CS 644: Introduction to Big Data

Daqing Yun
New Jersey Institute of Technology

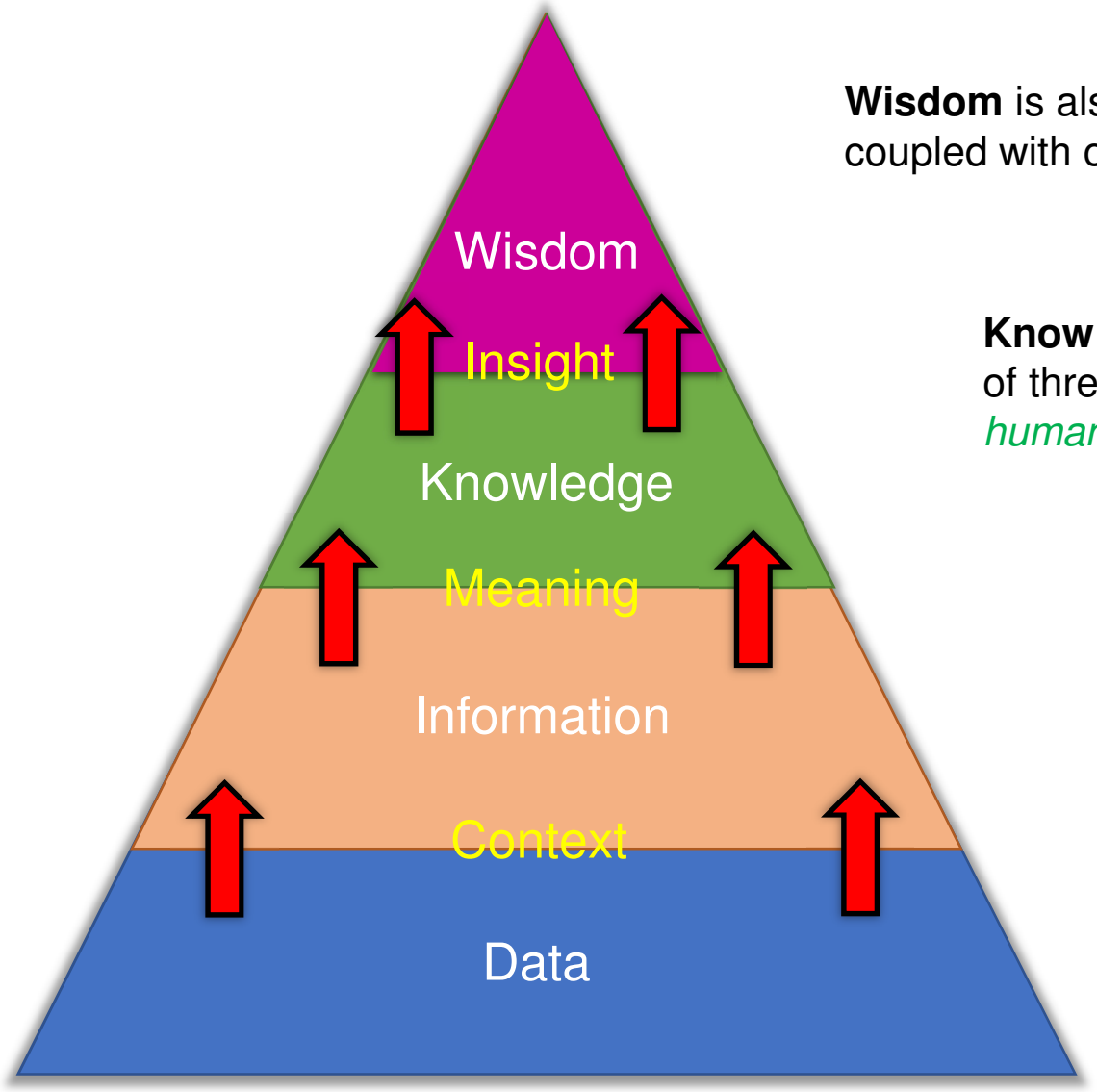
Summit SC@ORNL

https://phys.org/news/2018-06-summit-supercomputer.html

Outline

- The Big Picture
  - The Process of A Typical Big Data Analytics Project
- Data Mining and Machine Learning Fundamentals
  - What is Data Mining?
  - Data Mining Tasks
  - Motivating Challenges
  - What is Machine Learning?
  - When to use Machine Learning?
  - Process of Machine Learning
  - How to create Machine Learning models?
  - How to deploy Machine Learning solutions?
- How to Start?

Data, Information, Knowledge, and Wisdom



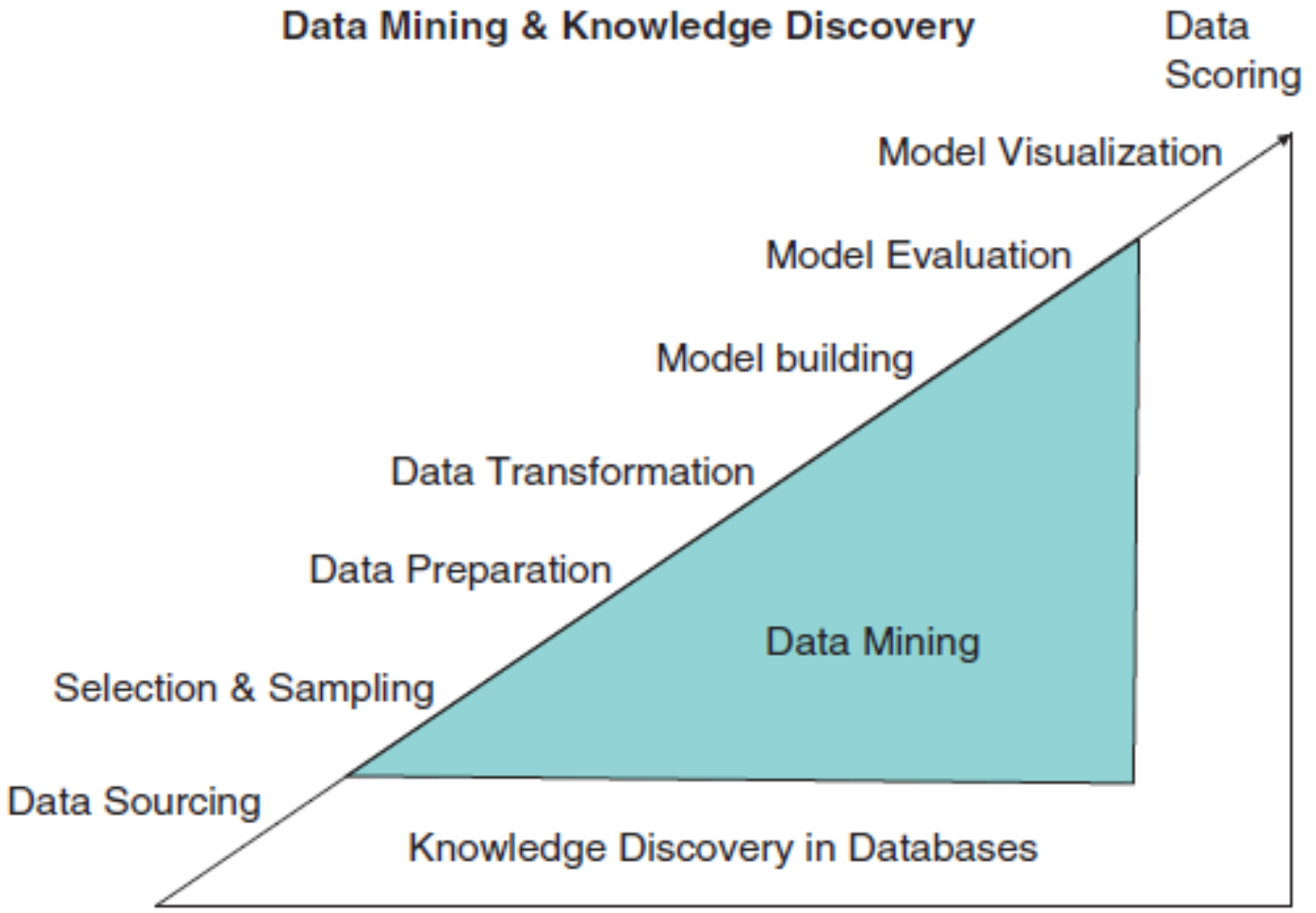
Wisdom is also the comprehension of what is true coupled with optimum *judgment* as to *action*

Knowledge is the result of the qualitative fusion of three elements; *information*, *experience*, and *human wisdom*

Information is the result of data analyzing for the purpose of finding *relations*, *indicators*, *correlations* upon which a decision can be made

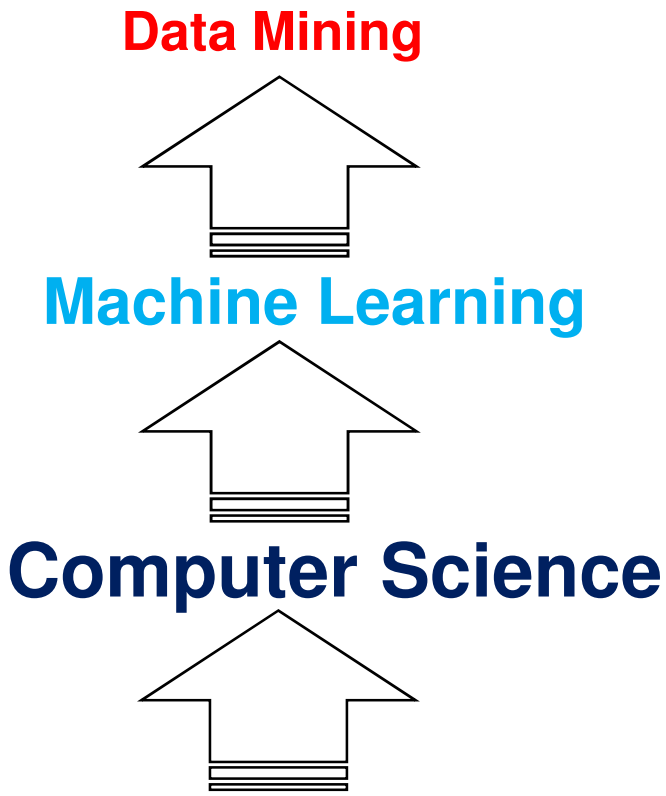
Data are typically the results of measurements and can be visualized using *graphs* or *images*

KDD vs. Data Mining



Machine Learning vs. Data Mining

- There is not a common agreement
- Machine learning focuses on designing algorithms that can *learn from historical data* and make predictions
- Data mining is a cross-disciplinary field that aims on *discovering properties* (useful information) of data sets
- Machine learning can be used for data mining



The Process

Data Sourcing/Representation
Where are the dataset from

Data Cleaning
To remove noise and inconsistent data

Data Integration
Multiple data sources may be combined

Data Selection/Reduction
Data relevant to the analysis task are retrieved from the database

Data Transformation/Consolidation/Reduction
Data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations

Data Preprocessing

Data Mining
The essential process where intelligent methods are applied to extract data patterns

Pattern Evaluation
To identify the truly interesting patterns representing knowledge

Knowledge Presentation
Visualization and knowledge representation techniques are used to present mined knowledge to users

Data Mining Fundamentals

What is Data Mining?

- Data mining is the *process* of (automatically) discovering *useful* information in (large) data repositories

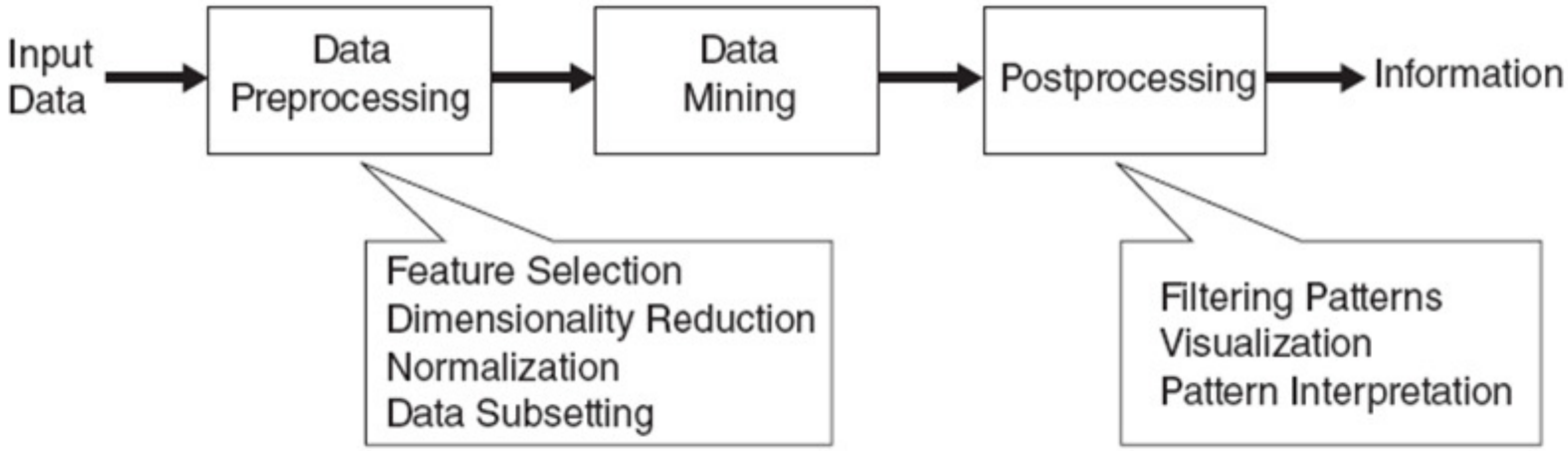
- Data mining is the discovery of *models* for data. However, a model can be one of several things, e.g., statistical models

What is Data Mining?

- Data mining techniques are deployed to handle large datasets in order to find *novel and useful patterns* that might *otherwise remain unknown*
- They also provide capabilities to predict the outcome of a future observation
  - e.g., predicting whether a newly arrived customer will spend more than $100 at a department store
- Not all information discovery tasks are considered to be data mining – *Information Retrieval*
  - e.g., looking up individual records using DBMS
  - e.g., finding particular Web pages via a query to Google

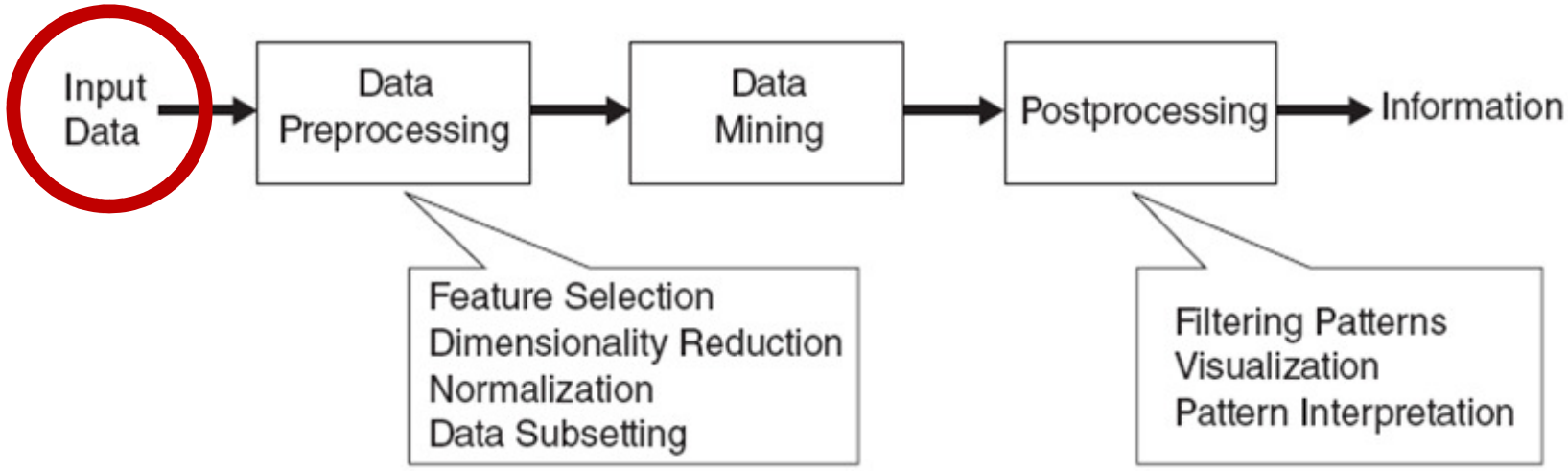
Data Mining vs. Knowledge Discovery

- Knowledge Discovery in Databases (KDD), the overall process of converting raw data into useful information, which consists of a series of transformation steps, from *preprocessing* to *postprocessing* of data mining results
- Data mining is an integral part of KDD



Data Mining -- Input Data

- Might be stored in a variety of formats
  - Flat files, spreadsheets, or relational tables, etc.
- Might reside in a centralized repository
- Or be distributed across multiple sites → data integration



Data Mining -- Preprocessing

- To transform the raw input data into an *appropriate* format for subsequent analysis
  - Fusing data from multiple sources, cleaning data to remove noise and duplicate observations, selecting records and features that are relevant to the data mining task at hand, etc.
- Could be laborious and time-consuming

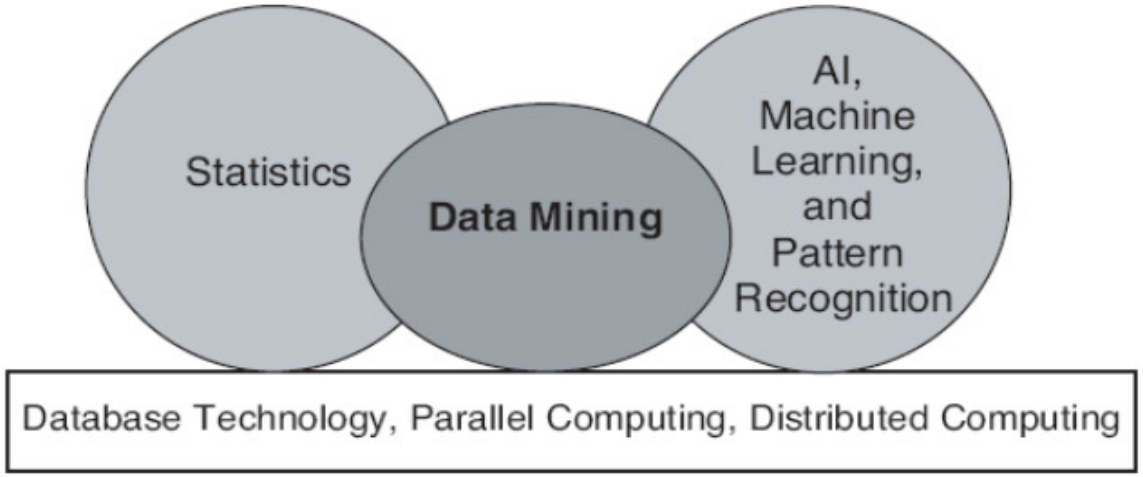
Data Mining -- Postprocessing

- Use data mining results in decision support systems
- Ensure that only valid and useful results are incorporated
  - e.g., in business applications, insights offered by data mining can be used for effective marketing promotions
  - e.g., visualization

The Origins of Data Mining

- Draws ideas from ML/AI, pattern recognition, statistics, and database systems

- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous and distributed nature

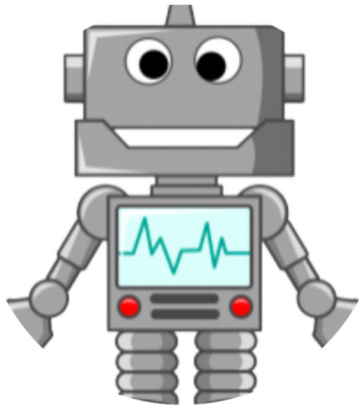


Systems

- **Conventional systems**
  - Classical (non-adaptive)
  - Needs a human programmer to redesign system's functions to accommodate changes in the environment
  - No learning capability ;-(

Systems

- **Nonconventional systems**
  - Adaptive, can detect changes in the surrounding, modify its behavior, and adjust its results accordingly, without human interference
  - Reason, learn, and connect perceptions to actions
  - Learning capability ;-)



Data Mining Systems

- Nonconventional systems
- Analyze data, extract patterns, and make predictions

Data Mining Tasks

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

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996
Tan, Steinbach, and Kumar. Introduction to Data Mining

Data Mining Tasks

- Classification – *predictive*
- Clustering – *descriptive*
- Association rule discovery – *descriptive*
- Sequential pattern discovery – *descriptive*
- Deviation/Anomaly detection – *predictive*
- Regression – *predictive*

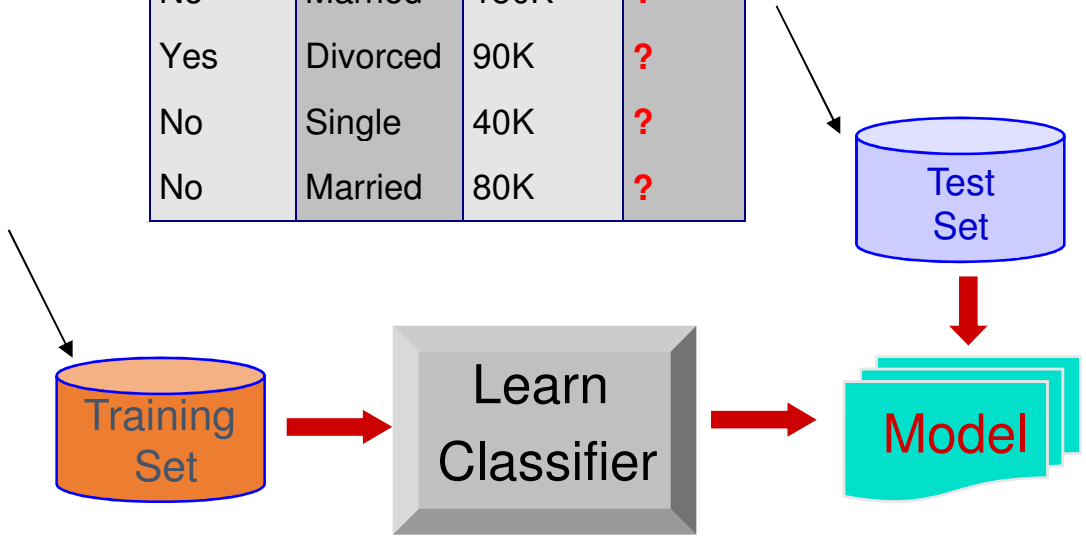
Classification -- Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*
  - One of the attributes is the (label) *class*
- Find a *model* for class attribute as a function of the values of other attributes
- Goal: *previously unseen* records should be assigned a class as accurately as possible
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

Classification -- Example

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification -- Application example

Direct Marketing

- Goal: reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product
- Approach:
  - Use the data for a similar product introduced before
  - Which customers decided to buy and which decided otherwise
    - This ***{buy, don't buy}*** decisions forms the *class attributes*
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model

From [Berry & Linoff] Data Mining Techniques, 1997

Clustering -- Definition

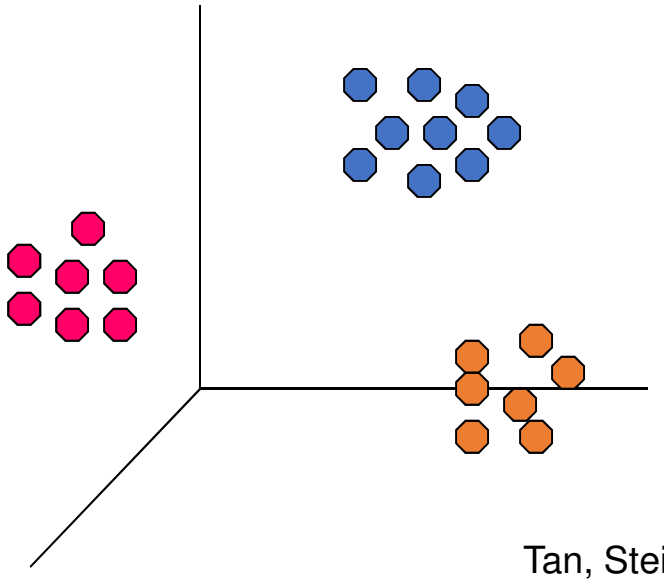
- Given a set of data points, each having a set of attributes, and a *similarity* measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another
- Similarity measures:
  - *Euclidean distance* if attributes are continuous
  - Other problem-specific measures

Clustering -- Example

- Euclidean distance based clustering in 3D space

Intra-cluster distances are minimized

Inter-cluster distances are maximized



Clustering -- Application example

Market Segmentation

- Goal: *subdivide* a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information
  - Find clusters of similar customers
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters

Association Rule Discovery -- Definition

- Given a set of records each of which contain some number of items from a given collection
- Produce dependency rules which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules discovered:

{Milk} → {Coke}

{Diaper, Milk} → {Beer}

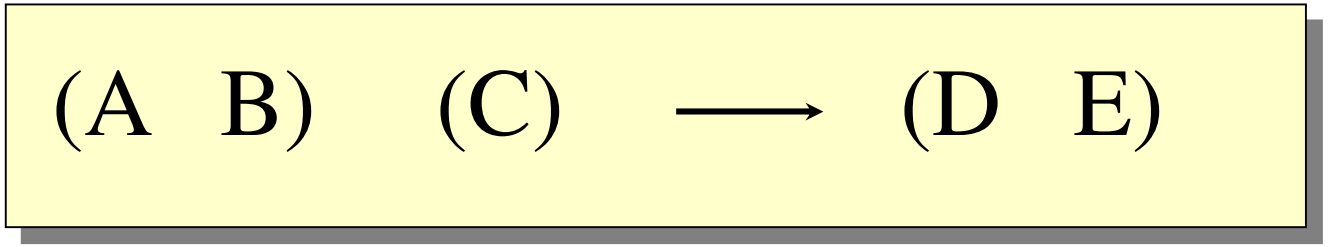
Association Rule Discovery -- Application example

Supermarket Shelf Management

- Goal: to identify items that are bought together by sufficiently many customers
- Approach: process the point-of-sale data collected with barcode scanners to find dependencies among items
- A classic rule:
  - If a customer buys diaper and milk, then he is very likely to buy beer
  - So, don’t be surprised if you find six-packs stacked next to diapers!

Sequential Pattern Discovery -- Definition

- Given a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events



Sequential Pattern Discovery - Application examples

- Customer shopping sequences
  - First buy a computer, then a CD-ROM, and then a digital camera, within 3 months
  - Introduction to Visual C, C++ Primer, Perl for Dummies
  - Shoes, Racket and Racketball, Sports Jackets
  - …
- Telephone calling patterns
- Weblog click streams
- DNA sequences and gene structures
- …

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Example applications:
  - Credit card fraud detection
  - Network intrusion detection

Typical network traffic at University level may reach over 100 million connections per day

Tan, Steinbach, and Kumar. Introduction to Data Mining

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

- Greatly studied in statistics, neural network fields

- Examples:
  - Predicting sales amounts of new product based on advertising expenditure
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices

Regression -- example

- Non Exact Math → Approximation → Uncertainty Modeling

- Given data set, input/output pairs with no function, i.e., no model, then how to find the *hypothesis*! (model)?

- Why do we need to do that?

- Example: suppose X is size of the house in Acres, and Y is the price of houses already sold in hundred thousands, and you have a house that has 2.5 Acres size, and you want to know how much you can get for it if you decide to sell it

y = f(x) ?

x	y
1	6
2	5
3	7
4	10

The task:

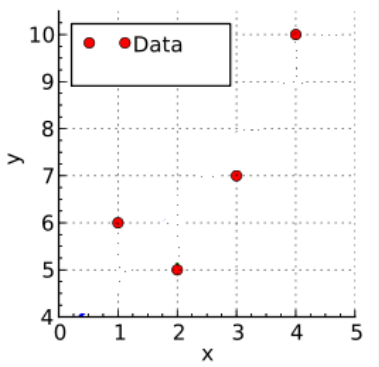
- Plot the data
- Use your eyes to suggest a *good* hypothesis
- Find the mathematical representation for the hypothesis
- Use it to predict the price for your house

Regression -- example

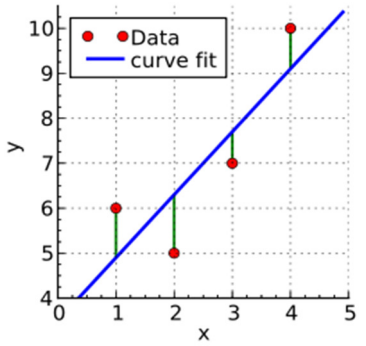
f ?

x	y
1	6
2	5
3	7
4	10

Plot



Approximate with a model



Model is a Line?

$$y = \beta_1 + \beta_2 x$$

$$\beta_1 + 1\beta_2 = 6$$
$$\beta_1 + 2\beta_2 = 5$$
$$\beta_1 + 3\beta_2 = 7$$
$$\beta_1 + 4\beta_2 = 10$$

$$\min S(\beta_1, \beta_2) = [6-(\beta_1+1\beta_2)]^2 + [5-(\beta_1+2\beta_2)]^2 + [7-(\beta_1+3\beta_2)]^2 + [10-(\beta_1+4\beta_2)]^2$$
$$= 4\beta_1^2 + 30\beta_2^2 + 20\beta_1\beta_2 - 56\beta_1 - 154\beta_2 + 210$$

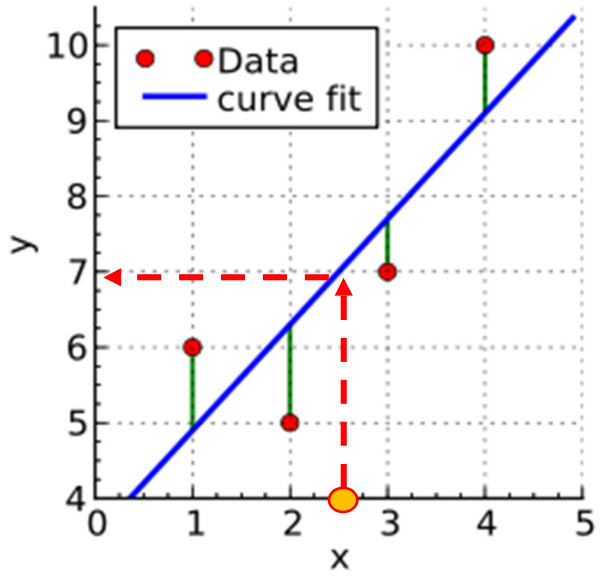
$$\begin{cases} \frac{\partial S}{\partial \beta_1} = 0 = 8\beta_1 + 20\beta_2 - 56 \\ \frac{\partial S}{\partial \beta_2} = 0 = 20\beta_1 + 60\beta_2 - 154 \end{cases}$$

$$\begin{cases} \beta_1 = 3.5 \\ \beta_2 = 1.4 \end{cases} \Rightarrow y = 3.5 + 1.4x \Rightarrow y = 3.5 + 1.4 \cdot (2.5) = 7$$

You will get 7 hundred thousand dollars if you decide to sell it now

Regression - Example

Can't you get more?



This is an example of using Polynomial Linear Regression for prediction, could this work with big data?

Regression -- example

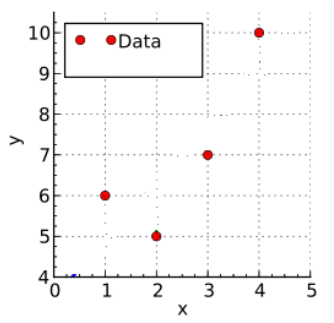
Modeling 2nd Degree Polynomial Regression

- Let's try quadratic model

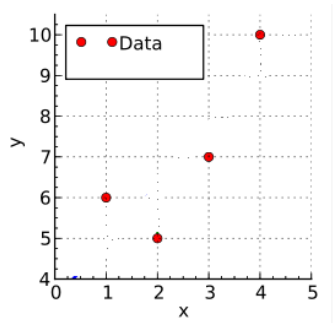
f ?

x	y
1	6
2	5
3	7
4	10

Plot



with a model



Quad. Model

$h = \beta_1 x^2$

$$6 = \beta_1(1)^2$$
$$5 = \beta_1(2)^2$$
$$7 = \beta_1(3)^2$$
$$10 = \beta_1(4)^2$$

$$\frac{\partial S}{\partial \beta_1} = 0 = 708\beta_1 - 498$$

$$\beta_1 = .703$$

$h = .703\ (2.5)^2 = 4.39$

You will get 4.39 hundred thousand dollars

Regression -- example

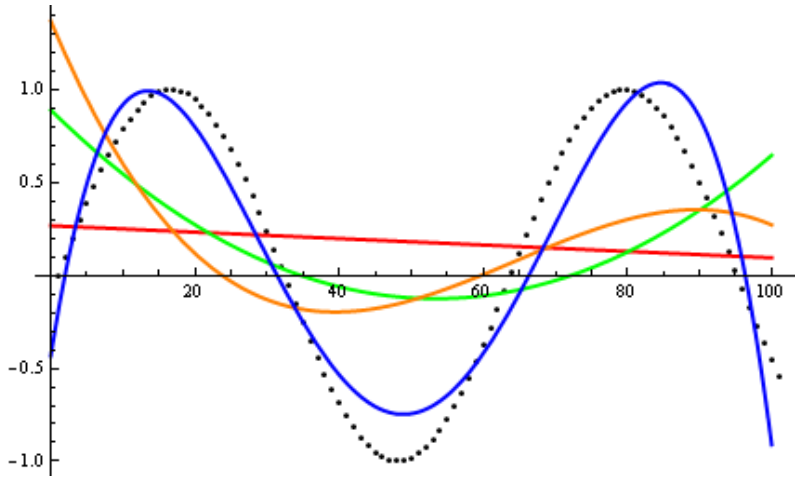
- Let's try higher degree polynomial models

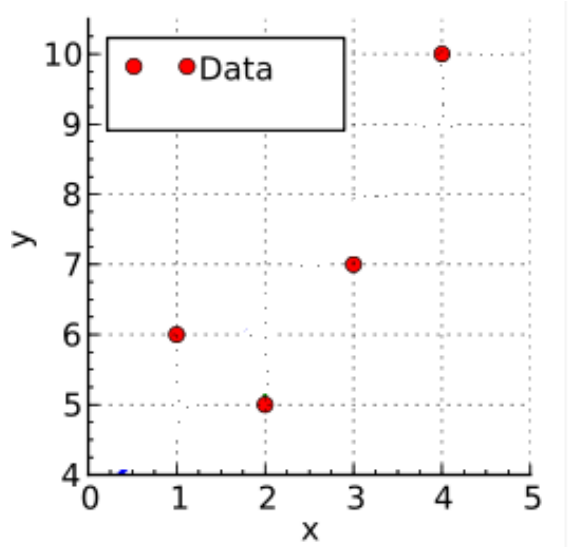
$$h = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$h = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

$$\vdots$$

$$h = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots\ldots\ldots + \beta_n x^n$$





The problem has become:
Finding the best solution (fit) and optimize it as much as you can

Motivating Challenges

- Scalability
- High dimensionality
- Heterogeneous and complex data
- Data ownership and distribution
- Non-traditional analysis

Scalability

- Advances of data generation and collection
- Datasets with size of gigabytes, terabytes, or even petabytes are becoming common
- Data mining algorithms must be scalable to handle exponential growth
- Require implementation of novel data structures to access individual records efficiently
- “Out-of-Core” algorithms may be necessary when datasets cannot fit into main memory

High Dimensionality

- Datasets with hundreds or thousands of attributes instead of the handful common a few years ago
  - In Bioinformatics, progress in microarray technology has produced gene expression data with thousands of features
  - Temporal or spatial datasets also tend to have high dimensionality
- Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data
- Computational complexity rapidly increases

Heterogeneous and Complex Data

- Traditional data analysis methods deal with datasets containing *attributes* of the *same type*
- Emergence of more complex (non-traditional) data recently
  - Web pages containing semi-structured text and hyperlinks
  - DNA data with sequential and three-dimensional structure
  - Climate data consists of time-series measurements (temperature, pressure, etc.) at various locations
- Mining such complex objects should take into consideration relationships in the data
  - Temporal and spatial autocorrelation
  - Graph connectivity
  - Semi-structured text, and XML documents

Data Ownership and Distribution

- Data might be geographically distributed among resource belonging to multiple entities
- Additional challenges
  - How to reduce amount of communication needed to perform the distributed computation?
  - How to effectively consolidate the data mining results obtained from multiple sources?
  - How to address data security issues?
  - …

Non-Traditional Analysis

- Traditional statistical approach “hypothesize-and-test”, which is labor-intensive
- Current data analysis tasks often require the generation and evaluation of thousands of hypotheses
- Automate the process of hypothesis generation and evaluation
- Datasets analyzed in data mining are typically not the result of a carefully designed experiments and often represent opportunistic sample of the data, rather than random sample
- Non-traditional types of data and data distributions are often involved

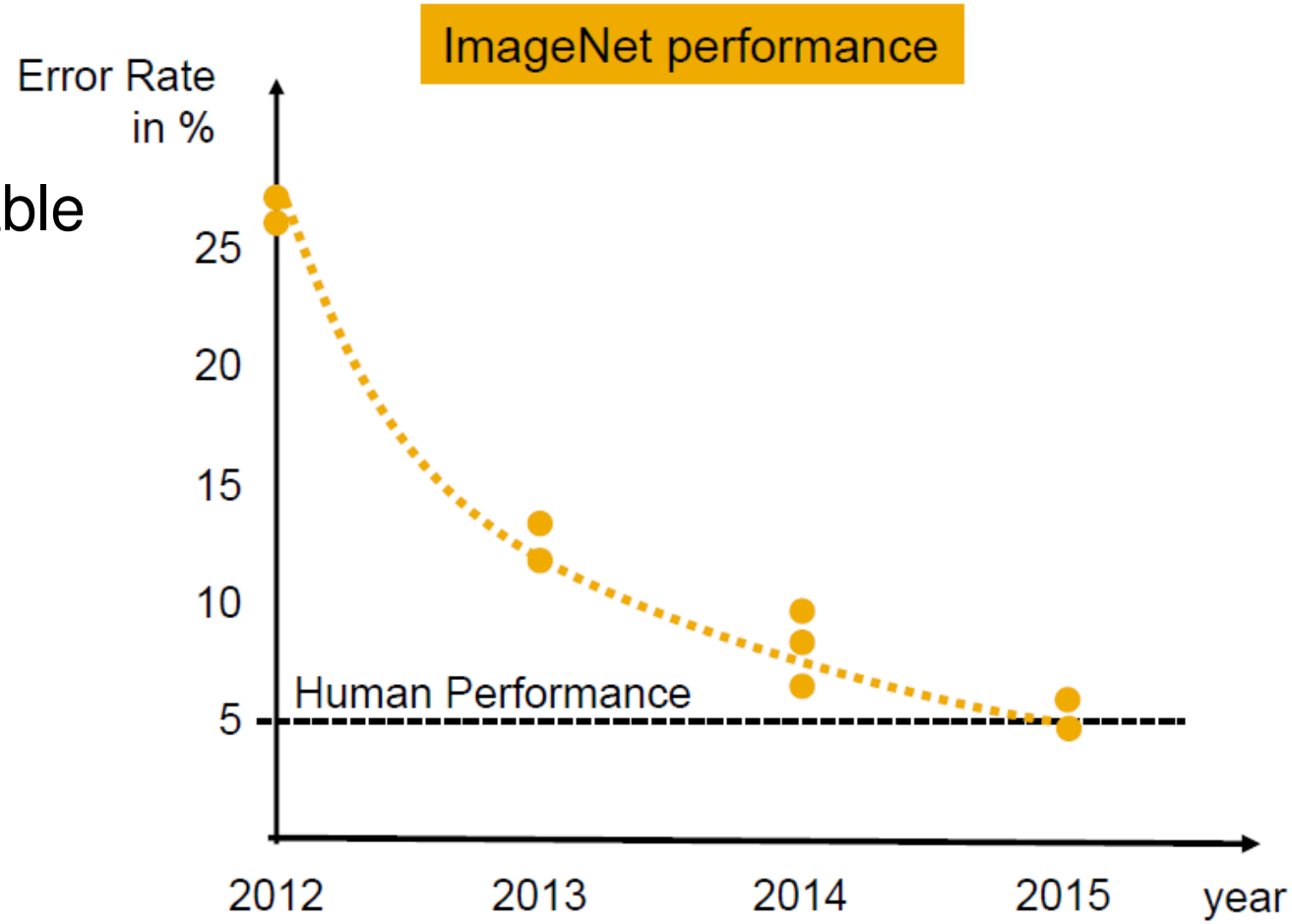
Machine Learning Fundamentals

Software is becoming intelligent

Computer Vision is surpassing human abilities

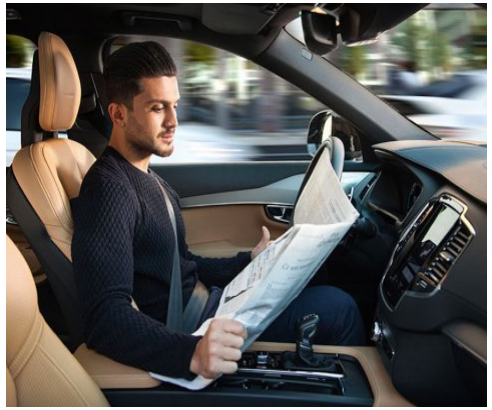
- Chair
- Dining table
- Person

- Dog
- Person
- Leaf



http://www.image-net.org/ https://open.sap.com/

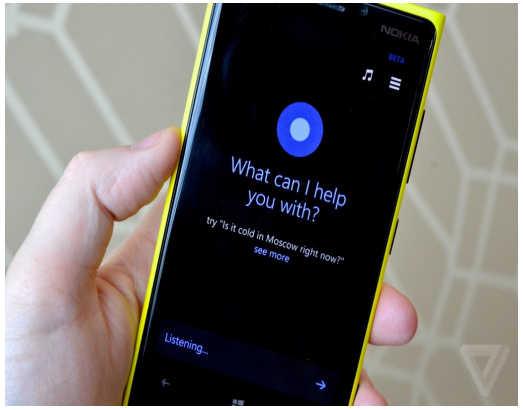
Machine learning example applications



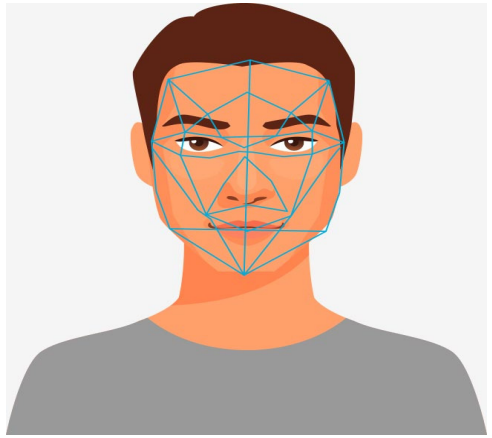
Self-Driving

Spam Detection

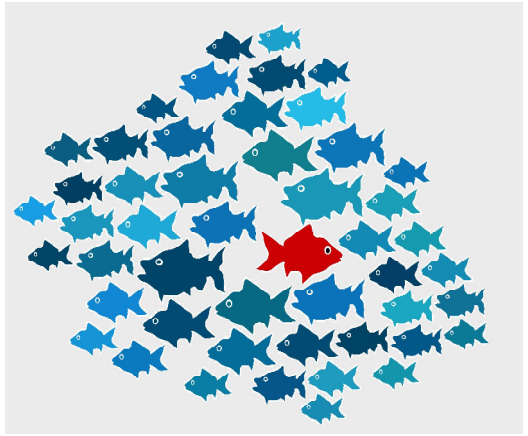
Fraud Detection



Voice Recognition

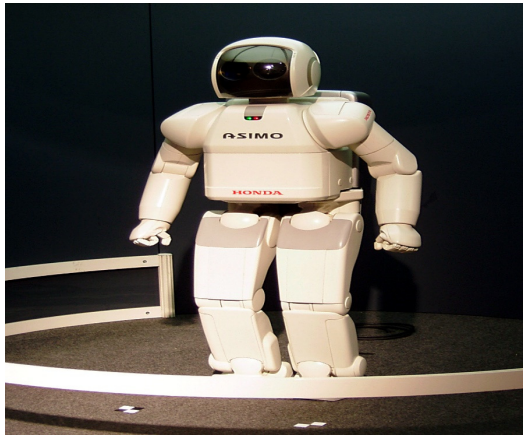


Face Recognition



Anomaly Detection

Sales Forecast



Robotics

Use predictive analytics to solve a variety of business challenges

- Churn reduction
- Customer acquisition
- Lead scoring
- Product recommendation
- Campaign optimization
- Customer segmentation
- Next best offer/action

- Predictive maintenance
- Load forecasting
- Inventory/Demand optimization
- Price optimization
- Manufacturing process optimization
- Quality management
- Yield management

- Fraud and abuse detection
- Collection and delinquency
- Credit scoring
- Operation risk modeling
- Crime threat analysis
- Revenue and loss analysis

- Cash flow and forecasting
- Budget simulation
- Profitability and margin analysis
- Financial risk modeling
- Employee retention modeling
- Succession planning

- Life sciences
- Healthcare
- Media
- Higher education
- Public section
- Social sciences
- Construction and mining
- Travel and hospitality
- Big data and IoT

Big pool of machine learning use cases

What is Machine Learning?

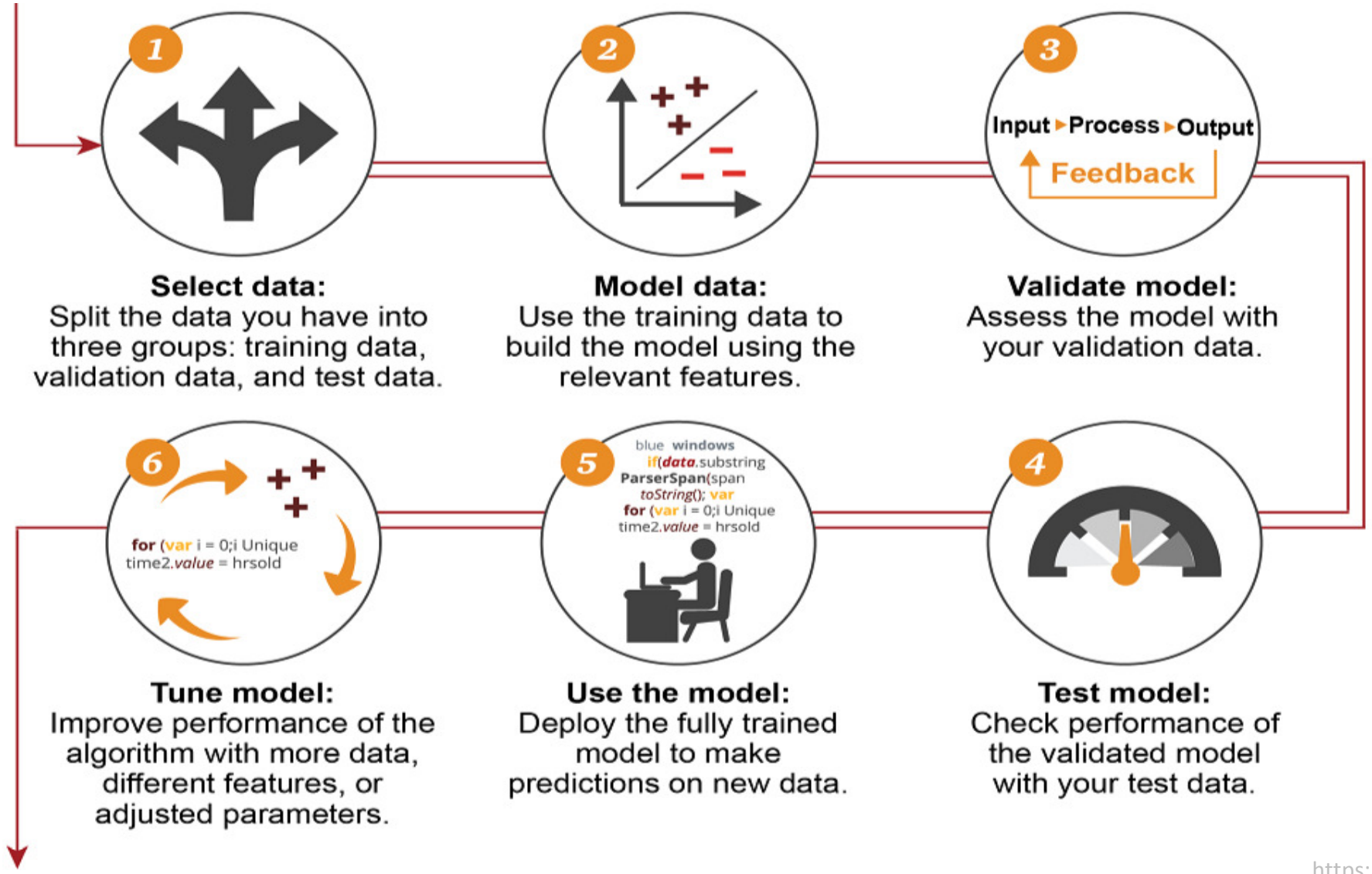
A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** if its performance at tasks in **T**, as measured by **P**, improves with **experience E**.

----Mitchell, T. (1997). Machine Learning, McGraw Hill

What is Machine Learning?

- Machine learning is the science of getting computers to act without being *explicitly programmed*
- Machine learning is a technique of data science that helps computers *learn from existing data* in order to *forecast future behaviors, outcomes, and trends*

How machine learning works



https://www.digitalpulse.pwc.com.au

An example of machine learning task

Let's say you rent cars.

How can you accurately predict demand for your product?

https://docs.microsoft.com/en-us/azure/machine-learning/studio/basics-infographic-with-algorithm-examples

Steps to build a machine learning solution

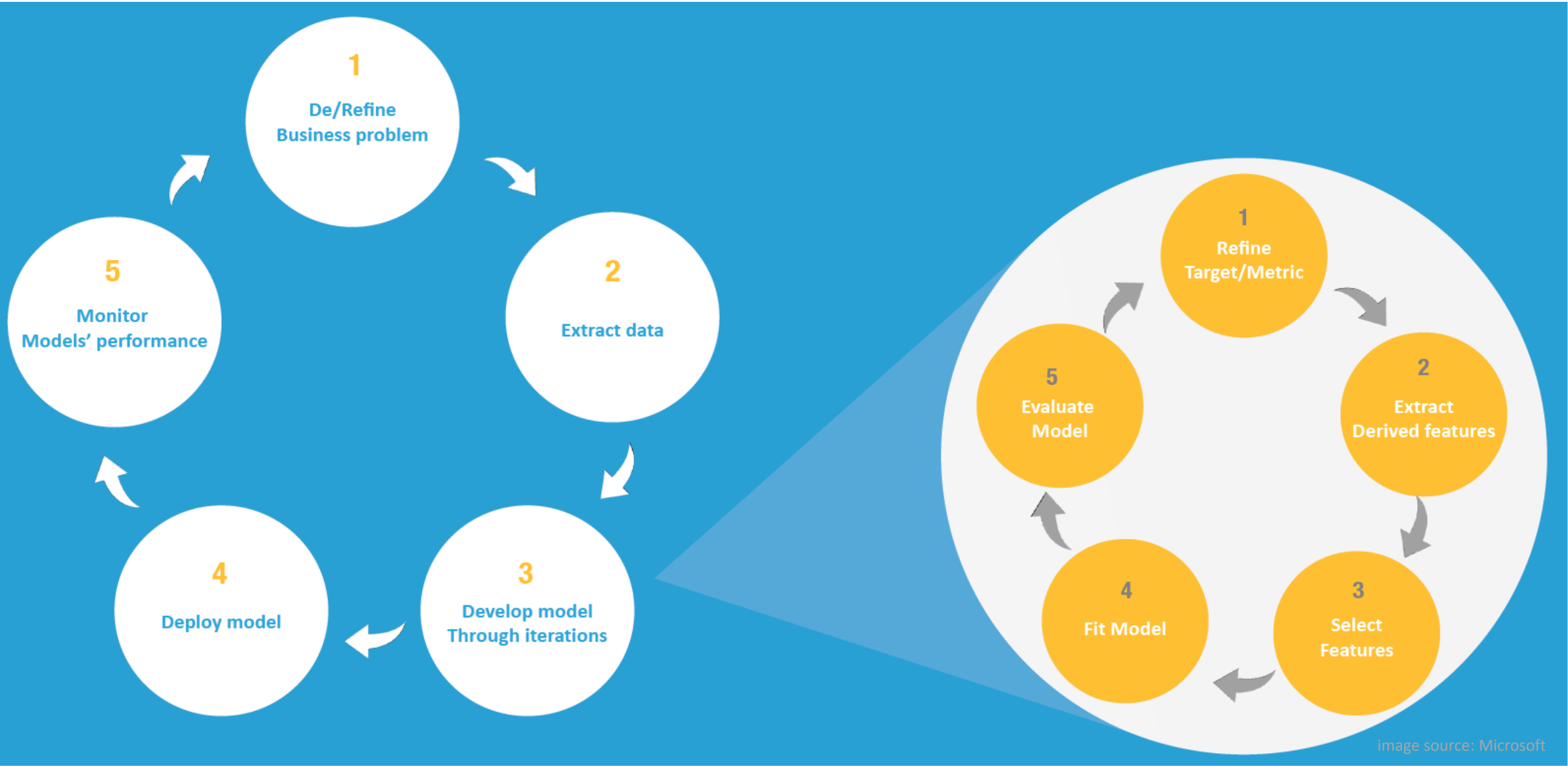


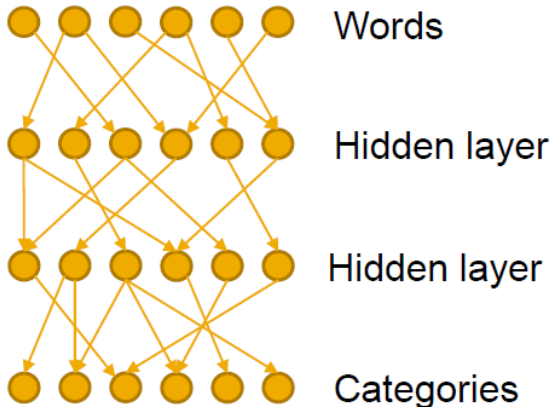
image source: Microsoft

Trained, not programmed

- **Traditional computer program**
  - Explicitly programmed to solve problems
  - Decision rules are clearly defined

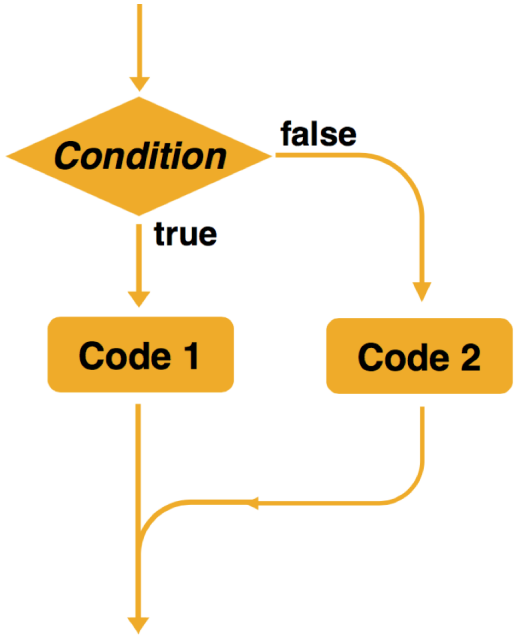
```
if (..) then {
   ..
} else {
   ..
}
```

- **Machine Learning**
  - Trained (i.e., learned) from examples
  - Decision rules are complex or fuzzy
  - Rules are not defined by human but learned by machine from data



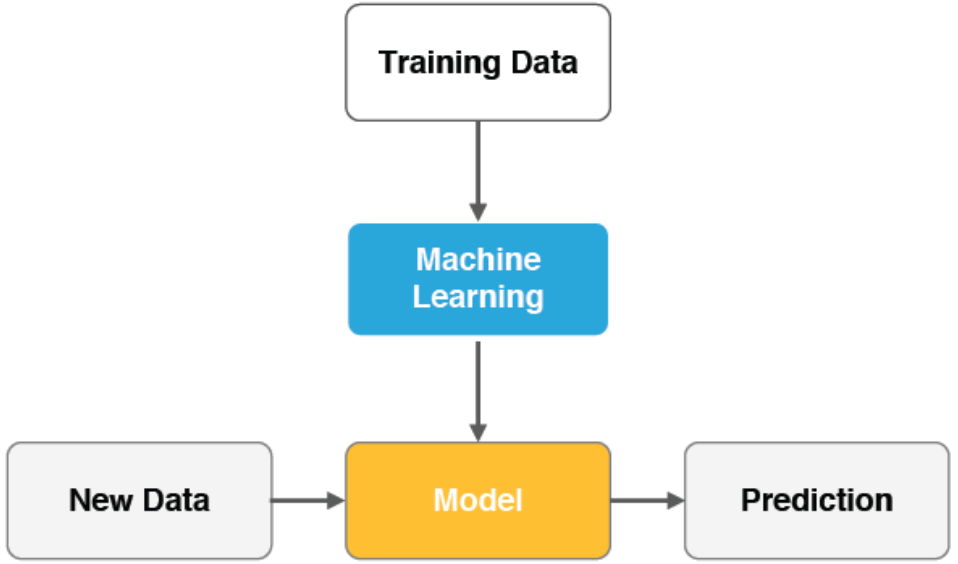
Difference between traditional approach and machine learning approach

Rule-based approach



- Explicitly programmed to solve problems
- Decision rules are clearly defined by humans

Machine learning approach



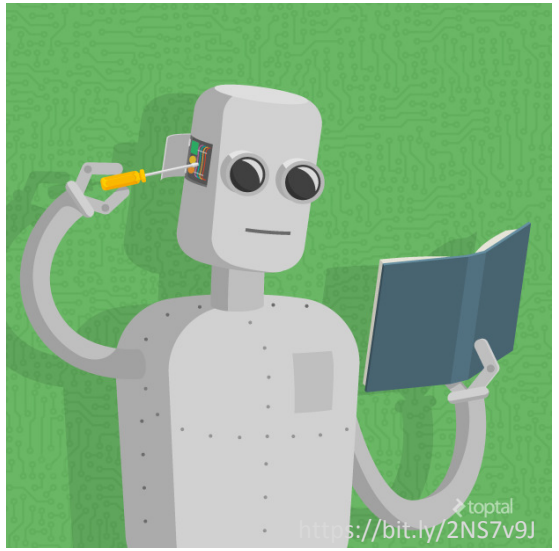
- Trained (i.e., learned) from examples
- Decision rules are complex and fuzzy
- Rules are not defined by humans but learned by machines from data

Summary

Machine learning uses historical data to make predictions

It is similar to data mining, but whereas data mining is the science of discovering unknown patterns and relationships in data; machine learning applies previously inferred knowledge to new data to make decisions in real-life applications

- Computers approximate complex functions from historical data
- Rules are not explicitly programmed but learned from data



When to use machine learning?

From business problem to machine learning problem: a recipe

1. Do you need machine learning?
2. Can you formulate your problem clearly?
3. Do you have sufficient examples?
4. Does your problem have a regular pattern?
5. Can you find meaningful representations of your data?
6. How do you define success?

When to use machine learning?

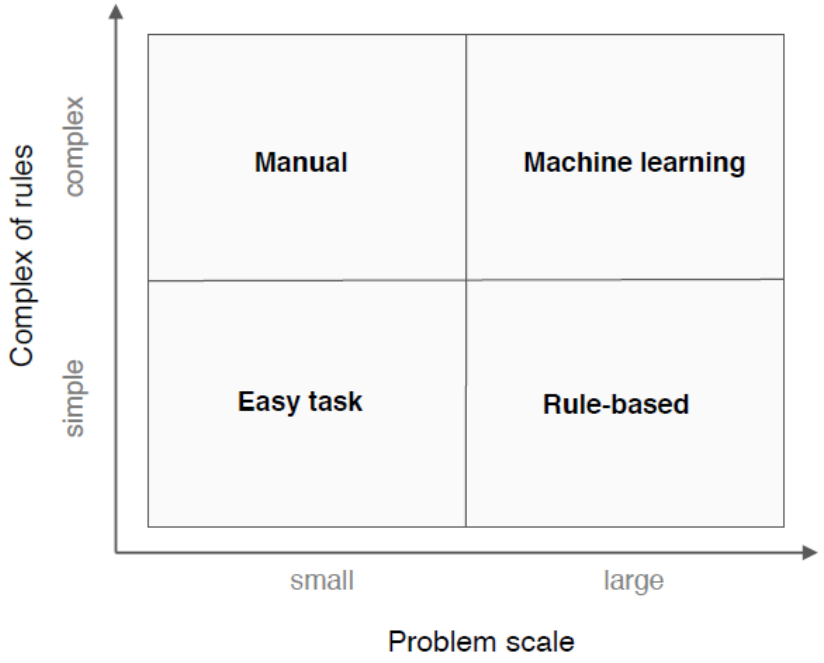
From business problem to machine learning problem: a recipe

1 Do you need machine learning?

- Do you need to automate the task?
- High volume tasks with complex rules and unstructured data are good candidates

Example: sentiment analysis

- High volume of reviews on the Web
- Unstructured text
- Human language is complex and ambiguous



When to use machine learning?

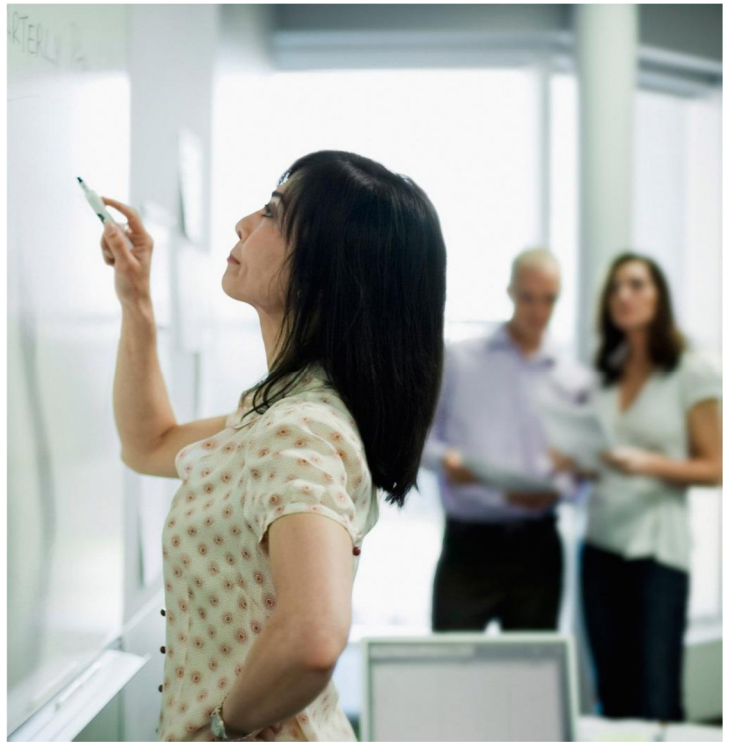
From business problem to machine learning problem: a recipe

2 Can you formulate your problem clearly?

- What do you want to predict given which input?
- Pattern: “given X, predict Y”
  - What is the input?
  - What is the output?

Example: sentiment analysis

- Given a customer review, predict its sentiment
  - Input: customer review text
  - Output: positive, negative, neutral



When to use machine learning?

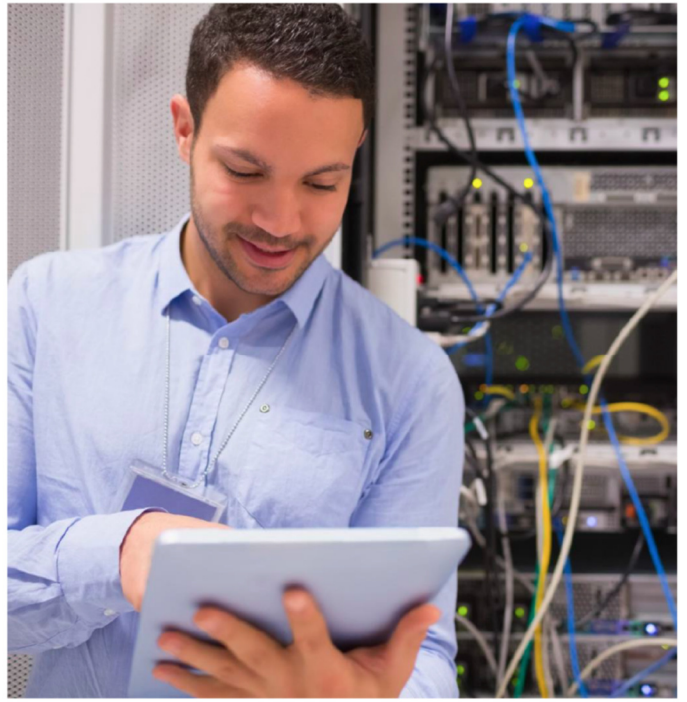
From business problem to machine learning problem: a recipe

3 Do you have sufficient examples?

- Machine learning always requires data
- Generally, the more data, the better
- Each example must contain two parts (supervised learning)
  - Features: attributes of the example
  - Label: the answer you want to predict

Example: sentiment analysis

- Thousands of customer reviews and rating from the Web



When to use machine learning?

From business problem to machine learning problem: a recipe

4 Does you problem have a regular pattern?

- Machine learning learns regularities and patterns
- Hard to learn patterns that are rare or irregular

Example: sentiment analysis

- Positive words like *good*, *awesome*, or *love it* appear more often in highly-rated reviews
- Negative words like *bad*, *lousy*, or *disappointed* appear more often in poorly-rated reviews

When to use machine learning?

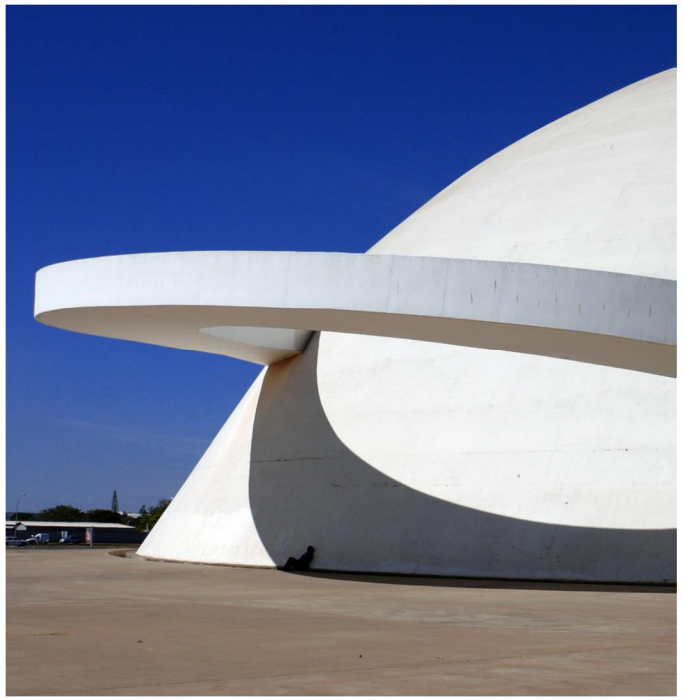
From business problem to machine learning problem: a recipe

5 Can you find meaningful representations of your data?

- Machine learning algorithms ultimately operate on numbers
- Generally, examples are represented as feature vectors
- Good features often determine the success of machine learning

Example: sentiment analysis

- Represent customer review as vector of word frequencies
- Label is positive (4-5 stars), negative (1-2 stars), neutral (3 stars)



When to use machine learning?

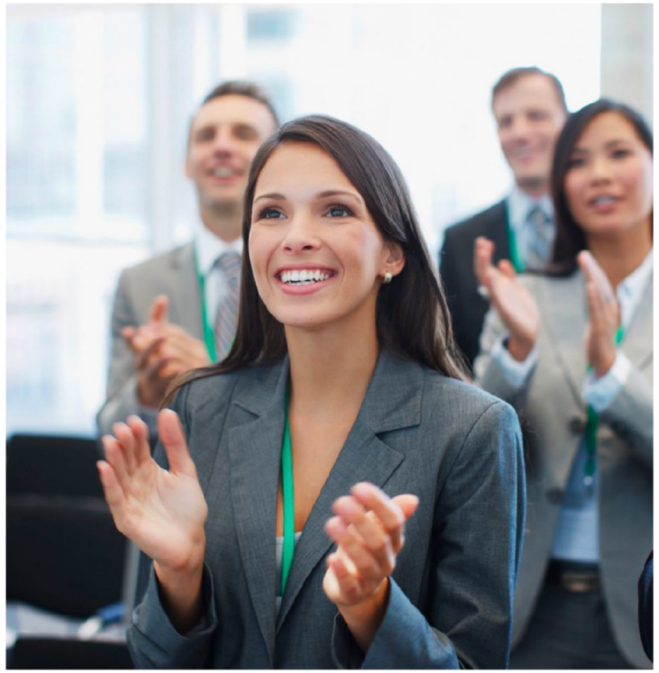
From business problem to machine learning problem: a recipe

6 How do you define success?

- Machine learning optimizes a training criteria
- The evaluation function has to support the business goals

Example: sentiment analysis

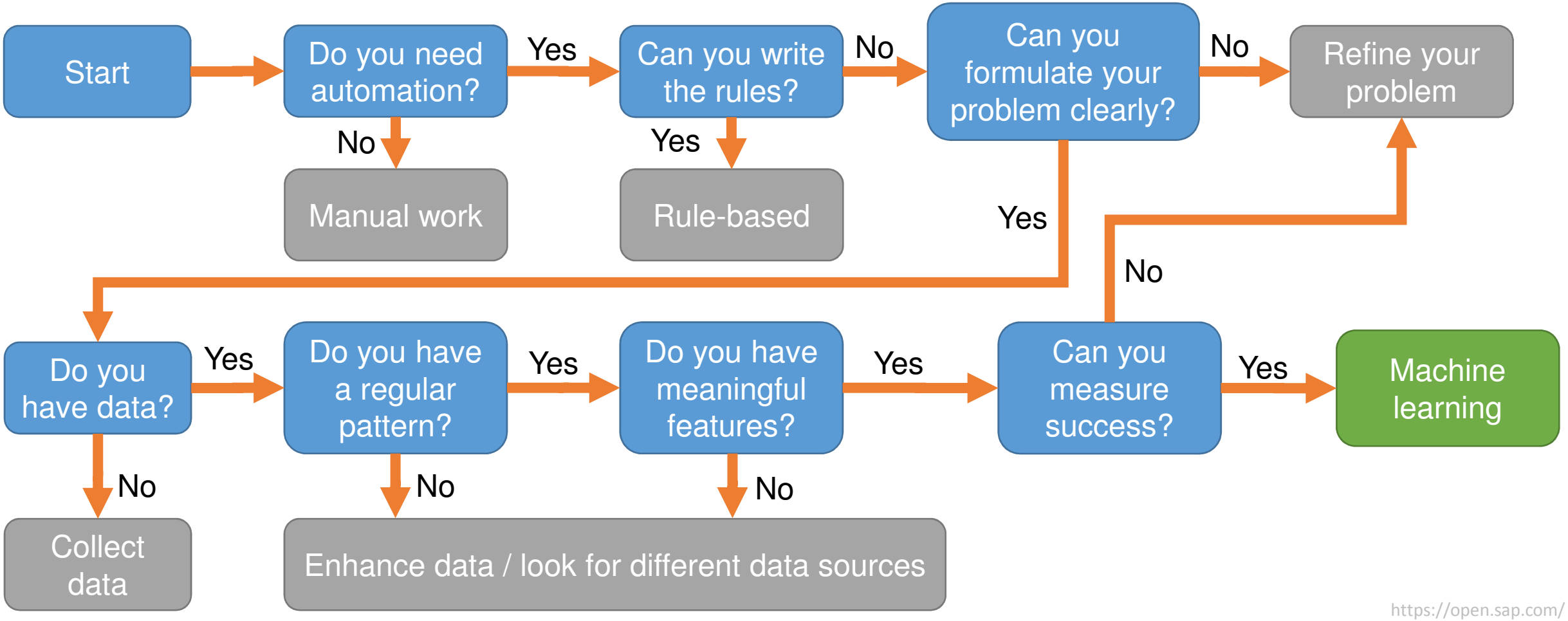
- Accuracy: percentage of correctly predicted labels



When to use machine learning?

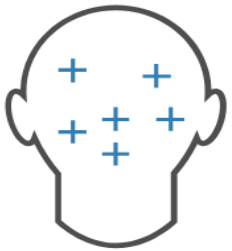
From business problem to machine learning problem: a recipe

The “cheat sheet”



Summary

- Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of variables, but *no existing formula or equation*. For example, machine learning is a good option if you need to handle situations like these:



Hand-written rules and equations are too complex—as in face recognition and speech recognition.

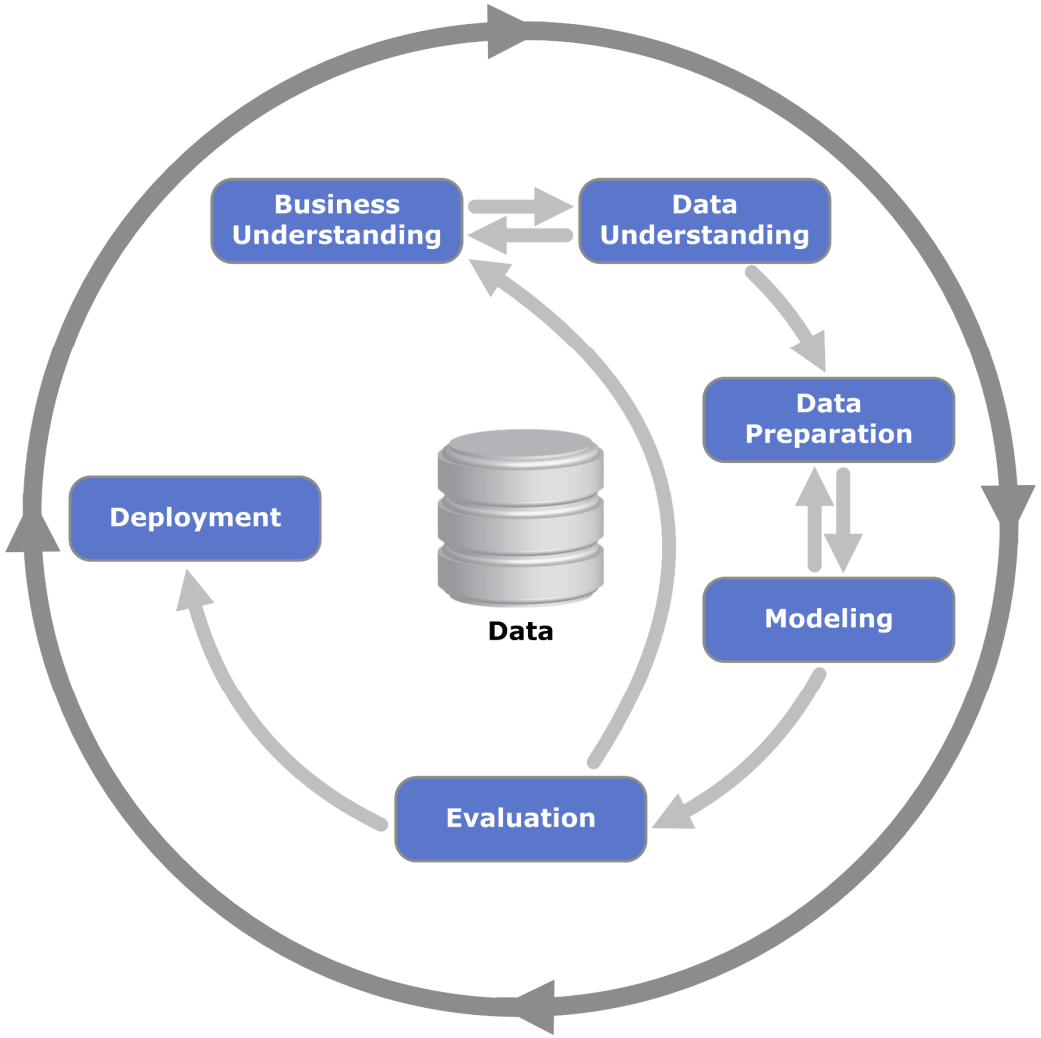
The rules of a task are constantly changing—as in fraud detection from transaction records.

The nature of the data keeps changing, and the program needs to adapt—as in automated trading, energy demand forecasting, and predicting shopping trends.

https://www.mathworks.com/discovery/machine-learning.html

The Process

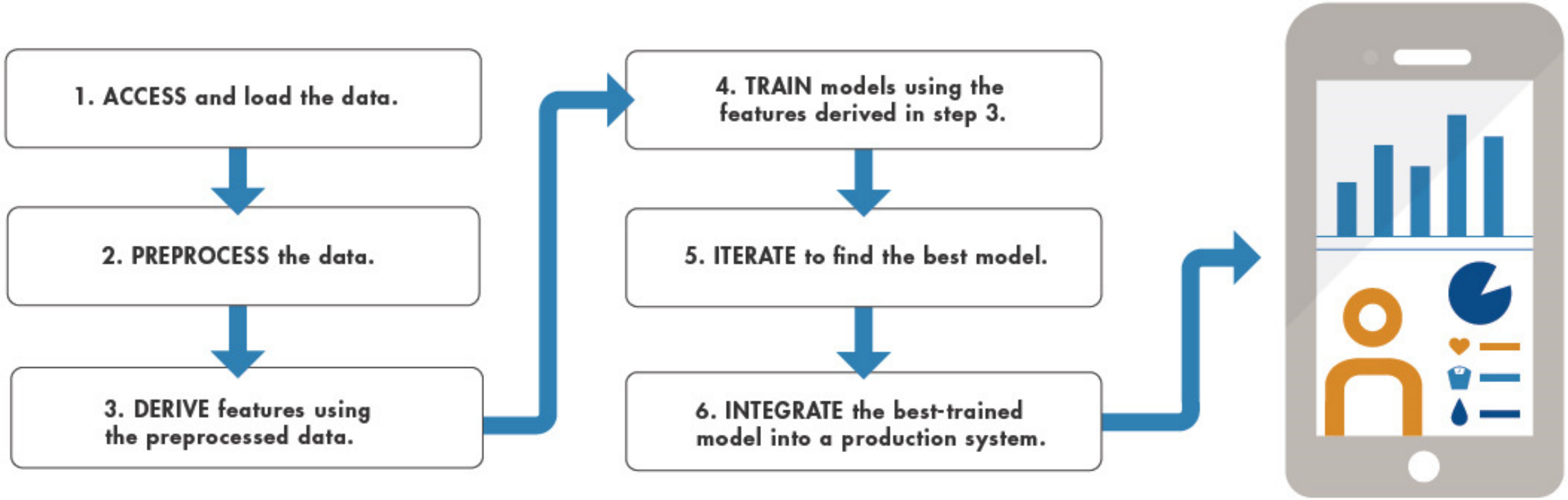
- Cross-industry standard process for data mining (CRISP-DM)



https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

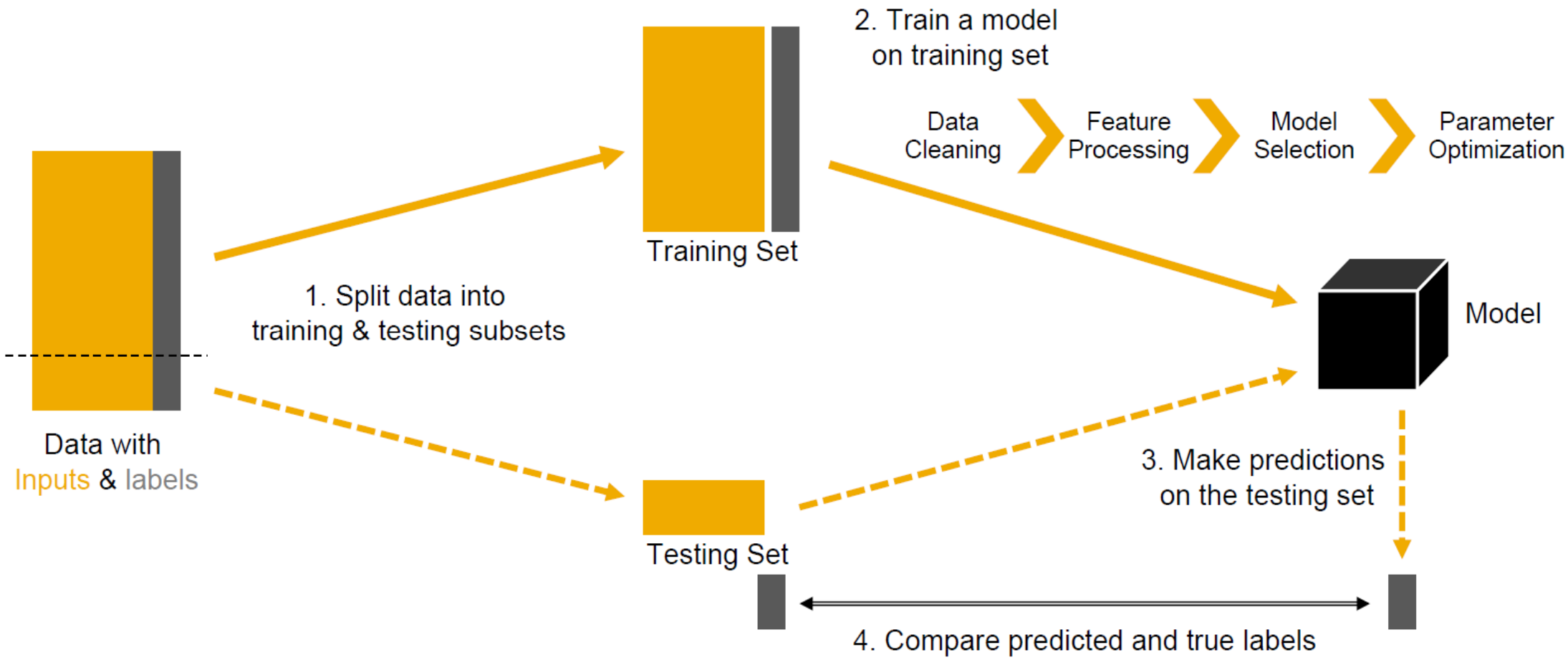
The Process

- A MathWorks perspective



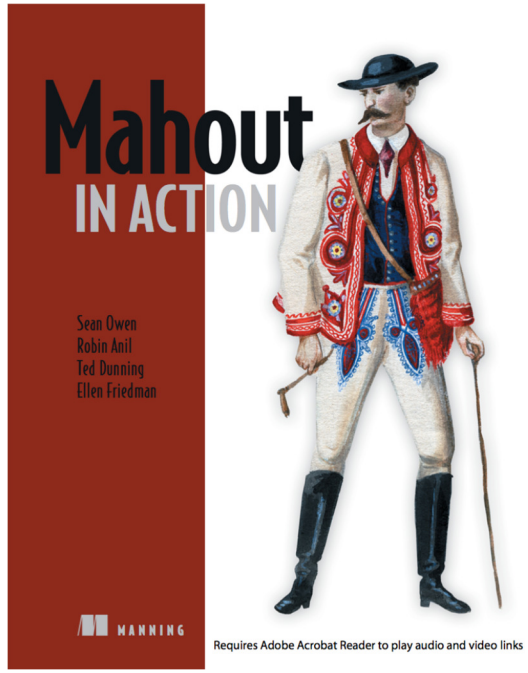
https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

How to create machine learning models?



How to deploy machine learning solutions?

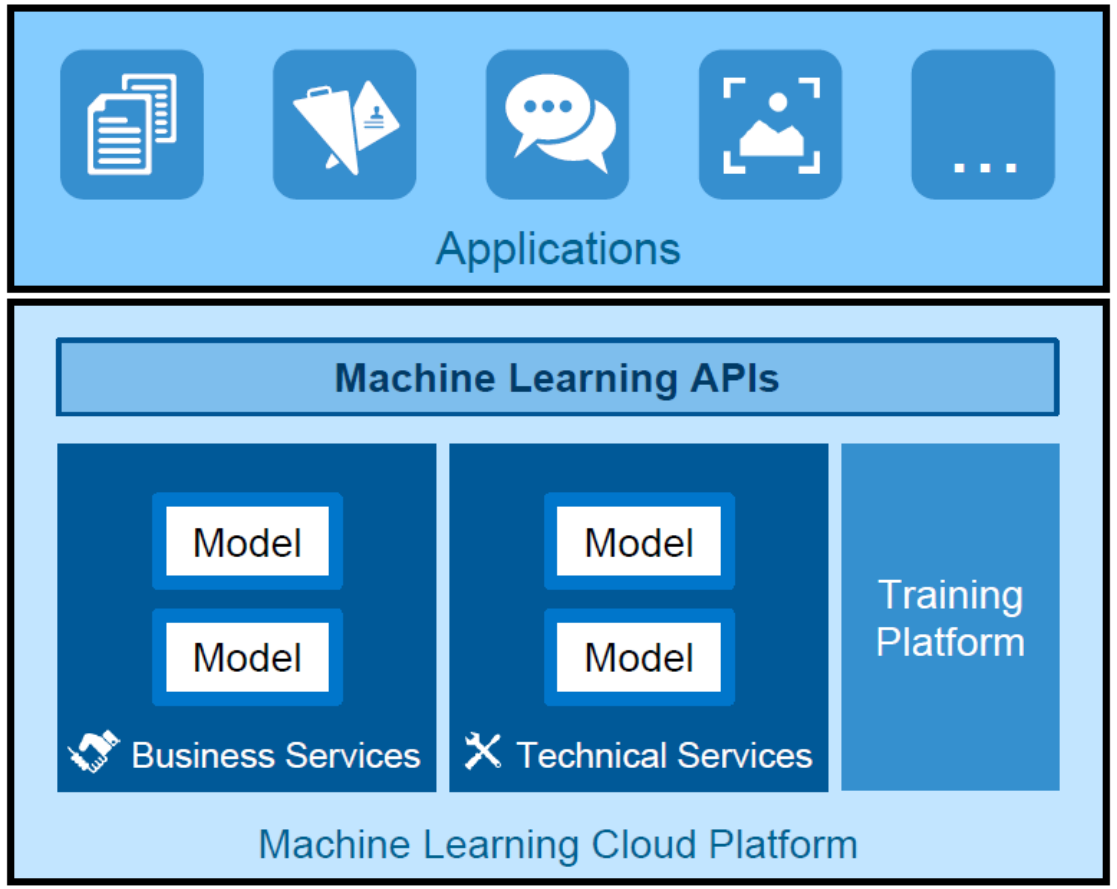
- Mahout https://mahout.apache.org/
- A scalable machine learning and data mining library
- Key components
  - Collaborative filtering, e.g., Weighted Matrix Factorization, SVD++, Parallel SGD
  - Classification, e.g., Logistic Regression, Naïve Bayes, Random Forest, Hidden Markov Models
  - Clustering, e.g., Canopy Clustering, K-means Clustering, Fuzzy K-means, Spectral Clustering



How to deploy machine learning solutions?

Machine learning architecture

- Cloud platforms for handling large data volume and integrating data sources
- Machine learning models deployed as micro-services
- Enterprise applications consume service through APIs
- Business services directly support enterprise applications
- Technical services are building blocks for new applications



Application Example: Natural Language Processing

Support Ticket Classification

Classify support tickets into categories so that they can be routed to corresponding agents

1

- Do you need machine learning?
  - High volume of support tickets
  - Human language is complex and ambiguous
- Can you formulate your problem clearly?
  - Given a customer support ticket, predict its service category
  - Input: customer support ticket; output: service category
- Do you have sufficient examples?
  - Large volume of customer support tickets with respective category from ticket support systems
- Does your problem have a regular pattern?
  - Common customer issues will have many tickets
  - Issues will correlate with common keywords, e.g., bill or payment will appear more often in support tickets with category payments
- Can you find meaningful representations of your data?
  - Represent customer support tickets as vector of word frequencies
  - Label is the service category of the customer support ticket
- How do you define success?
  - Measure percentage of correctly predicted service categories

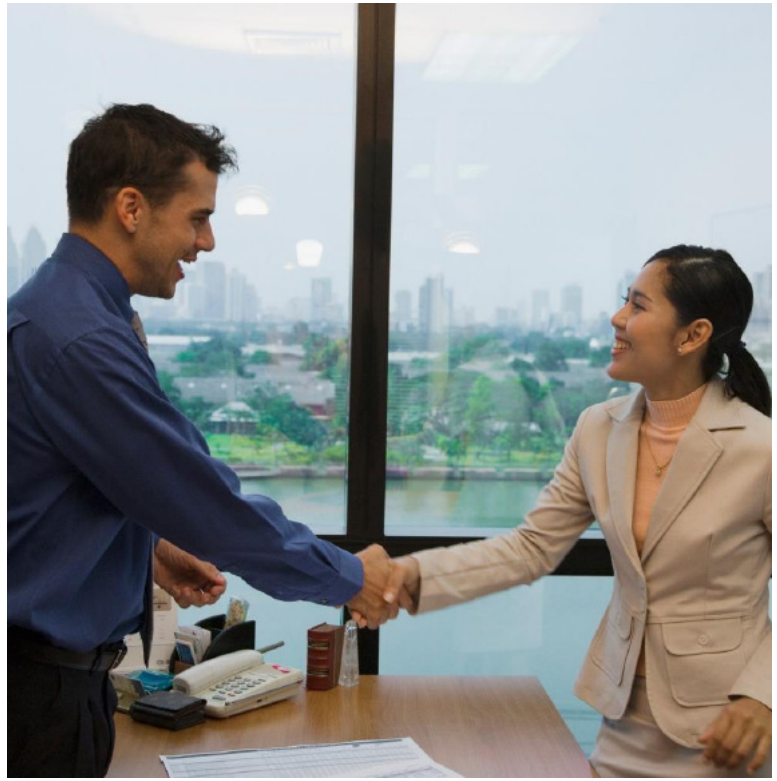
Application Example: Natural Language Processing

Recruiting – CV matching

Shortlist candidates during recruiting

2

- Do you need machine learning?
  - Hundreds of applications per job opening
  - Manual effort to read CVs and screen candidates
- Can you formulate your problem clearly?
  - Given a candidate's CV and a job description, predict suitability
  - Input: CV and job description; output: yes or no
- Do you have sufficient examples?
  - Large volume of previous job applications, job descriptions, and whether candidate was invited for interview
- Does your problem have a regular pattern?
  - Required skills in job description should match experience in CV
  - Good CVs has no typos, are neither too long nor too short, etc.
- Can you find meaningful representations of your data?
  - Represent CVs and job descriptions as vector of features that measure similarity and match
  - Label is whether the candidate was invited for interview
- How do you define success?
  - Measure precision and recall of correct predictions



Application Example: Computer Vision
Retail Shelf Analytics

Given a picture of a retail shelf, detect all products in the picture and compare with planned layout

3

- Do you need machine learning?
  - High manual effort to monitor store shelves every day
  - Detecting products in images not possible with simple rules
- Can you formulate your problem clearly?
  - Given a shelf image, first detect products and then compare their positions with planned shelf layout
  - Input: photo; output: bounding boxes of products
- Do you have sufficient examples?
  - Large volume of collected retail shelf images, manually labelled bounding boxes of products
- Does your problem have a regular pattern?
  - Product packaging has regular shape, colors, and logos
- Can you find meaningful representations of your data?
  - Represent photos as array of pixel values
  - Image patches with products are positive examples; random patches are negative examples
- How do you define success?
  - Measure precision and recall of predicted bounding boxes
  - Similarity of the detected layout to the true layout

Application Example: Computer Vision
Fashion Apparel Color Analysis

Analyze color trends in fashion apparel on social media to detect trends

4

- Do you need machine learning?
  - High volume of photos on the Web and social media
  - Detecting fashion apparel in images not possible with simple rules
- Can you formulate your problem clearly?
  - Given an image, first detect fashion apparel and then compute its color histogram
  - Input: photo; output: histogram of colors
- Do you have sufficient examples?
  - Large volume of social media images including fashion apparel, manually labelled bounding boxes, and color histograms
- Does your problem have a regular pattern?
  - Fashion apparel has regular, distinct shape
- Can you find meaningful representations of your data?
  - Represent photos as array of pixel values
  - Image patches with fashion apparel are positive examples; random patches are negative examples
- How do you define success?
  - Measure precision and recall of predict bounding boxes
  - Similarity of detected color histograms to ground truth

Takeaway question

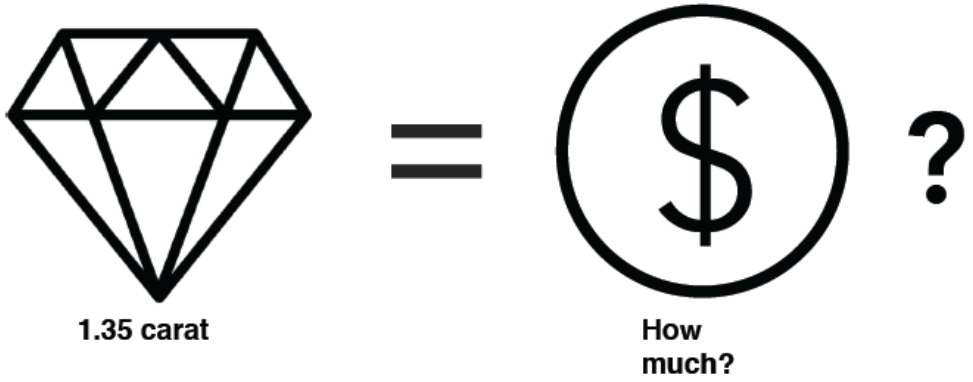
Example: predict the price of a diamond

Suppose I want to shop for a diamond. And you want to get an idea of how much it will cost. I take a notepad and pen into the jewelry store, and I write down the price of all of the diamonds in the case and how much they weigh in carats. Starting with the first diamond – it's 1.01 carats and $7,366. Now I go through and do this for all the other diamonds in the store.

Prices of diamonds in jewelry stores

Carats	Price ($)
1.01	7,366
0.49	985
0.31	544
1.51	9,140
0.37	493
0.73	3,011
1.53	11,413
0.56	1,814
0.41	876
0.74	2,690
0.63	1,190
0.6	4,172
2.06	11,764
1.1	4,682
1.31	6,172

How much will it cost to buy a 1.35 carat diamond?



Hint: recall the "price of house" example we have seen previously

https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-predict-an-answer-with-a-simple-model

Summary

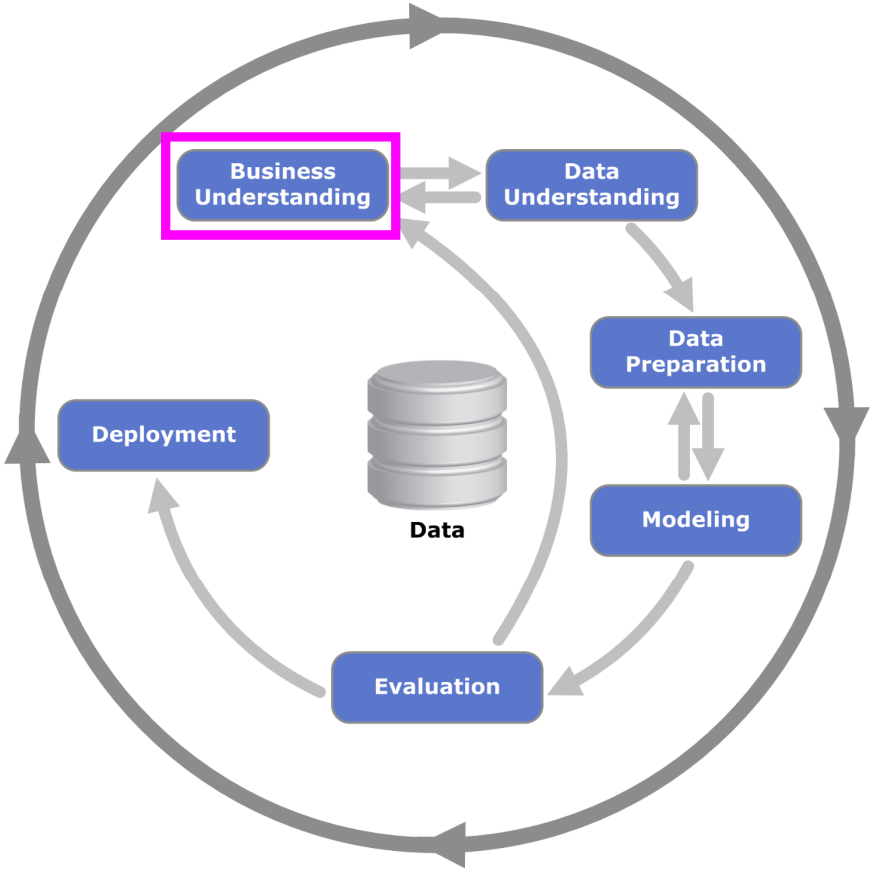
- Machine learning enables computers to learn from data
  - Computers approximate complex functions from historical data
  - Rules are not explicitly programmed but learned from data
- Intelligent applications are expected to have a significant effect on the future of knowledge work and employment
  - Large number of work activities can be automated with today's technology
- Identify if machine learning can solve your business problem
  - High-volume, repetitive tasks on unstructured data are good candidates
  - You cannot explicitly write the rules but you have example
- Machine learning comes with new architecture principles
  - Machine learning requires separate training and prediction step
  - Micro-services and APIs are a common architecture principle for ML services

How to Start?

How to start?
Understand the business

Tasks include

- Identifying your business goals
- Assessing your situation
- Defining your data mining goals
- Producing your project plan



https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

How to start?

Ask a question you can answer with data

- **Sharp questions** can be answered with a name or a number
  - What will my stock's sale price be next week?
  - Which car in my fleet is going to fail first?

- **Vague questions** cannot be answered with a name or a number
  - How can I increase my profits?
  - What can my data tell me about my business?

- How you ask a question is a clue to which algorithm can give you an answer

The 5 questions data science can answer

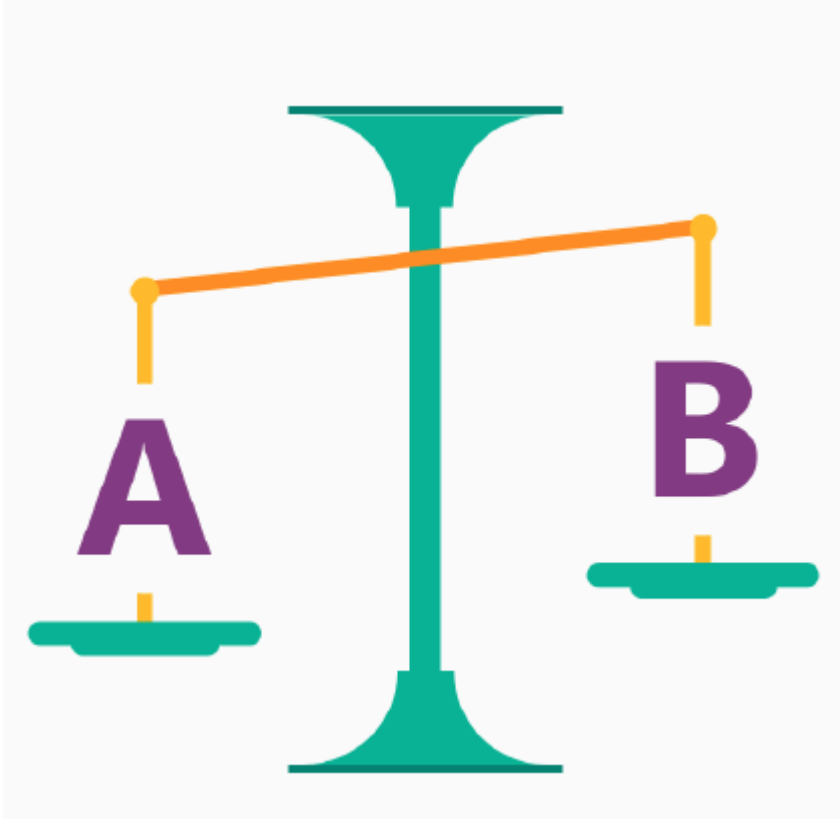
Surprise? but there are only five questions DS can answer

- Is this A or B?
- Is this weird?
- How much or how many?
- How is this organized?
- What should I do now?

https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Q1: Is this A or B?

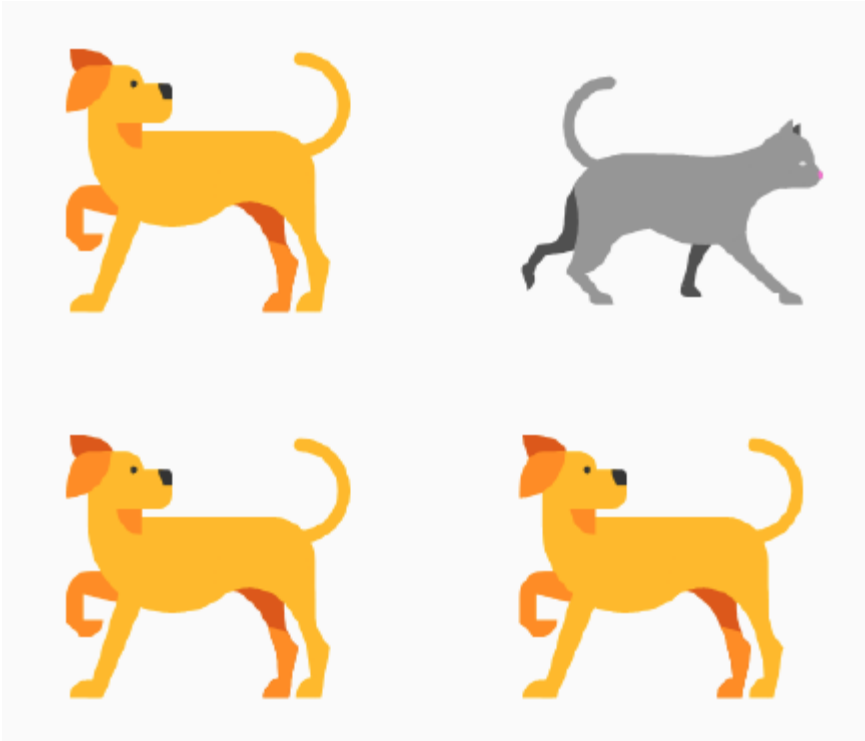
- Use Classification algorithms
- Will this tire fail in next 1000 miles?
  - Yes or No?
- Which brings in more customers?
  - A $5 coupon or a 25% discount?



https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Q2: Is this weird?

- Use Anomaly Detection algorithms
- Your credit company analyzes your purchase pattern, so that they can alert you to possible fraud
- Charges that are “weird” might be a purchase at a store where you do not normally shop or buying an unusually pricey item



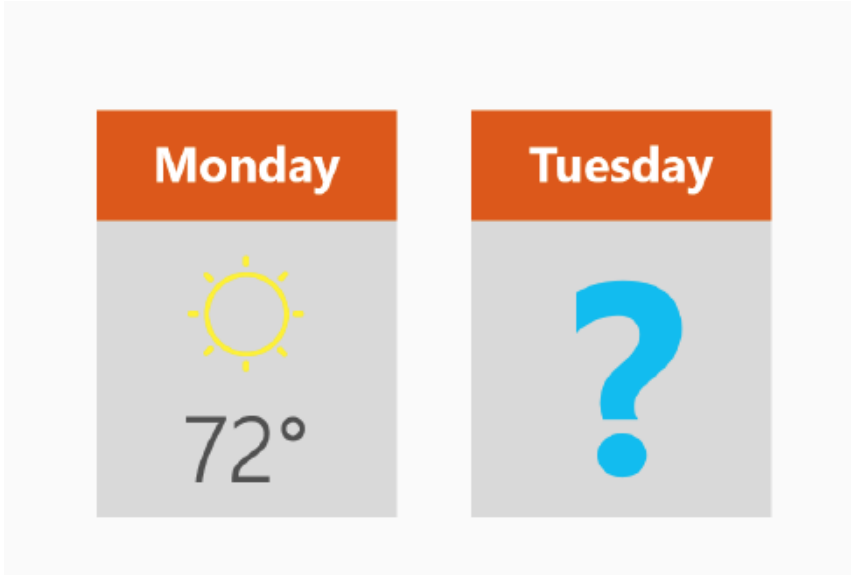
https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Q3: How much? or How many?

- Use Regression algorithms

- Regression algorithms make numerical predictions, such as
  - What will the temperature be next Tuesday?
  - What will my fourth quarter sales be?

- They help answer any question that asks for a number



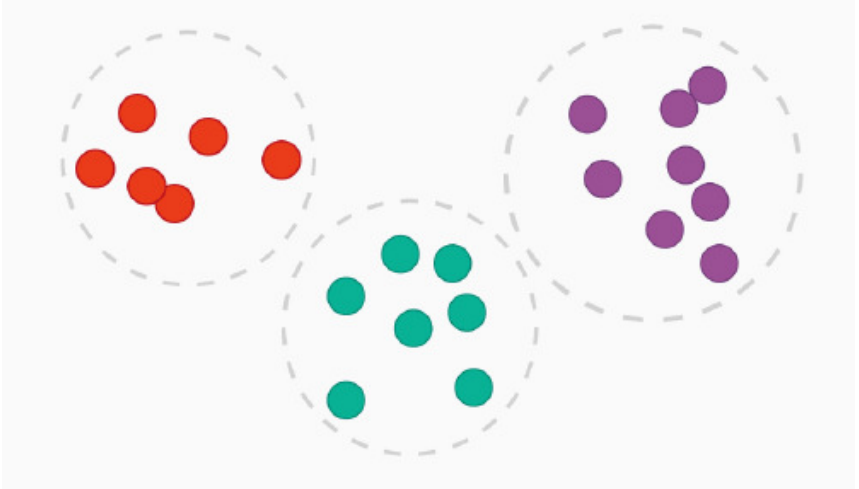
https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Q4: How is this organized?

- Use Clustering algorithms

- Common examples of clustering questions are:
  - Which viewers like the same types of movies?
  - Which printer models fail the same way?

- Sometimes you want to understand the structure of a data set



https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

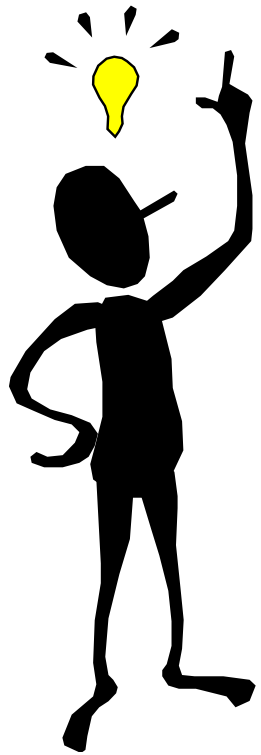
Q5: What should I do now?

- Use Reinforcement Learning algorithms

- Questions it answers are always about what action should be taken – usually by a machine or robot, e.g.,
  - For a self-driving car: at a yellow light, brake or accelerate?
  - For a robot vacuum: keep vacuuming, or go back to the charging station

https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

So, what do you want to find out?



Thanks ! ☺

Questions ?