



# Statistical Data Analysis

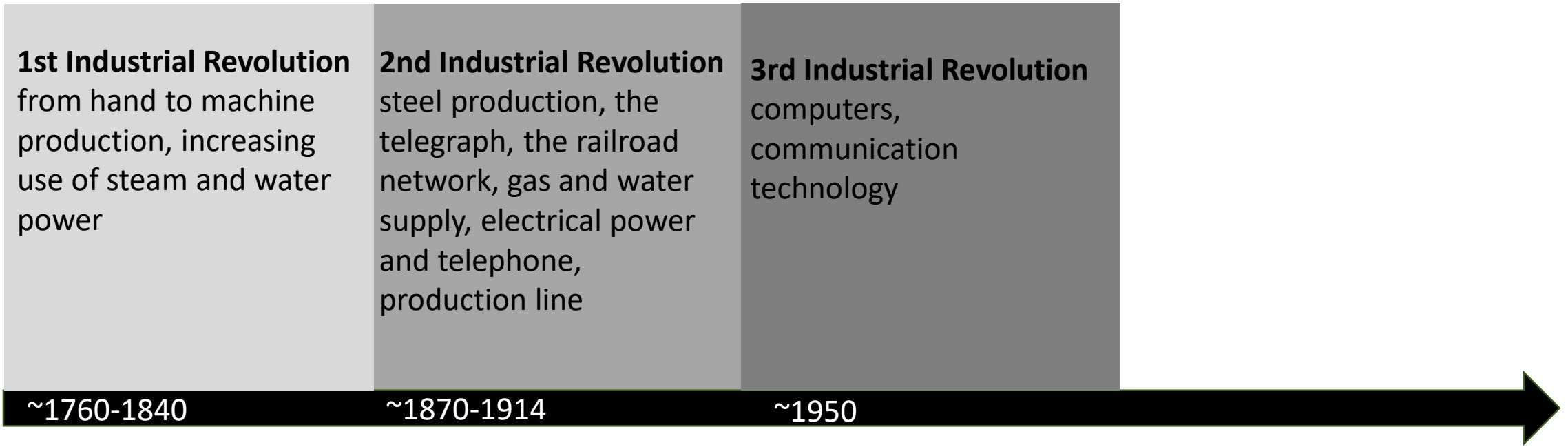
Industry 4.0 seminar series

Frédérique Oggier

Division of Mathematical Sciences, NTU

# Technological Revolution

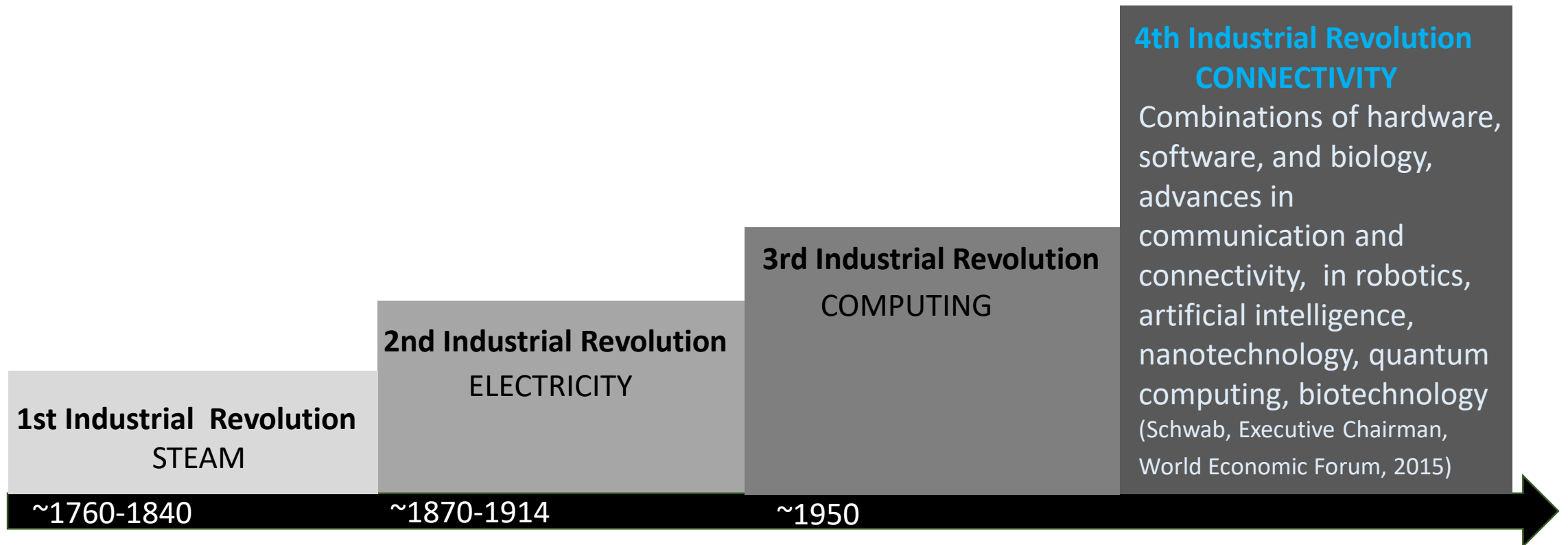
A period in which a technology (or several) is replaced in a short amount of time, creating innovations and abrupt changes in society.



Source: wikipedia

# Industry 4.0

Trend towards automation and data exchange in manufacturing technologies, inclusion of the internet of things, cloud computing, artificial intelligence.



Source: wikipedia

Our website uses cookies. By using this website, you have provided consent for us to continue using cookies to improve your user experience. Your privacy matters to us. To disable cookies, you may read our [Privacy Policy](#) for more information.



Manufacturing

## Singapore's advanced manufacturing avatar "Industry 4.0"

10 Jul 2017

BRANDED CONTENT

## Empowering the Singapore workforce for Industry 4.0

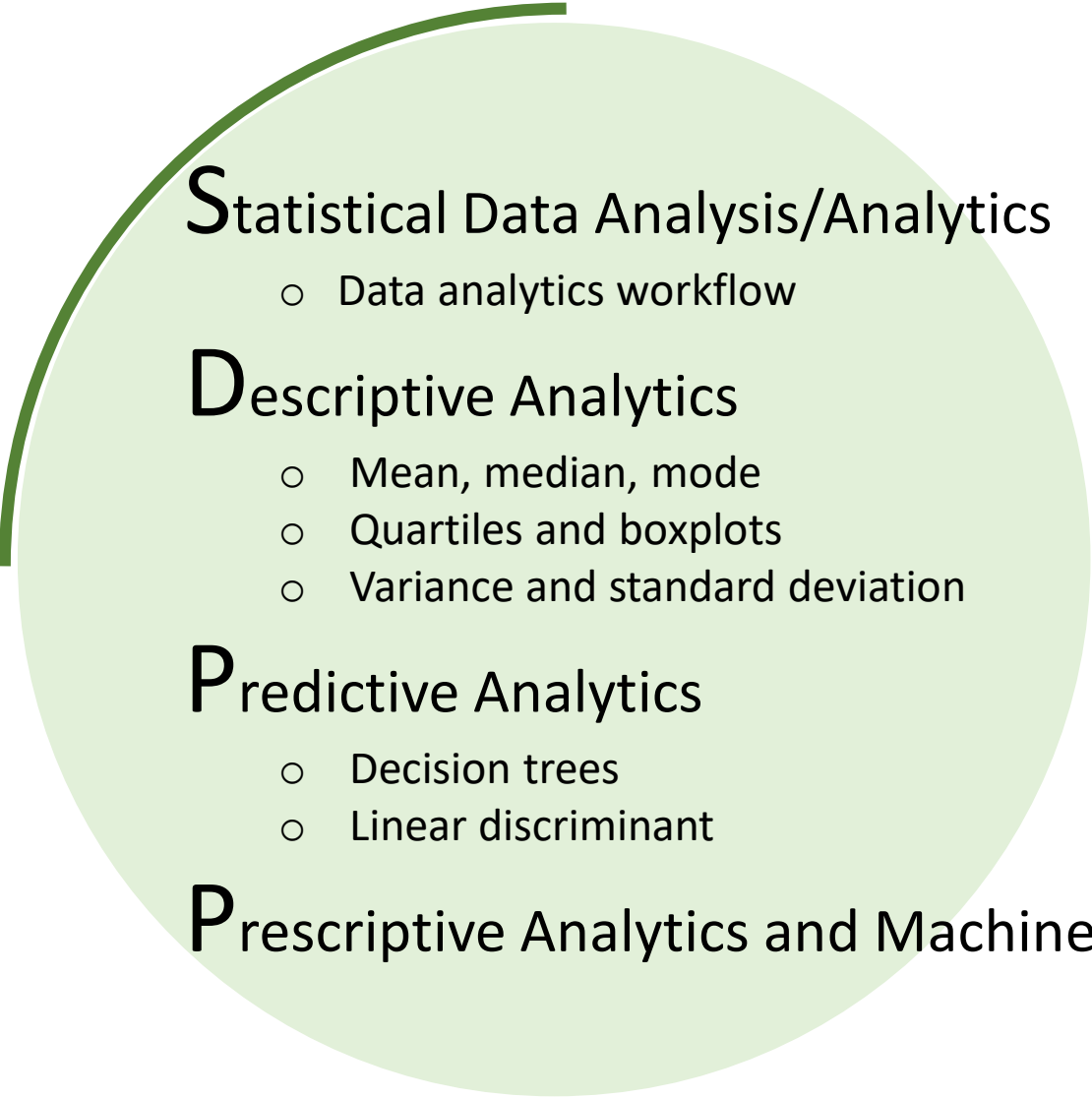


To survive and to lead in today's business environment, organisations need to innovate and transform how operations are run with smart technologies. PHOTO: SCHNEIDER ELECTRIC

“

Industry 4.0 follows the adoption of computers and automation and enhances it with smart systems powered by actionable data. This means data that is parsed and interpreted via **analytics**, to extract knowledge and insights which in turn support businesses and industries.

”



## Statistical Data Analysis/Analytics

- Data analytics workflow

## Descriptive Analytics

- Mean, median, mode
- Quartiles and boxplots
- Variance and standard deviation

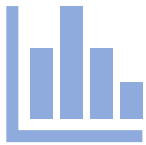
## Predictive Analytics

- Decision trees
- Linear discriminant

## Prescriptive Analytics and Machine Learning

# Data analytics

The discovery, interpretation, and communication of meaningful patterns in data.



## **Descriptive**

What has happened?



## **Predictive**

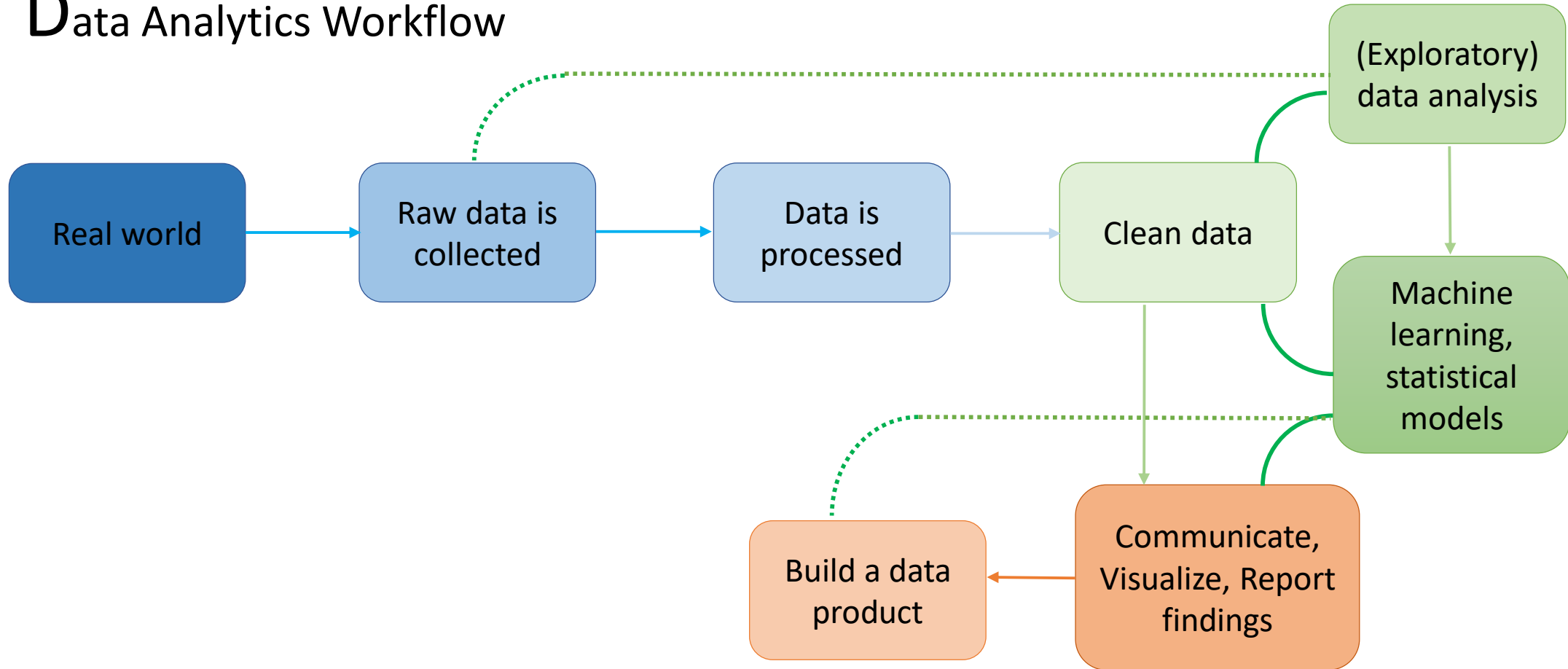
What will happen?



## **Prescriptive**

What should I do?

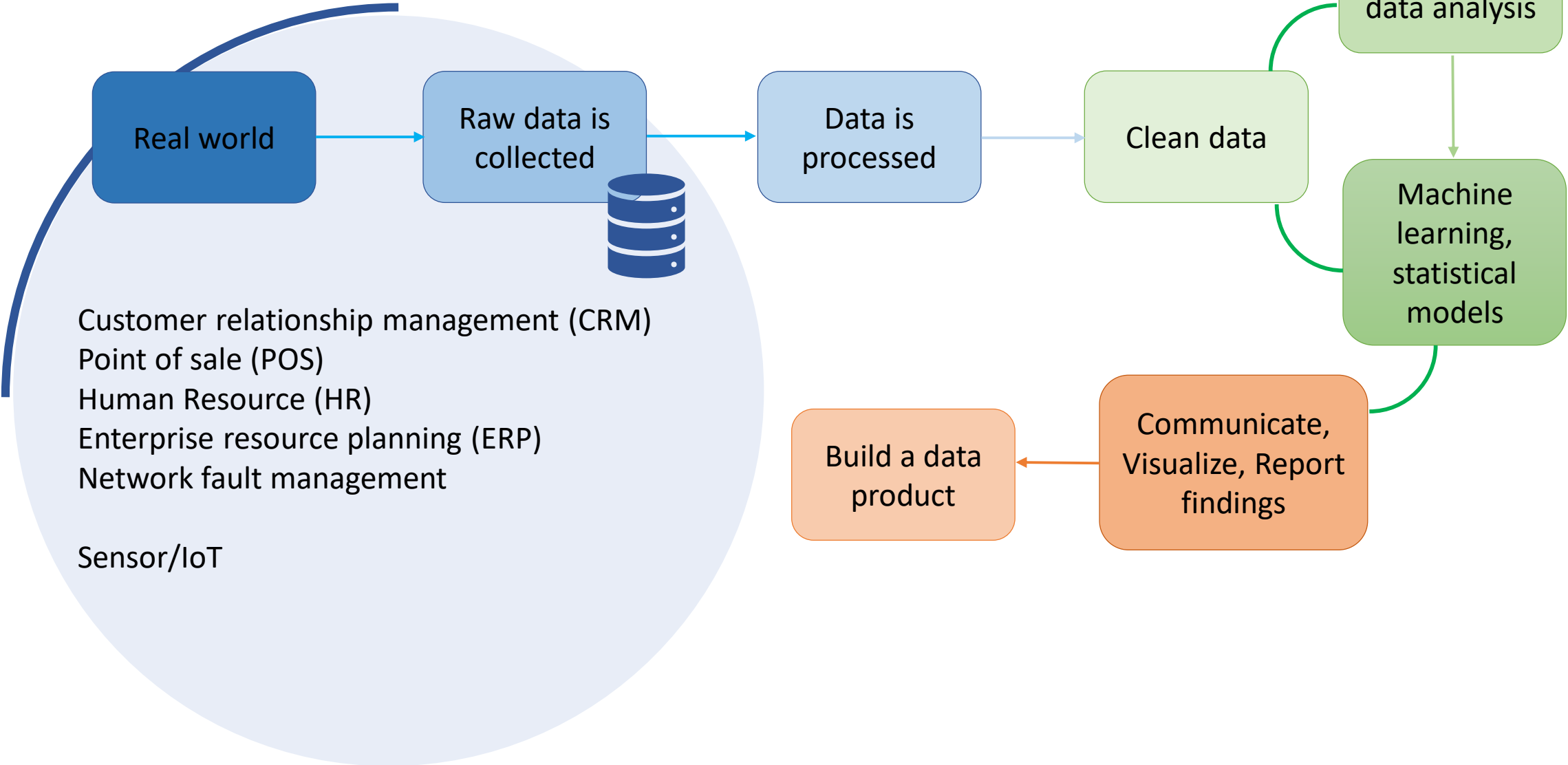
# Data Analytics Workflow



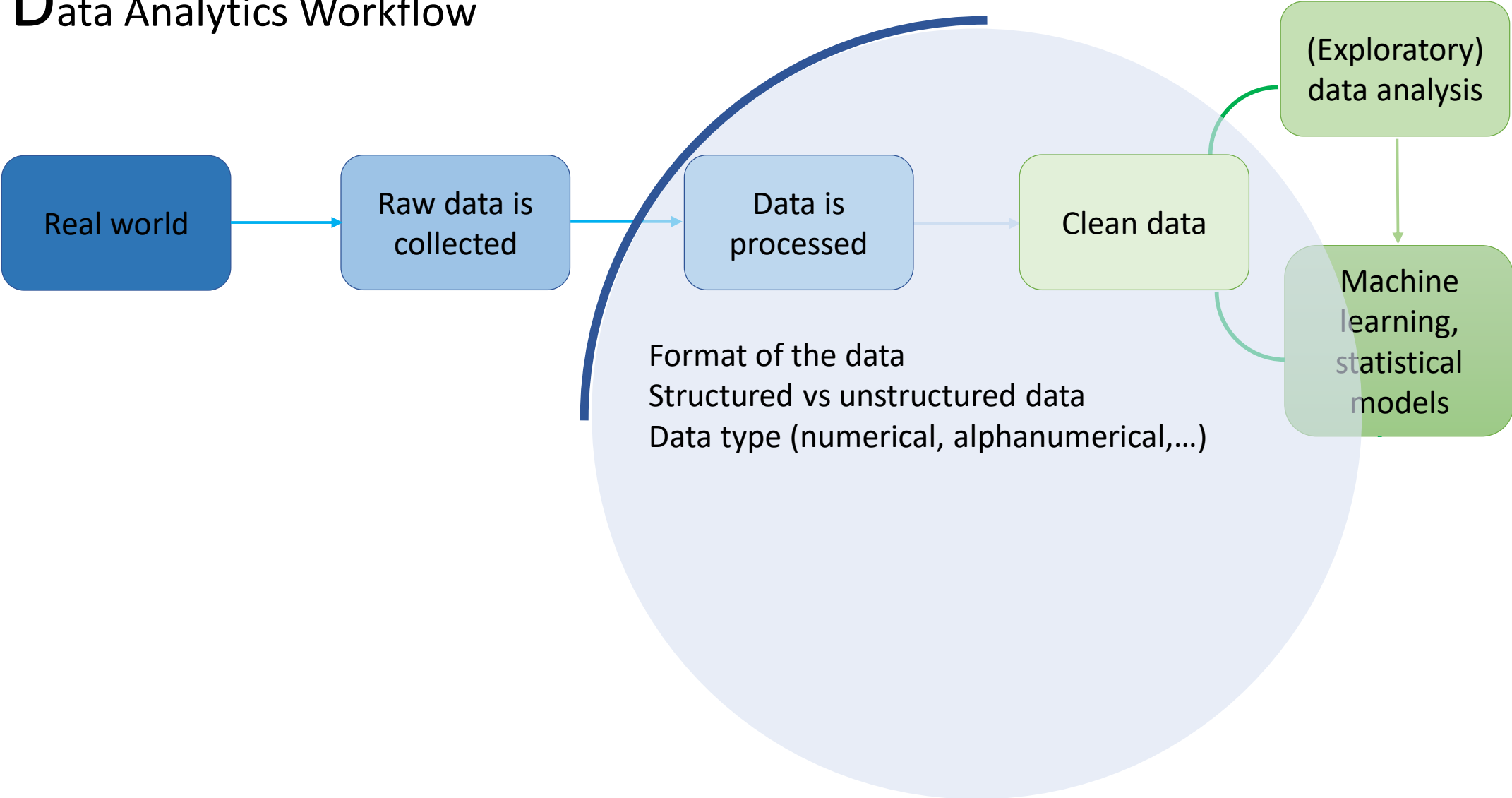
Source: "Doing Data Science" by R. Schutt and C. O'Neil)



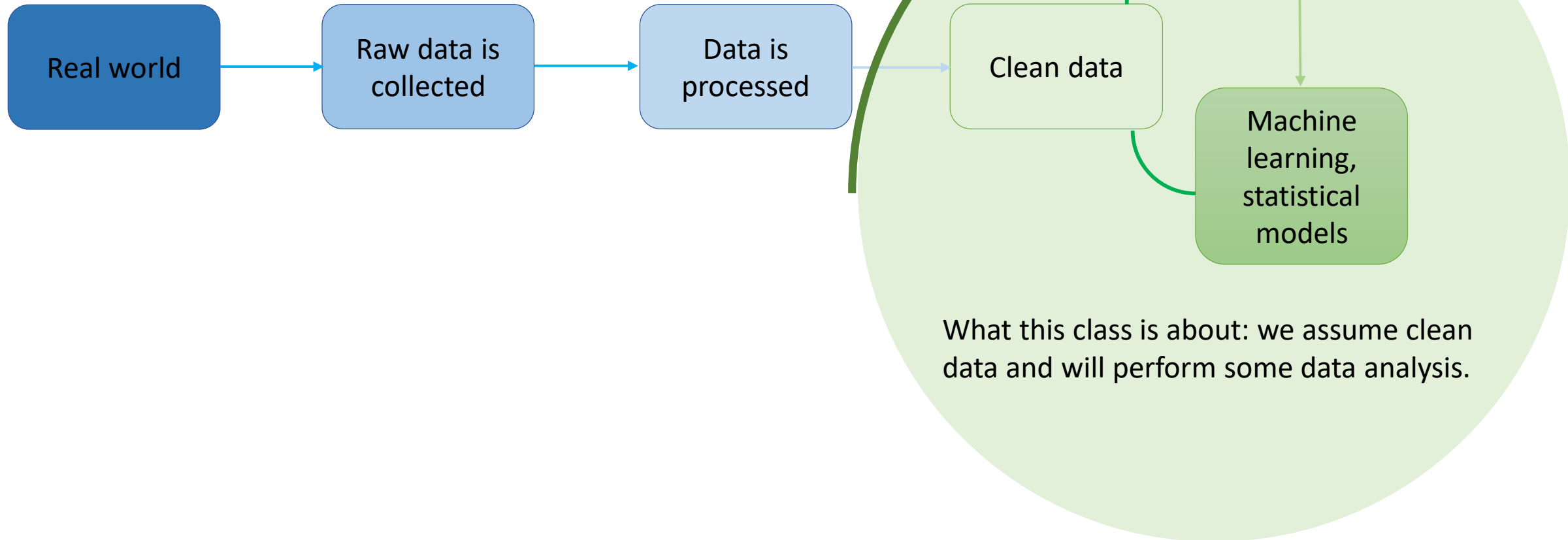
# Data Analytics Workflow



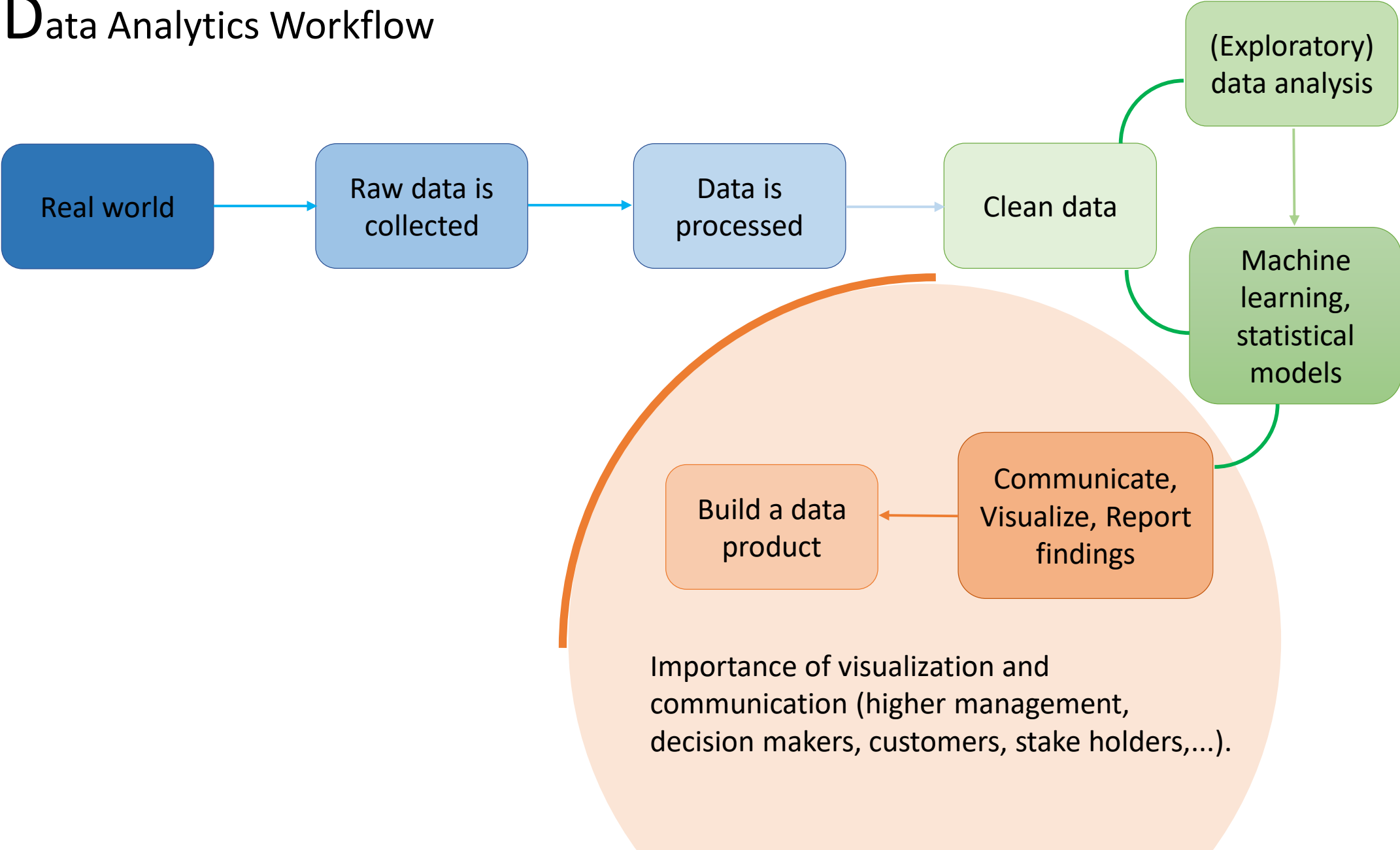
# Data Analytics Workflow



# Data Analytics Workflow



# Data Analytics Workflow



# Statistical Data Analysis/Analytics

- Data analytics workflow

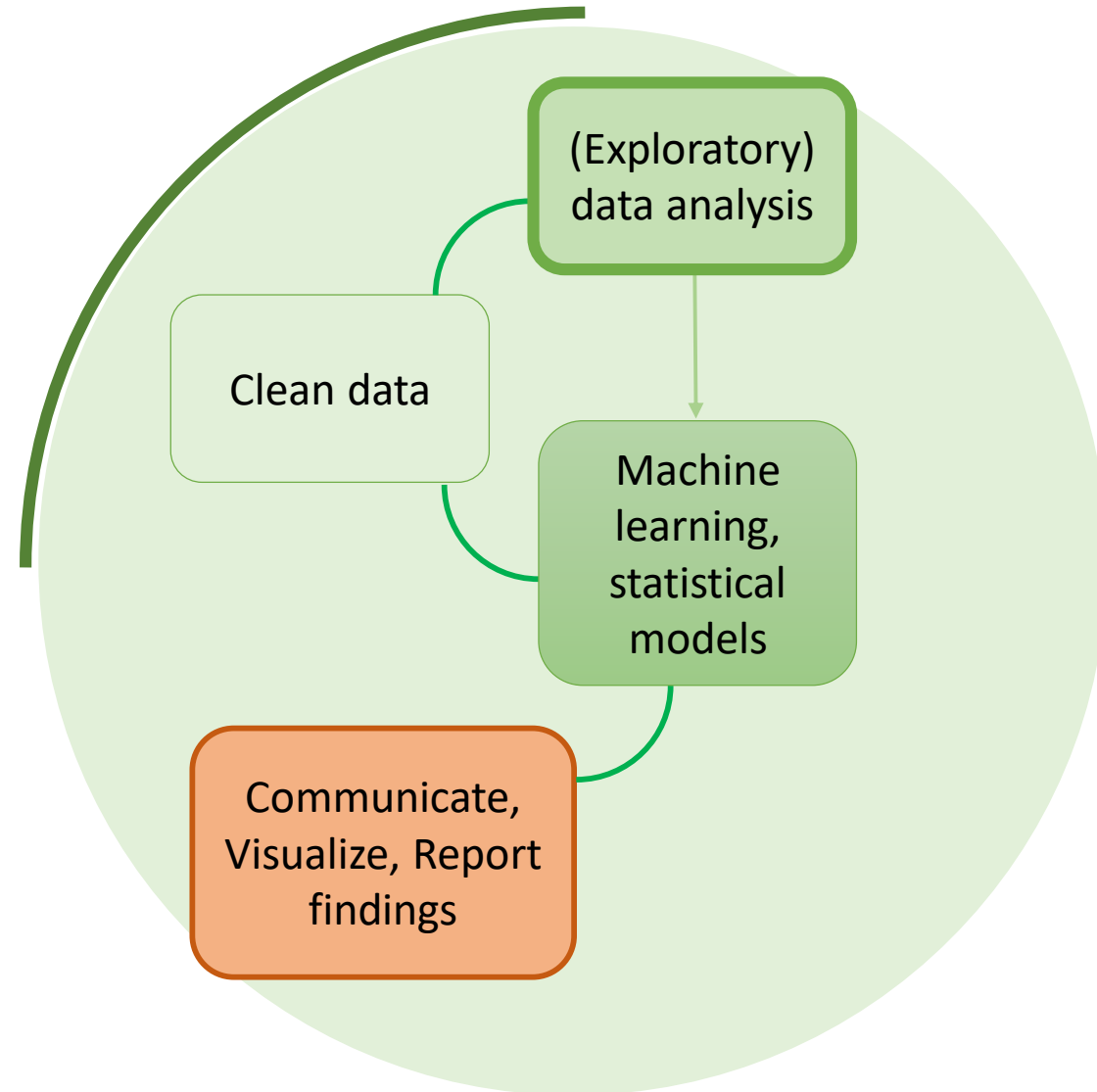
## Descriptive Analytics

- Mean, median, mode
- Quartiles and boxplots
- Variance and standard deviation

## Predictive Analytics

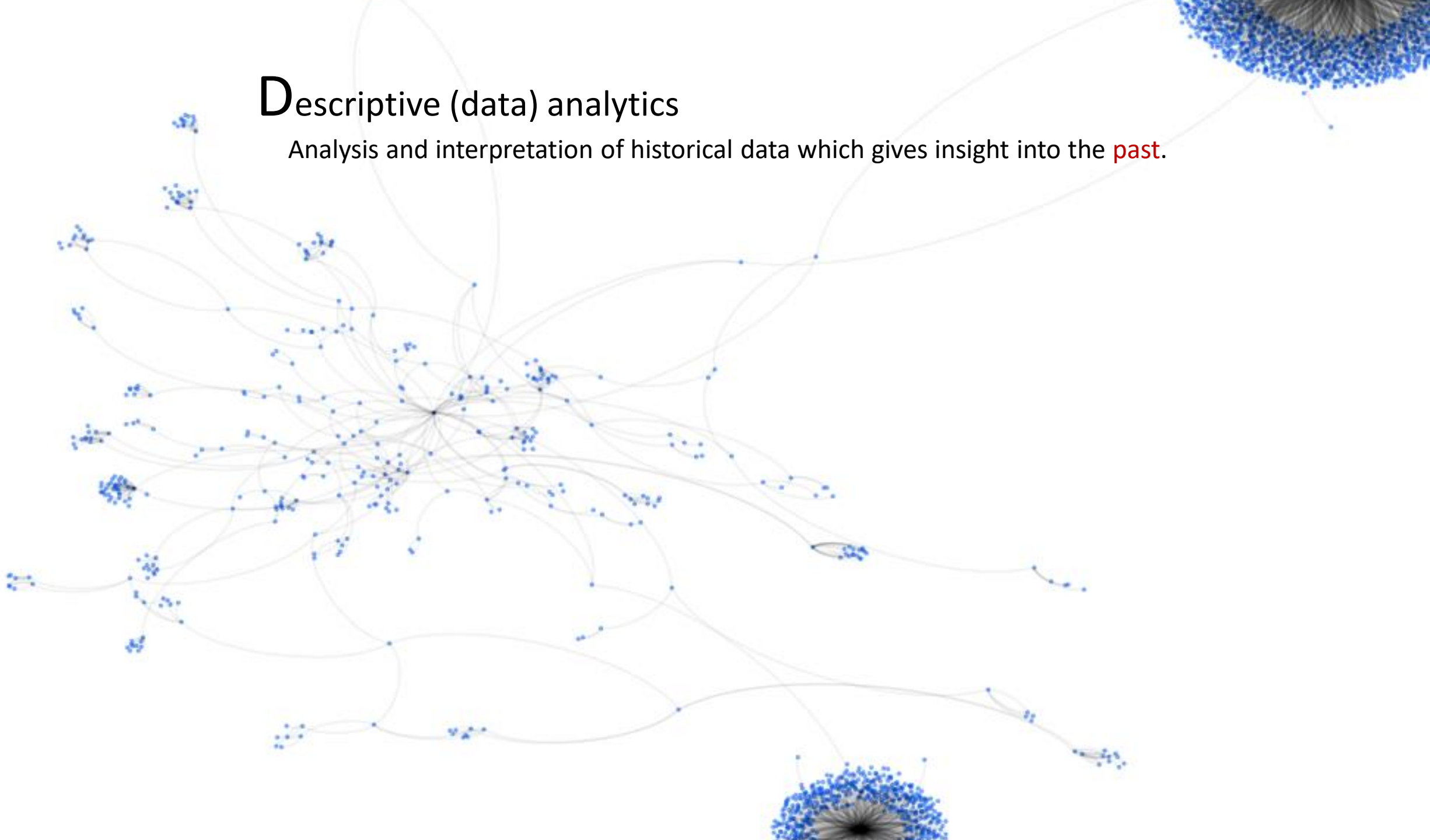
- Decision trees
- Linear discriminant

## Prescriptive Analytics and Machine Learning



# Descriptive (data) analytics

Analysis and interpretation of historical data which gives insight into the **past**.



A Singapore Government Agency Website

DEPARTMENT OF STATISTICS SINGAPORE

What's New Find Data Publications Our Services and Tools Standards Who We Are Careers

## SINGAPORE ECONOMY

### GDP 2018 AT CURRENT MARKET PRICES

# S\$491,175 Mil

Per Capita GDP at Current Prices: \$87,108

Real GDP Growth: 3.1%

Scroll down to see more

**2019 QS WORLD UNIVERSITY**

Tweets	Following	Followers	Likes
3,843	245	16.2K	1,077

TEMASEK

中文

## Building a disciplined institution

Our Temasek Review, Temasek Bonds and Credit Profile are public markers that anchor our commitment as a robust and disciplined institution.

Our Credit Profile provides a snapshot of our credit quality and the strength of Temasek's financial position.

[SEE CREDIT QUALITY](#)

<b>21x</b> Net portfolio value over total debt	<b>7x</b> Liquid assets over total debt	<b>22x</b> Dividend income over interest expense	<b>4x</b> Liquidity balance over total debt due in next 10 years
---	--	---	---

**Portfolio Performance**  
We track our total returns to our shareholder over various periods.

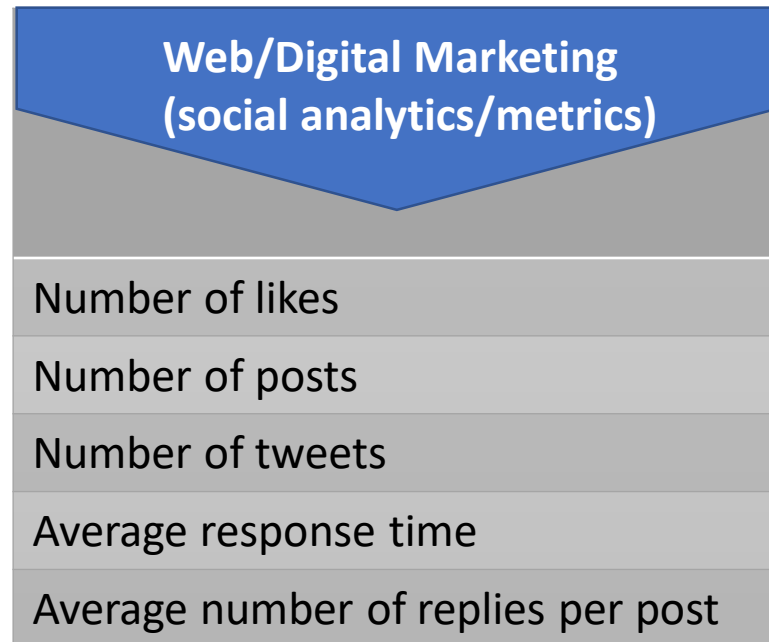
[VIEW OUR PERFORMANCE](#)

**S\$ Total Shareholder Return (%)**

Year	Return (%)
2015	15
2014	12
2013	7
2012	14
2011	1

As at 31 March 2015.

# Examples





# Descriptive Analytics: what for?

- For summarizing, describing different aspects of your business.
- For reviewing regional sales, customer attrition rate/churn, success of a marketing campaign.
- For summarizing social metrics, e.g. Facebook likes, tweets, followers, which in turn provide information about "hot" or "new" trends (e.g. food, travel, app), about the success of digital marketing.
- For reporting your company's production, operations, finance.
- For reporting faults (when? where? how many?)

# Mean

Given a list of numbers (data set), the quantity obtained by adding up all the numbers and then dividing by the size of the list (this is "the average" you are used to).

## Examples

List 1={1, 2, 4, 4, 4, 6, 7}: we add up the numbers  $1+2+4+4+4+6+7 = 28$ ,  
there are 7 numbers in the list, so the mean is  $28/7 = 4$ .

List 2={1, 2, 2, 3, 5, 5, 7, 9}: we add up the numbers  $1+2+2+3+5+5+7+9 = 34$ ,  
there are 8 numbers in the list, so the mean is  $34/8$ .

List 3={3, 1, 2, 5, 4, 6, 8}: we add up the numbers  $3+1+2+5+4+6+8 = 29$ ,  
there are 7 number in the list, so the mean is  $29/7$ .

# Median

Given a list of numbers (data set), the middle value in this list.

## Examples

List 1={1, 2, 4, 4, 4, 6, 7}: the middle value is 4 since {1, 2, 4, **4**, 4, 6, 7}, the median is 4.

List 2={1, 2, 2, 3, 5, 5, 7, 9}: there are two middle values, 3 and 5, so the median in this case is given by  $(3+5)/2 = 4$ .

List 3={3, 1, 2, 5, 4, 6, 8}: we start by sorting the list to get {1, 2, 3, 4, 5, 6, 8}, so the median is 4.

# Mode

Given a list of numbers (data set), the value that appears most often in this list.

## Examples

List 1={1, 2, 4, 4, 4, 6, 7}: the mode is 4 since 4 appears three times.

List 2={1, 2, 2, 3, 5, 5, 7, 9}: there are two modes (bimodal), since both 2 and 5 appear twice.

List 3={3, 1, 2, 5, 4, 6, 8}: there is no mode since no value appears more than once.

# Mean-Median-Mode

List	Mean	Median	Mode
{1, 2, 4, 4, 4, 6, 7}	4	4	4
{1, 2, 2, 3, 5, 5, 7, 9}	34/8	4	2,5
{3, 1, 2, 5, 4, 6, 8}	29/7	4	none

Advantage of the median over the mean: it is **less skewed** by a small proportion of extremely large/small values, thus it may give a better idea of what is a typical value.

Example: household income or assets.

Median personal income by educational attainment (2017)<sup>[2][13]</sup>

Measure	Some high school	High school graduate	Some college	Associate's degree	Bachelor's degree or higher	Bachelor's degree	Master's degree	Professional degree	Doctorate degree
Persons, age 25+, employed full-time	\$30,598	\$38,102	\$43,377	\$47,401	\$71,221	\$64,074	\$77,285	\$117,679	\$101,307

"As of mid-March 2017, Bezos was the world's third-richest man, according to Forbes, with a net worth of around \$73 billion, behind only Bill Gates and Warren Buffett."

Sources: wikipedia, <http://money.com/money/4738275/jeff-bezos-wealth-amazon-highest-paid-worker/>

# Quartiles (1st, 2nd, 3rd)

1st quartile ( $Q_1$ ) = the middle number between the smallest number and the median.

2nd quartile ( $Q_2$ ) = the median of the data.

3rd quartile ( $Q_3$ ) = the middle value between the median and the largest number.

If the list size is an odd number, you may either remove/use twice the median when computing the 1st and 3rd quartile.

## Examples

List 1={1, 2, 4, 4, 4, 6, 7}: the median (=2nd quartile) is 4 since {1, 2, 4, **4**, 4, 6, 7}.

If we consider {1, 2, 4}, {4, 6, 7}, the 1st quartile is 2, the 3rd quartile is 6.

If we consider {1, 2, 4, **4**}, {**4**, 4, 6, 7}, the 1st quartile is  $(2+4)/2=3$ , the 3rd quartile is  $(4+6)/2=5$ .

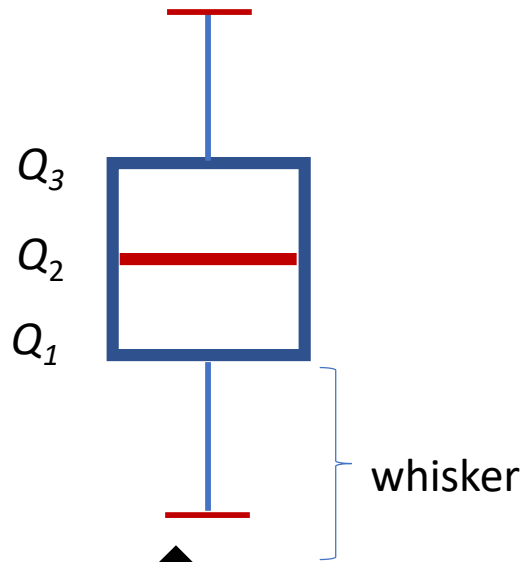
List 2={1, 2, 2, 3, 5, 5, 7, 9}: there are two middle values, 3 and 5, so the median (=2nd quartile) is given by  $(3+5)/2 = 4$ .

Split the list into {1, 2, 2, 3}, {5, 5, 7, 9}, the 1st quartile is 2, the 3rd quartile is 6.

# Boxplots

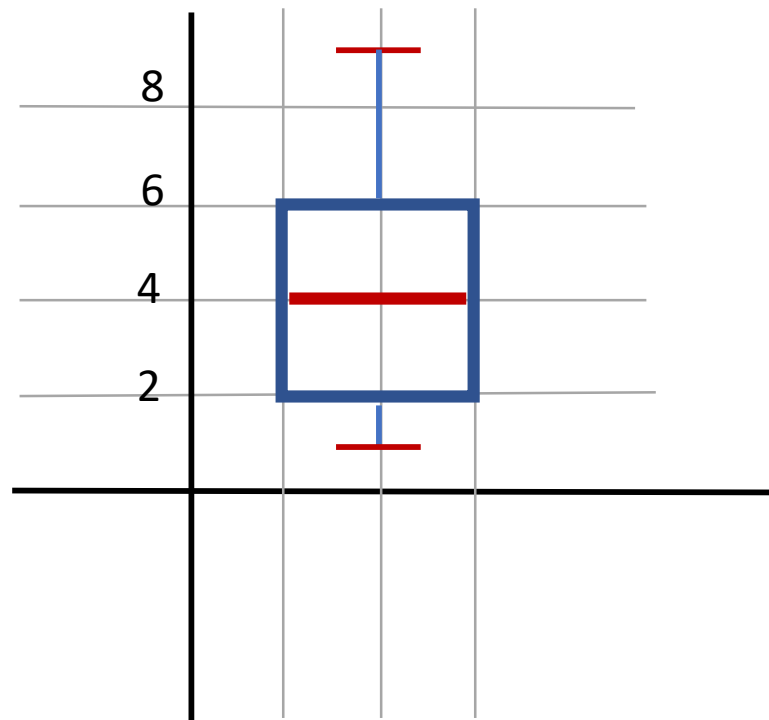
List 2={1, 2, 2, 3, 5, 5, 7, 9}: the median ( $Q_2$ ) is given by  $(3+5)/2 = 4$ .

Split the list into {1, 2, 2, 3}, {5, 5, 7, 9}, the 1st quartile ( $Q_1$ ) is 2, the 3rd quartile ( $Q_3$ ) is 6.



End of whisker

Variation 1: minimum and maximum of the list.



# Standard deviation and Variance

You are given a list of numbers, and you know how to compute the mean.

## Example

List 1={1, 2, 4, 4, 4, 6, 7}: we add up the numbers  $1+2+4+4+4+6+7 = 28$ , there are 7 numbers in the list, so the mean is  $28/7 = 4$ .

There are many lists (of 7 elements) that give the same mean of 4, e.g. {1,1,1,6,6,6,7}.

We would like to know how spread apart numbers are from the mean.

So we look at the differences of each number and the mean: {1-4=-3, 2-4=-2, 4-4=0, 4-4=0, 4-4=0, 6-4=2, 7-4=3}

- What do we do with negative differences that cancel out positive ones? (e.g. 1-4 and 7-4)

What about ignoring the sign (absolute value): { $|1-4|=3$ ,  $|2-4|=2$ ,  $|4-4|=0$ ,  $|4-4|=0$ ,  $|4-4|=0$ ,  $|6-4|=2$ ,  $|7-4|=3$ }

- The values 2 and 6 contribute to  $|2-4|=2$  and  $|6-4|=2$ , thus 4. If we had the values 1 and 5 instead of 2 and 6, we would still get  $|1-4|=3$  and  $|5-4|=1$ , also for a sum of 4. Are they equally spread out?

Consider the square differences: {(1-4)<sup>2</sup>=9, (2-4)<sup>2</sup>=4, (4-4)<sup>2</sup>=0, (4-4)<sup>2</sup>=0, (4-4)<sup>2</sup>=0, (6-4)<sup>2</sup>=2, (7-4)<sup>2</sup>=3}

- Now  $(2-4)^2+(6-4)^2 = 8$ , while  $(1-4)^2+(5-4)^2 = 10$ .



# Variance

Given a list of numbers and its mean, the quantity obtained by first summing the square differences of each number and the mean, and then dividing by the size of the list.

## Example

List 1={1, 2, 4, 4, 4, 6, 7}, mean = 4: the square differences are  $\{(1-4)^2=9, (2-4)^2=4, (4-4)^2=0, (4-4)^2=0, (4-4)^2=0, (6-4)^2=2, (7-4)^2=3\}$ , we sum them to get  $9+4+2+3$ , and finally we divide by 7 to find that the variance is:  $(9+4+2+3)/7 = 18/7$

# Standard deviation

Given a list of numbers and its variance, the quantity obtained by computing the square root of the variance.

## Example

List 1={1, 2, 4, 4, 4, 6, 7}, variance =  $18/7$ : its standard deviation is  $\sqrt{18/7}$ .

Both are measures of spread of the data (how far the data is spread with respect to the mean).

Variance versus standard deviation: The standard deviation remains in the same scale as the data.

# Case study: Bitcoin statistics

Daily return = the difference between each day's closing and opening price, as a percentage of the opening price = return of buying on the open and selling on the close.

## Bitcoin daily return descriptive statistics – April 28, 2013 to May 21, 2018

Mean	0.31%
Median	0.20%
Mode	0.00%
Range	64.61%
Standard Deviation	4.49%
Count	1,850
Min; Max	-22.93%; 41.68%

# Case study: Bitcoin statistics

Median:

it is positive (0.20%),

→ more days of positive return than negative.

smaller than the mean,

→ daily expected returns are positively skewed towards particularly positive values.

## Bitcoin daily return descriptive statistics – April 28, 2013 to May 21, 2018

Mean	0.31%
Median	0.20%
Mode	0.00%
Range	64.61%
Standard Deviation	4.49%
Count	1,850
Min; Max	-22.93%; 41.68%

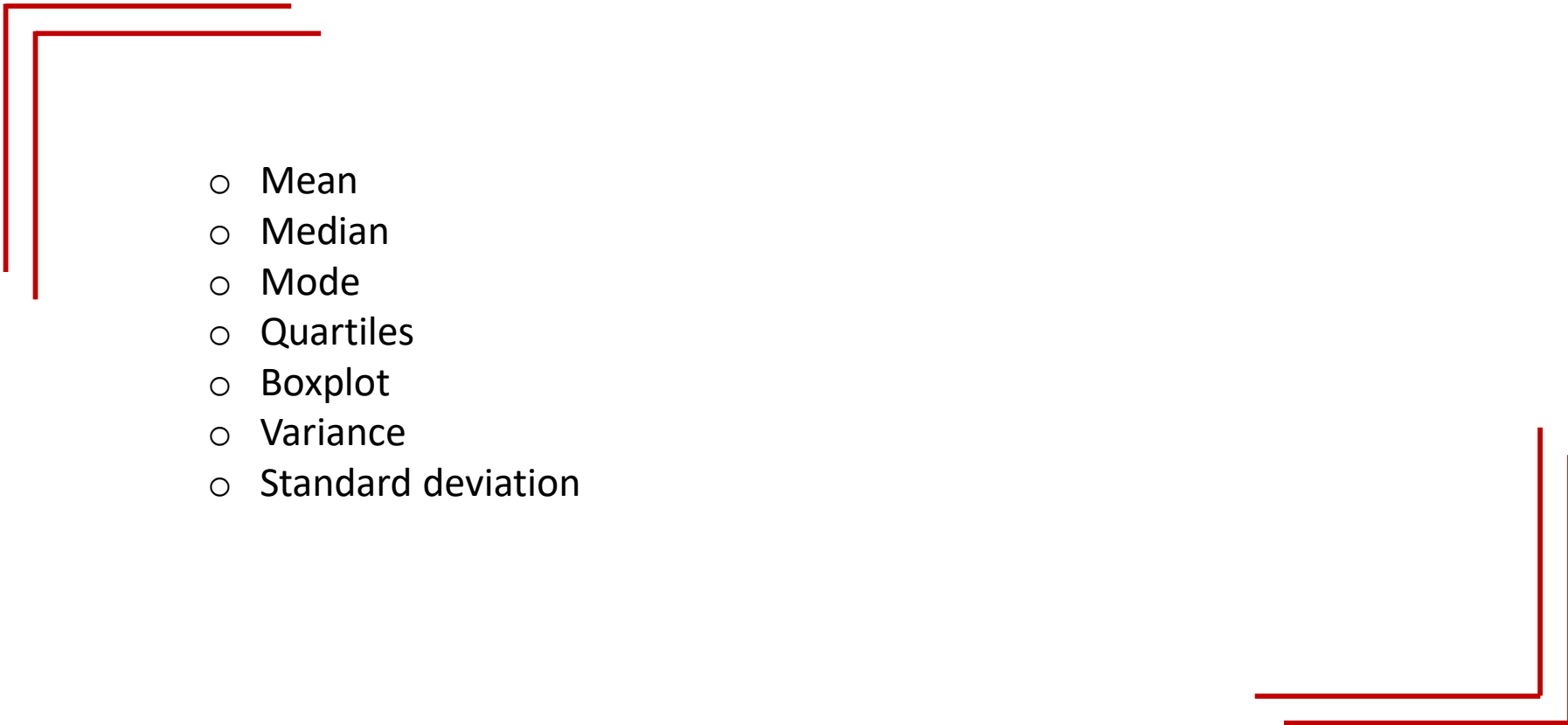
# Case study: Bitcoin statistics

Range = difference between the highest and lowest expected daily return.  
Its lowest value (Min) was -22.93%,  
its highest value (Max) was 41.68%.

## Bitcoin daily return descriptive statistics – April 28, 2013 to May 21, 2018

Mean	0.31%
Median	0.20%
Mode	0.00%
Range	64.61%
Standard Deviation	4.49%
Count	1,850
Min; Max	-22.93%; 41.68%

# Statistical tools for descriptive analytics: summary

- 
- Mean
  - Median
  - Mode
  - Quartiles
  - Boxplot
  - Variance
  - Standard deviation

# Questions (I)

Which concept of descriptive statistics do we need:

1. To compute the average of a data set?
2. To compute the spread of values within a data set?
3. To find the most frequent data within a data set?
4. To compute the middle of the data?

Which statistics are displayed in boxplots?

List one visualization technique seen in this course, and at least one that you might have encountered.

# Statistical Data Analysis/Analytics

- Data analytics workflow

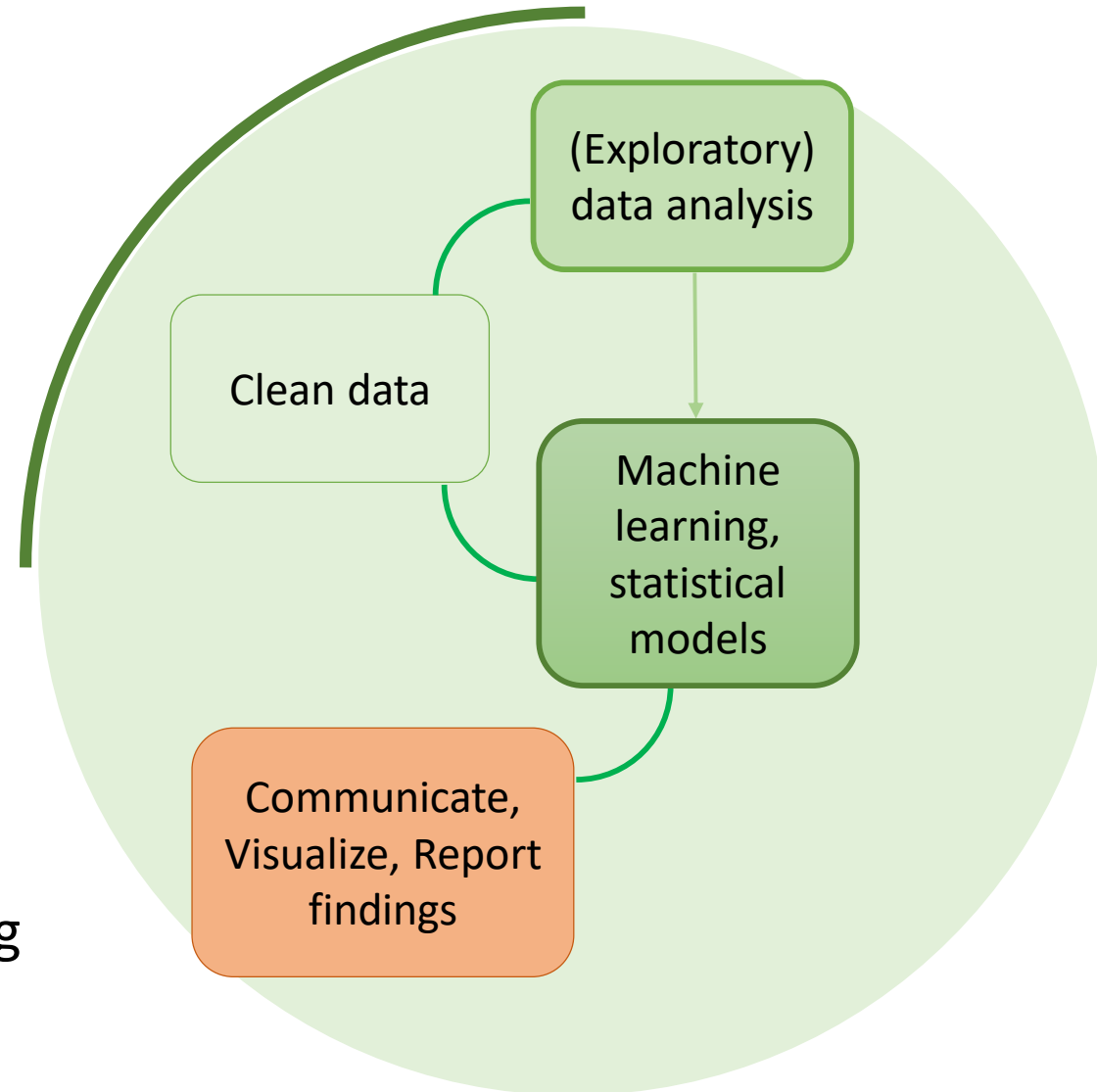
## Descriptive Analytics

- Mean, median, mode
- Quartiles and boxplots
- Variance and standard deviation

## Predictive Analytics

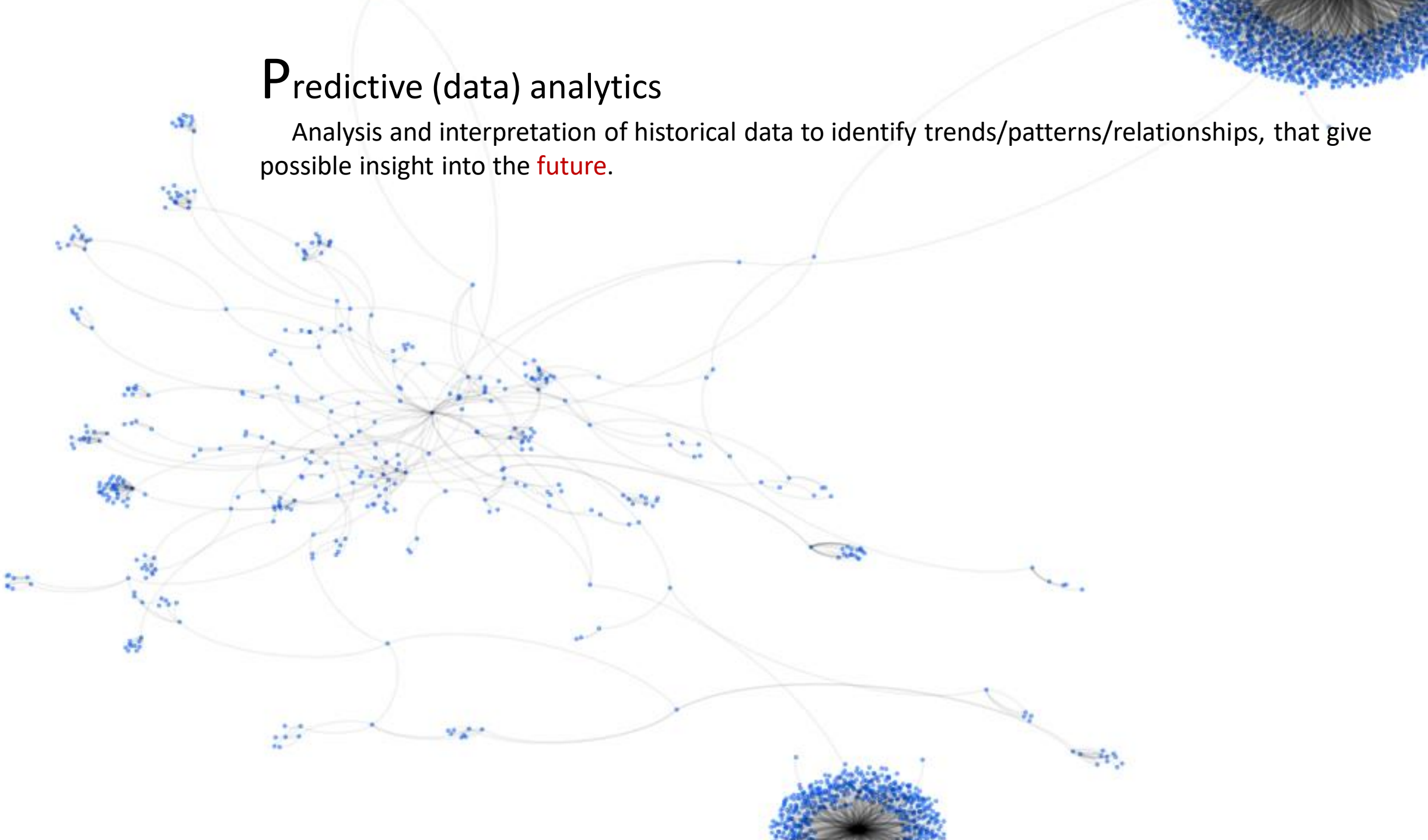
- Decision trees
- Linear discriminant

## Prescriptive Analytics and Machine Learning



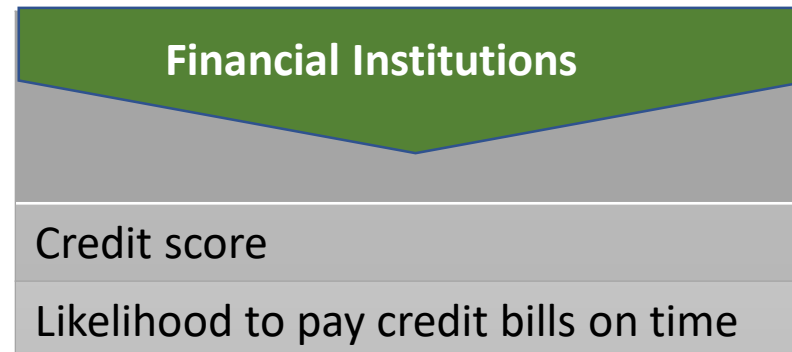
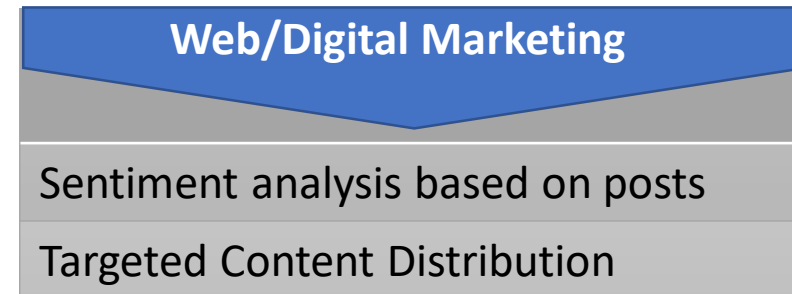
# Predictive (data) analytics

Analysis and interpretation of historical data to identify trends/patterns/relationships, that give possible insight into the **future**.





# Examples



# More examples

**Fault prediction** for a given piece of equipment helps manufacturers/engineers plan repairing, avoiding delays and further complications.

---

**Risk management** uses analytics to predict potential problems, calculate the likelihood of the problematic issues, and identify backup solutions and contingency plans.

---

Applications of predictive analytics **to health** include predicting epidemics or predicting the likelihood of a patient to end up in intensive care due to changes in e.g. environmental conditions.

**Demand forecasting** gives a better control of the inventory better and reduces the storage of products.

---

**Price optimization** looks for the best price for both buyers and sellers and can increase the seller's profit.

---

Predictive analytics improves **weather forecasting**.

---

Insurance firms can lessen losses within risk tolerances, thanks to predictive analytics for **risk assessment**.

# Some numbers

The current Airbus A350 model has a total of close to **6,000** sensors across the entire plane and generates **2.5 Tb** of data **per day** (*source: <https://www.datasciencecentral.com/profiles/blogs/that-s-data-science-airbus-puts-10-000-sensors-in-every-single> )*

Currently, each vehicle has an average of **60-100** sensors on board. Because cars are rapidly getting “smarter” the number of sensors is projected to reach as many as 200 sensors per car. These numbers translate to approximately **22 billion** sensors used in the automotive industry per year by 2020. (*source: <http://www.automotivesensors2017.com/>*)

Number of netflix users: **151.59 millions** in 2019  
(*source: [https://expandeddrablings.com/index.php/netflix\\_statistics-facts/](https://expandeddrablings.com/index.php/netflix_statistics-facts/)*)

Number of facebook users: **1.63 billion daily** active users on average, **2.45 billion monthly** active users as of September 30, 2019.  
(*source: <https://newsroom.fb.com>*)

# Predictive (data) analytics: what for?



For setting realistic goals for your business, for effective planning and restraining expectations.



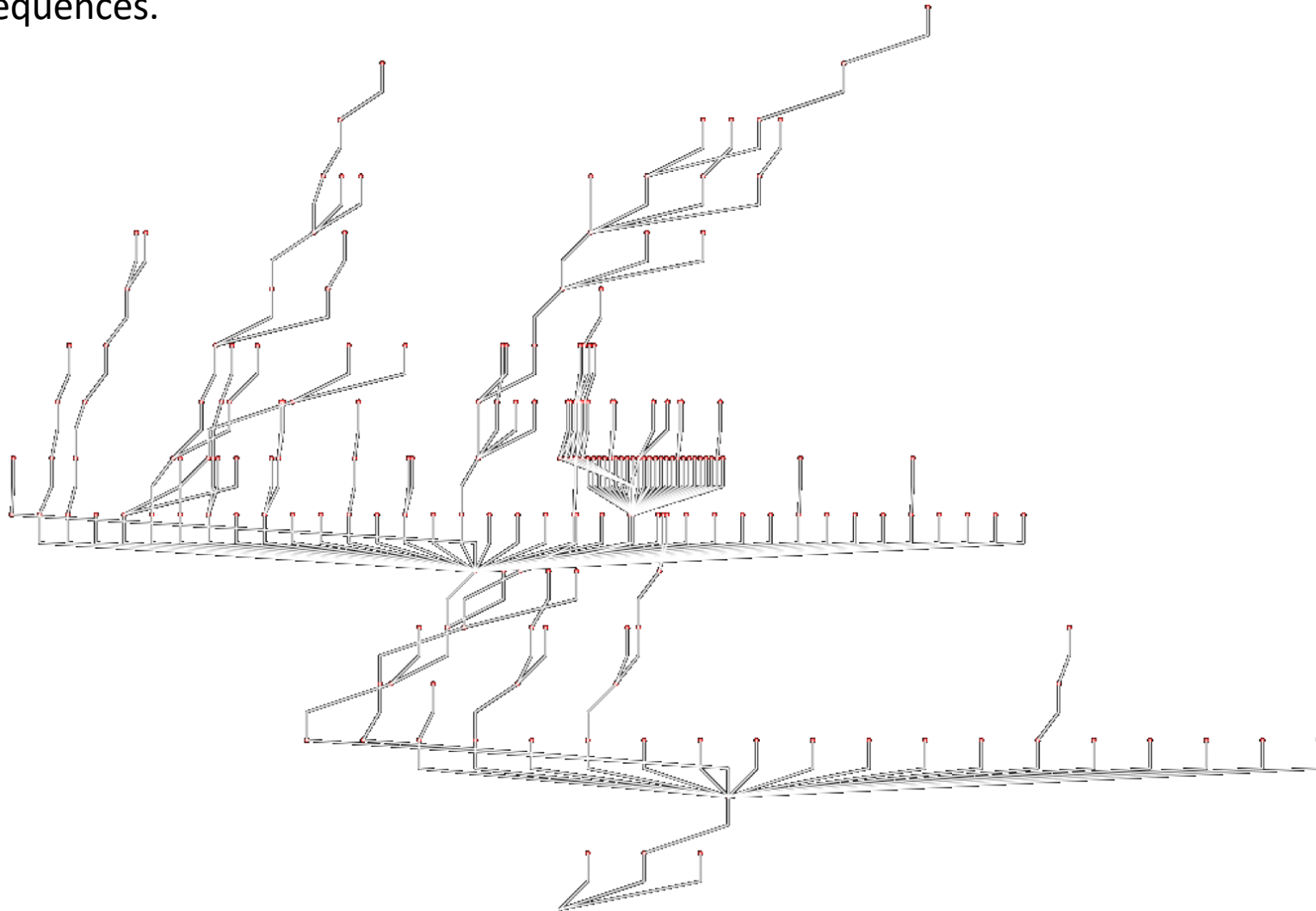
For designing an effective business strategy.



For detecting problems before they even occur (e.g. maintenance of machines/vehicles).

# Decision tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences.



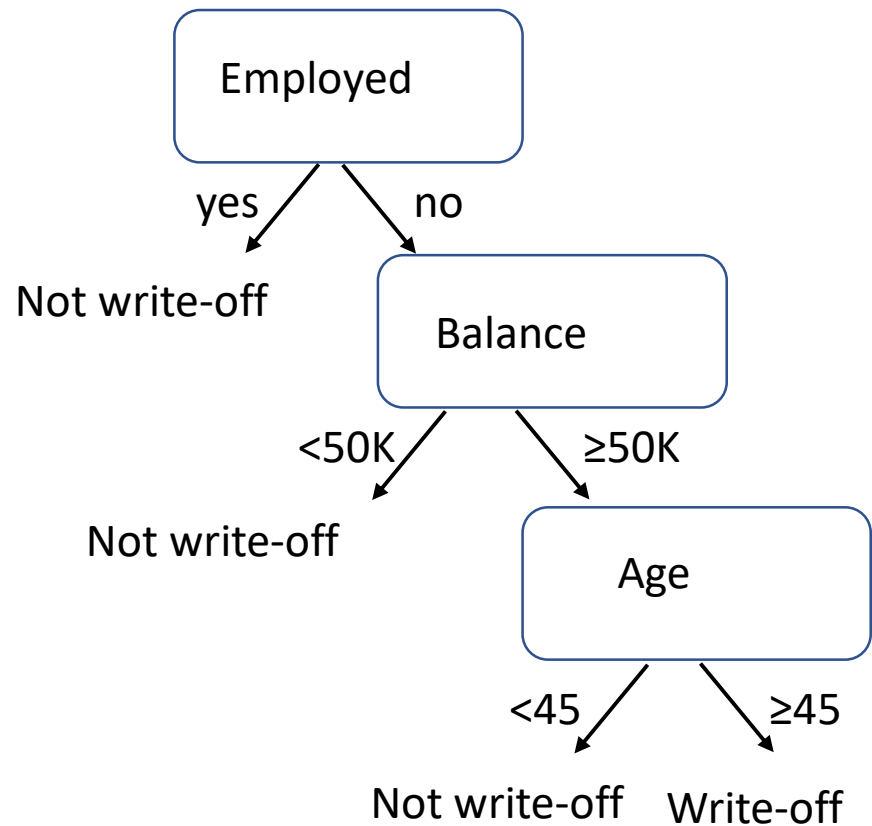
# Decision problem

- Context: credit
- Data available: historical customer data
- Question: how likely is a write-off for a given new customer?

<b>Name</b>	<b>Balance</b>	<b>Age</b>	<b>Employed</b>	<b>Write-off</b>
Mike	\$200'000	42	no	yes
Mary	\$35'000	33	yes	no
Claudio	\$115'000	40	no	no
Robert	\$29'000	23	yes	yes
Dora	\$72'000	31	no	no

(Source: Data Science for Business by Foster Provost, Tom Fawcett)

# Decision tree: attempt



Name	Balance	Age	Employed	Write-off	
Mike	\$200'000	42	no	yes	✗
Mary	\$35'000	33	yes	no	✓
Claudio	\$115'000	40	no	no	✓
Robert	\$29'000	23	yes	yes	✗
Dora	\$72'000	31	no	no	✓

Classification (or decision) tree:  
there are several variations of  
decision trees.

# Decision tree: purity

Question: which attributes to choose (balance, age, employed) to best distinguish write-offs from non-write-offs?

We would like to have groups as *pure* as possible.

The most common splitting criterion is called *information gain*. It relies on a purity measure called *entropy* (Shannon, 1948).

$p_i$  = relative percentage of property  $i$

entropy =  $-p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots$



# Examples

$p_i$  = relative percentage of property i

entropy =  $-p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots$

Name	Balance	Age	Employed	Write-off
Mike	\$200'000	42	no	yes
Mary	\$35'000	33	yes	no
Claudio	\$115'000	40	no	no
Robert	\$29'000	23	yes	yes
Dora	\$72'000	31	no	no

Write-off / non-write-off

2/5 write-off, 3/5 non-write-off

$-(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = \mathbf{0.97095}$

Employed / non-employed

2/5 employed, 3/5 not employed

$-(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = \mathbf{0.97095}$

Age <45 / Age ≥45

5/5 aged < 45, 0/5 aged ≥45

$-1 \log_2(1) = \mathbf{0}$

Age < 30, Age 30-40, Age > 40

1/5 aged < 30, 3/5 aged 30-40,  
1/5 aged >40

$-(1/5)\log_2(1/5) - (3/5)\log_2(3/5) - (1/5)\log_2(1/5) = \mathbf{1.3710}$

# Intuition

Entropy is a *purity measure*: if the attribute gives *pure* groups, the entropy is 0, if the attribute gives *impure* groups, the entropy is large (it is at most  $\log_2(\text{number of elements in the group})$ ).

## Example

Attribute: Age  $<45$  / Age  $\geq 45$ , entropy = 0

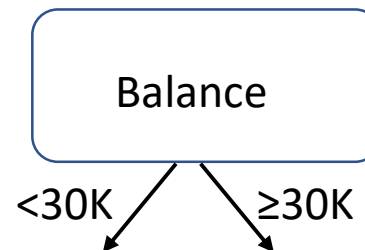
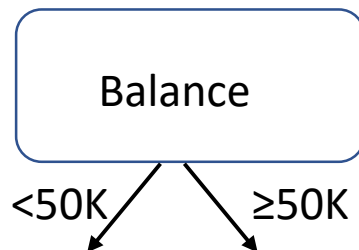
Attribute: Write-off / non-write-off, entropy = 0.97095 (it is at most 1)

Using entropy to measure *purity*, the concept of *information gain* (IG) is used to measure how informative an attribute is with respect to our target, namely how much gain in information the attributes gives us with respect to our target if we split groups (and thus branch out) accordingly:

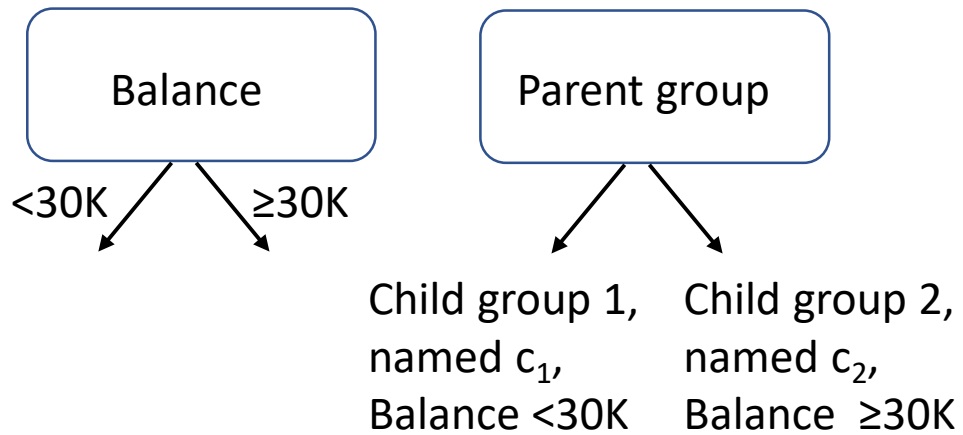
## Example

Target: Write-off /non-write-off

Which splitting gives the most information about write-off?



# Information gain



Say the attribute we split on (Balance) has  $k$  values ( $k=2$ ). The original group is called the parent group. The result of splitting on the attribute values gives  $k$  children groups. How much information has the splitting of this attribute provided?

That depends on how much *purer* the children are than the parent, namely how much would the knowledge of the attribute (Balance) increase our knowledge of the target (write-off)?

$$IG(\text{parent, children}) = \text{entropy}(\text{parent}) - [p(c_1)\text{entropy}(c_1) + p(c_2)\text{entropy}(c_2) \dots]$$

Proportion of members of  $c_1$  among the members of the parent group

Splitting off a group with one member gives a set which is pure, it may not be as good as splitting the parent set into two large relatively pure subsets.

# Examples

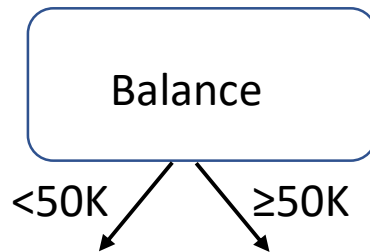
Target: Write-off / non-write-off,  
entropy = 0.97095 (it is at most 1)

(very impure)

$$IG(\text{write-off, children}) = 0.97095 - [p(c_1)\text{entropy}(c_1) + p(c_2)\text{entropy}(c_2)]$$

Name	Balance	Age	Employed	Write-off
Mike	\$200'000	42	no	yes
Mary	\$35'000	33	yes	no
Claudio	\$115'000	40	no	no
Robert	\$29'000	23	yes	yes
Dora	\$72'000	31	no	no

Example



$c_1 = \{\text{Mary, Robert}\}, c_2 = \{\text{Mike, Claudio, Dora}\}$

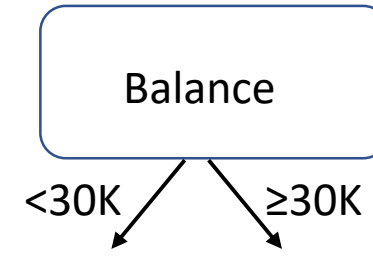
$p(c_1) = 2/5, p(c_2) = 3/5$

$\text{entropy}(c_1) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/2) = 1$

$\text{entropy}(c_2) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) = 0.91830$

$IG(\text{write-off, balance}) = 0.97095 - (2/5) - (3/5)0.91830$   
 $= 0.019970$

Example



$c_1 = \{\text{Robert}\}, c_2 = \{\text{Mike, Claudio, Dora, Mary}\}$

$p(c_1) = 1/5, p(c_2) = 4/5$

$\text{entropy}(c_1) = -(1)\log_2(1) = 0$

$\text{entropy}(c_2) = -(1/4)\log_2(1/4) - (3/4)\log_2(3/4) = 0.81128$

$IG(\text{write-off, balance}) = 0.97095 - (4/5)0.81128$   
 $= 0.32193$

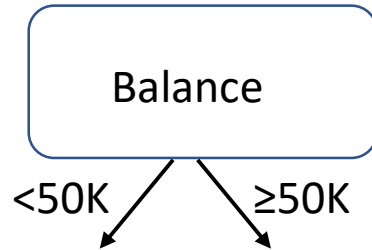
# Examples (continued)

Target: Write-off / non-write-off,  
entropy = 0.97095 (it is at most 1)

(very impure)

$$IG(\text{write-off, children}) = 0.97095 - [p(c_1)\text{entropy}(c_1) + p(c_2)\text{entropy}(c_2)]$$

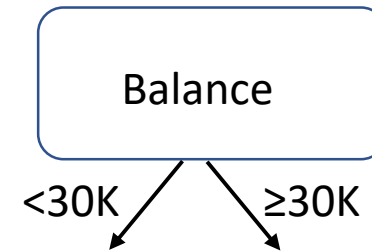
Example



$$IG(\text{write-off, balance}) = 0.97095 - (2/5) - (3/5)0.91830 = 0.019970$$

Information gain measures the change in entropy due to any amount of new information being added.

Example



$$IG(\text{write-off, balance}) = 0.97095 - (4/5)0.81128 = 0.32193$$

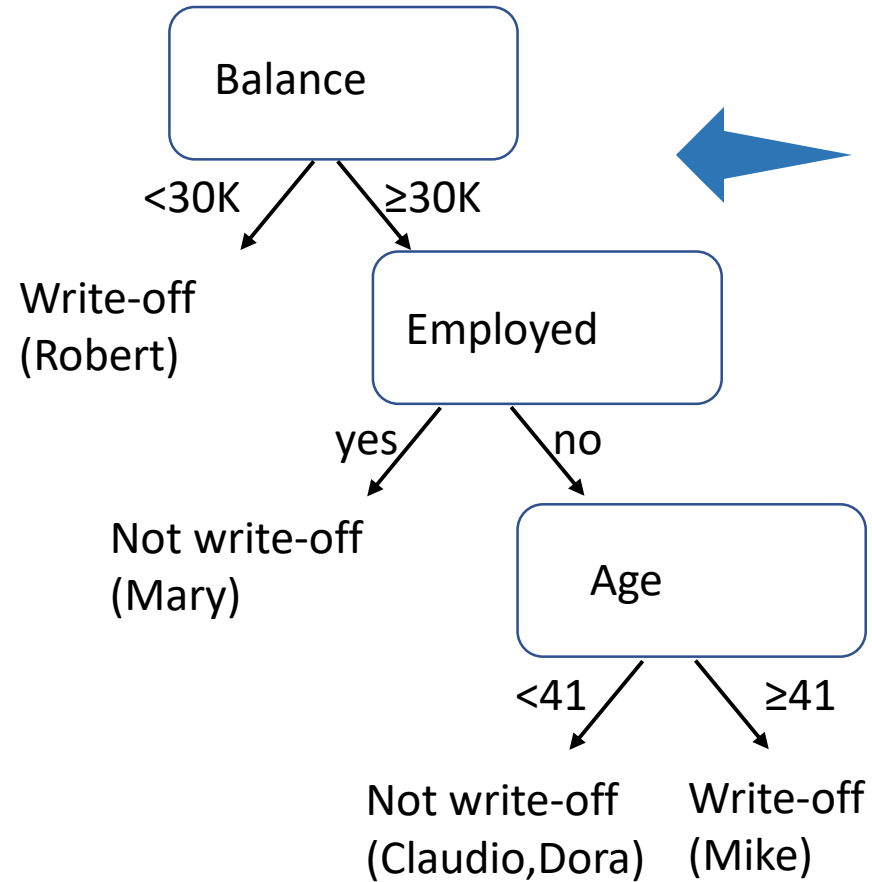
Children groups are purer with this attribute, their entropy is less, therefore the information gain is more.

# Decision tree: iterations

Tree induction (divide-and-conquer approach):

- start with the whole dataset,
- try to create the “purest” subgroups possible using the attributes, obtain the “split” that yields the largest information gain.
- The obtained subgroups are smaller versions of the initial problem.
- Take each data subgroup and recursively apply attribute selection to find the best split.
- This algorithm will stop eventually (all groups contain one element or all attributes have been exhausted).
- Best stopping criterion? (if not all groups are pure)

# Examples (continued)

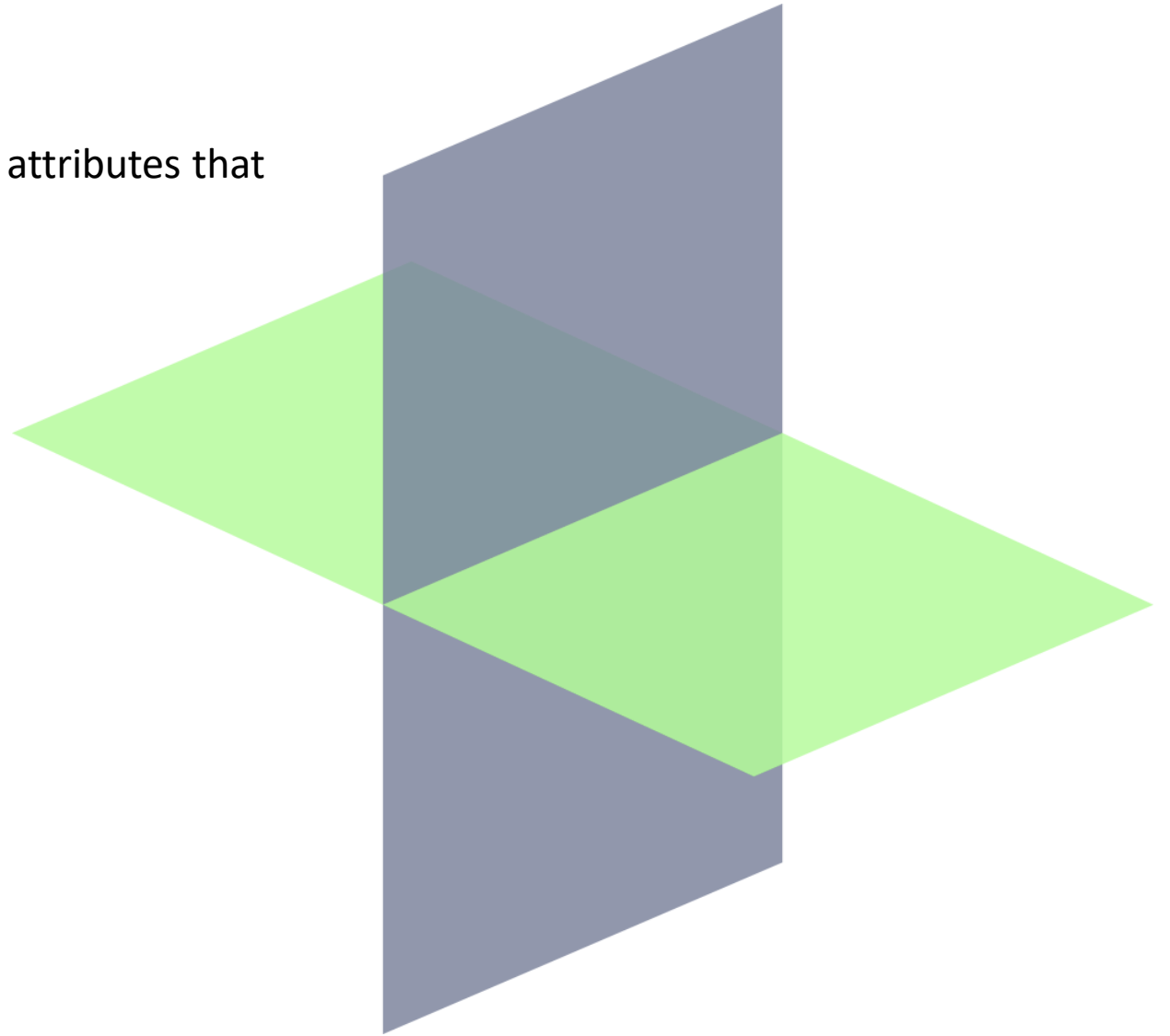


Name	Balance	Age	Employed	Write-off
Mike	\$200'000	42	no	yes
Mary	\$35'000	33	yes	no
Claudio	\$115'000	40	no	no
Robert	\$29'000	23	yes	yes
Dora	\$72'000	31	no	no

Name	Balance	Age	Employed	Write-off
Mike	\$200'000	42	no	yes
Claudio	\$115'000	40	no	no
Dora	\$72'000	31	no	no

# Linear discriminant analysis

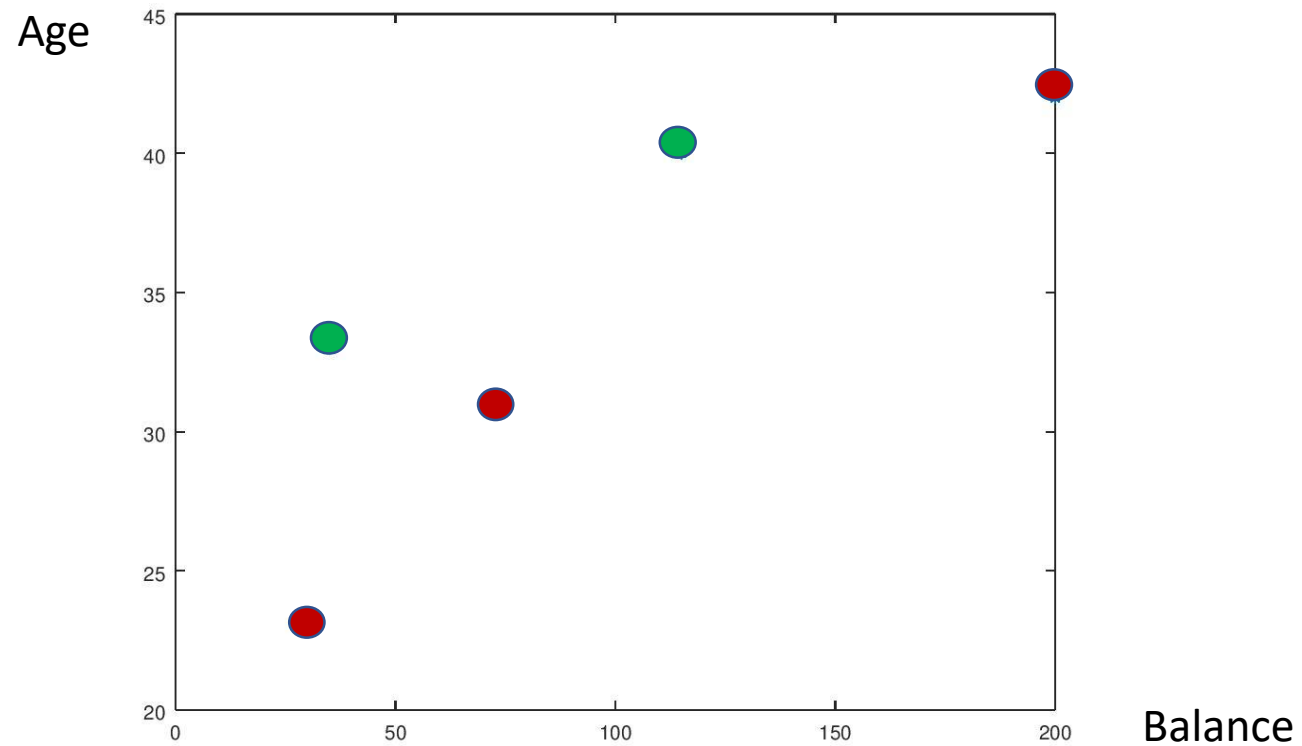
A method used to find linear combinations of attributes that characterizes two (or more) classes of objects.



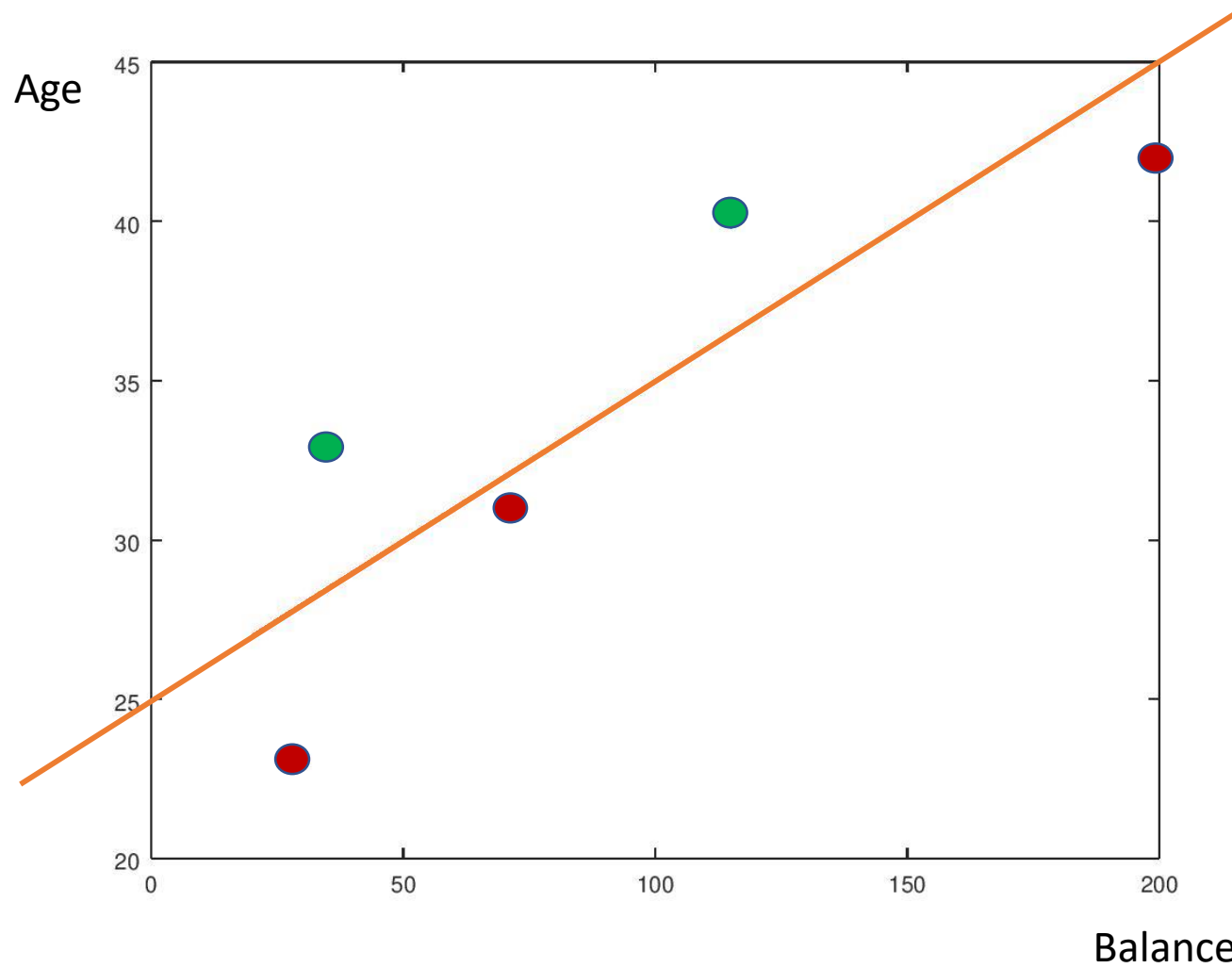


# Example: geometry

Name	Balance	Age	Employed	Write-off
Mike	\$200'000	42	no	yes ●
Mary	\$35'000	33	yes	no ●
Claudio	\$115'000	40	no	no ●
Robert	\$29'000	23	yes	yes ●
Sally	\$73'000	31	no	yes ●



# Example: geometry (continued)



$$Age = (1/10)(Balance) + 25$$

Balance = 0, Age = 25  
Balance = 200K, Age = 45

# Linear discriminant

$$\text{Age} - (1/10) \text{Balance} - 25 = 0$$

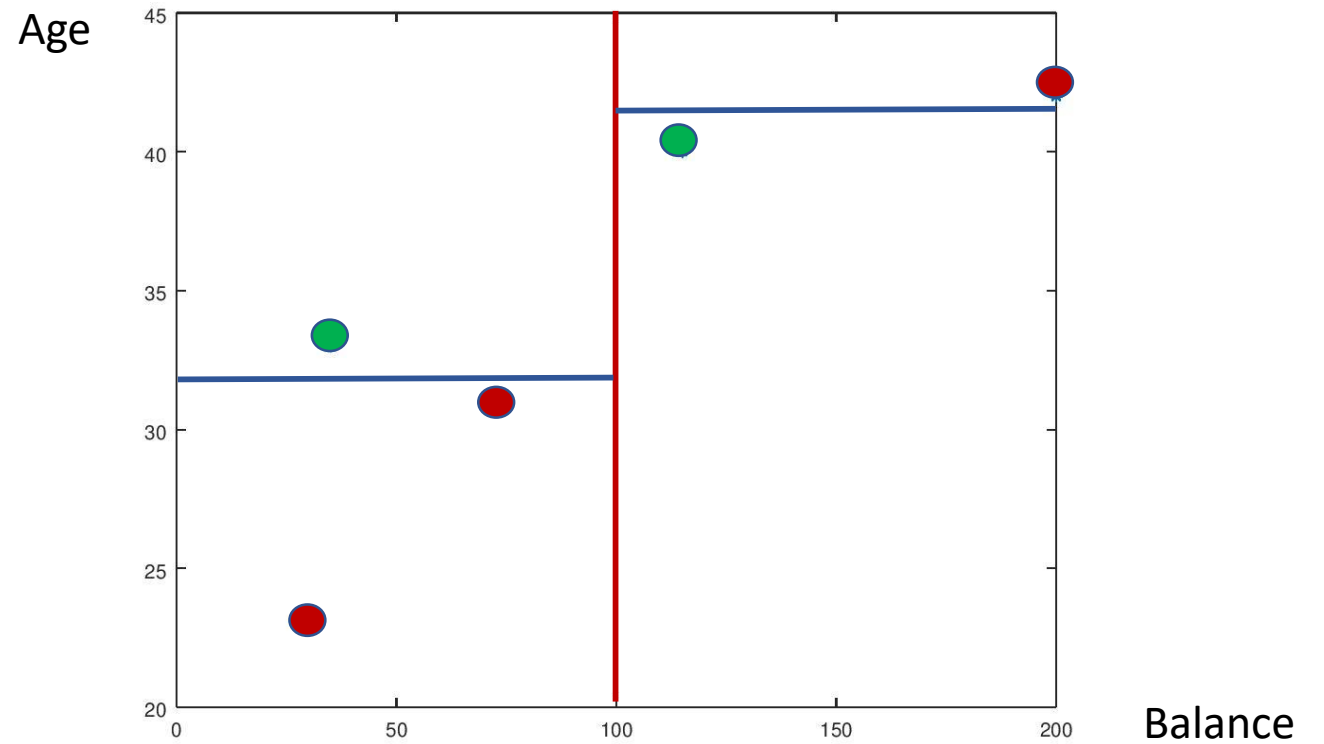
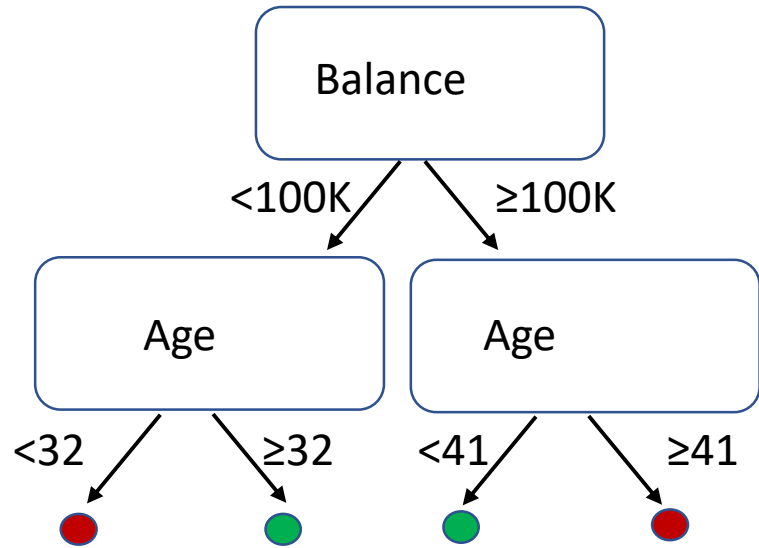
$$\text{Class}(x) = \begin{cases} \bullet & \text{if Age} - (1/10) \text{Balance} - 25 > 0 \\ \bullet & \text{if Age} - (1/10) \text{Balance} - 25 \leq 0 \end{cases}$$

Linear combinations = weighted sum of attributes

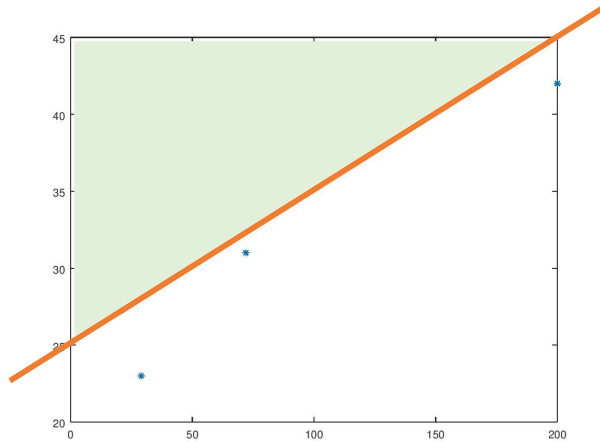
$$(\text{constant 1})(\text{attribute 1}) + (\text{constant 2})(\text{attribute 2}) + \text{constant 3} = 0$$

Linear discriminant: it characterizes (*discriminates*) between the groups (or classes, here two of them). Also called a linear classifier.

# Decision tree: geometry



# Linear discriminant vs Decision tree

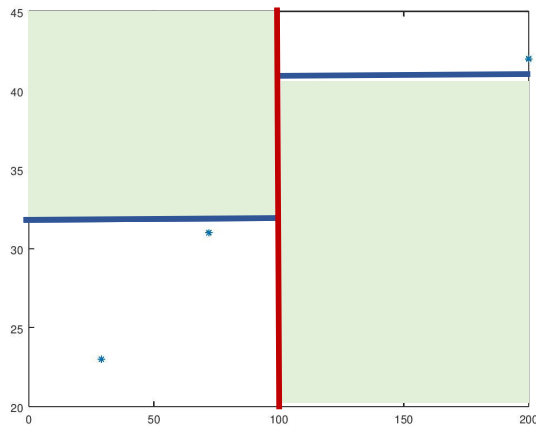


Boundaries:

- Perpendicular to the attribute axes for classification trees
- Any direction/orientation for linear classifiers

This is because

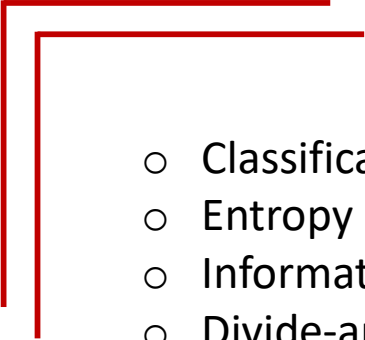
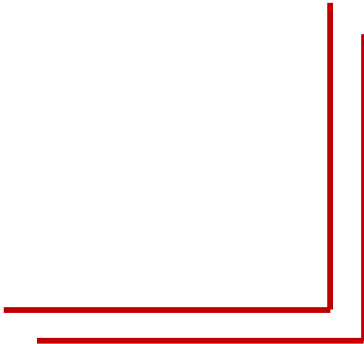
- A single attribute is selected at a time for trees
- Weighted combinations of all attributes are looked at for linear classifiers



Classification trees segments the space recursively into regions that could become arbitrarily small.

A linear classifier places a single boundary (with freedom in the orientation) for the whole space.

# Statistical tools for predictive analytics: summary

- 
- Classification (decision) tree
  - Entropy
  - Information gain
  - Divide-and-conquer algorithm
  - Linear discriminant
  - Weighted sum
- 

## Questions (II)

1. What does entropy measure?
2. What does information gain measure?
3. A set which is pure has (a) high, (b) low entropy.
4. Describe the geometry of decision trees and linear discriminant.

# Statistical Data Analysis/Analytics

- Data analytics workflow

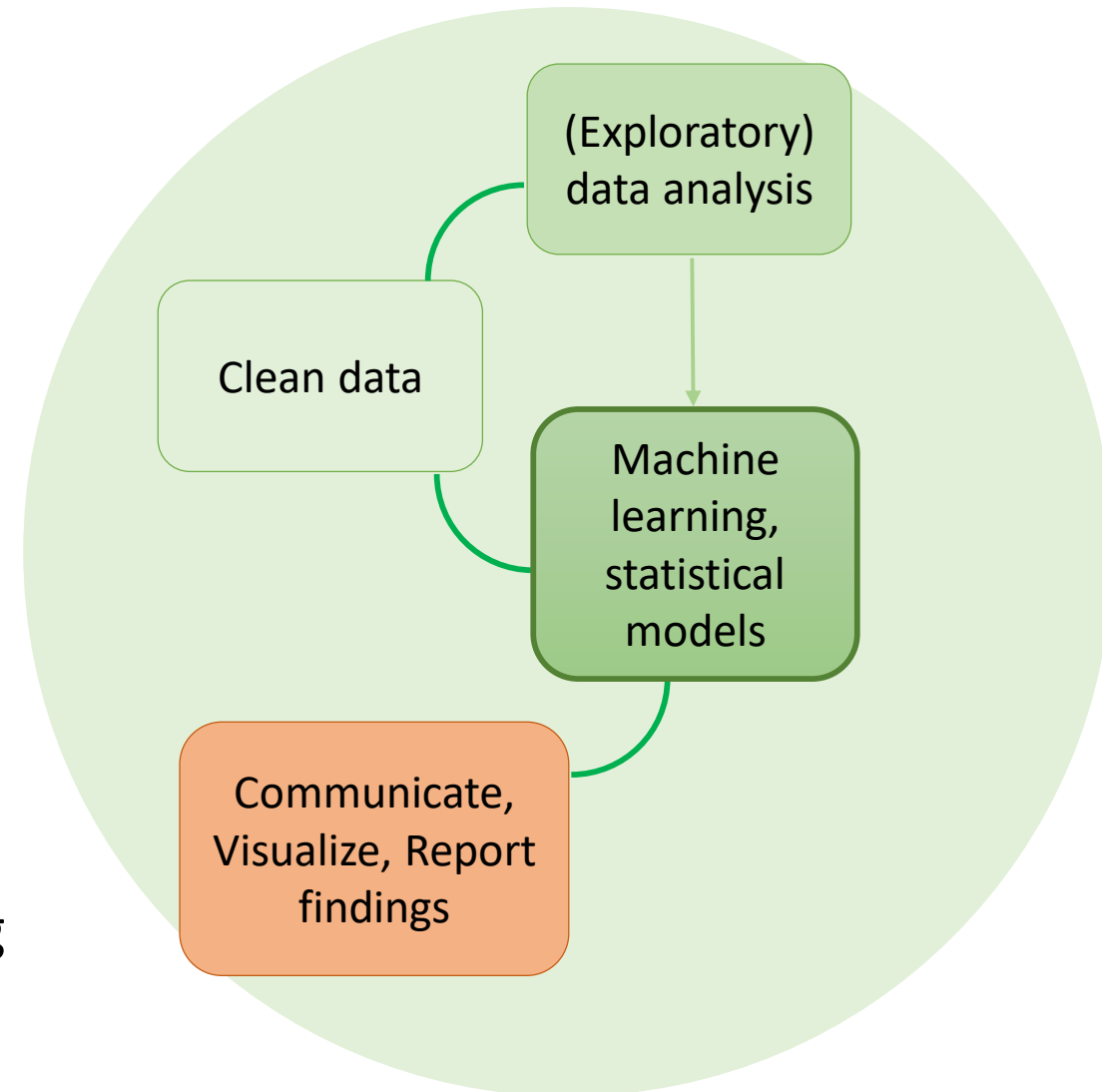
## Descriptive Analytics

- Mean, median, mode
- Quartiles and boxplots
- Variance and standard deviation

## Predictive Analytics

- Decision trees
- Linear discriminant

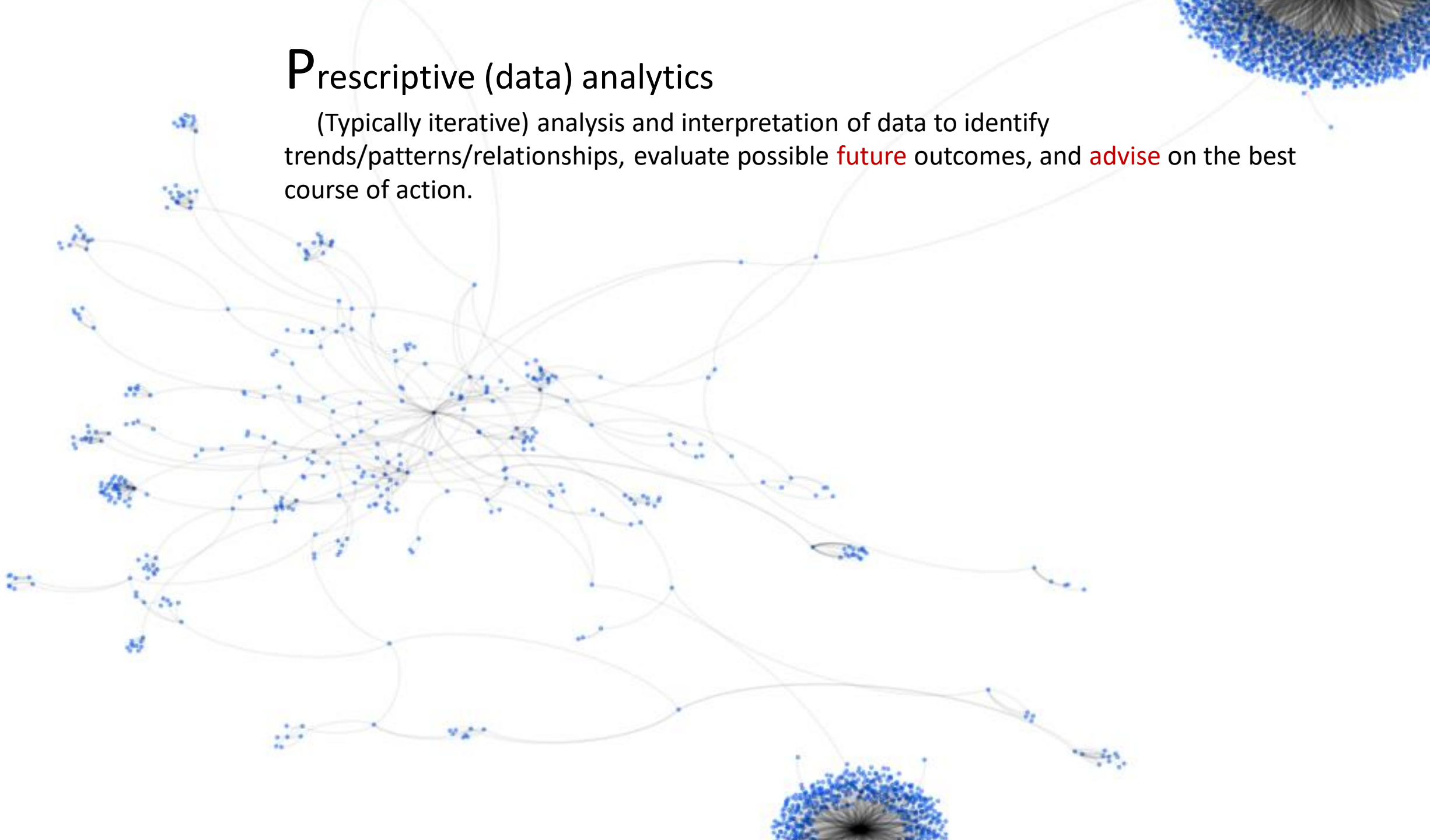
## Prescriptive Analytics and Machine Learning





# Prescriptive (data) analytics

(Typically iterative) analysis and interpretation of data to identify trends/patterns/relationships, evaluate possible future outcomes, and advise on the best course of action.



# Prescriptive (data) analytics: more



Relies on optimization (optimize the outcome), statistical models (simulate the future), descriptive and predictive analytics, is considered complex.



Typically takes in new data to re-predict and re-prescribe, to improve prediction accuracy.



May use different datasets, including historical data, but also for example real-time data feeds.



Uses *machine learning*.

# Examples

Aurora Health Care system saved \$6 million annually by using prescriptive analytics to reduce re-admission rates by 10%.

*Source: <https://www.marketwatch.com/press-release/healthcare-prescriptive-analytics-market-to-experience-significant-growth-during-the-forecast-period-2016-2022-2019-08-20>*

Prescriptive maintenance for aviation with Boeing's AnalytX platform.

(Source: <https://www.boeing.com/features/innovation-quarterly/nov2018/btj-analytics.page>)

# Prescriptive (data) analytics: what for?



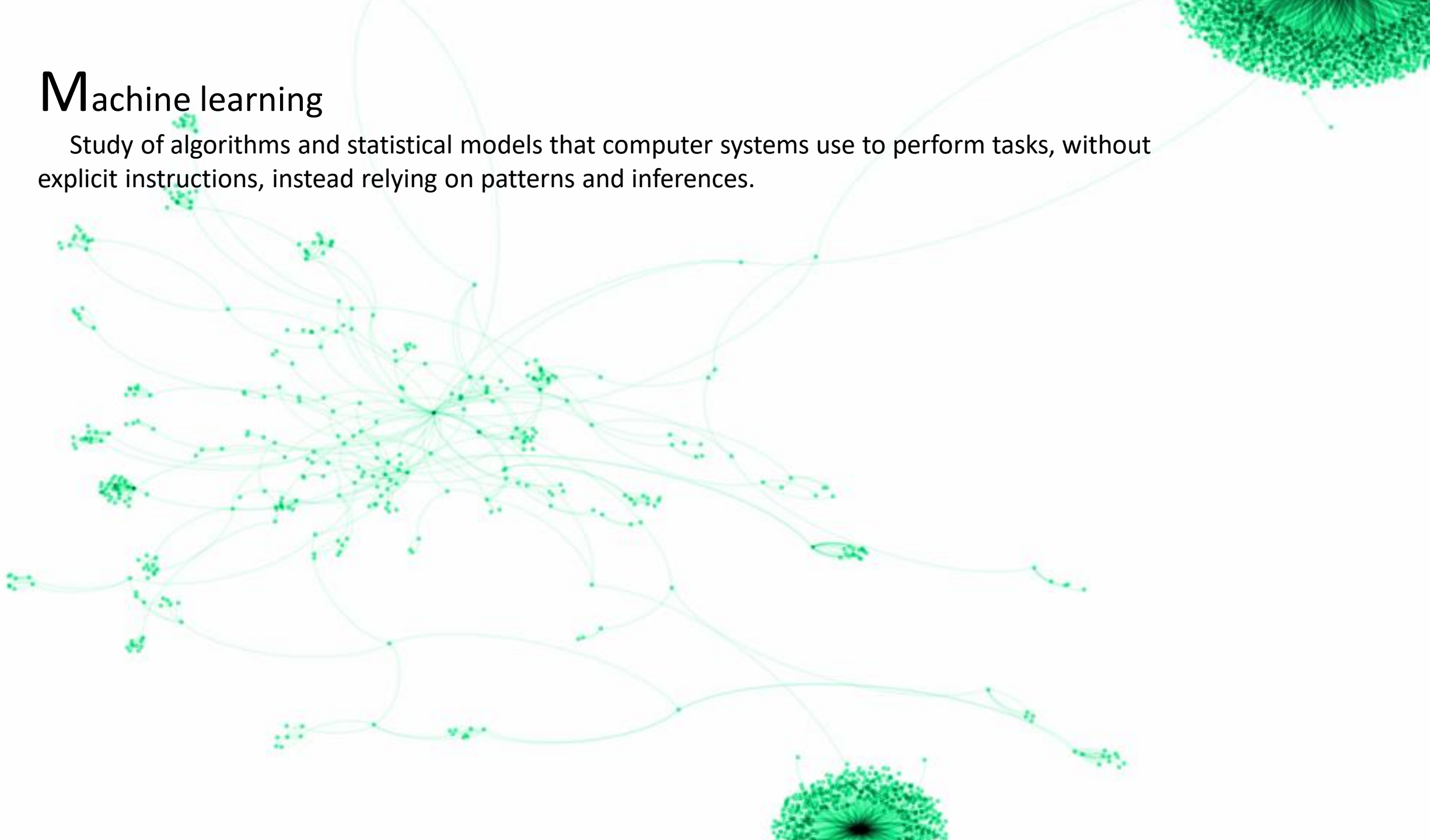
For providing users with advice on what action to take.



For taking advantage of future opportunities, mitigating a future risk, understanding the implications of different decisions.

# Machine learning

Study of algorithms and statistical models that computer systems use to perform tasks, without explicit instructions, instead relying on patterns and inferences.



# Supervised learning versus unsupervised learning

Supervised learning: you are given a ground truth (e.g. historical data), and the goal is to best evaluate the patterns/relationships available, to get the best prediction on new data.

Unsupervised learning: there is no ground truth, the goal is to infer structures within the data.

# Supervised learning

“ Supervised learning: you are given a ground truth (e.g. historical data), and the goal is to best evaluate the patterns/relationships available, to get the best prediction on new data. ”

Classification: predicts in which group to classify a new data point (discrete)

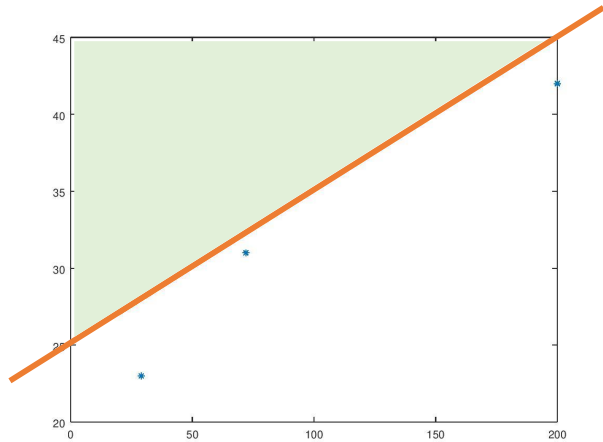
Regression: predicts a value attached to a new data point (continuous)

## Example

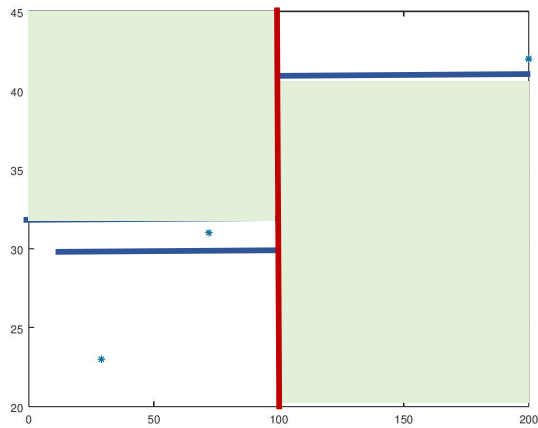
Given a real estate dataset of houses:

- At which price to sell? → Regression
- Is a house costly, affordable, cheap? → Classification

# Supervised learning: classification



Yes, decision trees and linear discriminant analysis can indeed be considered "machine learning" algorithms for supervised learning.

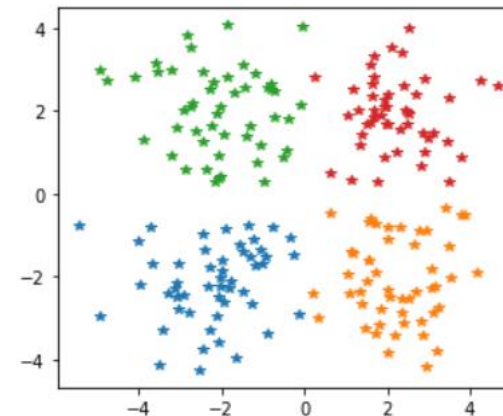
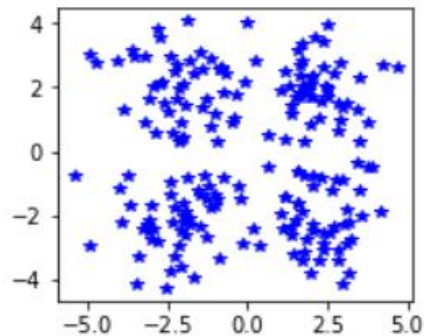




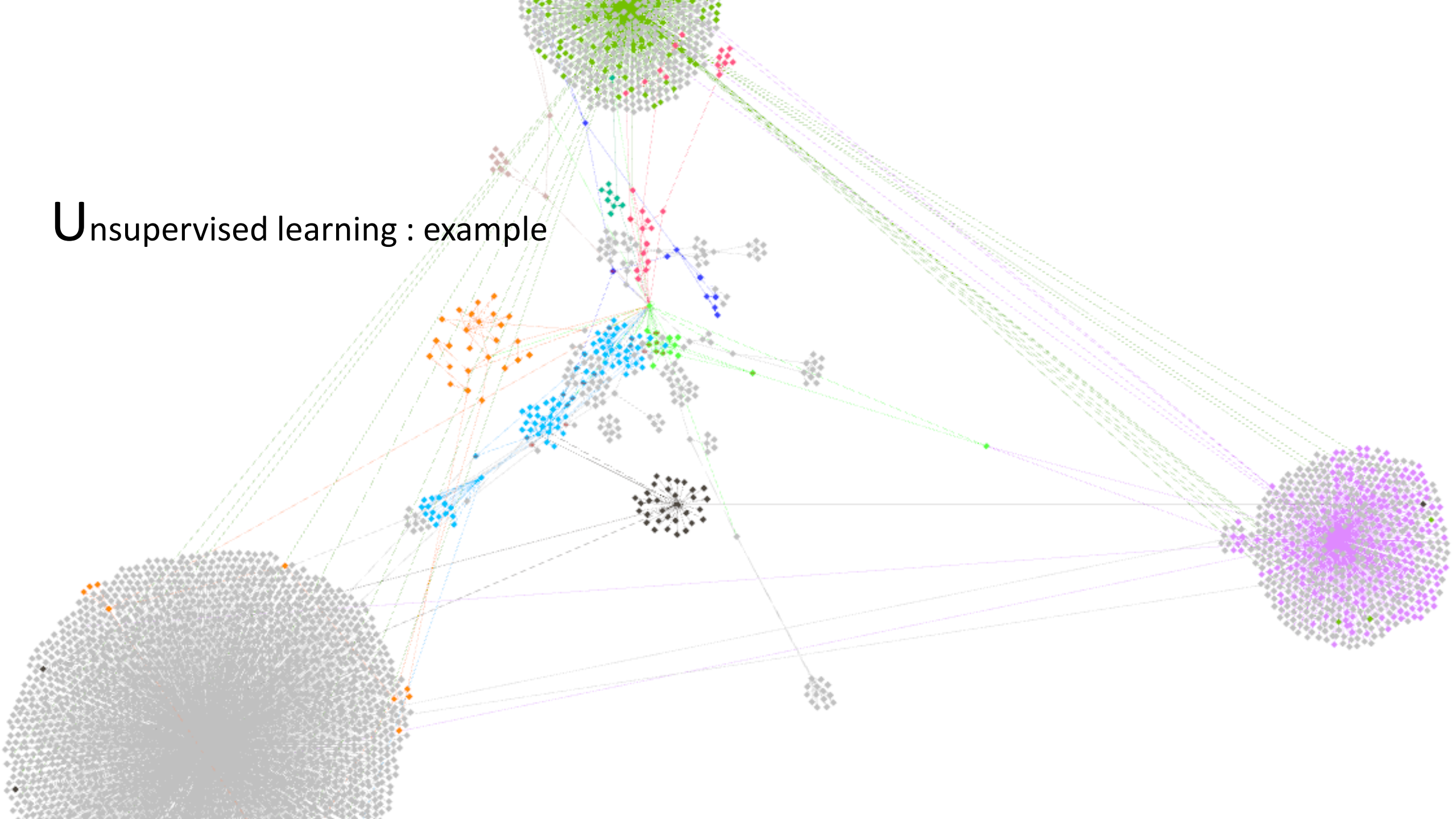
# Unsupervised learning

“ Unsupervised learning: there is no ground truth, the goal is to infer structures within the data. ”

Clustering: group together items/data points into clusters such that points in a cluster are more *similar* to each other than to those in other clusters?



Unsupervised learning : example



# Questions (III)

1. What is the main difference between supervised and unsupervised learning?
2. Classification is considered a task of (a) supervised learning, (b) unsupervised learning.
3. Clustering is considered a task of (a) supervised learning, (b) unsupervised learning.
4. Given a data base of customers, you need to decide whether a customer will leave or stay. Which learning task is involved?
5. Given a data base of faults in a system, you need to decide the likelihood of a new fault happening. Which learning task is involved?

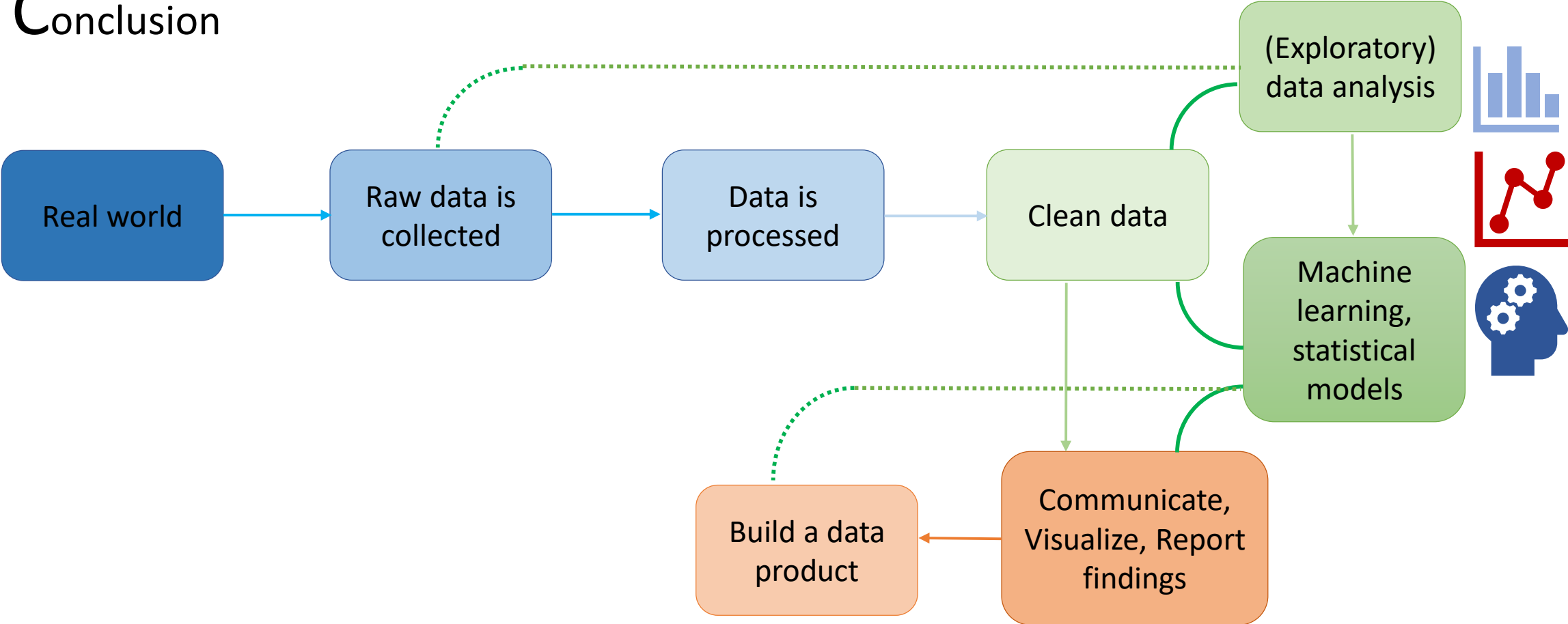
# Tools

Excel,  
Python, R  
SAP  
Google Analytics  
Tableau

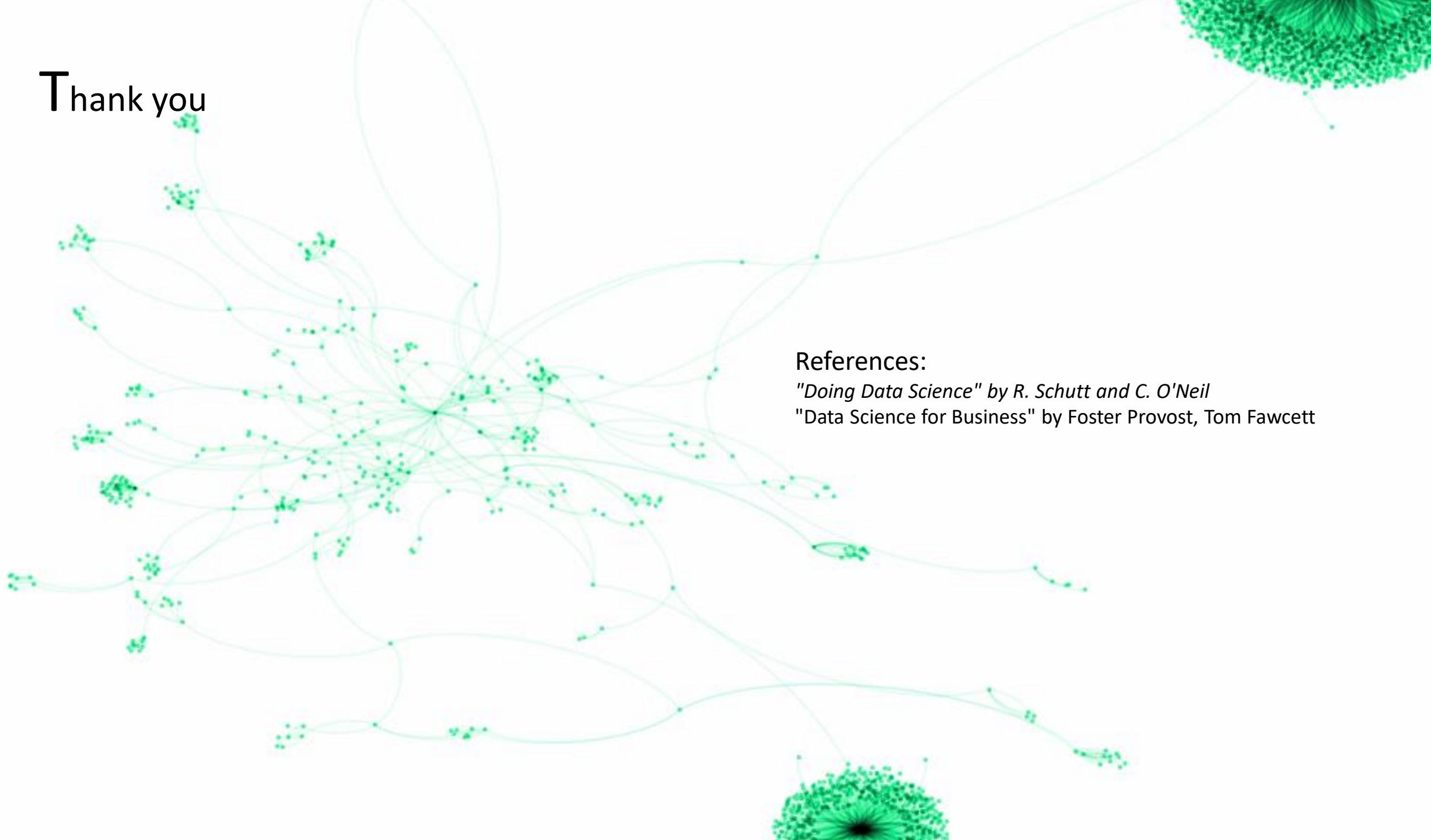
Popular Analytics Tools	Top Companies Using Them
<b>Open Source</b>	
R	Accenture, Cognizant, Google, Facebook, Citibank, Genpact, MuSigma, Fractal Analytics
Python	Alibaba, Google, Cognizant, TCS, Genpact, Gramener
Apache Spark	Uber, Pinterest, Ola, Facebook, Infosys, Wipro, Netflix
Apache Storm	Groupon, Twitter, Yahoo, Alibaba, Spotify, Flipboard
PIG & HIVE	Yahoo, Facebook, Twitter, Baidu, Uber, Flipkart
<b>Commercial</b>	
SAS	HSBC, Citibank, Google, Netflix, WNS, Genpact, Accenture, HDFC
Tableau	Barclays, Citibank, Gallup, Ogilvy, LA Times, Toyota, AOL, Dell, HP, Marico, Ashok Leyland
Excel	Almost every company known to mankind
Qlikview	TCS, Capgemini, Accenture, Cisco, Deloitte, Citibank
Splunk	Adobe, Nasdaq, Coca-Cola, Cognizant, Groupon, First Data, GoodData, ING, Intuit

Source: <https://analyticstraining.com/10-most-popular-analytic-tools-in-business/>

# Conclusion



Thank you



References:

*"Doing Data Science" by R. Schutt and C. O'Neil*

*"Data Science for Business" by Foster Provost, Tom Fawcett*

# Answers (I)

Which concept of descriptive statistics do we need:

1. To compute the average of a data set? [The mean.](#)
2. To compute the spread of values within a data set? [The variance or the standard deviation.](#)
3. To find the most frequent data within a data set? [The mode.](#)
4. To compute the middle of the data? [The median.](#)

Which statistics are displayed in boxplots? [The quartiles.](#) Then there are variations of [boxplots where different statistics are included in the whiskers, e.g. we saw the minimum and the maximum, but other statistics could be present as well in other variations.](#)

List one visualization technique seen in this course, and at least one that you might have encountered. [Boxplots, you may have encountered histograms, pie charts, scatter plots for example.](#)

# Answers (II)

1. What does entropy measure? It measures im/purity (of a set) or un/certainty.
2. What does information gain measure? A change of knowledge about the target given the knowledge of an attribute.
3. A set which is pure has (a) high, (b) low entropy. Low entropy (the lowest is 0).
4. Describe the geometry of decision trees and linear discriminant. Decision trees have boundaries which are perpendicular to the attribute axes, while for linear discriminant, there is one boundary with any possibly orientation.



# Answers (III)

1. What is the main difference between supervised and unsupervised learning? **Supervised learning relies on ground truth.**
2. Classification is considered a task of (a) supervised learning, (b) unsupervised learning. **It is typically considered as supervised learning.**
3. Clustering is considered a task of (a) supervised learning, (b) unsupervised learning. **It is typically considered as unsupervised learning.**
4. Given a data base of customers, you need to decide whether a customer will leave or stay. Which learning task is involved? **Classification (supervised learning).**
5. Given a data base of faults in a system, you need to decide the likelihood of a new fault happening. Which learning task is involved? **Regression (supervised learning).**