Introduction to Data Mining and Machine Learning

COMP90049
Knowledge Technologies

Hasti Samadi & Sarah Erfani & Karin Verspoor, CIS

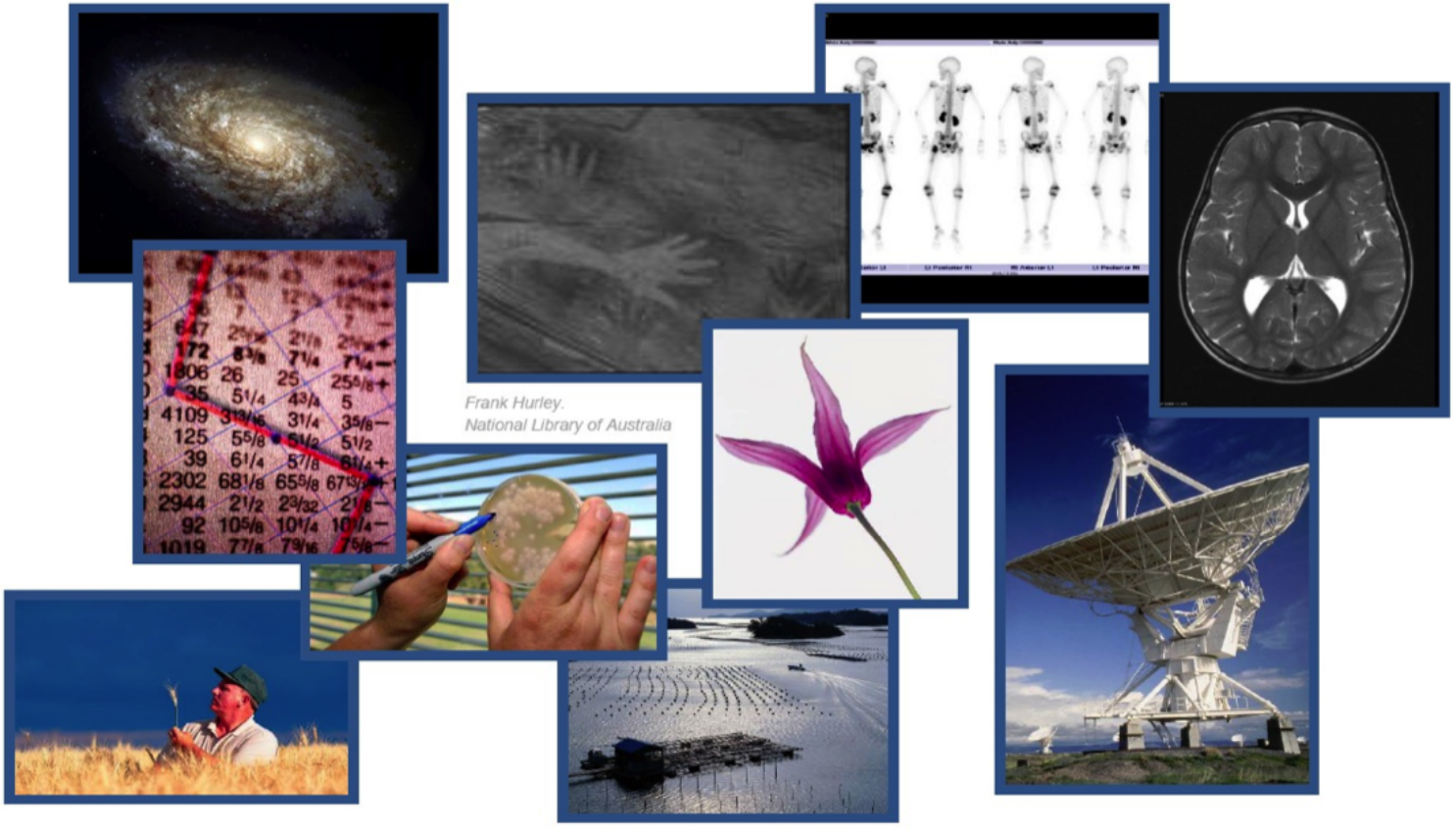
Semester 2, 2019

THE UNIVERSITY OF MELBOURNE

From Data to Wisdom

http://www.innovation.gov.au/Science/PMSEIC/Documents/DataForScience.pdf

Remember: Data is everywhere



Reminder: What is Knowledge?

Reminder: What is Knowledge?

Information interpreted with respect to a user's context to extend human understanding in a given area.

“Big Data”

"Big Data"

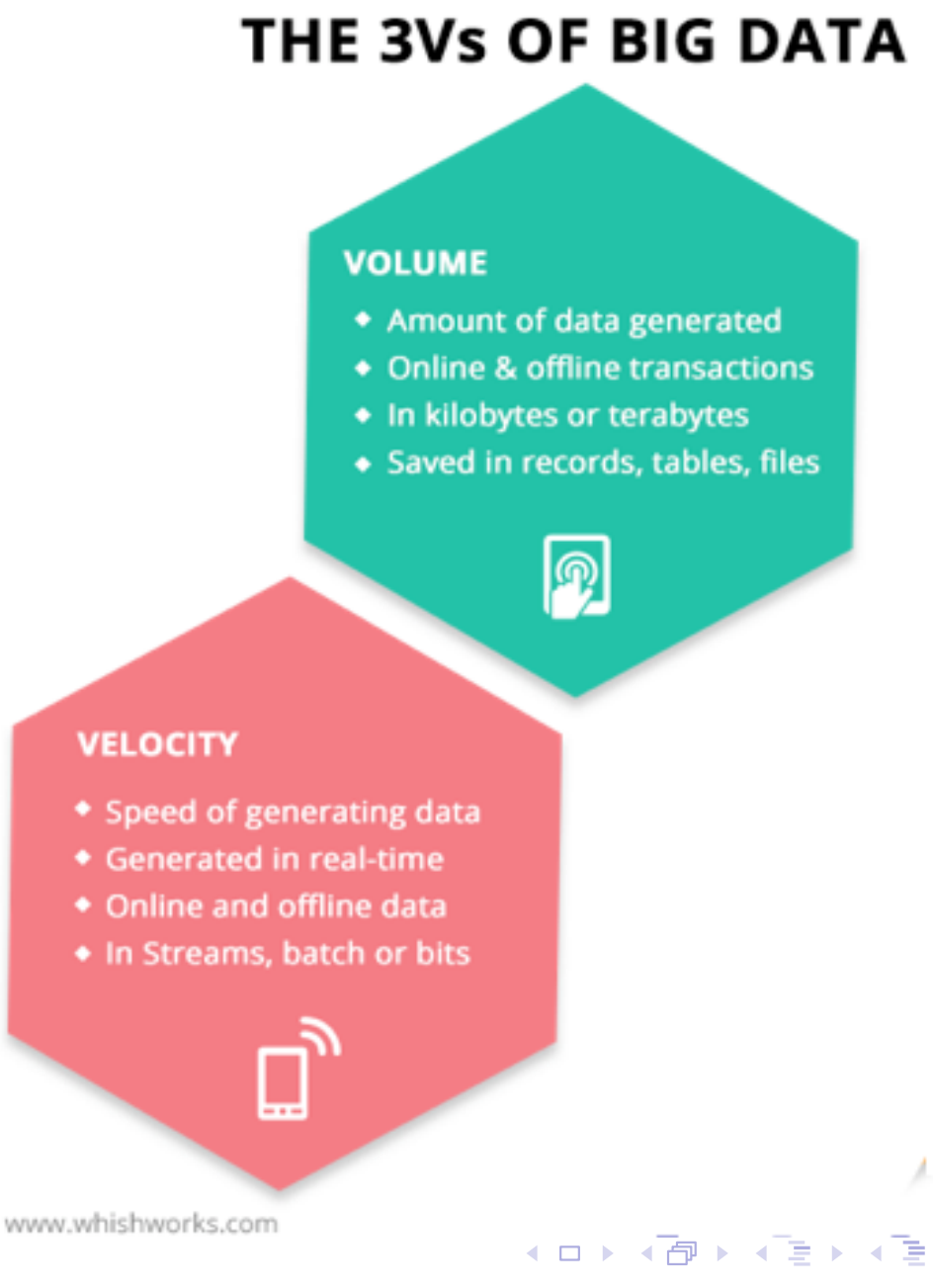
3 'V's:

- Volume

"Big Data"

3 'V's:

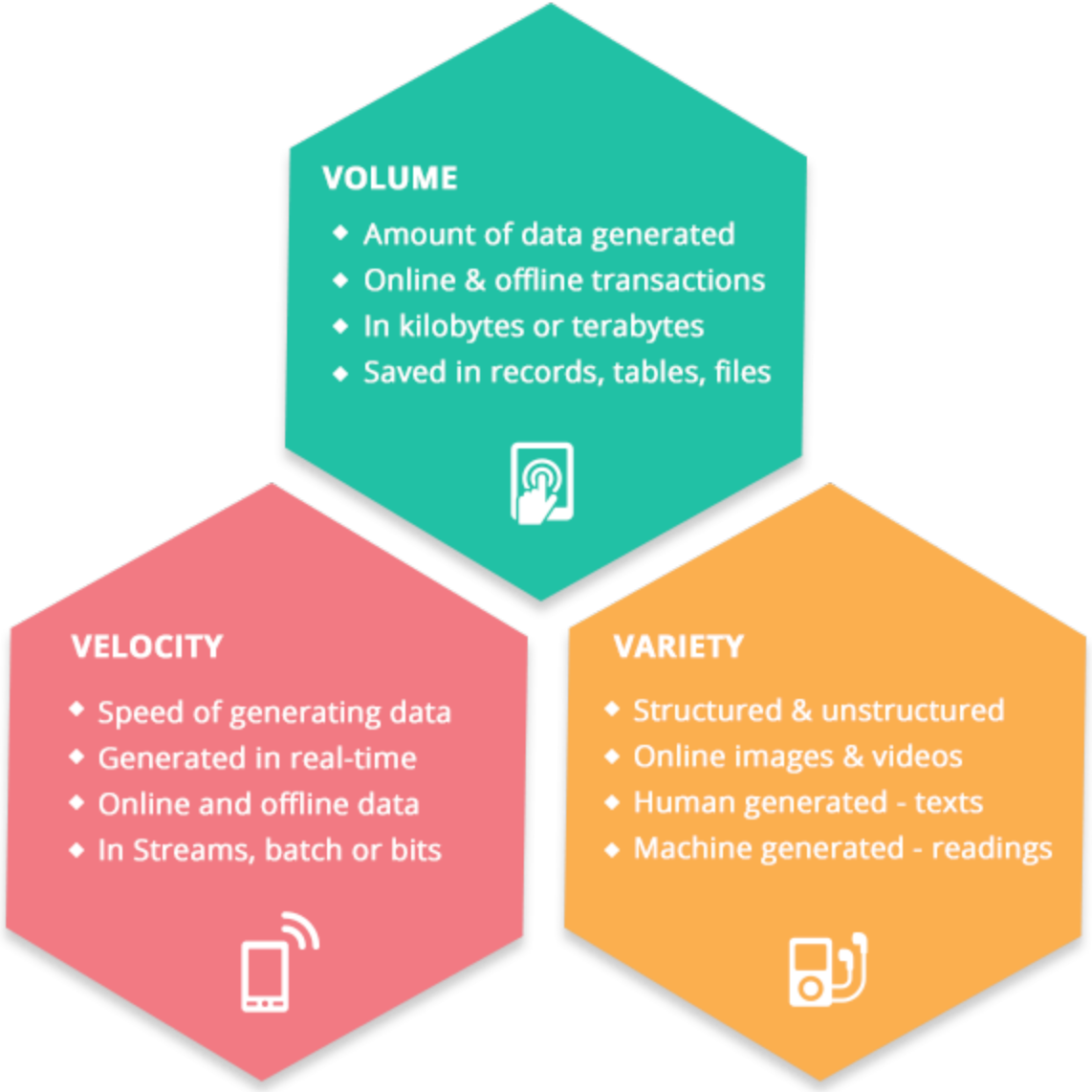
- Volume
- Velocity



"Big Data"

3 'V's:

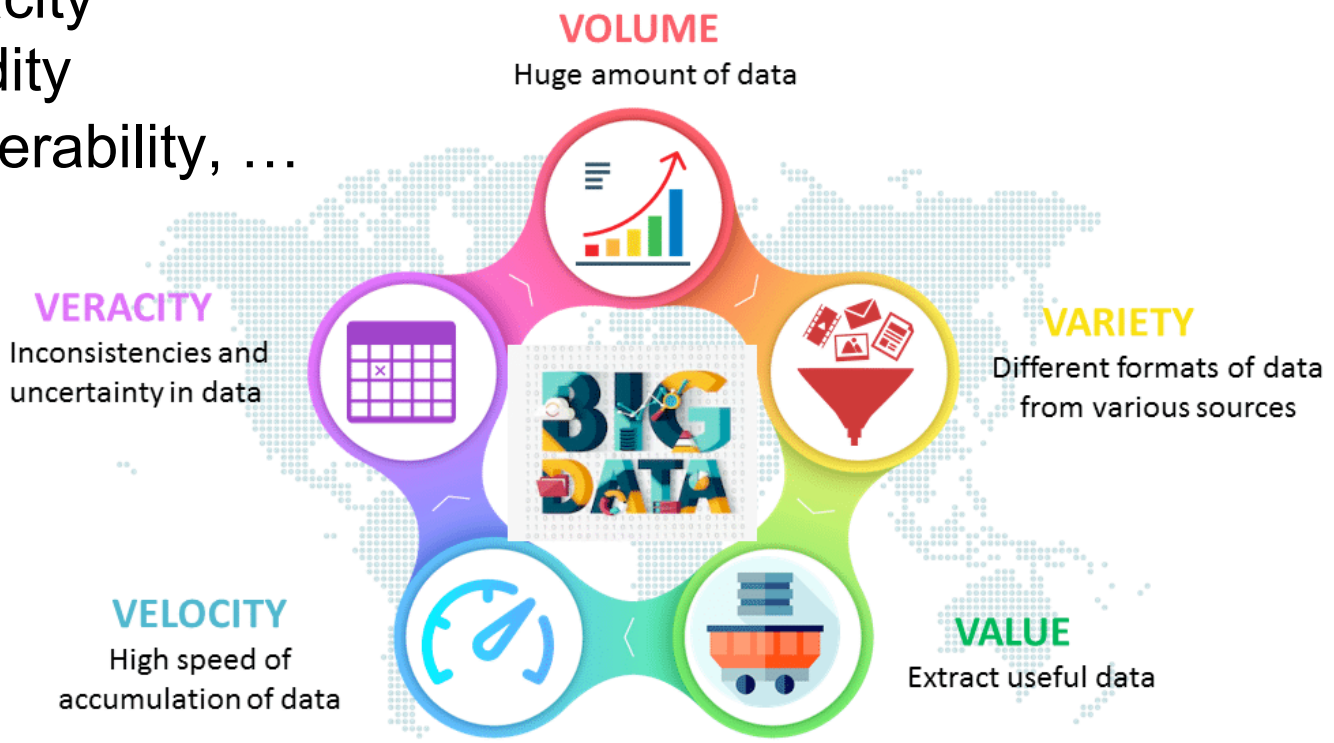
- Volume
- Velocity
- Variety



"Big Data"

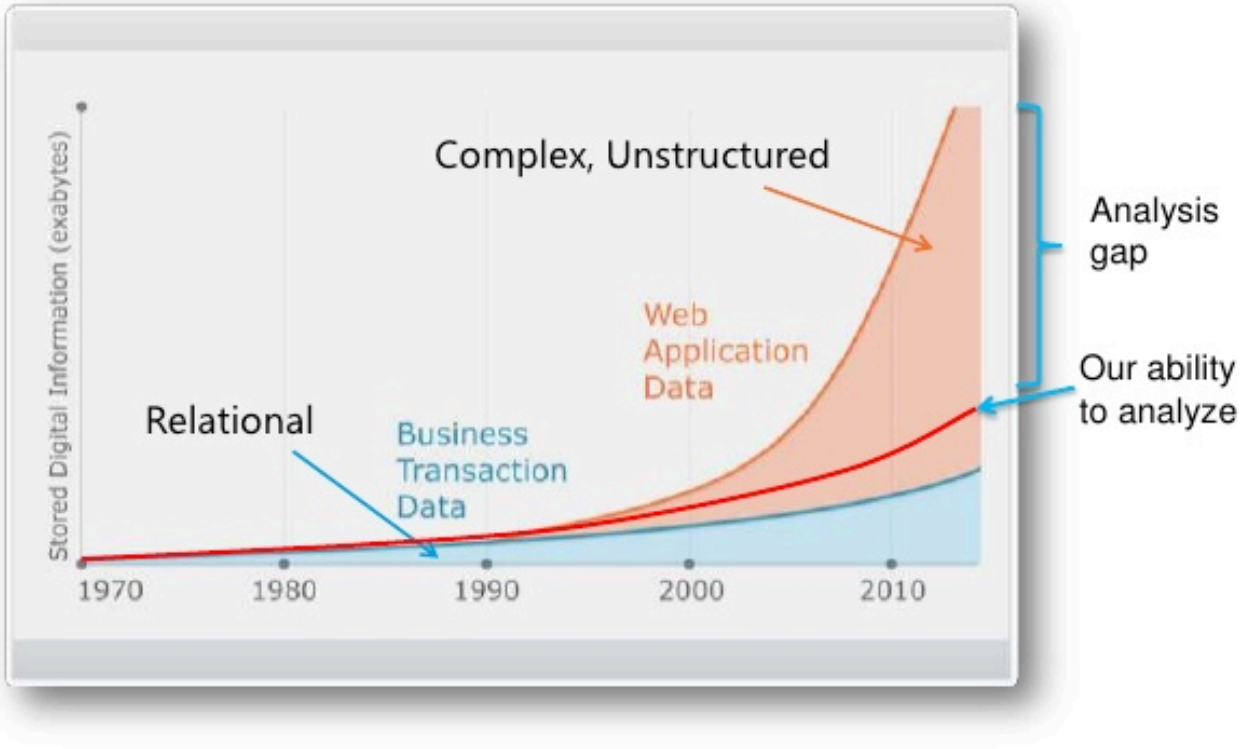
More 'V's:

- Value
- Veracity
- Validity
- Vulnerability, …



Importance of Problem

- Current computational methods cannot handle *magnitude* and ***dimensionality*** of the data
- Decision makers and Scientists need techniques to help making evidence based decisions



Source: An IDC White Paper - sponsored by EMC. As the Economy Contracts, the Digital Universe Expands. May 2009.

Machine learning definitions

Arthur Samuel (1959)

- "Field of study that gives computers the ability to learn without being explicitly programmed"

Machine learning definitions

Arthur Samuel (1959)

- "Field of study that gives computers the ability to learn without being explicitly programmed"

Tom Mitchell (1997)

- "how to construct computer programs that automatically improve with experience .... A computer program is said to learn from experience ... if its performance ... improves with experience..."

Machine learning definitions

Witten et al. (2011)

- "Data Mining is ... the process of discovering patterns in data.... *Machine Learning* [comprises] techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it.

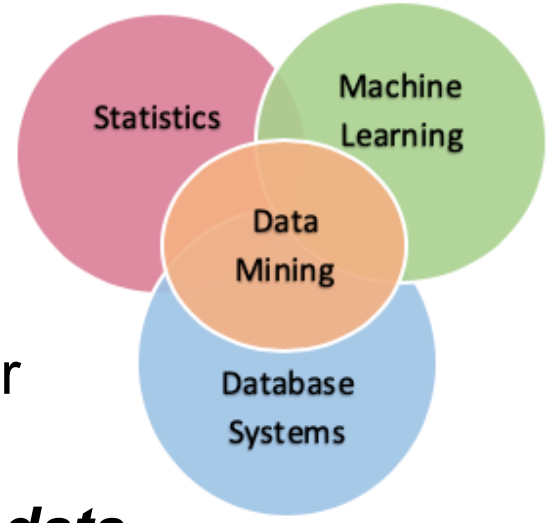
Machine learning definitions

Witten et al. (2011)

- "Data Mining is ... the process of discovering patterns in data.... *Machine Learning* [comprises] techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it.

Data Mining vs Machine Learning

The distinctions between Data Mining and Machine Learning are not cut-and-dried.



- Data mining is primarily about discovering something hidden in your data, that you did not know before.
  - ***Extracting Knowledge from data.***

- Machine learning emphasises algorithms used to generalise existing knowledge to new data, as accurately as possible.
  - ***Techniques used to learn from data.***

Data mining applications typically use a lot of machine learning techniques.

Data Mining vs Machine Learning vs Data Science

Machine learning tends to:

- be more concerned with theory than applications
- largely ignore questions of run time/scalability

Data mining tends to:

- be more concerned with (business) applications than theory
- talk a lot about databases and run time/scalability

Data science tends to:

- straddle data mining and machine learning, and have more focus on applications and interpreting/communicating data insights

Very fuzzy dividing line between the three

Types of machine learning algorithms

Supervised learning

- *Teach* the computer how to do something *(by example)*, then let it use its new-found knowledge to do it
- Labeled data: for given inputs, provide the expected output ("the answer")
- Infer a function mapping from inputs to outputs

Applications

- Classification

  predicting a label (*rainy* or *sunny*)

- Regression

  predicting a numeric quantity (*temperature*)

Example: Supervised Learning

Scenario:

You are in charge of developing the next “release” of Coca Cola, and want to be able to estimate how well received a given recipe will be

Example: Supervised Learning

Scenario:

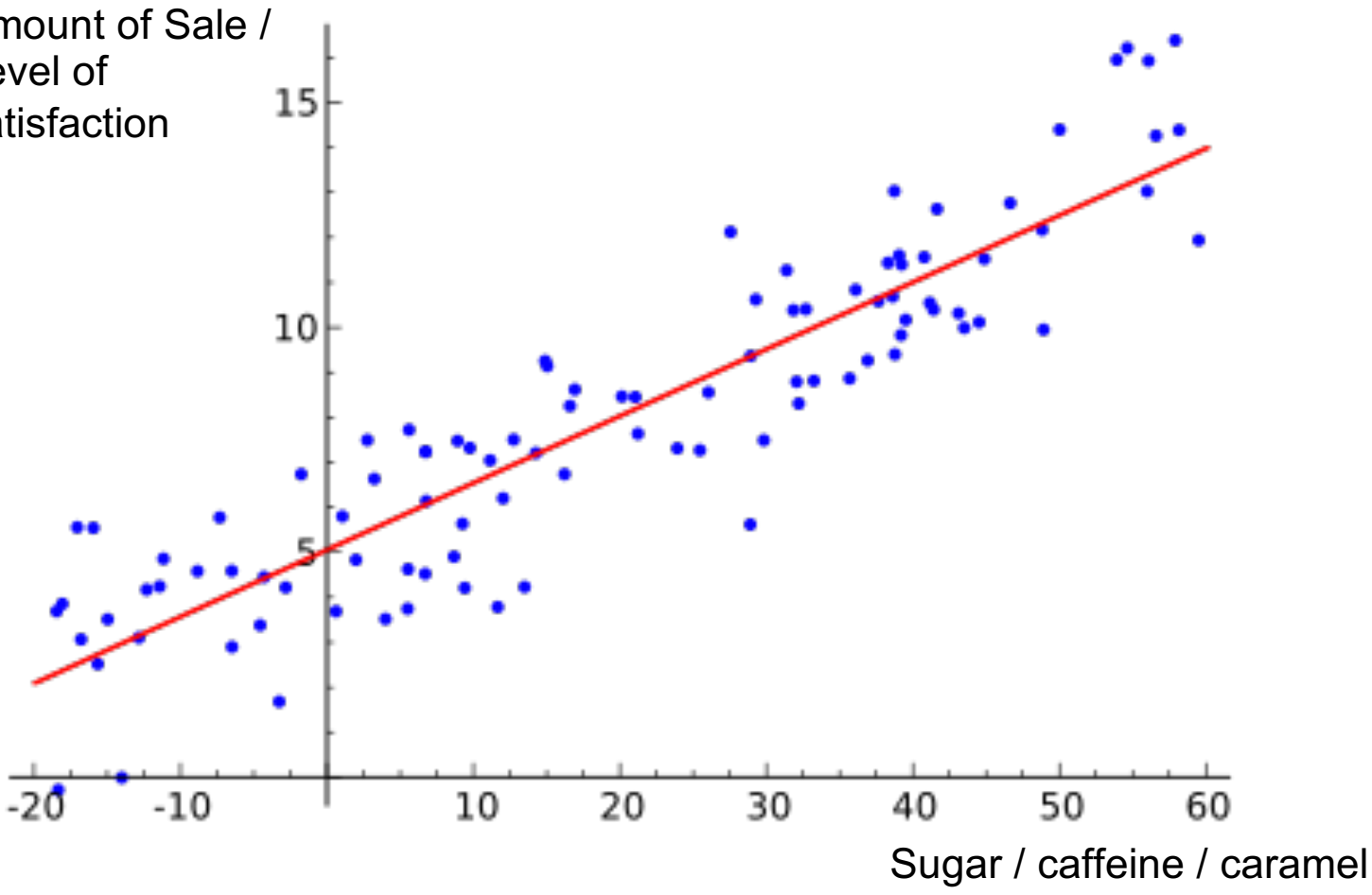
You are in charge of developing the next “release” of Coca Cola, and want to be able to estimate how well received a given recipe will be

Solution:

Carry out taste tests over various “recipes” with varying proportions of sugar, caramel, caffeine, phosphoric acid, coca leaf extract, ... (and any number of “secret” new ingredients), and estimate the function which predicts customer satisfaction from these numbers

Example: Supervised Learning (Regression)

Regression



Types of machine learning algorithms

Unsupervised learning

- Let the computer *learn how to do something*
- Unlabelled data: Don't give the computer "the answer"
- Determine structure and patterns in data

Applications

- Clustering
  grouping similar instances into clusters (marketing groups)
- Association
  detecting associations between features (Shelf layout)
- Recommender systems (Amazon, Netflix)
- Anomaly/outlier detection (Credit card fraud)

Example: Unsupervised learning applications

Scenario:

You are a supermarket manager, wishing to boost sales without increasing expenditure, but with lots of historical purchase data

Example: Unsupervised learning applications

Scenario:

You are a supermarket manager, wishing to boost sales without increasing expenditure, but with lots of historical purchase data

Solution:

Strategically position products to entice consumers to spend more:

- beer next to chips?
- beer next to bathroom cleaner?

Example: Unsupervised learning applications

Scenario:

You are a supermarket manager, wishing to boost sales without increasing expenditure, but with lots of historical purchase data

Solution:

Strategically position products to entice consumers to spend more:

- beer next to chips?
- beer next to bathroom cleaner?

Association Rule Discovery

Other types of machine learning algorithms

- Semi-structured Learning
- Reinforcement Learning

Some basic Machine Learning concepts

The input to a machine learning system consists of:

- ***Instances***: the individual, independent examples of a concept
- ***Attributes***: measuring aspects of an instance *also known as features*
- ***Classes***: things that we aim to learn

Example: Supervised Learning (Classification)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Given information about current weather conditions and the forecast, can we determine whether we will go out to play?

Classification (Instances/Training examples)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

INSTANCE 1

INSTANCE 2

Classification (Attributes/Features)

ATTRIBUTE 1	ATTRIBUTE 2			
Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Classification (Classes/Labels)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

CLASS

Attributes

Each instance is described by a fixed feature vector

Possible attribute types (levels of measurement):

- Nominal / categorical
- Ordinal
- Continuous / Numeric

Attributes

Each instance is described by a fixed feature vector

Possible attribute types (levels of measurement):

- Nominal / categorical

  `{male, female}`

- Ordinal
- Continuous / Numeric

Attributes

Each instance is described by a fixed feature vector

Possible attribute types (levels of measurement):

- Nominal / categorical

  `{male, female}`

- Ordinal

  `{low, medium, high}`

- Continuous / Numeric

Attributes

Each instance is described by a fixed feature vector

Possible attribute types (levels of measurement):

- Nominal / categorical

  `{male, female}`

- Ordinal

  `{low, medium, high}`

- Continuous / Numeric

  `weight, Age`

Nominal attributes

- Also called *categorical*, *enumerated*, or *discrete*
- Values are distinct symbols

  `{Sunny,Overcast,Rainy}`

  `{Head, Tail}`
- Values themselves serve only as labels or names
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

Ordinal attributes

- An explicit order is imposed on the values
  `{hot,mild,cool}` where `hot>mild>cool`
- No distance between values defined; addition and subtraction don't make sense
- But we can compare the values
  `temperature` < `hot`
- Distinction between nominal and ordinal not always clear (e.g. `outlook`)

Continuous attributes

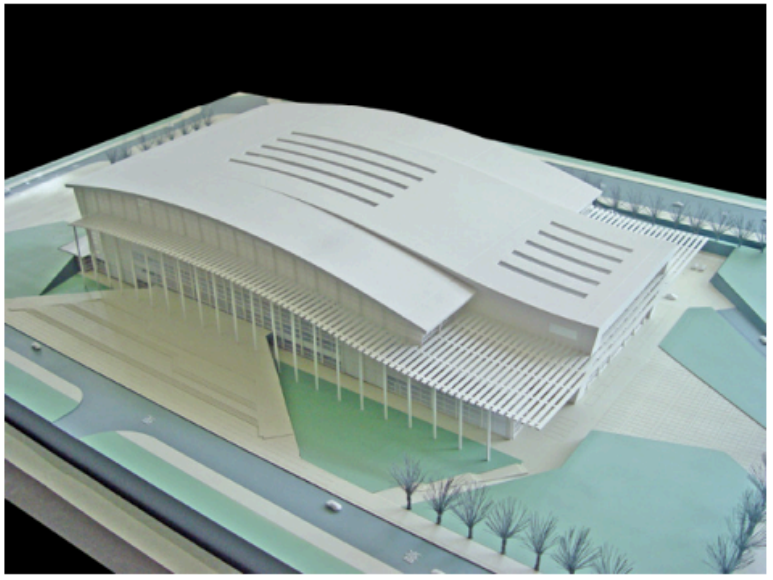
- Continuous features are real-valued with a well-defined zero point and no explicit upper bound
- Also called *numeric*
- Example: attribute `distance`

  Distance between an object and itself is zero

- All mathematical operations are allowed

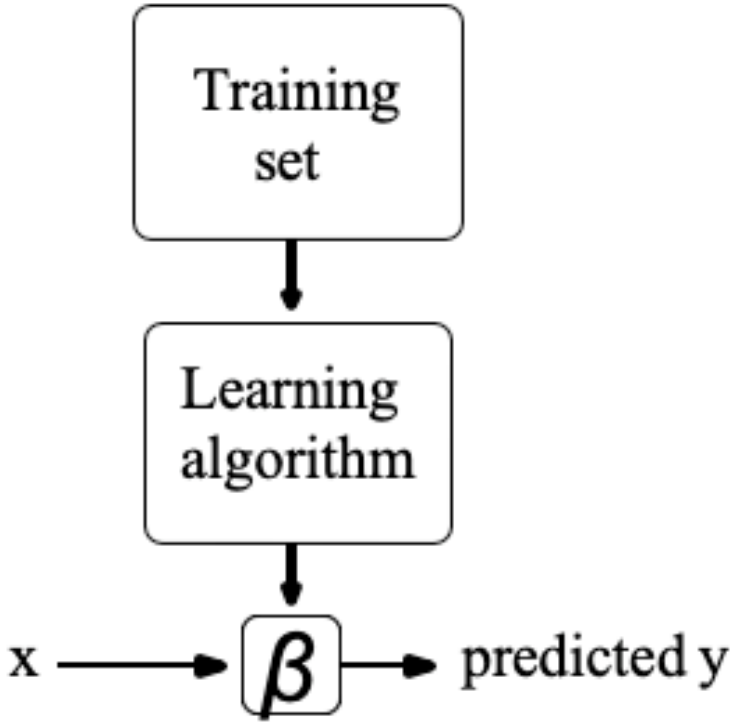
Models

- We talk about modelling data, building models.
- So, what makes a model a model?

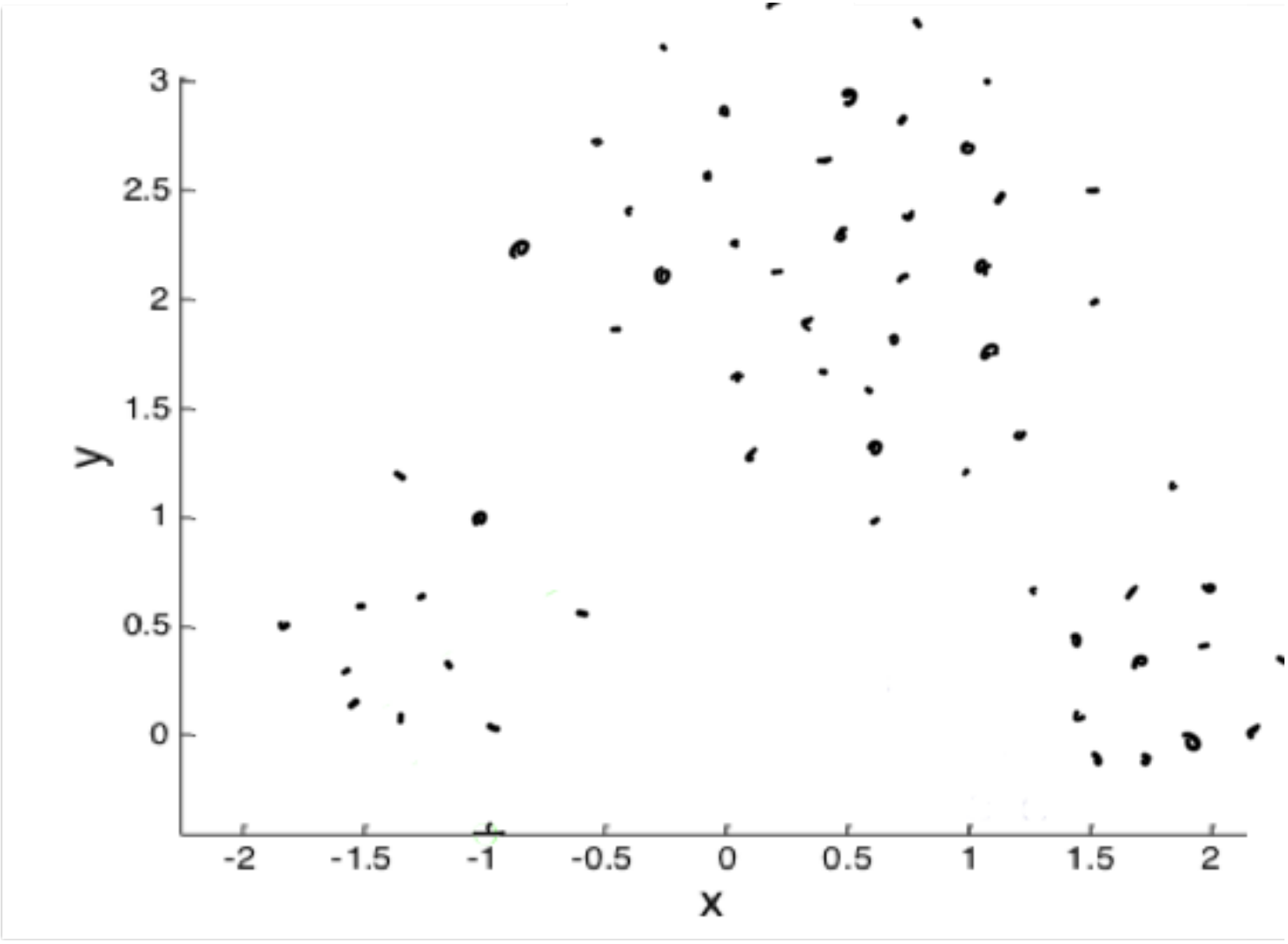


Models

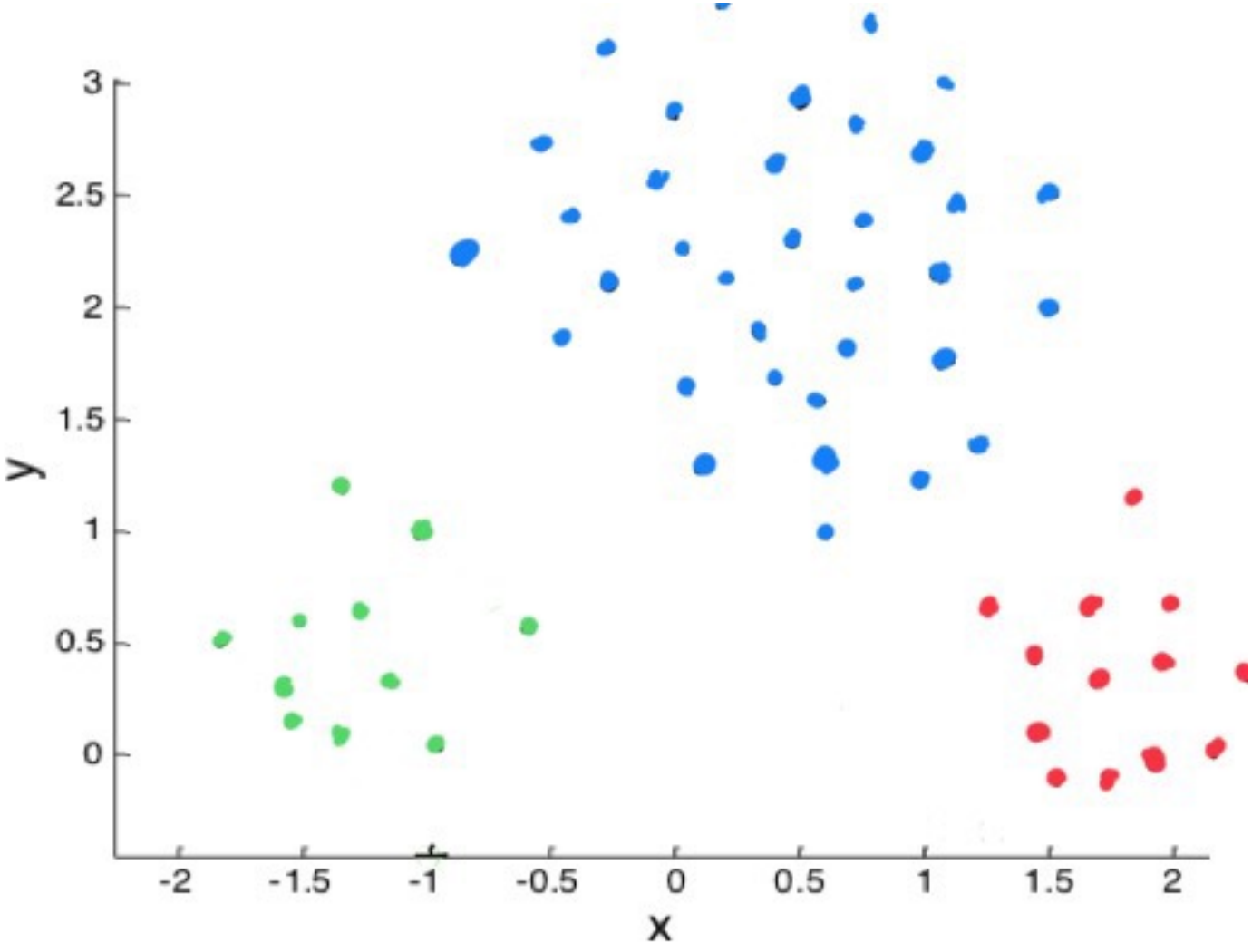
- A model is our attempt to understand and represent the reality through a particular lens: architectural, biological, or mathematical.
- A Machine Learning a Model is usually a mathematical representation of our dataset (training data).
- We can use different algorithms to build different models (with different levels of accuracy).
- In modelling there is a trade-off between simplicity and accuracy



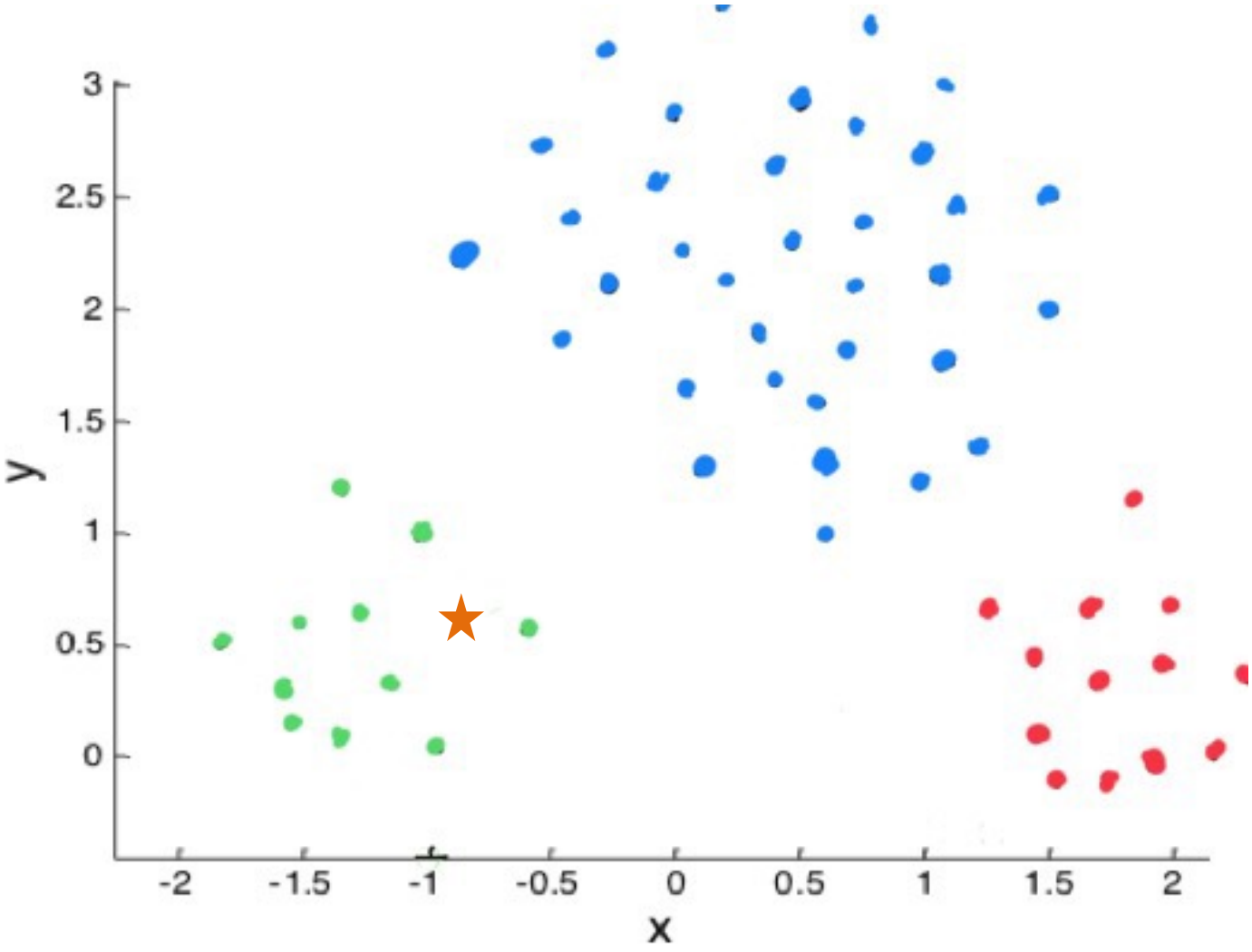
Models



Models



Models



Thought experiment

How might you approach data mining the Weather dataset?

- Methods
  - Using Supervised methods?
  - Using Unsupervised methods?

- Attributes
  - Are there regularities among the attributes?
  - Are there different ways you could make use of the attributes (e.g. different combinations? different thresholds?)?

Books and Websites

Introduction to Data Mining
Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. Addison Wesley.
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Data Mining: Practical Machine Learning Tools and Techniques
Ian Witten, Eibe Frank, Mark Hall
http://www.cs.waikato.ac.nz/ml/weka/book.html

Tom M. Mitchell. Machine Learning. McGraw-Hill, New York, USA, 1997

Tools and Resources

WEKA Toolkit
http://www.cs.waikato.ac.nz/ml/weka/index.html

List of more specific tools
http://www-users.cs.umn.edu/~kumar/dmbook/resources.htm

Data sets
UC Irvine Machine Learning Data Repository
http://archive.ics.uci.edu/ml/datasets.html