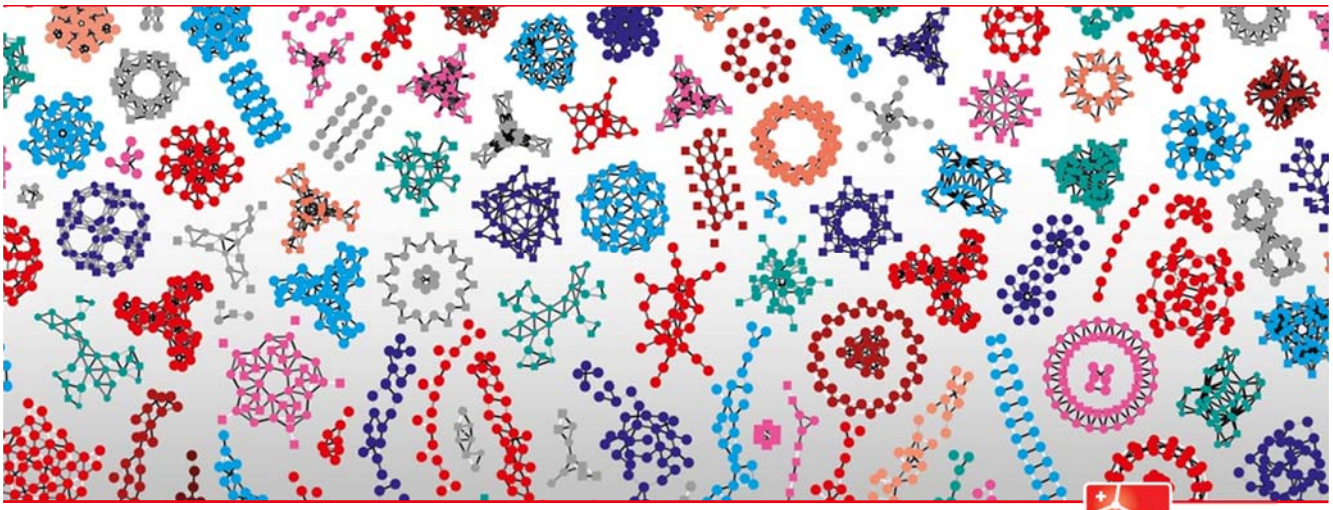


Intermediate statistics: data analysis in practice

21-22 April 2021

Isabelle Dupanloup, Rachel Marcone, Frédéric Schütz

Swiss Institute of Bioinformatics

www.sib.swiss

What do you think is easy and what do you think is complicated with data analysis ?

- Exploratory data analysis
- Data summarization: mean, median, SD, SEM, CI
- Graphics, tables
- Statistical tests
  - Student's t-tests (one- and two-sample, paired)
  - Wilcoxon
  - Fisher's exact test
  - Chi-square
  - ANOVA
- Correlation, linear regression
- Principal component analysis (PCA)

What are the general steps of data analysis ?

data.xls – LibreOffice Calc

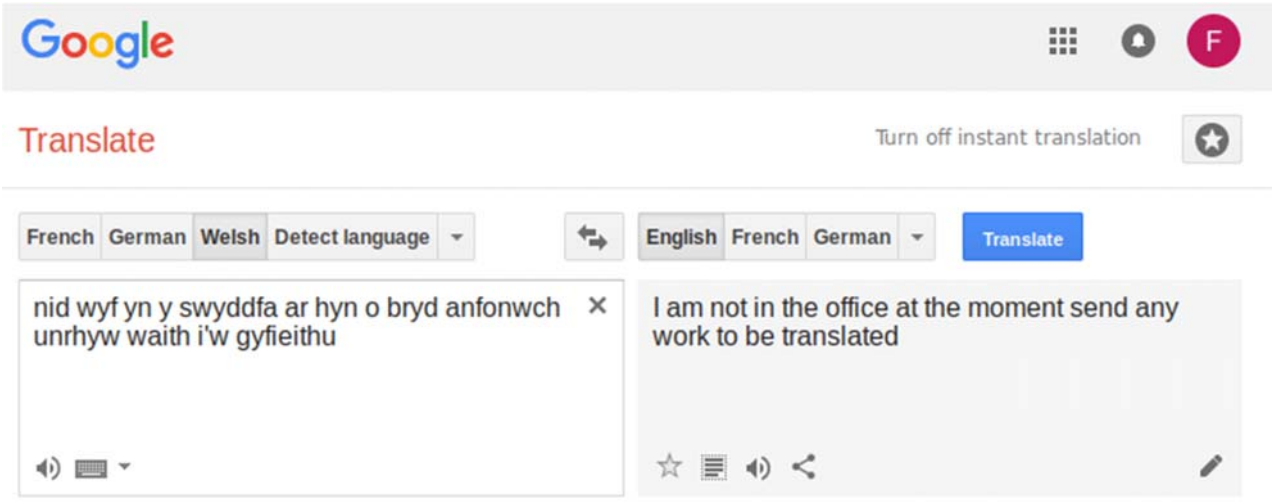
File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1																							
2		WT				0	0	1	2	3	4	5	6	8	10	12	14	16	17	18	19	21	22
3		HFD		HFD	WB no / ID	2-Nov-05	10-Nov-05	9-Nov-05	16-Nov-05	23-Nov-05	30-Nov-05	7-Dec-05	14-Dec-05	28-Dec-05	11-Jan-06	25-Jan-06	8-Feb-06	23-Feb-06	2-Mar-06	10-Mar-06	17-Mar-06	30-Mar-06	4-Apr-06
4	1	1	WT	HFD	224	23.1		23.6	24.4	25.6	25.3	25.1	25.2	26.2	29.1	29.5	29.8	30.7	30.5	31.2	31.5	31.4	31.9
5	1	3	WT	HFD	223	21.1		21.3	21.6	22.6	23.2	24	25.4	27.6	29.3	30.9	31.3	33.3	31.6	32.1	32.5	30.4	30.3
6	1	5	WT	HFD	229	20.2		22.6	23.7	25.1	25.7	26.2	26.6	27.6	29.0	29.9	29.7	30.4	29.6	30.3	31.5	30.5	30.9
7	1	7	WT	HFD	248	18.5		24.6	26.7	29.0	30.7	31.8	33.5	35.8	38.6	40.2	41.3	41.9	43	45	46.4	47.4	47.4
8	1	9	WT	HFD	254	17.6		23	27.1	29.5	30.3	30.8	31.8	33.2	35.1	35.7	36.6	37.9	37.4	39.2	39.3	40	40.8
9	1	11	WT	HFD	247	17.2		21.7	26.2	27.7	28.8	29.6	30.9	32.2	33.1	34.2	34.4	36.6	37.2	38.8	40.2	39.2	41.6
10	1	13	WT	HFD	256	16.4		22.9	25.0	27.0	27.9	29	29.4	30.9	33.4	35.8	37	39.3	39	41.8	43.2	47.1	48.5
11	1	15	WT	HFD	240	16.1		21.8	24.1	26.3	28.1	29.4	29.0	34.0	35.8	39.9	41.9	45.1	44.8	46.2	47.9	49.2	49.5
12	1	17	WT	HFD	234	15.7		22.8	23.6	25.3	25.6	26.2	26.6	31.0	33.1	34.2	36.5	37.3	36.7	35.8	37.3	38.7	39.7
13	1	19	WT	HFD	241	15.4		21.3	22.0	22.8	23.2	24.8	25.9	29.4	30.9	32.0	33.2	34	33.3	35.4	36.2	36.8	37.3
14	1	21	WT	HFD	243	15		21.5	23.1	24.6	24.7	26.9	29.2	33.7	36.6	39.1	41	42.4	42.8	41.1	44.4	46.1	47.4
15	1	23	WT	HFD	245	14.3		20.3	22.1	23.0	28.1	25.3	26.8	30.9	35.1	37.6	40.8	43.6	43.3	44.9	46.8	48.4	48.4
16	1	25	WT	HFD	280		15.5	16.8	18.1	18.9	20.2	20.5	21.1	22.3	24.8	25.6	26.1	28	27.5	28.7	29.5	28.7	29.5
17	1	27	WT	HFD	282		19.8	21.3	22.8	24.5	26.7	27.4	28.2	31.8	24.4	36.3	37.7	39.7	40.9	42.4	43.8	44.8	45.9
18	1	29	WT	HFD	283		19.6	20.1	20.6	21.2	21.7	23.1	23.4	24.7	27.2	28.4	28.4	30.5	30.6	30.9	32.1	31.1	31.3
19		WT				0	0	1	2	3	4	5	6	8	10	12	14	16	17	18	19	21	22
20		FEN-HFD			WB no / ID	2-Nov-05	10-Nov-05	9-Nov-05	16-Nov-05	23-Nov-05	30-Nov-05	7-Dec-05	14-Dec-05	28-Dec-05	11-Jan-06	25-Jan-06	8-Feb-06	23-Feb-06	2-Mar-06	10-Mar-06	17-Mar-06	30-Mar-06	4-Apr-06
21	2	1	WT	FEN-HFD	222	23.5		22.4	24.3	26.1	27.8	28.4	29.6	30.7	32	32.6	32.1	33.9	33.4	33.7	33.9	34.5	34.7
22	2	3	WT	FEN-HFD	250	21.7		23.5	24.5	25.6	26.8	27.3	28.6	29.9	31.5	32.2	33.0	35	35.4	35.2	36.5	35.5	35.6
23	2	5	WT	FEN-HFD	227	20.3		22.2	23.4	24.4	24.9	25.4	26.2	27.4	27.3	28.9	29.0	30.3	30.3	30.3	31.2	30.7	31.1
24	2	7	WT	FEN-HFD	226	19.5		21.1	22.6	23	24.1	24.1	24.4	26.8	27.5	29	28.3	28.9	29	27.9	28.5	28.7	27.9
25	2	9	WT	FEN-HFD	253	17.6		23.8	25.2	26.2	27.5	28.6	29.6	30.1	32.2	33.5	33.3	34.3	33.9	33.9	34.8	35	35.4
26	2	11	WT	FEN-HFD	252	17.5		21.9	23.2	25.2	25.9	26.9	28.9	30.1	33.7	34.9	35.2	36.4	37	38.6	39.1	40.5	39.6
27	2	13	WT	FEN-HFD	251	16.5		21.6	22.2	24	25.2	25.7	26.8	28.2	30.5	31.5	32.4	33.5	34.2	33.4	33.8	33.4	33.6
28	2	15	WT	FEN-HFD	249	16.3		22.8	24.2	25.6	25.8	27.2	28.2	29	31	31.3	31.3	31.7	33.8	33.7	34.6	35.8	35.9
29	2	17	WT	FEN-HFD	242	15.9		21.2	22.6	23.9	23.8	23.8	25.2	26.6	28.9	30.5	32	34	34.2	34.9	35	37.3	37.8
30	2	19	WT	FEN-HFD	244	15.7		20.7	22.3	23.1	24.3	25.4	26.5	28.5	31	31.4	32.3	33.3	32.3	33.5	33.3	33.8	33.1
31	2	21	WT	FEN-HFD	246	15.2		21	23	25.4	26.6	28.1	29.4	33.4	36.9	39.6	41.2	45.2	46.2	48.1	49.7	50.7	53.1
32	2	23	WT	FEN-HFD	236	14.4		19.9	21.7	23.8	24	23.9	24.3	26.5	28.6	29.6	31.2	33.3	32.9	34	34.2	34.9	34.4
33	2	25	WT	FEN-HFD	287		14.5	16.7	18.8	21.1	22.3	24	24.7	26.4	28.6	30.5	31.3	32.7	33	33.2	34.3	35.1	35.2
34	2	27	WT	FEN-HFD	284		20.2	20.9	21.6	22.2	22.7	23.2	23	25.1	25.2	26.8	27	29	29.2	29.4	29.8	30.1	30.4
35	2	29	WT	FEN-HFD	288		16.7	18.6	20.5	22.6	23.7	25.1	25.2	27.8	30.2	31.4	32	32.8	32.4	32.4	33.1	34.1	34.2
36		KO				0	0	1	2	3	4	5	6	8	10	12	14	16	17	18	19	21	22
37		HFD			WB no / ID	2-Nov-05	10-Nov-05	9-Nov-05	16-Nov-05	23-Nov-05	30-Nov-05	7-Dec-05	14-Dec-05	28-Dec-05	11-Jan-06	25-Jan-06	8-Feb-06	23-Feb-06	2-Mar-06	10-Mar-06	17-Mar-06	30-Mar-06	4-Apr-06
38	2	2	KO	HFD	206	22.2		25.8	26.8	28.7	29.9	30.8	31.9	31.6	33.6	34.2	35.1	36.7	37.1	37.6	38.1	39.8	40.1
39	2	4	KO	HFD	201	21.6		25.7	27.8	30.5	31.4	32.4	33.7	36.3	38.7	41.1	42.2	42.2	41.4	45.2	47.1	45.1	44.8
40	2	6	KO	HFD	203	21.4		25	26.5	27.6	28.6	29.5	30.2	32	35.4	36.8	37	39.3	41	41.1	38.2	40.9	41.8

new style BW sheet (2) / july 14 / july 28 / Aug 3 / aug 11 group making / Aug 17 / Aug22 / Making new groups 31 Aug / Body weights cohort1

Sheet 1 / 30 PageStyle_new style BW sheet (2) Sum=0 75%

Nid wyf yn y swyddfa ar hyn o bryd. Anfonwch unrhyw waith i'w gyfieithu



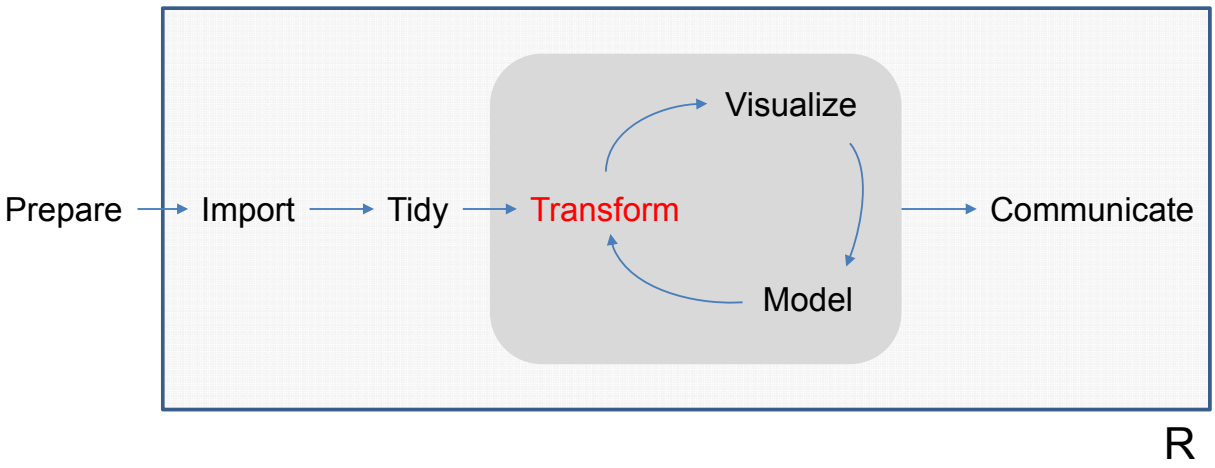
Some important rules of statistical analysis

Know the story around the data

Ask concrete questions

Always look at the data!

Data analysis workflow

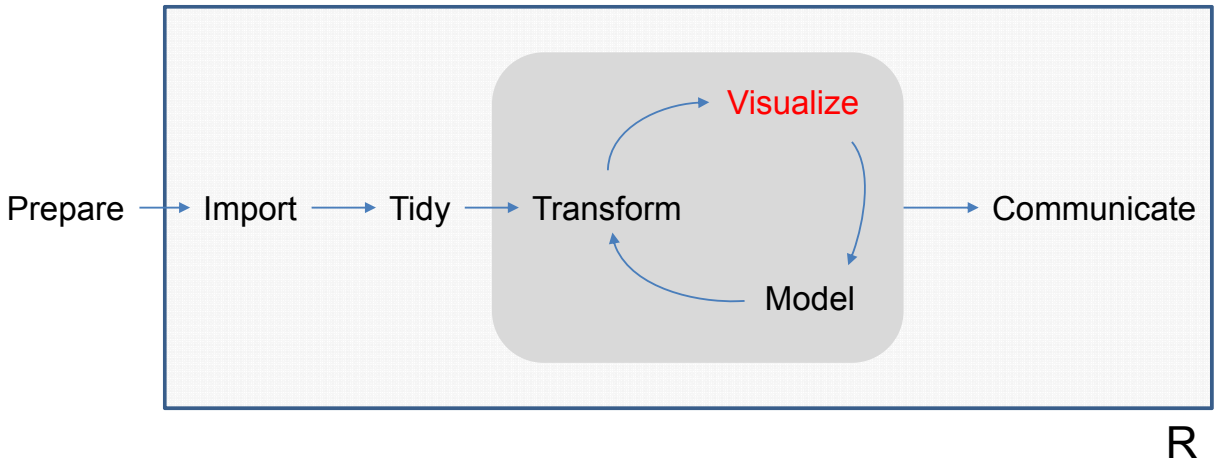


Should we really "transform" the data ?

Examples of transformations

- **Summarization**: mean, median, etc
- **Create new variables:**
  - Combine variables (height+weight → BMI)
  - Change the scale of a variable or units
  - Normalize
  - Log-transform
- **Aggregate data**

Data analysis workflow



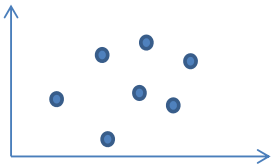
Adapted from Hadley Wickham

Dataset: four pairs of variables

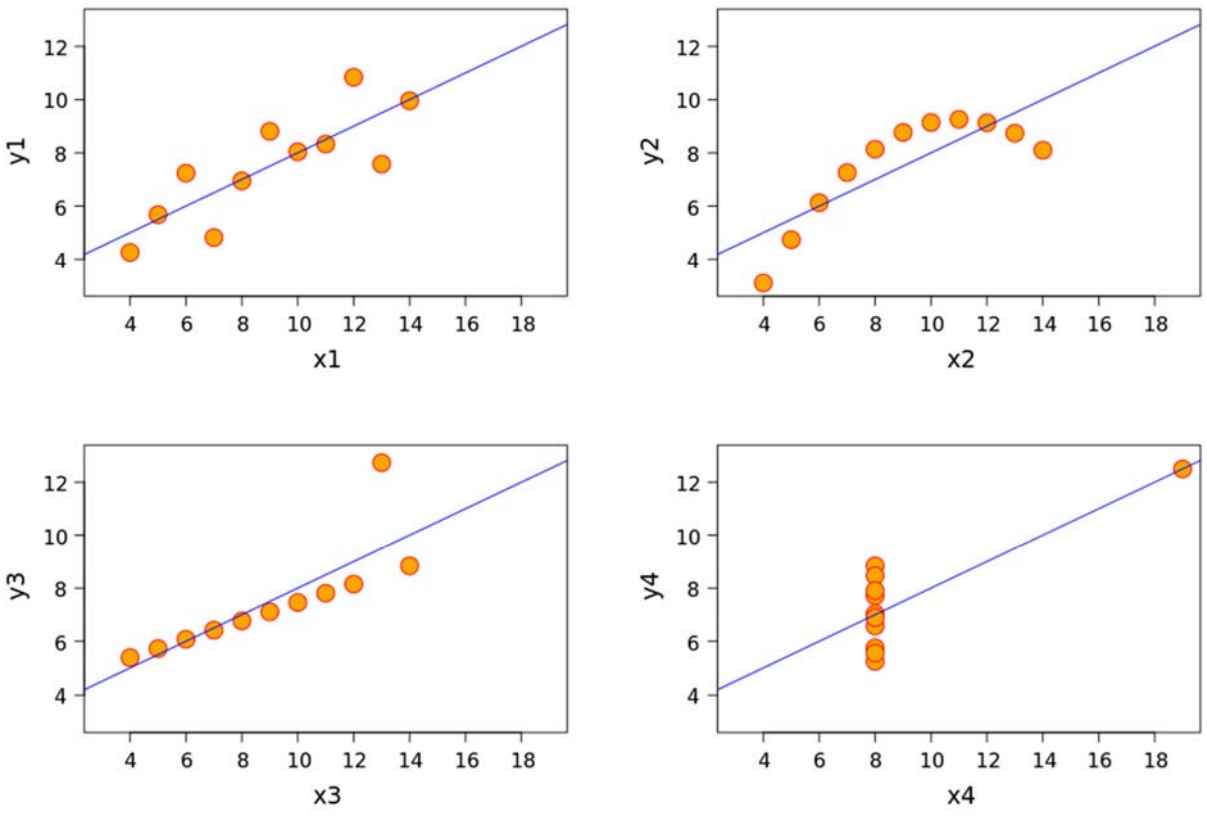
	X1	X2	X3	X4	Y1	Y2	Y3	Y4
Mean	9.0	9.0	9.0	9.0	7.5	7.5	7.5	7.5
Standard deviation	3.3	3.3	3.3	3.3	2.0	2.0	2.0	2.0

	X1 vs Y1	X2 vs Y2	X3 vs Y3	X4 vs Y4
Correlation	0.81	0.81	0.81	0.81
Regression line	Y = 3 + 0.5x	Y = 3 + 0.5x	Y = 3 + 0.5x	Y = 3 + 0.5x

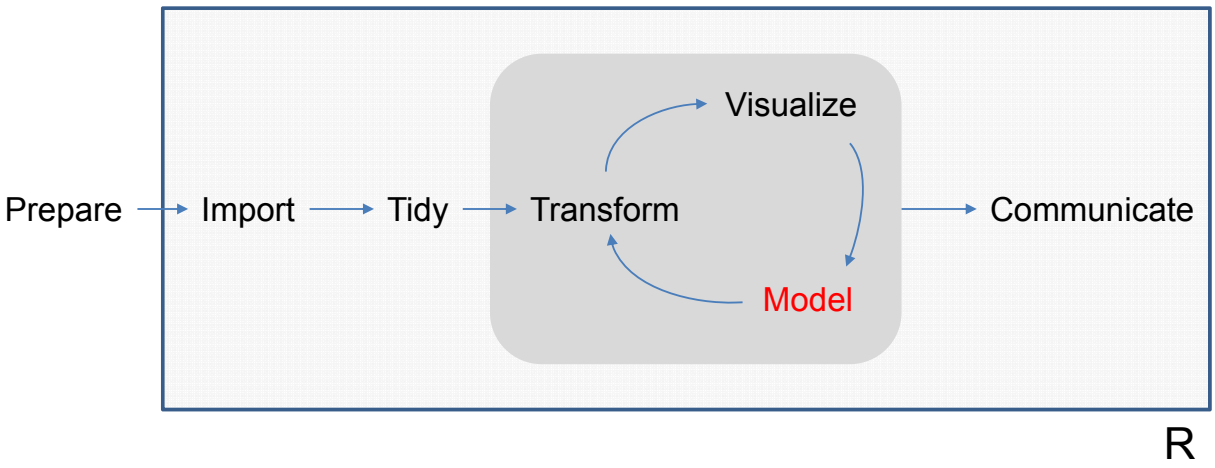
Can we guess what the scatterplots look like ?



The Anscombe dataset



Data analysis workflow



Adapted from Hadley Wickham

Here:

a model provides a summary/explanation of a dataset

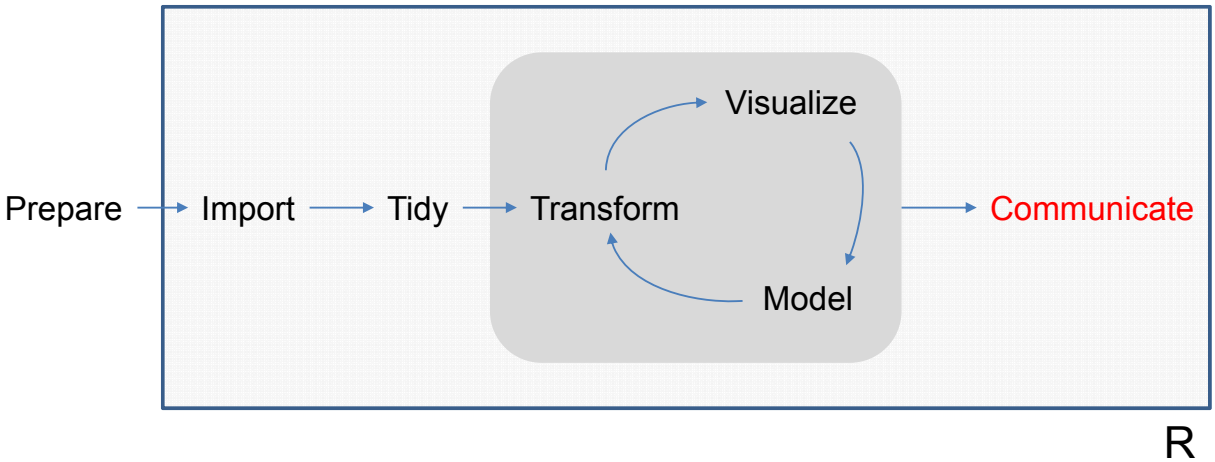
Explains true signal vs noise

One rule:

Keep it simple

(although this is not universally agreed)

Data analysis workflow



Adapted from Hadley Wickham

Flipped classroom ?

How we are going to work

- Work in groups of 5 (5 groups in total)
- Get a project to work on (it may or may not be the same for all groups)
- Work on the analysis within the group, and prepare an ***informal*** presentation of the results
- Don't discuss only the results, but also the methods and the issues you've encountered.
- Ask for feedback if you are stuck or if you hesitate

- Groups will then present and we will discuss the topics afterwards.

9:00	Welcome and introduction
10:00	Coffee break
10:15	Work in group (qPCR data)
12:00	Lunch
13:00	Presentation: knitr
13:30	Work in group
14:30	Coffee break
14:45	Presentation and discussion
15:30	Work in group (dataset 2)

First dataset: qPCR dataset

The qPCR dataset

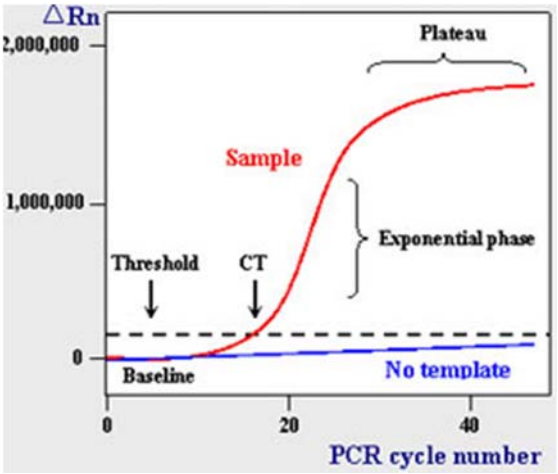
- Measure of gene expression in mice
- Two variables of interest:
  - Genotype: WT/KO mice for the MAF1 gene
  - Treatment: either normal food (control) or fasting (treated)
- We are interested in the expression of gene AKT
- 3 biological replicates for each group
- 3 technical replicates for each biological replicate

Our questions

- In MAF1 WT mice, is there a difference in AKT expression depending on the treatment ?
- Does the effect of the treatment depend on the MAF genotype (WT/KO) ?

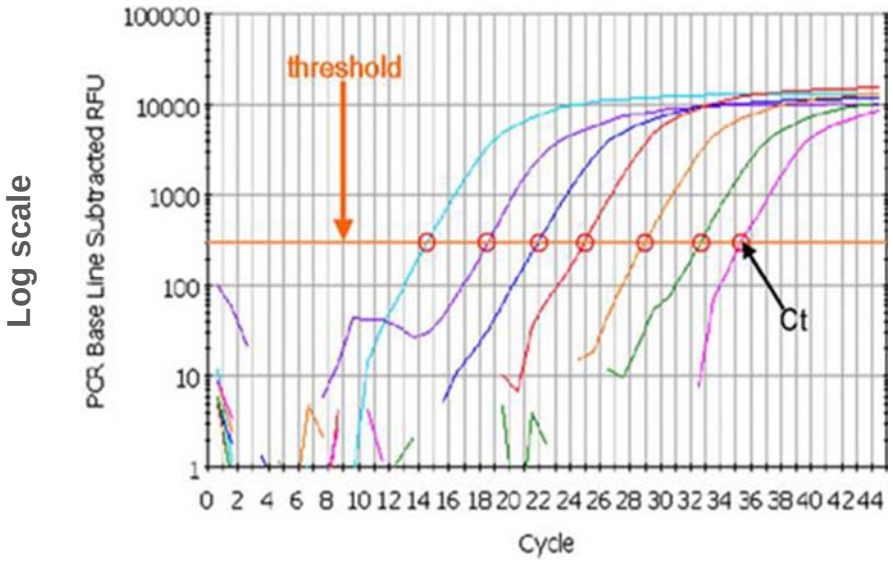
- Extension of the PCR technology
- Use fluorescence to measure the amount of expression at every cycle.
- Measure the number of cycles (Ct, “cycle threshold”) required until the fluorescence crosses a given threshold.
- This measure is made during the exponential phase.
- The threshold is arbitrary

- **Gene more expressed = lower Ct value**



http://www.rt-pcr.com/

A real example



Plateau zone

Background zone

http://bio-analyze.com/OurRealtimePrject1.htm

What the data looks like…

Well	Sample	Gene	Ct
A1	WT-Treated	A	35
A2	WT-Treated	A	36.82
A3	WT-Treated	A	34.34
A4	WT-Untreated	A	34.89
A5	WT-Untreated	A	35.29
A6	WT-Untreated	A	35.65
A7	KO-Treated	A	34.22
A8	KO-Treated	A	31.94
A9	KO-Treated	A	35.24
A10	KO-Untreated	A	29.57
A11	KO-Untreated	A	35.79
A12	KO-Untreated	A	33.77
…			

Well	Sample	Gene	Ct
B1	WT-Treated	B	29.69
B2	WT-Treated	B	29.14
B3	WT-Treated	B	26.6
B4	WT-Untreated	B	27.4
B5	WT-Untreated	B	32.44
B6	WT-Untreated	B	27.1
B7	KO-Treated	B	26.03
B8	KO-Treated	B	23.43
B9	KO-Treated	B	27
B10	KO-Untreated	B	26.2
B11	KO-Untreated	B	28.21
B12	KO-Untreated	B	24.19
…			

Normalization

- The measured expression must be normalized for difference between samples (e.g. amount of starting material).
- This is usually done using a *reference gene* (or *standard gene*, *housekeeping gene*), which should be expressed in all cells and have the same number of copies in all cells
- Examples of typical housekeeping genes:
  - Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)
  - Beta actin
- The assumptions are very stringent and not always satisfied…
- Common recommendations: take several (at least 3) housekeeping genes.

http://en.wikipedia.org/wiki/Housekeeping_gene

How to measure differential expression: the ΔΔCt method

The efficiency is assumed to be perfect (100% = exact doubling at every cycle), and the efficiencies for the gene of interest and the reference should be similar.

Ct_g = Ct for gene of interest

Ct_h = Ct for housekeeping gene

$$\Delta Ct_g = Ct_g - Ct_h$$

Difference between conditions 1 and 2:

$$\Delta\Delta Ct_g = \Delta Ct_{g2} - \Delta Ct_{g1}$$

Log fold change = $-\Delta\Delta Ct_g = \Delta Ct_{g1} - \Delta Ct_{g2}$

Fold change = $2^{-\Delta\Delta Ctg}$

(The base is 2 because a difference of one cycle represents a doubling of the amount of the material. In comparison to other assays, the base of the logarithm is not arbitrary here)

A typical setting

Condition 1

Biological replicate	Technical replicate	Gene	Ct
1	1	A	35
1	2	A	36.82
1	3	A	34.34
2	1	A	34.89
2	2	A	35.29
2	3	A	35.65
3	1	A	34.22
3	2	A	31.94
3	3	A	35.24
1	1	HKG	29.57
1	2	HKG	35.79
1	3	HKG	33.77
2	1	HKG	29.57
2	2	HKG	35.79
2	3	HKG	33.77
…			

Condition 2

Biological replicate	Technical replicate	Gene	Ct
1	1	A	29.69
1	2	A	29.14
1	3	A	26.6
2	1	A	27.4
2	2	A	32.44
2	3	A	27.1
3	1	A	26.03
3	2	A	23.43
3	3	A	27
1	1	HKG	26.2
1	2	HKG	28.21
1	3	HKG	24.19
2	1	HKG	29.69
2	2	HKG	29.14
2	3	HKG	26.6
…			