

Advanced Data Management (CSCI 490/680)

Introduction

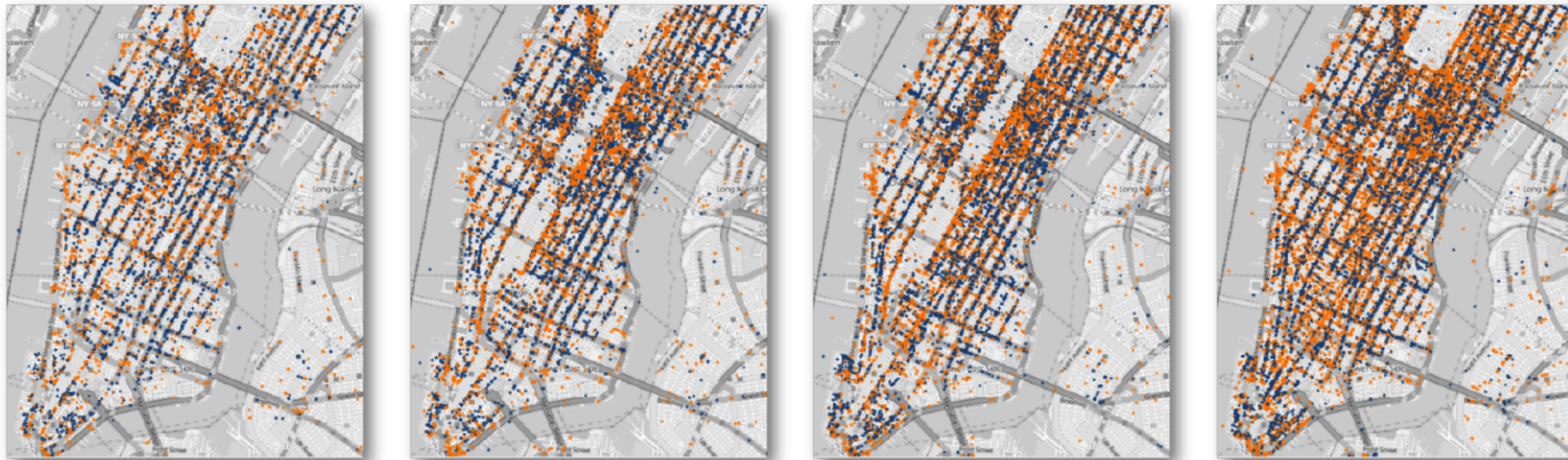
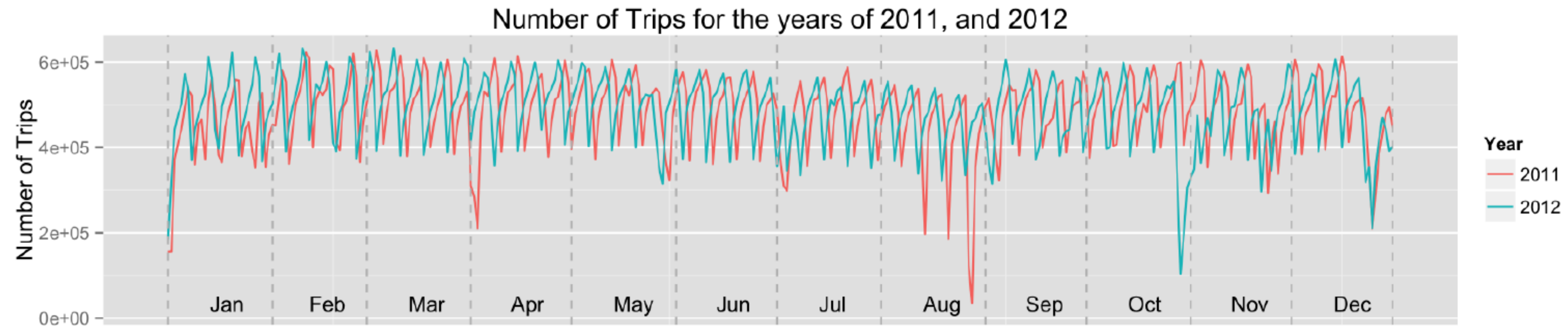
Dr. David Koop

NYC Taxi Data



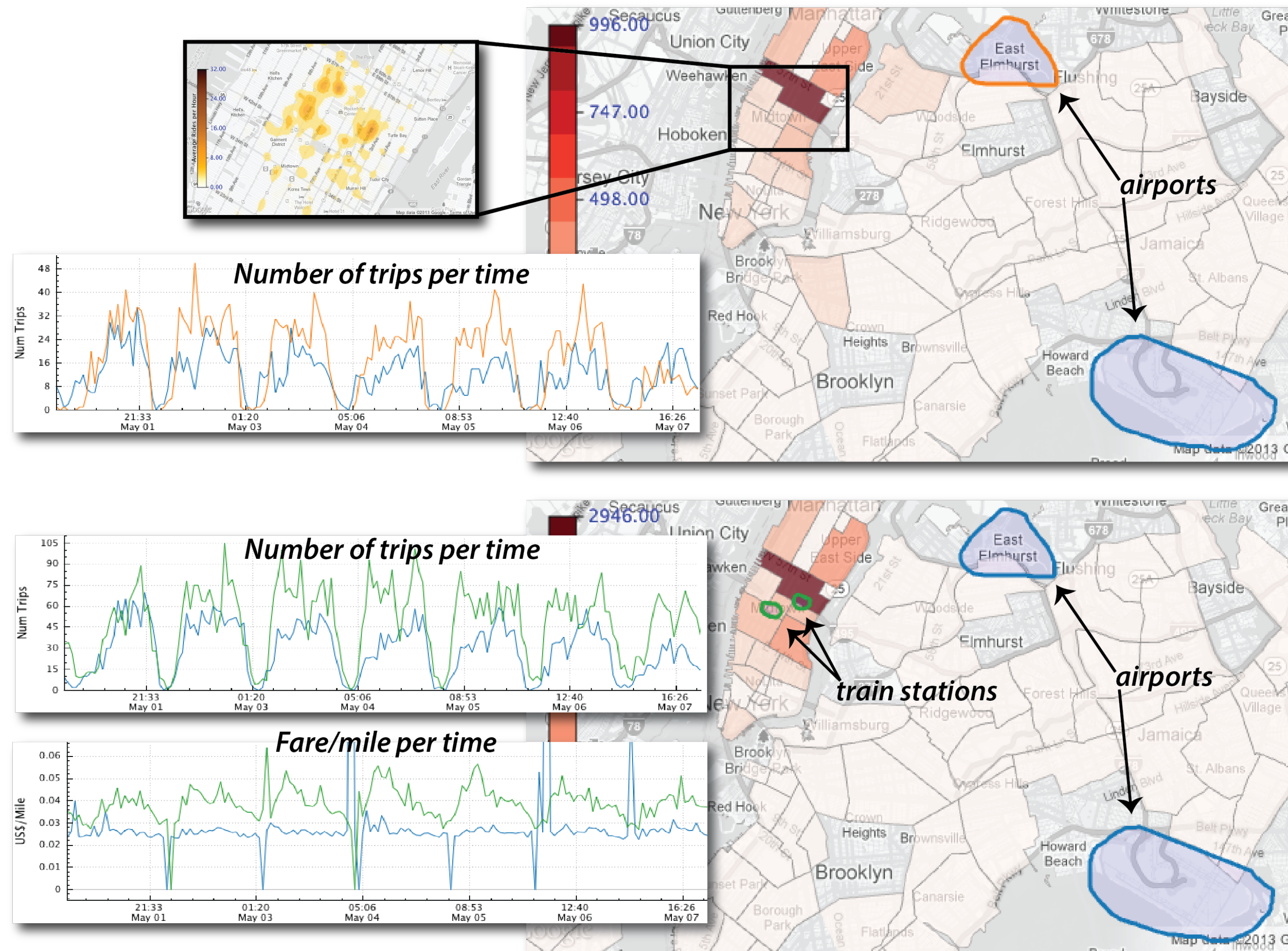
[Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance, T. W. Schneider]

NYC Taxi Data: Day analysis



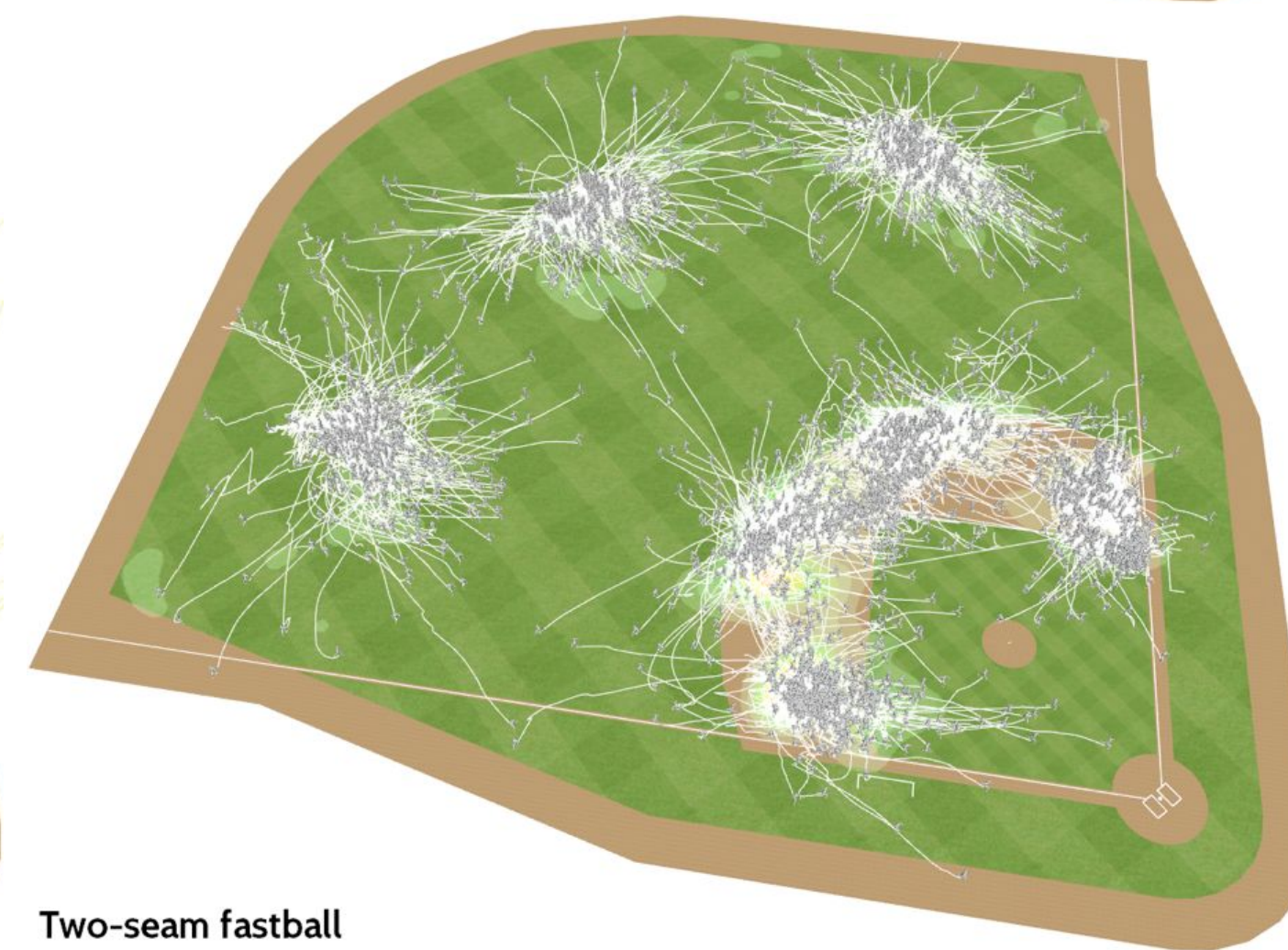
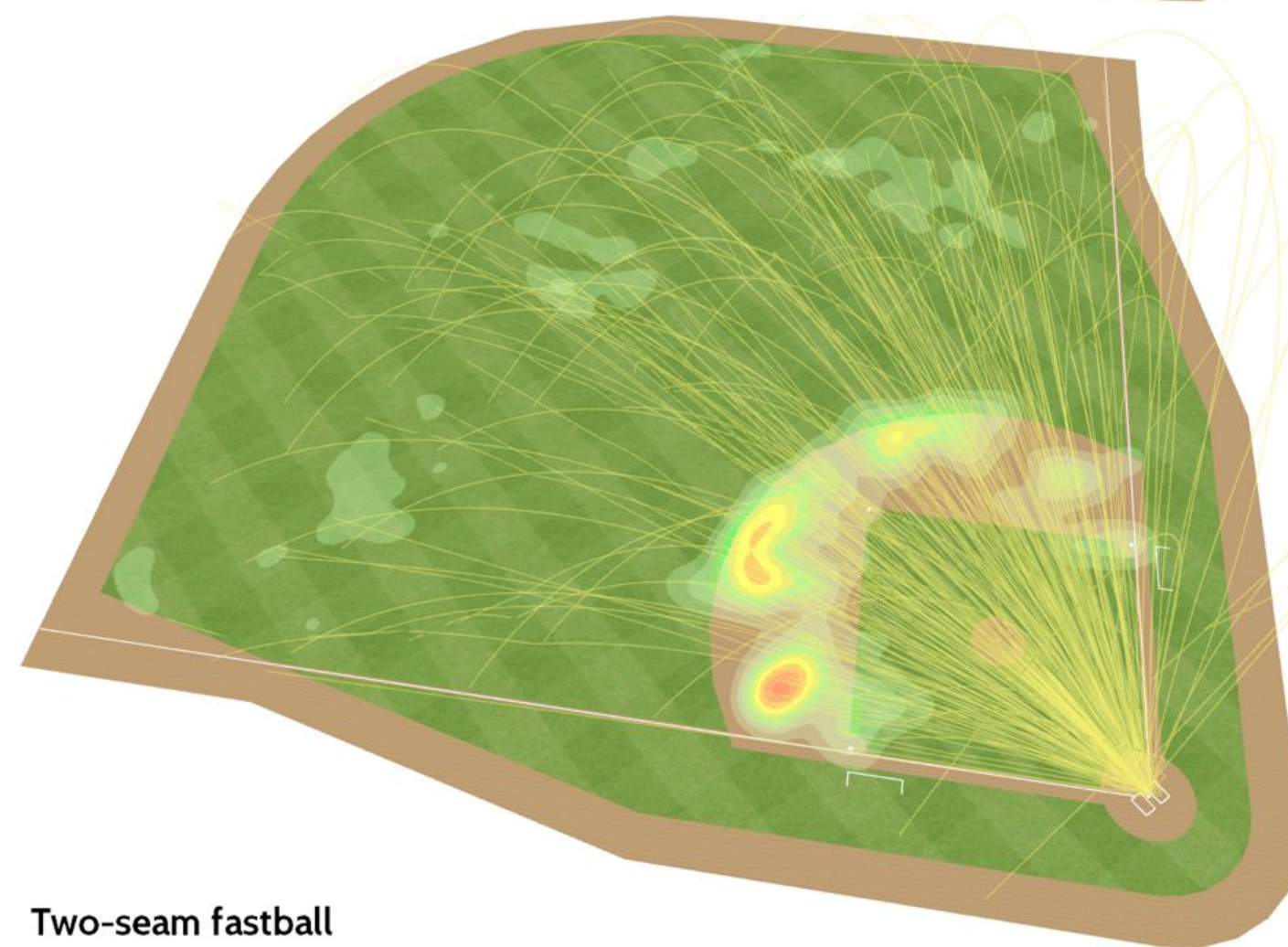
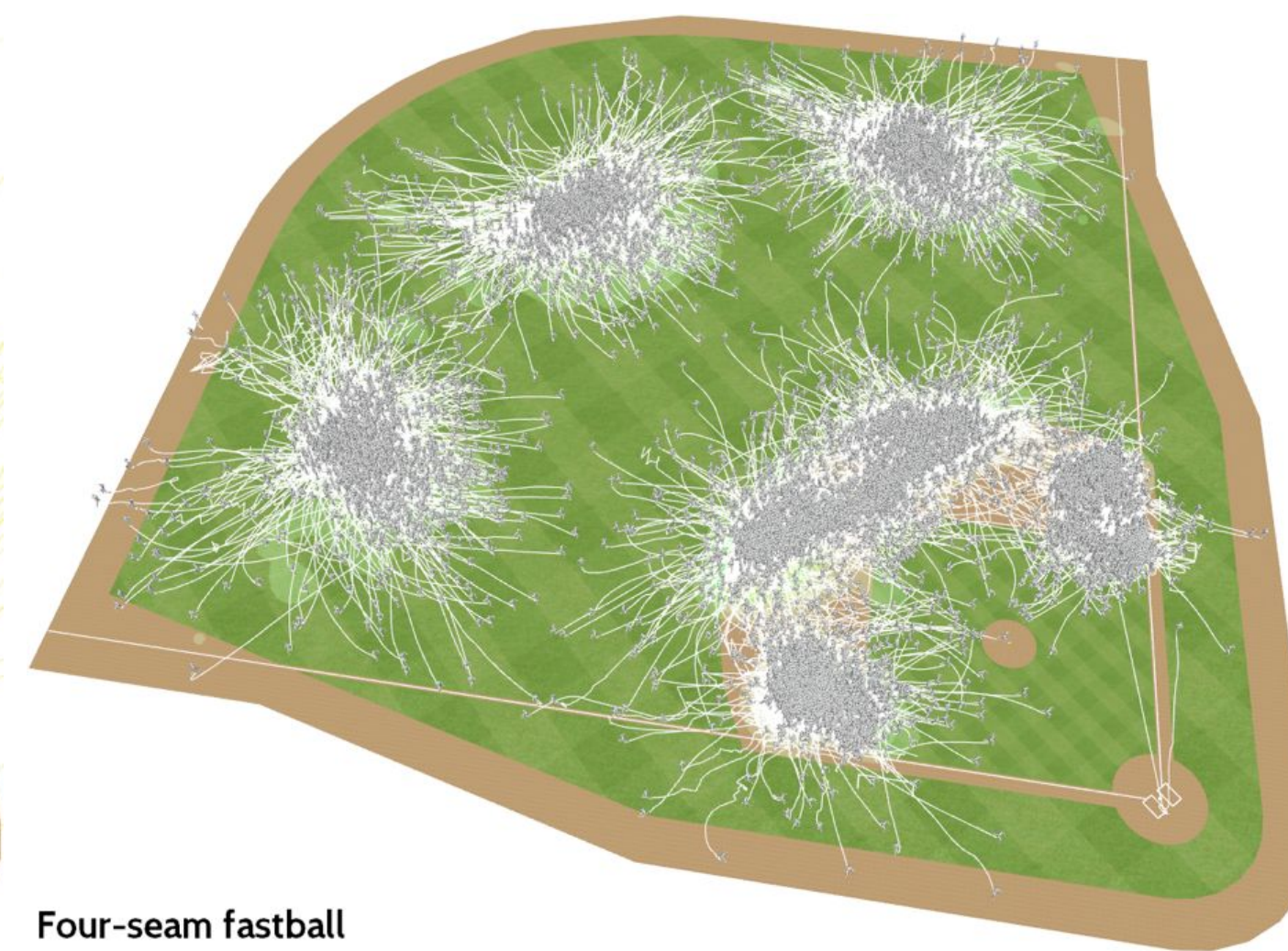
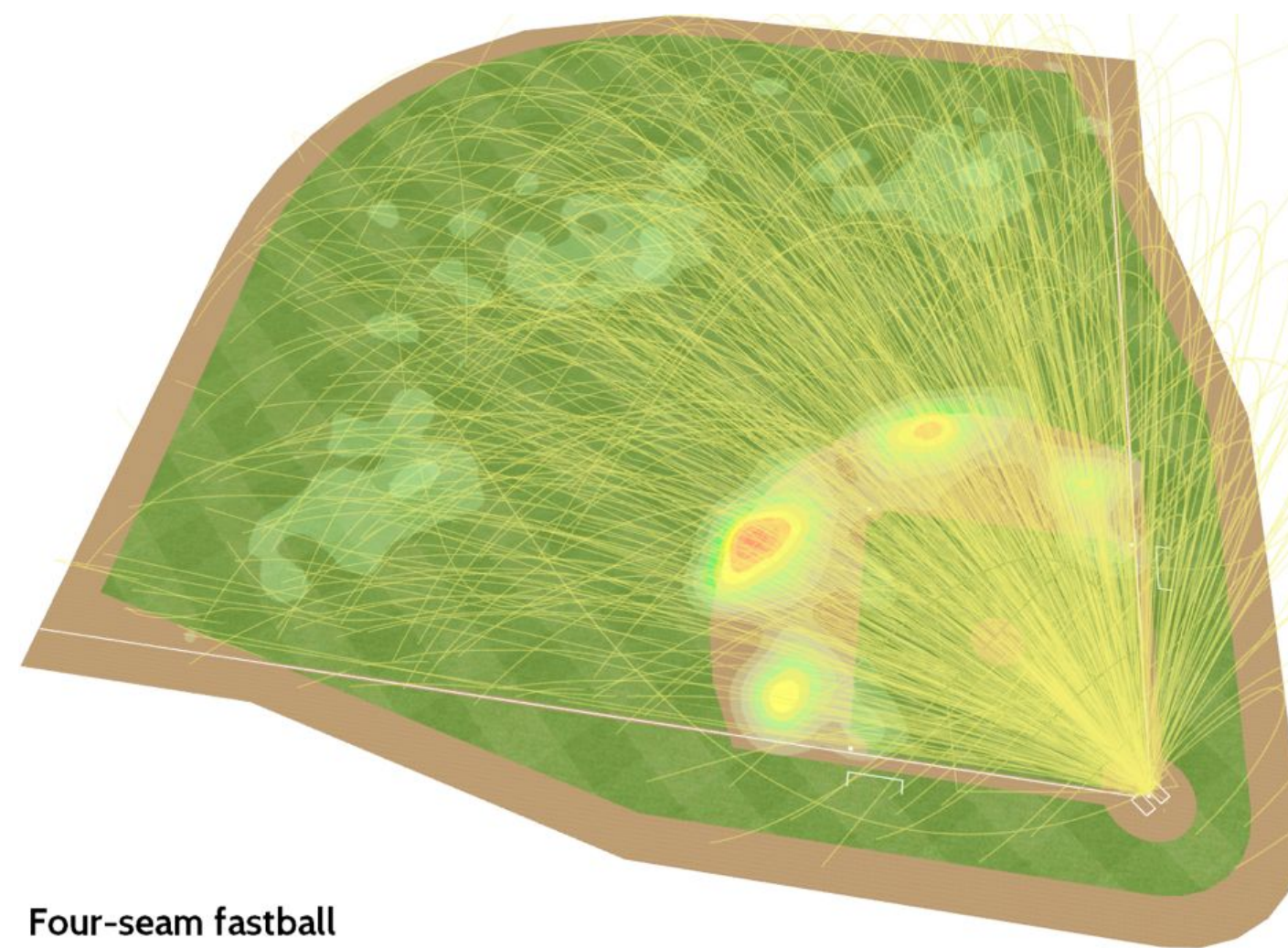
[Ferreira et al., 2013]

NYC Taxi Data: Region analysis



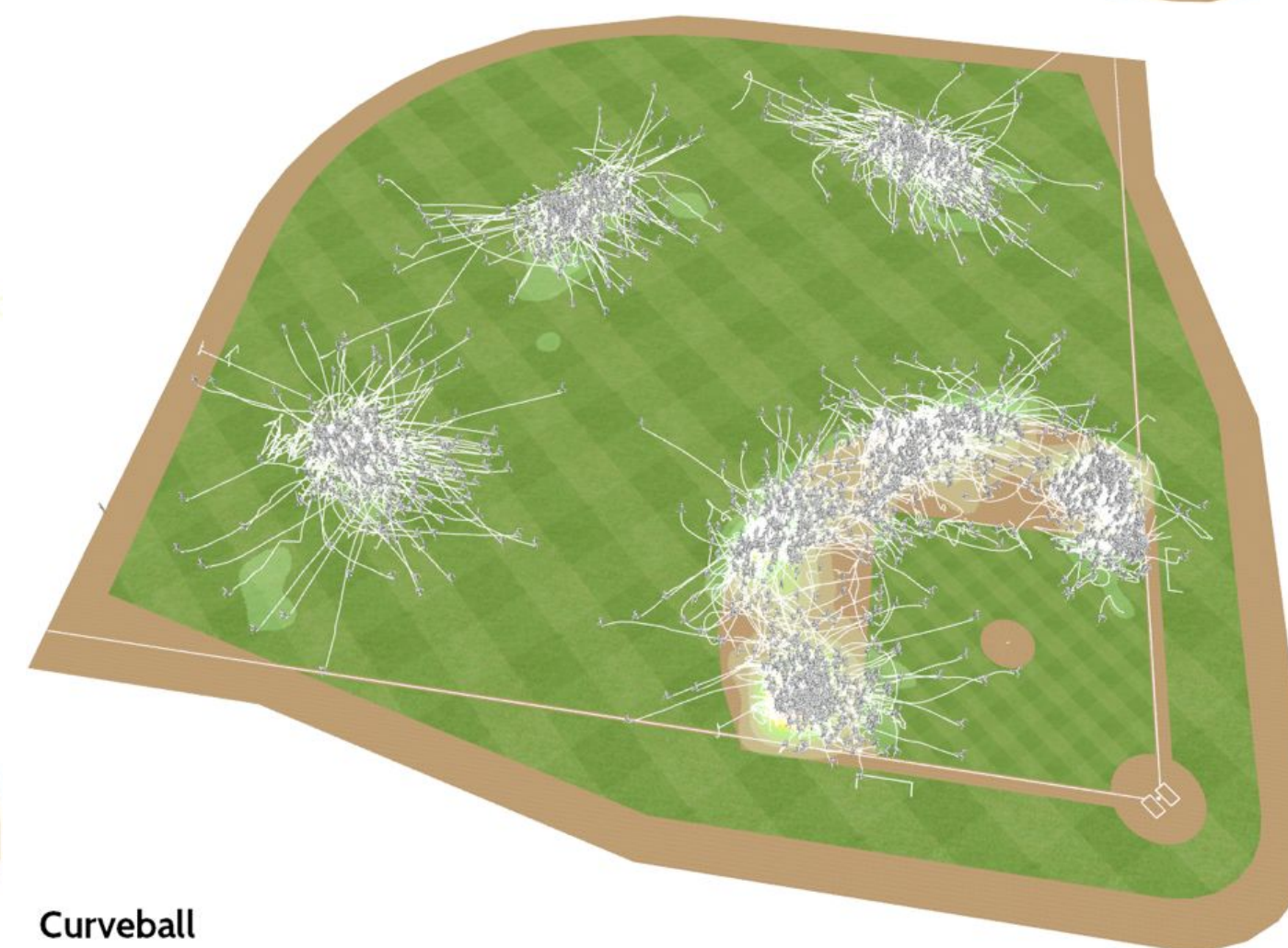
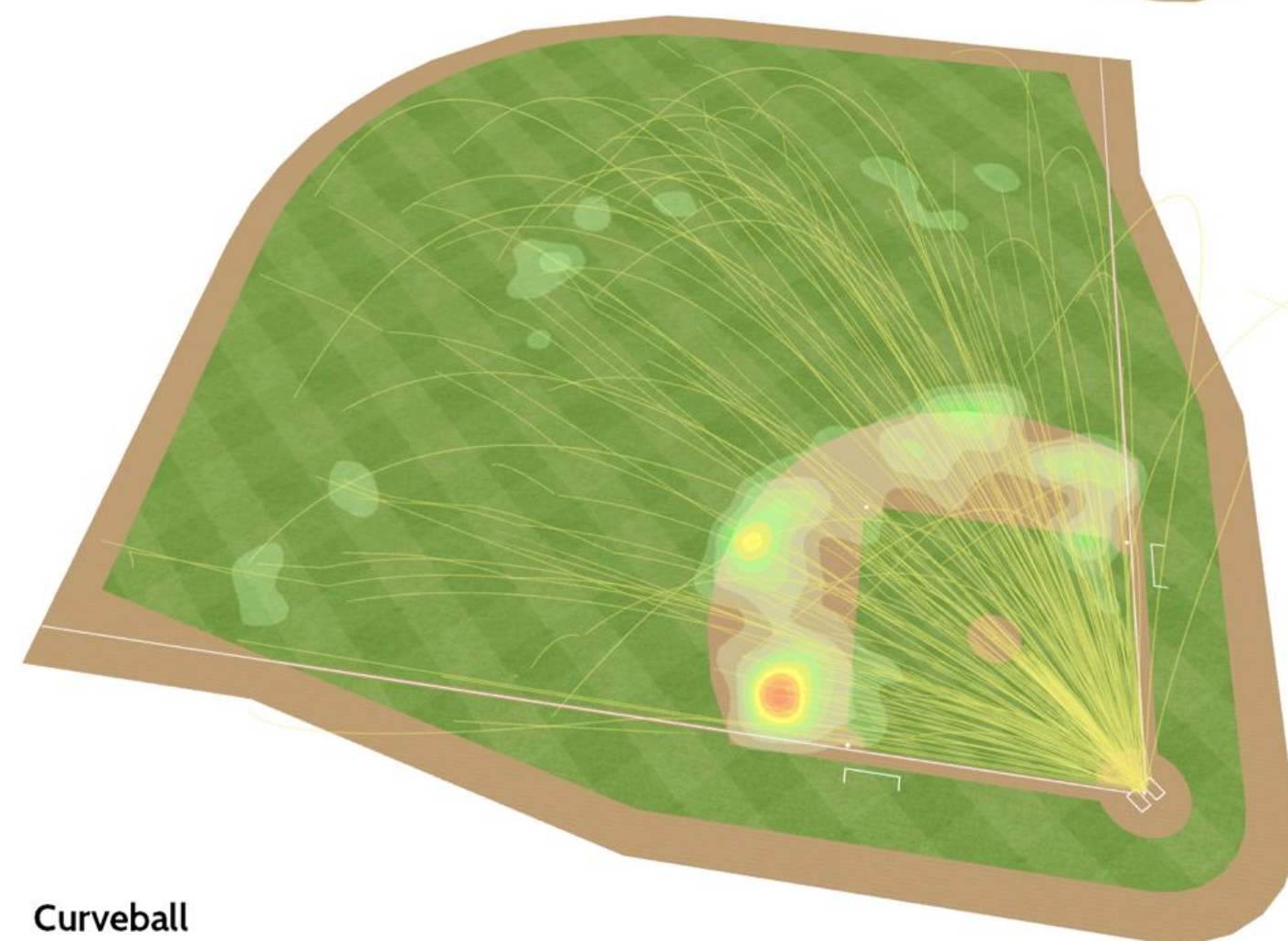
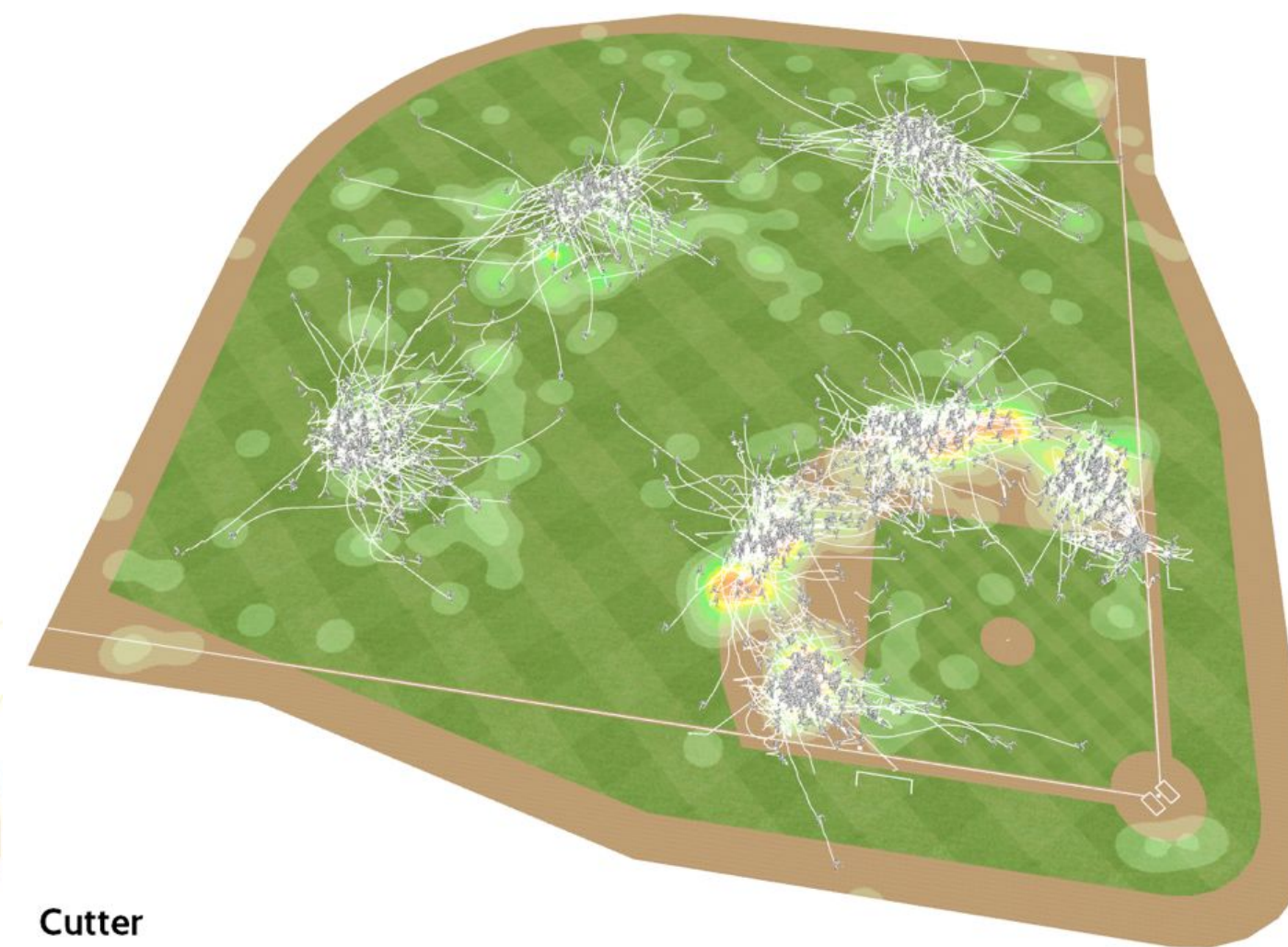
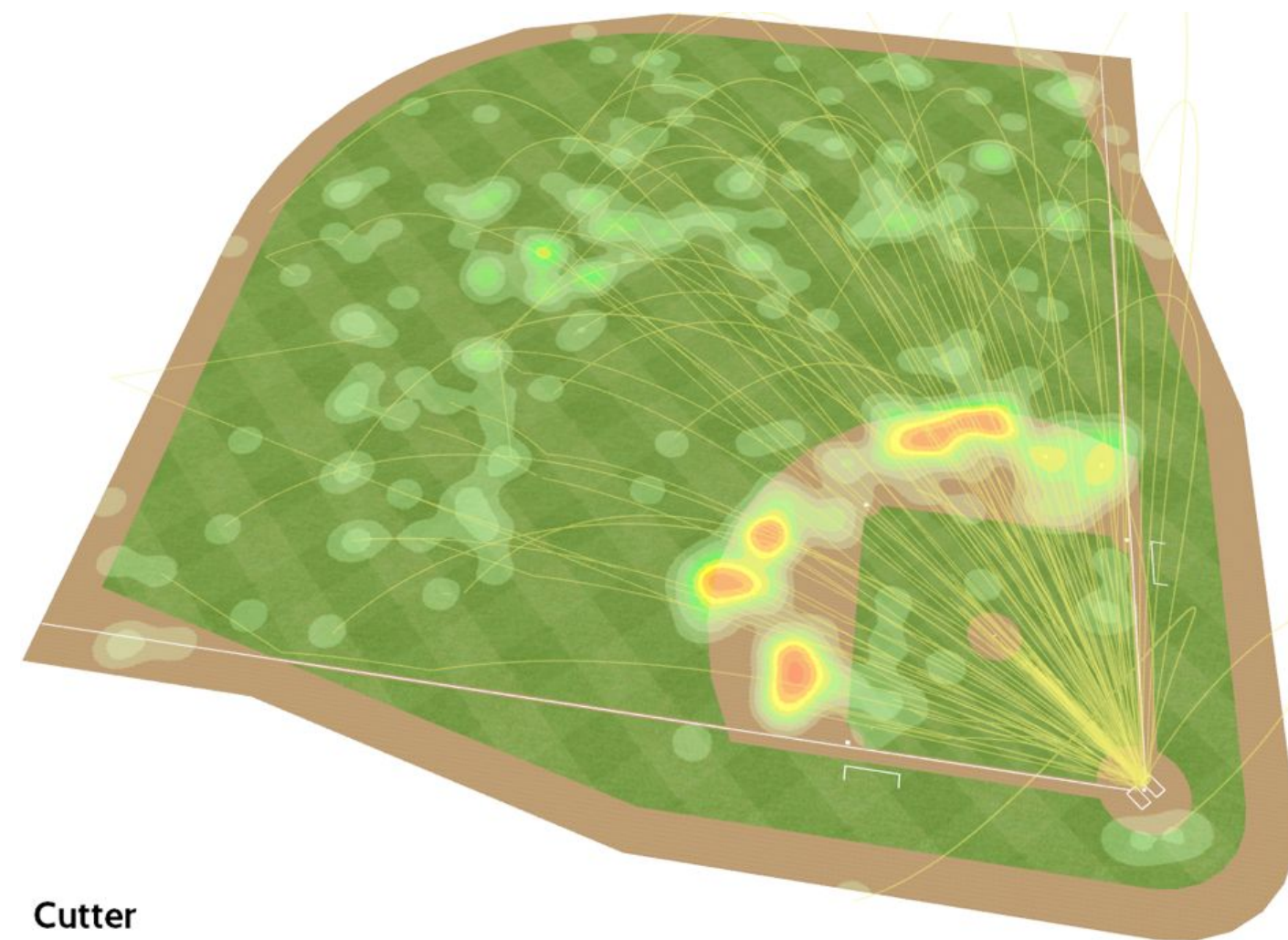
[Ferreira et al., 2013]

Baseball Data



[Deitrich et al., 2014]

Baseball Data

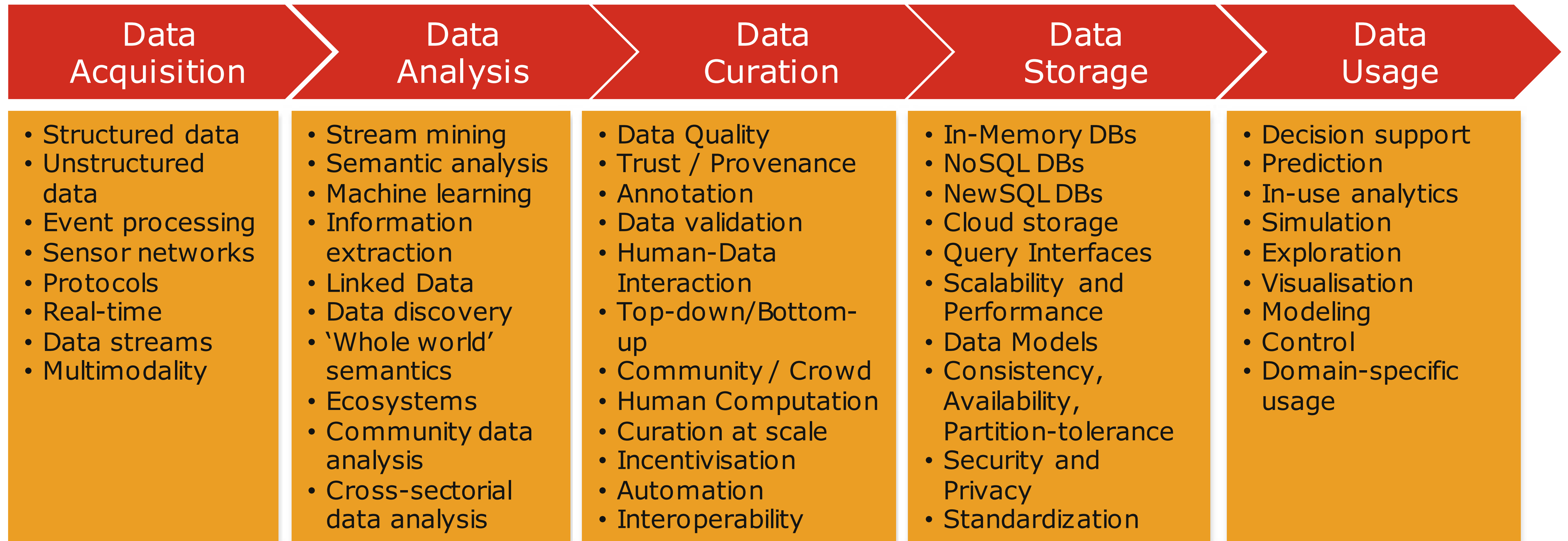


[Deitrich et al., 2014]

Real-time Analysis

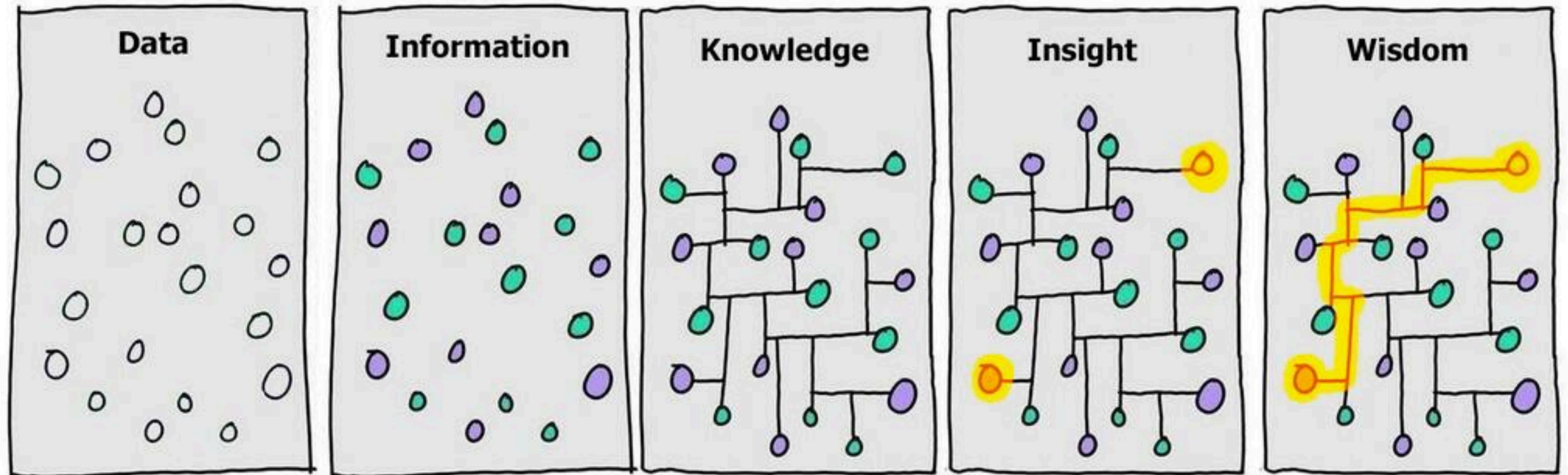
- Want to have results now
- How?
 - Faster machines
 - Clusters
 - Progressive techniques

What's involved in dealing with data?



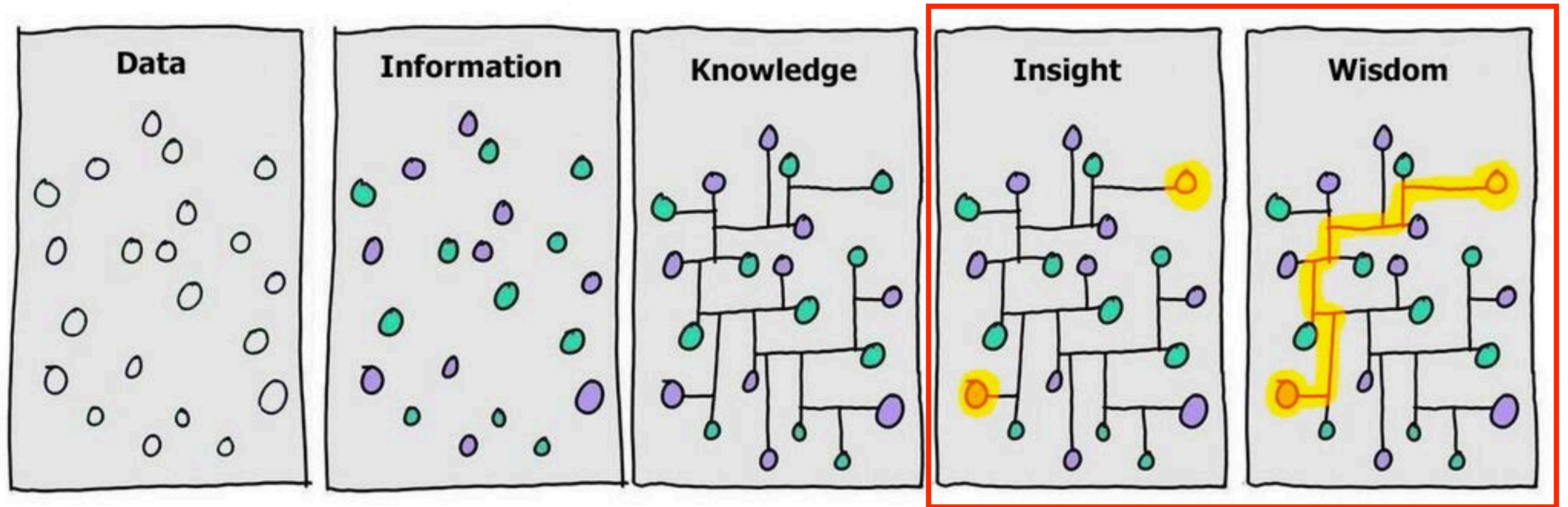
[Big Data Value Chain, Curry et al., 2014]

Data to Knowledge



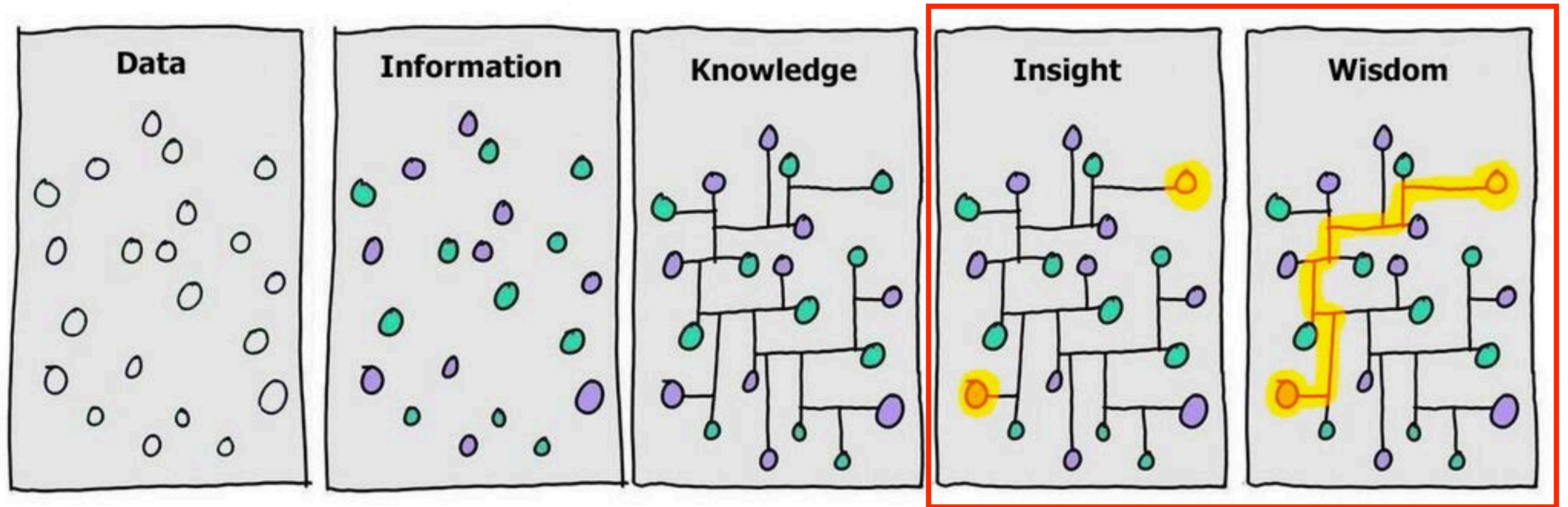
[D. Somerville, based on H. McLeod's original]

Data to Knowledge



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

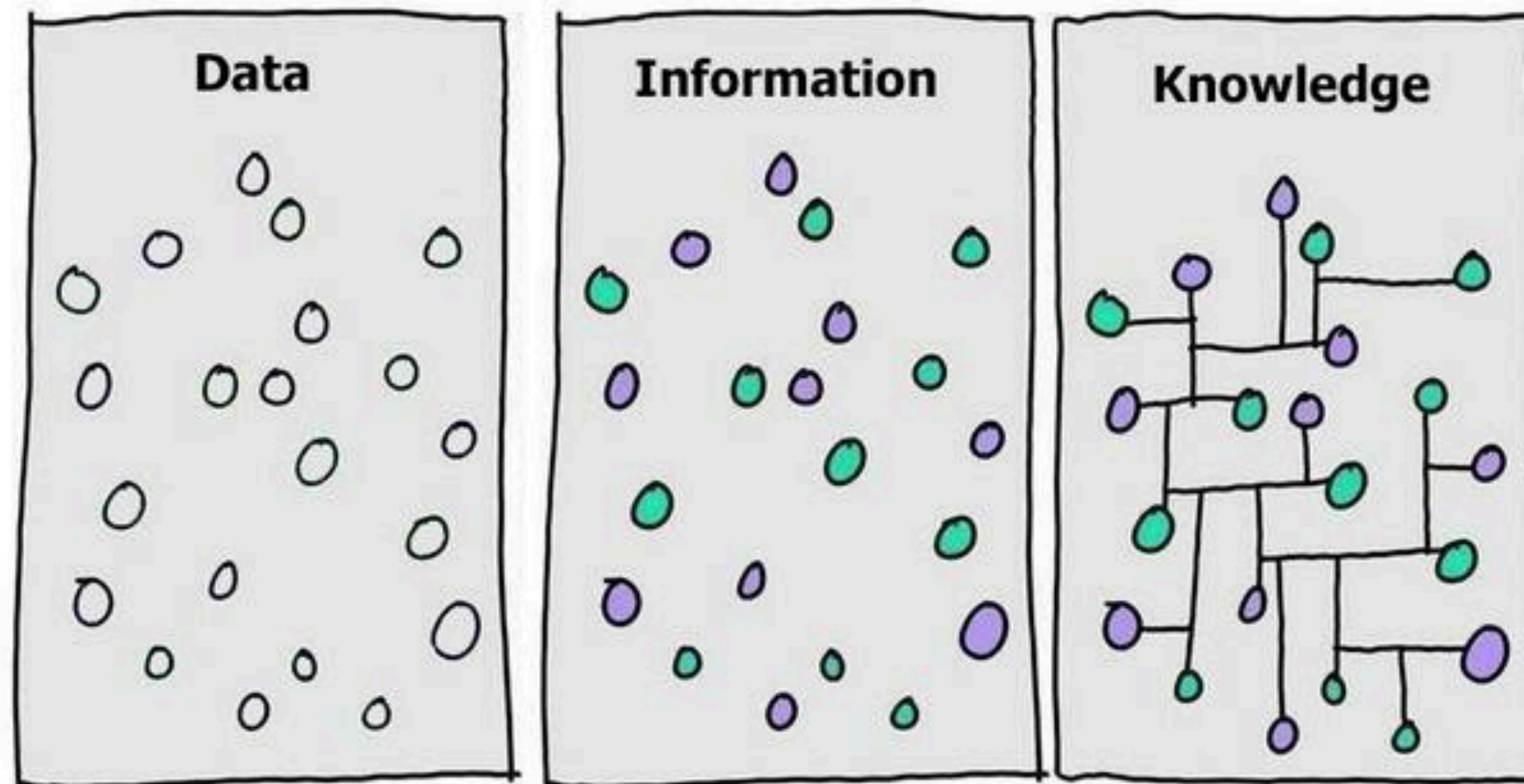
Data to Knowledge



Require People

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

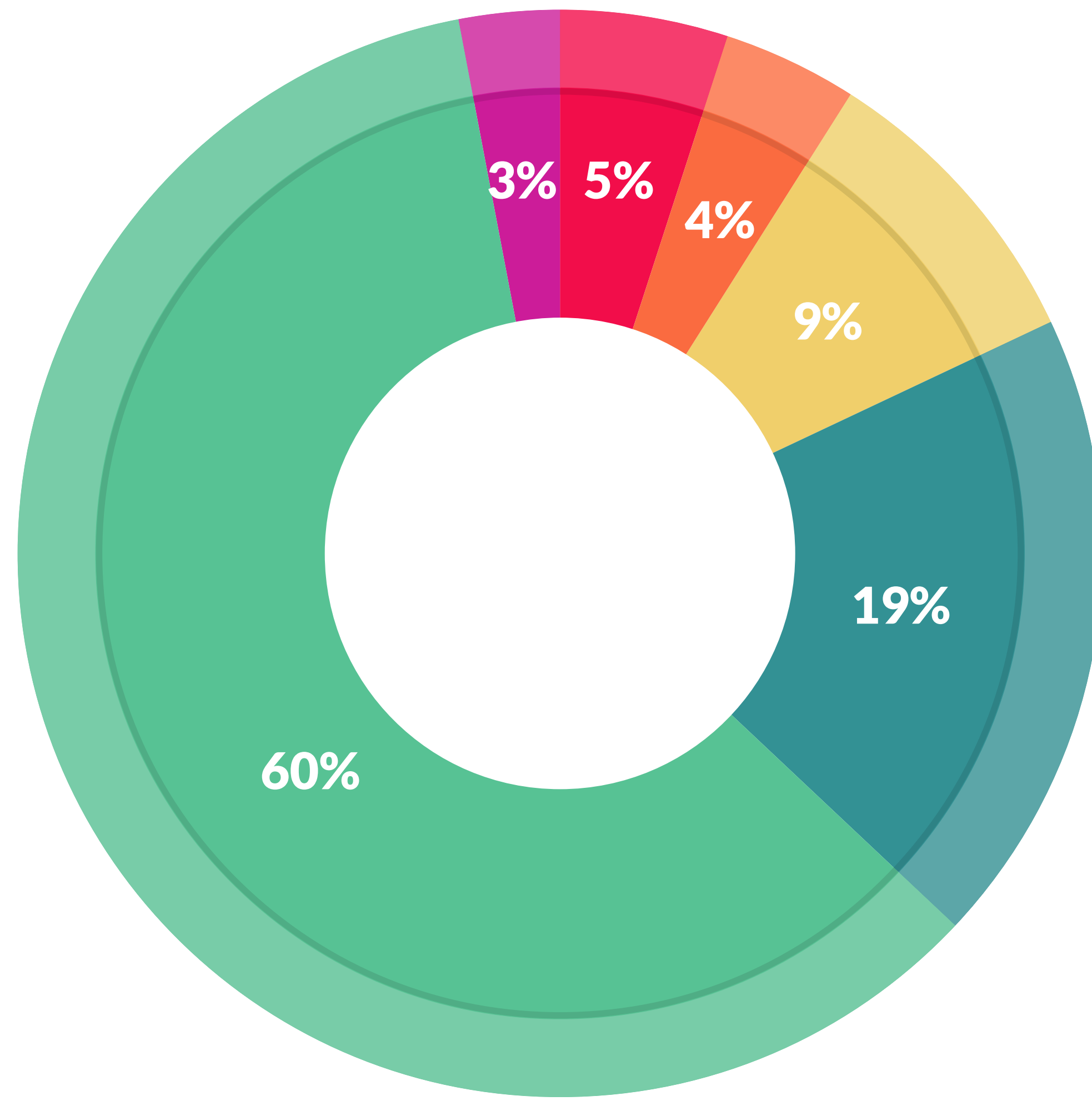
Data to Knowledge



Can computers do this for us?

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

How do data scientists spend their time?

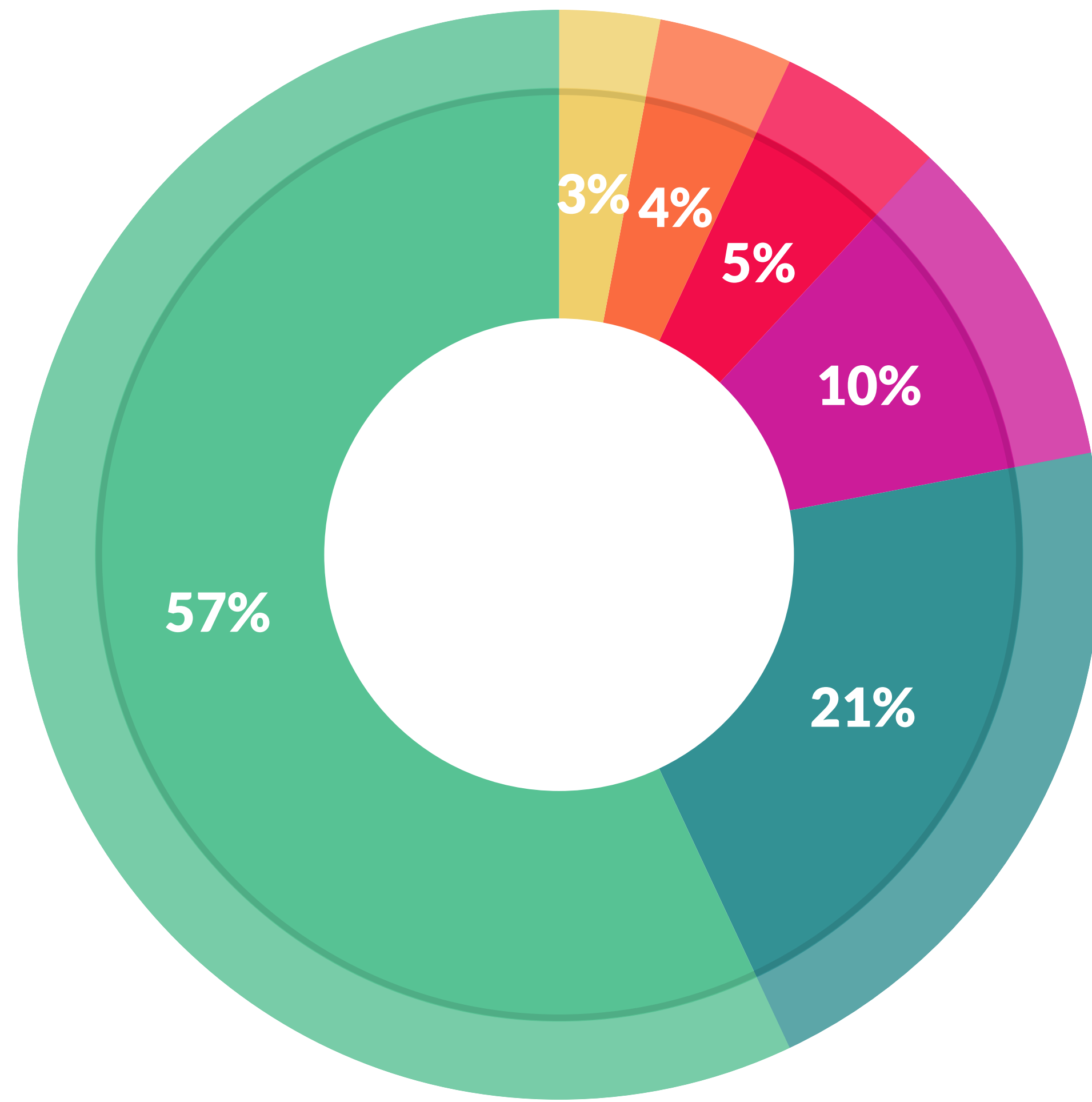


What data scientists spend the most time doing

- *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?

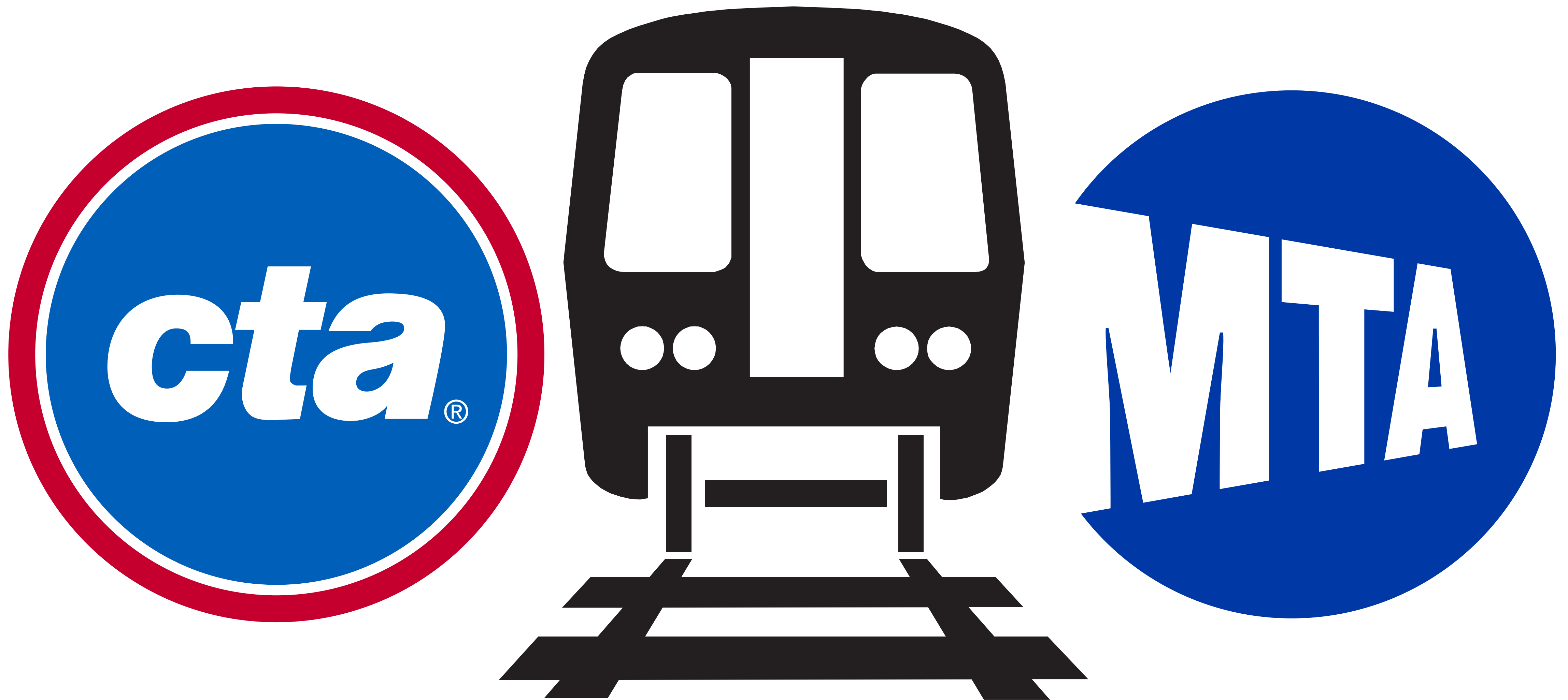


What's the least enjoyable part of data science?

- *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

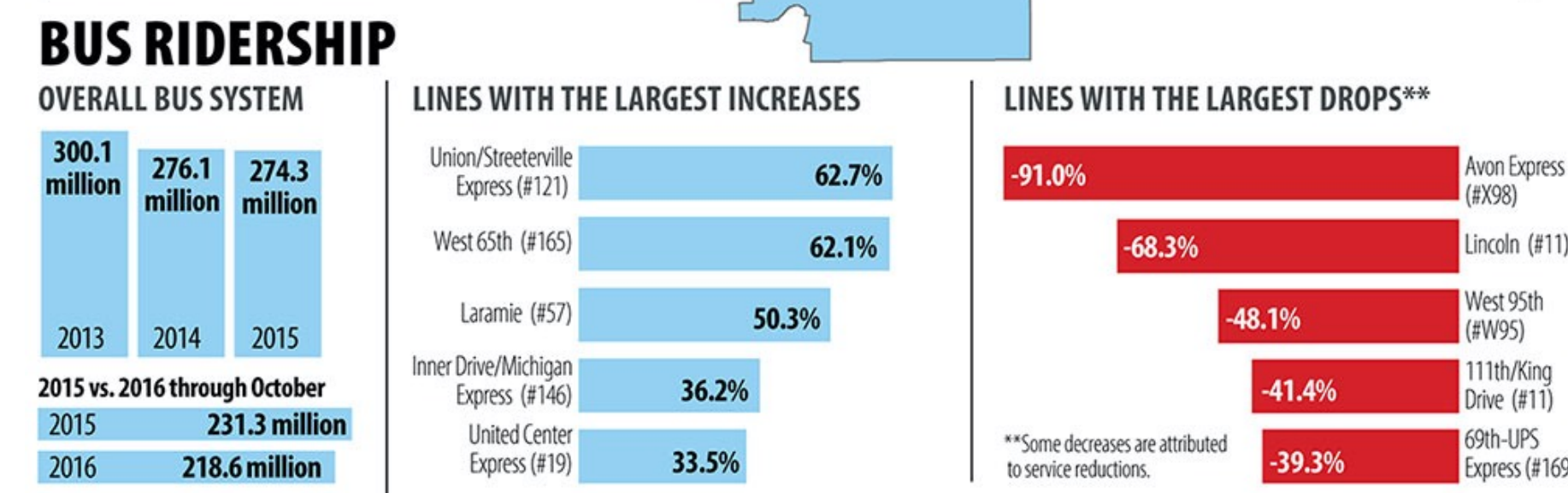
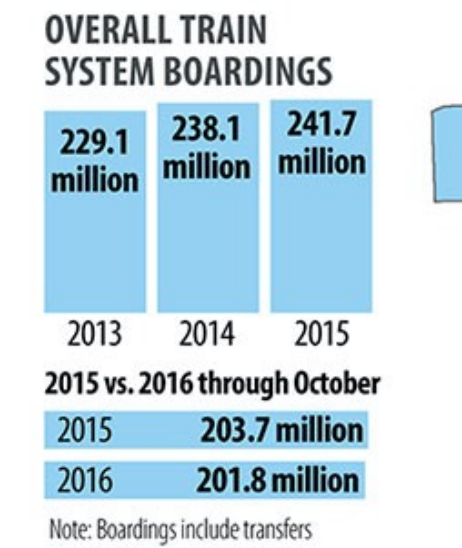
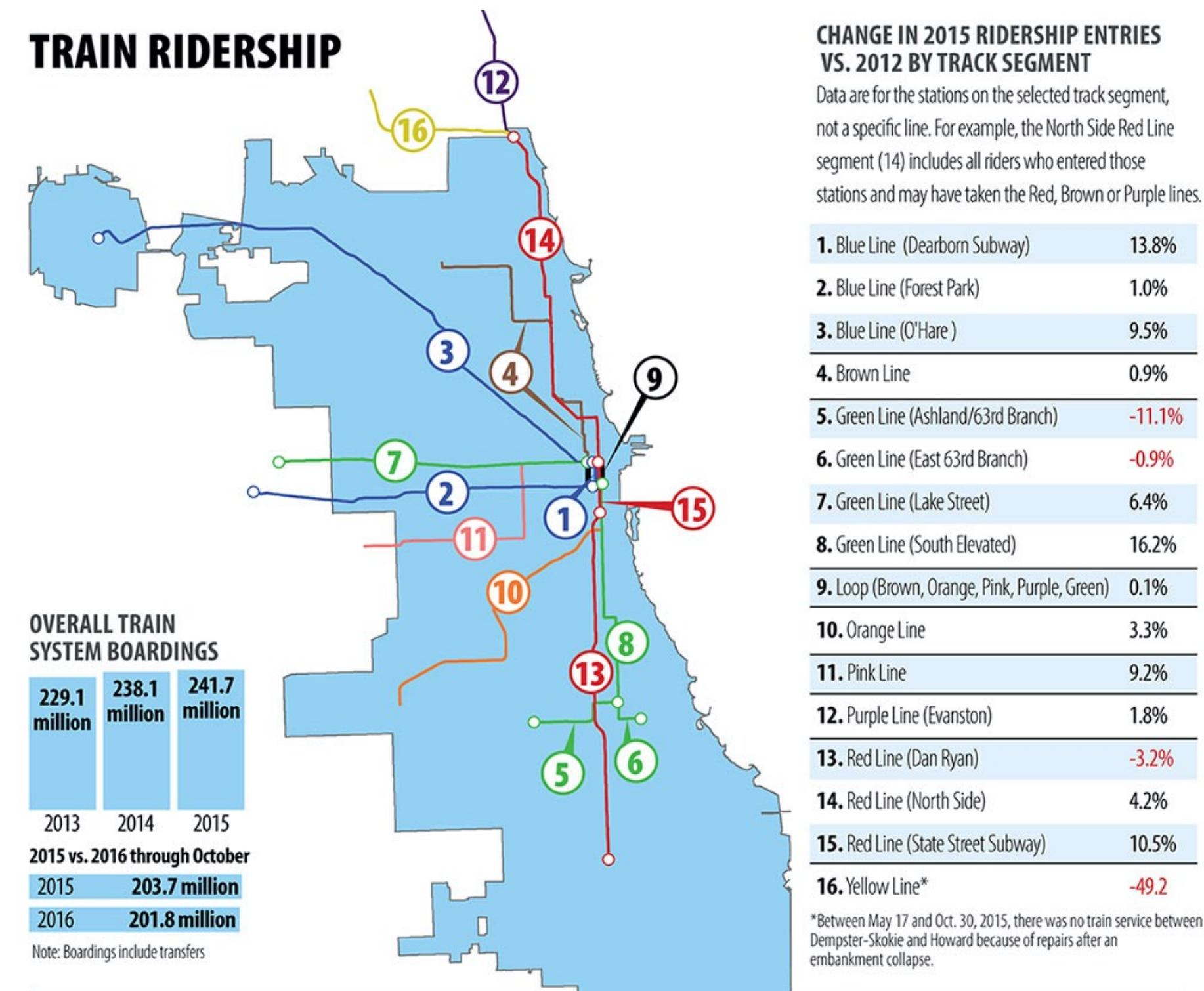
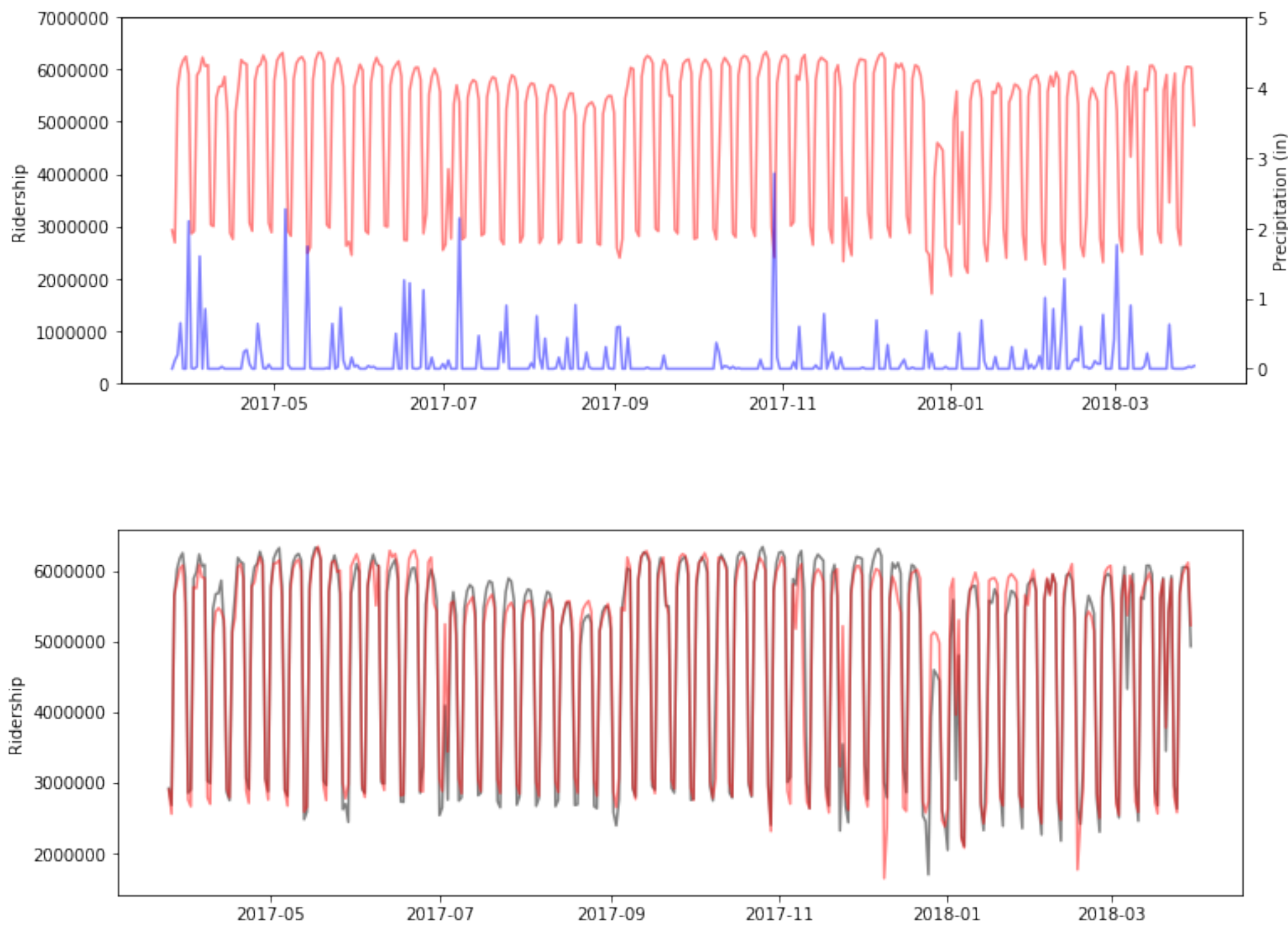
< Previous Next >

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
Showing Rows 1-100 out of 962,546				

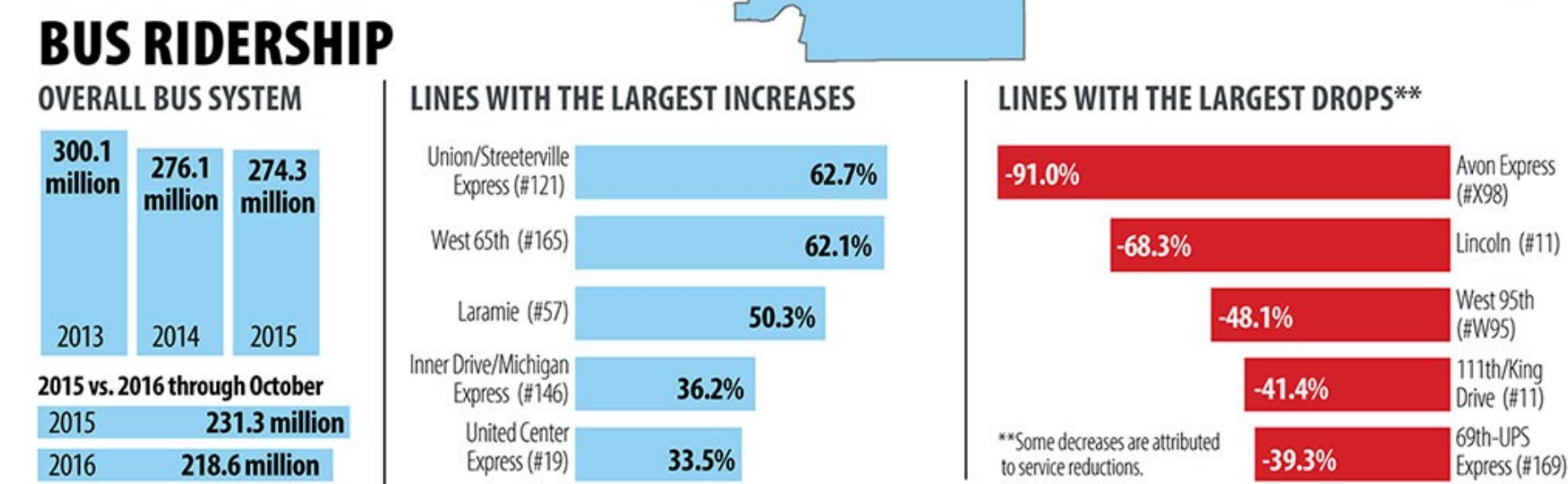
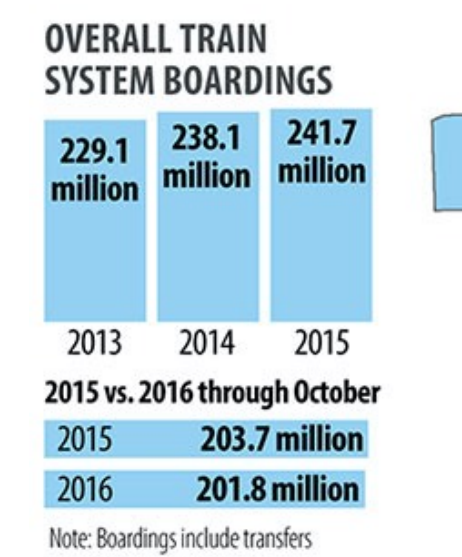
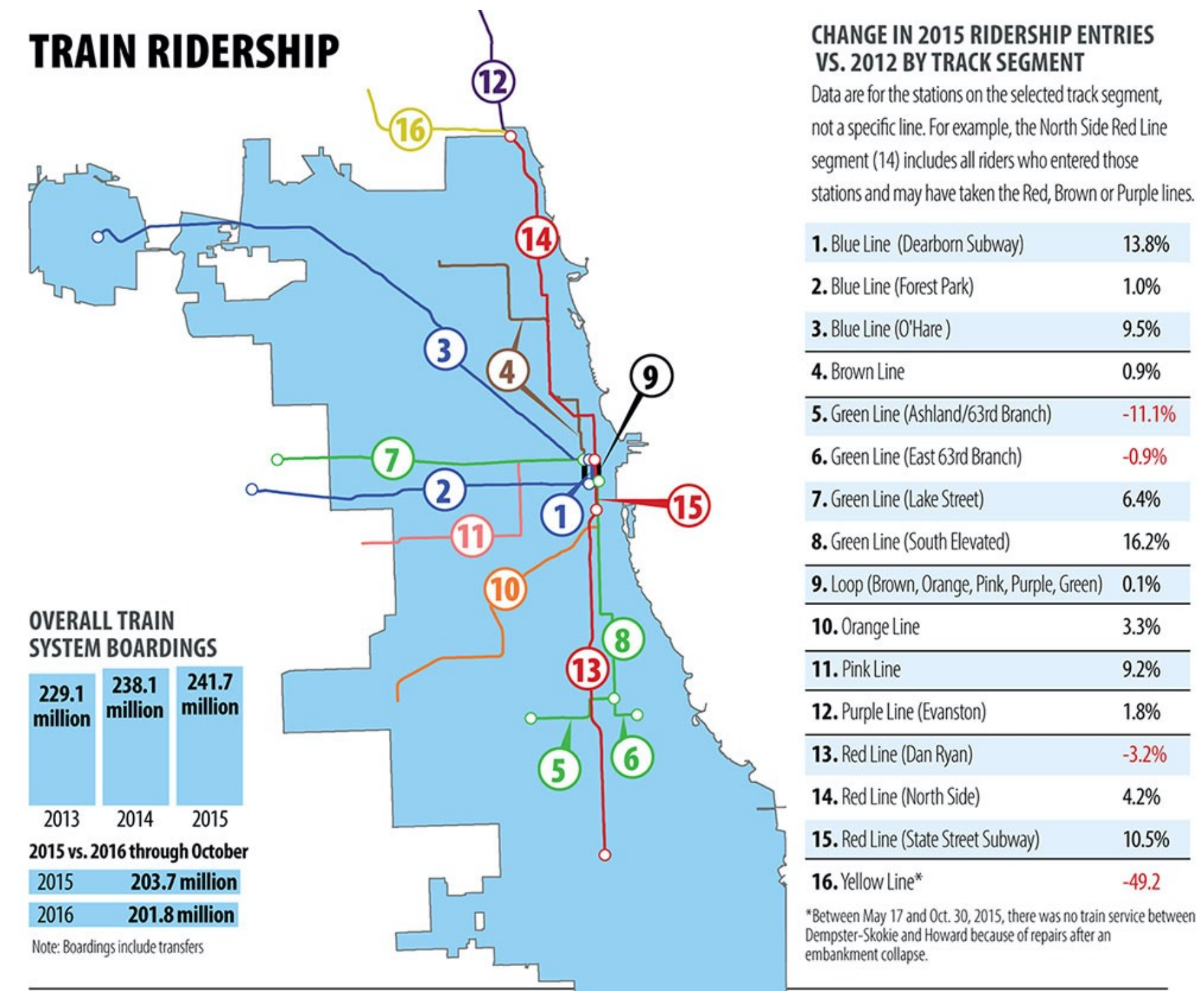
C/A,UNIT,SCP,STATION,LINENAME,DIVISION,DATE,TIME,DESC,ENTRIES,EXITS
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,03:00:00,REGULAR,0007331213,0002484849
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,07:00:00,REGULAR,0007331224,0002484861
A002,R051,02-00-00,59
ST,NQR456W,BMT,01/04/2020,11:00:00,REGULAR,0007331281,0002484936
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

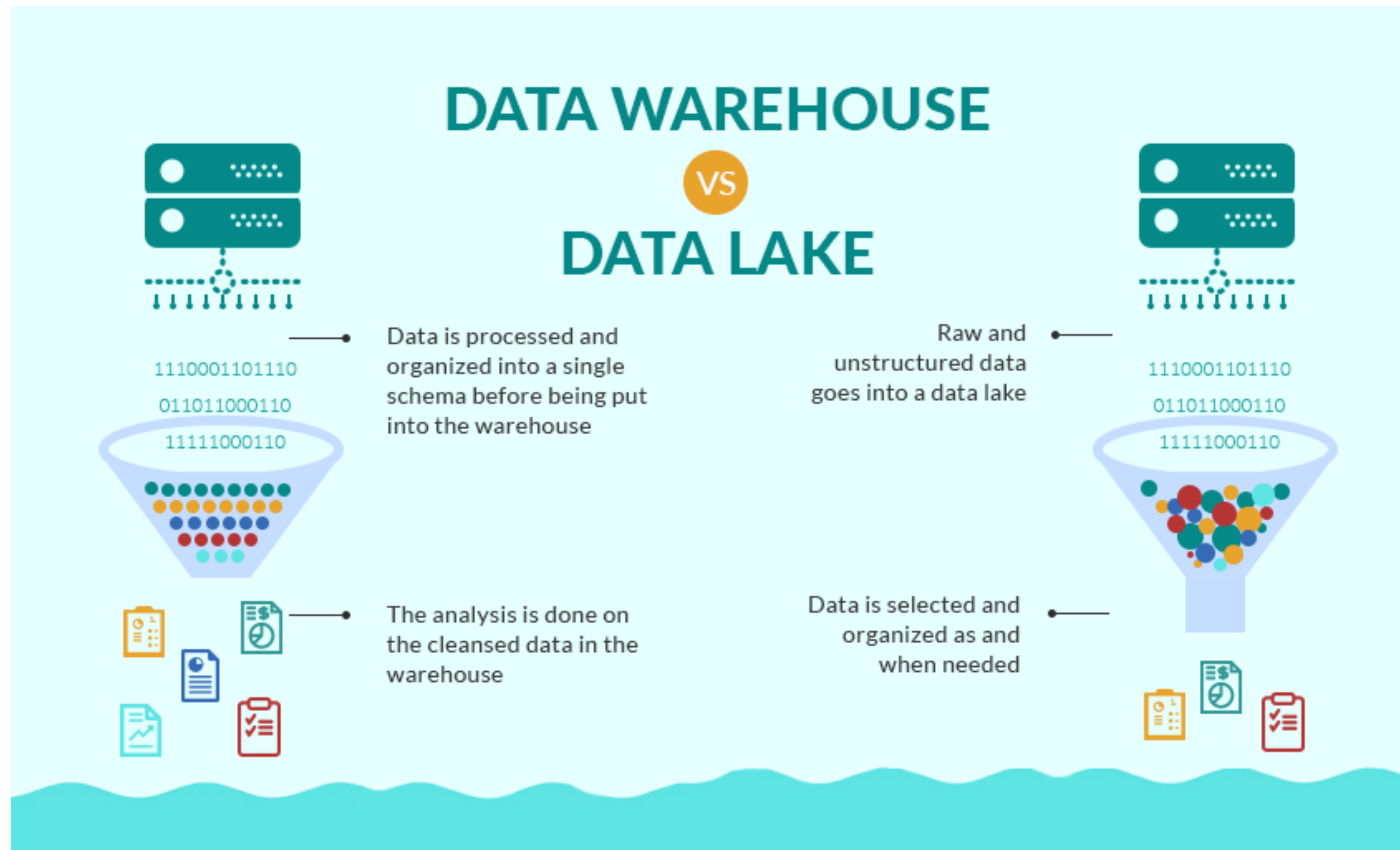
Then... Machine Learning Models & Pretty Visualizations



[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

CUSTOMER ANALYSIS >

customer

Random

Run Job

Preview

#	IMSI	CONTRACT_END	CONTRACT_START	SUBSCRIBER_AGE	RBC	STATUS
310T - 310.26T		Jan 2013 - Dec 2016	Jan 2000 - Dec 2014	0 - 15	2 Categories	
310170226812721		6/4/16	7/29/09			ACTIVE
310160900766700		3/28/15	10/6/13	1		ACTIVE
310170546822541		9/23/16	1/9/07	7		ACTIVE
310005432849230		5/29/15	2/14/01	13		ACTIVE
310026939721905		9/11/15	9/18/10	4		ACTIVE
310026015466952		8/27/15	3/13/06	8		ACTIVE
310170484724861		1/16/16	5/11/04			ACTIVE
310170765640471		05-Jul-2011	9/11/06	4		INACTIVE
310260310245556		12/24/15	3/28/01	13		ACTIVE
310150834295817		3/6/15	7/26/00	14		ACTIVE
310160464252516		9/25/15	4/4/04	10		ACTIVE
310120438750772		4/30/16	9/8/04	10		ACTIVE
310260195729676		1/16/15	1/3/04	11		ACTIVE
310026261822880		8/13/13	11/23/08	4		INACTIVE
310005667082048		8/4/16	10/22/14			ACTIVE
310170836020164		1/22/15	10/19/14	0		ACTIVE
310160772267782		11/21/15	12/28/14			ACTIVE
310170116249240		27-Sep-2011	2/9/09			INACTIVE
310026110612337		5/29/15	3/29/05	9		ACTIVE
310260681676970		11/17/16	5/21/07	7		ACTIVE
310004436630316		9/15/16	7/24/11			ACTIVE
310120423699542		2/27/15	6/29/11	3		ACTIVE
310120773194729		4/28/16	6/15/04	10		ACTIVE
310030295859214		2/7/15	3/24/12	2		ACTIVE
310012150088547		13-Jan-2009	12/10/05	3		INACTIVE
310120387060694		10/1/16	10/25/11	3		ACTIVE

Pattern Details CONTRACT_END

Hide Example Values

m / dd / yy

9/18/15
6/13/15
5/21/15
12/12/15
1/16/16

5.37k

dd - month-abbrev - yyyy

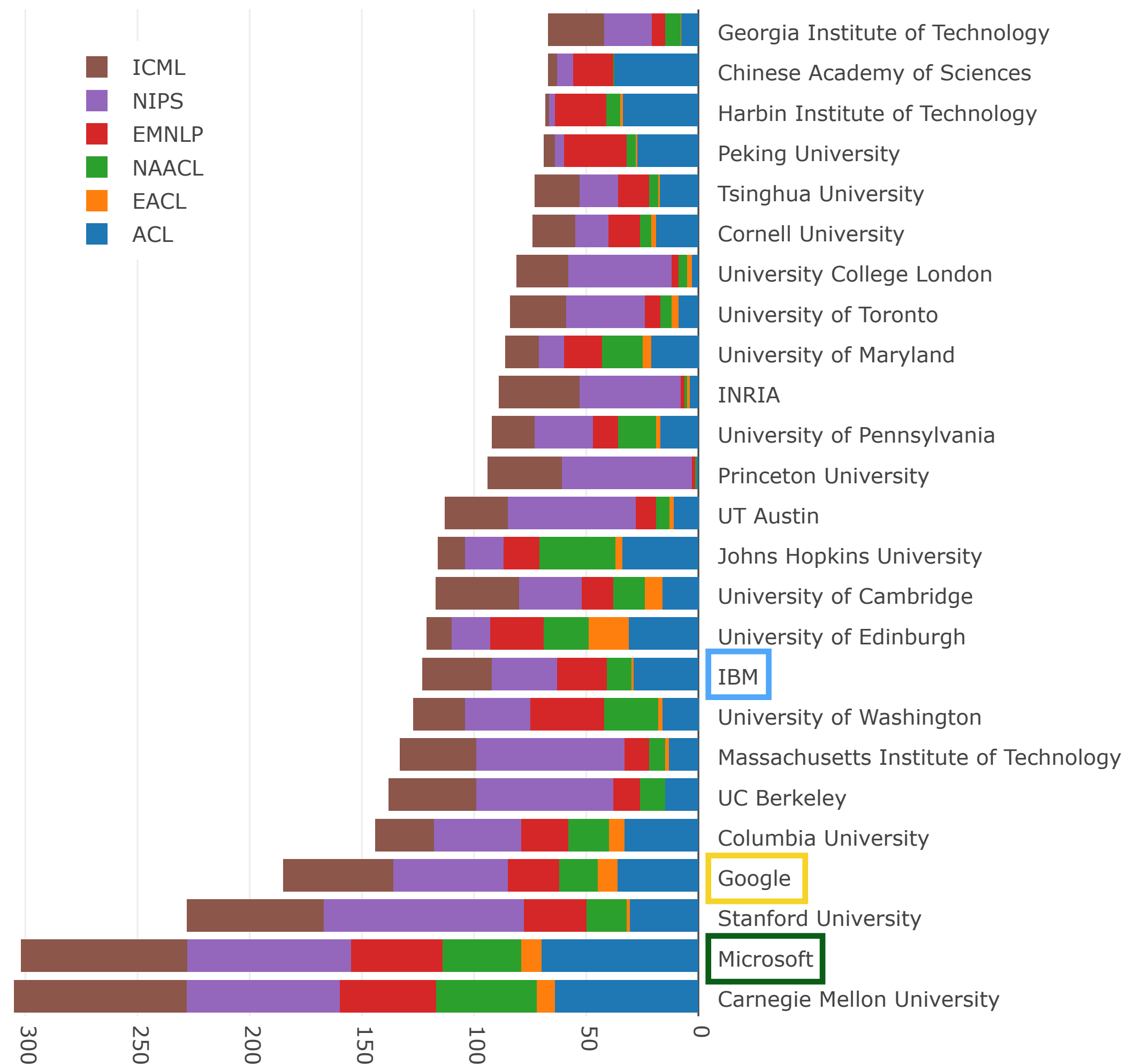
14-Nov-2012
11-Jul-2007
20-Jul-2010

1.99k

19 Columns
20,000 Rows
8 Data Types

Show only affected
Rows

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."

Me ↓

could be me...? →

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

Not me {

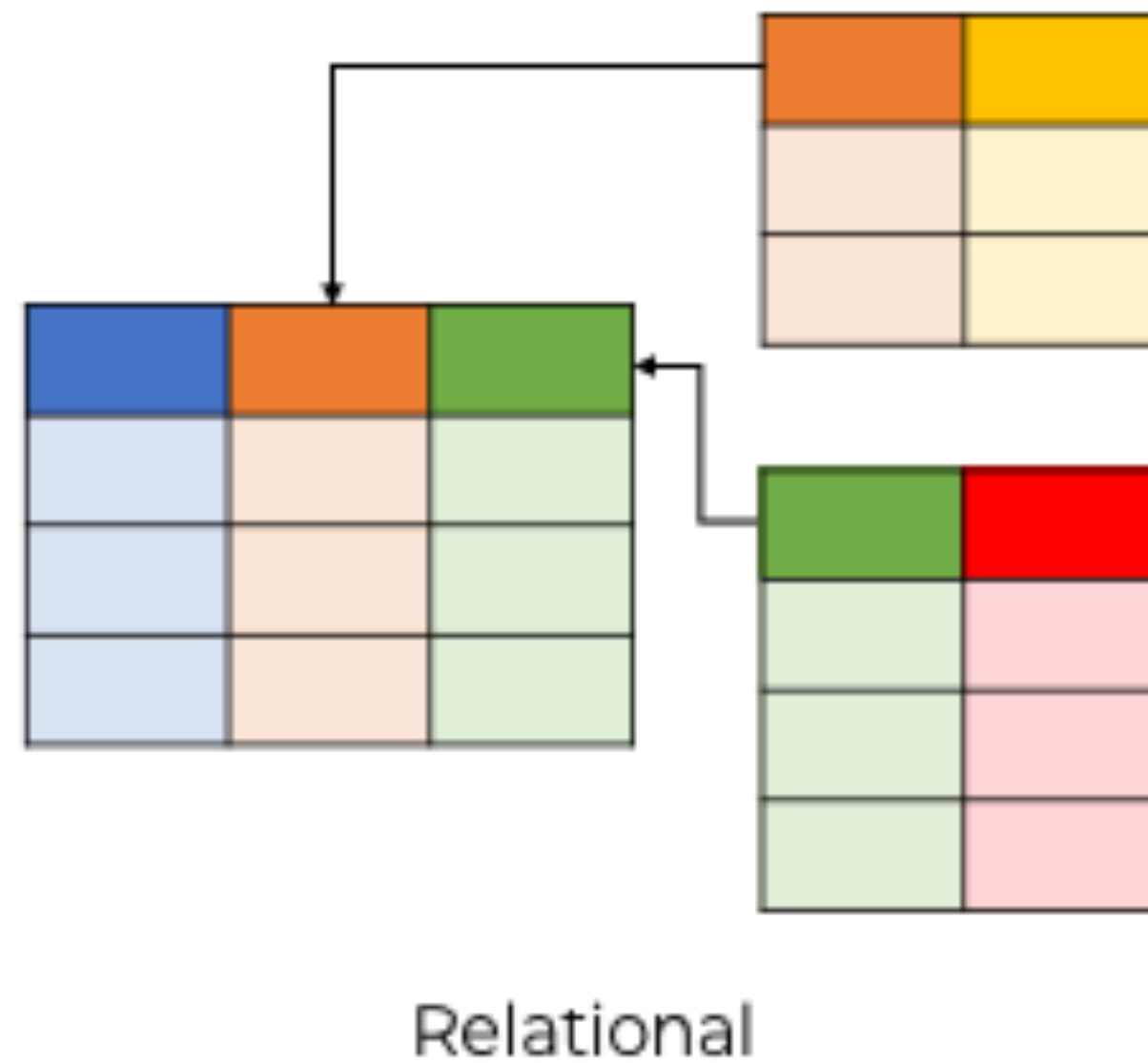
Definitely not me ←

People also search for

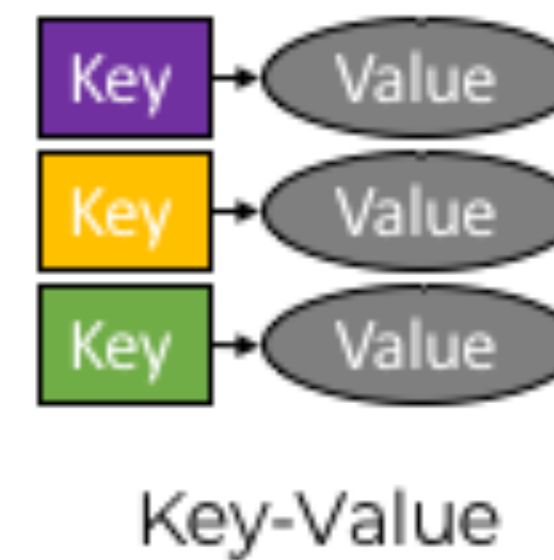
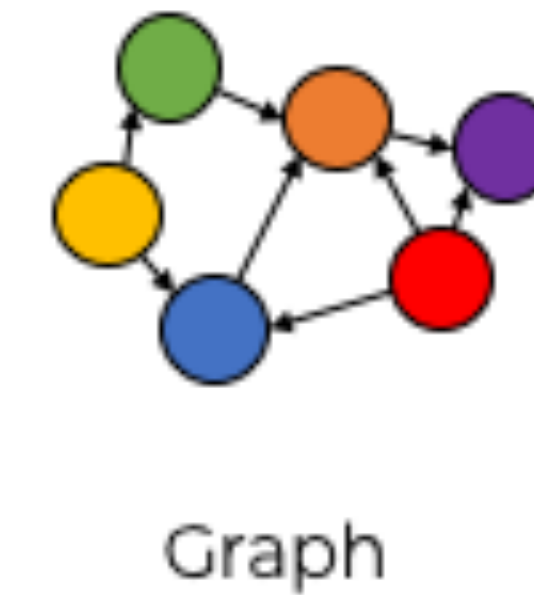


Data Storage

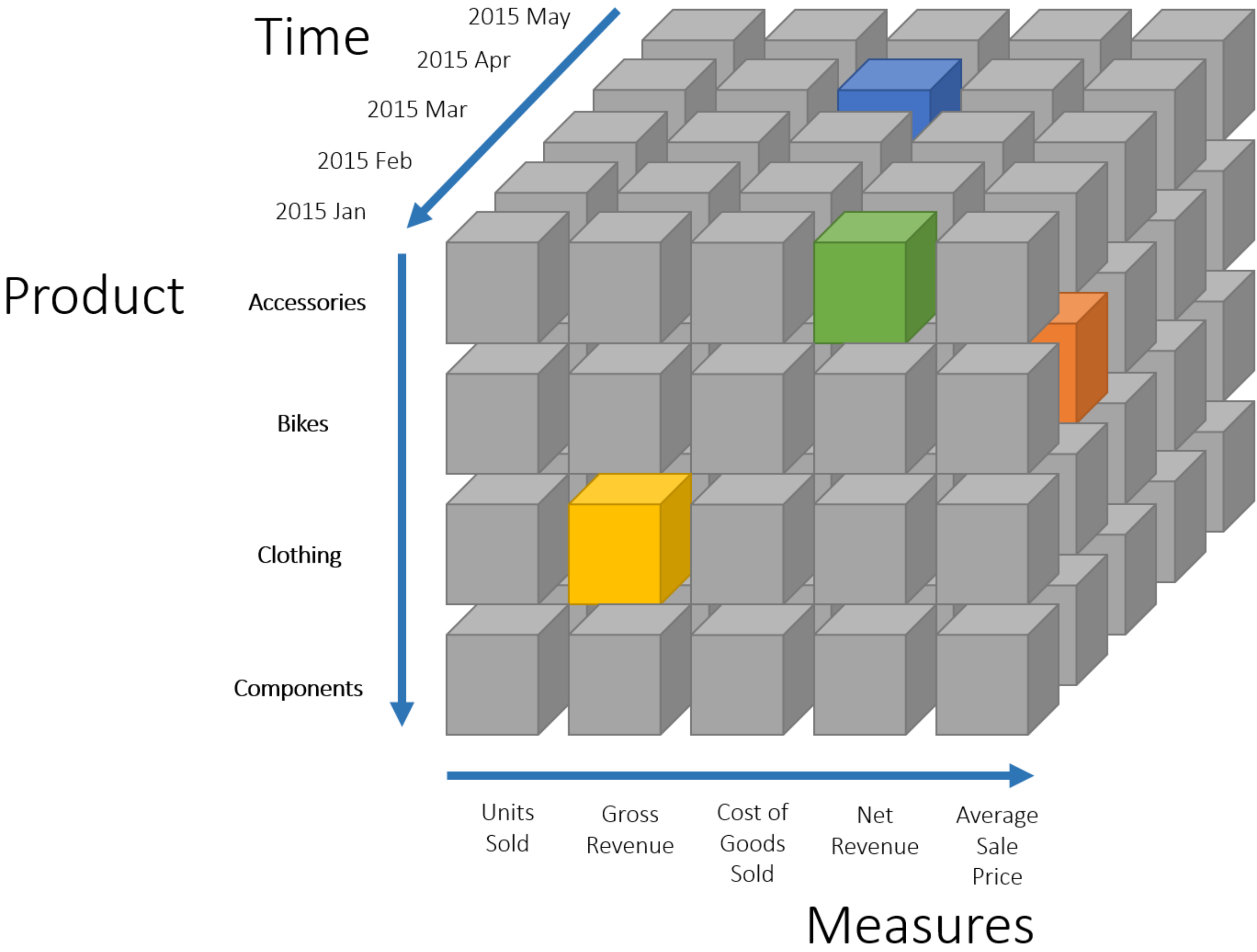
SQL DATABASES



NoSQL DATABASES



Data Cubes



[M. K. Hernandez]

Provenance and Reproducibility

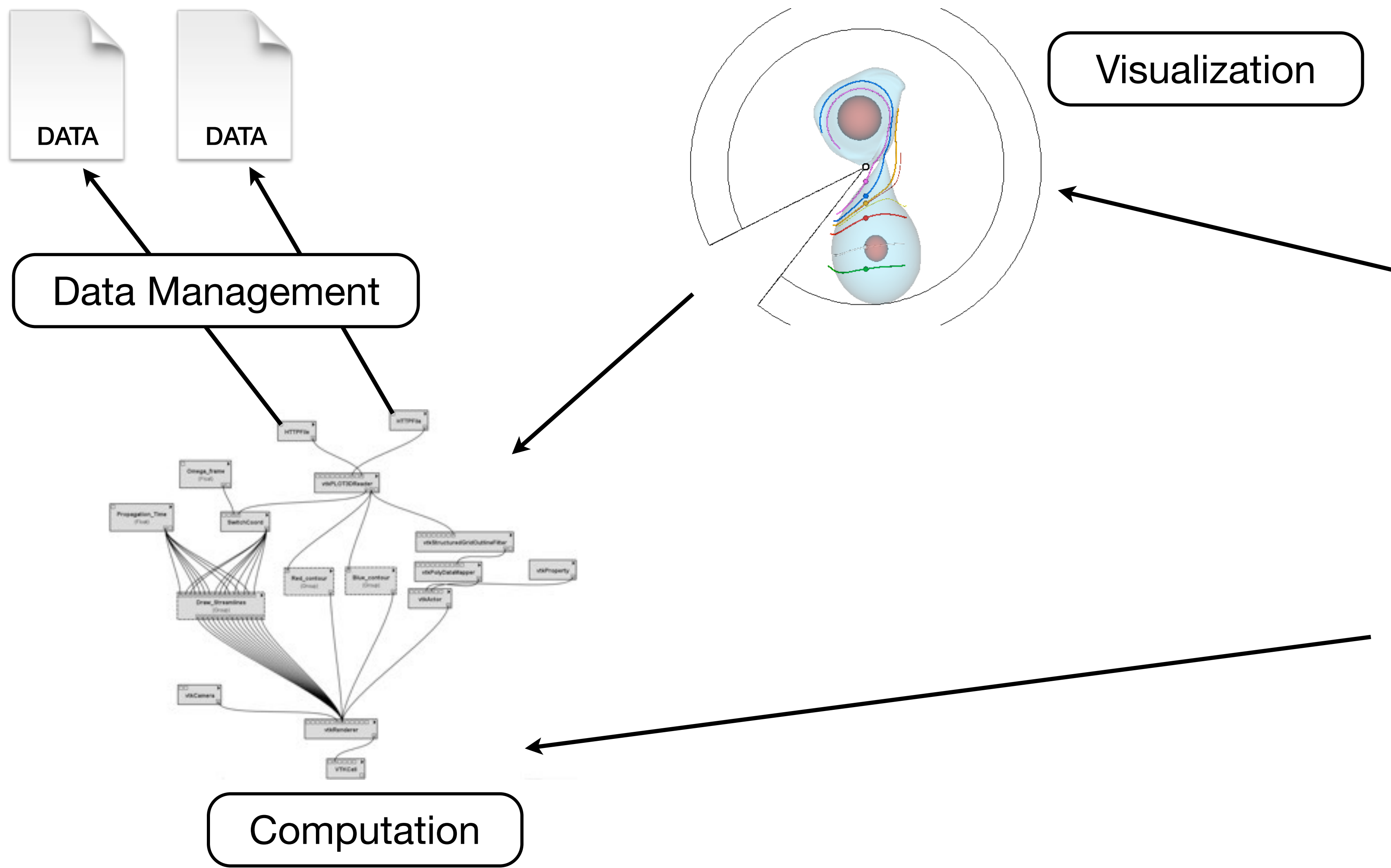


Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at <http://www.crowdclabs.org/vitrails/workflows/details/119/>.

ACKNOWLEDGMENTS

Our research has been funded by the National Science Foundation (grants IIS-0905385, IIS-0746500, ATM-0835821, IIS-0844546, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0334628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724), the Department of Energy SciDAC (VACET and SDM centers), and IBM Faculty Awards (2005, 2006, 2007, and 2008). E. Santos is partially supported by a CAPES/Postgraduate fellowship.

REFERENCES

- [1] L. Bayvel, S. Callahan, P. Cossano, J. Freire, C. Scheidegger, C. Silva, and H. Vo. Vitra: Enabling Interactive Multiple-View Visualizations. In *IEEE Visualization 2005*, pages 135–142, 2005.
- [2] S. P. Callahan, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Towards provenance-enabled parallelism. pages 120–127, 2008.
- [3] Chemical biospace. <http://ch.biospace.com/>.
- [4] NSF Center for Coastal Margin Observation and Prediction (CMOP). <http://www.ccmop.org>.
- [5] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of SIGMOD*, pages 1345–1350, 2008.
- [6] R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.
- [7] S. Finkel and J. Claiborn. Guest editors' introduction: Reproducible research. *Computing in Science Engineering*, 11(1):5–7, Jan-Feb. 2009.
- [8] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21, May-June 2008.
- [9] J. Freire, C. Silva, S. Callahan, E. Santos, C. Scheidegger, and H. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18. Springer Verlag, 2006.
- [10] R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047–1051, 2008.
- [11] IBM. OpenDX. <http://www.research.ibm.com/idx>.
- [12] Kiwari. Paraview. <http://www.paraview.org>.
- [13] Kiwari. The visualization toolkit. <http://www.vtk.org>.
- [14] Many Eyes Wikified. <http://wikified.research.ibm.com>.
- [15] M. McKoon. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081–1088, 2009.
- [16] A. R. Poon, T. Kelder, M. P. van Iersel, K. Hampers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7):2008.
- [17] D. D. Roore, C. Gobbe, and R. Stevens. The design and realization of the virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, 2009.
- [18] E. Santos, L. Lins, J. Ahrens, J. Freire, and C. Silva. Visumashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539–1546, 2009.
- [19] Swivel. <http://www.swivel.com>.
- [20] J. Tabbone and E. Santos. Visualizing a Journal that Serves the Computational Sciences Community. *Computing in Science & Engineering*, 12(3), 2010. To appear.
- [21] J. E. Tabbone. Scientific Visualization: A Necessary Chore. *Computing in Science & Engineering*, 9(6):76–81, 2007.
- [22] C. Upson, J. Thomas Fialhaber, D. Kanins, D. H. Laidlaw, D. Schögl, J. Vroom, R. Gervitz, and A. van Dam. The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, 9(4):30–42, 1989.
- [23] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKoon. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.
- [24] Visit Visualization Tool. <http://wci.lrii.gov/codes/visit>.
- [25] The Vitra Project. <http://www.vitrails.org>.

Provenance and Reproducibility

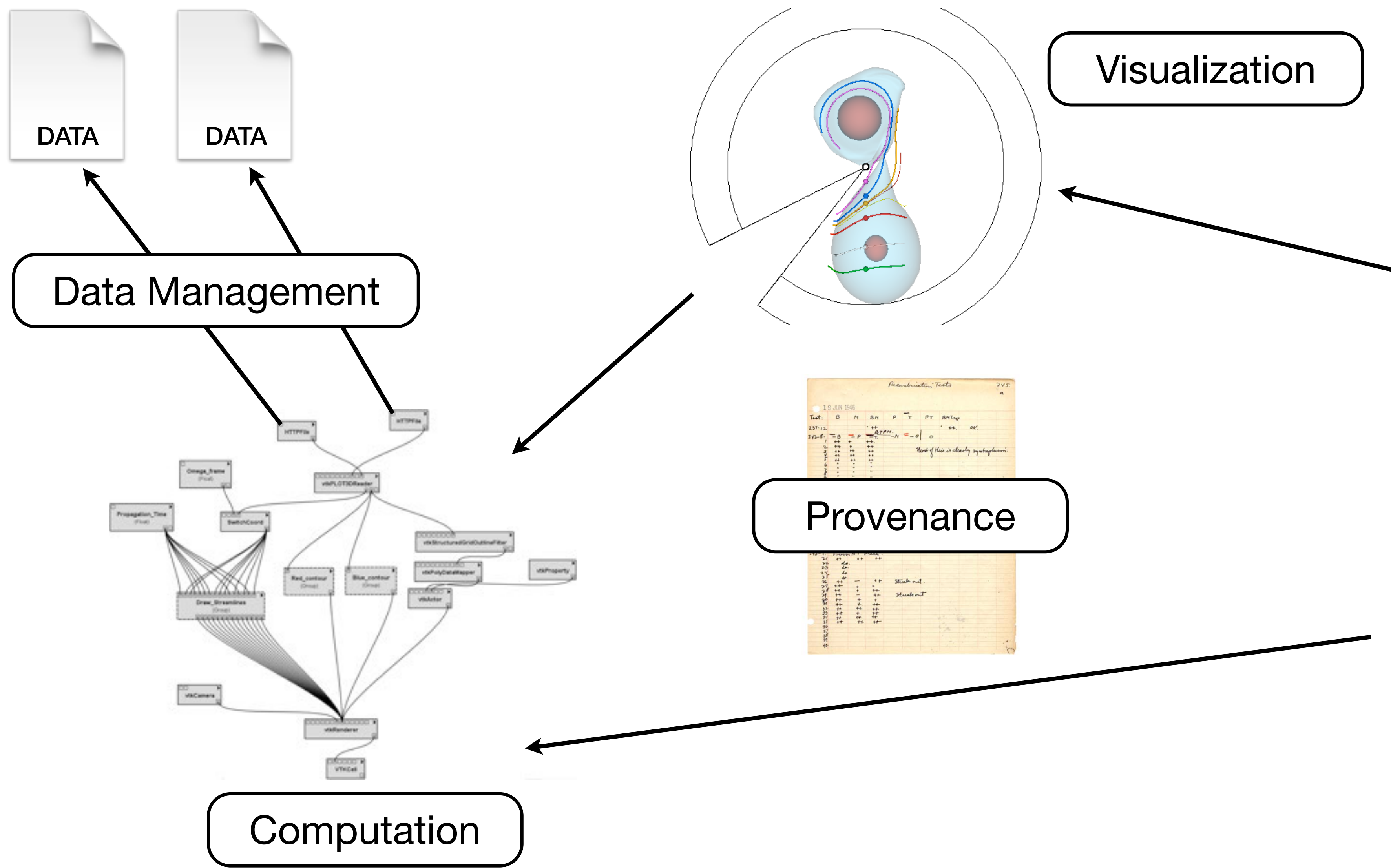


Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at <http://www.crowlabs.org/vistrails/workflows/details/119/>.

ACKNOWLEDGMENTS

Our research has been funded by the National Science Foundation (grants IIS-0905345, IIS-0746500, ATM-0835821, IIS-0844546, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0334628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724), the Department of Energy SciDAC (VACET and SDM centers), and IBM Faculty Awards (2005, 2006, 2007, and 2008). E. Santos is partially supported by a CAPES/Postgraduate fellowship.

REFERENCES

- [1] L. Bayvel, S. Callahan, P. Cossano, J. Freire, C. Scheidegger, C. Silva, and H. Vo. Vistrails: Enabling Interactive Multiple-View Visualizations. In *IEEE Visualization 2005*, pages 135-142, 2005.
- [2] S. P. Callahan, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Towards provenance-enabling panview, pages 120-127, 2008.
- [3] Chemical biospace. <http://cbi.chem.mcgill.ca/>.
- [4] NSF Center for Coastal Margin Observation and Prediction (CMOP). <http://www.ccmop.org>.
- [5] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of SIGMOD*, pages 1345-1350, 2008.
- [6] R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.
- [7] S. Fomel and J. Claiborn. Guest editors' introduction: Reproducible research. *Computing in Science Engineering*, 11(1):5-7, Jan-Feb. 2009.
- [8] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11-21, May-June 2008.
- [9] J. Freire, C. Silva, S. Callahan, E. Santos, C. Scheidegger, and H. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10-18. Springer Verlag, 2006.
- [10] R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047-1051, 2008.
- [11] IBM. OpenDX. <http://www.research.ibm.com/dx/>.
- [12] Kiwari. Panview. <http://www.panview.org>.
- [13] Kiwari. The visualization toolkit. <http://www.vtk.org>.
- [14] Many Eyes Wikified. <http://wikified.researchlabs.ibm.com>.
- [15] M. McKoon. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081-1088, 2009.
- [16] A. R. Poon, T. Kelder, M. P. van Iersel, K. Hampers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7):2008.
- [17] D. D. Roore, C. Gobbe, and R. Stevens. The design and realization of the virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561-567, 2009.
- [18] E. Santos, L. Lins, J. Ahrens, J. Freire, and C. Silva. Vismashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539-1546, 2009.
- [19] Swivel. <http://www.swivel.com>.
- [20] J. Tabbone and E. Santos. Visualizing a Journal that Serves the Computational Sciences Community. *Computing in Science & Engineering*, 12(3), 2010. To appear.
- [21] J. E. Tabbone. Scientific Visualization: A Necessary Chore. *Computing in Science & Engineering*, 9(6):76-81, 2007.
- [22] C. Upson, J. Thomas Fialhaber, D. Kanins, D. H. Laidlaw, D. Schögl, J. Vroom, R. Gervitz, and A. van Dam. The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, 9(4):30-42, 1989.
- [23] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKoon. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121-1128, 2007.
- [24] Vistrails Visualization Tool. <http://wci.lri.fr/govcodes/vistrails>.
- [25] The Vistrails Project. <http://www.vistrails.org>.

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

- Questionnaire
 - 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/cs680-2020sp>
 - Schedule, Readings, Assignments will be posted online
 - Check the web site before emailing me
- 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

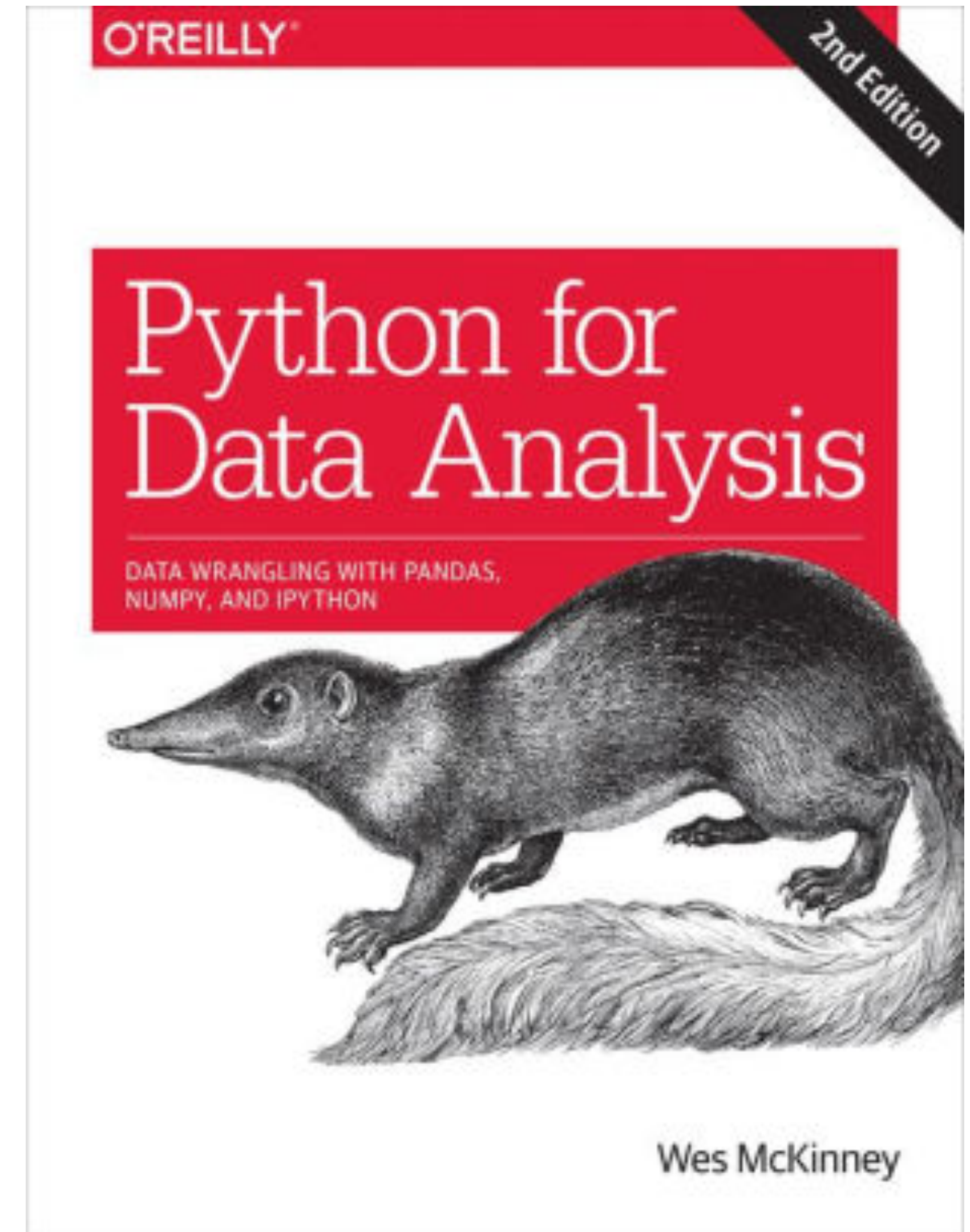
- Balance of techniques and research ideas
- Some background (Python) followed by topic areas and readings
- Programming assignments (~4)
- Two tests + final exam
 - Please check these dates now
- Topic areas:
 - Data Acquisition
 - Data Wrangling
 - Data Storage and Access
 - Cloud Storage and Scalable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

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
- Many websites
- Research papers



Course Material

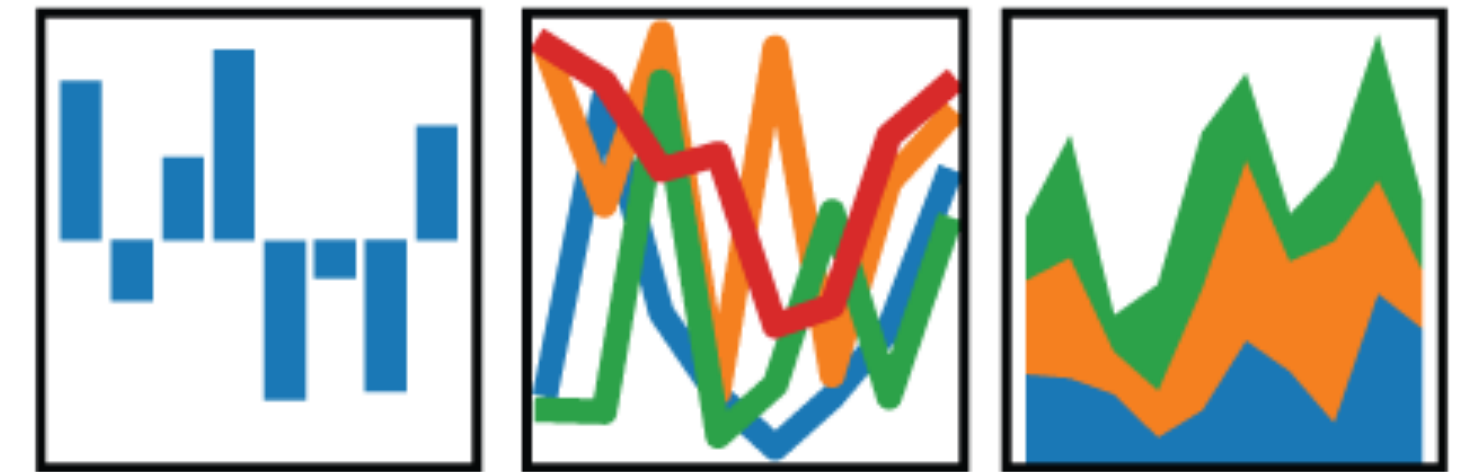


- Software:
 - Anaconda Python Distribution (<https://www.anaconda.com/distribution/>): makes installing python and python packages easier
 - JupyterLab: Web-based interface for interactively writing and executing Python code
 - JupyterHub: Access everything through a server

Course Material

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

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$


TRIFACTA

Office Hours & Email

- Scheduled office hours are open to all students
 - TTh: 1:30pm-2:30pm, or by appointment
- You do not need an appointment to stop 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 3.7):
 - <https://www.anaconda.com/distribution/>