When Apples cause Heart Problems – Causal Dependencies in Data Narrations

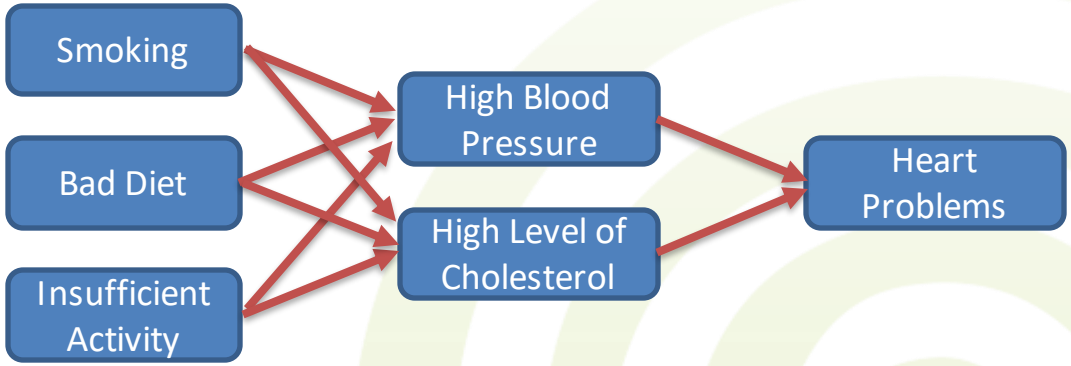
Denis Nagel

Braunschweig, 08.06.2021

Technische Universität Braunschweig

Narratives and Data Sets

- Narrative
  - Events and Entities embedded into a coherent argumentation structure
    - e.g. description of how risk factors can lead to a disease

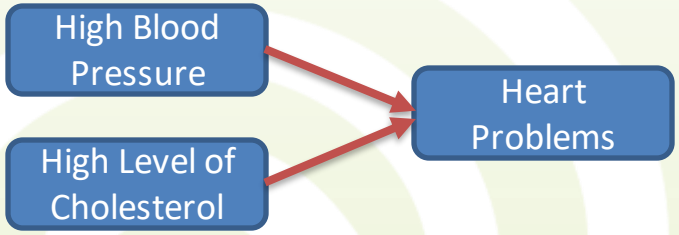


  - Explicitly expressed in newspaper articles, clinical studies, research papers and everyday discourse

Narratives and Data Sets

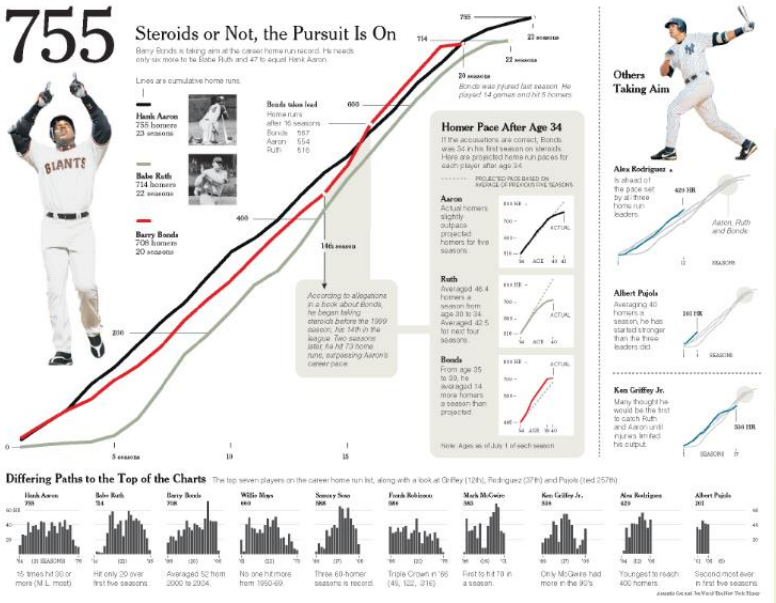
- Data Narration
  - A Narrative (or part of a narrative) that is implicitly expressed by the knowledge in a data set

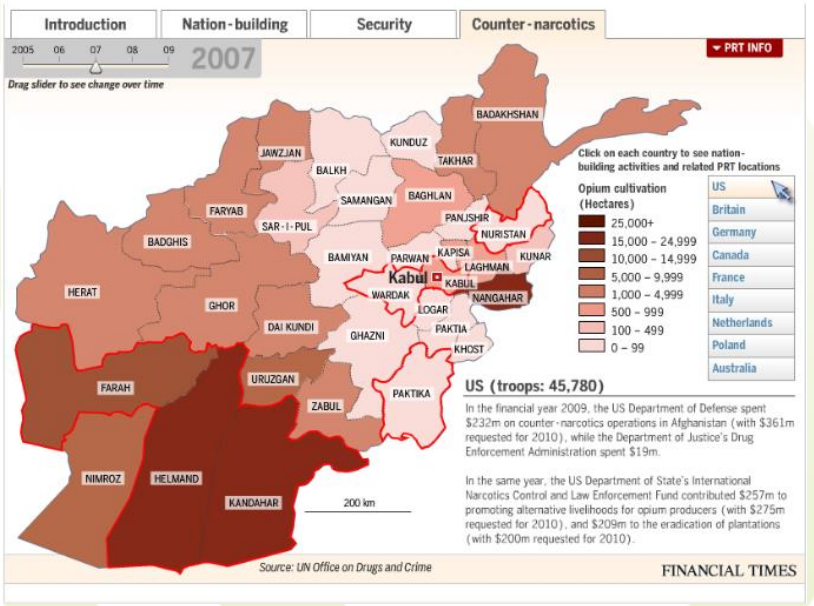
Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
67	120	229	1
37	130	250	0
41	130	204	0
56	120	236	0
62	140	268	3
57	120	354	0
63	130	254	2
53	140	203	1



Data Visualization

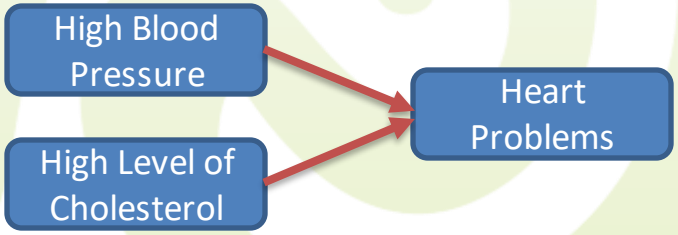
- Data Visualization used to present Data Narrations
  - Goal: Make Data understandable for Humans
  - Knowledge about intrinsic relations necessary





Narratives and Data Sets

- Finding Data Narrations for arbitrary Data Sets
  - Without careful manual data analysis any relation between entries in the data set might be possible
- Idea:
  - Use existing narratives as templates to provide sensible narratives
  - Reduces the problem of finding data narrations to the computation of narrative bindings



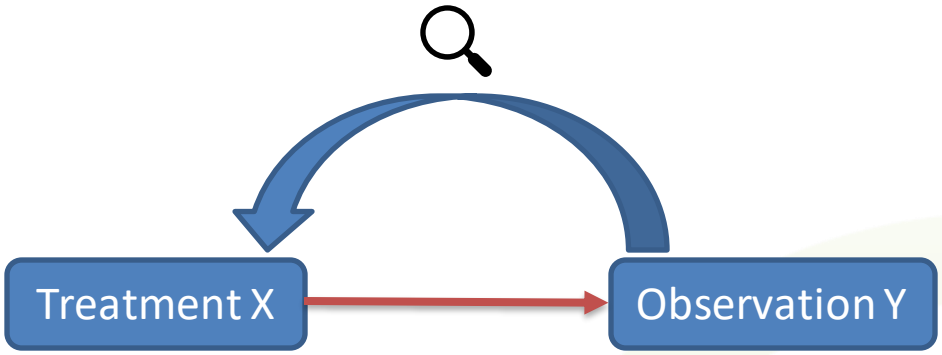
Cause and Effect

- Cause-Effect Relations

  - Y can be observed because X occurred
    - Does this mean if X is prevented Y can never happen?

Cause and Effect

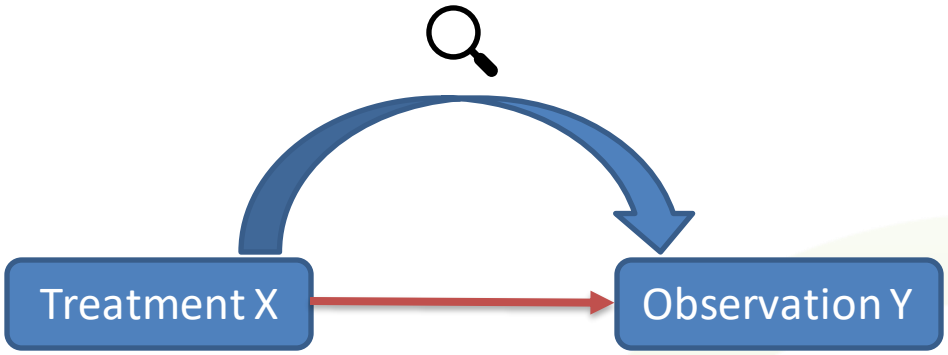
- Necessary Causes

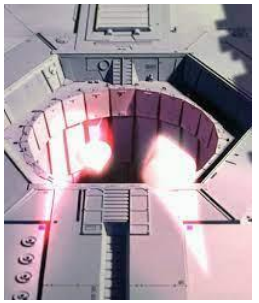


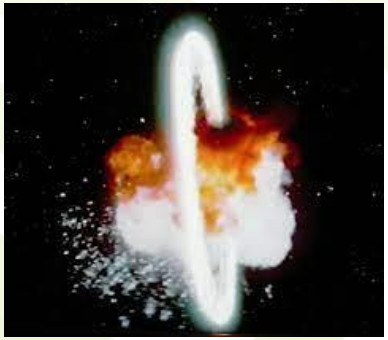
  - Y can ONLY be observed when X occurs
    - If X is prevented Y can be avoided

Cause and Effect

- Sufficient Causes



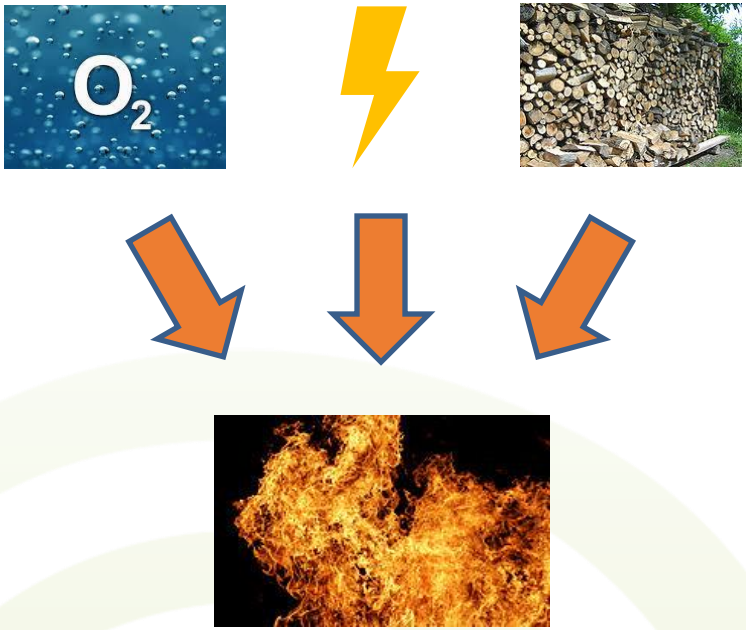


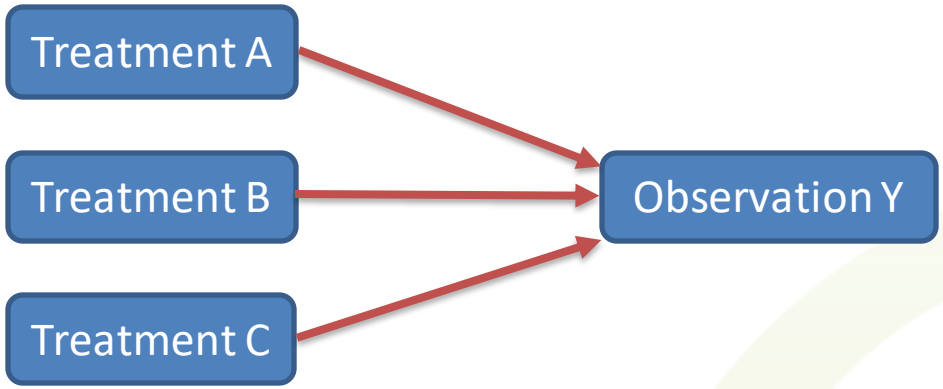


  - If X occurs then Y has to occur
    - If X is prevented Y can still occur

Cause and Effect

- Contributory Causes





  - A, B and C all contribute to the occurrence of Y
    - Each treatment might be necessary
    - Usually a single treatment is not sufficient

Causal Relations in Data Sets

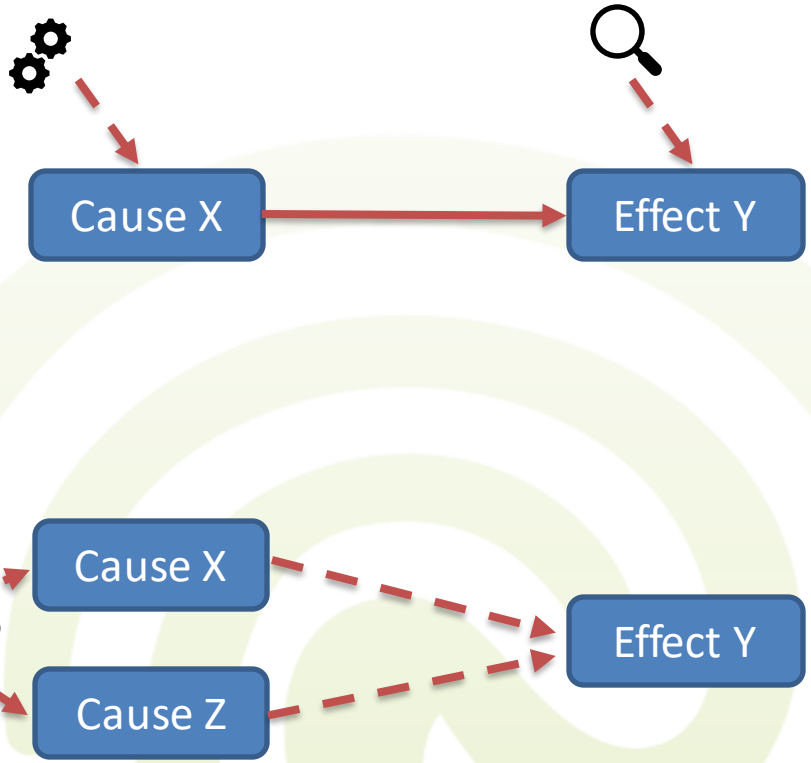
- Which factors cause Heart Disease?
  - Let us look into some data
  - Focus on observational data

Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
67	120	229	1
37	130	250	0
41	130	204	0
56	120	236	0
62	140	268	3
57	120	354	0
63	130	254	2
53	140	203	1

- Src.: V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Assessing Causality

- Differentiation between two tasks
  - Causal Inference
    - How much is Y affected by a change of X?
    - Causal Effect
  - Causal Discovery
    - Changing which variable actually affects Y?
    - Causal Relation



Causal Relations in Data Sets

- Which factors cause Heart Disease?

Possible Necessary Causes:
- Age ≥ 53
- Blood Pressure ≥ 120 mm Hg
- Cholesterol ≥ 203 mg/dL

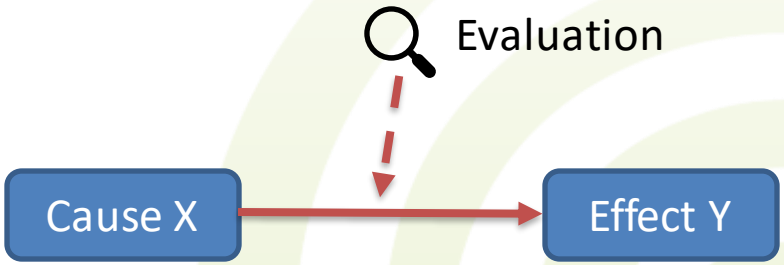
Possible Sufficient Causes:
- Age ≥ 67
- Blood Pressure ≥ 160 mm Hg
- Cholesterol ?

Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
67	120	229	1
37	130	250	0
41	130	204	0
56	120	236	0
62	140	268	3
57	120	354	0
63	130	254	2
53	140	203	1

- Is a high Cholesterol level even causal to Heart Disease?
  - According to the WHO it is

Causal Relations in Data Sets

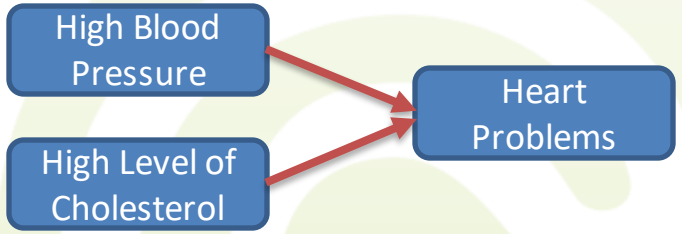
- Problems with Identifying Causality in Data Sets
  - Incomplete Sample: Are there exceptions we missed?
  - Missing Factors: Is there a hidden cause we do not know about?
    - According to the WHO smoking is causally related to CVD
  - The Temporal Component: Is there a delay between cause and effect?



  - Mono- vs. Multicausality: Do we have to look at a combination of factors? Which combinations are feasible?

Causal Relations in Data Sets

- By looking at the Data alone it is not possible to identify causal dependencies
  - Domain knowledge about the processes at the core of a causality is required
  - Why does a cause lead to an effect?
- Causal Inference/Discovery rely on causal models as ground truth
  - Again: External knowledge about existing causal relations is necessary



- But what about plausibilizing assumed causalities?

Testing Plausibility

- Trivial: Any Causality should have a correlation at its core

  Pearson Correlation Coefficient:
  - Age, Heart Disease → 0.22
  - Blood Pressure, Heart Disease → 0.17
  - Cholesterol, Heart Disease → 0.07

  – Only weak indication
    - Correlation ≠ Causality
    - Does not indicate a direct dependency

Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
(…)	(…)	(…)	(…)

Testing Plausibility

- Idea: Use a control group to measure the impact a factor has on an event
  - Applied regularly in clinical studies
- Metric: Relative Risk

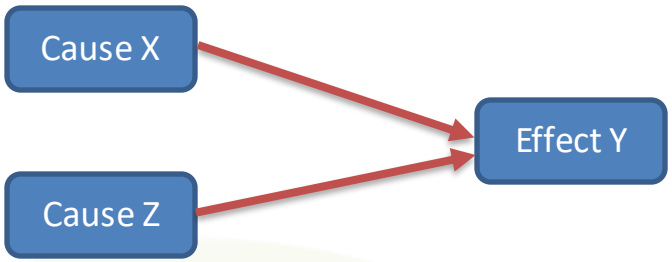
$$relRisk(C \rightarrow E) = \frac{P(E|C)}{P(E|\neg C)}$$

Relative Risk on Heart Disease:
- Age ≥ 60 → 1.44
- Blood Pressure ≥ 140 → 1.33
- Cholesterol ≥ 240 → 1.28

Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
(…)	(…)	(…)	(…)

Testing Plausibility

- Most Data Sets describe a multicausal environment
  - If the narrative describes contributory causes, further assumptions are possible



- Assumptions:
  - Removing X from the Experimental Group should negatively impact Y
    - relRisk((¬X ∧ Z) → Y) << 1.0
  - Cases in which Z occurs without X should have a lower impact than all cases in which Z occurs
    - relRisk((¬X ∧ Z) → Y) < relRisk(Z → Y)
  - Cases in which X and Z occur together should have a higher impact on Y than all cases in which Z occurs
    - relRisk((Z ∧ X) → Y) > relRisk(Z → Y)
  - If Z is associated with X, it should affect the probability of X
    - relRisk(Z → X) >> 1.0

Testing Plausibility

- Use Case: 2 Data Sets
  - Patient Data regarding Heart Disease from the V.A. Medical Center
  - Data on Forest Fires in Portugal

Age	Resting Blood Pressure (mm Hg)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
(...)	(...)	(...)	(...)

Month	Humidity (%)	Temperature (°C)	Area (ha)
mar	51	8.2	0
oct	33	18	2
(...)	(...)	(...)	(...)

Assumption:
- Raised Blood Pressure (≥ 140) and High Cholesterol Levels (≥ 240) are causal to the occurence of heart disease (≥ 1)
- High Age (≥ 60) is a confounder for both causes

Assumption:
- Low Humidity (≤ 30) and High Temperature (≥ 20) are causal to large forest fires (≥ 25)
- Summer Months are a confounder for both causes

Testing Plausibility

- Correlation and Relative Risk

PCC with Heart Disease:
- Age → 0.22
- Blood Pressure → 0.17
- Cholesterol → 0.07

Relative Risk on Heart Disease:
- Age → 1.44
- Blood Pressure → 1.33
- Cholesterol → 1.28

PCC with Burnt Area:
- Summer → 0.05
- Humidity → -0.08
- Temperature → 0.10

Relative Risk on Burnt Area:
- Summer → 1.32
- Humidity → 1.5
- Temperature → 1.55

Testing Plausibility

- Removing X from the Experimental Group should negatively impact Y
  - relRisk((¬X ∧ Z) → Y) << 1.0
    - X = Blood Pressure, Z = Cholesterol → 0.003
    - X = Humidity, Z = Temperature → 0.002
- Cases in which Z occurs without X should have a lower impact than all cases in which Z occurs
  - relRisk((¬X ∧ Z) → Y) < relRisk(Z → Y)
    - X = Blood Pressure, Z = Cholesterol → 0.003 vs. 1.28
    - X = Humidity, Z = Temperature → 0.002 vs. 1.55

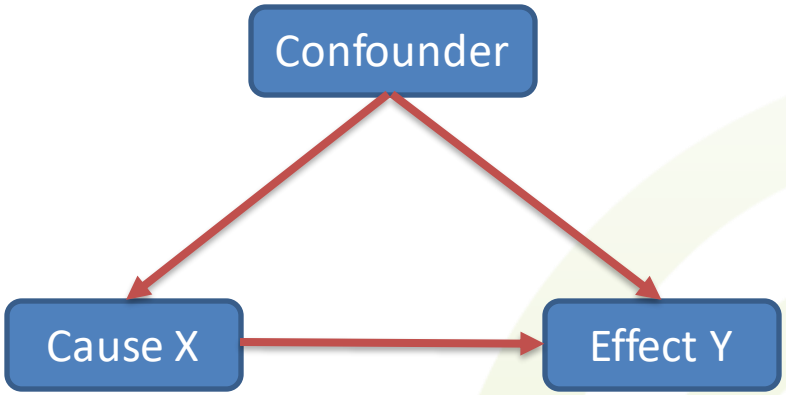
Testing Plausibility

- Cases in which X and Z occur should have a higher impact on Y than all cases in which Z occurs
  - relRisk((Z ∧ X) → Y) > relRisk(Z → Y)
    - X = Blood Pressure, Z = Cholesterol → 0.004 vs. 1.28
    - X = Humidity, Z = Temperature → 0.003 vs. 1.55

- If Z is associated with X, it should affect the probability of X
  - relRisk(Z → X) >> 1.0
    - X = Blood Pressure, Z = Cholesterol → 1.11
    - X = Humidity, Z = Temperature → 2.25

Testing Plausibility

- Problem: Confounders
  - Factors that affect both X and Y



  - Confounder 'shares' experimental group with X

How to Detect Confounders?

- Removing X from the Experimental Group should negatively impact Y
  - relRisk((¬X ∧ Z) → Y) << 1.0
    - Blood Pressure → 0.004
    - Cholesterol → 0.004
    - Humidity → 0.002
    - Temperature → 0.002

- Cases in which Z occurs without X should have a lower impact than all cases in which Z occurs
  - relRisk((¬X ∧ Z) → Y) < relRisk(Z → Y)
    - Blood Pressure → 0.004 vs. 1.44
    - Cholesterol → 0.004 vs. 1.44
    - Humidity → 0.002 vs. 1.32
    - Temperature → 0.002 vs. 1.32

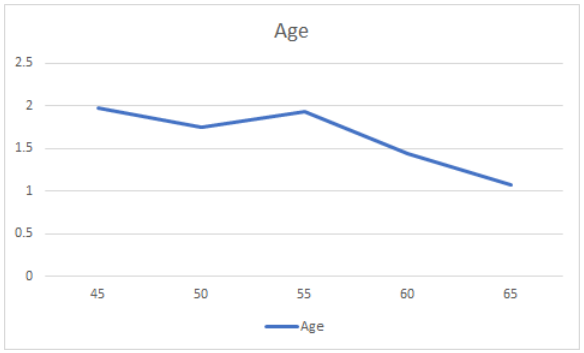
How to Detect Confounders?

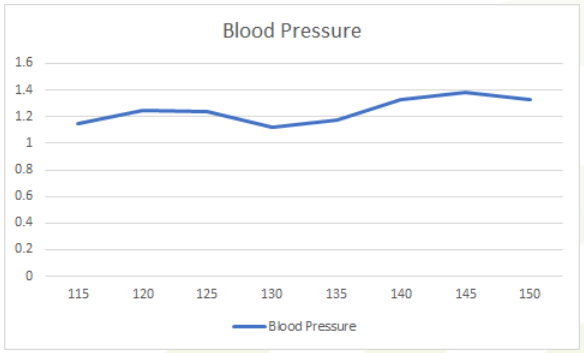
- Cases in which X and Z occur should have a higher impact on Y than all cases in which Z occurs
  - relRisk((Z ∧ X) → Y) > relRisk(Z → Y)
    - Blood Pressure → 0.004 vs. 1.44
    - Cholesterol → 0.004 vs. 1.44
    - Humidity → 0.003 vs. 1.32
    - Temperature → 0.002 vs. 1.32

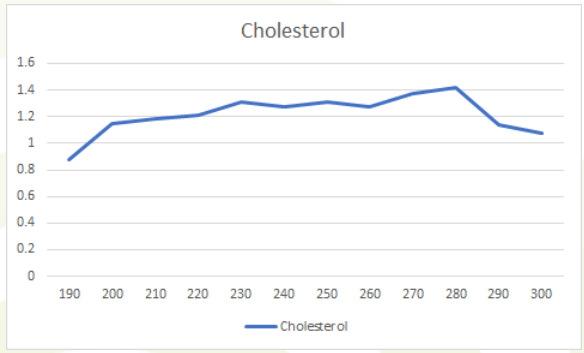
- As Z is associated with X, it should affect the probability of X
  - relRisk(Z → X) >> 1.0
    - Blood Pressure → 1.98
    - Cholesterol → 1.27
    - Humidity → 0.65
    - Temperature → 21.02

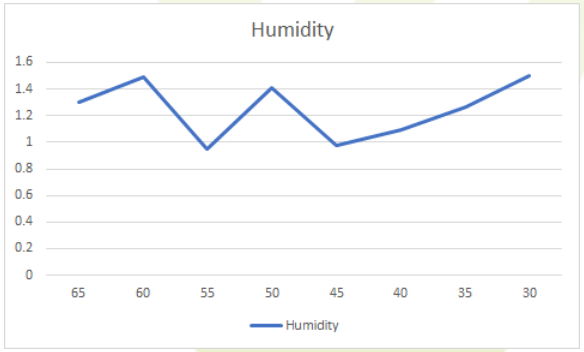
How to Detect Confounders?

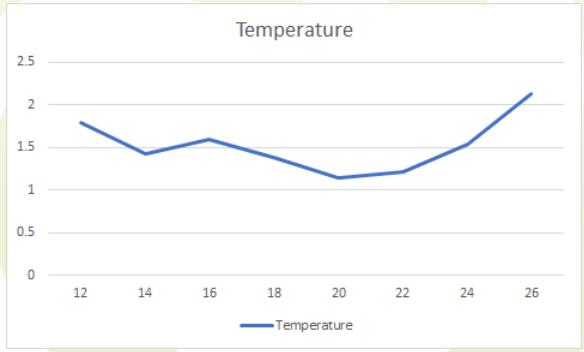
- What else can we do?
  - If the treatment is based on numerical values, its effect on the observation should increase with higher values
    - e.g. raising the blood pressure should increase the probability of heart diseases occuring











Conclusion

- Without external information...
  - It is only possible to look for hints that contradict a causal relation
  - Possibilities include application of correlation/relative risk
- Most real world data sets describe multicausal environments
  - Further Restrictions possible
  - Confounders are encountered regularly
  - Confounder detection not really applicable without a causal model
- Without domain knowledge it is not possible to create causal models
  - It might be possible to extract a causal model from existing narratives

Thank you for your Attention!

Technische Universität Braunschweig