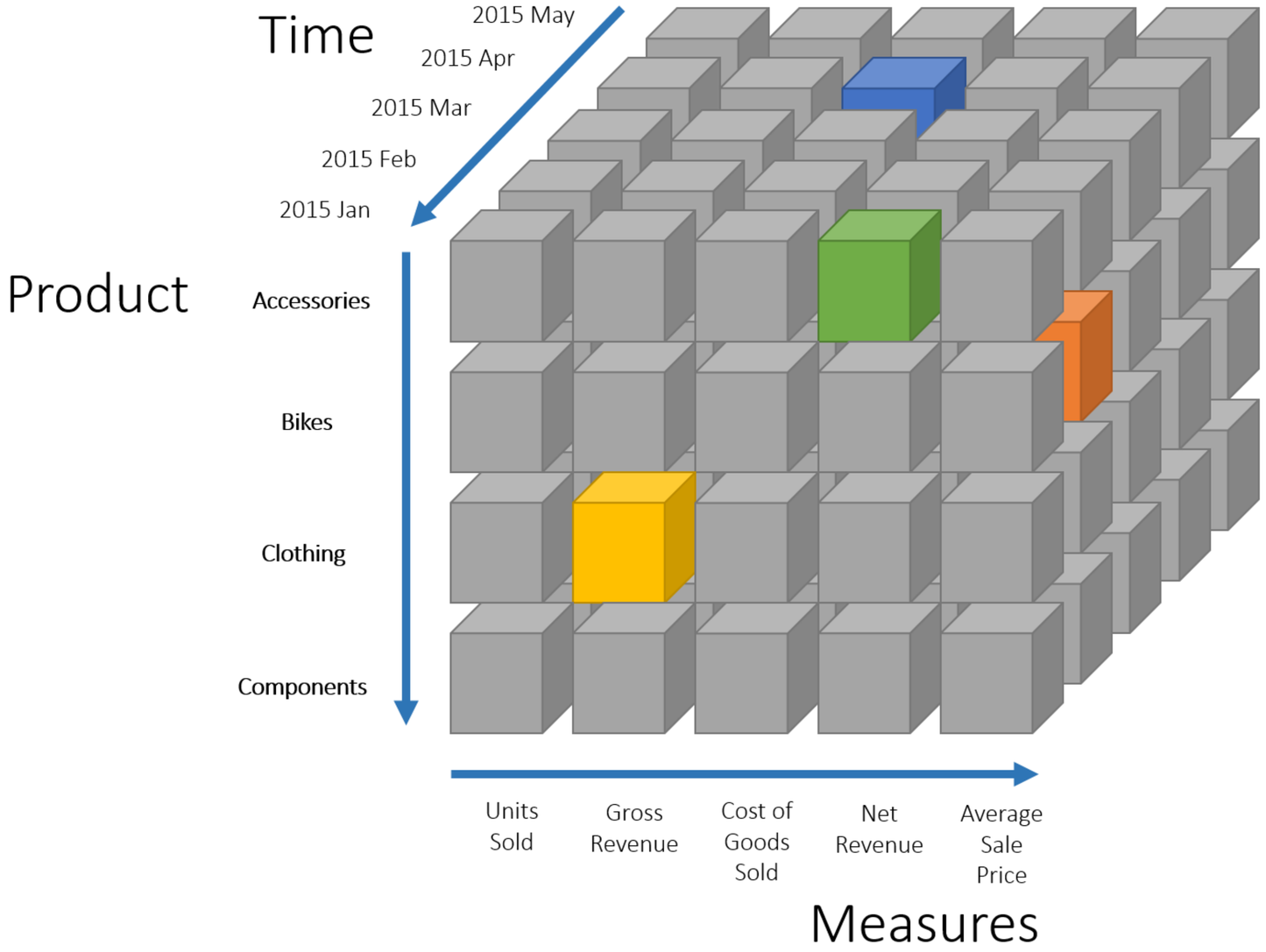


Data Storage



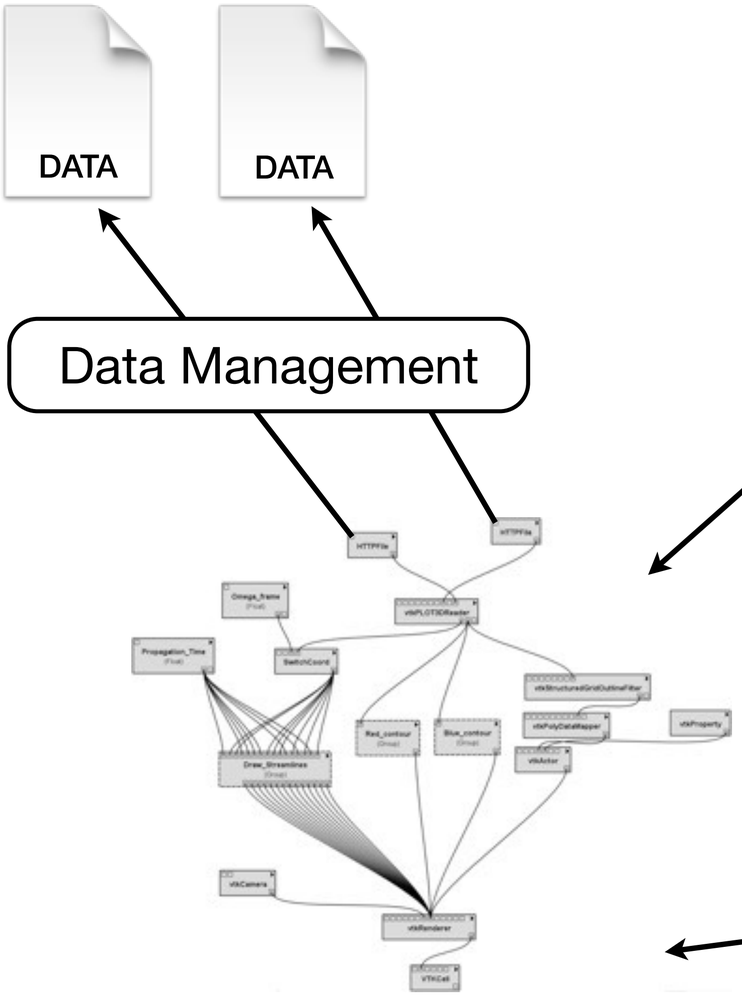
[V. Wilkinson]

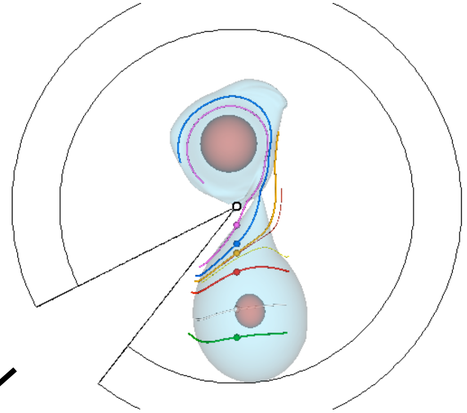
Data Cubes



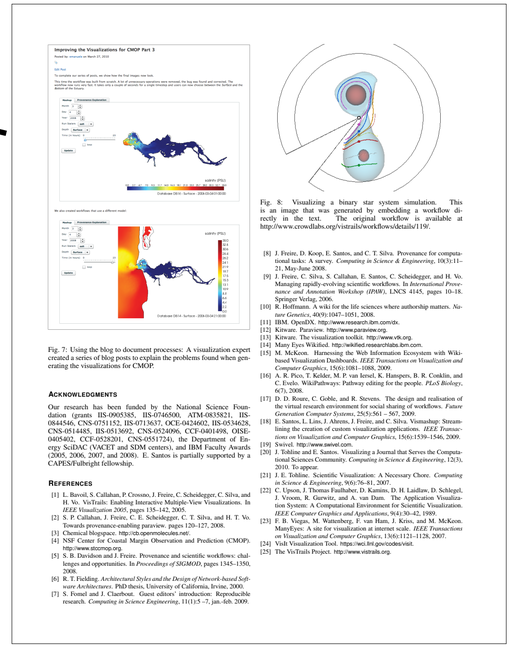
[M. K. Hernandez]

Provenance and Reproducibility



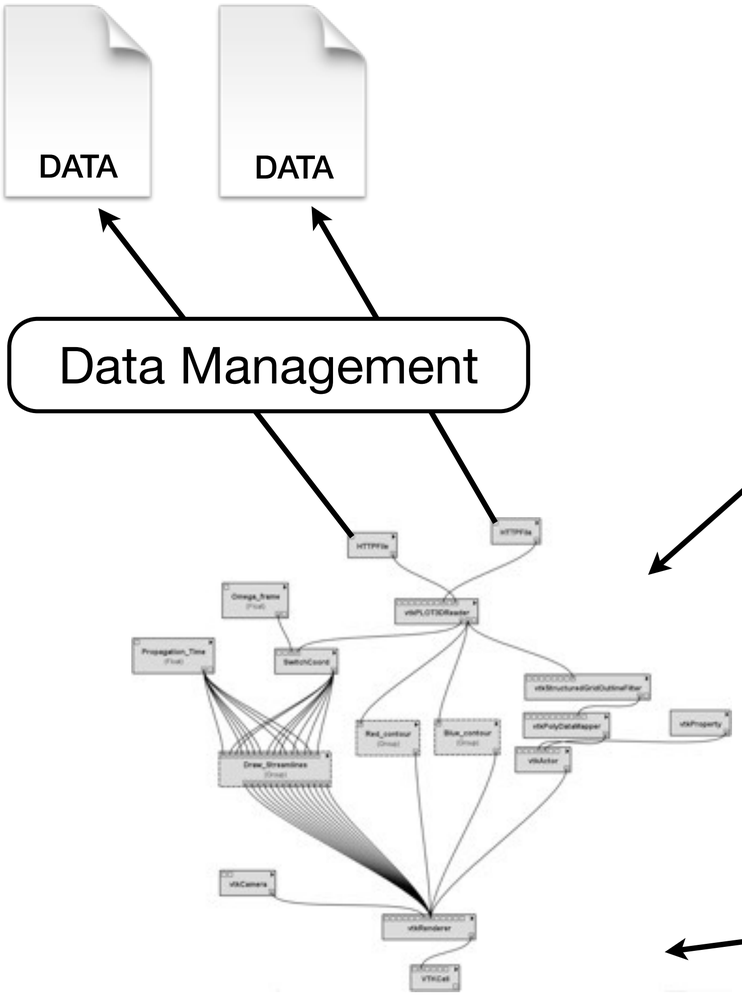


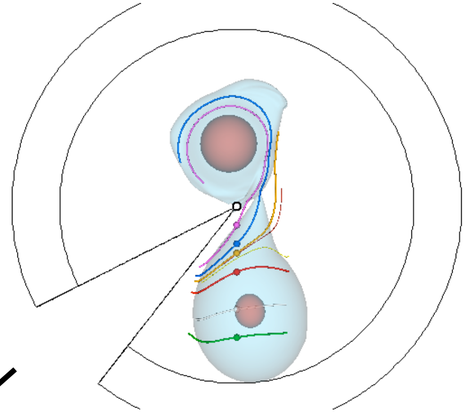
Visualization



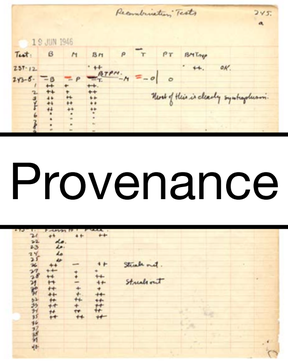
Paper

Provenance and Reproducibility

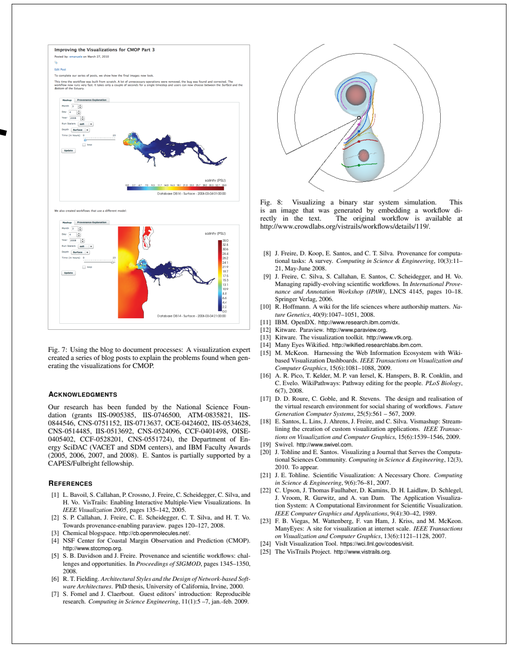




Visualization



Provenance



Paper

About Me

- Research Interests
  - Visualization
  - Computational Provenance
  - Geospatial Analysis
- Research Projects
  - Dataflow Notebooks
  - Geospatial Trajectory Data
  - Provenance for Web Applications
- See my web page for more information
  - http://faculty.cs.niu.edu/~dakoop/

About You

- Research Papers?
- Data Science?
- Python?
- Database Experience?
- Analytics Experience?
- Cloud Computing Experience?
- Anything you want to see covered?

About this course

- Course web page is authoritative:
  - http://faculty.cs.niu.edu/~dakoop/cs680-2021sp/
  - Schedule, Readings, Assignments will be posted online
  - Check the web site before emailing me
- Lectures via Zoom (link on Blackboard), recordings also available
- Course is meant to be more "cutting edge"
  - Still focus on building skills related to data management
  - Tune into current research and tools
- Requires student participation: readings and discussions

About this course

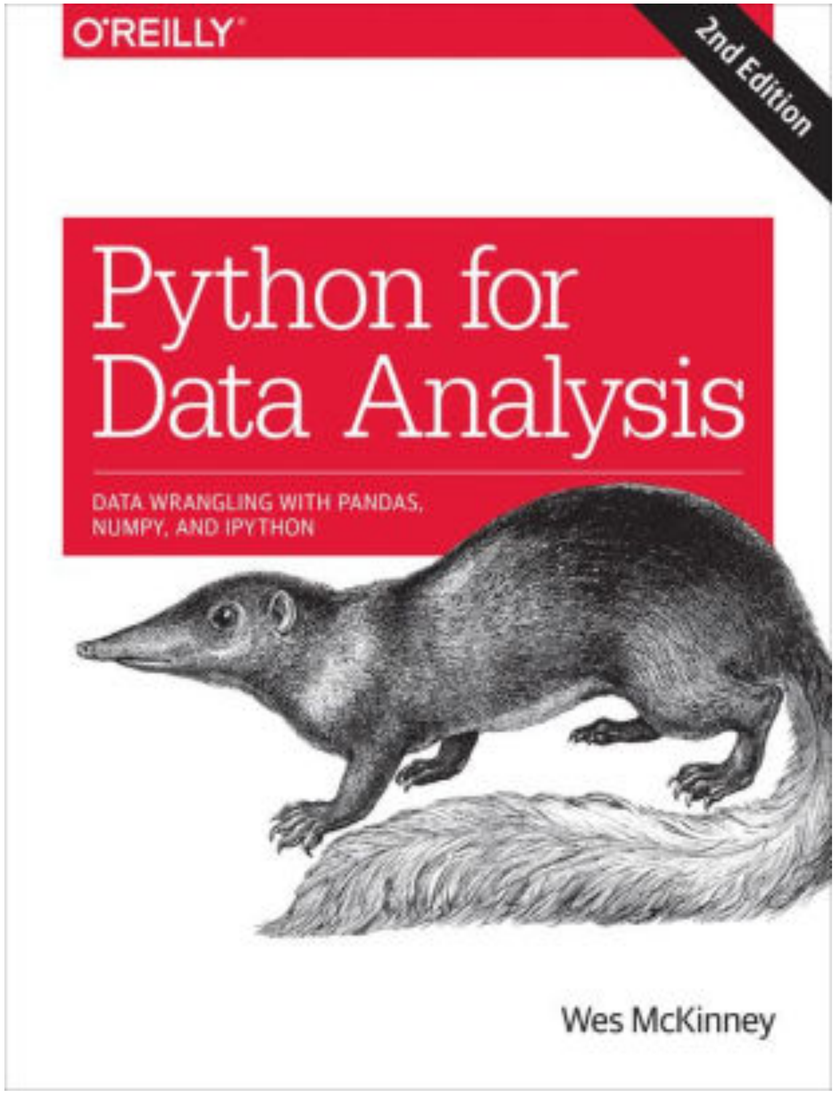
- Balance of techniques and research ideas
- Some background (Python) followed by topic areas and readings
- Programming assignments (~4)
- Two tests + final exam: Please check these dates now
- Topic areas:
  - Data Acquisition
  - Data Wrangling
  - Data Storage and Access
  - Cloud Storage and Scalable Data Management
  - Spatial, Graph, Time Series Data
  - Provenance and Reproducibility

About this course

- Course Registration:
  - Make sure you have registered for the course
  - Email me if you are not registered but are interested in taking the course
- Undergraduate (CS 490) and Graduate (CS 680)
  - Grad students have extra reading, exam questions, assignment tasks
- Review of course policies:
  - Plagiarism and academic honesty
  - If you have any concerns or questions, please email me as soon as possible
- If you are not sure if this course is a good fit, please email me or talk to me

Course Material

- Recommended: *Python for Data Analysis* by Wes McKinney, 2nd ed., 2017
  - Good reference for data science topics in Python
  - McKinney created the Pandas package
- Other texts:
  - Intro to Python, Deitel & Deitel
  - Python Data Science Handbook, J. VanderPlas
- Research papers
- Many websites

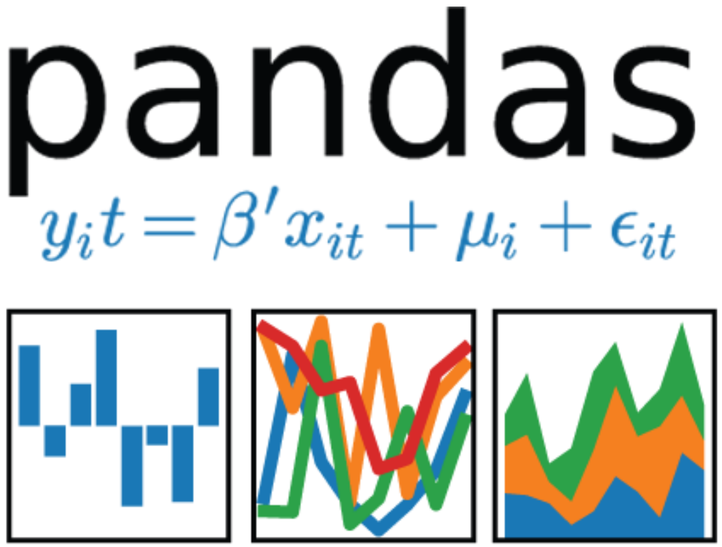


Course Material

- Software:
  - Anaconda Python Distribution (https://www.anaconda.com/distribution/): makes installing python and python packages easier
  - JupyterLab: Web-based interface for interactively writing and executing Python code
  - JupyterHub: Access everything through a server

Course Material

- Pandas:
  - Python library for data analysis
  - Many operations available
  - Efficient
- Trifacta Wrangler



TRIFACTA

Office Hours & Email

- Scheduled office hours are open to all students via Zoom
  - MW: 10:30am-11:30am, or by appointment (Prof. Koop's Office Hours)
- You do not need an appointment to zoom in during scheduled office hours
- If you need an appointment outside of those times, please email me with **specific details** about what you wish to discuss
- Many questions can be answered via email. Please **do not** schedule an appointment to ask a question that could be answered via email

Next Class

- Introduction to/review of Python
- Download and install anaconda distribution (Python 3.8):
  - https://www.anaconda.com/distribution/