Introduction to Business Analytics

By:
Dr. Abhishek Verma
IIM Rohtak


DATA
IS THE NEW OIL
Find it . Extract it . Refine it . Distribute it . Monetize it

What is Data Analytics?

The vast reservoir of data is the next big thing

Real impetus is the derived potential insights

DEFINITION:

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information.

INFORMS: Analytics is the scientific process of transforming data into insight for making better decisions.

DATA → INFORMATION → KNOWLEDGE → WISDOM

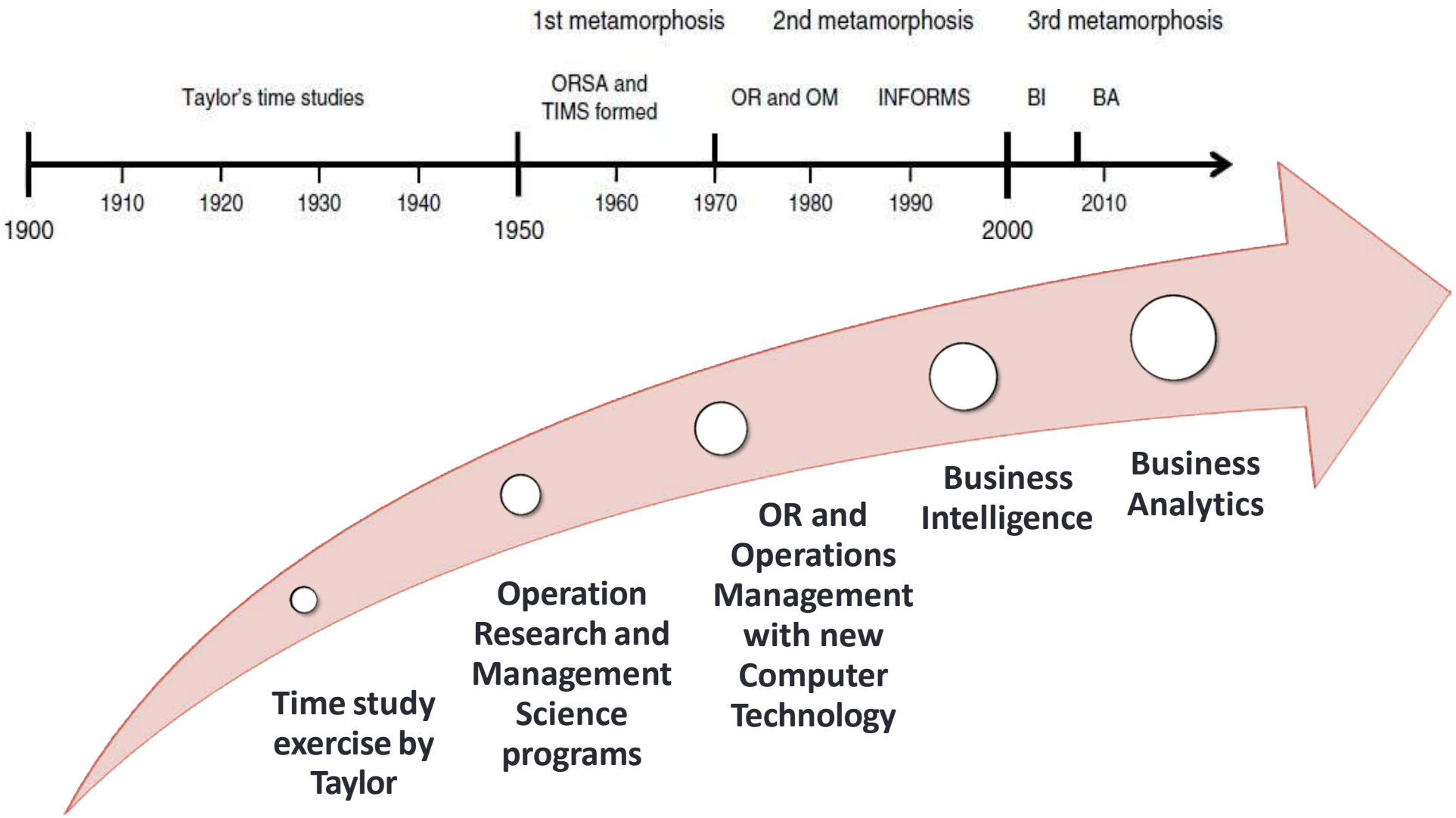
Companies and organization use it to make better business decisions

In sciences it is used to verify/ disprove existing models or theories.

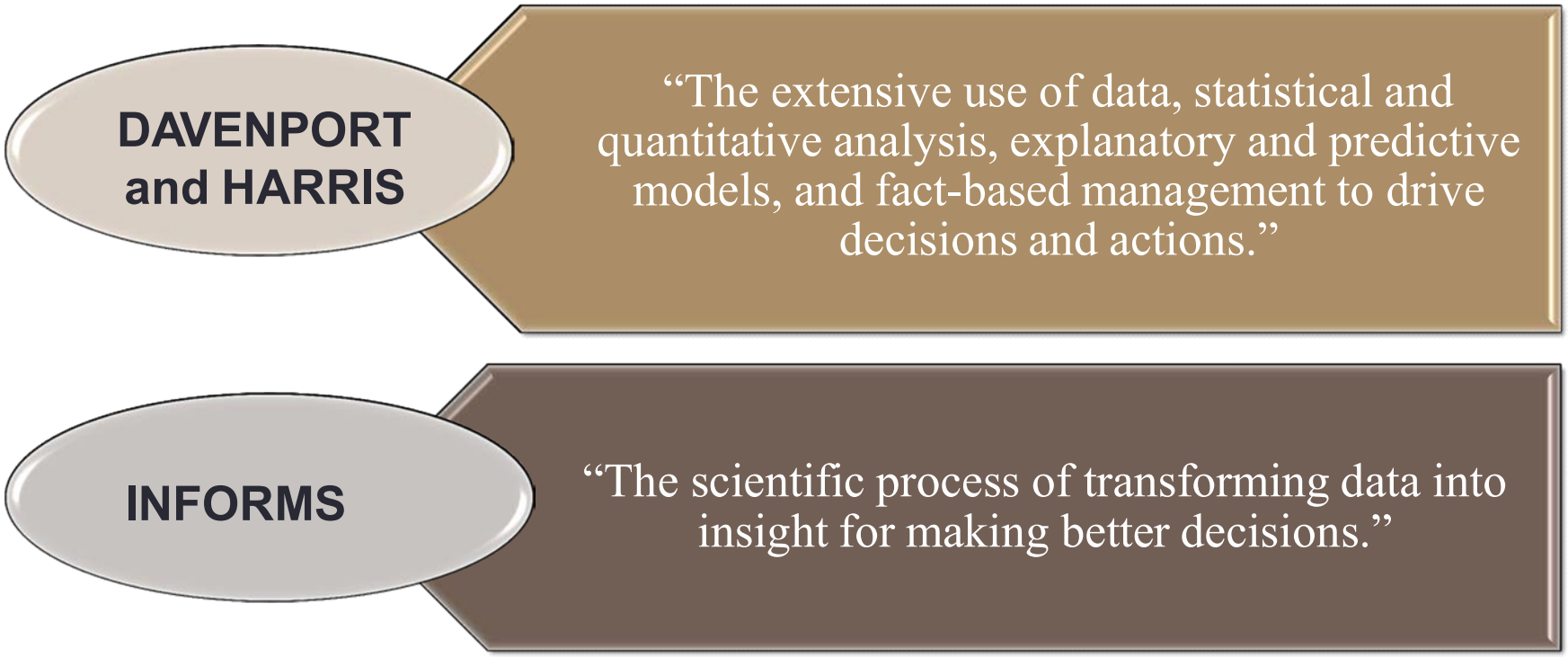
- Business analytics
- Web analytics
- Learning analytics
- Intelligent systems
- Pattern recognition
- Kernel methods
- Computing
- Simulation
- Optimization
- Operations Management
- Natural language processing

- Health care
- Environmental sciences
- Energy & power systems
- Transportation & logistics
- Emergency management
- Supply chain management
- Marketing
- Risk assessment
- Portfolio analysis
- Manufacturing
- Agriculture

History of Business Analytics



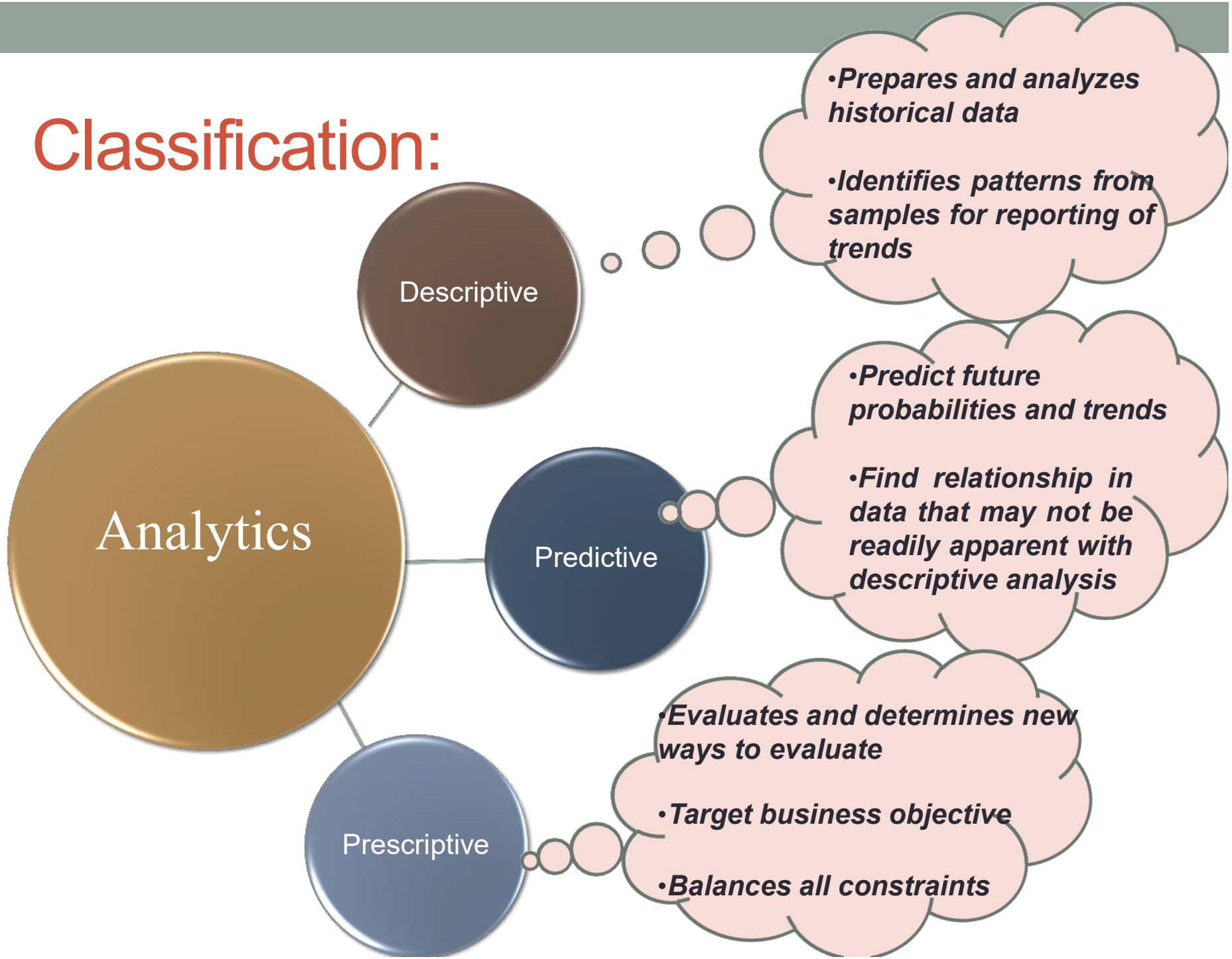
What is Business Analytics



INFORMS defines OR as '*A discipline that deals with the application of advanced analytical methods to help make better decisions'.*

BA extends OR by more broadly including the critical data transformation process to support OR models and decision making.

Classification:



Data Analytics

Descriptive Analytics

Describes what happened in the past

Used for reporting and dashboards, and for preliminary exploratory data analysis to understand the data

Customer segmentation, Clustering

Diagnostic Analytics

"Why did it happen?"

examines data or content to answer the question

Predictive Analytics

Uses models and data from the past to forecast the future

Causal Relationship not assumed

Churn Prediction, Customer Scoring

Prescriptive Analytics

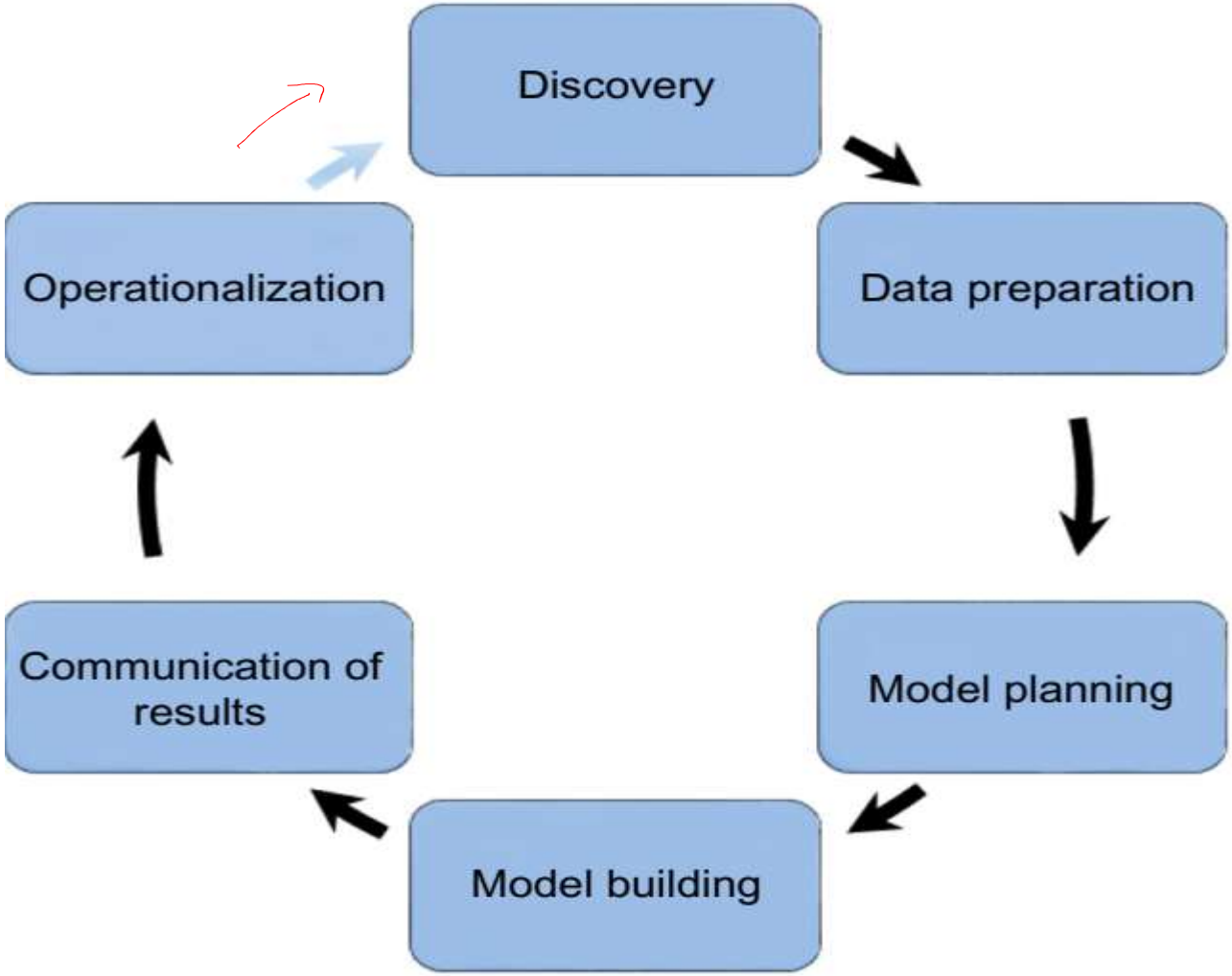
Prescribes actions to perform

Two approaches – Experimental Design, Optimization

Scenario Analysis, Decision Analytics

Classification	Questions	Examples from Business
Prescriptive	What is the best outcome? What if?	Optimisations Scenario testing Randomised tests
Predictive	What could happen? What is happening next? Why is this happening?	Statistical modelling Forecasting
Descriptive	What happened? How many, how often? What action is needed?	Standard and ad hoc reports Queries Alerts

The Data Analytics Lifecycle



The Data Analytics Lifecycle

1. ***Discovery***
   - This stage covers learning about the business problem and the approaches that have been attempted in the past (if any)
   - Assessment of the available data and resources, identification of the important stakeholders, and formulation of the initial hypotheses are included
2. ***Data Preparation***
   - This stage involves collection of the data necessary to address the problem and reformatting it to facilitate successful analysis
   - It often takes 60%–80% of the project time
3. ***Model Planning***
   - This stage covers preliminary data analysis, e.g., exploration of the relationships between different variables to assess which variables appear to be most important
   - Determination of possible models that may be applicable for addressing the business problem is included

The Data Analytics Lifecycle

4. Model Building.

This stage covers implementation and fine-tuning of the models

5. Communication of Results

This stage includes determination of whether the project has accomplished its goals, assessment of the business value of the proposed approach, and summary reports and presentations of key findings to the different stakeholders

6. Operationalization

This stage covers delivery of the technical documents and code, implementation of the model in a production environment, and a pilot project run

Why Analytics?

Deciding to buy a car

Analytical Approach	Non-Analytical approach
Nailing down constraints- time, money, five feature requirement and five wish-lists	Process may start by test driving a car irrespective of any criteria
Prioritize on requirements and wishes Eg: good mileage is high priority while emission low priority;	
Based on must haves and constraints, shortlist cars for test-drive	You either begin creating your own criteria as you go along- may be rejecting some car and loving others based on what you “feels” good.
Grade each vehicle with a 1-5 score on each requirement and wish. Requirement graded with an extra point .	
Take average of requirement and wish list	

A year later after buying a car, XYZ started complaining about his expenditure on fuel/mileage.

To this ABC asked him :

1)Did your office is farther now than one year earlier (job change if any)?

2) If the car is giving lower mileage than expected or advertised?

3)Didn't you buy a car with higher mileage at first place knowing long travels?

(XYZ answered NO to all the above questions)

XYZ: He didn't know the cost would be this high and burden some to him. He really liked the car when he drove it.

ADVANTAGE: Using data to drive decisions deliver a significantly higher chance of making a good, long-lasting decision over non data-driven approach.

Applications

- Managing knowledge from Big Data analytics in New Product Development
- Understanding and targeting customer segmentation
- Predict Security threats
- Predictive support: through sensors and machine generated data companies identify when a malfunction is likely to occur
- Understanding and optimizing business processes
- Natural language processing and text mining


Cost Saving
Time Saving
More Productivity
ADVANTAGES
Increased customer satisfaction
Competitive advantage
Expanded sales/profits

Advantages

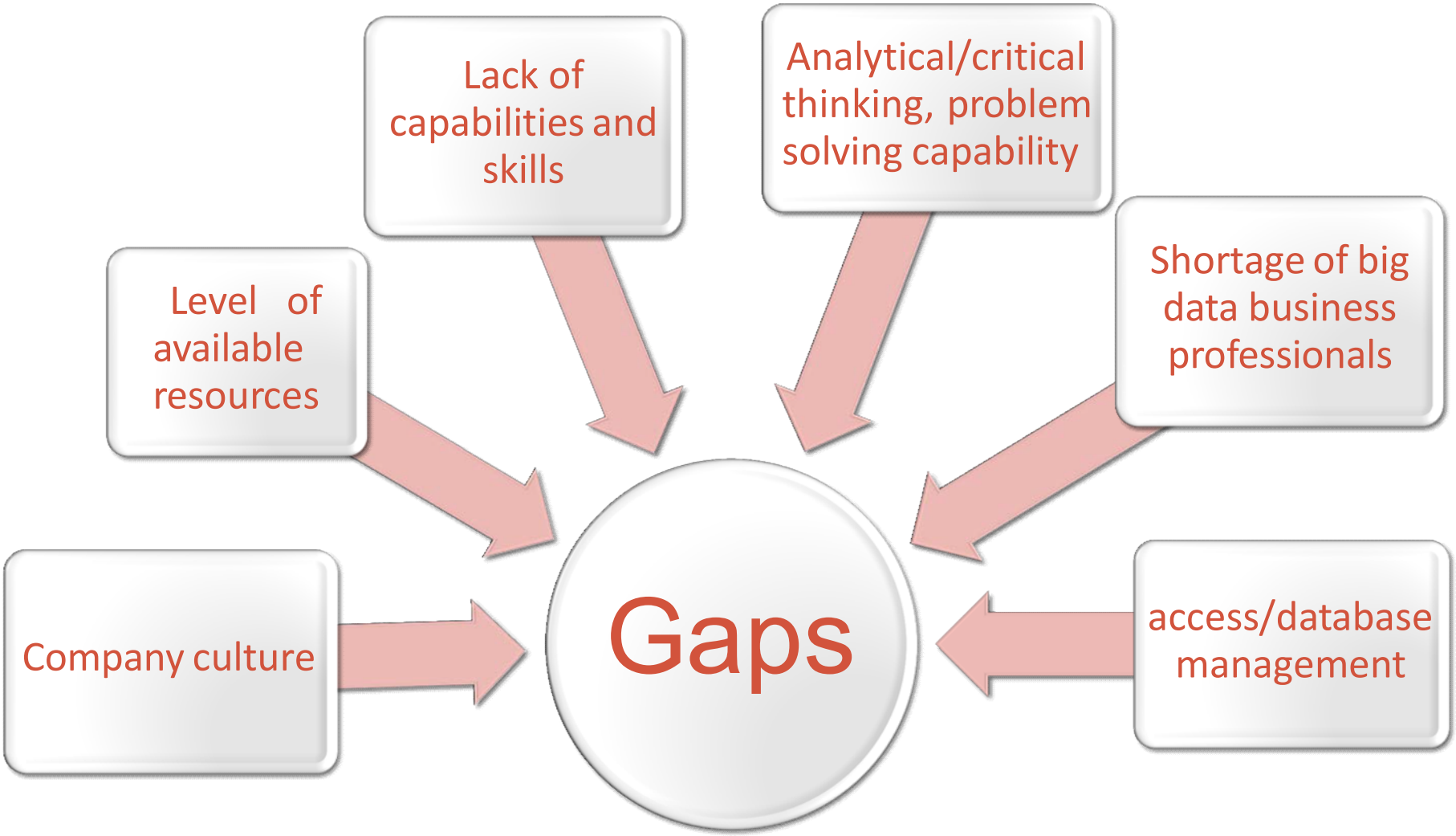
Time Saving

- Achieved through real-time monitoring and forecasting of events that impact business performance/operations

Cost Savings

- Significant cost savings over traditional analytical techniques achieved by adoption of Big Data due to usage of Hadoop clusters

Current Gaps



Adapting to a Technological Revolution

Decision makers need to :

- Plan for their information need
- Learn how to better incorporate statistical results into decision-making
- Digest a haystack of information by judging the accuracy and reliability of the results.

Organization needs a different infrastructure to incorporate BA

- Need of specialization and skill development
- Placing the right people in the right roles

Business Analytics need to evolve

- Expand the tool set
- Adapt the tools for corporations and for big data

BROAD APPLICATIONS OF DATAANALYTICS

Industry Specific

Optimize Funnel Conversion

Data analytics allows companies to track leads through the entire sales conversion process, from a click on an adword ad to the final transaction, in order to uncover insights on how the conversion process can be improved.

EXAMPLE:

CREDEM uses Data Analytics to predict which financial products or services a customer would appreciate, so it can better target consumers during the sales process. With these insights, the bank increased average revenue by 22 % and reduced costs by 9 %.

CREDEM BANCA

La forma e la sostanza.

Company	Industry
Credem	Finance

Behavioral Analytics:

With access to data on consumer behavior, companies can learn what prompts a customer to stick around longer, as well as learn more about their customer's characteristics and purchasing habits in order to improve marketing efforts and boost profits.

EXAMPLE:

McDonalds tracks vast amounts of data in order to improve operations and boost the customer experience. The company looks at factors such as the design of the drive-thru, information provided on the menu, wait times, the size of orders and ordering patterns in order to optimize each restaurant to its particular market

Company	Industry
McDonalds	Food & Beverages

Customer Segmentation

By accessing data about the consumer from multiple sources, such as social media data and transaction history, companies can better segment and target their customers and start to make personalized offers to those customers.

EXAMPLE:

Walmart combines public data, social data and internal data to monitor what customers and friends of customers are saying about a particular product online. The retailer uses this data to send targeted messages about the product, and to share discount offers. Walmart also uses data analysis to identify the context of an online message, such as if a reference to "salt" is about the movie or the condiment.

Walmart
Save money. Live better.

Company	Industry
Walmart	Retail

Predictive Support

Through sensors and other machine-generated data, companies can identify when a malfunction is likely to occur. The company can then pre-emptively order parts and make repairs in order to avoid downtime and lost profits.

EXAMPLE:

Southwest analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction or safety issue. This allows the airline to address potential problems and make necessary repairs without interrupting flights or putting passengers in danger.

Company	Industry
Southwest airlines	Travel

Market Basket Analysis & Pricing Optimization

By quickly pulling data together from multiple sources, retailers can better optimize their product selection and pricing, as well as decide where to target ads.

EXAMPLE:

P&G uses simulation models and predictive analytics in order to create the best design for its products. It creates and sorts through thousands of iterations in order to develop the best design for a disposable diaper, and uses predictive analytics to determine how moisture affects the fragrance molecules in dish soap, so the right fragrance comes out at the right time in the dishwashing process.

P&G

Procter & Gamble

Company	Industry
Procter & Gamble	Household Retail

Cont…

EXAMPLE:

Coca-Cola uses an algorithm to ensure that its orange juice has a consistent taste throughout the year. The algorithm incorporates satellite imagery, crop yields, consumer preferences and details about the flavours that make up a particular fruit in order to determine how the juice should be blended.

Company	Industry
Coca-Cola Co.	Food

Predict Security Threats

Big data analytics can track trends in security breaches and allow companies to proactively go after threats before they strike.

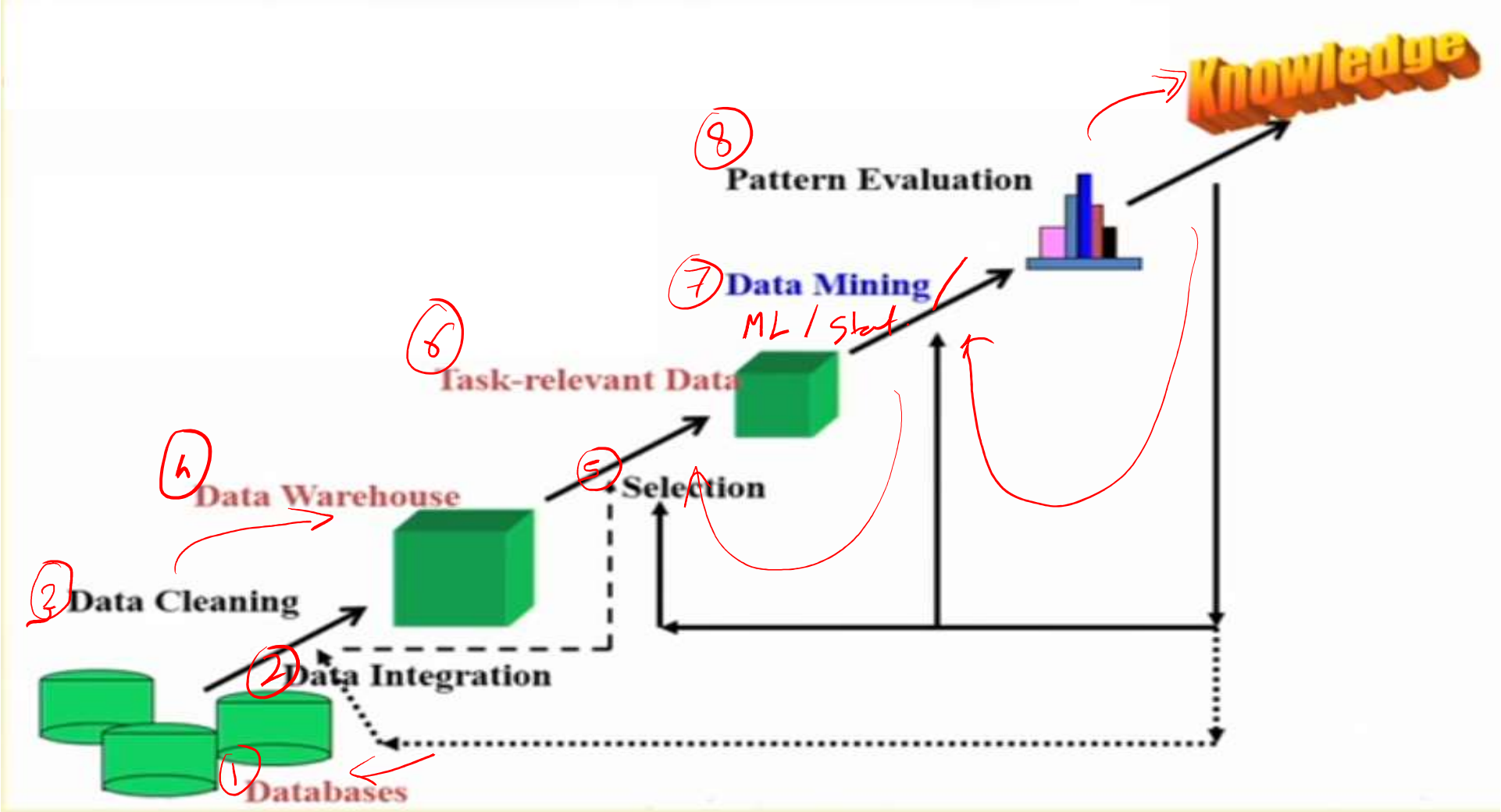
EXAMPLE:

amazon.com

With more than 1.5 billion items in its catalog, Amazon has a lot of product to keep track of and protect. It uses its cloud system, S3, to predict which items are most likely to be stolen, so it can better secure its warehouses.

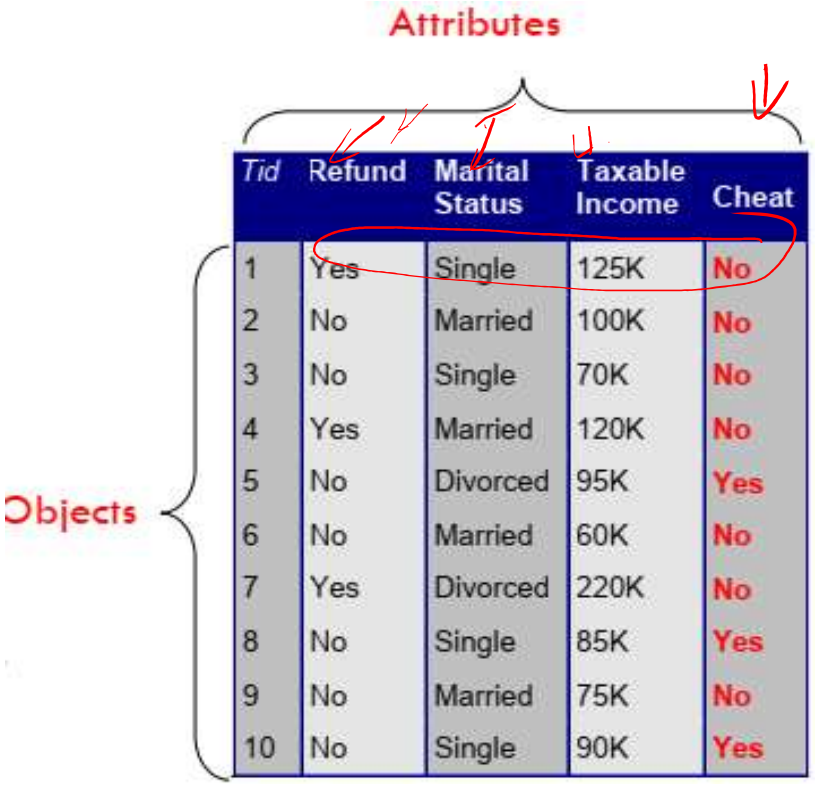
Company	Industry
Amazon	Online Retail

Knowledge Discovery in Data: Process



What is Data?

- a collection of number assigned as value to quantitative variable and/ or characters assigned as value to qualitative variables, or
- collection of records and their attributes
- An attribute is a characteristic of an object
  - Example: Colours of eyes, temperature, etc.
  - Attribute is also known as variable, feature, characteristics, fields, etc.
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity or instances



Tid	Refund	Marital Status	Taxable Income	Cheat
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

Types of Attributes

- Nominal
  - Used to assign individual cases to categories
  - Example: eye colour, ID number, Zip code, etc
- Ordinal
  - Used to rank order cases
  - Example: ranking (eg. movie on scale of 1-10), height (tall, medium, short), grades
- Interval
  - Example: Calendar dates, longitude, latitude
- Ratio
  - Same as interval variable but they have a "true zero"
  - Example: time, length, population, age

Properties of Attribute values

- The type of an attributes depends on which of the following properties it possess:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

- Nominal: Distinctness
- Ordinal: Distinctness, Order
- Interval: Distinctness, Order, Addition
- Ratio: all 4 properties

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countable infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: Binary attributes are special cases of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variable

***STRUCTURED DATA*:**

A data structure is a particular way of storing and organizing data in a computer so that so that it can be used efficiently.

data types: boolean, chart, float, double, array, set, queue, graph, etc.

UNSTRUCTURED DATA:

Unstructured data refers to information that either does not have a predefined data model or is not organized in a predefined manner.

Type of data sets

- Record Data
  - Data Matrix
  - Transaction data
- Graph Data
  - World wide web
  - Molecular structure
- Ordered
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic sequence data

Record Data

Data that consists of a collection of records, each of which consists of fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multidimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Data Matrix Example for Documents

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Doc1 - Ram can run faster than shyam
Doc2 - Shyam can " " Ram
Doc3 - Mohan is playing cricket

	Ram	can	run	faster	than	shyam	Mohan	is	playing	cricket
Doc1	1	1	1	1	1	1	0	0	0	0
Doc2	1	1	1	1	1	1	0	0	0	0
Doc3	0	0	0	0	0	0	1	1	1	1

Transaction Data

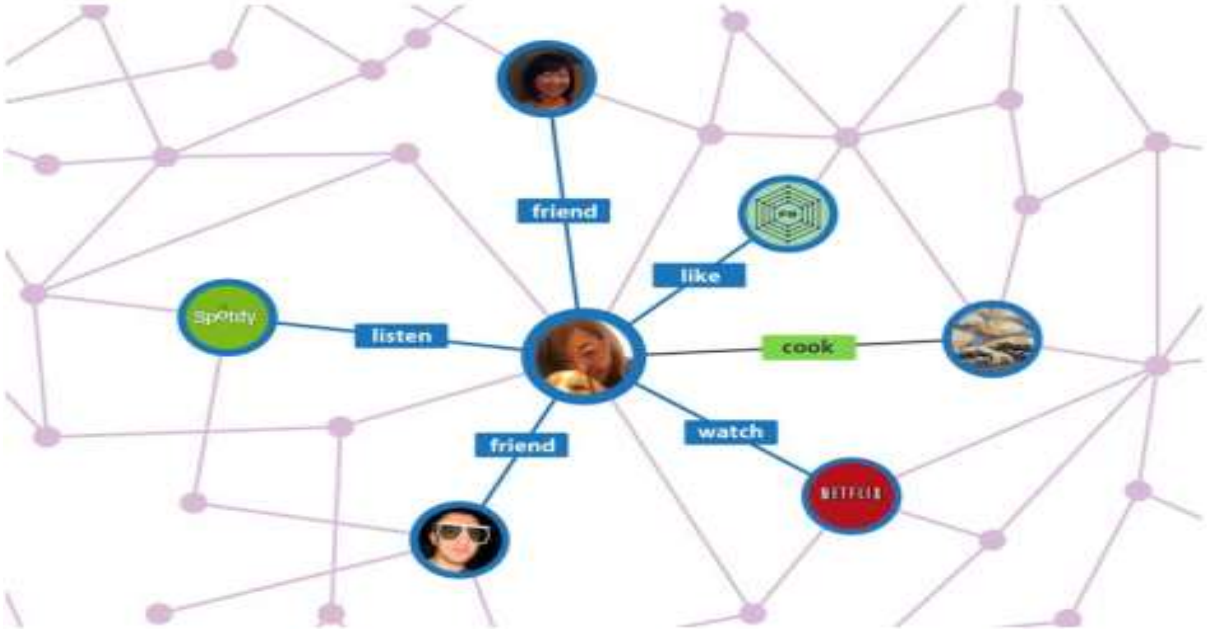
- A typical type of record data, then
  - Each record (transaction) involves a set of items

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

Market-Basket Dataset

Graph data

Example: Facebook graph and HTML links



Ordered data

Genetic sequence data

Species	Alignment of Amino Acid Sequences of β-globin
Human	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Monkey	1 VHLTPEEKNA VTTLWGKVNV DEVGGEALGR LLLVYPWTQR FFESFGDLSS
Gibbon	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Human	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
Monkey	51 PDAVMGNPKV KAHGKKVLGA FSDGLNHLDN LKGTFAQLSE LHCDKLHVDP
Gibbon	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFAQLSE LHCDKLHVDP
Human	101 ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH
Monkey	101 ENFKLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH
Gibbon	101 ENFRLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH

Data Quality

- What kind of data quality problems?
- How can we detect the problem with th data?
- What can we do about these problem?
- Examples of data quality problems:
  - Missing values
  - Noise and outliers
  - Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	~~90K~~	No

Data Quality: Missing Values

Reasons for missing values

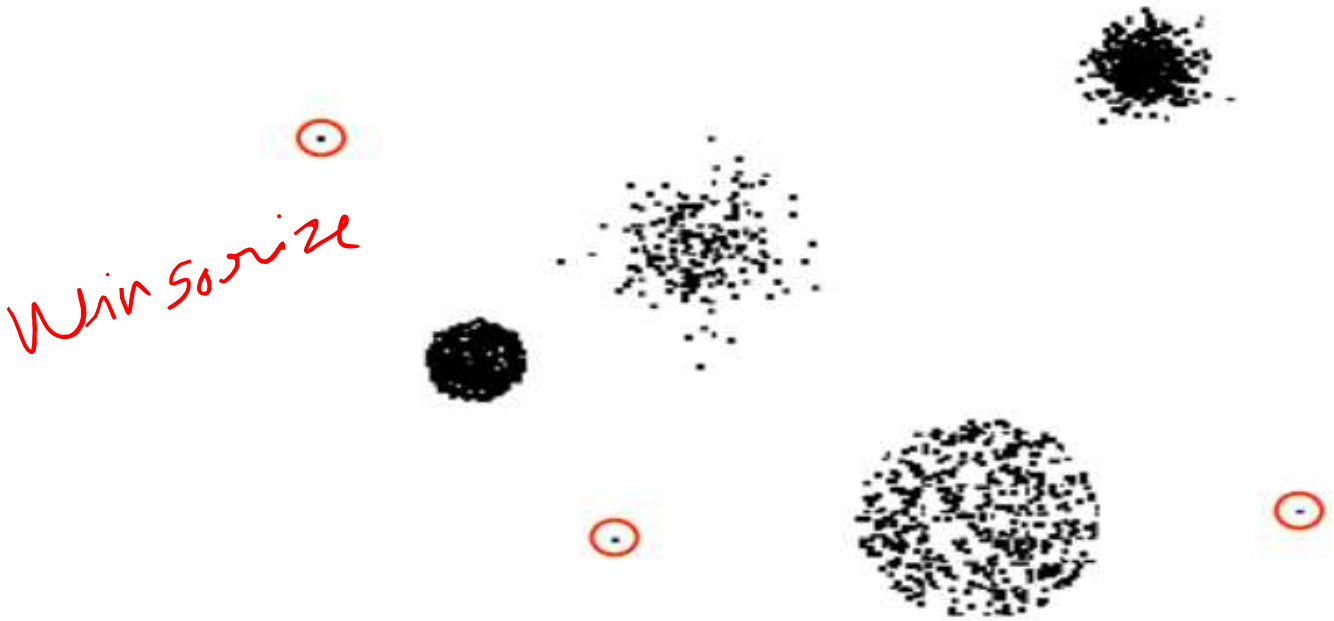
- Information is not collected
  (e.g., people decline to give their age and weight)
- Attributes may not be applicable to all cases
  (e.g., annual income is not applicable to children)

Handling missing values

- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

Data Quality: Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Data Quality: Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

- Imputation
- Outlier management
- Feature selection
  - PCA → Principle component Analysis
  - Correspondce Analyse.

Type of Learning Techniques

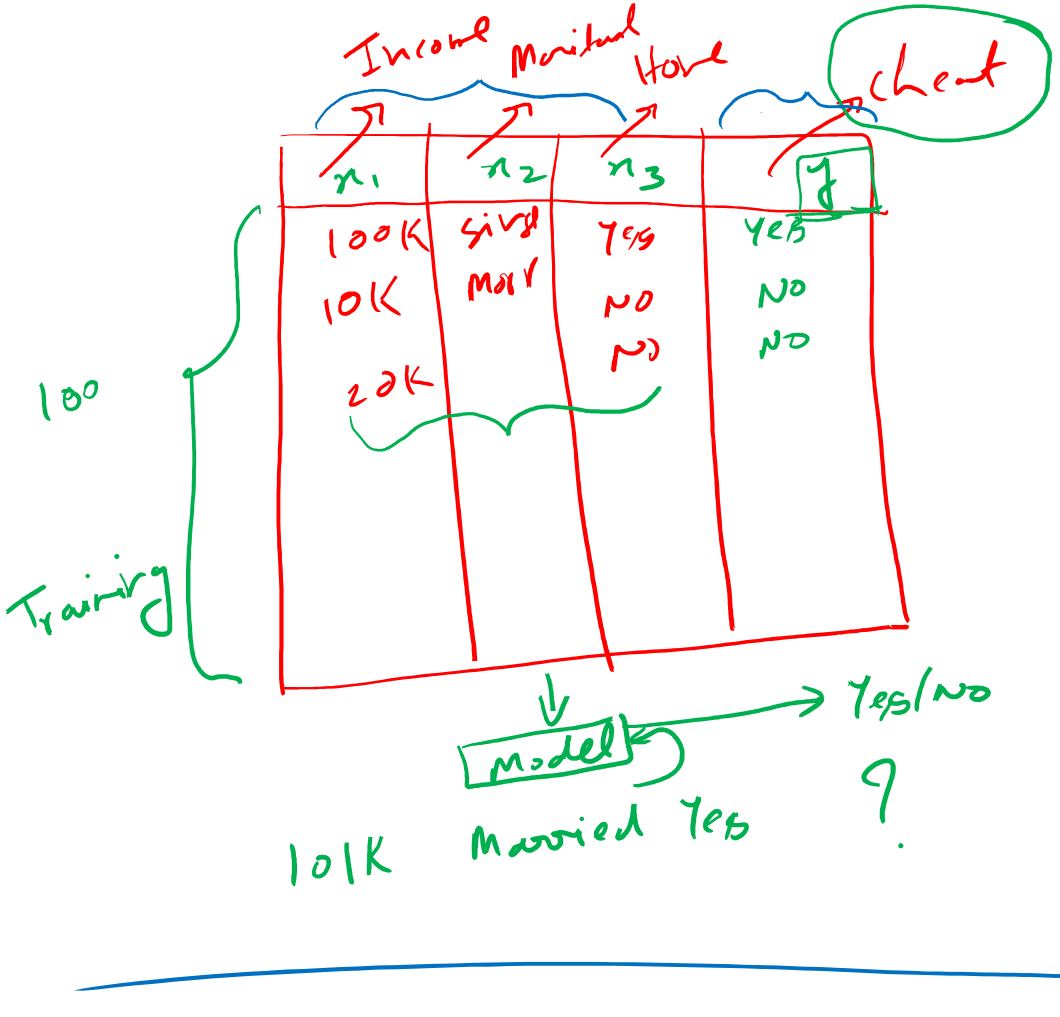
Two Major type of Learning Techniques:

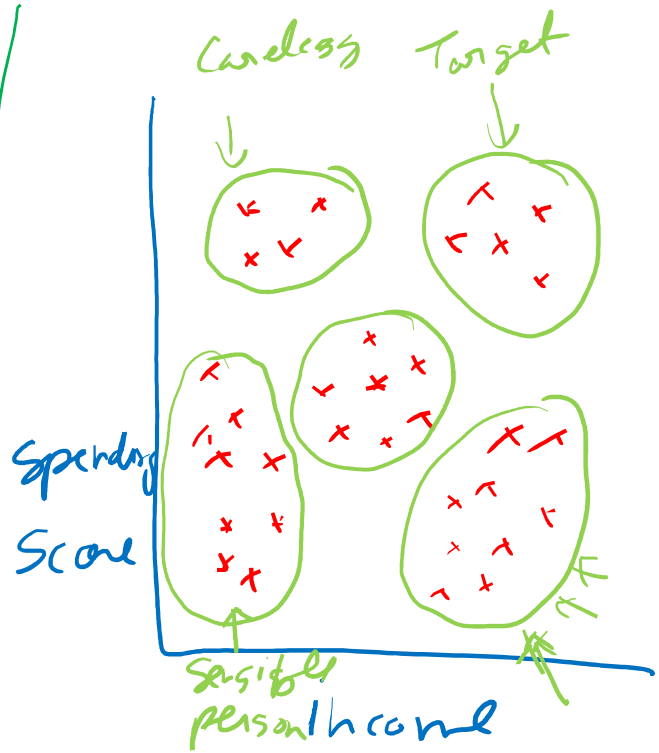
Supervised Learning

Unsupervised Learning

What is Supervised Learning?

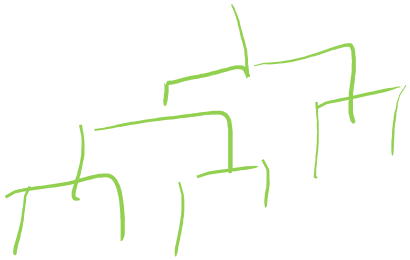
- ***In Supervised learning, you train the machine using data which is well "labeled."***
- ***It means some data is already tagged with the correct answer***
- ***It can be compared to learning which takes place in the presence of a supervisor or a teacher.***


Income
Marital
Home
Cheat
x_1
x_2
x_3
y
100K
10K
20K
Single
Mar
Yes
NO
NO
Yes
NO
NO
100
Training
Model
Yes/NO
101K
Married
Yes
?


Careless
Target
Spending
Score
Possible
Personal Income

Supervised learning algorithms

- **Classification**
  - k-Nearest neighbor
  - Linear classifiers,
  - Support vector machines (SVM),
  - Naive Bayes,
  - Decision trees,
  - Random forest



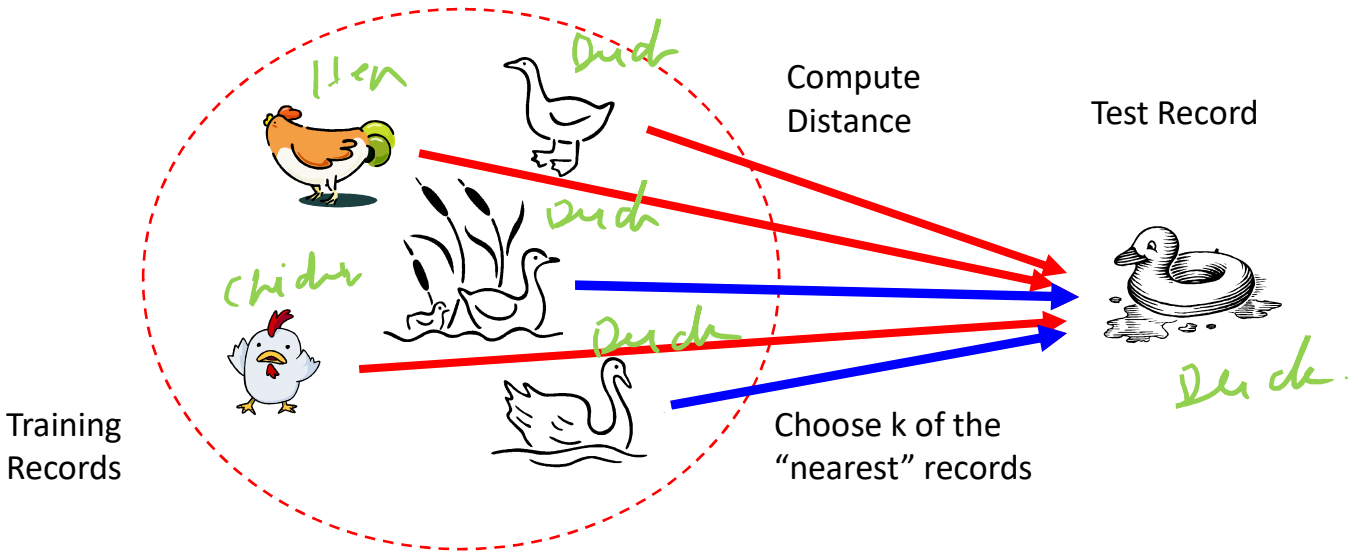
- **Regression**
  - Linear regression,
  - Logistical regression, and
  - polynomial regression

$3x_1 + 2.5x_2 + 4.8x_3 \Rightarrow y$

Classification

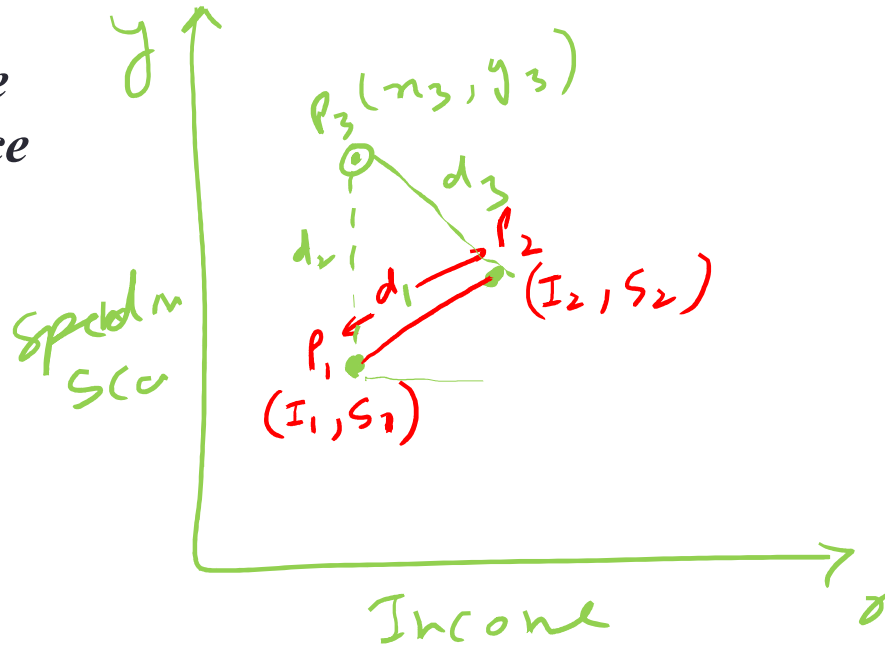
Basic idea:

If it walks like a duck, quacks like a duck, then it's probably a duck



Popular Distance Metric

Euclidean Distance
Manhattan Distance

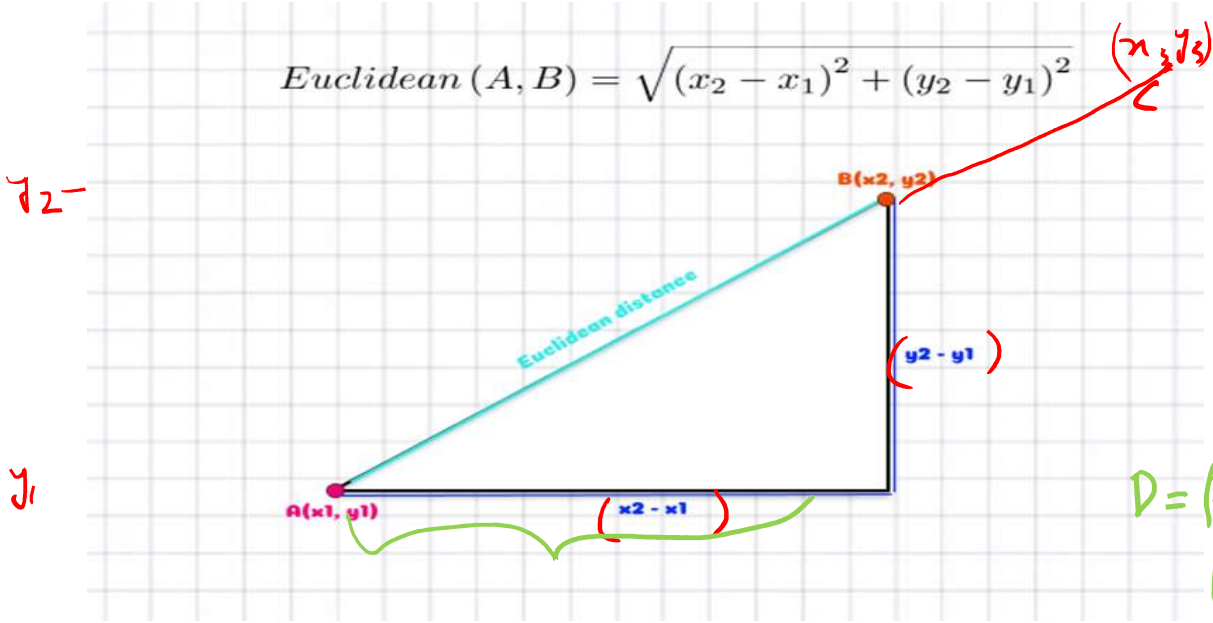


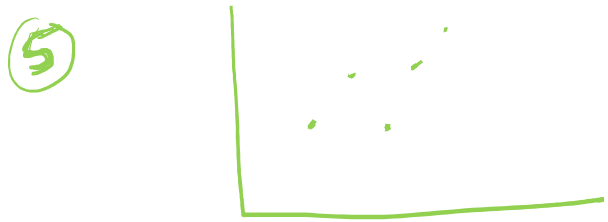
$$D = |x_2 - x_1| + |y_2 - y_1|$$

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \cdots}$$

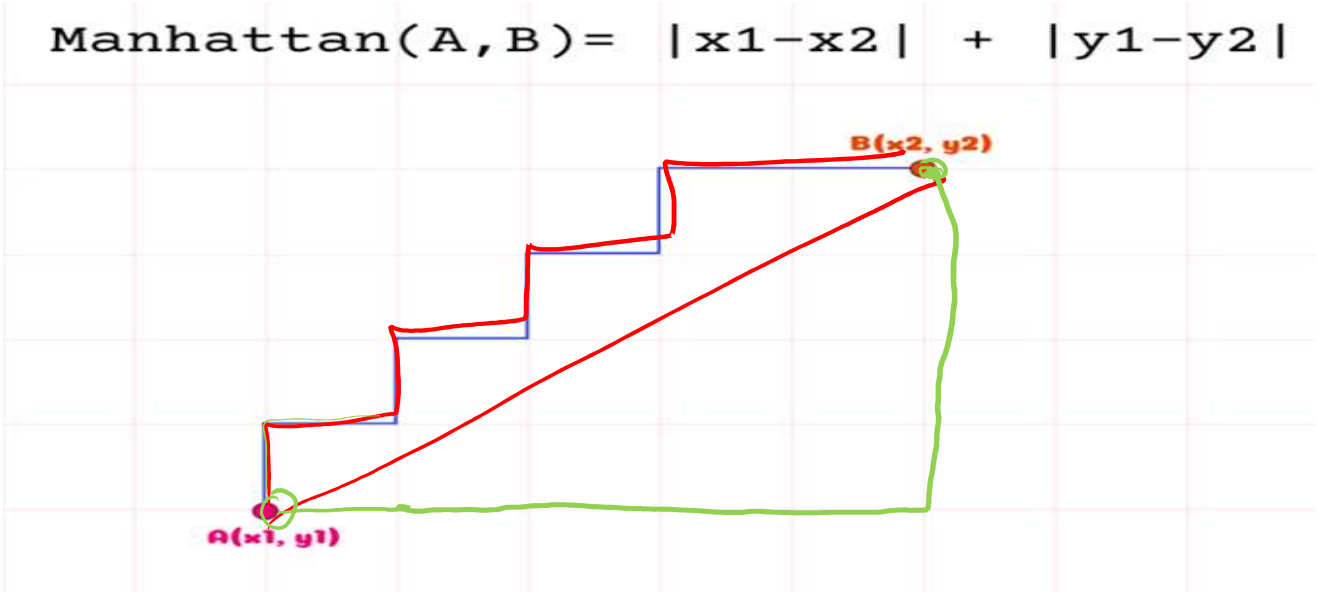
$$D' = \sqrt{(I_2 - I_1)^2 + (S_2 - S_1)^2}$$

Euclidean Distance represents the shortest distance between two data points.

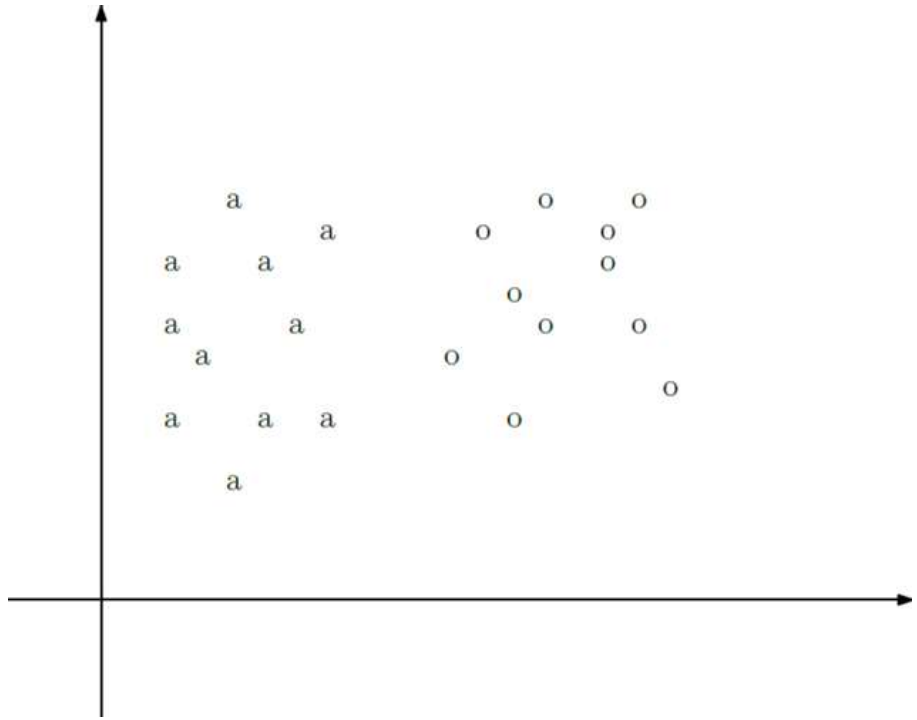




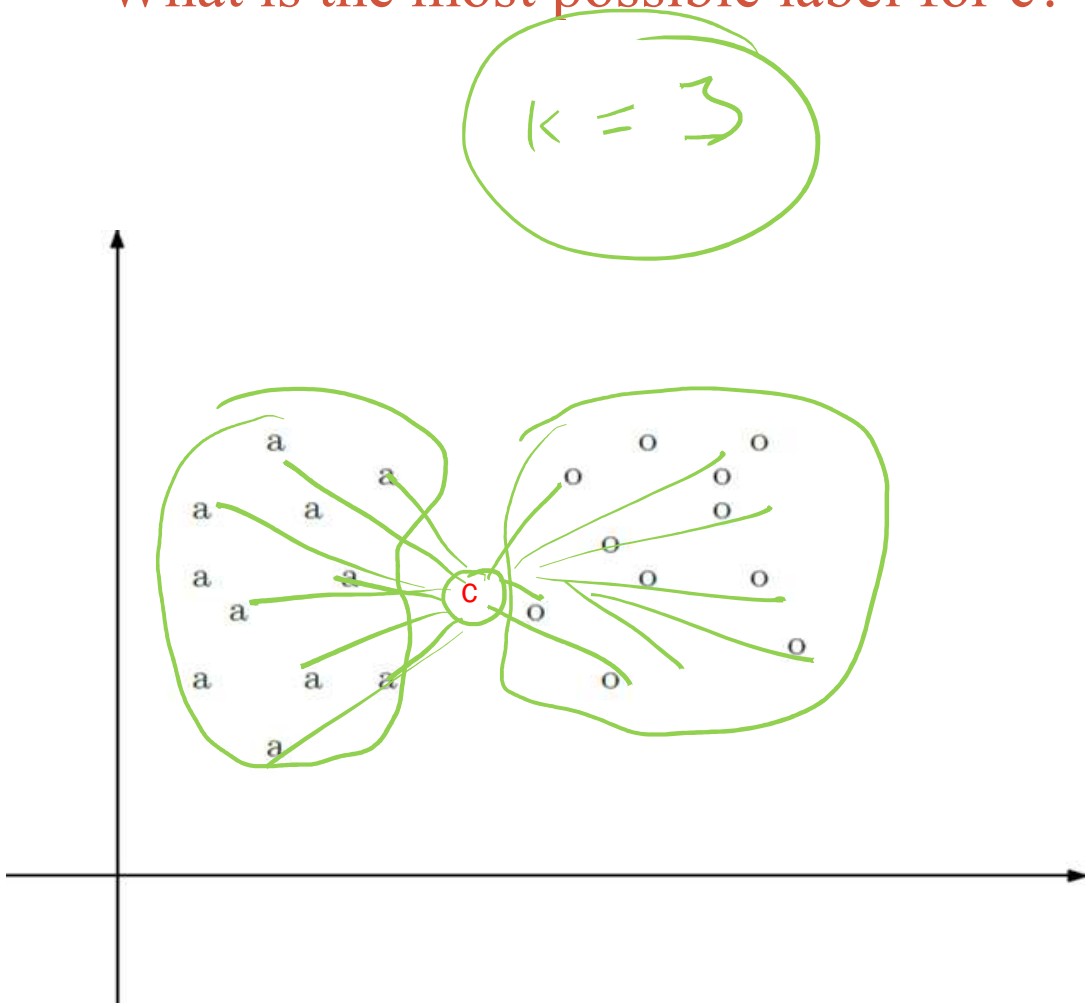
Manhattan Distance is the sum of absolute differences between points across all the dimensions.



(Example)



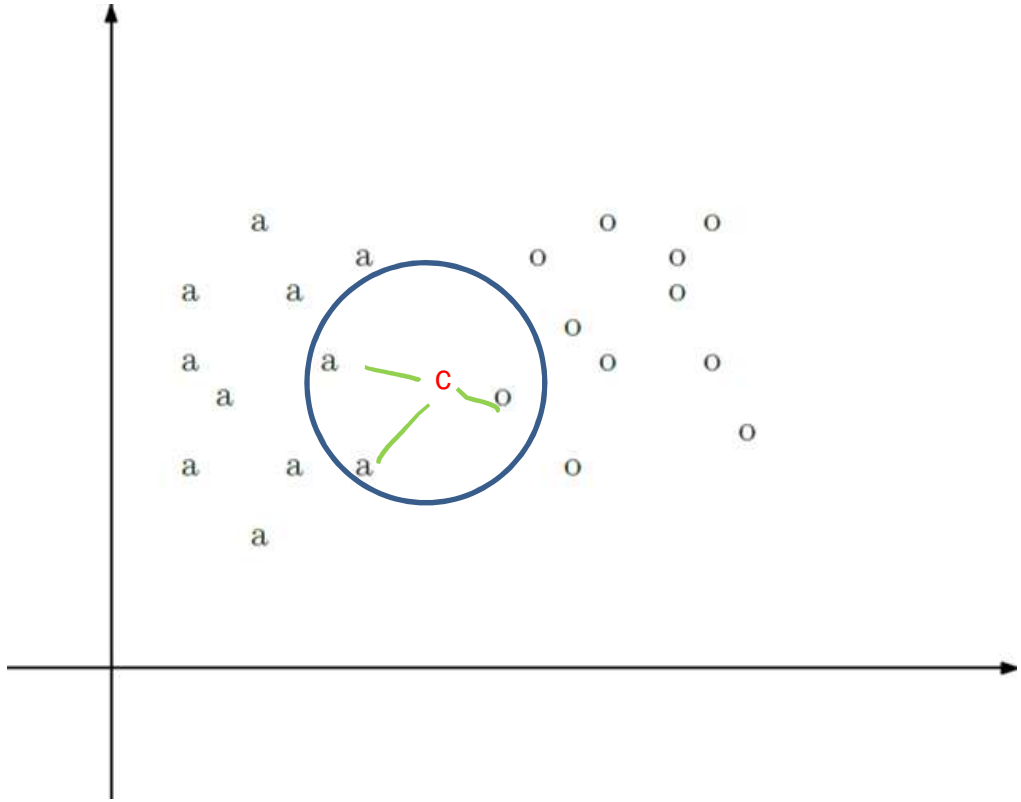
What is the most possible label for c?



What is the most possible label for c?

Solution: Looking for the nearest K neighbors of c.
Take the majority label as c's label
Let's suppose k = 3:

What is the most possible label for c?



What is the most possible label for c?

The 3 nearest points to c are: a, a and o.
Therefore, the most possible label for c is a.

Case: Predicting Admission Chances for UCLA

Prospective graduate students always face a dilemma deciding universities of their choice while applying to master's programs. While there are a good number of predictors and consultancies that guide a student, they aren't always reliable since decision is made on the basis of select past admissions.

What is Unsupervised Machine Learning?

Unsupervised learning is a technique, where you do not need to supervise the model.

Instead, you need to allow the model to work on its own to discover information.

It mainly deals with the unlabelled data.

How Unsupervised Machine Works?

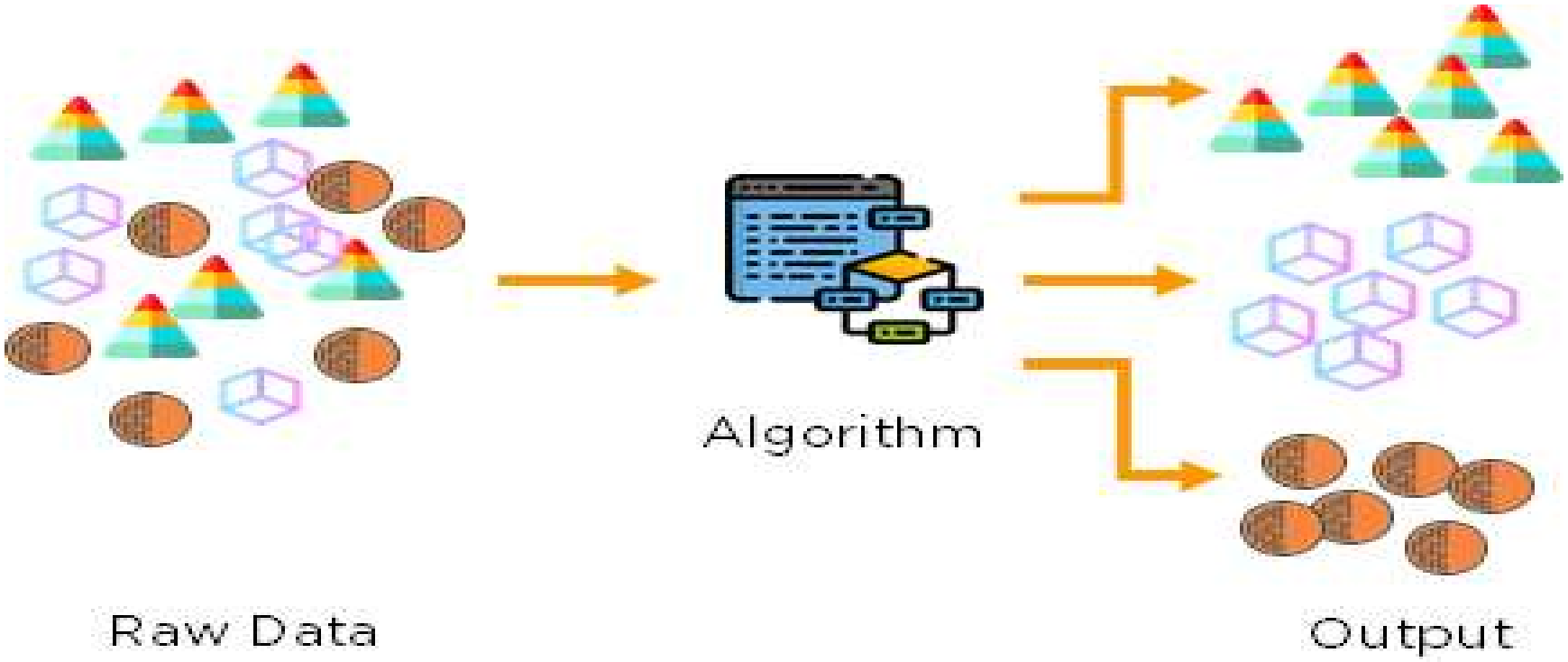


Image Source: https://medium.com/

Common unsupervised learning approaches

Unsupervised learning models are utilized for three main tasks:

- Clustering,
  - K-Mean Clustering,
  - Hierarchical Clustering
- Association

When are Analytics not practical?

- When there's no time
- When there's no precedent
- When history is misleading?
- When the decision maker has considerable experience
- When the variable can't be measured

?

QUESTIONS IF ANY?

THANK YOU ☺