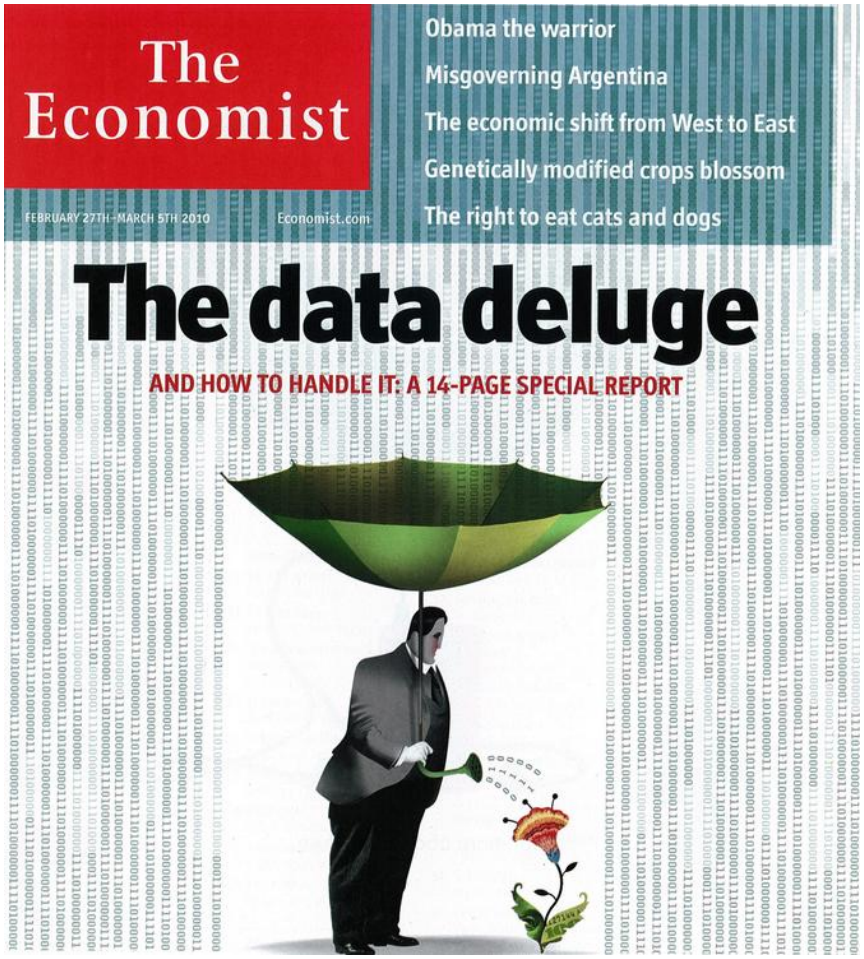
TOPICS IN DATA AND COMPUTATIONAL SCIENCE

If you are **not** enrolled in the **Master's in Data Science program,** please leave. This course is capped and can not be taken otherwise (However, this is likely to change next year. So you might try again in a year)

UGUR CETINTEMEL
TIM KRASKA
DAN POTTER

teaching datascience .org


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
The End of Science
The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.


Google

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

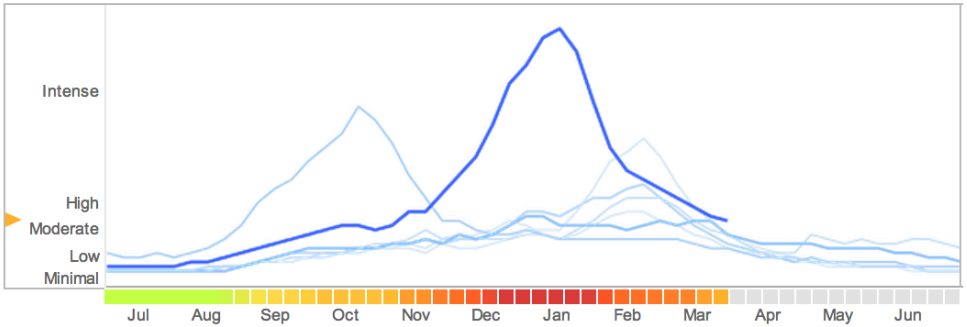

NETFLIX
MENU
tv

Explore flu trends - United States

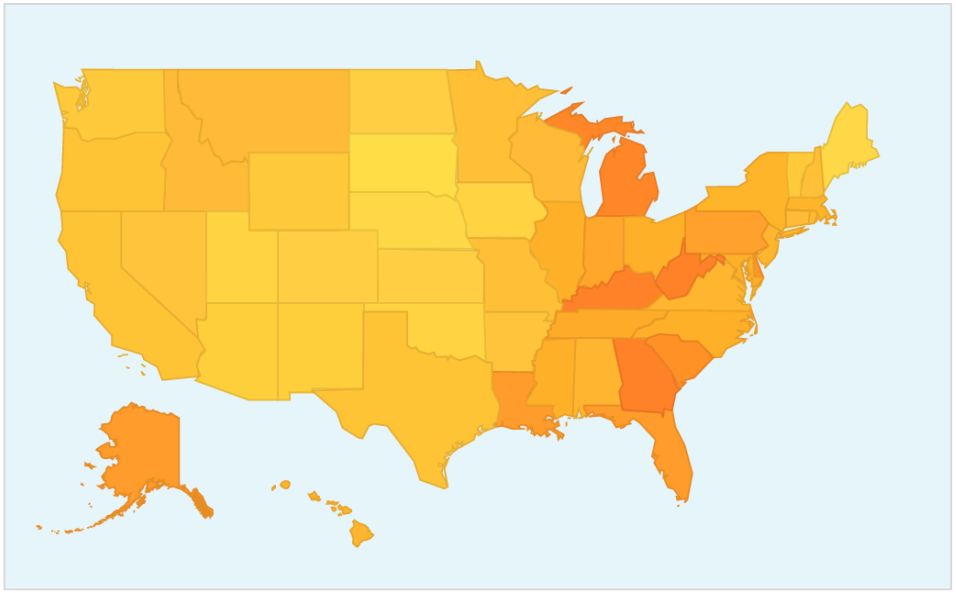
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

2012-2013 Past years ▼



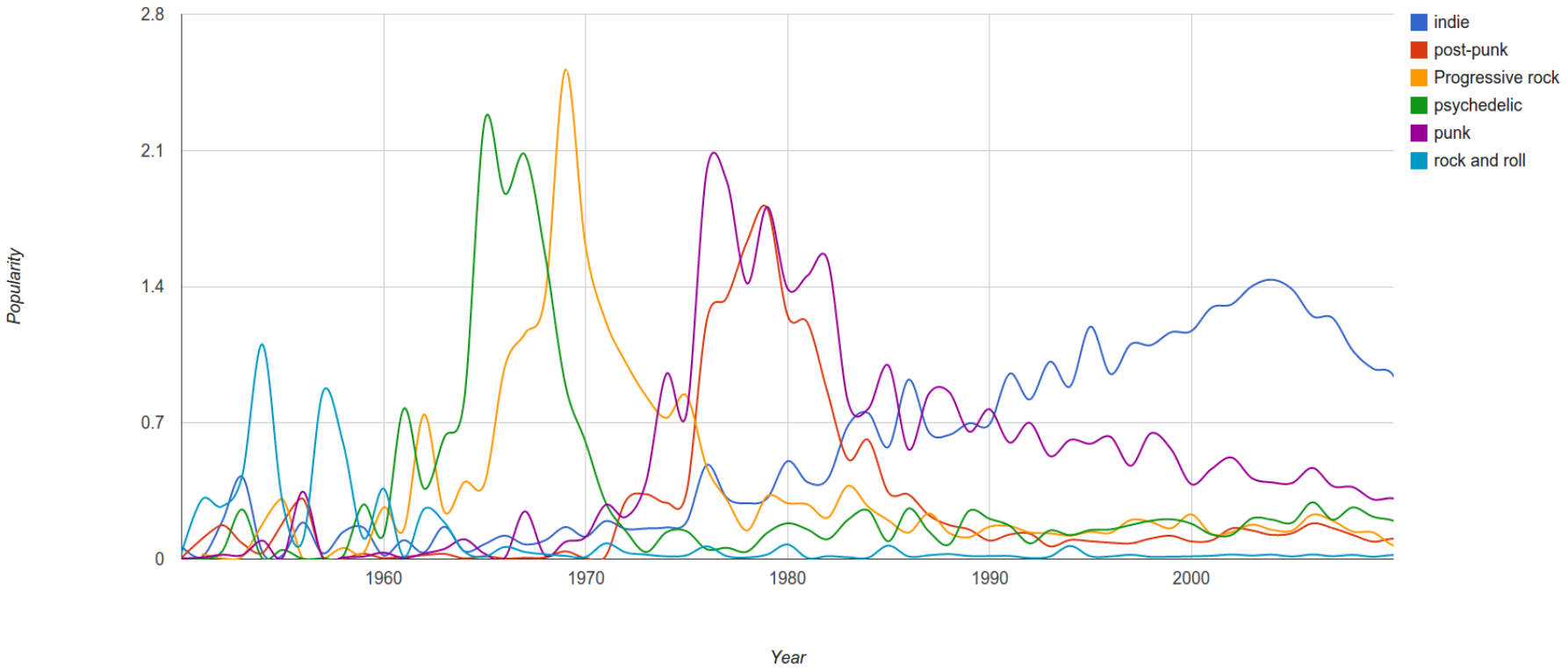
States | Cities (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.

source:
http://www.google.org/flutrends/us/#US

LAST.FM



"Since we have a massive amount of user tag data available we can easily correlate tags and years and measure "popularity" of a genre by counting the number of artists formed in a specific year."

Janni Kovacs, Last.FM

EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY

1) Convert all the digitized books in the 20th century into n-grams
(Thanks, Google!)
(http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

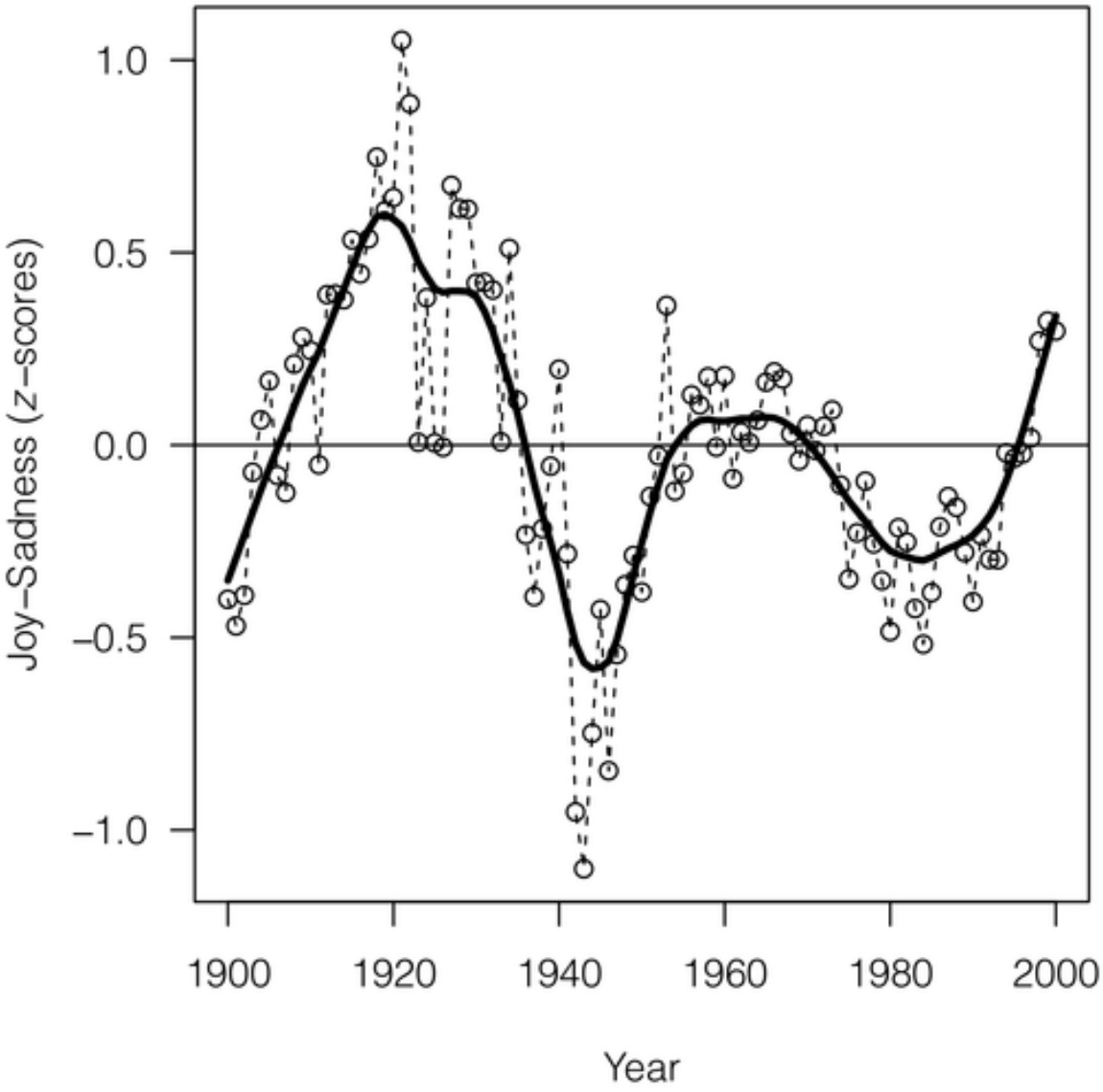
2) Label each 1-gram (word) with a mood score.
(Thanks, WordNet Affect)

3) Count the occurences of each mood word

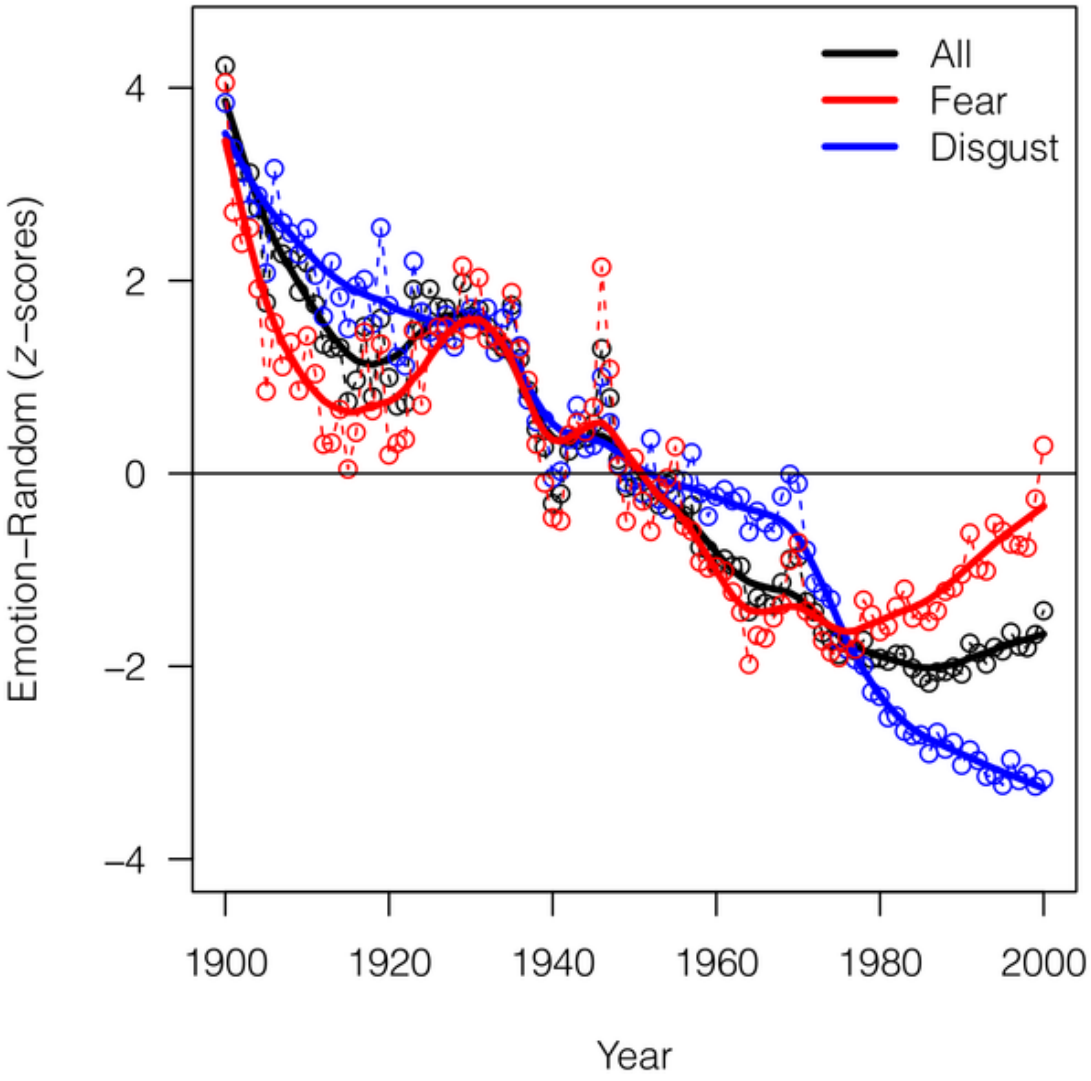
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

Flavor network and the principles of food pairing

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási

Affiliations | **Contributions** | **Corresponding authors**

Scientific Reports **1**, Article number: 196 | doi:10.1038/srep00196

Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.

bon appétit

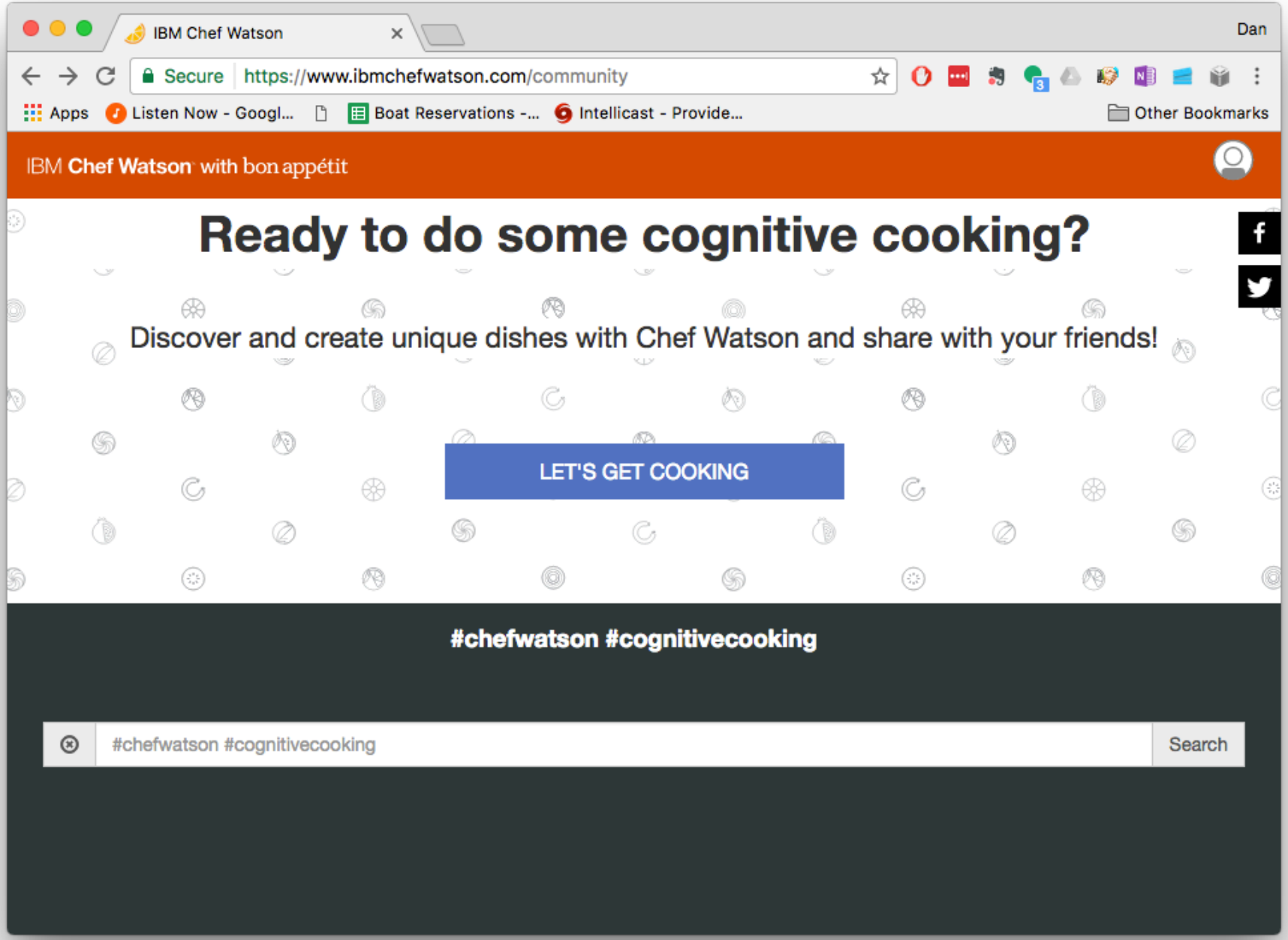
SUBSCRIBE

Search Bon Appétit

TRENDS + NEWS

We Spent a Year Cooking With the World's Smartest Computer—and Now You Can, Too

http://www.bonappetit.com/entertaining-style/trends-news/article/chef-watson-app


IBM Chef Watson
Dan
Secure
https://www.ibmchefwatson.com/community
Apps
Listen Now - Googl...
Boat Reservations -...
Intellicast - Provide...
Other Bookmarks
IBM Chef Watson with bon appétit
Ready to do some cognitive cooking?
Discover and create unique dishes with Chef Watson and share with your friends!
LET'S GET COOKING
#chefwatson #cognitivecooking
#chefwatson #cognitivecooking
Search

Estimating Sheep Pain Level Using Facial Action Unit Detection

Yiting Lu, Marwa Mahmoud and Peter Robinson
Computer Laboratory, University of Cambridge, Cambridge, UK

***Abstract*— Assessing pain levels in animals is a crucial, but time-consuming process in maintaining their welfare. Facial expressions in sheep are an efficient and reliable indicator of pain levels. In this paper, we have extended techniques for recognising human facial expressions to encompass facial action units in sheep, which can then facilitate automatic estimation of pain levels. Our multi-level approach starts with detection of sheep faces, localisation of facial landmarks, normalisation and then extraction of facial features. These are described using Histogram of Oriented Gradients, and then classified using Support Vector Machines. Our experiments show an overall accuracy of 67% on sheep Action Units classification. We argue that with more data, our approach on automated pain level assessment can be generalised to other animals.**

I. INTRODUCTION

Pain level assessment is critical to the welfare of sheep. Severe pain in sheep often indicates diseases, such as footrot [16] and mastitis [17]. Recognising and quantifying pain are essential to the subsequent treatment and pain alleviation [18]. Moreover, efficient and reliable pain assessment tools would help with early diagnoses.

Facial expressions are often used as an indicator of pain level in animals [2], [15]. The Sheep Pain Facial Expression Scale (SPFES) [1] has recently been introduced. It is a standardised measure to assess pain level based on facial expressions of sheep, and has been shown to recognise pain in sheep faces with high accuracy. However, training of scorers and the scoring process can be time-consuming, and individual bias may lead to inconsistent scores [1].

In this paper, we have used computer vision techniques to automate the analysis of facial expressions in sheep. Our approach can improve efficiency and ensure consistency in estimation of pain. We have deployed techniques that are widely used in human emotion recognition to address the problem of automatically assessing pain in sheep.

The overall pipeline of our sheep pain level estimation system is shown in Fig. 1. The main contributions of this paper can be summarised as follows:

1) Introducing a preliminary taxonomy for sheep facial Action Unit (AUs) based on the SPFES.
2) Presenting an automatic multi-level approach for estimating pain level in sheep by extending computer vision techniques that have been widely used in human emotion recognition.
3) Demonstrating that our approach can successfully classify 9 facial action units of sheep and can automatically estimate pain levels. We also show that our approach is generalisable across different dataset of sheep faces. Finally, we argue that - with their pain scales calibrated - the proposed automatic pain level estimation approach can be generallised to other animals, such as mice [12] [5], rabbits [14] and horses [13].

978-1-5090-4023-0/17/$31.00 ©2017 IEEE

We start by reviewing the related work in Section 2. A description of our dataset is discussed in Section 3. Our methodology is described in section 4 followed by the experimental evaluation in Section 5. Finally, conclusions and future work are presented in Section 6.

II. RELATED WORK

Analysing facial expressions of animals was first introduced by Langford *et al.* [4] to facilitate detection of pain level in mice. This approach has been advanced and generalised to many other animals. Yet, manual scoring is the usual practice and automatic assessment of pain level is still an underdeveloped area.

Recently, a standardised sheep facial expression pain scale SPFES was developed by McLennan *et al.* [1]. They showed that their approach is able to recognise sheep pain face with high degree of accuracy. Since manual labelling was used, they found that for different scorers, the accuracy of the pain assessment ranged from 60% to 75%. Their work is the basis of our sheep AU taxonomy.

Sotocinal *et al.* [5] attempt to automate animal pain assessment. They introduced a partially automated approach for pain level assessment on rats. A Haar feature cascade classifier is used for real time eye and ear detection. The classifier served as a pre-screening tool so that only frames detected with the key features are kept as candidates for manual assessment. They found such partially automated pain recognition largely solves the labour-intensive problem of manual scoring.

Yang *et al.* [6] analysed sheep faces and proposed a novel approach to localise sparsely distributed facial landmarks, which uses triplet-interpolated feature (TIF) extraction scheme under the cascade pose regression (CPR) framework [7]. They applied the TIF model on sheep, and reported good results regardless of sheep breed, head pose, partial occlusion, etc. Yet, their work assumed sheep face bounding boxes are known. In our work, we implement sheep face detection before applying the TIF model, then we use the localised sheep facial landmarks for later AU detection.

III. DATA

Unlike human AU analysis, facial expression recognition of sheep is still an underdeveloped area. Very few datasets are available on sheep and fewer include ground truth labels

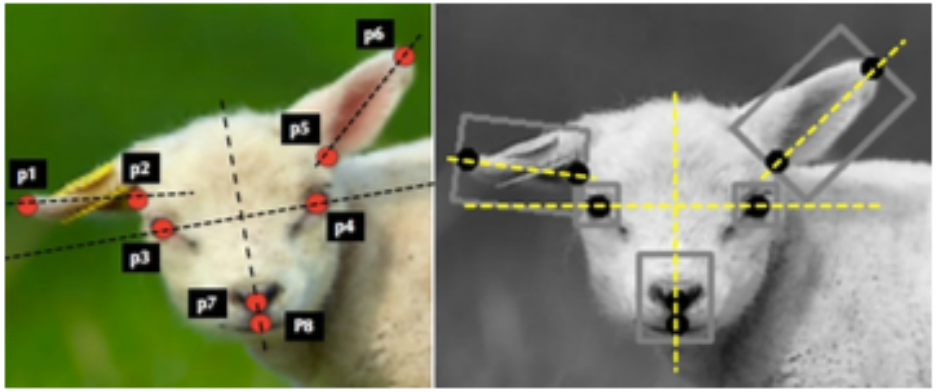


Fig. 3. Left: Localised facial landmarks (Note: the eight facial landmarks are labelled from p1 to p8) Right: Normalised sheep face marked with feature bounding boxes

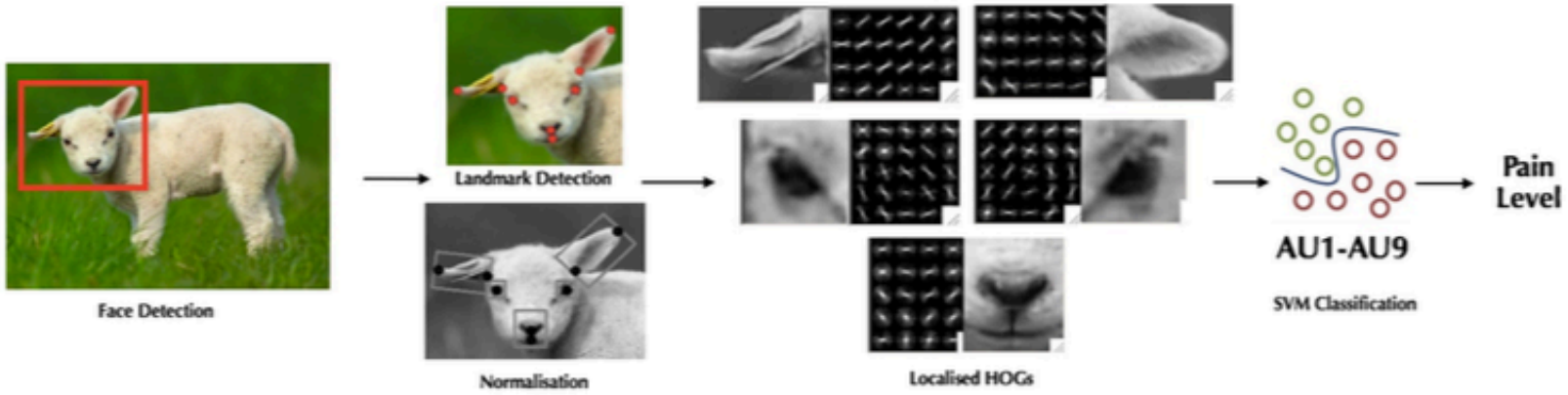
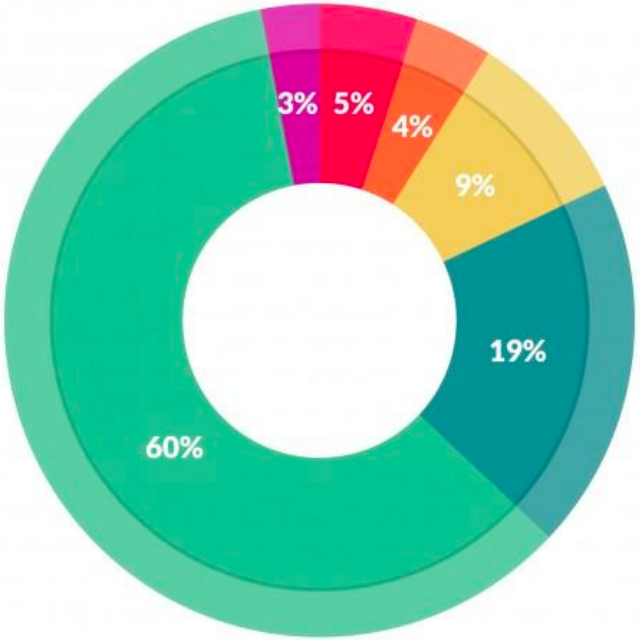


Fig. 1. The pipeline of our automatic approach to estimate pain level in sheep


CAUTION
Big Data
on board



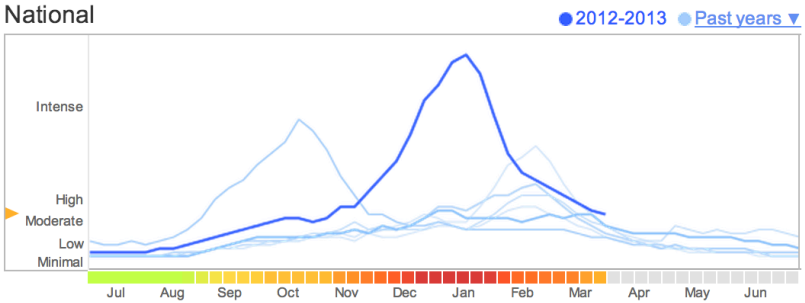
What data scientists spend the most time doing

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

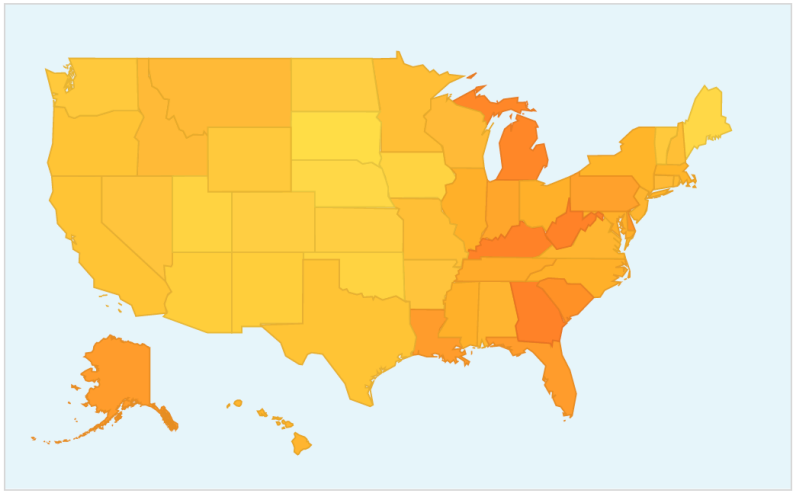
Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

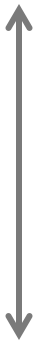
National



States | Cities (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.



flu risk

"Scientific hindsight shows that Google Flu Trends far overstated this year's flu season…."

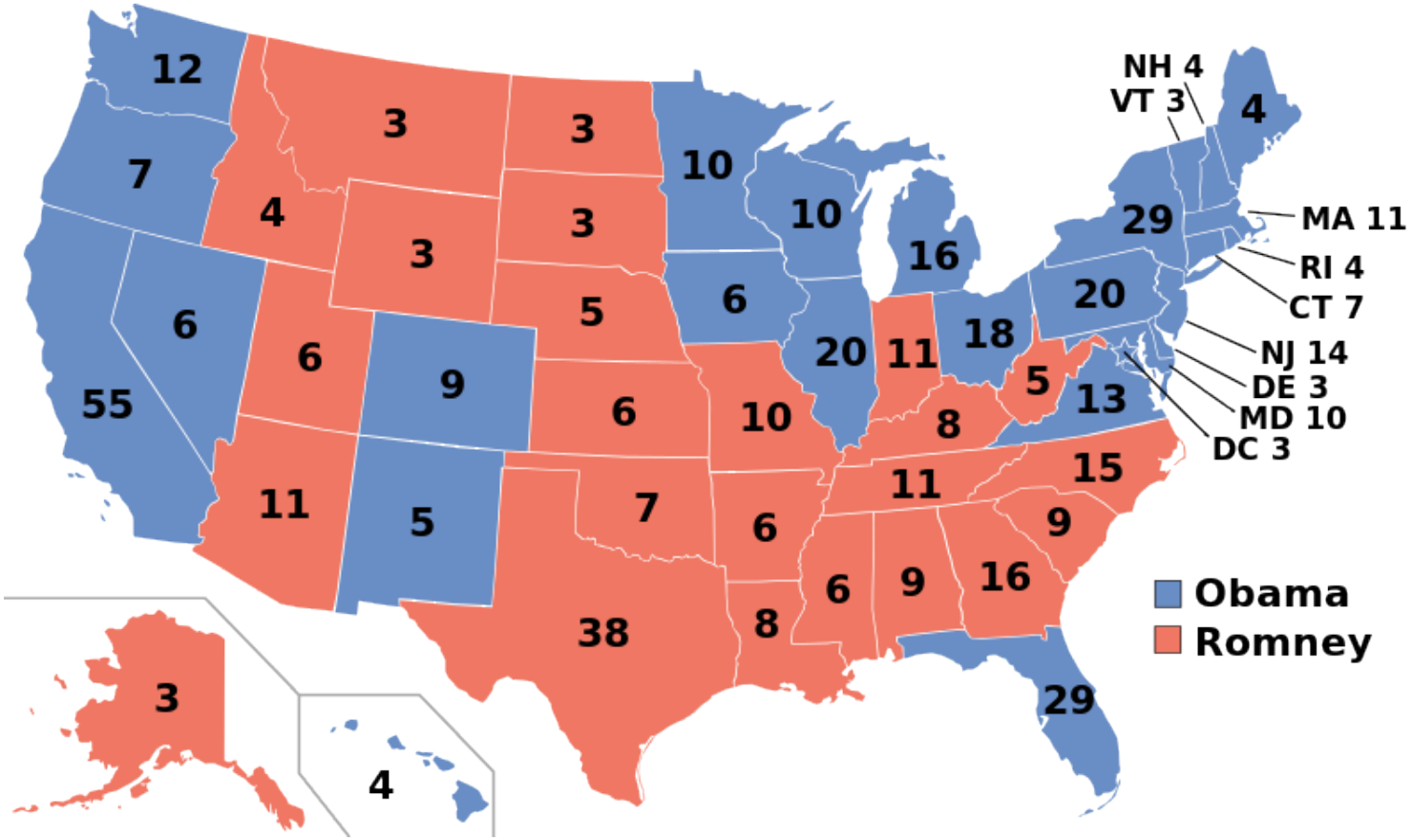
"Lots of media attention to this year's flu season skewed Google's search engine traffic."

David Wagner, Atlantic Wire, Feb 13 2013

source: http://www.google.org/flutrends/us/#US

NATE SILVER

"Silver, who made his name by using cold hard math to call 49 out of 50 states in the 2008 general election and all 50 in 2012"



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Nate Silver

source: randy stewart

"The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes."

Nate Silver, Oct. 26, 2012

fivethirtyeight.com

"…the argument we're making is exceedingly simple. Here it is: Obama's ahead in Ohio."

Nate Silver, Nov. 2, 2012

fivethirtyeight.com

"The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign."

Nate Silver, Nov. 10, 2012

DailyBeast

RELATED: OBAMA CAMPAIGN'S DATA-DRIVEN GROUND GAME

"In the 21st century, **the candidate with [the] best data**, merged with the best messages dictated by that data, **wins**."

Andrew Rasiej, Personal Democracy Forum

"…the **biggest win came from good old SQL** on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have **Hadoop** do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

Also a good read: http://fivethirtyeight.com/features/a-history-of-data-in-american-politics-part-2-obama-2008-to-the-present/

How Nate Silver Missed Donald Trump

The election guru said Trump had no shot. Where did he go wrong?

2.2k 318 980

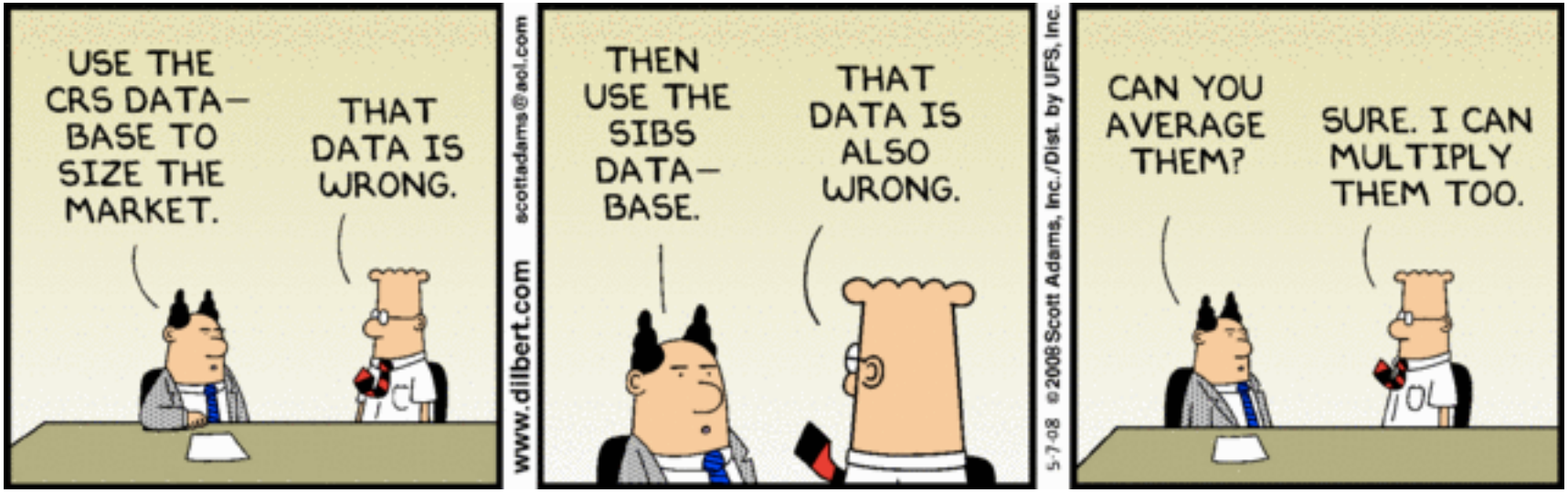
By *Leon Neyfakh*

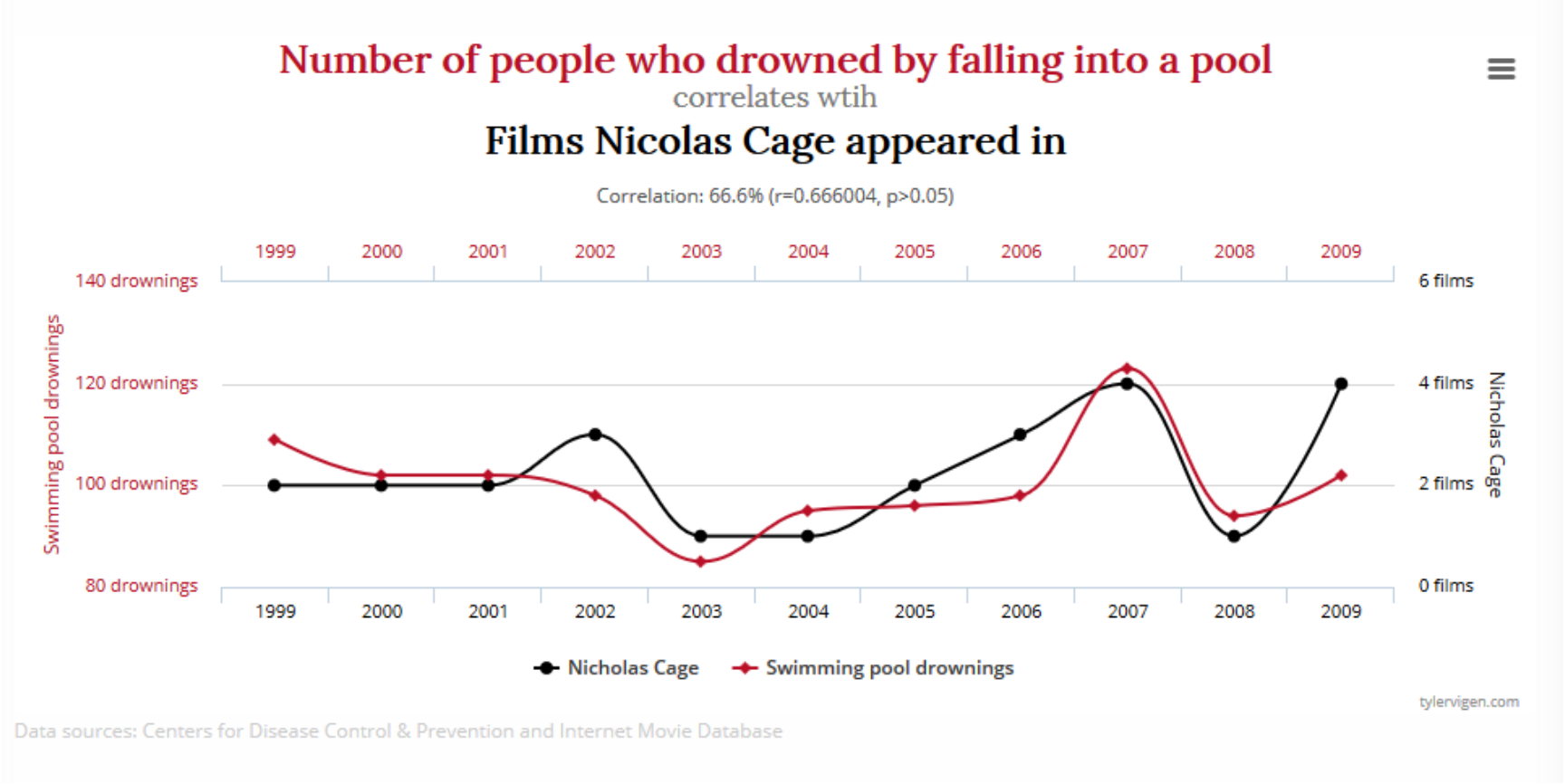
Polls whiz kid Nate Silver and presidential candidate Donald Trump.

Photo illustration by *Slate*. Images by Slaven Vlasic/Getty Images and Ethan Miller/Getty Images.

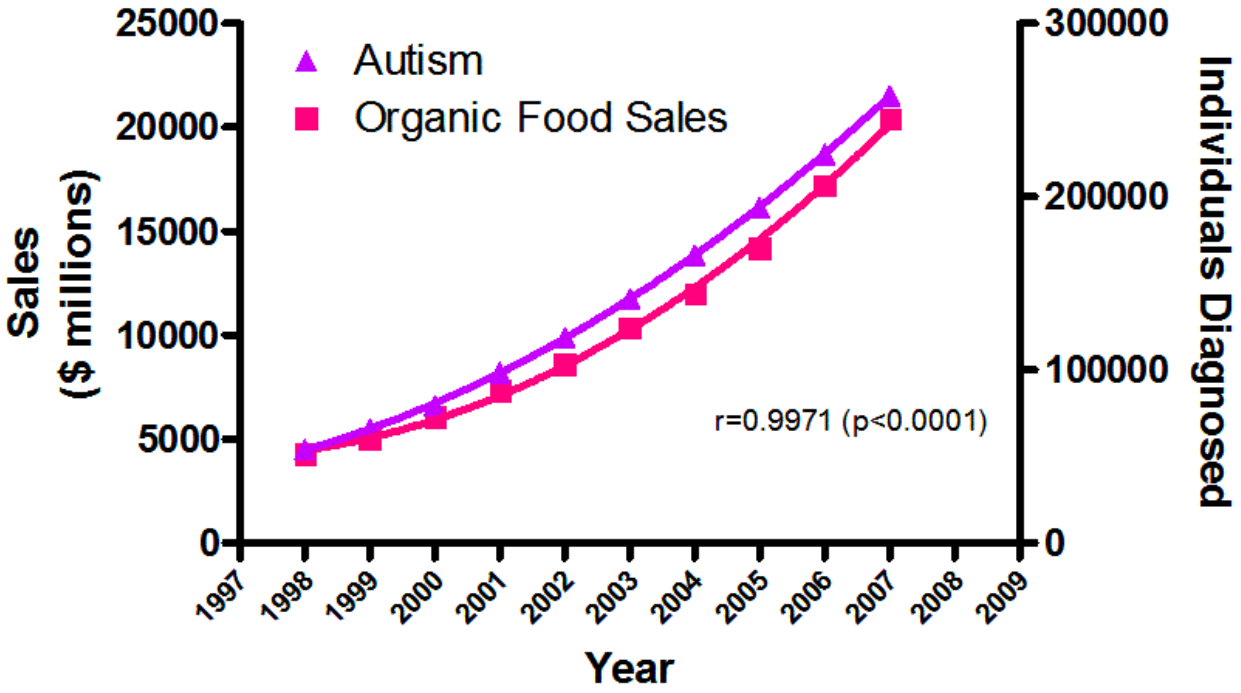
"If Silver's system depends largely on interpreting poll numbers, how reliable can that system be if the pre–Iowa and New Hampshire polls are basically worthless? **Garbage in, garbage out.**"

http://www.slate.com/articles/news_and_politics/politics/2016/01/nate_silver_said_donald_trump_had_no_shot_where_did_he_go_wrong.2.html


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
www.dilbert.com scottadams@aol.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5-7-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.


Number of people who drowned by falling into a pool
correlates wtih
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004, p>0.05)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

The real cause of increasing autism prevalence?



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

WHAT IS DATA SCIENCE?

Fortune

- "Hot New Gig in Tech"

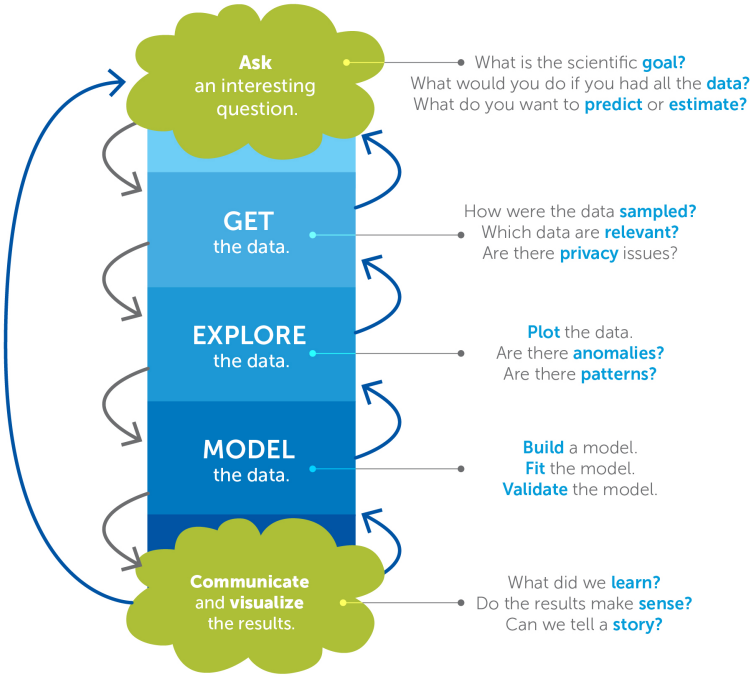
Hal Varian, Google's Chief Economist, NYT, 2009:

- "The next sexy job"
- "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—that's going to be a hugely important skill."

Mike Driscoll, CEO of metamarkets:

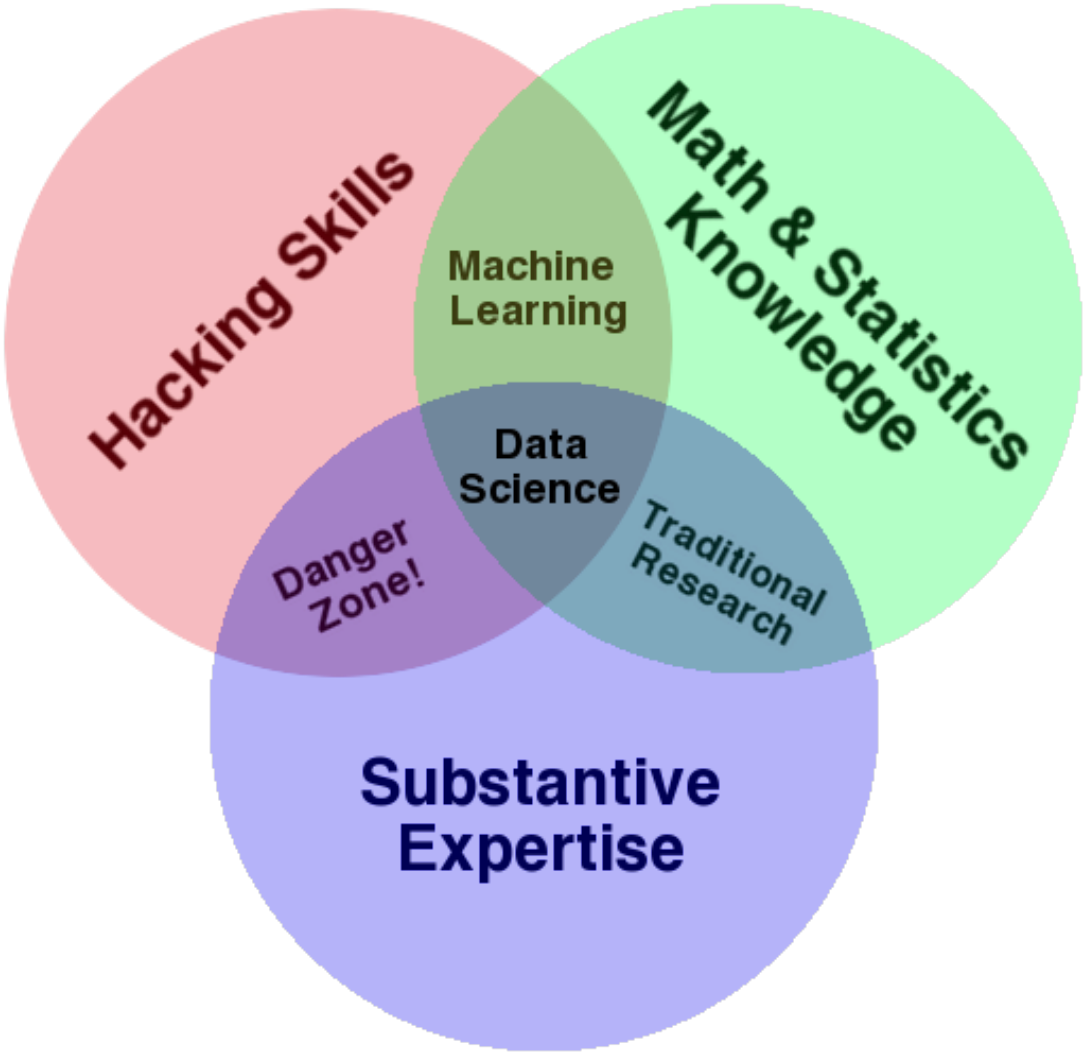
- "Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics."
- "Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

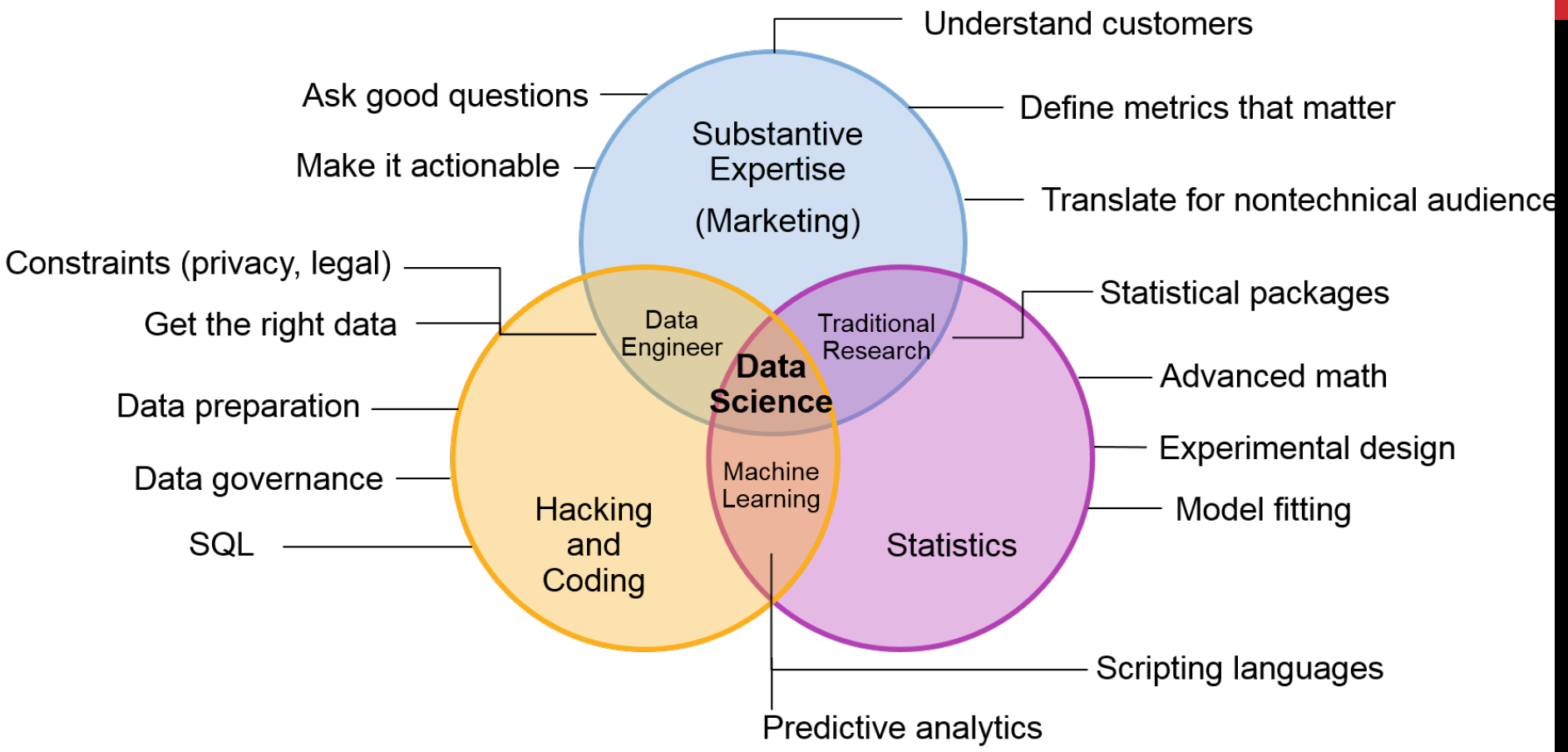
The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

DATA SCIENCE VENN DIAGRAM




Understand customers
Ask good questions
Define metrics that matter
Substantive Expertise (Marketing)
Make it actionable
Translate for nontechnical audience
Constraints (privacy, legal)
Statistical packages
Get the right data
Data Engineer
Traditional Research
Data Science
Advanced math
Data preparation
Experimental design
Machine Learning
Data governance
Model fitting
Hacking and Coding
Statistics
SQL
Scripting languages
Predictive analytics

WHAT DO DATA SCIENTISTS DO?

"They need to find nuggets of truth in data and then explain it to the business leaders"

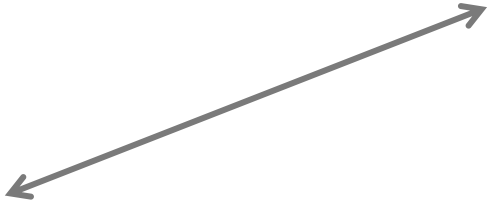
-- Richard Snee, EMC

Data scientists "tend to be "hard scientists", particularly physicists, rather than computer science majors. Physicists have a strong mathematical background, computing skills, and come from a discipline in which survival depends on getting the most from the data. They have to think about the big picture, the big problem."

-- DJ Patil, Chief Scientist at LinkedIn

MIKE DRISCOLL'S THREE SEXY SKILLS OF DATA GEEKS

"data wrangling"
"data jujitsu"
"data munging"



Data Wrangling

- parsing, scraping, and formatting data

Statistics

- traditional analysis

Visualization

- graphs, tools, etc.

DOING DATA SCIENCE (PETER HUBER)

1. **Inspection**
2. **Error checking**
3. **Modification**
4. **Comparison**
5. **Modeling and model fitting**
6. **Simulation**
7. **What-if analyses**
8. **Interpretation**
9. **Presentation of conclusions**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

DOING DATA SCIENCE (BEN FRY)

1. **Acquire**
2. **Parse**
3. **Filter**
4. **Mine**
5. **Represent**
6. **Refine**
7. **Interact**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

DOING DATA SCIENCE (COLIN MALLOWS)

1. **Identify data to collect and its relevance to your problem**
2. **Statistical specification of the problem**
3. **Method selection**
4. **Analysis of method**
5. **Interpret results for non-statisticians**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

A PRACTICAL DEFINITION

Data Science is about the whole processing pipeline to extract information out of data

Data Scientist understand and care about the whole data pipeline

A data pipeline consists of 3 steps:

1) **Preparing to run a model**

   Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

2) **Running the model**

3) **Communicating the results**

DATA SCIENCE IS ABOUT *DATA PRODUCTS*

- "Data-driven apps"
  - Spellchecker
  - Machine Translator
- Interactive visualizations
  - Google flu application
  - Global Burden of Disease
- Online Databases
  - Enterprise data warehouse
  - Sloan Digital Sky Survey

Data science is about building data products, not just answering questions

Data products empower others to use the data.

May help communicate your results (e.g., Nate Silver's maps)

May empower others to do their own analysis (e.g., Global Burden of Disease)

(Mike Loukides)

DISTINGUISHING DATA SCIENCE FROM…

Business Intelligence

Statistics

Data(base) Management

Visualization

Machine Learning

HUGE NUMBER OF RELEVANT AREAS

Stochastic/Statistics

Machine Learning

Databases

Distributed Systems

Networking

Cloud Computing

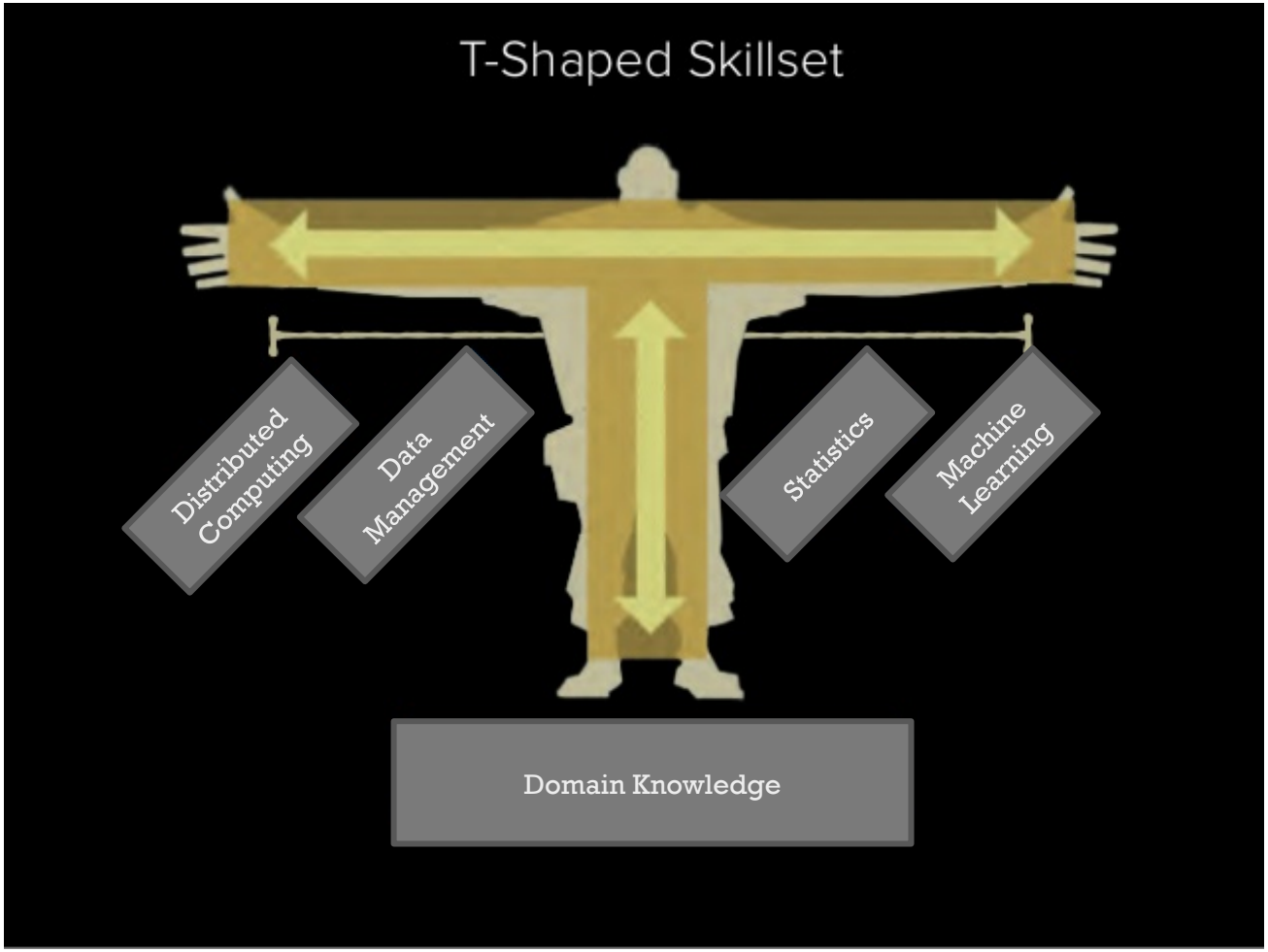
Natural Language Processing

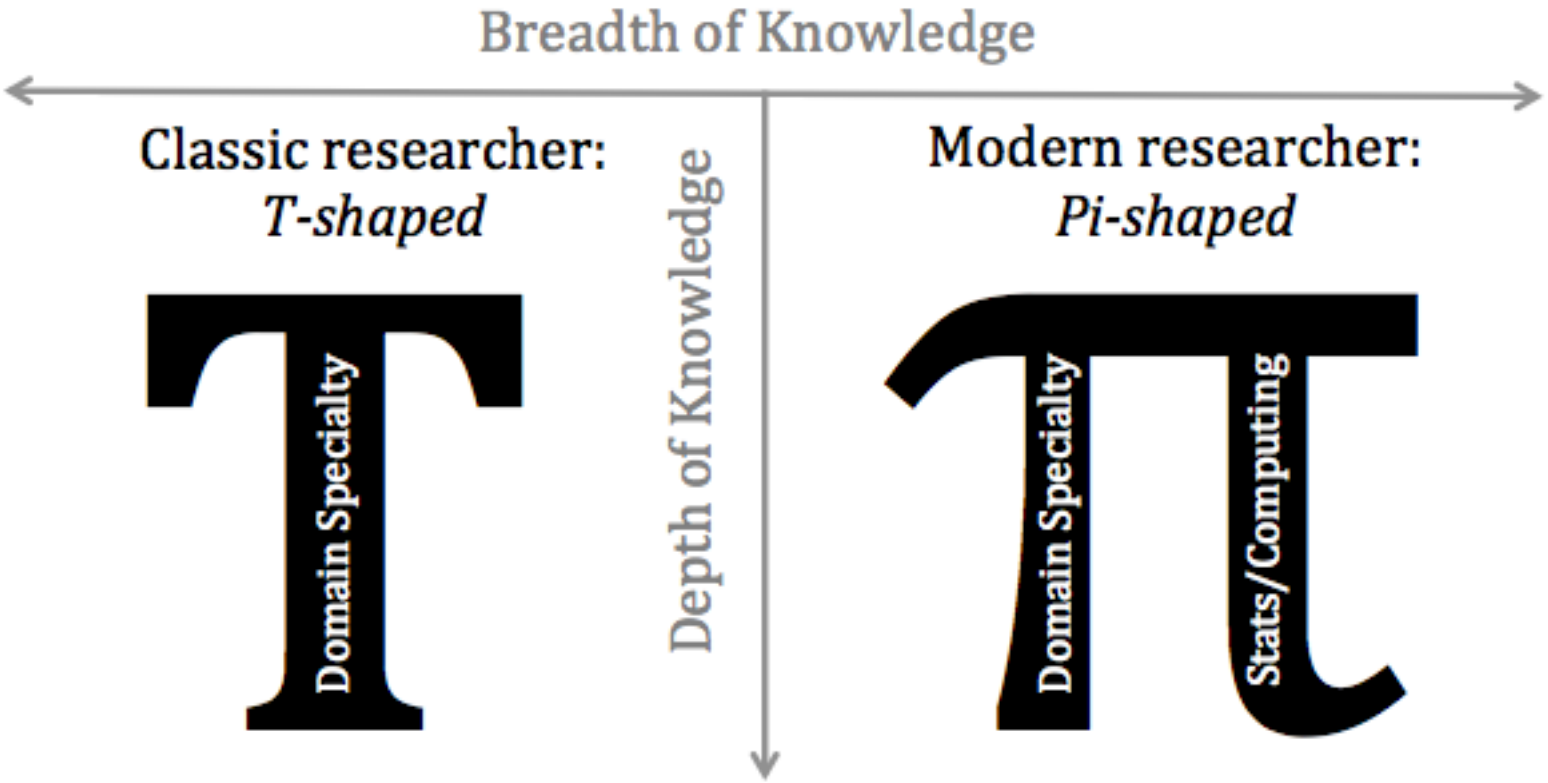
Visualization

…

"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

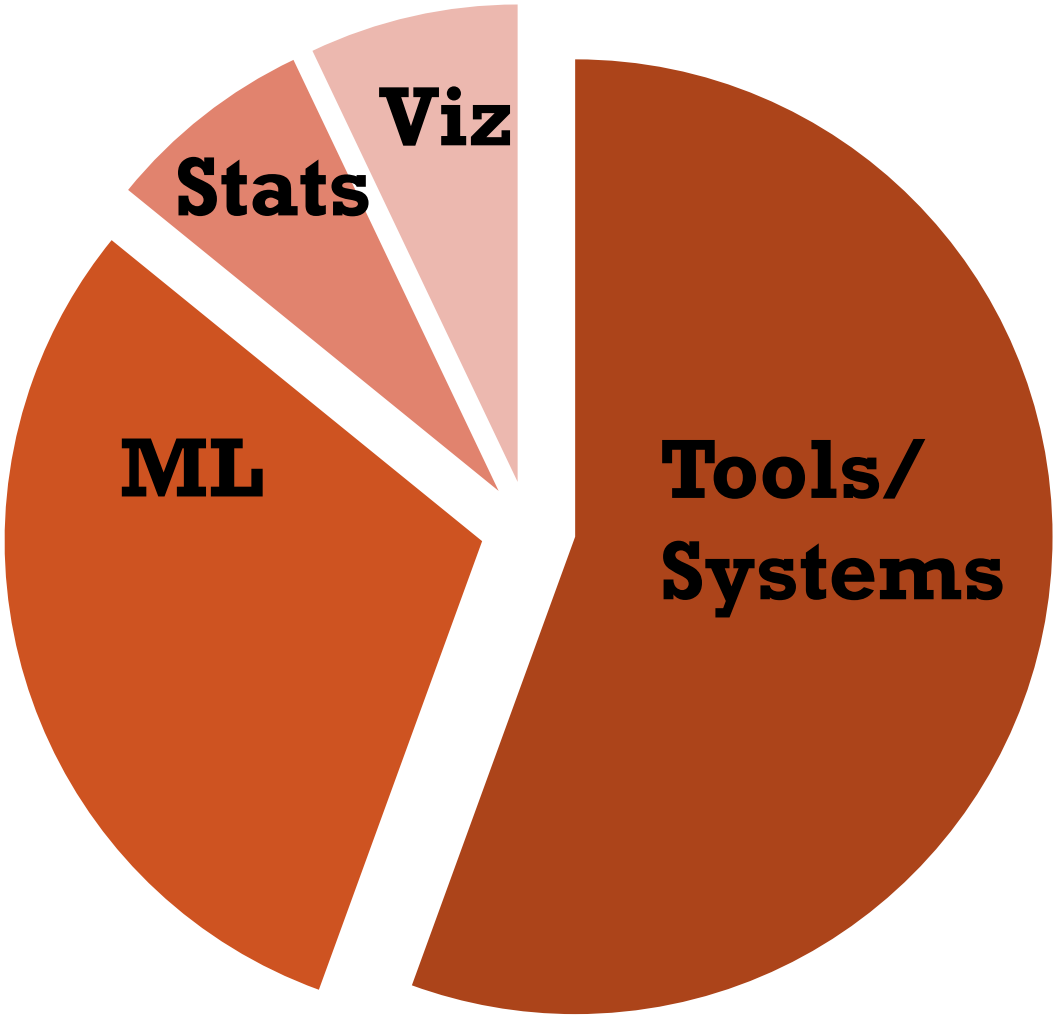
-- Aaron Kimball, CTO Wibidata


T-Shaped Skillset
Distributed Computing
Data Management
Statistics
Machine Learning
Domain Knowledge


Breadth of Knowledge
Classic researcher:
T-shaped
Domain Specialty
Depth of Knowledge
Modern researcher:
Pi-shaped
Domain Specialty
Stats/Computing

LOGISTICS

THIS COURSE



COURSE GOALS

- **You will learn the most important data science techniques**
  - Concepts and Abstractions
    - Relational model, classification, data cleaning, visualization, parallel processing, neural networks, data mining, …
  - Modern tools and system
    - Pandas, Jupiter, MapReduce, Tensorflow, Relational DBMSs, Hadoop, Spark, MongoDB, Google Data Cloud, …
- **You will be comfortable with producing**
  - Complete data pipelines and products
- **You will get**
  - significant breadth, and
  - some depth

ORGANIZATION

Lectures:

Tuesday and Thursday

8:30-10:20am in Lubrano (CIT 477).

Labs:

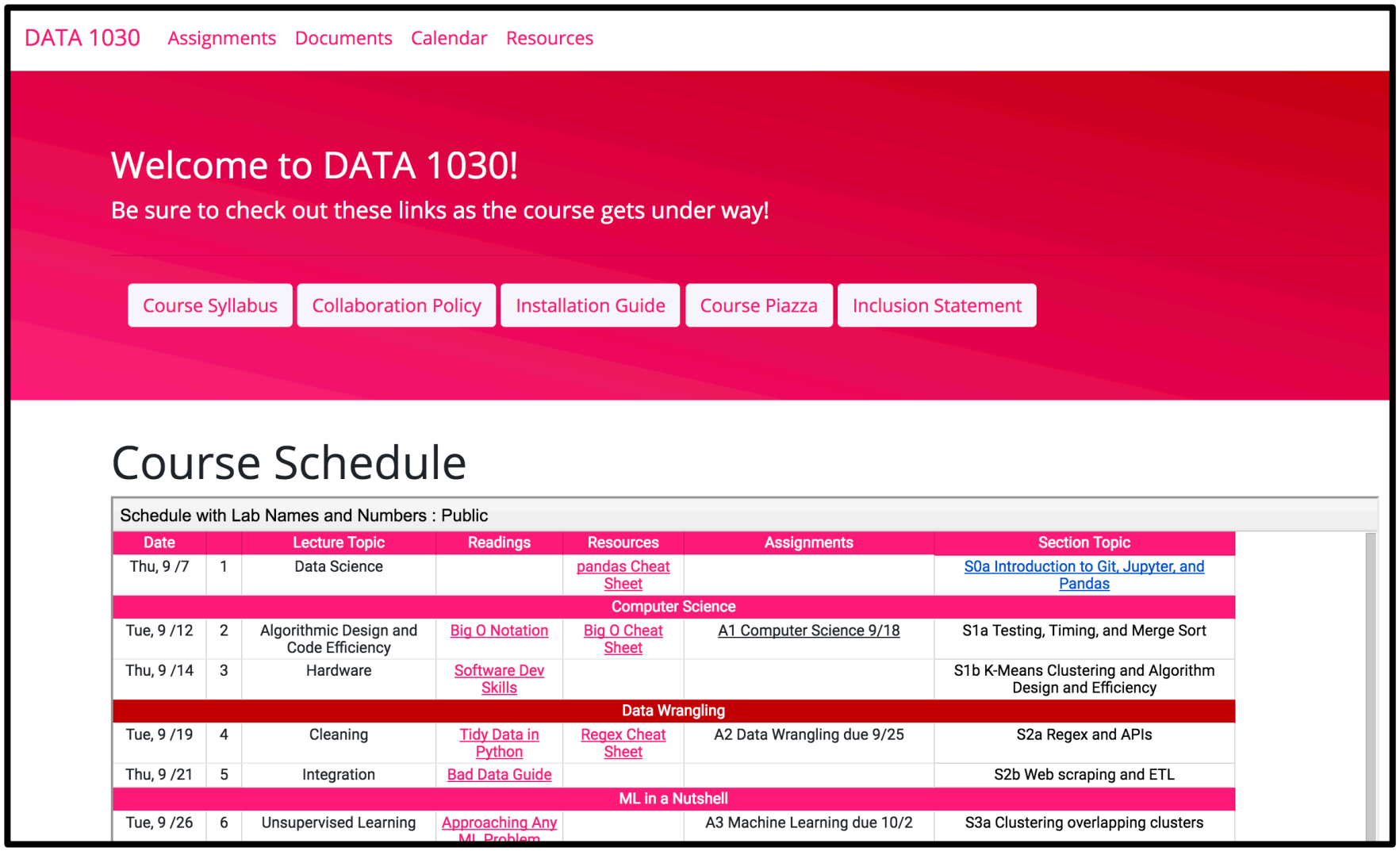
Tuesday and Thursday

12 pm - 2 pm in SWIG (CIT 241).

Starting Today!

COURSE WEB-PAGE

https://data1030.github.io/



DATA 1030 Assignments Documents Calendar Resources

Welcome to DATA 1030!

Be sure to check out these links as the course gets under way!

Course Syllabus | Collaboration Policy | Installation Guide | Course Piazza | Inclusion Statement

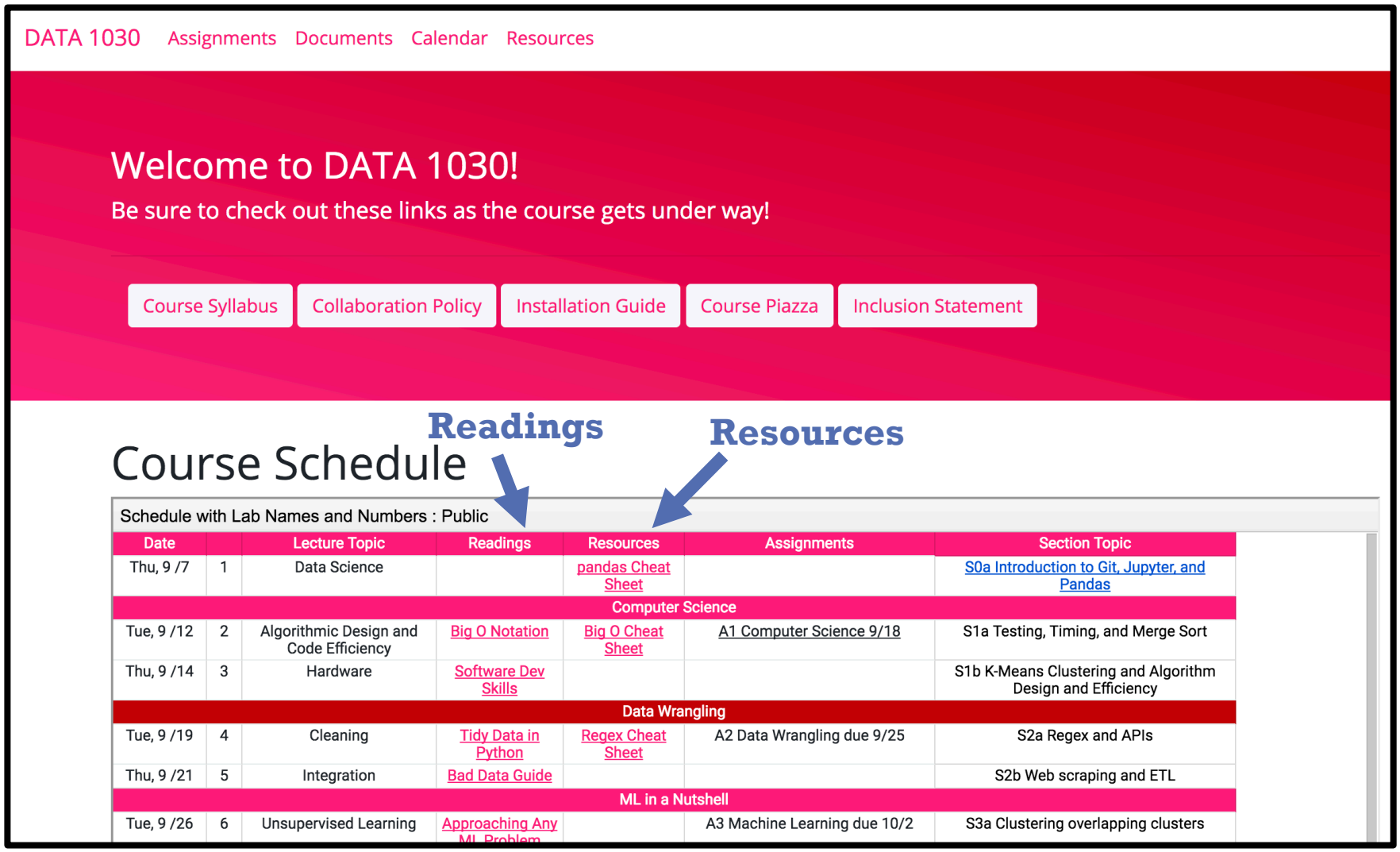
Course Schedule

Schedule with Lab Names and Numbers : Public

Date		Lecture Topic	Readings	Resources	Assignments	Section Topic
Thu, 9 /7	1	Data Science		pandas Cheat Sheet		S0a Introduction to Git, Jupyter, and Pandas
Computer Science						
Tue, 9 /12	2	Algorithmic Design and Code Efficiency	Big O Notation	Big O Cheat Sheet	A1 Computer Science 9/18	S1a Testing, Timing, and Merge Sort
Thu, 9 /14	3	Hardware	Software Dev Skills			S1b K-Means Clustering and Algorithm Design and Efficiency
Data Wrangling						
Tue, 9 /19	4	Cleaning	Tidy Data in Python	Regex Cheat Sheet	A2 Data Wrangling due 9/25	S2a Regex and APIs
Thu, 9 /21	5	Integration	Bad Data Guide			S2b Web scraping and ETL
ML in a Nutshell						
Tue, 9 /26	6	Unsupervised Learning	Approaching Any ML Problem		A3 Machine Learning due 10/2	S3a Clustering overlapping clusters

COURSE WEB-PAGE

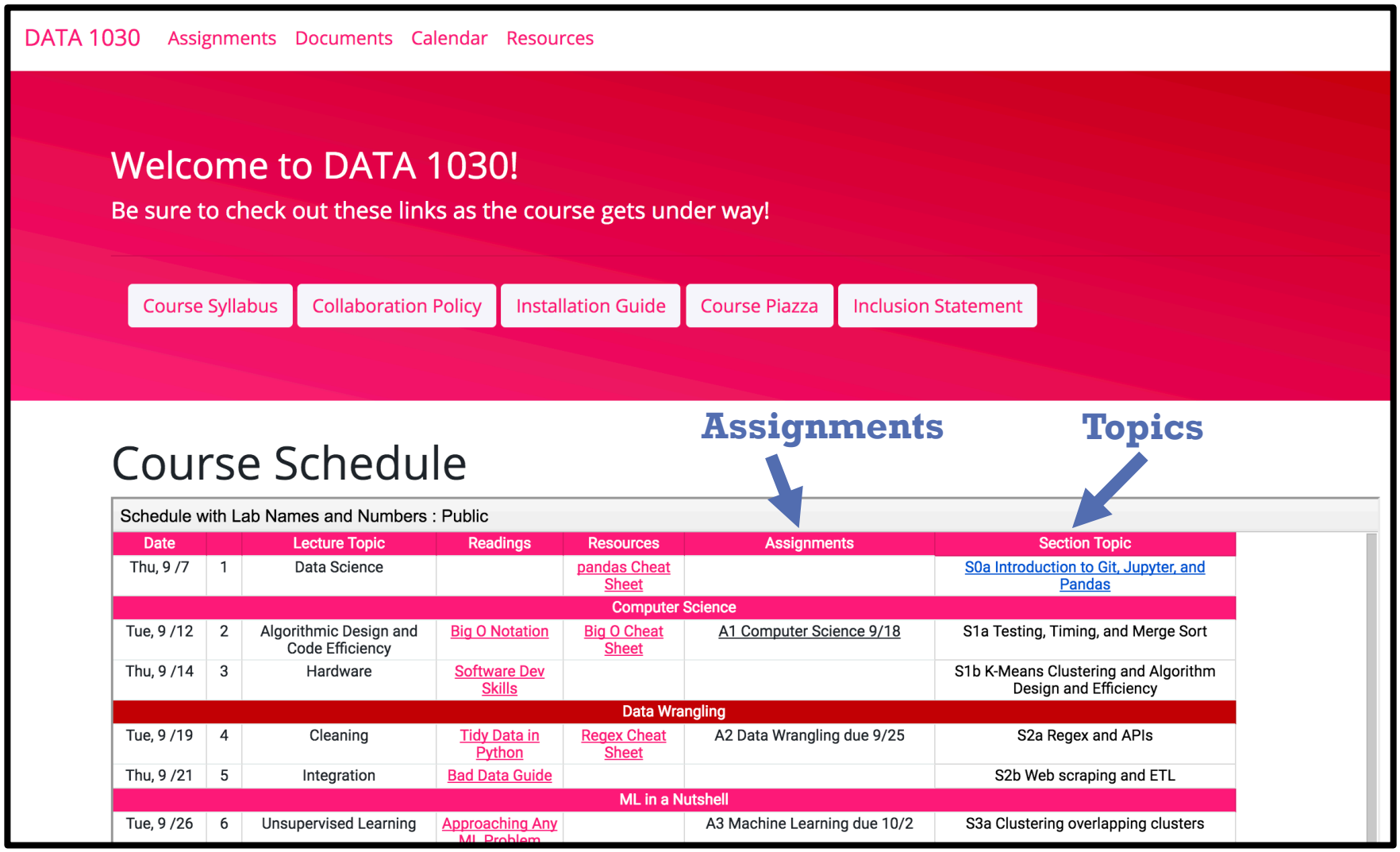
https://data1030.github.io/



Date		Lecture Topic	Readings	Resources	Assignments	Section Topic
Thu, 9 /7	1	Data Science		pandas Cheat Sheet		S0a Introduction to Git, Jupyter, and Pandas
Computer Science						
Tue, 9 /12	2	Algorithmic Design and Code Efficiency	Big O Notation	Big O Cheat Sheet	A1 Computer Science 9/18	S1a Testing, Timing, and Merge Sort
Thu, 9 /14	3	Hardware	Software Dev Skills			S1b K-Means Clustering and Algorithm Design and Efficiency
Data Wrangling						
Tue, 9 /19	4	Cleaning	Tidy Data in Python	Regex Cheat Sheet	A2 Data Wrangling due 9/25	S2a Regex and APIs
Thu, 9 /21	5	Integration	Bad Data Guide			S2b Web scraping and ETL
ML in a Nutshell						
Tue, 9 /26	6	Unsupervised Learning	Approaching Any ML Problem		A3 Machine Learning due 10/2	S3a Clustering overlapping clusters

COURSE WEB-PAGE

https://data1030.github.io/



Date		Lecture Topic	Readings	Resources	Assignments	Section Topic
Thu, 9 /7	1	Data Science		pandas Cheat Sheet		S0a Introduction to Git, Jupyter, and Pandas
Computer Science						
Tue, 9 /12	2	Algorithmic Design and Code Efficiency	Big O Notation	Big O Cheat Sheet	A1 Computer Science 9/18	S1a Testing, Timing, and Merge Sort
Thu, 9 /14	3	Hardware	Software Dev Skills			S1b K-Means Clustering and Algorithm Design and Efficiency
Data Wrangling						
Tue, 9 /19	4	Cleaning	Tidy Data in Python	Regex Cheat Sheet	A2 Data Wrangling due 9/25	S2a Regex and APIs
Thu, 9 /21	5	Integration	Bad Data Guide			S2b Web scraping and ETL
ML in a Nutshell						
Tue, 9 /26	6	Unsupervised Learning	Approaching Any ML Problem		A3 Machine Learning due 10/2	S3a Clustering overlapping clusters

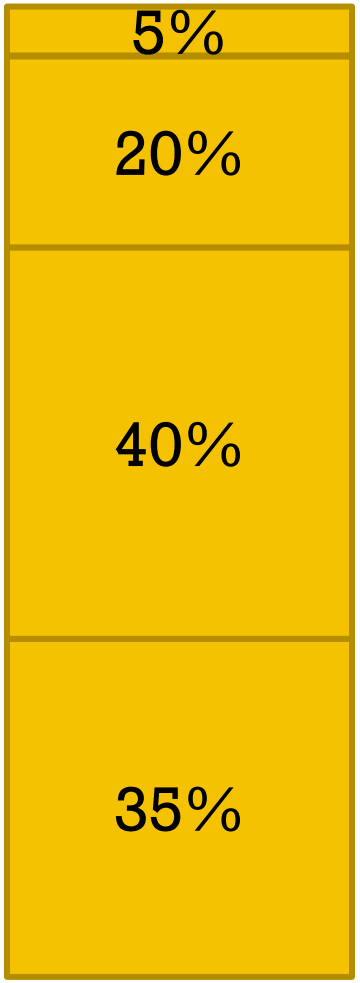
TOPICS

Date		Lecture Topic	Readings	Resources	Assignments	Section Topic
Thu, 9 /7	1	Data Science		pandas Cheat Sheet		S0a Introduction to Git, Jupyter, and Pandas
Computer Science						
Tue, 9 /12	2	Algorithmic Design and Code Efficiency	Big O Notation	Big O Cheat Sheet	A1 Computer Science 9/18	S1a Testing, Timing, and Merge Sort
Thu, 9 /14	3	Hardware	Software Dev Skills			S1b K-Means Clustering and Algorithm Design and Efficiency
Data Wrangling						
Tue, 9 /19	4	Cleaning	Tidy Data in Python	Regex Cheat Sheet	A2 Data Wrangling due 9/25	S2a Regex and APIs
Thu, 9 /21	5	Integration	Bad Data Guide			S2b Web scraping and ETL
ML in a Nutshell						
Tue, 9 /26	6	Unsupervised Learning	Approaching Any ML Problem		A3 Machine Learning due 10/2	S3a Clustering overlapping clusters
Thu, 9 /28	7	Supervised Learning	Supervised Learning	sklearn Roadmap		S3b Learning from disaster
Tue, 10 /3	8	Visualization				S3c Effective Visualization
Data Management						
Thu, 10 /5	9	Modeling			A4 SQL Databases due 10/16	S4a Modeling Data
Tue, 10 /10	10	SQL		SQL Cheat Sheet		S4b Loading Data
Thu, 10 /12	11	Query Optimization				S4c Querying Data

TOPICS

Big Data						
Tue, 10 /17	12	Abstractions and Systems	System Design Primer	Scaling Applications on Azure	A5 Big Data due 10/30	S5a
Thu, 10 /19	13	Scalability	Dynamo	Data Engineering Ecosystem		S5b
Tue, 10 /24	14	Cloud Platforms				S5c
Thu, 10 /26		In class Midterm				
Data Mining						
Tue, 10 /31	15	Decision Trees	Decision Trees Tutorial		A6 Decision Trees due 11/6	S6a
Thu, 11 /2	16	Ensemble Methods	Tree-Based Models			S6b
Tue, 11 /7	17	Recommender Systems			Project Proposal due 11/13	S6c
Thu, 11 /9	18	Performance Tuning				S6d Project Idea Feedback, Brain Storming
Neural Networks						
Tue, 11 /14	19	Introduction to NNs	Neural Network Playground		A7 Neural Networks due 11/22	S7a
Thu, 11 /16	20	Convolutional NNs		Neural Network Zoo		S7b
Tue, 11 /21	21	Recurrent NNs				S7c
Thu, 11 /23		Thanksgiving (No Class)				
Selected Topics						
Tue, 11 /28	22	Final exam				
Thu, 11 /30	23	Data Privacy				

DELIVERABLES / GRADING GUIDELINE



Participation

Exams

Assignments & Labs

Project

COLLABORATION POLICY

Collaboration in this course is governed by a liberal collaboration policy, found on the resources section of the class website.

You **must** sign the collaboration policy before any of your work can be graded. Failure to meet the standards set by the collaboration policy can result in failure and possible disciplinary action.

THE "LITTLE" EXAMS (15%)

2 of them:

10/26/2017

(data management, text-processing, visualization, basic ml)

11/28/2017

(more advanced machine-learning, deep learning)

LABS AND ASSIGNMENTS

- **Every class is followed by a lab at noon**
- **Labs and assignments are the same thing**
- **Every topic has an assignment**
- **Every assignment is graded**

LATE DAY POLICY

3 late days on any assignments, excluding the final project and labs.

Once these three late days have been exhausted, **you will receive NO credit for your late hand-in.**

For medical extensions, please directly **contact Dan Potter**. Acceptable excuses include, but are not limited to:

- 1. Illness (with a doctor's note)
- 2. Family emergencies
- 3. Religious holidays

Regardless of circumstance, please contact the HTAs at least 48 hours in advance of the due date to arrange for an extension. HTAs reserve the right to reject extensions introduced with less than 48 hours before the due date. Extenuating circumstances will be evaluated on a case-by-case basis.

piazza

Sign-up page: https://piazza.com/brown/

(if you are not already signed up)

All questions regarding assignments, projects, etc. should go through piazza (no email please)

Always, always search piazza first. Somebody might have already asked the same question.

If not, ask the question first on piazza publicly unless it concerns your project or reveals too much of a solution.

You can send private messages

TA HOURS

On the course calendar

Policy: Try to use piazza first (even before going to TA hours), unless it is specific to your project

INSTRUCTOR HOURS

Please visit us for questions about our respective lectures. See Dan for special situations/circumstances. Questions about assignments labs and your project should **always first go to piazza and afterwards to the TAs** during the office hours. Only if you are still not happy with the answer, come to us.

PROJECT

Groups of 2-3 students (match-making on piazza)

Pre-Project (2/8 – 3/21) (10%)

- Goal: Get you started to explore a data set
- Mid-term

Final-Project (4/2 – 5/12) (30%)

- Can but does not have to build on the pre-project
- Goal: Demonstrate that you master the data pipeline from cleaning, model building, to presenting the result by taking a data set and deriving some *interesting* insight.
- You will write three blog posts (e.g., on wordpress) about your progress, interesting insights, tools you learned, etc. The blog posts will be also graded

Topic: Up to you

i>clicker®

- You can get one from the Friedman Center in the Sciences Library.
- Don't forgot to register your clicker in **Canvas** (do not use the iclicker web-page)
- Clicker participation counts 4% towards the class, but answers are not graded for correctness
- If you are asked to pay a fee, you did something wrong!!!

ANATOMY OF A PROJECT

DOLLA DOLLA BILLS Y'ALL

CSCI1951A: Data Science Spring 2016 Final Project

Adam Hoff, Angelia Wang, Chris Grimm, Athyuttam Eleti

Explored potential arbitrage between Amazon and eBay

ANATOMY OF A PROJECT

Explored a hypothesis: Amazon and eBay aren't optimal markets and thus have arbitrage. Can we identify and predict arbitrage possibilities and item prices?

Used services like CamelCamelCamel to get Amazon and eBay time series price data

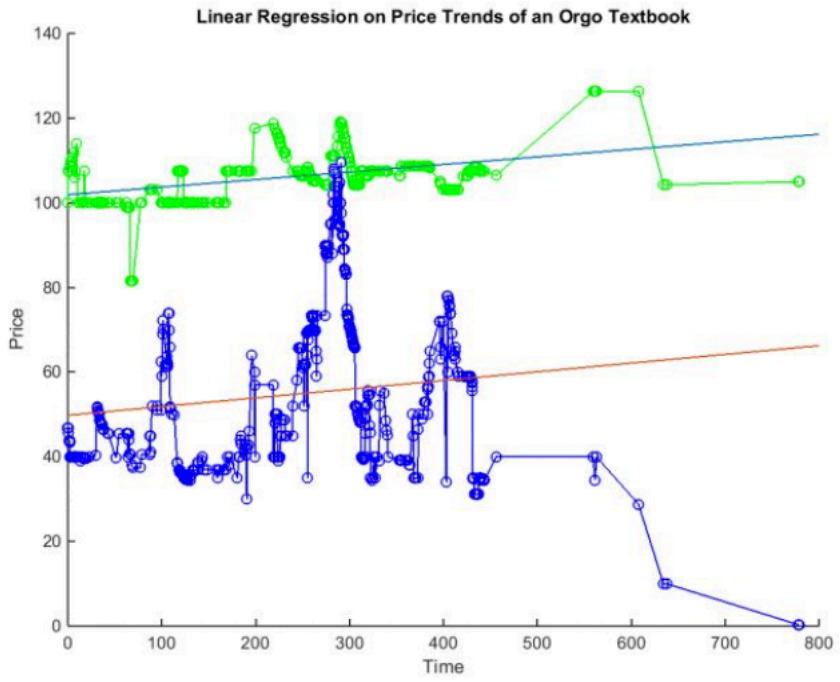
Examined both Auctions and Buy It Now on eBay

Had to build integration system that paired Amazon and eBay items

amazon.com ebay

ANATOMY OF A PROJECT

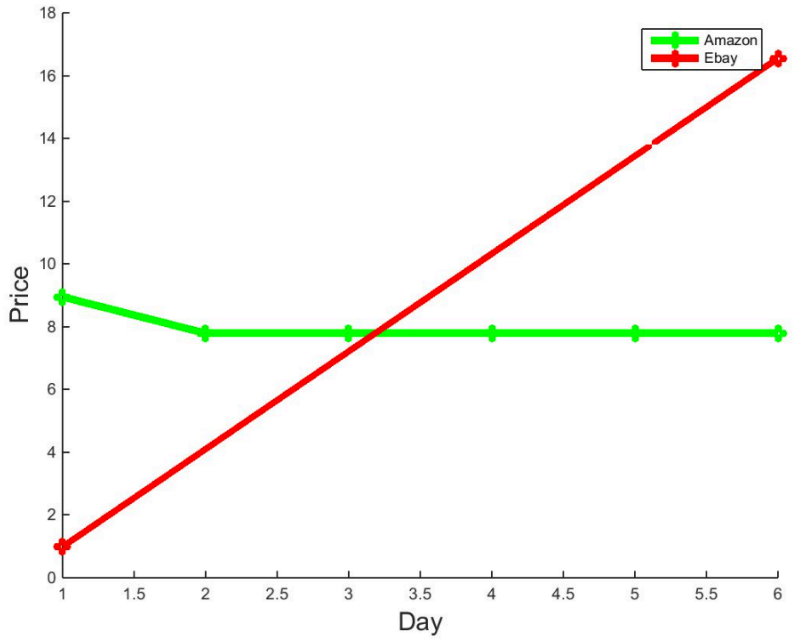
Used a linear regression to predict prices and model price differences



ANATOMY OF A PROJECT

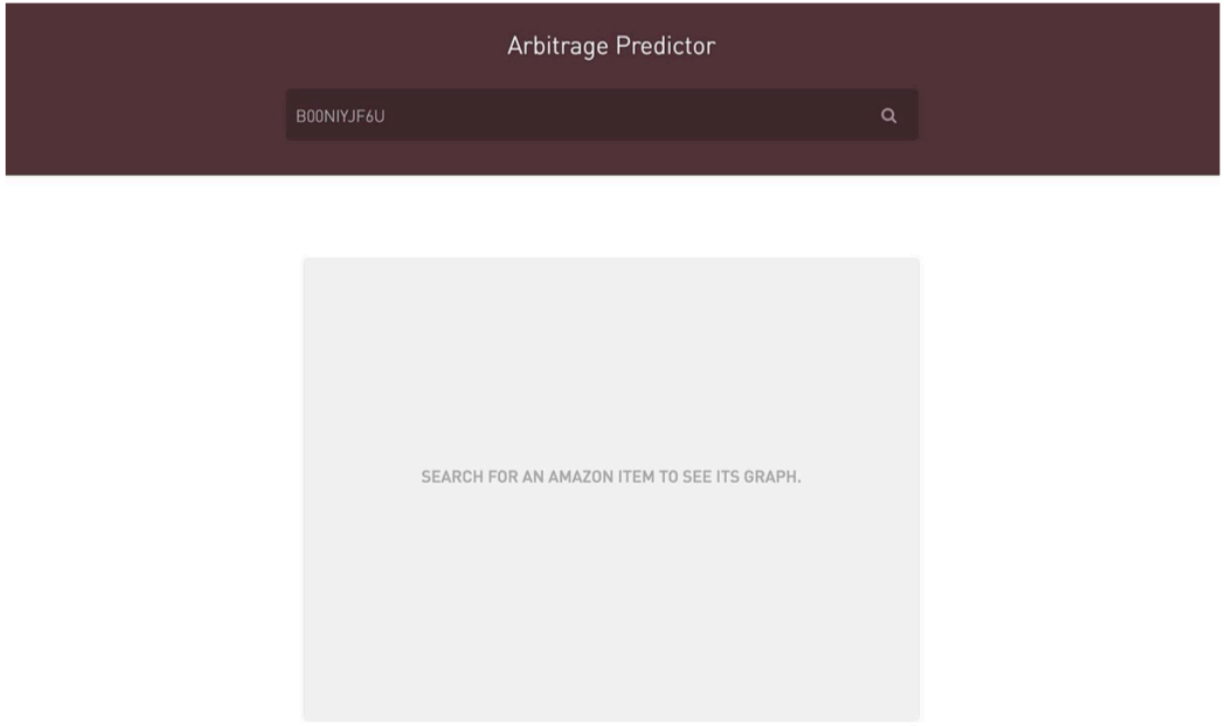
Built an arbitrage predictor with staggering 90% accuracy

Could predict future opportunities from past data



ANATOMY OF A PROJECT

As part of the capstone, built a system to enter Amazon items and predict future prices



EXAMPLE FINAL PROJECT

BREWREPORT

Beer Name | Search

Harpoon IPA

Harpoon Brewery
American IPA
82 on BeerAdvocate
5.90% abv.

Extremely Floral | Fairly Hoppy | Fairly Easy | Fairly Citrus | Fairly Carbonated | Slightly Bitter | Somewhat Pine | Somewhat Earthy | Somewhat Malty | Somewhat Grapefruit | Somewhat Orange

TASTES LIKE

Sierra Nevada Pale Ale Sierra Nevada Brewing Co.
Phin & Matt's Extraordinary Ale Southern Tier Brewing Company
Redhook Long Hammer IPA Redhook Ale Brewery

QUOTES

"O- This was the first IPA I had (besides DFH's 60 and 90) where I realized "Hey, I like IPAs"

-Immortale25

"It's more balanced than many IPAs but still has a pleasing citric hops flavor"

-wahhmaster

"as far as East Coast American IPAs go, I really like this one"

-irishevans

YOU MIGHT LIKE

Sweetwater 420 Extra Pale Ale
SweetWater Brewing Company

Lagunitas PILS (Czech Style Pilsner)
Lagunitas Brewing Company

Fat Tire Amber Ale
New Belgium Brewing

PRE-PROJECT

- **Goal: get ready to explore a dataset**
- **Begins with a pre-proposal, due mid February**
- **You'll be paired with a mentor TA**
- **Midterm report will show your progress so far**
  - Should have access to data and be partially cleaned
- **Get started early! (Can't stress this enough)**

FINAL PROJECT

- **Built on top of your pre-project**
- **Goal: demonstrate that you master the data pipeline from cleaning, model building, to presenting the result by taking a data set and deriving some *interesting* insight**
- **Weekly updates in the form of blog posts about your progress, insights, tools, etc. (We grade these.)**

PROJECT

We will give you more links (e.g., how to find data, etc.) and details about the projects in the next few weeks.

NEXT STEPS

Get the software stack installed. Go to the lab.

Read/Watch:

- Alon Halevy, Peter Norvig, and Fernando Pereira: The Unreasonable Effectiveness of Data
- http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
- http://www.ted.com/talks/gary_flake_is_pivot_a_turning_point_for_web_exploration.html