Simplicial depth measure

What are depth measures?

Depth measures are algorithms used to identify how deep a certain point, relative to a given data set, lies within the data cloud, providing center-outward ordering of points in any dimension. Depth measures, commonly used for clinical trials and in finance, highlight the most superficial points called anomalies and the deepest point i.e. multivariate median: the middle value in a frequency distribution.

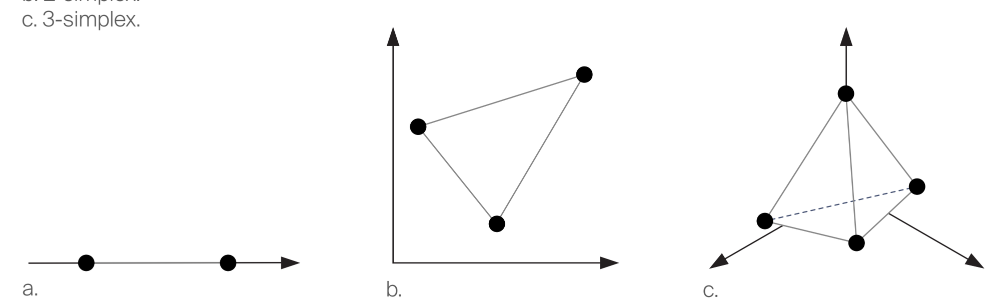
Why simplicial?

This poster illustrates the simplicial depth measure method. It is called "simplicial" because the process is based on the creation of simplices containing the points of the data cloud.

What is a simplex?

The simplex is the simplest possible closed figure in any given space. The figure changes with the number of dimensions of the space.

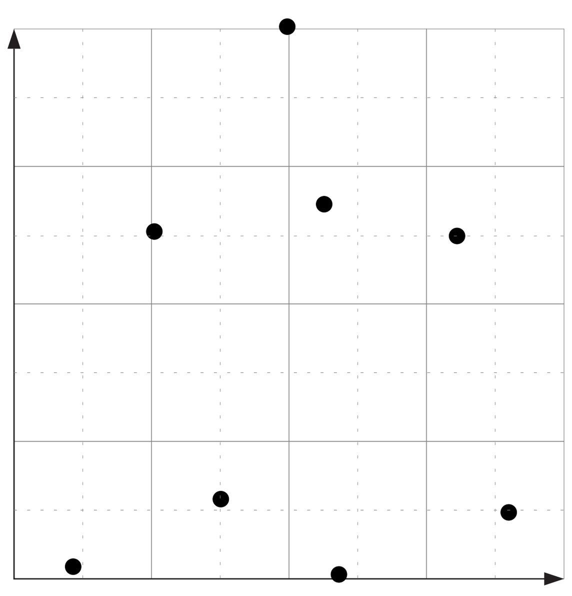
a. 1-simplex.
b. 2-simplex.
c. 3-simplex.



1. Arrange the dataset

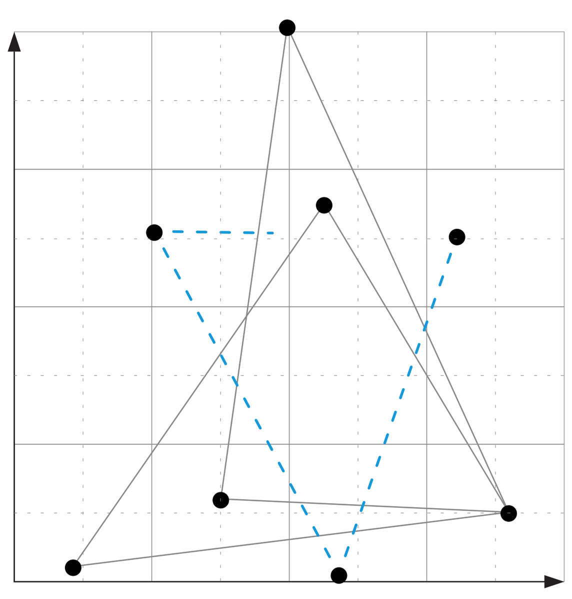
Order every data point in the space according to its **features**. Given that the data is bivariate, the space is a Cartesian plane.

To simplify this part of the representation, these steps show only eight points of the training set.



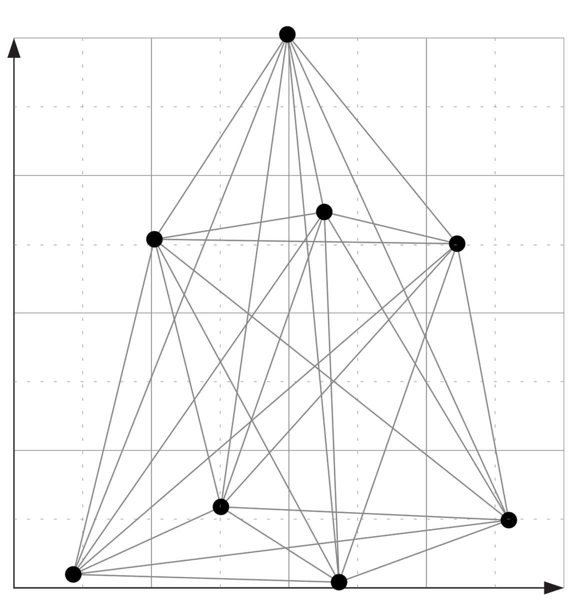
2. Draw the simplices

Calculate all the simplices in the space. In this case every simplex is obtained using three data points. The number of simplices is given by the **matrix (N r)**, where *N* is the total number of data points in the space and *r* is the number of vertices needed to draw a simplex.



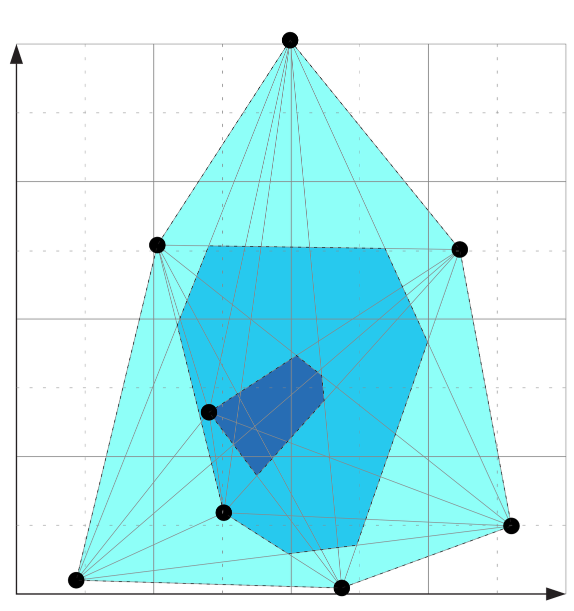
3. Count the number of simplices

The depth of each data point is given by the percentage of simplices in which it is included: **the number of simplices in which the point sits, divided by the total number of simplices**. The perimetral points need to be included in the count.



4. Draw the depth map

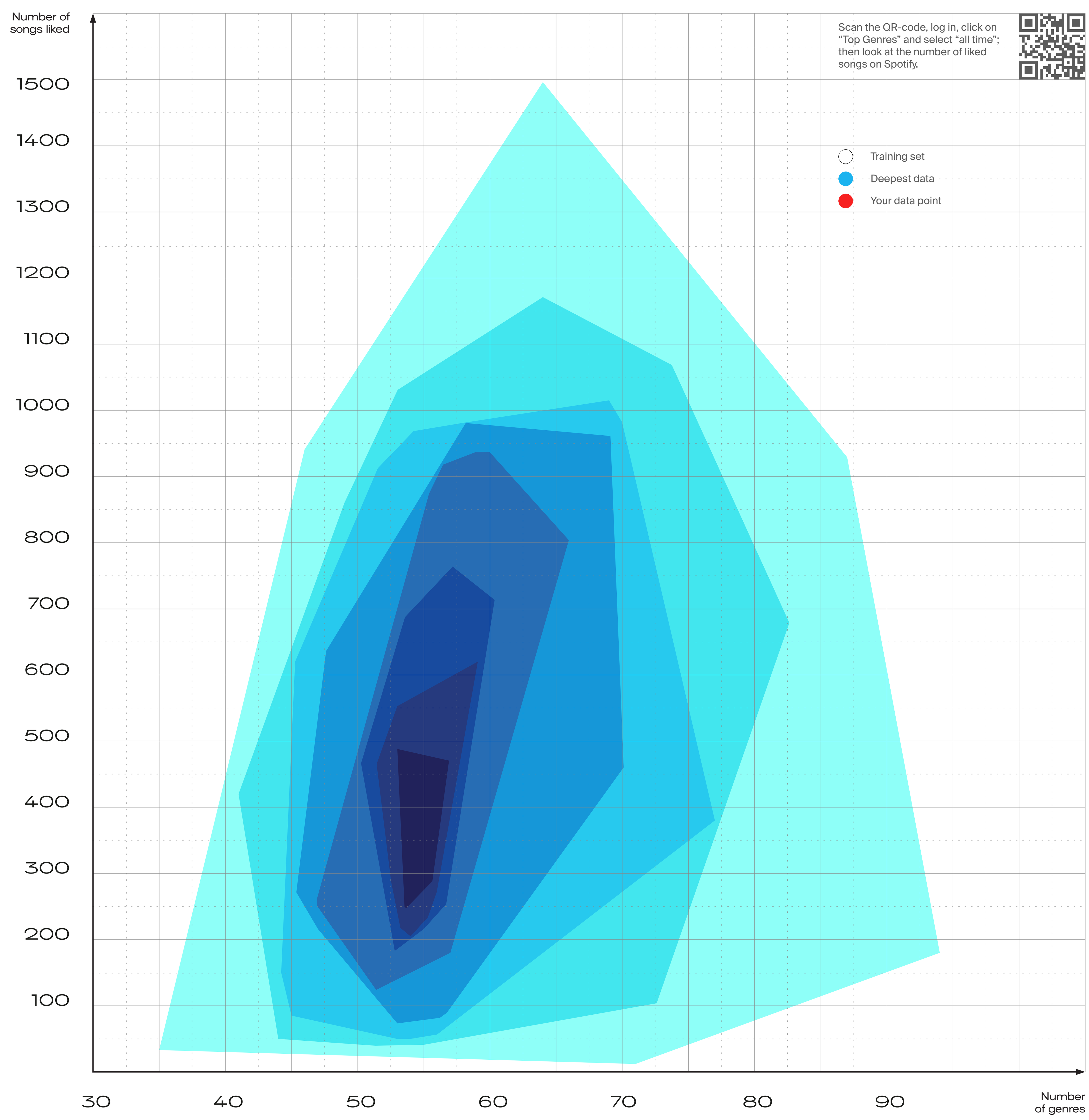
To create the different **layers of depth**, connect the data points that have the same range of depth. The area of a layer must also contain all the layers below, creating a progression of concentric shapes.

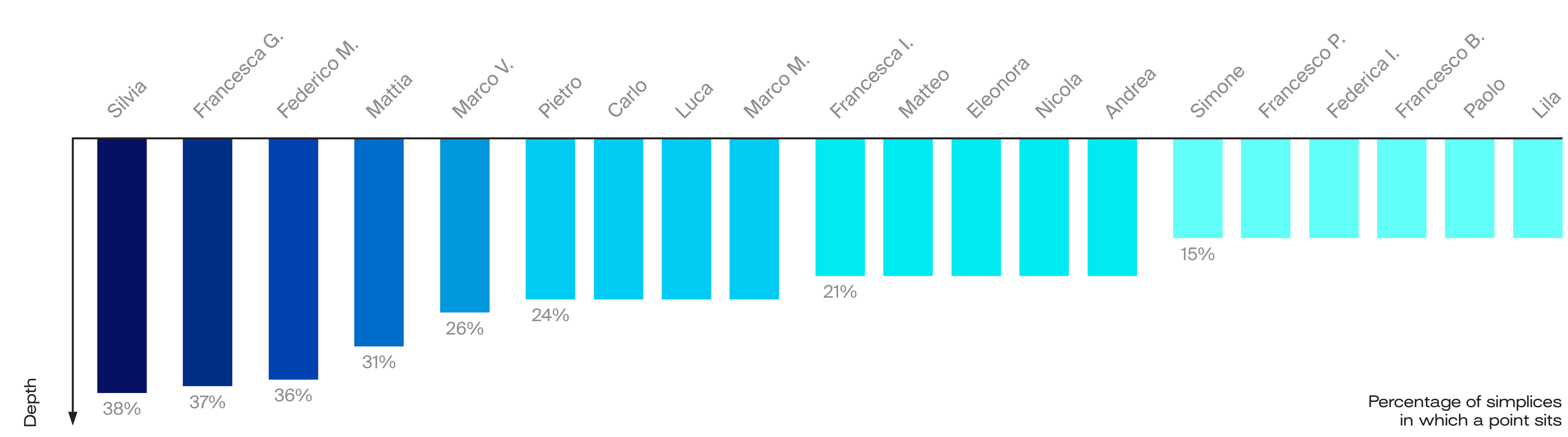


5. Compare your data to the training-set

Place your data point in the depth map to compare your use of Spotify to the training set.

This step will represent all the data points of the training set.

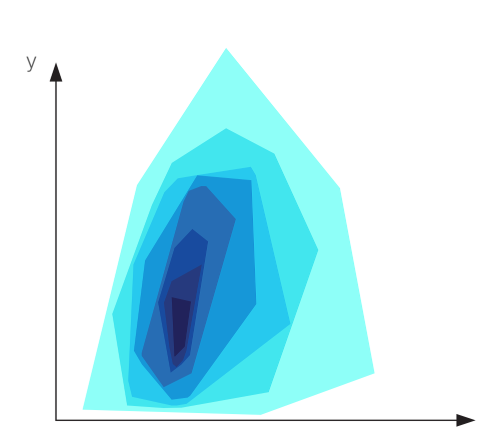


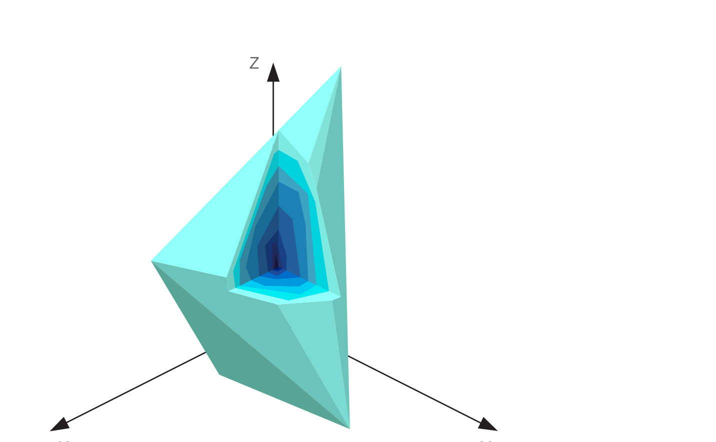


Dimensional Curse

A particular feature that sets the depth measure apart from other measurements in statistics is its ability to withstand the so-called "**dimensional curse**". It is in fact possible to calculate the depth measure of a data set in **any number of dimensions**.

x. Number of genres.
y. Number of songs liked.
z. Account age (weeks).





VISUAL EXPLANATIONS OF STATISTICAL METHODS

Simplicial depth measure

AUTHORS

Francesco Battistoni
Carlo Boschis
Federica Inzani
Federico Meani
Mattia Mertens
Ottavia Robuschi

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

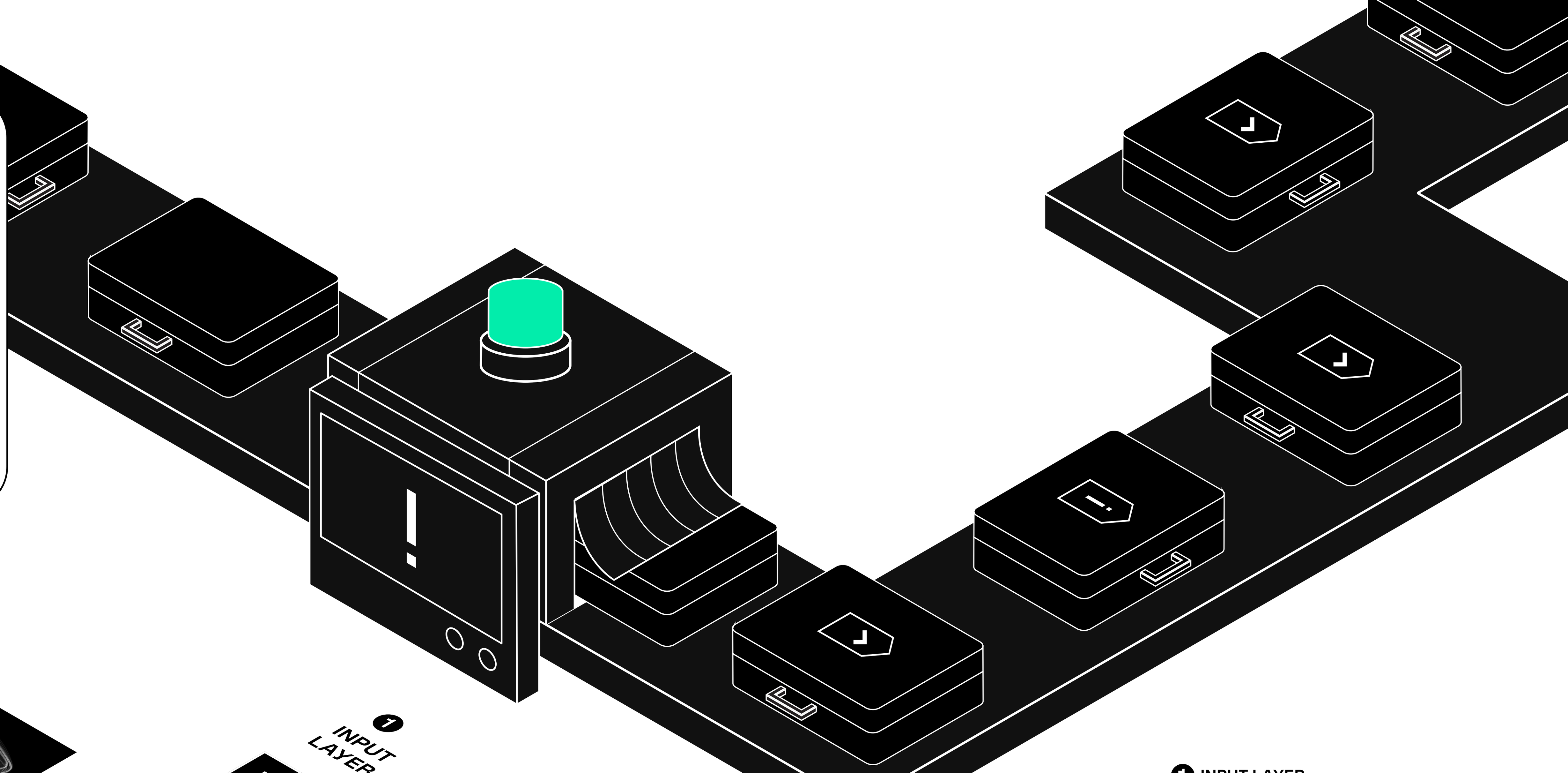
LM in Communication Design
A.A. 2021/2022

DEN-
SITY
GN+

POLITECNICO
MILANO 1863
SCHOOL OF DESIGN

NEURAL NETWORKS

How can a machine understand if the contents of your luggage are dangerous or not? In this poster we will try to go through the inner workings of a neural network trained to detect hazardous items. This kind of system is already in use in various industries like security, health and transports. Let's have a look inside this black box and try to understand how a machine can see.

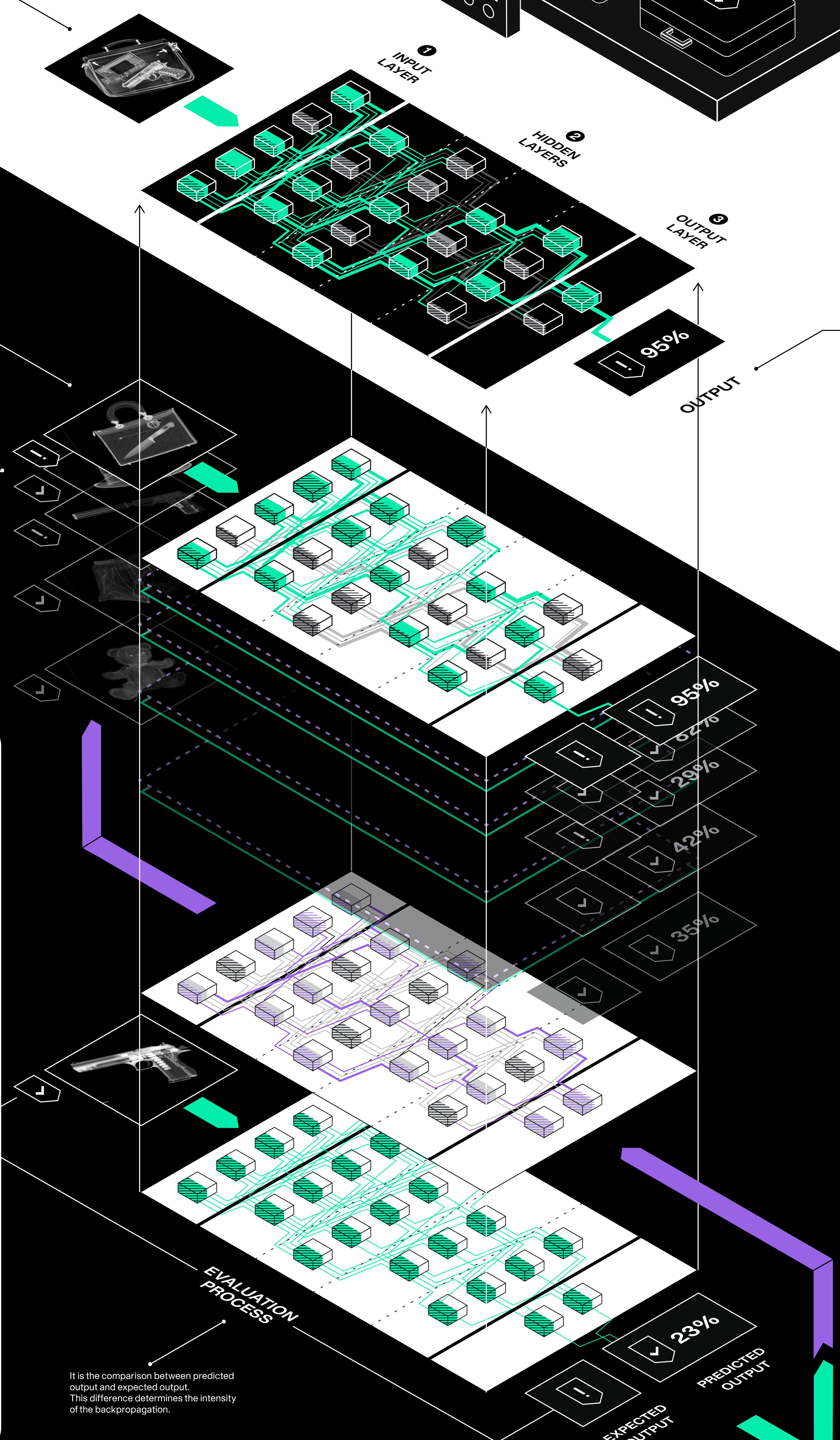


INPUT IMAGE
The image that will be decomposed in order to be submitted to the network.

NEURON DISPOSITION
The disposition of the neurons in the network can vary depending on the task. The number of neurons and the shape of the connections is what the developers choose, usually starting from a set of standard models.

TRAINING DATA
The dataset is composed of thousands of images classified as "safe" or "not safe"; the machine detects, without knowing the label, which category an image belongs to. The set of algorithms in which we use a labeled dataset is called supervised learning.

LABEL
A label expresses one or more feature of the image. It is usually assigned by humans. During the training this tag is hidden from the network, it will be used to check the output during the evaluation phase.



1 INPUT LAYER
It is the first layer of the neural network which passes the raw information to subsequent layers without performing any computational tasks.

2 HIDDEN LAYERS
The hidden layer consists of one or several layers and acts as the connection between the input and output layer. These layers perform all the computational work.

3 OUTPUT LAYER
The output layer is responsible for producing the predicted output of neural networks.

The output consists of a symbol and a percentage value. The symbol indicates whether the package is considered safe or not, while the percentage indicates the confidence level of the forecast.

TRAINED MODEL

Thanks to the training phase, the model is ready to perform a determined computational task with a high level of accuracy.

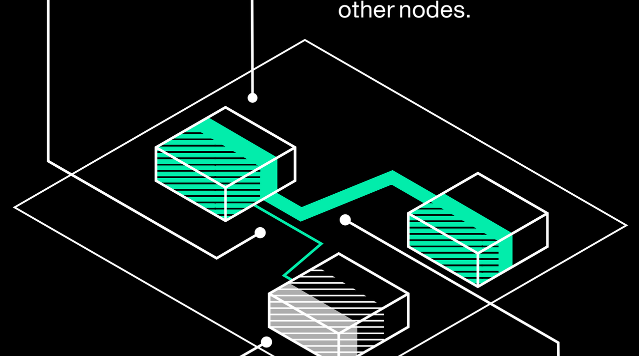
TRAINING PHASE

During the training phase, a neural network is fed thousands of labeled images, learning to classify them.

FORWARD PROPAGATION

CONNECTION:
It is the link between neurons in which the signal passes through.

NEURON
It is the basic unit of computation in a neural network. The role of the neuron is to receive signal from other nodes.



ACTIVATION FUNCTION
It defines when a neuron fires. It's a threshold: the signal is sent onward only if the aggregate signal crosses it. For each neuron, you choose the one that works best in a particular scenario.

WEIGHT
Each connection has an associated weight. During the learning process it increases or decreases depending on the strength of the signal at a connection.

BACKPROPAGATION

ADJUSTMENTS
During the learning process the thresholds and the weights are adjusted according to the signal it receives from the backpropagation.

LABEL CLASSIFICATION

NOT SAFE

SAFE

It is the comparison between predicted output and expected output. This difference determines the intensity of the backpropagation.

Depending on the output of the evaluation, the signal goes backwards into the network adjusting the weights and thresholds until training data with the same labels consistently yield similar outputs.

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Neural networks

AUTHORS

Mattia Casarotto
Davide Chiappini
Andrea De Simone
Emanuele Ghebaur
Francesca Gheli
Hanlu Ma
Raffaele Riccardelli

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DENSITY DESIGN

POLITECNICO MILANO 1863
SCHOOL OF DESIGN

HOW TO UNDERSTAND MILLIONS OF WORDS IN A MINUTE

SENTIMENT ANALYSIS

"The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral."

Oxford Dictionaries

ABOUT

Sentiment Analysis allows estimation of proportions for categories in a target population without classifying each individual document.

Key advantage is its flexibility:
1. Repetition both over and in real-time
2. No need to encode every word manually
3. Results adapt to pattern changes.

PROCEDURE

The procedure starts with the construction of the training set. With the results from the training set (T), the key information is used to begin the sentiment analysis process (S).

1. T → 2. S

DATA EXTRACTION

Sentiment Analysis begins with gathering relevant text data of the topic. Some possible sources are:

- online articles
- blogs
- social media
- customer reviews

DATASET

all collected data

T *(100-500 documents)*

S *(500.000 documents)*

THE CONSTRUCTION OF THE TRAINING SET

START
THE EXAMPLE

1 CATEGORY DEFINITION

Label each text document into categories. Documents must be divided into so that:
- No document belongs to several categories
- No uncategorized documents

Ex. label the texts "positive" or "negative". Off-topic documents, different languages, and spam are removed. This is done manually by data scientists.

Negative sentiment

Positive sentiment

Discarded documents

2 TEXT PREPROCESSING

Transform collected texts into data variables to be computed by simplifying text into a short list of meaningful unigrams. This is done automatically through software.

Meaningful words

Off-topic words

Punctuation

Stopwords

"...stayed here few years... again an extraordinary adventure"

"We really love this awesome place, it's always a nice experience. We will certainly come here again"

Nice Person, 4 days go

3 UNIGRAM PREPROCESSING

Lower-case, and lack of punctuation. Documents are converted into unigrams

Unigrams: A one word sequence
Ex. Running or runner are reduced to RUN
Unigrams deletion: if appears in fewer than 1% and more than 99% of all documents

4 UNIGRAM COUNTING

A set of polarized unigrams *(ex: good VS bad)* are chosen to regulate positive and negative sentiment ratio of the training set. Unigrams are counted on the entire training set and both categories.

UNIGRAM-SET

UNIGRAM A "GOOD" · UNIGRAM B "AWFUL" · UNIGRAM C "FINE" · UNIGRAM D "BAD" · UNIGRAM E "GREAT"

Unigrams ocurrence on negative documents: 40%, 80%, 20%, 70%, 18%

Unigrams ocurrence on positive documents: 80%, 10, 20%, 20%, 70%

5 RESULT OVERVIEW

This determines:
- Overall proportion of positive and negative documents in the traing set sample.
- Occurrence and proportion of each unigram in each category.

This is a key information to perform the sentiment analysis

EXAMPLE: UNIGRAM D "BAD" appears in:

70% — 70% of the positive documents

20% — 20% of the negative documents

Finished Training Set T — NEXT STEP — Sentiment Analysis S

SENTIMENT ANALISYS

1 TARGET DATASET ANALYSIS

The algorithm analyzes the target dataset and checks all of the documents containing all the unigrams.

UNIGRAM D "BAD" 33%

Unigram D appears in 33% of the Dataset

2 ESTABLISHING ASSUMPTIONS

- Sum of all positive and negative documents equals total documents in the dataset
- Sum of positive and negative documents with unigram equals all documents with unigram in dataset

POSITIVE DOCS + NEGATIVE DOCS

(POS) + (NEG) = 1

%(POS) + %(NEG) = %

3 TAKING VALUES OF THE TRAINING SET

UNIGRAM D "BAD": 20% → 0.2*(POS); 70% → 0.7*(NEG)

$0.2*(POS)+0.7*(NEG) = 0.33 \longrightarrow 0.2*(POS)+0.7*(1-POS) = 0.33$

4 ADJUSTING PROPORTIONS

The algorithm adjusts the proportions of the categories until the number of the documents that contain a unigram will match the dataset, while the sum of the proportions stays the same.

Meaning:

Percent of documents with the unigram in each category is fixed

Proportion of categories is dependent on the presence of the unigram in the dataset

What if we had different datasets?

*What if unigram "Bad" appears in **33%** of the dataset?* — 33%

*What if unigram "Bad" appears in **50%** of the dataset?* — 50%

The algorithm adjusts proportions to match...

The algorithm finds an optimal match

74% · 26% — *That means that **74%** of the documents are **positive** and **26%** are **negative***

40% · 60% — *That means that **40%** of the documents are **positive** and **60%** are **negative***

*The results are different. **The more often unigram "Bad" appears, the more negative documents exist.***

RESULTS

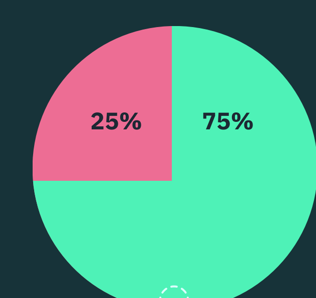
By repeating this process, it's possible to find with high accuracy the ratio of positive and negative documents of the whole data set. The more Unigrams analyzed, the higher accuracy ratio.

UNIGRAM D "BAD": 33% — 74% · 26%

UNIGRAM E "GREAT": 57% — 75% · 25%

UNIGRAM A "GOOD": 70% — 75% · 25%

DATASET POSITIVE AND NEGATIVE SENTIMENT



The sentiment analysis of dataset = ratios found by unigram

VISUAL EXPLANATIONS OF STATISTICAL MODELS

Sentiment Analysis

AUTHORS

Anel Alzhanova
Soraya Astaghforellahi
Maria Camila Coelho
Beatrice Foresti
Yaqing Luan
Nelly Saad
Severin Alois Schwaighofer

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Andrea Benedetti
Tommaso Elli
Beatrice Gobbo

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DENSITYDESIGN

POLITECNICO MILANO 1863

PAGE RANK

Become part of the algorithm, discover its functioning by finding the most relevant pages by yourself!

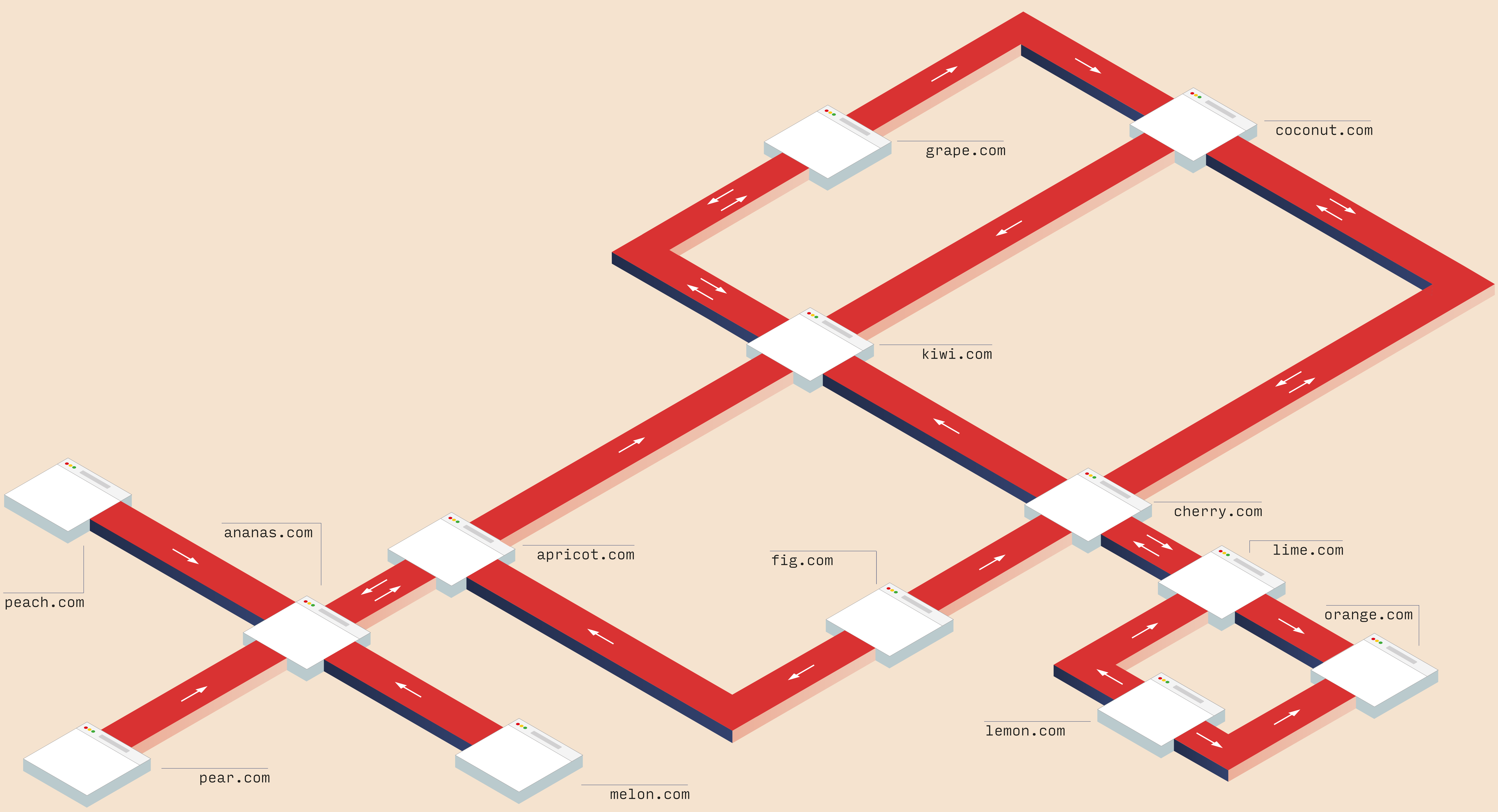
Where do Random Walkers meet?

Can an algorithm predict the most important pages of the web? Google's Page Rank can, and so can you! Page Rank estimates the probability of a web page to be visited by a user. It makes an evaluation of the most relevant pages based on their links with the other pages.

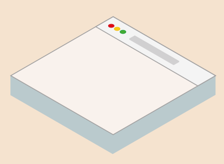
A "Random Walker" is also considered in the process: hypothetical "drunk" users of the web randomly moving between the pages, always ending up on the best linked ones. These are the most relevant pages and they will be the first results proposed by the search engine.

Rules of the game

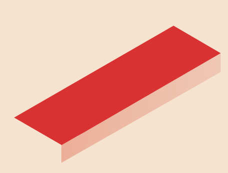
1. Take a sticker strip of random walkers from the pocket
2. Pick any web page as your starting point
3. Move between the pages following the directions on the paths and paste a sticker on each web page you pass through
4. If you can't move to any other directions,don't worry! If you still have other stickers you can choose another web page and restart from it!
5. When you have finished your stickers,make a step back and look at the poster! Where is the most of the random walkers? Try to guess which pages are the most relevant one.
6. Reveal the second layer by turning the handle. When you have read the results, please turn it again in order to put the other layer back up for the next player.



Legend:



Web Page



Link

Link Direction

Random Walker

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Page Rank Algorithm

AUTHORS

Letizia Agosta
Aurora Antonini
Martina Bombardieri
Elena Busletta
Camilla De Amicis
Federico Lucifora

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DEN-
SITY
GN+

POLITECNICO
MILANO 1863
SCHOOL OF DESIGN

PAGE RANK

Become part of the algorithm, discover its functioning by finding the most relevant pages by yourself!

Where do Random Walkers meet?

Can an algorithm predict the most important pages of the web? Google's Page Rank can, and so can you! Page Rank estimates the probability of a web page to be visited by a user. It makes an evaluation of the most relevant pages based on their links with the other pages.

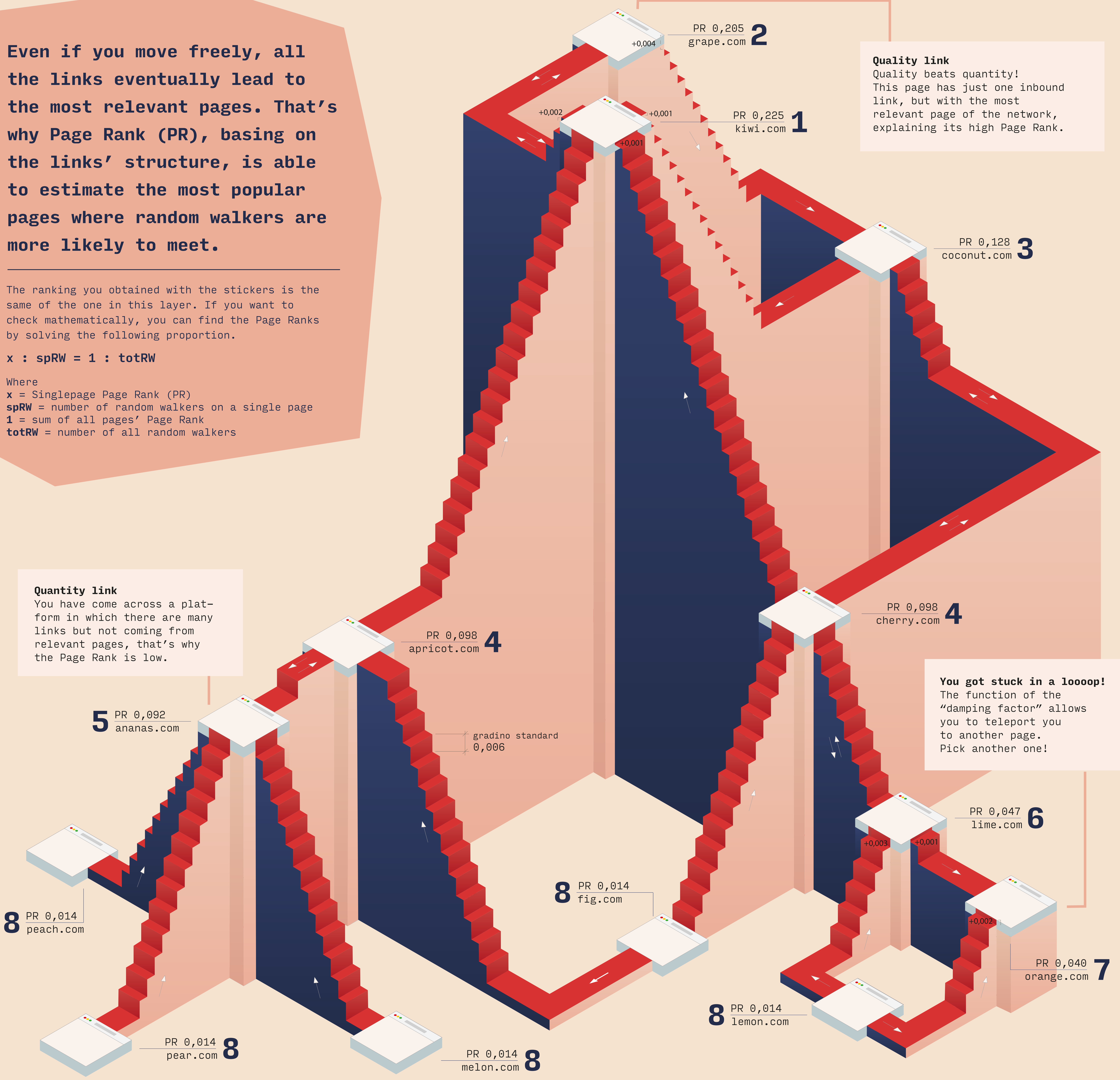
A "Random Walker" is also considered in the process: hypothetical "drunk" users of the web randomly moving between the pages, always ending up on the best linked ones. These are the most relevant pages and they will be the first results proposed by the search engine.

Even if you move freely, all the links eventually lead to the most relevant pages. That's why Page Rank (PR), basing on the links' structure, is able to estimate the most popular pages where random walkers are more likely to meet.

The ranking you obtained with the stickers is the same of the one in this layer. If you want to check mathematically, you can find the Page Ranks by solving the following proportion.

x : spRW = 1 : totRW

Where
x = Singlepage Page Rank (PR)
spRW = number of random walkers on a single page
1 = sum of all pages' Page Rank
totRW = number of all random walkers



Legend: Web Page — Link — Link Direction — Random Walker

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Page Rank Algorithm

AUTHORS

Letizia Agosta
Aurora Antonini
Martina Bombardieri
Elena Busletta
Camilla De Amicis
Federico Lucifora

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DEN-SITY GN+

POLITECNICO MILANO 1863
SCHOOL OF DESIGN

Hierarchical Clustering

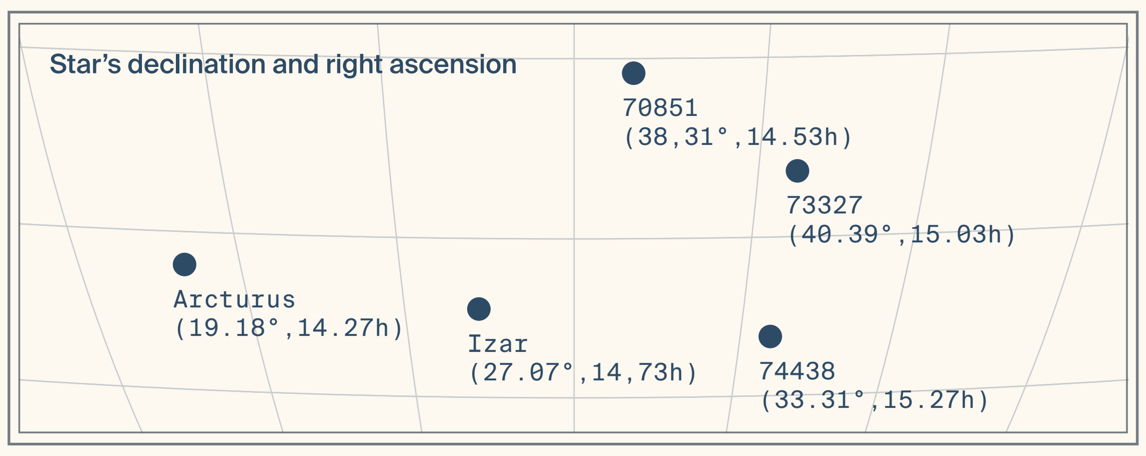
In search of data-driven constellations

Since time immemorial, mankind has looked at the sky, gazing at heavenly bodies and connecting the closest ones to draw figures in the sky. This is how constellations originated. What would constellations look like if, instead of humans, it was an algorithm that searched for patterns in the celestial vault?

Hierarchical clustering is a method used in descriptive statistics to determine a **hierarchy of clusters**, which are collections of samples **based on similarities** among their features. It is an **unsupervised learning** method, this means it reveals spontaneous patterns found in the dataset instead of relying on human-defined groupings. For this reason, it's the perfect tool to find new data-driven constellations.

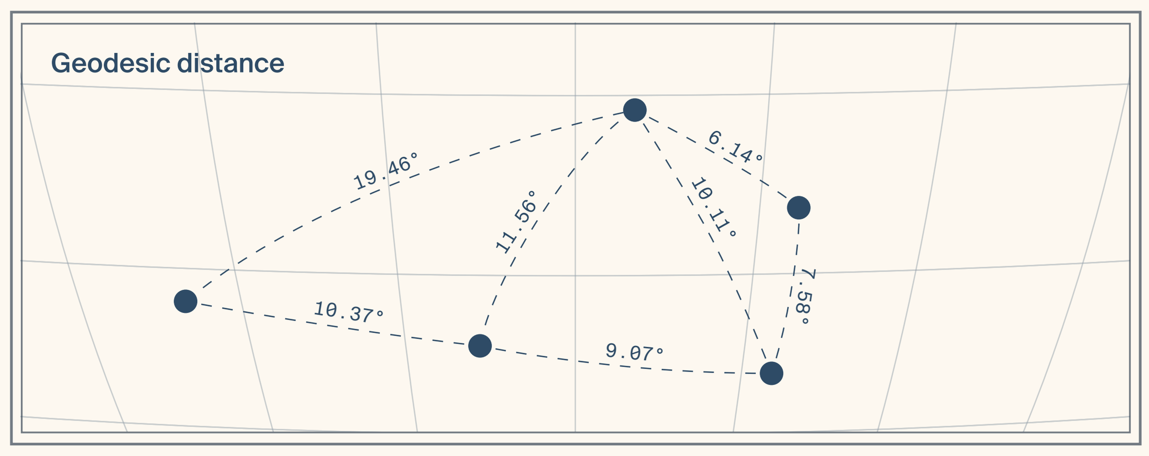
Key concepts of hierarchical clustering

Observation

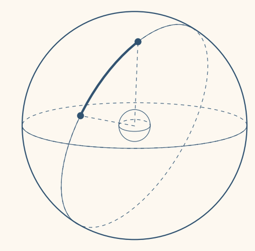


When looking for similarities, first we choose which features to compare and then we make sure these features are commensurable, to assess their similarity. For example, we could group stars according to their luminosity or hue. In our case, we consider *declination* and *right ascension* as features to compare; they are the **spherical coordinates** of any star on the celestial vault, similar to longitude and latitude on Earth.

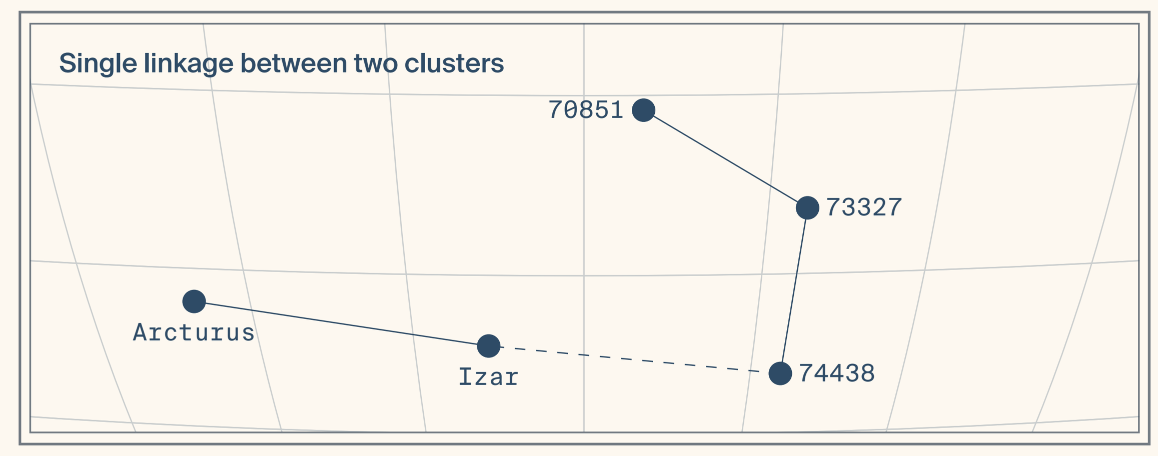
Distance



The similarity of two samples can be quantified in different ways according to the nature of the data being analised and to the scope of the analysis. The aim of distance measure is to find similar data objects and to group them in the same cluster. In this example, we will be using **geodesic distance** which measures distance along a non-flat surface.



Linkage



Since clusters contain multiple data objects, we have different options to measure distance, e.g. between the centres of each cluster, between the furthest points of each cluster. In our case, we use the **single-linkage criterion**; it states that two clusters are as similar as their most similar elements — or as close as their closest stars, in our example.

Legend

Groups of circles and lines are **constellations** and represent a cluster

Circles represent stars

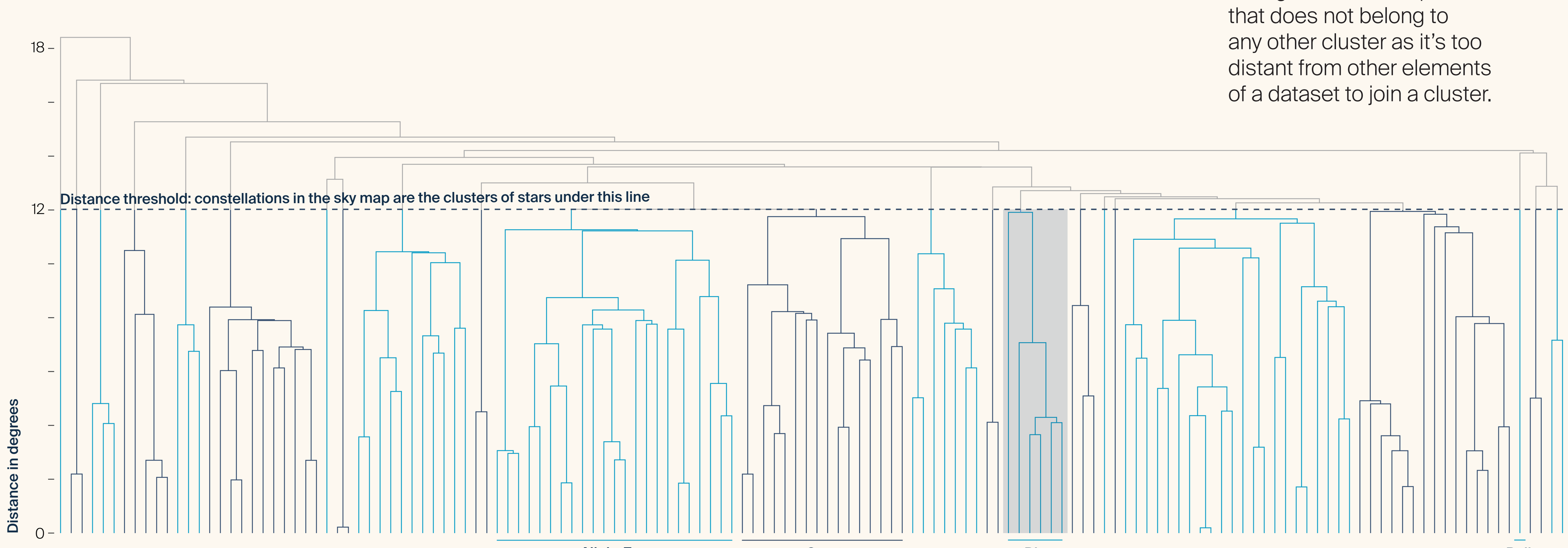
Lines link stars in the same constellation

Mind the gap

These stars are 12.02° apart, since our pruning threshold is 12° these two constellations are not clustered together.

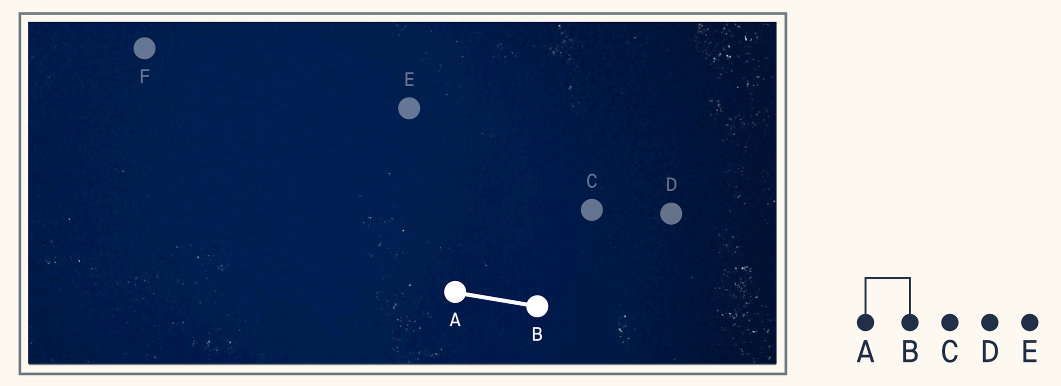
Singleton

A singleton is a sample that does not belong to any other cluster as it's too distant from other elements of a dataset to join a cluster.

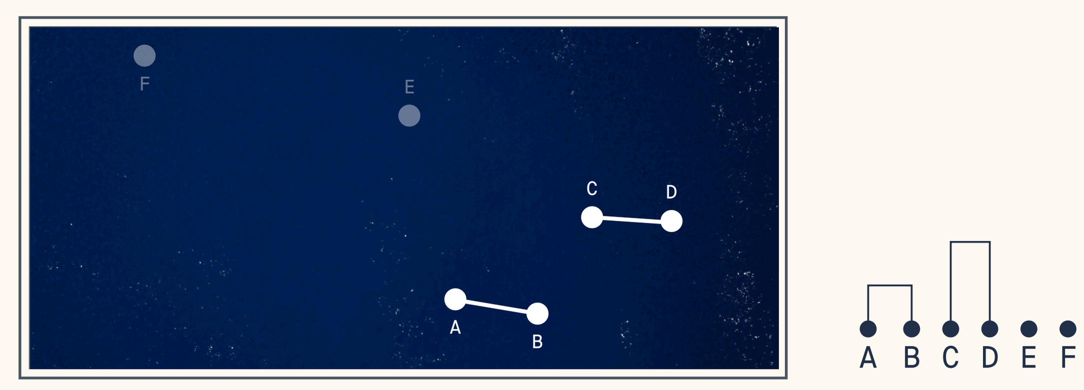


Building a hierarchy

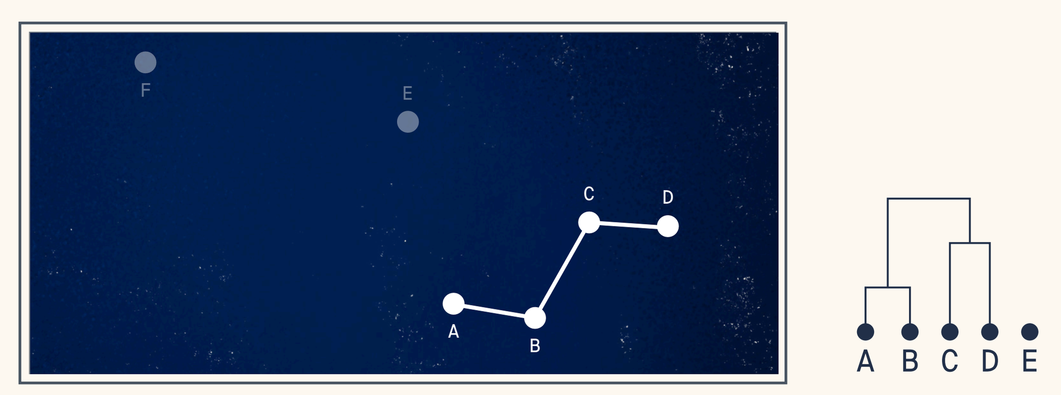
Here we illustrate how hierarchical clustering can be applied to our example, while visualising the results through a dendrogram, a diagram that codifies information about our samples' similarity and their hierarchy.



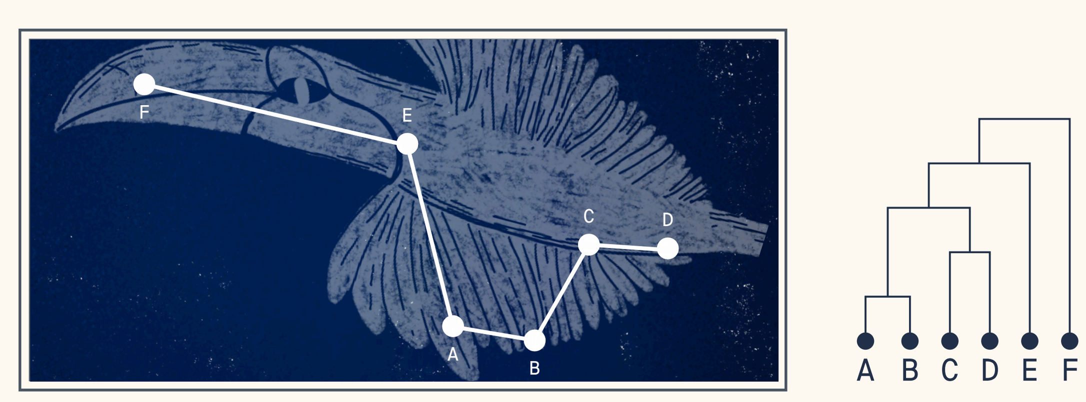
1. The distance between every possible pair of stars is computed. Once the two closest stars are found, they are grouped into a cluster. On the dendrogram the samples are connected by a line, this means they have been clustered together.



2. The previous step is repeated, a new cluster is formed and drawn on the dendrogram. We can see that the vertical lines of this cluster have a different height as it is proportional to the distance between the two stars.



3. Distances are computed once more but now the two closest elements are not stars but the two clusters that have just been formed. Looking at the dendrogram on the right we can see how a hierarchy is starting to form.



4. The last data point can now join the clusters to its left. Their order on the horizontal axis does not codify any information about the analysis; most often it is simply the order that maximises the dendrogram's readability.

Pruning

At the end of the process, all 142 stars are collected in a single, large cluster and a hierarchy is defined for the entire dataset. In our case, we would end up with one large constellation, while we need to define multiple, distinct ones instead.
The selection of clusters is called pruning, from the idea of cutting branches off the dendrogram and picking the resulting subtrees as clusters. The pruning criterion depends once again on the scope of the analysis. In this example, we set a distance threshold of **twelve degrees** and selected the twenty-three resulting subtrees as constellations.

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Hierarchical clustering

AUTHORS

Alice Bocchio
Michele Bruno
Maria Celeste Casolino
Luca Draisci
Virginia Leccisotti
Barbara Roncalli
Sara Zanardi

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

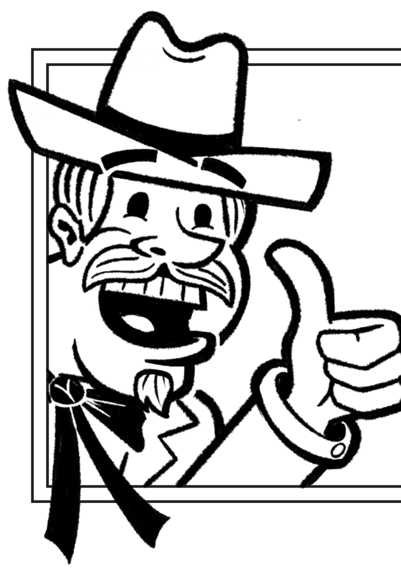
LM in Communication Design
A.A. 2021/2022

DENSITY DESIGN

POLITECNICO MILANO 1863
SCHOOL OF DESIGN

ALL ABOUT CONTROL CHARTS

How statistics can set your chicken farm to be a successful one



WHY YOU NEED CONTROL CHARTS

Control charts are visual tools used to monitor processes. They can detect possible issues of a production chain, allowing operators to take action and ensure consistent quality of the products. Here is my exclusive guide on how to apply one to improve your chicken farm's efficiency. Trust me, I've been in this business since 1924.

HOW TO BUILD ONE

1 LEARN FROM THE FINEST

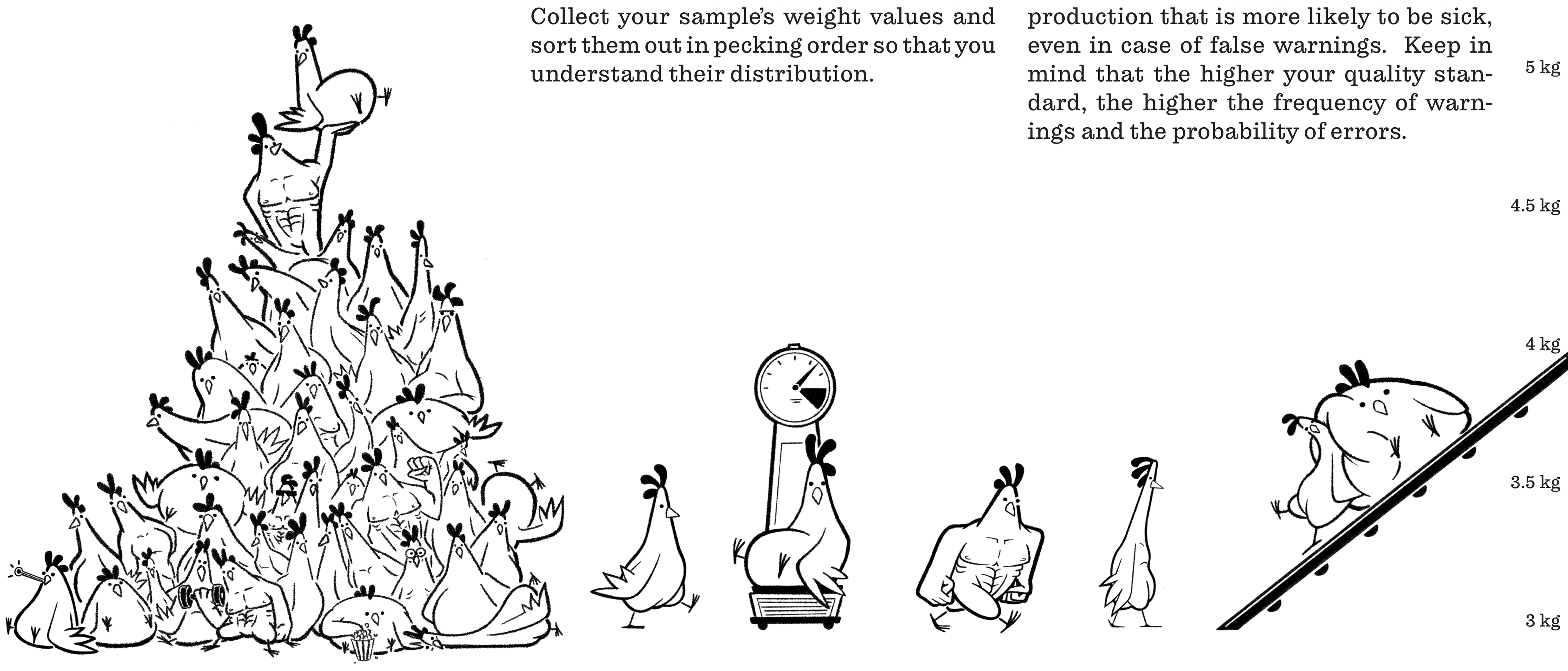
You first need to select a sample of your production that satisfies your desired quality level. A sample of chickens that are healthy and ready to be sold will be your starting point of this first phase.

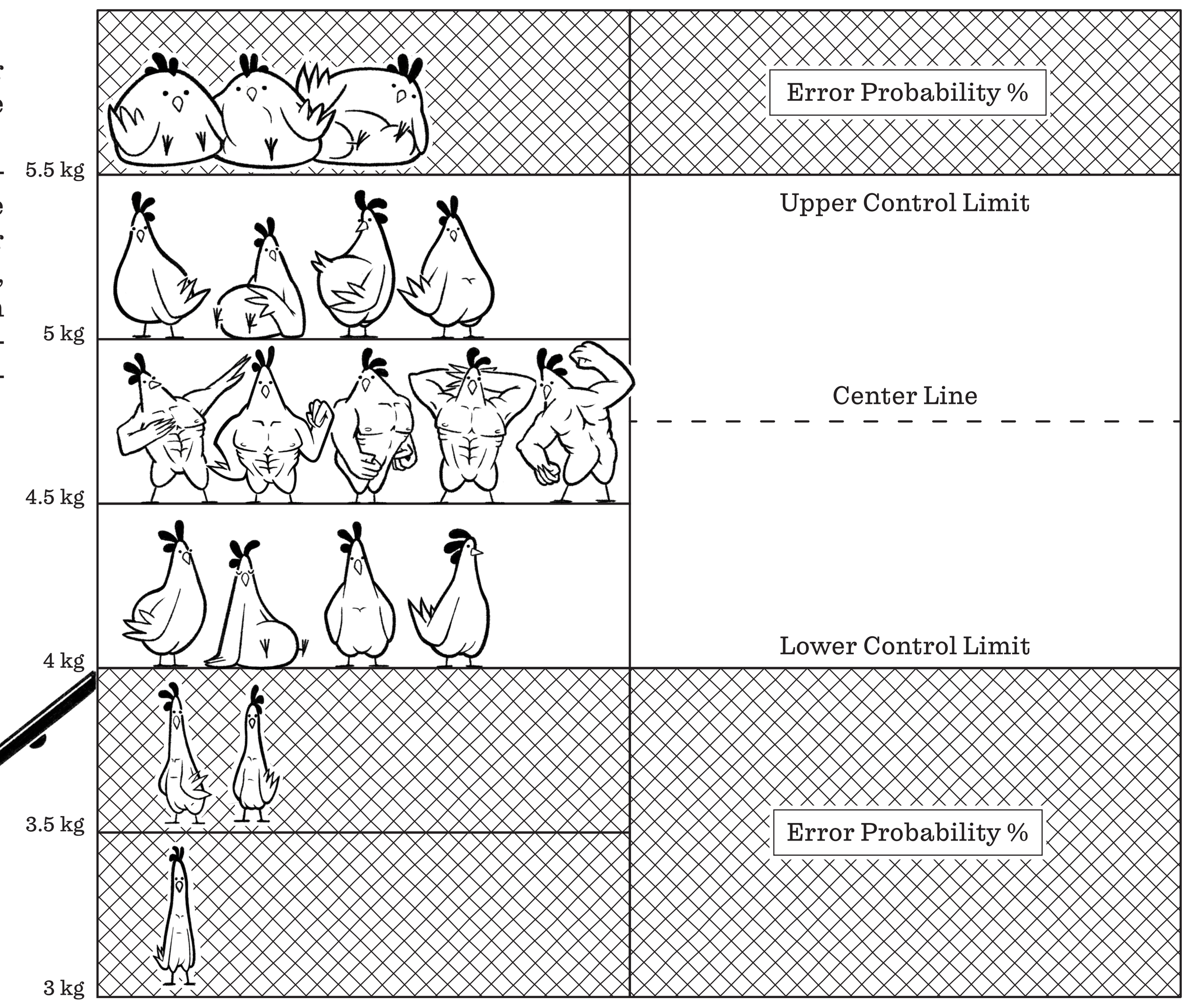
2 CHOOSE A PARAMETER

Now you want to determine a feature that will allow you to quickly evaluate your production. From my experience in the poultry business, I know that the most efficient way to monitor the chickens' health is to keep an eye on their weight. Collect your sample's weight values and sort them out in pecking order so that you understand their distribution.

3 SET YOUR GOLDEN RULE

Even though all the chickens from your sample are healthy, this might not be the case in your future production.
You want to set limits that exclude chickens that may not be suitable for the market, discarding a percentage of your production that is more likely to be sick, even in case of false warnings. Keep in mind that the higher your quality standard, the higher the frequency of warnings and the probability of errors.





HOW TO USE IT

4 LET'S GET GOING!

Now the chart is ready to be used to monitor your production! From now on your chickens' weights will be collected and displayed on the chart over time.

In-Control Point

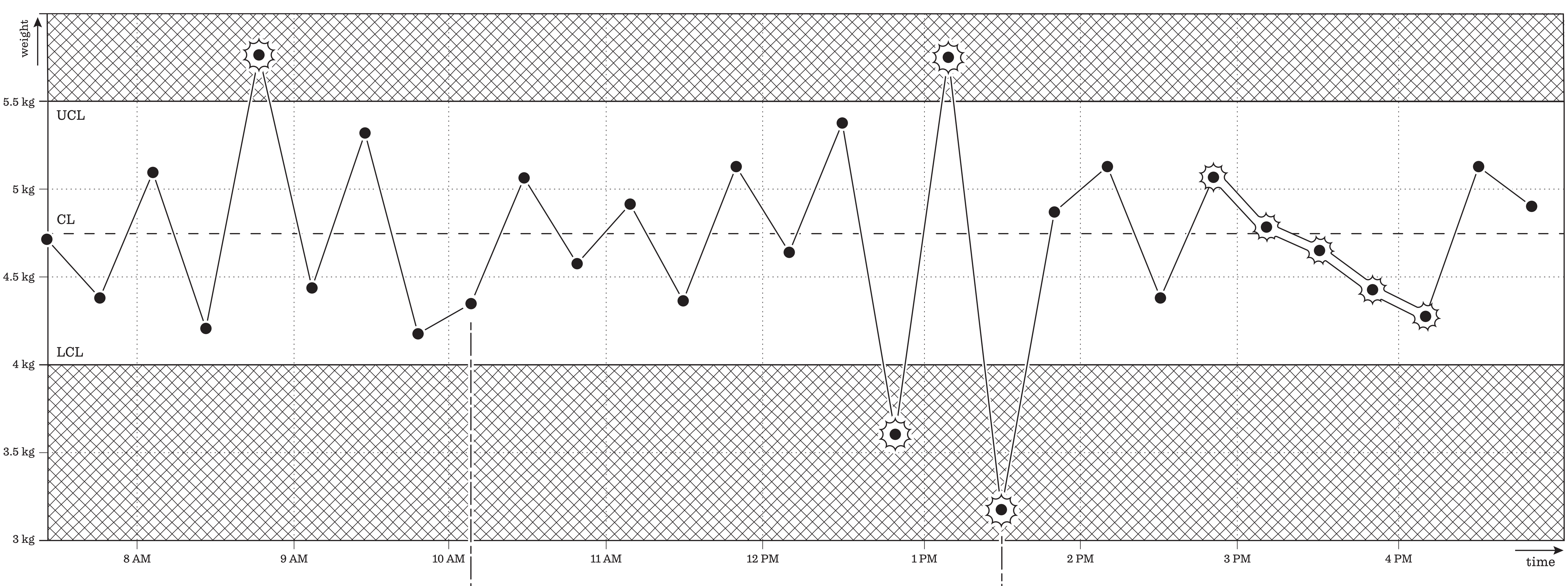
When a value falls inside the desired range, we say that it is "in control". When the graph shows a series of in-control points randomly distributed, it means that the process is stable and you are getting the expected results.

Out-of-Control Point

An out-of-control point is identified when a data value point falls outside the control limits. In this scenario, the process is subject to some kind of unwanted behaviour (special cause variation), and it needs to be inspected.

Unstable Trend

You can also have an out-of-control situation if the points are inside the thresholds. This happens whenever your data points start forming a pattern that is no longer random.



MEANWHILE AT THE FARM...

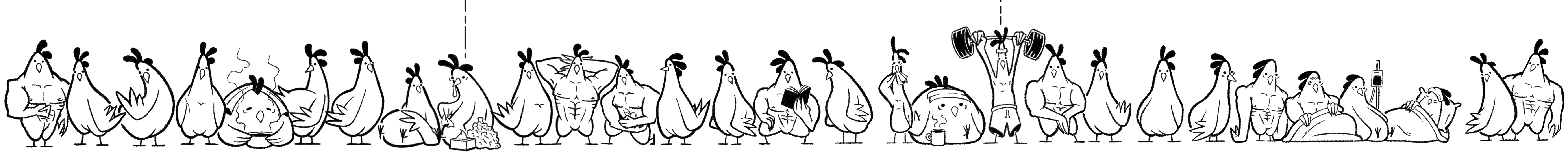
Always remember that the control chart doesn't show the actual health of the chickens, but only its prediction, based on the weight.

Missed Warning

There's a chance that a data point falls within the control limits when it should be out of the thresholds. The wider the distance between the control limits, the higher the chance of a missed warning.

False Warning

Sometimes there may be some false warnings: this happens when the chart shows that the process is not stable when it is. This error always ruffles a few feathers.



VISUAL EXPLANATIONS OF STATISTICAL METHODS

Control Charts

AUTHORS

Bharath Arappali
Lorenzo Bernini
Chiara Caputo
Irene Casano
Yasmine Hamdani
Marco Perico
Davide Perucchini

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DENSITY GN+

POLITECNICO MILANO 1863
SCHOOL OF DESIGN

BOOTSTRAP*

HOW TO EVALUATE ACCURACY OF **THINGS**** ABOUT LOTS OF **STUFF*****
WHEN ONLY HAVING **NOT SO MUCH STUFF******

* A statistical algorithm extremely useful when only a small set of data is available. It gives a precise idea of the accuracy of the estimate

** Estimate of statistical parameters
*** A huge population
**** A small sample

START

We want to estimate the mean height of all the blueberry-loving people in Milan, but we only know the height of 10 of them. **How do we get there?**

STEP 01 **SAMPLE**

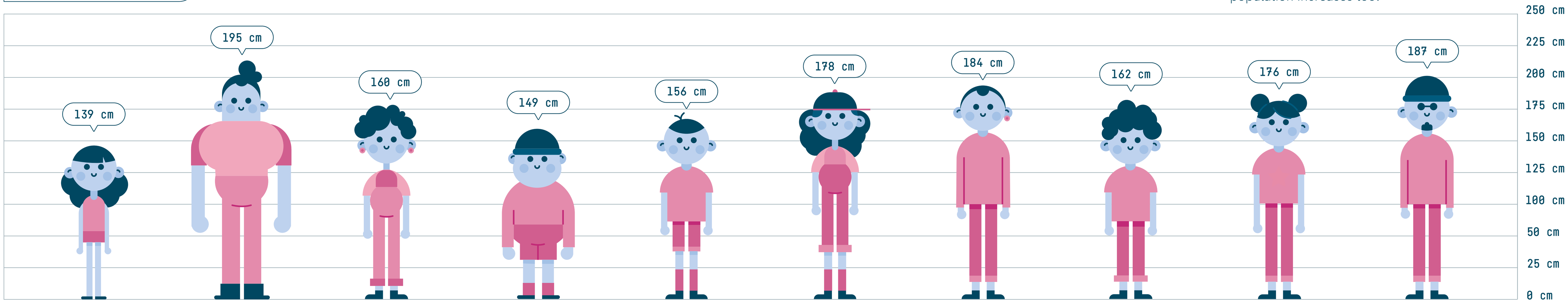
WE TAKE THE DATA FROM THE SAMPLE

As we all know, blueberry-loving people are very *very* shy, so we only managed to tackle 10 of them for our super-duper important research about the link between height and blueberryness. We'll call them our **sample**.

The bigger the obtainable sample, the better the estimate. In our case 10 people will do just fine.

▼ **Law of Large Numbers**
As the sample size increases, its ability to give an accurate representation of the whole population increases too.

STARTING SAMPLE

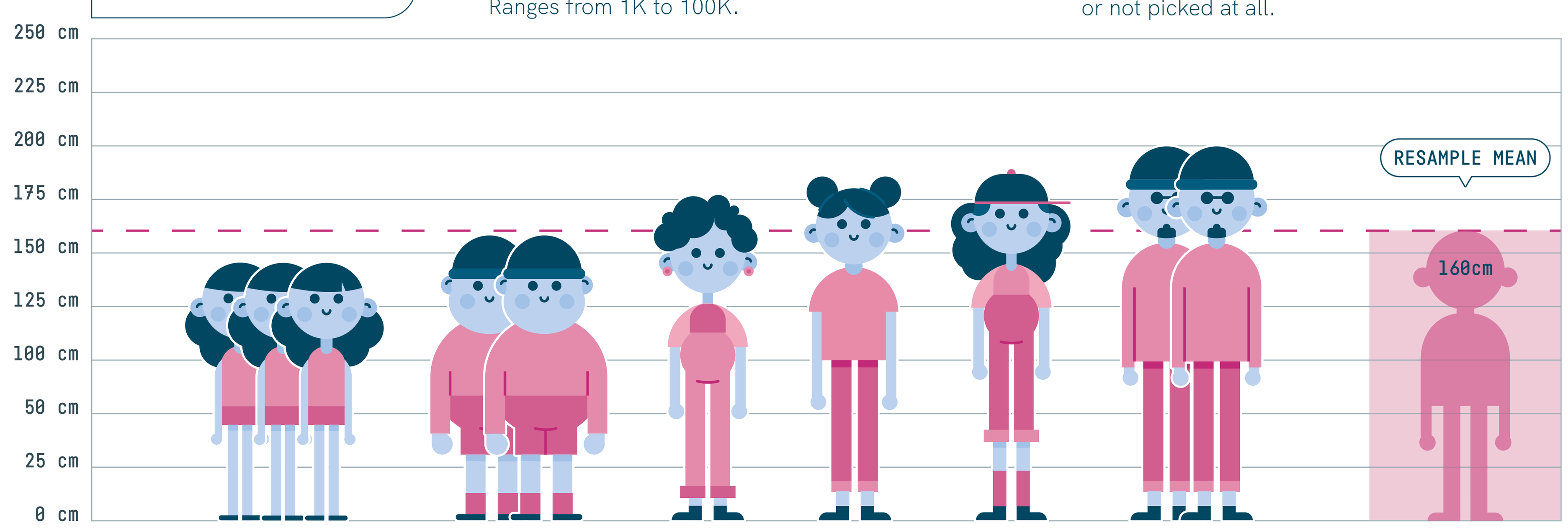


STEP 02 **RESAMPLE**

RESAMPLE #0001

◄ **Recommended number of resamples**
Ranges from 1K to 100K.

▼ **Resampling with replacement**
Replacement means that repetition is allowed, each value can be picked more than once or not picked at all.



WE RESAMPLE OUR DATA AND CALCULATE THE MEAN *MANY MANY MANY* TIMES

We could just calculate the mean of the starting sample, but **we wouldn't have any information about the accuracy of the estimate**.

Instead, we **resample with replacement** many (*many!*) times: each resample is made of 10 height values, randomly picked from the original 10.

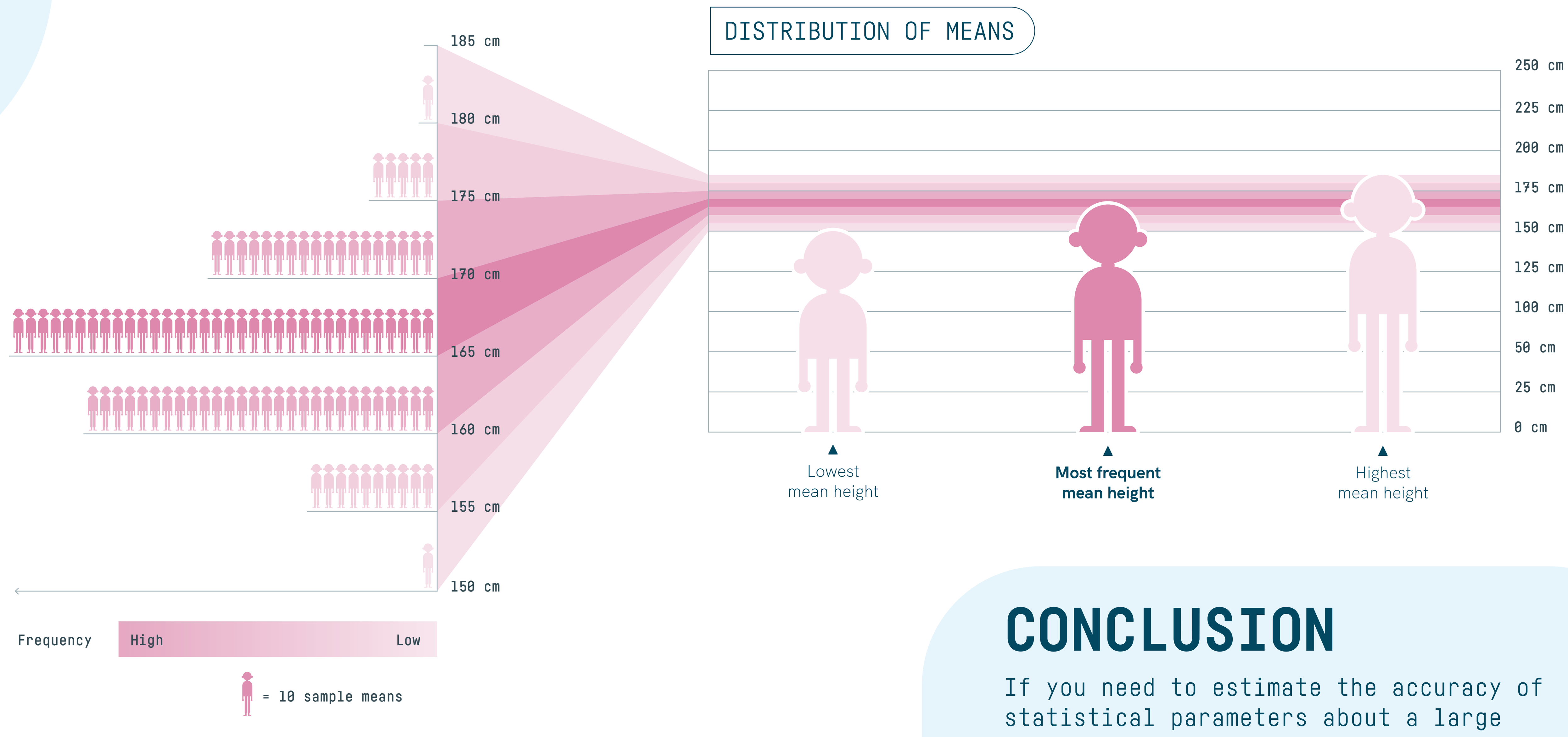
We calculate the **mean height value of each resample**, and we store them away for later. (in case we're hungry).

STEP 03 **DISTRIBUTION**

WE DRAW THE DISTRIBUTION OF THE MEAN HEIGHT VALUES

The distribution shows the **estimated mean height** and the **accuracy of the estimate**. The frequency is expected to be higher for values near the **real population mean**, while lower for further values.

DISTRIBUTION OF MEANS



CONCLUSION

If you need to estimate the accuracy of statistical parameters about a large population, but only have a small sample, the bootstrap algorithm is the way to go.

CAUTION! overconsumption of blueberries might result in your skin turning blue.

VISUAL EXPLANATIONS OF STATISTICAL METHODS
Bootstrap

AUTHORS
Daniele Dell'Orto
Martina Francella
Octavian Husoschi
Martina Melillo
Matteo Pini
Alessandro Quets
Shan Huang

FACULTY
Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS
Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3
LM in Communication Design
A.A. 2021/2022
DENSITY DESIGN
POLITECNICO MILANO 1863
SCHOOL OF DESIGN

the magic of classification trees

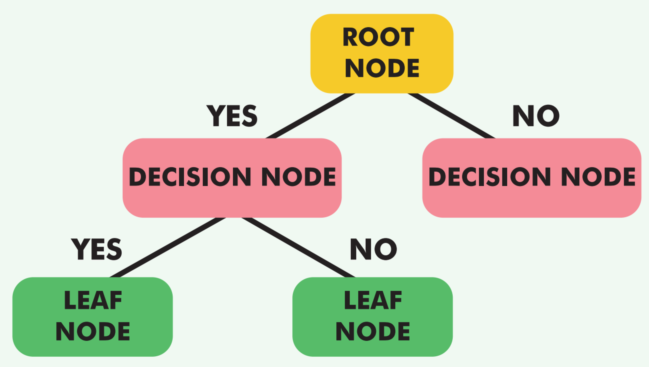
How can machines make accurate predictions? Classification trees can be used to predict possible outcomes to a decision based on observations of a certain item's features.
For example, a classification tree could be used to predict the drinkability of new samples from a water source by comparing certain qualities (e.g., color, pH, and hardness) against a database of previous samples.

Join Wilfred the wizard while he prepares a potion. To finish it, he needs more magical mushrooms. By analyzing the individual features of mushrooms in his vast catalogue of specimens, Wilfred wants to create a classification tree to identify if new mushrooms are magical.

training phase

1 first things first

The goal of a classification tree is to develop a model to predict the category (in this case: magic or non-magic), of an element (mushroom) based on its features or variables (cap color, stem color, and their height) by learning rules inferred by previous data.



The classification starts with a root node and then the algorithm will divide the mushrooms into smaller groups by asking a series of yes or no questions on their features.

2 shroomy dataset

To build the classification tree, we use the following data from the wizard's catalogue.

While most of the data will be used to build the tree, **a portion are kept aside** to be used to check the accuracy of the mushroom magic analyses during the testing phase later on.

cap colour	stem colour	stem height	is it magic?
pink	pink	tall	no
pink	green	tall	yes
green	pink	tall	no
yellow	green	short	yes
pink	green	short	yes
pink	pink	short	no
pink	yellow	short	no
yellow	green	tall	yes
yellow	pink	tall	yes
pink	yellow	tall	no
green	yellow	tall	yes
green	green	tall	yes
yellow	pink	short	yes
yellow	yellow	tall	no

3 make it binary

The model can't understand categorical data (e.g., pink, green, or yellow stem colour) so it must be transformed into **binary information**, by asking yes or no questions.

This way we can prepare Wilfred's data from his observations on the features of his mushrooms for use.

For example, is the cap pink?

cap colour	binary answer
pink	yes
green	no
yellow	no
pink	yes

4 how to split the data?

How does the algorithm decide where to divide the mushrooms? **The algorithm tries to create groups that are as pure as possible** - ideally groups of exclusively magical and non-magical mushrooms.

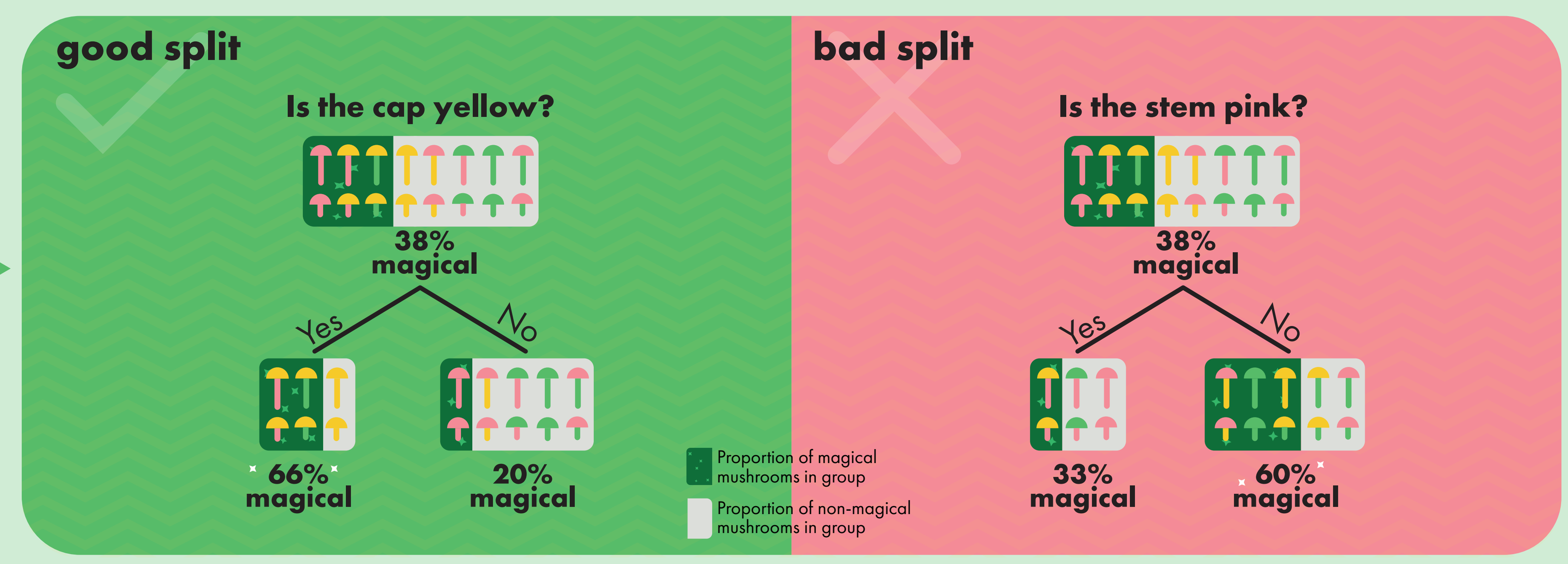
A good split would put all the magical mushrooms in one node and all non-magical mushrooms in another.

A bad split would divide the mushrooms but keep the same ratio of magical and non-magical mushrooms in each group.

By using this logic, the whole tree is created.

Of these two questions, this one gives a better divide.

This is because there are almost exclusively magical or non-magical mushrooms on either side, meaning that a yellow cap is a good indicator of a magical mushroom.



testing phase

5 checking the accuracy

To ensure its accuracy, we must test our tree. To do this, we use the testing mushrooms we put aside at the very start. We want to see how successfully the tree sorts these known mushrooms.

Ideally, the testing mushrooms would give a similar accuracy to the training mushrooms:

100%
TRAINING ACCURACY
Accuracy of predicting initial sample that built tree

95%
TESTING ACCURACY
Accuracy of predicting testing sample of known magicness

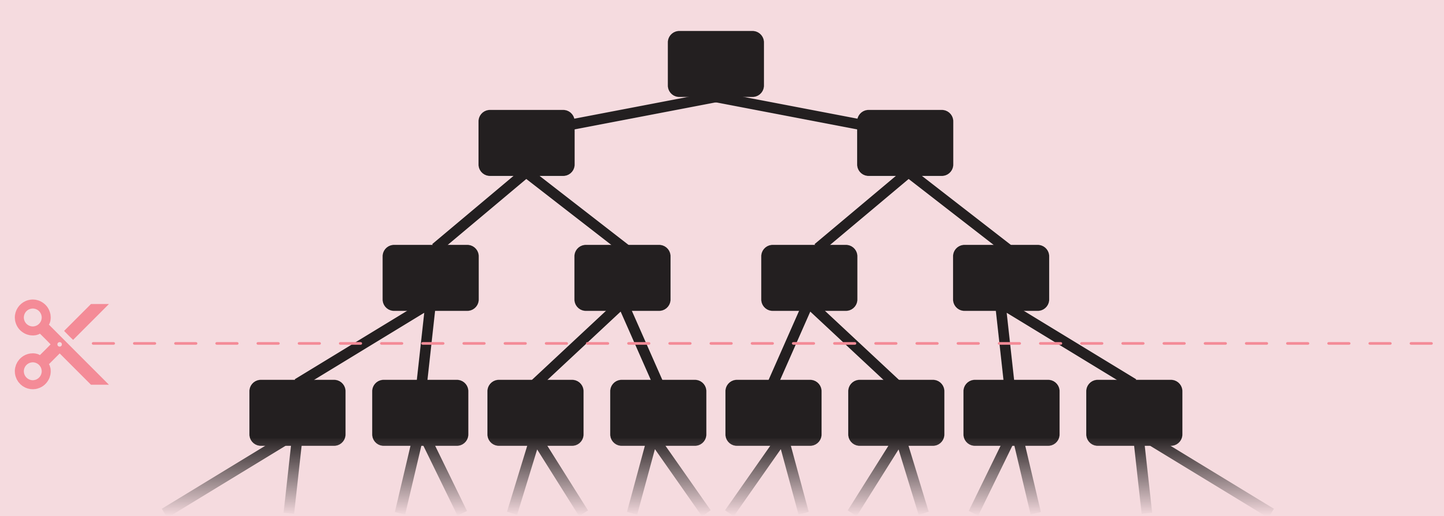
If the testing accuracy is significantly lower than the training accuracy, it means it is **overfit**. This is a problem because it means that the tree is too much adapted to the initial database and won't predict well when the wizard wants to identify a new mushroom.

100%
TRAINING ACCURACY
Accuracy of predicting initial sample that built tree

60%
TESTING ACCURACY
Accuracy of predicting testing sample of known magicness

6 pruning

If the tree is overfit, we can fix this by pruning it. That means that we cut the questions that do not help the tree classify information, and helps reduce its size and complexity.

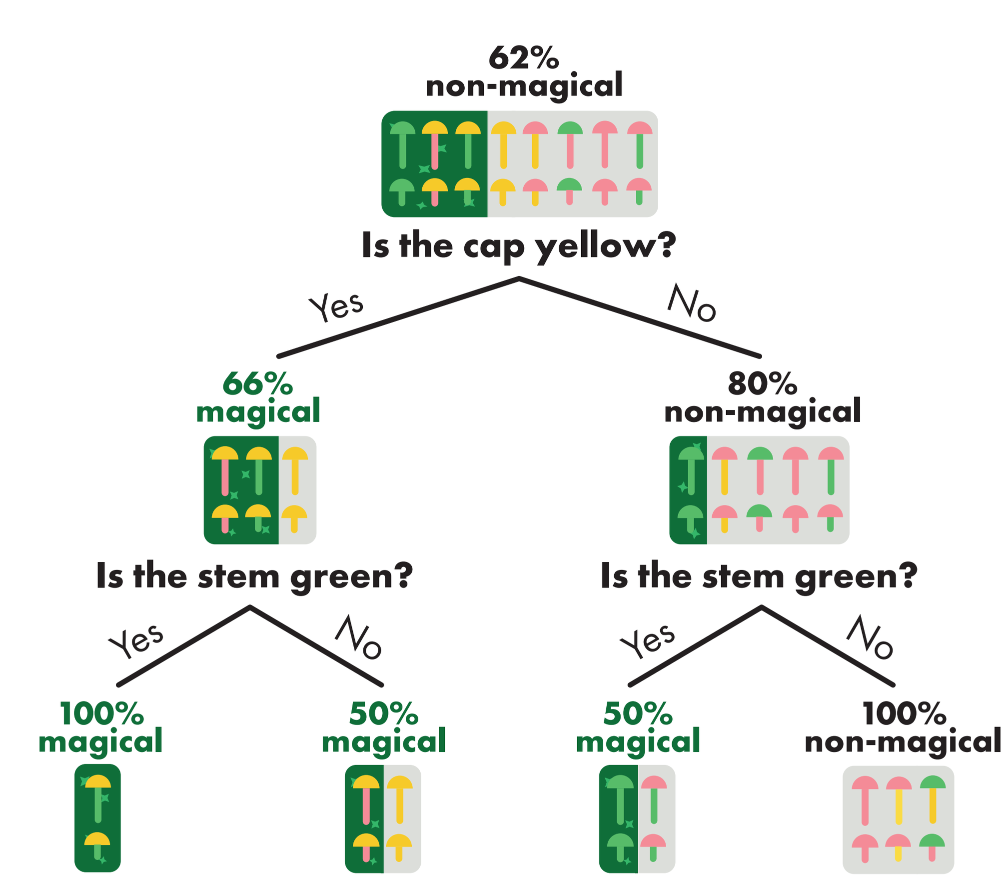


the result

The tree created using the sample data

Proportion of magical mushrooms in group

Proportion of non-magical mushrooms in group

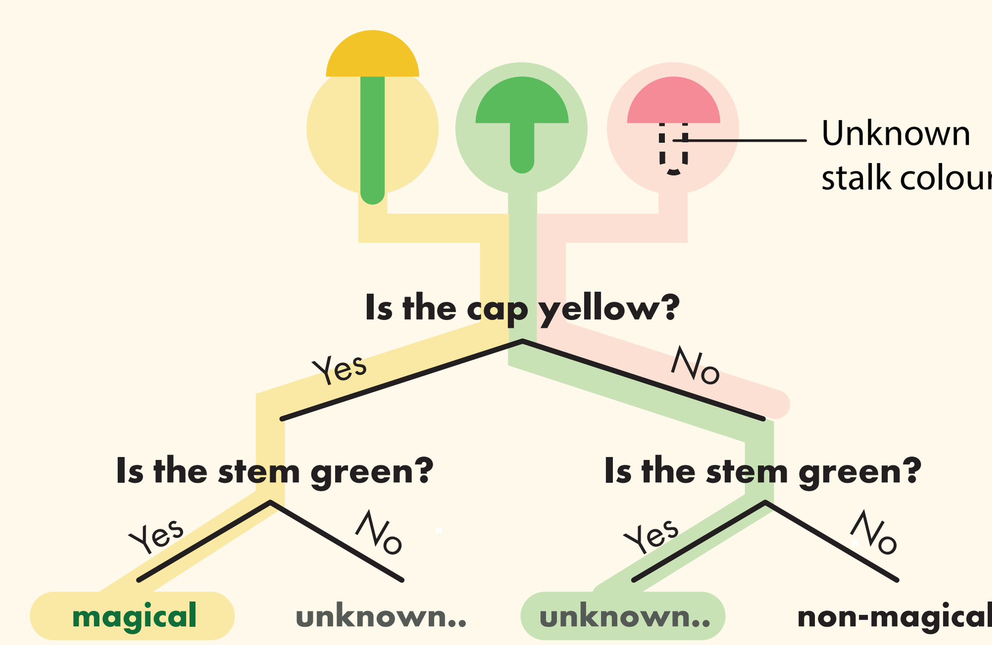


prediction phase

we've found some new mushrooms!

Once the pruning is done, the algorithm is more accurate and is ready to be applied to new unknown mushrooms.

Wilfred wants to discover if his new discoveries are magical or not.



Most likely magical! As the original group had 100% magical mushrooms.

Impossible to tell if magical or not, as the original group only had 50% magical mushrooms.

Most likely non-magical. Classification trees can be used with incomplete data, as the sample can follow the tree for as long as it can. In the original tree, this group had 80% non-magical.

As our original sample was very small, the reliability of this tree would be low. A successful tree needs many more samples.

VISUAL EXPLANATIONS OF STATISTICAL MODELS
Classification Trees

AUTHORS
Marina Fernández
Andrea Llamas
Amanda Cestaro
Leo Gamberini
Regina Salviato
Qi Yu
Renata Martínez

FACULTY
Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo

Final Synthesis Design Studio
Sect. C3
LM in Communication Design
A.A. 2021/2022
DENSITY DESIGN
POLITECNICO MILANO 1863