Data Science

Transforming Data into Knowledge and Vision
Understanding the Power and Beauty of Data

Introduction

William Yu, PhD
Economist
UCLA Anderson Forecast
Week 1

UCLAAnderson
SCHOOL of MANAGEMENT

APPLY | FOR COMPANIES | GIVE

Google Custom Searcl

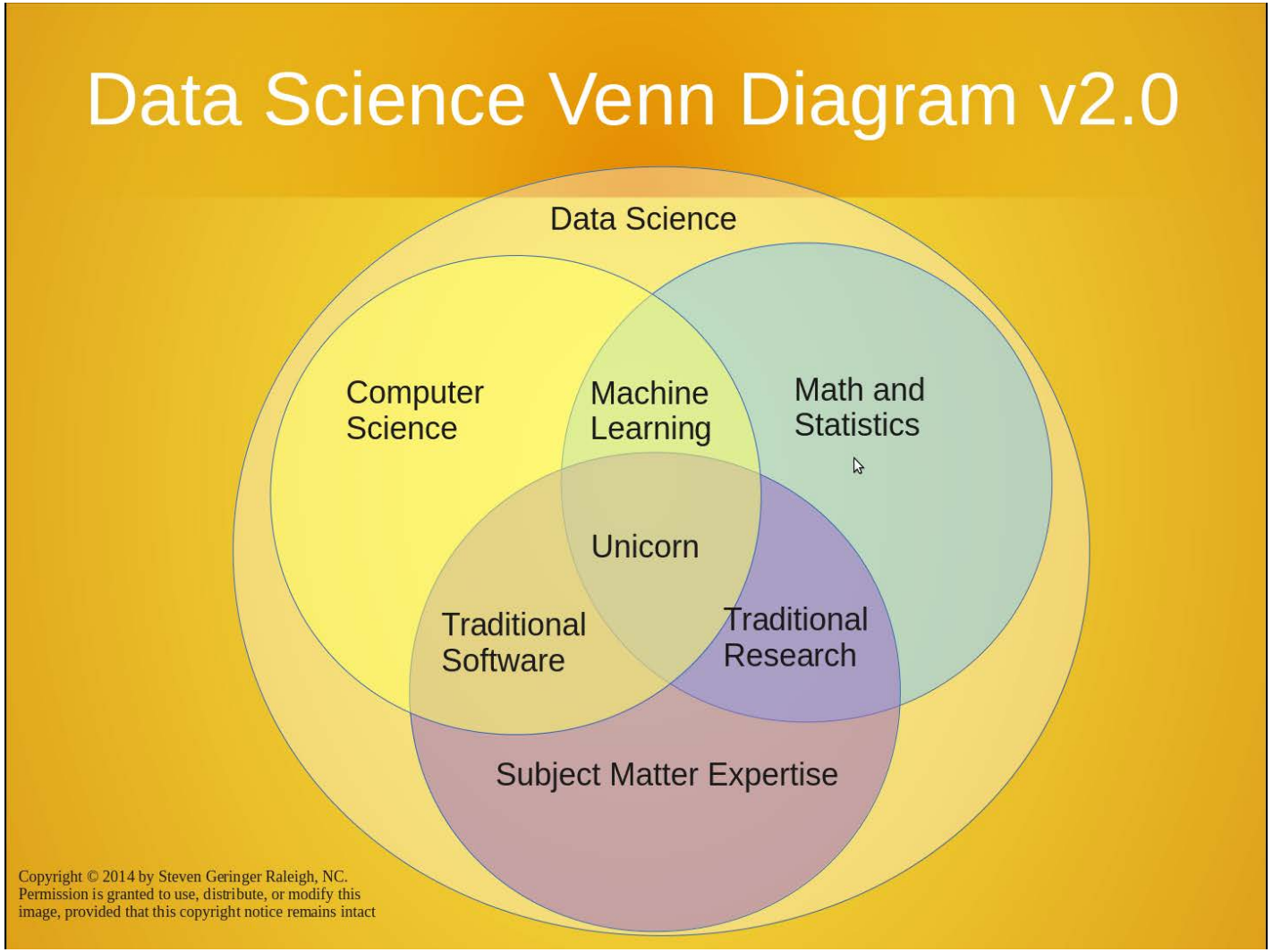
Home > Centers > UCLA Anderson Forecast > About Us

UCLA Anderson Forecast

- Overview
- About Us
  - *Edward Leamer*
  - *Jerry Nickelsburg*
  - *David Shulman*
  - *William Yu*
- Events
- Research
- Projects and Partnerships
- Membership
- Sponsorship
- Join Email List
- Contact Us

ABOUT THE FORECAST

Data science = programming + statistics + business knowledge



Key to be a good data scientist: GET YOUR HANDS DIRTY!!

Beyond programming, problem-solving skills are important!

Tools available to data scientists

- Data storage – MySQL, Oracle, SQL Server, Hbase, MongoDB, Redis
- Data querying – SQL (Structured Query Language), R, Python, and Java
- Data analysis – R, Python, SAS, Stata, Matlab, EViews
- Data visualization – Tableau, R, Python, and JavaScript
- Cloud – Amazon AWS, Microsoft Azure, and Google Cloud
- Hadoop Big Data – Spark, PIG, Hive, HDFS MapReduce

R vs Python

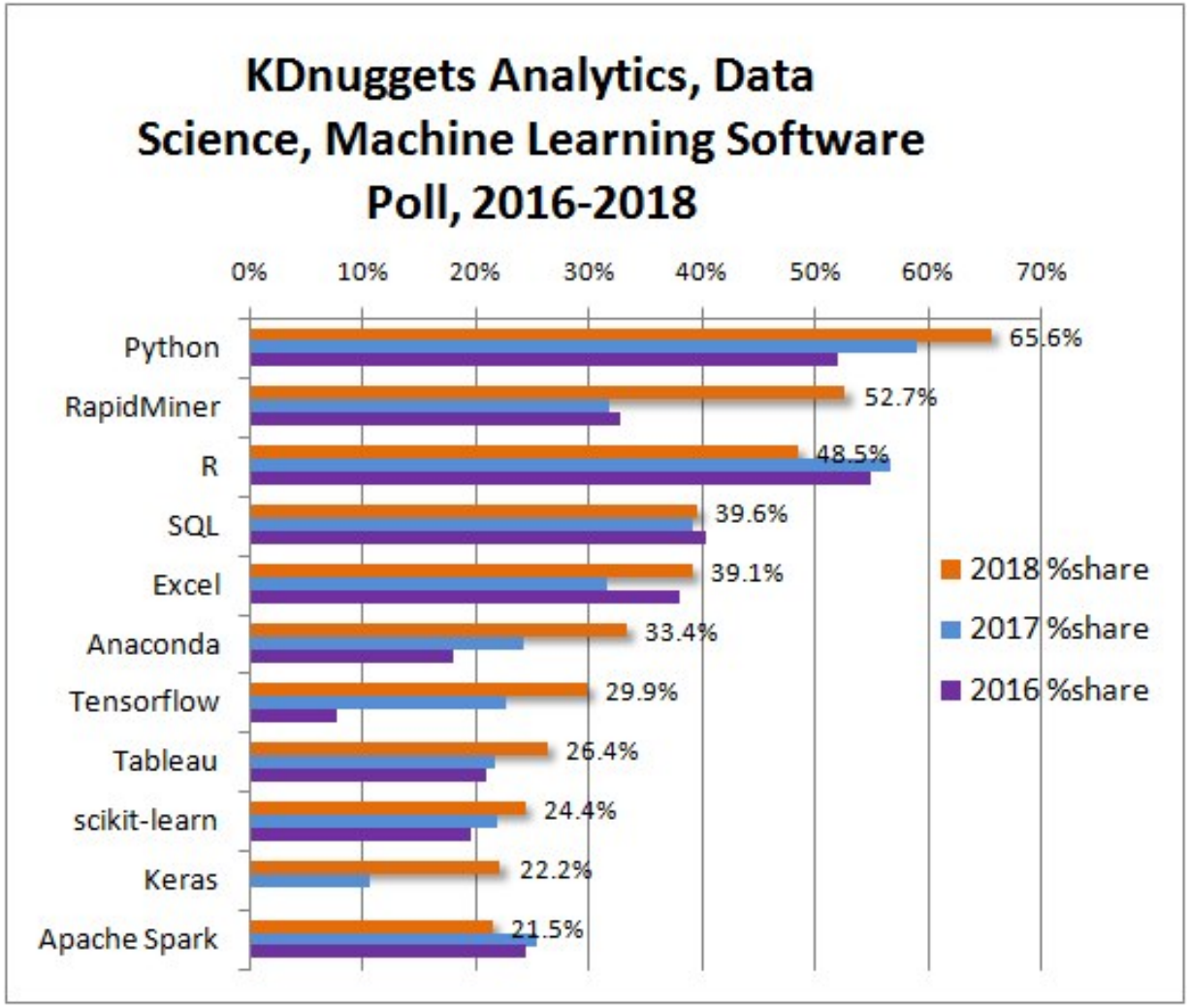
R

- Open source and free
- A language for statistics
- More than 10,000 packages:
  - Data cleaning and management
    - dplyr
  - Exploratory data analysis
    - ggplot2
  - Interactive data visualizations
    - Shiny
  - Machine learning
    - Caret
  - Communication
    - Jupyter Notebook, RMarkdown

Python

- A wide variety of use
- Only two versions: Python 2.7 and 3.6
- Few packages:
  - Data cleaning and management
    - pandas, SciPy, NumPy, StatsModels
  - Exploratory data analysis
    - Pandas, matplotlib
  - Interactive data visualizations
    - Bokeh, Plotly
  - Machine learning
    - scikit-learn
  - Communication
    - Jupyter Notebook
  - Other
    - Beautiful Soup (Web scraping)
    - Tweepy (Twitter scraping)

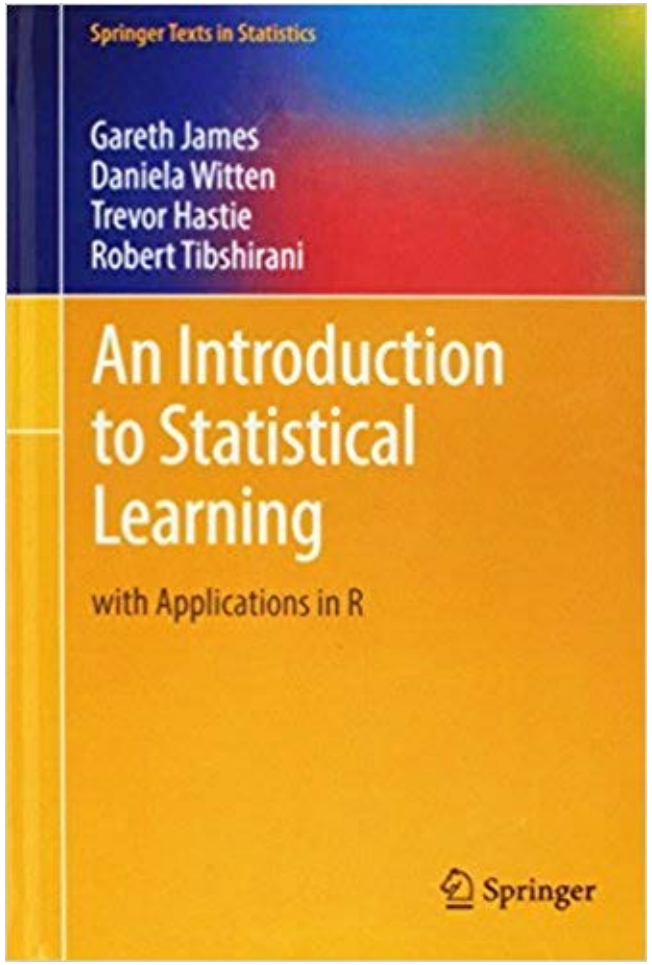
Top analytics, data science, machine learning tools

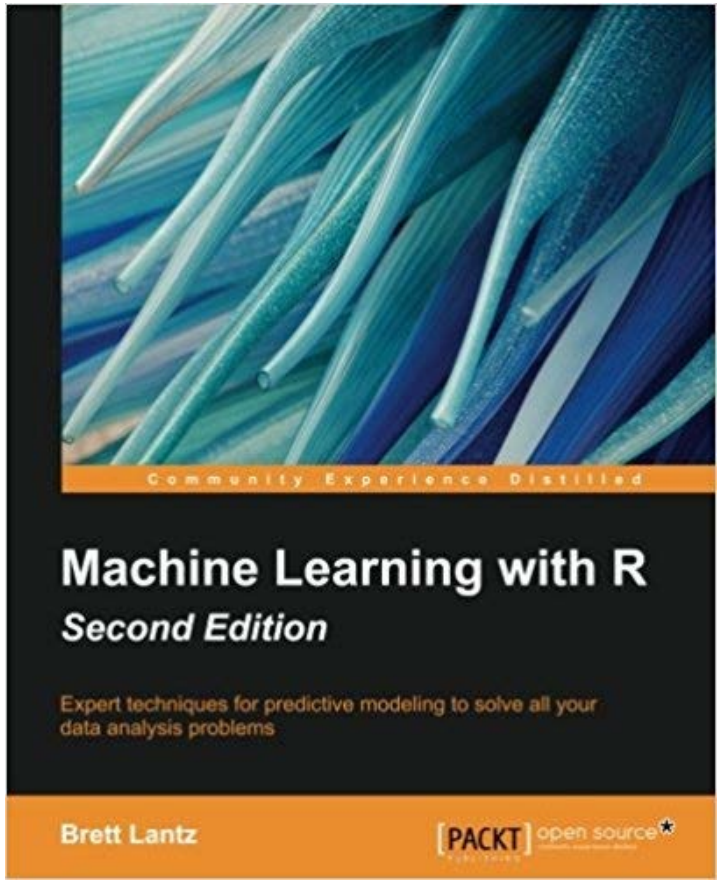


Detailed ranking:
https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html/2

Textbooks

- An Introduction to Statistical Learning with Applications in R
- By James, Witten, Hastie, and Tibshirani
- Book website:
- http://www-bcf.usc.edu/~gareth/ISL/
- Book free download PDF file:
- http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf





https://github.com/stedy/Machine-Learning-with-R-datasets

Introduction to *R*

- Download R 
  - https://cran.r-project.org/bin/windows/base/ for Windows
  - https://cran.r-project.org/bin/macosx/ for Mac
- Download R Studio, a more productive platform for R 
  - https://www.rstudio.com/products/rstudio/download/ (select the free one!)
- Why learning R?
  - Once you learn one computer language, it is much easier to learn others.
  - R is open source.
  - R platform is organic with huge potential growth via its packages.
  - R provides advanced statistical, data analytical, and time series econometrics tools.
- Resources for learning R
  - A short introduction to R https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
  - A introduction to R https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
  - Elementary Statistics with R http://www.r-tutor.com/elementary-statistics
  - UCLA IDRE Resources on R http://www.ats.ucla.edu/stat/r/

Introduction to **Anaconda Distribution**

https://www.anaconda.com/download/

- Python and R distribution.
- The most popular Python data science platform
- Provide a simple and convenient online resource and community.
- Provide tools like Spyder (IDE/editor) and Jupyter Notebook, etc.

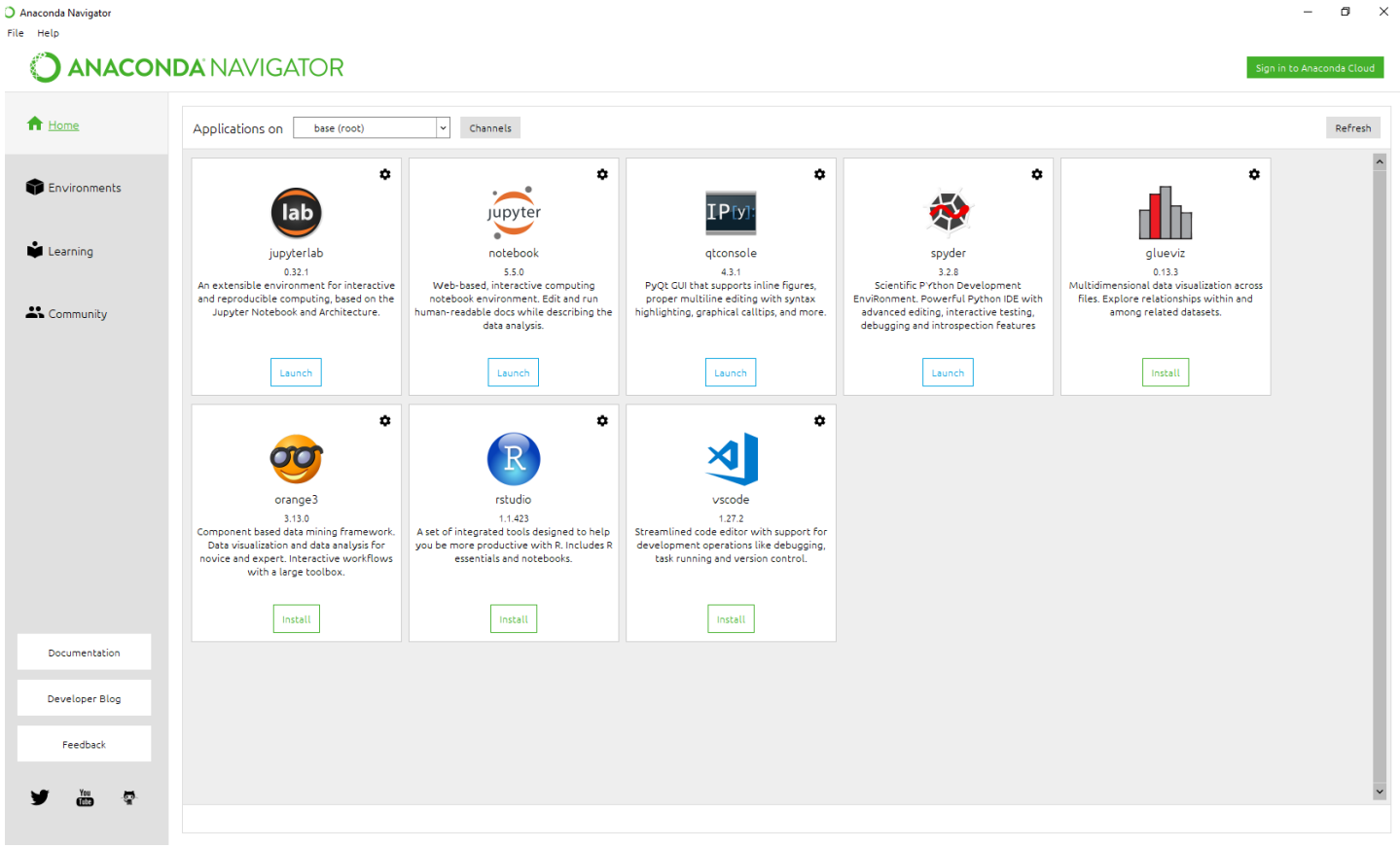
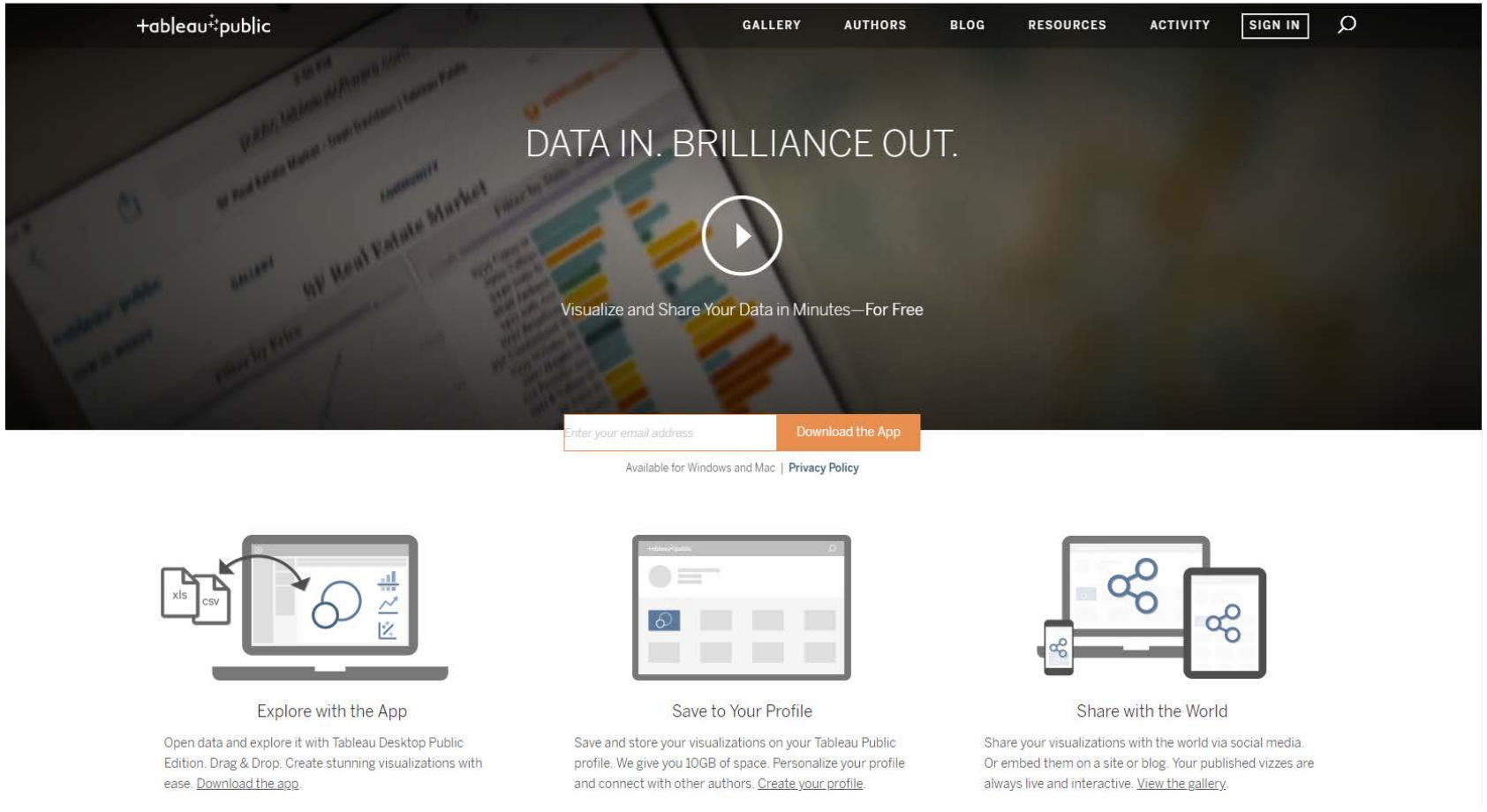


Tableau Public
the wizard of data visualization

https://public.tableau.com/en-us/s/



Useful resources

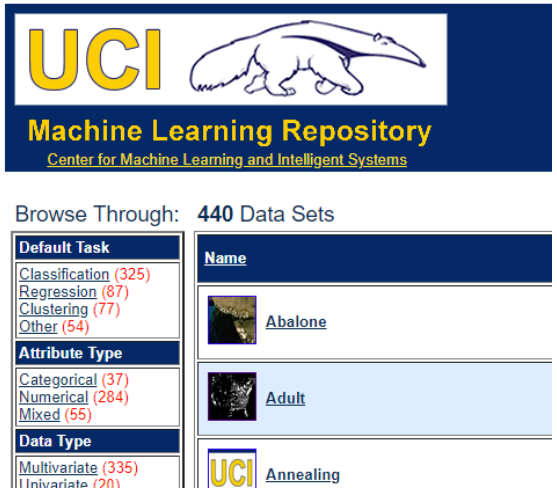
Public Datasets

https://github.com/awesomedata/awesome-public-datasets

https://www.kdnuggets.com/websites/index.html

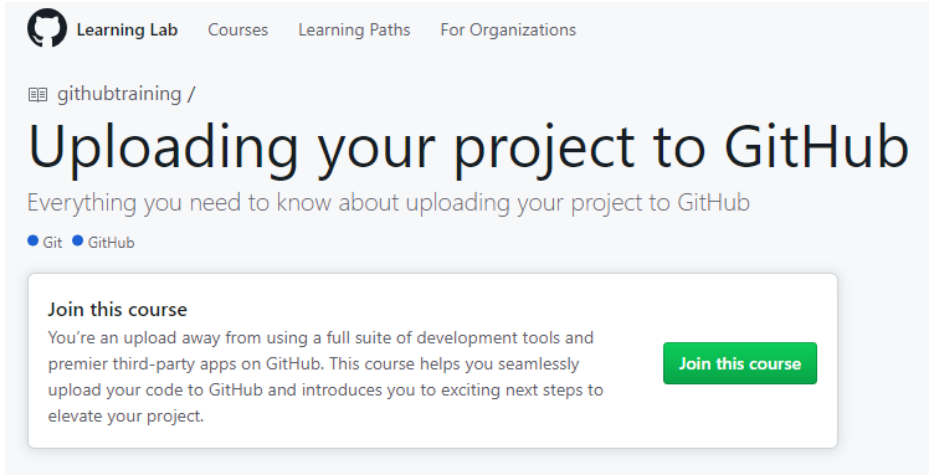
UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets.html

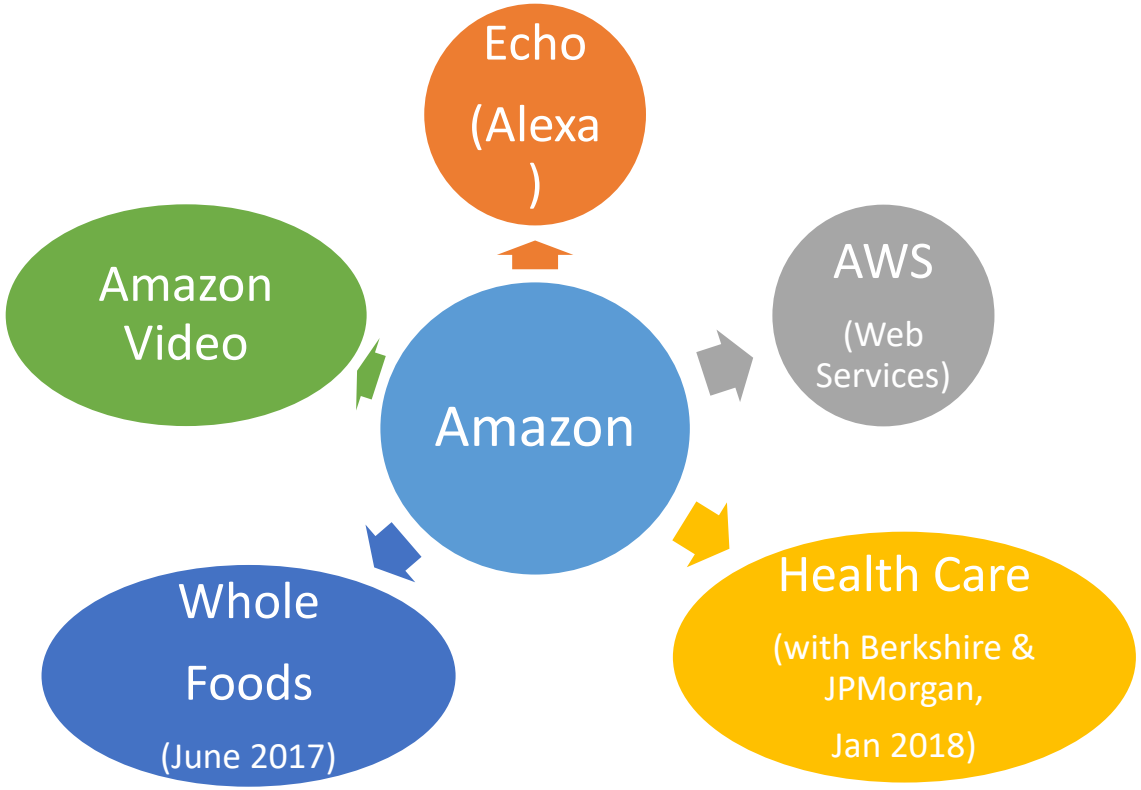


Github

https://github.com/



Why businesses are afraid of becoming Amazon's competitors?



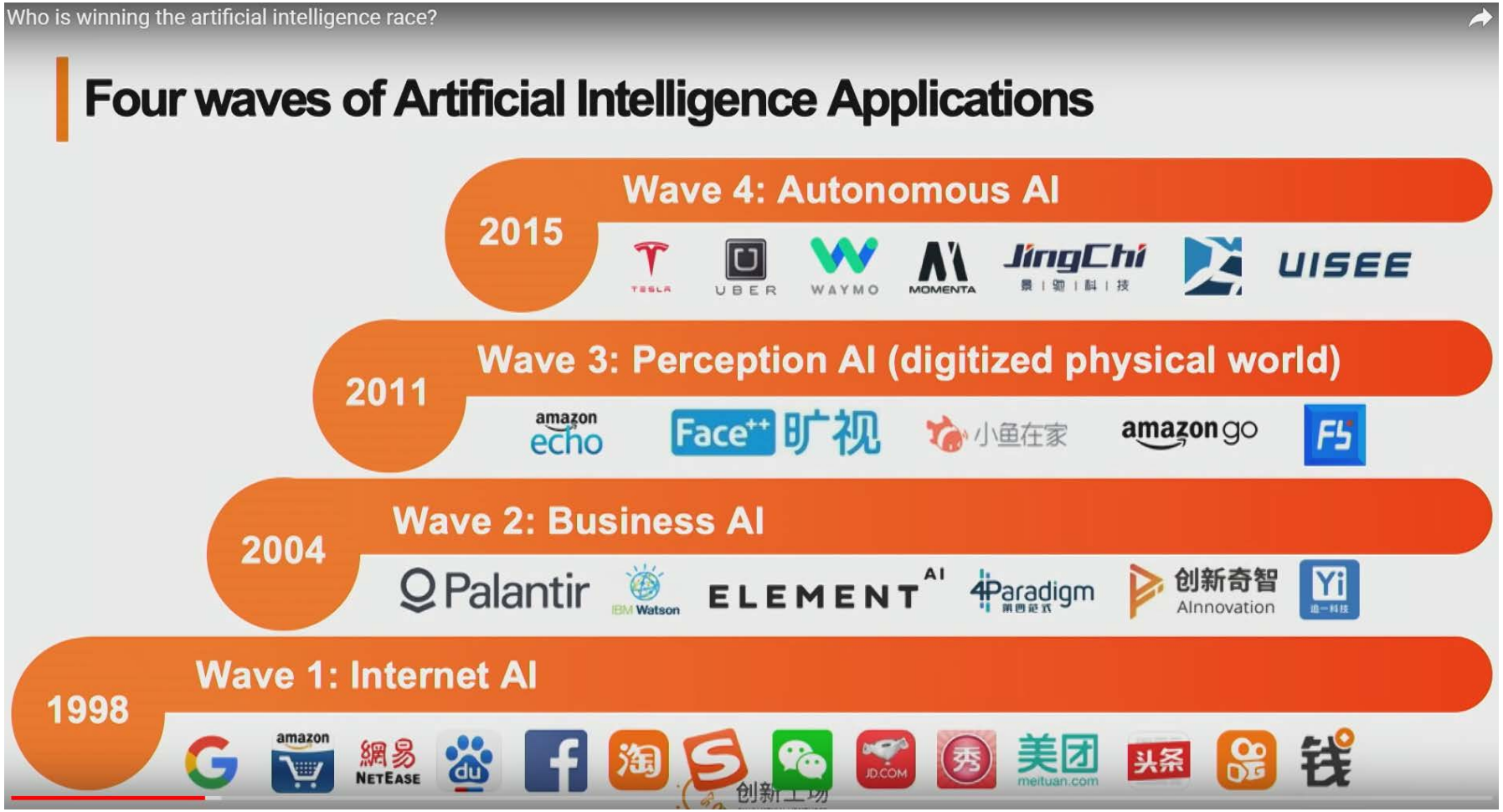
Ans: The weapon behind Amazon's aggressive expansion is its data analytics prowess!

Markets reward those companies with growth potential and data analytics prowess

Company	Market Capitalization ($ Billion)	Price-Earning Ratio
Amazon	$884	229
Oracle	$193	54
Google	$825	50
Facebook	$600	34
Walmart	$258	29
Apple	$938	18
Time Warner	$77	15
AT&T	$234	6

- In addition to instincts, experiences, theories, businesses could make better decisions and management by *analyzing data*.

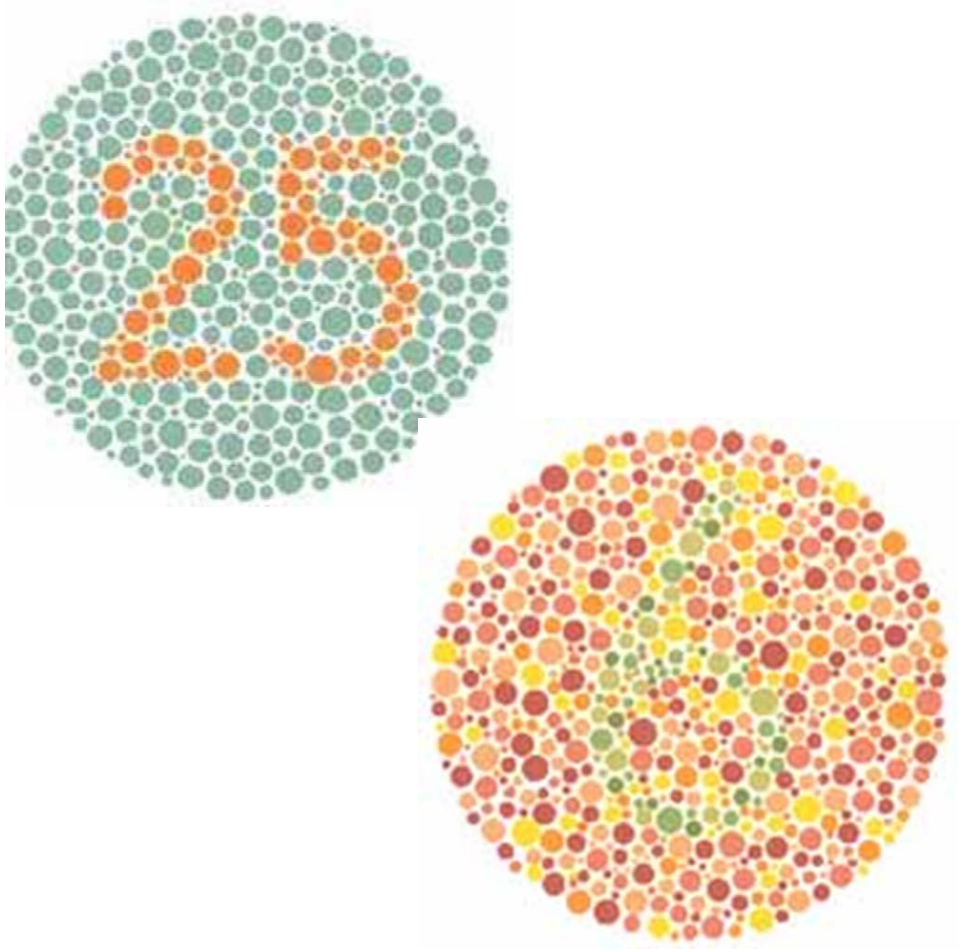
Data is the new oil. Data is the driver of AI.



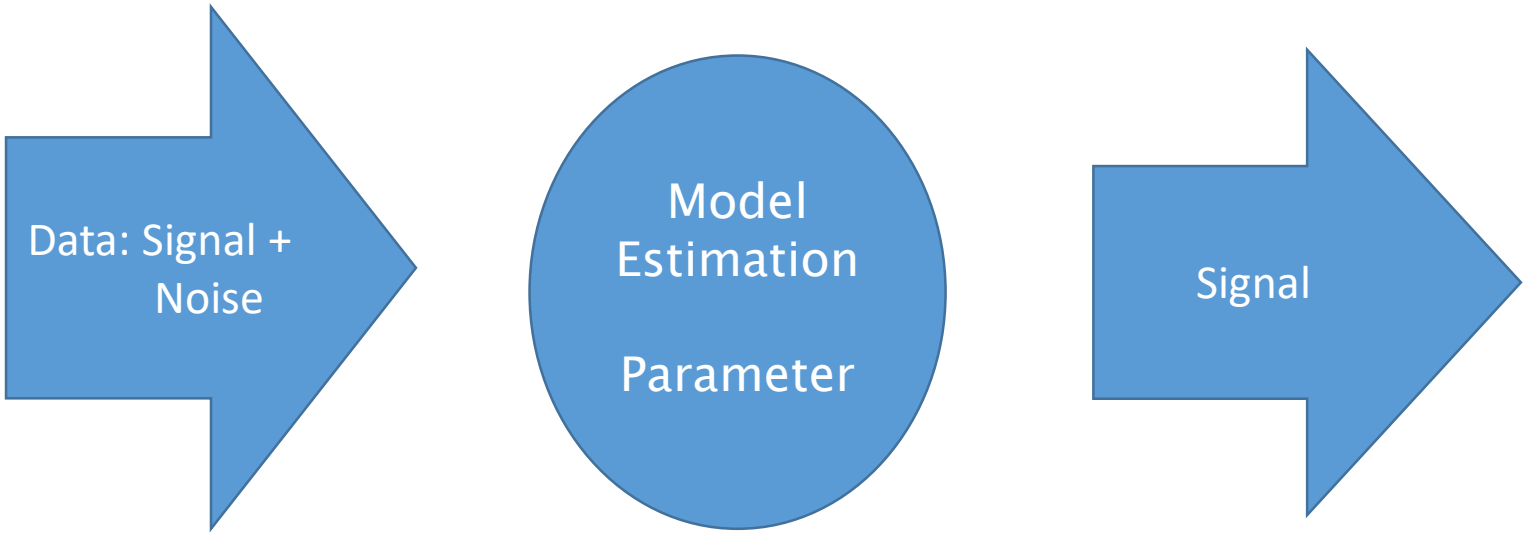
Source: From Kai-Fu Lee's speech at Stanford

How to transform data into knowledge and vision?

- Facing astronomical data, note that some are important, while others are not.
- Within important data, there are two components:
- (1) **Information / signal / pattern**
- (2) **Noises**
- Once the pattern is found, you can make forecast and decision.



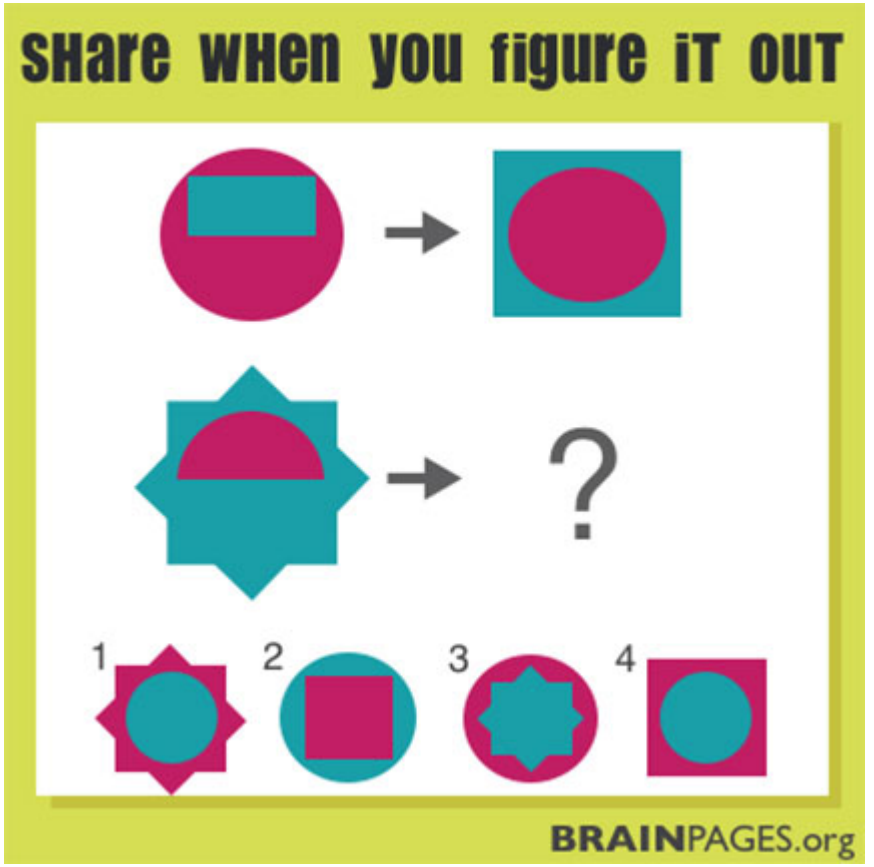
Data, model, machine learning, AI



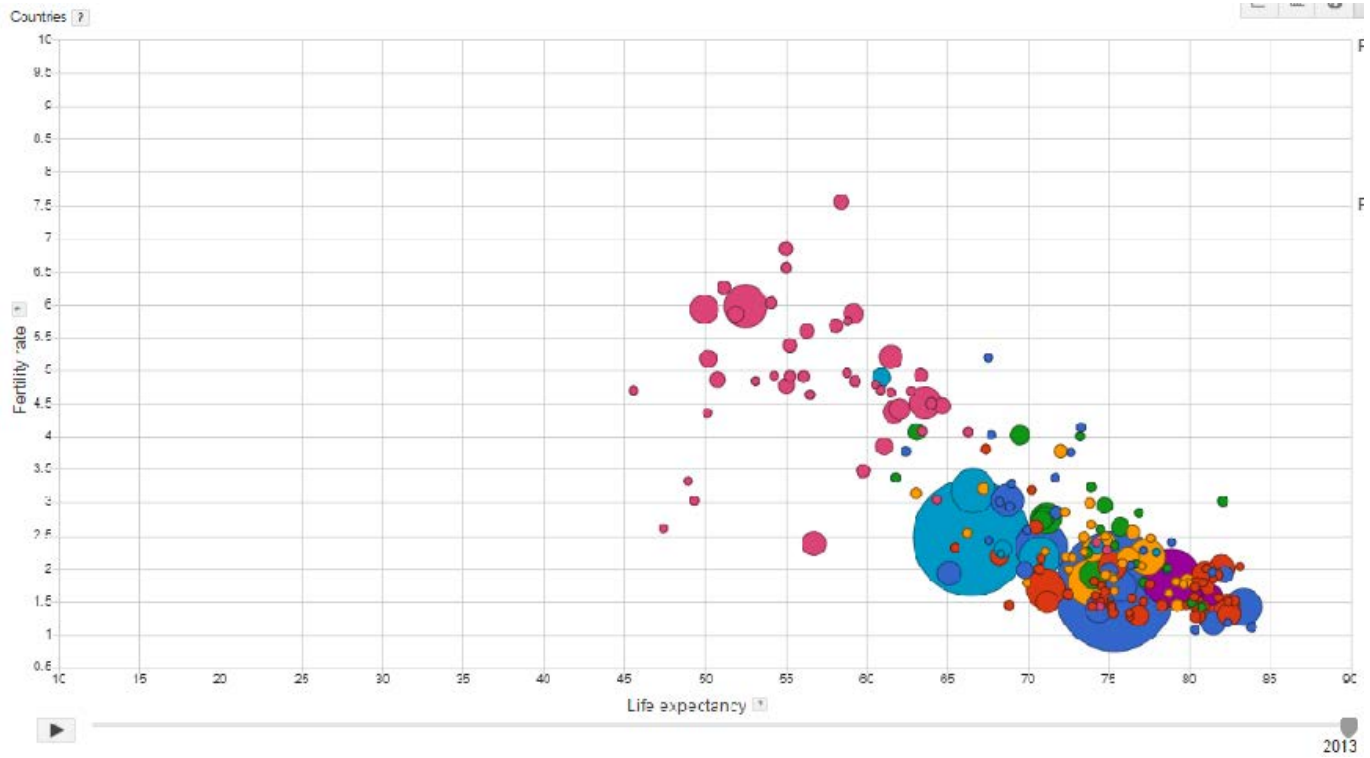
- Using a set of models performed by computer, it is called Machine Learning that learns from data to figure out pattern and signal.
- AI: By feeding or adding the data daily to the model/machine, it is making updated parameters. In other words, by training the computers, it is improving its forecasts or recognizing patterns rapidly. AI comes from data.
- A good model/analytics is that it extract all available signal from data. No signal will be left in noise. Noise is like white noise, so called an I.I.D. (Independent, Identical Disturbance)

Can pattern-seeking skills be learned?

- Pattern recognition is a gift, but it can also be learned.
- How?
- (1) Visualization / charts
- (2) Models



Correlation is essential in data analytics



- Google Public Data (https://www.google.com/publicdata/directory)
- Check correlation between GDP per capita and life expectancy

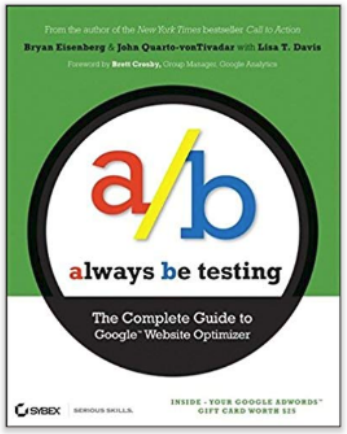
Correlation is used widely in marketing tools

Always Be Testing: The Complete Guide to Google Website Optimizer

by Bryan Eisenberg (Author), John Quarto-vonTivadar (Author), Brett Crosby (Foreword), Lisa T. Davis (Contributor)

28 customer reviews

Look inside



ISBN-13: 978-0470290637
ISBN-10: 0470290633
Why is ISBN important?

Have one to sell? Sell on Amazon

Add to List

Share <Embed>

Paperback $7.50 - $18.81 | Other Sellers See all 7 versions

Buy used $7.50

Buy new **$18.81**

Only 1 left in stock - order soon.
Ships from and sold by u_pick.

List Price: ~~$29.99~~
Save: $11.18 (37%)

9 New from $17.49

Get it as soon as July 23 - 30 when you choose **Standard Shipping** at checkout.

Deliver to Losangeles 90001

$18.81 + $3.99 shipping

Add to Cart

Turn on 1-Click ordering

More Buying Choices

56 used & new from $4.50

9 New from $17.49 | 47 Used from $4.50

See All Buying Options

prime student College student? Get FREE shipping and exclusive deals LEARN MORE

Frequently bought together

Total price: $55.97

Add all three to Cart

Add all three to List

These items are shipped from and sold by different sellers. Show details

This item: Always Be Testing: The Complete Guide to Google Website Optimizer by Bryan Eisenberg Paperback $17.79

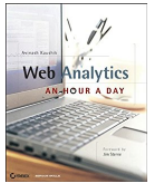
Landing Page Optimization: The Definitive Guide to Testing and Tuning for Conversions by Tim Ash Paperback $16.98

Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity by Avinash Kaushik Paperback $21.20

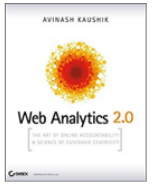
Customers who bought this item also bought

Page 1 of 3

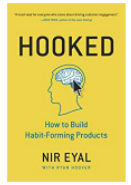
Landing Page Optimization: The Definitive Guide to…
› Tim Ash
139
Paperback
$16.98 prime



Web Analytics: An Hour a Day
› Avinash Kaushik
103
Paperback
113 offers from $1.95



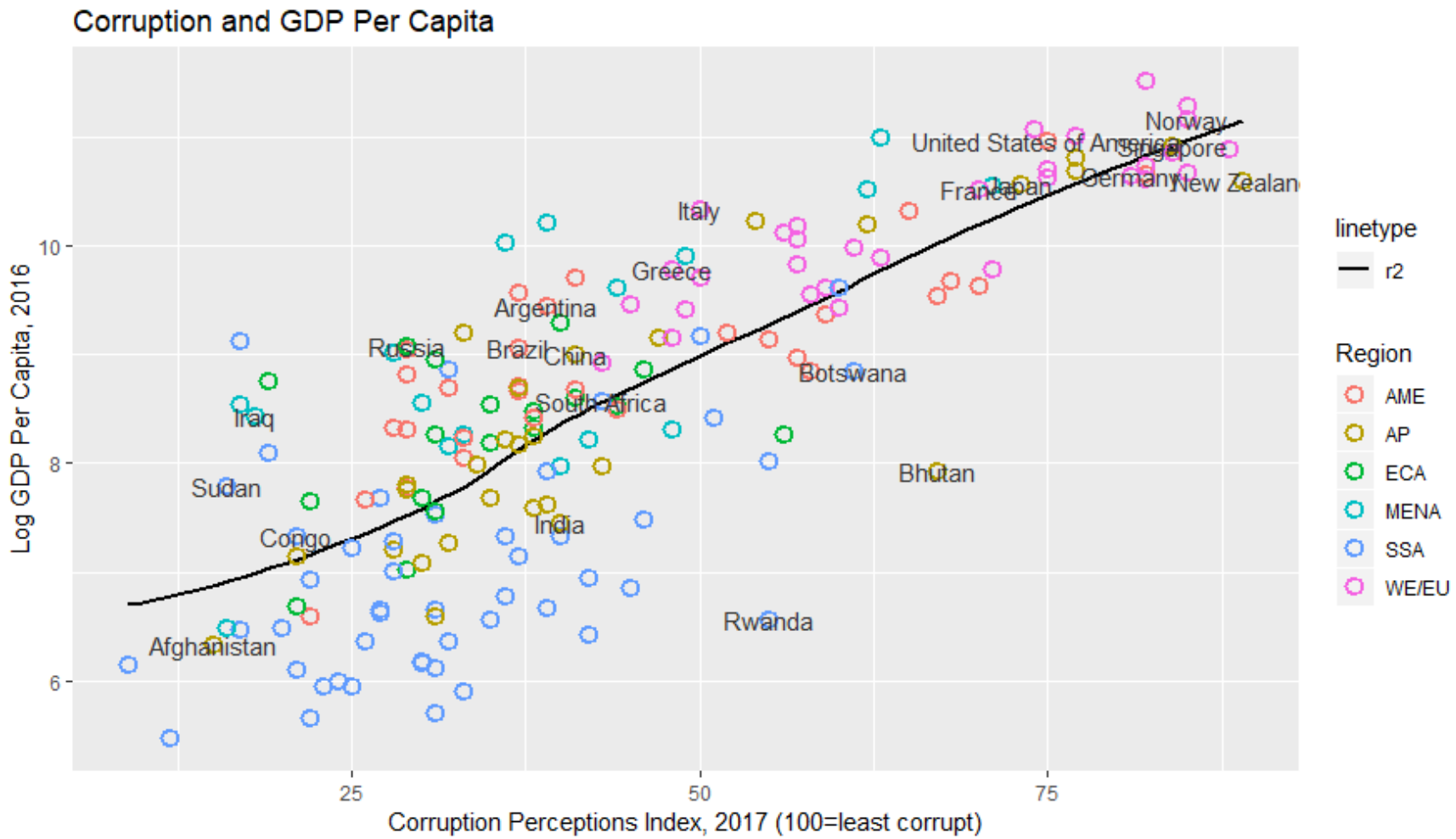
Web Analytics 2.0: The Art of Online Accountability and Science of Customer…
› Avinash Kaushik
138
Paperback
$21.20 prime



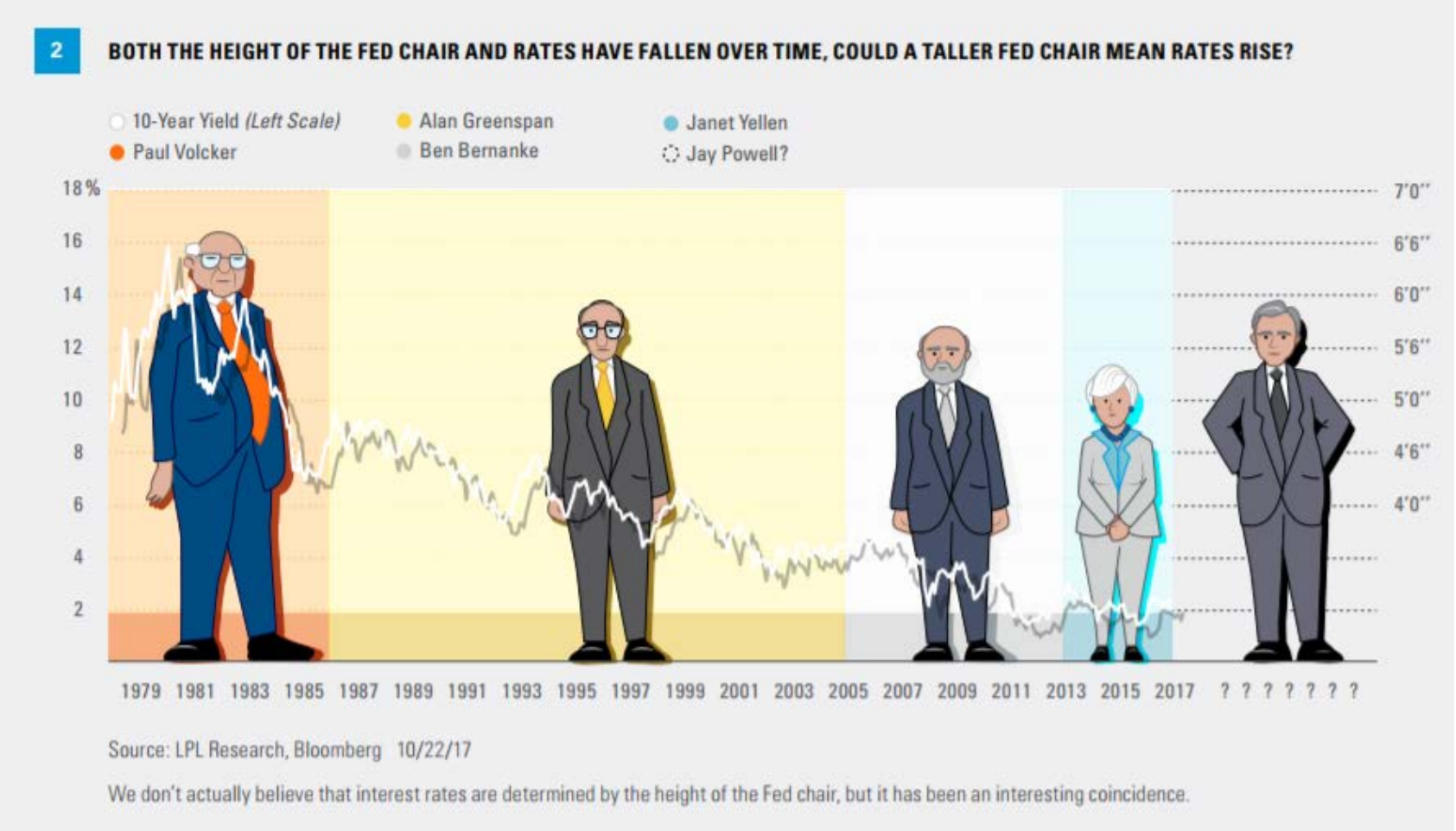
Hooked: How to Build Habit-Forming Products
› Nir Eyal
1,092
Hardcover
$16.04 prime

Be Like Amazon: Even a Lemonade Stand Can Do It
› Jeffrey Eisenberg
98
Hardcover
$14.93 prime

R package: ggplot2 on correlation



Correlation is very important but be careful about Spurious Correlation: http://tylervigen.com/spurious-correlations



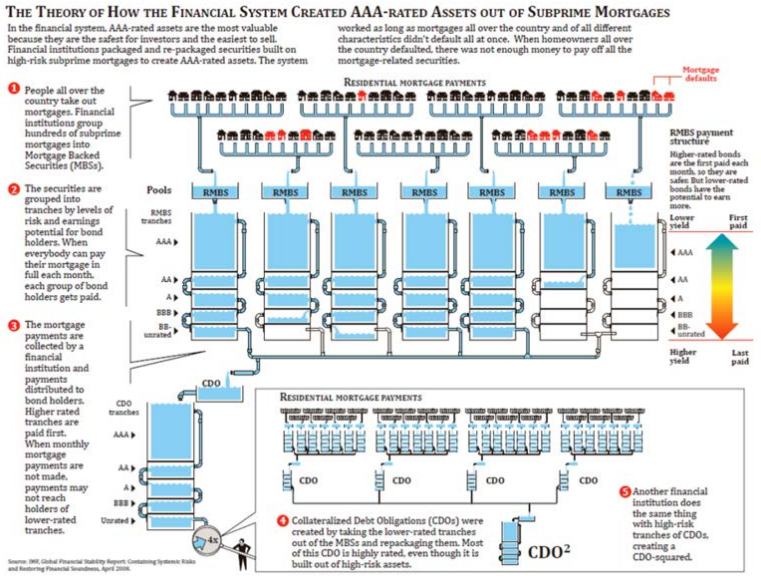
Data science and human decision/judgement

- Human beings are **pattern seeking** and **story telling**
- To have economic/statistical analytical tools, models, and theories
- To have the understanding of the context
- It is both *Science* and *Art*
- Parsimony Principle
  - **KISS**
  - Keep It Sophisticatedly Simple

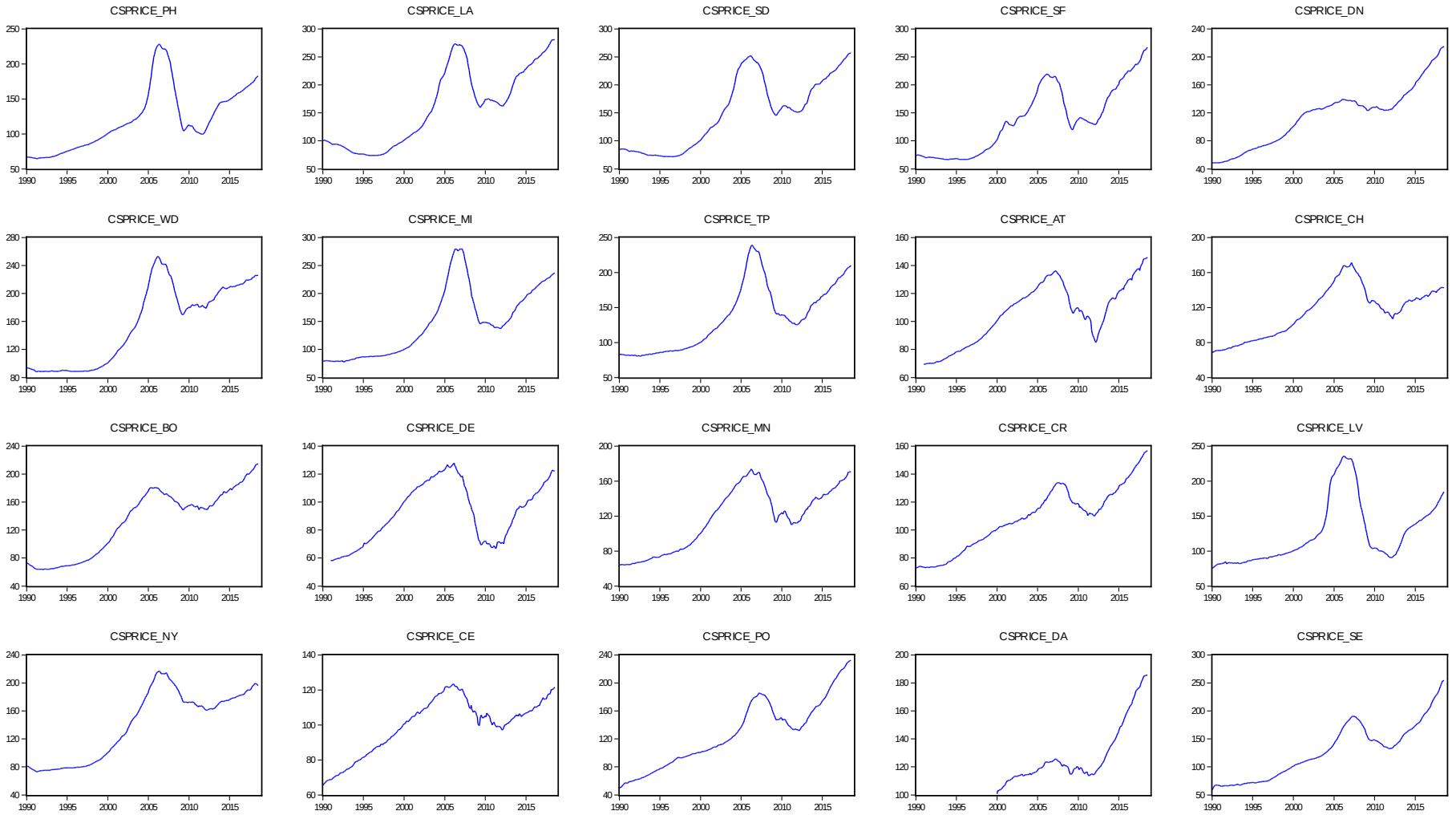
A pattern without a good story

An Irish boy born in ‘90,

height from birth to age 11



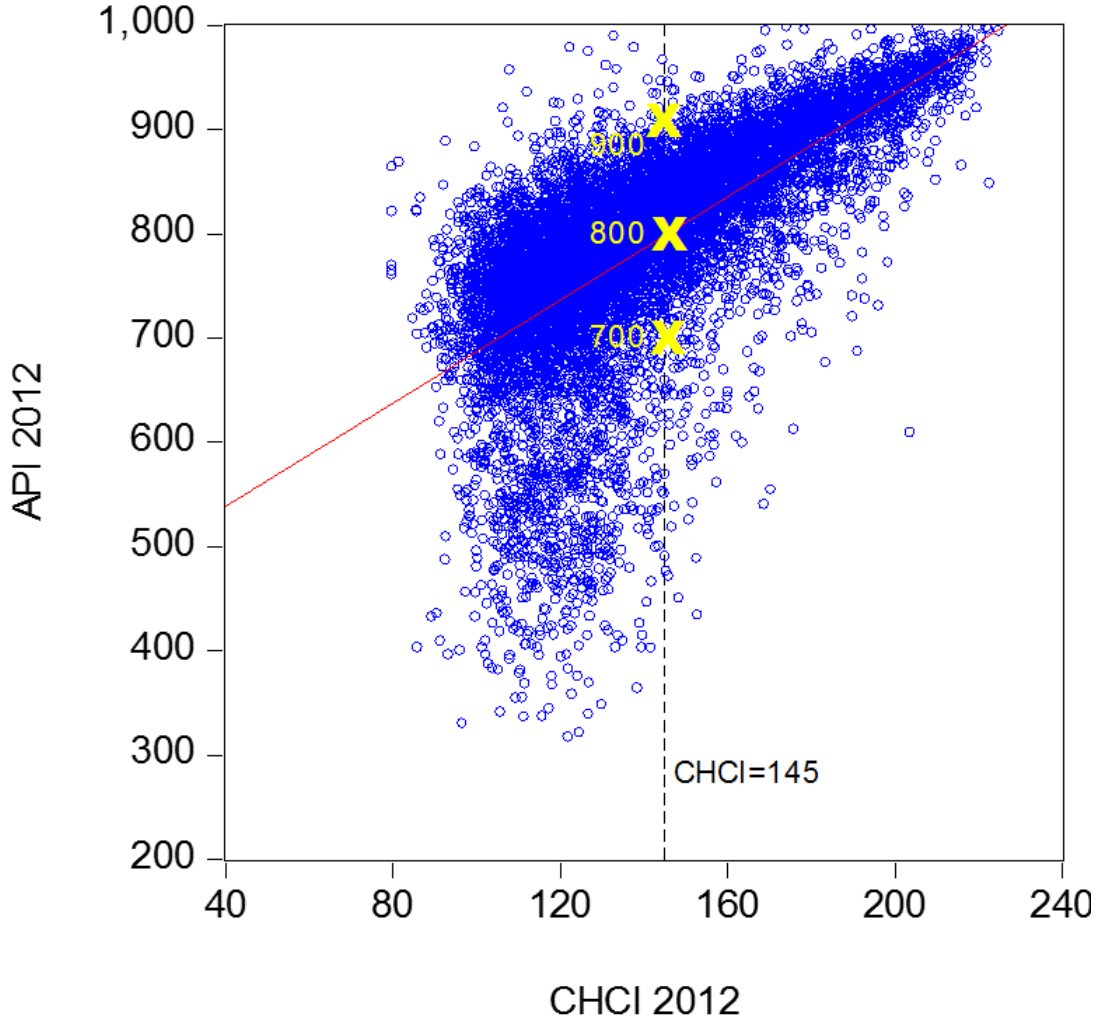
Case-Shiller home price index



A pattern with a good story

2012 CHCI (City Human Capital Index) and

API (Academic Performance Index)
for 10,900 public schools in California



Data: http://www.anderson.ucla.edu/centers/ucla-anderson-forecast/projects-and-partnerships/city-human-capital-index

Machine learning models

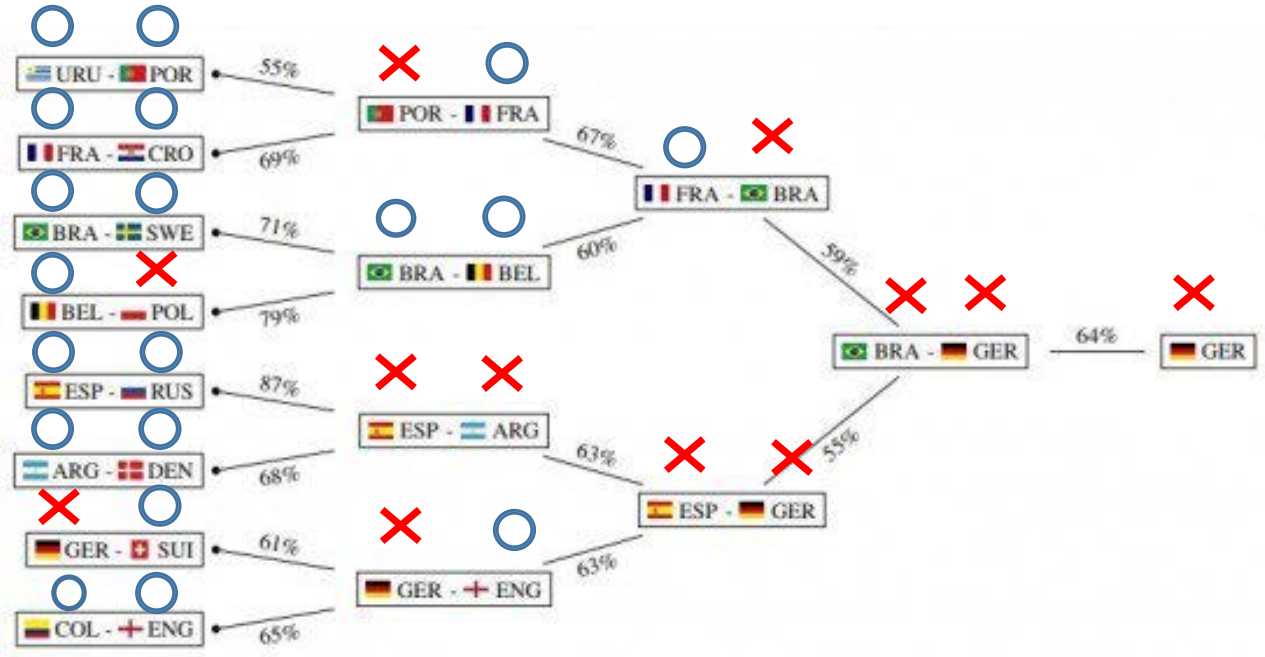
- Supervised learning:
  - Linear regression
  - Logistic regression
  - Linear discriminant analysis
  - Classification & regression trees
  - Naïve Bayes
  - K-nearest neighbors (KNN)
  - Support vector machines
  - Bagging and random forests
  - Artificial Neural Network (ANN)
- Unsupervised learning:
  - Principal components analysis (PCA)
  - K-means clustering

Q1. Did machine learning predict World Cup winner?
No! It didn't.

https://www.technologyreview.com/s/611397/machine-learning-predicts-world-cup-winner/

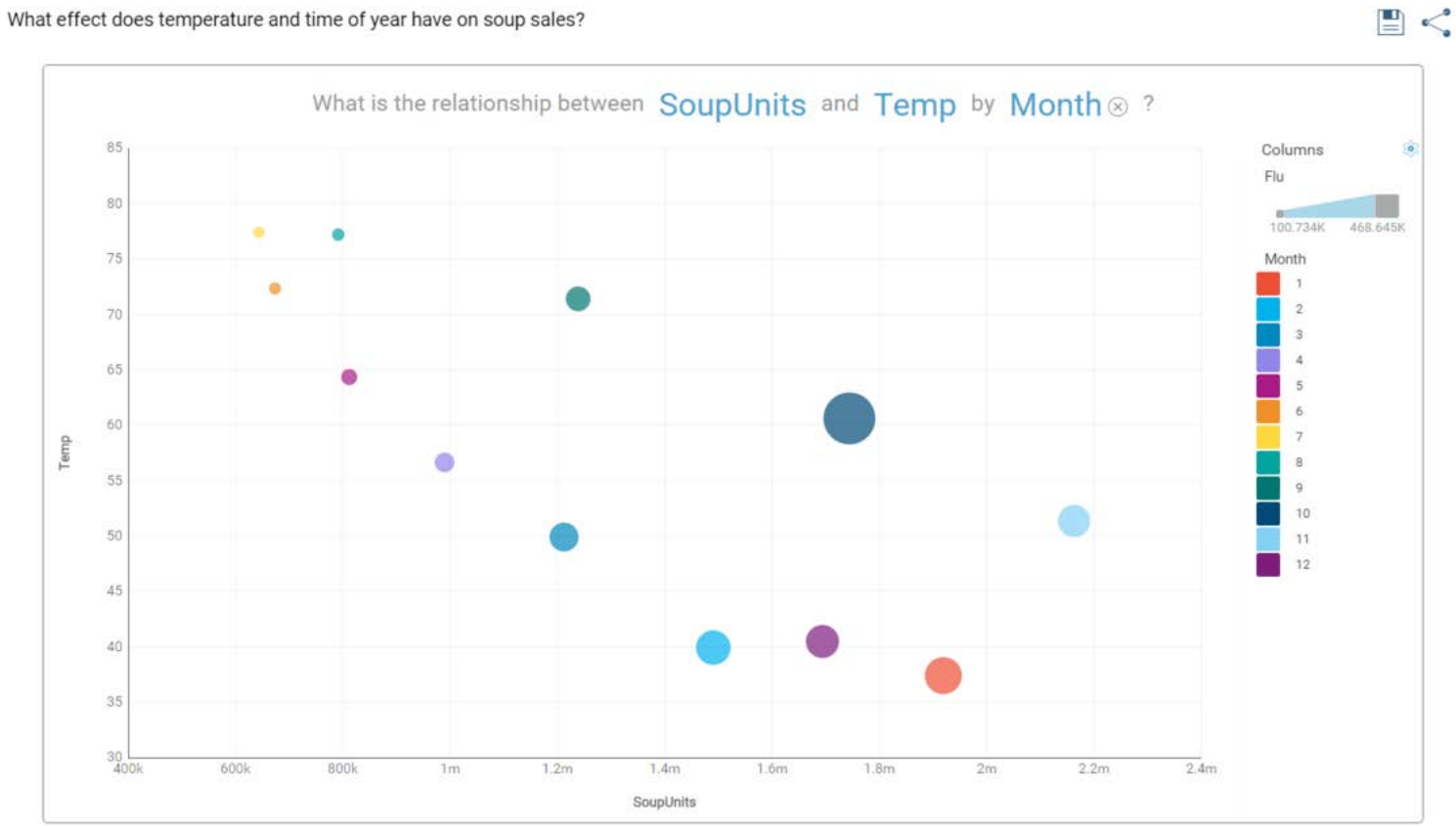
Why not?

Q2. Can machine learning predict stock prices tomorrow?



Case #1: Why did IBM buy The Weather Company in 2015?

- 195,000 personal weather stations.
- How minor changes to weather, such as temperature, might affect consumers buying behavior, helping retailers to adjust their supply chains and inventory management.
- [Watson Analytics]



Case #2: What predicts a long-term relationship?

http://www.wsj.com/articles/be-my-financially-compatible-valentine-1455286835

The research (Credit scores and committed relationships) uses the address and credit information from Equifax.

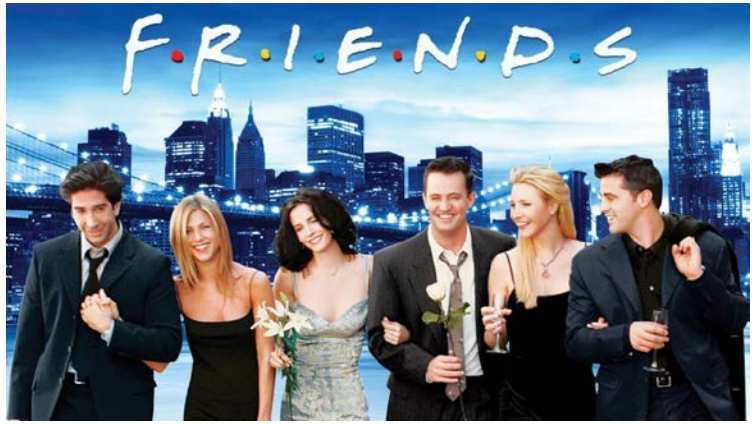
http://www.federalreserve.gov/econresdata/feds/2015/files/2015081pap.pdf

Table 5: Credit Score Levels and Spousal Relationship Formation and Dissolutions

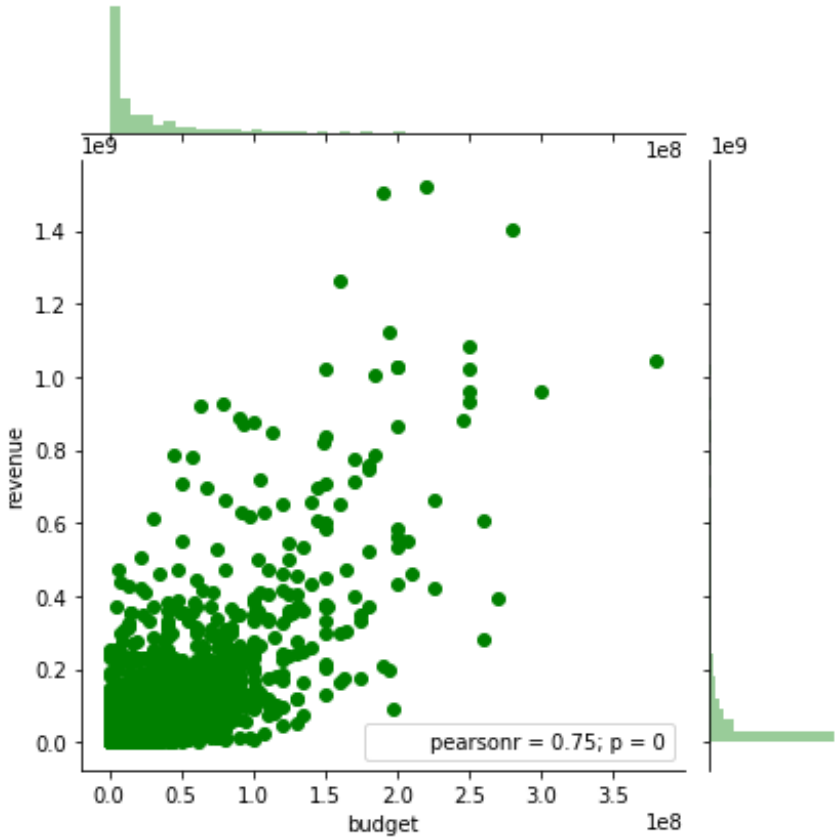
	Including non-primary sample individuals		Primary sample individuals only		
	Relationship Formation	Relationship Dissolution first 6 years	Relationship Dissolution		
			2nd year	3rd or 4th year	5th or 6th year
	(1)	(2)	(3)	(4)	(5)
Score/100	0.125***	-0.371***	-0.339***	-0.491***	-0.438***
	(0.001)	(0.002)	(0.017)	(0.020)	(0.029)
	[1.144]	[0.677]	[0.729]	[0.630]	[0.668]
White	0.320***	-0.690***	-0.334***	-0.397***	-0.242**
	(0.006)	(0.006)	(0.067)	(0.078)	(0.117)
	[1.377]	[0.502]	[0.917]	[0.904]	[0.943]
College	-0.180***	2.534***	3.002***	2.002***	1.119***
	(0.009)	(0.019)	(0.164)	(0.204)	(0.312)
	[0.967]	[1.323]	[1.393]	[1.246]	[1.131]
Log(income)	0.196***	-0.543***	-0.496***	-0.438***	-0.240***
	(0.004)	(0.005)	(0.046)	(0.056)	(0.082)
	[1.105]	[0.781]	[0.802]	[1.033]	[0.986]
Age gap		0.079***	0.088***	0.096***	0.062***
		(0.001)	(0.004)	(0.005)	(0.008)
		[1.082]	[1.331]	[1.354]	[1.205]
Community div. pop. share		2.817***	2.711***	1.944***	1.834***
		(0.037)	(0.333)	(0.409)	(0.613)
		[1.136]	[1.149]	[1.089]	[1.082]
Char. at the time of matching					
White gap			0.356***	0.190**	0.351**
			(0.079)	(0.097)	(0.147)
			[1.068]	[1.035]	[1.064]
College gap			0.659***	0.246*	-0.132
			(0.115)	(0.142)	(0.219)
			[1.092]	[1.033]	[0.983]
Log(income) gap			0.012	0.073**	-0.031
			(0.032)	(0.036)	(0.065)
			[1.005]	[1.033]	[0.986]
Control for					
Age poly.	Yes	Yes	Yes	Yes	Yes
Yearly FE	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes
N	11,400,150	1,872,801	41,685	29,188	20,518

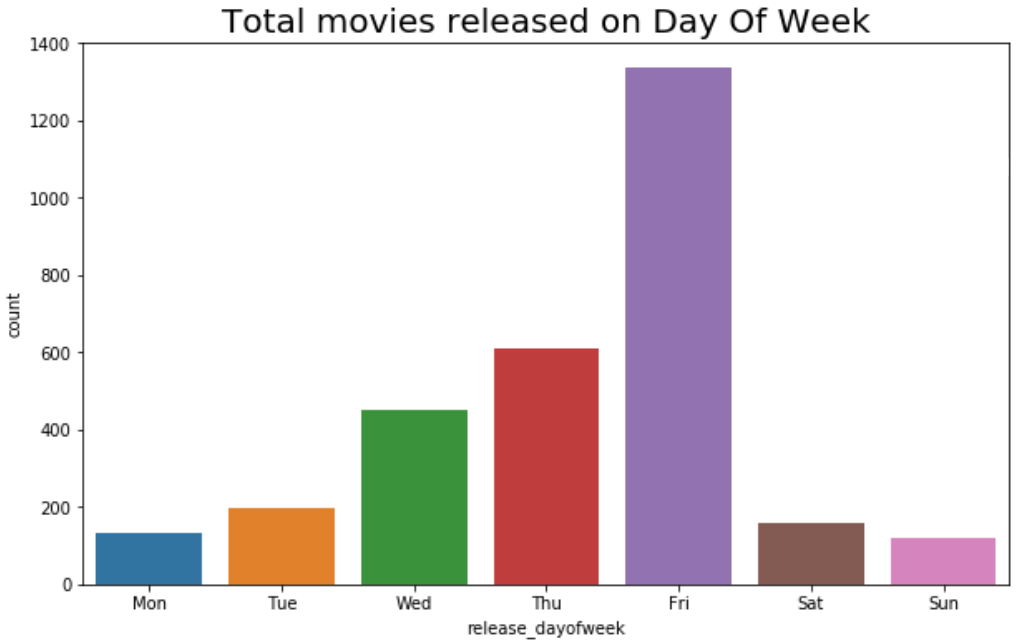
Note. Standard errors are reported in parentheses. Odds ratios of the logistic regressions are reported in brackets. ** denotes the estimate is statistically significant at the 95-percent level, and *** denotes the estimate is statistically significant at the 99-percent level. Odds ratios are

Case #3: Why some movies/TVs are more popular?



Blockbuster Formulas -- Predicting Box Office Revenues





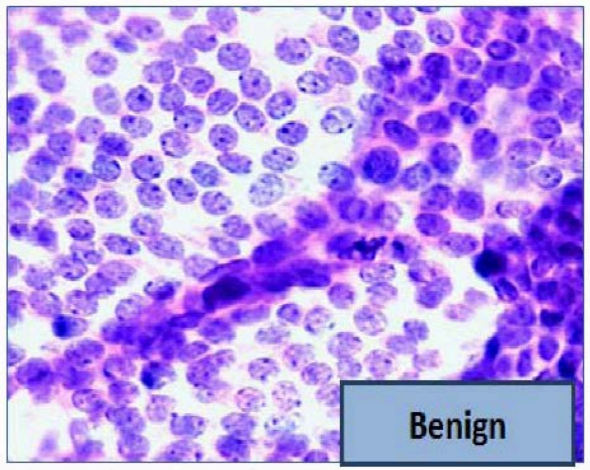
Predictors: Budget, popularity,
genres: action (+), adventure (+), comedy (+),
family (+), documentary (-), drama (-), foreign (-)

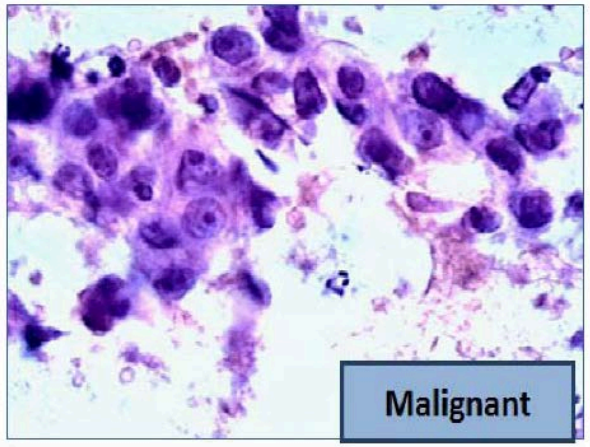
Case #4: Facebook is mastering targeting ads

How Facebook ads helped elect Trump?
https://www.cbsnews.com/news/how-facebook-ads-helped-elect-trump/

- "Infrastructure…so I started making ads that showed the bridge crumbling…that's micro targeting…I can find the 1,500 people in one town that care about infrastructure. Now, that might be a voter that normally votes Democrat," he says. Parscale says the campaign would average 50-60,000 different ad versions every day, some days peaking at 100,000 separate iterations – changing design, colors, backgrounds and words – all in an effort to refine ads and engage users.

Case #5: Breast Cancer Tumor Classification: A Glimpse at the Intersection of Machine Learning and Healthcare by Alec Hon, 11/2018

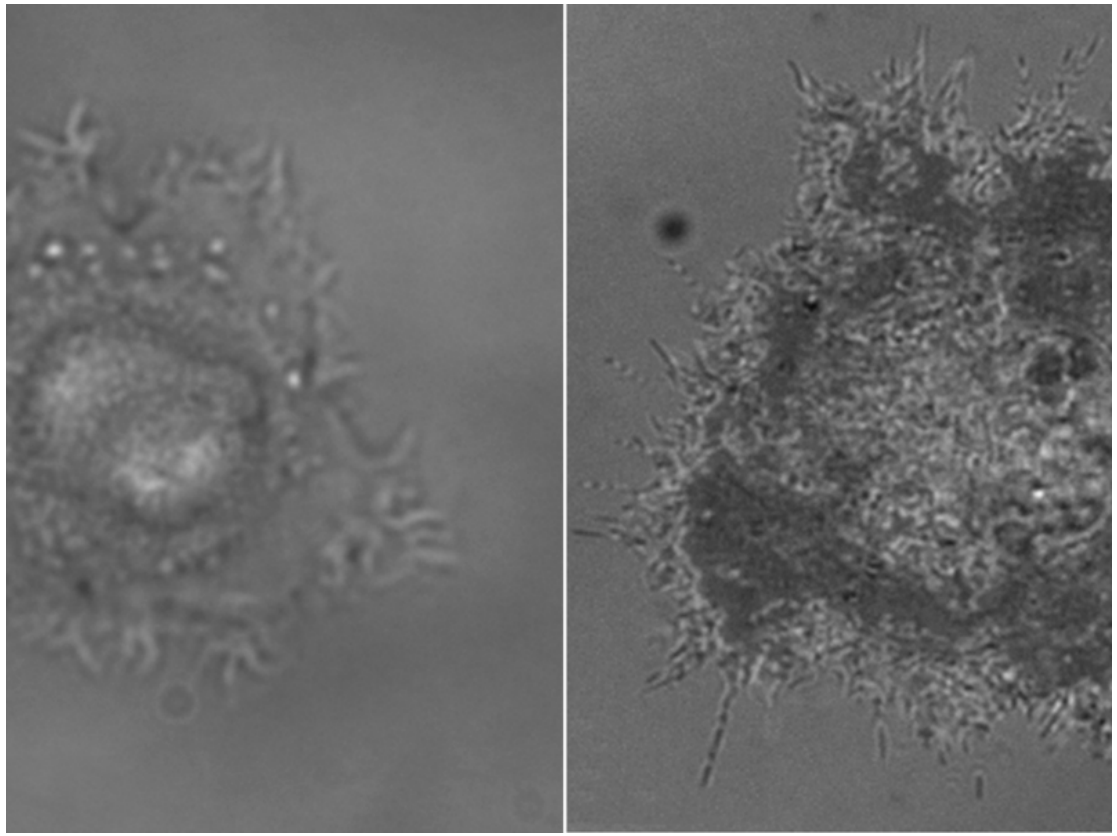




Source: https://www.semanticscholar.org/paper/Classifying-breast-cancer-types-based-on-fine-data-Ahmad-Yusoff/60e88562a14a03e7c9be580f965180f4c9b83d08

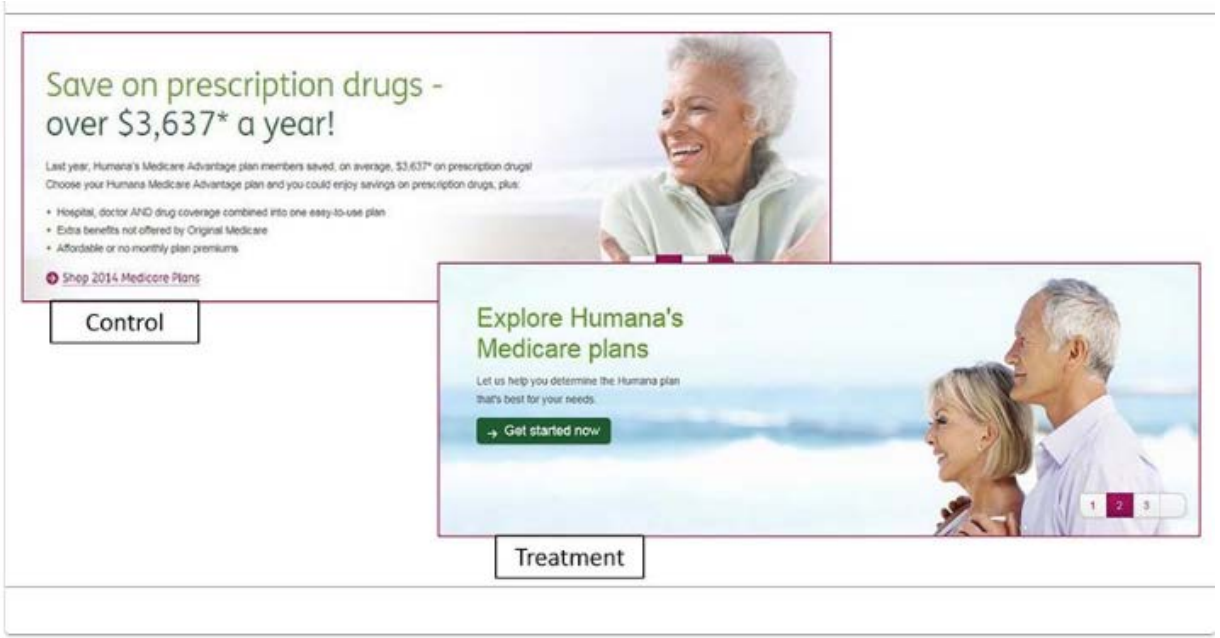
Healthy Cell

Malignant Tumor



https://medicalxpress.com/news/2013-12-malignant-healthy-cells-characteristic-fractal.html

Case #6: A/B testing at scale



A simpler design plus a stronger CTA led to 433% more clickthroughs.

An efficient bandit algorithm for real-time multivariate optimization

Assignments

- Download R, RStudio, and Tableau.
- Start to read these two textbooks: JWHT and Lantz.
- Browse these books for review of Statistics, R, Python.
  - "Fundamental Statistics," by Bryan Burnham https://sites.google.com/site/fundamentalstatistics/unit-1-the-basics
  - "Using R for Introductory Statistics," by John Verzani.
  - "Introduction to Statistical Thought," by Michael Lavine.
  - "The Art of R Programming," by Norman Matloff.
  - "Python Crash Course," by Eric Matthes.
- Read those introductory sources for R.
- Download Anaconda distribution.
- Set up a Github account (optional).