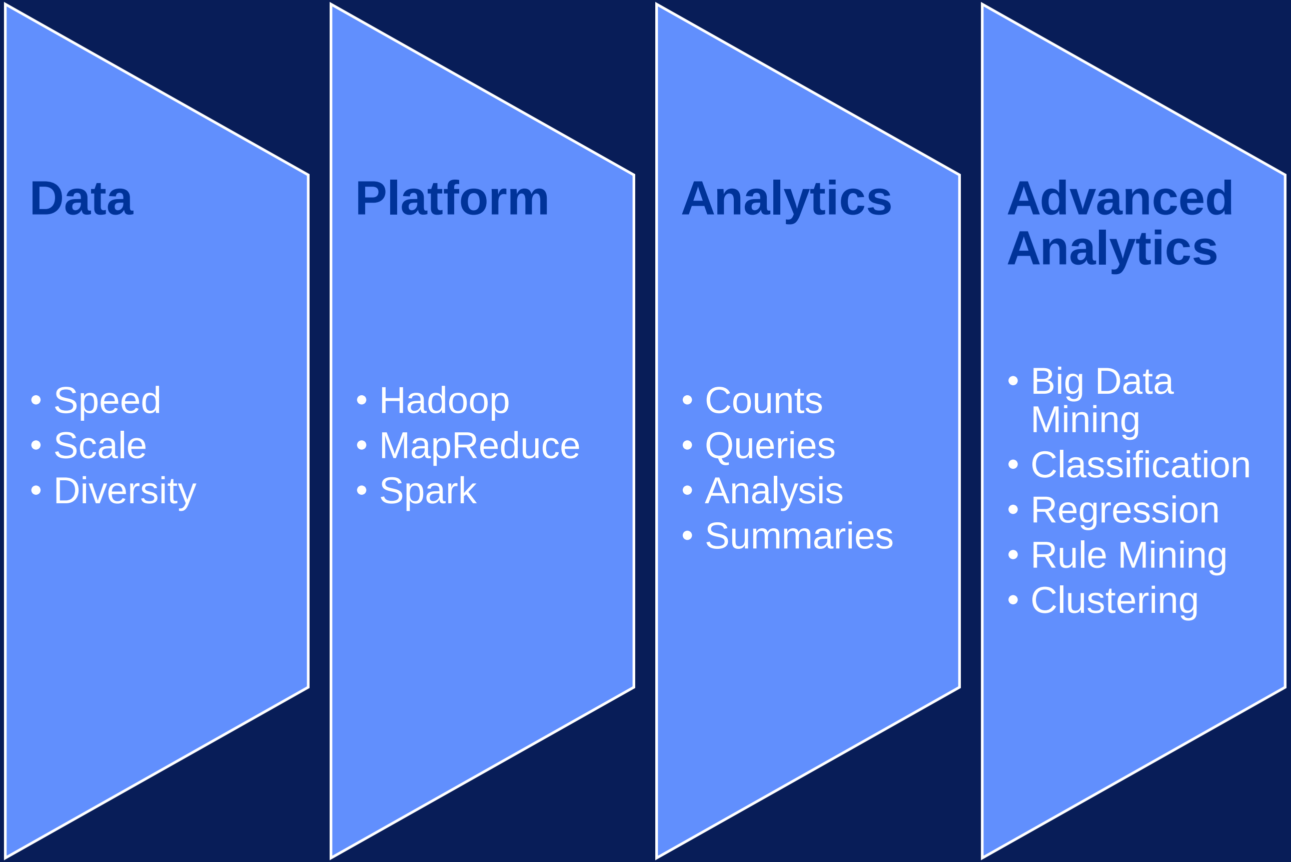
Intro to Machine Learning Lecture #1

Lesson #1

Intro

Real-time, High Performance Platform



Data

- Speed
- Scale
- Diversity

Platform

- Hadoop
- MapReduce
- Spark

Analytics

- Counts
- Queries
- Analysis
- Summaries

Advanced Analytics

- Big Data Mining
- Classification
- Regression
- Rule Mining
- Clustering

What are Big Data Challenges?

Taking this class and/or Big Data specialization!

Finding Talent

Gathering data from different sources

Understanding tools and platforms

Throughout the Big Data Specialization and in the Capstone

In this class we will learn about Data Mining tools and techniques for Big Data

Transforming Data Into Insight For Making Better Decisions

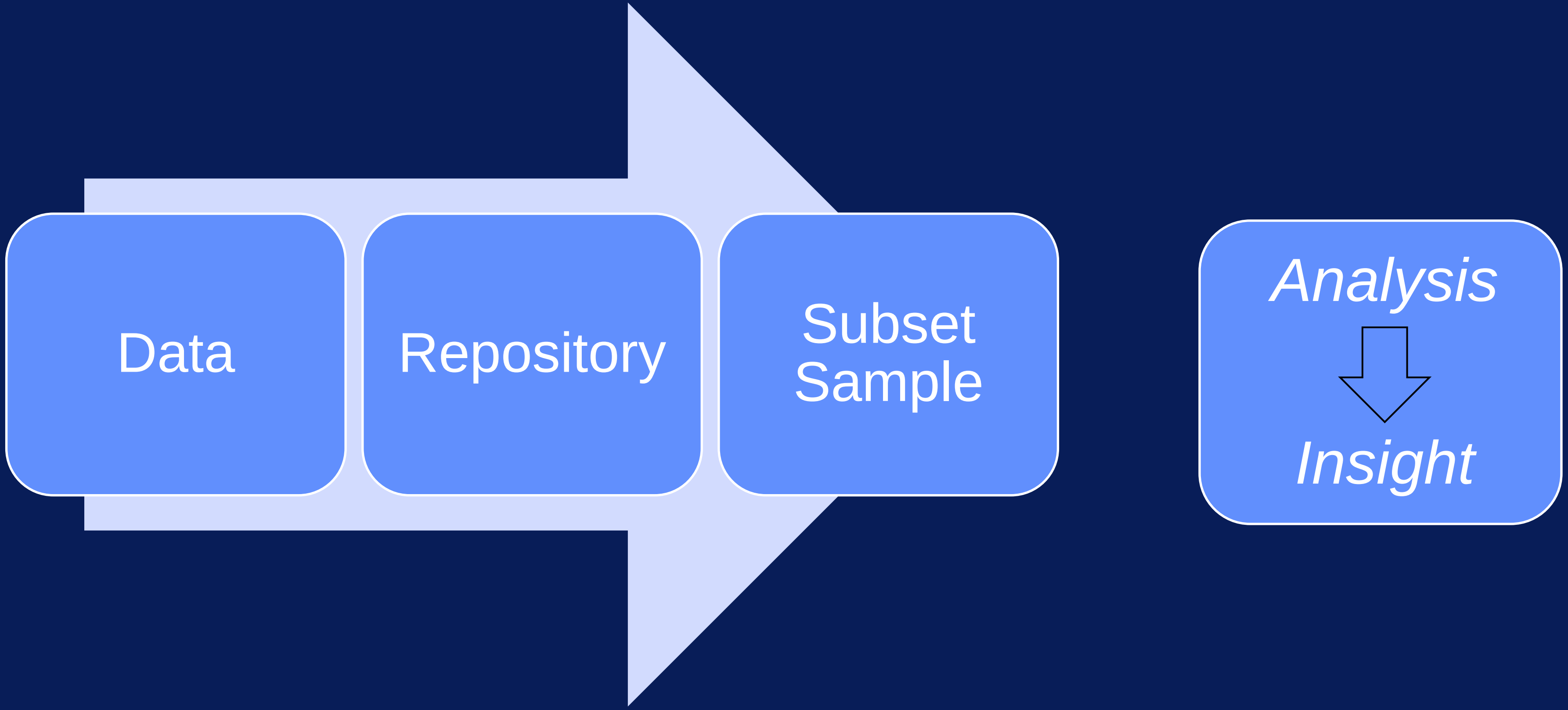
Analysis

Data Driven Decision

Data

Insight

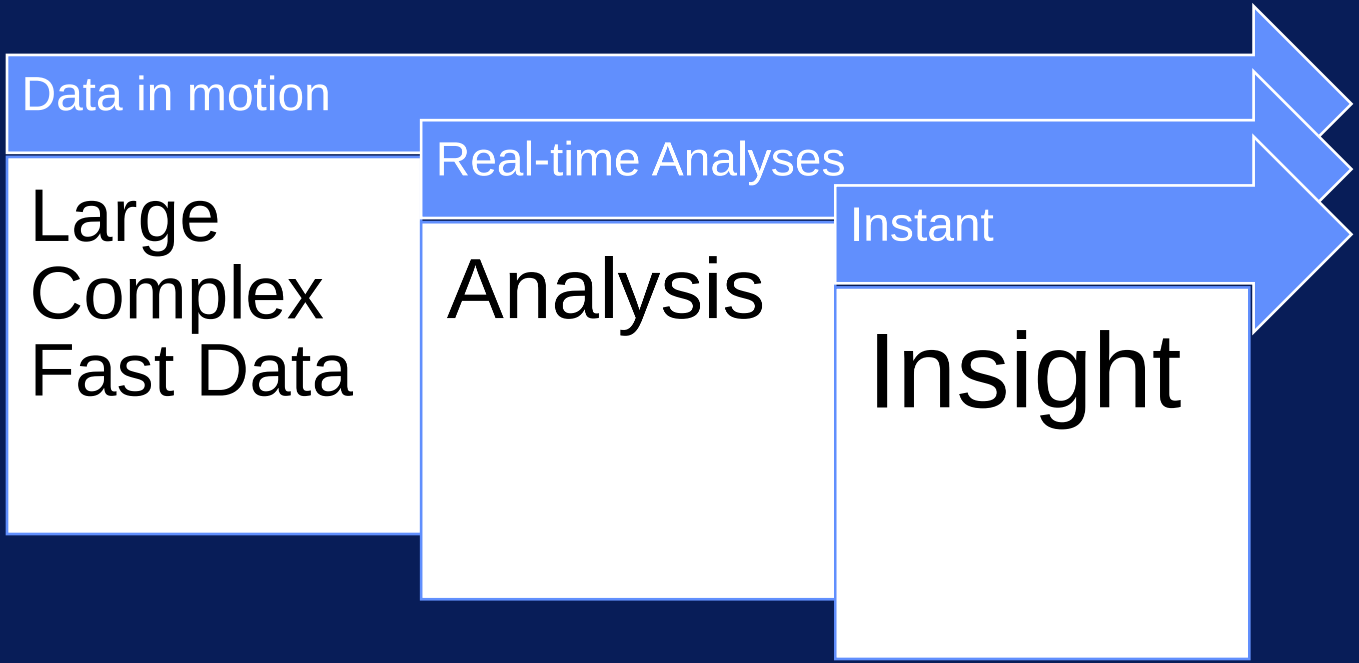
Action



Traditional Analytics Approach

Big Data Approach

Paradigm shift



Data in motion

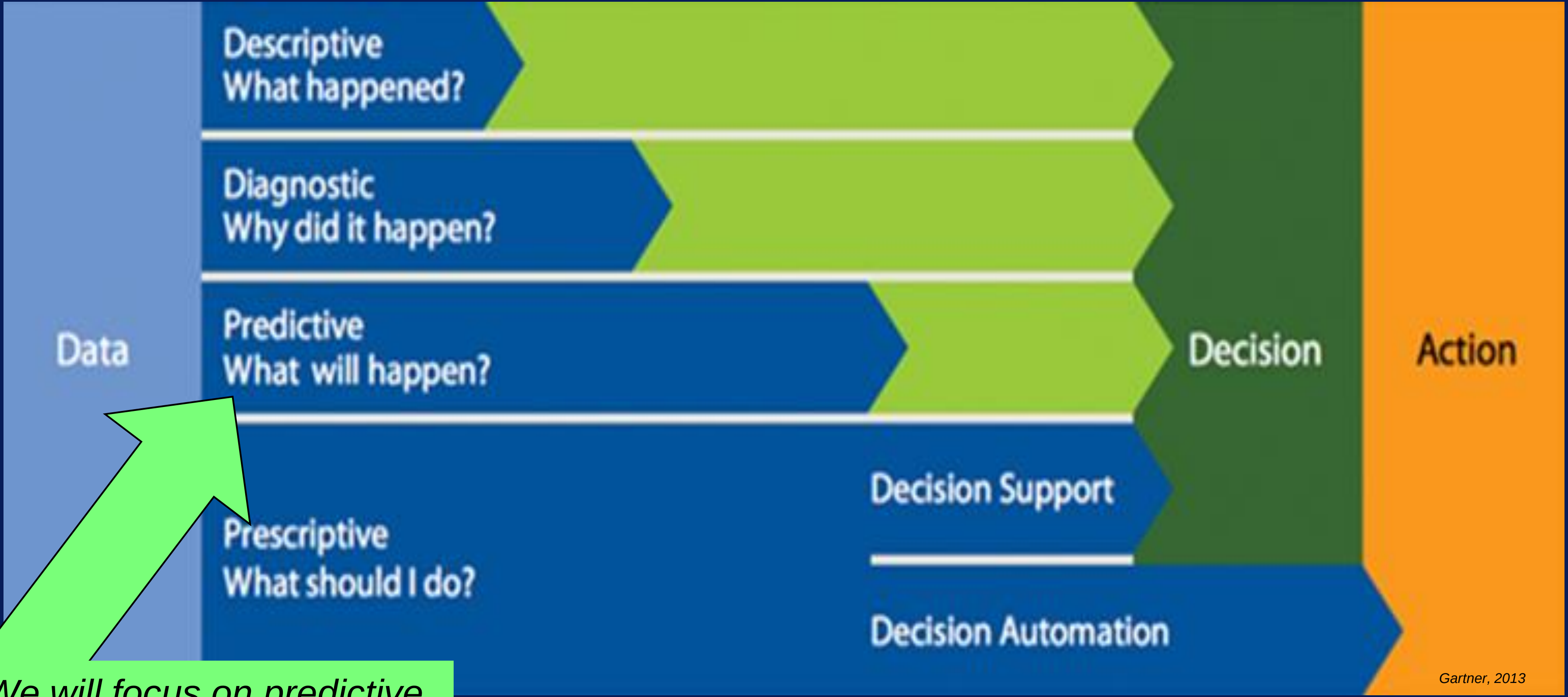
Large
Complex
Fast Data

Real-time Analyses

Analysis

Instant

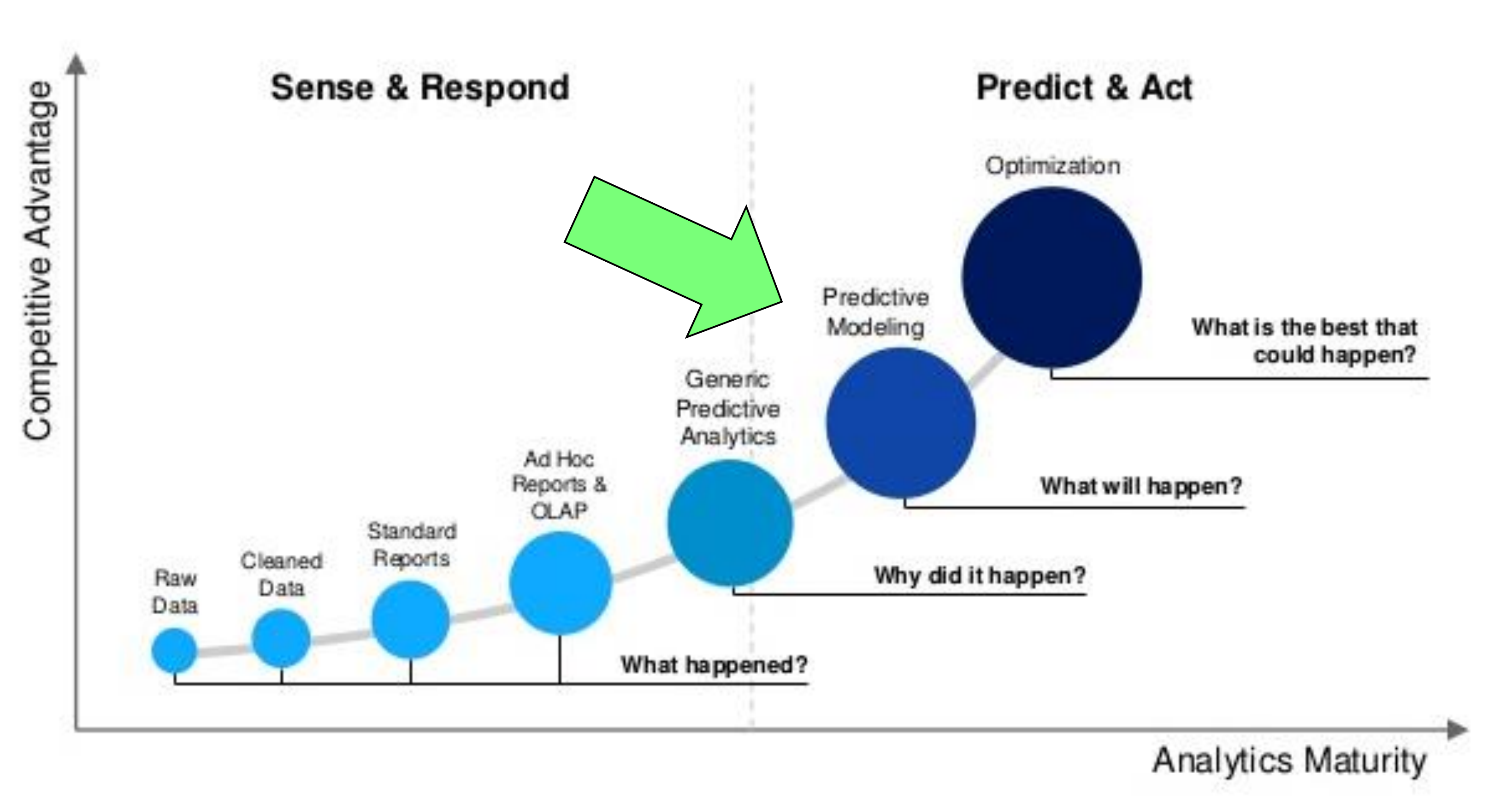
Insight



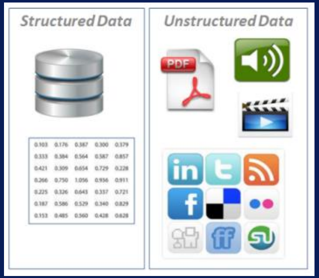
Gartner, 2013

We will focus on predictive methods in this class

Analytics Maturity Levels


Sense & Respond
Predict & Act
Competitive Advantage
Analytics Maturity
Raw Data
Cleaned Data
Standard Reports
Ad Hoc Reports & OLAP
Generic Predictive Analytics
Predictive Modeling
Optimization
What happened?
Why did it happen?
What will happen?
What is the best that could happen?

Big Data and Machine Learning



Websites
Email
RSS Feeds
Document Management Systems
Social Media
Databases
File Folders

Insight

Big Data Applications

- Data’s journey from general purpose business application to specific Big Data
- Healthcare, manufacturing, marketing
- What do they all have in common?

Big Data Applications

- Data's j ral purpose busines pecific Big Data
- Healthc g, marketi
- What d common?

At the core of most of Big Data applications is built in Machine Learning

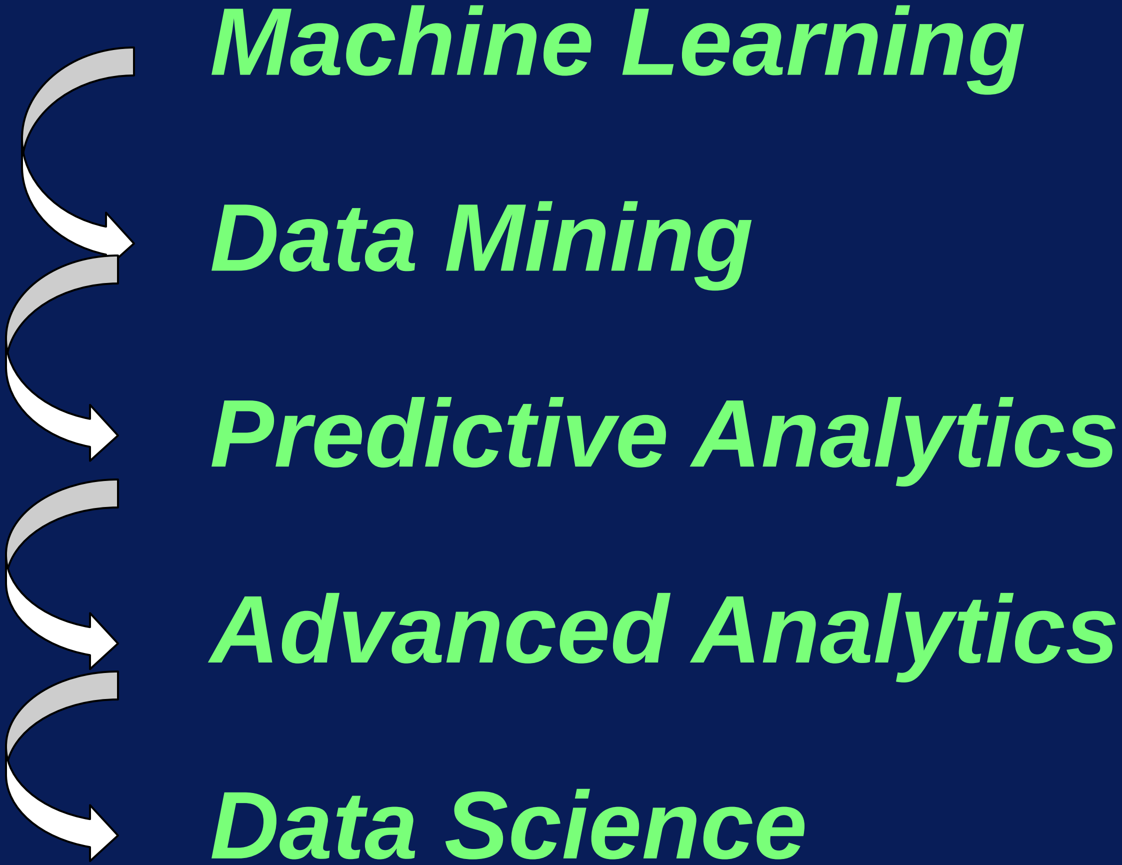
Lesson #2

Intro to Machine Learning

Data Explosion

“We are drowning in data, but starving for knowledge!”

(John Naisbitt, 1982)



Machine Learning

Data Mining

Predictive Analytics

Advanced Analytics

Data Science

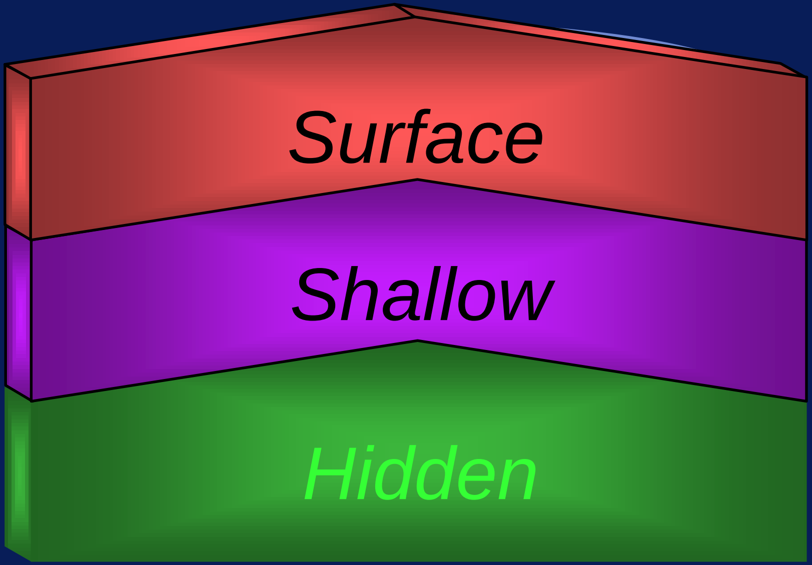
Extraction or “mining” of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Many definitions and descriptions of data mining

Extraction of implicit, previously unknown, unexpected, potentially extremely useful information from data

Top-Down Methodology

Bottom-Up Methodology

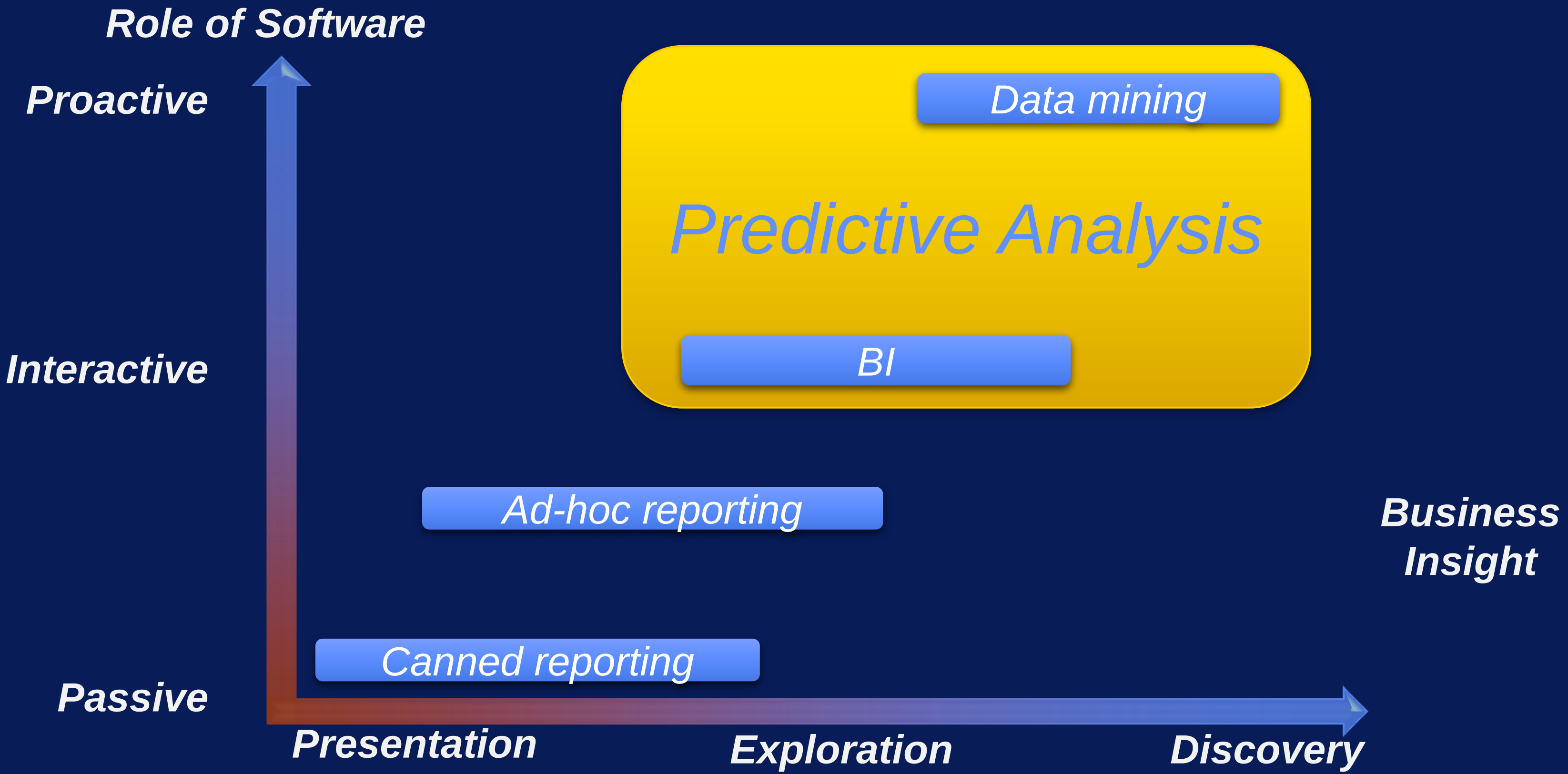


Analytical Tools

SQL tools for simple queries and reporting

Statistical & BI tools for summaries and analysis

Data Mining methods for knowledge discovery



Role of Software

Proactive

Interactive

Passive

Data mining

Predictive Analysis

BI

Ad-hoc reporting

Canned reporting

Presentation

Exploration

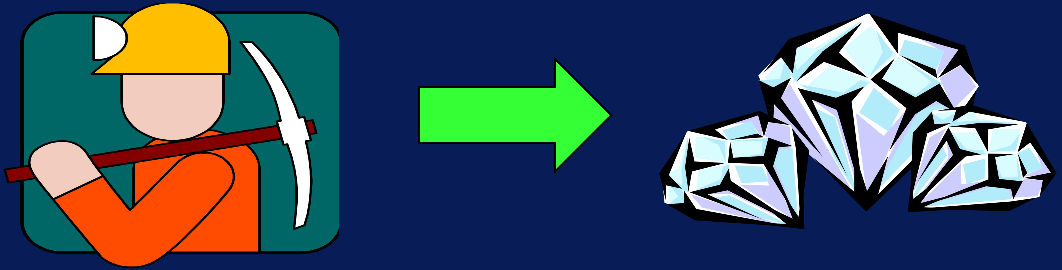
Discovery

Business
Insight

What Is Data Mining?

AI = Artificial Intelligence

Combination of AI and statistical analysis to discover information that is “hidden” in the data



What can be hidden in data?

Associations

Sequences

Classifications

Forecasting

Anomalies

Grouping/Clusters/Segments

Data Mining is NOT

Data Warehousing

Query processing

Expert Systems

Online Analytical Processing (OLAP)

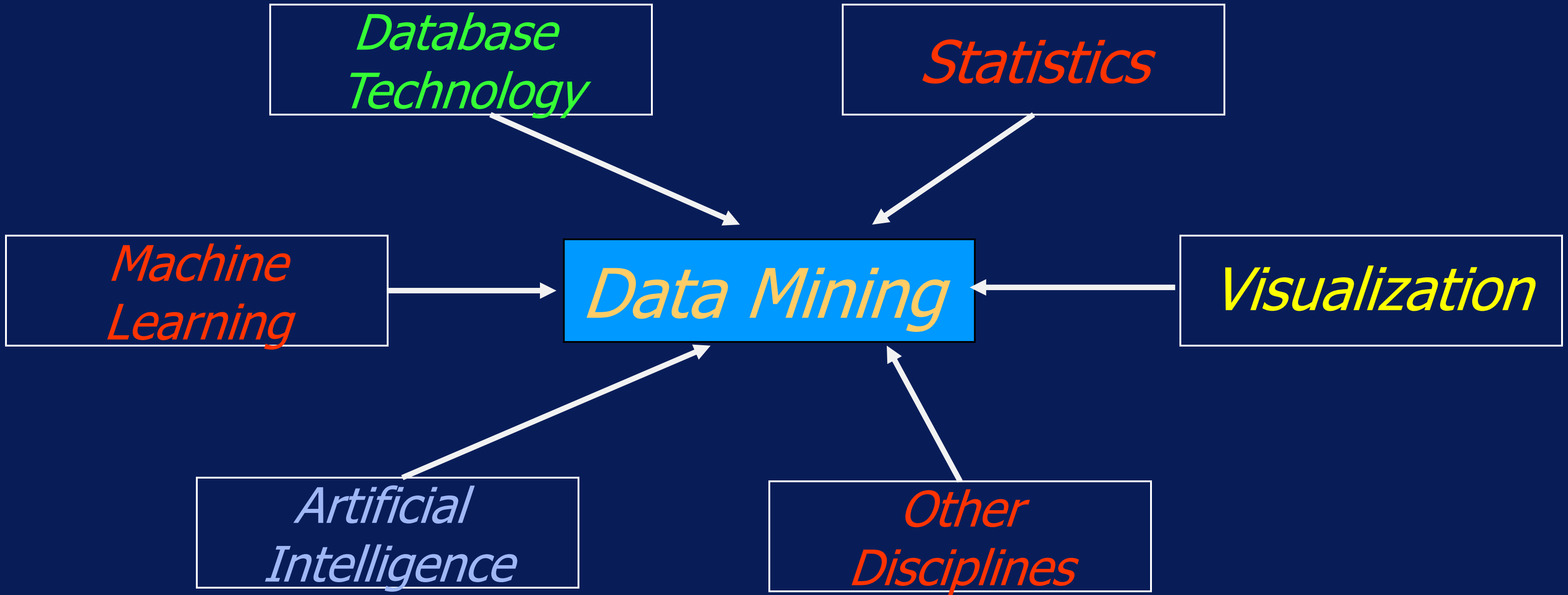
Generic Statistical Analysis Tool

Data visualization

Business Intelligence Tools (BI tools)

Generic Workflows

Multidisciplinary Field



Database Technology

Statistics

Machine Learning

Data Mining

Visualization

Artificial Intelligence

Other Disciplines

Lesson #3

History of Data Mining

History

Emerged late 1980s

Flourished in 1990s

Roots traced back along three family lines

Classical Statistics

Artificial Intelligence

Machine Learning

Statistics

Foundation of most methods

- Regression analysis, standard distribution/deviation/variance, cluster analysis, confidence intervals

Building blocks

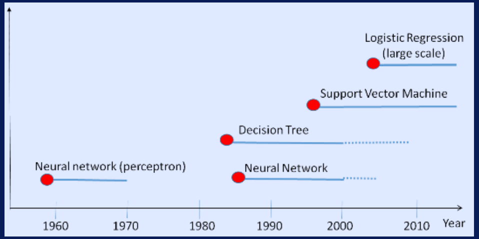
Artificial Intelligence (AI)

Heuristics vs. Statistics

Human-thought-like processing

Supercomputers

Machine Learning



Union of Statistics and AI

Blends AI heuristics with advanced Statistical Analysis

Terminology

Gold Mining

Knowledge extraction

Knowledge Discovery Databases (KDD)

Information harvesting

Business intelligence

Predictive Analytics

Data Science

Lesson #4

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 vs. Unsupervised

Learning in a presence of an expert/teacher

Training data set is labeled with a class value

Goal: Predict a class or value label

No knowledge of the output class/value

Data is NOT labeled

Goal: learn patterns/groupings

How Does Machine Learning Work?

Explore Data

Finds Patterns

Performs Predictions

What Form of Insight can Data Mining Discover?

Predictive Modeling

- Classification, Regression, Forecasting

Descriptive Modeling

- Cluster analysis/segmentation

Discovering Patterns and Rules

- Association/Dependency rules
- Sequential or Temporal sequences

Deviation detection

Many Data Mining Applications

From

Science including Chemistry, Physics, Medicine, Bioscience

Pharmaceutical, Insurance, Health care, Personalized Medicine

Energy, Sustainability, Smart City

To

Financial Industry, Banks, Businesses, E-commerce

Market analysis and management

Risk analysis and management

To more recently

Sports and Entertainment

Improve ability to classify and treat cancer, tumors, diseases

Adjust credit scores as transactions are occurring to account for risk fluctuations

Apply inferred customer social relationships to prevent churn

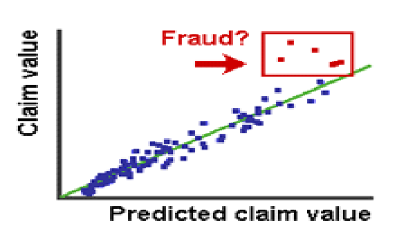
Increase revenue and customer satisfaction by discovering passengers who are likely to miss their flight

Hospital ***Loan officer*** ***Call Center*** ***Airline***

Lesson #5

Add a better picture

Data Mining 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 Mining Tasks

Association (correlation and causality)

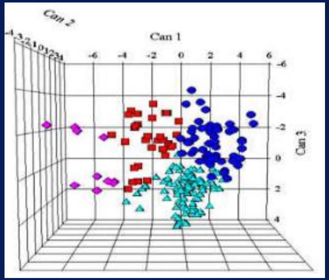
Multi-dimensional interactions and associations

Example:

age(X, “20-29”) ^ income(X, “60-90K”) -> buys(X, “TV”)

Customer(area code) ^ buys(X) ->offer(type) ^ product(cost)

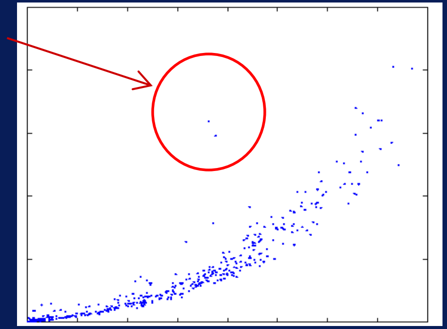
Data Mining Tasks



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

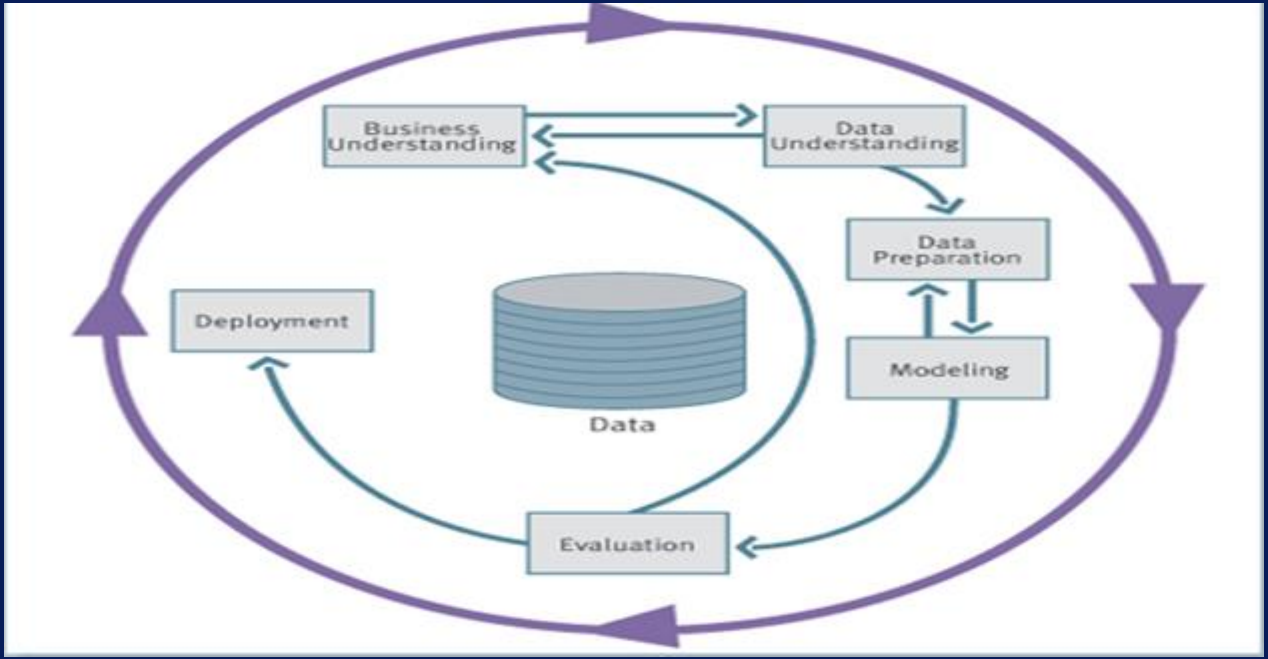
Data Mining Tasks



Outlier analysis

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

CRISP-DM - Cross Industry Standard Process for Data Mining



Lesson #6

Evaluation

Error on the training data

vs.

Performance on future/unseen data

Evaluation

Simple solution

- Split data into training and test set
- Re-substitution error
    - error rate obtained from the training data

Three sets

- training data, validation data, and test data

Training and Testing

Test set

Set of independent instances that have not been used in formation of classifier in any way

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 Methods

Holdout

- ½ training and ½ testing (2/3&1/3)

Repeated Holdout Method

- Random sampling – repeated holdout

Cross-validation

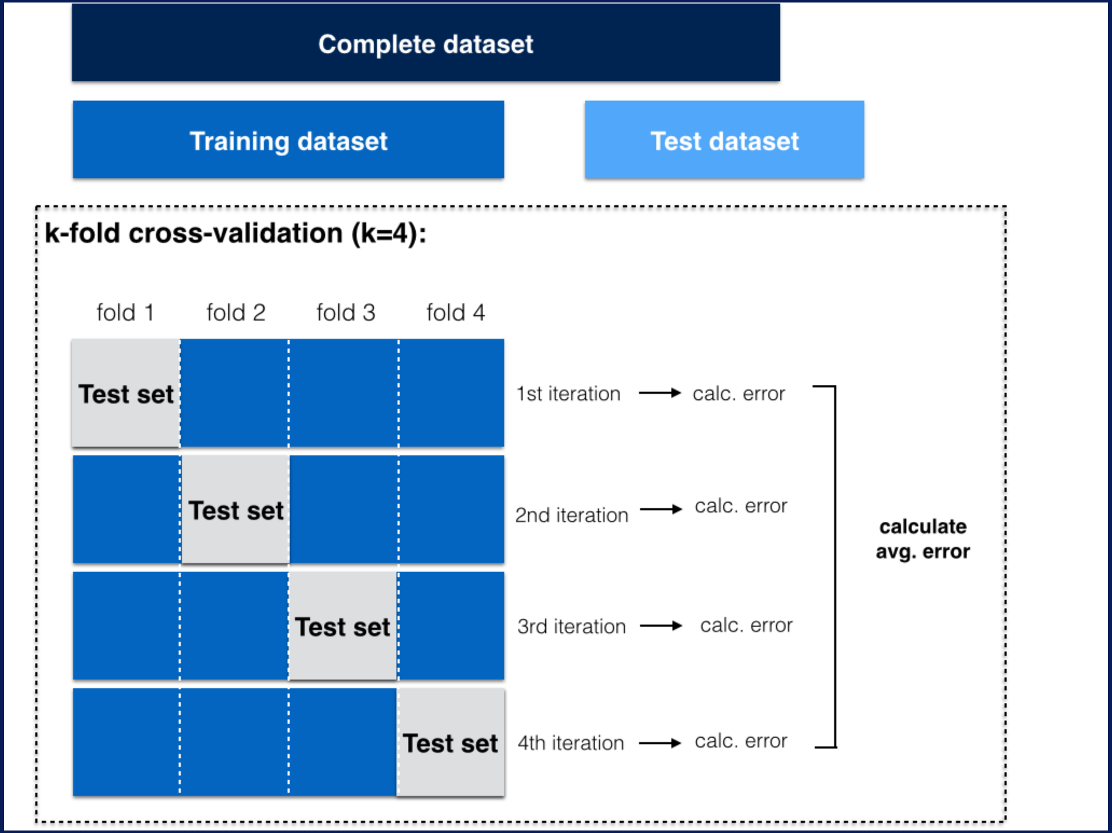
- Partition in K disjoint clusters
- Train k-1, test on remaining

Leave-one-out Method

Bootstrap

- Sampling with replacement

Cross-validation Error Estimation



Lesson #7

Data Mining Challenges

Computationally expensive to investigate all possibilities

Dealing with noise/missing information and errors in data

Mining methodology and user interaction

Scalability for some methods

Data Mining Heuristics and Guide

Appropriate attributes/input representation

Minimal attribute space

Adequate evaluation function(s)

Extracting meaningful information

Not overfitting

Open Source Data Mining Tools

- Python
- R
- WEKA
- KNIME
- Orange
- RapidMiner
- Rattle
- Mahout
- MlLib

Summary

Discovering interesting patterns from large amounts of data

CRISP-DM Industry standard

Learn from the past

- High quality, evidence based decisions

Predict for the future

- Prevent future instances of generalized patterns
- Adapt to changing circumstances

Next lecture

- Classification
- Clustering
- Association Rule Learning