Course: Introduction to Digital Agro
Lead instructor: Dr. Maria Pukalchik
Lecture: 02

Skoltech Digital Agricluture Laboratory

# Data and basic statistics in agriculture and environmental sciences

Know your data, make conclusions, develop models, build decisions

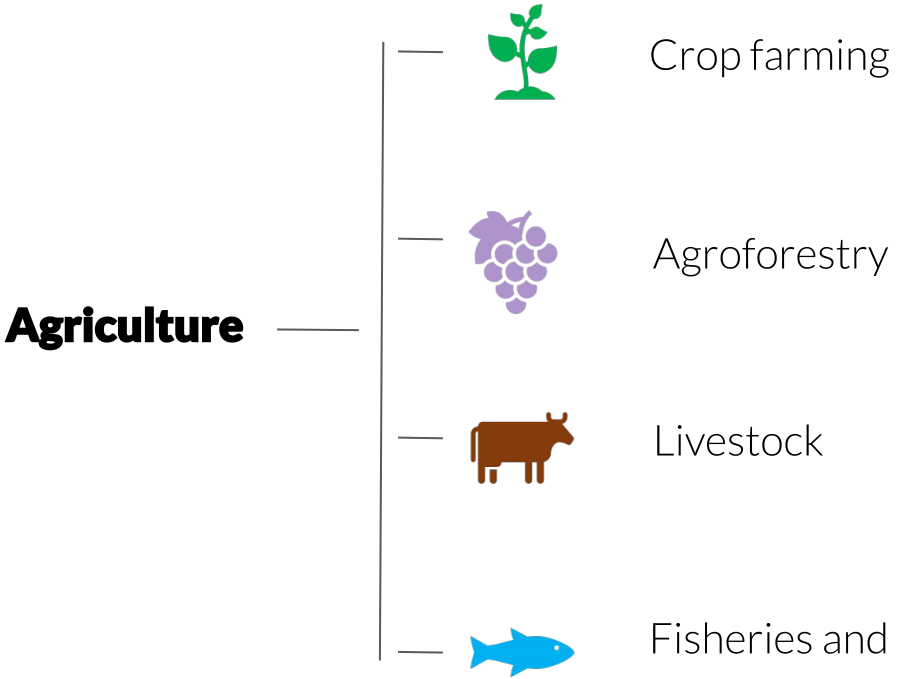
# Themes of the lecture

1. What is the agriculture?
2. What kind of data usually support agricultural studies, examples of R&D projects.
3. Common objectives and tasks of the agriculture: today and the future.
4. Preprocessing of data, description and tools for primary analysis – remember the basics.

# What is the agriculture?



Agriculture is the art and science of cultivating the soil, growing crops and raising animals. It includes the preparation of plant and animal products for people to use – for food, fuel, fiber or medicine (*and their distribution to markets*).




In general, to work on the issues of any agriculture sector, researcher/developer needs to get the information about:

- The object itself
- The environmental characteristics

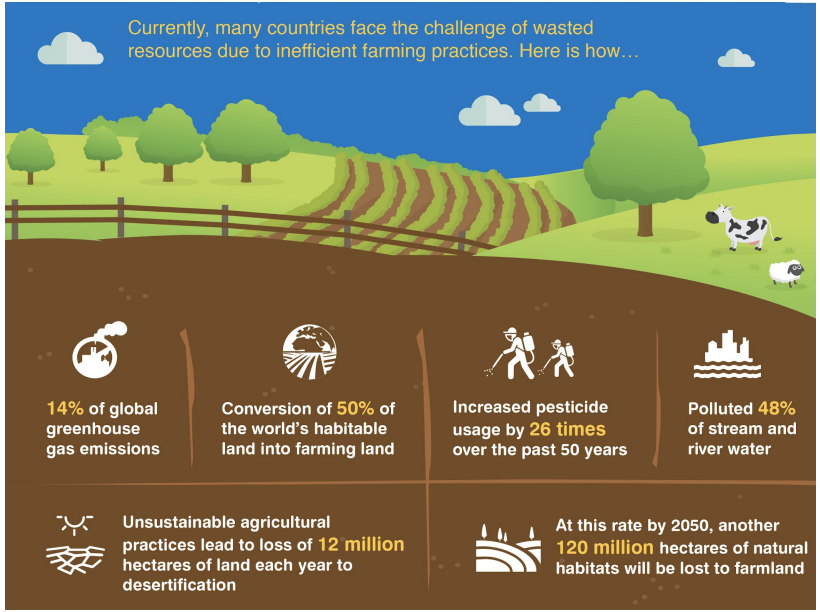
# What is the environment?



In general, 'environment' simply means 'surroundings'.

- ➔ Natural: water (navigable waters, the waters of the contiguous zone, and the ocean waters any other surface water, ground water, drinking water supply), land surface (soil) or subsurface strata, ambient air + community of living organisms. Simply: all natural resources.
- ➔ Man-made: created by human.

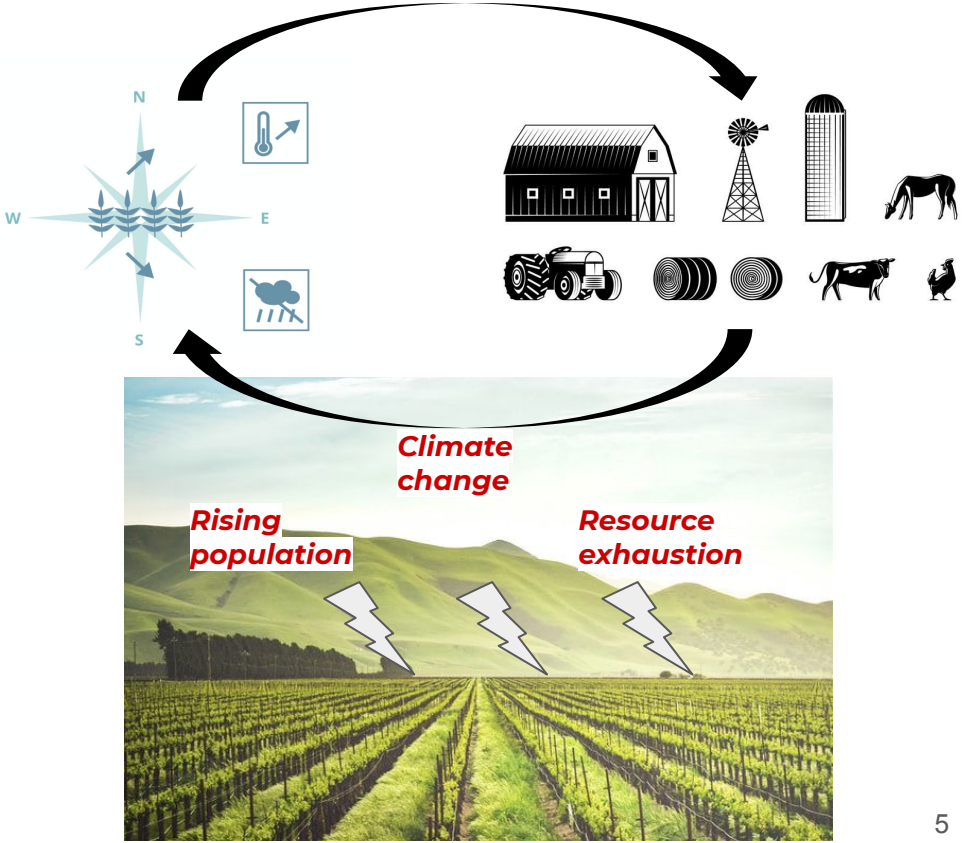
Environmental characteristics are important.



# What is the environment? Examples related to Agro: outdoor farming

- Weather conditions (temperature, insolation, amount of precipitations) and climate in global;
- Soil;
- Availability of water resources;
- Topography;
- Biotic factors: pests,microorganisms.

Mostly uncontrolled and affected each other.



# What is the environment? Examples related to Agro: indoor farming

Man-made conditions in man-made space:

- Temperature,
- Levels of light and shade,
- Irrigation,
- Fertilizer application,
- Atmospheric humidity.

Fully controlled, but provided by usage of mentioned natural resources as well and affecting natural environments.

# Why is this important?



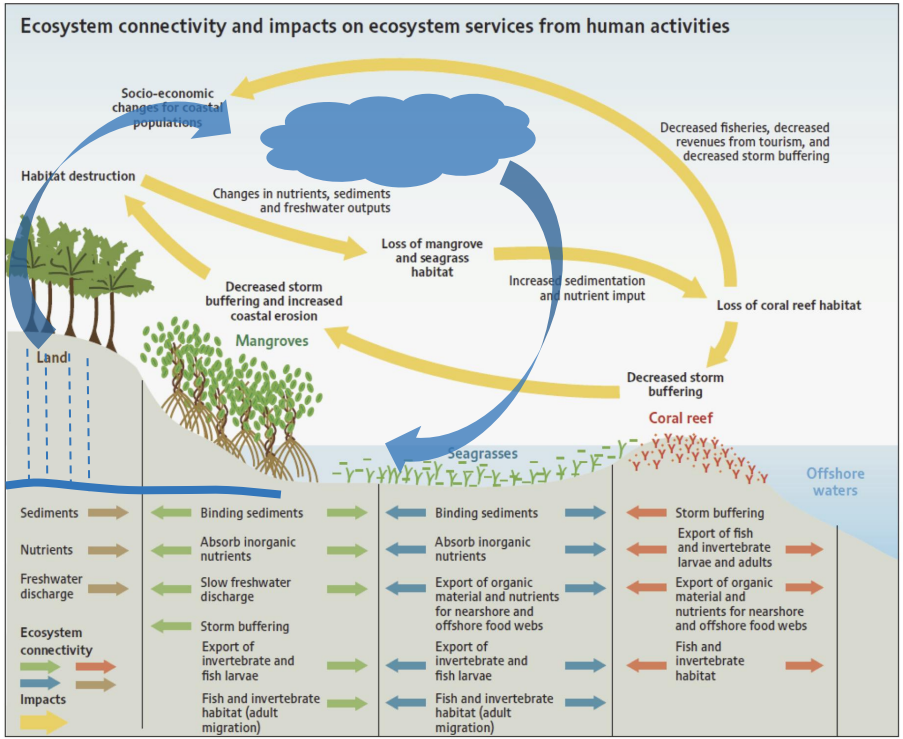



**Figure 1. An example of ecosystem connectivity, showing mangroves, seagrasses and coral reefs.**

From "Approach for reporting on ecosystem services", prepared by *United Nations Environment Programme (UNEP)*

Key points:

- There are many factors are influencing on agricultural activity and management.

Should we consider them all?! (Spoiler: **yes**).

- All of the human activities influence significantly on the natural environment. This impact is mostly **negative**.

**All of agriculture sectors require protection, regeneration and adequate management in all of the agricultural landscapes**

# Strategies of ongoing and future research/R&D projects in the agriculture: focus and tools

# The Data: examples of large projects in agriculture sub-sectors. Crop farming.

- Name of the project: SolACE
- Where, Who, When

*European Union, INRA France, 2017-2021, Supported by EU HORIZON2020*

SolACE

Solutions for improving Agroecosystem and Crop Efficiency for water and nutrient use

- Amount of financial support

*€ 7 192 148*

- Main objective

*"SolACE's overarching goal is to help European agriculture facing the challenge to deal with more frequent combined limitations of water and nutrients in the coming decades, through the design of novel crop genotypes and agroecosystem management innovations to improve water and nutrient (i.e. N and P) use efficiency"*

- Example of tasks

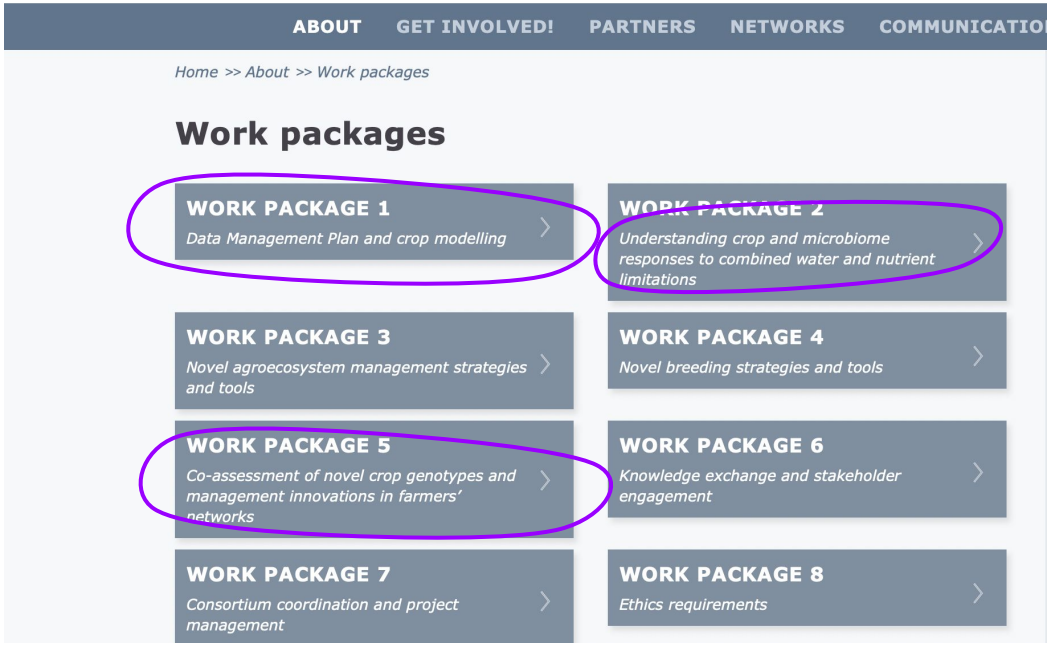
*"Complementary approaches, from data mining, modelling, phenotyping in high throughput platforms and field conditions, to experiments in research stations and farmers' networks in contrasted pedo-climatic zones. Through the co-design and co-assessment with the end-users of the selected novel breeding and management strategies to increase the overall system resource use efficiency".*

https://cordis.europa.eu/project/id/727247

# The Data: examples of large projects in agriculture sub-sectors. Crop farming. SolACE workflow.



SolACE

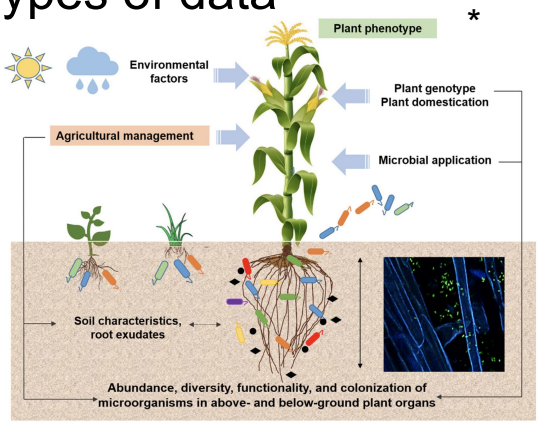


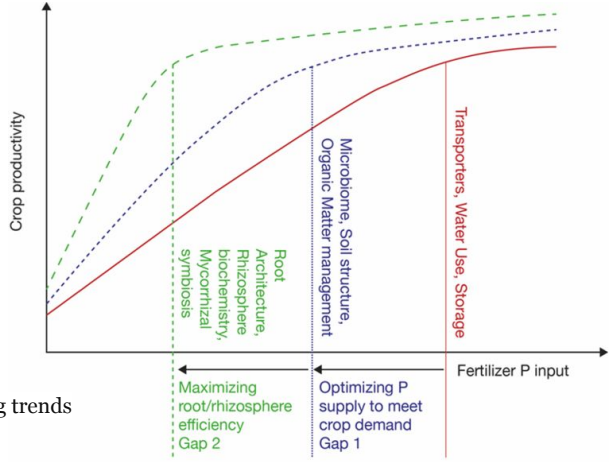

Multidisciplinarity:
Soil science, agronomy, biotechnology, microbiology.

# The Data: examples of large projects in agriculture sub-sectors. Crop farming. Types of data

Types of data :

- Root images for modeling and recognition: millions of images that are currently being analysed and will generate values of root traits, in particular dynamic traits (e.g. root growth rate, emergence, maximum length, apical diameter);
- Soil characteristics and management practice, including historical;
- Microbiome characteristics.




*A review on the plant microbiome: Ecology, functions and emerging trends in microbial application, 2019, in Journal of Advanced Research

# The Data: examples of large projects in agriculture sub-sectors. Crop farming.

- Name of the project: APOLLO

*Advisory platform for small farms based on earth observation*

- Where, Who, When

*European Union, DRAXIS ENVIRONMENTAL S.A., 2016-2019, Supported by EU HORIZON2020*

- Amount of financial support

*€ 2 135 785*

- Main objective

*“ To develop a commercial platform that will provide a suite of farm management advisory services specifically designed to address the needs of small farmers”*

- Example of tasks

*“Based on the agricultural parameters calculated, a suite of farm management services (tillage scheduling, irrigation scheduling, crop growth monitoring, and crop yield estimation) will be developed, and will be delivered through a web and mobile interface. The service requirements will be elaborated in close collaboration with end users. ”*

*https://cordis.europa.eu/project/id/687412*

# The Data: examples of large projects in agriculture sub-sectors. Crop farming. APOLLO workflow.

The **six APOLLO services** support farmers at all stages of the growing cycle. The services are available on the APOLLO platform over the internet through the desktop, or via the dedicated APOLLO mobile/tablet application.

### Tillage Scheduling

Know when to till for best results, avoiding soil degradation and saving energy.

### Irrigation Scheduling

Find out when and how much to water your crops, reduce waste and avoid over-irrigating.

### Crop Growth Monitoring

Keep an eye on the state and health of crops from emergence to harvest

### Crop Yield Estimation

Analyse field productivity and make better-informed decisions on whether to sell or store.

### Weather Forecast and Alerts

Weather forecasts and major weather events alerts.

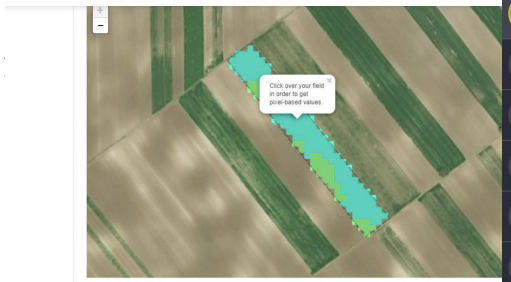
### Farm Management Zoning

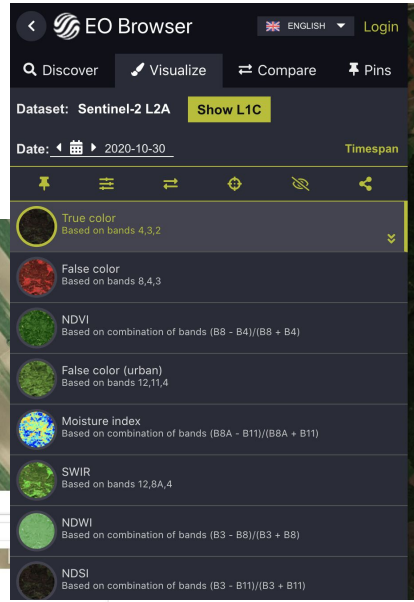
Analyse the field's past to better understand its future.

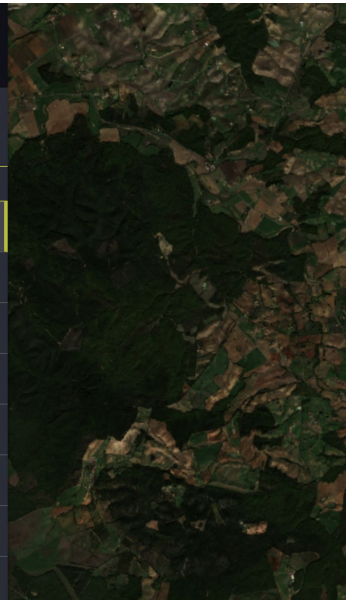
# The Data: examples of large projects in agriculture sub-sectors. Crop farming. APOLLO workflow.

Types of data (according to the publishing list):

- Open-source satellite images;
- Weather characteristics
- Agronomic data;
- Soil characteristics;
- Plant characteristics.



# The Data: examples of large projects in agriculture sub-sectors. Agroforestry.



- Name of project: BREEDCAFS


- Where, Who, When

*European Union, C.I.R.A.D. EPIC, France, 2017-2021, Supported by EU HORIZON2020*

- Amount of financial support

*€ 6 368 786*

- Main objective

*" To design and test coffee varieties, better adapted to agroforestry systems and climate change and maintaining a robust defense system to biotic and abiotic stresses"*

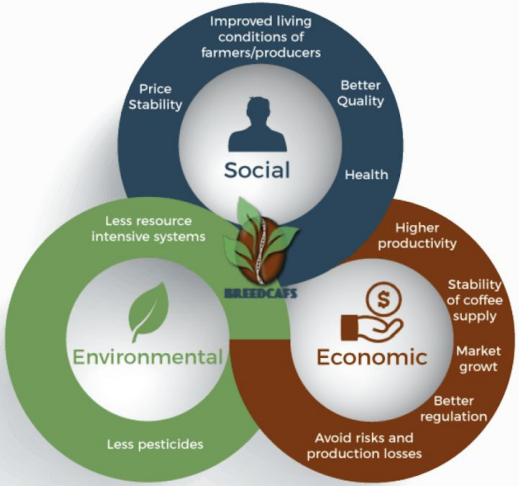
- Example of tasks

*"Based on combination of extensive phenotyping with metabolomic and transcriptomic analysis, analytical and predictive tools for Coffee Metabolic Networks will be developed, leading to marker aided rapid selection and a new approach for breeding of perennial crops "*

https://cordis.europa.eu/project/id/727934

# The Data: examples of large projects in agriculture sub-sectors. Crop farming. BREEDCAFS workflow (on the work package 2 example).

## Expected Impacts




"The main objective of WP2 is to provide a bioinformatic toolkit, leading to a better understanding of coffee physiology and its (epi)genetic basis alongside the development of a novel methodology for the generation of coffee hybrids adapted to agroforestry.

The specific objectives of the WP are to:

1. Develop the BREEDCAFS database for storage of environmental, phenotypic, and molecular data
2. Link the database to user-friendly visualisation as well as analytical and predictive tools
3. Develop / upgrade coffee analysis software tools for RNAseq, epigenetic and genetic data (epi-GBS and GBS)
4. Build statistical models to identify best molecular and/or epi-genetic predictors for coffee quality and yield in AFS
5. Predict best parental crosses for the creation of elite F1 hybrids suited for agroforestry based on developed models
6. Prediction of performance of F1 hybrids at an early developmental stage

# The Data: examples of large projects in agriculture sub-sectors. Livestock.


- Name of the project:

*Creation of a method to evaluate and a system to monitor regional dairy cattle breeding in the republic of Tatarstan*

- Where, Who, When

*Tatarstan, **Skoltech**, 2018-2019, Supported by the Government of Tatarstan.*
*PI: Prof. Philipp Khaitovich*

- Main objective

*“ Development an original selection index that can be used to calculate the genetic index of the pedigree stock in the region and select the stock to breed genetically improved livestock, giving a more accurate prediction of the livestock yield indices improving the milk-yield rates and other utility characteristics of the dairy cattle in the region”*

- Example of tasks

*“Data cleaning and genotyping of dairy females for the development of a genomically-based breeding programme”*

# The Data: examples of large projects in agriculture sub-sectors. Fisheries and aquaculture.



- Name of the project: GAIN

*Advisory platform for small farms based on earth observation*

- Where, Who, When

*European Union, UNIVERSITA CA' FOSCARI VENEZIA, Italy, 2017-2021,*
*Supported by EU HORIZON2020, in collaboration with IBM*

- Amount of financial support

*€ 6 109 648*

- Main objective

*" Develop and optimize sustainable feeds, without increasing the pressure on land and fish stocks"*

- Example of tasks

*"Improve the management of finfish and shellfish farms, in terms of FCR, fish welfare and reduction of wastes, through the use of sensors, biomarkers, Big Data, IoT (Internet of Things) and predictive mathematical models"*


Data from :

- moored sensors,
- satellite,
- autonomous sensors,
- KPIs,
- simulation models

*https://cordis.europa.eu/project/id/7773300*

All spheres of agriculture are based on the complex of disciplines, including life sciences, environmental sciences. Using data are complex, multivariable and multidimensional.

Not only agricultural process ***per se*** should be considered, but also market entry and distribution: where, how and to whom to sell.

Most of the research and R&D projects are finalized as models, recognition tools or SDS.

# Predictive/modeling tasks in agriculture. Examples.



I) Process-based modeling solutions: process-based model is the mathematical representation of one or several processes characterizing the functioning of well-delimited biological systems of fundamental or economical interest.

Usually, such models consist of a set of ordinary or partial differential equations that define the essence of each process (temporal patterns of key parameters), as well as their inputs and outputs, as a function of first principles or else empirical knowledge (Buck-Sorlin G., 2013).

Examples:

- Express-methods for properties estimation, e.g. quality and quantity of soil carbon by mass-spectrometry;
- Modeling of yield;
- Soil erosion dynamics;
- Fish spawning dynamics.

Normally, these types of tasks include parametrization process, based on descriptive and advanced statistics.

# Predictive/modeling tasks in agriculture. Examples.



II) Data-driven solutions: machine learning/artificial intelligence approaches: having a large enough training set, modeling of parameters and their distribution.

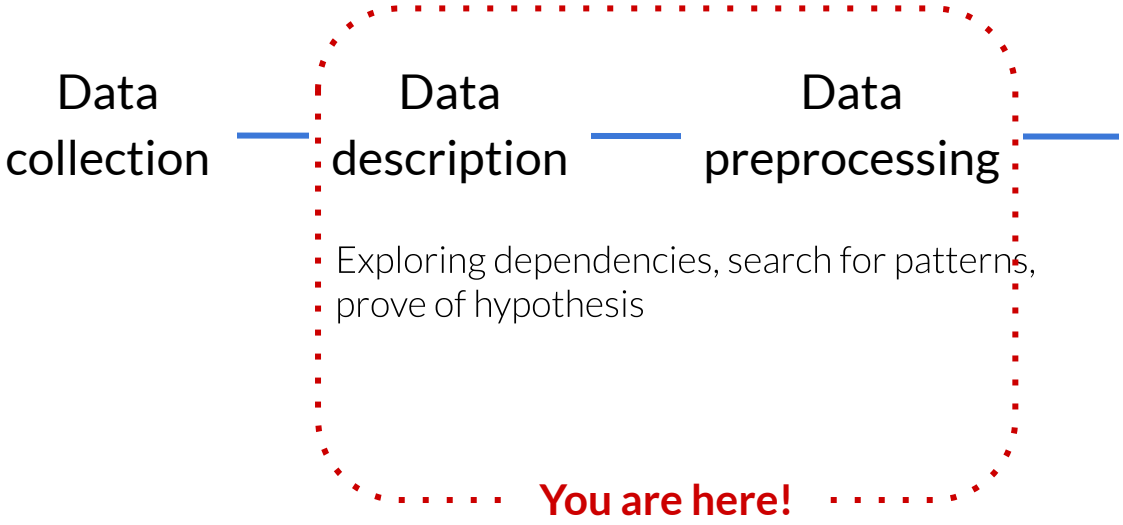
Tools: computer vision, artificial neural networks, machine learning, robotics, support decision systems.

Examples:

- Prediction of surface and groundwater quality;
- Prediction of plant response to the contamination – oil, heavy metal, microplastic, etc.;
- Prediction of novel organic fertilizer influence on soil, plants and microorganisms, etc.
- Classification of habitus: plants, animals.

Normally this types of task include large amount of different variables, not all of them are significant. To reduce dimensions, basic pre-processing routine is required.

# Principal steps in agriculture research/R&D projects



# 5 minutes of history

# Rothamsted Experimental Station, someones Student, sir Ronald Fisher and Analysis of Variance and Experimental Design

https://www.rothamsted.ac.uk

**Statistical significance, industrial quality control, efficient design of experiments .. where all of these concepts came from?**

Rothamsted Experimental Station, now is known as Rothamsted Research, has been founder in 1843.
It is one of the oldest agricultural institutions in the world, but additionally it is one of the most significant birthplaces of modern statistics.
Many famous statisticians worked in Rothamsted, such as sir Fisher or Gosset.

# Gosset and beer

Barley, *Hordéum vulgáre*

This or this?


# Our old buddy, Student.. Who are you?

Student's t-test – one of the most popular basic statistical tools. It is an any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis. Example: whether the means of two data-sets significantly differ from each other or not, when the data available are very limited in number.

William S. Gosset developed t-test as an economical approach to monitor the quality of famous stout beer – Guinness. Since policy company forbade to reveal any work concepts, Gosset was compelled to publish his statistical work under pseudonym of the name Student.

William Sealy Gosset
(1876-1937)

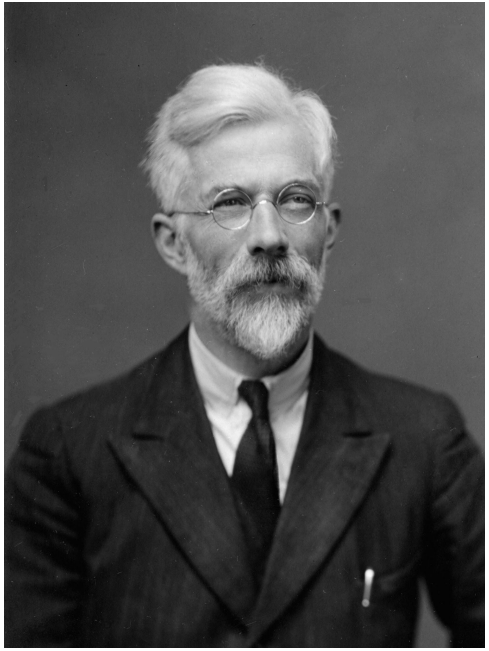
# “Statistical Methods and Scientific Inference”

In Rothamsted Fisher provided series of the plant-breeding experiments to answer the question:

How to provide greater information with less investments of time, effort, and money? How to conduct experiment properly?

Fisher introduced:

- Principle of randomization – random selection of units of materials supported by control need to be used to diminish the effects of variability in experimental materials;
- Concept of analysis of variance, a.k.a. ANOVA – enabled experiments to answer several questions at once. Fisher’s principal idea was to arrange an experiment as a set of partitioned sub-experiments that differ from each other in one or more of the factors or treatments applied in them.

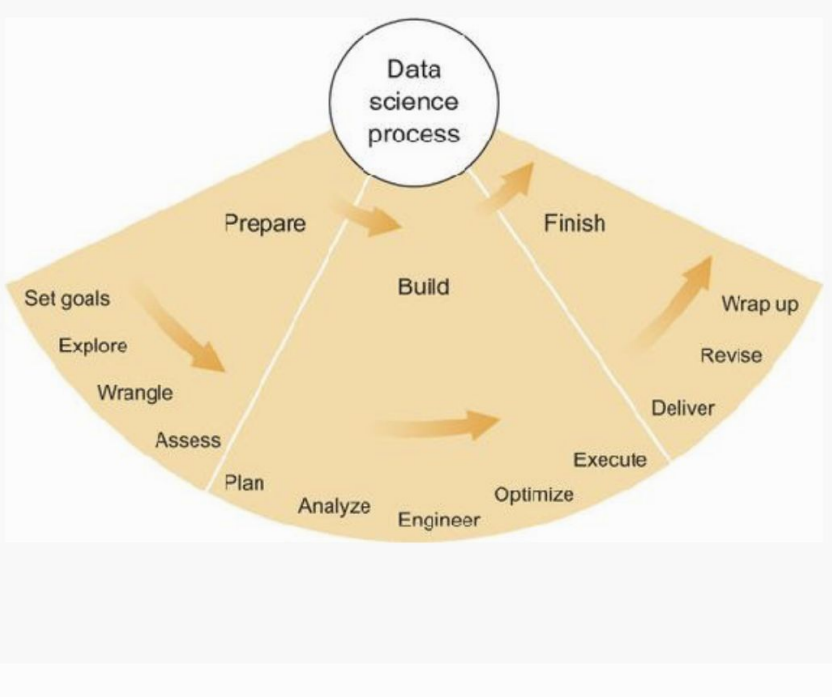


Sir Ronald Aylmer Fisher
(1890-1962)

# Data exploring and preprocessing workflow

# Environmental data exploring and analysis starter-pack






- Exploring the data structure: identification of types of variables;
- Exploratory data analysis: missing values, range, mean, median, trimmed mean, quartiles;
- Distribution type check-up;
- Statement of hypothesis;
- Search of correlations;
- Dimensionality reduction;
- Search of outliers and outliers treatment.

Godsey, Brian. *Think Like a Data Scientist: Tackle the data science process step-by-step*. Manning Publications Co., 2017.

# General population and the sample

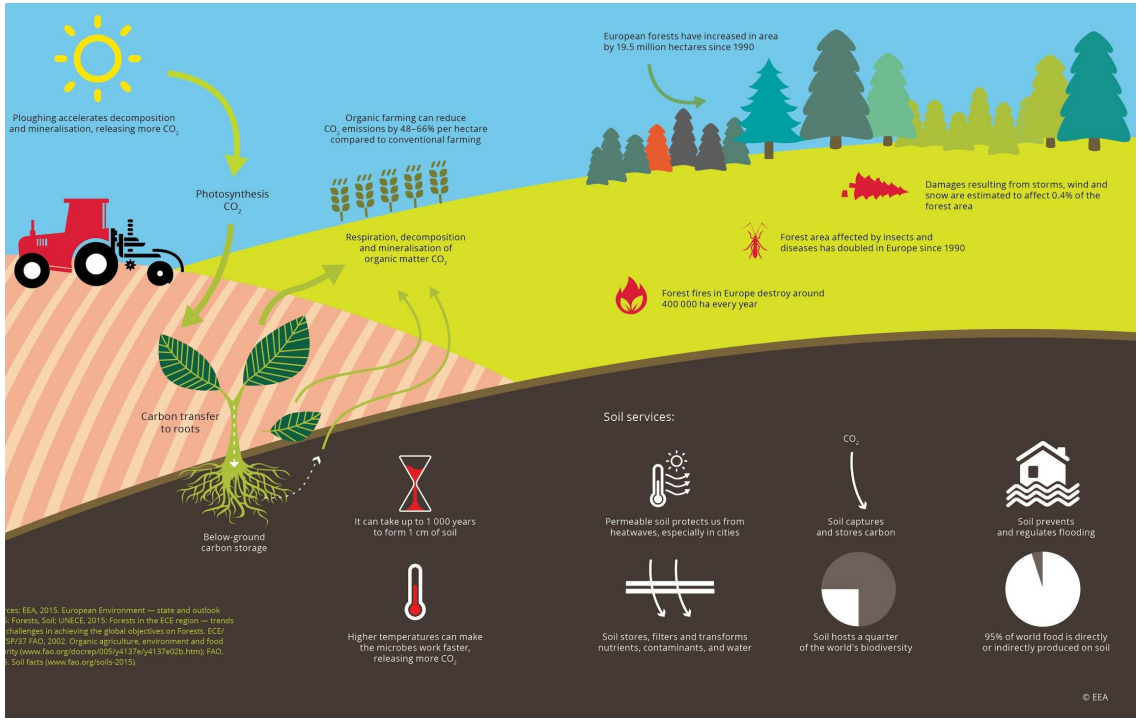
The set of all the objects that we want to make a conclusion about is **general population**.

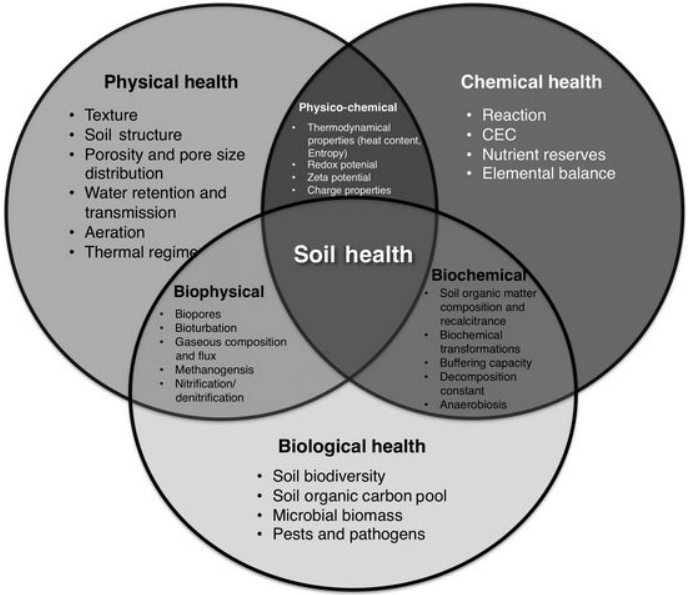
A set of observations selected from the general population in accordance with special declared rules – is a **sample**.

A statistical inference is a decision, estimate, prediction, or generalization about the population based on information contained in the sample.


Digital Elevation Model of the field and sample points

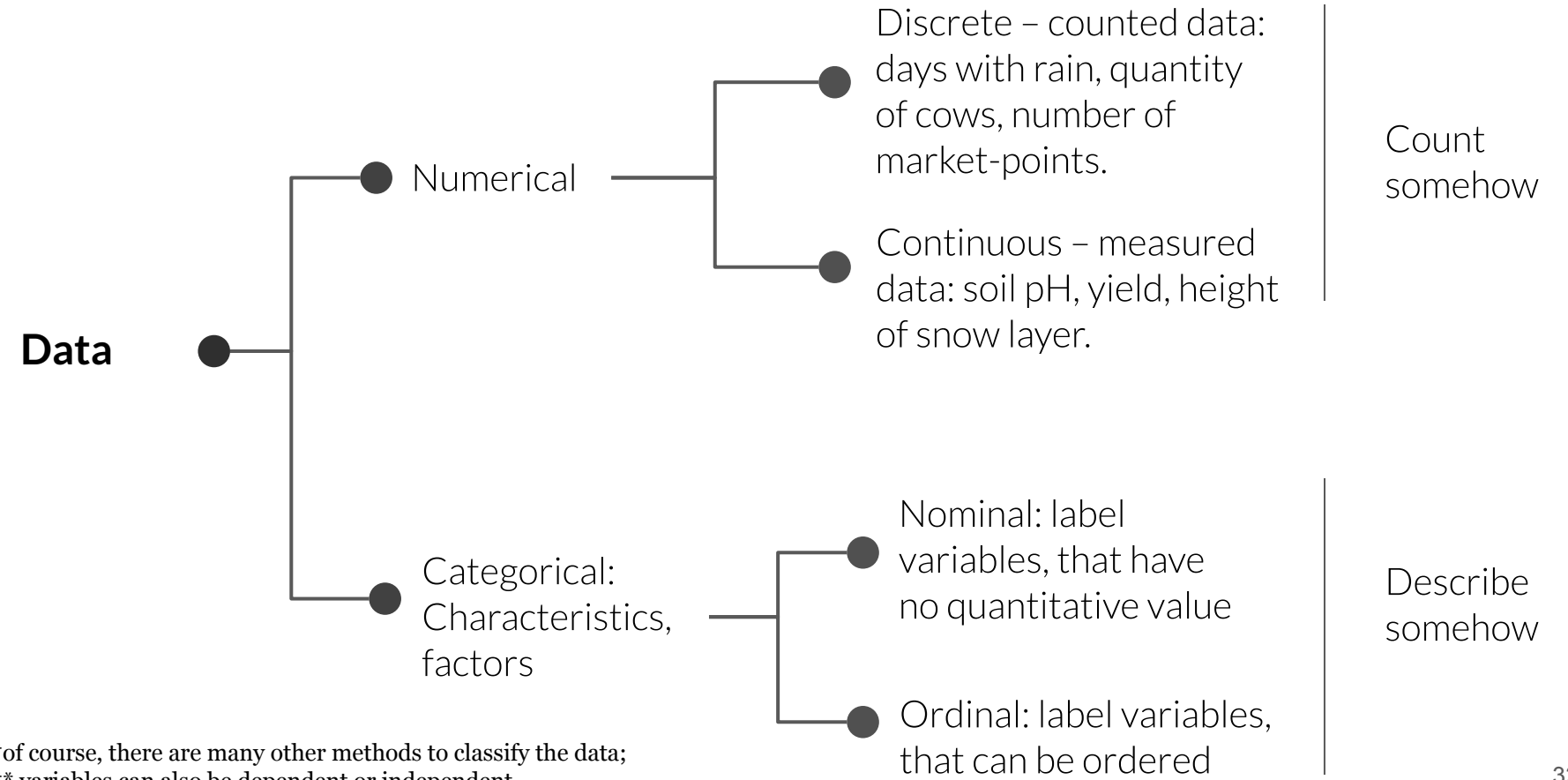
# Why so many samples?




From “Soil Health and Climate Change: An Overview” by Rathan Lal.

# Types of data in statistics*




*of course, there are many other methods to classify the data;
** variables can also be dependent or independent

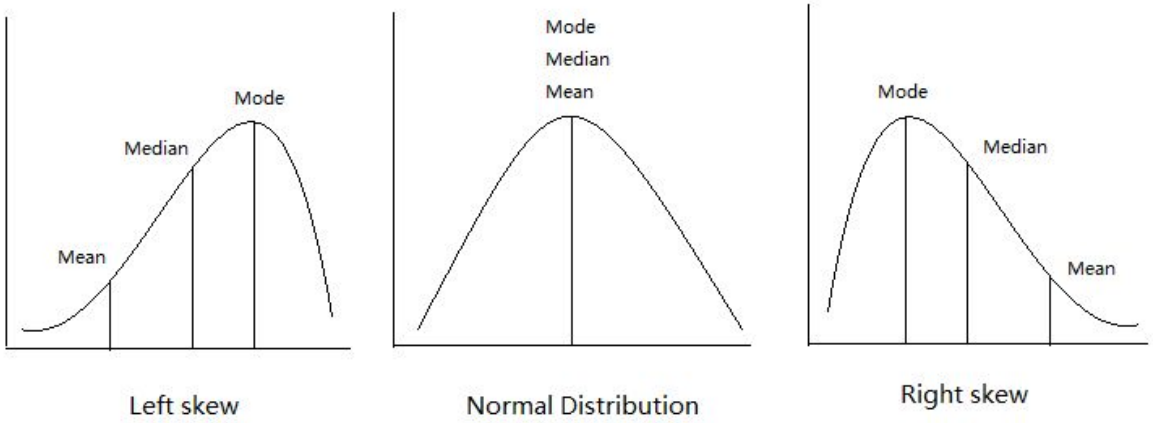
# Measures of central tendency: mean, median, mode

The **sample mean** you have to find the sum of all the observations in the sample and divide the sum by the number of observations

The **median** is a value that divides sorted data into two equal parts.

---

The **mode** is the most frequent value or values in a sample.



# Measures of dispersion (range, quartiles, SD, variance)



**Dispersion** – the extent to which data values differ from one another.

**Range** – the distance between the biggest and the smallest values.

**Variance** is the mean squared deviation of a variable from its mean. The higher the variance, the larger the variability of the data. Adding a constant just shifts the distribution and doesn't affect variability of the data.

**Standard deviation** – square root of variance. Adding a constant just shifts the distribution and doesn't affect variability of the data.

**Quartiles** divide the data into some equal groups of observations. Interquartile range (IQR) is the difference between the first and the third quartile.

**Percentiles** divide the data into hundred equal parts. Median is equal to both second quartile and fiftieth percentile.

# Example: temperature/precipitations observations

~/Desktop/icr-ms/data_proc/

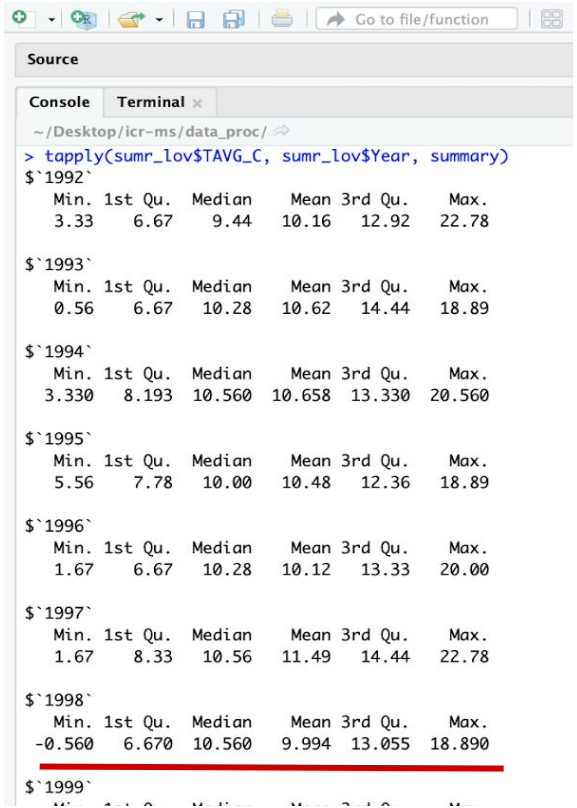
```
> summ_tov
```

| | STATION | NAME | LATITUDE | LONGITUDE | ELEVATION | Month | Year | PRCP | TAVG_C |
|---|---|---|---|---|---|---|---|---|---|
| 1 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 5.08 | 4.44 |
| 2 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 4.44 |
| 3 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 8.33 |
| 4 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 10.56 |
| 5 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 8.33 |
| 6 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 11.67 |
| 7 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 7.22 |
| 8 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 4.06 | 5.56 |
| 9 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 11.11 |
| 10 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 10.00 |
| 11 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 11.11 |
| 12 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 18.33 |
| 13 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 22.78 |
| 14 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 22.78 |
| 15 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 18.89 |
| 16 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 12.78 |
| 17 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 1.52 | 7.22 |
| 18 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 7.22 |
| 19 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 10.00 |
| 20 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 15.56 |
| 21 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 16.67 |
| 22 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 4.06 | 11.67 |
| 23 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 3.89 |
| 24 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 6.67 |
| 25 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | 0.00 | 6.11 |
| 26 | RSM00022127 | LOVOZERO, RS | 68 | 35.033 | 162 | 6 | 1992 | NA | 4.44 |

Data: average daily summer (June, July, August) weather observations – air temperature and amount of precipitations – from the station near Lovozero (Murmansk region) from 1992 to 2019.

Useful tip: such type of data (climate data) can be extracted from the **world** database of the National Oceanic and Atmospheric Administration (NOAA).

# Example: temperature/precipitations observations

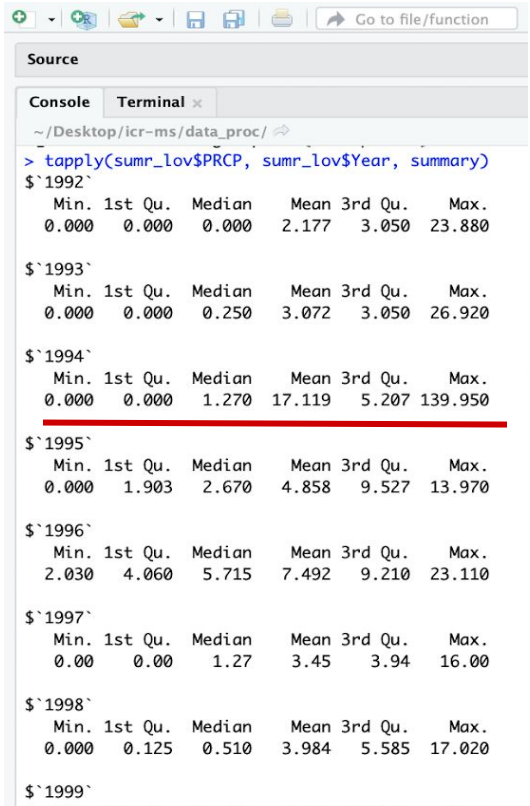


Source

Console Terminal

~/Desktop/icr-ms/data_proc/

```
> tapply(sumr_lov$TAVG_C, sumr_lov$Year, summary)
$`1992`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.33    6.67    9.44   10.16   12.92   22.78

$`1993`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.56    6.67   10.28   10.62   14.44   18.89

$`1994`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.330   8.193  10.560  10.658  13.330  20.560

$`1995`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.56    7.78   10.00   10.48   12.36   18.89

$`1996`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.67    6.67   10.28   10.12   13.33   20.00

$`1997`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.67    8.33   10.56   11.49   14.44   22.78

$`1998`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.560   6.670  10.560   9.994  13.055  18.890

$`1999`
```



Source

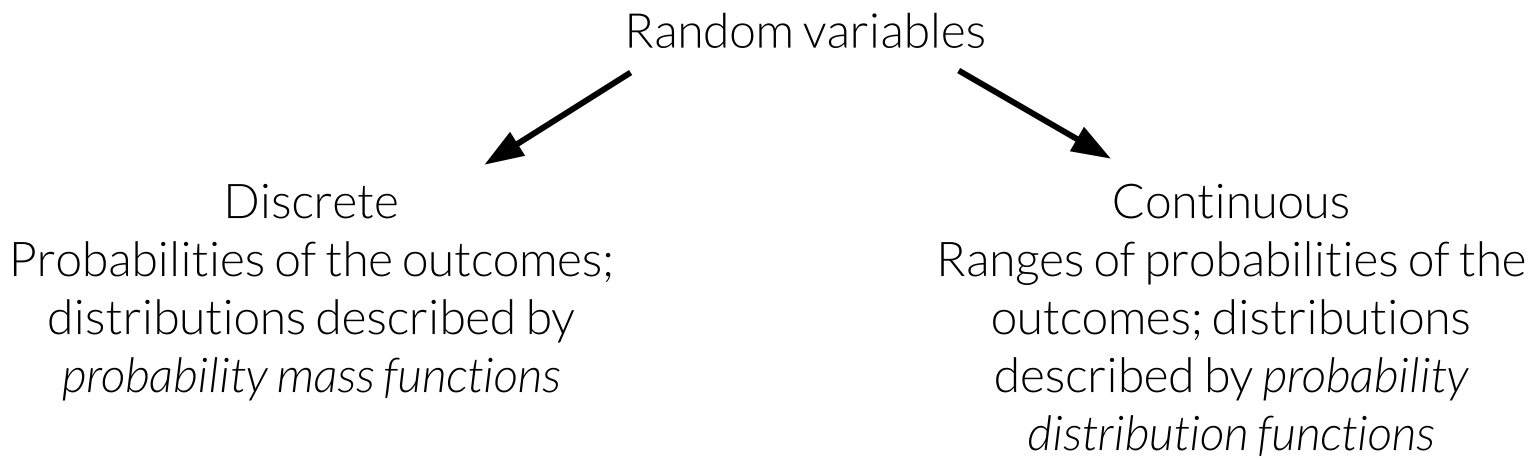
Console Terminal

~/Desktop/icr-ms/data_proc/

```
> tapply(sumr_lov$PRCP, sumr_lov$Year, summary)
$`1992`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   2.177   3.050  23.880

$`1993`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.250   3.072   3.050  26.920

$`1994`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.270  17.119   5.207 139.950

$`1995`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.903   2.670   4.858   9.527  13.970

$`1996`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.030   4.060   5.715   7.492   9.210  23.110

$`1997`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    1.27    3.45    3.94   16.00

$`1998`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.125   0.510   3.984   5.585  17.020

$`1999`
```

Quick observation of the data, quick conclusions to be made.

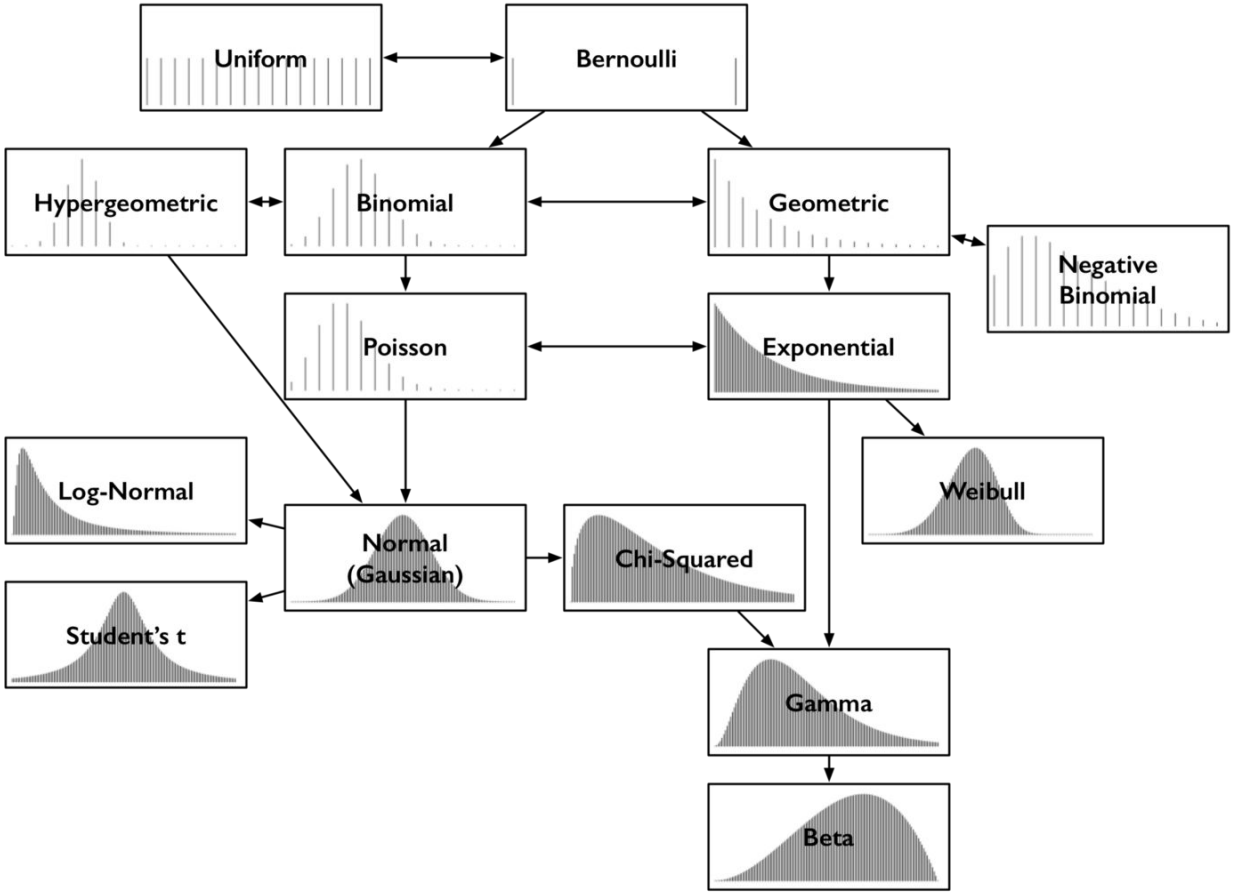
# Statistical distribution types




A **probability distribution** is a mathematical function that provides the probabilities of the occurrence of various possible outcomes in an experiment. Probability distributions are used to define different types of random variables in order to make decisions based on these models.

# Probability distributions

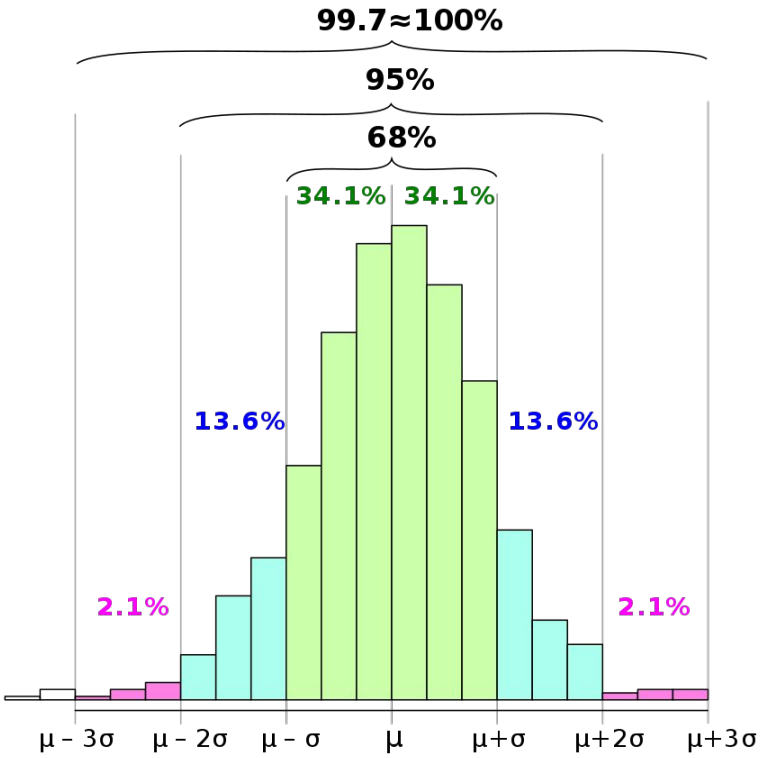
- There is a huge amount of them, but just several more frequent;
- The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space.
- Each distribution is characterised by its own probability density function;
- Once you know a distribution, you can choose statistics to describe observations.




Cloudera blog on Medium

# (Rare) normal distribution and 3 sigma rule






https://en.wikipedia.org/wiki/68–95–99.7_rule

Observations correspond to normal distribution are distributed according to the probabilistic rule :

$$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$$

To check normality use statistical tests: such as the Shapiro-Wilk or Kolmogorov-Smirnoff. Remember, though, for small sample sizes (n < 20), the tests are unlikely to detect non-normality and for larger sample sizes (n > 50), the tests can be too sensitive.

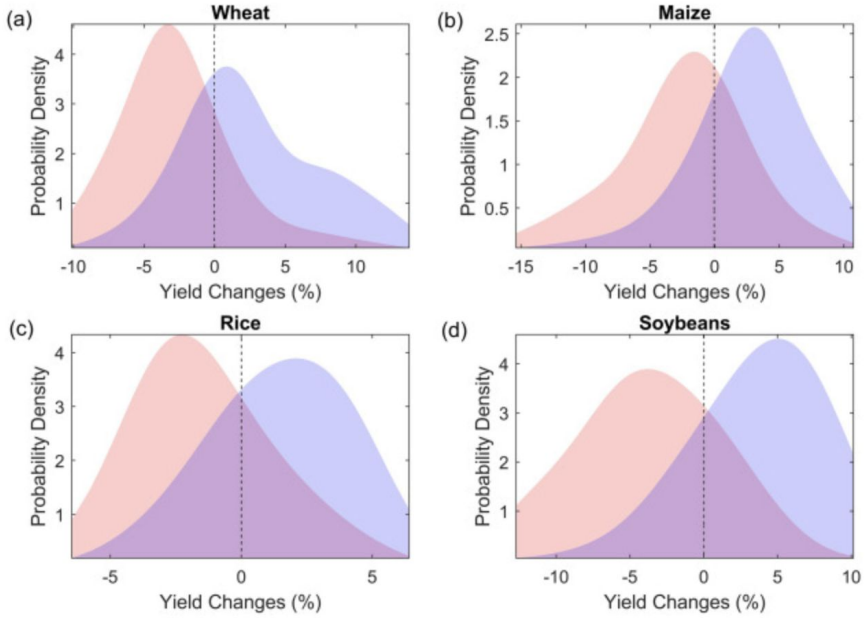
For quick check use histograms, boxplots, Q-Q (quantile-quantile) plots. If distribution is skewed, use non-parametric statistics. Forcing distribution to normal is possible but not recommended for most descriptive tasks.

# Distributions are different and often non-linear, but the information about it is meaningful by itself



Distributions more often asymmetric, they may have not only 1 peak.

- Skewness – is a measure of the asymmetry of data distribution.
- Kurtosis – sharpness of a distribution pattern. If all individuals are concentrated around the mode, their kurtosis is high and vice versa.




Download : Download high-res image (368KB)

Download : Download full-size image

Fig. 3. Conditional probability distribution of yield changes (%) relative to its long-term mean under moderate drought (red) and wet (blue) conditions for (a) wheat, (b) maize, (c) rice and (d) soybeans.

Leng G, Hall J. Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. Science of the Total Environment. 2019 Mar 1;654:811-21.

# Hypothesis test

Hypothesis test: what is the probability of measuring or detecting an effect **is true and not random**? The hypothesis that an **apparent effect is due to chance** is called the **null hypothesis**.

1. Specify the null hypothesis: parameter equals zero, greater than or equal to zero or that a parameter is less than or equal to zero.
2. Specify significance level (**α**-level): the probability of rejecting the null hypothesis when it is true. For most environmental studies **α**=0.05, sometimes 0.01.
3. Estimation of probability value – p-value. This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
4. Compare probability with **α**-level: if the data analysis results in a p-value below the **α**-level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected.
5. Corrections to avoid type I (false negative) and type II errors (false positive).

# Statistical tests

**Distribution of data is *approximately* normal**

Z-test

A z-score is calculated with population parameters such as "**population** mean" and "**population** standard deviation" and is used to validate a hypothesis that the sample drawn belongs to the same population.

t-test

T-test compares the **mean** of two given **samples**.

ANOVA (analysis of variance)

ANOVA compares **multiple samples** with a single test.

**Data is non-normally distributed**

Wilcoxon Signed Rank test

Test compares two related samples, matched samples or repeated measurements on a single sample to assess whether their population mean **ranks** differ.

Kruskal-Wallis/Mann-Whitney test

Test compares the **medians** of two or more samples to determine if the samples are from different populations.

# Example

In the paper “Multi-country evidence that crop diversification promotes ecological intensification of agriculture” (*Nature Plants, 2016*)

**Objective**: to prove that border plants would promote natural enemy activity—principally by providing nectar—and thereby reduce pest densities and the need for insecticides without a yield penalty.

Yield is higher, pesticides use is less, farming is more advantageous

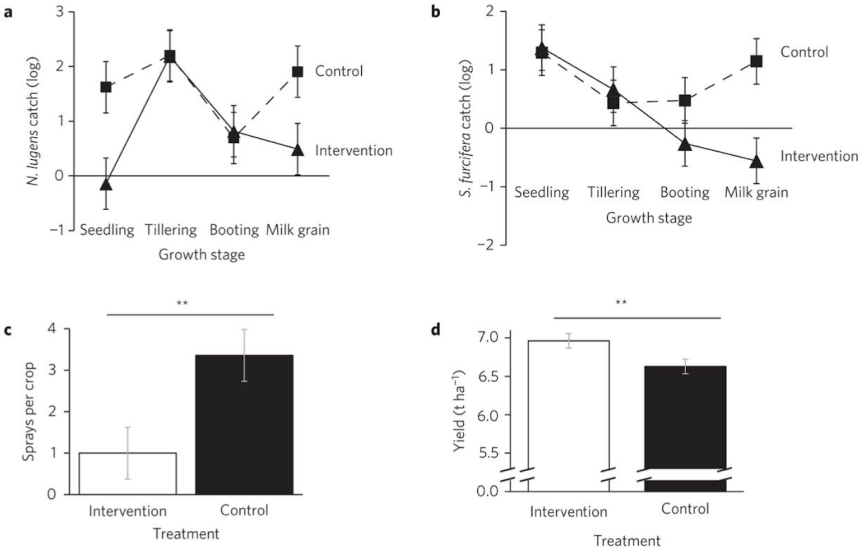
# Example

In the paper “Multi-country evidence that crop diversification promotes ecological intensification of agriculture” (***Nature Plants, 2016***)

**Reported result**: inexpensive intervention significantly reduced populations of two key pests, reduced insecticide applications by 70%, increased grain yields by 5% and delivered an economic advantage of 7.5%. Simple diversification approach can contribute to the ecological intensification of agricultural systems.

Figure 1: Multi-site, multi-year comparison of diversification of rice (intervention) with conventional practice (control).

From: Multi-country evidence that crop diversification promotes ecological intensification of agriculture




**a**, Rice brown planthopper catch (*Nilaparvata lugens*) (log-predicted mean from GLMM) (significant treatment by date effect, $F = 6.26$, d.f. = 3, 198.4, $P < 0.001$). **b**, Equivalent values for whitebacked planthopper (*Sogatella furcifera*) (significant treatment by date effect, $F = 9.63$, d.f. = 3,201.1, $P < 0.001$), **c**, Mean number of insecticide applications per annum ($F = 14.20$, d.f. = 1, 13, $P = 0.002$). **d**, End-of-season grain yield ($F = 12.31$, d.f. = 1, 13, $P = 0.004$). Error bars, s.e.m. **$P < 0.01$ using ANOVA.

# Correlation

**Covariance** measures how much two variables change together. The greater the covariance value corresponds to the stronger the connection between the two variables.

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

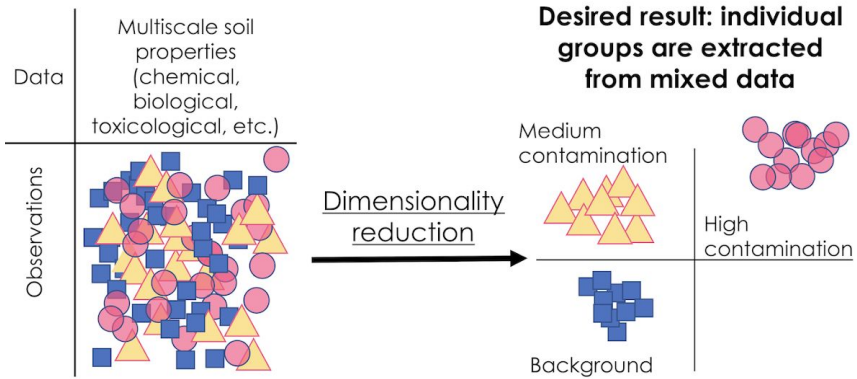
**Correlation** measures the strength of the relation of one variable to another. Correlated parameters move together in some way.

- Correlation coefficients range between -1 (strong negative correlation) to 1 (strong positive correlation) via 0 (no correlation). **Pearson Correlation Coefficient** – for linearly dependent normally distributed parameters; **Spearman Correlation Coefficient** – for non-linear relationships. **Kendall's Tau Correlation Coefficient** is also based on variable ranks but does not take into account the difference between ranks.
- Correlation is used as a basic quantity for many modelling techniques;
- **Multicollinearity** occurs when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy (linear regression is not robust, but decision/boosted trees – are).

# Dimensionality reduction

Problem statement: the data is multiscale and crowded. What to do?

> **feature extraction** and feature engineering: transformation of raw data into features suitable for modeling;
> **feature transformation**: transformation of data to improve the accuracy of the algorithm;
> **feature selection**: removing unnecessary features.



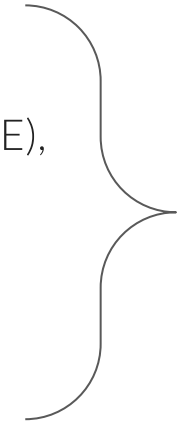
# Feature extraction DR



Aim: to operate with the multidimensional dataset and simplify the interconnections by projecting the initial high-dimensional data to 2D space. Can be unsupervised and supervised; linear and non-linear. Most know DR techniques are:

- Principal Component Analysis (PCA)

- Multidimensional Scaling (MDS),
- Isometric Feature Mapping (Isomap),
- t-distributed Stochastic Neighbor Embedding (t-SNE),
- Locally Linear Embedding (LLE),
- Hessian Locally-Linear Embedding (HLLE),
- Modified Locally Linear Embedding (MLLE),
- Local Tangent Space Alignment (LTSA),
- Spectral Embedding (SE)



**non-linear**

# Principal component analysis (PCA)



PCA is a factor model in which the factors are based on the total variance. It aims to **reduce the number of inter-correlated variables to a smaller set** which explains the overall variability almost as well.

No variable is designated as dependent, and no observation grouping is a priori assumed. The function of the PCA is to **extract important information** from datasets and to express this information through a set of new orthogonal variables. These new variables can be used in further analysis e.g. regression, cluster analysis, etc.

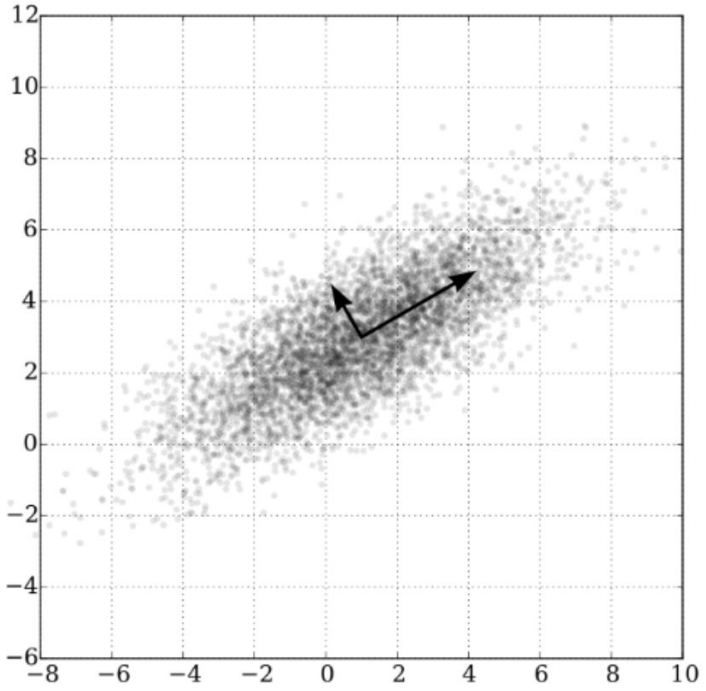



Image is from [wiki-page about the PCA]

# Principal component analysis (PCA) limitations

- Variables should be normally distributed and have appropriate number of cases;
- Multiple variables should be measured at the continuous level (although ordinal variables are very frequently used).
- Relationships between all variables need to be linear. In practice, this assumption is somewhat relaxed (even if it shouldn't be) with the use of ordinal data for variables.
  *Non-linear methods allow for nonlinear data projections and may better capture nonlinear patterns than PCA.
- No significant outliers. Outliers are important because these can have a disproportionate influence on your results.

# PCAs' general points and definitions



- Principal component – is a **linear combination** of optimally weighted initial variables. Weight vector is the eigenvector which is satisfies the principle of least squares. PCs are orthogonal.
- PCA is performed on a square symmetric matrix. Most frequently it is based on either the covariance matrix (variables scales are similar) or the correlation matrix (variables scales are different).
- PCA includes normalization procedure.
- General points:

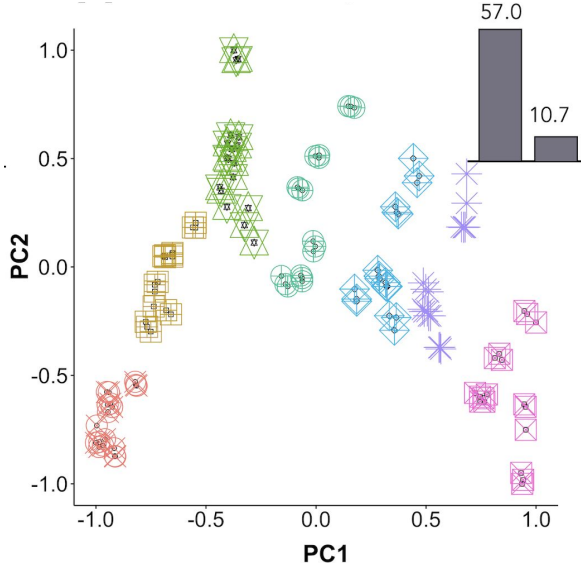
➔ Eigenvalue – column sum of squared loadings for a factor, i.e., the latent root. It conceptually represents that amount of variance accounted for by a factor.

➔ Variance explained by PCs – is computed from eigenvalues. It is the measure of variance attributed to each principal component.

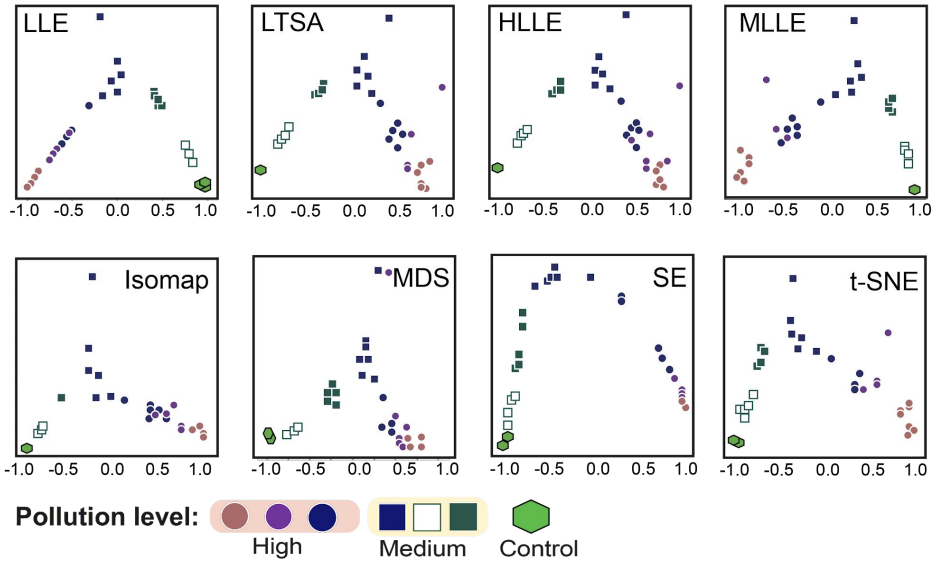
➔ Component loadings – correlations between individual variables and the PC's (pay attention to coefficients out of the range between –0.3 and 0.3).

➔ Component score – is of all row and columns, which can be used as an index of all variables and can be used for further analysis.

# DR as visualization tool




PCA VS non-linears



# Outliers detection

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. **Outliers matter!**

To detect outliers use:

- Z-Score or Extreme Value Analysis (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (PCA, Least median squares regression)
- Proximity Based Models (non-parametric)
- High Dimensional Outlier Detection Methods (high dimensional sparse data)

# Useful links

1. 2030 Agenda: Sustainable Development Goals
2. DeGroot M. H., Schervish M. J. Probability and statistics. – Pearson Education, 2012. (PDF)
3. OnlineStatBook (web-page)
4. Verzani J. Using R for introductory statistics. – CRC press, 2018. (PDF)
5. Statistics in Python (web-page)
6. Top 20 Python libraries for data science in 2018 (yeah, long time ago..)
7. Introduction to statistics (Stepic online free course)
8. Explained visually
9. Scikit-learn library (python)