Visualization of scientific data

Methods of Scientific Working (for Crop Sciences) (3502-440)

21 Jan 2025

Table of contents

1 Why visualization?

> *A picture is worth 10,000 words* – Anonymous

This proverb explains the utility of visualization: Complex data or results can be simplified by visualization to enhance understanding and insight. For this reason, it is a means to summarize information. Visualization can also lead to new insights by presenting data in a different way, therefore it is also a means for inference and exploration to discover new (and possibly unexpected) patterns in the data, explanations of the data and may lead to new scientific hypotheses.

The goal of this chapter is not to provide a list of tips for a good visualization of scientific data, but to provide an introduction into **visual thinking** In this approach, we closely follow the concepts of the information scientist and graphics designer Edward Tufteˆ.[His website is at http://www.edwardtufte.com.]

The value of visualization is demonstrated by some simple data in Table 1 and their visualization in Figure 1.

Table 1 *– A simple data set of pairs of numbers.*

x	y
1	3.5
2	11.8
3	23.9
4	39.6
5	58.5
6	80.5
7	105.4
8	133.0
9	163.7
10	196.8
20	662.0
30	1,345.9
40	2,226.8
50	3,290.5
60	4,527.2
70	5,929.1
80	7,489.9
90	9,204.3

The visualization of the data shows that there is a linear relationship between the values, but only after a log-log-transformation. This indicates that the data in the table were generated from a mathematical function of the form

$$f(x) = k \cdot x^a \tag{1}$$

where k is a constant and m is the slope in the log-log plot.

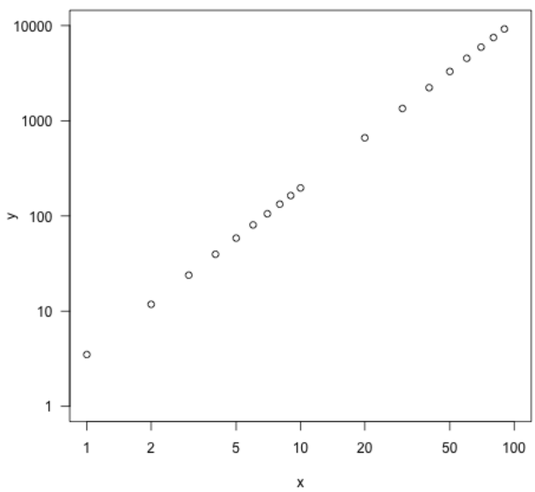


Figure 1 *– Visualization of the data in Table ref:tab:simpledata.*

Visualization extends very much beyond the plotting of scientific diagrams. It merges with computer graphics, photography, animations and so on. As a consequence of the data explosion in and outside of science a whole new field of information graphics evolved that is concerned with the optimal (and truthful) representation of information. Nevertheless, visualization has been an important tool since the beginning of era of the New World (beginning with the discovery of the Americas by Columbus), and in particular since the industrial revolution. Even in artistic paintings, useful information was stored and visualized to improve understanding.

One of the first visualizations that contains a large number of data points by Charles Joseph Minard (1781-1870; Wikipedia), a French Civil Engineer (Figure 2). This classical graphics shows the changes in Napoleon'#s army during and after the march to Moscow (Figure 2). It is a **band graph** for illustration of flows that are also called **Sankey diagrams** (Wikipedia).

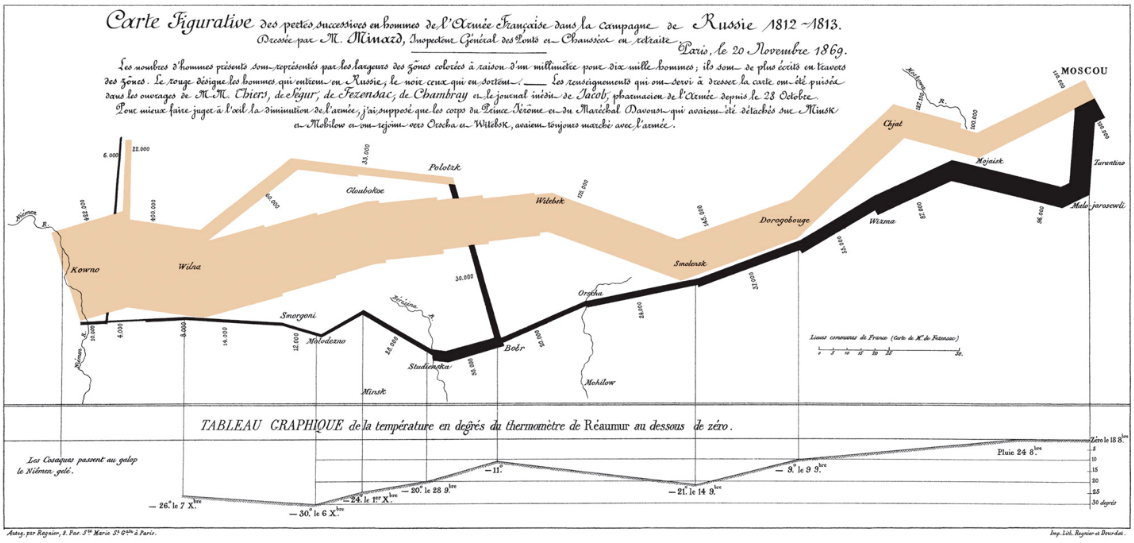


Figure 2 *– The change in the size of Napoleon's army during his march to Moscow in 1812 and his subsequent defeat.*

Minard's figure integrates various types of data:

Numerical information The size of Napoleon's Army.
Geographical information The size of his army in different regions.
Temporal information The size of his army at different time points during his conquest.

In Figure 2 multidimensional data are reduced on to two dimensions, without much loss of information. One important aspect of a good information graphic is that it contains raw data that allow to check whether their representation as graphical elements is correct. One way to measure the effectiveness of a graphic is to express it as its **information content**. How many data points are included in the graph? The information content of a graph is useful decide whether it is better to present data as a table or as a figure. As a rule of thumb, the leading information scientist and graphics designer Edward Tufte postulated that a data set with less than 20 data points should be presented as a table.

Another approach to how both temporal and spatial information can be visualized is shown in Figure 3. Again, four dimensions are represented in two dimensions.

2 The information content of graphics

One of the important decisions in the context of visualization is whether to present results as table or figure. This can be demonstrated with the following example. A simple way to express the information content can be shown with a normal distribution. With the statistics package R, one can easily simulate 1,000 normally distributed data points with a mean of 0 (Listing 1).

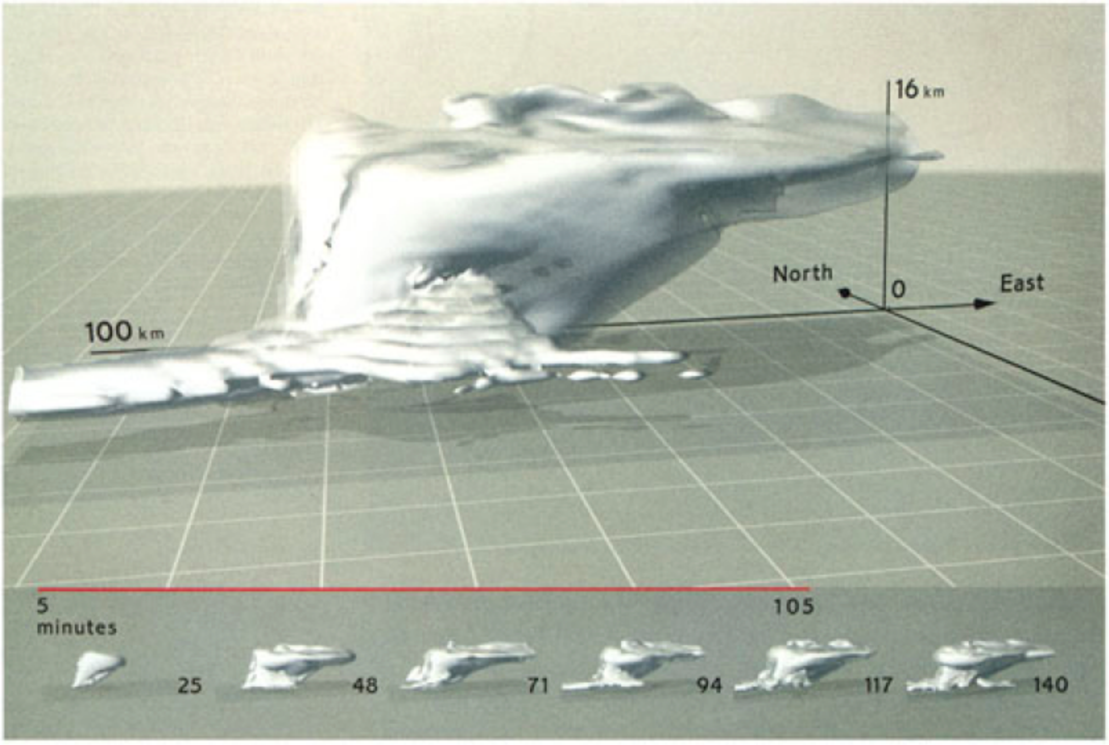


Figure 3 *– Modelling of a storm. Four dimensions are presented in two dimensions with the time axis shown on the bottom of the figure. Note the similarity of the design with the Youtube viewing window (e.g., the red line at the bottom). Source: Tufte (1997)*

Listing 1 Simulation of a normal distribution with R

a <- rnorm(1000,0) a [1] -0.0210952179 -1.0912065990 -1.1395054396 -1.3189259016 ... [8] -0.2912739666 -0.1470279242 1.7375103224 1.2480682763 ... ... [988] 0.0045077573 -0.1658796133 -0.2793556344 -0.1170640982 ... [995] 2.0871831247 -0.4179269672 -0.4178945557 -0.3895868971 ...

Obviously, the raw data are not suitable for visualization. If we want to summarize and simplify the data, there are several possibilities. We can calculate the mean and standard deviation, which is $\mu = 0.0236$ and $\sigma = 1.01241$.

Three different representations of the same data set that allow to extract different types of information (Figure 4). A **boxplot** provides information about the mean, median, quartiles and outliers. The **histogram** shows the frequency distribution, which allows a rapid scan whether the distribution is normally distributed or skewed. The red line in Figure 4 c is the expected normal distribution with a mean of 0 and a standard deviation of 1 and allows a visual comparison between observed and expected distributions. For this reason, it allows a first inference. A more formal inference can be conducted with a test of normality and the results of such a test can be summarized with the following sentence.

````
```{R}
a <- rnorm(1000,0)
shapiro.test(a)
```
````

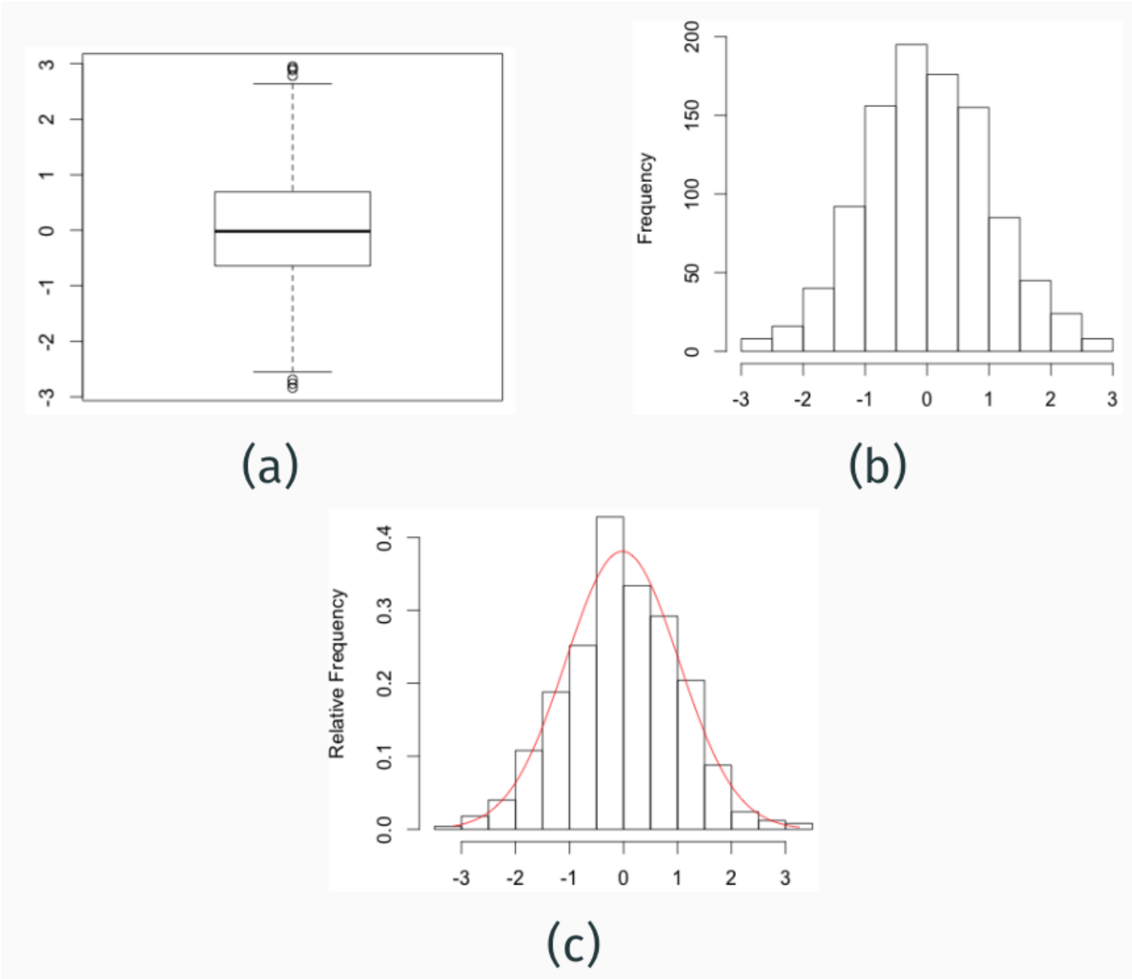


Figure 4 *– Different plot types for the same dataset.*

```
    Shapiro-Wilk normality test

data:  a
W = 0.99869, p-value = 0.6797
```

The observed variation of 1,000 values shows a mean value of 0 (Standard deviation = 1) and does not significantly differ from a normal distribution, because $p > 0.05$, which is usually taken as the significance level of rejecting a null hypothesis, H_0. In this example, the null hypothesis is that data are normally distributed.

Which of the available possibilities to present the data are best? It depends on the context (i.e. how much space is available), or which aspect of the data is particular interesting. For example, if the data are expected to be normally distributed but the observed data are not, it may be interesting to show the histogram, but not if the observed data do not reject the normality test.

3 Improving the presentation of data

The following example shows how scientific data can be visualized without losing their information content. The data and figures were taken from a review of cancer survival probabilities from the website of Edward Tufte.[1]. Figure 5 shows the original data from the Lancet article. Figure 6 an improved version of the table by Edward Tufte.

Figure 6 is an improved version of the table by Edward Tufte by changing mainly the type and the size of the font.

A presentation with Powerpoint is shown Figure 7, where the data have been mixed up to obfuscate them.

There is a third way of making the data on cancer survival rates more accessible: by combining graphical elements with a tabular structure of the original data. The last version in Figure 8 probably results in an optimal structure.

4 Which type of visualization is appropriate?

Are data better represented as a *picture* or a *scientific plot*?

Possible criteria for decision are:

- Ease of understanding
- Background knowledge required
- Complexity: How much time does it take to understand the figure?
- **Data-to-ink ratio**

As an example: The relationship of head volume to body volume during different stages of human development (Figure 9).

A more ‘scientific’ approach is to plot the number for the body height and the arm length as shown in Figure 10.

The drawing conveys the key message more efficiently without much explanations required and the plot is more accurate, but requires more *a priori* knowledge and time to understand.

[1] https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0000Jr

	Relative survival rate, % (SE)			
	5 years	10 years	15 years	20 years
Cancer site				
Oral cavity and pharynx	56·7 (1·3)	44·2 (1·4)	37·5 (1·6)	33·0 (1·8)
Oesophagus	14·2 (1·4)	7·9 (1·3)	7·7 (1·6)	5·4 (2·0)
Stomach	23·8 (1·3)	19·4 (1·4)	19·0 (1·7)	14·9 (1·9)
Colon	61·7 (0·8)	55·4 (1·0)	53·9 (1·2)	52·3 (1·6)
Rectum	62·6 (1·2)	55·2 (1·4)	51·8 (1·8)	49·2 (2·3)
Liver and intrahepatic bile duct	7·5 (1·1)	5·8 (1·2)	6·3 (1·5)	7·6 (2·0)
Pancreas	4·0 (0·5)	3·0 (0·5)	2·7 (0·6)	2·7 (0·8)
Larynx	68·8 (2·1)	56·7 (2·5)	45·8 (2·8)	37·8 (3·1)
Lung and bronchus	15·0 (0·4)	10·6 (0·4)	8·1 (0·4)	6·5 (0·4)
Melanomas	89·0 (0·8)	86·7 (1·1)	83·5 (1·5)	82·8 (1·9)
Breast	86·4 (0·4)	78·3 (0·6)	71·3 (0·7)	65·0 (1·0)
Cervix uteri	70·5 (1·6)	64·1 (1·8)	62·8 (2·1)	60·0 (2·4)
Corpus uteri and uterus, NOS	84·3 (1·0)	83·2 (1·3)	80·8 (1·7)	79·2 (2·0)
Ovary	55·0 (1·3)	49·3 (1·6)	49·9 (1·9)	49·6 (2·4)
Prostate	98·8 (0·4)	95·2 (0·9)	87·1 (1·7)	81·1 (3·0)
Testis	94·7 (1·1)	94·0 (1·3)	91·1 (1·8)	88·2 (2·3)
Urinary bladder	82·1 (1·0)	76·2 (1·4)	70·3 (1·9)	67·9 (2·4)
Kidney and renal pelvis	61·8 (1·3)	54·4 (1·6)	49·8 (2·0)	47·3 (2·6)
Brain and other nervous system	32·0 (1·4)	29·2 (1·5)	27·6 (1·6)	26·1 (1·9)
Thyroid	96·0 (0·8)	95·8 (1·2)	94·0 (1·6)	95·4 (2·1)
Hodgkin's disease	85·1 (1·7)	79·8 (2·0)	73·8 (2·4)	67·1 (2·8)
Non-Hodgkin lymphomas	57·8 (1·0)	46·3 (1·2)	38·3 (1·4)	34·3 (1·7)
Multiple myeloma	29·5 (1·6)	12·7 (1·5)	7·0 (1·3)	4·8 (1·5)
Leukaemias	42·5 (1·2)	32·4 (1·3)	29·7 (1·5)	26·2 (1·7)

Rates derived from SEER 1973–98 database (both sexes, all ethnic groups).[12]
NOS=not otherwise specified.

Table 4: **Most recent period estimates of relative survival rates, by cancer site**

Figure 5 *– (a) The original table from the Lancet publication.*

Estimates of relative survival rates, by cancer site

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Figure 6 *– A slightly improved table by E. Tufte.*

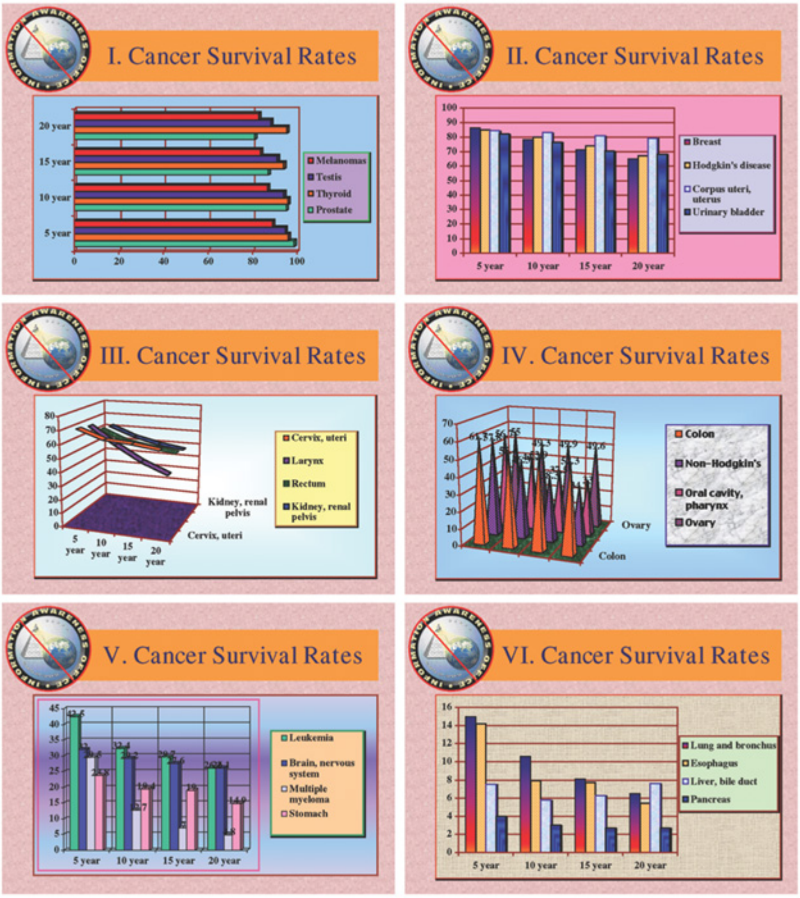


Figure 7 *– A Powerpoint presentation of the data in Table Figure 5.*

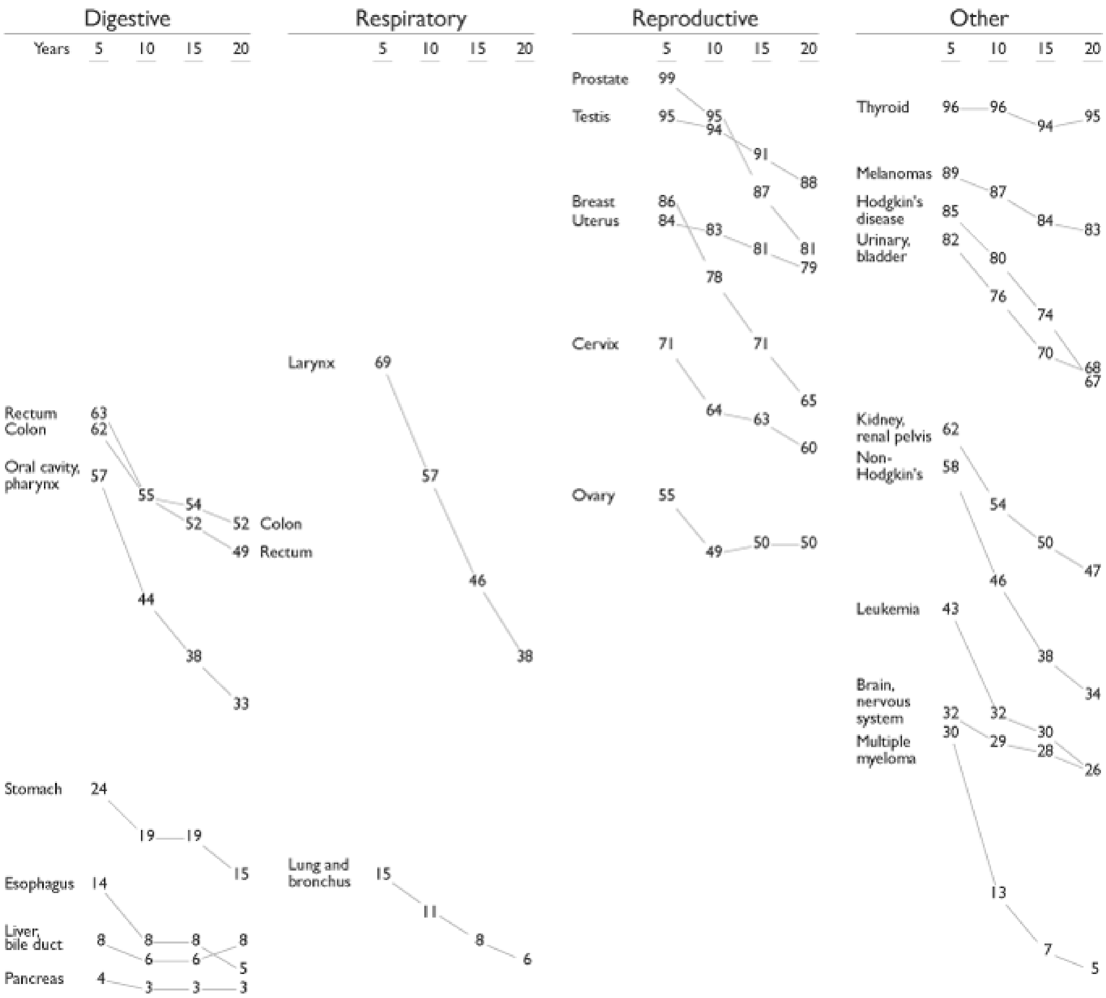


Figure 8 *– Graphical representation of a table. label:fig:cancertable4*

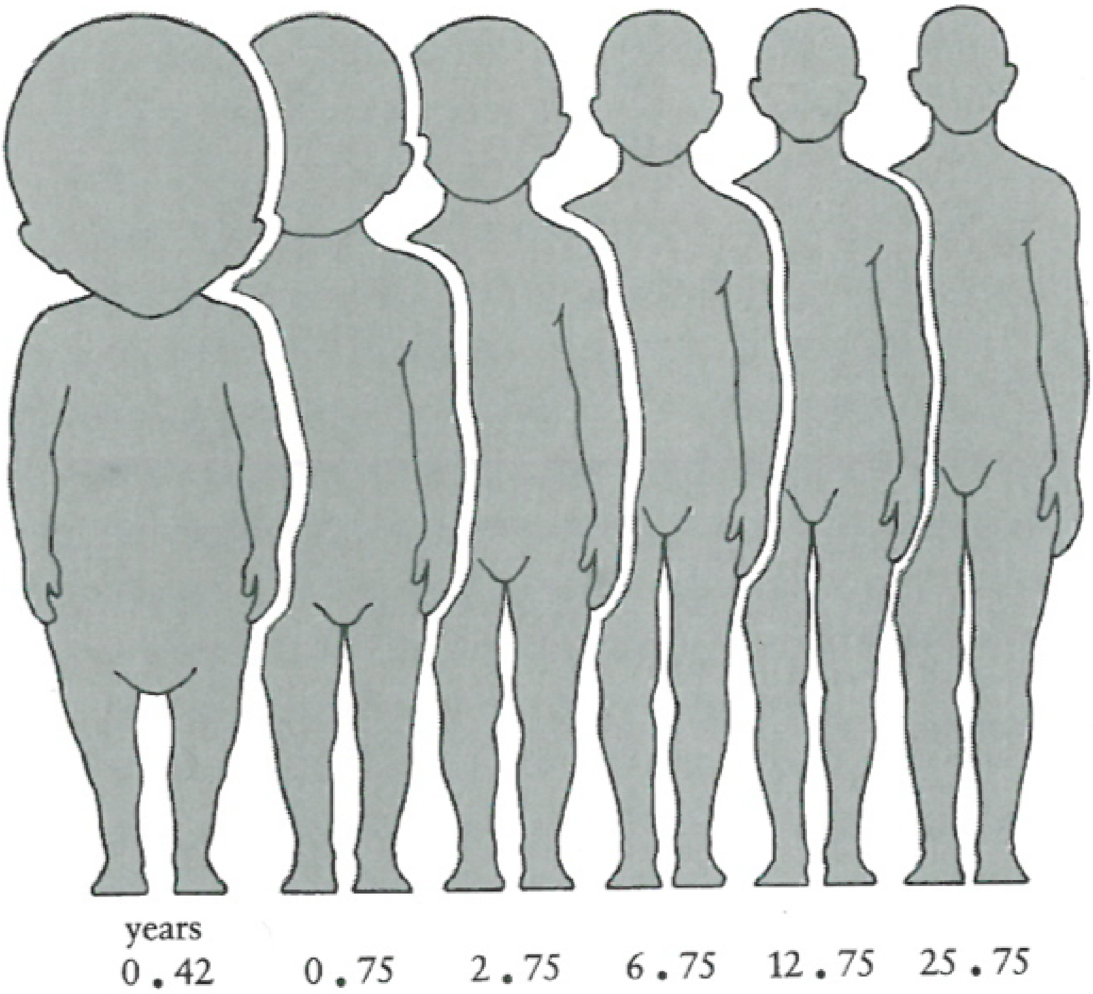


Figure 9 *– Change of body shape with increasing age in human development. Source: Tufte (2001)*

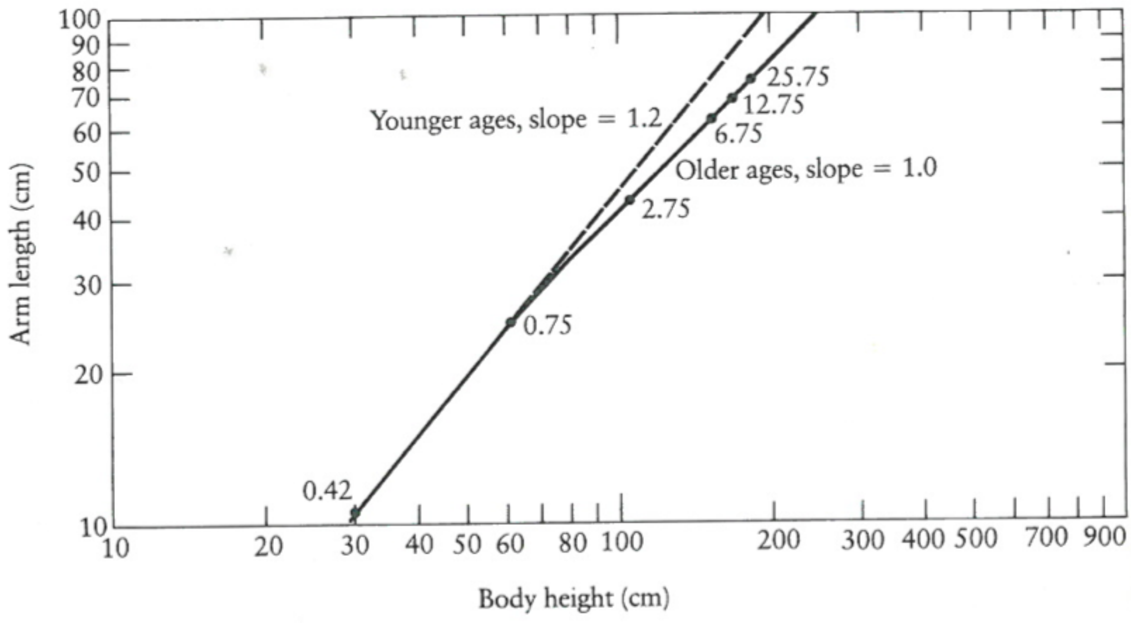


Figure 10 *– A plot of arm length against body height. Relationship of arm length and body length during different stages of human development. Source: Tufte (2001)*

Sometimes it may be appropriate to present two different types of visualizations. In Figure 11, the expansion of a fireball produced by a nuclear explosion is shown. The series of photographs indicates the growth and the shape of the fireball, but it is difficult to recognize whether the expansion is linear or exponential. Here, the plot provides the clear and unambigous information that the explosion is exponential.

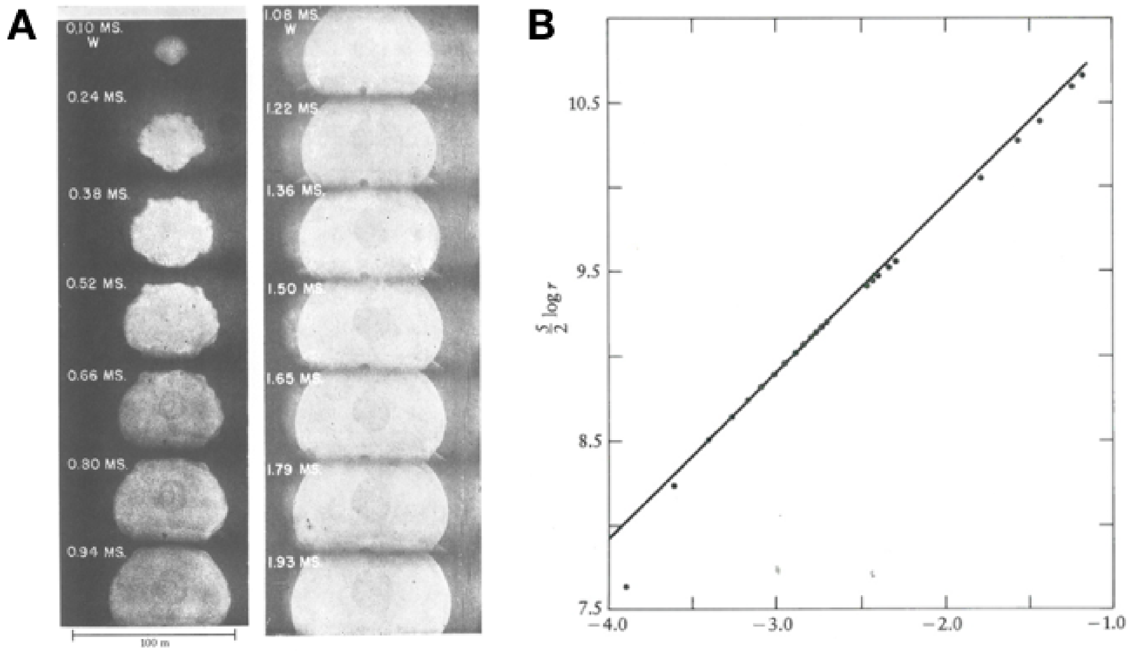


Figure 11 *– (a) Series of photographs of a nuclear explosion. (b) Plot of time since explosion against the width of the fireball (Measured as radius of the shock wave). The solid line indicates the theoretical result. Source: Tufte (2001)*

Plots can be annotated to include additional information and to make understanding easier. This is particularly interesting and useful if the general relationship of variables shown in the plot is very strong and unambigous (Figure 12 a). Additional annotations allow to check out details about data points (i.e., to learn about the identity of outliers) or to identify general patterns in the data.

More complex visualizations help to quickly identify 'hidden' aspects of the data. For example, Figure 12 b allows quickly to identify the unique position of humans as well as their rapid evolution with respect to the relative brain size.

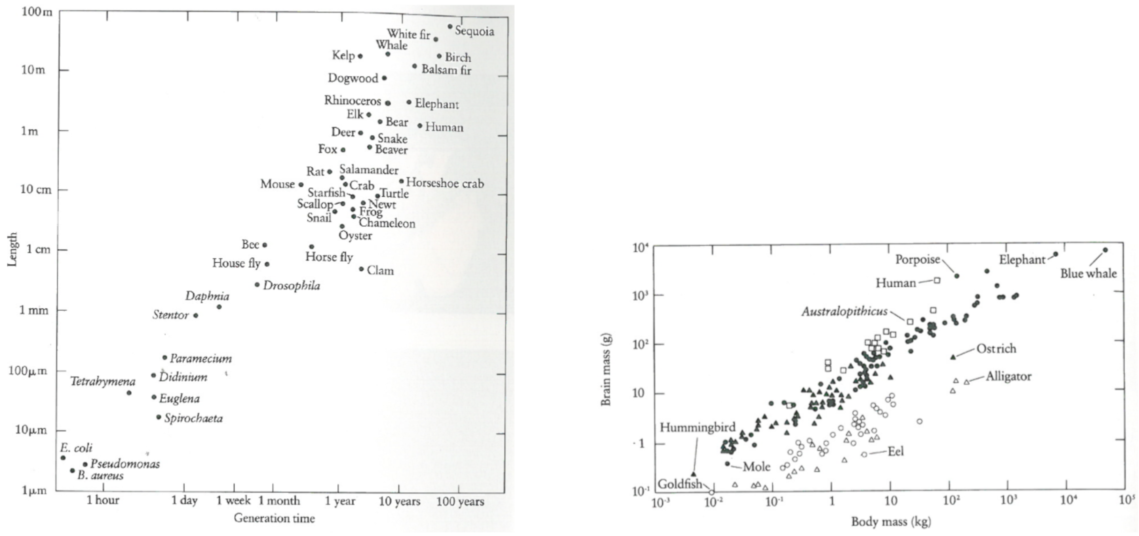
The visualization of data can also be used to easily identify deviations from a hypothesis. For example, it has been hypothesized by ecologists that there exists a relationship between species diversity and the size of the species. They developed a model that models the number of species, S, varies as the inverse square of the linear dimension, l,

$$S \propto l^{-2} \tag{2}$$

It has been shown that many groups of animals support this relationship. However, the available data sets deviate from this relationship for small data size. Thus the graphical visualization of the hypothesis and the data immediately point to deviations from the model.

5 Modern visualizations

The availability of powerful computers, large amounts of data and new algorithms for data analysis allows new ways of presenting data. One approach is the Worldmapper project which uses publicly



(a) (b)

Figure 12 – *(a) Relationship between the generation time and the length of an organism at the time of reproduction. (b) Brain size of vertebrates plotted against body size on a log-log graph. Primates are open squares; other mammals are solid dots, birds are solid triangles, bony fishes are open circles, and reptiles are open triangles. Source: McMahon and Bonner (1983)*

available data for creating instructive maps (Figure 14). In this type of map, the size of countries is changed according to the relative value of a data of interest.

Importantly, this representation works only well by using a *reference*, i.e., the well known world map. A simple demonstration of how important such a reference is, can be done by turning the map upside down. Then it becomes much more difficult to understand.

The increasing availability of public data allows comparisons between entities such as countries. By using modern visualization techniques interesting comparisons can be made that are useful for outlining relationships that previously unknown or unexpected such as the lack of a strong relationship between health spending and life expectancy (Figure 15). Furthermore, outliers become immediately obvious.

6 Application of visualization techniques in science

Scientific publishing is generally quite conservative. For this reason, classical approaches to visualization (plots, etc) are frequently used. However, there is a growing awareness for visualization of complex data, particularly in high-profile journals such as Nature and Science, who employ graphical designers to redraw and improve figures that are provided by authors of scientific papers. One consequence is that particularly in journals where the lengths of articles tend to be limited, figures become very - sometimes overly - complex (Figure 16).

As scientific data grow in their complexity, printed representation is not sufficient anymore and the data become web-based. Good examples are so-called genome browsers that allow to zoom through genome sequence data and to retrieve information (Figure 17).

There are limitations which amount of data can be meaningfully implemented into a picture. They are often shown in high-profile publications, but are likely essentially unusable if there is not a parallel, digitally enhanced figure that allows the browsing through the data in an internet browser (Figure 18).

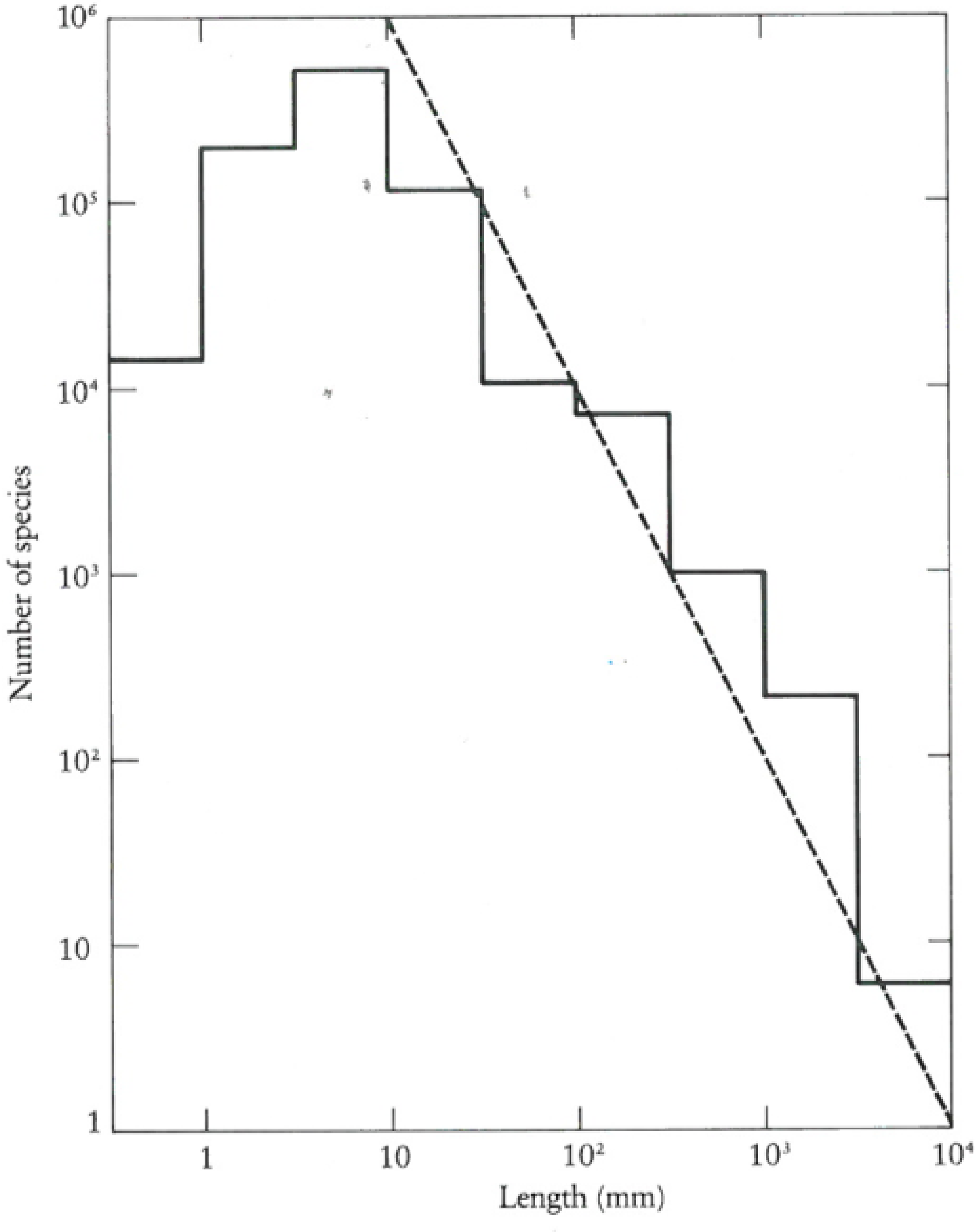


Figure 13 *– Number of species of all terrestrial animals classified according to their length. It should be noted that the numbers used are very rough estimates. The dashed line shows the expectation of an inverse proportion to the square of the length. Source: McMahon and Bonner (1983).*

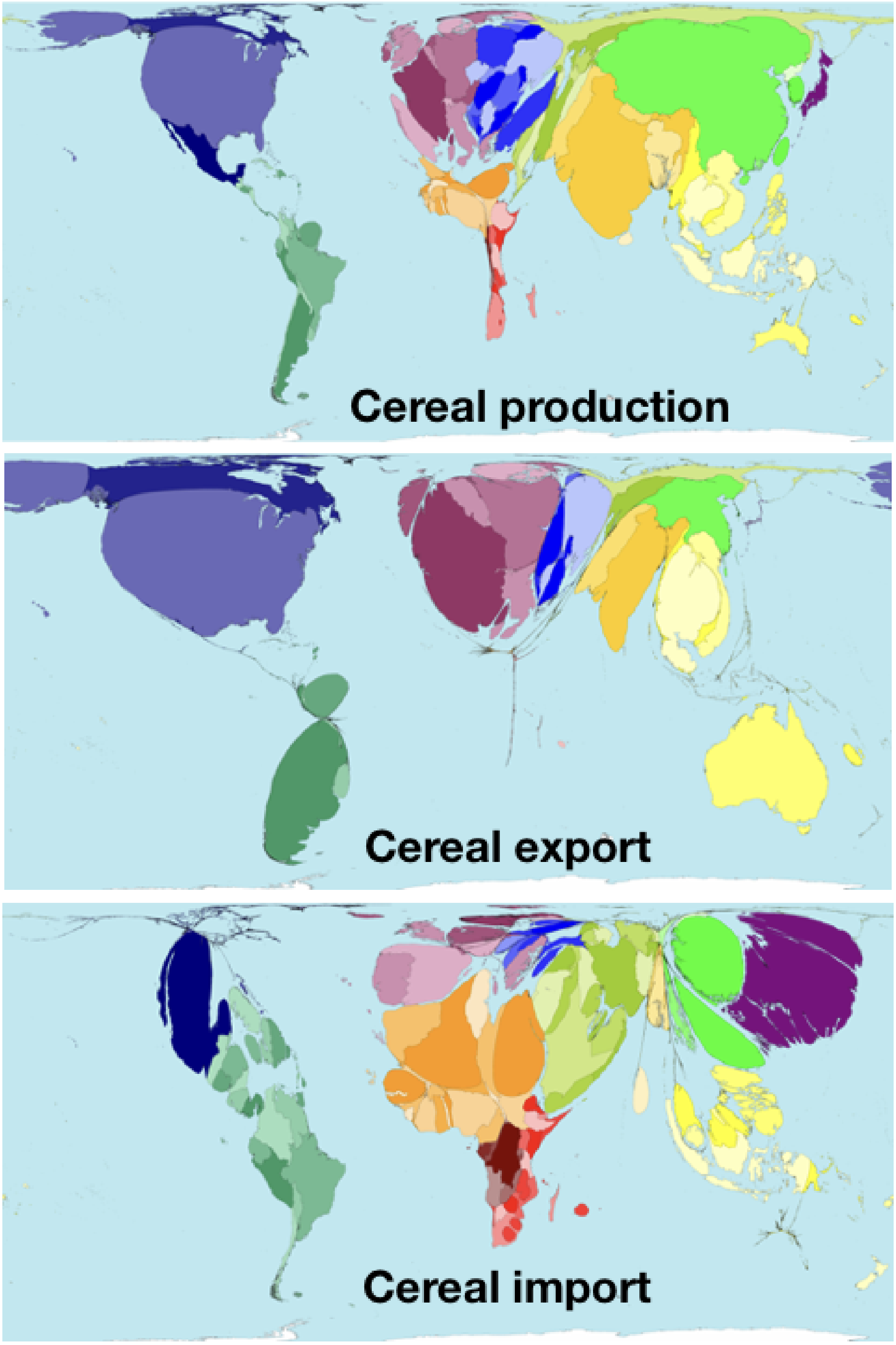


Figure 14 – *Visualizations of different data sets regarding the production and trading of cereals from the Worldmapper project. The original data are from FAO and the figures from http://www.worldmapper.org.*

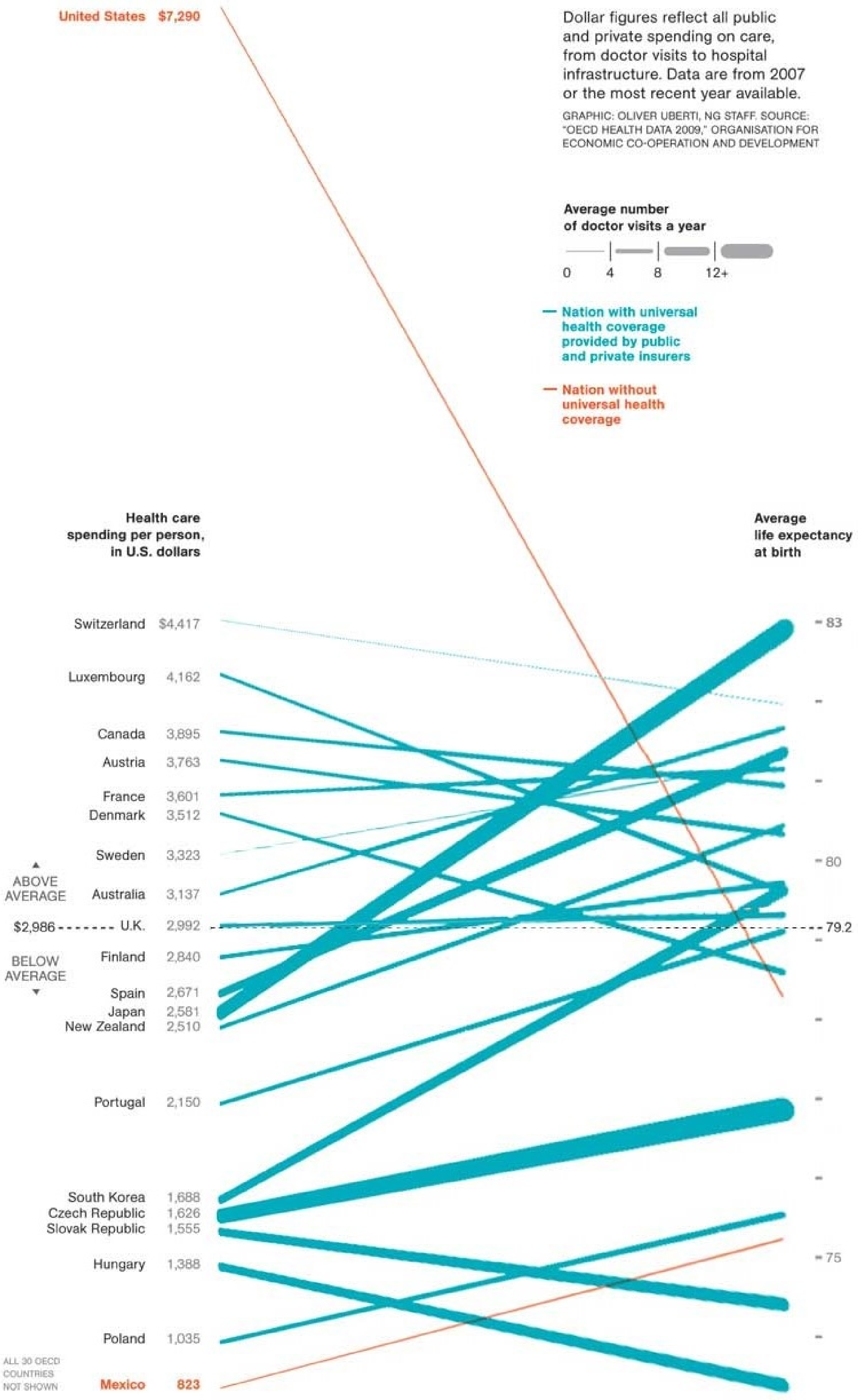


Figure 15 *– Relationship between health spending (left side) and life expectancy (right side) in different countries. The data for the United States and Mexico are shown in orange. Source: National Geographic.*

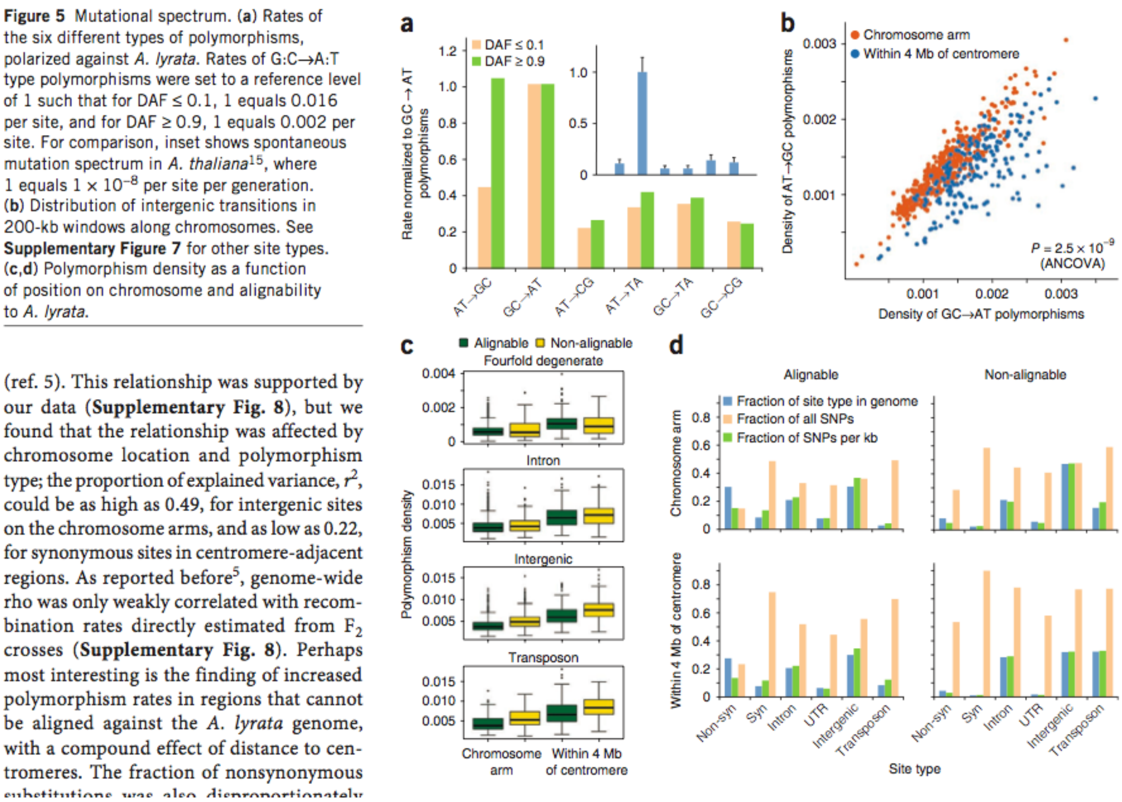


Figure 16 *– Composite figure in a paper from Nature Genetics. Source: Cao et al. (2011)*

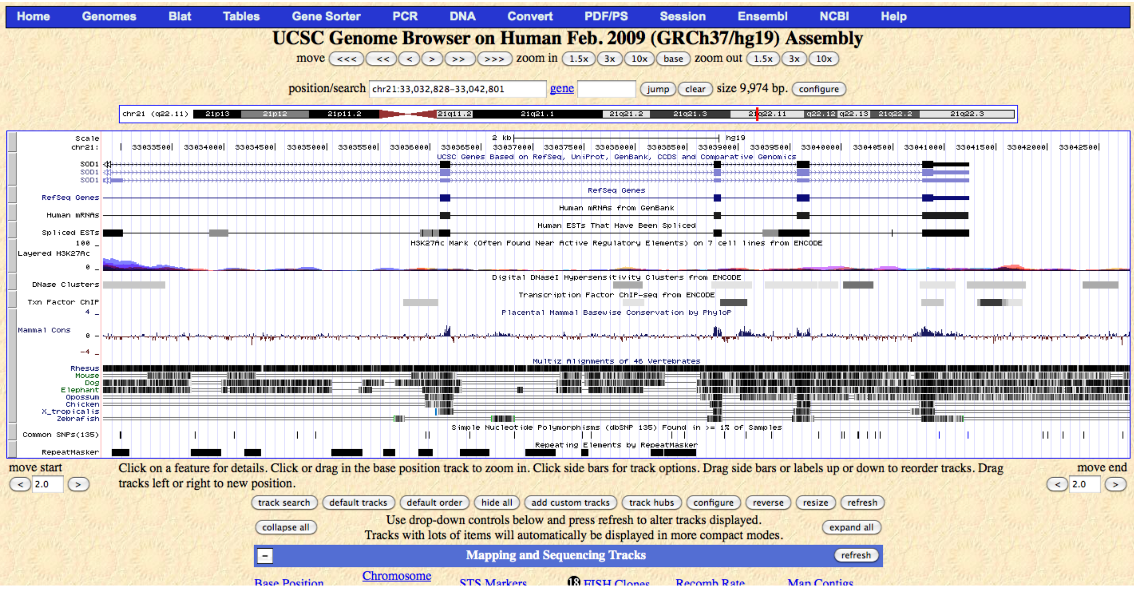


Figure 17 *– Screenshot of the UCSC genome browser (http://genome.ucsc.edu). This browser allows to navigate through the human genome. Similar browsers exist for plants.*

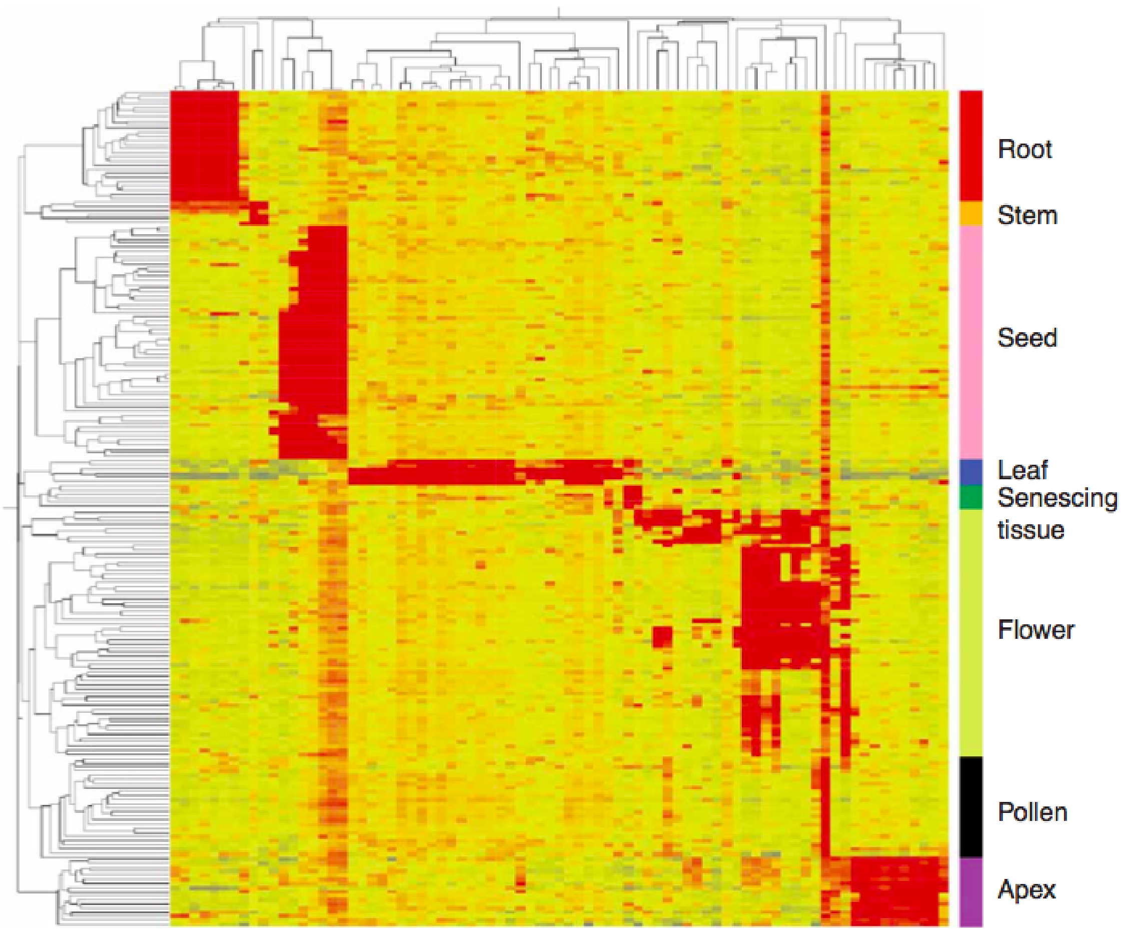


Figure 18 – *Heat map of tissue specific marker genes in the model plant* Arabidopsis thaliana*. Source:Schmid et al. (2005)*

7 Summary

- The main purpose of this lecture is to provide a basic introduction into the importance and diversity of visualization approach in science.
- One can follow established rules for the visualization (or presentation) of scientific results, but it can also be viewed as a creative process.
- The key requirements of visualizations is that it should be correct, truthful (i.e., not manipulative) and user-friendly.

8 Key concepts

- ◻ Data to ink ratio (Tufte)
- ◻ Glass slippers (Bergstrom and West: Calling Bullshit)
- ◻ Principle of proportional ink (Bergstrom and West: Calling Bullshit)

9 Further reading

9.1 General reading

- Edward Tufte, *The Visual Display of Quantitative Information*. 2nd Ed. Graphics Press. (2004) - The classic book on visualization.
- T. A. McMahon and J. T. Bonner, *On Size and Life*. Scientific American Library (1983) - A very beautiful book on the size relationships in living organisms with an excellent application of visualization principles.

9.2 Practical tips

- Vandemeulebroecke et al. (2019) - An excellent tutorial on how to visualize quantitative information. Read this to be prepared for writing your master thesis.
- Rougier et al. (2014) - Ten simple rules for better figures https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
- Claus Wilke wrote an extremely useful book about practical aspects of data visualization with corresponding code in R: https://serialmentor.com/dataviz/
- Nature Methods Points of View articles on visualization: http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

There are also some very good internet resources:

- Website of Edward Tufte: http://www.edwardtufte.com
- Worldmapper: http://www.worldmapper.org
- R Graphics Gallery: http://addictedtor.free.fr/graphiques/
- Hans Rosling and his gapminder program: http://www.gapminder.org - Also check out his videos!

For your master thesis, using the R package to plot the figures is a good start. There are numerous free introductions into R graphics.

10 Discussion questions

Based on the chapter 7 *Data Visualization*, Calling Bullshit.

1. Why are concept maps such as the subway map and Venn diagrams susceptible to misuse?
2. What is the problem of **binning** data for visualization?
3. What needs to be considered when plotting data with two different y-axes?
4. Which arguments play a role when plotting absolute vs. relative values in different types of plots like bar plots, line plots.

   Think about the roles of:

   - Effect size
   - Sample size
   - Types of comparison

11 In class exercises

11.1 Some example plots from the literature

Discuss the following points:

- Key message and underlying visualization concept
- Advantage of concept
- Disadvantage of concept

Corona vaccinations and deaths in Europe

Source: https://twitter.com/Vuckolino/status/1480570696275316739?s=20

How could the plot improved or plotted differentially to identify countries that are outliers in the relationship between vaccination status and corona-related deaths?

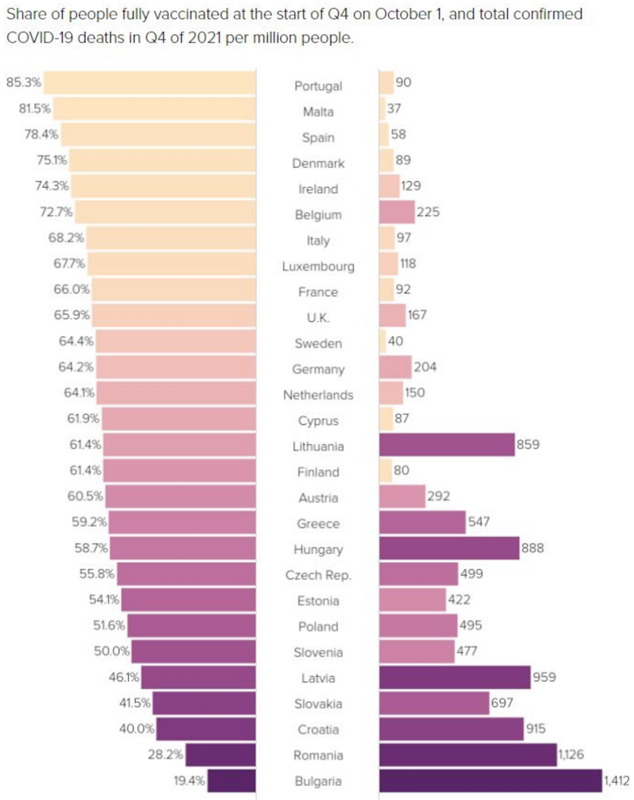


Figure 19 *– Comparison of vaccination status with corona deaths.*

12 Gene families in different crops

The image shows the distribution of gene families of the banana in other crops.

What are the key messages of this figure?

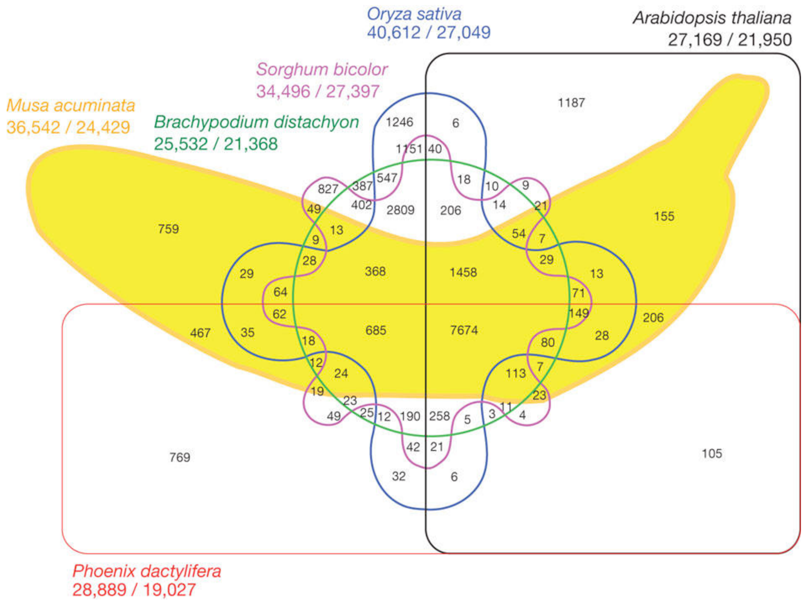


Figure 20 – *Six-way Venn diagram showing the distribution of shared gene families (sequence clusters) among* M. acuminata, P. dactylifera*,* Arabidopsis thaliana*,* Oryza sativa*,* Sorghum bicolor *and* Brachypodium distachyon *genomes. Source: D'Hont et al. (2012)*

13 Anthocyanin accumulation in strawberries

Source: Low temperature inhibits anthocyanin accumulation in strawberry fruit by activating FvMAPK3-induced phosphorylation of FvMYB10 and degradation of Chalcone Synthase 1 Wenwen Mao, Yu Han, Yating Chen, Mingzhu Sun, Qianqian Feng et al. The Plant Cell, koac006, https://doi.org/10.1093/plcell/koac006 (2022)

Could the message be improved by a different type of graph?

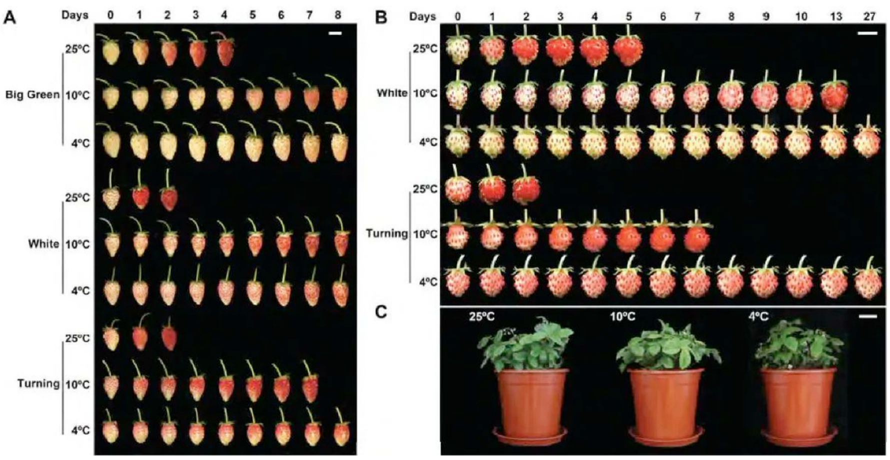


Figure 21 – *Low temperature inhibits anthocyanin accumulation in strawberry fruit by activating FvMAPK3-induced phosphorylation of FvMYB10 and degradation of Chalcone 1. Source: Mao et al. (2022)*

References

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* **43**:956–963. doi:10.1038/ng.911

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci A-M, Weissenbach J, Ruiz M, Glaszmann J-C, Quétier F, Yahiaoui N, Wincker P. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**:213–217. doi:10.1038/nature11241

Mao W, Han Y, Chen Y, Sun M, Feng Q, Li L, Liu L, Zhang K, Wei L, Han Z, Li B. 2022. Low temperature inhibits anthocyanin accumulation in strawberry fruit by activating FvMAPK3-induced phosphorylation of FvMYB10 and degradation of Chalcone Synthase 1. *The Plant Cell* **34**:1226–1249. doi:10.1093/plcell/koac006

McMahon TA, Bonner JT. 1983. On size and life. Scientific American Library.

Rougier NP, Droettboom M, Bourne PE. 2014. Ten Simple Rules for Better Figures. *PLoS Comput Biol* **10**:e1003833. doi:10.1371/journal.pcbi.1003833

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of Arabidopsis thaliana development. *Nature Genetics* **37**:501–506. doi:10.1038/ng1543

Tufte E. 1997. Visual explanations: Images and quantities, evidence and narrative. Cheshire, Connecticut: Graphics Press.

Tufte ER. 2001. The visual display of quantitative information, 2nd ed. Cheshire, Connecticut: Graphics Press.

Vandemeulebroecke M, Baillie M, Margolskee A, Magnusson B. 2019. Effective Visual Communication for the Quantitative Scientist. *CPT: Pharmacometrics & Systems Pharmacology* **8**:705–719. doi:10.1002/psp4.12455