

Data Science using Python

Manaranjan Pradhan

A Data Enthusiast

LinkedIn: <https://www.linkedin.com/in/manaranjanpradhan>

He writes blogs at www.awesomestats.in

Complete the following steps...

- Create a working directory on your desktop / laptop, where you can store all data and the programs of this workshop
- Download datasets from the link below and store in the above directory
 - https://github.com/manaranjanp/Python_DS_2days
 - Download the complete repository as zip file and unarchive it
- Install Anaconda Distribution of Python on your desktop / laptop
 - <https://www.continuum.io/downloads>
- Bookmark the following web site. Lab guides are available on this site. (We will access online during hands on sessions).
 - <http://www.awesomestats.in/python>

H1N1 Pandemic

عربى 中文 English Français Русский Español

 World Health Organization



[Home](#) [Health topics](#) [Data](#) **Media centre** [Publications](#) [Countries](#) [Programmes](#) [Governance](#) [About WHO](#) [Search](#)

Media centre

World now at the start of 2009 influenza pandemic

Dr Margaret Chan
Director-General of the World Health Organization
Statement to the press by WHO Director-General Dr Margaret Chan
11 June 2009

Ladies and gentlemen,

In late April, WHO announced the emergence of a novel influenza A virus.

This particular H1N1 strain has not circulated previously in humans. The virus is entirely new.

The virus is contagious, spreading easily from one person to another, across country to another. As of today, nearly 30,000 confirmed cases have been reported in 74 countries.

This is only part of the picture. With few exceptions, countries with large numbers of cases are those with good surveillance and testing procedures in place.

Spread in several countries can no longer be traced to clearly-defined clusters of human-to-human transmission. Further spread is considered inevitable.


TRENDING: Wearable Tech // Archaeology // Military & Spy Tech // Zika virus // OurAmazingPlanet

2009 Swine-Flu Death Toll 10 Times Higher Than Thought

By Bahar Gholipour, Staff Writer | November 26, 2013 05:00pm ET

35
[Share](#)

10
[Tweet](#)



The swine-flu pandemic of 2009 may have killed up to 203,000 people worldwide—10 times higher than the first estimates based on the number of cases confirmed by lab tests, according to a new analysis by an international group of scientists.

What if Flus can be predicted?

CDC Home
CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

A-Z Index A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

H1N1 Flu

Will help

- Allocate resources optimally
- Increase awareness
- Better manage and control spread

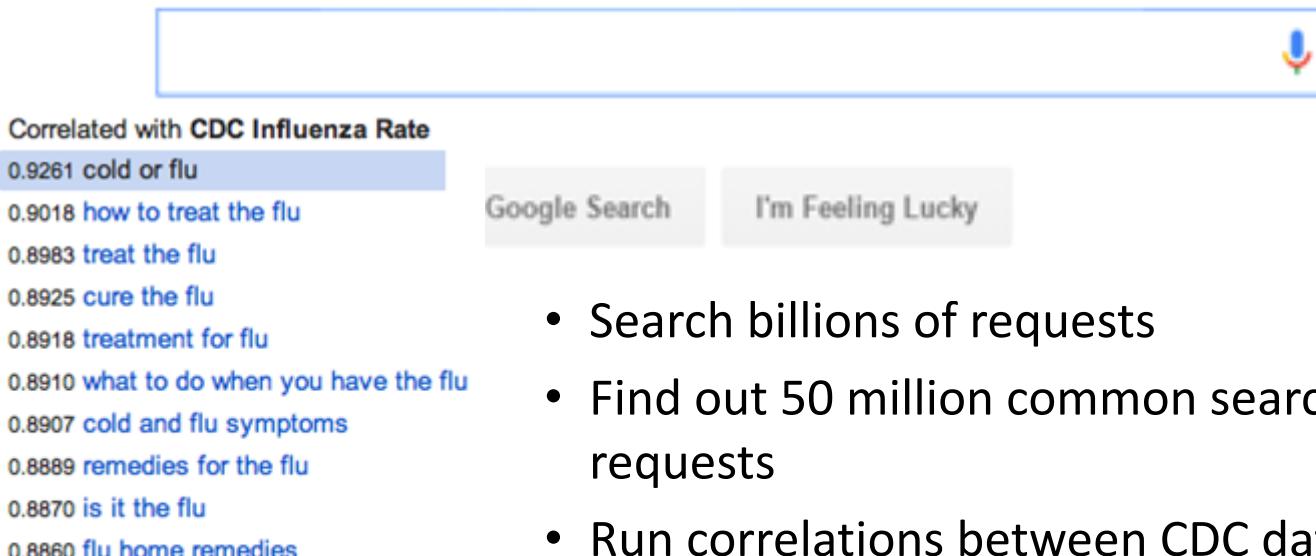
Better Prepared

Flu Vaccines Manufacturers

- Optimize production and distribution



Predicting Flu Trends

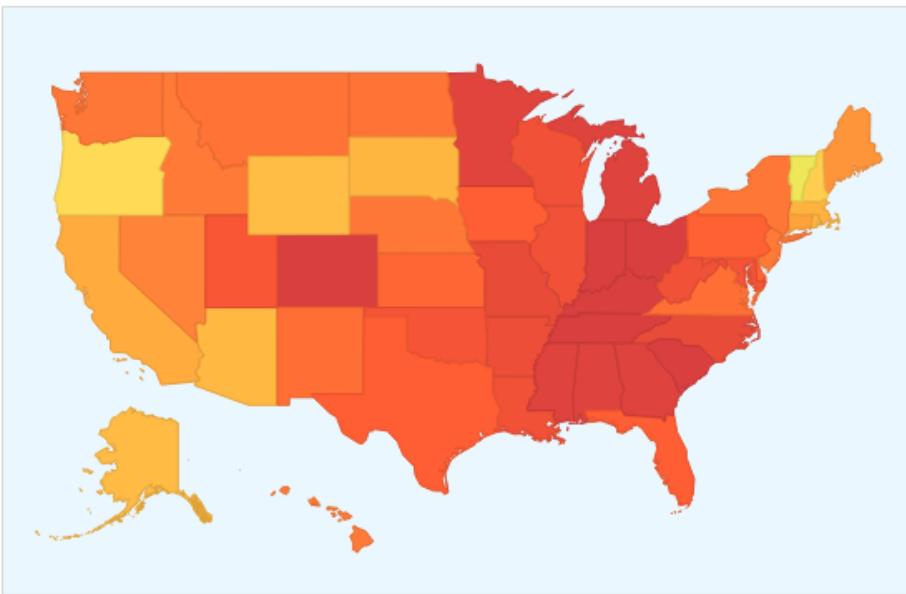


- Search billions of requests
- Find out 50 million common search requests
- Run correlations between CDC data and search terms
- Filter out 50 common terms with high correlations > 0.9

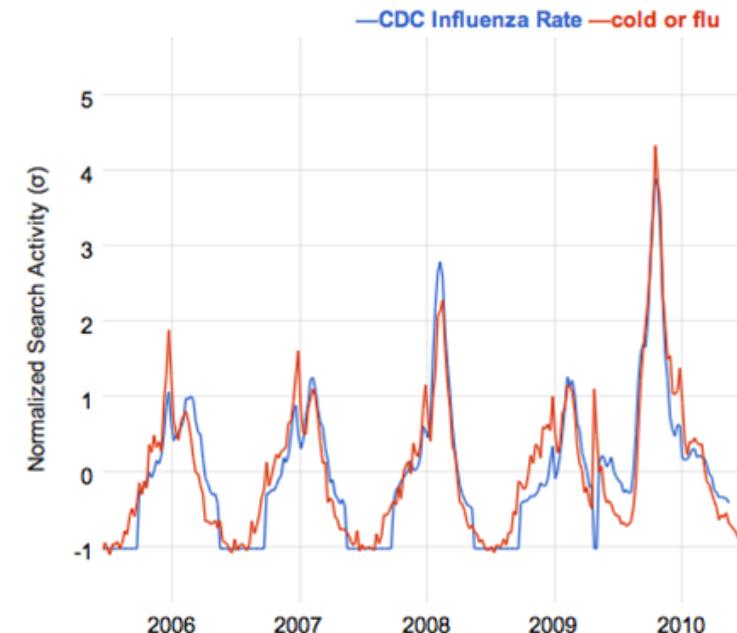
$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon$$

where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, β_0 is the intercept,

Google Flu Trends



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through December 27, 2014.



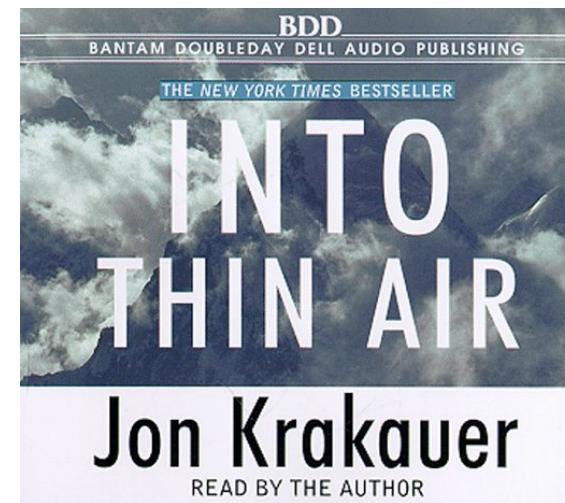
Google Flu Trends predictions were 97% accurate comparing with CDC data

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

Into Thin Air

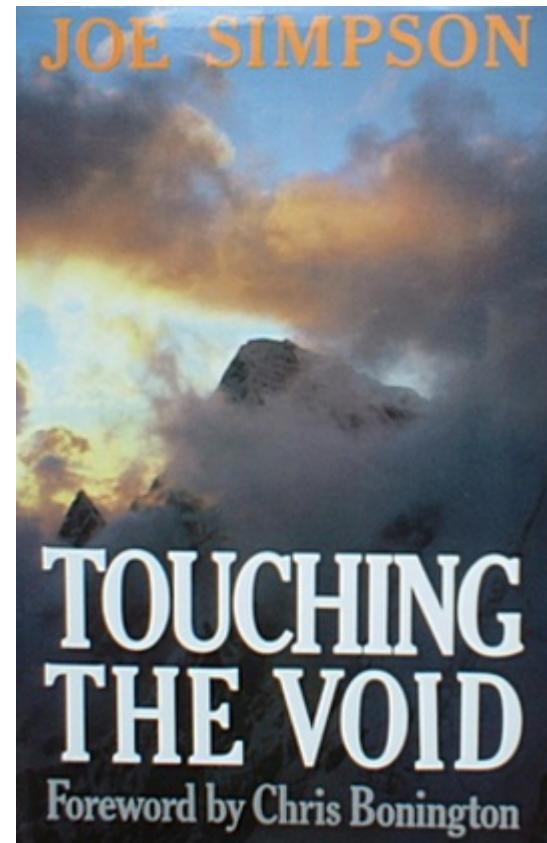


It details the author's presence at [Mount Everest](#) during the [1996 Mount Everest disaster](#), when eight climbers were killed and several others were stranded by a "rogue storm"



Touching the Void

Touching the Void is a 1988 book by [Joe Simpson](#), recounting his and [Simon Yates'](#) successful but disastrous and nearly fatal climb of the 6,344-metre (20,813 foot) [Siula Grande](#) in the [Peruvian Andes](#) in 1985.



Customers who bought this item also bought...

[Look inside](#) ↴

#1 NATIONAL BESTSELLER
A Personal Account of the Mt. Everest Disaster

INTO THIN AIR

"Ranks among the great adventure books of all time." — *THE WALL STREET JOURNAL*

Jon Krakauer
AUTHOR OF *INTO THE WILD* AND *TIGER DREAMS*

Paperback – October 19, 1999
by Jon Krakauer (Author, Photographer), Randy Rackliff (Illustrator), Daniel Rembert (Contributor), & 2 more
★★★★★ 2,414 customer reviews
#1 Best Seller in Mountain Climbing

See all 65 formats and editions

Kindle \$3.40	Hardcover \$18.66	Paperback \$10.36	Mass Market Paperback from \$0.01
------------------	----------------------	----------------------	--------------------------------------

Read with our [free app](#)

667 Used from \$0.01
82 New from \$4.49
51 Collectible from \$6.37

483 Used from \$0.01
117 New from \$6.22
11 Collectible from \$9.70

487 Used from \$0.01
18 New from \$4.59
19 Collectible from \$3.00

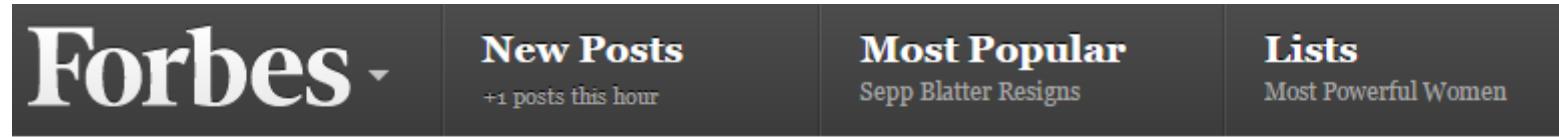
National Bestseller

A bank of clouds was assembling on the not-so-distant horizon, but journalist-mountaineer Jon

Customers Who Bought This Item Also Bought

 LOOK INSIDE!	 LOOK INSIDE!	 LOOK INSIDE!	 LOOK INSIDE!	 LOOK INSIDE!
Into the Wild » Jon Krakauer ★★★★★ 2,304 #1 Best Seller in Travelogues & Travel Essays Paperback \$7.34	Under the Banner of... » Jon Krakauer ★★★★★ 1,361 Paperback \$10.03 Get it by Tomorrow	Missoula: Rape and the... Jon Krakauer ★★★★★ 361 Hardcover \$18.09 Get it by Tomorrow	Touching the Void: The... » Joe Simpson ★★★★★ 315 Paperback \$11.22 Get it by Tomorrow	Buried in the Sky: The... Peter Zuckerman and Amanda Paddan ★★★★★ 225 Paperback \$10.63 Get it by Tomorrow

What customers bought together



4/06/1998 @ 12:00AM



Diaper-beer syndrome

IT'S PART OF the folklore of data processing. A retail chain put all its checkout-counter data into a giant digital warehouse and set the disk drives spinning.

Out popped a most unexpected correlation: sales of diapers and beer.

Evidently, young fathers would make a late-night run to the store to pick up Pampers and get some Bud Light while they were there.

Capitalizing on the discovery, the store placed the disparate items together. Sales zoomed.

Recommendations are key to personalization

Amazon's recommendation secret

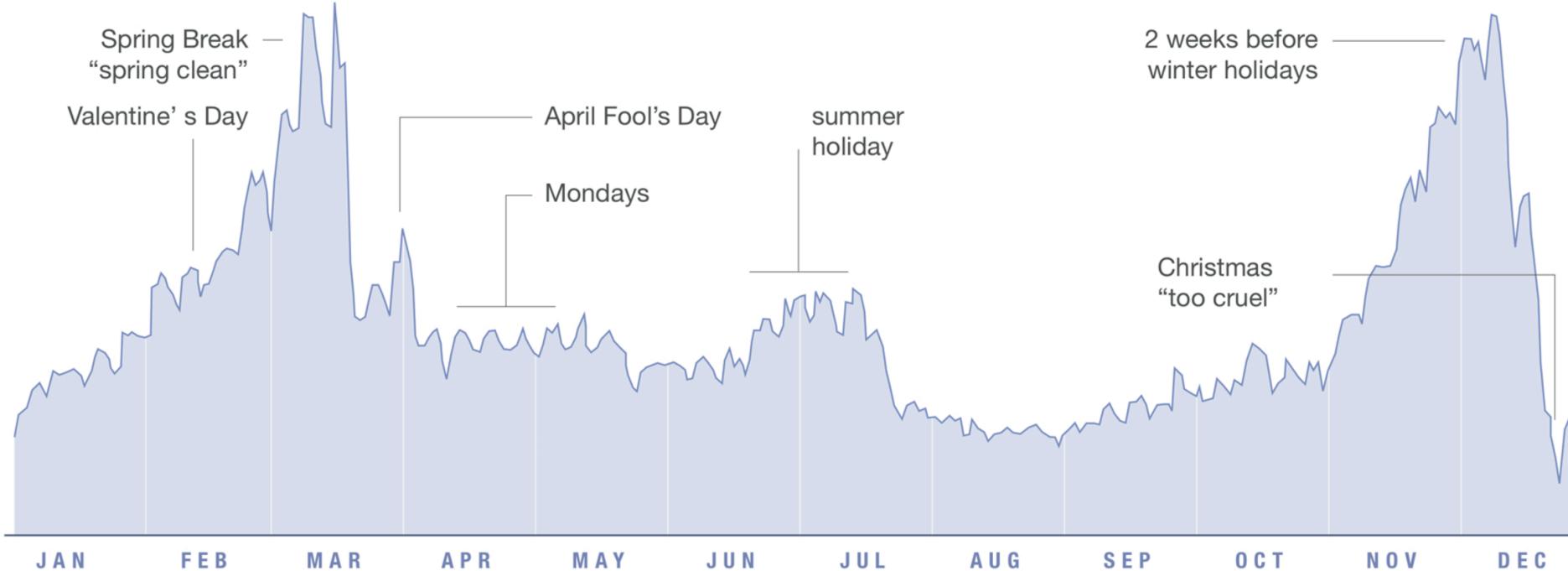
Judging by Amazon's success, the recommendation system works. The company reported a 29% sales increase to \$12.83 billion during its second fiscal quarter, up from \$9.9 billion during the same time last year. A lot of that growth arguably has to do with the way Amazon has integrated recommendations into nearly every part of the purchasing process from product discovery to checkout.

<http://fortune.com/2012/07/30/amazons-recommendation-secret/>

Facebook relationship breakups...

Peak Break-Up Times

According to Facebook status updates



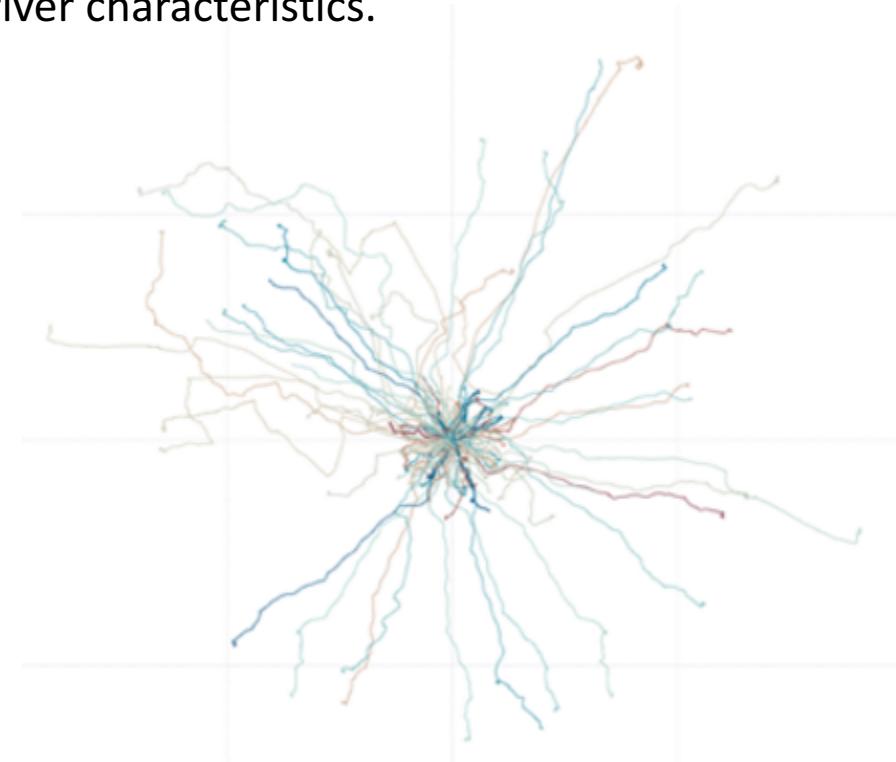
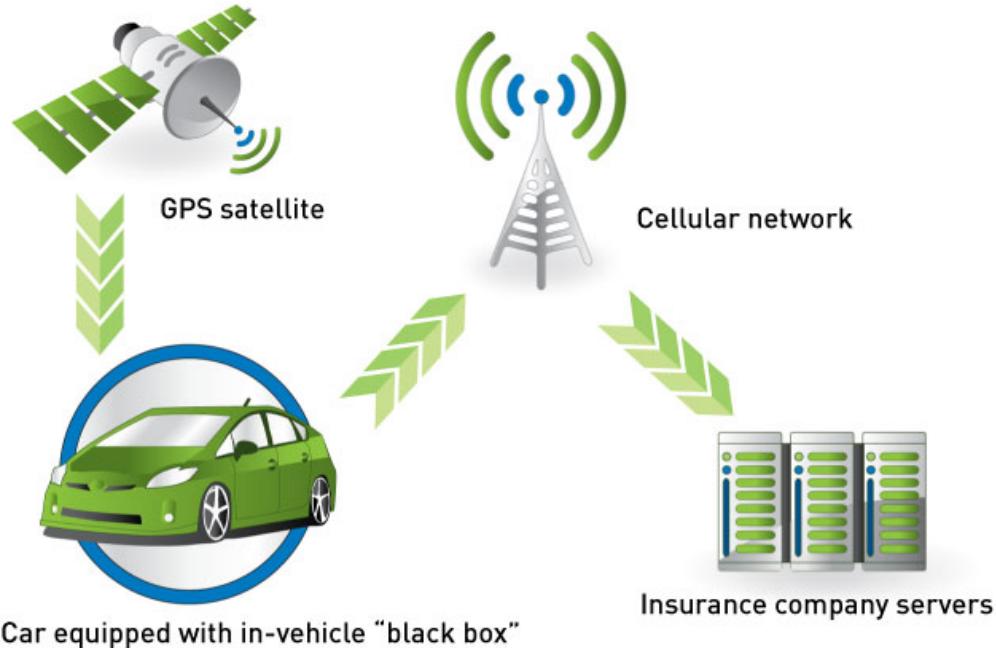
Knowing when breakup happens can help business...

- More traffic to online dating sites after breakup surge
- Increase in demand for relationship lawyers and counselors.
- Higher demand for housing as they move out
- Brand switching (people change the brand of beer after breakup)

With emergence of new technologies and
big data explosion, can we do more or
better analytics.

Better Risk Assessment

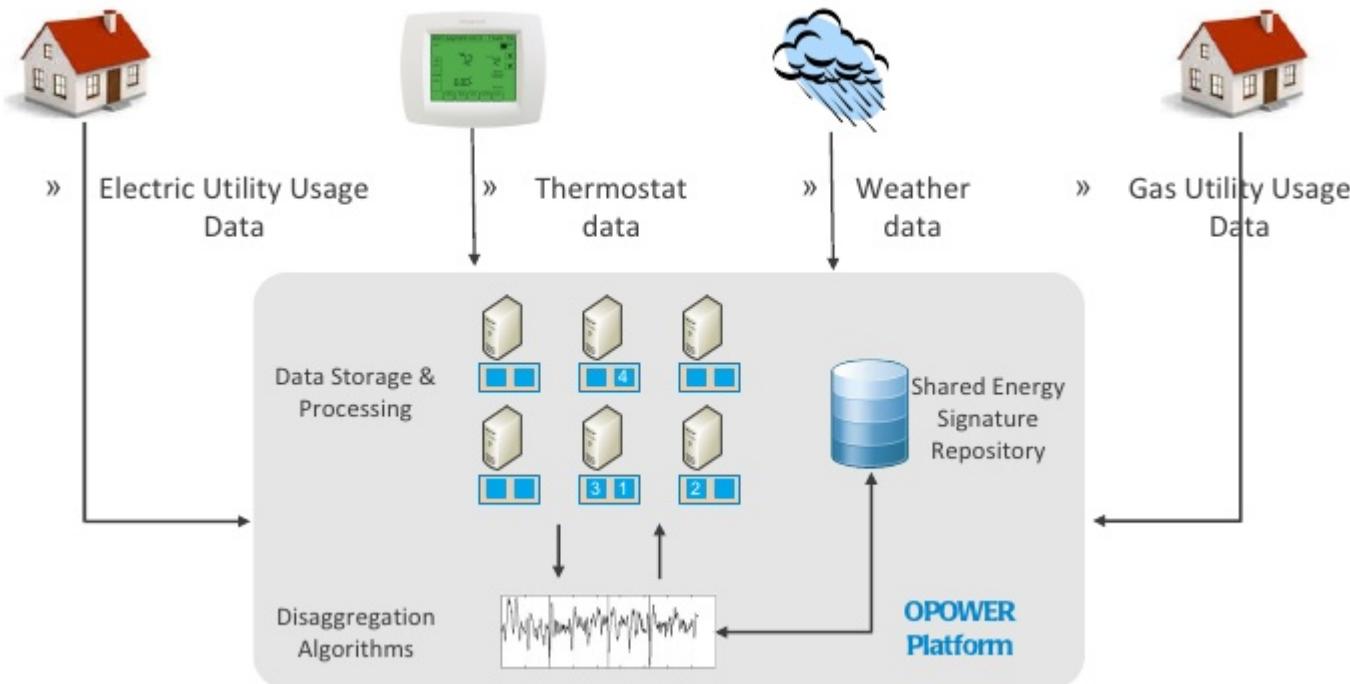
Auto insurance industries using telematics data to profile driver characteristics.



For automobile insurers, telematics represents a growing and valuable way to quantify driver risk. Instead of pricing decisions on vehicle and driver characteristics, telematics gives the opportunity to measure the quantity and quality of a driver's behavior.

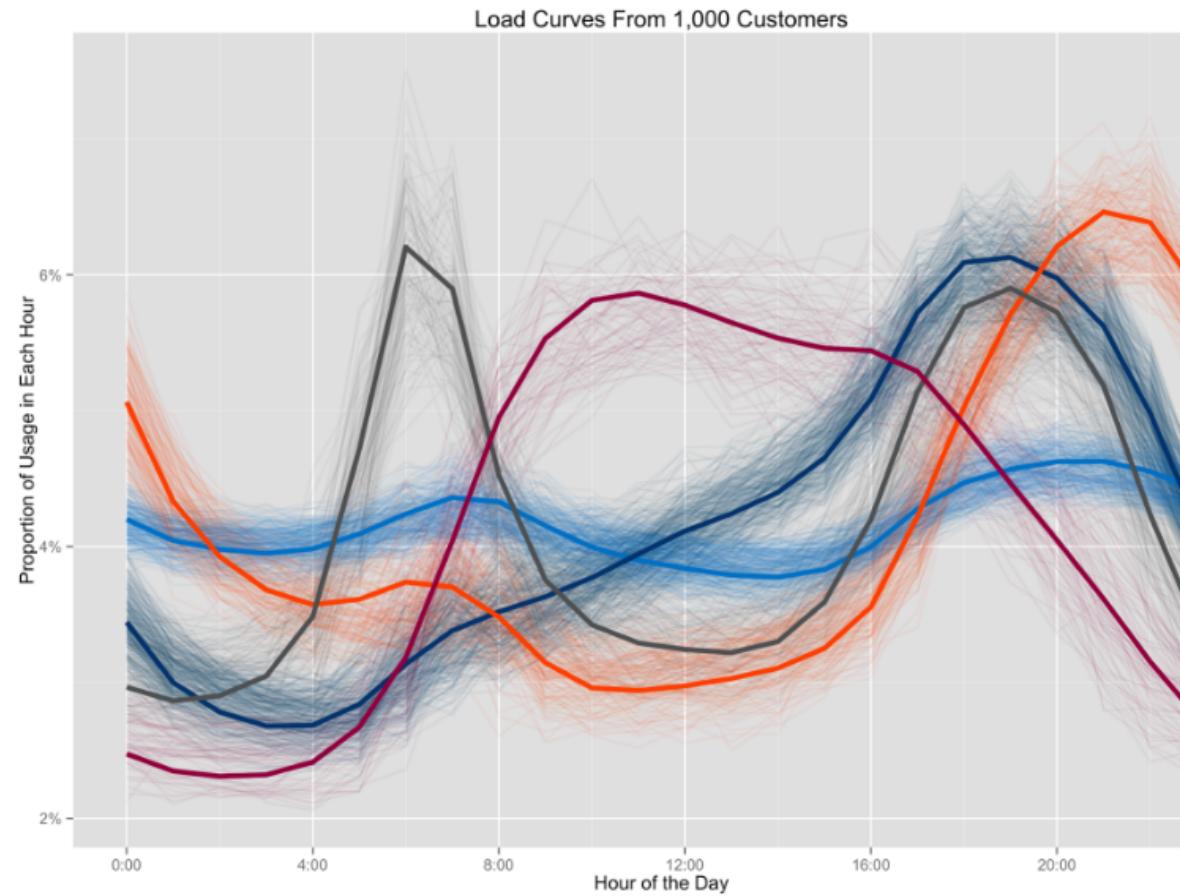
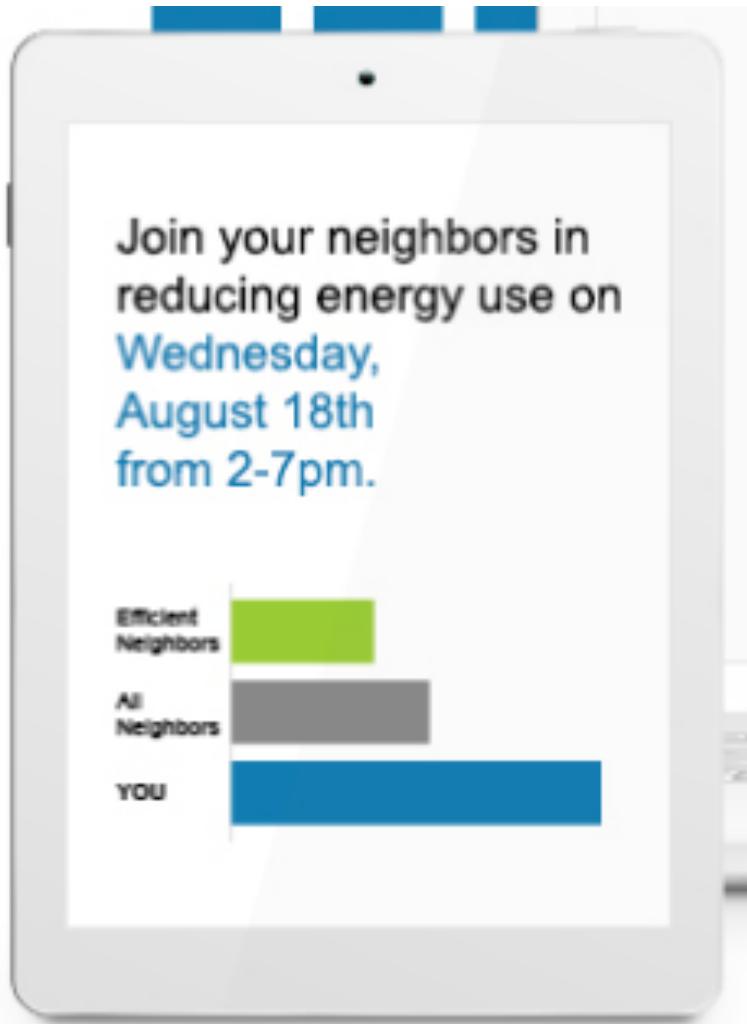
<https://www.kaggle.com/c/axa-driver-telematics-analysis>

Monitoring Energy Consumption



OPower collects and analyses more than **100 billion meter readings per year**

Customer Profiling

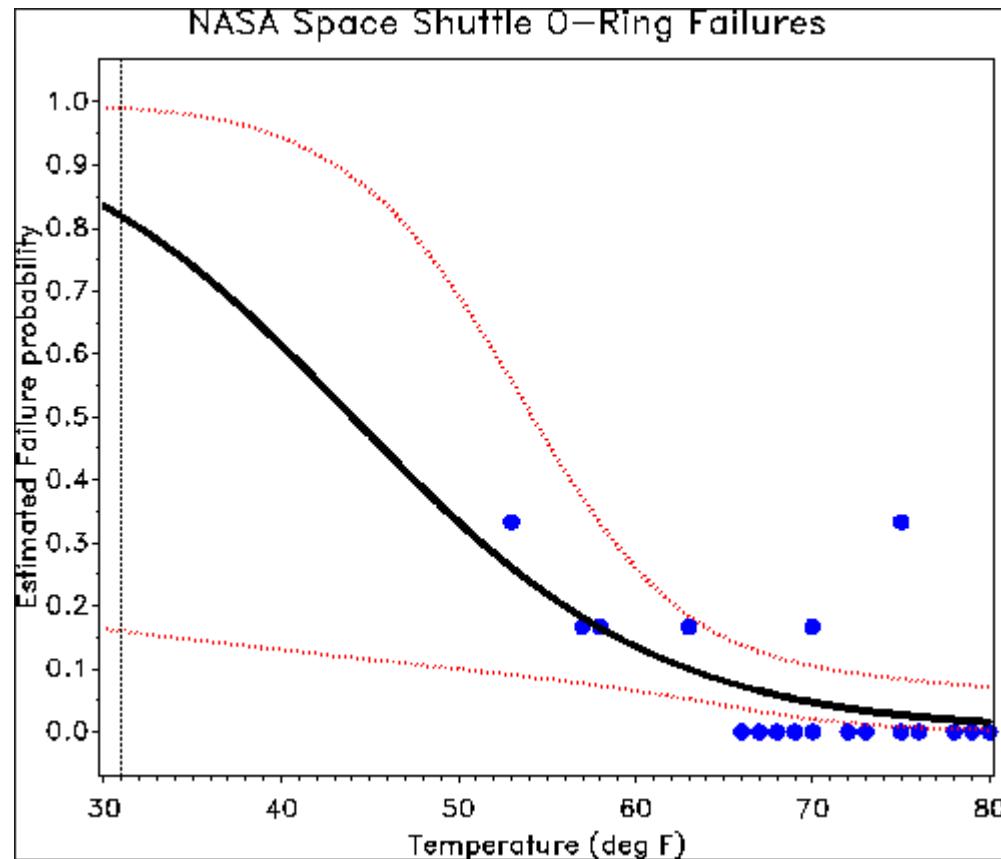


Source: Opower (2014)

Applications across industries

Cost of not using data and analytics

NASA Could have predicted Challenger Disaster



Challenger launched on its 10th mission on Jan. 28, 1986.
See more at: <http://www.space.com/10677-challenger-tragedy-overview.html#sthash.sR63YmVz.dpuf>





**“Whoever unlocks the reams
of data and uses it
strategically will win.”**

Angela Ahrendts, CEO of Burberry

Data Scientists will be the sexiest job of 21st century
Harvard Business Review 2012

Analytics Stages

Stages of Analytics

Understanding what happened and why happened by exploring past data.

Descriptive

Product sales patterns or factors influencing product sales.

Learning from past data and predicting what may happen in future and likelihood of happening in future.

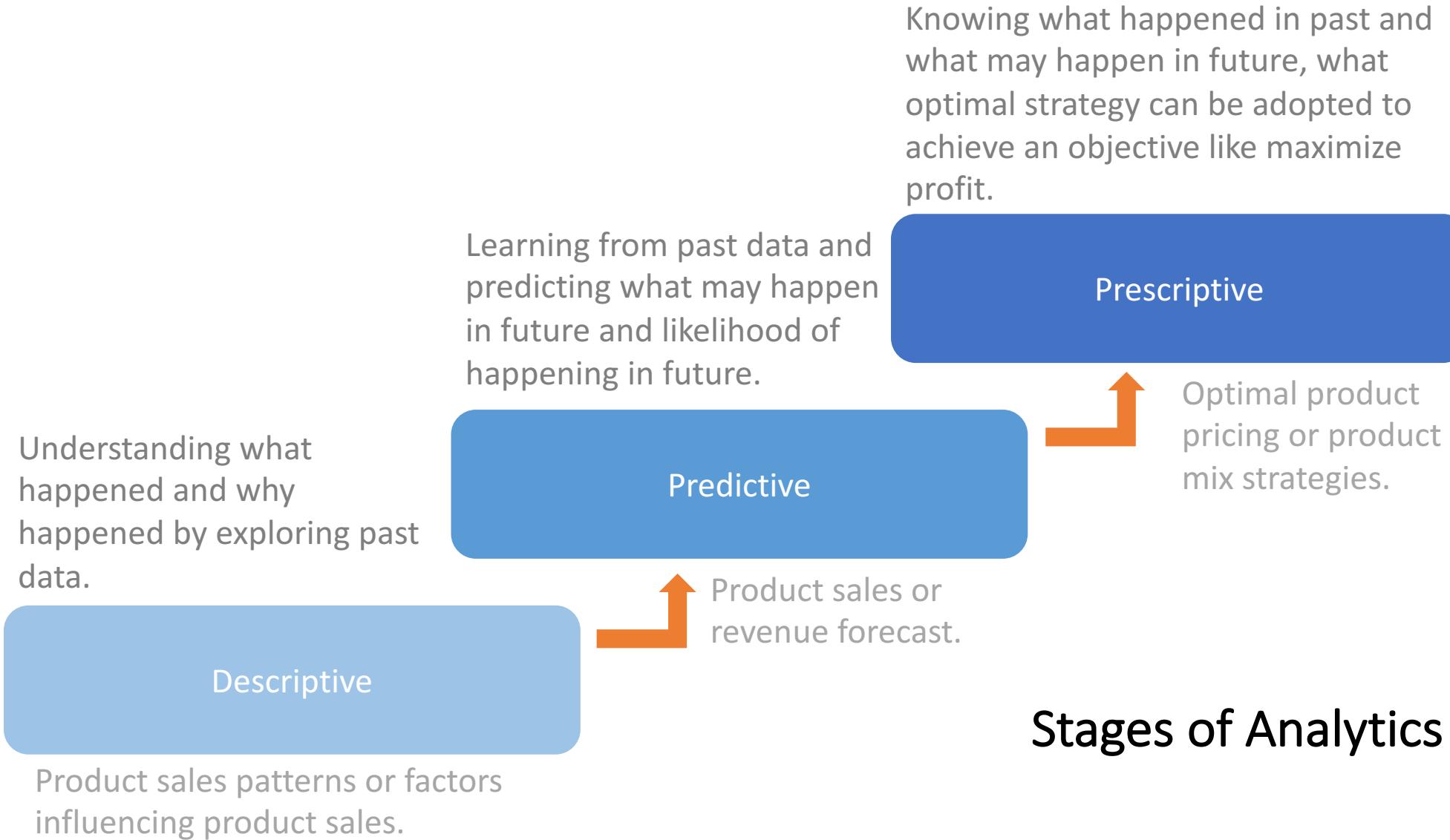
Predictive

Product sales or revenue forecast.

Prescriptive

Knowing what happened in past and what may happen in future, what optimal strategy can be adopted to achieve an objective like maximize profit.

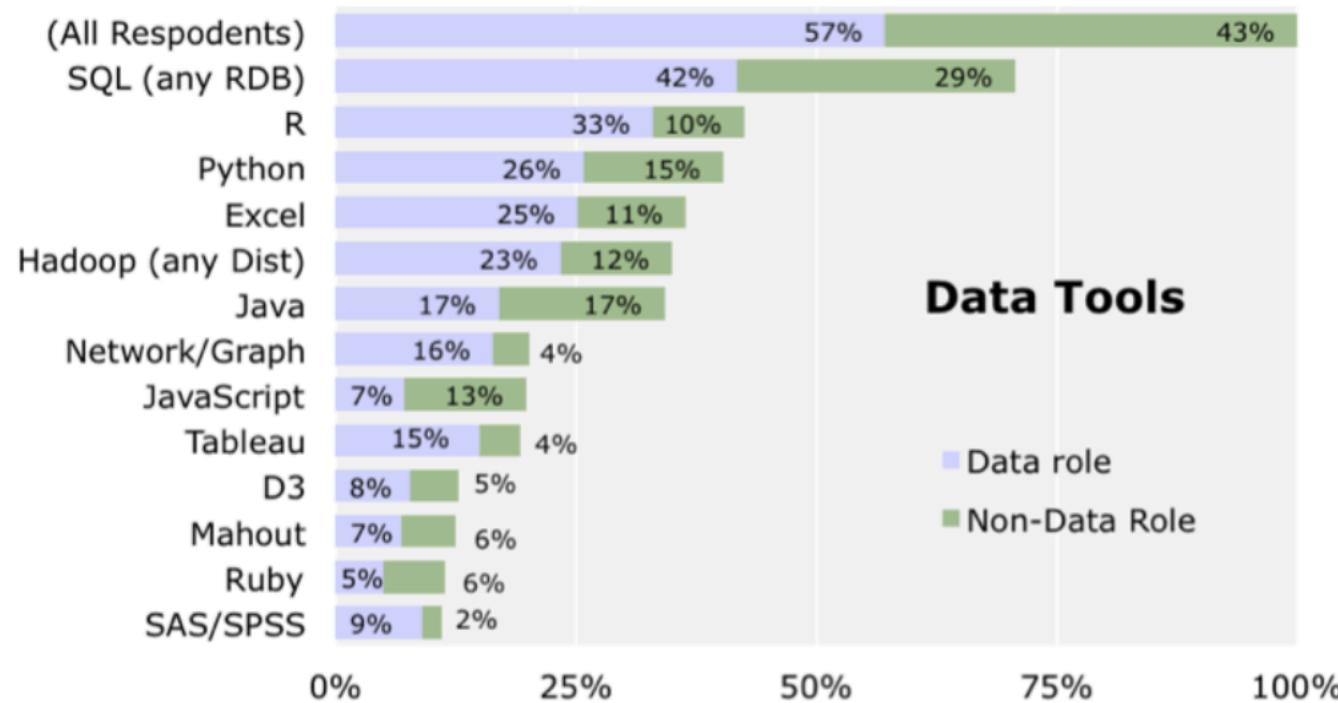
Optimal product pricing or product mix strategies.



Stages of Analytics

What Tools are available?

Python Stack For Data Science@



<http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html>

Python

- Multi-purpose
 - Web Developments
 - Scripting
 - Server Side Developments
 - Statistical Learnings & Machine Learnings
- Object Oriented
- Interpreted
- Strongly typed and Dynamically typed
- Focus on readability and productivity



R Vs. Python

R	Python
Built for Statistical Analysis.	General Purpose Language. Main objective is productivity and readability.
Primarily used in academics and research. Enterprise have started adopting it for analysis.	Has a very strong presence in enterprises for large number of software developments. Easier adoption in enterprises as strong development experience already exists.
Integration with other enterprise systems are not straightforward.	Integration with other enterprise systems or applications are easier.

Python Stack For Data Science

Efficient storage of arrays and matrices. Backbone of all scientific calculations and algorithms.



Library for scientific computing. Linear algebra, statistical computations, optimization algorithm.



seaborn



Plotting and visualization

High-performance, easy-to-use data structures for data manipulation and analysis. Pandas provide the features of dataframe, which is very popular in the area of analytics for data munging, cleaning & transformation.



IDE or Development environment for data analysis in python.

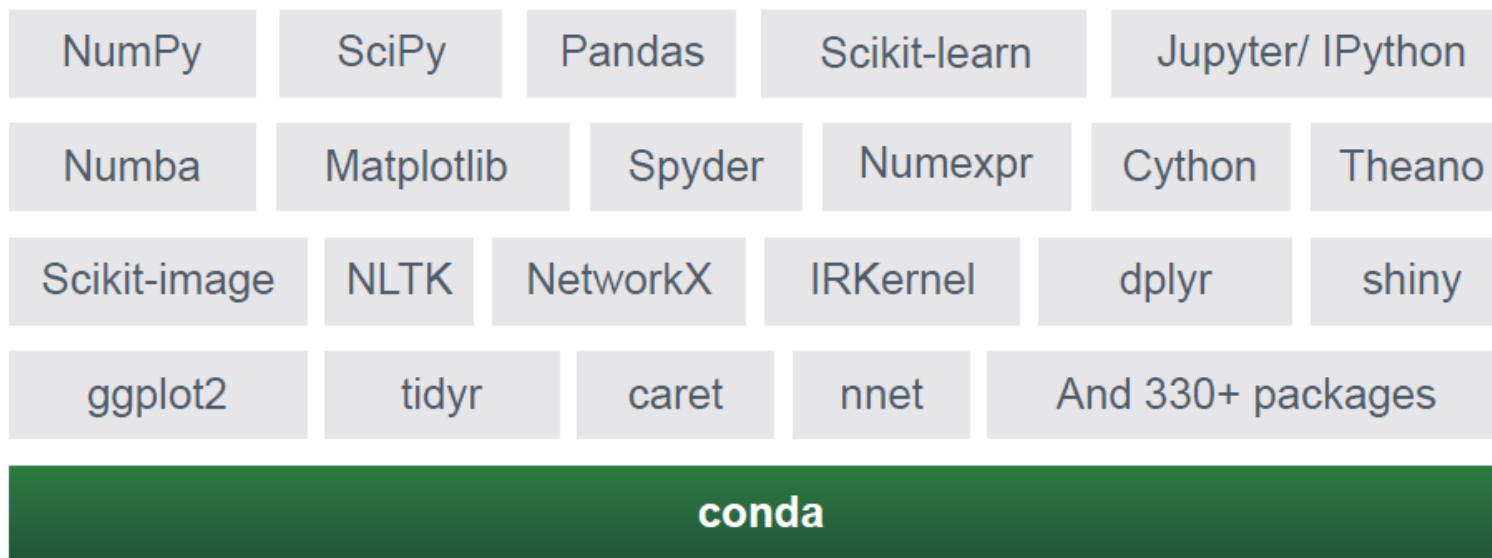


Machine learning library. Collection of ML algorithms.

Python Distribution



ANACONDA®



**Game-Changing
Enterprise Ready
Python Distribution**

- 2 million downloads in last 2 years
- 200k / month and growing
- conda package manager serves up 5 *million* packages per month
- Recommended installer for IPython/Jupyter, Pandas, SciPy, Scikit-learn, etc.

Source: Continuum Analytics

Download link: <https://www.continuum.io/downloads>

What we will learn in this workshop?

Exploratory	Descriptive	Predictive
<ul style="list-style-type: none">••Data Loading••Data Cleanup••Data Transformations••Basic Statistics	<ul style="list-style-type: none">••Visualization••Understanding Distributions••Hypothesis Testing	<ul style="list-style-type: none">••Regression••Classification••Clustering••Model Evaluation••Prediction

DESCRIPTIVE

Data Loading
and
Manipulation

Data Cleaning &
Transformation

Visualization

Understanding
Distributions &
Hypothesis Tests

PREDICTIVE

Regression

Classification

Clustering

Model
Evaluation &
Prediction

Start Jupyter notebook

- For MAC
 - Click on Anaconda Navigator and click on “launch notebook”
 - Or go to command prompt and enter
 - **jupyter notebook --ip=***
- For Windows
 - Go to command prompt and enter
 - **jupyter notebook --ip=***

Start a jupyter notebook



The screenshot shows the Jupyter Notebook interface. At the top, there are tabs for "Files", "Running", "Clusters", and "Conda". Below the tabs, a message says "Select items to perform actions on them." On the left, there is a file browser with a sidebar containing icons for home, refresh, and search. The main area lists several directories: "anaconda", "Applications", "Desktop", "Documents", "Downloads", and "metastore_db". On the right, there is a "New" button with a dropdown menu. The dropdown menu includes options: "Text File", "Folder", "Terminal", "Notebooks", "Python [conda root]" (which is highlighted), "Python [default]", and "Spark 2.1.0". There are also "Upload" and "New" buttons at the top of the dropdown menu.

Click on new to start new notebook. For every hands on exercise, start a new notebook.

Let's Start

NumPy

- Library for mathematical and numerical routines like Matlab
- Provides basic routines
 - Manipulating large arrays and matrices of numeric data.
- Foundational library for all statistical and machine learnings
 - Pandas and SciPy
- Using NumPy library

import numpy as np

Pandas

- Recent API based on Numpy, Optimized for performance
- Easy to work with messy and irregularly indexed data
- Adopts concepts of R language dataframes
- The two basics structures of pandas
 - Series 1d array
 - DataFrame 2d array
- Typical Data Munging Activities
 - Filtering, selecting data
 - Aggregating, transforming data
 - Joining, concatenating, merging data
 - Descriptive basics statistics

Pandas

columns		id	country	isOver	amount
index		▼	▼	▼	▼
a	►	P255	Afg	True	300000
b	►	P31256	Fr	False	22354
c	►	P2245	Cor	False	12478
d	►	415	Som	False	Nan
e	►	P332	Esp	True	4789123

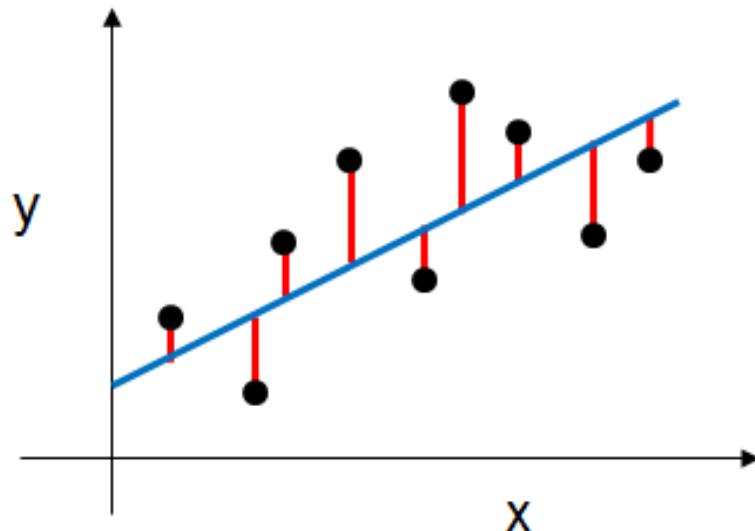
Table like structure

- 2D data structure
- Row and column index
- Size mutable: insert or delete columns
- SQL like transformations – select, groupby, aggregations, filtering, joining etc.

Linear Regression

- Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) - wikipedia
- The dependent variable must be a continuous variable
- The relationship is assumed to be linear
- Is a supervised learning
- Use cases
 - Understand marketing effectiveness
 - Pricing and promotions on sales of a product
 - Evaluate trends and make estimates

Linear Regression



Error of the
Estimated
line is :

$$SS_{residuals} = \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

Model Prediction
↓
Observed Result

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- y is the response
- β_0 is the intercept
- β_1 is the coefficient for X_1 (the first feature)
- β_n is the coefficient for X_n (the nth feature)

The β values are called the **model coefficients**:

Regressions finds the line which minimizes the **sum of squared residuals**. The method is called **OLS** (Ordinary least square)

Linear Regression – Model Evaluation

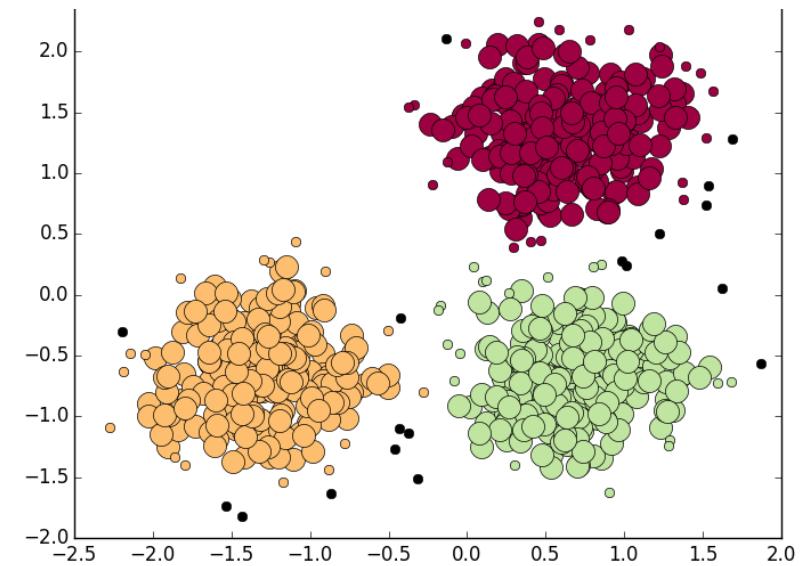
- MAE – Mean Absolute Error
- MSE – Mean Squared Error
 - Penalizes larger residuals
- RMSE – Root Mean Squared Error
- R Square – How much variance explained by the model
 - 1 – all variance explained
 - 0 – no variance explained. It is better not have the model at all.
 - Closer to 1 – is better model.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|.$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

Clustering

- Determine the intrinsic grouping in a set of unlabeled data
 - A *cluster* is a collection of objects which are “similar”
 - Objects belonging to different clusters are dissimilar
- Unsupervised Learning
- *Use Cases*
 - *Marketing*: finding groups of customers with similar behaviour either based on their characteristics or past purchase patterns
 - *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost or attributes
 - Identifying frauds
 - *Biology*: classification of plants and animals given their features;



Classification

- **Classification** is identifying to which of a set of categories a new observation belongs
- The number categories can be two or more
- Is a supervised learning
- Most widely used ML technique
- Use Cases
 - Predict if a customer would churn – churn analysis
 - Predict if the loan applicant would default or not
 - NLP – spam mail or not a spam mail
 - Sentiment Analysis
 - Document classification
 - If a patient has a risk of a disease or not

Classification – Model Evaluation

Confusion Matrix

	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

		Predicted Class	
		Positive	Negative
True Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

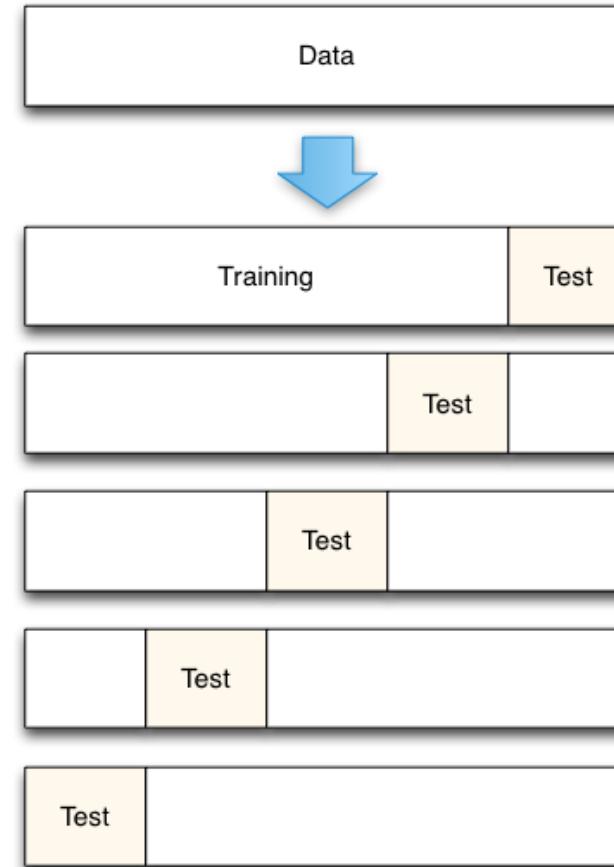
Sensitivity: $TP/(TP + FN)$

Precision: $TP/(TP + FP)$

Specificity: $TN/(TN + FP)$

Accuracy: $(TP + TN)/(TP + TN + FP + FN)$

Train - Test Split & Cross Validation



Wisdom of the crowd!

Francis Galton

- Galton promoted statistics and invented the concept of correlation.
- In 1906 Galton visited a livestock fair and stumbled upon an intriguing contest.
- An ox was on display, and the villagers were invited to guess the animal's weight.
- Nearly 800 gave it a go and, not surprisingly, not one hit the exact mark: 1,198 pounds.
- Astonishingly, however, the average of those 800 guesses came close - very close indeed. It was 1,197 pounds.

