Big Data Mining: A Gentle Introduction

Big Data Mining

PUCV

Dr. Héctor Allende-Cid

*Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

hector.allende@pucv.cl

September, 2017

Motivation from Real World Problems

- Large Hadron Collider experiments: 25 PB annual rate.
- Sloan Digital Sky Survey (SDSS): 200 GB per night.
- Google processes 24 PB each day.
- AT&T transmits 30 PB each day.

Other examples

Aplications

- Biology/Medicine applications.
- Seismology.
- Meteorology.
- Astronomy.
- Financial Data.
- Neuro-Science.
- etc.

Biology/Medicine

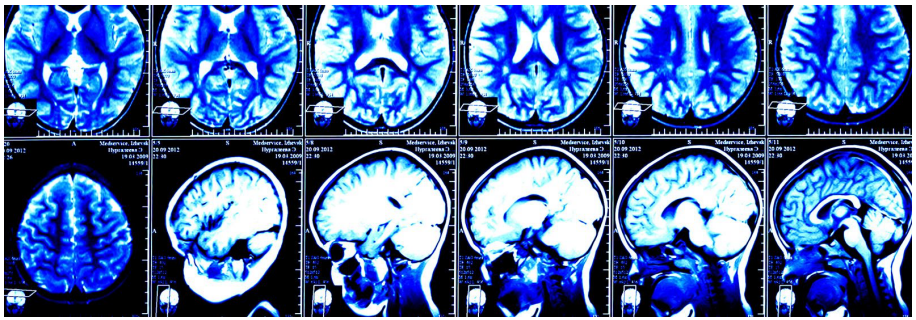
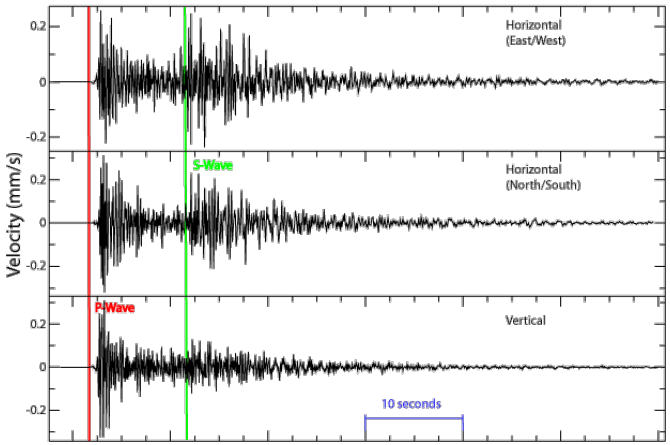
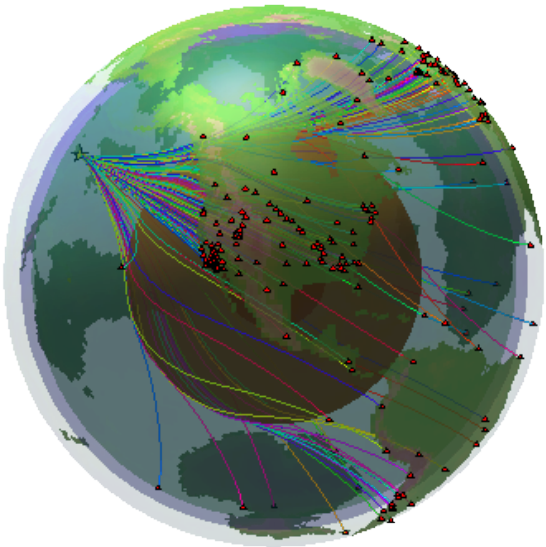


Figure: Medicine

Seismology





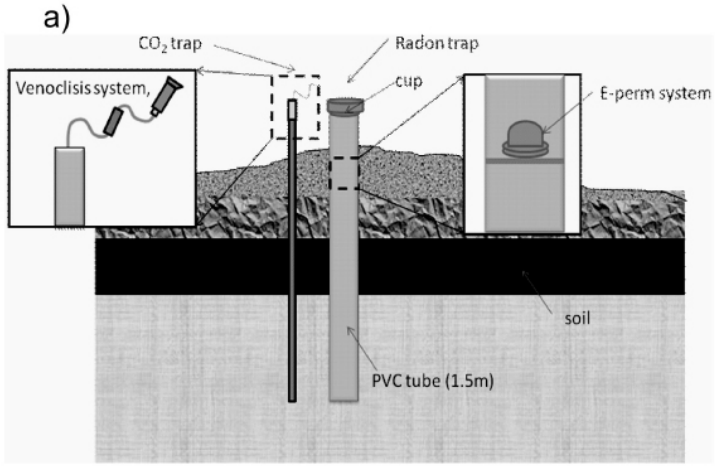
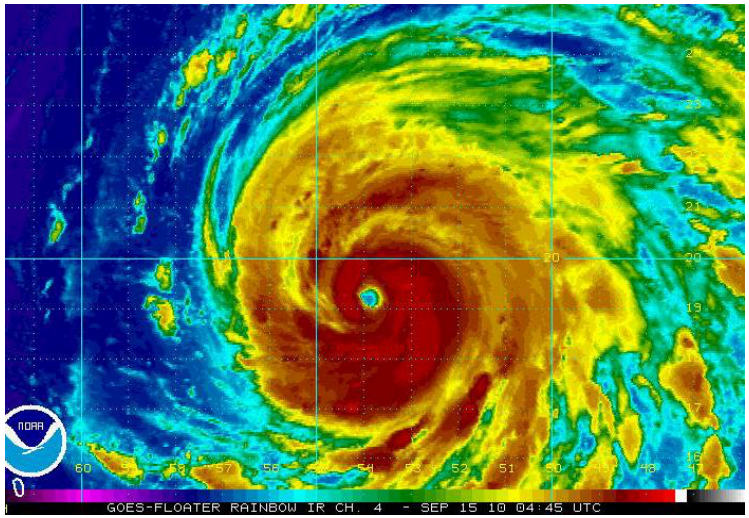


FIGURE 3. a) Example of radon and CO_2 in soil station deployment, used in this study. b) collecting data.

Figure: Seismology

Meteorology



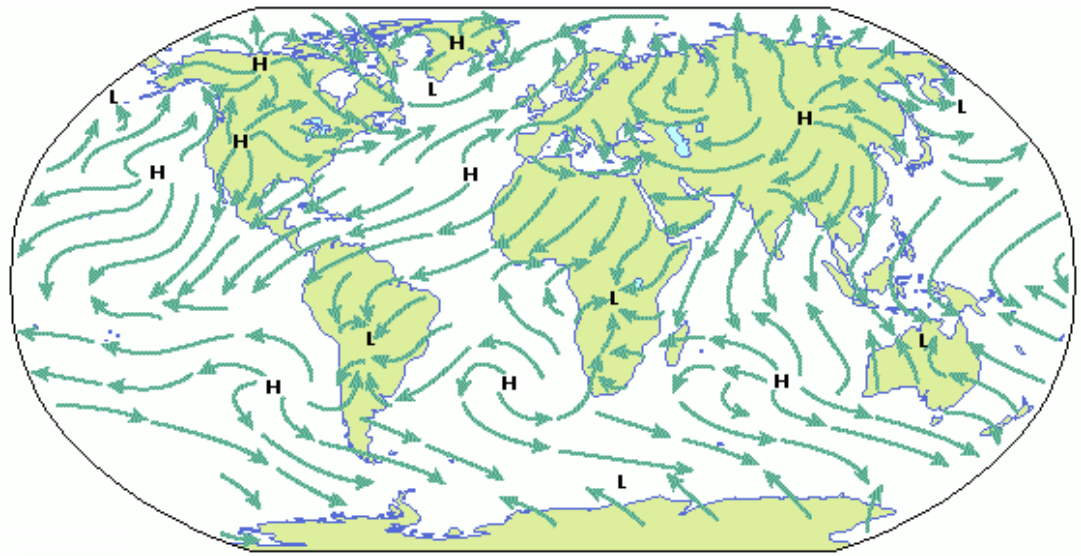
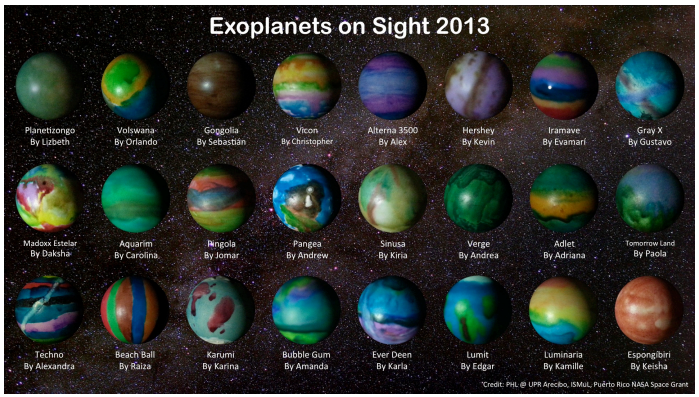


Figure: Meteorology

Astronomy



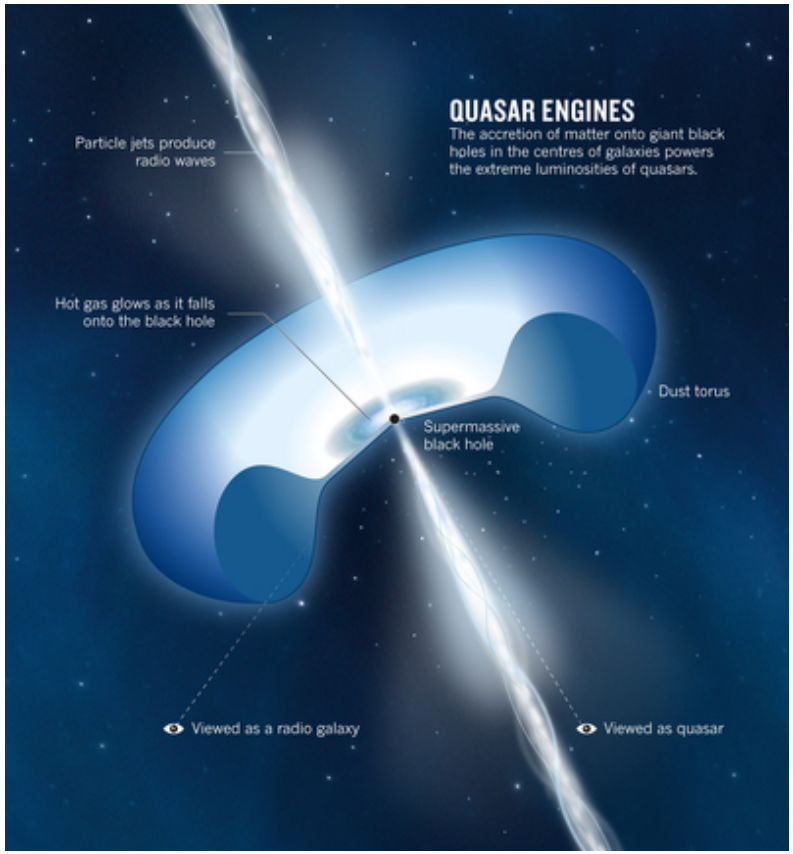
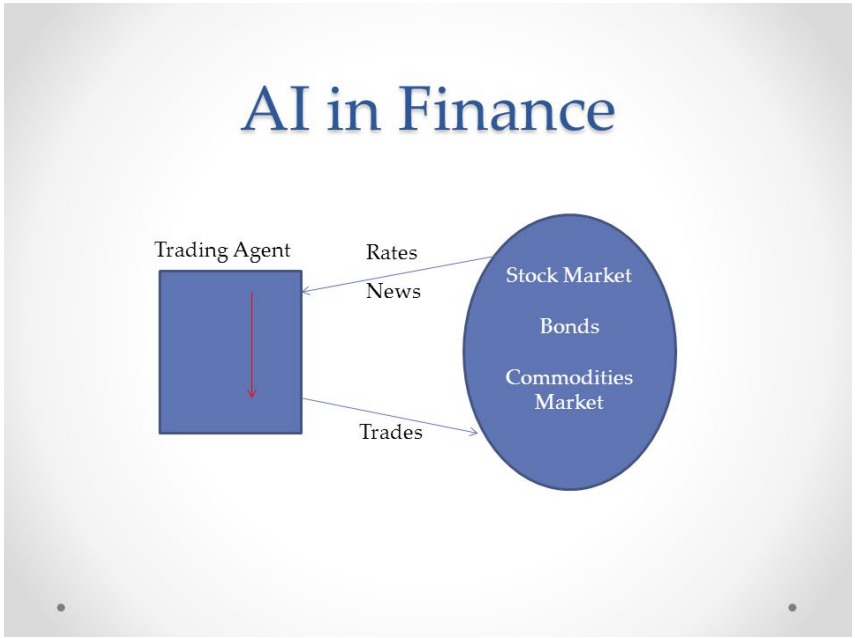


Figure: Astronomy

Finance



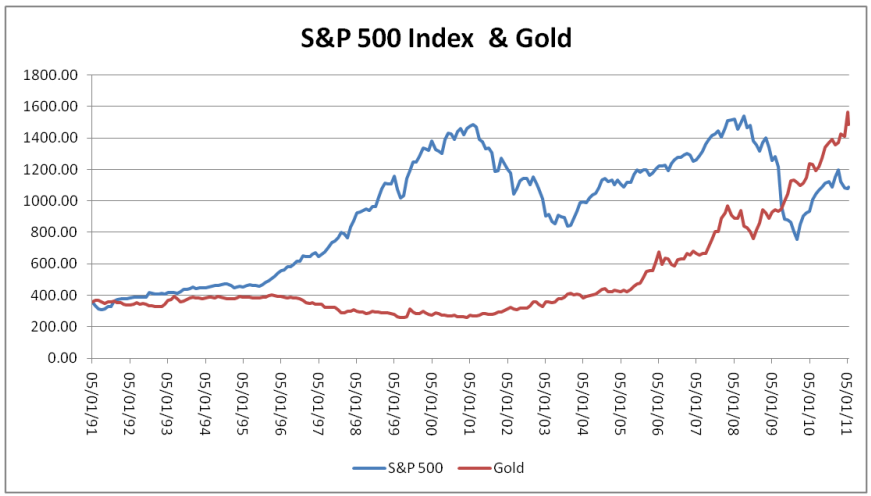
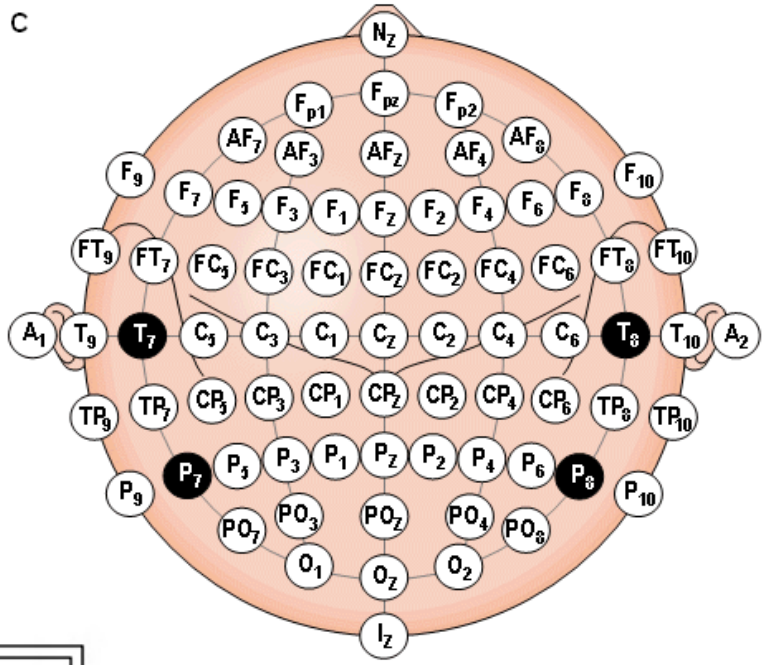


Figure: Financial Data

Neuroscience



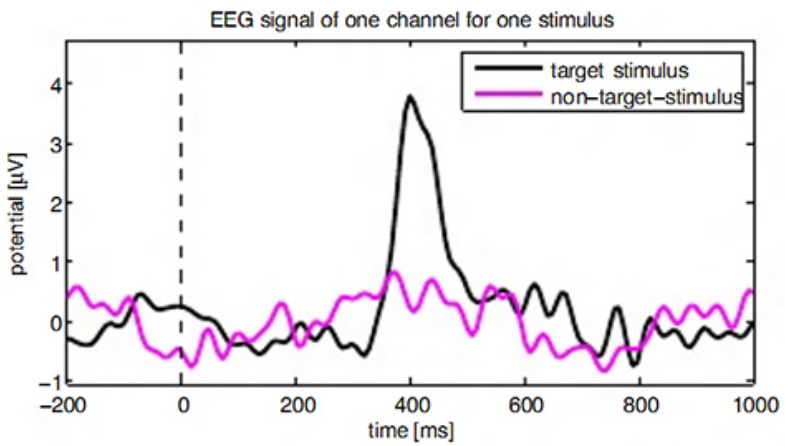
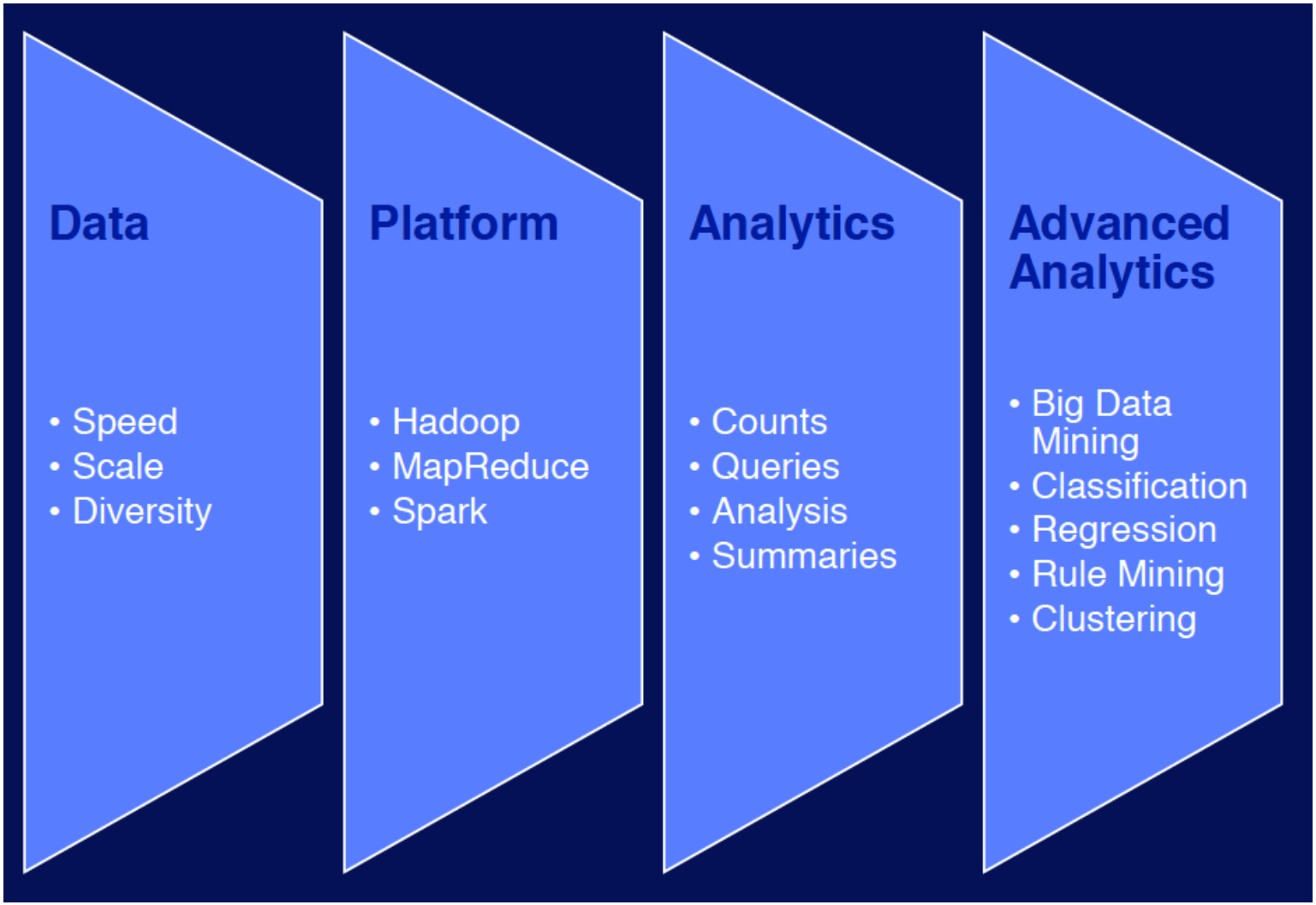


Figure: Neuroscience

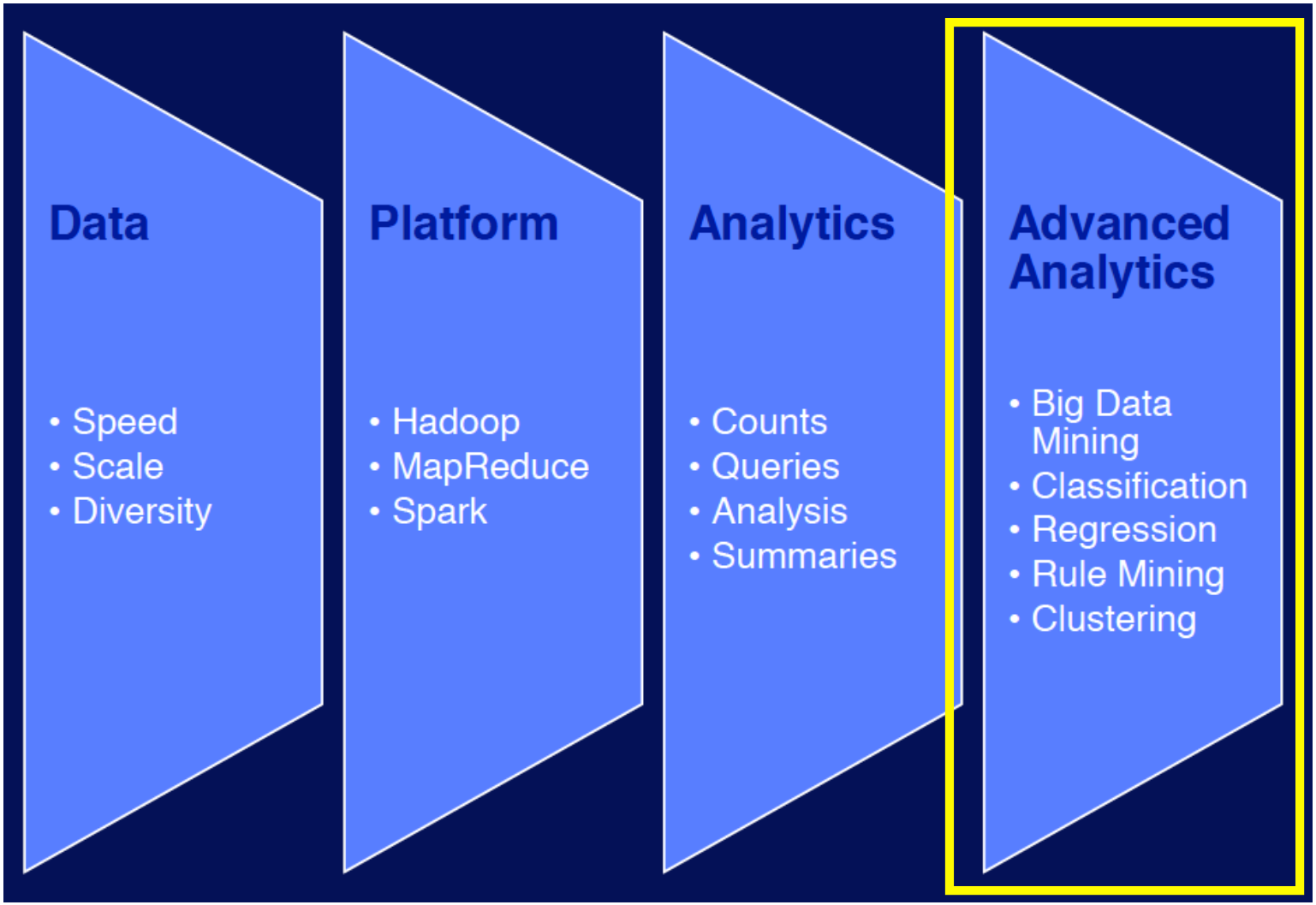
What is a Data Scientist?

https://www.youtube.com/watch?v=iQBat7e0MQs

Advanced Analytics



Advanced Analytics



What are the Big Data Challenges?

Figure: Big Data Challenges

Data → Insight → Action

Figure: Transforming Data into Insight for Making Better Decisions

Approaches (1/2)

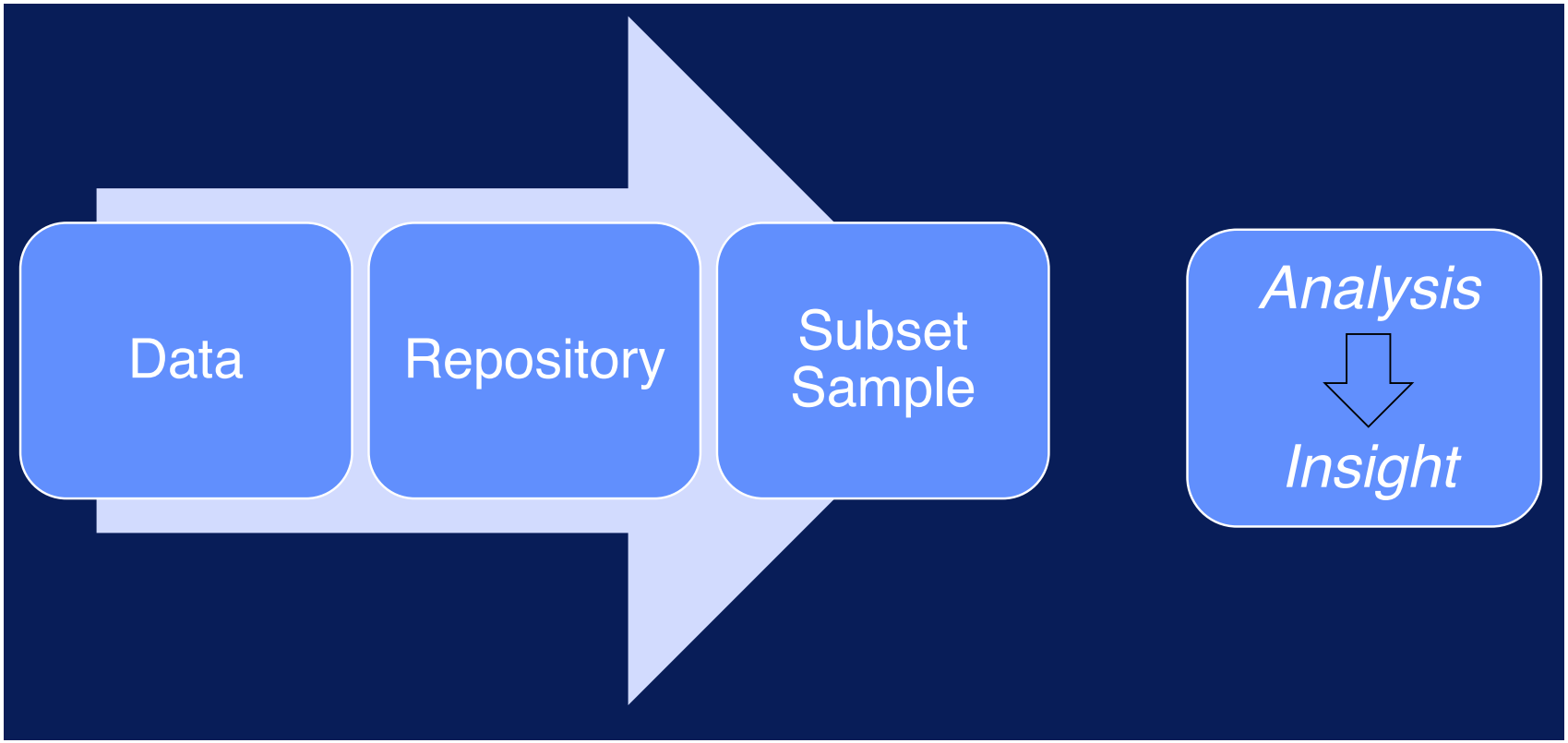


Figure: Traditional approach

Approaches (2/2)

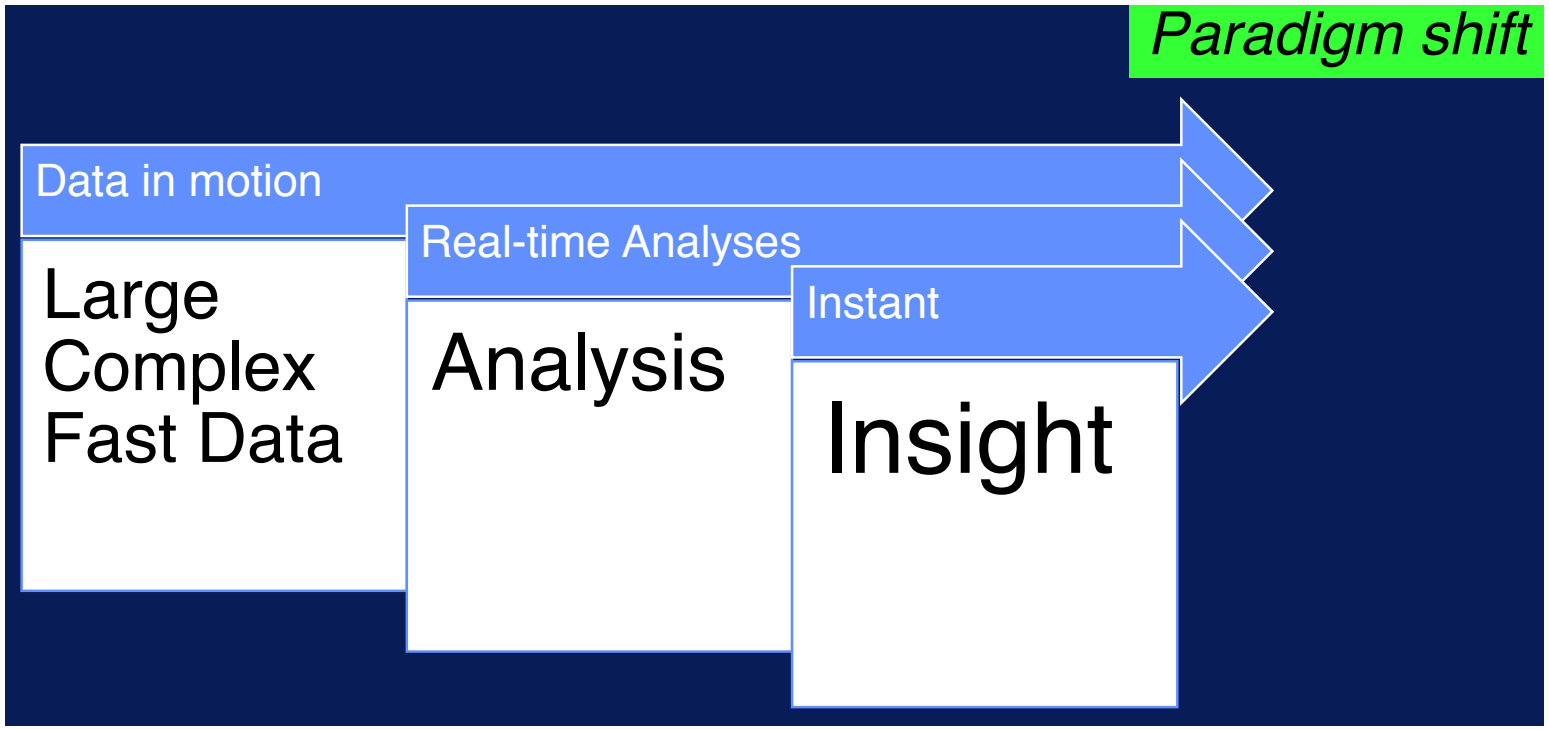
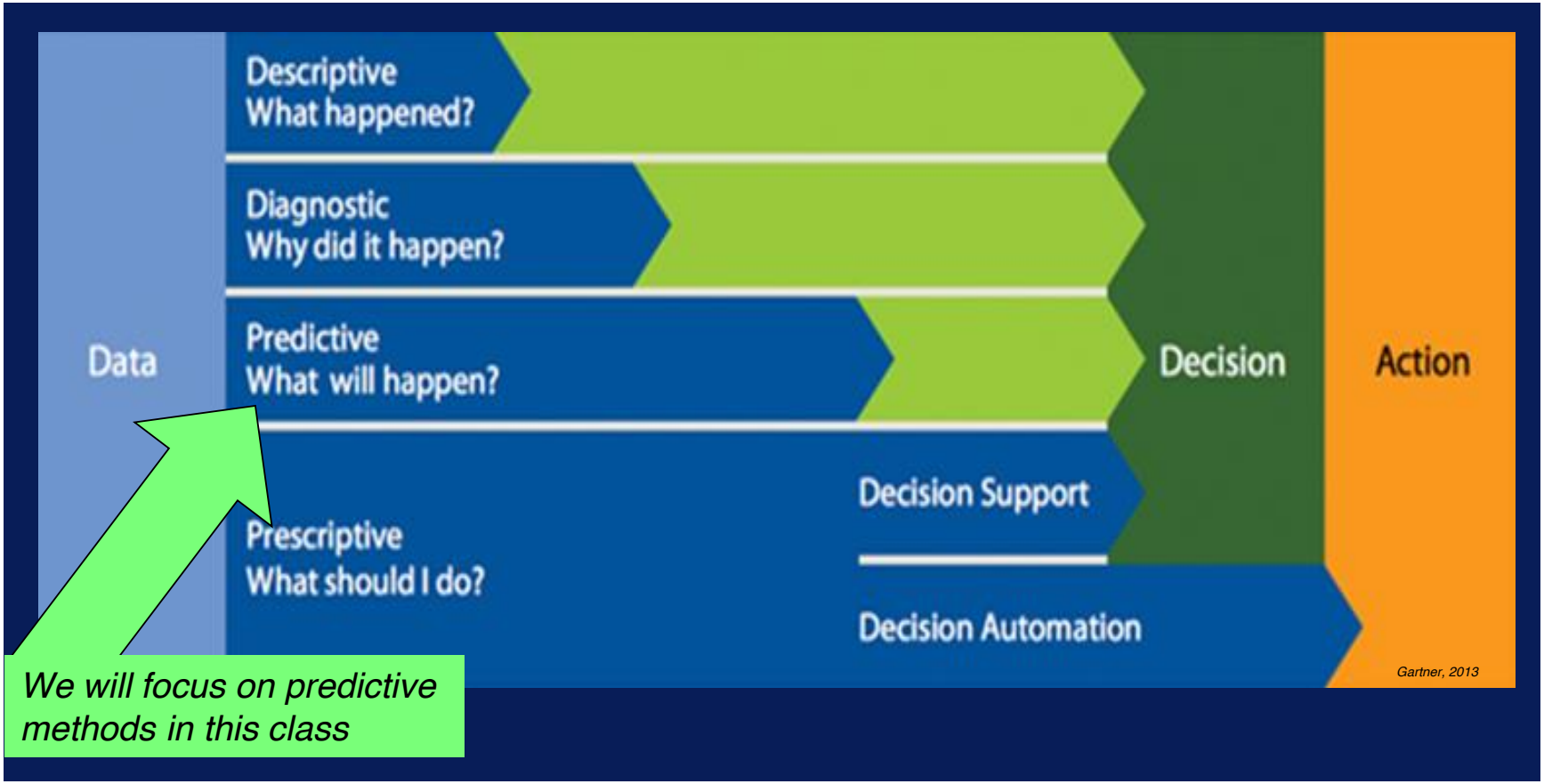


Figure: Big Data Approach

Focus


Descriptive
What happened?
Diagnostic
Why did it happen?
Predictive
What will happen?
Prescriptive
What should I do?
Data
Decision Support
Decision Automation
Decision
Action
Gartner, 2013
We will focus on predictive methods in this class

Maturity Level

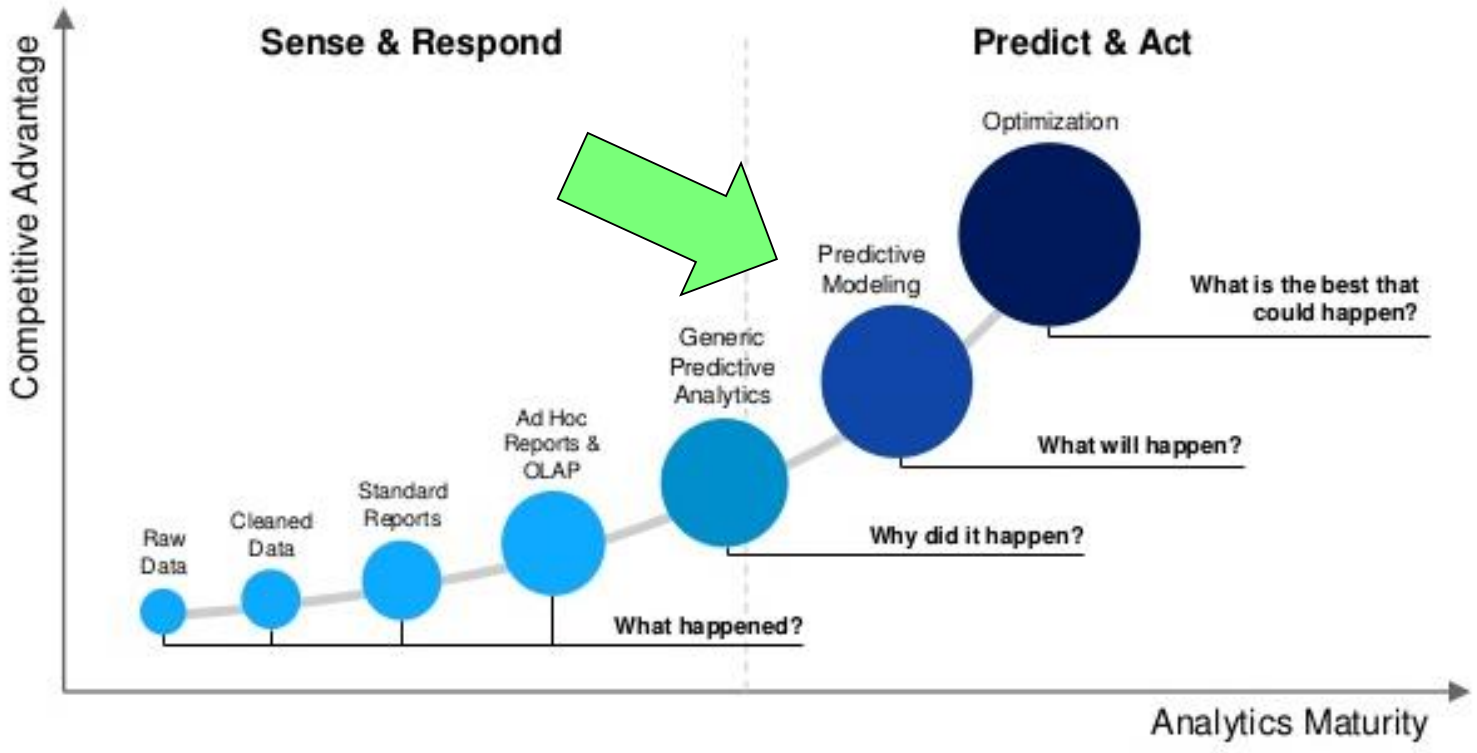


Figure: Maturity Levels

BD and ML

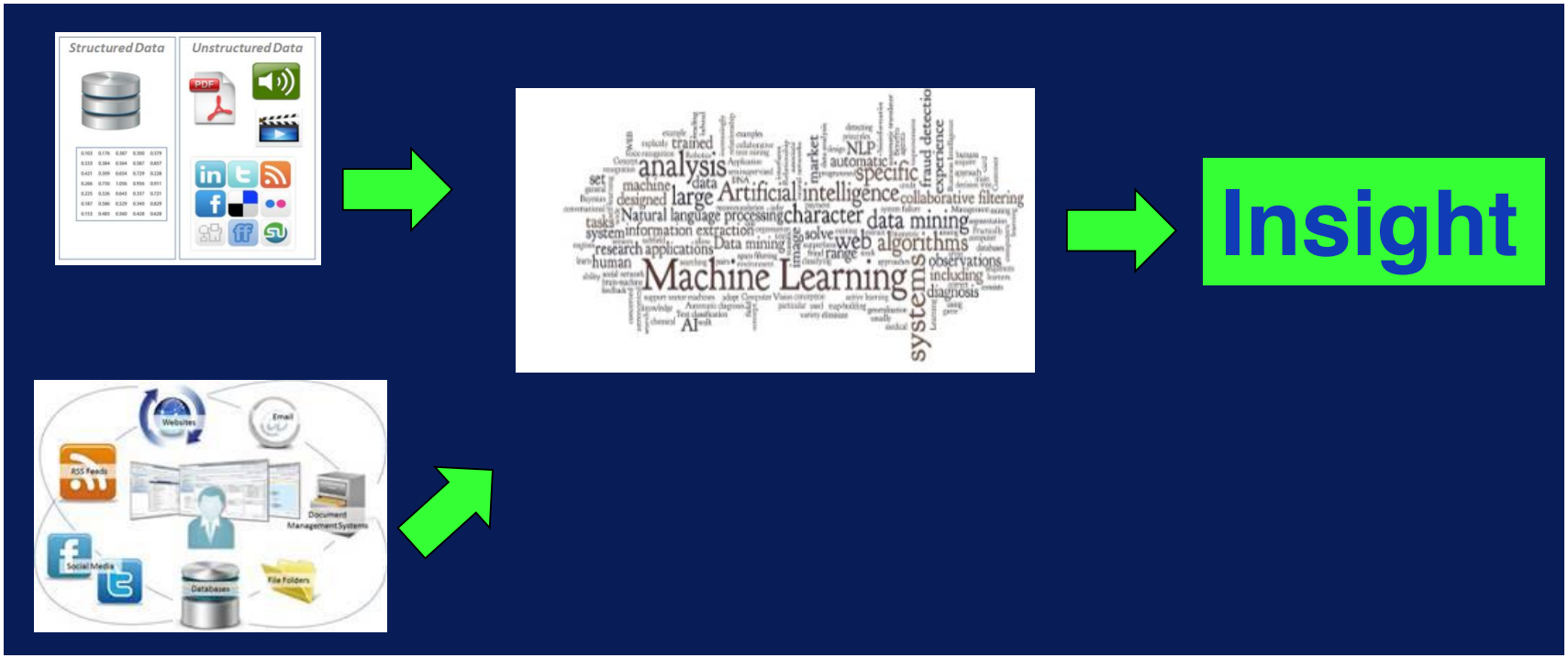


Figure: Big Data and Machine Learning

Why do Machine Learning on Big Data?

Traditional analytics tools are not well suited to capturing the full value of big data.

The volume of data is too large for comprehensive analysis, and the range of potential correlations and relationships between disparate data sources — from back end customer databases to live web based clickstreams —are too great for any analyst to test all hypotheses and derive all the value buried in the data.

Basic analytical methods used in business intelligence and enterprise reporting tools reduce to reporting sums, counts, simple averages and running SQL queries. Online analytical processing is merely a systematized extension of these basic analytics that still rely on a human to direct activities specify what should be calculated.

Machine learning is ideal for exploiting the opportunities hidden in big data.

It delivers on the promise of extracting value from big and disparate data sources with far less reliance on human direction. It is data driven and runs at machine scale. It is well suited to the complexity of dealing with disparate data sources and the huge variety of variables and amounts of data involved. And unlike traditional analysis, machine learning thrives on growing datasets. The more data fed into a machine learning system, the more it can learn and apply the results to higher quality insights.

Articles

SHARE

SHARE 138

TWEET

PIN

COMMENT 0

EMAIL

SPONSOR CONTENT MARTIN HACK, SKYTREE

USE DATA TO TELL THE FUTURE: UNDERSTANDING MACHINE LEARNING

Image: manoftaste.de/Flickr

LATEST NEWS

ELECTION 2016
Clinton to America: What If Twitter Were the Situation Room
21 HOURS

MAPS
Fascinating App Shows You How Misleading Maps Can Be
1 DAY

Big Data: IoT

- "IoT will produce a treasure trove of big data - data that can help cities predict accidents and crimes, give doctors real-time insight into information from pacemakers or biochips, enable optimized productivity across industries through predictive maintenance on equipment and machinery, create truly smart homes with connected appliances and provide critical communication between self-driving cars. The possibilities that IoT brings to the table are endless."
- "In an IoT situation, machine learning can help companies take the billions of data points they have and boil them down to what?s really meaningful. The general premise is the same as in the retail applications - review and analyze the data you've collected to find patterns or similarities that can be learned from, so that better decisions can be made."

Similar concepts

Figure: Similar concepts used (sometimes) indistinctively

The Big Picture

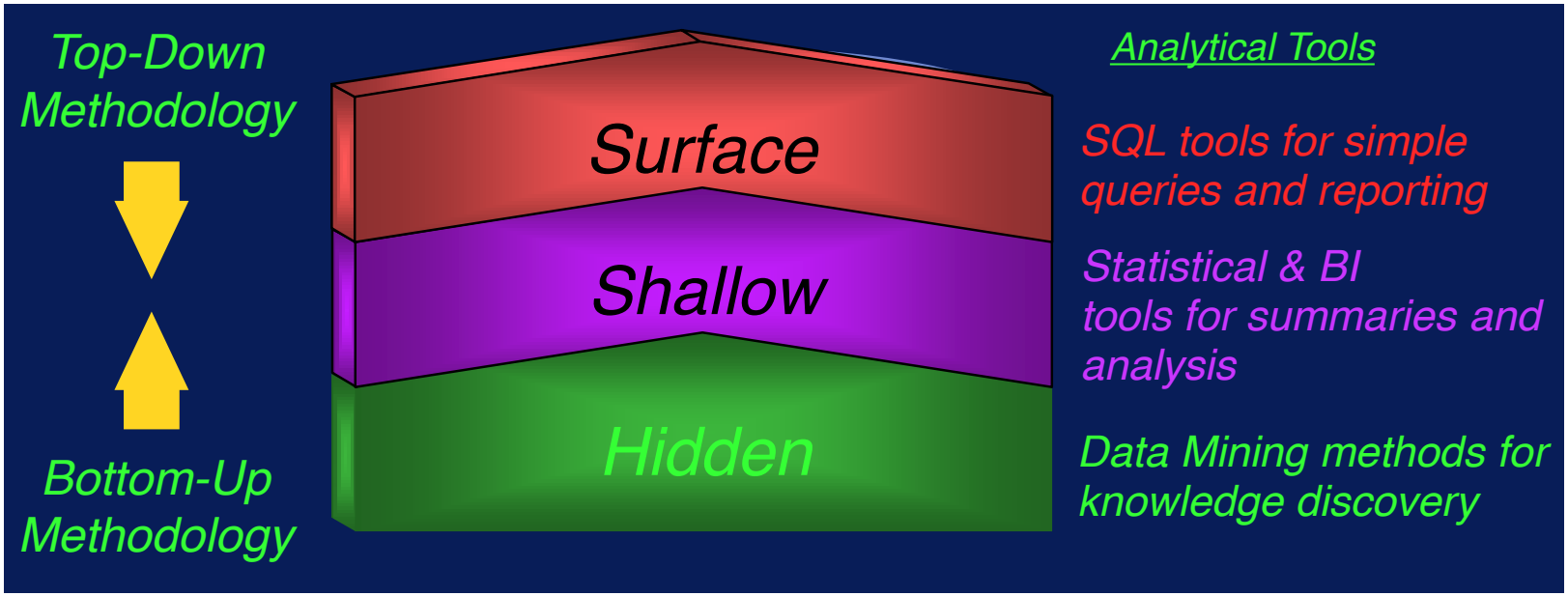
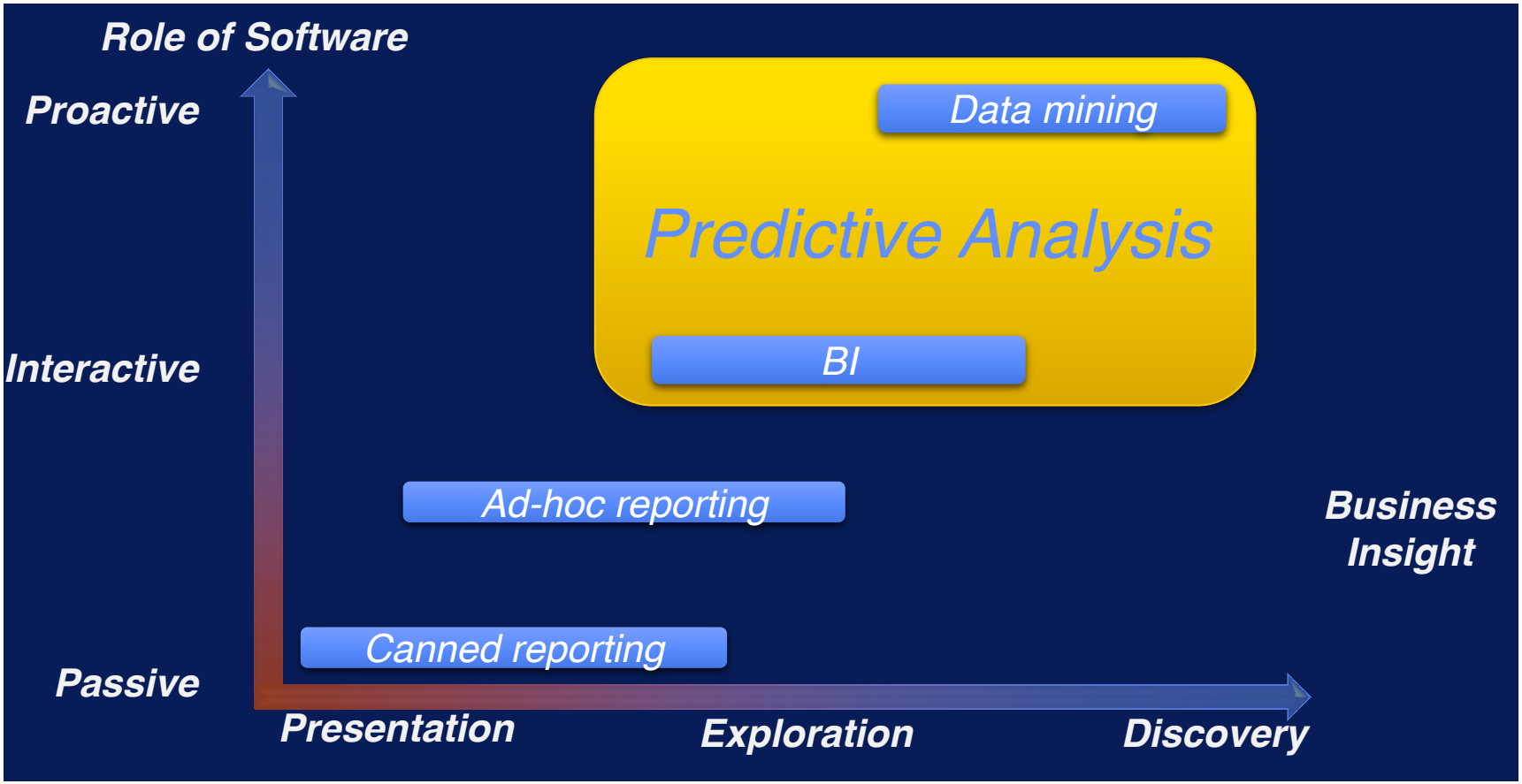


Figure: How to approach it.

Predictive Analysis



What is Data Mining?

Combination of AI and statistical analysis to discover information that is "hidden" in the data.

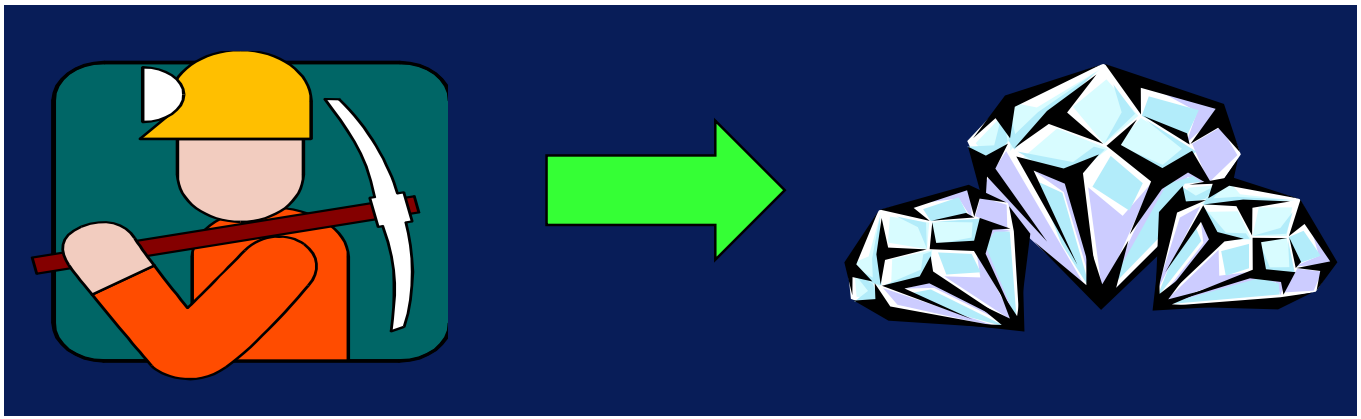


Figure: Data "miner" / Data Scientist

What can be hidden in the data?

- Associations
- Sequences
- Classification
- Forecasting
- Anomalies
- Grouping/Clusters/Segments

Data Mining

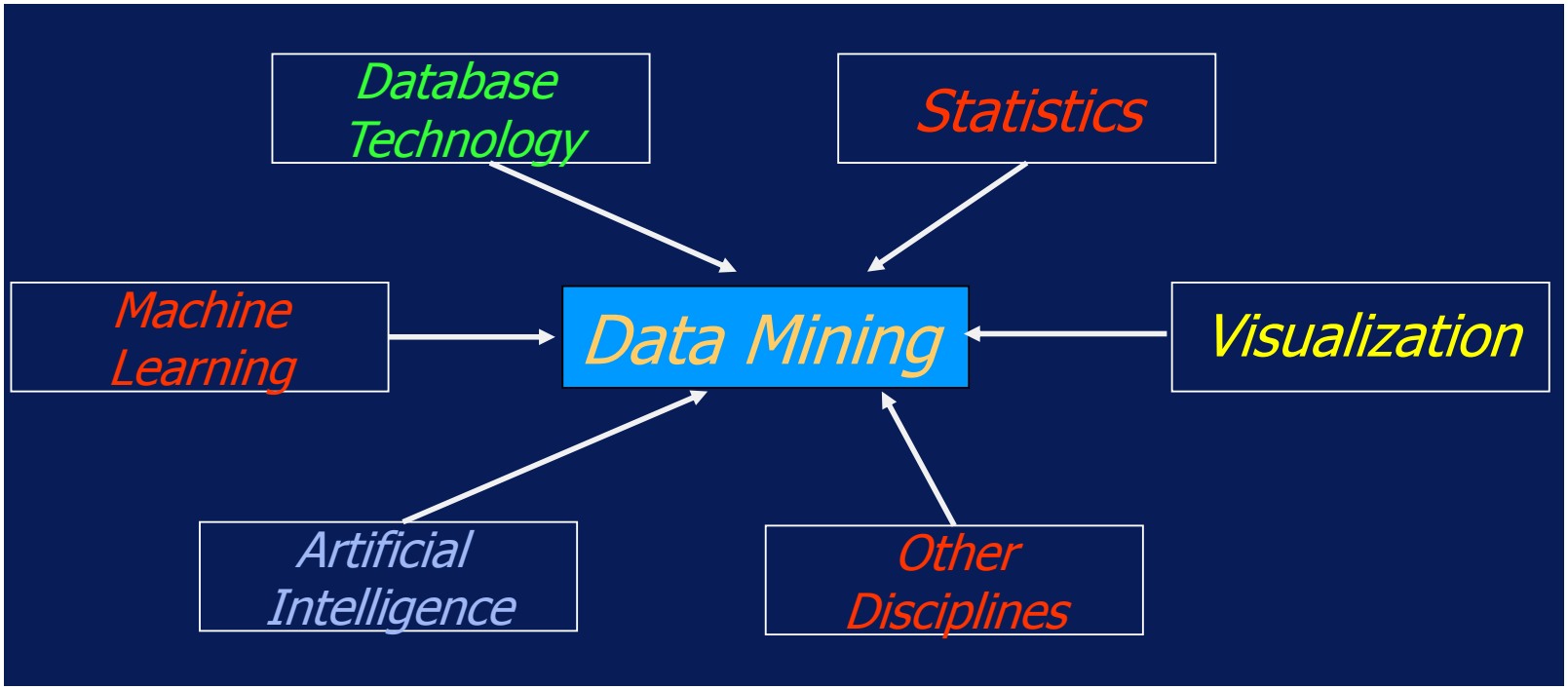


Figure: Data Mining and related fields

History of Data Mining

Emerged late 1980s
Flourished in 1990s
Roots traces along three family lines:

- Classical Statistics
- Artificial Intelligence
- Machine Learning

Classical Statistics

- Foundations of most methods: Regression analysis, standard distribution/deviation/variance, cluster analysis, confidence intervals.
- Building blocks

Artificial Intelligence

- Heuristics vs Statistics
- Human-thought-like processing
- Meta-heuristics

Machine Learning

- Algorithmic reasoning with Statistics
- Blends AI heuristics with advanced Statistical Analysis
- Learning from Data $\leftarrow$ Important

Data Science Road Map

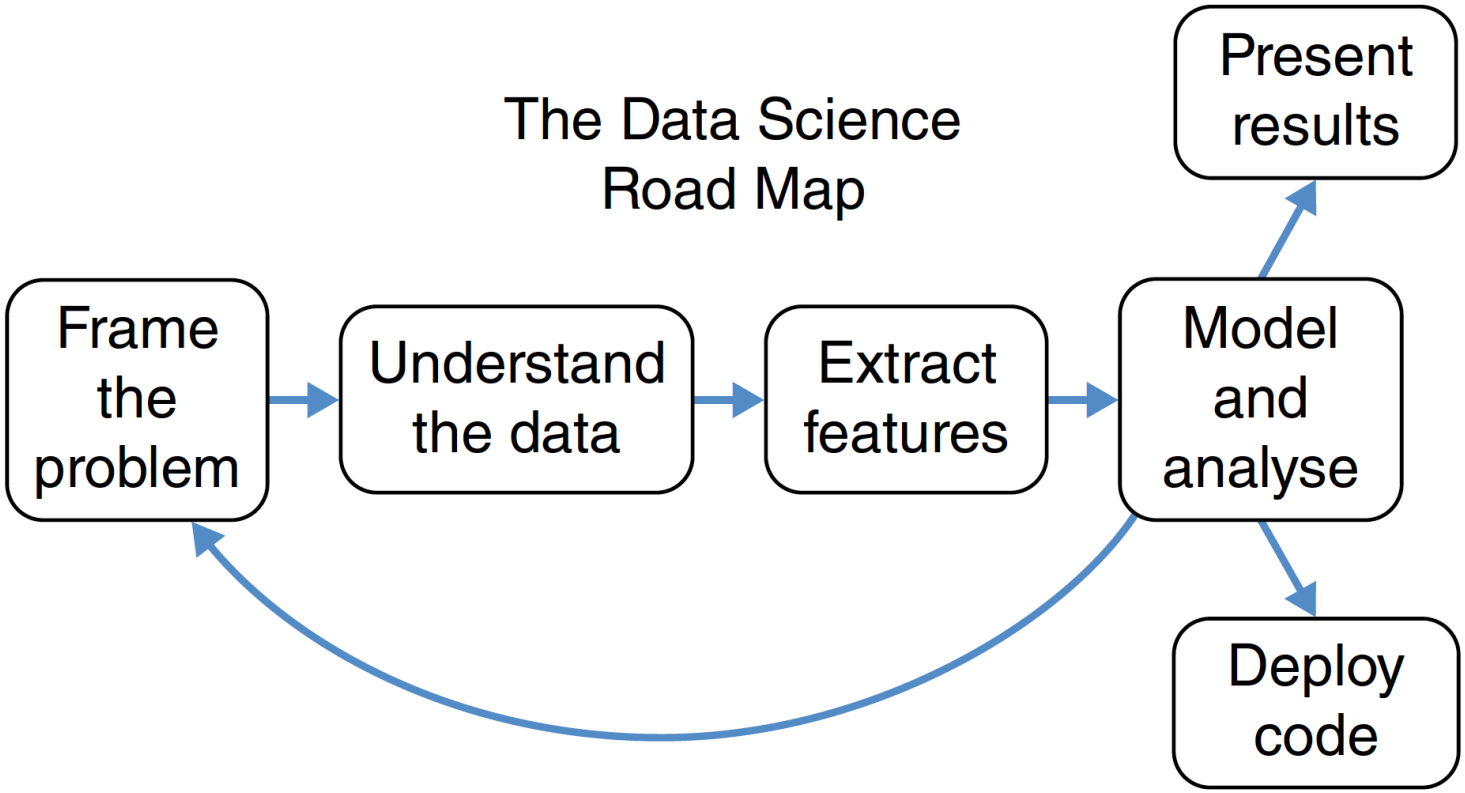


Figure: Data Science Road Map

Data Science Steps: Frame the problem

- Ask the right question.
- Understand the "business" use case and craft well-defined analytics problems.
- Define your clientes (Humans or machines).
- Define your "statement of work" (SOWs)

Data Science Steps: Understand the Data (Basic Questions)

- Have a battery of standard questions:
  - How big is the data?
  - Is it the entire dataset?
  - Is it representative of the entire problem?
  - Could there be outliers or a lot of noise?
  - Are there any fields with unique identifiers?
  - Are there missing values?
- Describe the data as soon as possible.
- Don't get too excited to get to the analytics part.

Data Science Steps: Understand the Data (Data Wrangling)

- From "raw" format to a more suitable form for conventional analytics.
- Is the main area where data scientists need skills that statisticians don't have.
- It implies accesing to databases, handle Big Data techniques to process them, performance tricks (distributed and parallel computing), etc.
- Structured vs Unstructured Data.

Data Science Steps: Understand the Data (Exploratory Analysis)

- "Poking" around the data.
- Visualization.
- Calculate correlation between variables. (or similar things)
- Most Machine Learning algorithms are not interpretable visually, so visualization is an important aspect.

Data Science Steps: Extract Features

- Most of the ML models, work with numerical vectors that describe the object under study.
- Necessary to form "tabular data".
- Expert vs All you can create (and then choose from them)
- Most creative part of Data Science.

Data Science Steps: Model

- Almost all Data Science Projects involve Machine Learning.
- The modelling stage is quite "simple": Take a battery of models, plug the data and see which one works the best.
- Carefully tune the model to fulfill your expectations (or the client's)

Data Science Steps: Present Results

- Paper?
- Written report with tables and figures.
- It is important depending on the expertise of your client, to use easy-to-follow language.

Data Science Steps: Deploy Code

If your clients are "computers", the resulting model falls into two categories:

- Batch Analytics Code
- Real-time Code (Online)

The deliverables are:

- The Code
- Documentation
- Unit tests (real-time) or sample input datasets (batch)

Data Science Steps: Iterating

- Data Science is deeply iterative
- Try to obtain results as fast as possible.
- Automate the analysis in a single script, but make as modular as possible.

Machine Learning

Machine Learning Definition

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.

Learning from data

Use observed data to estimate some unknown phenomenon.

Taxonomy

- Predictive Methods:
  Use some variables to predict some unknown or future values of other variables.
- Descriptive Methods:
  Find human - interpretable patterns that describe the data.

Supervised vs. Unsupervised

- Supervised:
  Learning in the presence of an expert/teacher.
  Training data set is labeled with a class value
  GOAL: Predict a class or value label
- Unsupervised:
  No knowledge of the output class/value
  Data is NOT labeled
  GOAL: Learn patterns/groupings

Supervised Learning

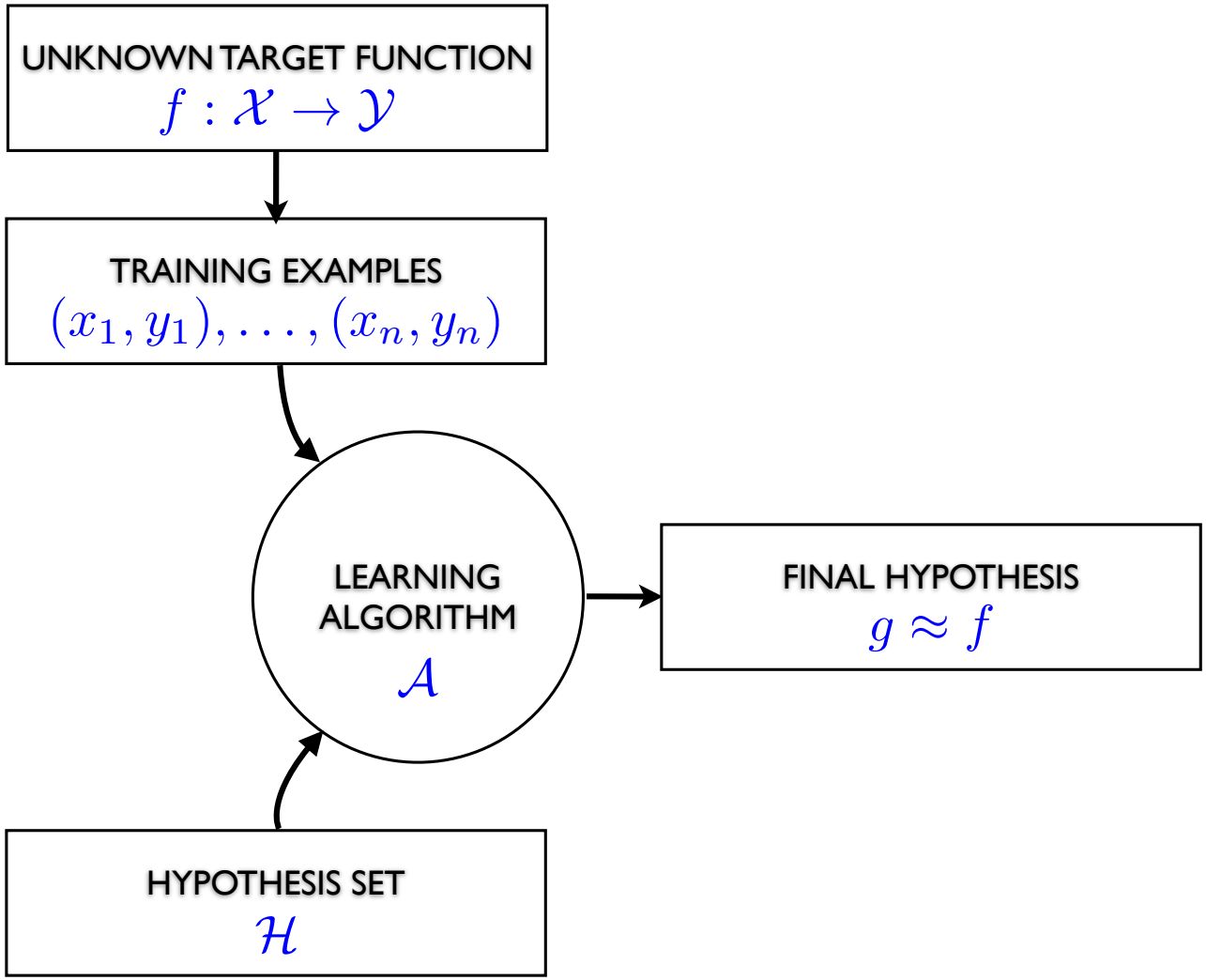


Figure: Supervised Learning

How does Machine Learning work?

Figure: Supervised Learning

What forms of Insight can DS discover?

- Predictive Modeling: Classification, Regression, Forecasting.
- Descriptive Modeling: Cluster analysis/segmentation
- Discovering Patterns and Rules: Association/Dependency rules, Sequential or Temporal sequences.
- Deviation detection: Outliers, abnormal events.

Data Science Tasks

Classification and Prediction

- Finding models (functions) that describe and distinguish classes or concepts for future prediction
- Example: Classify countries based on climate, or classify cars based on gas mileage
- Model representation: IF-THEN rules, decision tree, classification rule, neural network
- Prediction: Predict some unknown or missing numerical values.

Data Science Tasks

Association (correlation and causality)

- Multi-dimensional interactions and associations
- Example:
- age(X,"20-29") and income(X,"60-90K") $\rightarrow$ buys(X,"TV")
- Customer(area code) and buys(X) $\rightarrow$ offert(type), product(cost)

Data Science Tasks

Cluster Analysis

- Class label is unknown: Group data to form new classes
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity.

Data Science Tasks

Outlier analysis

- Data object that does not comply with the general behaviour of the data
- Mostly considered as noise or exception, but is quite useful in fraud detection, rare events analysis

Evaluation

Error on the training data vs. Performance on future/unseen data

- Simple solution: split data into training and test set
- Three sets: Training data, validation data, and test data.
- Test set: Set of independent instances that have not been used in the training process. (Assumption: Data contains representative samples of the underlying problem)

Evaluation

- Significance tests: Statistical reliability of estimated differences in performance
- Performance measures: Number of correct classifications, Accuracy of probability estimates, Error in numeric predictions.

Error estimation

- Holdout: $\frac{1}{2}$ training and $\frac{1}{2}$ testing ($\frac{2}{3}$ and $\frac{1}{3}$)
- Repeat Holdout Method: Random sampling - repeated holdout
- Cross-validation: Partition in K disjoint clusters, train $K-1$, test on remaining.
- Leave-one-out Method
- Bootstrap: Sampling with replacement.

Cross validation

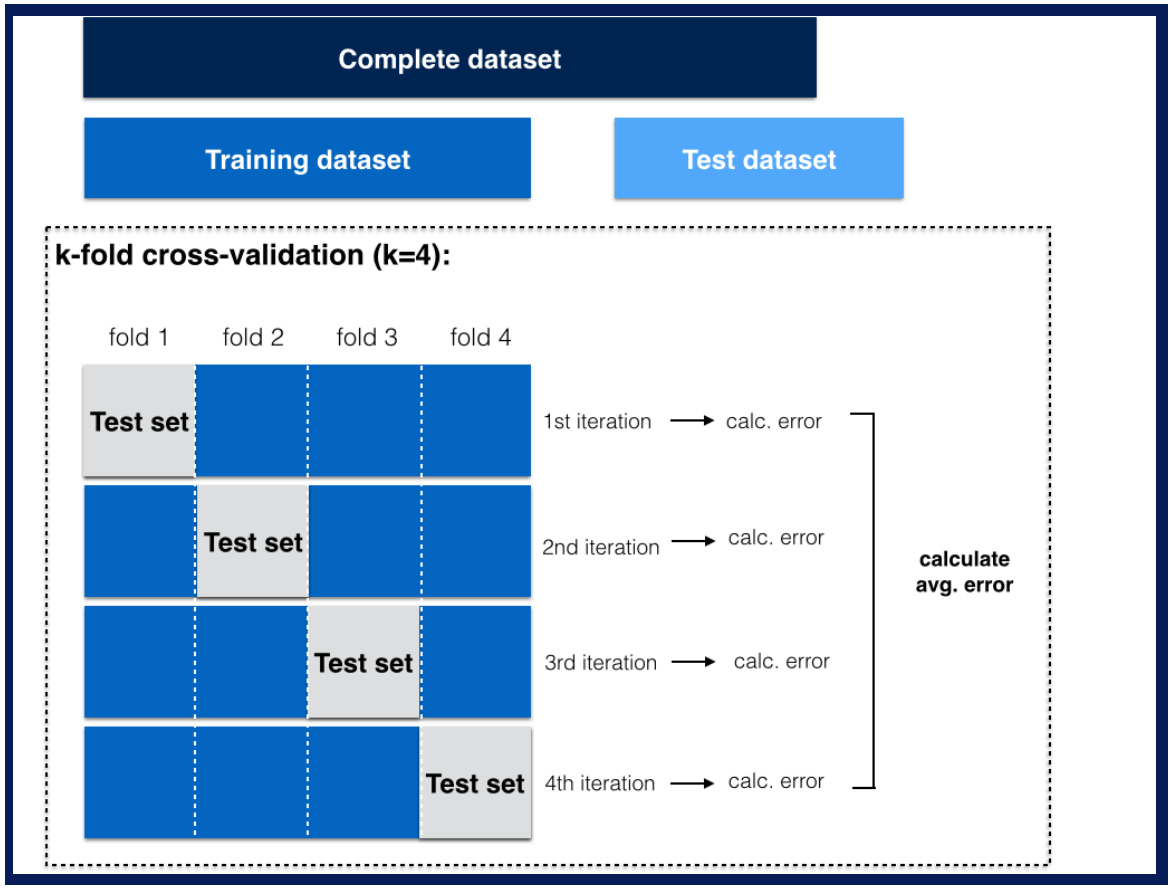


Figure: Crossvalidation process

Data Science challenges

- Computationally expensive to investigate all possibilities
- Dealing with noise/missing information and errors in data
- Mining methodology and user interaction
- Scalability of some methods.

Data Science Heuristics and Guide

- Appropiate attributes/input representation
- Minimal attribute space
- Adequate evaluation function(s)
- Extracting meaningful information
- Not overfitting

Open Source Data Science Tools

- **Python**
- R-Project
- WEKA
- KNIME
- Orange
- RapidMiner
- Rattle
- Mahout
- **Apache Spark - MLib**

Python - Scikit Learn

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search ×

scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

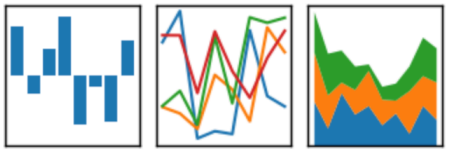
Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Fork me on GitHub

overview // get pandas // documentation // community // talks

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NUMFocus sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project.

A Fiscally Sponsored Project of
NUMFOCUS
OPEN CODE = BETTER SCIENCE

0.19.0 Final (October 2, 2016)

This is a major release from 0.18.1 and includes number of API changes, several new features, enhancements, and performance improvements along with a large number of bug fixes. We recommend that all users upgrade to this version.

Highlights include:

- `merge_asof()` for asof-style time-series joining, see here
- `.rolling()` is now time-series aware, see here
- `read_csv()` now supports parsing `Categorical` data, see here

VERSIONS

Release
0.19.0 - October 2016
download // docs // pdf

Development
0.20.0 - December 2016
github // docs

Previous Releases
0.18.1 - download // docs // pdf
0.18.0 - download // docs // pdf
0.17.1 - download // docs // pdf
0.17.0 - download // docs // pdf
0.16.2 - download // docs // pdf
0.16.1 - download // docs // pdf
0.16.0 - download // docs // pdf
0.15.2 - download // docs // pdf
0.15.1 - download // docs // pdf
0.15.0 - download // docs // pdf
0.14.1 - download // docs // pdf
0.14.0 - download // docs // pdf
0.13.1 - download // docs // pdf
0.13.0 - download // docs // pdf
0.12.0 - download // docs // pdf

ABOUT PANDAS

Apache Spark - MLib

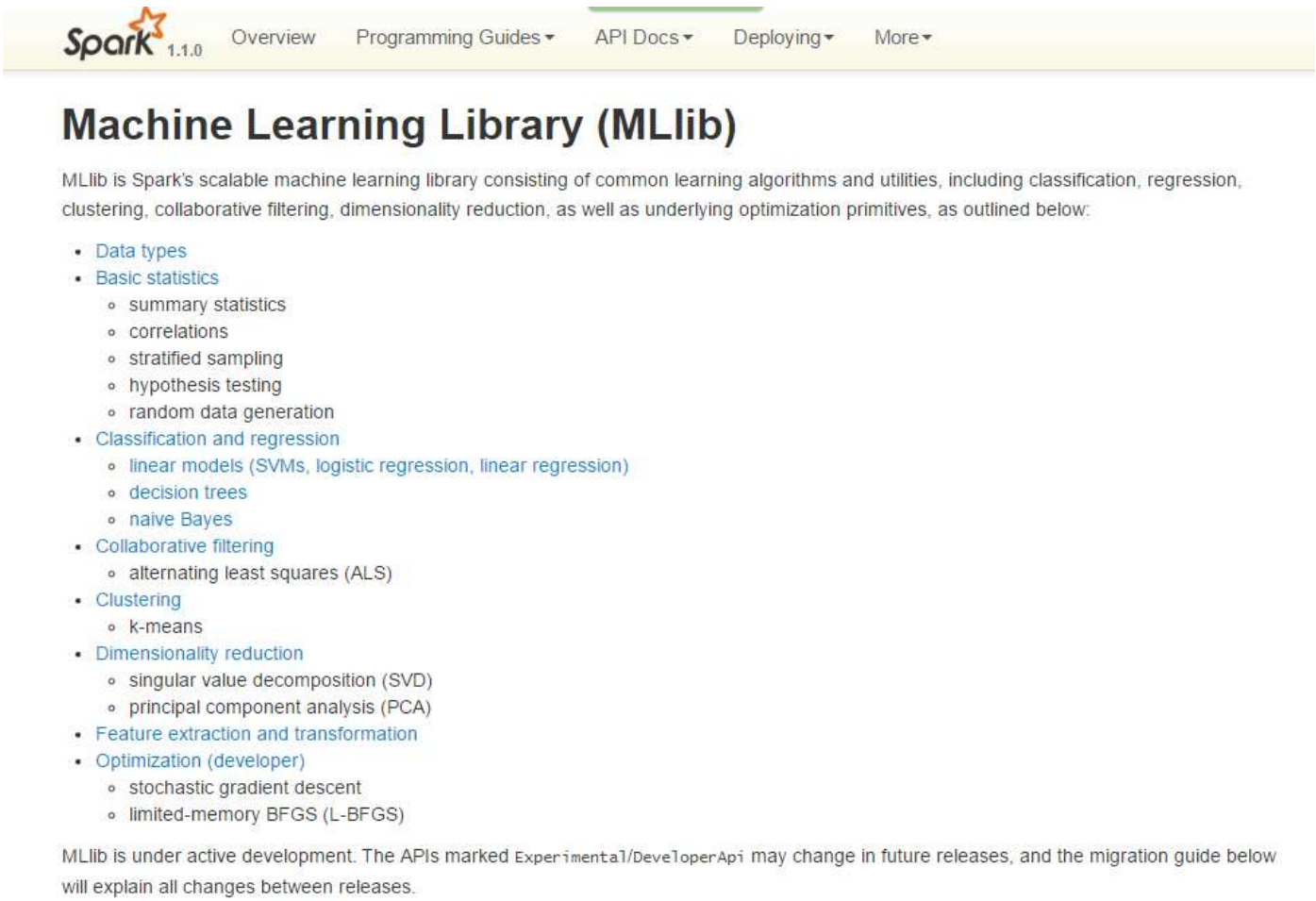


Figure: Machine Learning in Spark

Some Regression Algorithms

- Artificial Neural Networks
- Support Vector Regression
- Adaptive Neuro-Fuzzy Inference System
- Ridge Regression
- Deep Neural Networks (LSTM, Dynamic Memory Networks)

Some Classification Algorithms

- Classification Trees
- Naive Bayes
- K-Nearest Neighbors
- Artificial Neural Networks
- Support Vector Machines
- Ensemble of Classifiers
- Deep Neural Networks (Conv Nets)

Some Clustering Algorithms

- K-Means
- SOM (Self-Organizing Maps)
- Hierarchical Clustering
- DBSCAN (Density-based spatial clustering of applications with noise)
- Neural Gas
- k- Shared Nearest Neighbors

Interesting Links

- http://machinelearningmastery.com
- http://www.kaggle.com
- https://www.analyticsvidhya.com/
- http://www.datasciencecentral.com/
- Some youtube channels:
  - Siraj Raval
  - Two Minute Papers
  - Data Science Dojo
  - Open Data Science
- JULIA language http://julialang.org

Python Notebooks (Jupyter Notebooks)

https://ipython.org/notebook.html

Easiest way to start

https://www.docker.com/
https://github.com/jupyter/docker-stacks/tree/master/pyspark-notebook

How to install Docker and PySpark Notebook locally

- Go to http://www.docker.com/products/overview and download Docker installation file (Windows, MacOsX, Linux)
- Install the application with downloaded file.
- Start Docker
- In a terminal (or command prompt) download the PySpark 2.0 container: (docker pull jupyter/pyspark-notebook)'
- To run for the first time, execute (sudo docker run -d -p 8888:8888 –name PySpark2.0 jupyter/pyspark-notebook:latest start-notebook.sh –NotebookApp.token=")
- To stop the running container (docker stop PySpark2.0)
- To start it again (docker start PySpark2.0)