DATA ANALYSIS

BASIC ANALYSIS METHODS

CONTINGENCY/PIVOT TABLES

Contingency table: examines the relationship between two categorical variables via their relative (cross-tabulation).

Pivot table: a table generated by applying operations (sum, count, mean, etc.) to variables, possibly based on another (categorical) variable.

Contingency tables are special cases of pivot tables.

	Large	Medium	Small
Window	1	32	31
Door	14	11	0

Type	Count	Signal avg	Signal stdev
Blue	4	4.04	0.98
Green	1	4.93	N.A.
Orange	4	5.37	1.60

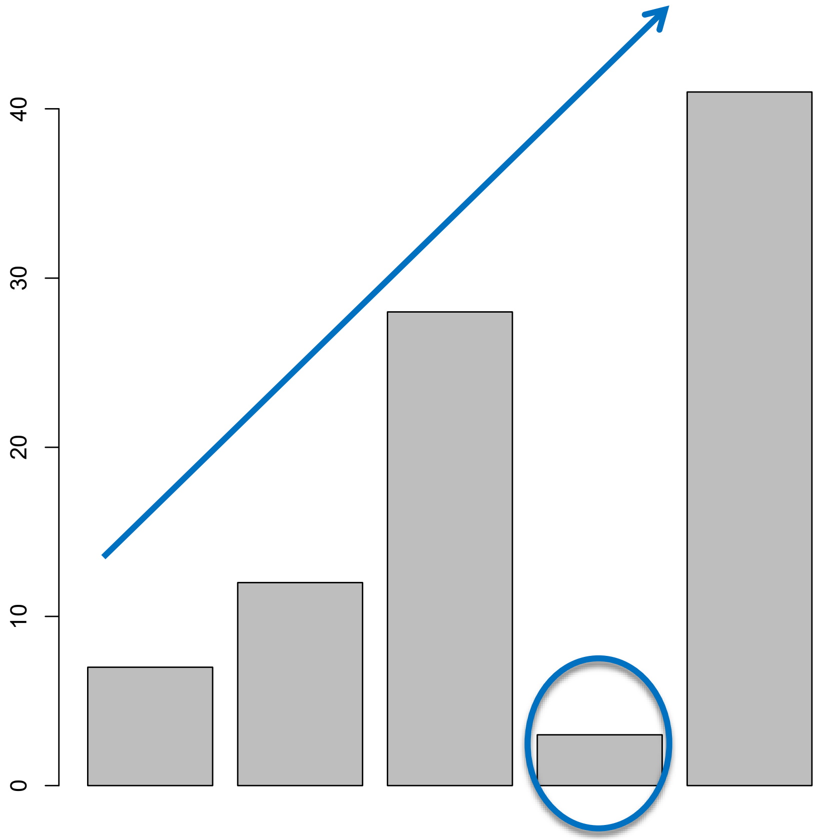
ANALYSIS THROUGH VISUALIZATION

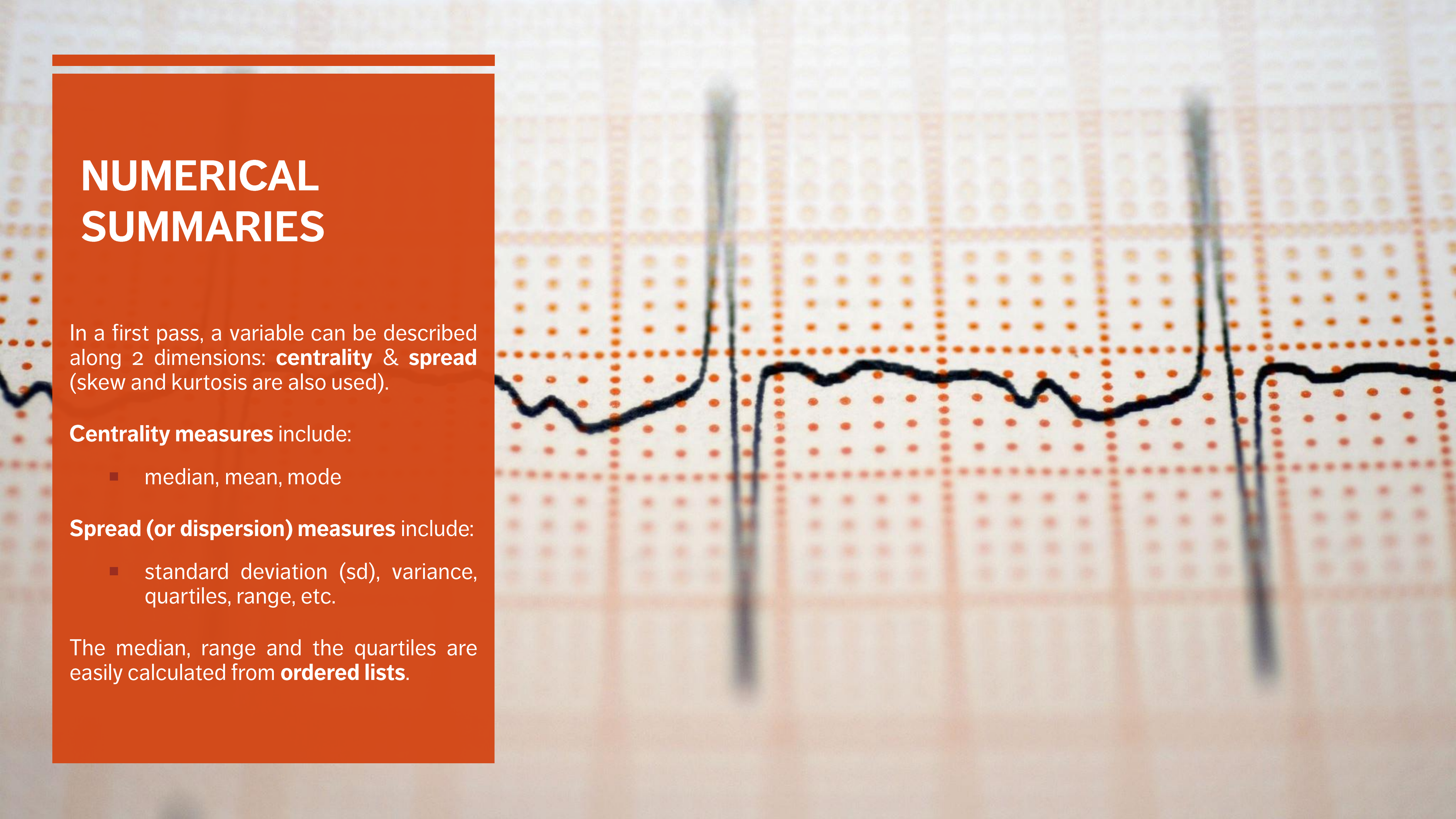
Analysis (broad definition):

- identifying patterns or structure
- adding meaning to these patterns or structure by interpreting them in the context of the system.

Option 1: use analytical methods to achieve this.

Option 2: visualize the data and use the brain's analytic power (perceptual) to reach meaningful conclusions about these patterns.





NUMERICAL SUMMARIES

In a first pass, a variable can be described along 2 dimensions: **centrality** & **spread** (skew and kurtosis are also used).

Centrality measures include:

- median, mean, mode

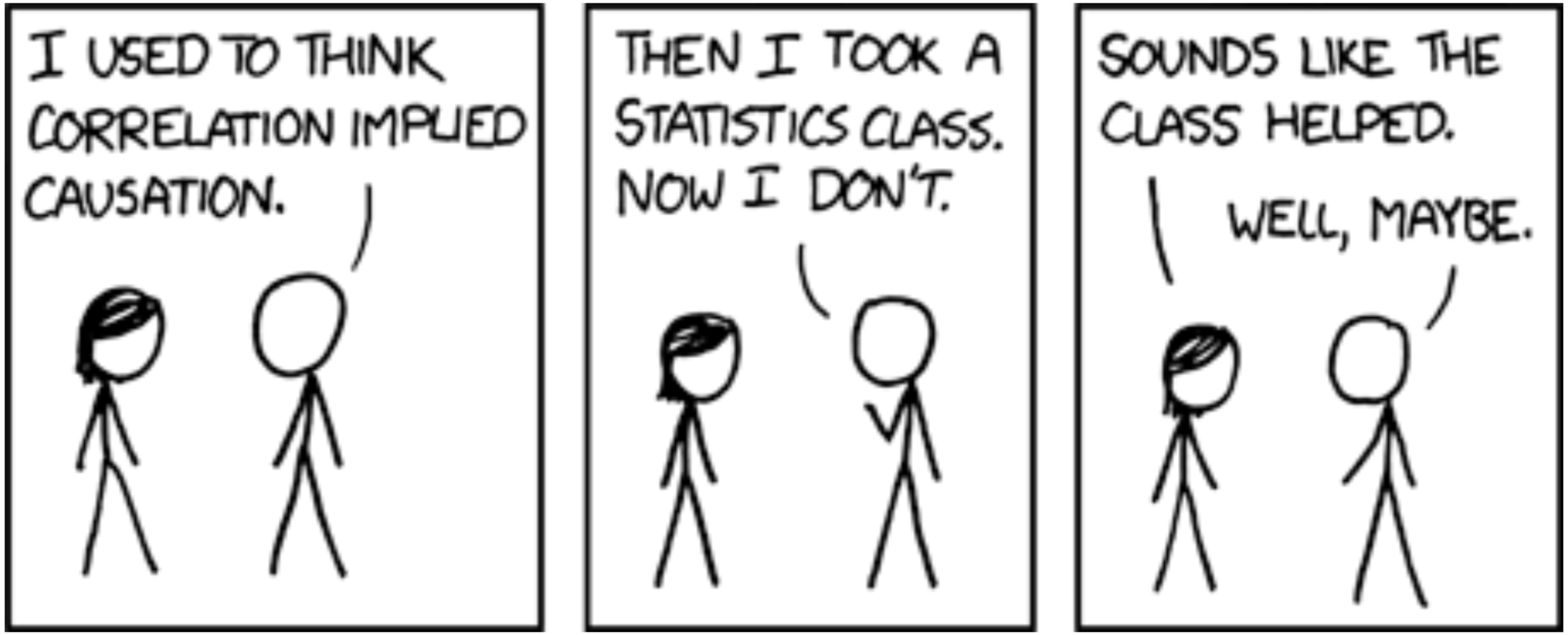
Spread (or dispersion) measures include:

- standard deviation (sd), variance, quartiles, range, etc.

The median, range and the quartiles are easily calculated from **ordered lists**.

[https://xkcd.com/552]

CORRELATION


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

LINEAR REGRESSION

The basic assumption of **linear regression** is that the dependent variable y can be approximated by a linear combination of the independent variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

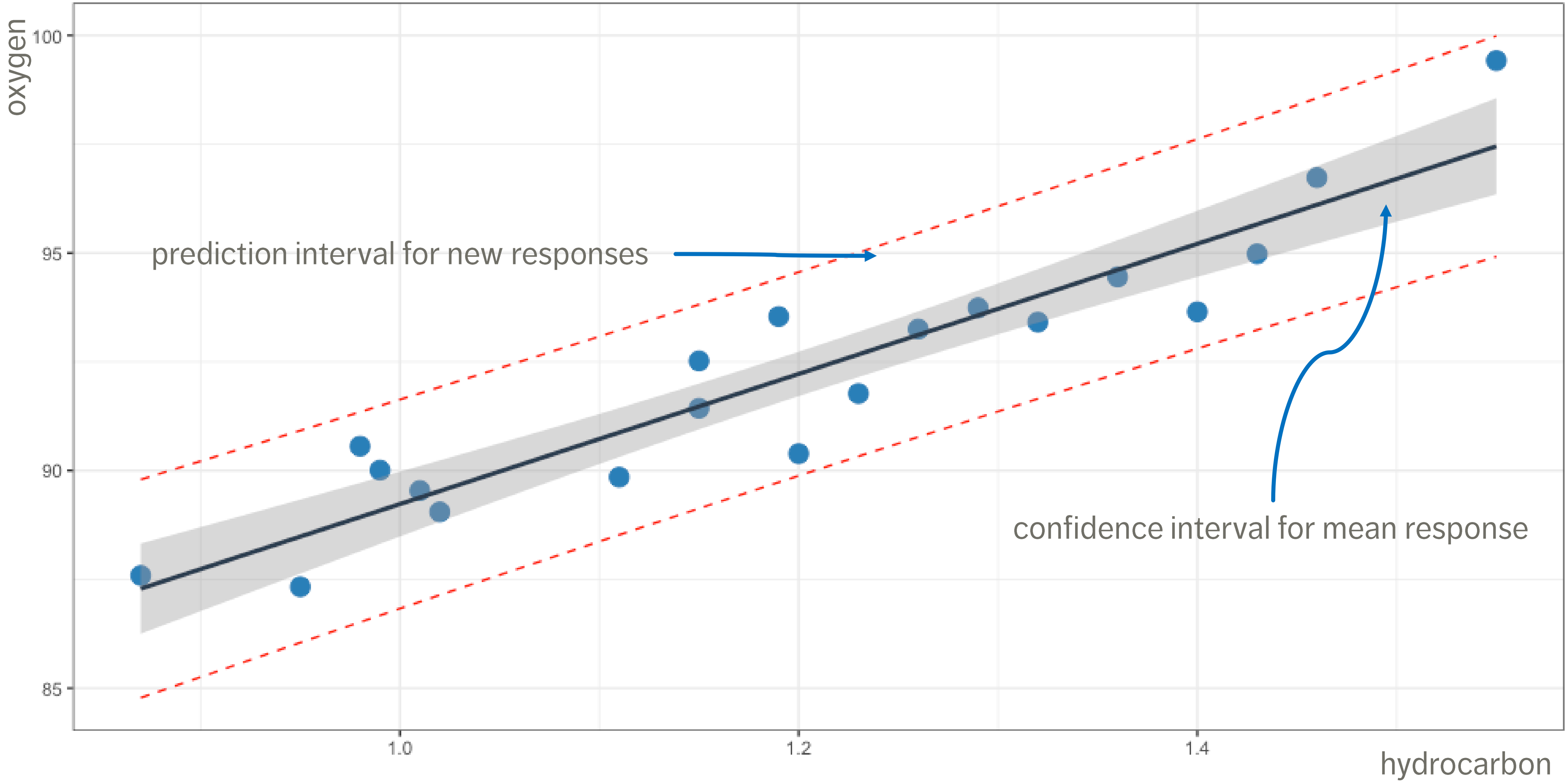
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be determined based on the **training set**, and for which

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \qquad E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\middle|\mathbf{X}\right) = \sigma^2\mathbf{I}.$$

Typically, the errors are also assumed to be **normally distributed**:

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

$$\text{oxygen} = 14.95 \times \text{hydrocarbon} + 74.28$$



prediction interval for new responses

confidence interval for mean response

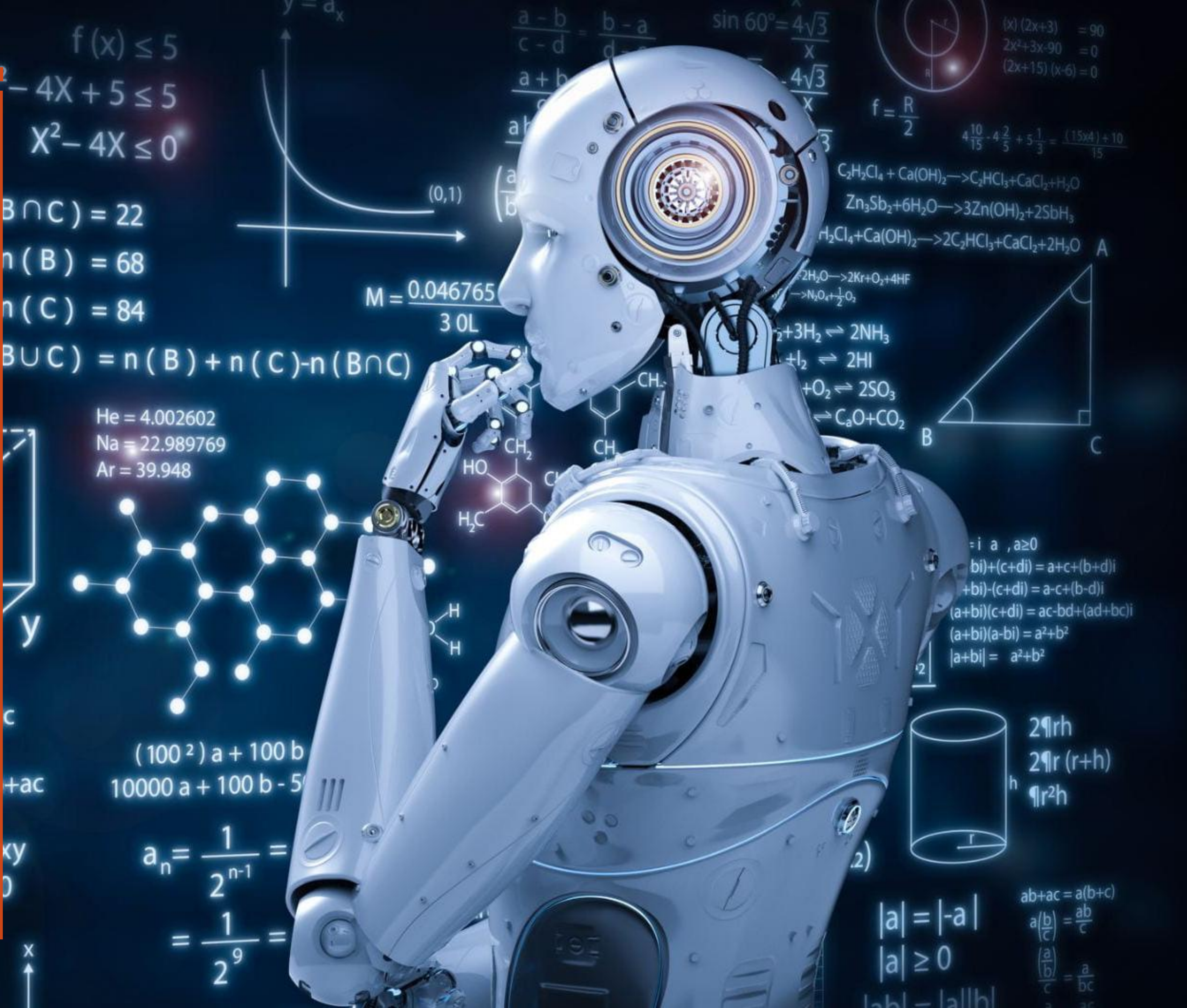
MACHINE LEARNING TASKS

Classification, class probability estimation: which clients are likely to be repeat customers?

Clustering: do customers form natural groups?

Association rule discovery: what books are commonly purchased together?

Others: **value estimation** (how much is a client likely to spend in a restaurant); **profiling and behaviour description**; **link prediction**; **data reduction**; **influence/ causal modeling**; **similarity matching** (which prospective clients are similar to a company's best clients?), etc.



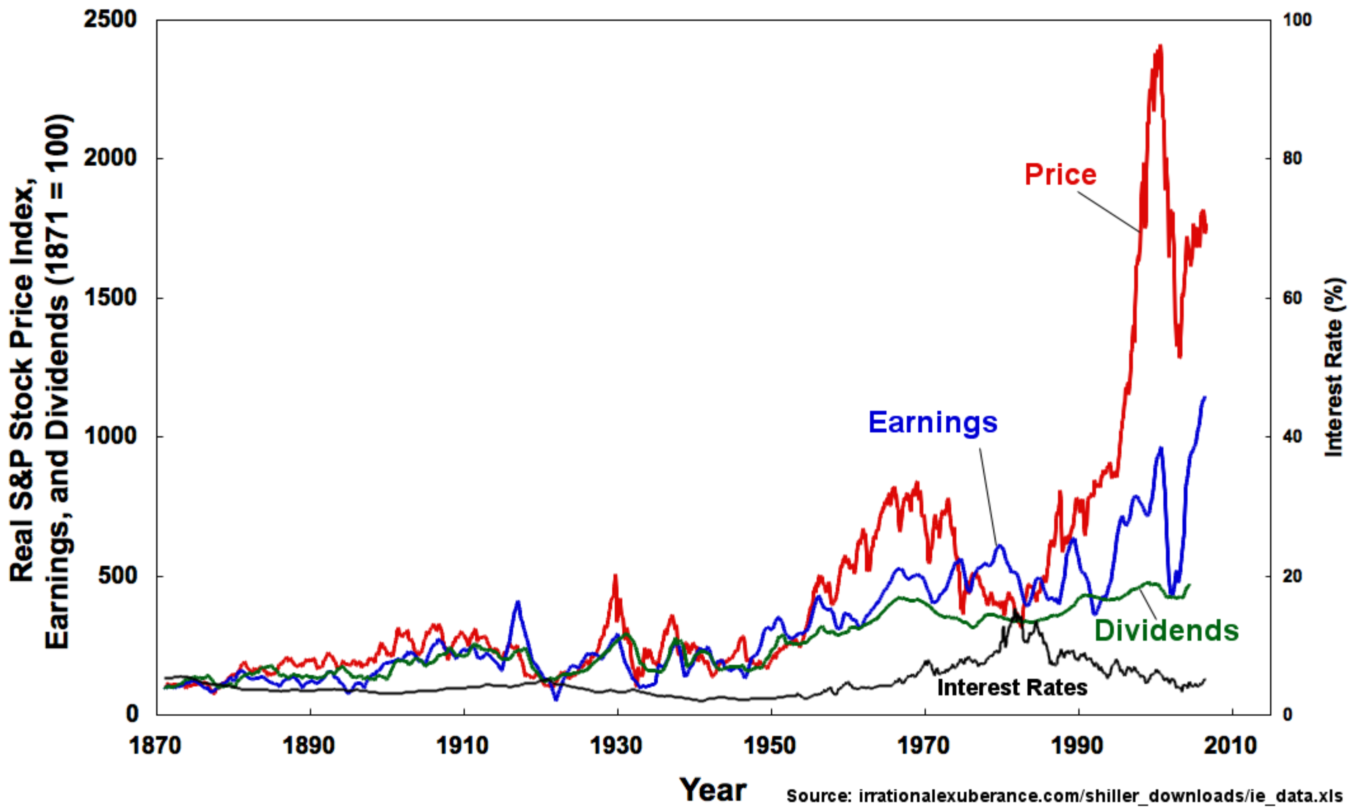
TIME SERIES ANALYSIS

A simple **time series:**

- has two variables: time + 2nd variable
- the second variable is *sequential*

What is the **pattern of behaviour** of this second variable over time? Relative to other variables?

Can we use this to **forecast the future behaviour** of the variable?



Source: irrationalexuberance.com/shiller_downloads/ie_data.xls

data-action-lab.com

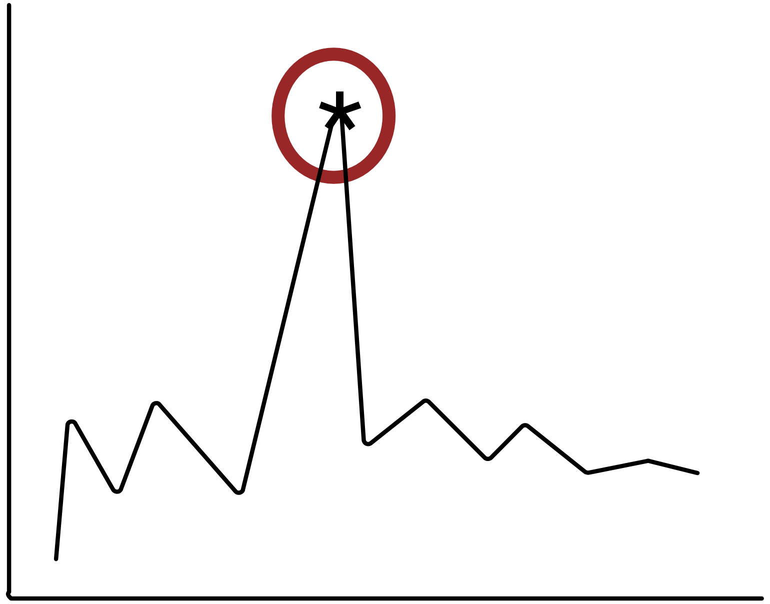
ANOMALY DETECTION

Anomaly: an unexpected, unusual, atypical or statistically unlikely event

Wouldn't it be nice to have a data analysis pipeline that alerted you when things were out of the ordinary?

Many different analytic approaches to take!

- clustering
- classification
- ensemble techniques, etc.



DATA ANALYSIS

BASIC ANALYSIS METHODS