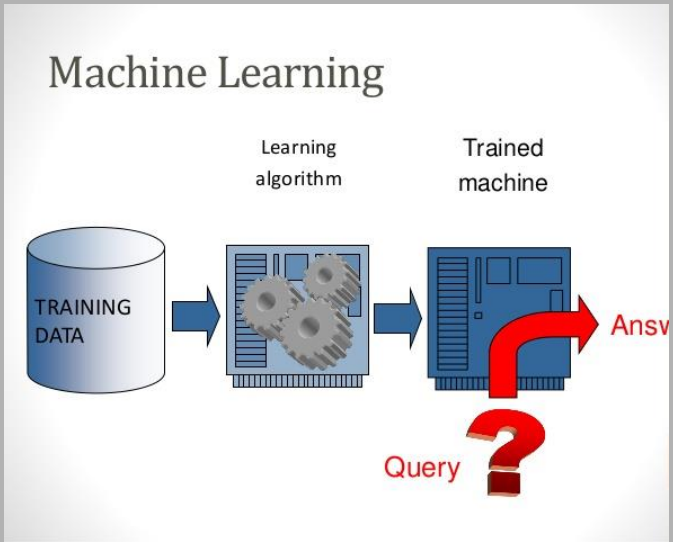
TECHNOPOLIS
THESSALONIKI ICT BUSINESS PARK

TECHNOPOLIS cluster

Μηχανική Μάθηση - Machine Learning

Από την Θεωρία στην Πράξη



Μάϊος 2017

Ioannis Vlahavas,
Dept. of Informatics,
Aristotle University of Thessaloniki

Every Minute of the Day

MasterCard Processes 74B Transactions a Year (140K/min)

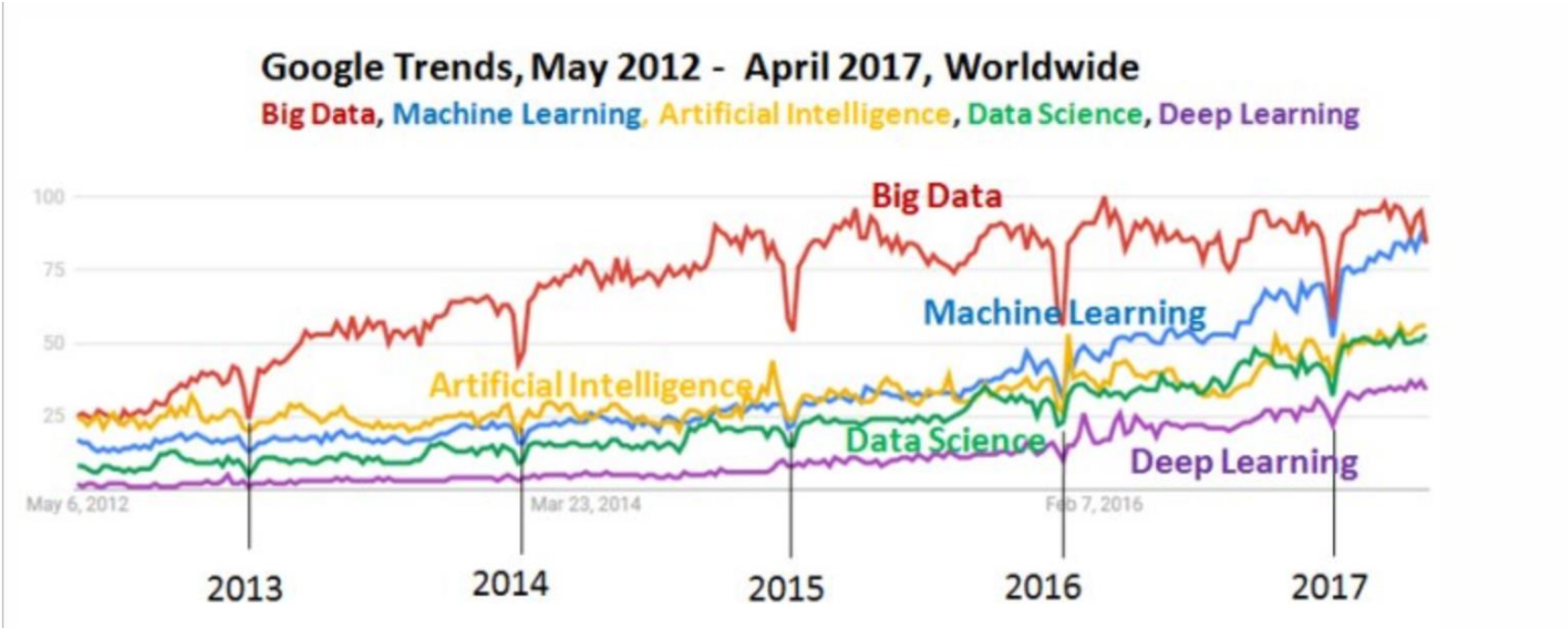
Huge Amounts of Data

- We are drowning in the deluge of data that are collected world-wide, while starving for knowledge at the same time
- *Big Data*

Wordle

**"Mining needle in a haystack.
So much hay and so little time"**

Google Trends, May 2012 - April 2017, Worldwide

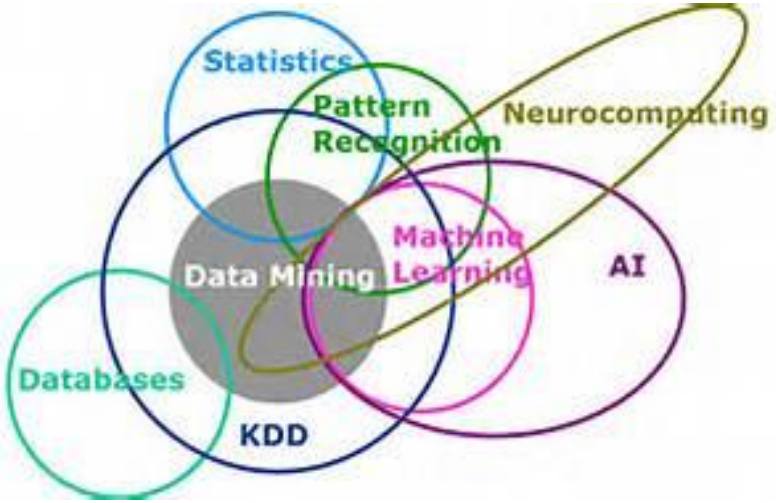


"Big Data" vs "Machine Learning" vs "Artificial Intelligence" vs "Data Science" vs "Deep Learning" search terms.

Concepts and Terminology

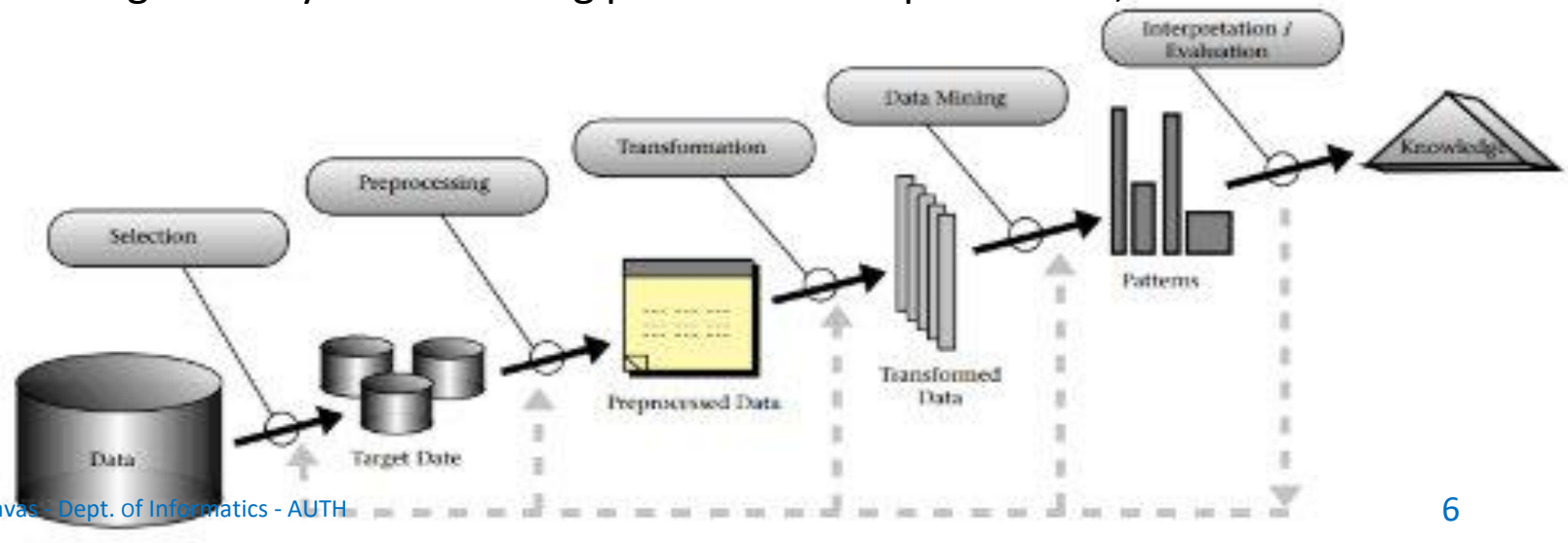
- When you are first exposing to Data Mining and Machine Learning, we think that is magic. Make significant predictions with accuracy? Sorcery!
- Curiosity, however, quickly leads you to discover that everything is above board, and sound scientific and statistical methods bear the responsibility.
- But this ends up leading to more questions in the short term. Knowledge Discovery in Databases (KDD), *Machine Learning, Data Mining, Statistics, Data Science.*

- The concepts and terminology are overlapping and seemingly repetitive at times.



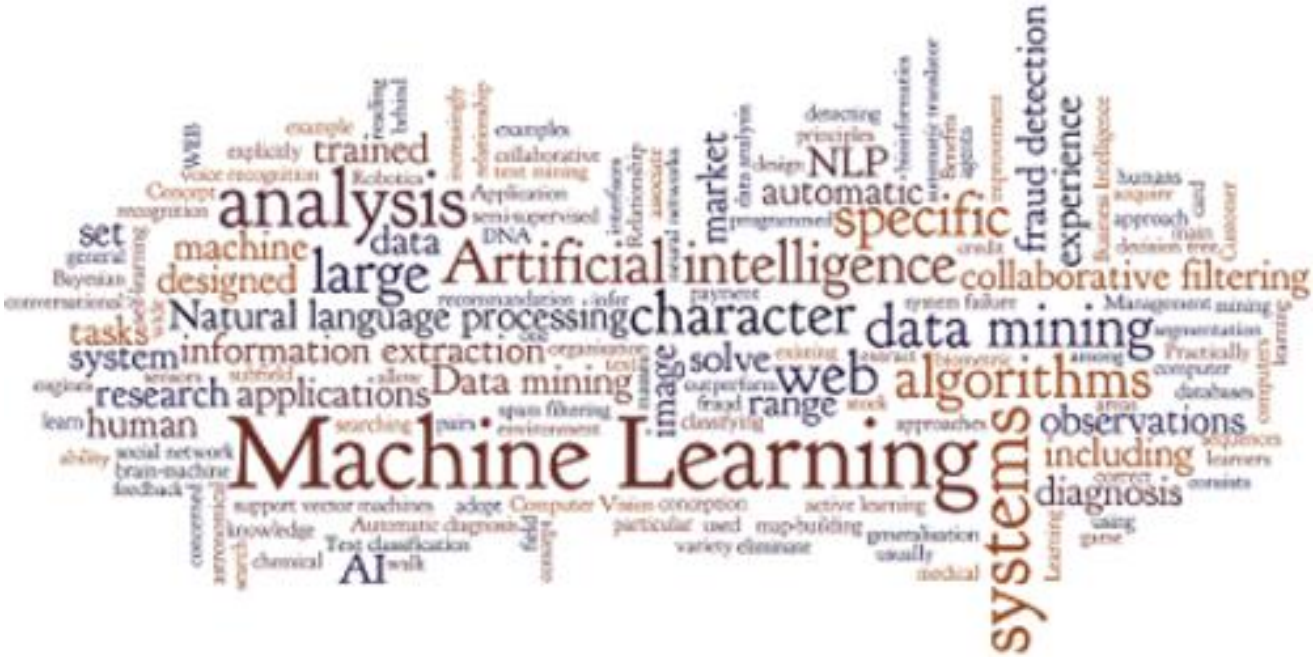
Knowledge Discovery in Databases (KDD)

- Is a multi-faceted process comprising 7 steps.
- Encompasses a variety of tasks that are not statistical in nature
  - Covers the entire process of data analysis, including data cleaning and preparation and visualization of the results, and how to produce predictions in real-time, etc.
  - Tasks are of central importance and when combined can account for more than 60% of all KDD time spent
- Includes a single step called “Data Mining" which is the application of Machine Learning methods used in the mining of data
  - Helping to distinguish between random noise and significant findings, and providing a theory for estimating probabilities of predictions, etc.



Terminology cont.

- The KDD Process was actually intended to describe a Knowledge Discovery in Databases process,
  →which kind of eventually became ***Data Mining***
  →which kind of eventually became ***Data Science***s



Data Mining VS Statistics

- Statistics is the analysis, interpretation and presentation of numeric facts or data
  - The field of statistics employs numerous statistical methods for accomplishing these goals

- Data mining is a multi-disciplinary field, the origins of which grew out of database technology, machine learning, artificial intelligence and statistics, among other fields
  - Data mining is the process of extracting hidden and previously unknown patterns from raw data, with the intent of turning these vast amounts of data into useful information

Machine Learning VS Statistics

- Machine Learning requires no prior assumptions about the underlying relationships between the variables. You just have to throw in all the data you have, and the algorithm processes the data and discovers patterns, using which you can make predictions on the new data set.
  - Machine learning treats an algorithm like a black box, as long it works
  - It is generally applied to high dimensional data sets
  - the more data you have, the more accurate your prediction is
  - In machine learning, proper validation is important to know which algorithm performs best on the data
  - is a subfield of computer science and artificial intelligence
- In contrast, statisticians must understand how the data was collected, statistical properties of the estimator (p-value, unbiased estimators), the underlying distribution of the population they are studying and the kinds of properties you would expect if you did the experiment many times.
  - You need to know precisely what you are doing and come up with parameters that will provide the predictive power
  - Statistical modeling techniques are usually applied to low dimensional data sets
  - In statistics, validation is necessary to be sure whether the conclusions drawn from the data are true or not.
  - is a subfield of mathematics.

Theory: Machine learning

- It gives "computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)
- As of 2016, is a buzzword
- Is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is *difficult or unfeasible*
- Allow data scientists to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.
- Is typically classified into three broad categories

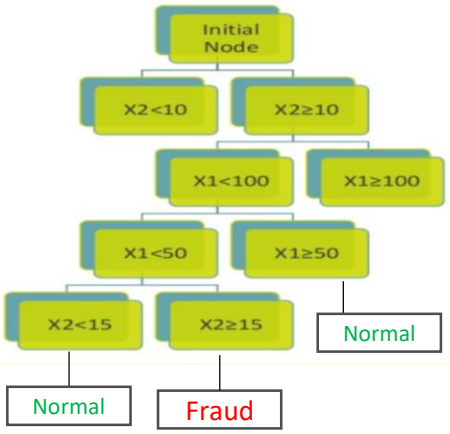
Theory: Supervised learning

- The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule (model) that maps inputs to outputs.
  - Inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes
- Model is used for prediction
- In commercial use, this is known as predictive analytics
- According to the desired output we distinguish two types of supervised learning:
  - Classification, when the outputs are discrete
  - Regression, when the outputs are continuous

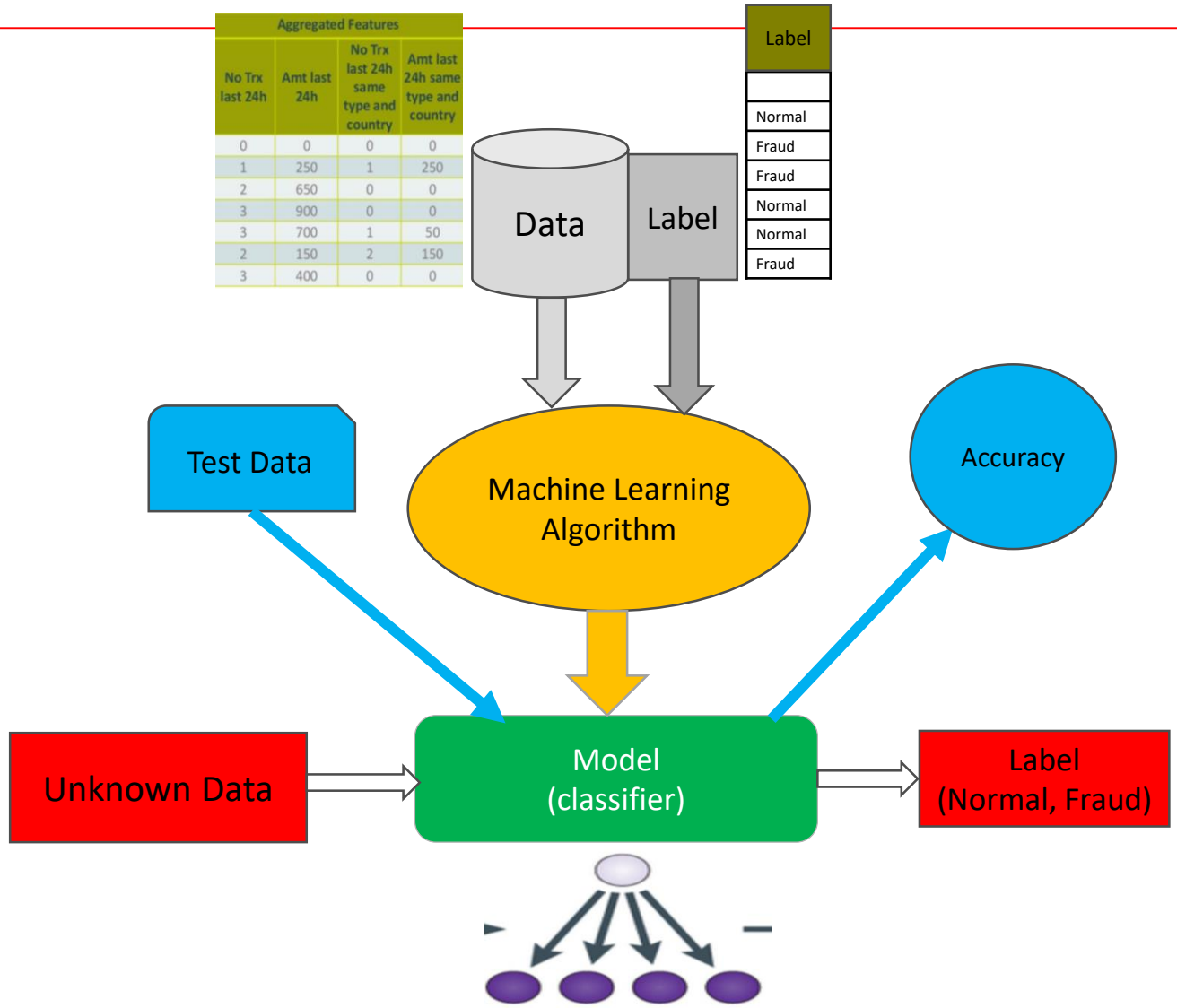
Theory: Supervised learning

- Example applications include
  - email filtering,
  - detection of network intruders or malicious insiders,
  - optical character recognition (OCR)
  - recommender systems
  - sentiment analysis
  - online advertising
  - computer vision

- Algorithms: Decision Trees, Nearest Neighbor, Neural Networks, Rule Learning, Support Vector Machines, ….

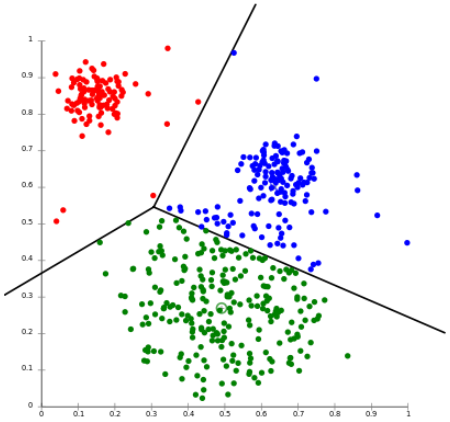


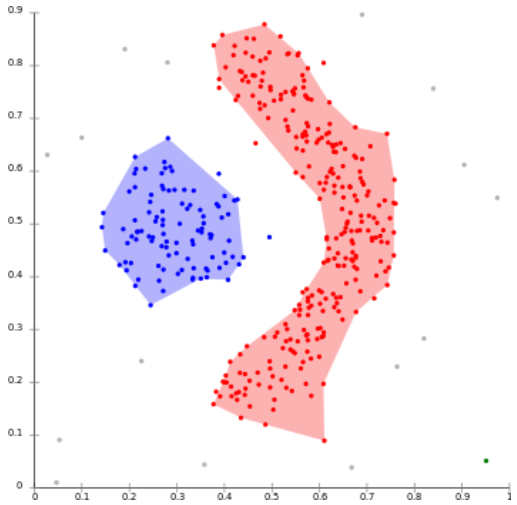
How Machine Learning works?



Theory: Unsupervised learning

- No labels are given to the learning algorithm, leaving it on its own to find hidden patterns in data
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task
- Algorithms: K-means, EM, DBSCAN, …





Theory: Reinforcement learning

- Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent).
  - The program is provided feedback in terms of rewards and punishments as it navigates its problem space.
  - Algorithms: Q-Learning, SARSA, R-Learning

Machine Learning (ML) Algorithms – concl.

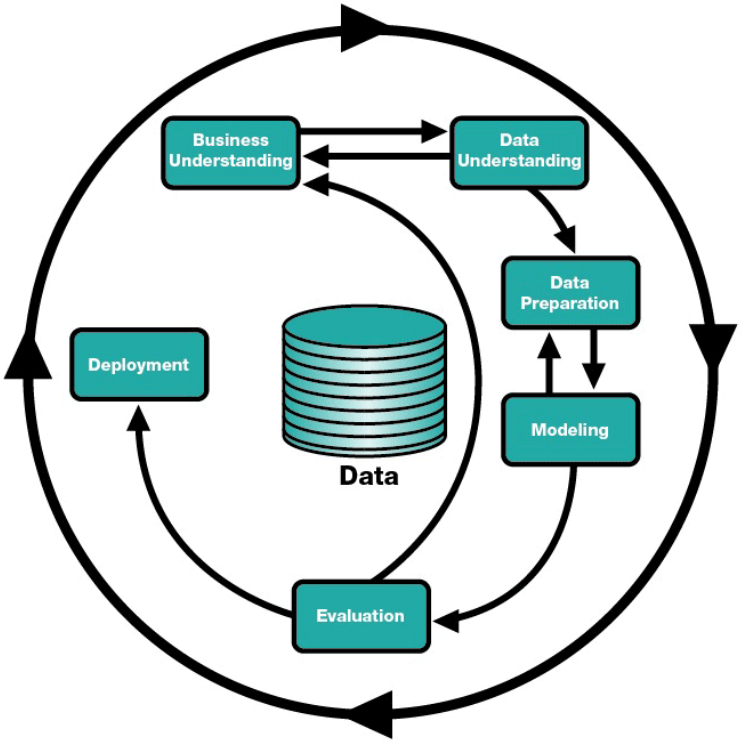
- Hundreds of ML algorithms available
- The best one does not exist (No-free lunch theorem)
- Some have better performances under certain conditions
  - e.g. Several studies have reported that Random Forest is the most accurate for fraud detection

- Which Software should we use?
  - Weka, DBminer?
  - Python?
  - R appears to be the standard between data scientists
    - Open source (Free) software and developed by academics

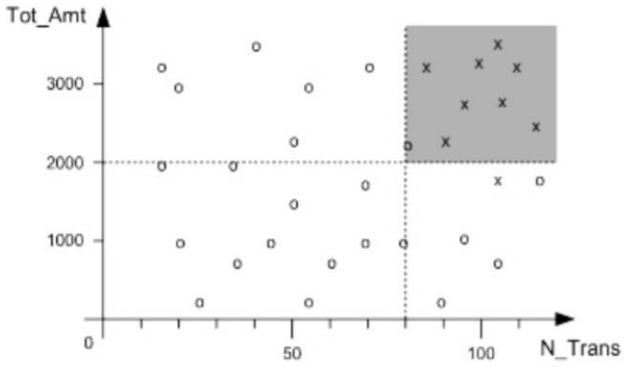
- Because finding patterns is hard, often not enough training data is available, and also because of the high expectations it often fails to deliver

CRISP-DM

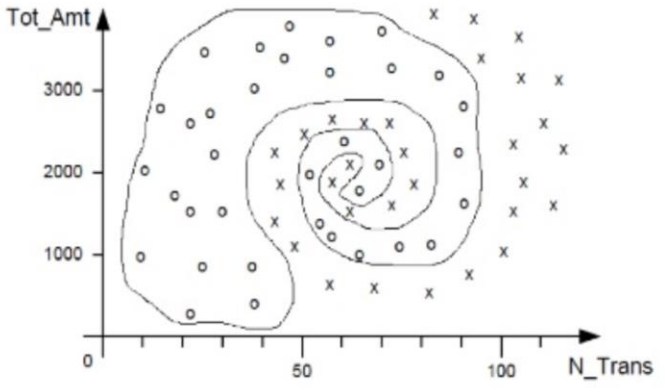
- **CRISP-DM (Cross Industry Standard Process for Data Mining)**, is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.
- CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects.



Example: Fraud detection-Expert Rules approach



Rule: IF N_Trans > 80 AND Tot_Amt > 2000 THEN fraud

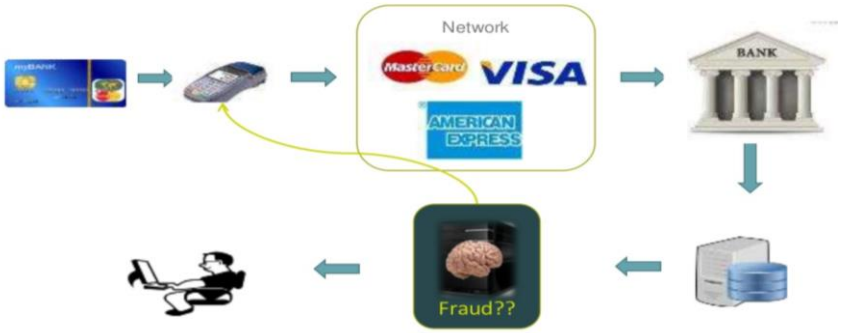


Rule: ?? We can learn this by means of **Machine Learning**

Example: Fraud detection-Machine Learning Approach

- Machine Learning can learn automatically rules able to find fraudulent patterns

IF COUNTRY=USA & LANGUAGE=EN & HAD_TEST=TRUE & NB_TX>10 & GENDER=MALE & AGE> 50 & ONLINE=TRUE & AMOUNT>1,000 & BANK=XXX
THEN class = fraud



Outlier (or Anomaly) Detection. Examples of Real World

Anomalies translate to significant (often critical) real life entities, e.g.

- Network /Cyber Intrusions
  - A web server involved in *ftp* traffic

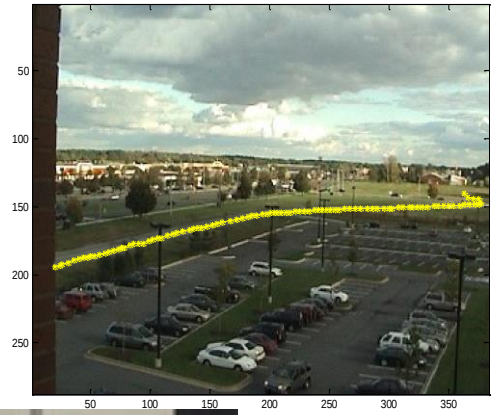
- Credit Card Fraud
  - An abnormally high purchase made on a credit card

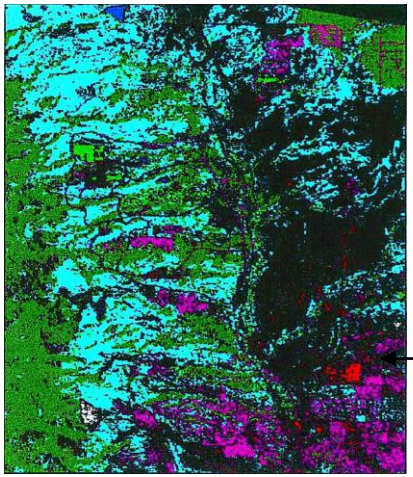
- Healthcare Informatics / Medical diagnostics
  - Detect anomalous patient records
- Industrial Damage Detection
  - Detection of faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, abnormal energy consumption, etc.

- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- …

Image Processing

- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large

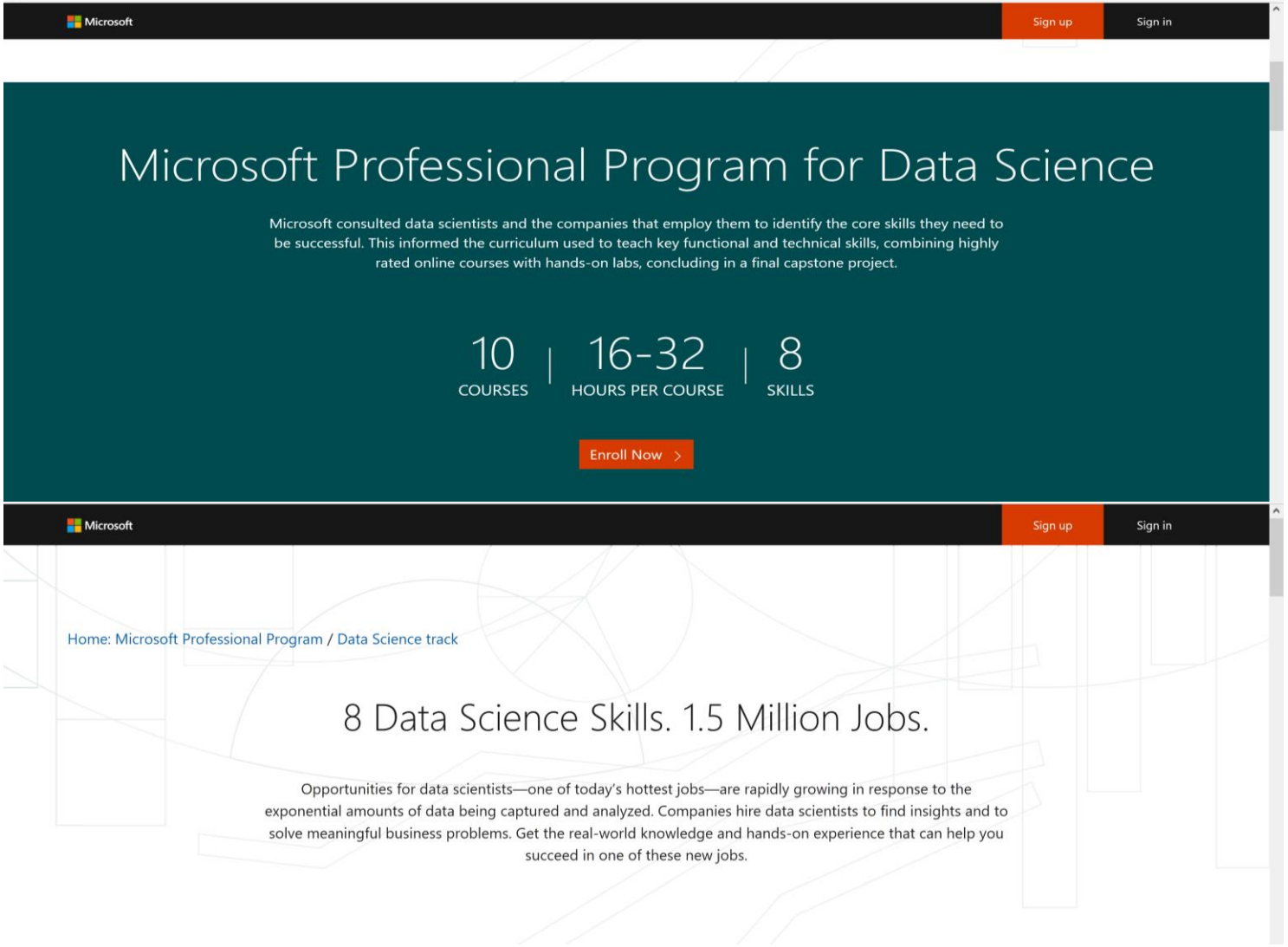




ML Applications – concl.

- Machine Learning (CS 229) is the most popular course at Stanford. Why? Because, increasingly, machine learning is eating the world.
- Machine learning is a powerful artificial intelligence tool that enables us to crunch petabytes of data and make sense of a complicated world. And it's transforming a wide variety of industries. It's solving previously unsolved problems.
- You may have heard that today's tech companies are using machine learning to identify and filter email spam (Google), blacklist and penalize spam blogs so that users get good search results (also Google), recommend products specifically for you (Amazon), and fight fraud (IBM).
- Text Mining, Energy, Business Intelligence, Marketing
  - *…The next talks*

Microsoft Professional Program for Data Science



Microsoft Professional Program for Data Science

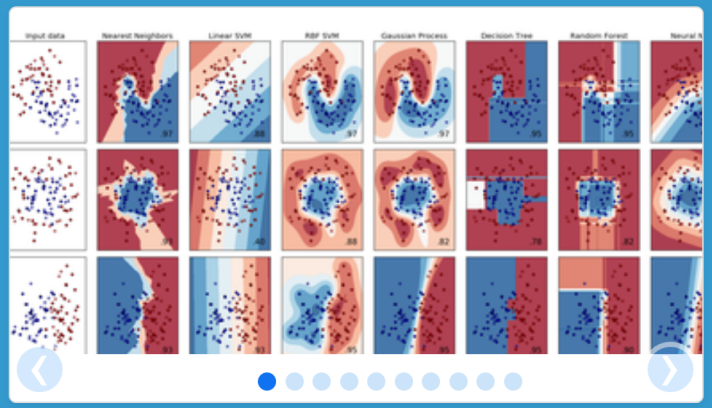
Technologies you can learn

T-SQL | Microsoft Excel | PowerBI | Python | R | Azure Machine Learning | HDInsight | Spark

- Principles of Machine Learning: Learn how to build, evaluate, and optimize machine learning models; including classification, regression, clustering, and recommendation.
  - Learn how to apply machine learning to solve common predictive problems, including text analytics, spatial data analysis, image processing, and time series forecasting.
- Explore Transact-SQL. Go from your first SELECT statement through to implementing transactional programmatic logic. Focus on querying and modifying data in Microsoft SQL Server or Azure SQL Database.
- Explore tools in Excel that enable the analysis of more data than ever before, with improved visualizations and more sophisticated business logic.
- Power BI is a suite of *business analytics tools* that deliver insights throughout your organization. Connect to hundreds of data sources, simplify data prep, and drive ad hoc analysis. Produce beautiful reports, then publish them for your organization to consume on the web and across mobile devices.

Microsoft Professional Program for Data Science

- Learn key concepts and techniques used to perform data science; including statistical analysis, data cleansing and transformation, and data visualization with:
- **The R programming language** (a free software environment for statistical computing and graphics)
  - Analyzing Big Data with Microsoft R Server
- Programming with **Python** for Data Science,
  - Libraries: *pandas*, IPython, statsmodels and scikit-learn.
  - *More work is still needed to make Python a first class statistical modeling environment, but we are well on our way toward that goal*
- Microsoft Azure Machine Learning
  - Learn how to use Spark in Microsoft Azure HDInsight to create predictive analytics and machine learning solutions. Find out how to cleanse and transform data, build machine learning models, and create real-time machine learning solutions using Python, Scala, and R with Apache Spark
  - Apache Hadoop is an open-source software framework used for distributed storage and processing of big data sets using the MapReduce programming model
- Apache Spark. It was developed in response to limitations in the MapReduce cluster computing paradigm
  - offers over 80 high-level operators that make it easy to build parallel apps and you can use it *interactively* from the Scala, Python and R shells



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

http://scikit-learn.org/stable/

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

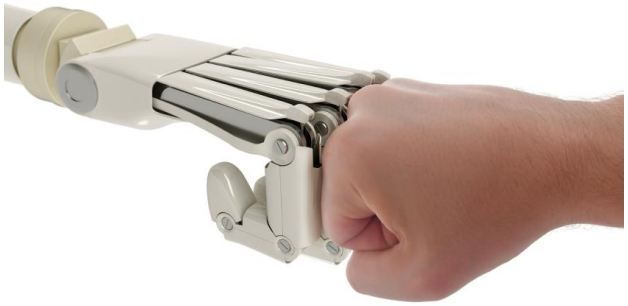
Preprocessing

Feature extraction and normalization.

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

Conclusion

- The solution is to combine machine-learning algorithms with data collected by human analysts
- People can still play a role validating the results



Questions?