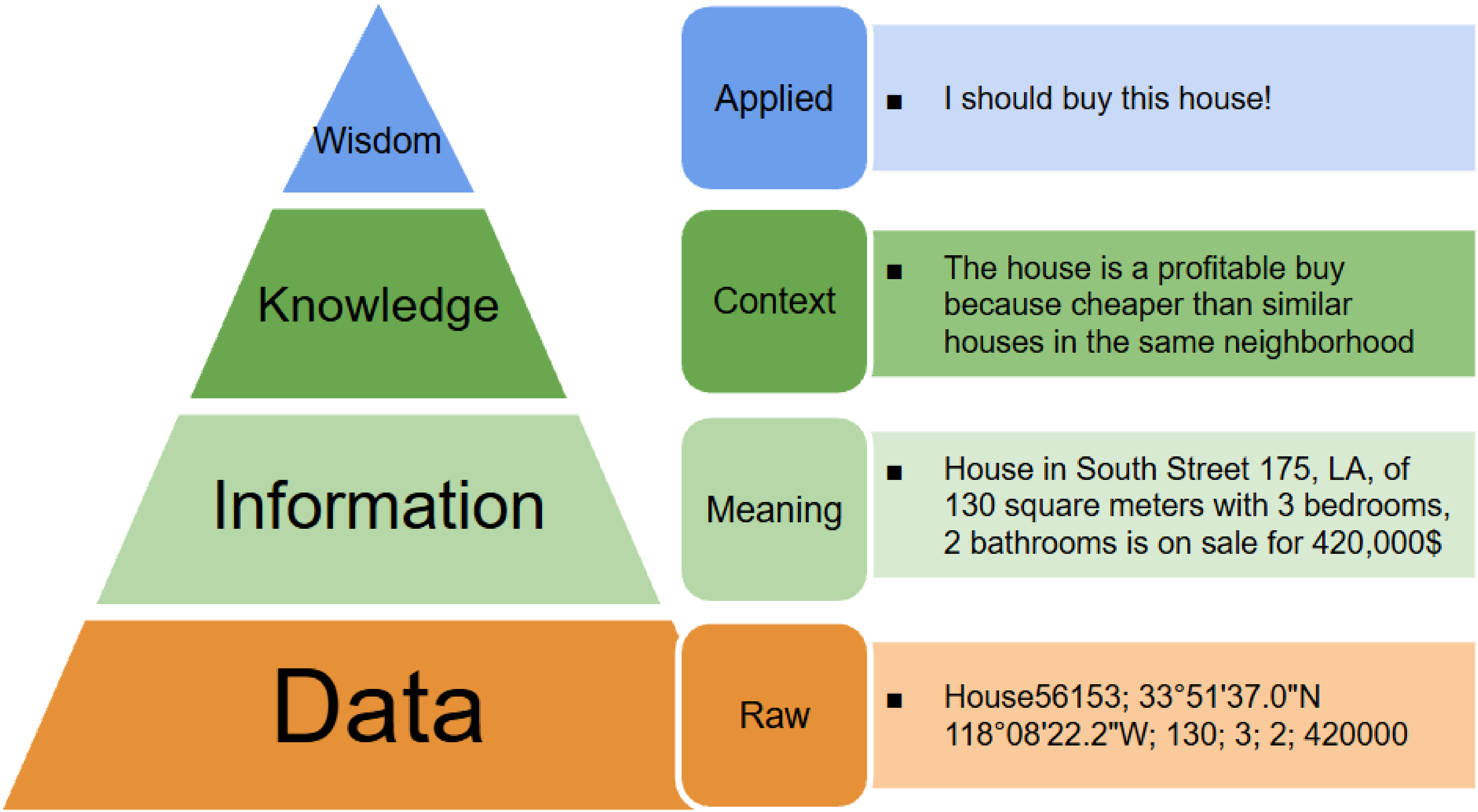
Data Analytics

Danilo Montesi
Stefano Giovanni Rizzo

From data to wisdom

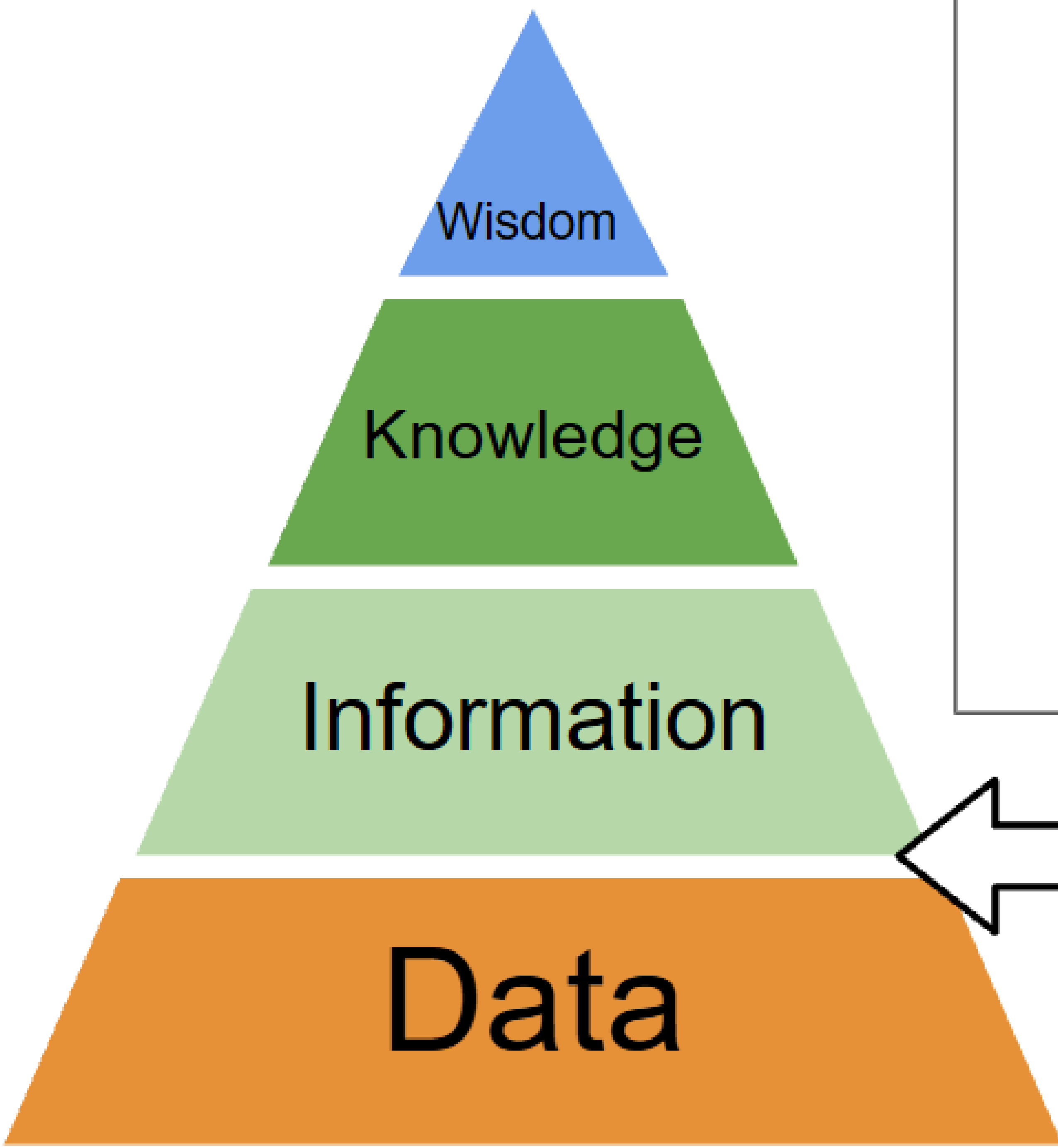
DIKW Pyramid: Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge



So far in this course

- **Semi-structured data**
  - Storing and querying data without having a rigid schema.
  - How semi-structured data relates to structured data and can be queried using a query language for structured data (SQL)
- **Information retrieval**
  - Retrieve a subset of documents with respect to user's information need
  - Searching in the WWW
  - Ranking documents by their relevance to a text query and user's feedback

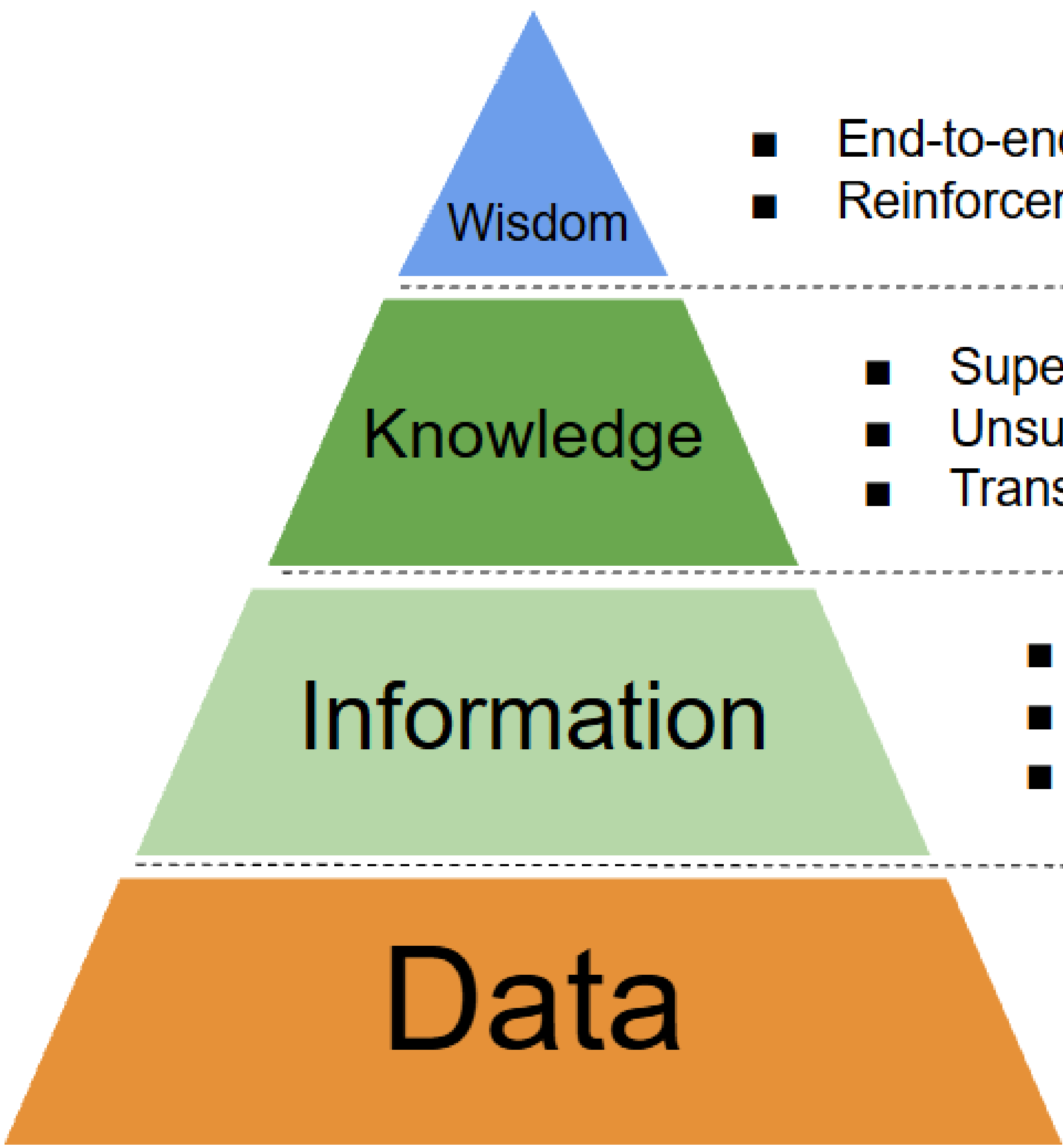
Where in the pyramid?



- We were still in a low level:
  - Semi-structured data can provide information only by manually defining complex queries.
  - Information retrieval searches and ranks information without extracting it from data: documents are already “information” in natural language.

What we will see

Starting from the bottom:



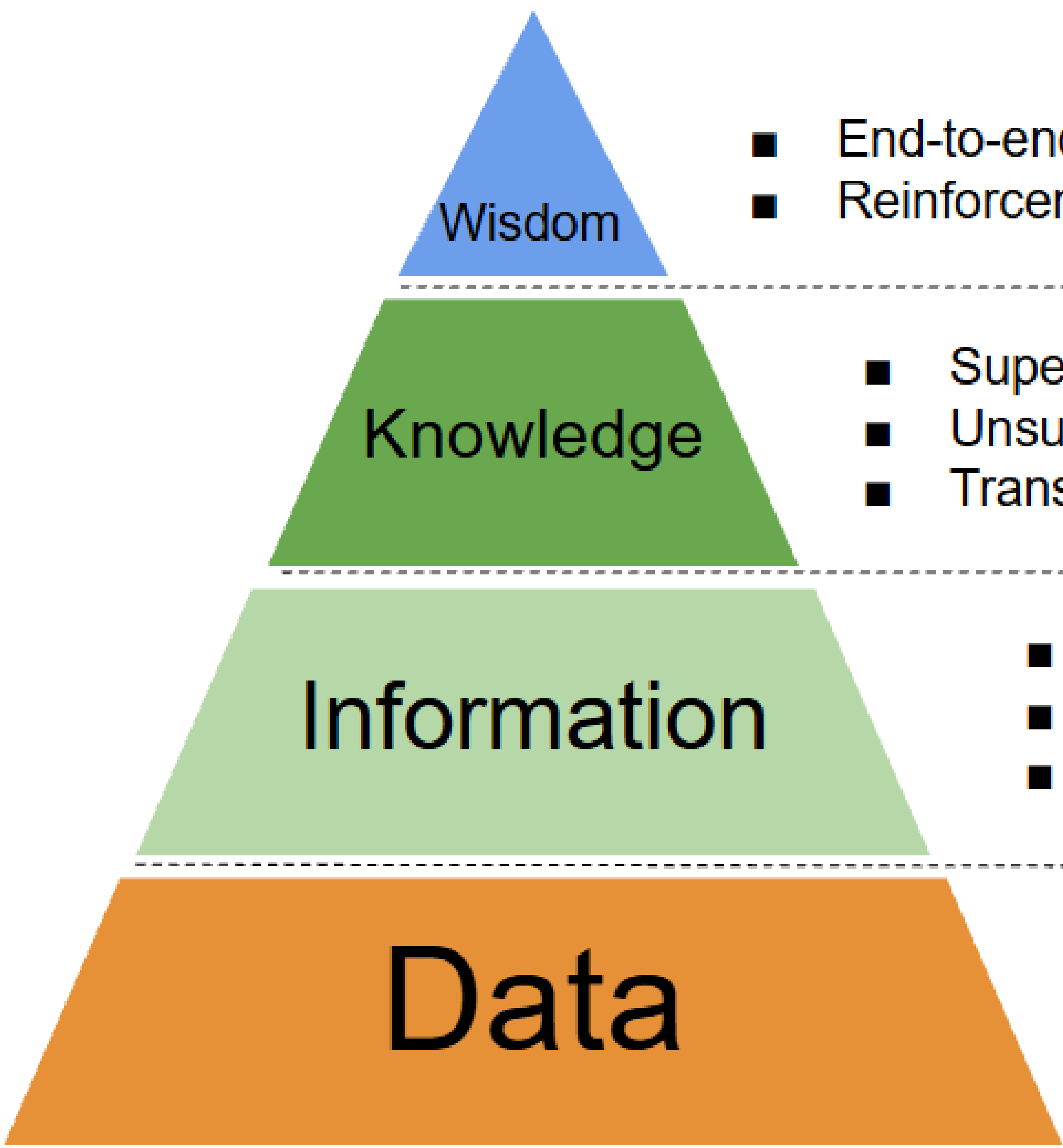
- End-to-end Deep Learning
- Reinforcement learning

- Supervised learning: regression, classification
- Unsupervised learning: clustering, embedding
- Transfer learning: changing tasks

- Data Analysis
- Data Warehouse
- Decision support systems

What we will see

Starting from the bottom:



- End-to-end Deep Learning
- Reinforcement learning

- Supervised learning: regression, classification
- Unsupervised learning: clustering, embedding
- Transfer learning: changing tasks

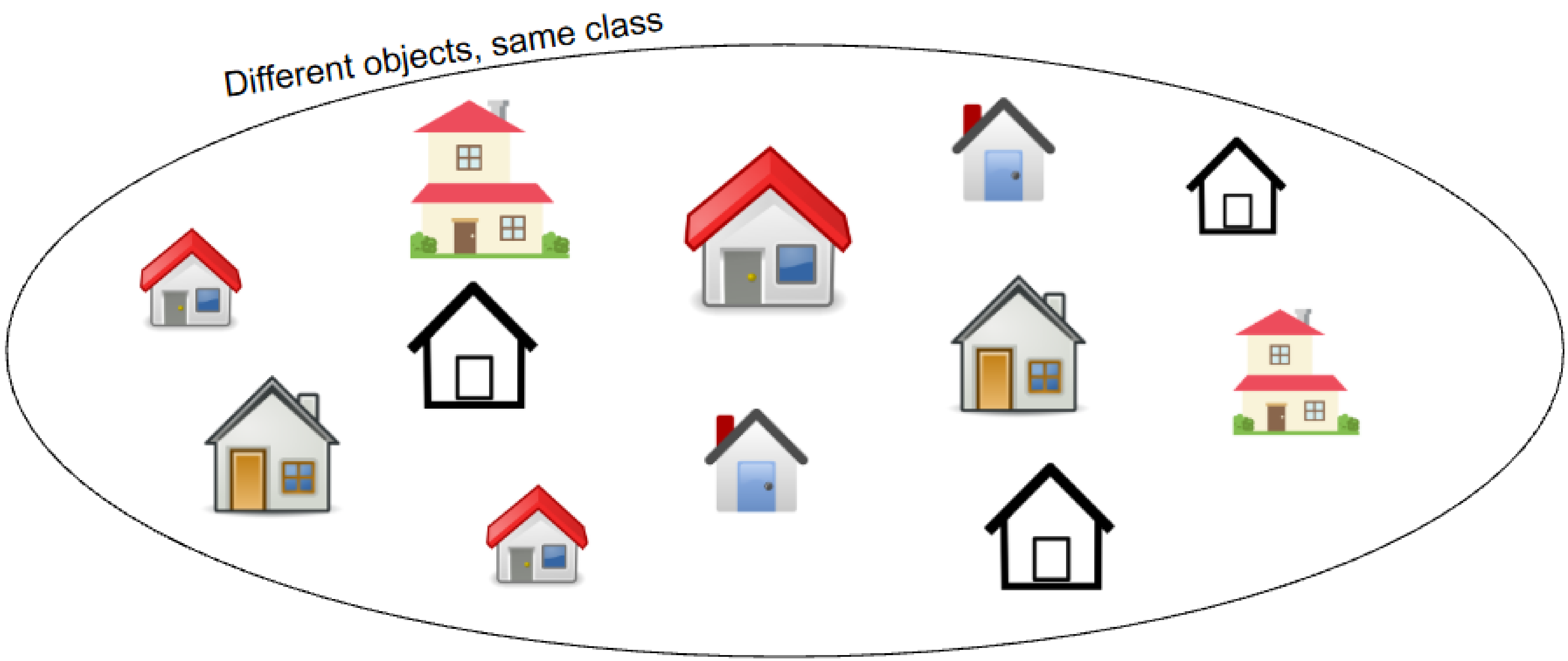
- **Data Analysis** ⟵
- Data Warehouse
- Decision support systems

Data Analysis

- The scope of Data Analysis is to **extract** basic **information** from collections of data.
- Extracted information can be:
  - Summarized information: e.g. the average from a set of numerical values
  - Association information: e.g. the relation between two sets of values (houses' price vs. square meters)
- Conceptual foundation is descriptive **statistics**

Concepts: Population

- A ***population*** is a collection of objects we are interested in, for example:
  - All the houses in Los Angeles
  - All the students in the university
  - All the receipts from a grocery shop



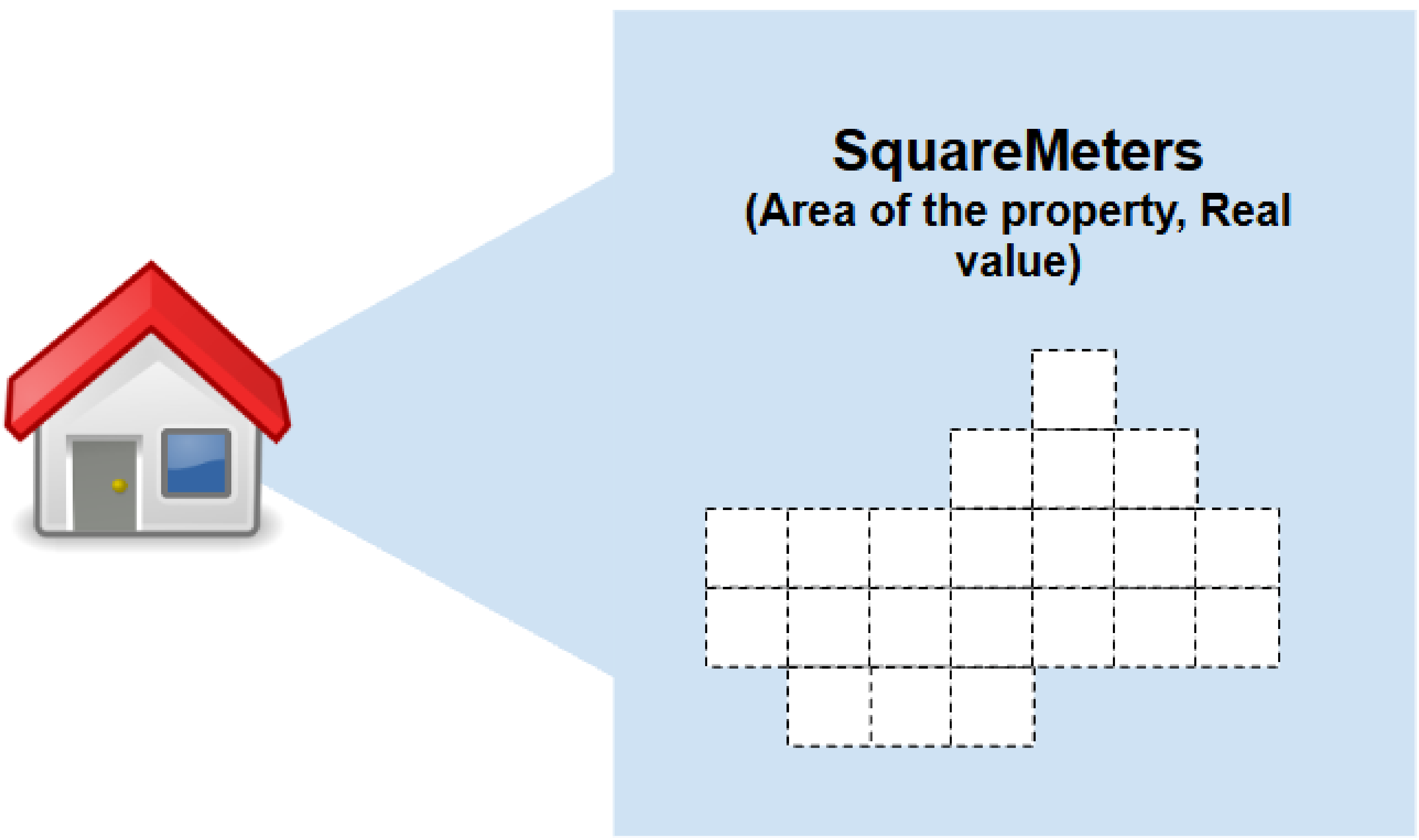
Concepts: Record

- A ***record*** (or observation, case) is a tuple of values that characterize an element of a population

City	Latitude	Longitude	Bedrooms	SquareMeters	Price
Los Angeles	**33°51'37.0"N**	**118°08'22.2"W**	**3**	**130**	**420000**

Concepts: variable

- A ***variable*** (or field, feature) it's the name for a record's value and has a common meaning and type for all the records in the population



Concepts: type of variable 1/2

- We can classify variable depending on the type of the values they can take
- Most important distinction is between
  - **numerical** variables (quantitative): if we can apply arithmetic operations on them
  - **categorical** variables (qualitative): otherwise
- Example:
  - The city (e.g. Los Angeles, New York, Rome) is a **categorical** variable
  - The price (e.g. 420000) is a **numerical** variable

Concepts: type of variable 2/2

- Numerical variables can be:
  - **Discrete**, if values can be counted
  - **Continue**, if they are the results of a continuous measure
- Categorical variables can be:
  - **Ordinal**, if a natural order exists on the possible values (e.g. school grades: A,B,C,D)
  - **Nominal**, otherwise (e.g. colors)

A dataset

- Finally, the collection of records (a dataset) takes the form of a single table

City	Latitude	Longitude	Bedrooms	SquareMeters	Price
Los Angeles	33°51'37.0"N	118°08'22.2"W	3	130	420000
Los Angeles	33°50'17.7"N	118°09'12.6"W	2	60	380000
Los Angeles	33°49'32.3"N	118°08'44.1"W	5	230	2500000
...	...	...	...	...	...
Albuquerque	35°12'08.1"N	106°58'31.1"W	2	105	190000
Albuquerque	35°15'17.0"N	106°59'26.8"W	4	225	440000
Albuquerque	35°14'22.0"N	106°26'26.2"W	2	140	220000
Albuquerque	35°32'23.0"N	106°38'21.2"W	3	150	250000

Descriptive statistic

- Descriptive statistic provides synthesizing indicators to identify, with a single value, **statistical properties** of a population..
- ...with respect to a **single variable**:
  - Centrality indicators: arithmetic mean, mode, median
  - Variation indicator: variance, standard deviation
- ...with respect to **multiple variables**:
  - Covariance
  - Correlation

Centrality: arithmetic mean

- Let X be a **numerical** variable of our dataset (we can't extract the mean from categorical values!)
- **n** is the **number of records** in our population
- $\mathbf{X_i}$ is the i-th record

$$mean = \frac{\sum_{i=1}^{n} X_i}{n}$$

Arithmetic mean: properties

- Suppose you have a record with a missing data (e.g. the price)

City	Latitude	Longitude	Bedrooms	SquareMeters	Price
Los Angeles	33°51'37.0"N	118°08'22.2"W	3	130	**???**

- You can keep the record without affecting the variable mean:
  - A solution is to r**eplace the missing data with the mean** for that variable
  - Adding a record with a mean value will not change the arithmetic mean for the whole dataset

Centrality: median

- Given a population of *sorted* values (e.g. the column "SquareMeters" sorted by its value):

(30,34,37,37, … ,91,91,91,91,91, … ,525,600,670)

x_1 $x_{n/2}$ x_n

- The *median* is the value in central position ($x_{n/2}$=91)
- Median age in Italy

Median: properties

- Median is a **robust** indicator: anomalies such as very large or very small values do not affects much the median value.
- This was not true for mean, which is much more sensitive to anomalies.
- Consider the following example:

2,3,3,4,5,6,6 {Median:4; Mean:4}

2,3,3,4,5,6,80 {Median:4; Mean:14.6}

- Median is still 4 (value in the central position) while mean has shifted from 4 to 14.6!

Centrality: mode

- Given a set of values of a variable (e.g. the column "bedrooms"):

  (1,2,4,2,5,3,2,2,3,4,1,3,4,2,6,2,1,3,1,2)

- First we count the occurrences of a value, that is the **frequency** of that value
  - e.g. "1" is repeated 4 times, "2" is repeated 7 times, "3" is repeated 4 times ...

- The *mode* is the value with higher frequency in the set of observations (i.e. the value "2" in the above example)

Mode: properties

- Unlike mean and median, mode also makes sense on **categorical** data:
  mode(Rome,Rome,Los Angeles,Albuquerue): Rome
- In a **voting** system (e.g. a set of many different classifiers) the mode determines the winning final result.
- While robust to **anomalies** like the median, it makes sense also when there is **no linear order** on the possible values (e.g. points in the plane).

Centrality: comparison

- We consider the following set of observations (values) on the variable bedrooms:

(1, 2, 2, 3, 4, 7, 9)

  - **Arithmetic mean** (sum of values of a data set divided by number of values): (1+2+2+3+4+7+9)/7=**4**
  - **Median** (middle value): (1, 2, 2, **3**, 4, 7, 9) = **3**
  - **Mode** (most frequent value): (1, **2**, **2**, 3, 4, 7, 9) = **2**

Centrality: in summary

- We have seen three **centrality** indicators: arithmetic mean, median and mode.
- All these indicators provides a different way of "**summarizing**" a set of values into a single, synthesizing value
  - **Arithmetic mean** is also useful to replace missing or wrong data without changing its overall distribution, but its value it's not drawn from the available data and it's sensitive to anomalies.
  - **Median** is an actual value from the observations and it's robust to anomalies but needs ordinal data.
  - **Mode** is also an actual value, the most frequent one. Robust to anomalies, does not need ordinal data and can be applied on categorical variables.

Variation: squared deviation

- Squared deviation: it measures the difference between each value x_i and the mean of the observations $\bar{x}$

$$dev = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- The more the values are far from the mean, the higher the deviation. In the following sample the mean is 15:

$$\text{dev}(\mathbf{4,6,10,40}) = \sum_{i=1}^{n} (x_i - \mathbf{15})^2$$

$$= (4-15)^2+(6-15)^2+(10-15)^2+(40-15)^2 = 121+81+25+625$$

$$= \mathbf{852}$$

Variation: variance

- The squared deviation is **affected by the number of observations**: the more values we have, the higher the deviation tend to be.
- The **variance** (often represented with s^2, σ^2, or Var) normalizes squared deviation by the number of observations:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n} dev$$

- In the previous example: s^2(**4,6,10,40**) = 852/4 = **213**

Variation: standard deviation

- Variance and squared deviation consider the **squared difference between values and the mean**, in order to have non-negative differences.
- This leads to large values that do not reflect the estimated deviation from the mean.
- **Standard deviation** (often represented with s, σ, or Stdev) is the **square root** of the variance.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Unlike squared deviation and variance, standard deviation is expressed in **the same units** as the data.

Variation: comparison

- We use the same example on the three variation indicators to observe their difference:

(**4,6,10,40**)

- Square deviation: dev(4,6,10,40) = 852
- Variance: Var(4,6,10,40) = 213
- Standard deviation: Stdev(4,6,10,40) = 14.6

Other single variable indicators

- **Minimum** (min): it's the minimum value in the observations
- **Maximum** (max): it's the maximum value of the observations
- **Range**: it's the difference between the maximum and the minimum value

Multiple variables indicators

- In descriptive statistics, the *association measures* allow to describe the **relation between two variables**, looking for associations.
- For example, they are useful to determine:
  - If the **price** of the houses is associated with the **square meters**
  - If the smoke is associated with heart diseases
  - If the budget on advertising is associated with the number of sales
- We will see two measures:
  - **Covariance**: measure the strength of the relation between two variables
  - **Correlation**: measure the strength of the relation between -1 and 1

Association measures: covariance

- The **covariance** quantifies the strength of the relation between two variables X and Y
- More specifically, it is the mean of the products of the values deviations:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n}$$

Association measures: covariance

- The **covariance** quantifies the strength of the relation between variables X and Y
- More specifically, it is the **mean** of the **products** of the values **deviations**:

Deviation of x_i from the mean of x

Deviation of y_i from the mean of y

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})\times(Y_i-\bar{Y})}{n}$$

Arithmetic mean

Product is **high** if deviations are large at the same time (in absolute terms)

Meaning of covariance

- The idea behind: the sum of products of the two deviation is high if, every time X deviate from its mean, Y also deviate from its mean **accordingly**.
- Why?
  - If Y does not deviate, its deviation is low and so will be the product
  - If Y deviates randomly (sometimes in a direction, sometimes in another direction) the summation will "cancel out"
- Note: a **positive** covariance means X and Y deviates in the **same direction** (directly proportional), a **negative** covariance means they deviates in **opposite directions** (inversely proportional)

Covariance limit

- A problem with covariance is that its value is affected by the **unit of measure**:
  - If values are large, the covariance tends to be large (even if X and Y are not much related)
  - If values are small, the covariance tends to be small (even if X and Y are strongly related)
- For example, given **the same set** of prices, we can **decrease** the covariance by a factor 1000, by simply expressing the price as **thousands of $**!

Correlation

- In order to overcome the problem of the unit measure, we use the **correlation**.
- The correlation solve this problem producing a result which is independent from unit measure, because it takes into account the standard deviations of X and Y:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Stdev}(X) \times \mathrm{Stdev}(Y)}$$

- Dividing the covariance by the product of the two standard deviations we ensure a value between **-1** and **1**.

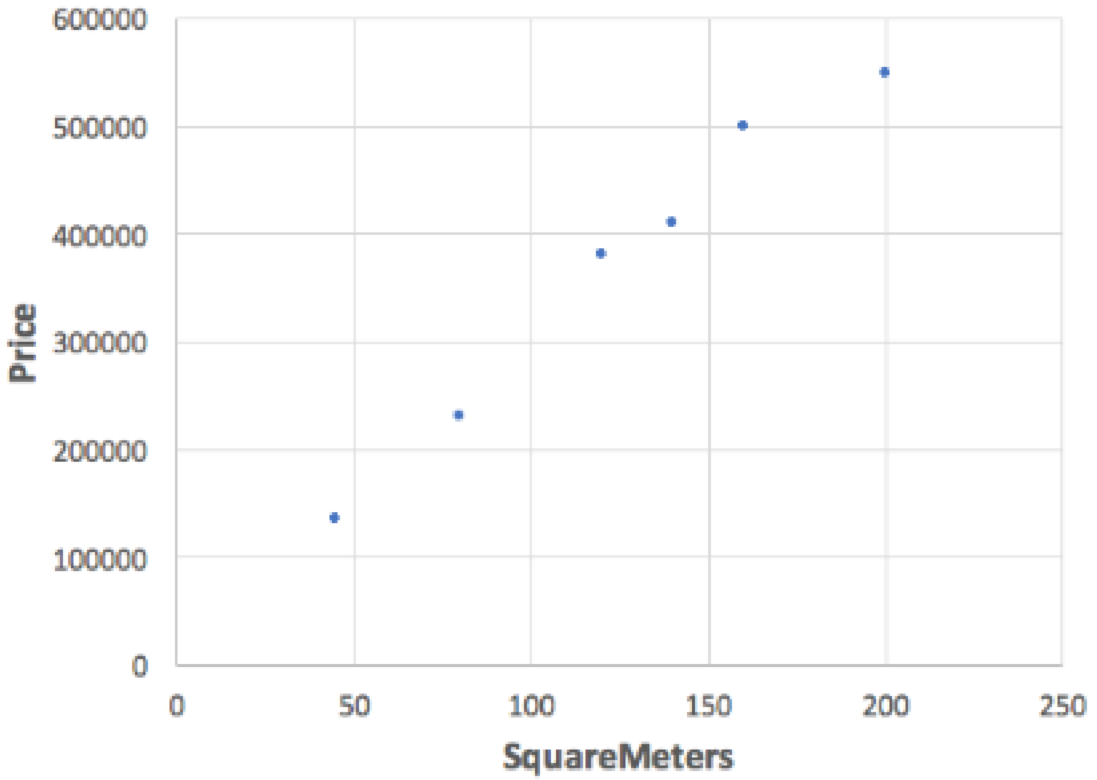
Meaning of correlation

- A value of correlation is close to -1 if the two variables tend to vary in opposite direction (inversely proportional)
- A value of correlation is close to 1 if the two variables tend to vary in the same direction (directly proportional)
- A value of correlation is close to 0 if the two variables have independent variations (at least for **linear** relations!)

Scatterplot and regression

- Apart from computing association measures, it is useful to visualize the pairs of values from the two variables in an XY plane:

SquareMeters	Price
120	380,000
200	550,000
80	230,000
160	500,000
45	135,000
140	410,000



- While association measures tell us **if** an association exists (correlation here is 0.99!!), **regression** models estimate the **actual function** that relates the two variable: in this case

$$\textbf{Price}(\textbf{SquareMeters}) = 3000 * \textbf{SquareMeters}$$

- But hold on! We will/have see regression models, in the machine learning section.

Association measures: summary

- Looking at two different variables at the same time help us understand if there is any relation between the two:
  - Covariance tell us if and how much X and Y vary accordingly
  - Correlation, in addition, it's not affected by the variables unit measures.

- **Warning**: this association measures works only if the relation is **linear** (i.e. the points form a straight line in the XY plane)
  - Correlation can be 0 even if exists a strong but **non-linear** relation between X and Y.

References

Jackie Nicholas

Introduction to Descriptive Statistics

Mathematics Learning Centre
University of Sydney
2010