fhs Fachhochschule Salzburg University of Applied Sciences

Big Data

A Lot of Opportunities

For Producing Wrong Results

Data Science Symposium 2014

MMag. Dr. Günther Eibl

Influence of „Big“ on Analysis Mistakes

- With Big Data
  - → Memory and computational costs get important
  - → Kind of mistakes are mostly known
  - → Easier to fall into a trap
  - → Mistakes may have greater effects
  - → Study typical (known) analysis effects

Outline: The Data Analysis Process

- Pre-analysis
  - Identify goals and constraints
  - Obtain data and its background
  - Treat data issues
- Analysis
  - Descriptive analysis
  - Modeling
- Reporting

Data Collection: General Issues

- Two main types of studies
  - Prospective study
    - Experiments with purposeful design
  - Retrospective study
    - Data are easy to get or already there
- Big Data similar to retrospective study
  - Data come from sensors or tracking devices, Web pages, Facebook accounts,..
- Drawbacks of a retrospective study
  - Maybe not representative → selection bias (missing data)
  - Controls are typically unavailable
  - Data scientists not involved in data collection → interpretation issues

→ Validity of results reduced

Data Collection: Big-Data-Specific

- Tools that are based on big data can be easily gamed
  - Wrong entries in Facebook
  - Google bombing (spamdexing)
- Robustness and repeatability of results
  - Google: changes in data collection due to live system
- Echo-chamber effect:
  - When data source is itself a product of big data → opportunities for vicious cycles
  - Example
    - Google translate compares parallel text from different languages
    - What, if one of the texts (e.g. a Wikipedia text in a rare language) already stems from Google translate?

Data Collection: Miss Important Inputs

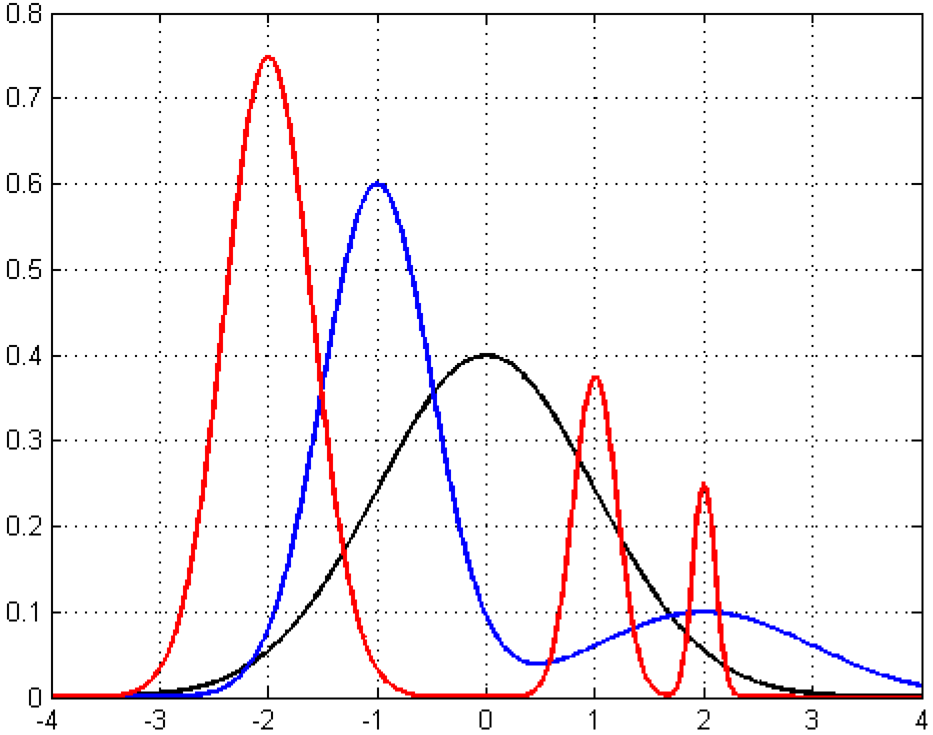
- Miss to collect important input variable (causes)
  - → Suboptimal prediction
  - → Wrong results, if not analyzed properly
- Remedy: Obtain background/domain knowledge
  - Example traffic flow

Descriptive Analysis

- Purpose
  - Get a „feeling“ for the data
  - Know the data domain: where are your data
  - Identify outliers (Boxplots)
  - See the distribution of values
  - ….
- Check domain knowledge
- Too few descriptive analyses
  - Time constraints
  - Value of the descriptive analysis underestimated
  - Too much trust in automatic analysis tools
  - „The data are the model“
- Result: Wrong assumptions → Wrong results

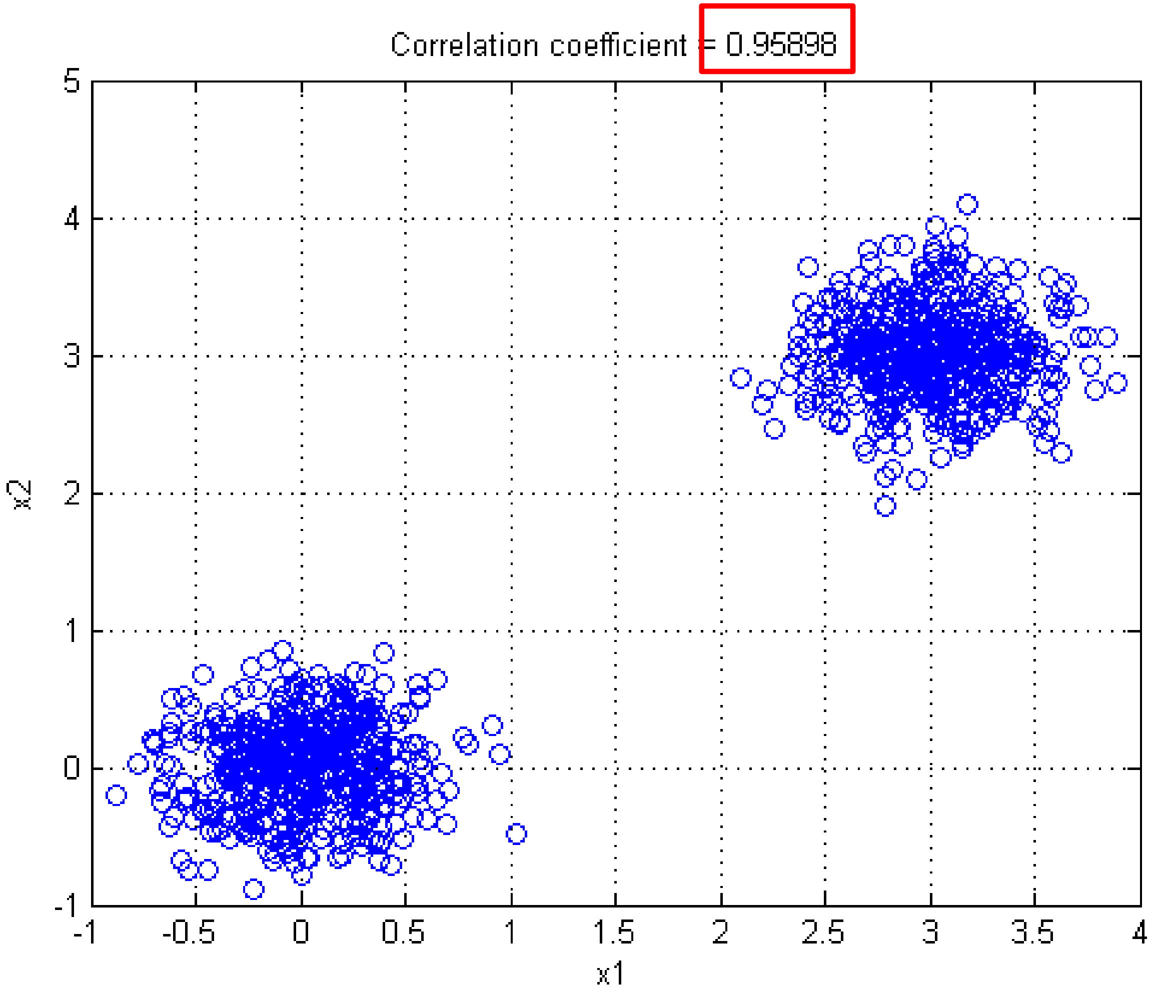
Automatic listing of means (and standard deviations)

- Still present mistake: only list mean values
- Example: mean = 0



Automatic listing of correlation coefficients

- „Common cause“-like effect

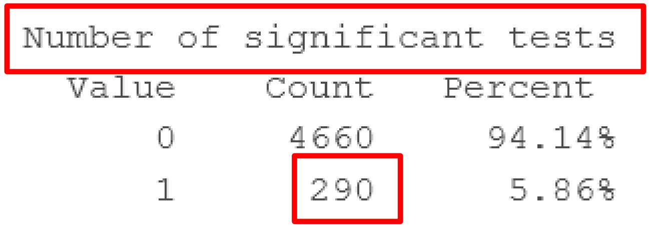


Hypothesis Testing: Multiple Test Problem

- Multiple comparison problem example
  - Does a vitamin have a beneficial influence?
  - On what?
  - On anything available
    - On weight, prevention of diseases, scores of IQ-tests, earnings,…
- Even worse: Compare everything with everything
- No clear goals identified → Research question: can we find anything?
- Answer for Big Data: yes, plenty!
- Great for the report, but useless for the product

Multiple Test Problem Example

- Sample size N = 10000
- 100 statistically independent normal distributions
- Compare all with all for differences in means → 4950 t-tests



```
Number of significant tests
    Value       Count     Percent
        0        4660      94.14%
        1         290       5.86%
```

Multiple Test Problem Example: Remedy

- Does a higher sample size help? No
- Recap Pre-analysis
  - Identify **goals**
  - Obtain **background** knowledge
- Formulate research questions
  - Determine the outcome
  - Determine possible influences
  - Compare outcome with each of the possible influences
  - Adjust family-wise alpha-error rate
    - Simplest method: Bonferroni-correction (conservative)
    - For each t-test use α = 0.05/“number of tests“

      $= 0.05/4950$

      $\approx 10^{-5}$
- No significant test remains in our test example

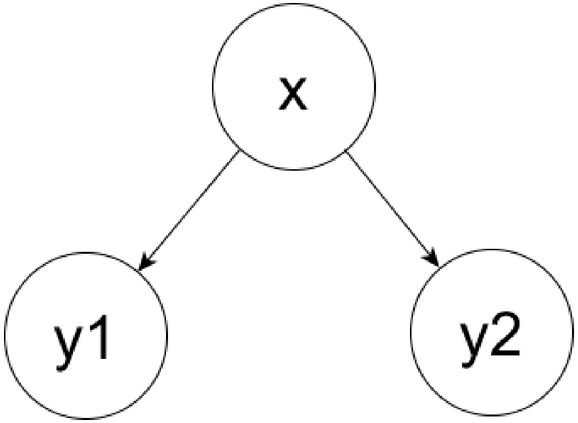
Correlation

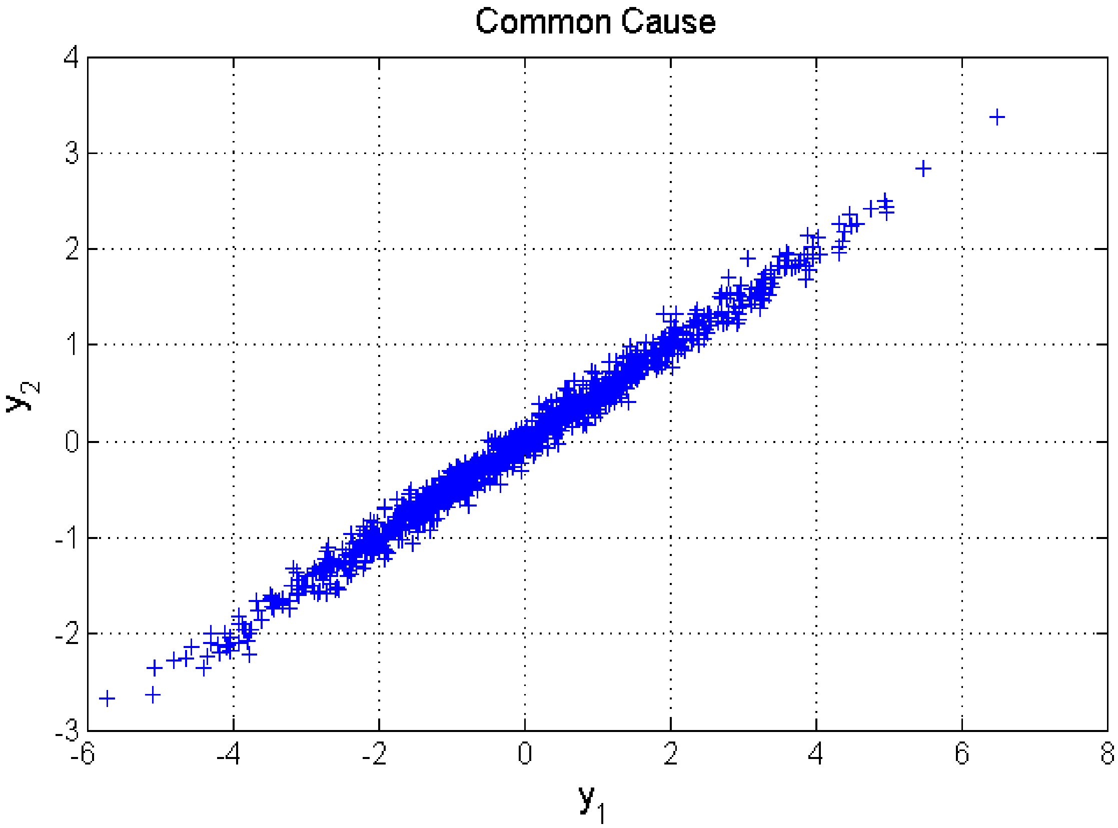
- Correlation vs. causality
  - From 2006 to 2011 the United States murder rate correlated well with the market share of Internet Explorer
    - Both went down sharply.
  - From 1998 to 2007 the number of new cases of autism diagnosed was extremely well correlated with sales of organic food
    - Both went up sharply

Causality

- Needed for research questions such as
  - **Why** did it happen?
  - What is the best that can happen?
- Correlation is not enough
- And correlation ≠ causality!

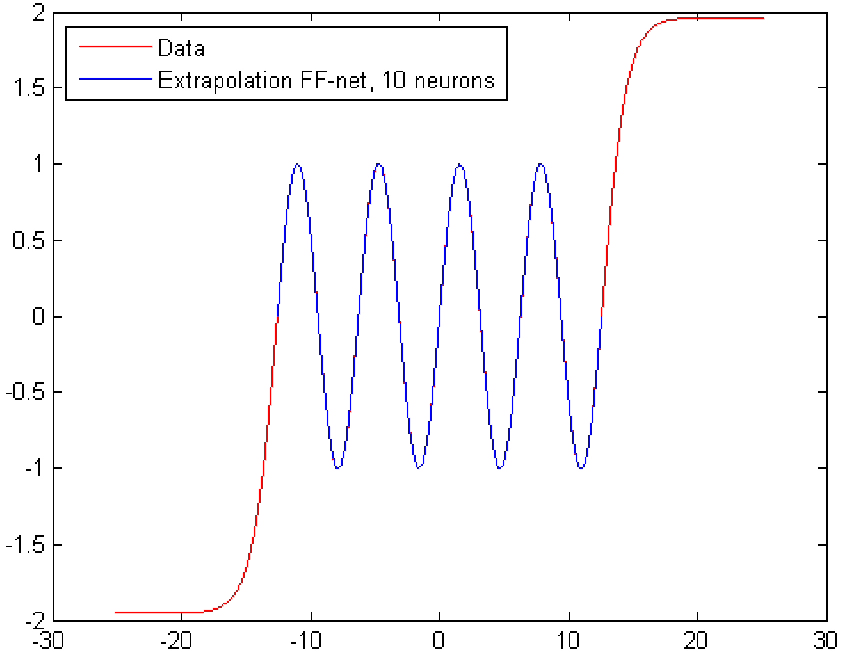
Correlation ≠ Causality: Common Cause





High-Dimensionality of Input Space

- Well-known: Curse of dimensionality → problems with fitting the model
- Extrapolation: what happens, if you evaluate your model outside your data domain?



- Remedy: avoid extrapolation, know your data!
- Description of a high-dimensional data domain more difficult

Reporting: Interpretation

- Who is more important?
- Who has the higher impact?
- The importance of a variable is hard to assess.
  - Maybe in the context of a linear model: biggest coefficient
  - Importance for a special group of people only
  - Interactions effect only in combination with other inputs
- Performance indexes
  - Do they measure what they are supposed to measure?

Summary

- Big Data mostly does not create new dangers
- But well-known mistakes can have more effect
- Treated topics
  - Selection bias
  - Miss important variables
  - Too little descriptive analysis
  - Multiple testing problem
  - Confusing correlations and causality
  - Curse of dimensionality
  - Extrapolation
  - Performance indexes
- Many things wait for being discovered
- Big Data can be a big help
- Hopefully, new results are real

Enjoy your results…..

…you will never see them again