

STAT121 / AC209 / E-109 CS109 Data Science

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

Verena Kaynig
vkaynig@seas.harvard.edu

Outline

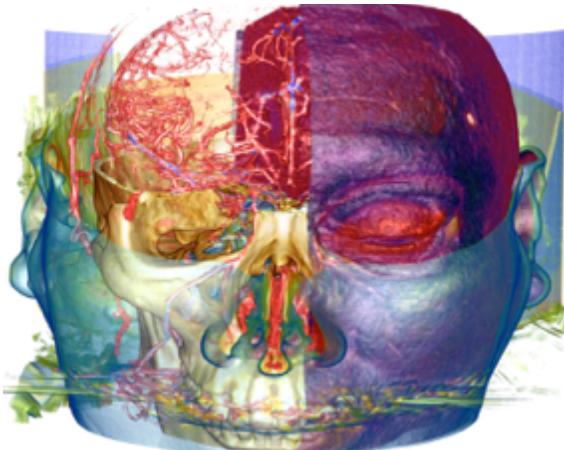
- What?
- Why?
- Who?
- How?

Outline

- What?
- Why?
- Who?
- How?

Data Science

To gain insights into data through computation, statistics, and visualization



A Data Scientist Is...

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock

“Data Scientist = statistician + programmer + coach + storyteller + artist”

- Shlomo Aragmon

Nate Silver

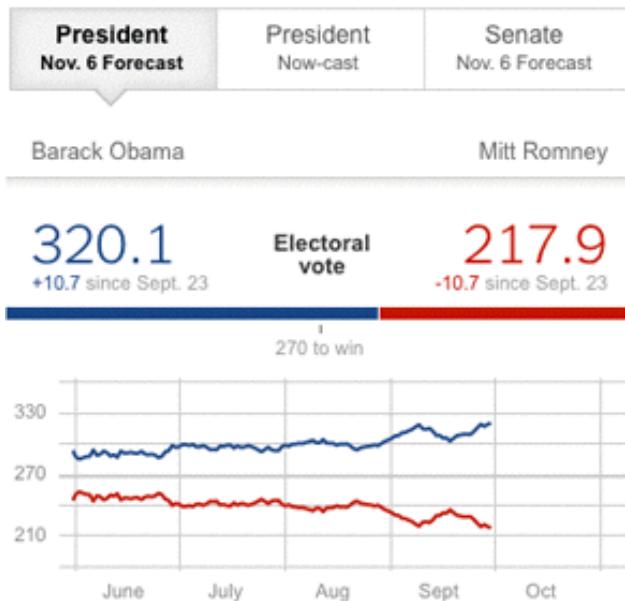


“Nate Silver won the election”

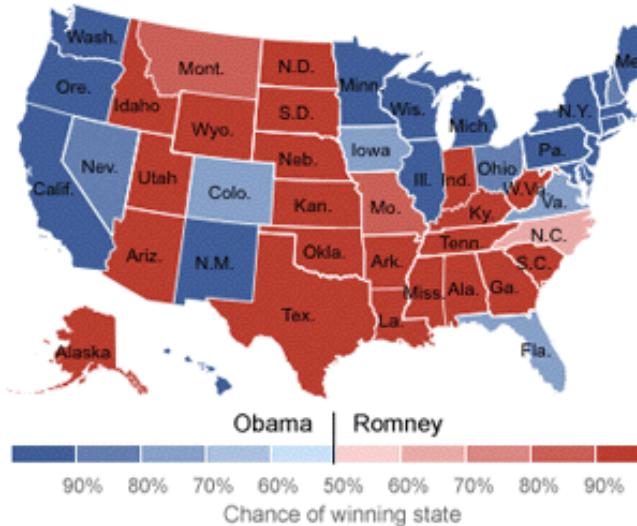
– Harvard Business Review

FiveThirtyEight Forecast

Updated 12:27 AM ET on Oct. 1

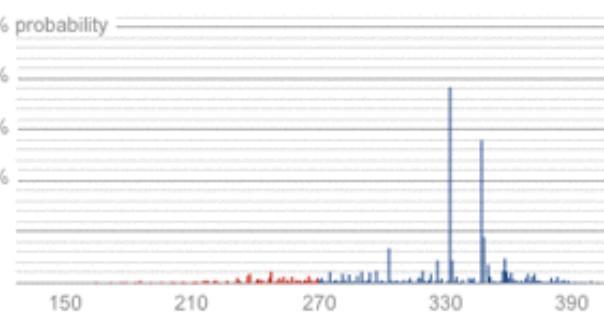


State-by-State Probabilities



Electoral Vote Distribution

The probability that President Obama receives a given number of Electoral College votes.



#natesilverfacts



Ben Hamner @benhamner

7 Nov

#natesilverfacts: Nate Silver doesn't update according to priors, priors update according to Nate Silver @mattcutts

[Expand](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)



DLDahly Epidemiology @statsepi

7 Nov

#natesilverfacts Nate Silver's models fit the test data even better than the training data.



citizenrobot @citizenrobot

7 Nov

Nate Silver knows when GRR Martin will finish the Winds of Winter

#NateSilverFacts



William Chen @wzchen

11 Nov

There is no such thing as missing data, only data that Nate Silver has not chosen to reveal to you. **#natesilverfacts**

Retweeted by Rodrigo Aldecoa and 1 other

[Expand](#)

[Reply](#) [Retweeted](#) [Favorite](#) [More](#)

Is Election Predictor Nate Silver A Witch? Probably. And Quantified Self Data Will Make You One Too



JOSH CONSTINE

Wednesday, November 7th, 2012

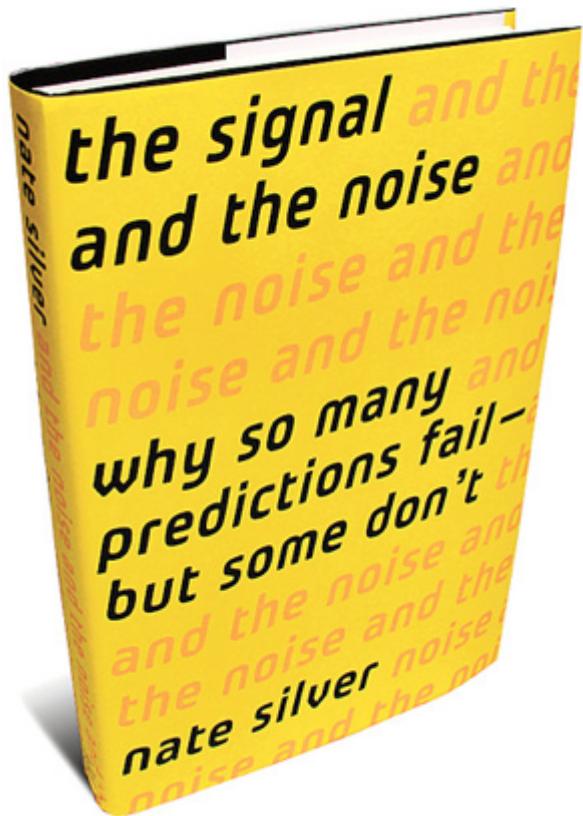
7 Comments



Scientists are yesterday's wizards and demigods. And Nate Silver is a scientist. One whose ability to **predict the outcome of elections** is so precise, it's nearly indistinguishable from magic. That's why **IsNateSilverAWitch.com** is so funny. But really what his flawless prediction of the presidential election signifies is the coming of age of the quantified universe.

<http://techcrunch.com/2012/11/07/nate-silver-as-software/>

Nate Silver on Pundits



Silver: “Pundits are no better than a coin toss.”

Stewart: “Do you foresee a coin getting its own show?
The coin toss show?”

<http://www.thedailyshow.com/watch/wed-october-17-2012/nate-silver>

Some Key Principles

- *use many data sources* (the plural of anecdote is not data)
- *understand how the data were collected* (sampling is essential)
- *weight the data thoughtfully* (not all polls are equally good)
- *use statistical models* (not just hacking around in Excel)
- *understand correlations* (e.g., states that trend similarly)
- *think like a Bayesian, check like a frequentist* (reconciliation)
- *have good communication skills* (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

Netflix Prize

The screenshot shows the official Netflix Prize website. At the top, the Netflix logo is visible. Below it, a large yellow banner features the text "Netflix Prize" on the left and a large red "COMPLETED" stamp on the right. Underneath the banner, there's a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The main content area has a dark background with a blurred image of two people. On the left, there's a section titled "Movies For You" with a list of movie recommendations. On the right, a white callout box with a red border contains the word "Congratulations!" in blue text. Below this, there's a paragraph of text about the purpose of the prize, followed by another paragraph about the award ceremony and links to the algorithm, leaderboard, and forum. The overall design is clean and professional, with a focus on the achievement of the competition.

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

NETFLIX

Browse | Recommendations | Friends | Queue | Buy DVDs

Home | Genres | New Releases | Previews | Netflix Top 100 | Critic Reviews

Movies For You

Randy, the following movies were chosen based on your interest in:

[Now Playing: Columbine](#)
[Family Guy: Season 1](#)
[Fahrenheit 9/11](#)

The Big One

All Discs Guaranteed

You really liked it...

Now owned for just \$5.99

Shop for more titles as low as \$4.99

Original art

Critics' Choice

OTL

Lewis Black: Right and Sane

meet \$nIdBent
>startDataAS
sistanceBD->
Cleva:
Season 2
Disc Series
Daniel Kna...
rivetingly crea...
series conti...

Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.



17,700 Movies
in the
Netflix Competition

Todd.Holloway@gmail.com 03/25/2007

Netflix Prize Progress



HBR, Oct 2012

The screenshot shows a web browser window displaying a blog post on the Netflix Tech Blog. The title of the post is "3 Years Later..." and the subtitle is "The Netflix Tech Blog". The main text of the post discusses the evaluation of new recommendation methods offline, noting that the accuracy gains measured did not justify the engineering effort required to bring them into production. The post is dated Friday, April 6, 2012, and is written by Xavier Amatriain and Justin Basilico. The sidebar on the right contains links to other Netflix blogs and an RSS feed.

Friday, April 6, 2012

3 Years Later... The Netflix Tech Blog

by Xavier Amatriain and Justin Basilico (Personalization Science and Engineering)

In this 3-part series, we will open the dialogue on the most valued Netflix asset: our recommendation system. In Part 1, we will relate the Netflix Prize to the broader recommendation challenge, outline the external components of our personalized service, and highlight how our task has evolved with the business. In Part 2, we will describe some of the data and models that we use, and discuss our approach to continually improving at scale. In Part 3, we will bring it all together, showing how we are using Embarcadero's tools to help us build great tech, so please stay tuned!

In 2006 we announced the Netflix Prize, a machine learning and data mining competition for movie rating prediction. We offered \$1 million to whoever improved the accuracy of our existing system called Cinematch by 10%. We had a solid system and needed to improve its recommendation quality. We had a few members who won the Netflix Prize. However, we had trouble with many more members who gave us the source code. They gave us the source code. We looked at the two underlying algorithms with the best performance in the ensemble: *Matrix Factorization* (which the community generally called SVD, *Singular Value Decomposition*) and *Restricted Boltzmann Machines* (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome some limitations, for instance that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we overcame those challenges, we put the two algorithms into production, where they are still used as part of our recommendation engine.

Links

- Canada Blog
- Netflix America Latina Blog
- Netflix Brazil Blog
- Netflix DACH Blog
- Netflix France Blog
- Netflix Germany Blog
- Netflix UK & Ireland Blog

[Open positions at Netflix](#)

Netflix Website

[Facebook](#) [Twitter](#) [LinkedIn](#) [YouTube](#)

Netflix UI Engineering

[RSS Feed](#)

About the Netflix Tech Blog

This is a Netflix blog focused on technology and technology issues. We'll share our perspectives, decisions and challenges regarding the software we build and use to create the Netflix service.

Xavier Amatriain and Justin Basilico, 2012

Some Challenges

- *massive data* (500k users, 20k movies, 100m ratings)
- *curse of dimensionality* (very high-dimensional problem)
- *missing data* (99% of data missing; *not* missing at random)
- *extremely complicated set of factors that affect people's ratings of movies* (actors, directors, genre, ...)
- *need to avoid overfitting* (test data vs. training data)

Kaggle

www.kaggle.com/competitions

DRB | GitLab The Hub Feedly MD Syntax Add to Pinboard My Pinboard Instapaper libx bit.ly Orrick Box Timesheet LaTeX symbols Lore 3G Mobile Hotspot

kaggle Host Competitions Scripts Jobs Community Sign up Login

Welcome to Kaggle's data science competitions.

New to Data Science?
[Tutorials on the Titanic competition »](#)

Want to learn from other's code?
[Kaggle's top rated scripts »](#)

Download Choose a competition & download the training data.

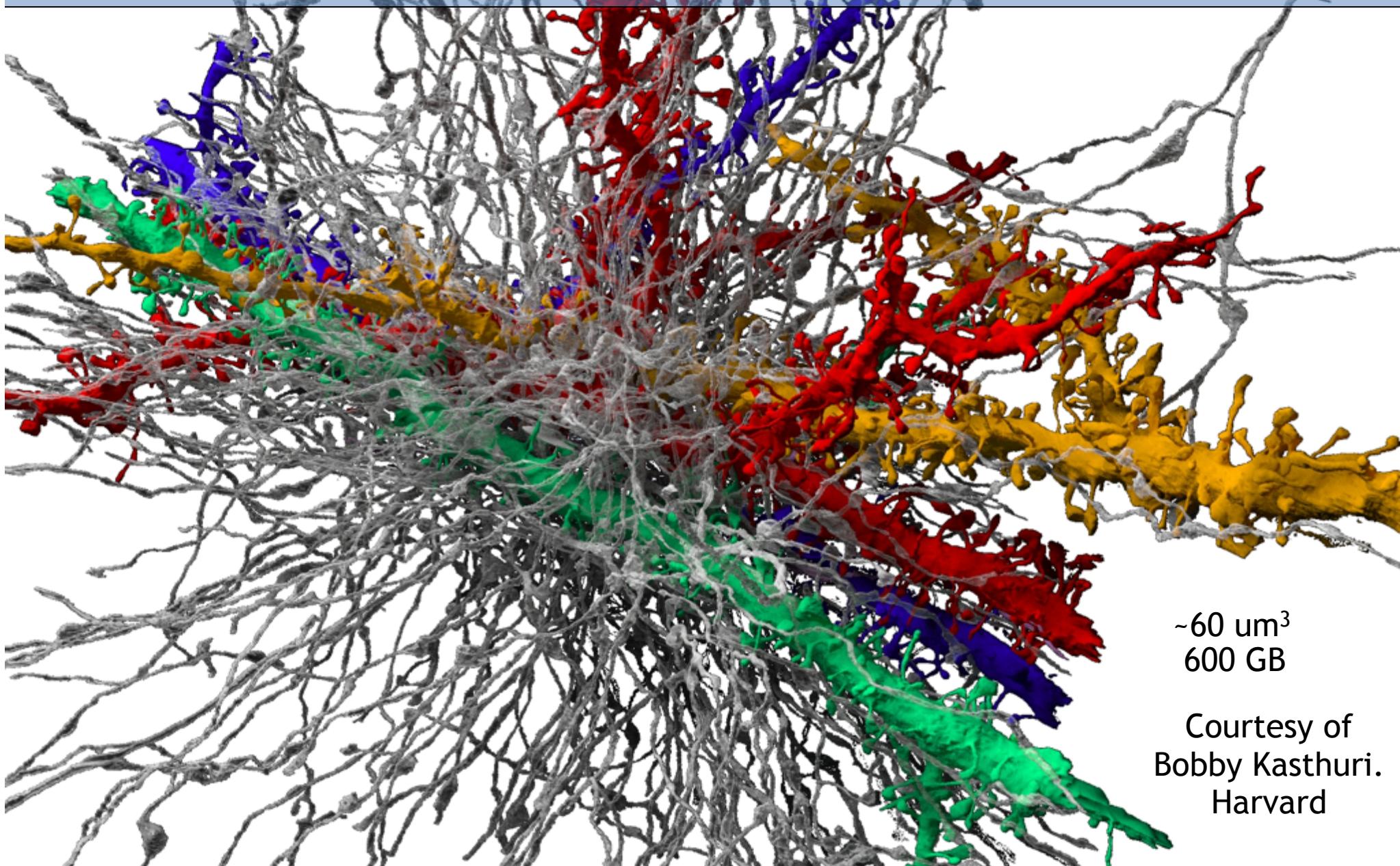
Build Build a model using whatever methods and tools you prefer.

Submit Upload your predictions. Kaggle scores your solution and shows your score on the leaderboard.

Active Competitions		Active Competitions		
All Competitions		Springleaf Marketing Response Determine whether to send a direct mail piece to a customer	46 days	964 teams
		Western Australia Rental Prices Predict rental prices for properties across Western Australia	2 months	10 teams \$100,000
		Coupon Purchase Prediction Predict which coupons a customer will buy	27 days	690 teams
			39 days	

The Connectome

How is the mammalian brain wired?

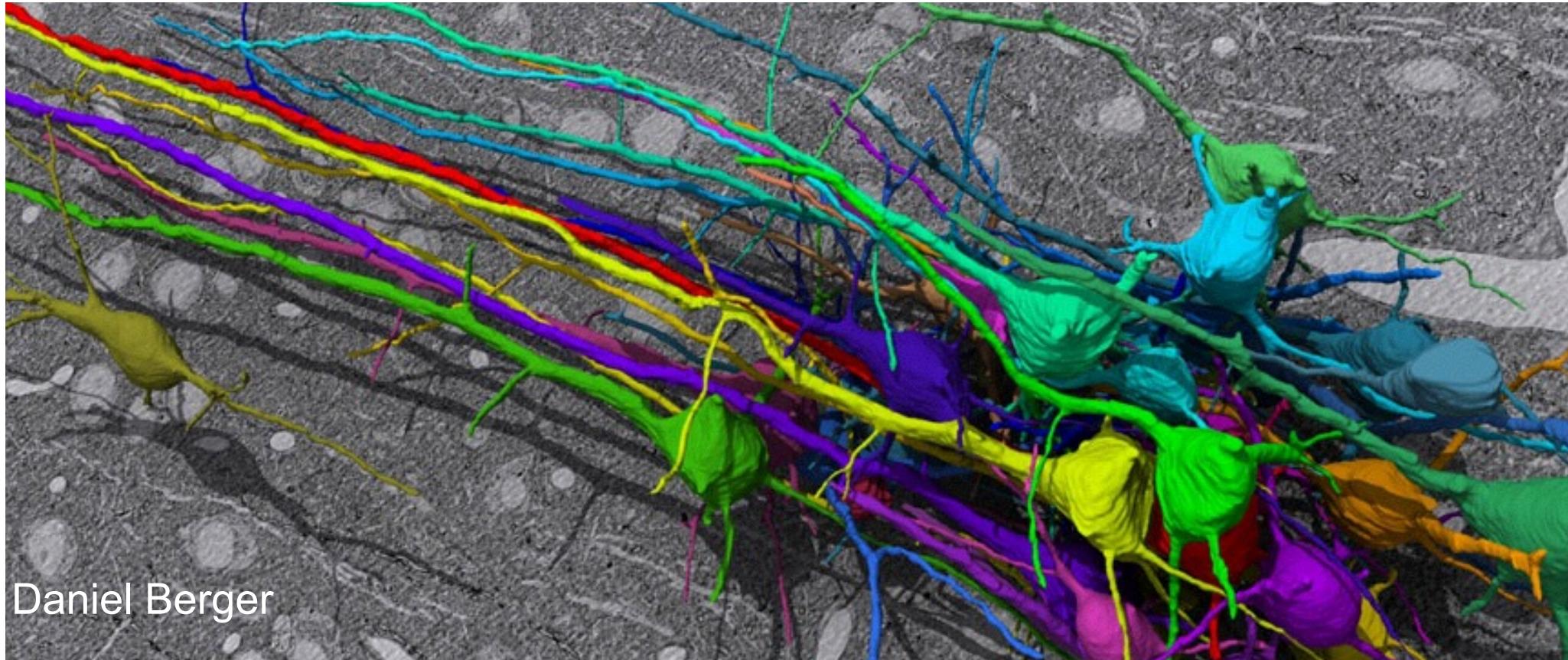


~60 μm^3
600 GB

Courtesy of
Bobby Kasthuri.
Harvard

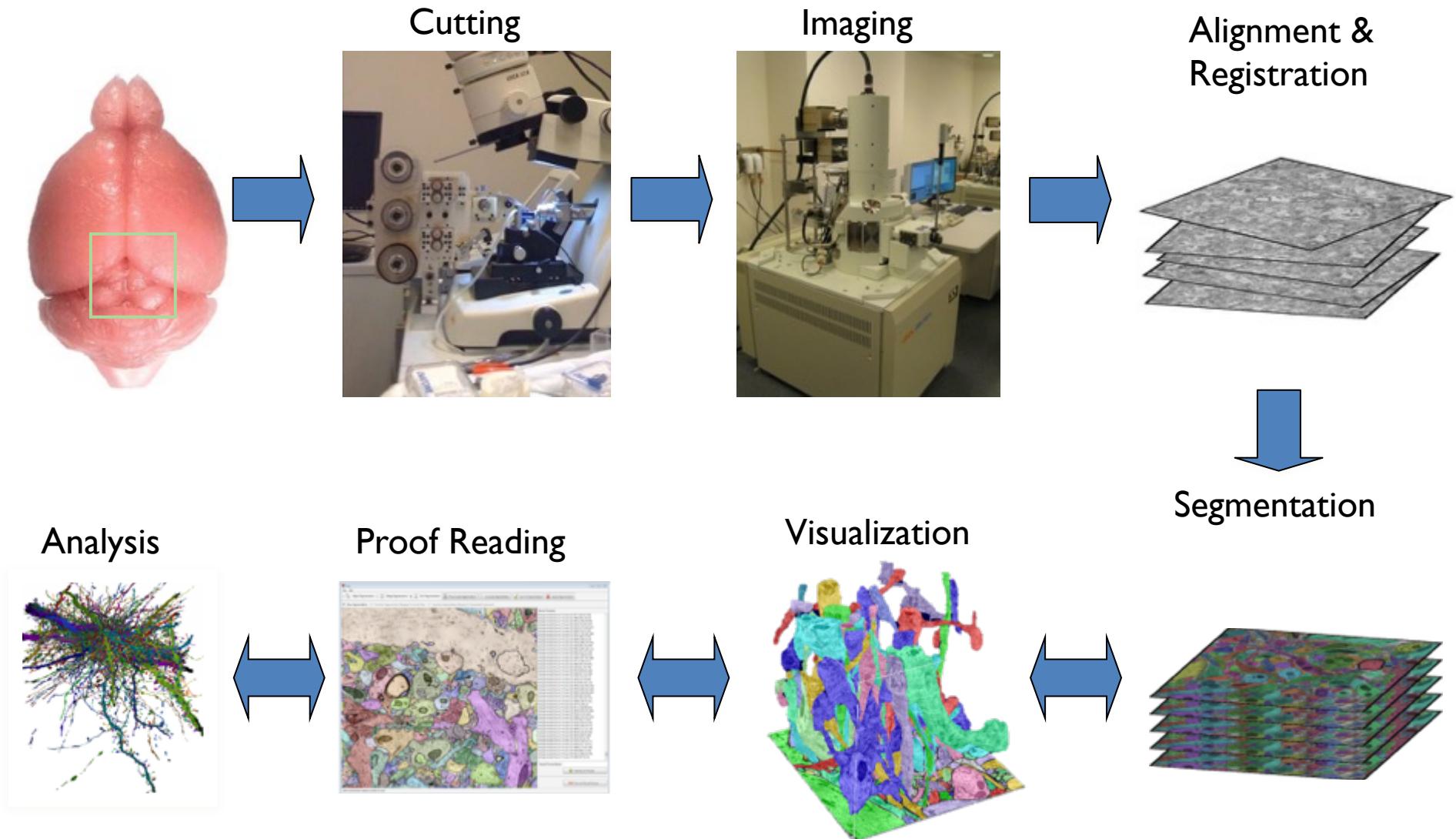
The Data Challenge

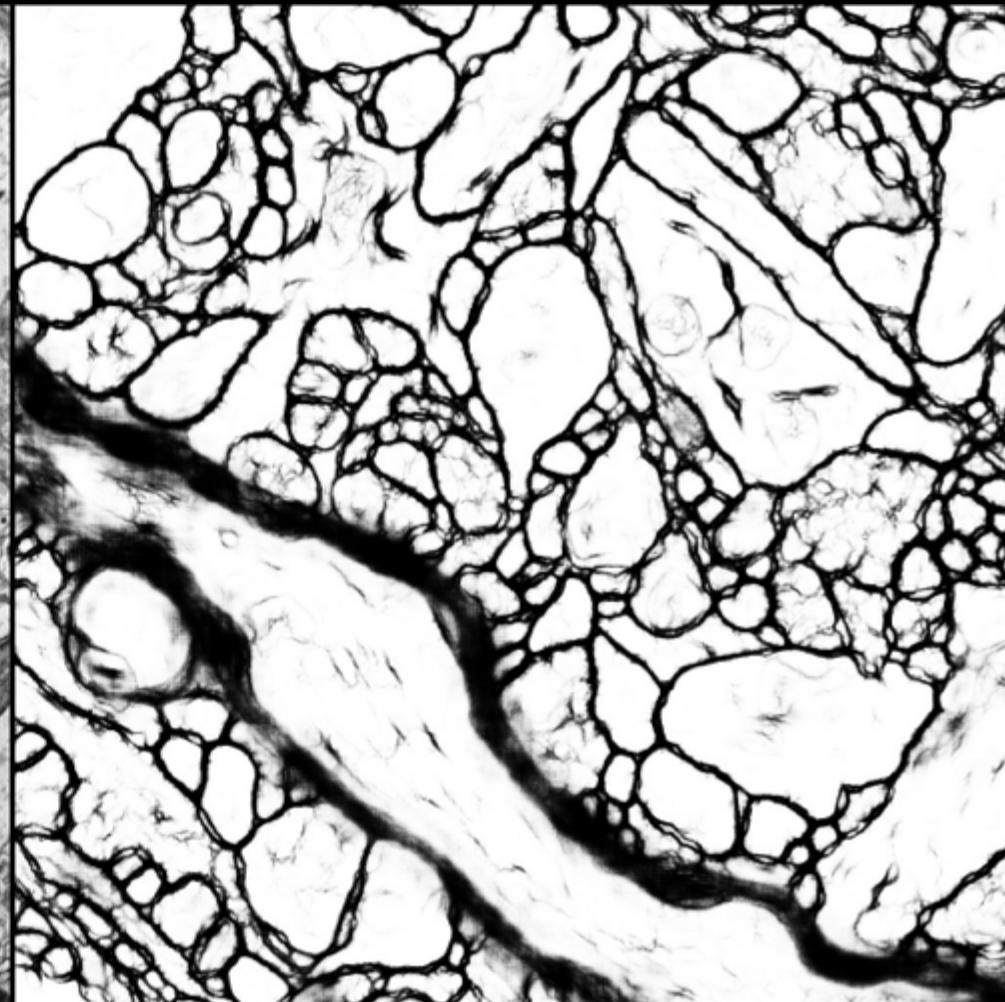
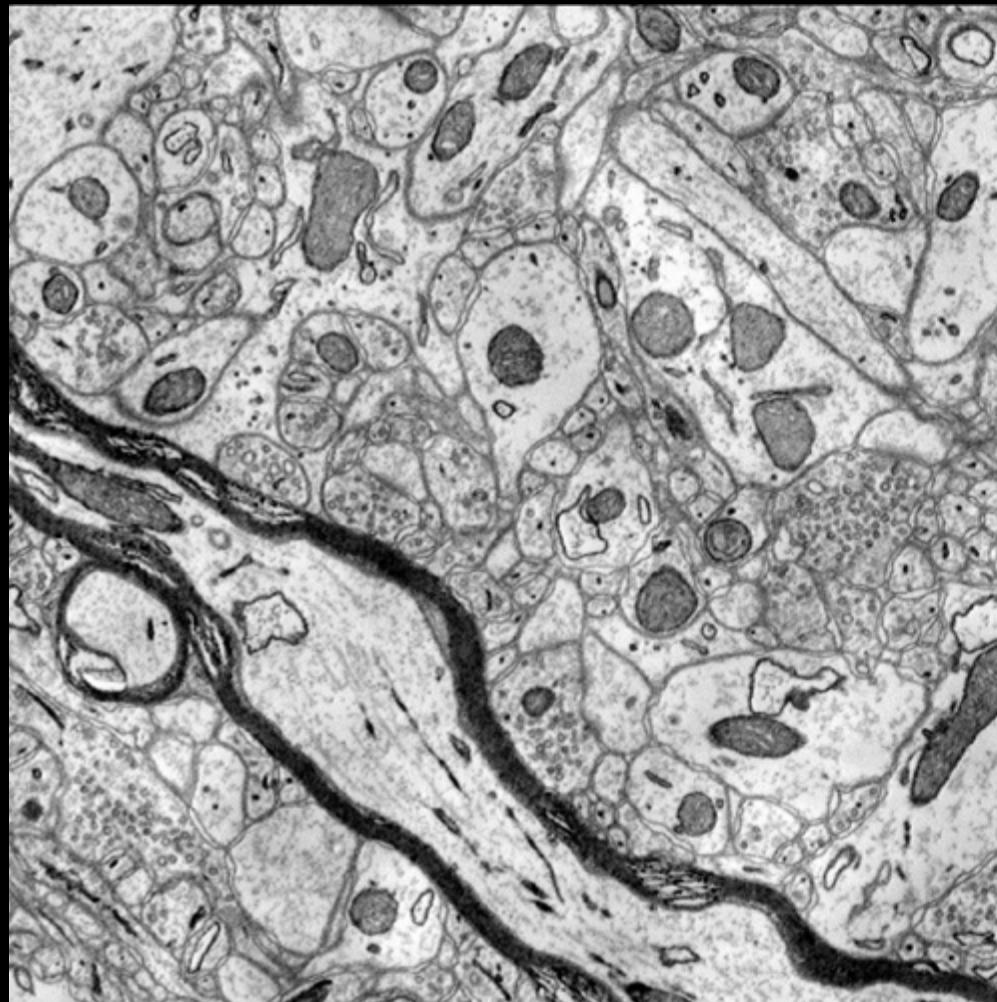
- Pixel resolution: 3-5 nm; Slice thickness: 30-50 nm
- 1 mm³: 40 Gpixels × 25,000 slices = **~1 PByte**

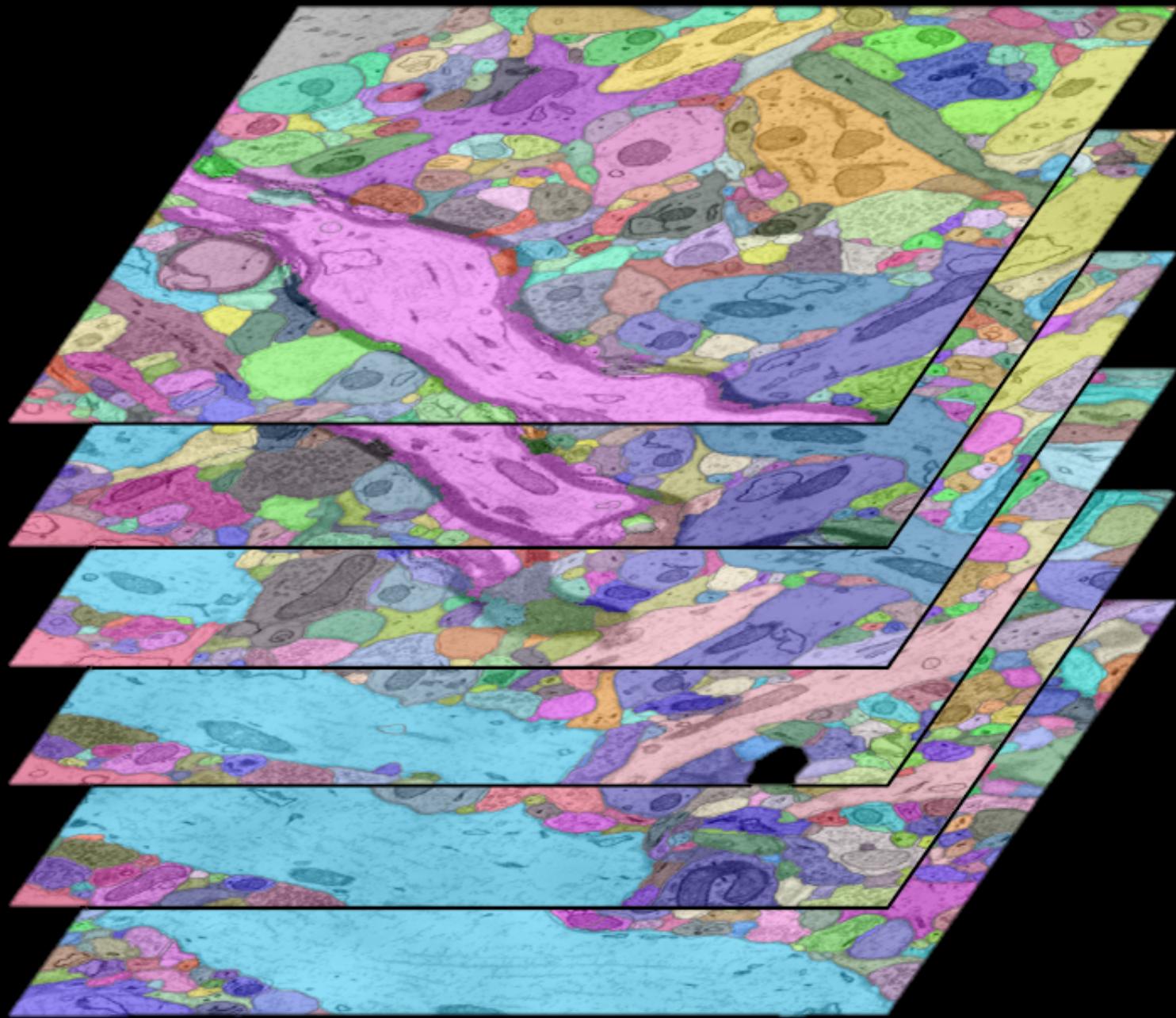


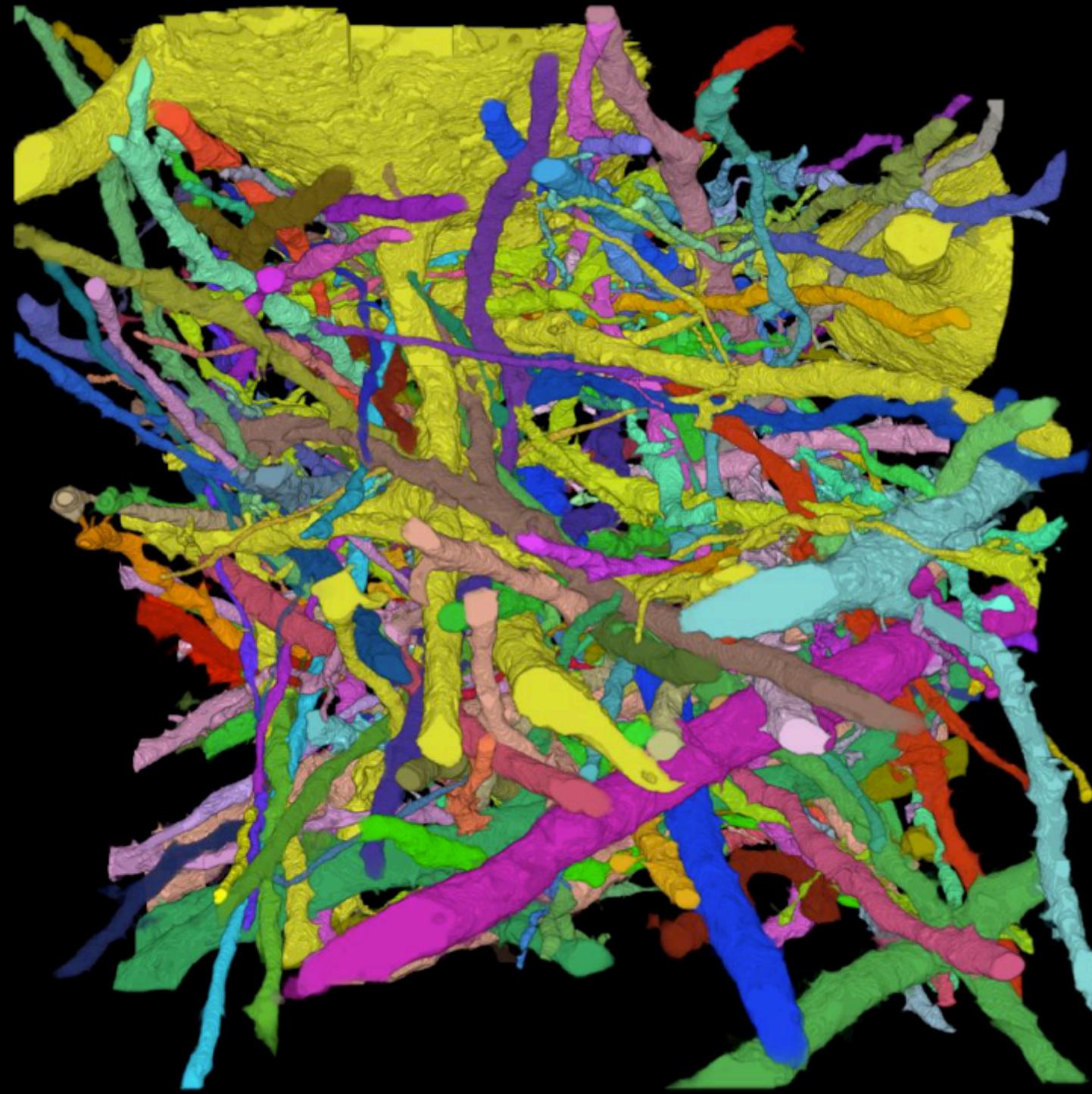
Daniel Berger

Connectome Workflow

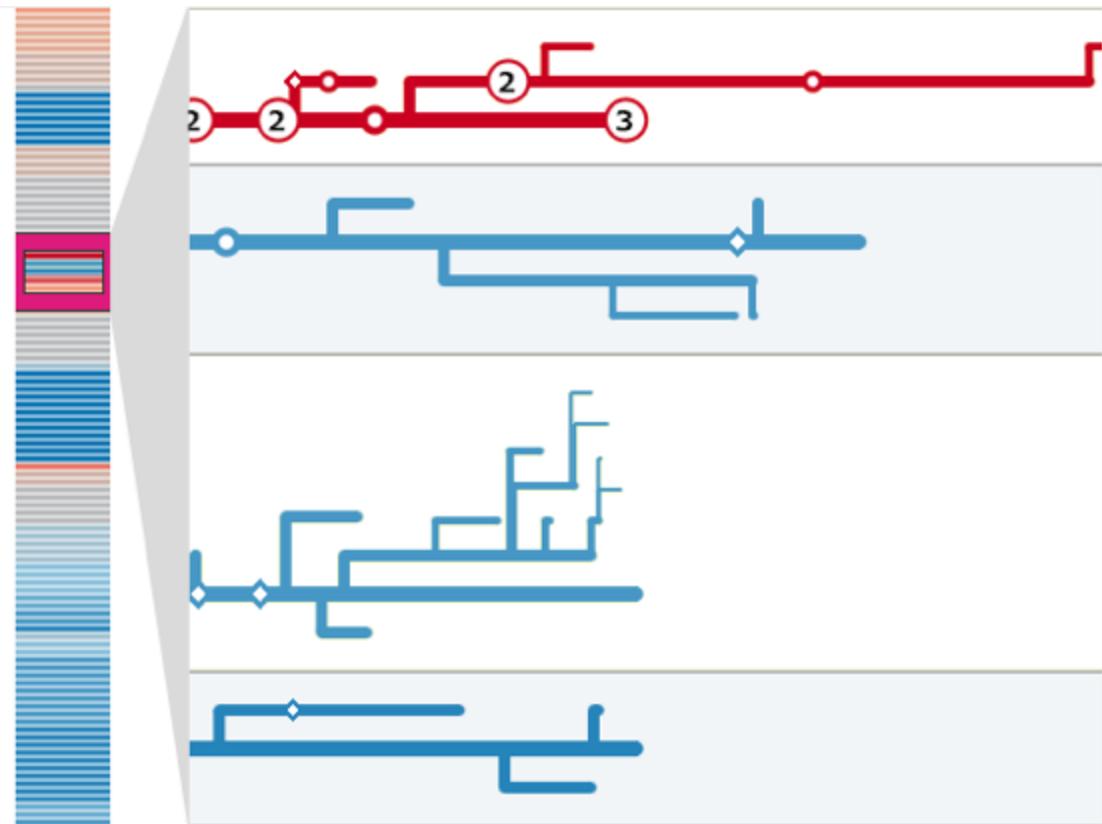
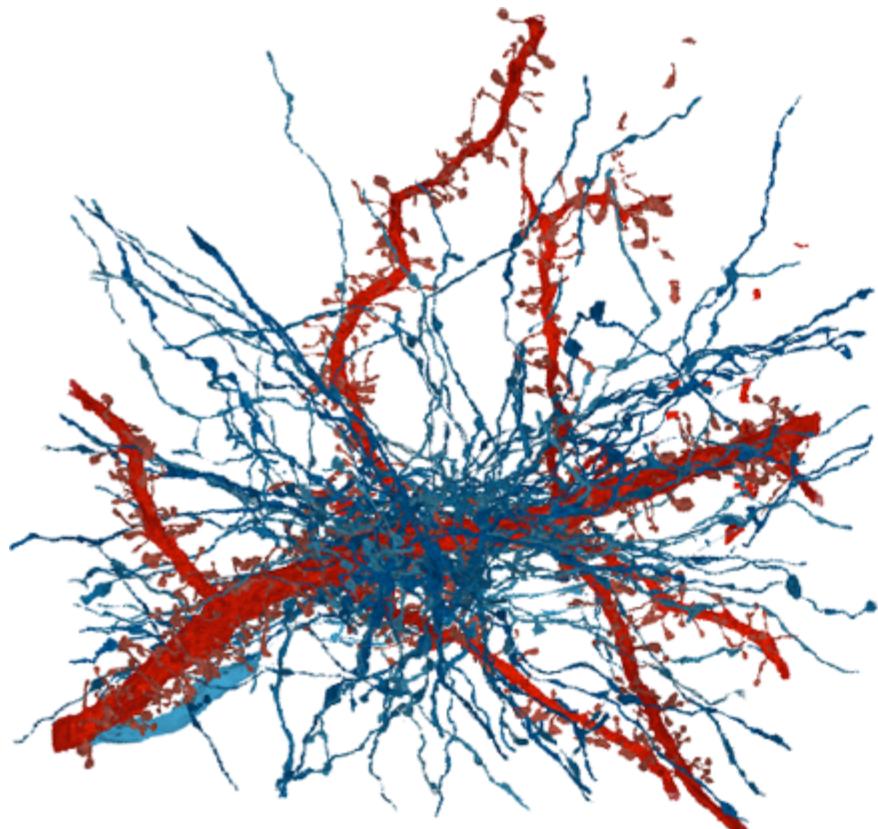








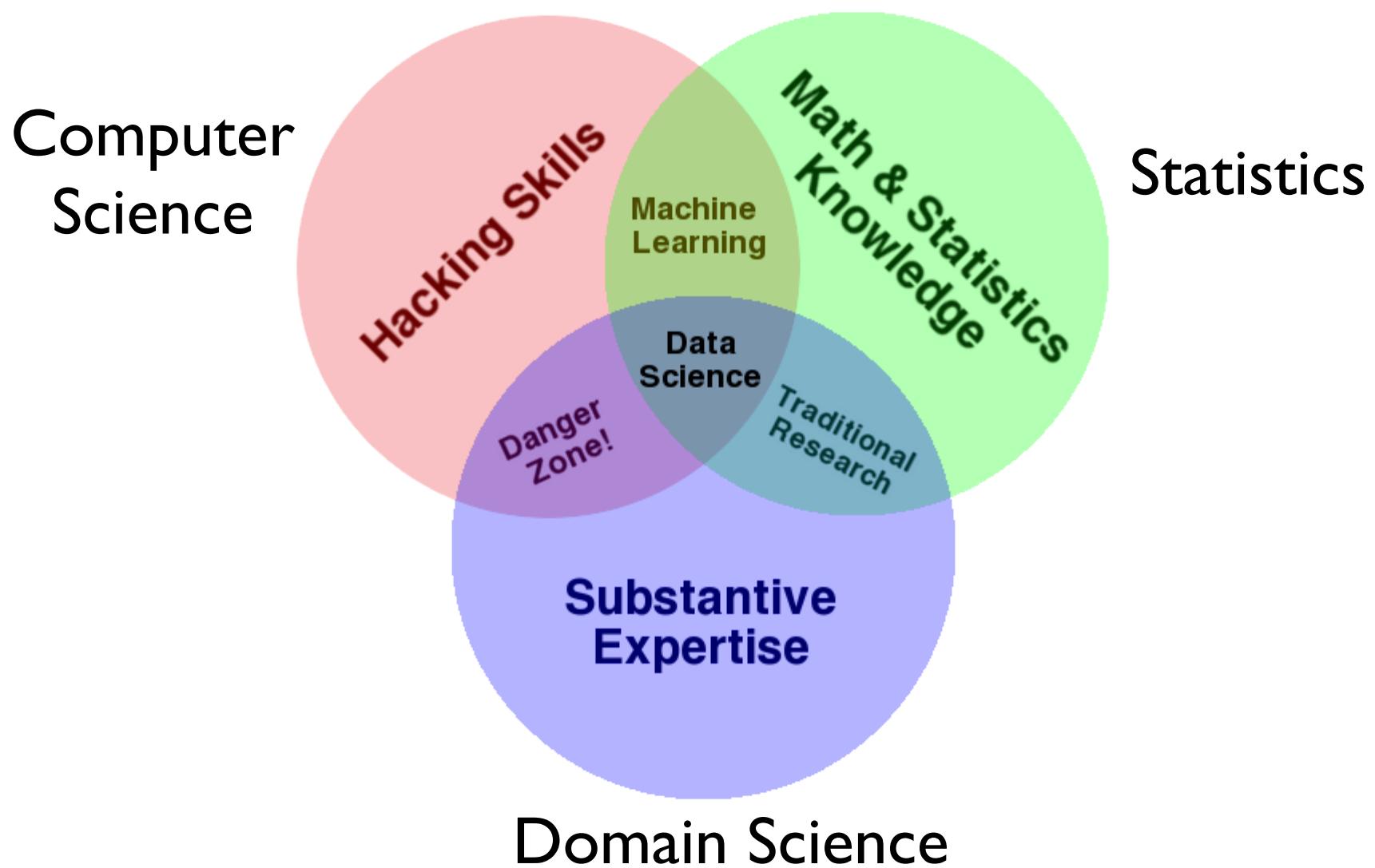
Analysis



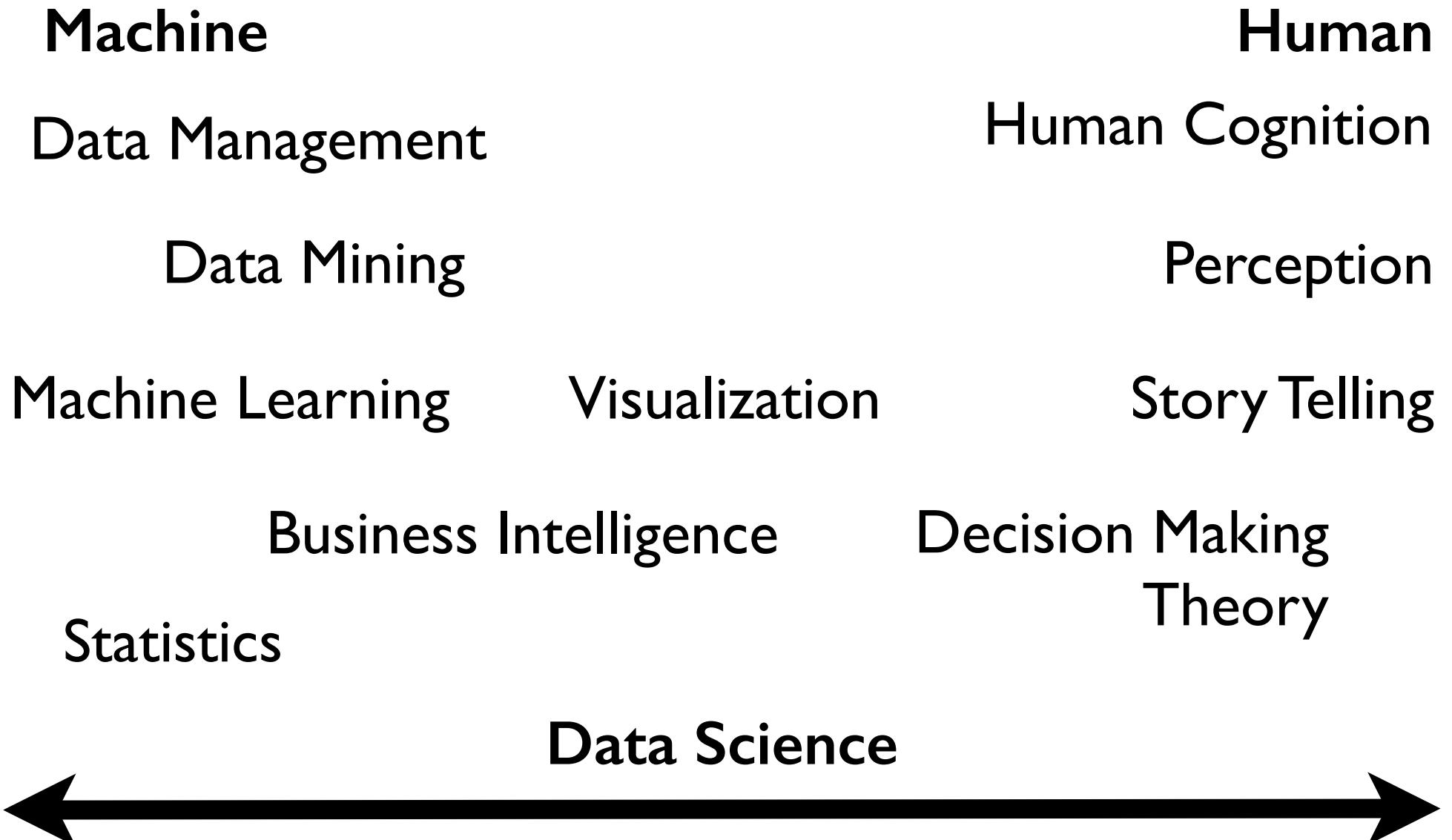
K. Al-Awami, et al.,

["NeuroLines: A Subway Map Metaphor for Visualizing Nanoscale Neuronal Connectivity"](#)
IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 2369-2378
2014

Data Science



Drew Conway

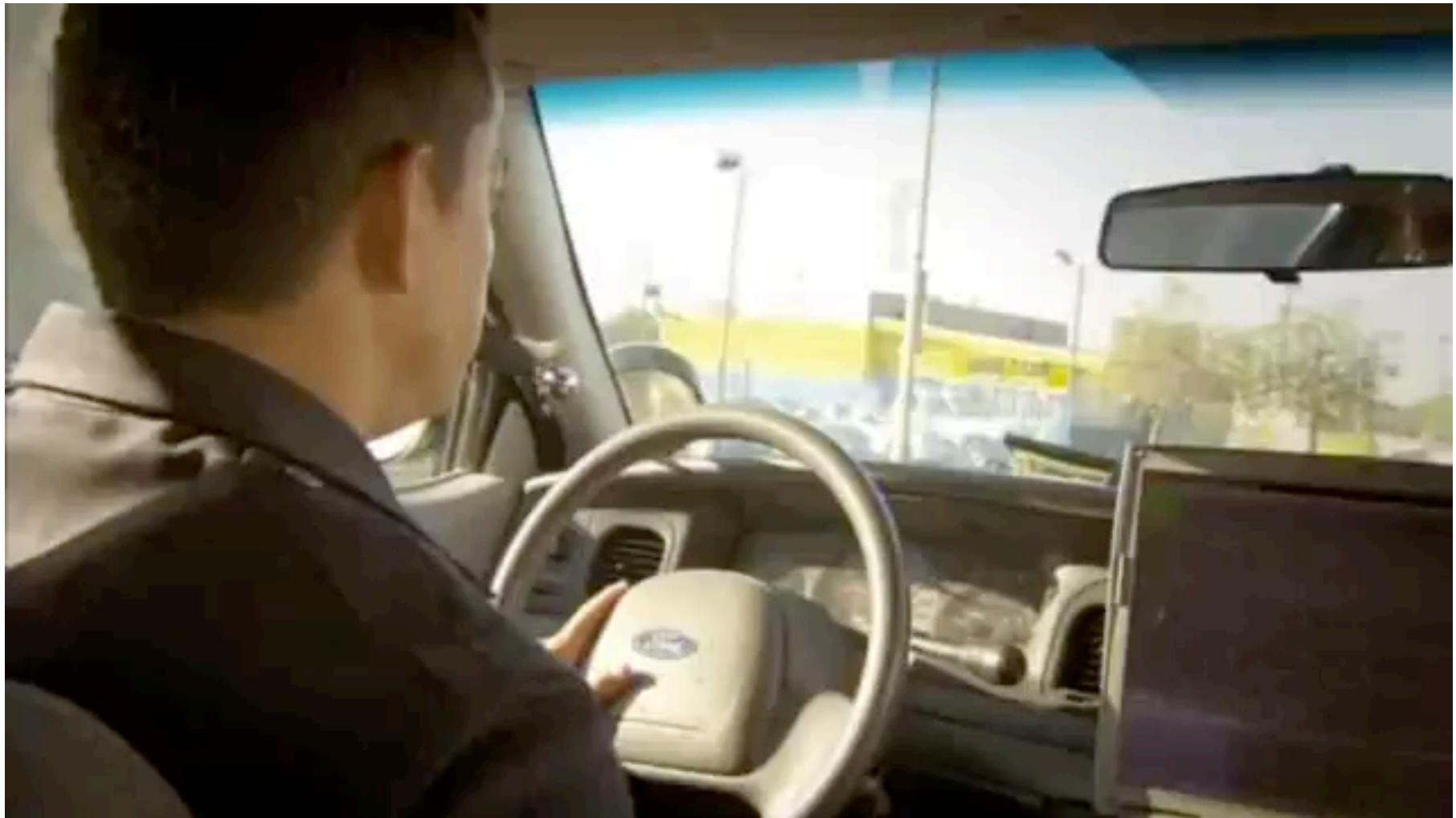


Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"

Outline

- What?
- Why?
- Who?
- How?

The Age of Big Data

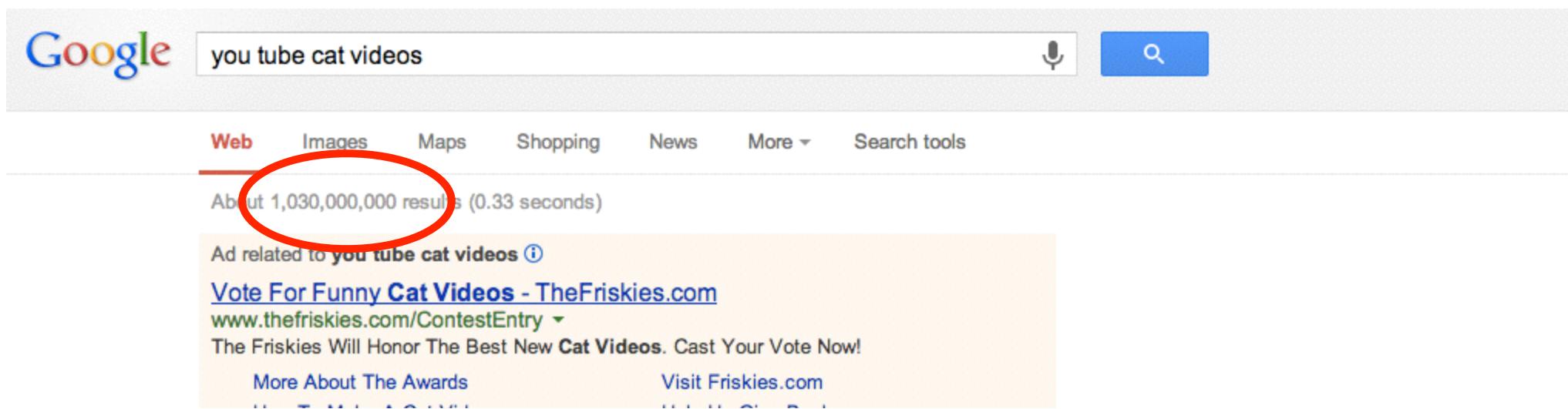


BBC, 2013

Big Data

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

Eric Schmidt, Google (and others)



A screenshot of a Google search results page. The search query "you tube cat videos" is entered in the search bar. Below the search bar, the "Web" tab is selected, while "Images", "Maps", "Shopping", "News", "More", and "Search tools" are also listed. A red circle highlights the text "About 1,030,000,000 results (0.33 seconds)". Below this, an advertisement for "TheFriskies.com" is displayed, encouraging users to vote for funny cat videos. The ad includes a link to "Vote For Funny Cat Videos - TheFriskies.com" and the URL "www.thefriskies.com/ContestEntry". The text "The Friskies Will Honor The Best New Cat Videos. Cast Your Vote Now!" is also visible. At the bottom of the ad, there are links for "More About The Awards" and "Visit Friskies.com".

In one second on the Internet there are...



THE BIG V'S OF **BIG DATA**

Turning Information Overload Into Big Sales

In the emerging market of Big Data, three "V" words have often been used to describe the issues at hand with information overload in our digital world.

THE EXISTING V'S

Big data has brought both great opportunity and change to the technological industry. Data scientists traditionally look at the existing V's, the ones that have classically been utilized to understand key variables of any data set.

VOLUME



Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.



VOLUME

Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.

IN ONE
DAY

2,500,000,000,000,000,000 BYTES ARE
CREATED IN THE DIGITAL UNIVERSE

ZETTABYTE =
1 SEXTILLION
BYTES

ZETTABYTES

2012

2015

2020

2.7
ZETTABYTES

7.9
ZETTABYTES

35
ZETTABYTES

The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.



ZETTABYTES

ZETTABYTES

ZETTABYTES

The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

VARIETY

In today's multi-faceted Internet culture, the great volume of data is also extremely varied in its form. So many variables can be thrown at a company that the true value of information can often be lost in the sea of data.



PURCHASE
TRANSACTIONS



WEBSITE
TRAFFIC



REWARDS
PROGRAMS



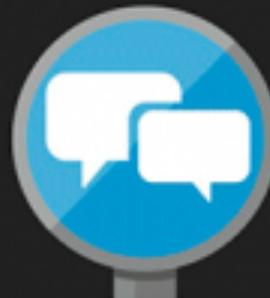
QUARTERLY
BUSINESS REPORTS



TWITTER



FACEBOOK



BLOG CONTENT

VELOCITY

Information is being created at a faster pace than ever before. The varied channels of big data are each increasing their output of content, daily.



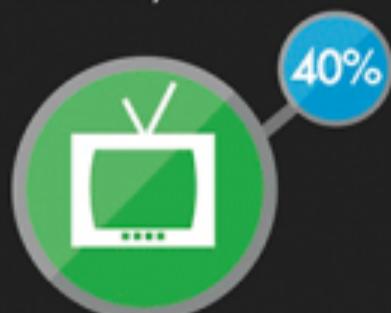
USERS GENERATE 2.7 BILLION LIKES ON FACEBOOK PER DAY



of the data in the world today has been created in the last two years alone



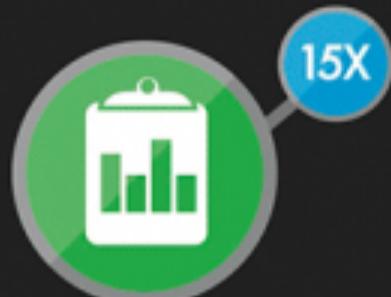
NEW TWEETS ARE CREATED BY ACTIVE USERS EACH DAY



40% of tweets are related to television and are beginning to be implemented in TV ratings



OF VIDEO IS UPLOADED TO YOUTUBE EVERY MINUTE

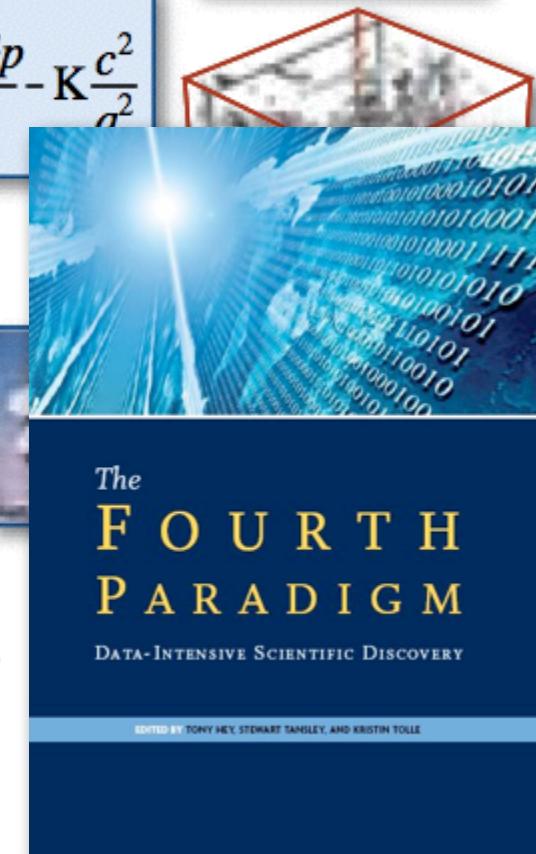


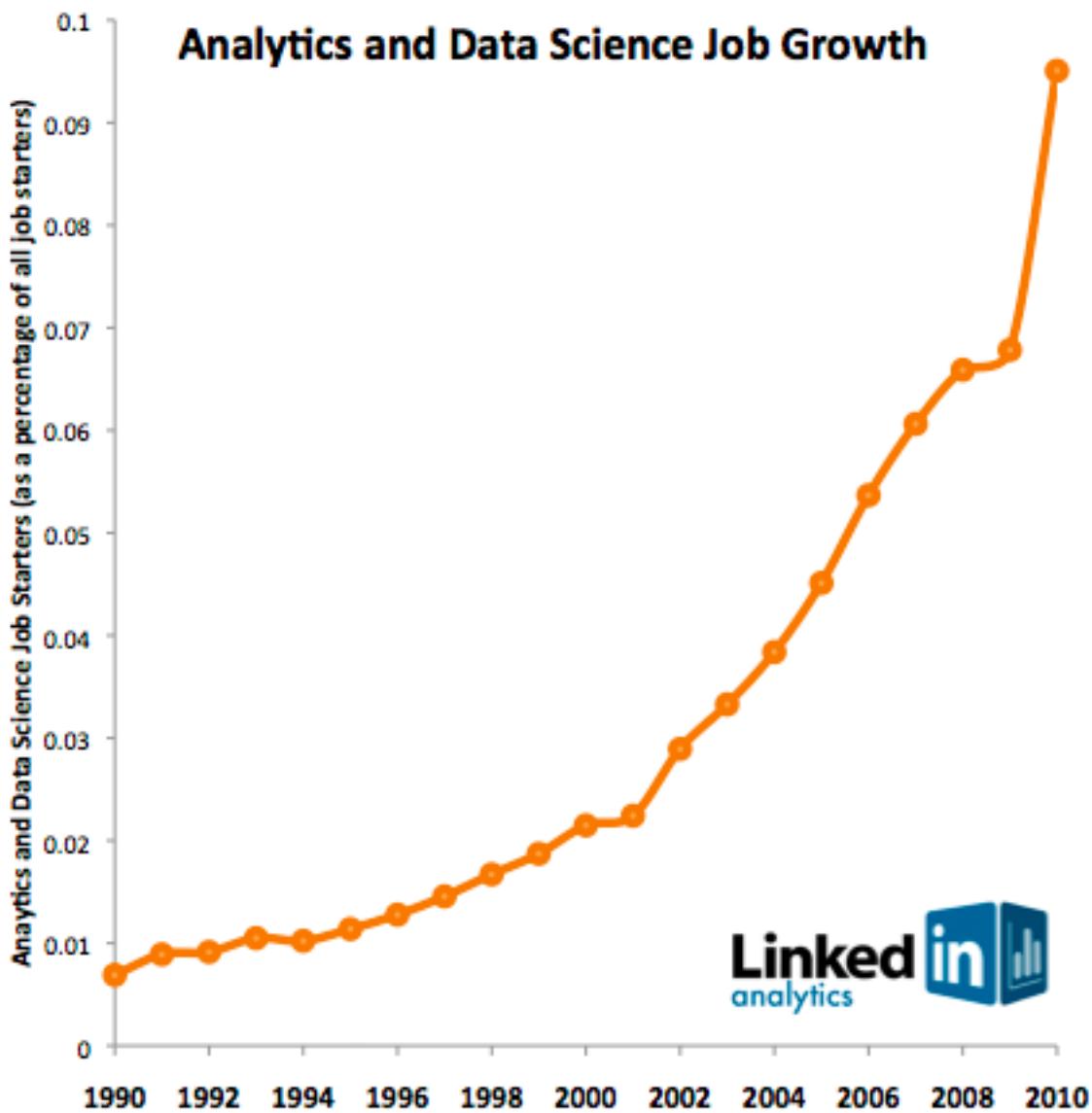
In 7 years, 15x the amount of data that exists today will be created every single year

Science Paradigms

- Thousand years ago:
science was empirical
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a computational branch
simulating complex phenomena
- Today: **data exploration (eScience)**
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K \frac{c^2}{\sigma^2}$$





“By 2018, the US could face a shortage of up to 190,000 workers with analytical skills”

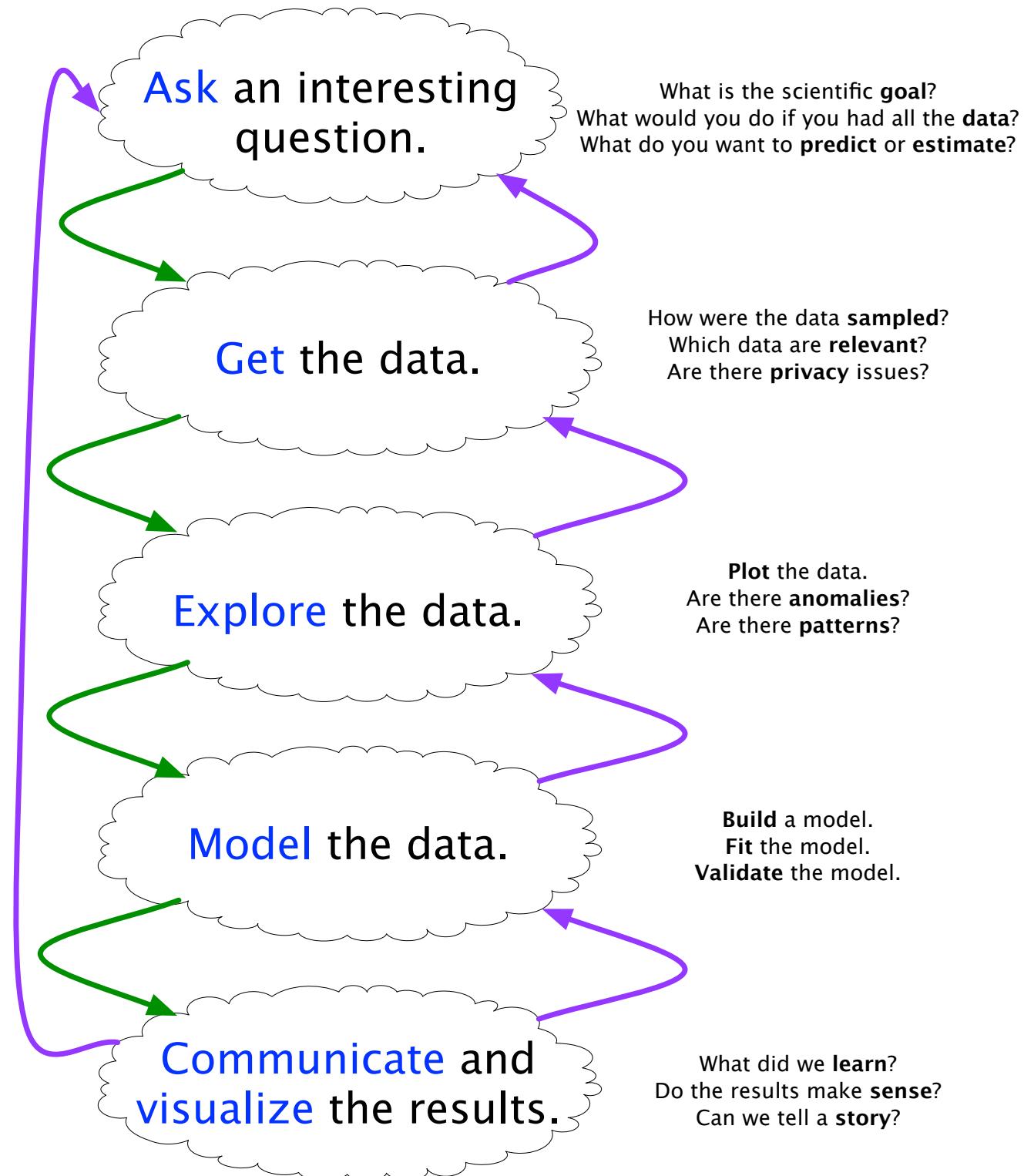
McKinsey Global Institute

“The sexy job in the next 10 years will
~~be statisticians.~~” *Data Scientists?*

Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

Hal Varian Explains...

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data.”** – Hal Varian



IPython Notebooks

<http://nbviewer.ipython.org/>

Home FAQ IPython Bookmarklet

IPython Notebook Viewer

A Simple way to share your IP[y]thon Notebook as Gists.

Share your own notebook, or browse others¹

Enter a gist number or url Go!

IP[y]: Notebook 01 Documenting your Research Journey

File Edit View Insert Cell Kernel Help

Documenting your Research Journey

The purpose of this code is to show how IPython notebooks can be used to document your GPU and the CPU. We compare the performance of each method using the system to document.

load image

```
In [1]: import PIL  
import PIL.Image  
  
image = PIL.Image.open("cinque_terre.jpg")  
image_array_rgb = numpy.array(image)  
  
r_original,g_original,b_original = numpy.split(image_array_rgb,  
a_original  
= numpy.ones_like(r_original)  
rgba_original = numpy.concatenate((r_original,  
  
figsize(6,4)  
matplotlib.pyplot.imshow(rgba_original);  
matplotlib.pyplot.title("rgba_original"))
```



Probabilistic Programming

Why would I want samples from the posterior, anyways?

We will deal with this question for the remainder of the book, and it is an understatement to say we can perform amazingly useful things. For now, let's finish with using posterior samples to answer the follow question: what is the expected number of tests at day t , $0 \leq t \leq 237$? Recall that the expected value of a Poisson is equal to its parameter, λ , so the question is equivalent to what is the expected value of λ at time t ?

In the code below, we are calculating the following: Let ℓ index a particular sample from the posterior distributions. Given a day t , we average over all λ_ℓ on that day t , using $\lambda_{\ell,t}$ if $t < t_1$ else we use $\lambda_{2,t}$.

```
matplotlib.legend.Legend at 0x1000000000000000  
Expected number of test messages received
```



XKCD Plot With Matplotlib

XKCD plots in Matplotlib

Out [1]:

CHECK IT OUT!



Sometimes when showing schematic plots, this is the type of figure I want to display. (But drawing it by hand is a good matplotlib. The problem is, matplotlib is a bit too precise. Attempting to duplicate this figure in matplotlib leads to:

Non Parametric Regression

Covariance function

The behavior of individual realizations from the GP is governed by the covariance function. The Matern class of functions is a flexible choice.

```
In [24]: from pygp.gp_cov_func import matern  
import numpy as np  
C = Covariance(eval_fun=maters.matern32, diff_degree=1.4, ampr=4., scale=1., rank_limit=1000)  
  
subplots(1,2)  
contourf(X, Y, C(X,Y).view(ndarray), origin='lower', extent=(-1,-1,1,1), cmap=mcm.bone)  
  
subplots(1,2)  
pcolor(X, Y, C(X,Y).view(ndarray), 'k-')
```

Exploring R formula

Let's test that with a fake design matrix

```
In [6]: show_x(np.random.normal(size=(10, 10)), 'My design')
```



Outline

- What?
- Why?
- Who?
- How?

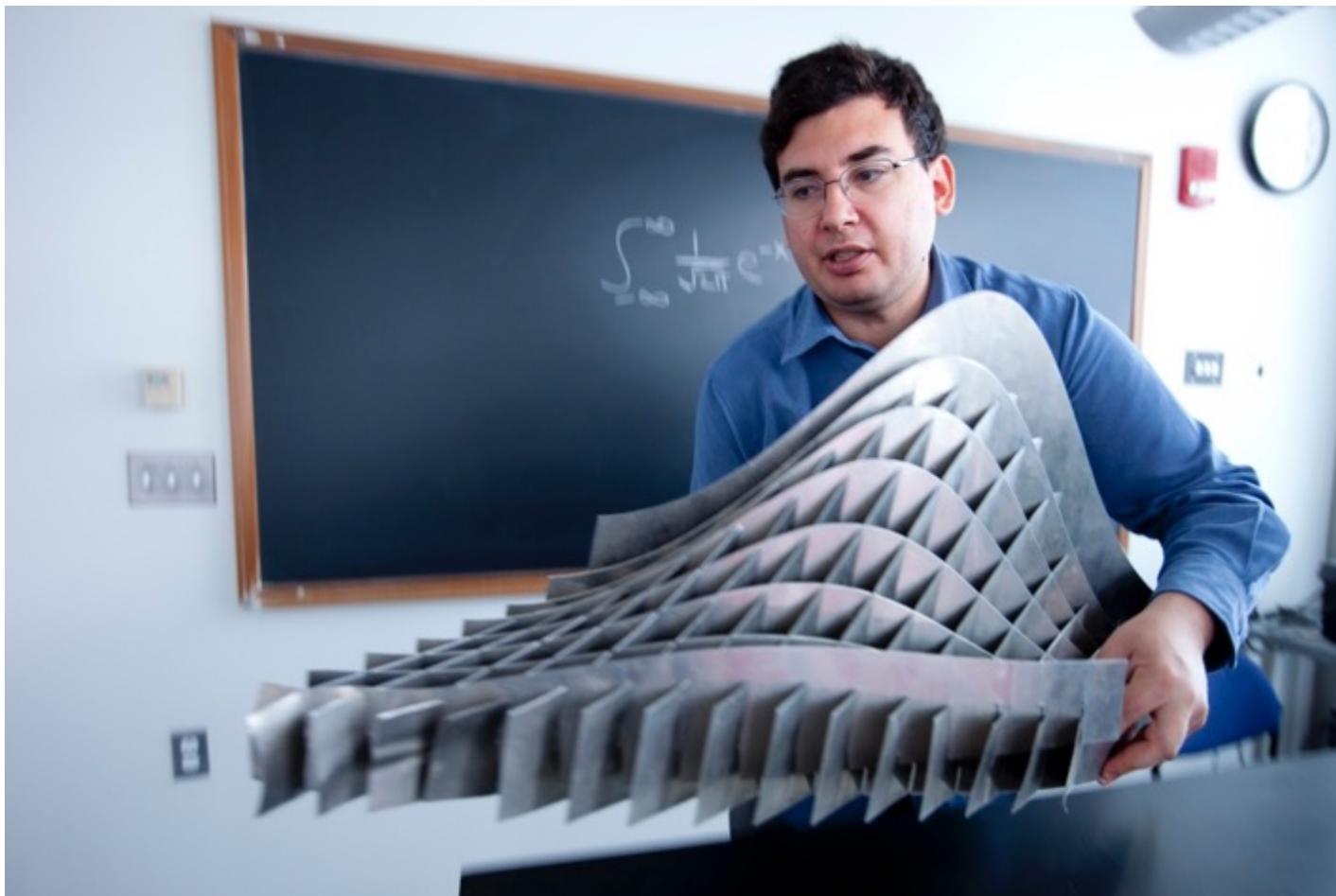
Hanspeter Pfister

An Wang Professor of Computer Science, SEAS
Director, Institute for Applied Computational Science
pfister@seas.harvard.edu / @hpfister



Joe Blitzstein

Professor of the Practice in Statistics,
Co-Director of Undergraduate Studies in Statistics
blitz@fas.harvard.edu, twitter @stat110, SC 714



Verena Kaynig-Fittkau

Lecturer and research scientist at IACS

vkaynig@seas.harvard.edu, NW B164



Rahul Dave

Head TF and Lecturer at IACS
rahuldave@gmail.com, NW B164



CS 109 Staff

Andrew Reece

Antonio Coppola

Austen Novis

Brian Feeny

Dana Katzenelson

Giri Gopalan

Irma Nomani

Jacob Dorabialska

Joseph Song

Kathy Li

Lawrence Kim

Leandra King

Luis Campos

Marcus Way

Michael Ma

Michael Packer

Nelson Santos

Richard Kim

Rick Wei-Jong Lee

Sail Wu

Stephen Klosterman

Xintao Qiu

Yingzhuo (Diana) Zhang

Yuhao Zhu

About You

Outline

- What?
- Why?
- Who?
- How?

CS109 Key Facets

- *data munging/scraping/sampling/cleaning* in order to get an informative, manageable data set;
- *data storage and management* in order to be able to access data quickly and reliably during subsequent analysis;
- *exploratory data analysis* to generate hypotheses and intuition about the data;
- *prediction* based on statistical tools such as regression, classification, and clustering; and
- *communication* of results through visualization, stories, and interpretable summaries.

Act I: Predictions

- Data Collection, “Munging”, and Storage
- Exploratory Data Analysis (EDA)
- Classification & Regression
- Cross Validation
- Dimensionality Reduction
- Effective Communication & Writing

Act II: Recommendations

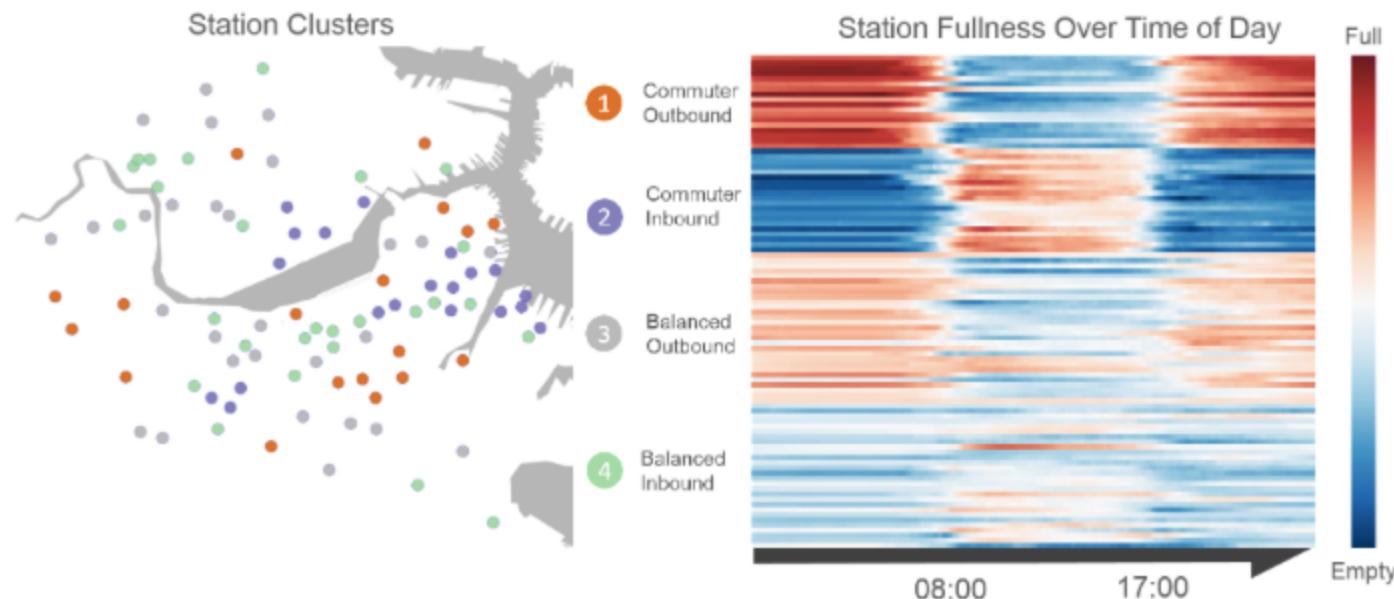
- Support Vector Machines
- Decision Trees & Random Forests
- Bagging & Boosting
- Machine Learning Best Practices
- MapReduce, Amazon's EC2, and Spark

Act III: Clustering & Text

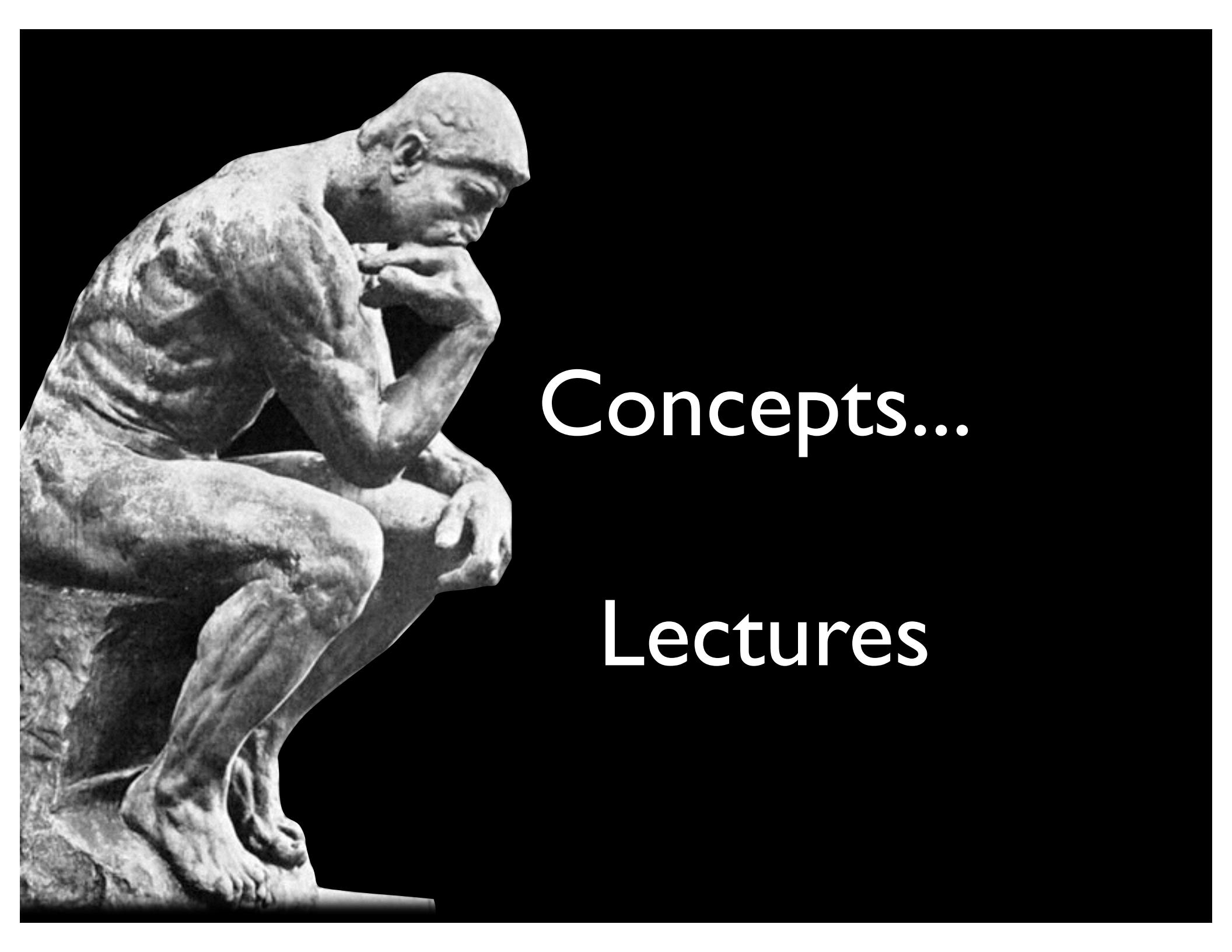
- Bayesian Thinking & Naive Bayes
- Text Analysis: LDA & Topic Modeling
- Clustering
- Effective Presentations
- Deep Learning
- Guest Lecture: Experimental Design



CS109 Data Science



Predicting Hubway Stations Status by Lauren Alexander, Gabriel Goulet-Langlois, Joshua Wolff

A black and white photograph of Auguste Rodin's sculpture 'The Thinker'. The statue depicts a man in a crouched, contemplative pose, resting his chin on his hand. The background is solid black.

Concepts...

Lectures

...and Skills
Sections



Sections

- Introduce tools & skills; available as lab notebooks and videos
- Mandatory, except for DCE students
- First (group) section this Friday!
 - 10am-12pm in MD G115
- Regular sections first week as office hours to get help with Python, Git, and HW0

Section Schedule (TBD)

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 AM			Rahul, NW-B150		
10:00 AM	Leandra	Ima	Steve		Luis
11:00 AM	NW-B150	NW-B150	NW-B150		NW-B150
12:00 PM					
1:00 PM	Diana				Michael Ma
2:00 PM	NW-B150	Lecture	Lawrence	Lecture	NW-B150
3:00 PM	Joseph	NW-B103	NW-B150	NW-B103	Michael Packer
4:00 PM	NW-B150		Antonio		NW-B150
5:00 PM		Sail, Nelson	NW-B150		
6:00 PM	Austen, Dana	NW-B150 and B166	Richard		
7:00 PM	NW-B150 and B166	Kathi	NW-B150		
8:00 PM		NW-B150			

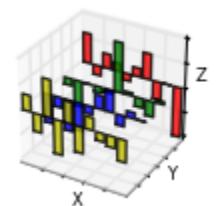
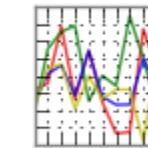
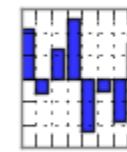
Homework

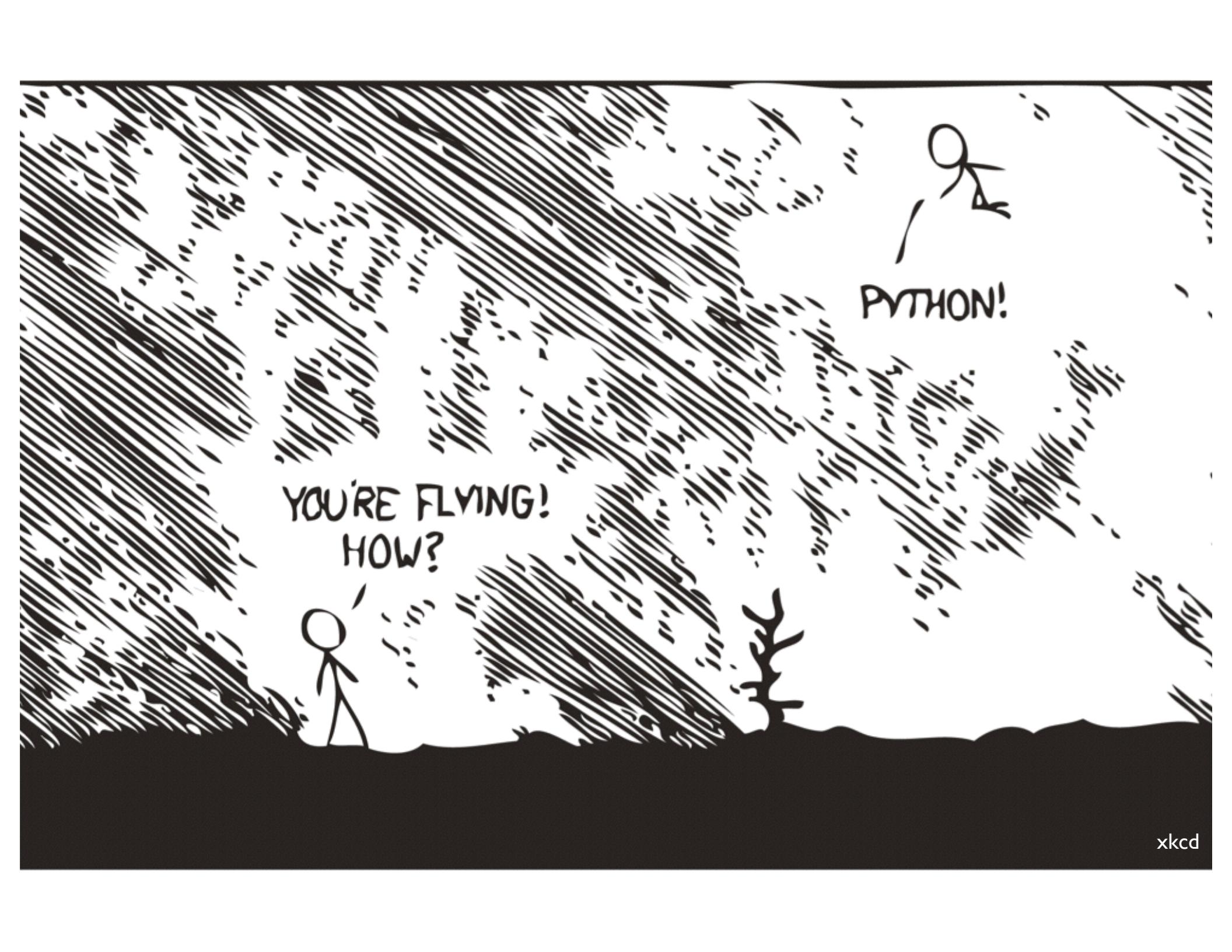
- Real-World focus
 - Scrape and wrangle messy data
 - Apply sophisticated statistical analysis
 - Visualize and communicate results
- Election data, music charts, recommendations, etc.

Programming

IP[y]: IPython
Interactive Computing

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





YOU'RE FLYING!
HOW?

PYTHON!

The screenshot shows a web browser window for the Piazza platform. The URL is piazza.com/cs109/huzz8a158o7j0?cid=1695. The page title is "Piazza". The navigation bar includes links for DRB | GitLab, The Hub, Feedly, MD Syntax, Add to Pinboard, My Profile, Logout, Patrick Box, Timesheet, LaTeX symbols, Lore, 3G Mobile Hotspot, and a search bar. The user profile is Hanspeter Pfister.

The main content area displays a note titled "note" with a star icon. The note text reads: "I This class has been made inactive. No posts will be allowed until an instructor reactivates the class." It has 111 views and an "Actions" button.

On the left, there's a sidebar with sections for "PINNED" posts and weekly summaries:

- PINNED**:
 - Instr data science talks 2/28/14: Dear all, I hope you are enjoying the new semester! It's been great to hear that many of you are going further into data science.
 - Instr Reminder: Weathering the D... 1/21/14: Welcome back and happy new year! A quick reminder that there's a science symposium coming up that we are organizing.
 - Instr Data Science Symposium, J... 12/13/13: We hope to see many of you at the data science symposium in January. Don't forget the final research presentations!
 - Instr Staff E-mail Addresses 1/22/15: If you want to contact your grading TF, here are the addresses (WIP) Alexander Lex - alexander.lux@seas.harvard.edu; Nicholas Rau - nicolas.rau@seas.harvard.edu.
- WEEK 4/19 - 4/25**:
 - Small research grants for projects explained 4/22/15: NASA recently announced a new data API for piles of their information. Some of you will be interested in attending the presentation.
- WEEK 3/8 - 3/14**:
 - Data Science Fellow at DrivenData 3/9/15: Join DrivenData as a Data Science Fellow! DrivenData is a social enterprise that aims to bring the power of data closer to people.
- WEEK 11/23 - 11/29**:
 - Job at OpportunitySpace 11/27/14: Hi all, I took this class two semesters ago and started a company that leverages data to make our cities better. We're looking for a few more people to join us.
- WEEK 11/16 - 11/22**:
 - Interesting article on Big Data and HSPH 11/16/14: www.hbr.org/2013/11/big-data-is-changing-healthcare.aspx

At the bottom, there are sections for "Average Response Time" (20 min), "Special Mentions" (Anonymous answered Big Data Rocks! in 5 min. 1 year ago), and a footer with links to Online Now, This Week, and a stats summary (1 post, 7 comments).

Grades

- No exams!
- 50% Homework
- 40% Projects (3-4 person teams)
- 10% Participation (Piazza & Sections)
- 10 point scale, holistic grading

Projects

DRB | GitLab The Hub Feedly MD Syntax Add to Pinboard My Pinboard Instapaper libx bit.ly Orrick Box Timesheet LaTeX symbols Lore 3G Mobile Hotspot +

[View on GitHub](#)

Predicting Hubway Stations Status

Lauren Alexander, Gabriel Goulet-Langlois, Joshua Wolff

Video Overview Analysis Prediction

[tar.gz](#) [.zip](#)

CS109 Class Project

95 Stations
1,300 Bikes
500,000 Trips

Project Overview

Motivation

A major challenge for bicycle sharing programs like Hubway is network imbalance. Some

Policies

- HWs due on Thursdays, 11:59 pm EST
- 6 late days for HW (no questions asked)
- Cannot submit HW later than 2 days
- Regrading requests within 7 days in writing
 - Grade may improve or go down

Collaboration Policy

- Work you turn in must be your own
- Projects are a 3-4 person team effort
 - With project group peer assessment
- Acknowledge all help and code you used
- Harvard Honor Code

Is this course for me ???



Prerequisites

Programming experience

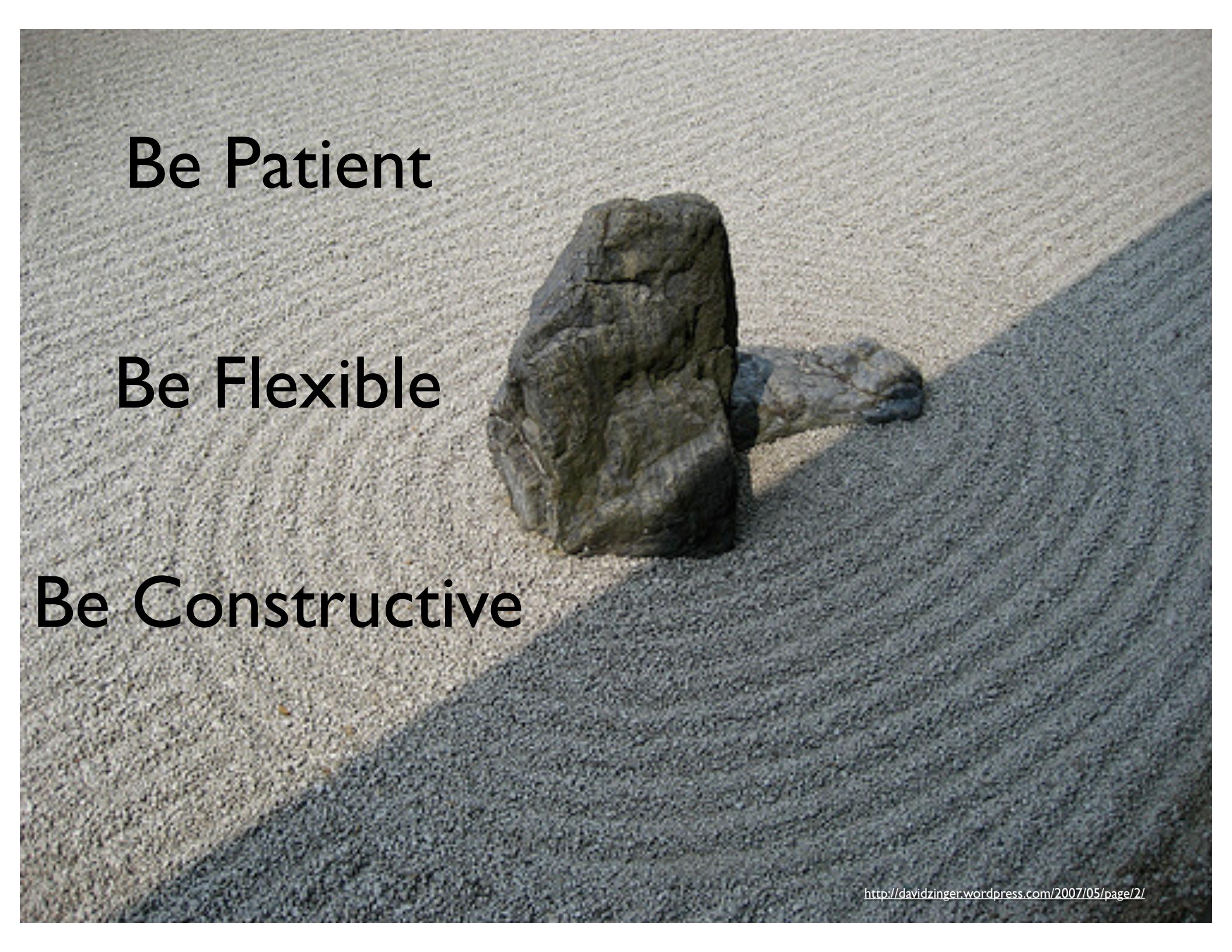
- CS50 and/or C, C++, Java, Python, etc.

Basic statistical knowledge

- STAT100, ideally STAT110

Willingness to learn new software & tools

- This can be time consuming
- You will need to read online documentation



Be Patient

Be Flexible

Be Constructive

Next Steps

- HW 0, mandatory, needs to be submitted!
 - Good test of your basic skills
 - Complete the survey by tomorrow! Needed to be able to submit HW 0
 - Installation of several Python frameworks
 - Not graded, do it as soon as possible
- Read syllabus carefully

Important Links

- Create a github account at <http://github.com>
- Then fill in our survey at <http://goo.gl/forms/bJwajS8zO8>
- HW 0 document at [https://github.com/cs109/2015lab1/
blob/master/hw0.ipynb](https://github.com/cs109/2015lab1/blob/master/hw0.ipynb)
- Week 1 notebooks at <https://github.com/cs109/2015lab1>
- HW repositories will be created for you on github. See HW 0 for details.

