Technology Fundamentals for Analytics

Jason Kuruzovich

Overview

- Introductions
- Let’s Get Excited about Data Science
- What do Data Scientists Do?
- Course Syllabus
- Lab 1. Development Environment

Next Startup Tech Valley Event is September 2 at 5:30 at Brown’s Revolution Hall

See more at www.startuptechvalley.org

Me

- Director of the Severino Center for Technological Entrepreneurship
- Associate Professor of Business Analytics
- Research on online markets and entrepreneurship

You

- What is your background?
- What are you looking for out of this class/program?
- What type of job do you want?
- What are your hobbies?

There have been profound changes in technology and the information processes define our society

Internet 0.1 beta (18th Century)

Internet 0.1 Beta (18th Century)

Chain of towers or optical telegraph capable of transmitting 1-3 symbols per min

Towers 5-20 KM apart

“We create as much information in two days now as we did from the dawn of man through 2003.”

-Eric Schmidt, Former CEO of Google

Information Economy

TRADITIONAL PRODUCTION PROCESS

INFORMATION BASED BUSINESS PROCESS

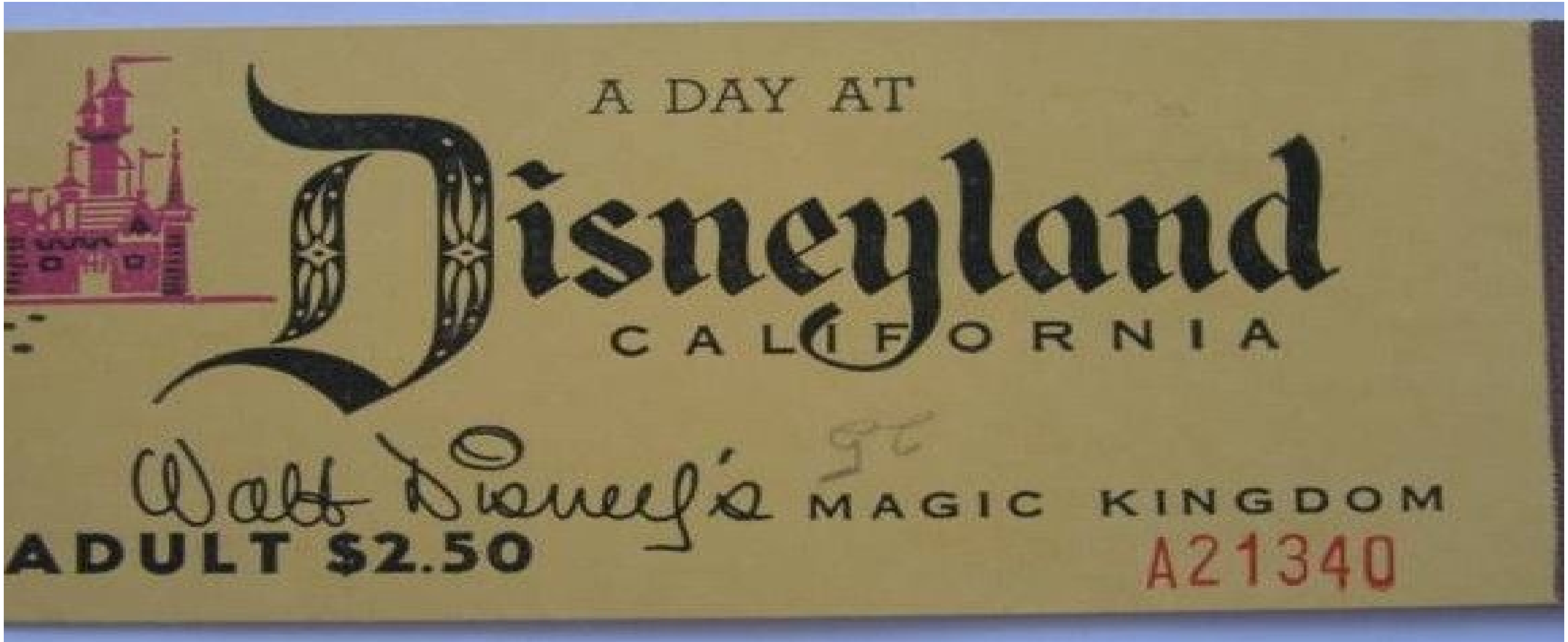
INFORMATION TECHNOLOGY

What is driving the growth of information and data?

Consider the evolution of a company like Disney…

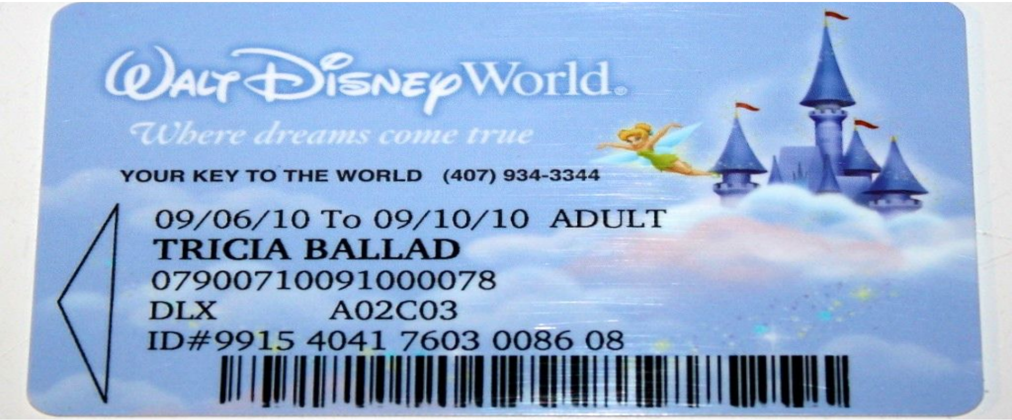
Disney

ROLE OF DATA: How many tickets did we sell?



Disney – Data Warehouse Stage

ROLE OF DATA: How much did our customers spend? How can we understand different customer types?



Walt Disney World.

Where dreams come true

YOUR KEY TO THE WORLD (407) 934-3344

09/06/10 To 09/10/10 ADULT

TRICIA BALLAD

07900710091000078

DLX A02C03

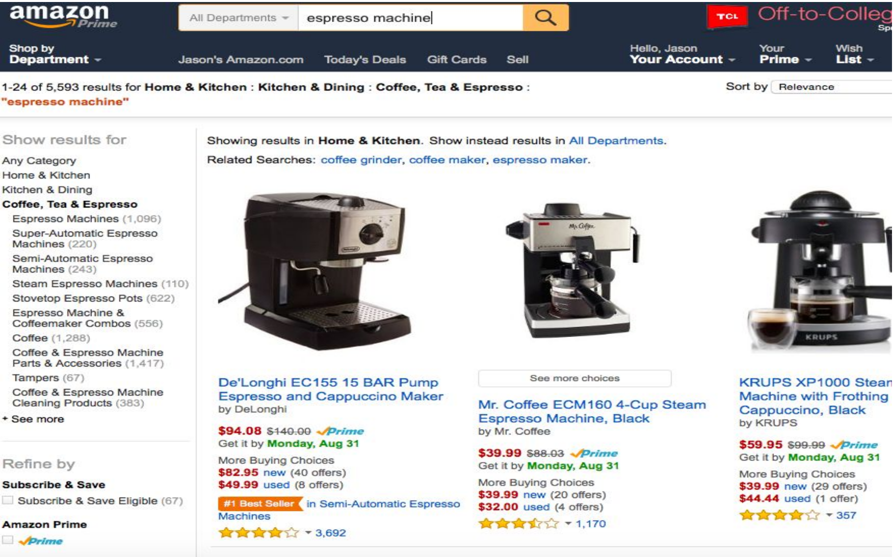
ID#9915 4041 7603 0086 08

Disney – Big Data

ROLE OF DATA: What path did customers take through the park, when did they leave? How long did they stand in line? When did they spend money on souvenirs and where? How often did they go to the bathroom and did they have to wait? How long did they spend at dinner in the Mexican pavilion compared with the German pavilion? How does the speed of entry correlate with tipping behavior?

Electronic Commerce

- TRADITIONAL SHOPPING
- TRAFFIC AND POS DATA



espresso machine

1-24 of 5,593 results for **Home & Kitchen** : **Kitchen & Dining** : **Coffee, Tea & Espresso** : "espresso machine"

Show results for

Any Category
Home & Kitchen
Kitchen & Dining
Coffee, Tea & Espresso
Espresso Machines (1,096)
Super-Automatic Espresso Machines (220)
Semi-Automatic Espresso Machines (243)
Steam Espresso Machines (110)
Stovetop Espresso Pots (622)
Espresso Machine & Coffeemaker Combos (556)
Coffee (1,288)
Coffee & Espresso Machine Parts & Accessories (1,417)
Tampers (67)
Coffee & Espresso Machine Cleaning Products (383)
+ See more

Refine by

Subscribe & Save
□ Subscribe & Save Eligible (67)

Amazon Prime
□ Prime

Showing results in **Home & Kitchen**. Show instead results in All Departments.

Related Searches: coffee grinder, coffee maker, espresso maker.

De'Longhi EC155 15 BAR Pump Espresso and Cappuccino Maker
by DeLonghi
$94.08 ~~$140.00~~ Prime
Get it by **Monday, Aug 31**
More Buying Choices
$82.95 new (40 offers)
$49.99 used (8 offers)
#1 Best Seller in Semi-Automatic Espresso Machines
★★★★☆ 3,692

Mr. Coffee ECM160 4-Cup Steam Espresso Machine, Black
by Mr. Coffee
$39.99 ~~$88.03~~ Prime
Get it by **Monday, Aug 31**
More Buying Choices
$39.99 new (20 offers)
$32.00 used (4 offers)
★★★☆☆ 1,170

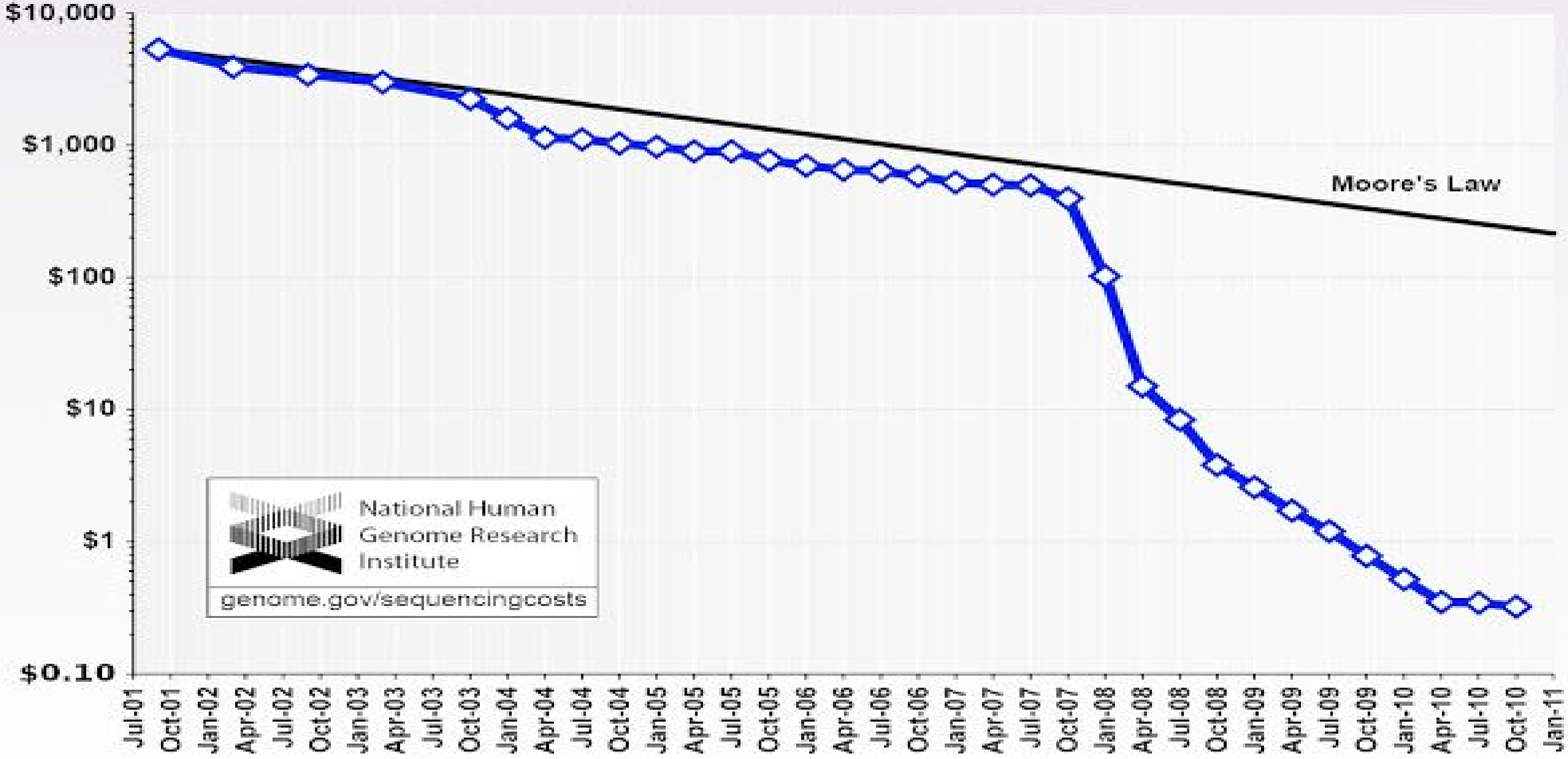
KRUPS XP1000 Stean Machine with Frothing Cappuccino, Black
by KRUPS
$59.95 ~~$99.99~~ Prime
Get it by **Monday, Aug 31**
More Buying Choices
$39.99 new (29 offers)
$44.44 used (1 offer)
★★★★☆ 357

- EVERY CLICK
- HISTORICAL PURCHASE
- REMARKETING

What if the online and offline merge?

https://youtu.be/uiDMlFycNrw

Cost per Megabase of DNA Sequence



genome.gov/sequencingcosts

Big Data and Astronomy

The Murchison Widefield Array is the first Square Kilometre Array precursor to enter full operations, generating a vast torrent of information that needs to be stored for later retrieval by researchers.

http://www.skatelescope.org/news/pawsey-centre/

“To store the Big Data the MWA produces, you’d need almost three 1 TB hard drives every two hours”

Social Media

The Facebook Social Network

Facebook

- 300 Petabyte Data warehouse
- 500+ Terabytes new data each day

http://www.adweek.com/socialtimes/orcfile/434041

Online Video

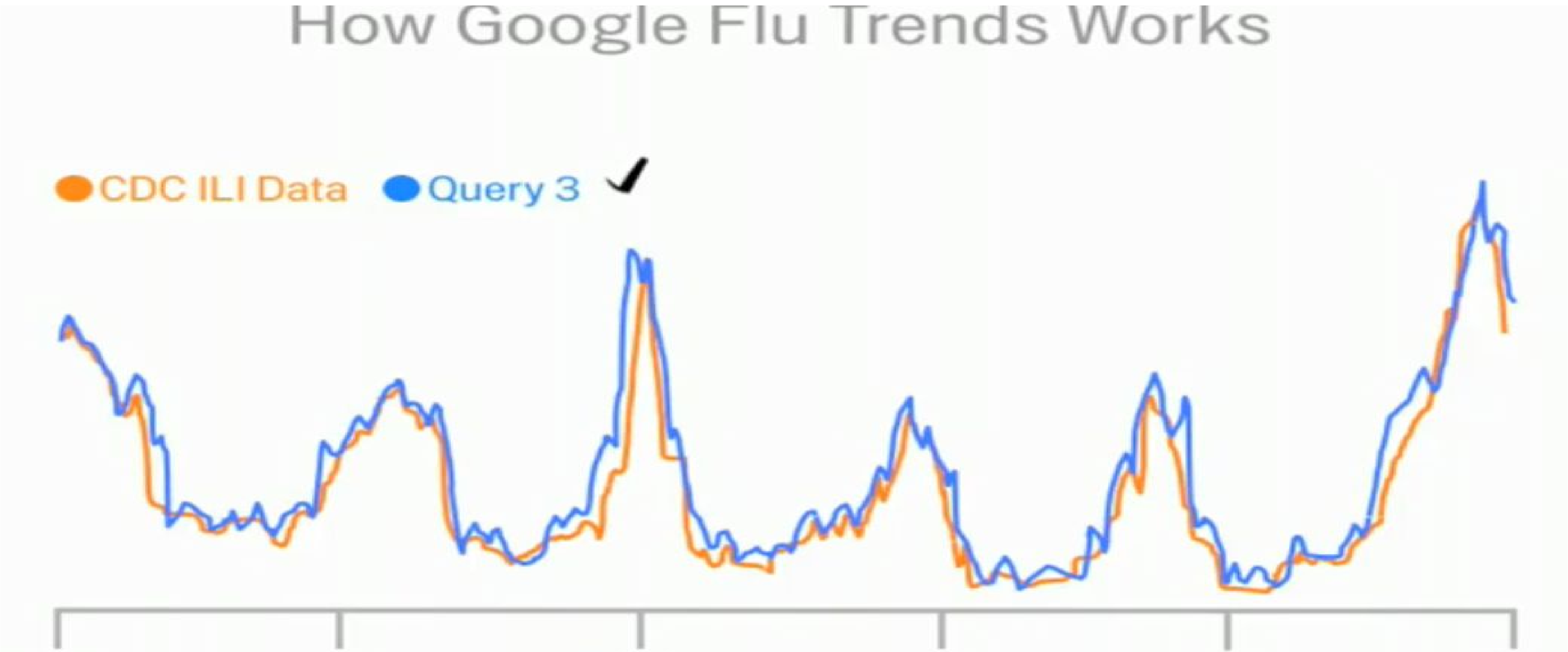
YouTube

- 60 hours of video per minute
- 4 billion views per day
- 800 million unique users

http://www.jeffbullas.com/2012/05/23/35-mind-numbing-youtube-facts-figures-and-statistics-infographic/#LhtL2qC4Mfj83EBD.99

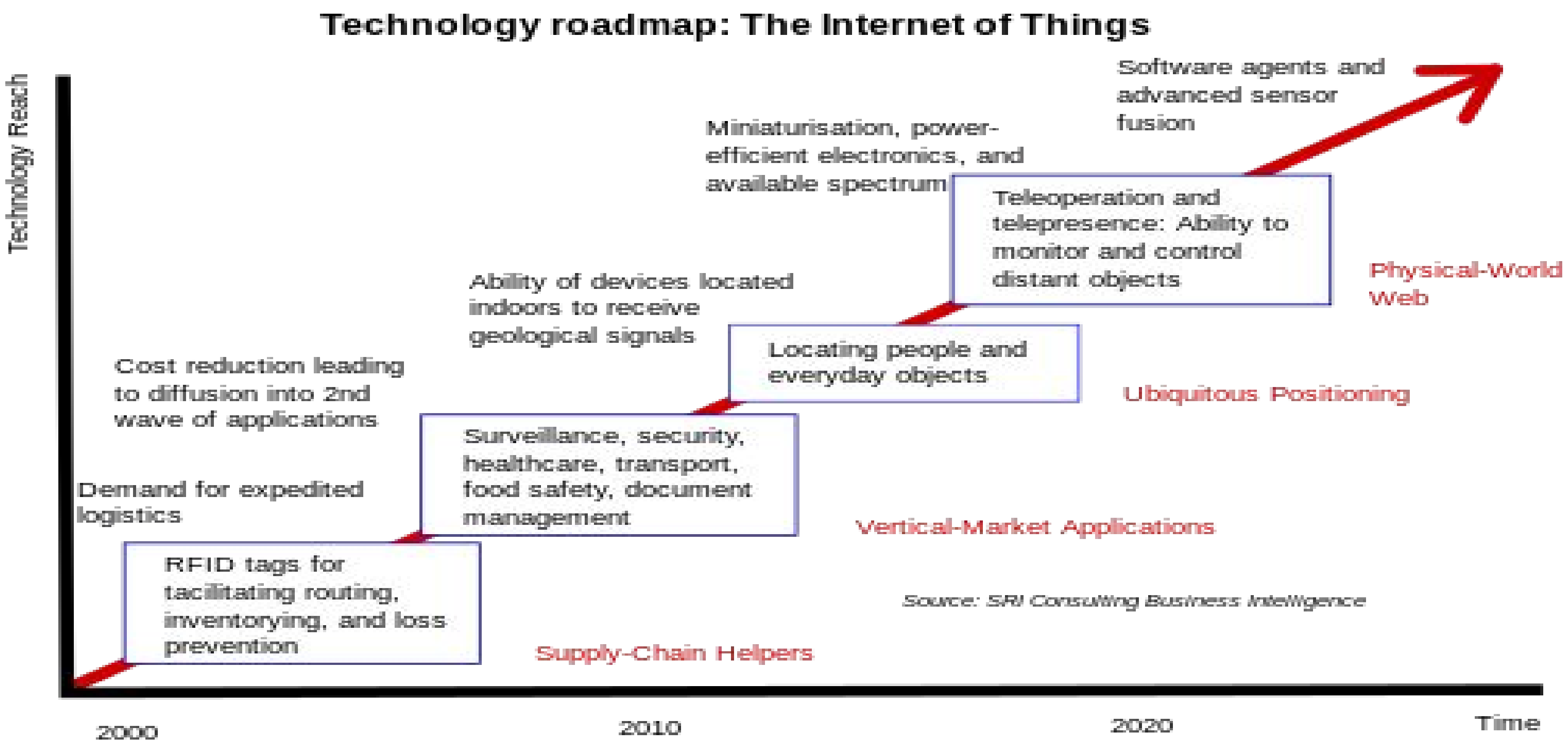
Search

Google Flu Trends



http://www.google.org/flutrends/about/how.html

Internet of Things


Technology roadmap: The Internet of Things
Technology Reach
Software agents and advanced sensor fusion
Miniaturisation, power-efficient electronics, and available spectrum
Teleoperation and telepresence: Ability to monitor and control distant objects
Physical-World Web
Ability of devices located indoors to receive geological signals
Locating people and everyday objects
Ubiquitous Positioning
Cost reduction leading to diffusion into 2nd wave of applications
Surveillance, security, healthcare, transport, food safety, document management
Demand for expedited logistics
Vertical-Market Applications
RFID tags for facilitating routing, inventorying, and loss prevention
Source: SRI Consulting Business Intelligence
Supply-Chain Helpers
2000
2010
2020
Time

https://commons.wikimedia.org/wiki/File%3AInternet_of_Things.svg
By SRI Consulting Business Intelligence/National Intelligence Council [Public domain], via Wikimedia Commons

Each day GE gathers “50 million pieces of data from 10 million sensors, off equipment worth $1 trillion.”

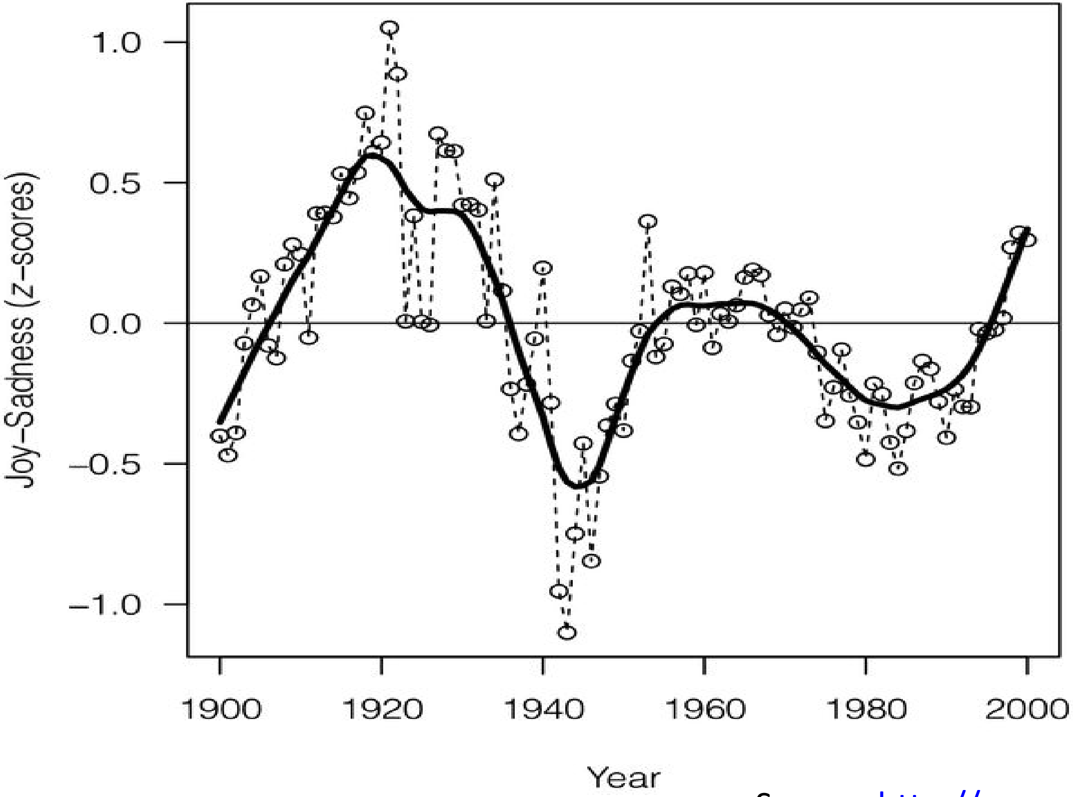
Product Company

Data Services Company

http://bits.blogs.nytimes.com/2014/10/09/ge-opens-its-big-data-platform/?_r=0

Almost the entire corpus of literature is now digital…

The Expression of Emotions in 20th Century Books

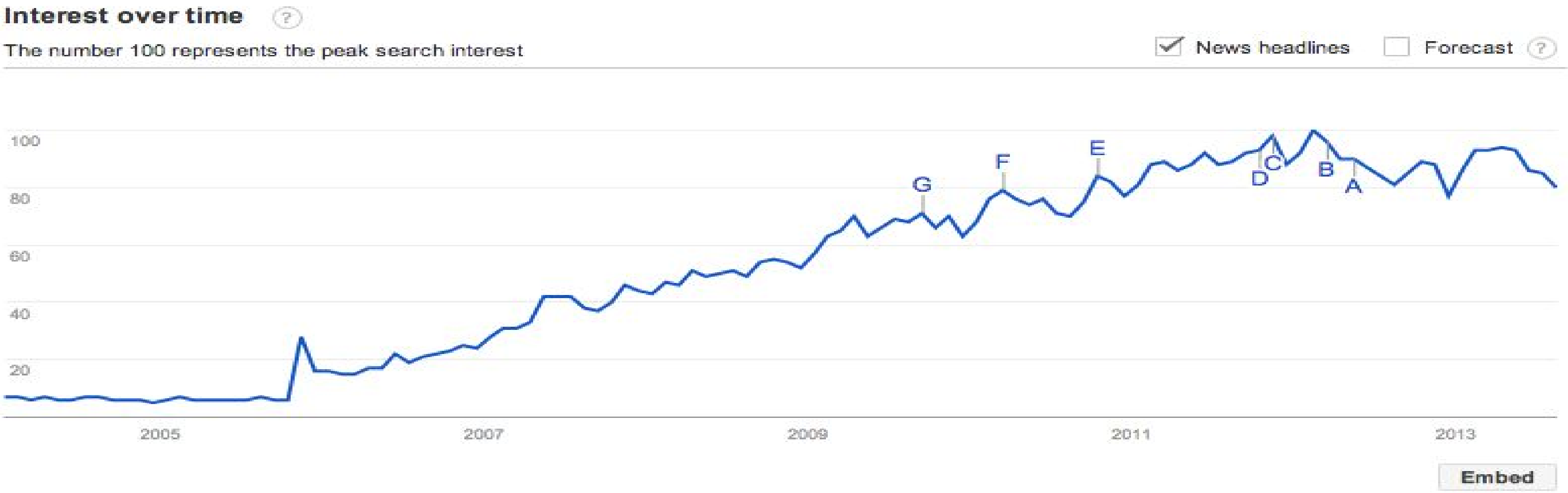


“using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative

Source: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

What do we mean by "Analytics"?

(Term Frequency in News)



Source: http://www.google.com/trends/explore?q=analytics#q=analytics&cmpt=q

What do we mean by “Analytics”?

Analytics as Data Apps

- The web is full of “data-driven apps.”

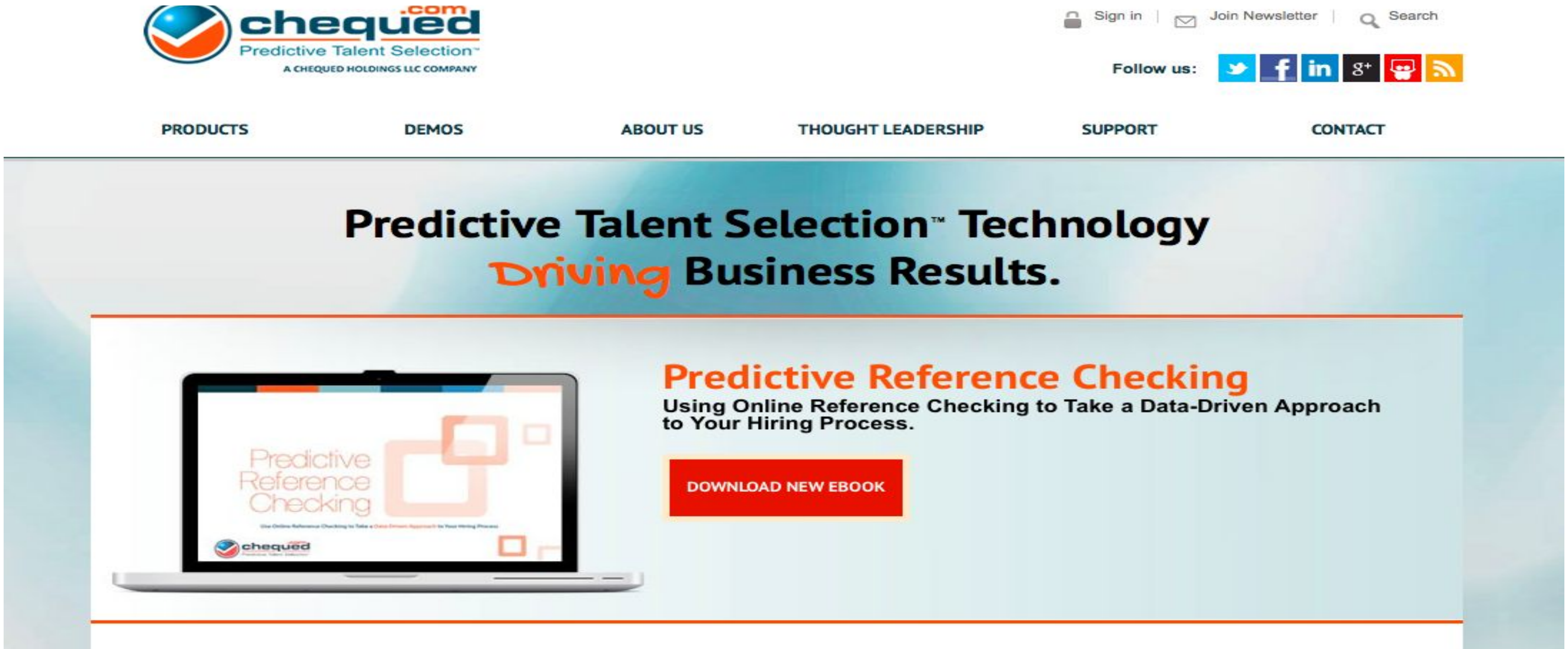
“The thread that ties most of these applications together is that data collected from users provides added value. Whether that data is search terms, voice samples, or product reviews, the users are in a feedback loop in which they contribute to the products they use. That’s the beginning of data science.”

Prediction is a Great Business Model! Even if it doesn’t use analytics!!!

THE

FARMER'S ALMANAC

Predict Who is Going to Be a Good Hire

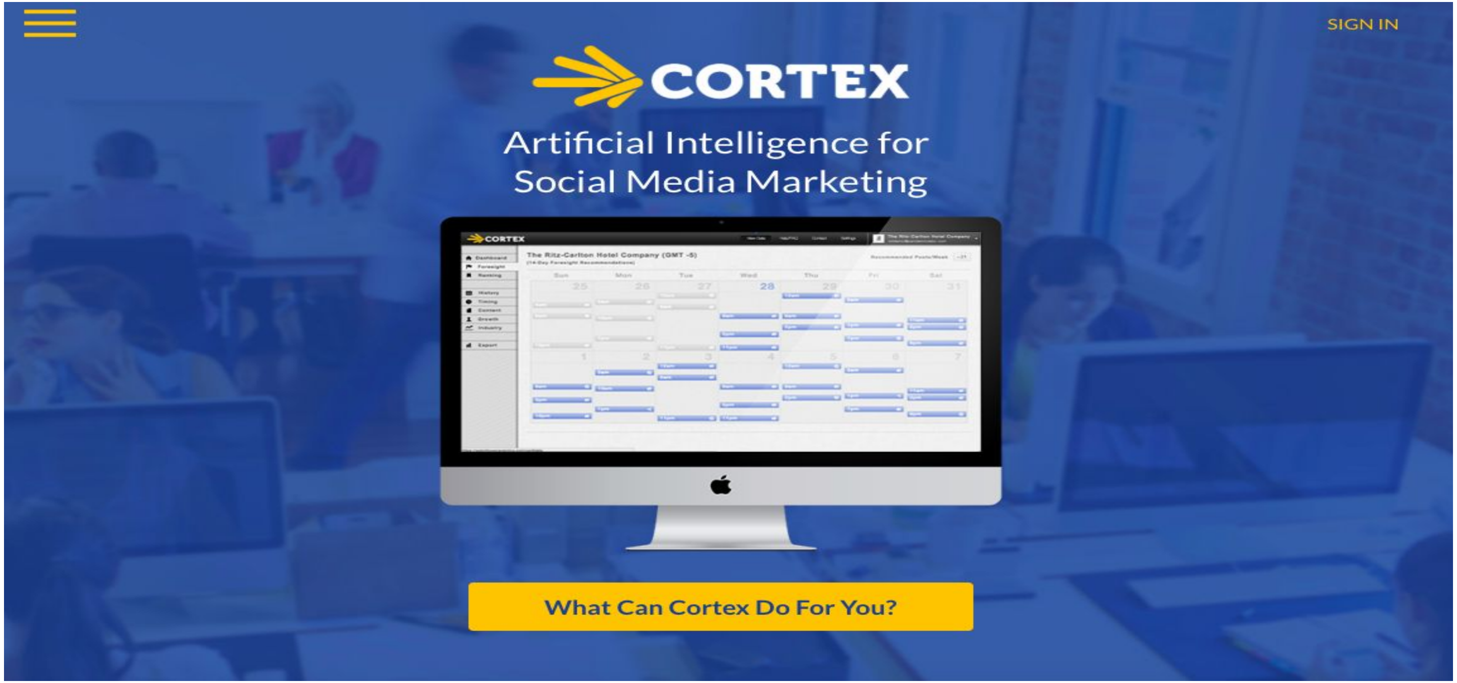


Predictive Talent Selection™ Technology Driving Business Results.

Predictive Reference Checking

Using Online Reference Checking to Take a Data-Driven Approach to Your Hiring Process.

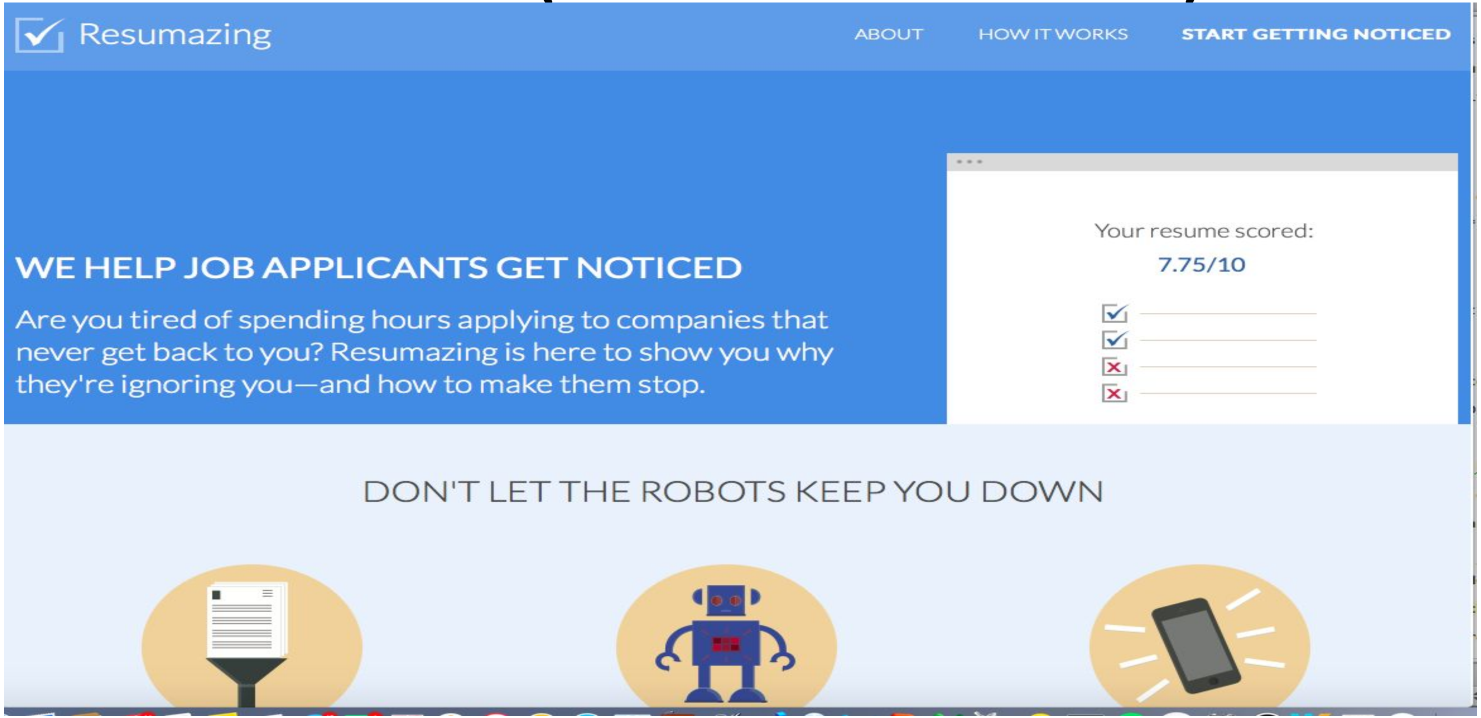
Predict the Best Type/Time of Post



SIGN IN
CORTEX
Artificial Intelligence for Social Media Marketing
The Ritz-Carlton Hotel Company (GMT -5)
What Can Cortex Do For You?

www.meetcortex.com

Predict what Terms Can Help Get You Hired (RPI STARTUP!!)



WE HELP JOB APPLICANTS GET NOTICED

Are you tired of spending hours applying to companies that never get back to you? Resumazing is here to show you why they're ignoring you—and how to make them stop.

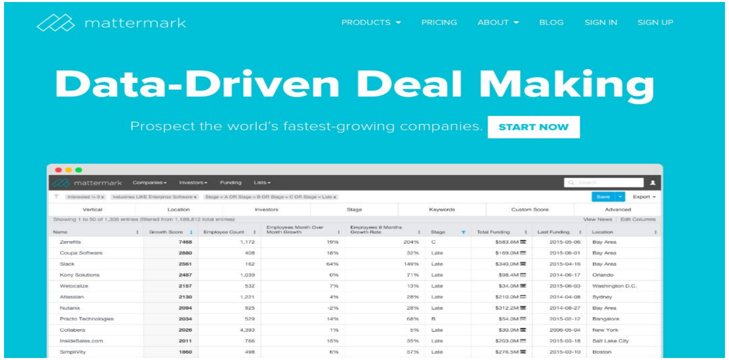
Your resume scored:

7.75/10

☑
☑
☒
☒

DON'T LET THE ROBOTS KEEP YOU DOWN

www.resumazing.net



Data-Driven Deal Making

Prospect the world's fastest-growing companies.

Interested != 0 x Industries LIKE Enterprise Software x Stage = A OR Stage = B OR Stage = C OR Stage = Late x

Showing 1 to 50 of 1,306 entries (filtered from 1,188,812 total entries)

Name	Growth Score	Employee Count	Employees Month Over Month Growth	Employees 6 Months Growth Rate	Stage	Total Funding	Last Funding	Location
Zenefits	7468	1,172	19%	204%	C	$583.6M	2015-05-06	Bay Area
Coupa Software	2880	408	18%	32%	Late	$169.0M	2015-06-01	Bay Area
Slack	2561	162	64%	149%	Late	$340.0M	2015-04-16	Bay Area
Kony Solutions	2487	1,039	0%	71%	Late	$98.4M	2014-06-17	Orlando
Welocalize	2157	532	7%	13%	Late	$34.0M	2015-06-03	Washington D.C.
Atlassian	2130	1,221	4%	28%	Late	$210.0M	2014-04-08	Sydney
Nutanix	2094	925	-2%	28%	Late	$312.2M	2014-08-27	Bay Area
Practo Technologies	2034	529	14%	68%	B	$54.0M	2015-02-12	Bangalore
Collabera	2026	4,393	1%	5%	Late	$30.0M	2006-05-04	New York
InsideSales.com	2011	766	15%	35%	Late	$203.0M	2015-03-18	Salt Lake City
SimpliVity	1860	498	6%	57%	Late	$276.5M	2015-03-10	Boston

Track Company Growth

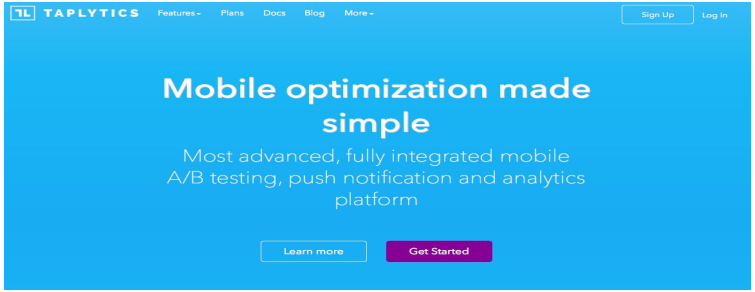
http://mattermark.com

Get the insights you need to stop customer churn.

Framed uses machine learning to analyze your analytics and tell you which customers are about to buy or abandon your product. Sign up today to find out how.

Find which Customers Will Churn

http://framed.io

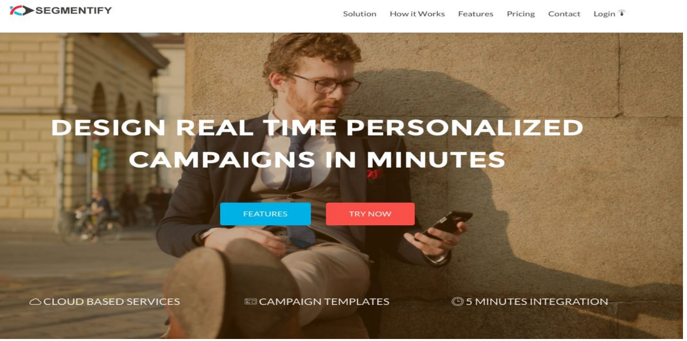


Mobile optimization made simple

Most advanced, fully integrated mobile A/B testing, push notification and analytics platform

A/B Testing for Mobile Apps

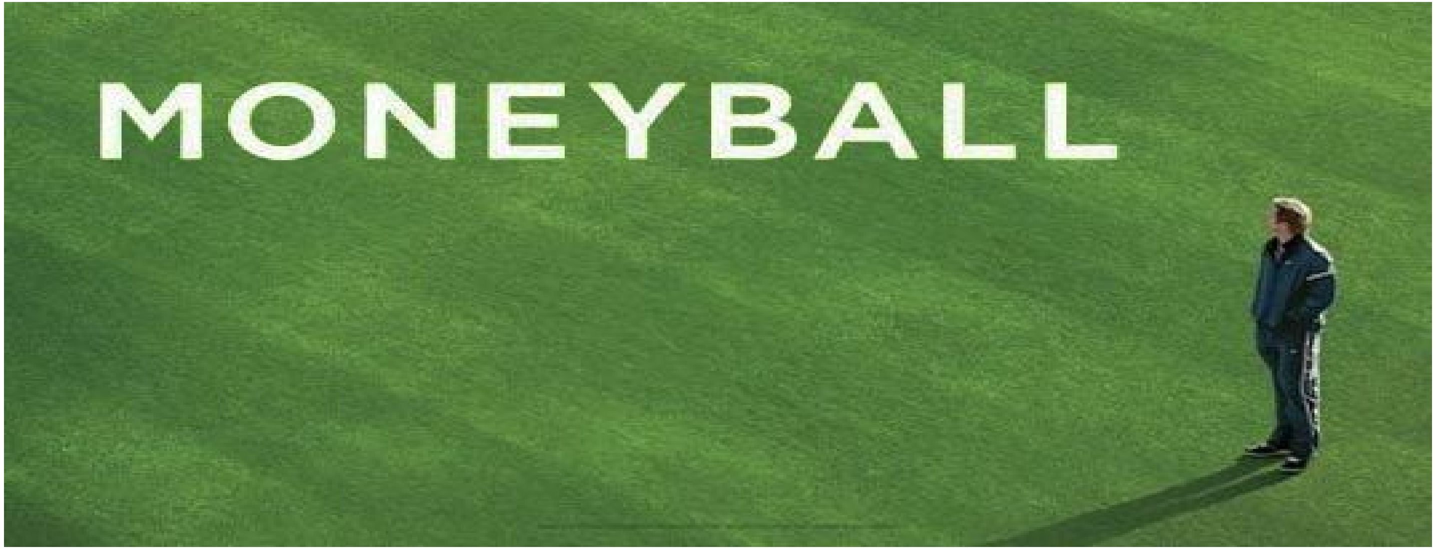
https://taplytics.com



DESIGN REAL TIME PERSONALIZED CAMPAIGNS IN MINUTES

Segment Customers

http://www.segmentify.com



Predict the players who are the best value…

http://www.youtube.com/watch?v=AiAHlZVgXjk

Predict Where to Advertise

- "The 2012 campaign took advantage of advanced set-top-box monitoring technology to figure out what shows the voters they wanted to reach were watching and when, resulting in a smarter and cheaper — if potentially more invasive — way to beam commercials into their homes. The system gave Obama a significant advantage over Mitt Romney, according to Democrats and many Republicans (at least those who were not on Romney's media team)."

Source: http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html?pagewanted=all&_r=0

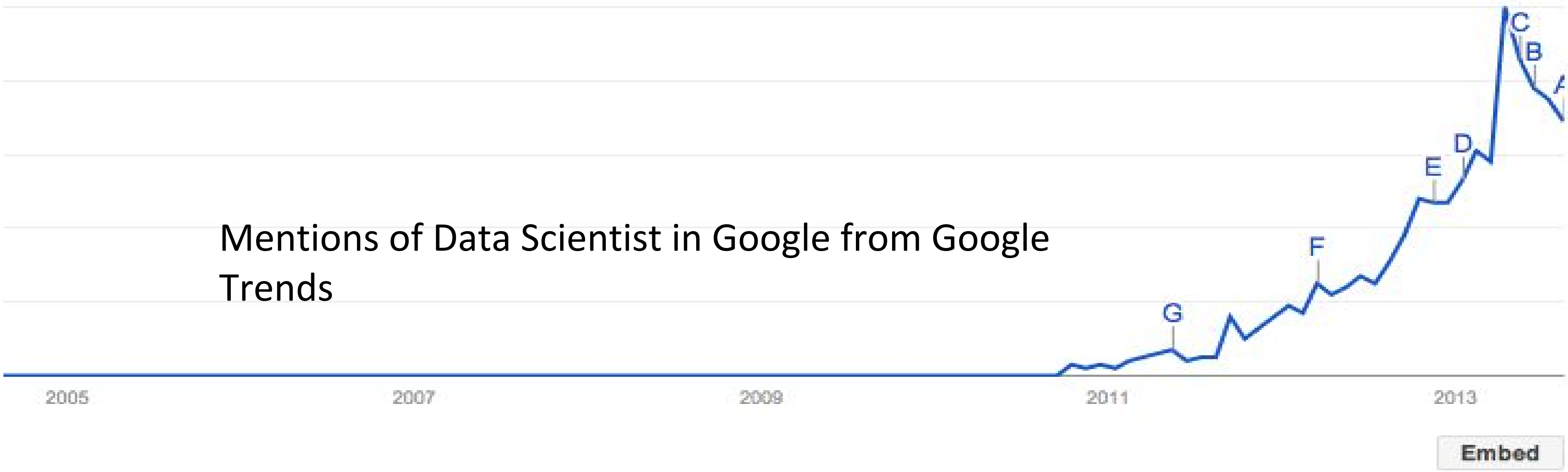
“**Analytics** is the discovery and communication of meaningful patterns in data.”

-Wikipedia

https://en.wikipedia.org/wiki/Analytics

The goal of this course will be to provide the technical foundation to enable students to become data scientists.

What do “Data Scientists” do?



Mentions of Data Scientist in Google from Google Trends

Data Scientist

- “A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

  - Hilary Mason, chief scientist at bit.ly

Data Scientist

Data science requires skills ranging from traditional computer science to mathematics to art. Describing the data science group he put together at Facebook (possibly the first data science group at a consumer-oriented web property), Jeff Hammerbacher said:

"... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization."

“**Analytics** is the discovery and communication of meaningful patterns in data.”

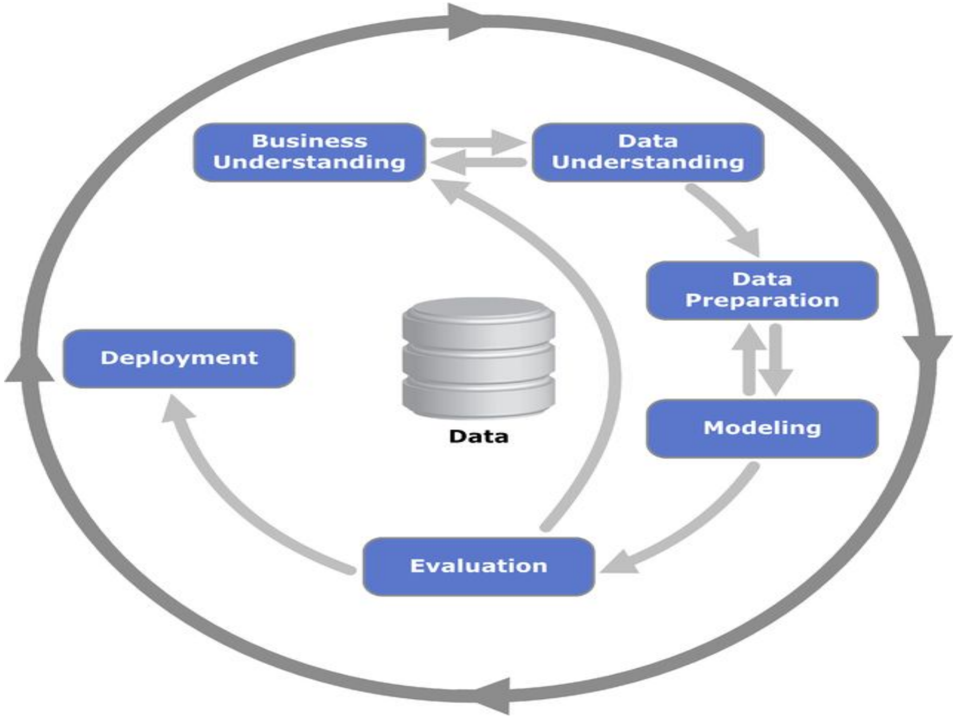
-Wikipedia

Data scientists do **analytics**.

https://en.wikipedia.org/wiki/Analytics

Well….how do you “do analytics”?

THE CRISP-DM PROCESS MODEL



Cross Industry Standard Process for Data Mining

https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Business Understanding

- Project objectives and requirements
  - Are we trying to *predict* or *understand*?
  - Do we want to reduce churn, segment customers, A/B test designs…etc.
- Converting knowledge into a data mining problem definition and a preliminary plan
  - What data is needed/accessible?

Data Understanding

- Data properties, type, distribution,
  - Understand data
- Data quality
  - Are there issues with the data? Is it as we would expect?

Visualization can be your friend for data understanding

Data Preparation

- Clean data
  - Missing values? Outliers?
- Match data from different sources
  - Common key? If not can we match text fields.
- Feature creation
  - Does data need to be processed to enable new insights?
- Aggregate, sample and subset
  - Do we want to do analysis on entire dataset, an aggregation?

Modeling

- Feature selection
  - Which of the identified features should be included in the model
- Splitting (fold) data
  - How should data be split to train the algorithm
- Algorithm selection and tuning
  - What algorithm should be used and what parameters set

Evaluation

- Statistical output (Understanding)
  - What data are meaningful in predicting the variable of interest
- Prediction
  - How well does the data predict the desired outcome in the training and test datasets

Deployment

- Implement prediction in a business process or in an application
  - Product recommendation
  - Job applicant
  - Customer segmentation

Data Analytics in the Wild

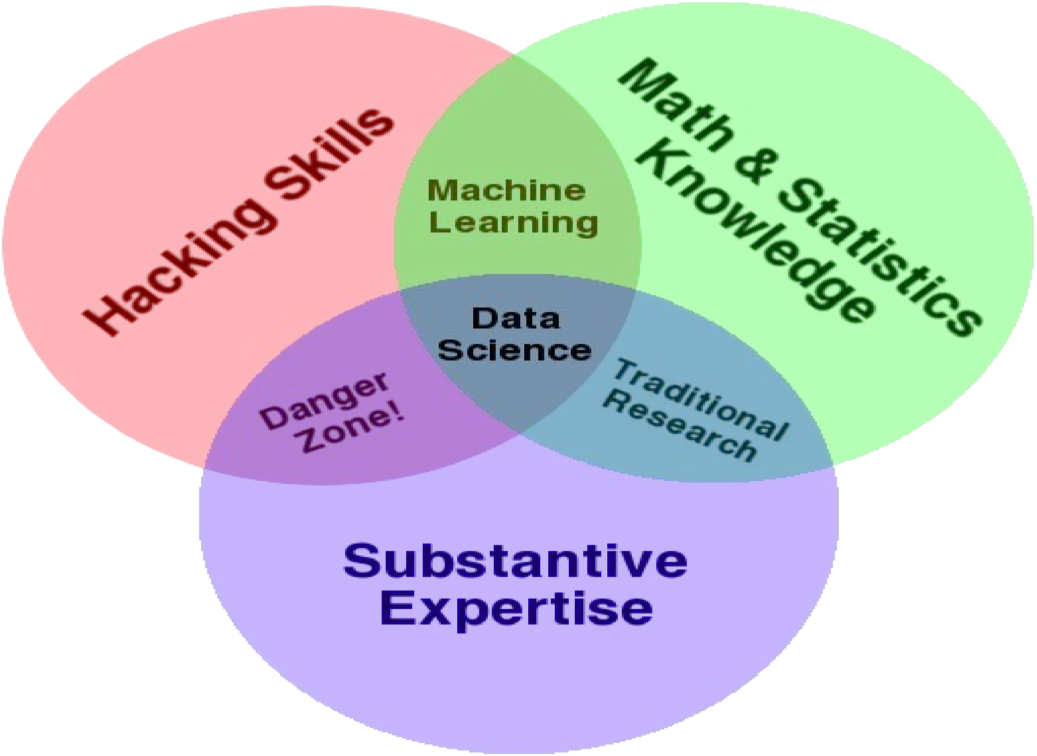
In a small group of 3-4, quickly discuss what you select for the data analytics in the wild assignment

Then go through the process (at a high level) for business understanding (and even a bit of data understanding) for the most interesting one in the group

Be prepared to report back to the class (~1 minute)

What skills are needed as a data scientist?

Data Science Venn Diagram



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Key Tools of the Data Scientist

- Data Munging - parsing, scraping, and formatting data
- Statistics - traditional analysis you're used to thinking about
- Visualization - graphs, tools, etc.

8 Skills to Get You Hired as a Data Scientist

1. Basic Tools
2. Basic Statistics
3. Machine Learning
4. Multivariable Calculus and Linear Algebra
5. Data Munging
6. Data Visualization & Communication
7. Software Engineering
8. Thinking Like A Data Scientist

http://blog.udacity.com/2014/11/data-science-job-skills.html

Abstractions vs Tools

- *Abstractions* of data science
  - Matrices and linear algebra
  - Relations and relational algebra
  - MapReduce
  - Feature selection in Visualization
- Tools
  - Python
  - R
  - SAS
  - SQL/MySQL
  - Hadoop (MapReduce)
  - Tableau (Visualization)

Syllabus and Lab