Visualization

3502-440 Methods of Scientific Working for Crop Science

WS 2024/2025

Prof. Dr. Karl Schmid
Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

Background

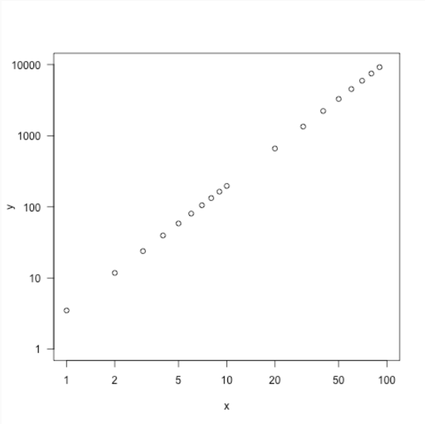
Modern visualizations

Background

A picture is worth a thousand words

- Summarize complex data
- Table or figure?
- Which designs for a figure are possible?
- Data mining: Find new patterns if figures.

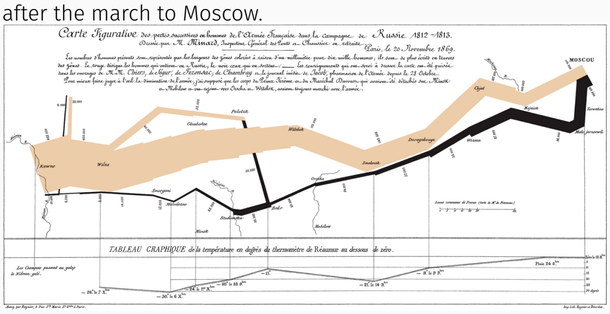
x	y
1	3.5
2	11.8
3	23.9
4	39.6
5	58.5
6	80.5
7	105.4
8	133.0
9	163.7
10	196.8
20	662.0
30	1,345.9
40	2,226.8
50	3,290.5
60	4,527.2
70	5,929.1
80	7,489.9
90	9,204.3



Visualization

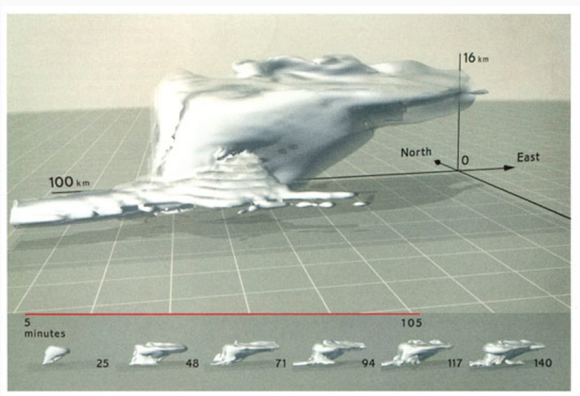
One of the first visualizations containing a large number of data points by Charles Joseph Minard (1781-1870), a French Civil Engineer.

The graphic shows the changes in Napoleon's army during and after the march to Moscow.



Visualization of four dimensions in two dimensions

Development of a storm over time

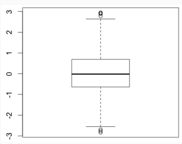


E. Tufte: Visual Explanations: Images and Quantities, Evidence and Narrative, 1997

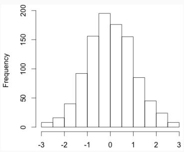
Information content of graphics

```
> a <- rnorm(1000,0)
> a
   [1] -0.0210952179 -1.0912065990 -1.1395054396 -1.3189259016 -1.2876563450 -0.6368083081 -0.3961465716
   [8] -0.2912739666 -0.1470279242  1.7375103224  1.2480682763  0.7357591312 -0.0605690848  0.7640946370
 ...
 [988]  0.0045077573 -0.1658796133 -0.2793556344 -0.1170640982 -0.5272327964  0.2952162296  2.4884583004
 [995]  2.0871831247 -0.4179269672 -0.4178945557 -0.3895868971  0.5237880959  2.1611690885
```

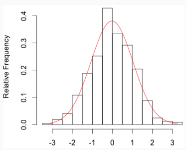

Information content of graphics



(a)



(b)



(c)

Visualization improves understanding of data

1. Original scientific publication
2. Modification of table by Edward Tufte
3. Converting the table into a graph by Edward Tufte
4. A typical powerpoint visualization of the same data
5. An even better visualisation by Dave Nash

Answer the following questions

1. Which cancer type has the highest survival probability after 5 years and after 20 years?
2. Which cancer type has the lowest survival probability after 5 years and after 20 years?
3. For which cancer type does the survival probability between 5 and 20 years drop the most?
4. For which tissue type is the average mortality (over cancer types) the highest and for which the lowest?

	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.

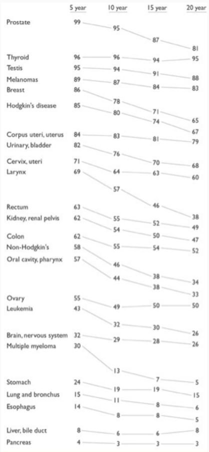
Table 4: **Most recent period estimates of relative survival rates, by cancer site**

Herman, The Lancet, 2002

Estimates of relative survival rates, by cancer site

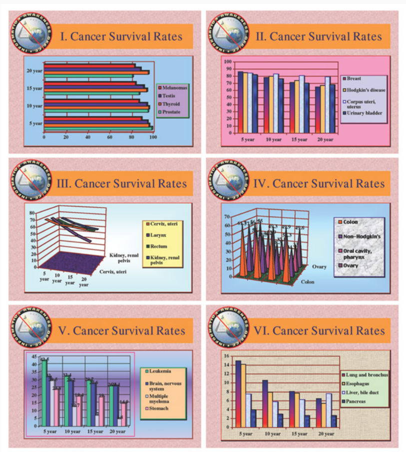
	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
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

www.edwardtufte.com



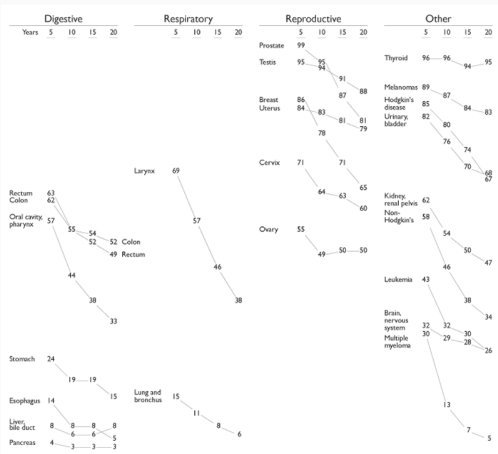
www.edwardtufte.com

13 / 30



www.edwardtufte.com

14 / 30

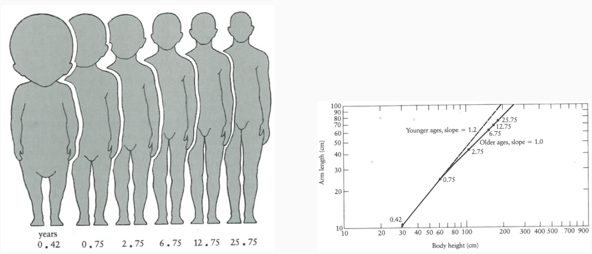


www.edwardtufte.com

15 / 30

Which type of visualization is appropriate?

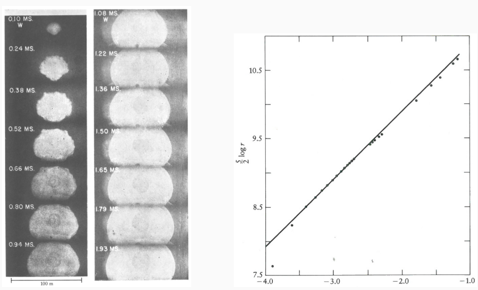
- Ease of understanding
- Background knowledge required
- Complexity: How much time does it take to understand the figure?
- Data to ink ratio



(d) (e)

Figure 1: (a) Change of body shape with increasing age in human development. (b) A plot of arm length against body height. Relationship of arm length and body length during different stages of human development.

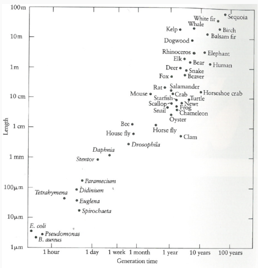
When are two types of information required?



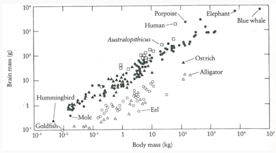
(a) (b)

Figure 2: (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.

Annotation of scientific graphics



(a)



(b)

Figure 3: (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.

Finding deviations from a biological model

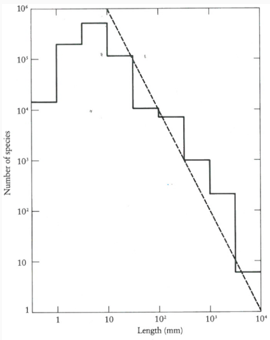
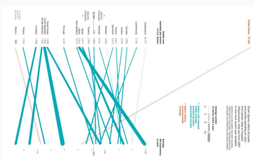


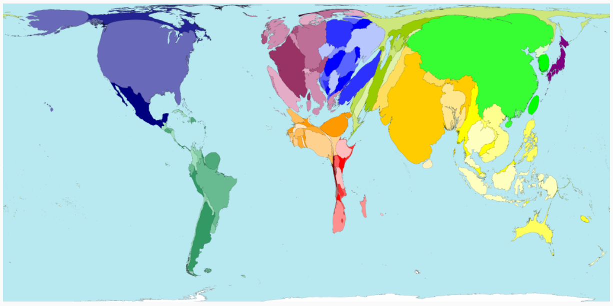
Figure 4: Number of species of all terrestrial animals classified according to their length. It should be noted that the numbers used are very rough estimates. The dashed line shows the expectation of an inverse proportion to the square of the length.

Relationship between health spending and life expectancy



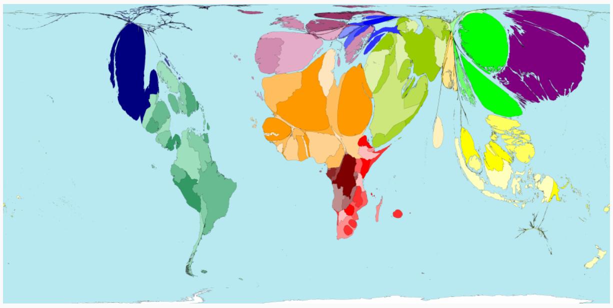
National Geographic

World cereal production



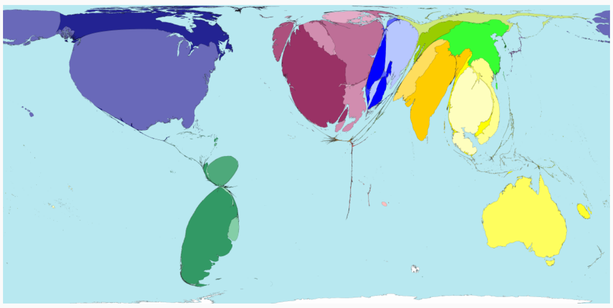
www.worldmapper.org

World cereal imports



www.worldmapper.org

World cereal exports



www.worldmapper.org

High complexity of scientific representations

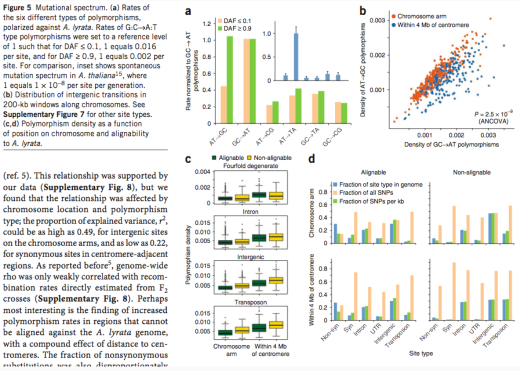


Figure 5: Composite figure in a recent paper from Nature Genetics. Source: Cao et al. (2011)

High complexity of modern scientific representations

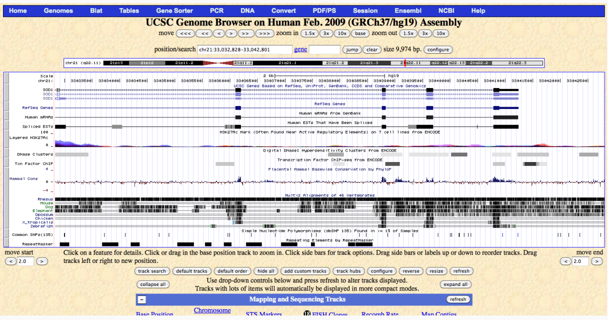


Figure 6: Screenshot of the UCSC genome browser (`genome.ucsc.edu`).

High complexity of modern scientific representations

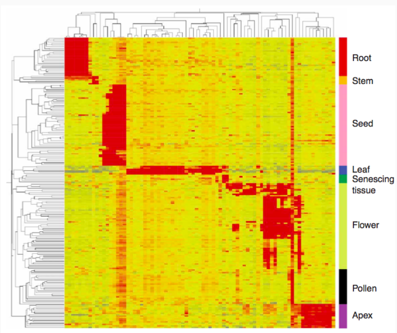


Figure 7: Heat map of tissue specific marker genes in the model plant *Arabidopsis thaliana*. Source: Schmid et al. (2005)

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.

Sources for reading about the importance of visualization

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

What are the advantages and disadvantages of visualization techniques?

References i

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, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–963.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, 37(5):501–506. Number: 5 Publisher: Nature Publishing Group.