Data, Math and Methods

Week 7, Statistics & Probability

~ Vítek Růžička

Today

Lecture:

- Probability and **modelling** from measurements
- Different random functions – through **probability density functions**
- **Statistics** – analyzing data
- **Visualizations**

Programming:

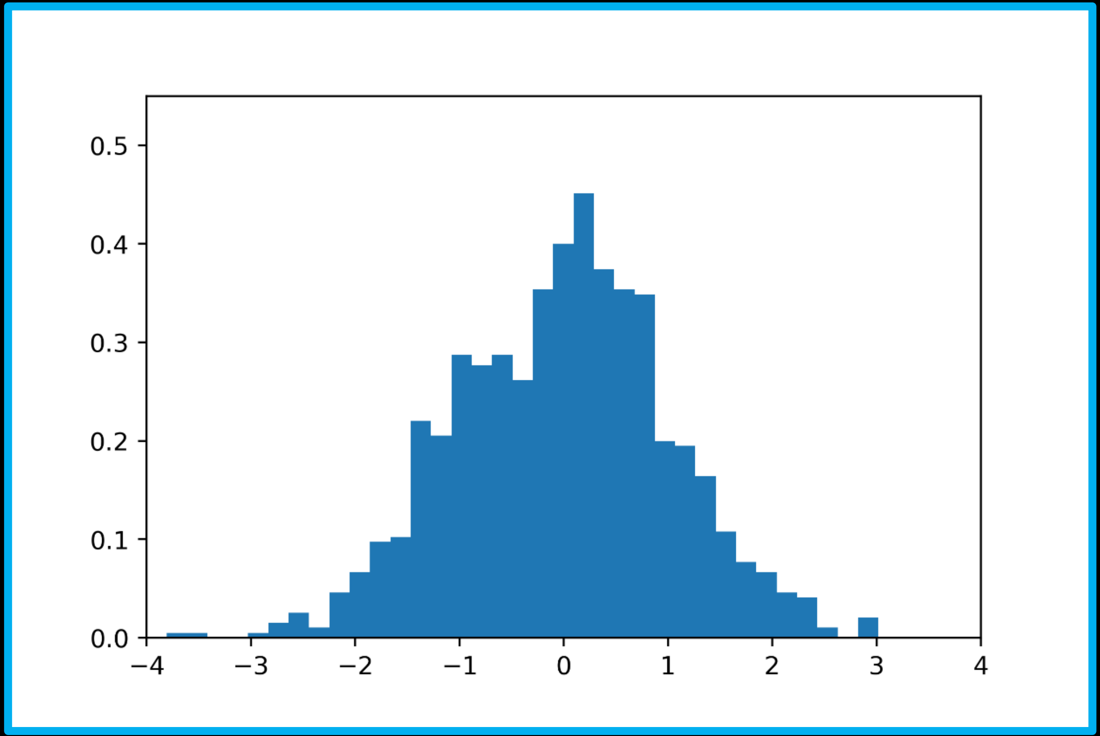
- Statistical analysis of data
- Using random function to generate textures

Probability

- What is the purpose of studying probability and statistics?

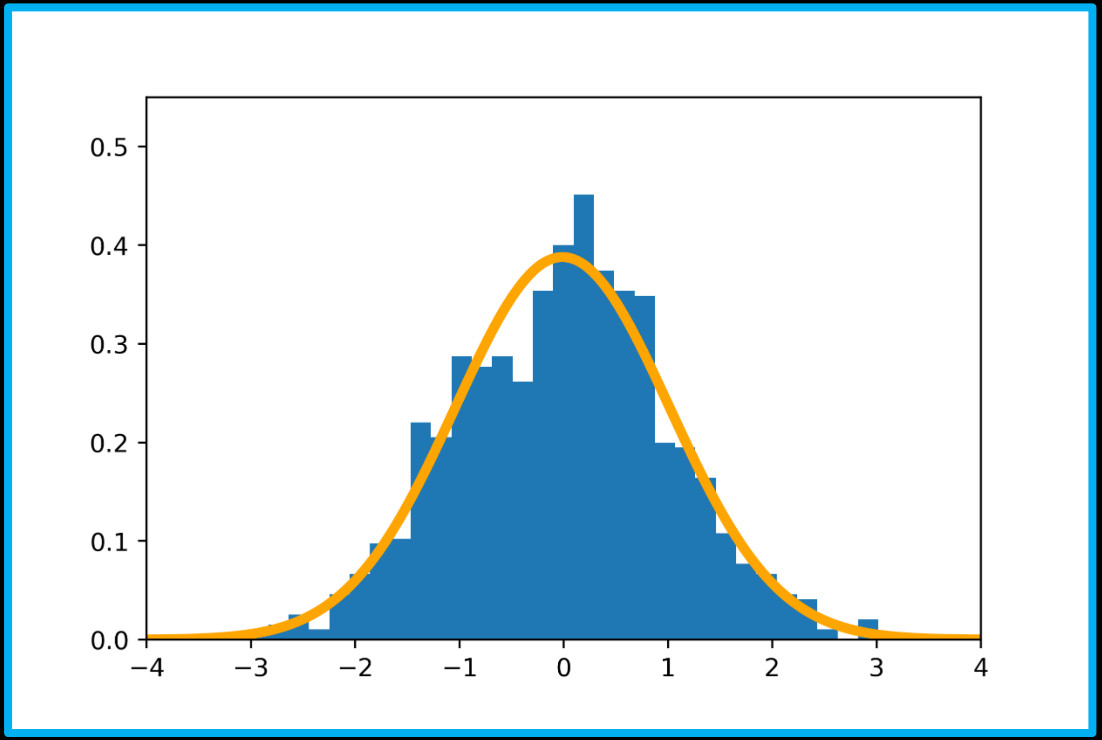
- Exploring **events** in the real world and trying to **model** them. Then when we have a model, trying to predict how they could continue. (*This is a sort of machine learning researcher's view btw :D*)

Modelling and Sampling



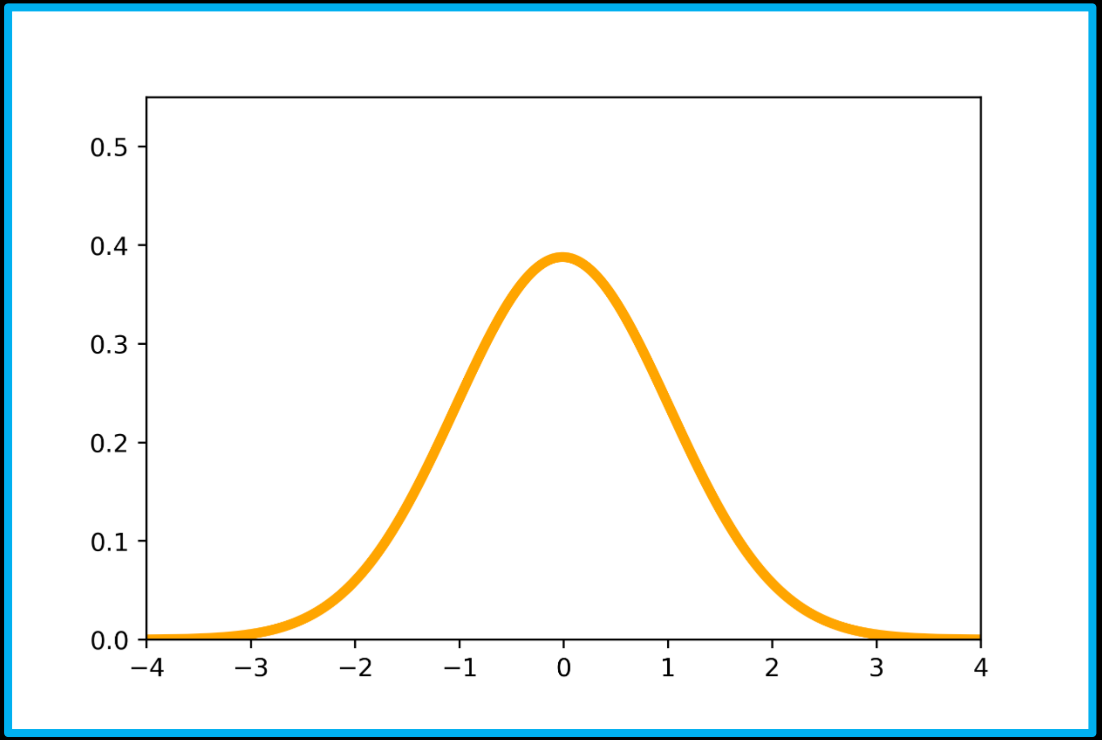
- We have **events** occurring in the real world
  - We can measure the number of occurrences (how many times the event happened with what strength)

Modelling and Sampling



- We would like to create some theories on how these events behave – ideally **create models** which correspond to them
  - Model is an abstraction of the real world

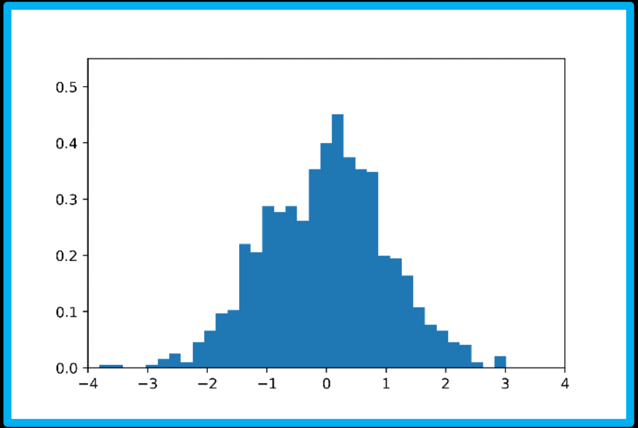
Modelling and Sampling



- We then end up with a model – we could use that model to predict what sort of results we can expect

Modelling and Sampling

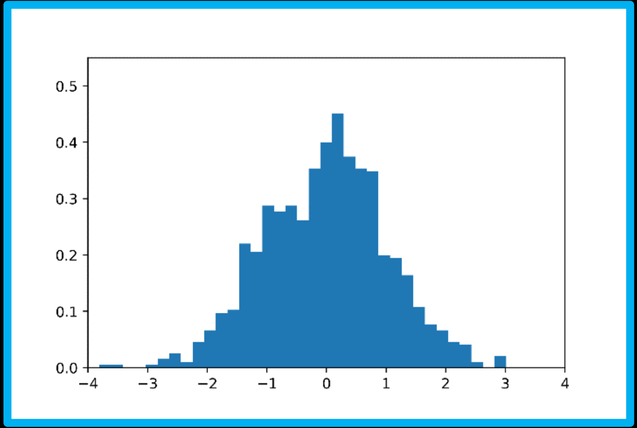
Measurements

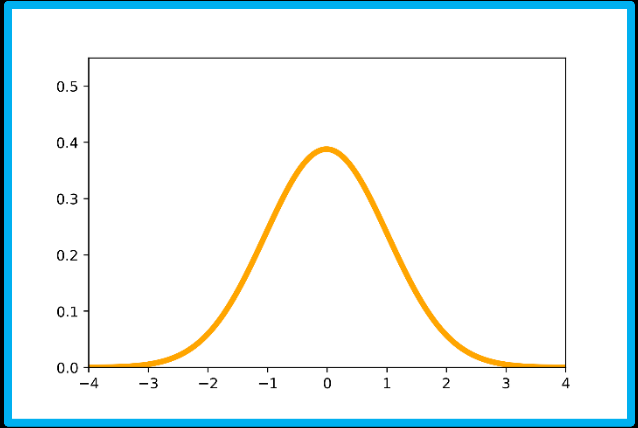


Modelling and Sampling

- **Modelling**: going from measurements to an abstracted model. Fitting a mathematical function which works in a “good-enough” way the same way as what we are seeing.

Measurements



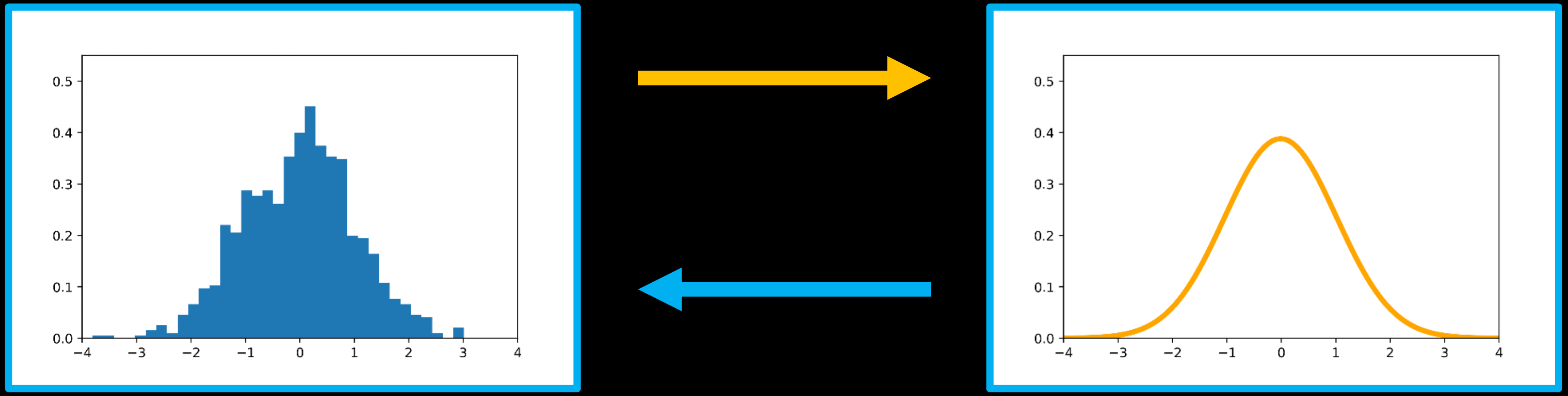


Model

Modelling and Sampling

- **Modelling**: going from measurements to an abstracted model. Fitting a mathematical function which works in a “good-enough” way the same way as what we are seeing.

Measurements

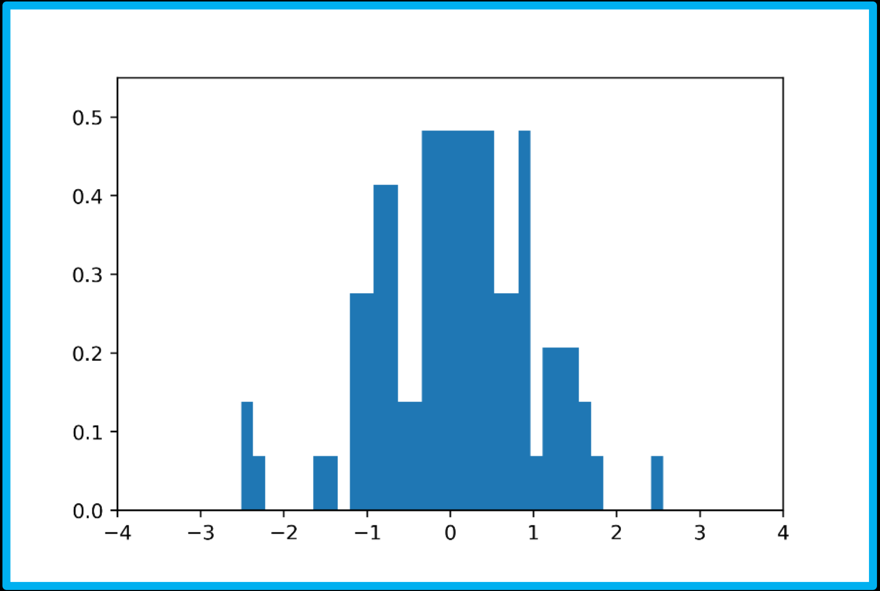


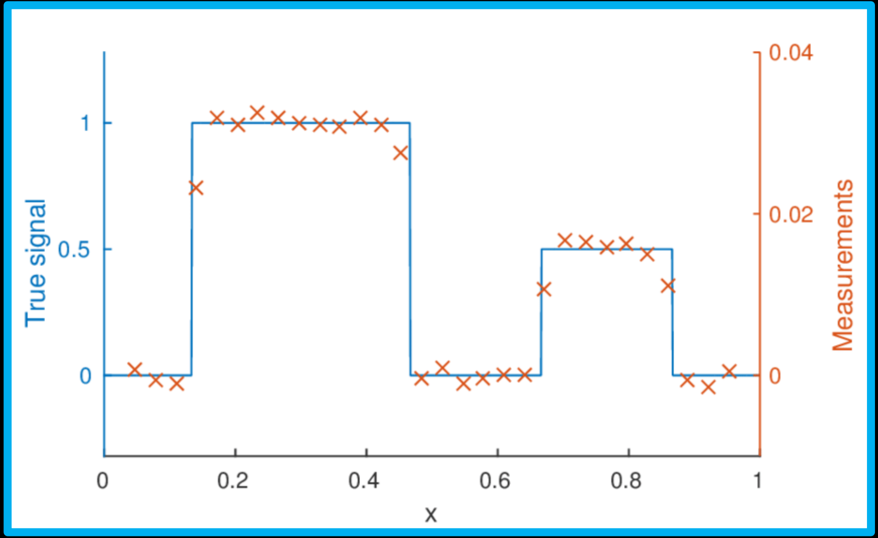
Model

- **Sampling**: going from existing model (which we think mirrors the real world) to predict how it’s going to be.
  - This could be used for simulation for example.

Real world is noisy

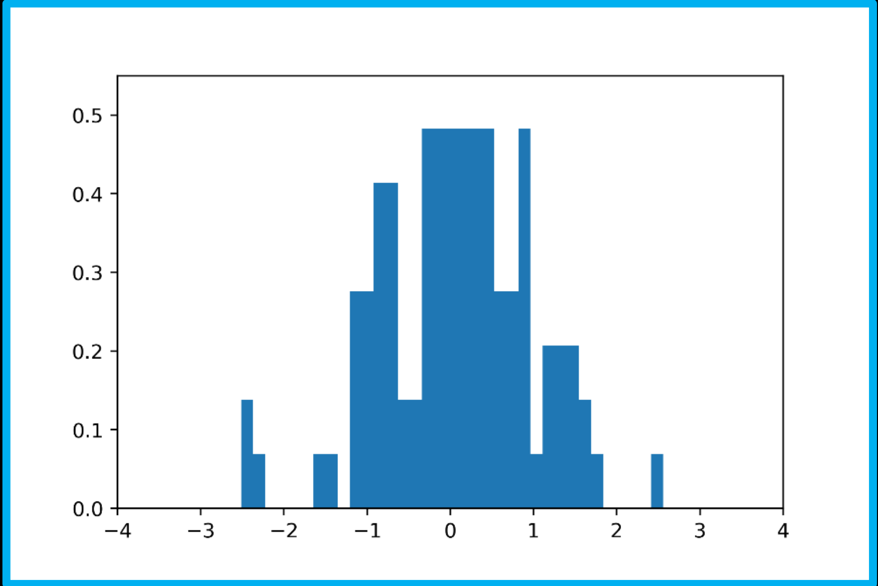
- There are a lot of **outliers** in the real world.
  - Maybe because we may have noisy measuring devices …
  - Maybe because real world tends to be diverse …

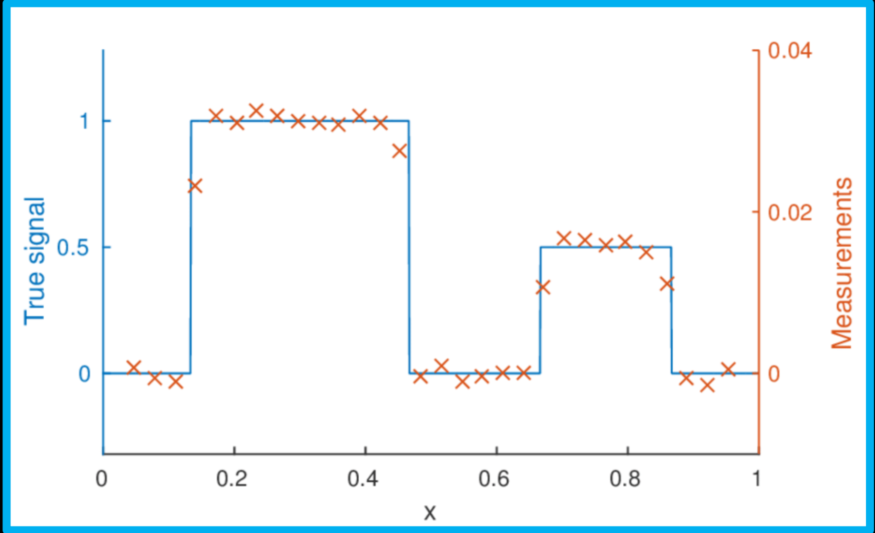


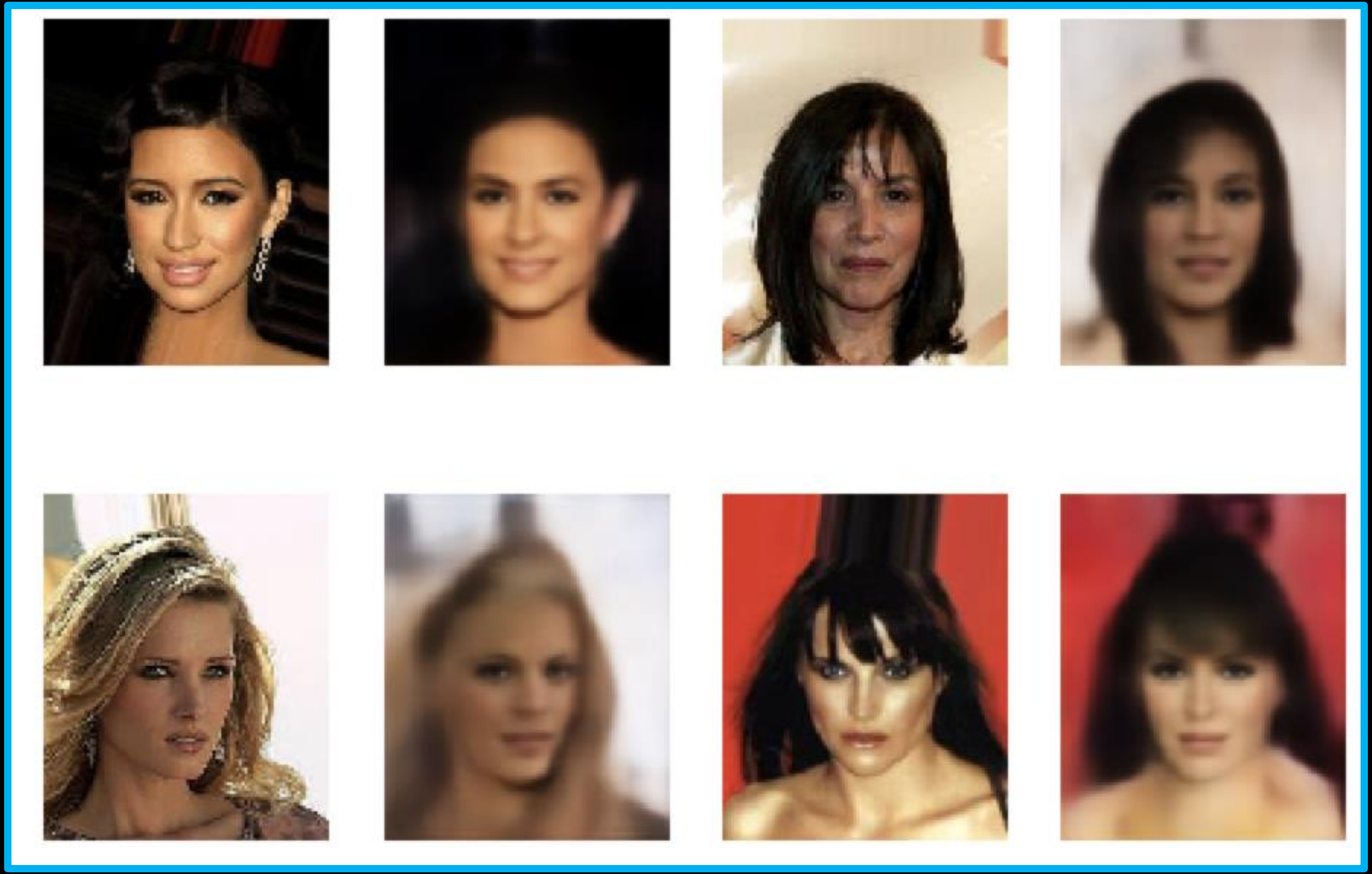


Real world is noisy

- This good!
- But we want to somehow make our models we are creating **robust to random perturbation / to noise**.
  - This usually means detecting the **outliers** and modelling the functions while ignoring them.
  - This however goes both ways. We would ideally like to have models which can produce their own outliers too … Models which don't just include the average sample.



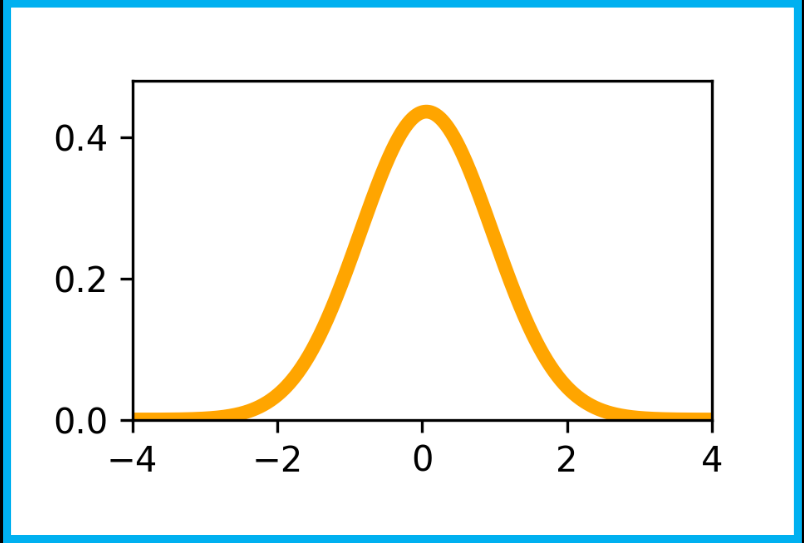




- Examples of two generative machine learning models – one uses the average features (VAE), while the other one allows for learning of details (GAN)

Some terminology

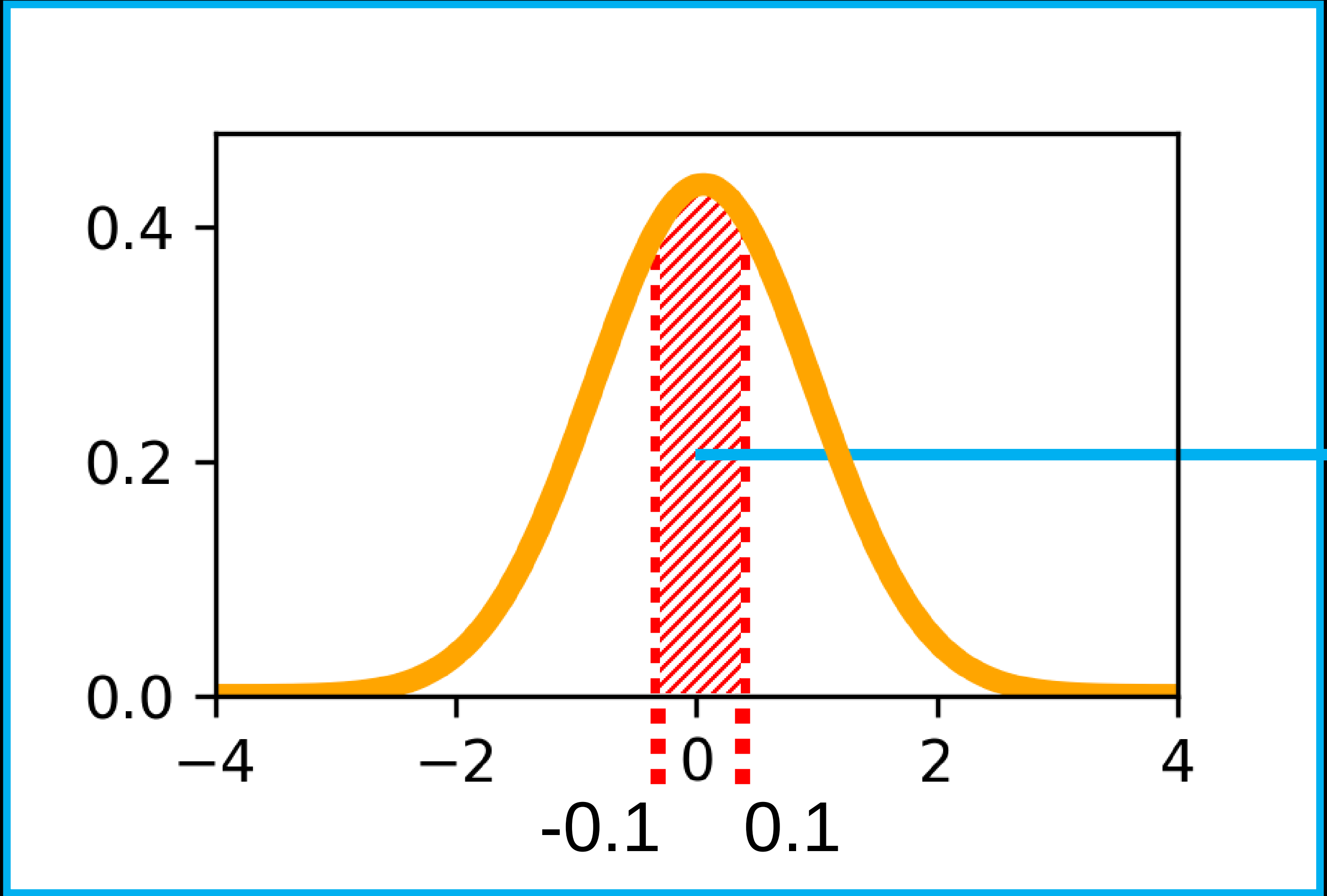
- We usually plot the function we think that corresponds to the model:
  - **Normal (=Gaussian) distribution:**



- We call this function the **probability density function (pdf)**
- We call one ***independent*** source of random behavior as a **random variable** (*PS: soon after that the terminology becomes hellish*)

pdf = probability density function

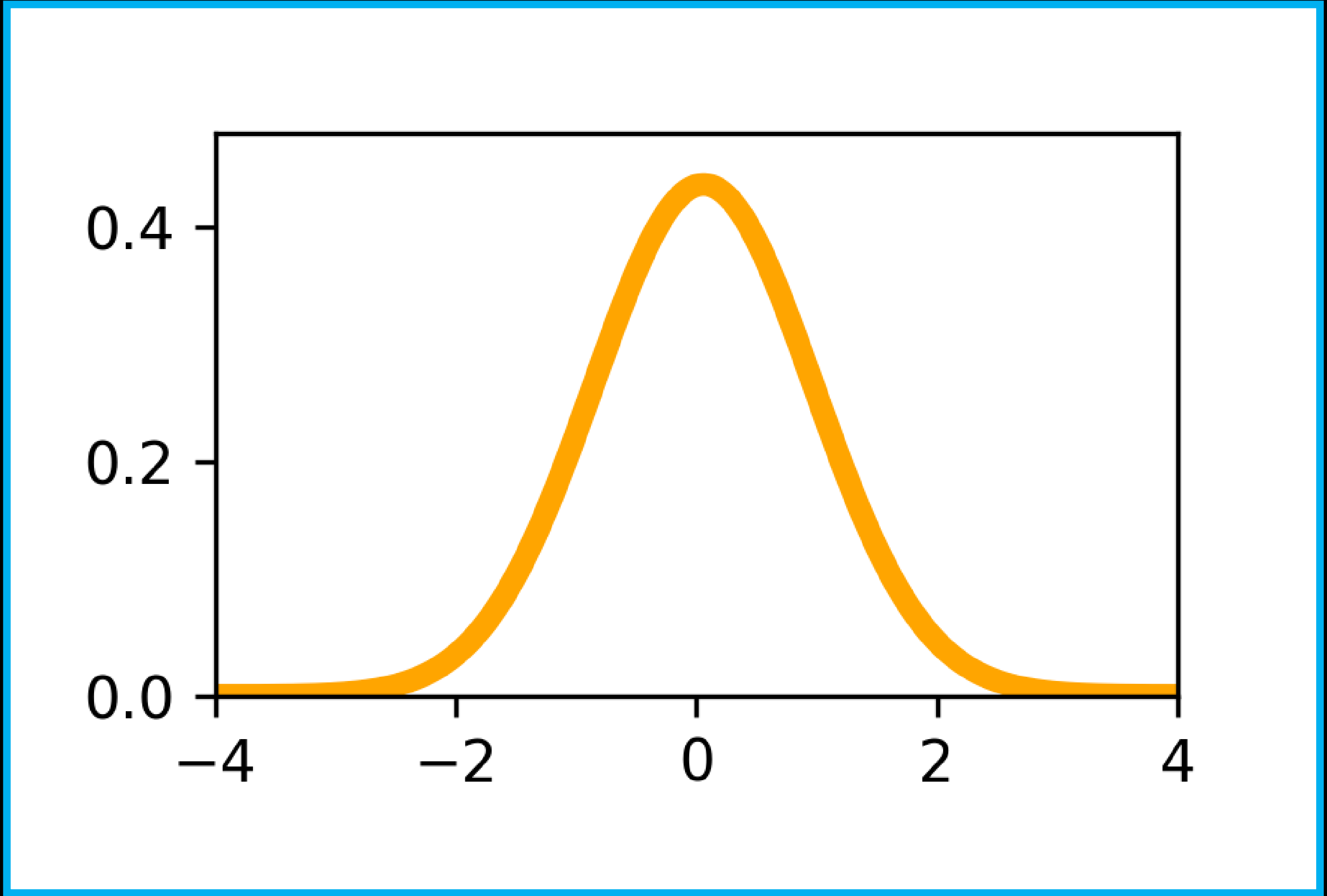
How to read p.d.f.?



- This **area** under the function corresponds to the probability of the value we measure to fall in between *-0.1* and *0.1*
- It can be calculated by integration
- *PS: Already includes the notion of imprecision of instruments – probability of the value to be exactly one number would be 0*

pdf = probability density function

How to read p.d.f.?

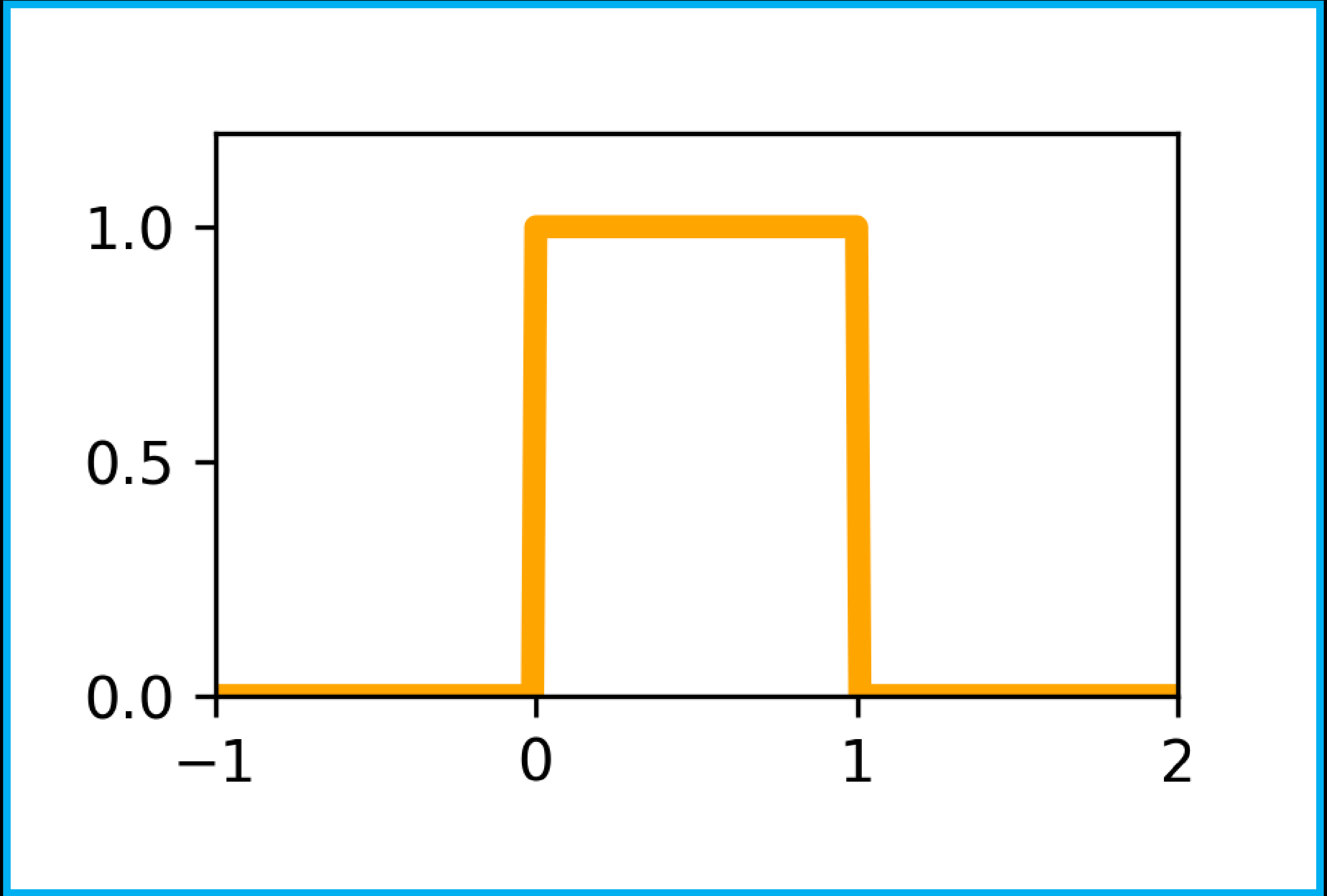


Intuitively we can understand it as:

- The largest probability is to get numbers **around 0**
- The further **away from 0**, the smaller probability

pdf = probability density function

How to read p.d.f.?



Intuitively we can understand it as:

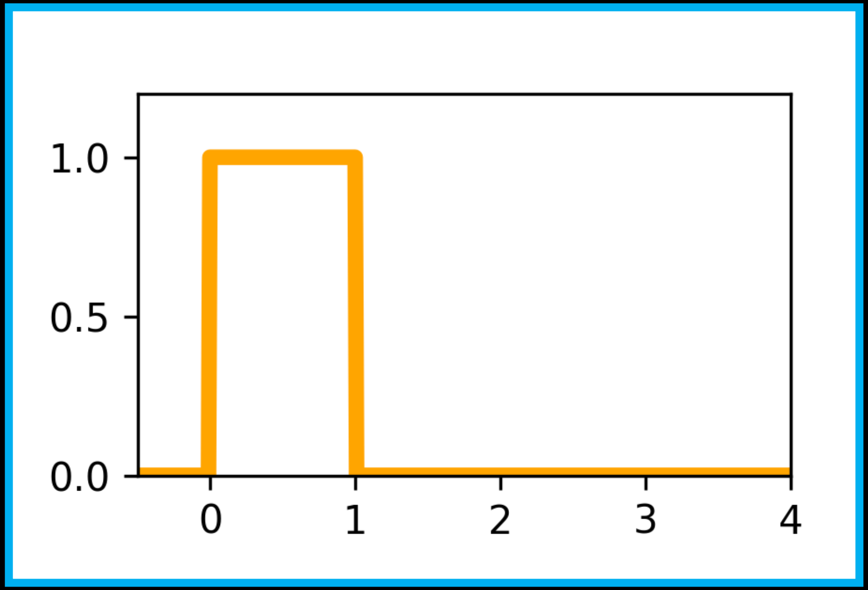
- Here we can get numbers in **between 0 and 1** with the same probability
- This is called **Uniform probability density function**

Fitting a model

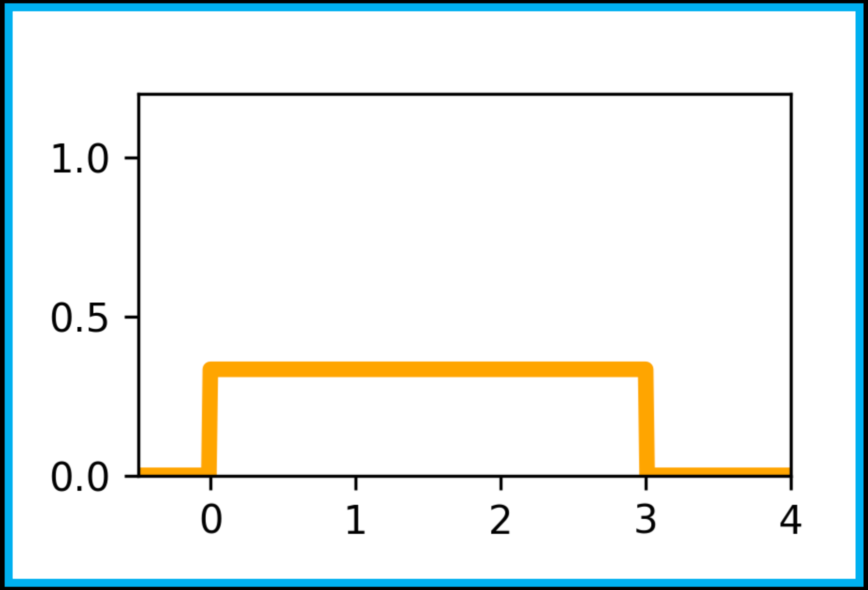
- These models have **parameters**, which influence what is the **shape of their probability density function**
- We can have one instance of the function with different parameters – one of these will usually be better for modelling the underlying event

Model parameters

- **Uniform probability density function**:
  - **U(a,b)** – numbers **between a,b** all with the same probability



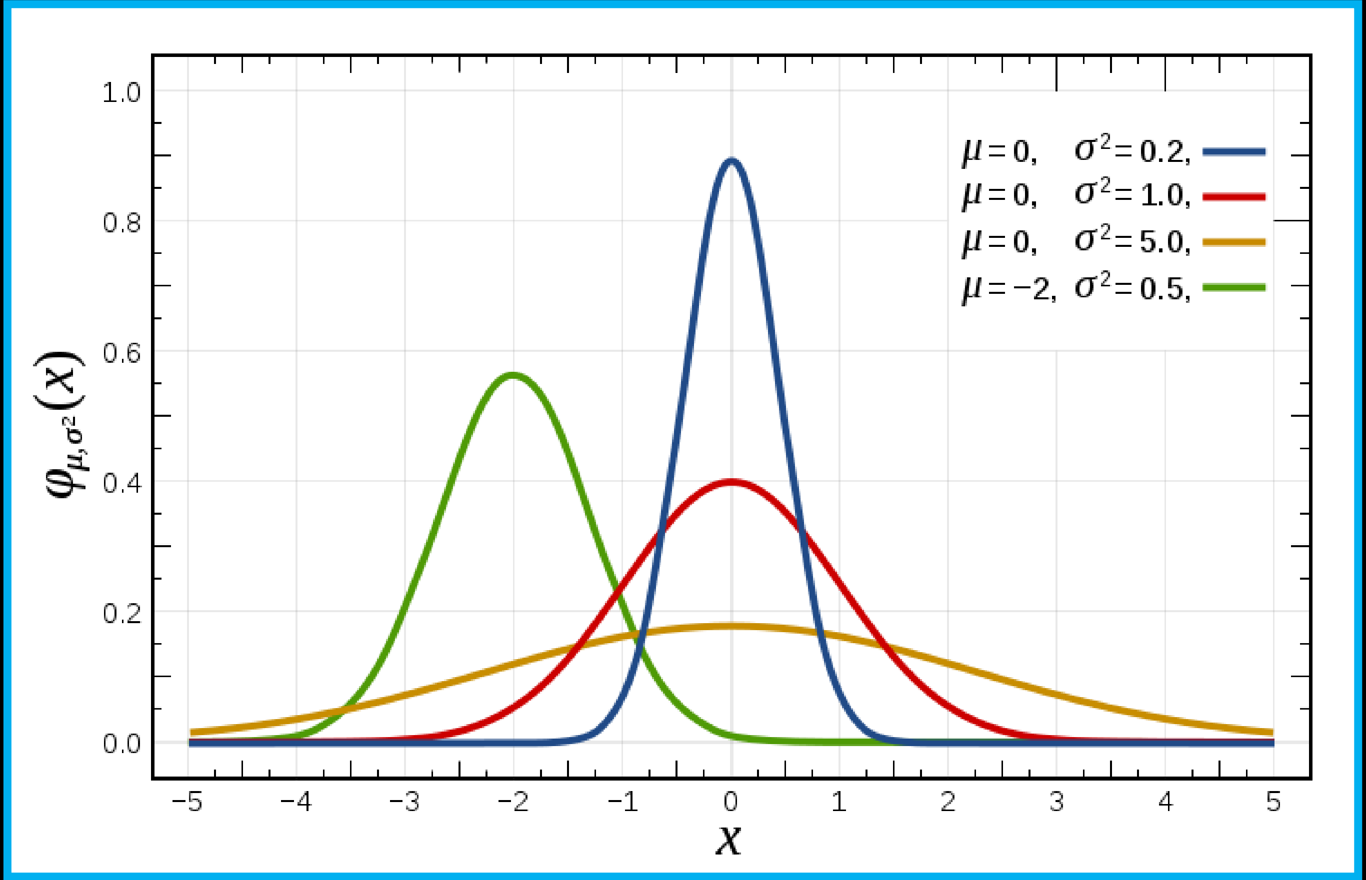
U(0,1)



U(0,3)

Model parameters

- **Normal (=Gaussian) probability density function**:
  - **N(μ, σ)** – numbers which are around mean **μ** with standard deviation of **σ**



Intuitively:

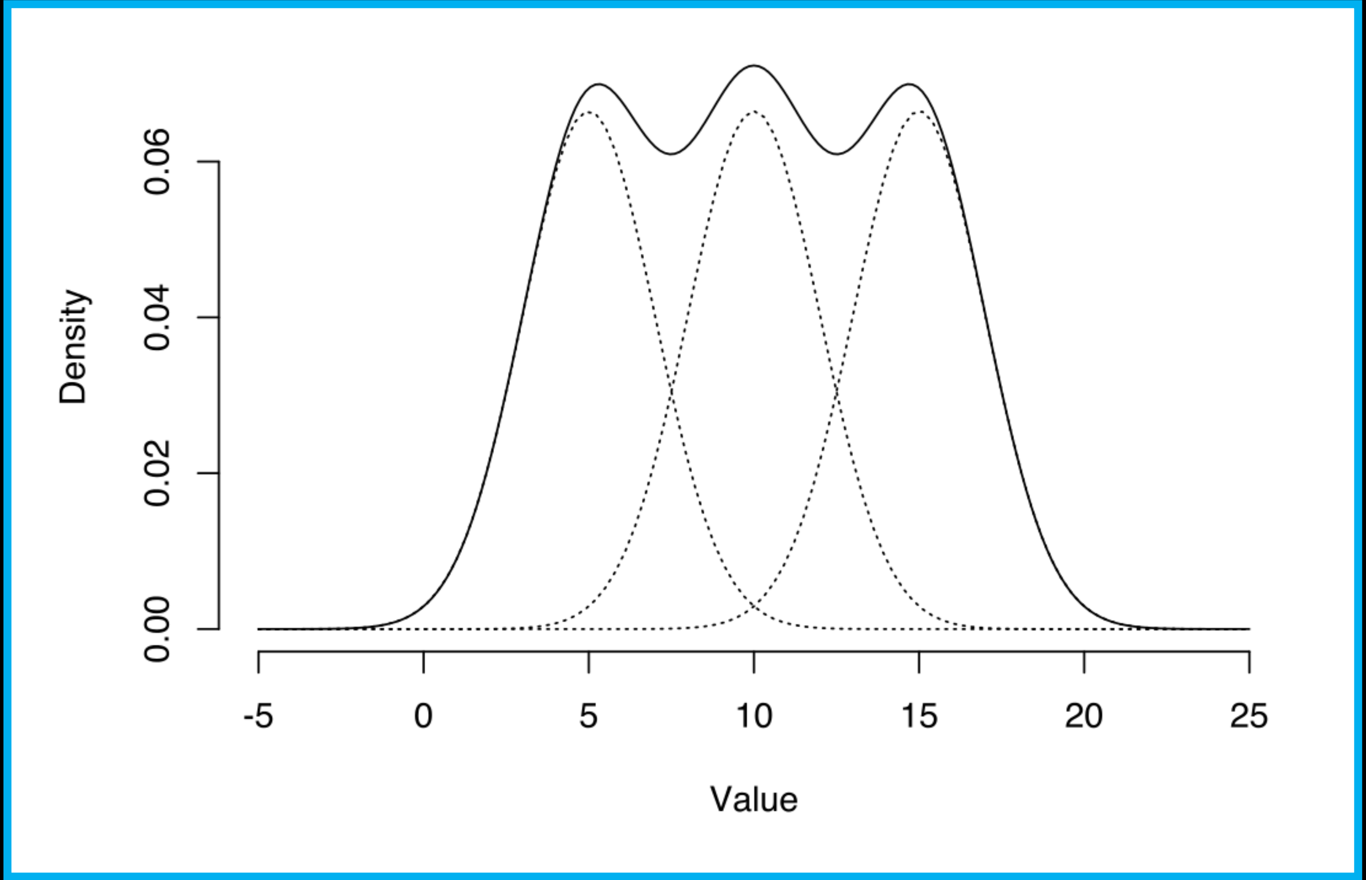
- The largest probability is to get numbers **around μ**
- How fast it drops is given by **σ**

Model parameters

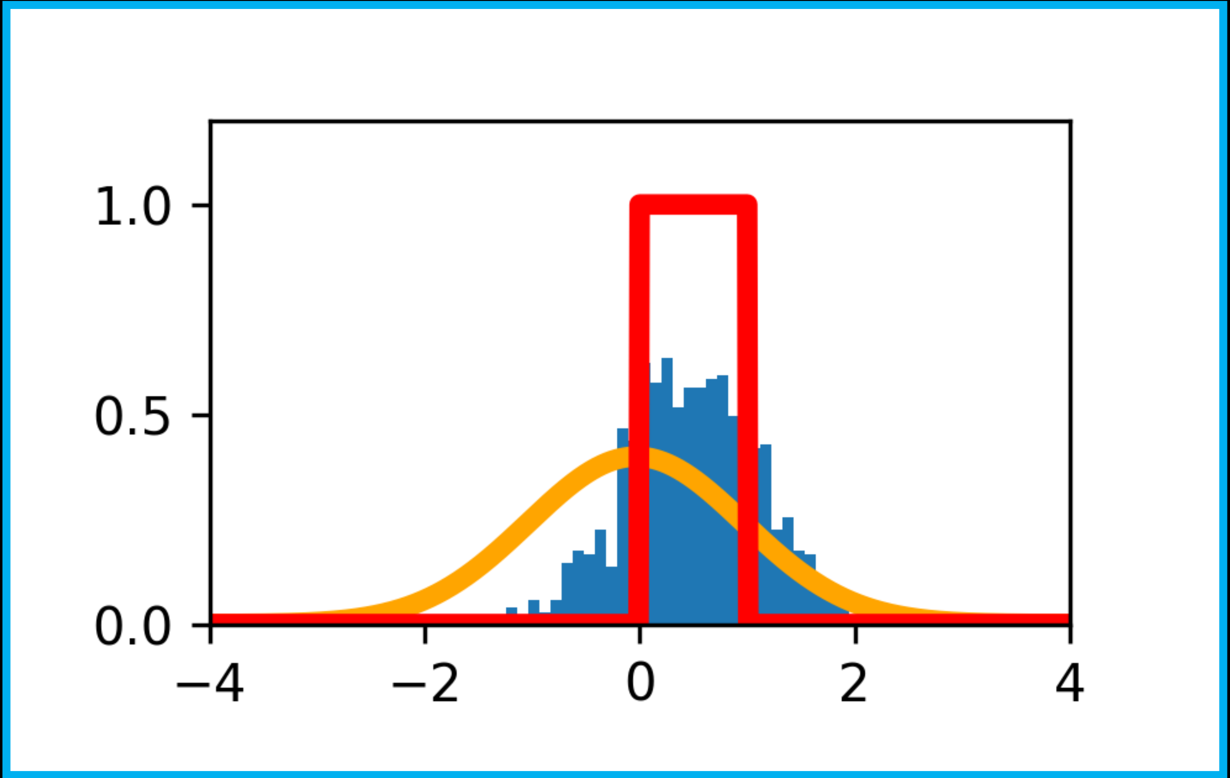
- **Mixture of normal distributions as a probability density function**:
  - Each of the normal distributions has their own parameters **N(μ, σ)**

Intuitively:

- To model more complex variables



Modelling



The task becomes to **pick the best model** corresponds to the measurements we are seeing.

And also to **find the best parameters** for this model.

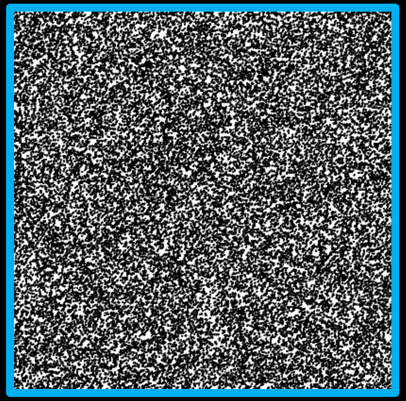
Note that these are generated values, in real life it usually isn't that simple.

U(0,1) vs. N(0,1) vs. infinite possibilities of the parameters

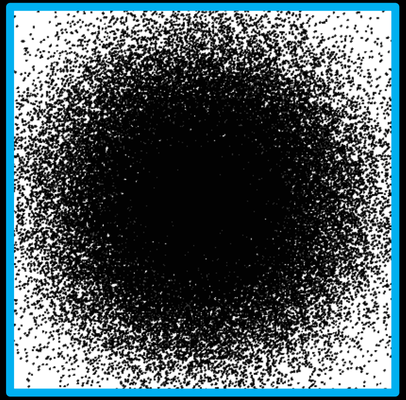
What does this means for us?

- Knowing the probability density functions can help us **understand** what a **random generator** will result with:

  - **Uniform** probability distribution – U(0,1)

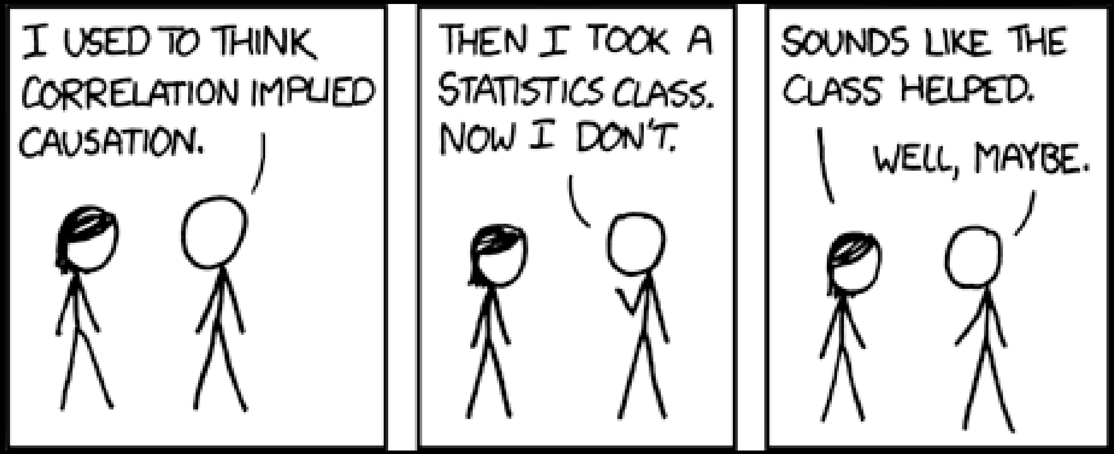
  

  - **Normal** probability density function – N(0,1)

Correlation vs Causality

- Sometimes (often!) we have **multiple events** happening at the same time. Without going too much into details, we can try to check for how one is influencing the other – aka try to see if there is **causality** (*one causes the other*) or only correlation (they are related, but one *doesn't cause the other*)



xkcd.com/552/

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

Correlation vs Causality

- Excerpt from *"Introduction to Probabilities, Graphs, and Causal Models"* by Judea Pearl (ch1 here):

  - (...) Observation that causal utterances are often used in situations that are plagued with uncertainty. We say, for example, "*reckless driving causes accidents*" or "*you will fail the course because of your laziness*" (Suppes 1970), knowing quite well that the antecedents merely tend to make the consequences more likely, not absolutely certain.
  - Connected with this observation, we note that probability theory is currently the official mathematical language of most disciplines that use causal modeling, including economics, epidemiology, sociology, and psychology. In these disciplines, investigators are concerned not merely with the presence or absence of causal connections but also with the relative strengths of those connections and with ways of inferring those connections from noisy observations.

Pause imminent

- After the pause, we will look into some **practical uses of probability**.

Pause 1

Statistics

- We can apply Probability theory with Statistics
- You might already know some of the approaches we can take to analyze some data which we measured in the real world …

Statistics

- We can use some statistical formulas to describe measurements:
  - **Average** of a list of numbers
  - **Standard deviation**
  - **Median** of a list of numbers
  - **Minimum** and **maximum**

Average and standard deviation:

- Given a **list of numbers X** (aka list of observations we made from the real world – maybe using some measuring device …), we can calculate:

```
X = [2, 0, 1]
N = len(X) = 3
```

- **Average**: $\mu = \frac{1}{N}\sum x_i$ $= \frac{1}{3}(2 + 0 + 1) = 1$

- **Standard deviation**: $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ $= \sqrt{\frac{(2-1)^2 + (0-1)^2 + (1-1)^2}{3}}$

In practice

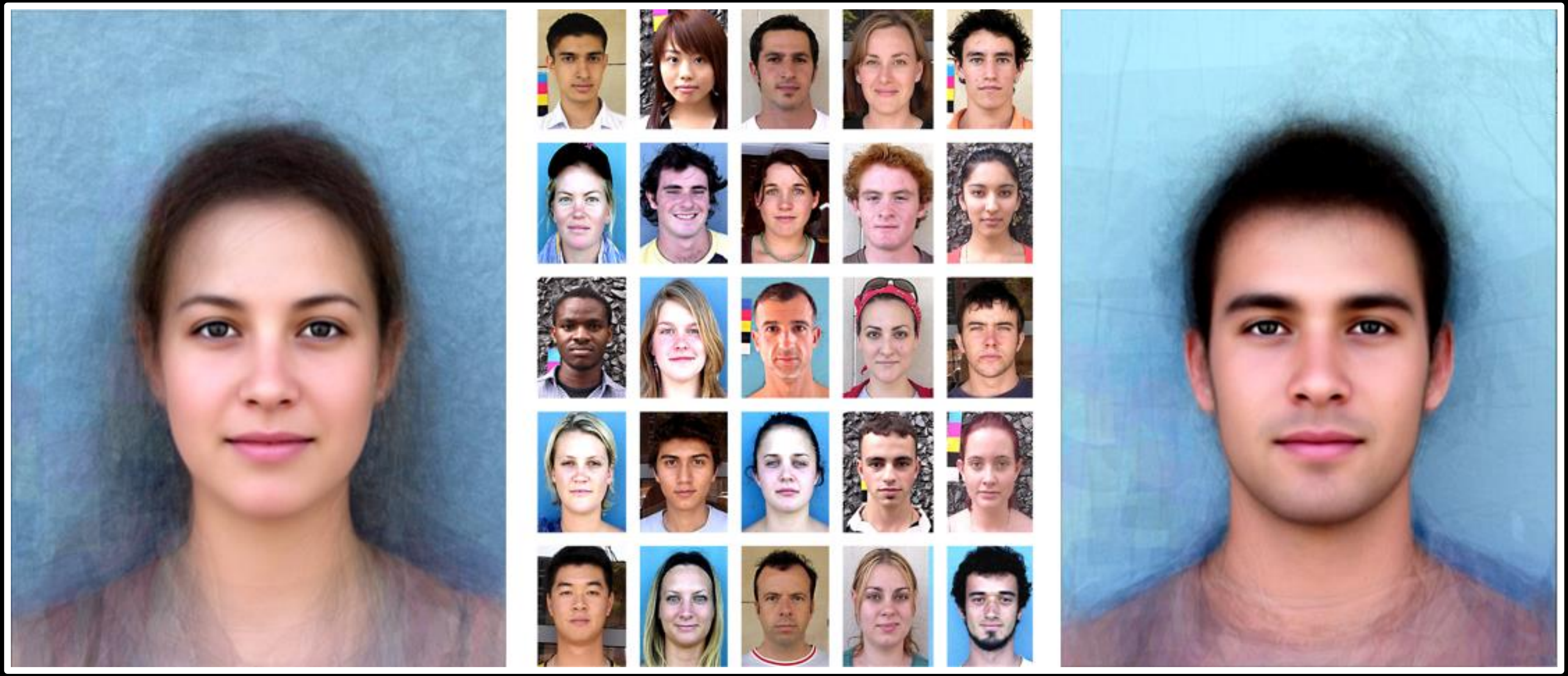
- In practice these statistics are often useful to compress the amount of information we are reading – instead of looking at a lot of point on a plot, we can look at the more concise message by observing these statistics

Average visually

- When working with images, we can illustrate these concepts visually:

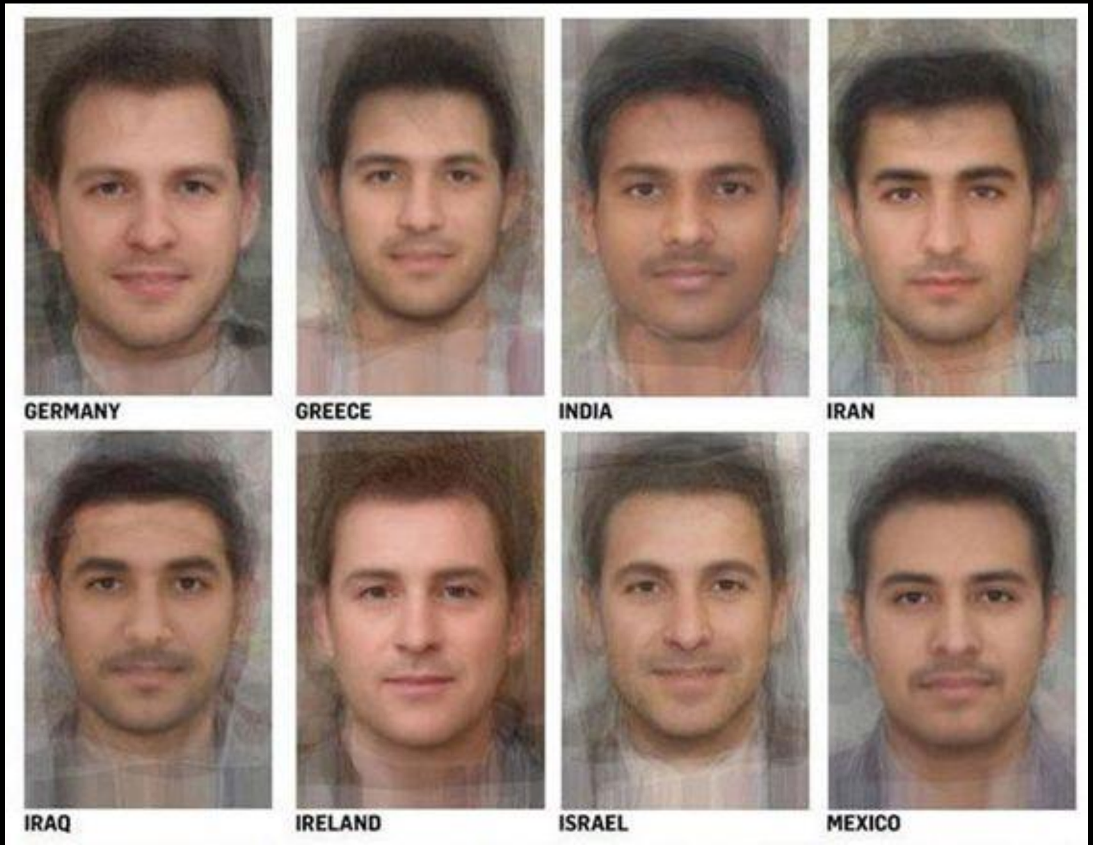
Average visually

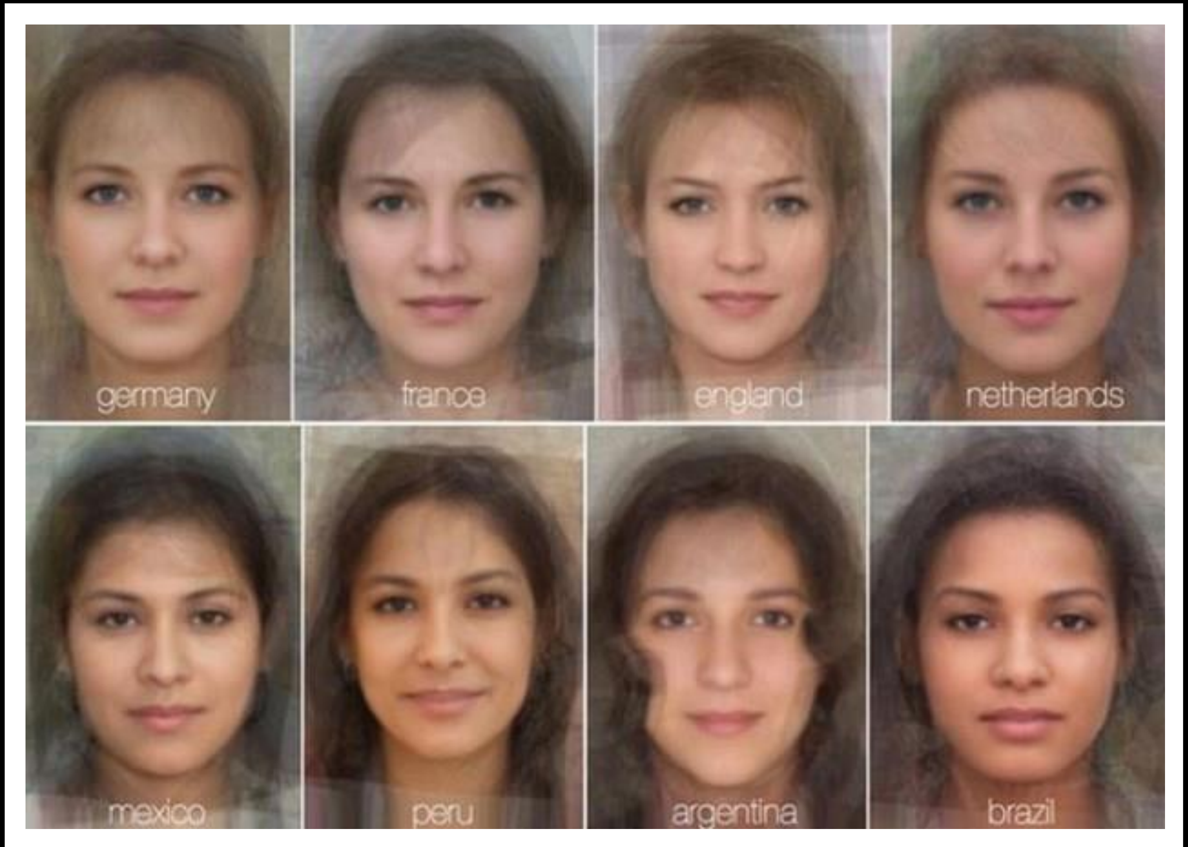
- When working with images, we can illustrate these concepts visually:
- Average human face



Average visually

When calculating statistics, the **choice of the dataset** matters:

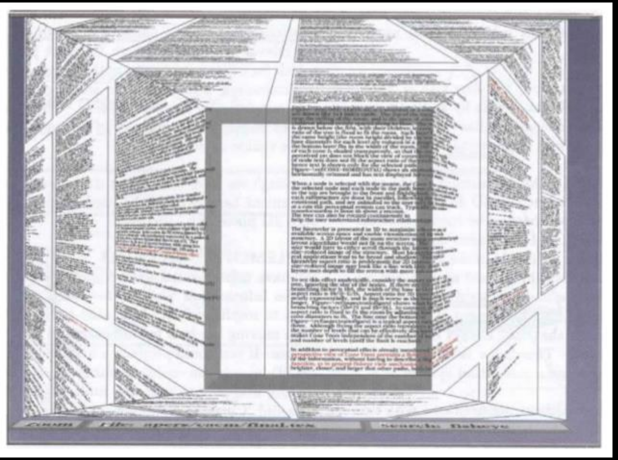


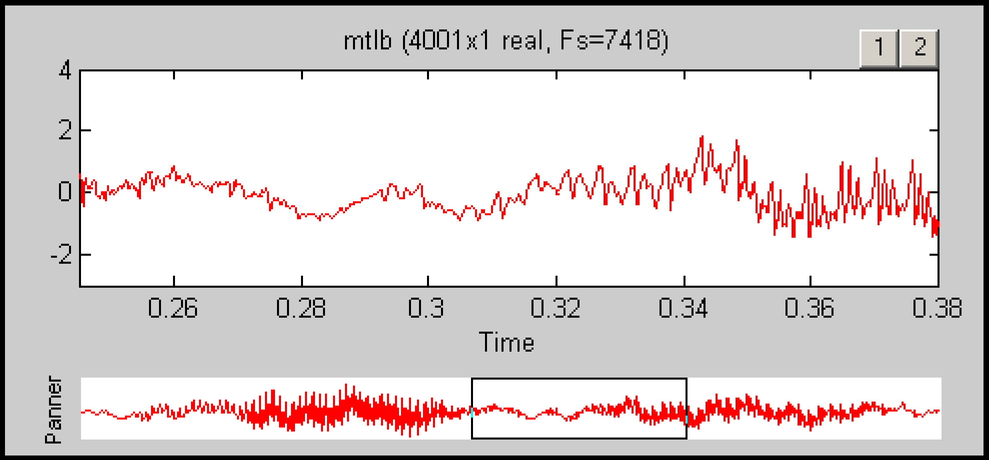


Concepts: Overview + Detail

- Task of visualization: show what is important (**detail**), but also somehow tell us about the context (**overview**).

- Interface design research - *"A review of overview+detail, zooming, and focus+ context interfaces."* A. Cockburn, …, (paper from 2009)



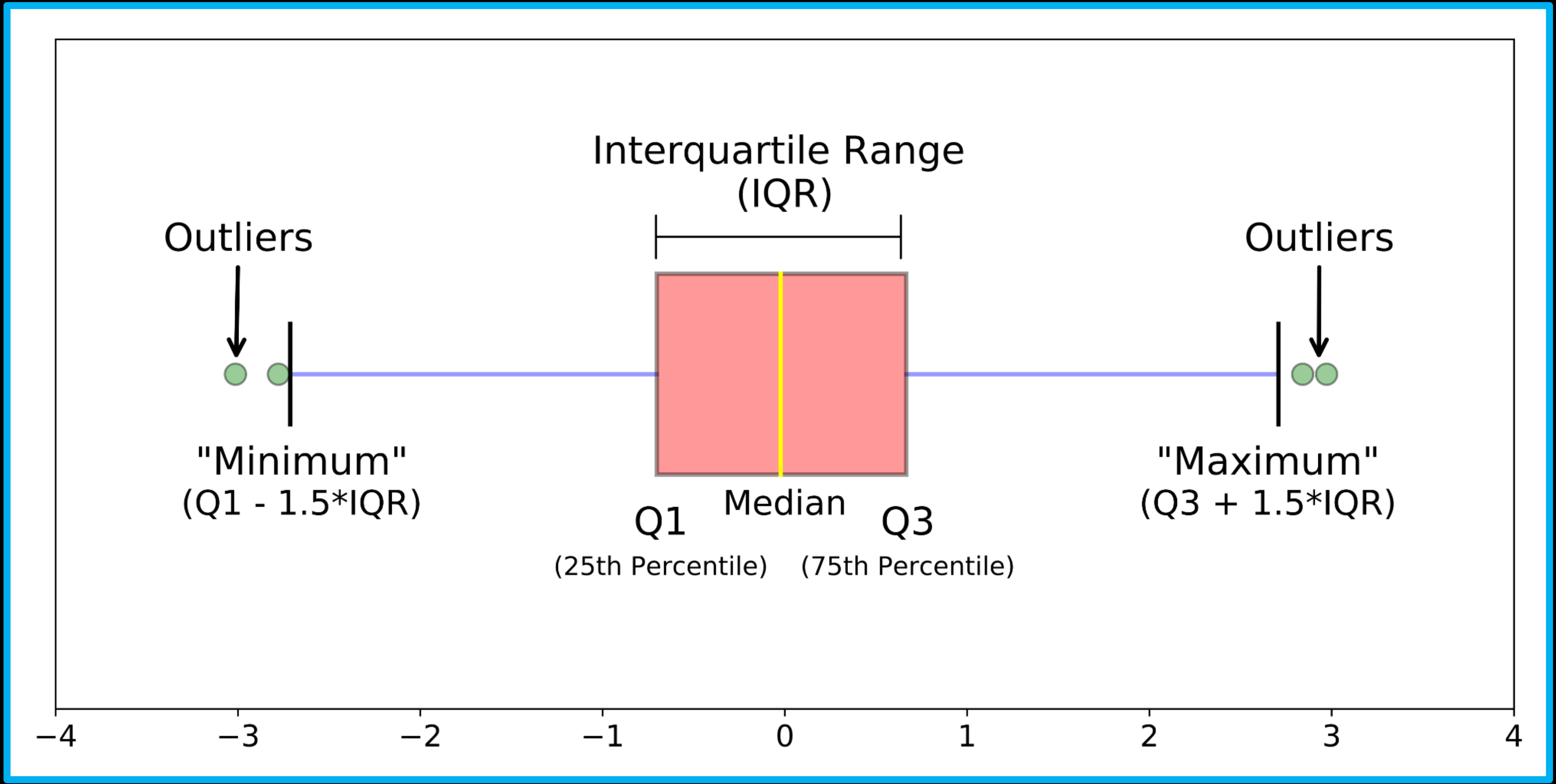


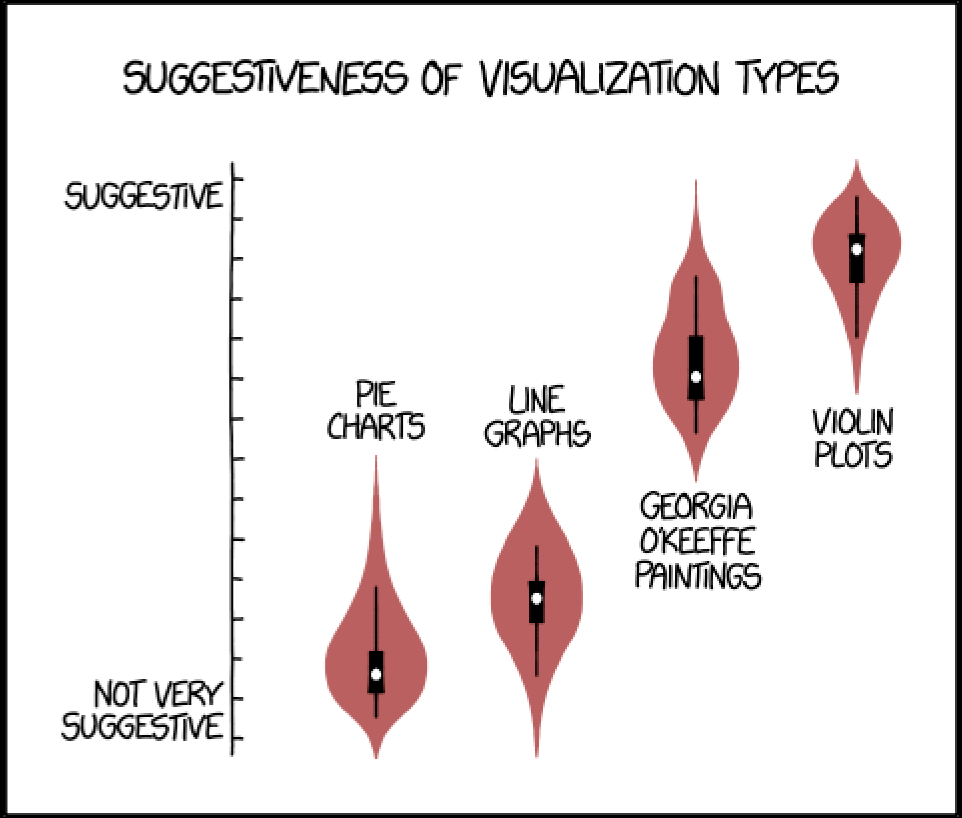
Ways of visualizing

- There are many types of plots we can use to show the data
  - Sometimes showing the **raw measurements** (just bunch of dots) might be the best …
  - … but most of the cases we want to somehow **process the data** so that we can have a better understanding when we see it

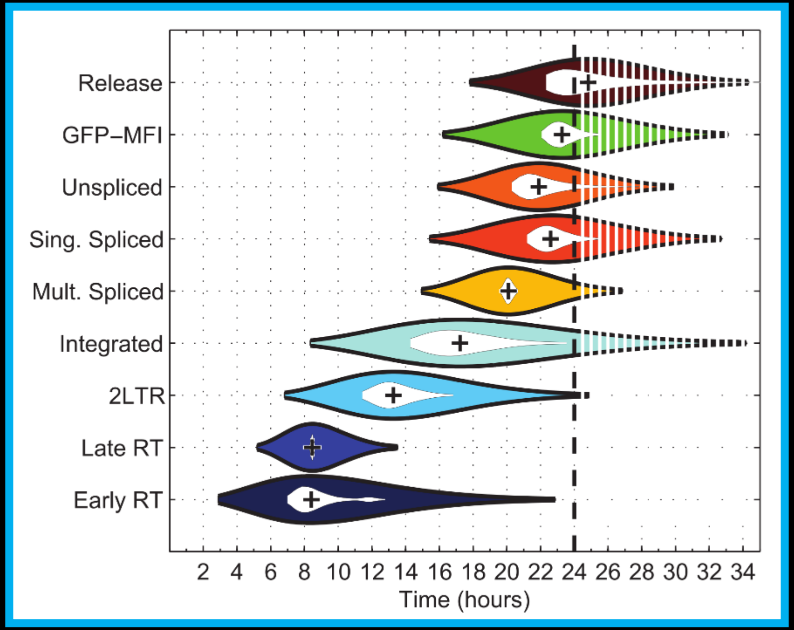
  - **Extracting the relevant information**

Box Plot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Median
Q1
(25th Percentile)
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4



xkcd.com/1967/

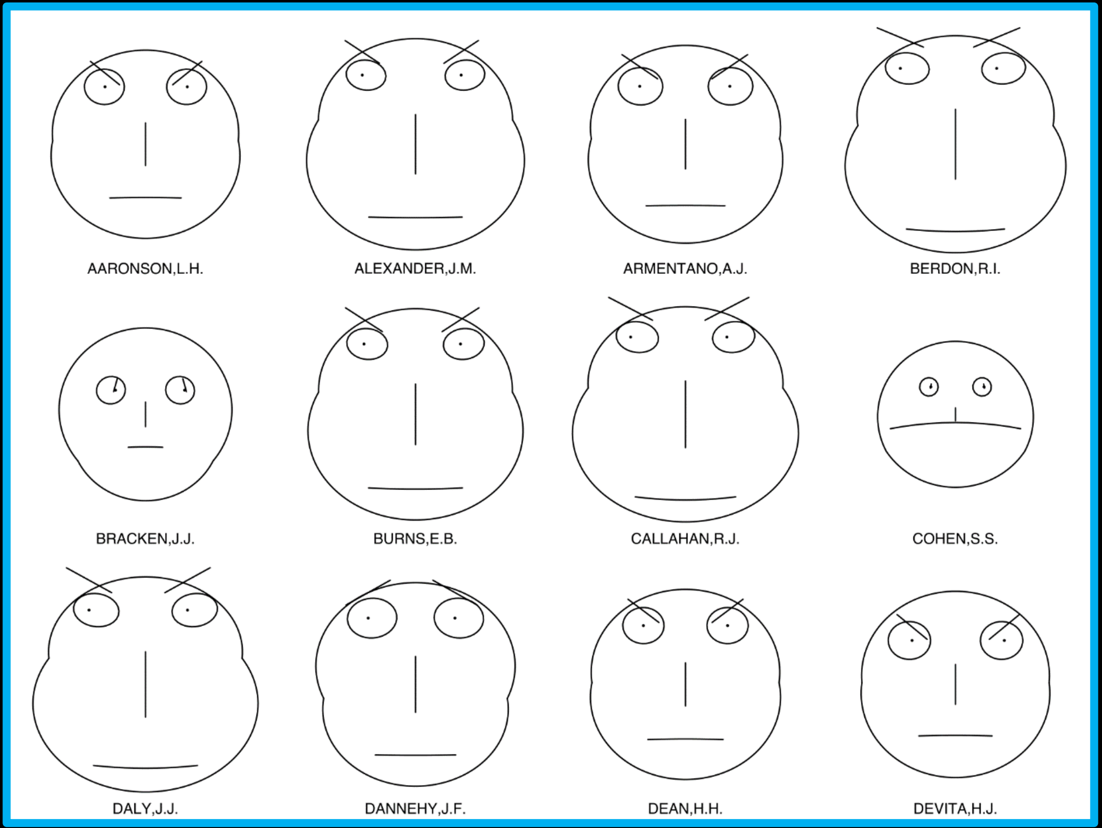


(PS: they are really used in research)

Other visualizations

- Fun one: **Chernoff faces**
  - Example: lawyer's ratings of twelve judges
  - Real world example in ArcGis

- Example of using a **glyph** as a visualization:
  - It turns out that visualizing data with many dimensions (*each data point is actually a full vector of measurements*) … is pretty hard!



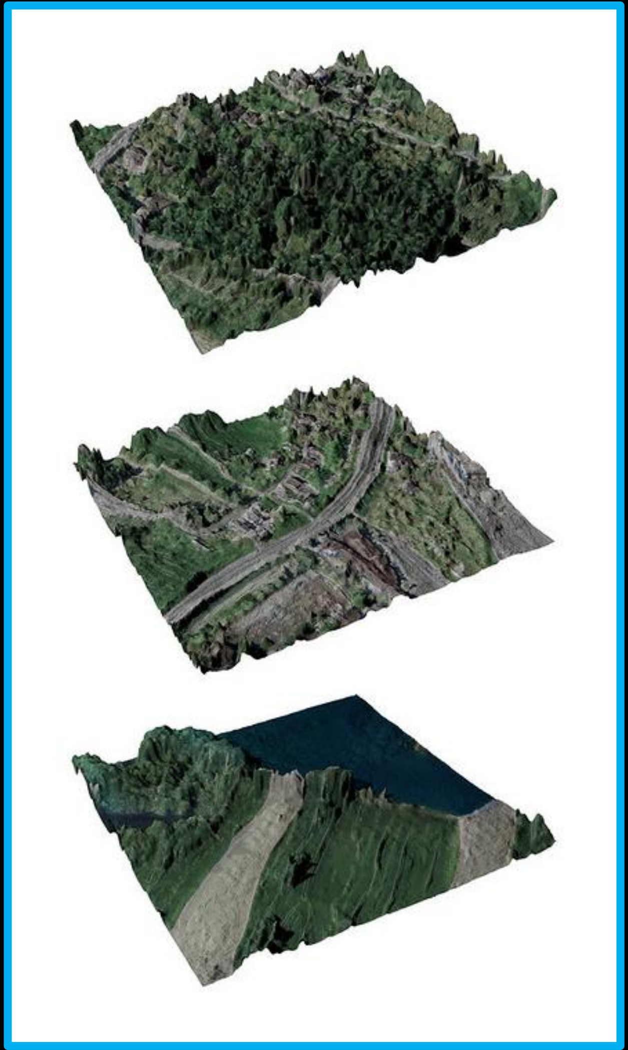
Sampling from random functions

- Sampling from random functions we can also:
  - Create 2D textures!
  - Generate terrains from them!

- Random functions:
  - Uniform, Normal/Gaussian
  - Perlin noise (procedural)

Links: Terrain generation

- Cool example of simulating effects such as erosion to alter the generated terrain:
  - https://www.youtube.com/watch?v=eaXk97ujbPQ
- Using machine learning to generate new landscapes and assign it heightmaps -> terrain:
  - Uncanny Valleys: Generative landscape
- Research paper from 2017 with interactive terrain generation:
  - https://www.youtube.com/watch?v=NEscK5RCtlo



Links: Texture generation

- Real described texture dataset: https://www.robots.ox.ac.uk/~vgg/data/dtd/
- Machine Learning generated: https://www.youtube.com/watch?v=KL6U6iasUxs

Pause 2

Programming

- **Random functions in Python**
  - Sample random numbers and plot then to show the probability density function
  - Do the same in 2D -> Generate textures!

- **Statistical analysis of data / images**
  - Describe data with statistics

The end