Data Science using Python

Manaranjan Pradhan

A Data Enthusiast

LinkedIn: https://www.linkedin.com/in/manaranjanpradhan
He writes blogs at www.awesomestats.in

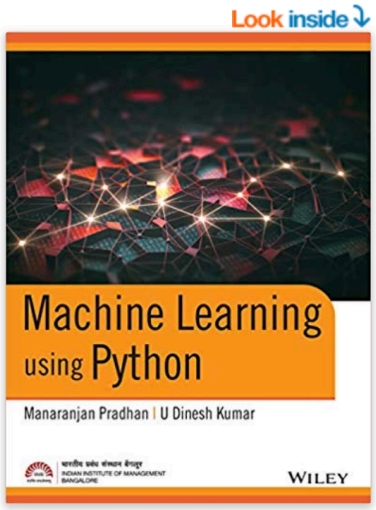
Complete the following steps…

- 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/**ML_Python**
  - Download the complete repository as zip file and unarchive it
- Install latest Anaconda Distribution for Python (3.6+) on your desktop / laptop
  - https://www.anaconda.com/download/

About Me

- Worked for HP and iGate for 12+ years
- Working as freelancer for last 5 years
- Mostly conduct workshops on Spark, Databricks, Machine Learning and Deep Learning
- https://www.linkedin.com/in/manaranjanpradhan

Manaranjan Pradhan (Manu)



Machine Learning using Python Paperback – 2019

by U Dinesh Kumar Manaranjan Pradhan (Author)

★★★★☆ 7 customer reviews

› See all 2 formats and editions

Kindle Edition	Paperback
₹ 423.20	₹ 529.00 prime
Read with Our Free App	2 New from ₹ 529.00

This book is written to provide a strong foundation in machine learning using Python libraries by providing real-life case studies and examples. It covers topics such as foundations of machine learning, introduction to Python, descriptive analytics and predictive analytics. Advanced machine learning concepts such as decision tree learning, random forest, boosting, recommended systems, and text analytics are covered. The book takes a balanced approach between theoretical understanding and practical applications. All the topics include real-world examples and provide step-by-step approach on how to explore, build, evaluate, and optimize machine learning models.

Course Outline

Introduction to Data Science and Setting up data analysis environment	Introduction to Data Science Setting up Python Environment for Data Analysis Overview of Data Analysis Stack - Numpy, Pandas, Matplotlib, scipy and Scikit-learn
Accessing and preparing data with Pandas	Loading data from Different Sources Data manipulation - Filtering, Grouping, Ordering, Joining Dealing with missing Data
Data Exploration, Visualizations & Statistical Analysis	Histograms, Bar charts Density Plots, Box Plots, Scatter Plots, Heat Maps Understanding Basic Statistics, Distributions, Correlations
Algorithms for Regression and Classification Problems	Understanding loss function and gradient descent approach for loss minimization Linear Regression, Logistic Regression, Decision Trees, KNN Model Optimization & Parameter Tuning
Clustering	K-means clustering Finding optimal number of clusters
Model Evaluation	Creating Training, validation and Test Data Sets Cross validations Understanding Evaluation Metrics: RMSE, R-square, ROC, Confusion Matrix, Precision, Recall, Accuracy etc.

H1N1 Pandemic

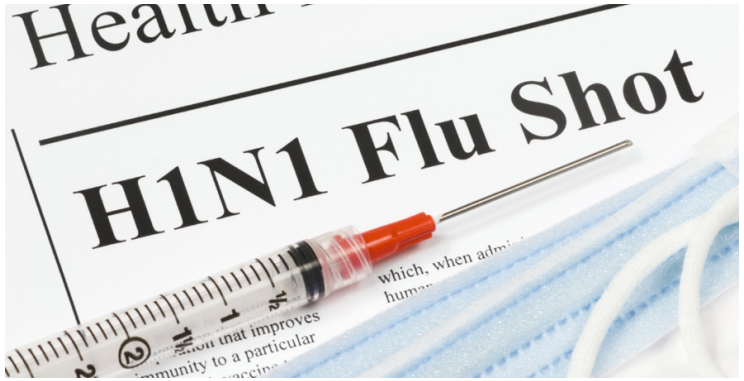
What if Flus can be predicted?

Flu Vaccines Manufacturers

- Optimize production and distribution

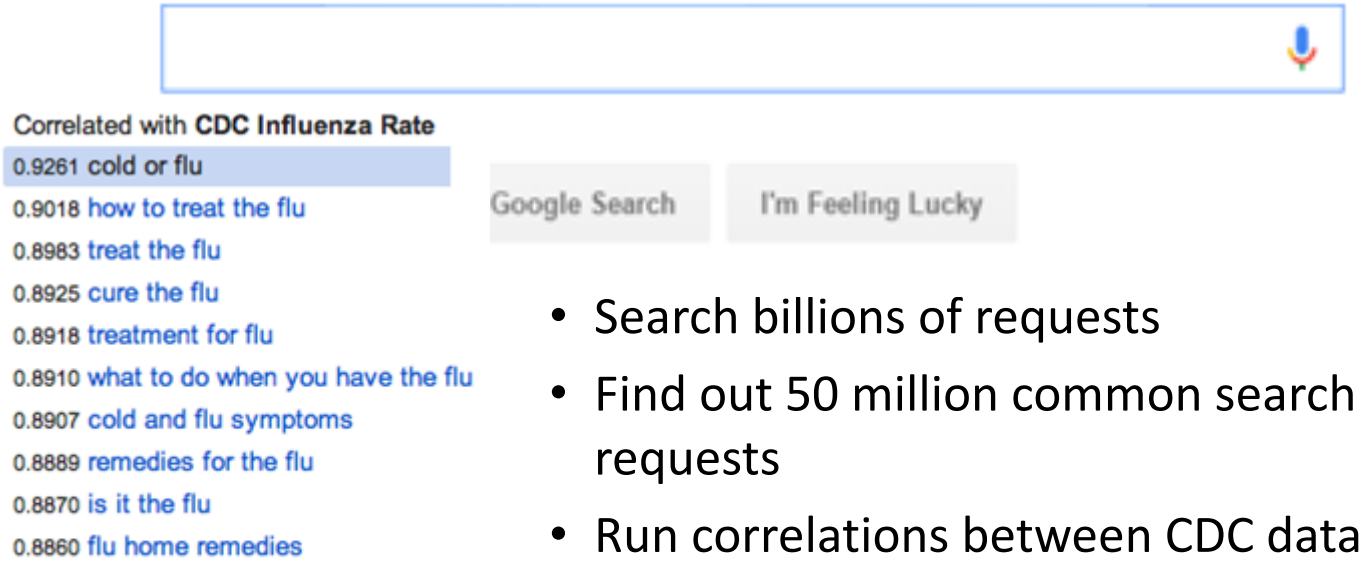
Will help

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



<u>Better Prepared</u>

Predicting Flu Trends

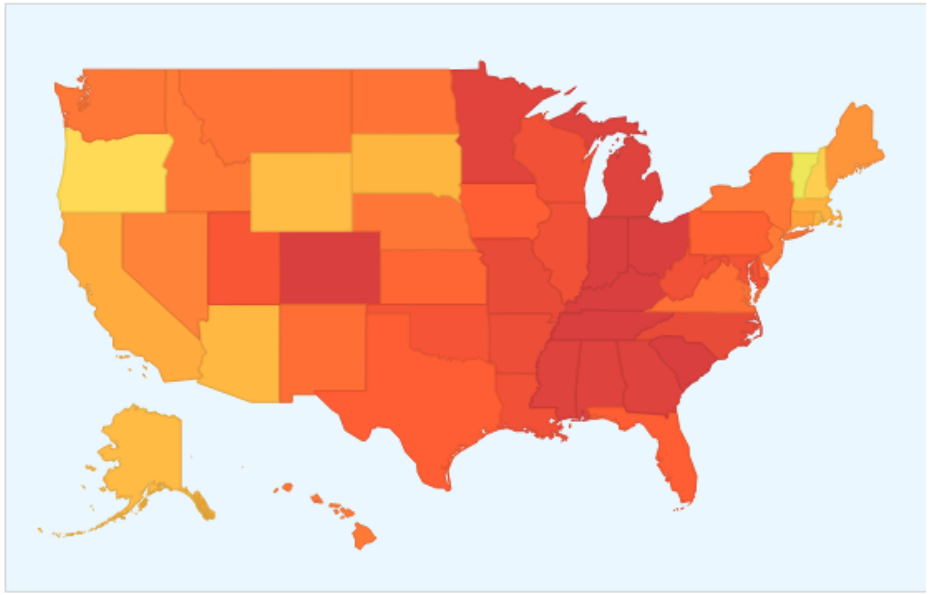


- Search billions of requests
- Find out 50 million common search requests
- Run correlations between CDC data and search terms
- Filter out 50 common terms with high correlations > 0.9

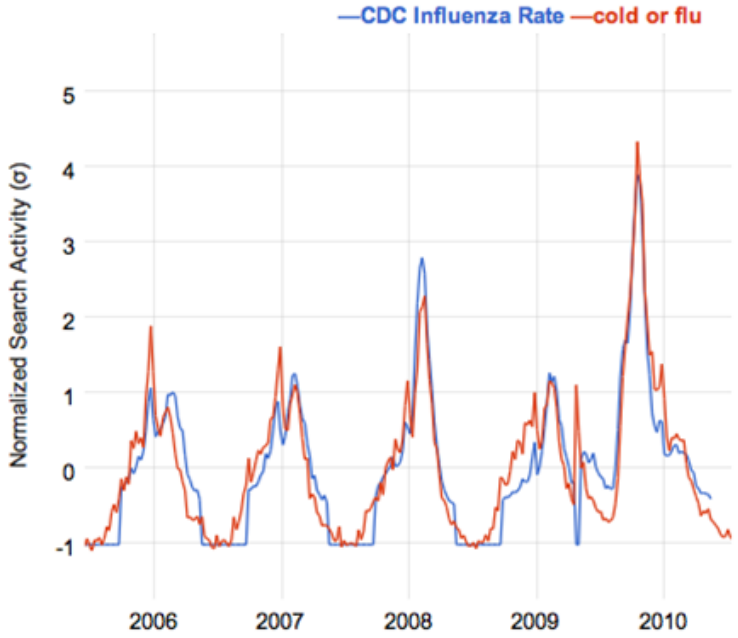
$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

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

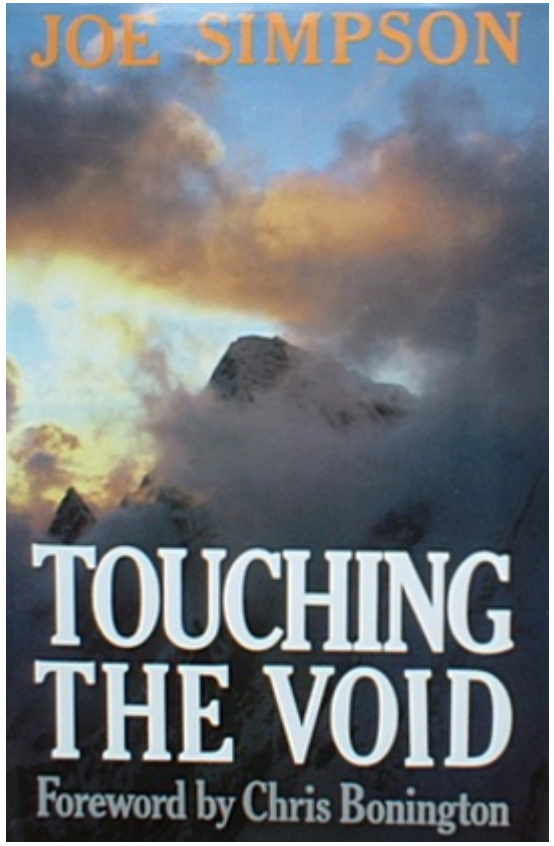
https://en.wikipedia.org/wiki/Google_Flu_Trends

Into Thin Air

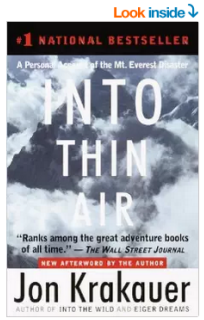
It details the author's presence at Mount Everest during the 1996 Mount Everest disaster, when eight climbers were killed and several others were stranded by a "rogue storm"

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 PeruvianAndes in 1985.



Customers who bought this item also bought…



Into Thin Air: A Personal Account of the Mt. Everest Disaster

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

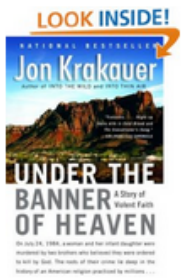
Kindle	Hardcover	Paperback	Mass Market Paperback
$3.40	$18.66	$10.36	from $0.01
Read with our free app	667 Used from $0.01 82 New from $4.49 51 Collectible from $6.37	483 Used from $0.01 117 New from $6.22 11 Collectible from $9.70	487 Used from $0.01 18 New from $4.59 19 Collectible from $3.00

National Bestseller

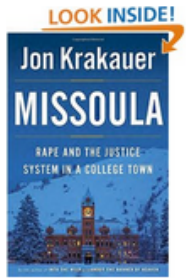
A bank of clouds was assembling on the not-so-distant horizon, but journalist-mountaineer Jon

Customers Who Bought This Item Also Bought

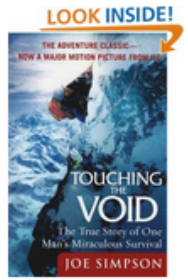
Into the Wild
› Jon Krakauer
2,304
#1 Best Seller in Travelogues & Travel Essays
Paperback
$7.34 Prime



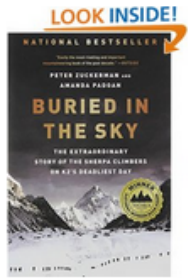
Under the Banner of…
› Jon Krakauer
1,361
Paperback
$10.03 Prime
Get it by Tomorrow



Missoula: Rape and the…
Jon Krakauer
361
Hardcover
$18.09 Prime
Get it by Tomorrow

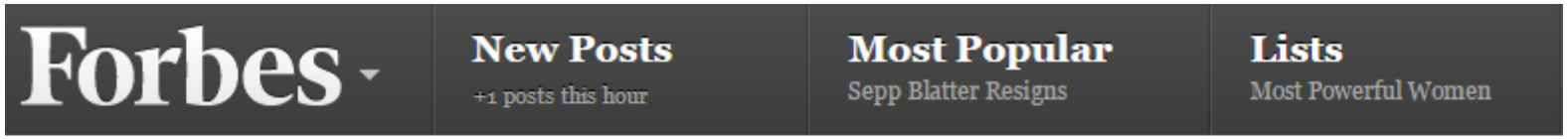


Touching the Void: The…
› Joe Simpson
315
Paperback
$11.22 Prime
Get it by Tomorrow



Buried in the Sky: The…
› Peter Zuckerman
225
Paperback
$10.63 Prime
Get it by Tomorrow

What customers bought together



4/06/1998 @ 12:00AM

Share

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.

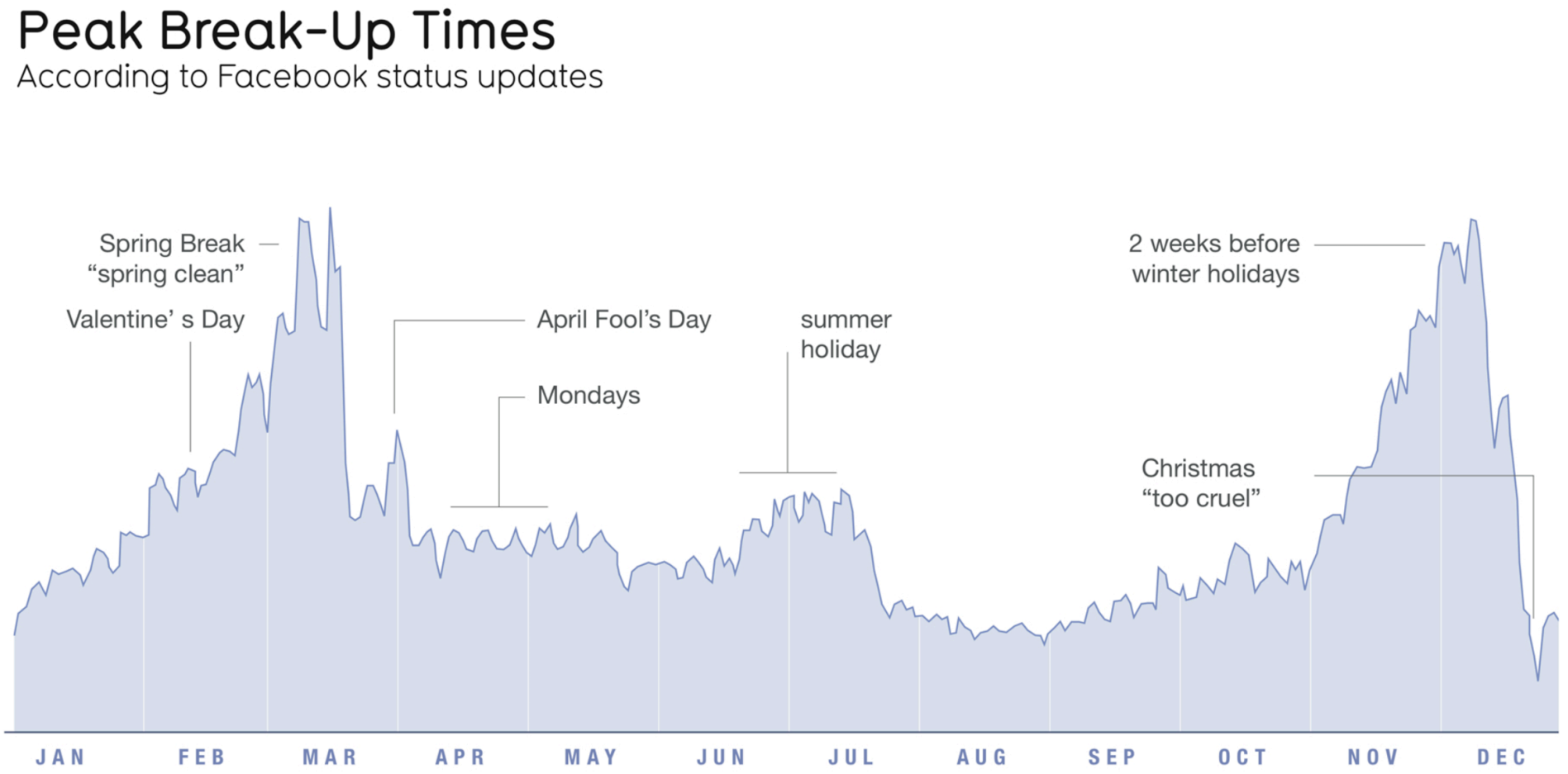
Recommendations are key to personalization

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…



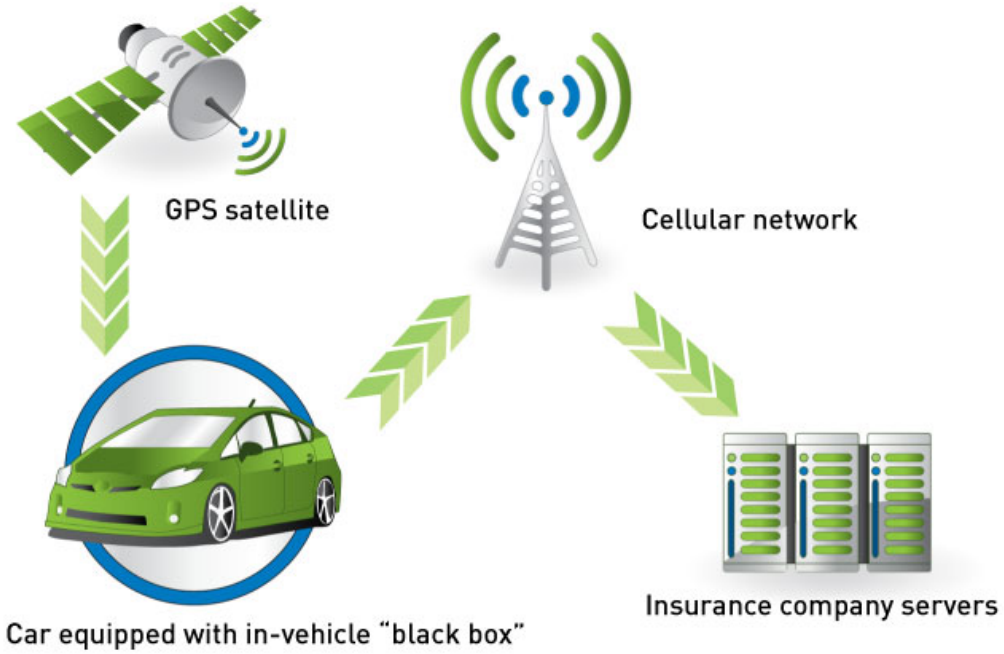
Knowing when breakup happens can help business…

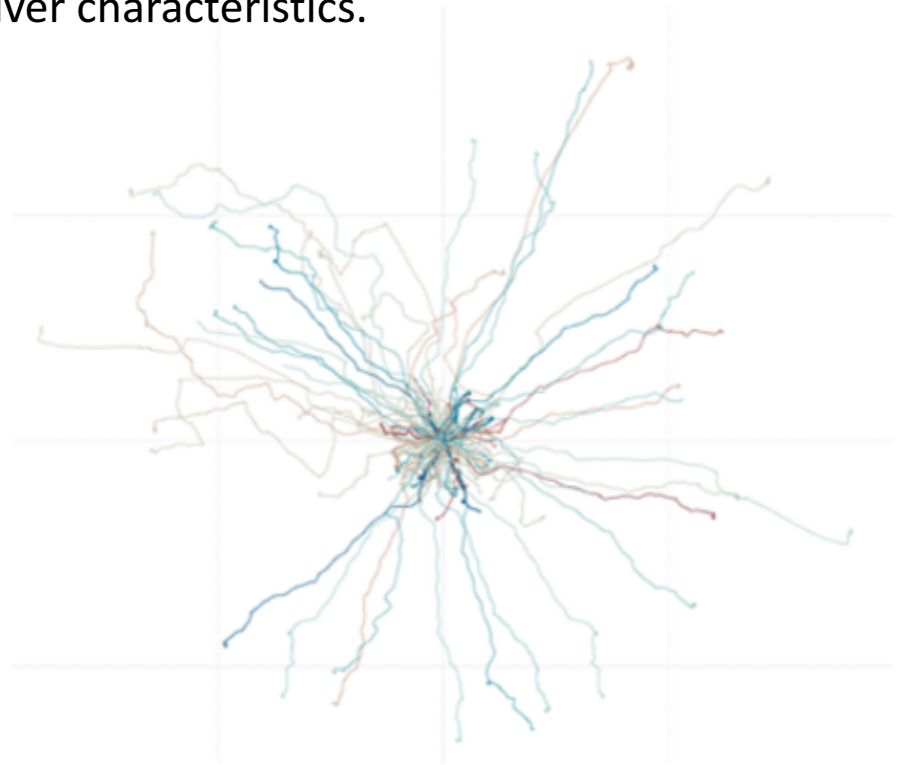
- More traffic to online dating sites after breakup surge
- Increase in demand for relationship lawyers and counselors.
- Higher demand for housing as they move out
- Brand switching (people change the brand of beer after breakup)

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.

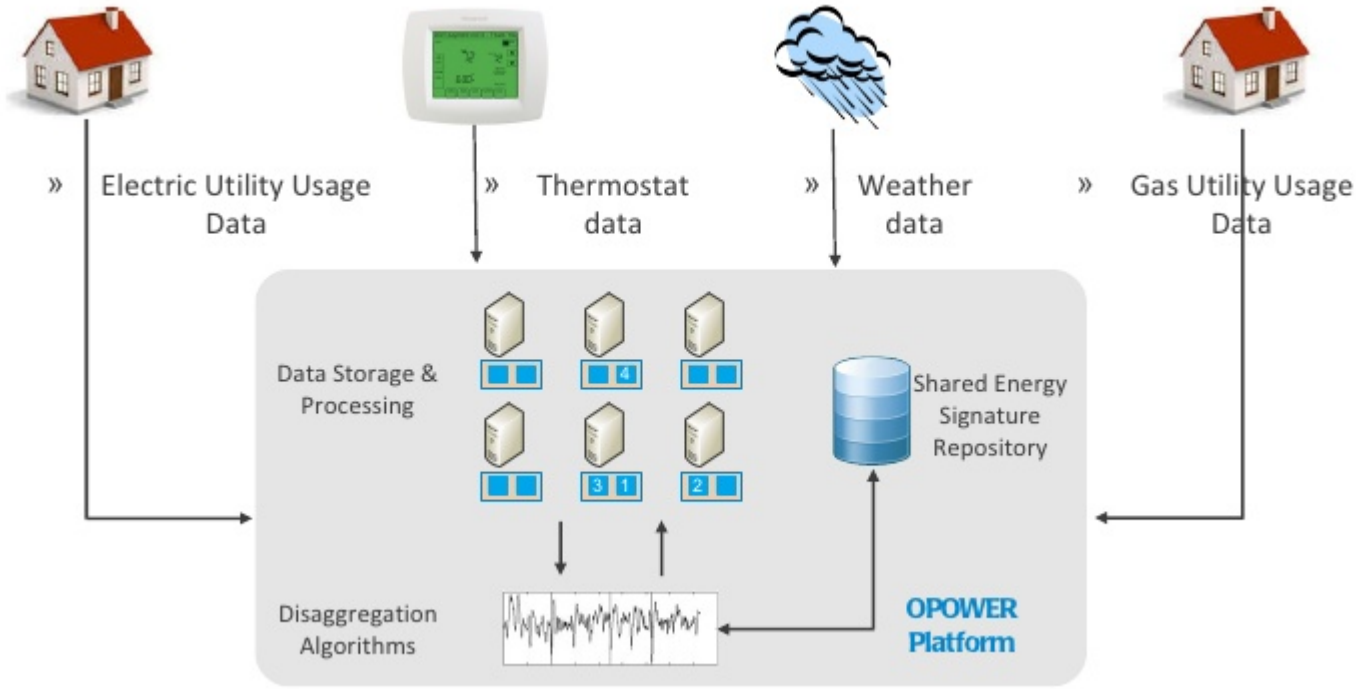




For automobile insurers, telematics represents a growing and valuable way to quantify driver risk. Instead of pricing decisions on vehicle and driver characteristics, telematics gives the opportunity to measure the quantity and quality of a driver's behavior.

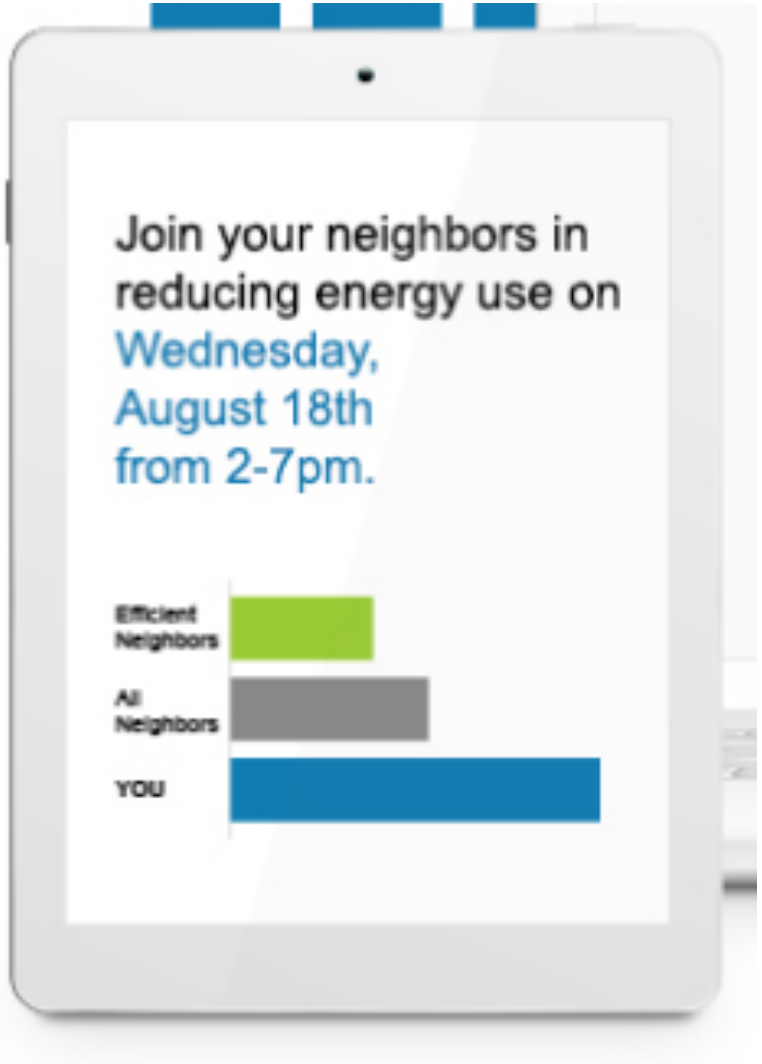
https://www.kaggle.com/c/axa-driver-telematics-analysis

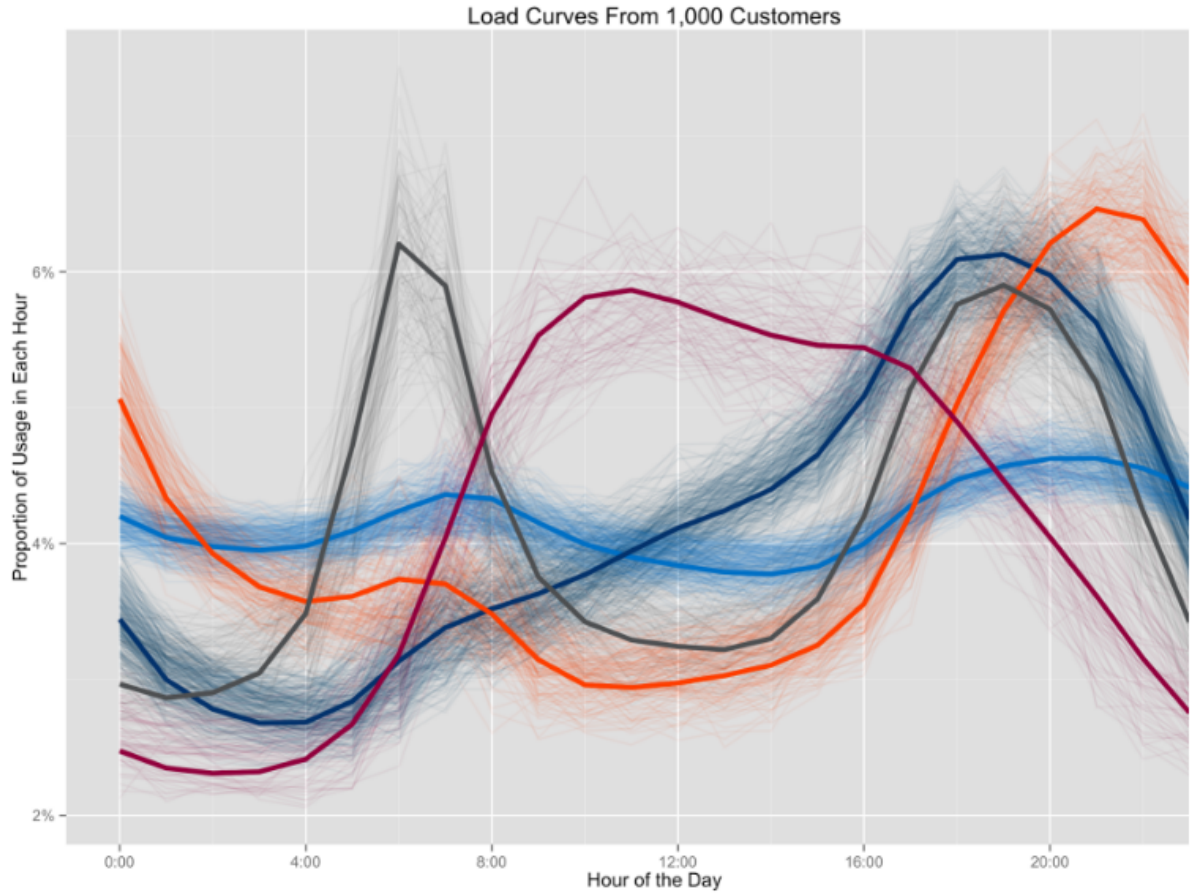
Monitoring Energy Consumption



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

Customer Profiling



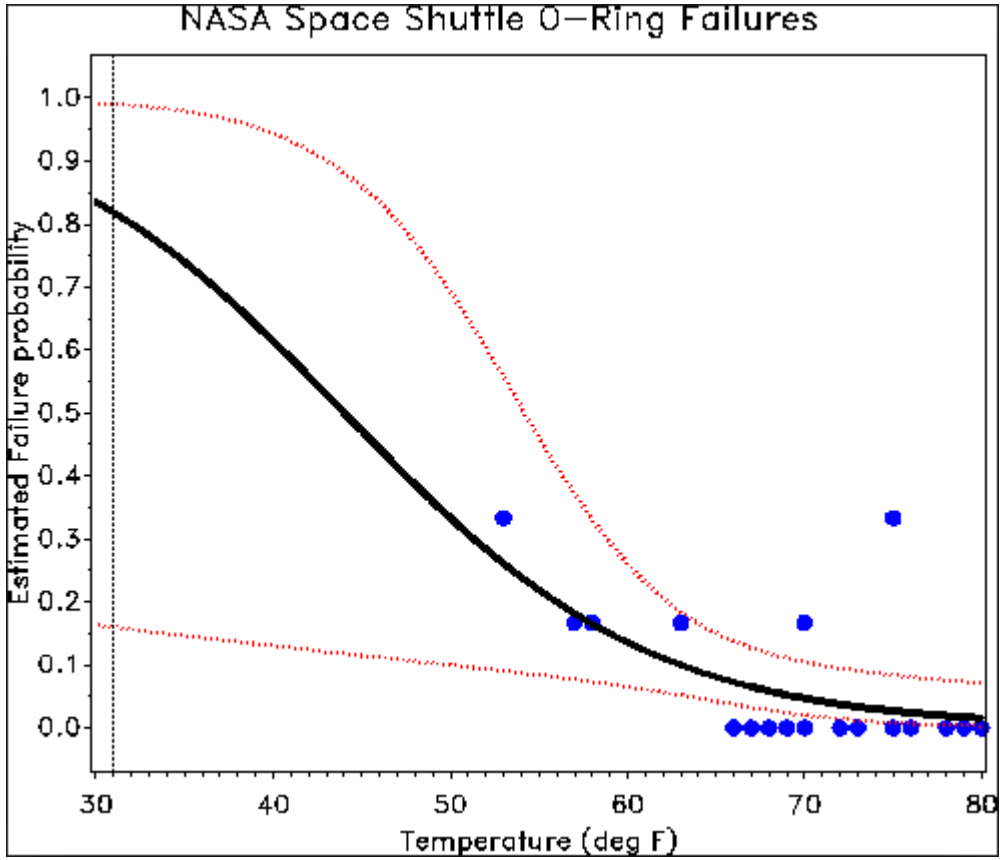


Source: Opower (2014)

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.”

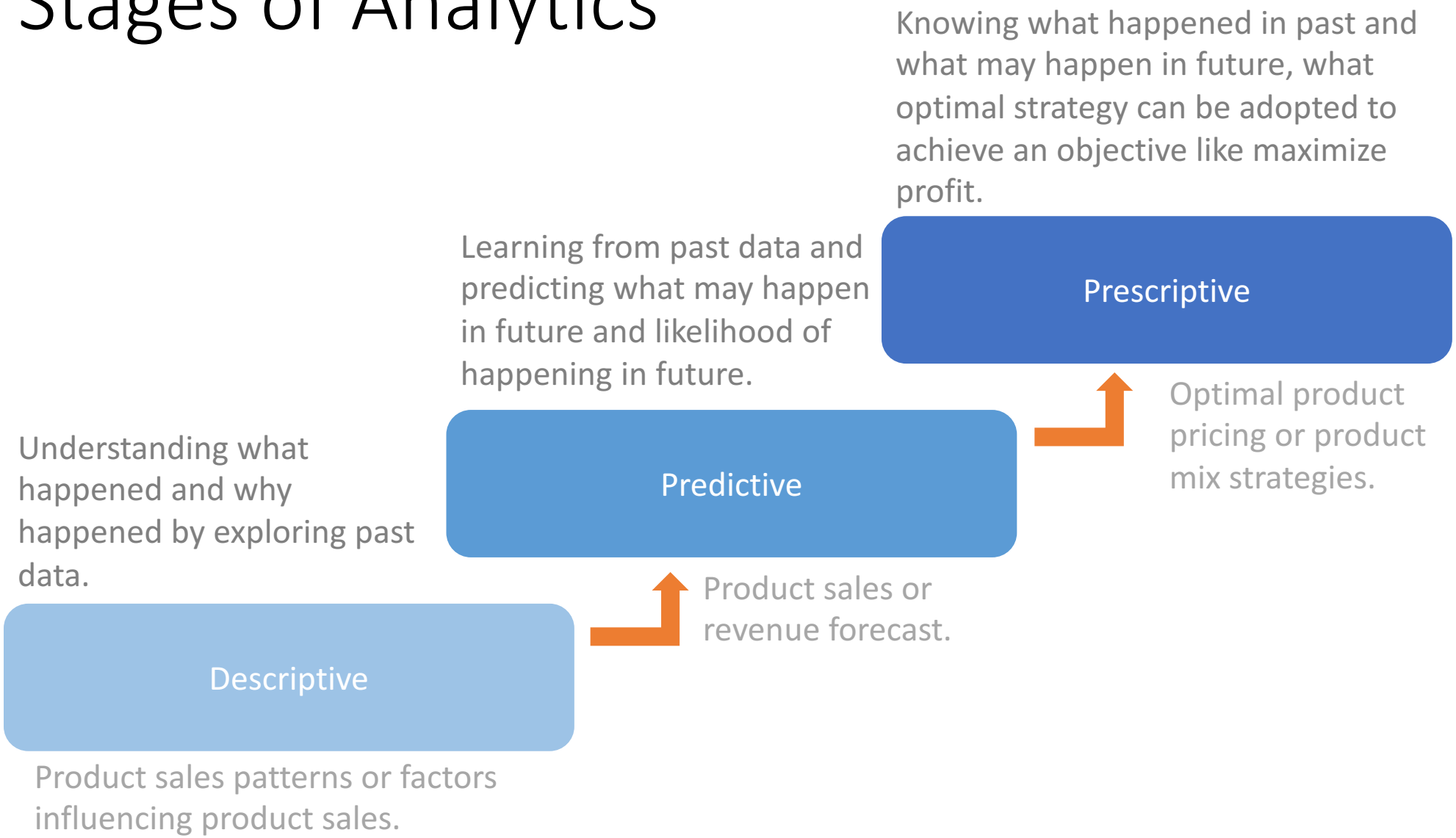
Angela Ahrendts, CEO of Burberry

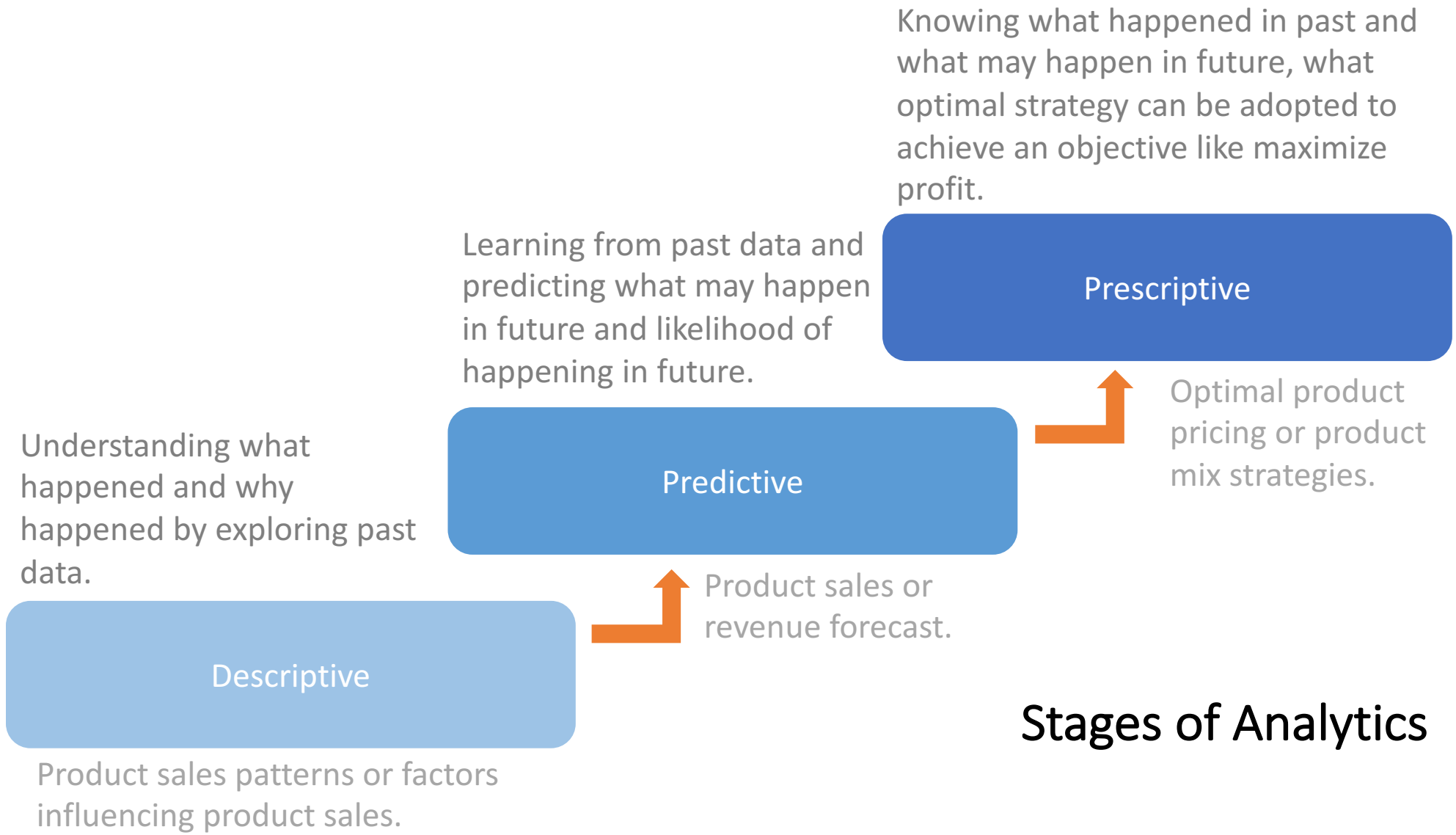
Data Scientists will be the sexiest job of 21st century

Harvard Business Review 2012

Analytics Stages

Stages of Analytics

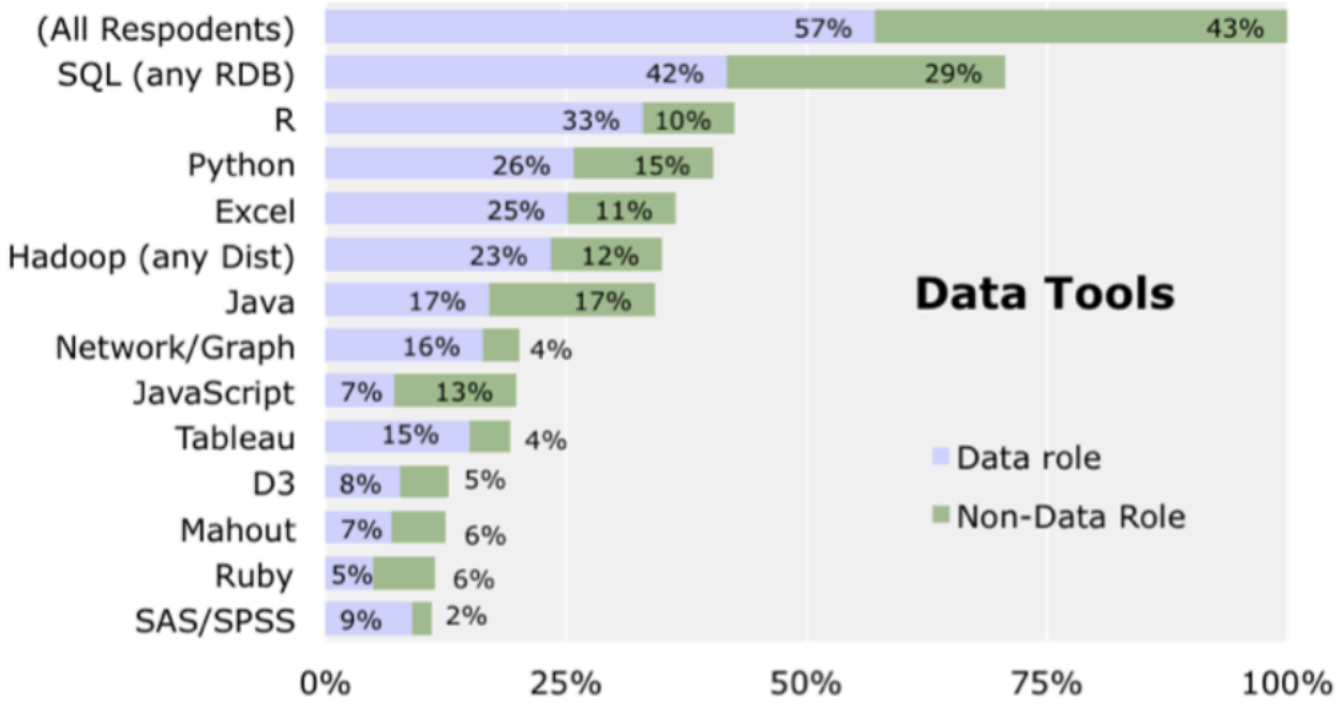




Stages of Analytics

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.

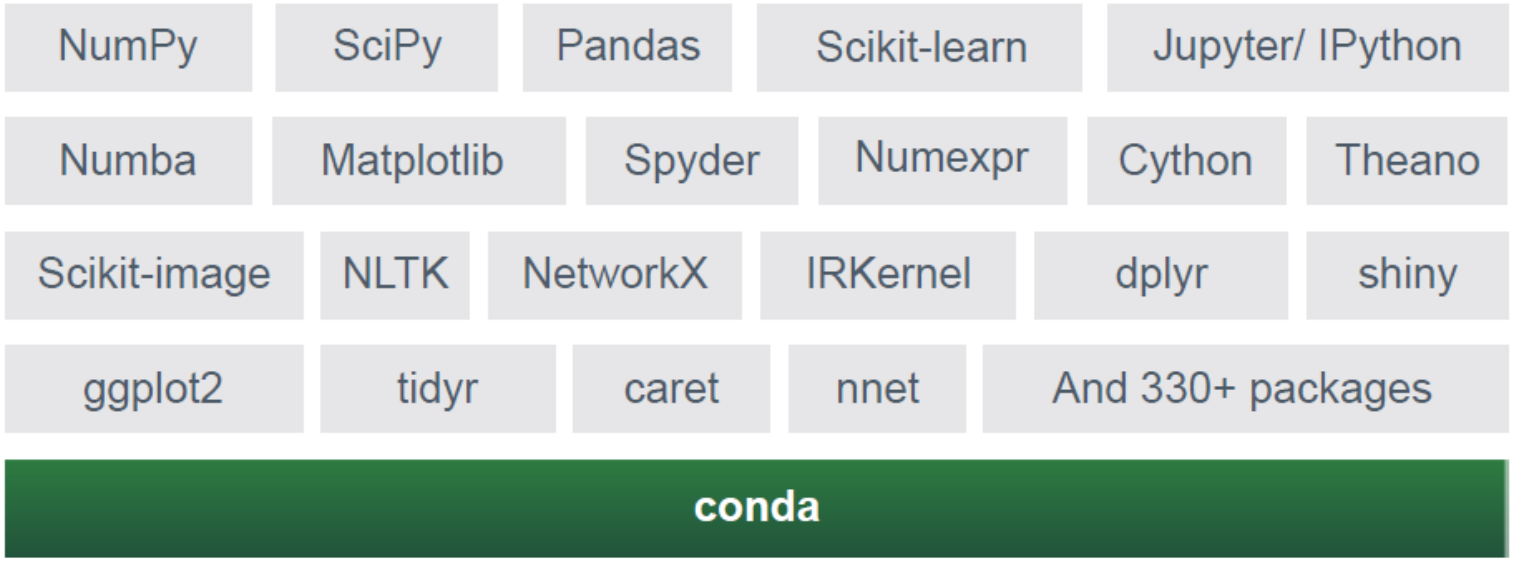
Plotting and visualization

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



Game-Changing Enterprise Ready Python Distribution

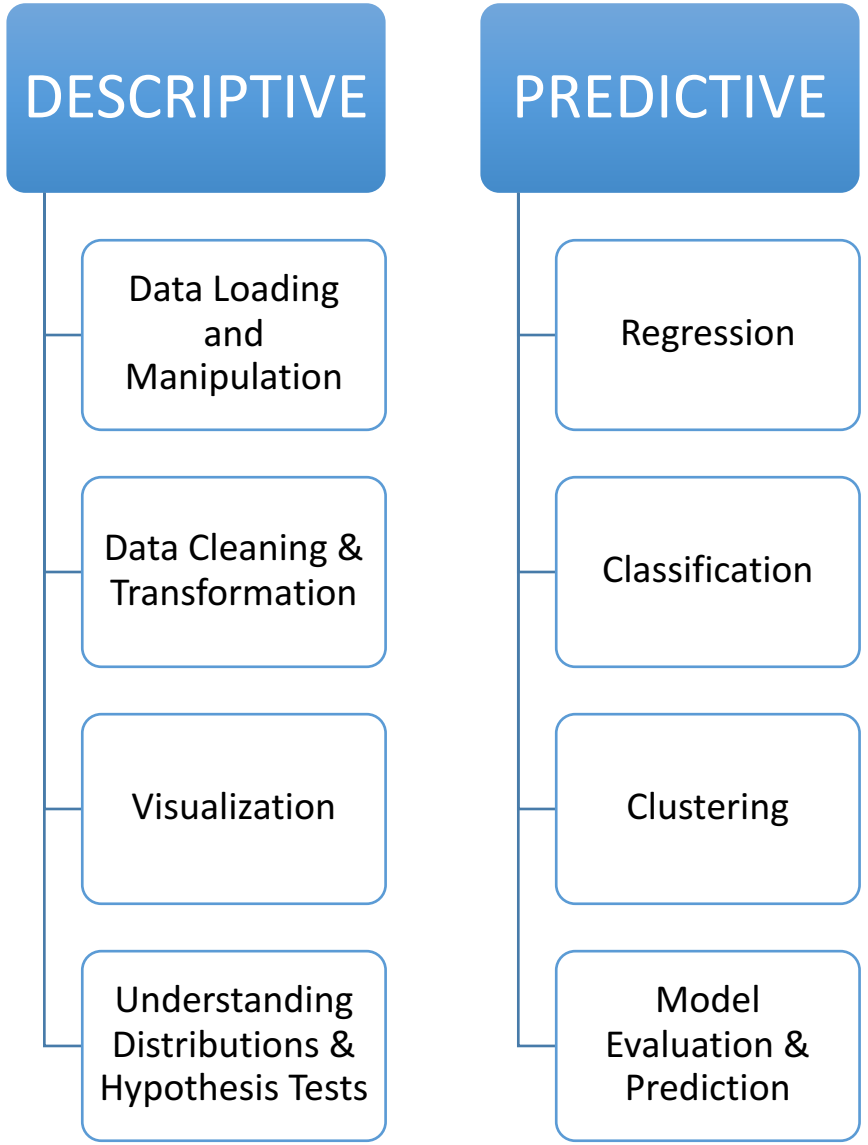
- 2 million downloads in last 2 years
- 200k / month and growing
- conda package manager serves up 5 *million* packages per month
- Recommended installer for IPython/Jupyter, Pandas, SciPy, Scikit-learn, etc.

Source: Continuum Analytics

[Download link: https://www.continuum.io/downloads](https://www.continuum.io/downloads)

What we will learn in this workshop?

Exploratory	Descriptive	Predictive
••Data Loading ••Data Cleanup ••Data Transformations ••Basic Statistics	••Visualization ••Understanding Distributions ••Hypothesis Testing	••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

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

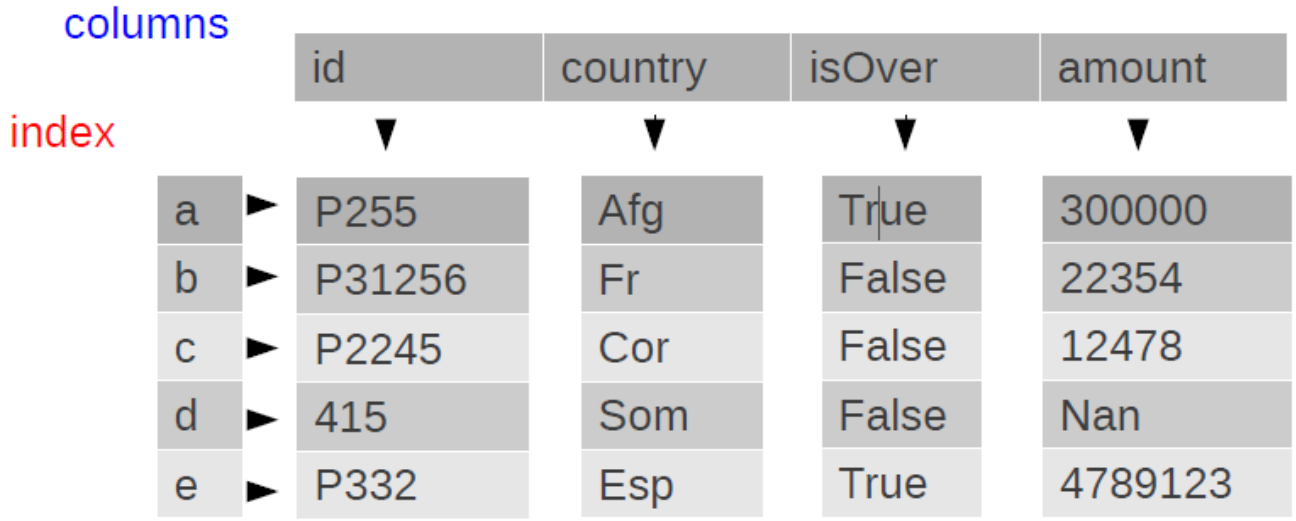


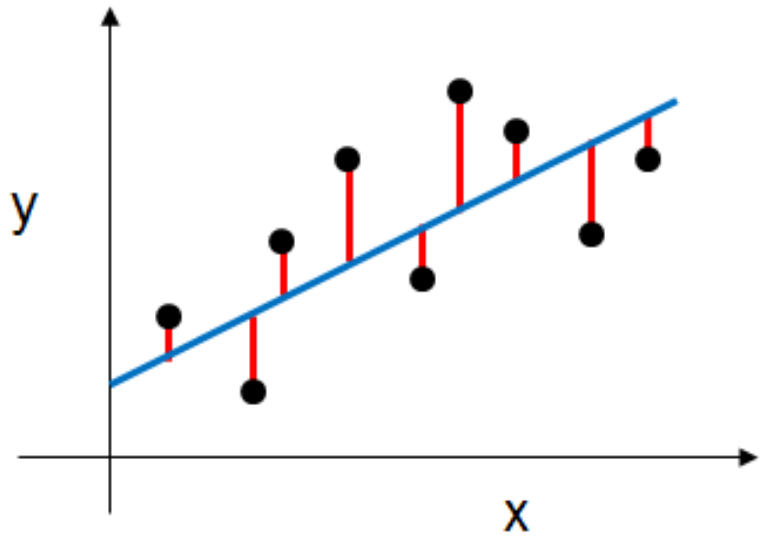
Table like structure

- 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



$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n$$

- 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**:

Error of the Estimated line is :

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

Model Prediction

Observed Result

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

Linear Regression – Model Evaluation

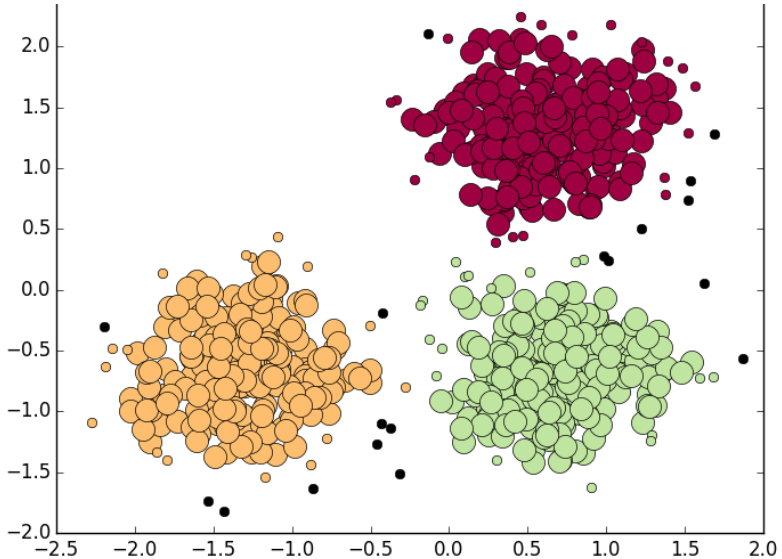
- MAE – Mean Absolute Error
- MSE – Mean Squared Error
  - Penalizes larger residuals
- RMSE – Root Mean Squared Error
- R Square – How much variance explained by the model
  - 1 – all variance explained
  - 0 – no variance explained. It is better not have the model at all.
  - Closer to 1 – is better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

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

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

Clustering

- Determine the intrinsic grouping in a set of un-labeled 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	Predicted Class: Negative
True Class	Positive	**T**rue **P**ositive (TP)	**F**alse **N**egative (FN)
True Class	Negative	**F**alse **P**ositive (FP)	**T**rue **N**egative (TN)

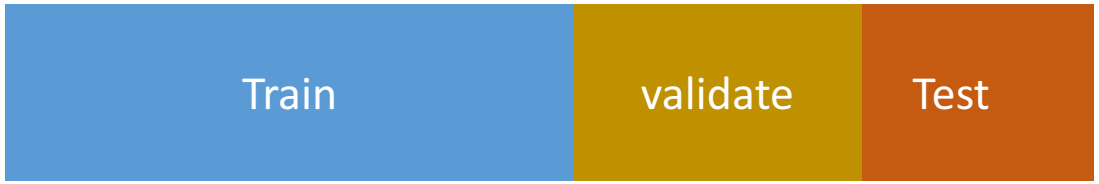
Sensitivity: $TP/(TP + FN)$

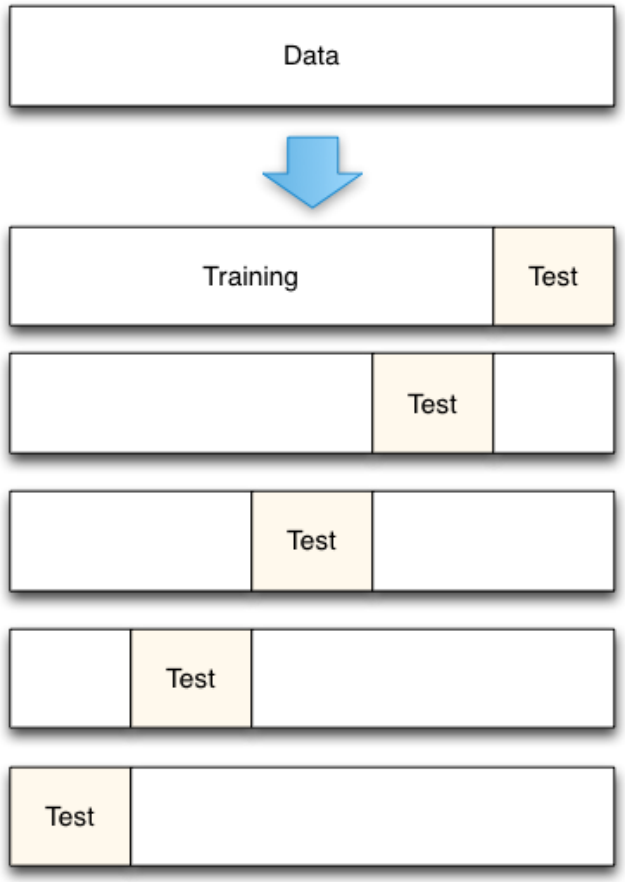
Precision: $TP/(TP + FP)$

Specifity: $TN/(TN + FP)$

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

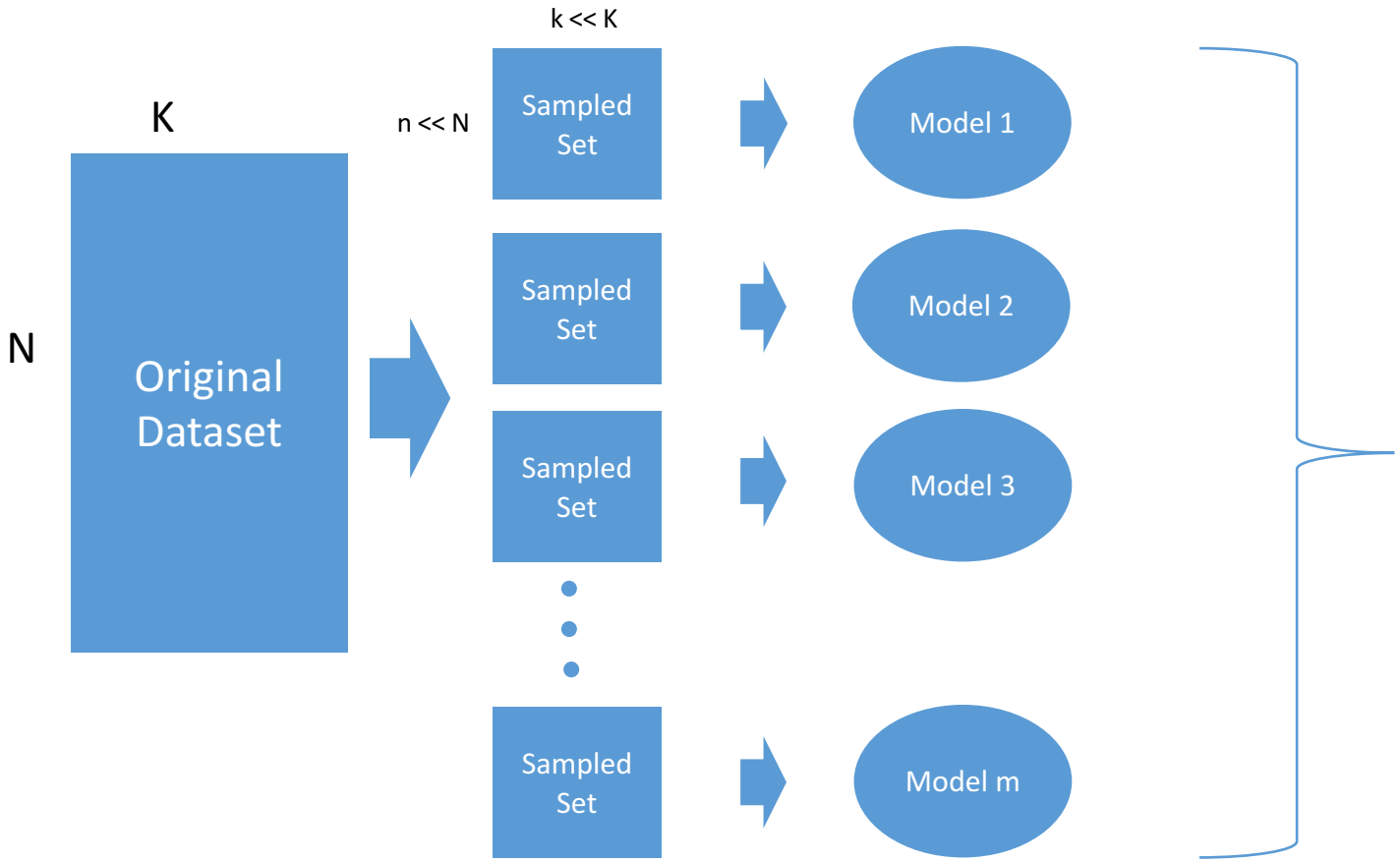
Train - Test Split & Cross Validation





Train - Test Split & Cross Validation

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteration 1	train	train	train	train	test
Iteration 2	train	train	train	test	train
Iteration 3	train	train	test	train	train
Iteration 4	train	test	train	train	train
Iteration 5	test	train	train	train	train



Aggregating Results

- Taking a simple or weighted average for Regression
- Majority Voting for Classification

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.