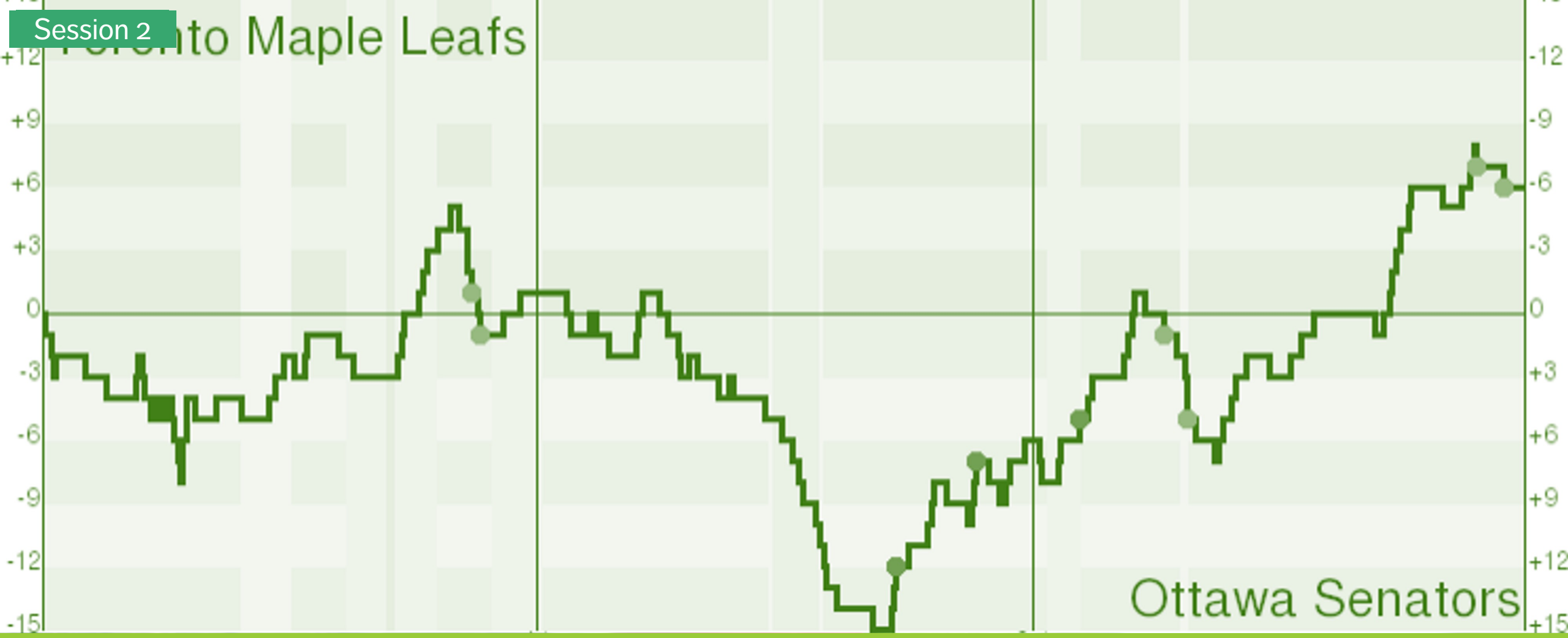
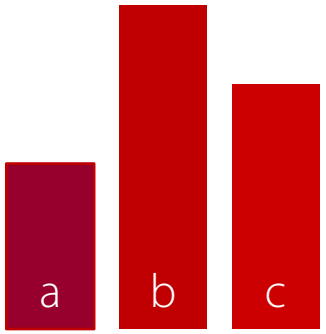


6. Getting Insight From Data

Analytics Modes

Descriptive



Show **what** happened

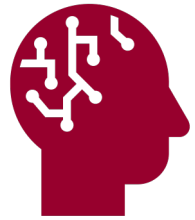
Diagnostic

Explain **why** something happened

Predictive

Guess **what will** happen

Prescriptive



Suggest **what should** happen

Low Value
Low Difficulty

High Value
High Difficulty

Asking the Right Questions

Data science is about asking and answering questions:

- **Analytics:** "How many clicks did this link get?"
- **Data Science:** "Based on this user's previous purchasing history, can I predict what links they will click on the next time they access the site?"

Data mining/science models are usually **predictive** (not **explanatory**): they show connections, but don't reveal why these exist.

Warning: not every situation calls for data science, artificial intelligence, machine learning, statistics, or analytics.

The Wrong Questions

Too often, analysts are asking the **wrong questions:**

- questions that are **too broad** or **too narrow**
- questions that **no amount of data could ever answer**
- questions for which **data cannot reasonably be obtained**

The **best-case scenario** is that stakeholders will recognize the answers as irrelevant.

The **worst-case scenario** is that they will erroneously implement policies or make decisions based on answers that have not been identified as misleading or useless.

Roadmap to Framing Questions

Understand the problem (opportunity vs problem)

What initial assumptions do I have about the situation?

How will the results be used?

What are the risks and/or benefits of answering this question?

What stakeholder questions might arise based on the answer(s)?

Do I have access to the data necessary to answering this question?

How will I measure my ‘success’ criteria?

Yes/No Trap

Examples of **bad** questions:

- Are our revenues **increasing** over time? **Has it** increased year-over-year?
- Are most of our customers from **this demographic**?
- **Does this project have** valuable ambitions to the broader department?
- **How great** is our hard-working customer success team?
- How often do you **triple check** your work?

Examples of **good** questions:

- What's the **distribution** of our revenues over the past three months?
- Where are our **top 5** high-spending cohorts from?
- What are the **different benefits** of pursuing this project?
- What are **three good** *and* **bad traits** of our customer success team?
- Do you **tend to** do quality assurance testing on your deliverables?

Question Audit Checklist

1. Did I avoid creating any yes/no questions?
2. Would anyone in my team/department understand the question irrespective of their backgrounds?
3. Does the question need more than one sentence to express?
4. Is the question 'balanced' - scope is not too broad that the question will never truly be answered, or too small that the resulting impact is minimal?
5. Is the question being skewed to what may be easier to answer for my/my team's particular skillset(s)?

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

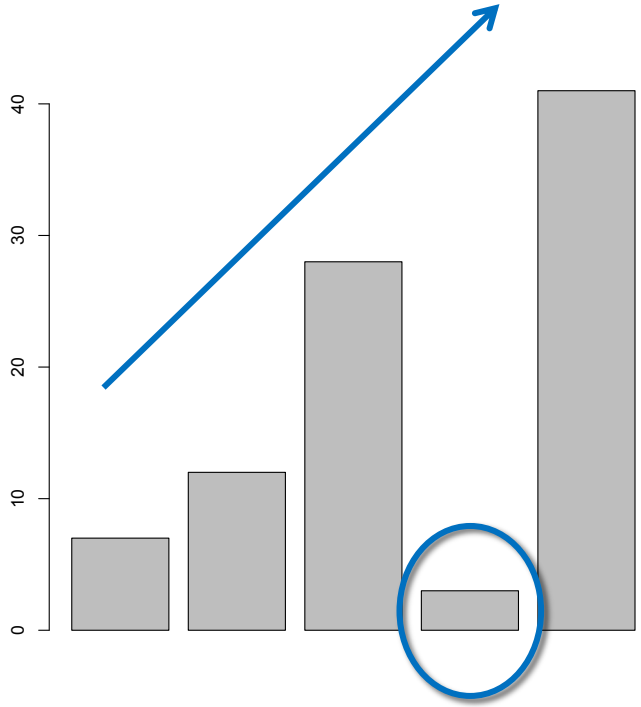
Analysis Through Visualization

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 sometimes).

Centrality measures include:

- median, mean, mode (less frequently)

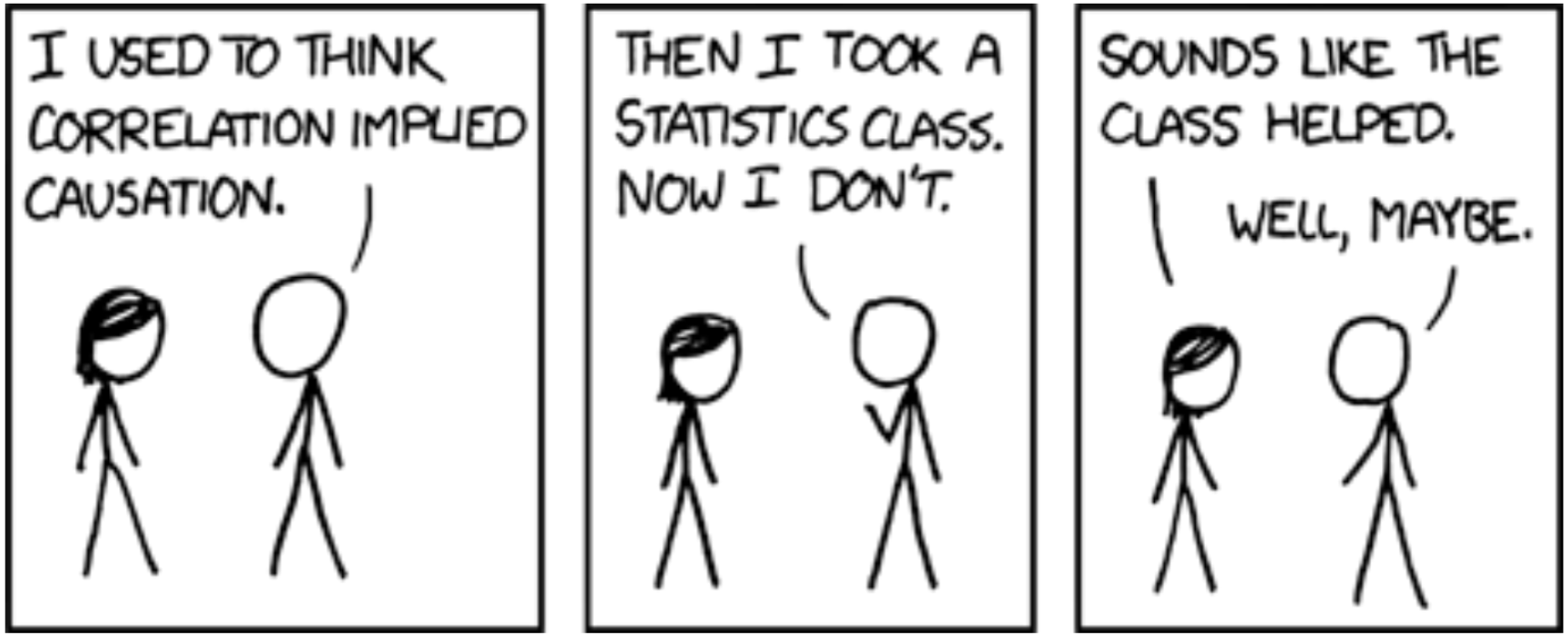
Spread (or dispersion) measures include:

- standard deviation (sd), variance, quartiles, inter-quartile range (IQR), range (less frequently)

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

[https://xkcd.com/552]

Correlation



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:

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

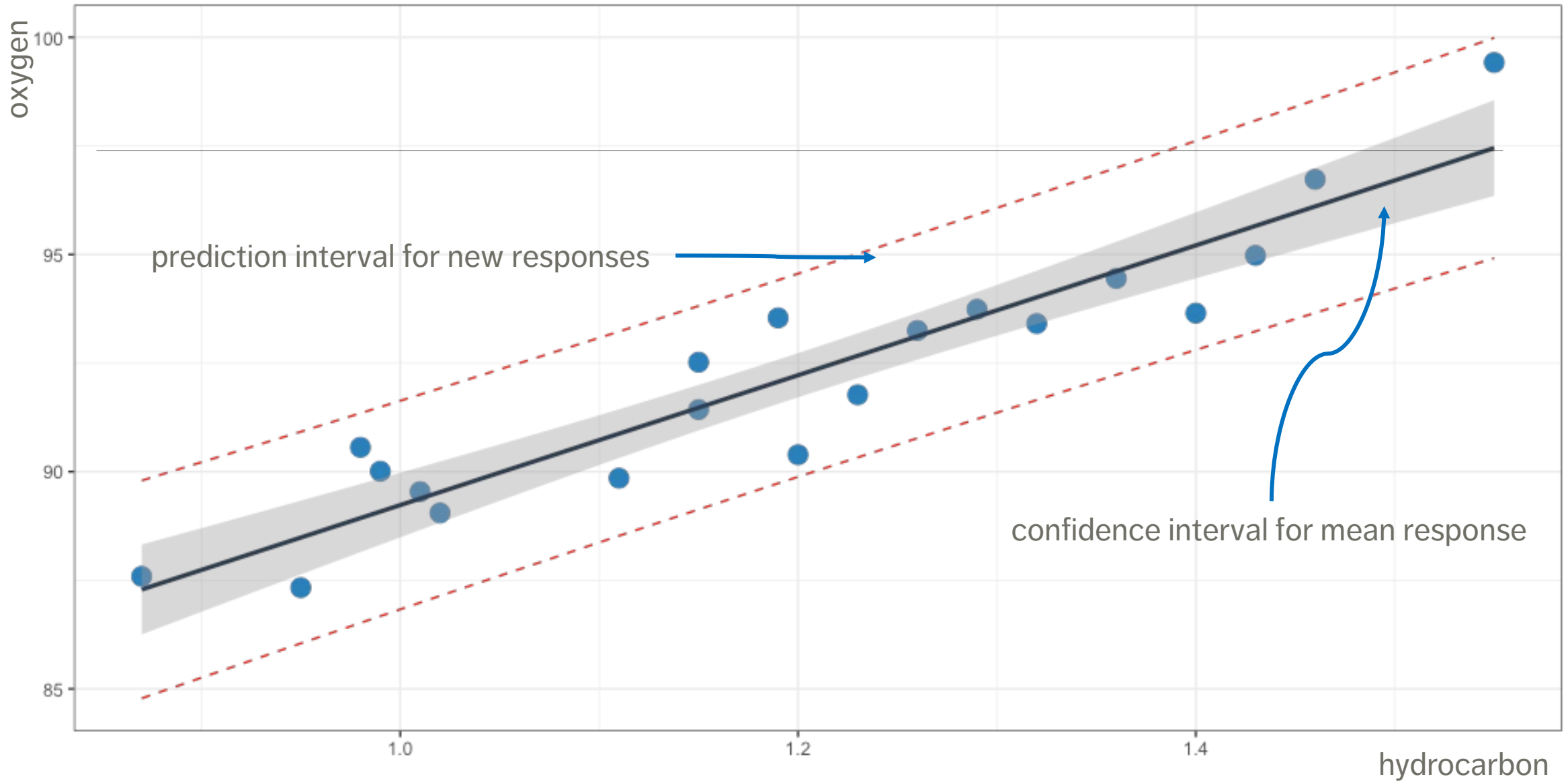
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be determined based on the **training set**, and for which

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

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

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

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



Machine Learning Tasks

Classification and **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:

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

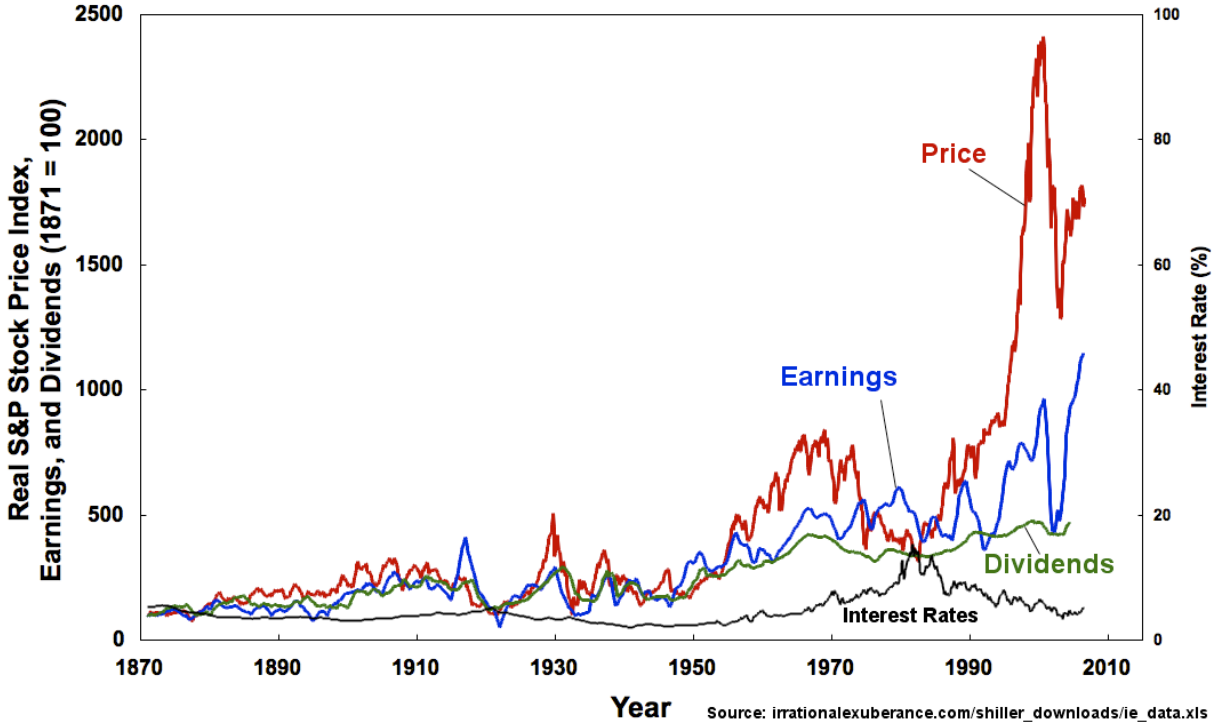
Time Series Analysis

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 ?



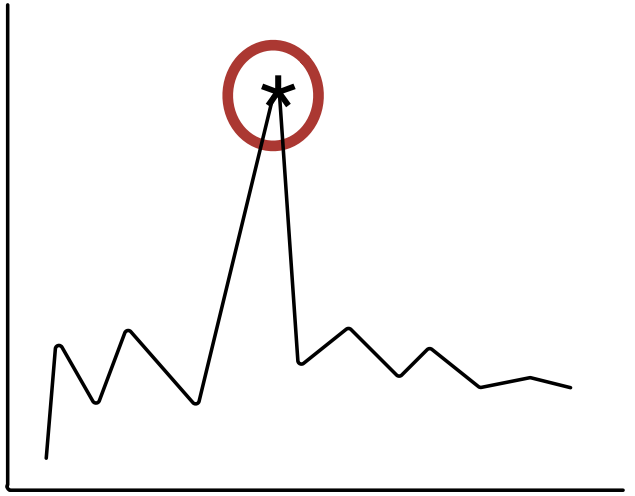
Anomaly Detection

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.



Suggested Reading

Getting Insight From Data

Data Understanding, Data Analysis, Data Science
Data Science Basics

Getting Insight From Data

- Asking the Right Questions
- Basic Data Analysis Techniques
- Common Statistical Procedures in R
- Quantitative Methods

***Probability and Applications** (advanced)

***Introductory Statistical Analysis** (advanced)

***Survey Sampling** (advanced)

***Regression Analysis** (coming soon)

Exercises

Getting Insight From Data

1. Do the exercise in Asking the Right Questions.

2. Recreate the examples of Common Statistical Procedures in R.

3. The file cities.txt contains population information about a country's cities. A city is classified as "small" if its population is below 75K, as "medium" if it falls between 75K and 1M, and as "large" otherwise. Locate and load the file into the workspace of your choice. How many cities are there? How many are there in each group? Display summary population statistics for the cities, both overall and by group.