Data Science and some (classical) challenges

Søren Højsgaard
Department of Mathematical Sciences
Aalborg University

Prelude

- Data science and statistics – somewhat synonymous ??
- Getting large datasets from different sources organised the right way is not something statisticians are typically good at.
- Some of the difficulties / pitfalls that has ”haunted” statistics for 100 years still exist in modern data science
- Many of these topics are not taught to (stats) students anymore; probably not to data scientists either…

- Just want to mention a few…

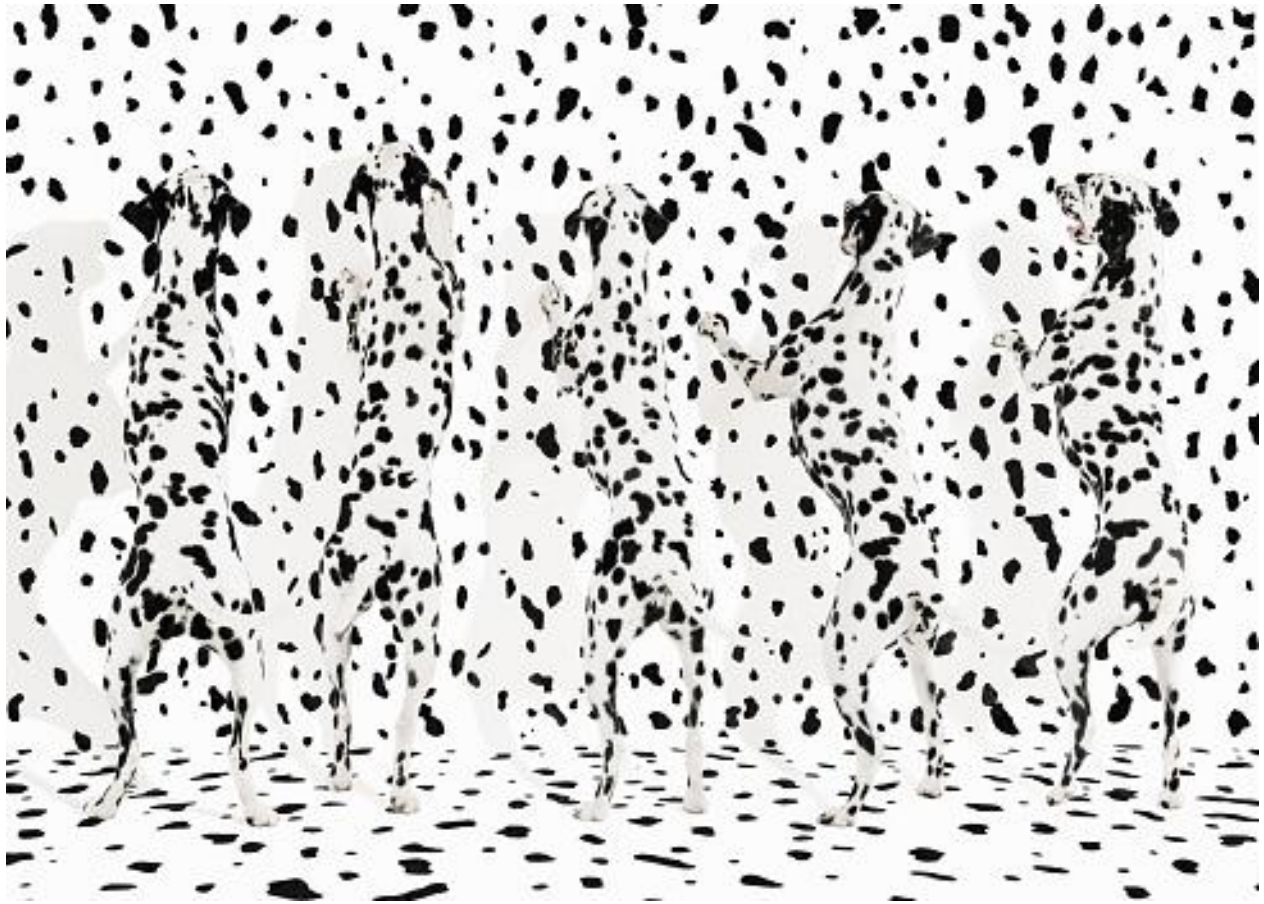
bigData versus smallData

- Q: What can we do with bigData that we can not do with smallData?
- A: Huh – well – We can explore; seek inspiration for further analyses; look for patterns in data…

”You should always enjoy your empirical findings – because you will (probably) never see them again…”

Finding patterns in data…

More patterns…



A classic: treatment and control

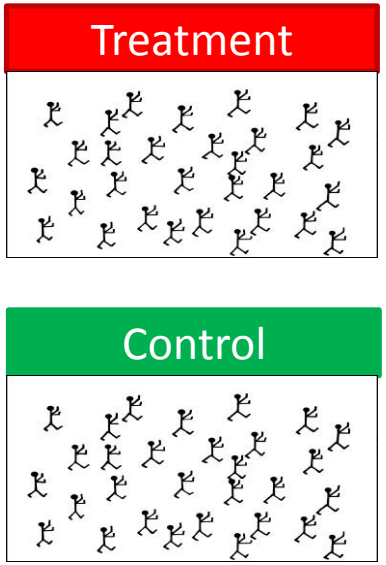
- Test medicine against high blood pressure

A classic: treatment and control

- Hypotheses:
  - Scientific hypothesis: There is an effect of treatment.
  - Statistical hypothesis:
    - Null hypothesis: H0: ”There is no effect of treatment”.
    - Alternativ hypothesis: HA: ”There is an effect of treatment”
- To investigate hypothesis we conduct a study:
  - Take 100 randomly selected high-blood-presure-patients from a population.
  - Allocate 50 to treatment and 50 to control (placebo).
- Notice: First we ask questions, then we collect data to answer them…

How to proceed

- Simple approach:
  - Measure blood pressure after 4 weeks (or measure before start and after 4 weeks and calculate difference within each patient).
  - Compare group means of treatment and control groups with t-test or similar.



- There are many alternatives - not to be discussed

Traditional thinking in statistics

- We have a theory about the world (a map).
- We have data (landscape).
- We don’t use data for ”proving” that we are right; we use data for ”proving” that we are wrong…

”If the map does not match the landscape then the map is wrong…”

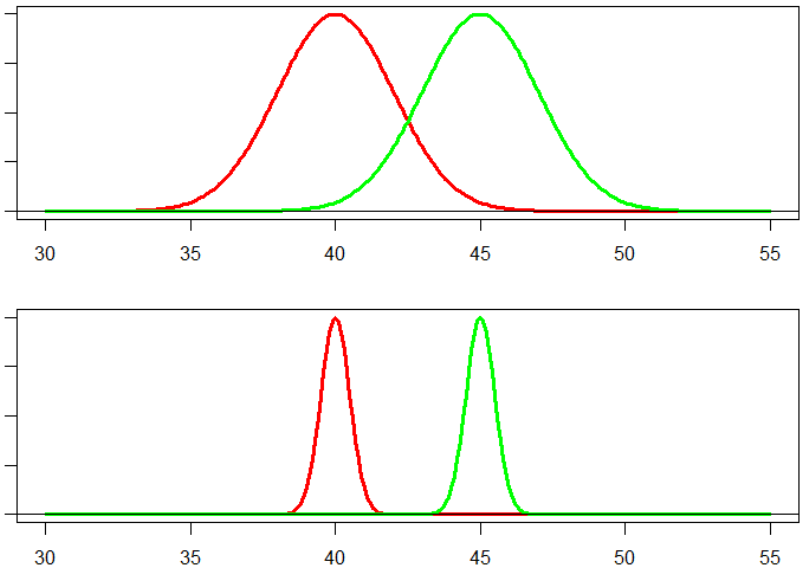
Traditional thinking…

- To reject H0 (”No effect of treatment”) is the strong conclusion of an experiment.
- If we see a significant difference it is (probably) because it is for real.
- Not rejecting H0 is not the same as accepting H0 as ”the truth”.

”Absence of evidence of an effect is not the same as evidence of an absence of an effect…”

Sample size…

- We could well end up not rejection H0 because of a small sample size
- With bigData we are more likely to discover a treatment effect – if it is there,
- bigData reduces uncertainty…
- Thumbs up!!!



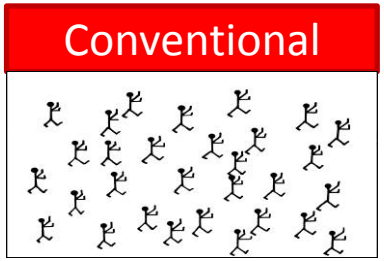
Allocation mechanishm is important

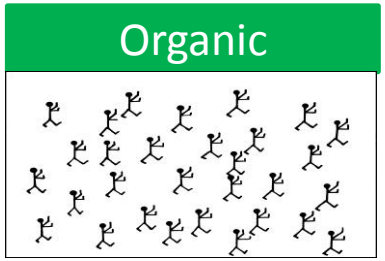
- What if – accidently – most control patients were old and most treatment patients were young?
- A treatment effect would be confounded with age: Can not say whether difference is due to treatment or age.
- No amount of bigData helps us here - thumbs down!!!

- Need to collect data the right way – and understand how it is collected.
- Avoid this e.g. by randomization:
  - Allocation is independent of characteristics of patients.
  - 50-50 chance of patient going to treatment or control group.

Example: Is organic food healthier than conventional food ?

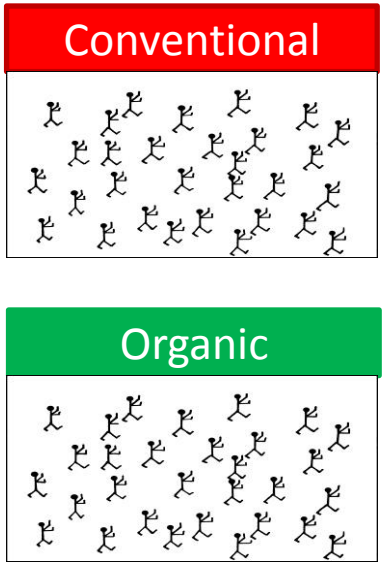
- Take two fields; grow organic carrots on one field, conventional carrots on the other.
- Feed carrots to pigs/rats
- Register e.g. reproductive ability.
- Will bigData help us detect a difference?
- And - what is bigData here?
- More pigs/rats?





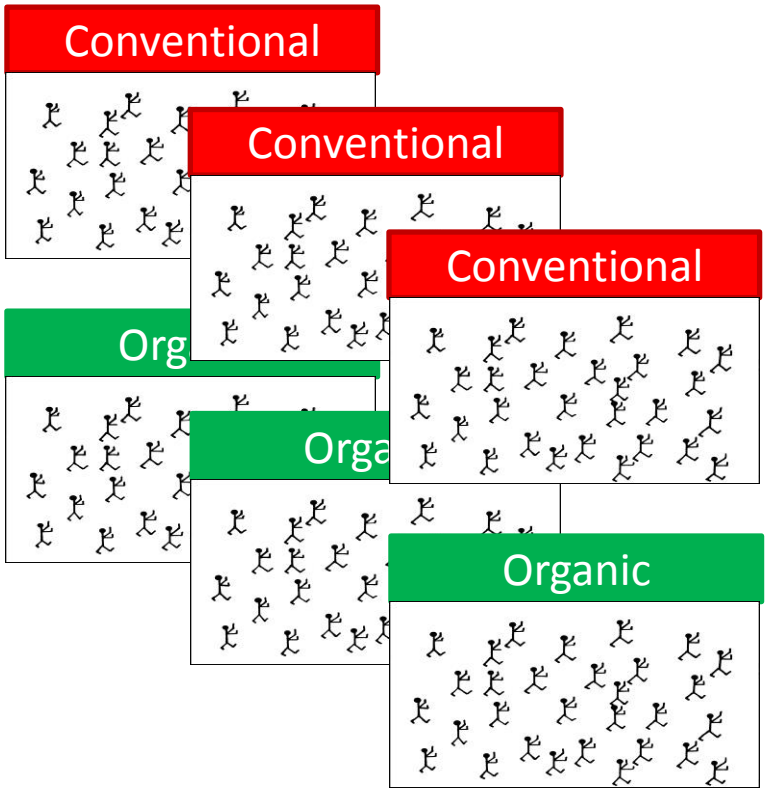
The point is…

- Notice: This is a plant experiment – NOT an animal experiment.
- The relevant source of variation is the field-to-field variation; not the animal-to-animal variation
- Since we have one field only per treatment we have no replications
- Increasing number of pigs/rats per field will give a larger dataset (bigData)
- But no information about relevant source of variation.



A remedy…

- The field-to-field variation is the relevant one.
- So – need more fields per treatment
  - Very common pitfall
  - Need to understand sources of variation
  - Data has to be ”big” in the right way.



Somewhat scary experience from previous job

- Denmark has long tradition for systematic collection of data in dairy industry.
- Big data is big issue; automatic collection of loads of online data on individual cows.
- Dairy industry gets lots out of their databases – but they could get more…

- Cattle people [CP]: We have this large database and would like to use it better:
- Data analyst / statistician [DA]: Excellent idea.
- [CP]: We think that data science / machine learning / pattern recognition / … can help us.
- [DA]: Probably; what specifically are you trying to obtain?
- [CP]: Well, we want to improve reproduction and reduce lameness and mastitis problems

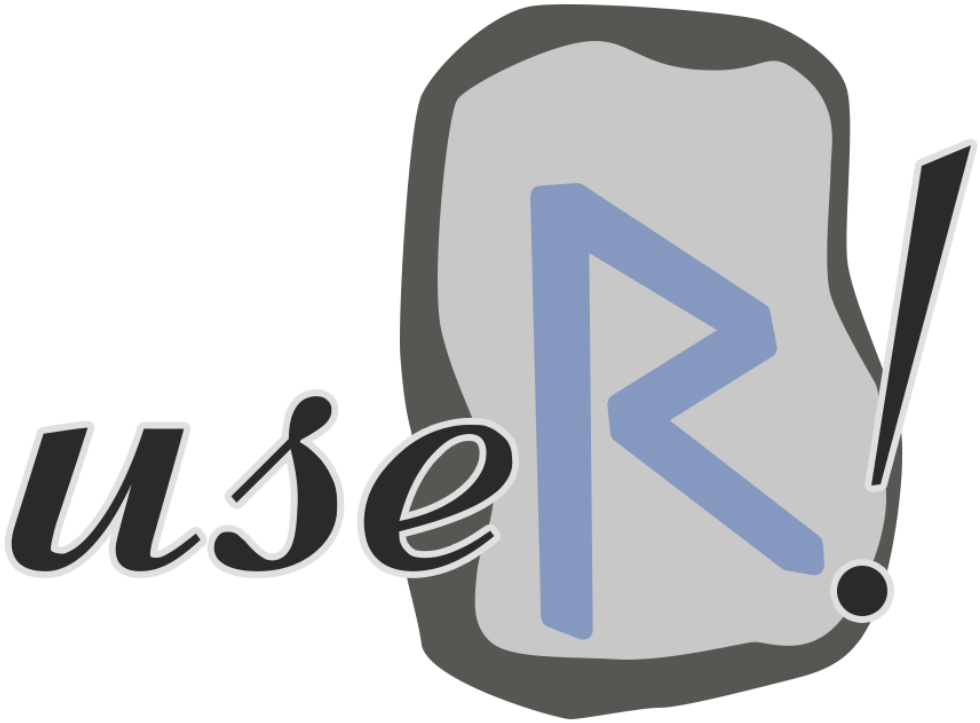
- [DA]: Interesting; when you say ”improve reproduction” what do you have in mind.
- [CP]: What do you mean?
- [DA]: We have to be more specific about where we want to go…
- [CP]: But we thought that modern data science could help us here…
- [DA]: Huh – yes and no…

Moore's Law of Big Data:

"The Amount of Nonsense Packed Into the Term "BIG DATA" Doubles Approximately Every Two Years."

-Mike Pluta, 2014-08-10

Speaking of data science




```
> sessionInfo()
[1] "June 30 - July 3, 2015"
[2] "Aalborg, Denmark"
```

Speaking of …

- R (www.r-project.org) is the statistics system language of choice for most statisticians
- Aalborg University hosts useR!2015 http://user2015.math.aau.dk/
- 700 statisticians / data scientists meet in Aalborg on June 30. – July 3.
- There is room for more – only hotels are limited..

We see patterns everywhere…