Mechatronic Engineering program

Python for machine learning and data science:
2: Scientific method, exploratory data analysis

Ziemowit Dworakowski

AGH University of Krakow

Example no 1



1) A => B (A causes B)
2) B => A (B causes A)
3) C => B and C => A (something causes A and B)
4) A ! => B and B !=> A (coincidence)



Example no 1

1) A => B (A causes B)
2) B => A (B causes A)
3) C => B and C => A (something causes A and B)
4) **A ! => B and B !=> A** **(coincidence)**

Maybe gather more data
and observe whether the relations is maintained?

Example no 1

1) A => B (A causes B)
2) B => A (B causes A)
3) **C => B and C => A** **(something causes A and B)**

Maybe calculate more features and check for other correlated features?

Maybe try to artificially cause A or B and see if the other follows?

Example no 1

1) A => B (A causes B)
2) **B => A** **(B causes A)**

Maybe look for time-dependent relations? If something was observed first, it might have been the root cause

Maybe look for physical model to infer causality based on expert knowledge?

Scientific method (basic route):

Observation → Formulation of a possible cause (hypothesis) → Try to falsify/confirm the hypothesis

- Make predictions based on hypothesis, check if they are fulfilled



- Think what evidence would render the hypothesis untrue, look for this evidence



- Look for potential biases that may have influenced the reasoning



- Check to which extent (under which conditions) the hypothesis holds



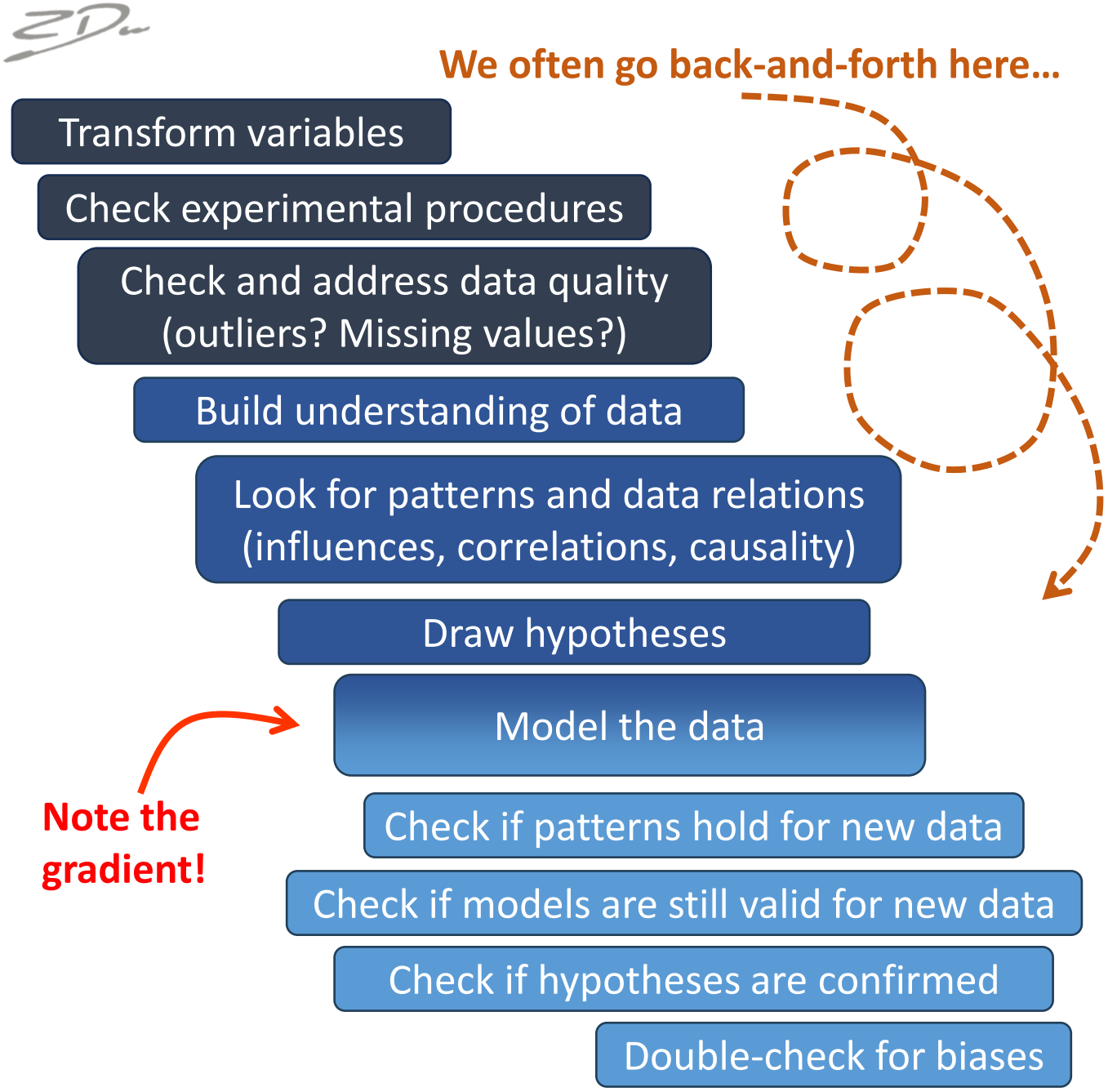
New observations!

This concept works in most of the typical scenarios:

We observe relations between features → We think one can be predicted based on the other → We build regression system and test it on **independent data subset** → We check if the results are statistically significant

We see that income in one group of occupations tends to be higher → We gather more data and check how likely it is that this relation was pure coincidence

We see which damage cases were recognized by a damage classifier and claim it can regognize up to 90% of cases → We gather more examples of different damage cases → We check if our hypothesis holds



Step 0: Initial research question or goal (optional)

Step 1: Initial Data Analysis

Step 2: Exploratory Data Analysis

Step 3: Confirmatory Data Analysis

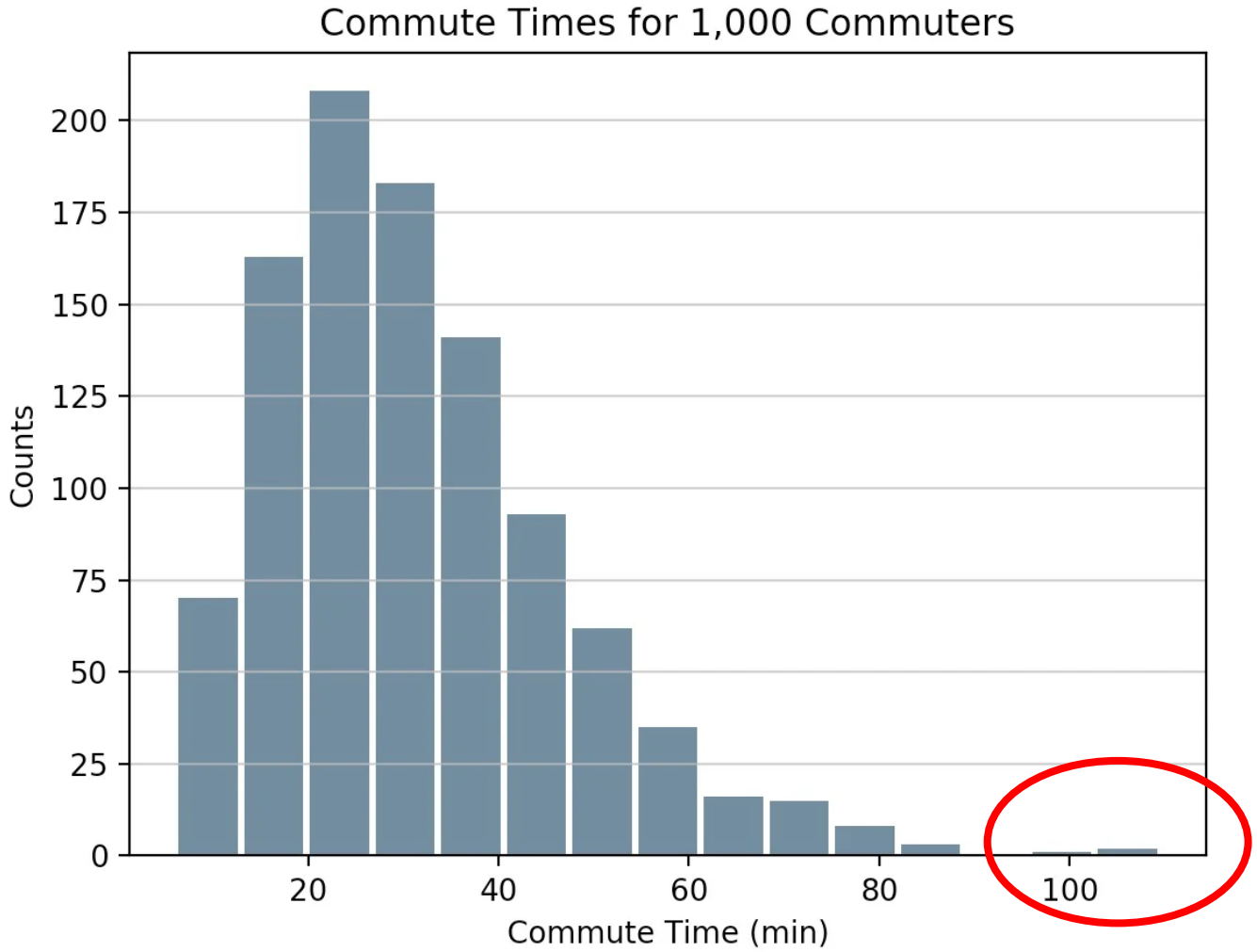
The more time we spend in the early steps, the less data we need for the later steps!

Different kinds of outliers and anomalies

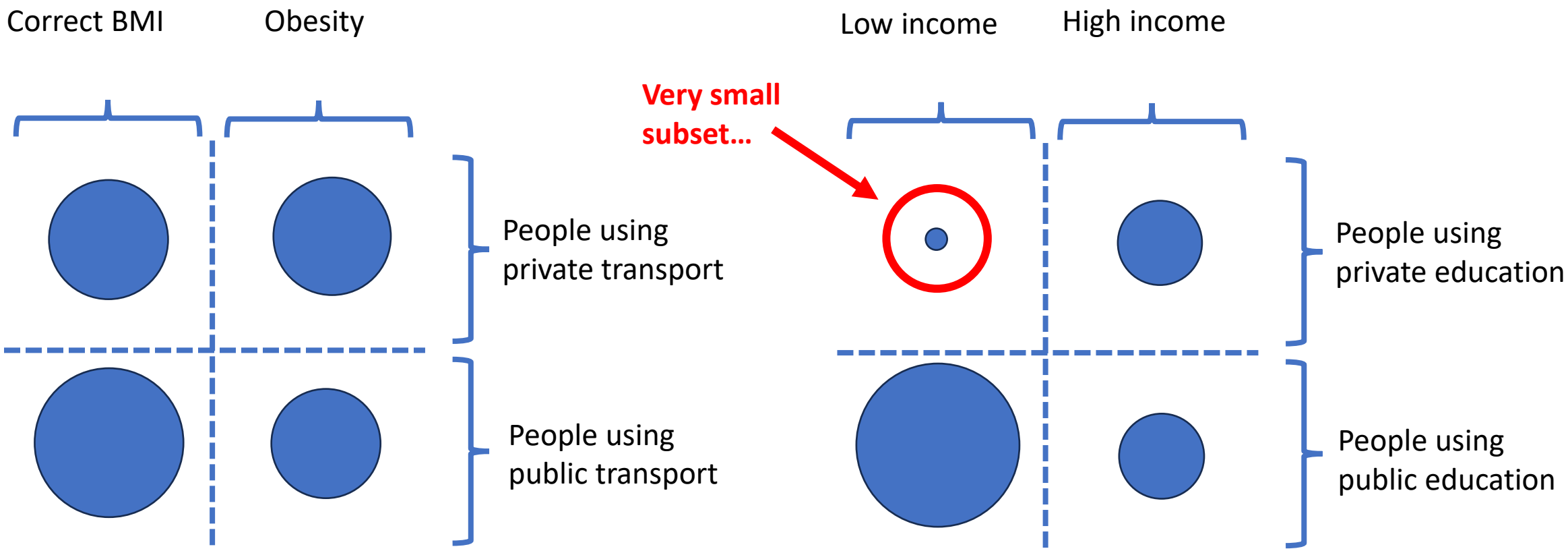
Employment area	Earnings / mo	Age	Education	Opened term deposit?
Agriculture	3900 $	34	MSc	Yes
Agriculture	2400 $	45	College	Yes
Law	5400 $	41	**Illiterate**	Yes
Transportation	3000 $	30	College	No
Science & Ed	4200 $	36	MSc	No
Food industry	5100 $	32	BSc	Yes

0.7 % of people are marked as „Illiterate”
Some of them are definitely educated.
What to do with such data?

Different kinds of outliers and anomalies



Different kinds of outliers and anomalies



Outliers can be found also in features' relations to each other. Each feature value, when considered separately, can be perfectly typical, but their combination can form an anomaly

Outliers and anomalies:

Measurement errors
„We didn't measure correctly"

Labeling errors
„We didn't understood what we measured"

Context shifts
„We measure correctly, we label correctly – but the meaning of the label or data is different"

Rare but valid cases
„We measured and labeled correctly – it only feels weird"

E.g.: *We measure level of vibration of a machine and its correlation with damage. The damage is assessed and logged differently by different operators*

E.g.: *We measure level of vibration as a function of speed and load. 99.9% of measurements are taken under steady state. 0.1% are taken in run-up or run-down*

Outliers and anomalies: What to do with them?

Can we provide consistent procedure for outlier detection?

Can we maintain this consistency for future samples too?

Measurement errors
„We didn't measure correctly"

Labeling errors
„We didn't understood what we measured"

Context shifts
„We measure correctly, we label correctly – but the meaning of the label or data is different"

Rare but valid cases
„We measured and labeled correctly – it only feels weird"

If 2 x , Remove. Otherwise, Remove only from training, keep in testing

Accept and try to fill undersampled regions, if 2 x build **anomaly detector** to flag anomalies

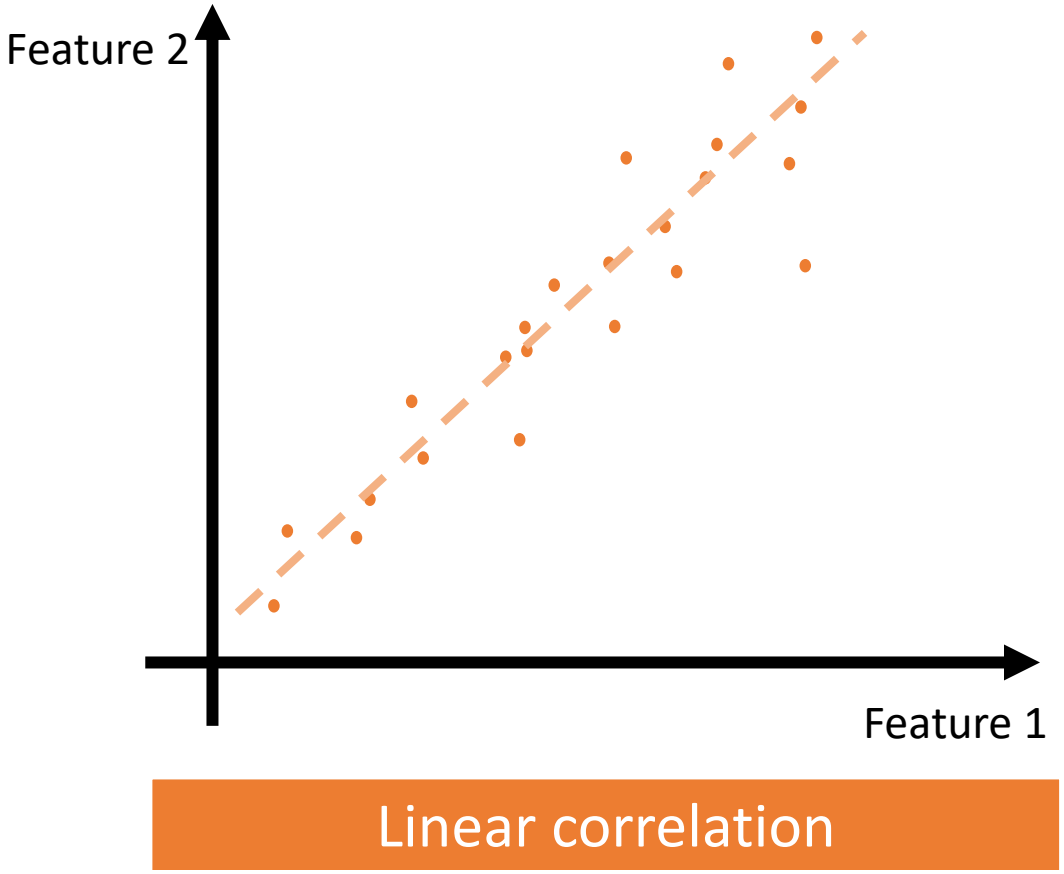
Should we remove them?

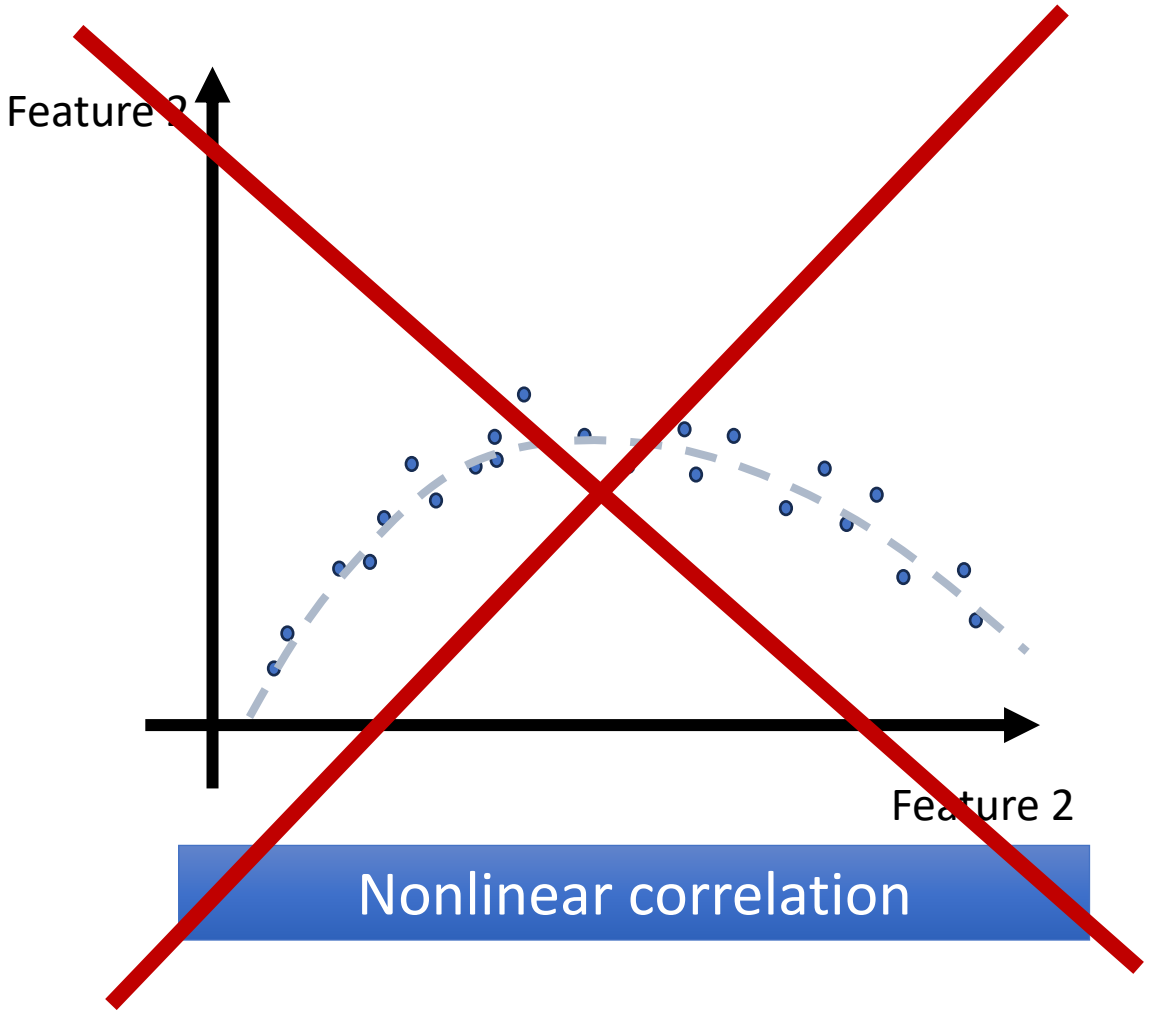
It depends:

- In general outliers caused by measurement errors or mislabeling should probably be deleted from training dataset. If we can provide a well-described methodology for deleting them, we can also do so for testing dataset. Otherwise it is better to leave them for testing – to actually measure expected final accuracy

- Anomalies that are just rare but correctly gathered examples of data (because of rarity of operational states or shifting context of measurement) are not actually outliers, they are still important for us! However, they will decrease efficiency of training because they will result in poorly-sampled regions of feature space. We can either **accept** them or **include an anomaly detector** to mark them in the future.

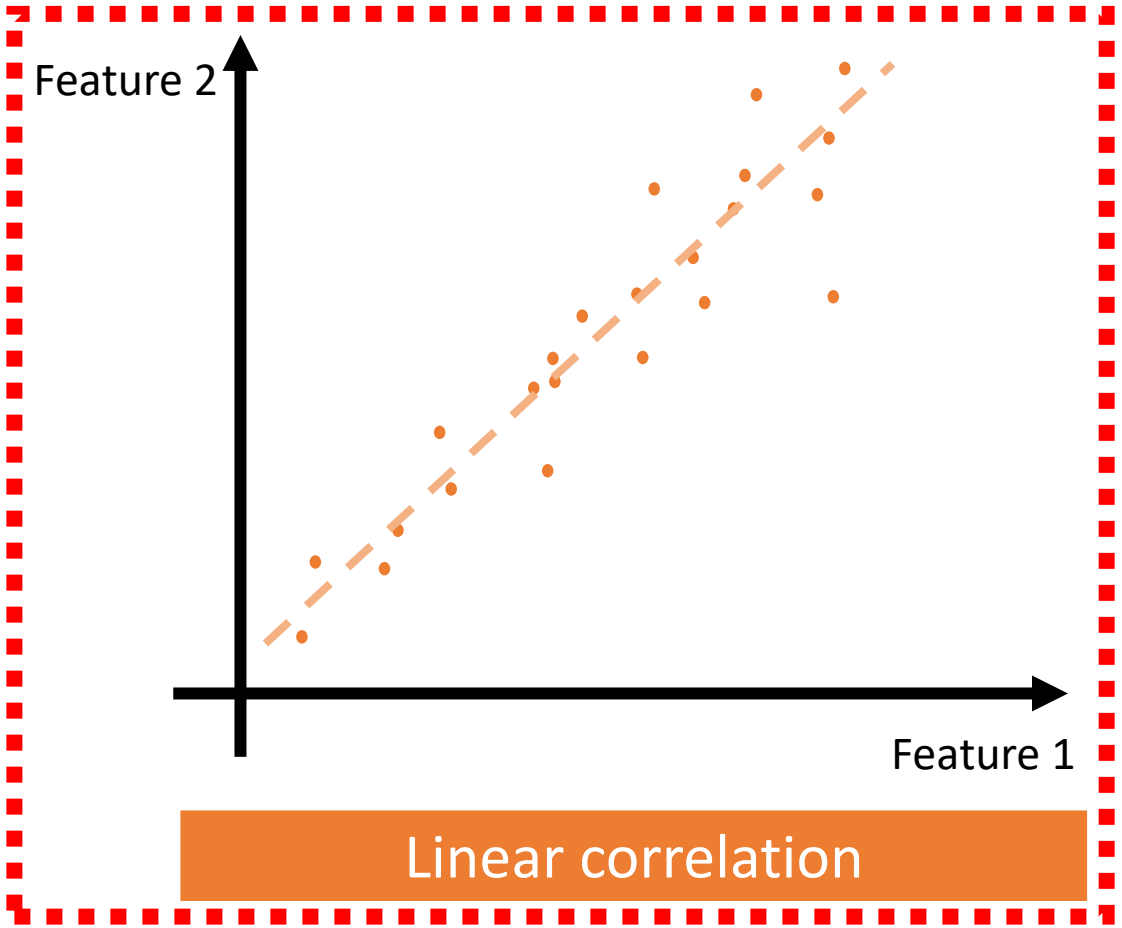
Features' correlations

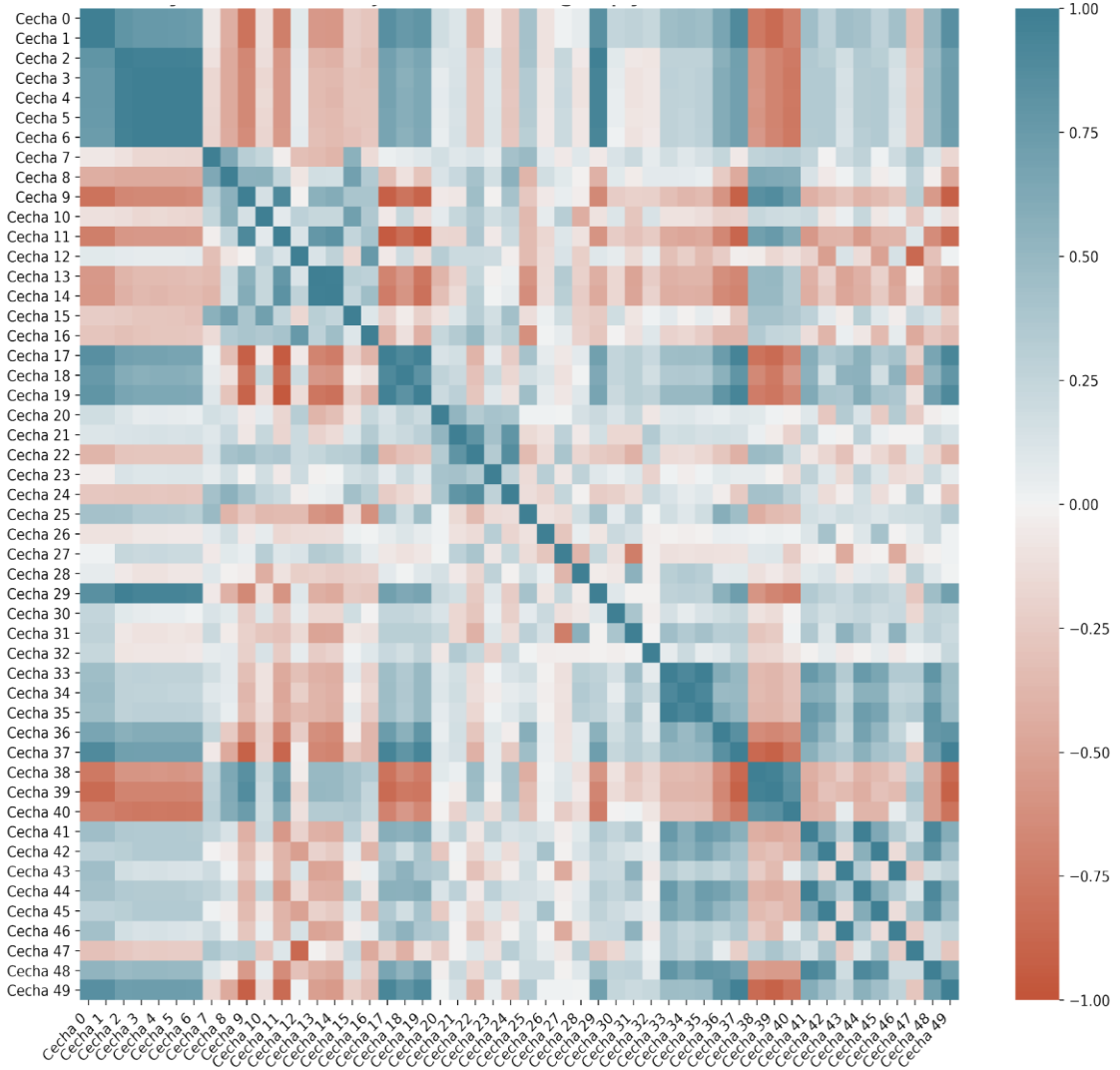




These we can't easily measure…

Features' correlations





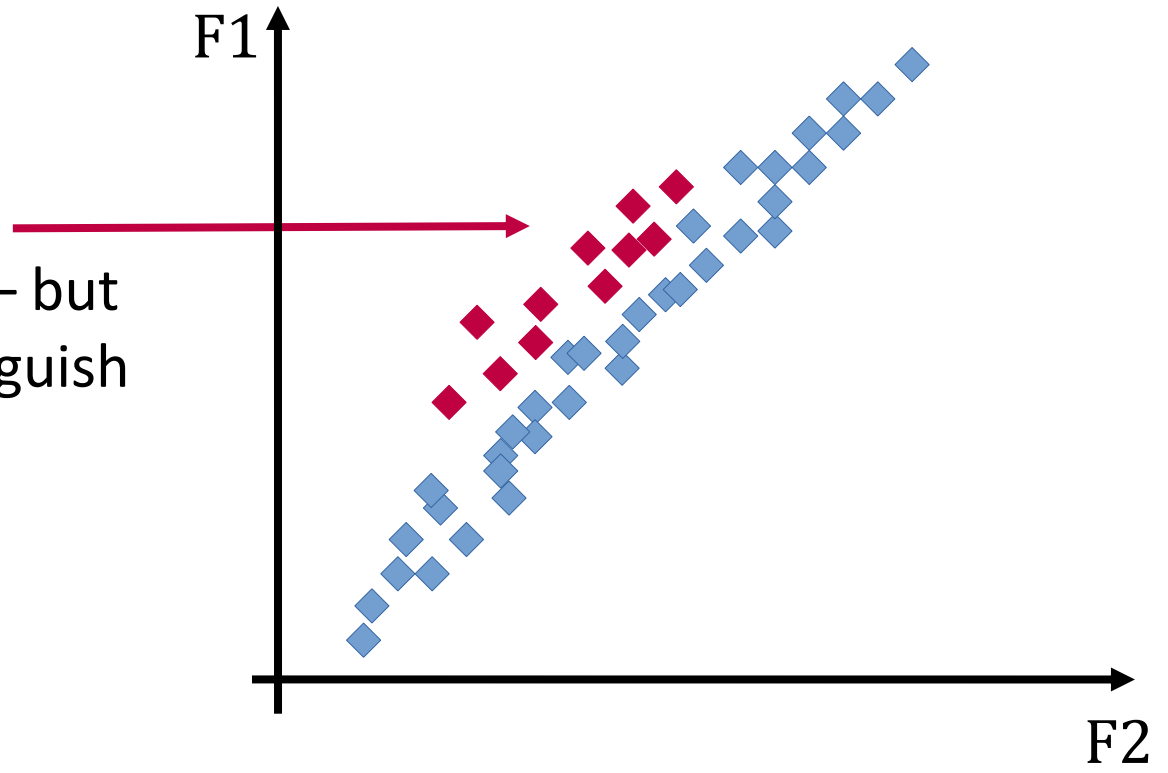
These we can easily measure…

…using Pearson correlation coefficient

Features' correlations

We want the features to **be highly correlated** with target value (in regression) but we also want the features **to not be correlated** with each other *(to reduce redundancy in our dataset).* And there is a risk in that:

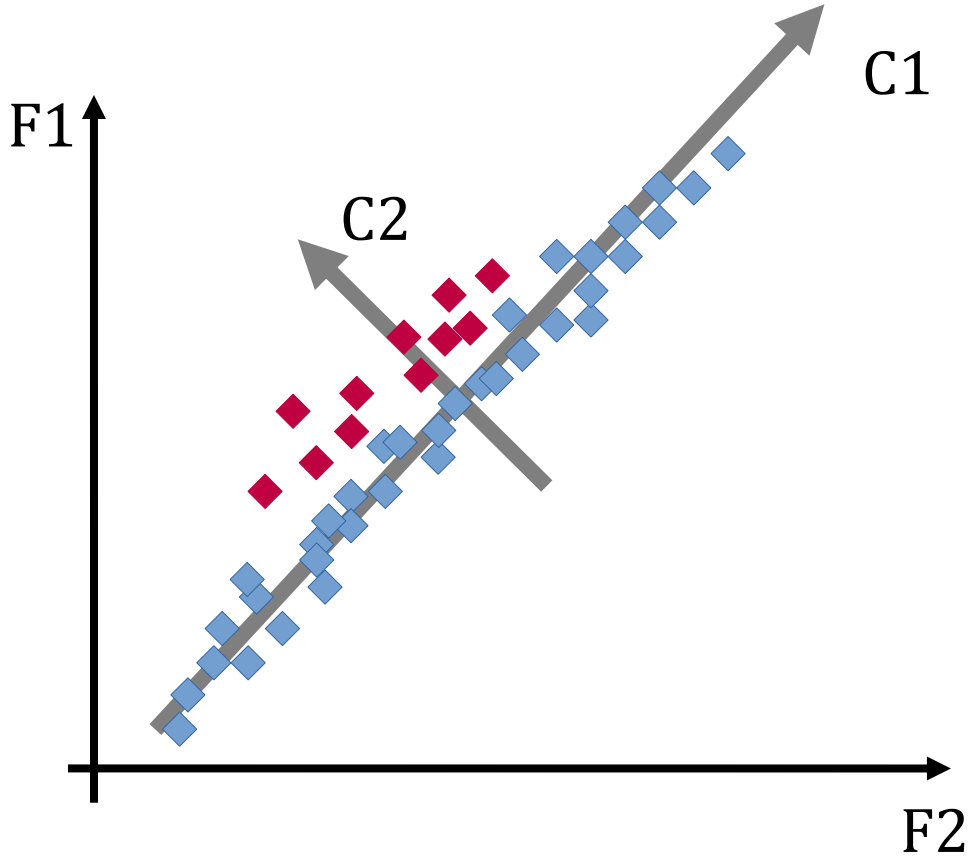
Example:
F1 is strongly correlated with F2 – but they both are necessary to distinguish classes from each other



Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

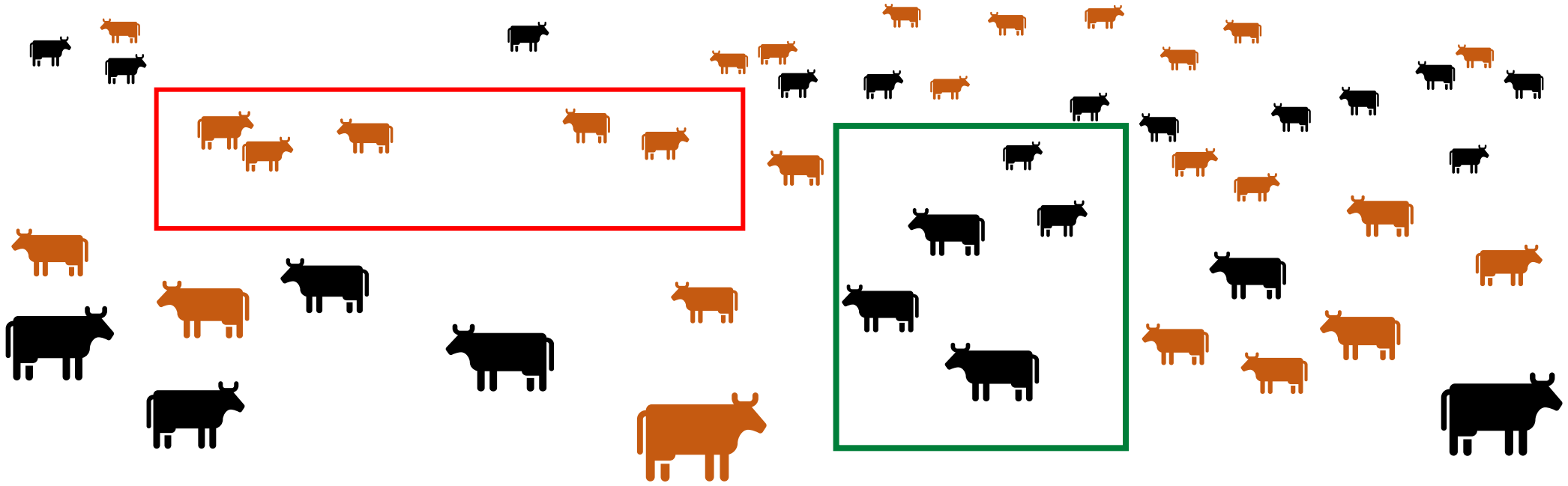
PCA maps data into new space where each dimension **(principal component)** is orthogonal to all others and maximizes remaining variance



And then we just take the first few principal components and use them for classification or regression.

Note! This still poses a risk of getting rid of important information!

Hypotheses testing



How many cows do we need to see to confirm hypothesis that they are usually brown?

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)

2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)

3) Consider experimental data (e.g.: we observed 5 brown cows)

4) Calculate how likely it is, that a test result returned 5 brown cows if the underlying probability of a brown cow is **at most** 90%
(That we obtained such results despite hypothesis being not true)

$$p = 0.9^5 = 0{,}59$$

5) If this probability (**p-value**) is lower than **0.05**, we say that the test was statistically significant (It is unlikely to get this data given false hypothesis)

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence
(while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005
5	0	More than 50% brown	50% or less are brown	0,031
4	1	More than 90% brown	90% or less are brown	0.328
			better: 80% are brown	0.409
0	5	More than 90% brown	90% or less are brown	0.00001?
			better: 0% are brown	1

$p = 0.9^5 = 0{,}590$

$p = 0.9^{50} = 0{,}005$

$p = 0.5^5 = 0{,}031$

$p = 5 \cdot (0.9^4 \cdot 0.1) = 0.328$

$p = 5 \cdot (0.8^4 \cdot 0.2) = 0.409$

~~$p < 0.1^5 = 0.00001$~~

$p < 1^5 = 1$

Hypotheses testing

Low enough p-value tells us only that data support hypothesis in a statistically significant way. It is **not** a final confirmation of the hypothesis.

It works under the assumption that the tests are independent to each other (*often not true! We can just happen to observe one pasture with cows of the same color and it will tell us nothing regarding the whole cow population*)

High p-value **does not** tell us that the hypothesis is false.

p-value **should not** be used for early stopping of the experiment or to select a subset of data to confirm a hypothesis <- this is *p-hacking*

Things to remember:

1. Explain an overview on scientific method
2. Explain steps of data analysis (IDA, EDA, CDA)
3. Explain sources for anomalies, explain how they affect model preparation
4. What does it mean that data are correlated? How do we measure correlation?
5. How does PCA work?
6. What are the risks associated with looking blindly into correlation information or using PCA?
7. How do we test hypotheses?
8. Explain what a p-value is and how is it used. Note what p-value does **not** allow.