Business Analytics & Machine Learning

Data Preparation and Causal Inference

Prof. Dr. Martin Bichler

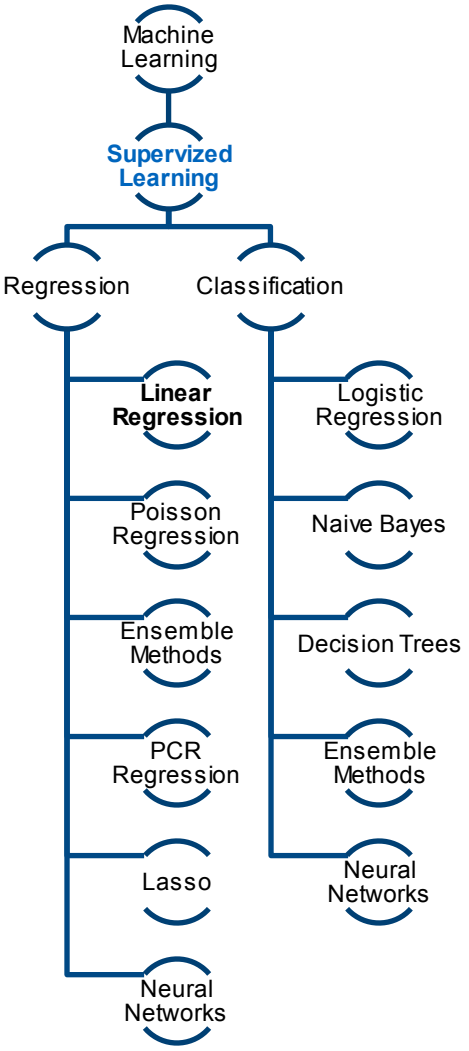
Department of Computer Science

School of Computation, Information, and Technology

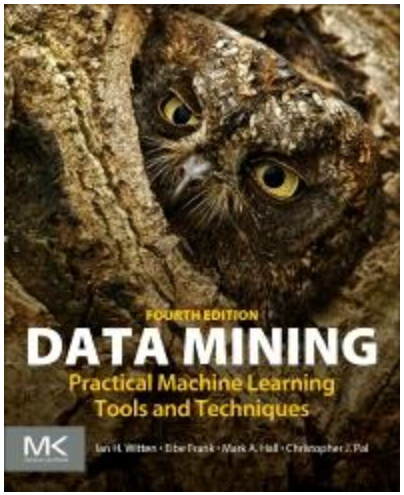
Technical University of Munich

Course Content

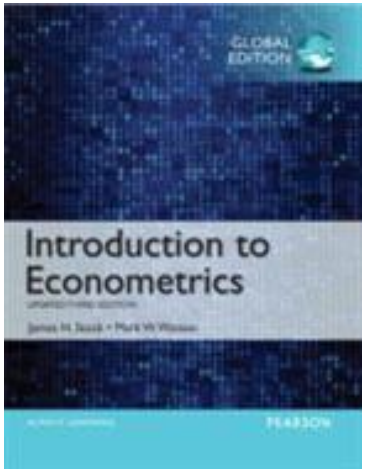
- Introduction
- Regression Analysis
- Regression Diagnostics
- Logistic and Poisson Regression
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- **Data Preparation and Causal Inference**
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- Dimensionality Reduction
- Association Rules and Recommenders
- Convex Optimization
- Neural Networks
- Reinforcement Learning



Recommended Literature



- **Data Mining: Practical Machine Learning Tools and Techniques**
  - Ian H. Witten, Eibe Frank, Mark A. Hall, Ch. Pal
  - http://www.cs.waikato.ac.nz/ml/weka/book.html
  - Section: 6.1, 8.1, 8.2



- **Introduction to Econometrics**
  - James H. Stock and Mark W. Watson
  - Chapter 9, 10, 13

- **R for Data Science**
  - Hadley Wickham, Garett Grolemund https://r4ds.had.co.nz/
  - Chapters 9, 12, 13, 15

- Online material on causal inference in econometrics
  - Empirical Economics by Esther Duflo, MIT Economics
  - http://web.mit.edu/14.771/www/emp_handout.pdf

Agenda for Today

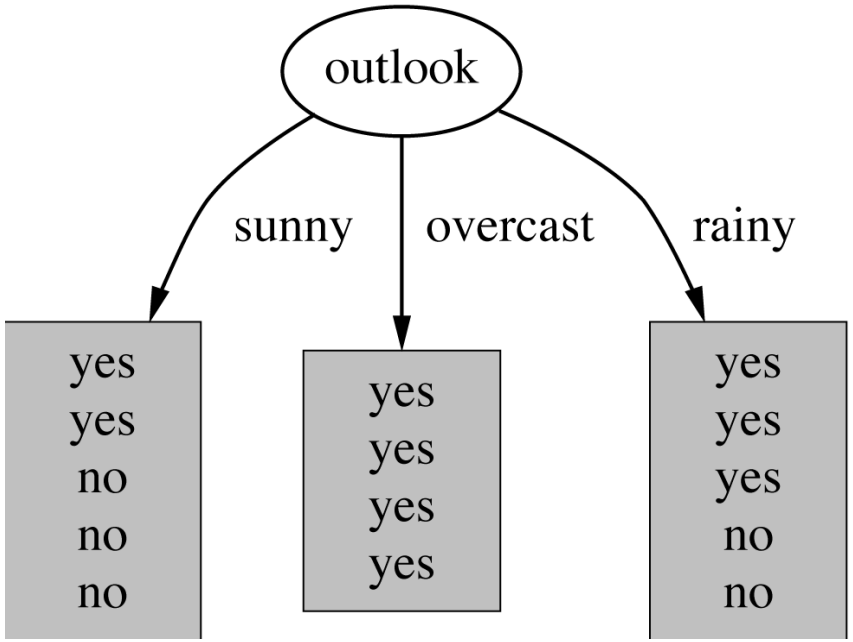
- **Pruning of decision trees**
- CRISP-DM process model
- Data preparation
  - Data cleaning
  - Balanced training data
  - Discretization
  - Feature selection
- Causal inference

Decision Tree Algorithms

At each node, available attributes are evaluated on the basis of separating the classes of the training examples.

A goodness function is used for this purpose.

Typical goodness functions:
- information gain (ID3/C4.5)
- information gain ratio
- gini index (CART)



Computing Information (last Class)

Information is measured in **nits**.

- Given a probability distribution, the info required to predict an event, i.e. if play is yes or no, is the distribution's **entropy**.
- **Entropy** (with the natural logarithm) gives the information required in nits. (This can involve fractions of nits!)

- Formula for computing the entropy:

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \ln p_1 - p_2 \ln p_2 \ldots - p_n \ln p_n$$

Pruning

Prepruning tries to decide *a priori* when to stop creating subtrees.

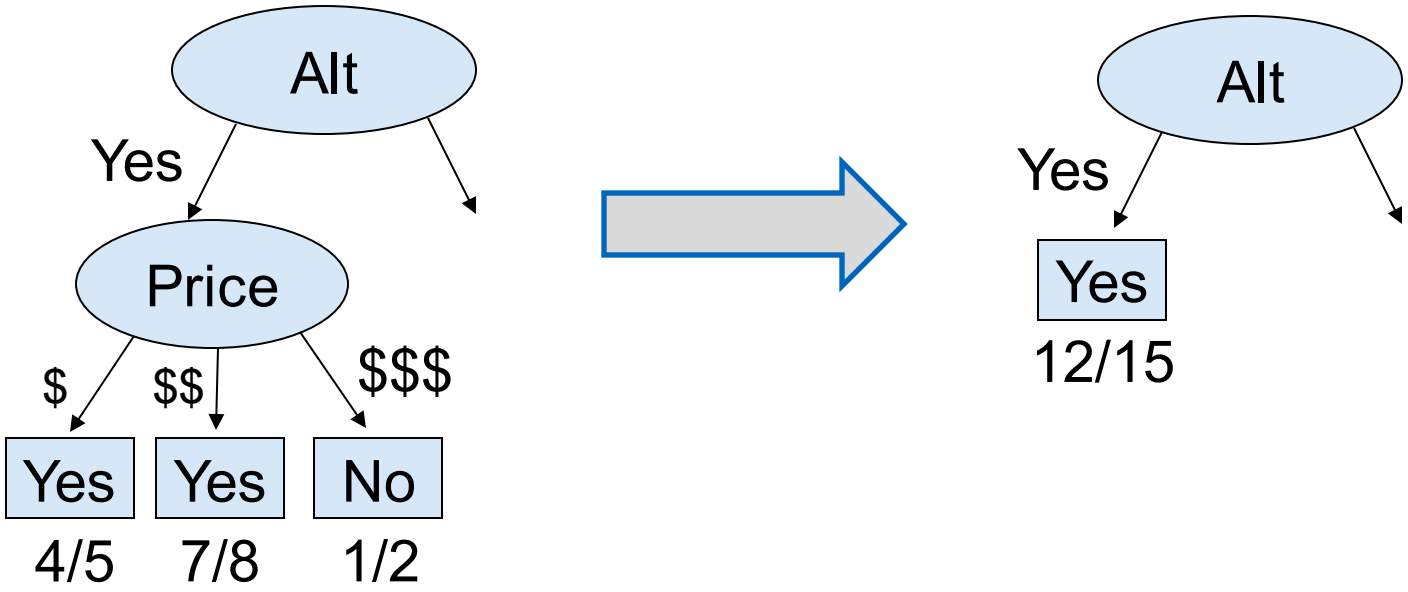
- halt construction of decision tree early
- use same measure as in determining attributes,
  e.g., halt if $\text{gain}(S, a) <$ threshold
- most frequent class becomes the leaf node
- this turns out to be fairly difficult to do well in practice

Postpruning simplifies an existing decision tree.

- construct complete decision tree
- then prune it back
- used in C4.5, CART
- needs more runtime than **prepruning**

Postpruning

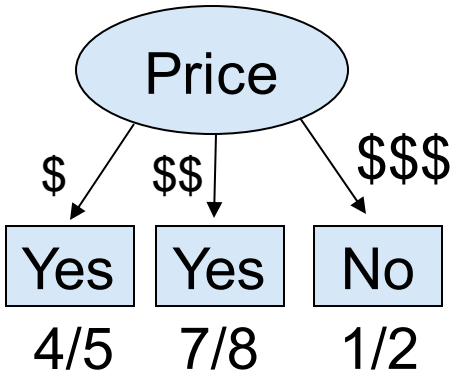
Subtree replacement replaces a subtree with a single leaf node (main method).



When to Prune a Tree?

To determine if a node should be replaced, compare the error rate estimate for the node with the combined error rates of the children.

Replace the node if its error rate is less than combined rates of its children.



err(3/15, 15) = 0.28
5/15 err(1/5,5) + 8/15 err(1/8, 8) + 2/15 err(1/2,2) = 0.33

When to Prune a Tree?

Prune only if it reduces the estimated error.

- Error on the training data is NOT a useful estimator.

Use a hold-out set for pruning (“reduced-error pruning”).

- Limits data you can use for training.

C4.5’s method:

- Derive confidence interval from training data.
- Use a heuristic limit for the error rate, derived from the confidence interval for pruning.
- Shaky statistical assumptions (because it is based on training data), but works well in practice.

Approximation of the Binomial Distribution

"Observed error rate" $f = \frac{errors}{n}$ with

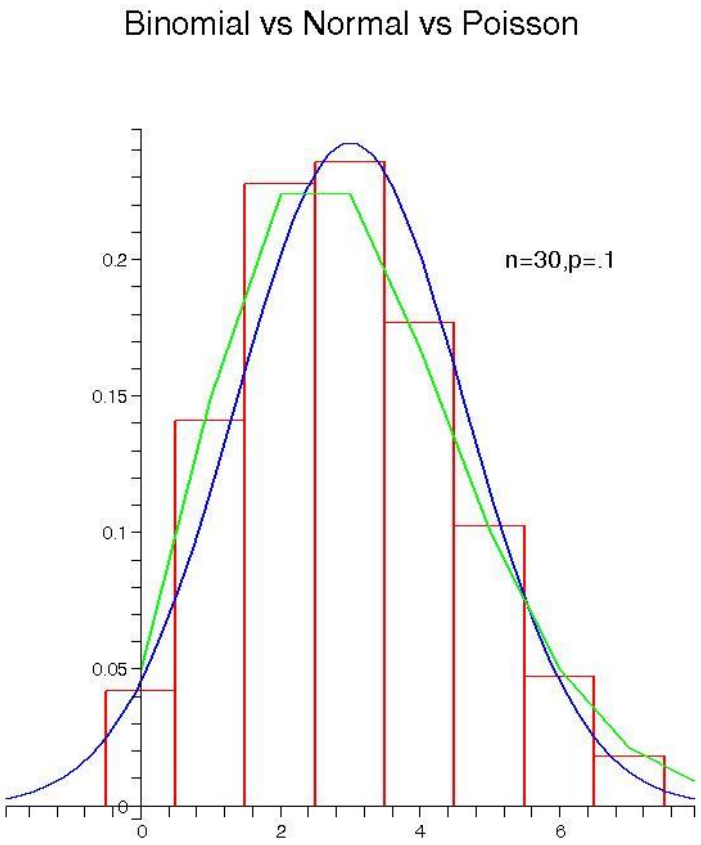
n = number of trials

This random variable can be modeled as a Bernoulli process.

$$B(k|p,n) = \binom{n}{k} p^k q^{n-k}$$

For large enough n, f follows a Normal distribution.

- Central Limit Theorem



Binomial PDF and Normal approximation for n=30 and p=0.1.

Central Limit Theorem Revisited

The **central limit theorem** states that the standardized average of any population of i.i.d. random variables X_i with mean μ_X and variance σ^2 is asymptotically $\sim N(0,1)$, or

Asymptotic Normality implies that
$P(Z \leq z) \to \Phi(z)$ as $n \to \infty, or\ P(Z \leq z) \approx \Phi(z)$

$$Z = \frac{\bar{X} - \mu_X}{\sigma/\sqrt{n}} \sim N(0,1)$$

Using the Confidence Interval of a Normal Distribution

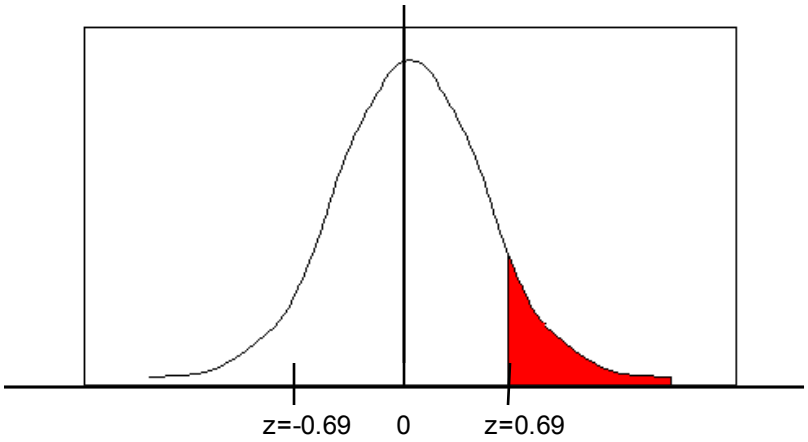
C4.5 uses a heuristic limit for the error rate, derived from the confidence interval of the error rate for pruning.

x% confidence interval $[-z \leq X \leq z]$ for random variable with 0 mean is given by:
$\Pr[-z \leq X \leq z] = x$

With a symmetric distribution:
$\Pr[-z \leq X \leq z] = 1 - 2\Pr[X \geq z]$

Confidence Limits

Confidence limits c for the standard normal distribution with 0 mean and a variance of 1:



$c = \Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
25%	0.68
40%	0.25

There is a 25% probability of X being > 0.68

$$\Pr[-0.68 \leq X \leq 0.68]$$

To use this we have to reduce our random variable f to have 0 mean and unit variance.

Transforming f

Standardized value for observed error rate f : $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{f-p}{\sqrt{p(1-p)/n}}$

(Standardization: subtract mean and divide by the standard deviation)

Binomial conf. interval: $\Pr\left[\frac{f-p}{\sqrt{p(1-p)/n}} > z\right] = c$

Solving for p provides limits for the confidence factor c:

$$p = \left(f + \frac{z^2}{2n} \pm z * \sqrt{\frac{f}{n} - \frac{f^2}{n} + \frac{z^2}{4n^2}}\right) \Big/ \left(1 + \frac{z^2}{n}\right)$$

You prune the tree stronger

- If c goes down → z goes up and also p goes up
- If n goes down → p goes up
- with p as an estimator for the error rate

C4.5's Method

Error estimate e for a node (:= upper bound of confidence interval):

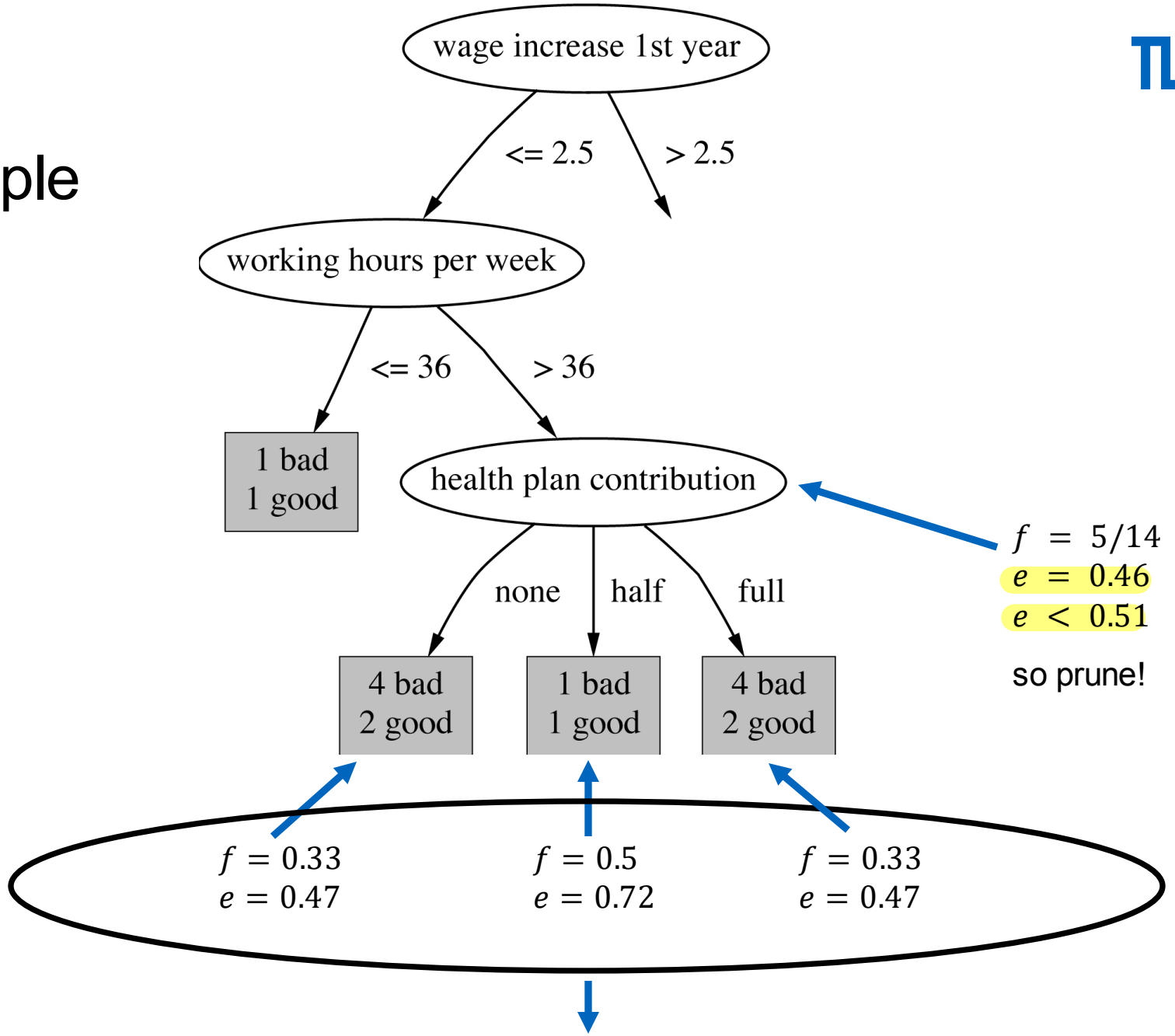
$$e = p = \left(f + \frac{z^2}{2n} + z * \sqrt{\frac{f}{n} - \frac{f^2}{n} + \frac{z^2}{4n^2}} \right) \Big/ \left(1 + \frac{z^2}{n} \right)$$

- If confidence limit $c = 25\%$ then $z = 0.69$ (from Normal distribution)
- f is the error on the training data
- n is the number of instances covered by the node

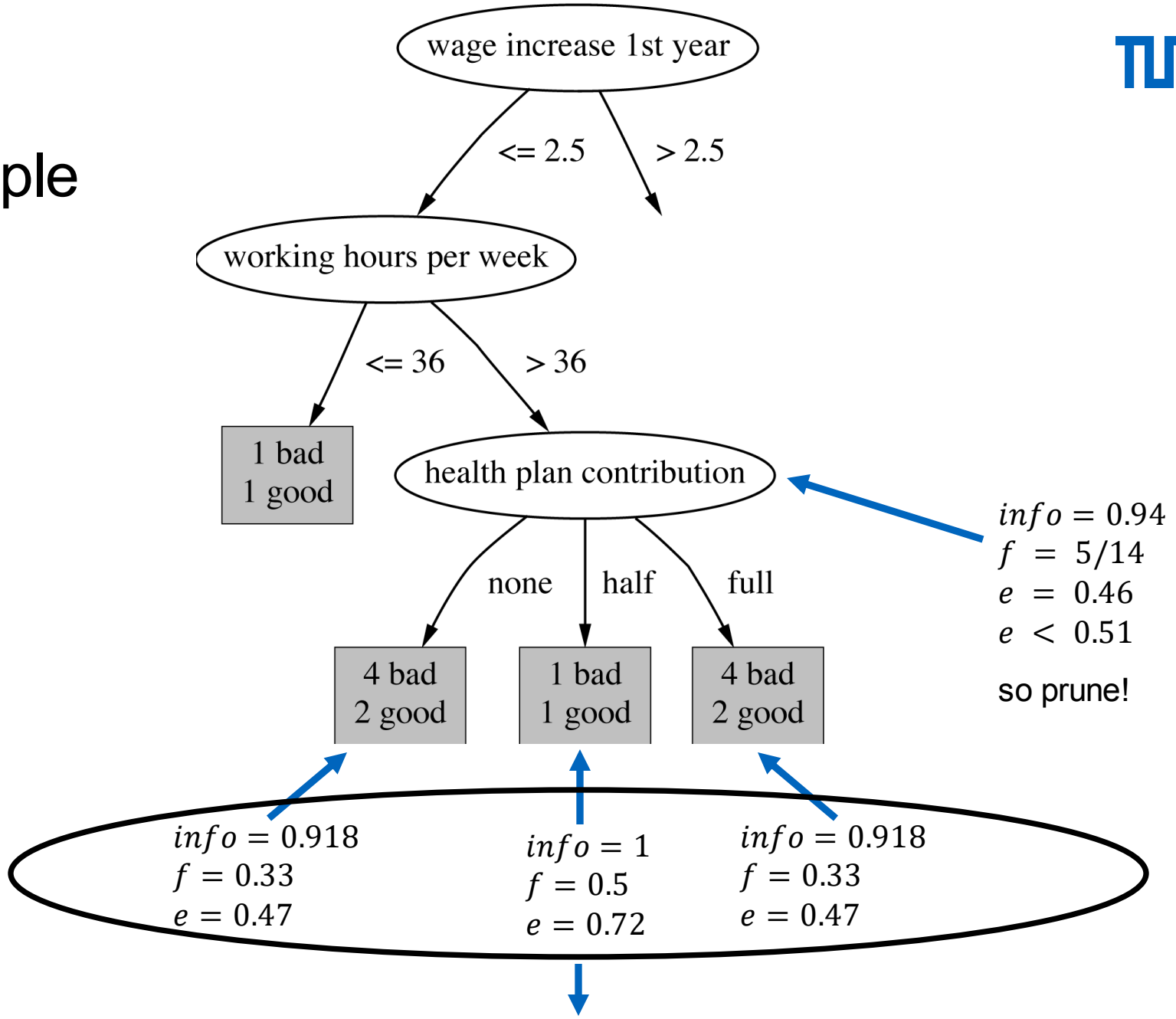
Even with pos. information gain, e might increase as well.

Error estimate for subtree is weighted sum of error estimates for all its leaves.

Example

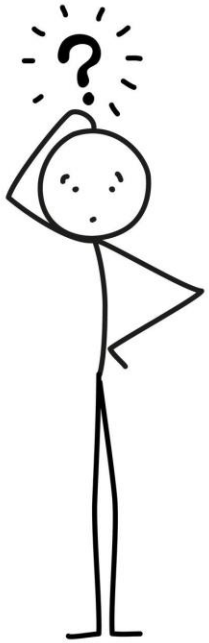

wage increase 1st year
<= 2.5
> 2.5
working hours per week
<= 36
> 36
1 bad
1 good
health plan contribution
none
half
full
4 bad
2 good
1 bad
1 good
4 bad
2 good
f = 5/14
e = 0.46
e < 0.51
so prune!
f = 0.33
e = 0.47
f = 0.5
e = 0.72
f = 0.33
e = 0.47
Combined and weighted using ratios 6:2:6 gives 0.51

Example



Info for sub-nodes = 0.92997, Combined and weighted using ratios 6:2:6 gives 0.51

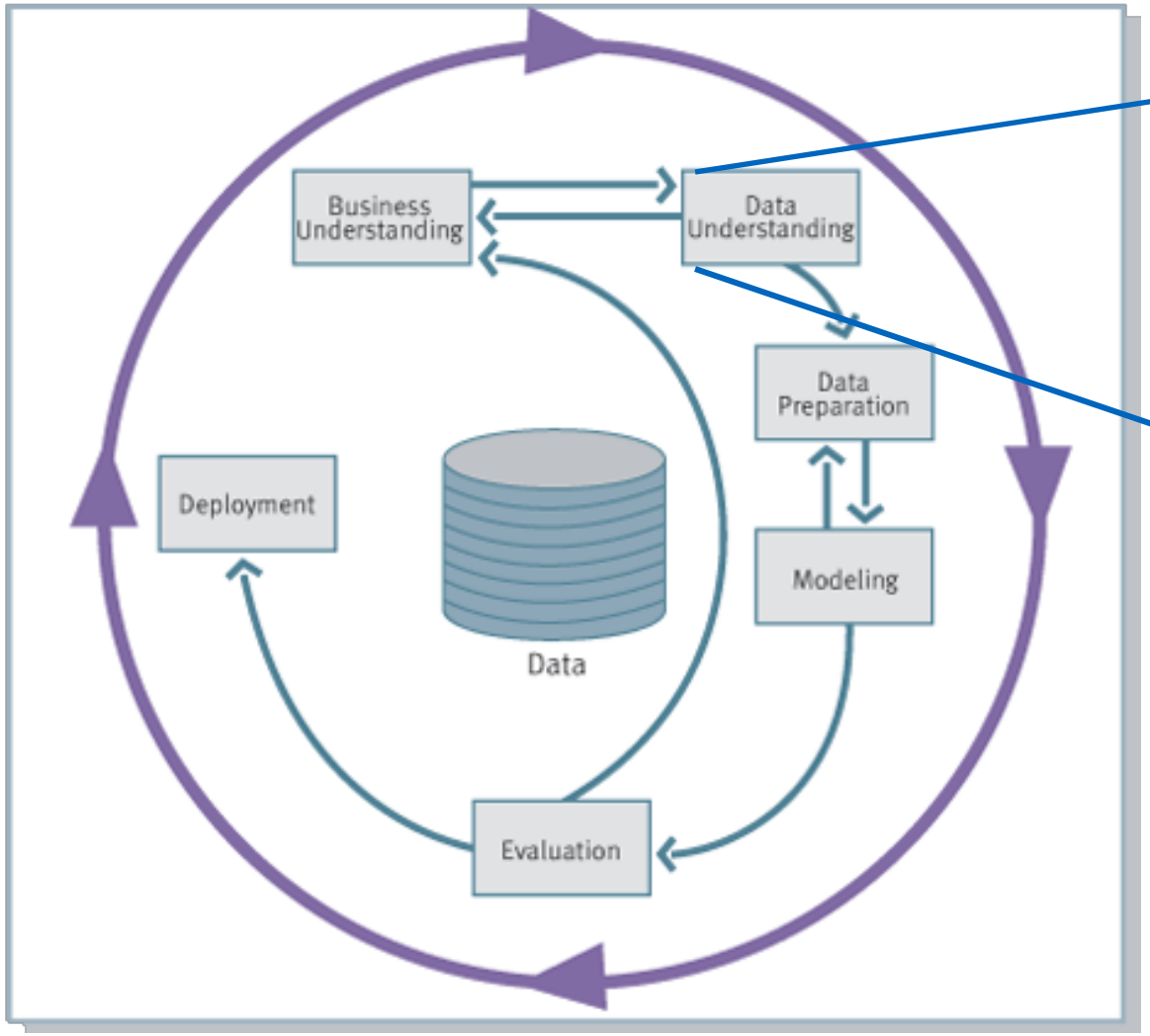
Do you remember the pseudo code for decision tree learners discussed in the last class?



Outline for today

- Pruning of decision trees
- **CRISP-DM process model**
- Data preparation
  - Data cleaning
  - Balanced training data
  - Discretization
  - Feature selection
- Causal inference

Knowledge Discovery Process



Data Understanding

Data Understanding: Quantity

Number of instances (records):

- Rule of thumb: 5,000 or more desired (nice to have)
- if less, results are less reliable, use special methods (boosting, …)

Number of attributes:

- Rule of thumb: start out with less than 50 attributes
- if many attributes, use attribute selection

Number of targets:

- Rule of thumb: >100 for each class
- if very unbalanced, use stratified sampling

Data Understanding: Quantity

- Analyzing the attributes:
  - How many attributes, and which?
  - What kind of attributes? (nominal, ordinal, interval, ratio)
  - How many missing values per attribute?

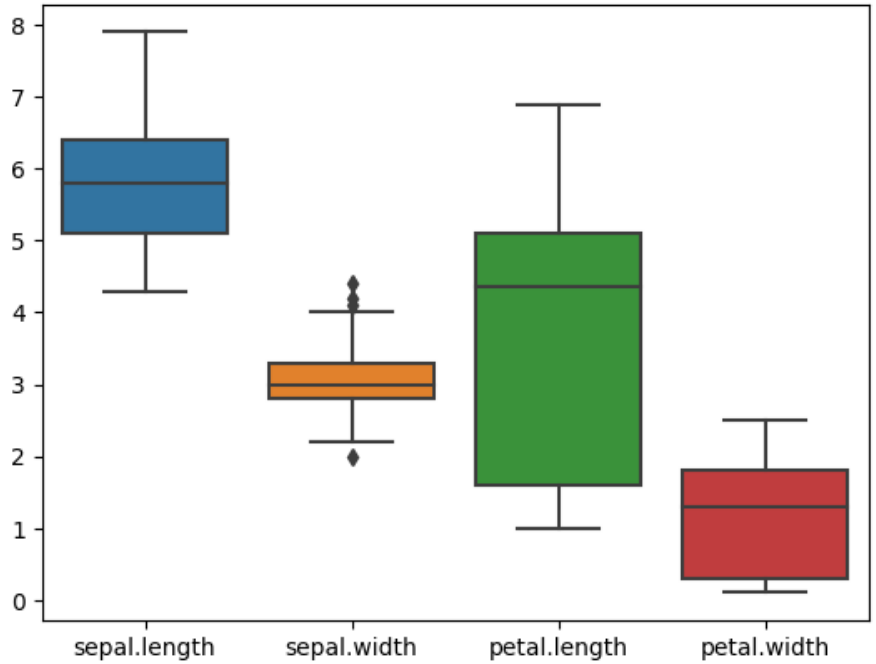
- Computing Statistics (per attribute):
  - Counts
  - Means
  - Variance
  - Ranges
  - Quartiles
  - Etc.

Data Understanding: Visualization

„Boxplot“:

- Compactly visualizes the „five-number summary“
  1. Sample minimum (min)
  2. Lower quartile (25%)
  3. Median (50%)
  4. Upper quartile (75%)
  5. Sample maximum (max)

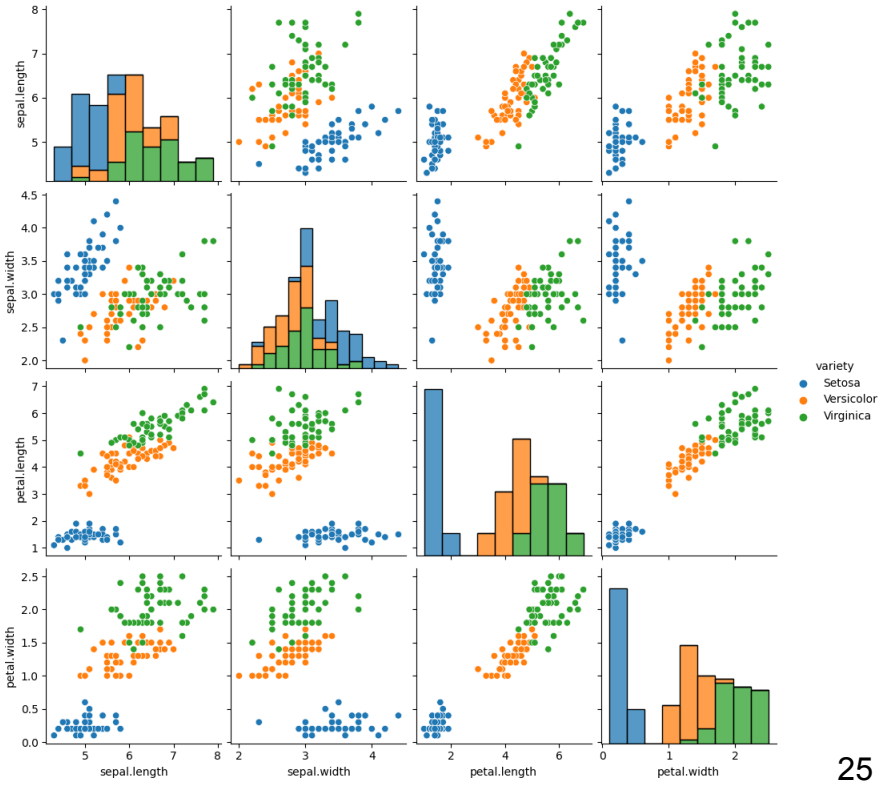
- „modified Boxplots“ also show outliers (modified min/max to certain ranges)



Data Understanding: Visualization

„Scatterplot-Matrix“:

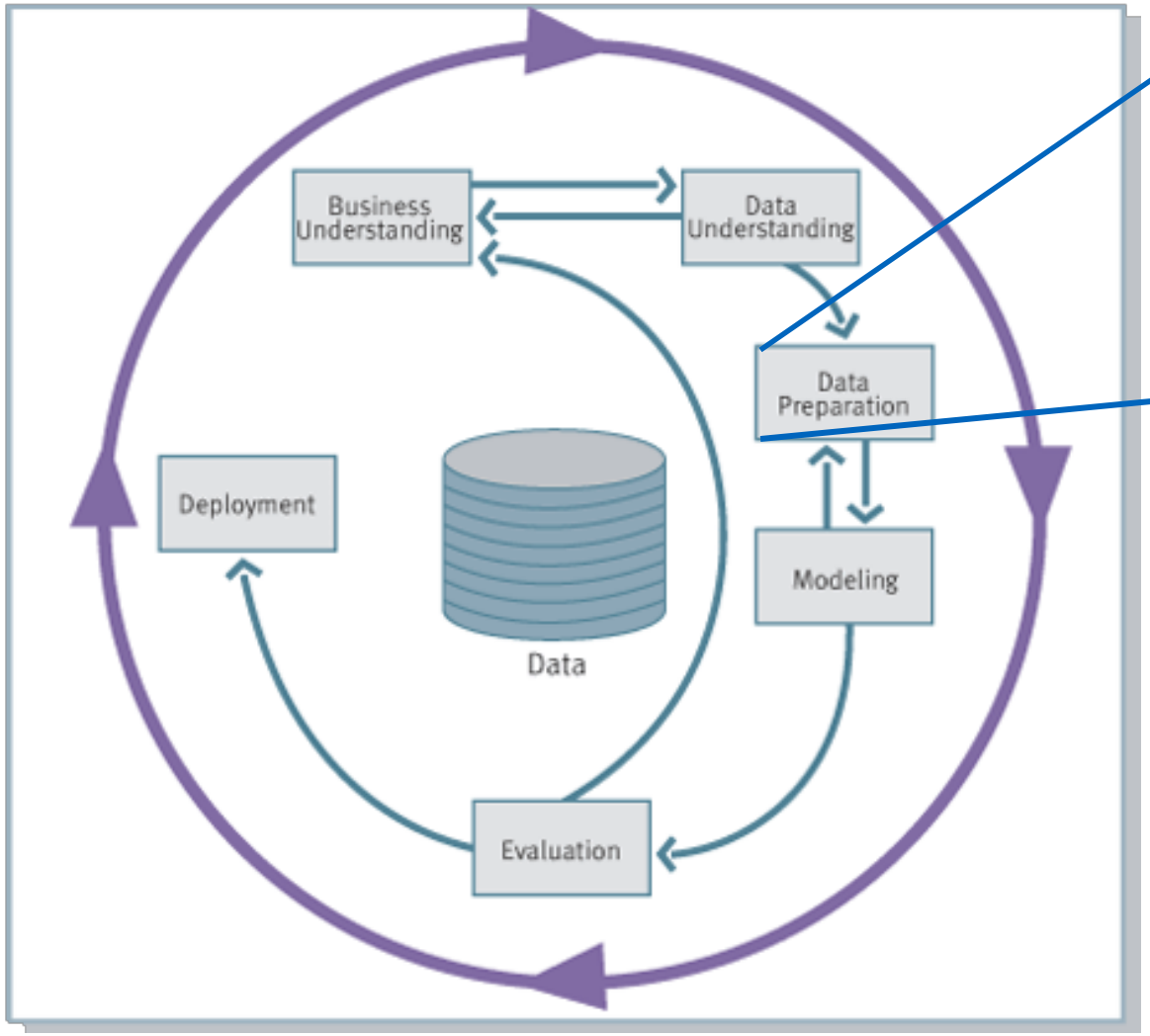
- Compactly visualizes relationsships between attributes and their distribution
  - Each plot shows relationships between two variables
  - Diagonal plots show distributions (can be histogram or density)



Outline for today

- Pruning of decision trees
- CRISP-DM process model
- **Data preparation**
  - Data cleaning
  - Balanced training data
  - Discretization
  - Feature selection
- Causal inference

Knowledge Discovery Process



Data Preparation
(Witten, Frank: Chapter 8)

Data Preparation estimated to take 70-80% of the time and effort

Tidy Data (see Tutorial slides for more)

- Tidying up a dataset is usually the first step of data cleaning.
- An input table is often optimized for something other than analysis (e.g. data entry, storage, fast processing, compliance with required formats, interfaces with other systems…)
- It can sometimes be ambiguous what constitutes a feature based on context:
  e.g. “Address” vs “Street | Number | ZIP | City | Country”;
- Dates need to be converted.
- Data needs to be scaled appropriately.
- Often relational data needs to be converted into a single table.

moodle_posts.csv

Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv

Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

Data Cleaning: Missing Values

Missing data can appear in several forms:

- <empty field> “0” “.” “999” “NA” …

Standardize missing value code(s)

Dealing with missing values:

- Ignore records with missing values
- Treat missing value as a separate value
- Imputation: fill in with mean or median values

Remark: the tutorial will provide you with more details about tidy data and

Conversion: Ordered to Numeric

Ordered attributes (e.g. Grade) can be converted to numbers preserving **natural** order, e.g.

- A → 4.0
- A- → 3.7
- B+ → 3.3
- B → 3.0

Why is it important to preserve **natural** order?

- To allow meaningful comparisons, e.g. Grade > 3.5

This is referred to as ***ordinal* encoding**.

Conversion: Nominal, Few Values

Multi-valued, unordered attributes with small no. of values:

- e.g. $Color = Red, Orange, Yellow, \ldots$
- for each value v create a binary dummy variable C_v , which is 1 if $Color = v$, 0 otherwise

ID	Color	…
371	Red	
433	yellow	

ID	C_red	C_orange	C_yellow	…
371	1	0	0	
433	0	0	1	

Using $|v|$ dummy variables is called **one-hot encoding** (OHE).

- OHE would lead to problems in the linear regression (wrt. to matrix inversion).

Using $|v| - 1$ variables to describe $|v|$ values is called **dummy variable encoding**.

Data Cleaning: Discretization (aka. Binning)

Discretization reduces the number of values for a continuous attribute.

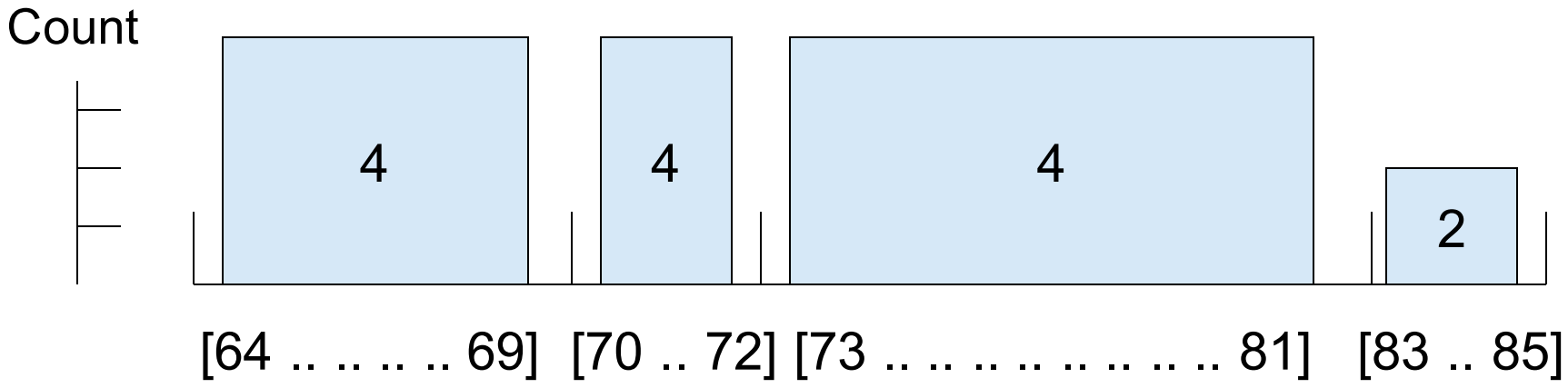
Why?

- some methods can only use nominal data
  - e.g., in ID3, Apriori, most versions of Naïve Bayes, CHAID
- some methods assume a normal distribution for any numerical which is not always appropriate → discretizing gets around this implicit assumption
- discretization can be useful to generate a summary of the data (e.g. histograms)
- Historically, discretization was often useful to improve the runtime of certain models (e.g. decision trees) when computational recourses were scarce.

Note that binning introduces some arbitrariness to the analysis. Nowadays, you can usually work directly with continuous data, modern implementations of decision trees, random forests, or gradient boosting do not require such preprocessing.

Discretization: Equal-Frequency

Temperature values:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Equal Height = 4, except for the last bin

Discretization: Class Dependent (Supervised Discretization)

For example, based on information gain of the class variable (see C4.5)

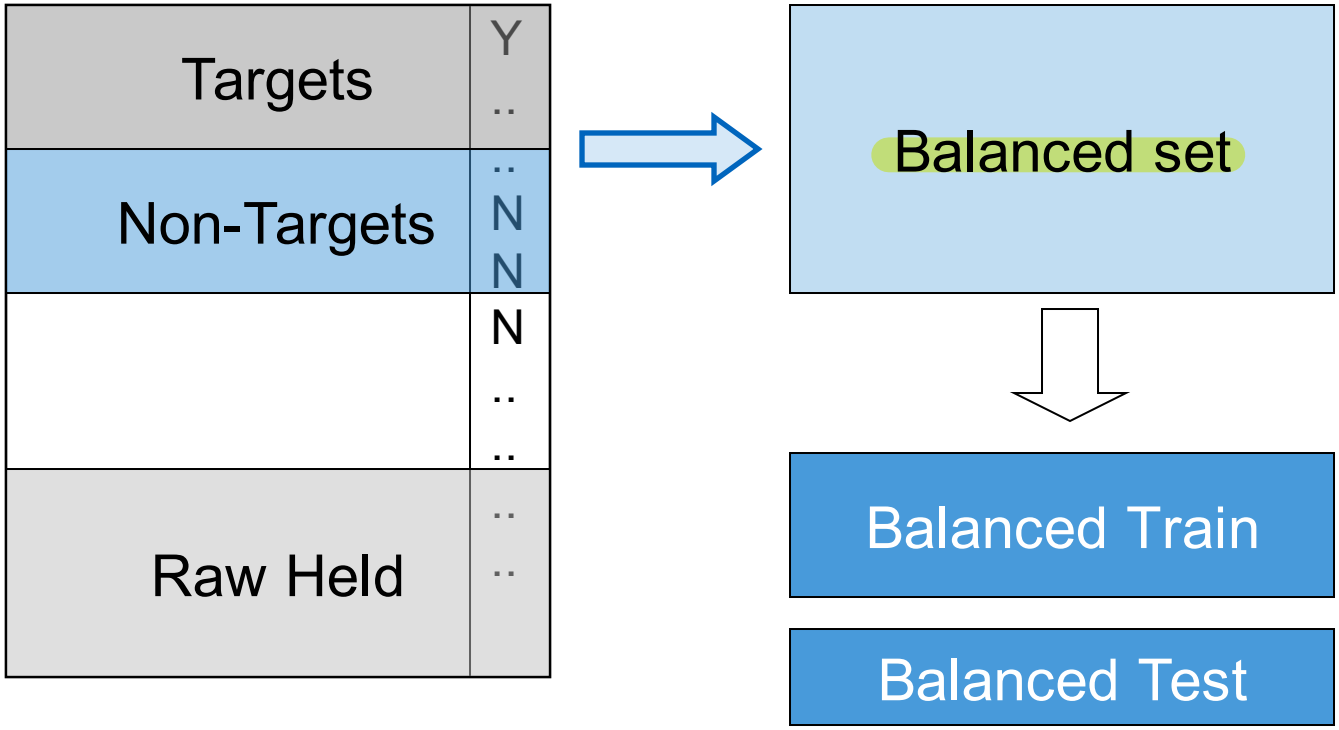
64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	**No**	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**Yes**	**No**

- treating numerical attributes as nominal discards the potentially valuable ordering information
- alternative: transform the k nominal values to $k-1$ binary attributes
- the $(i-1)^{th}$ binary attribute indicates whether the discretized attribute is less than i

Unbalanced Target Distribution

- sometimes, classes have very unequal frequency
  - Churn prediction: 97% stay, 3% churn (in a month)
  - Medical diagnosis: 90% healthy, 10% disease
  - eCommerce: 99% don't buy, 1% buy
  - Security: >99.99% of Germans are not terrorists
- similar situation with multiple classes
- majority class classifier can be 97% correct, but useless

Building Balanced Train Sets

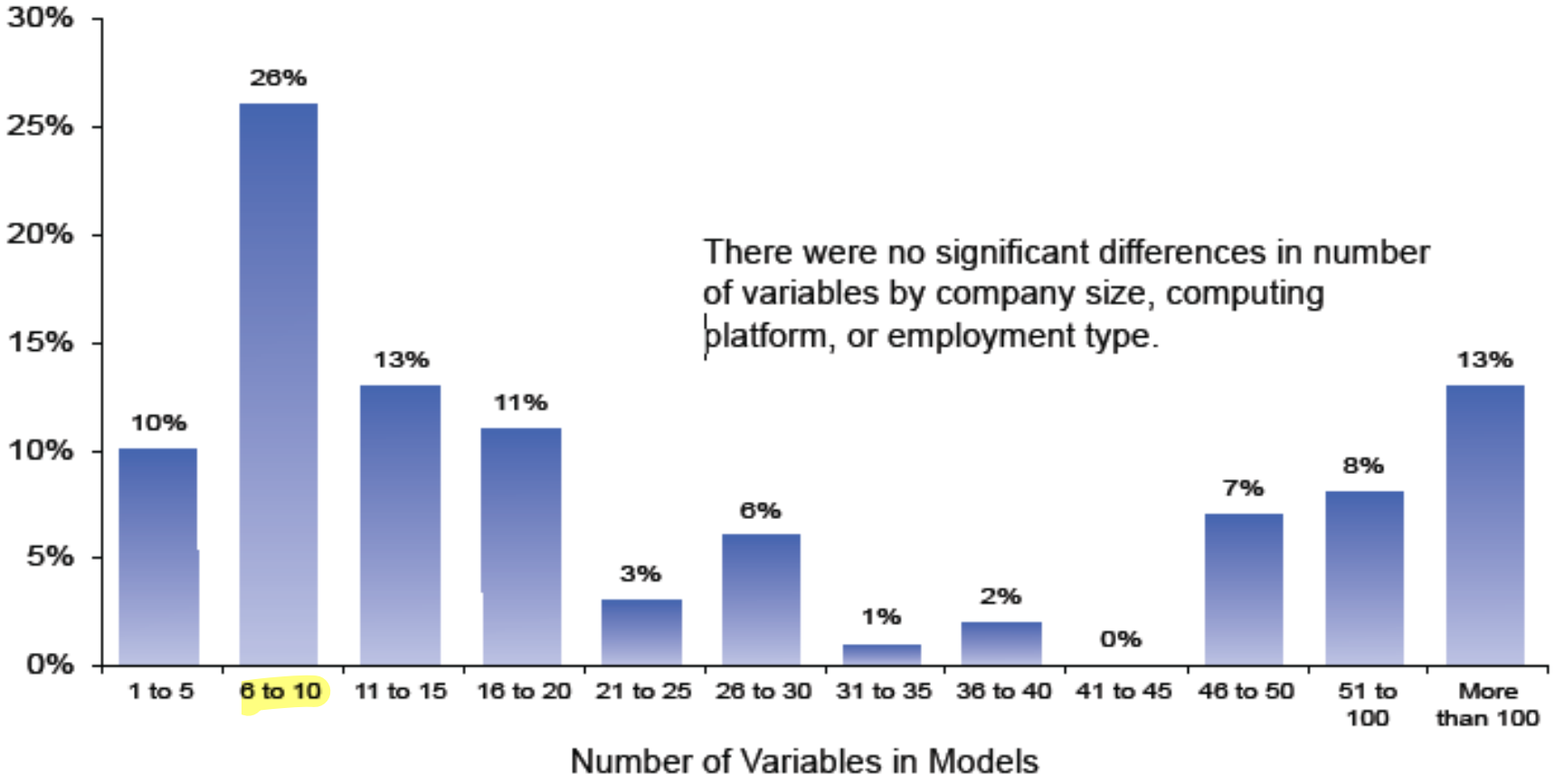


Subset Selection

- aka feature / variable / field selection
- if there are too many attributes, select a subset that is most relevant
  - companies often have > 2000 attributes per customer
- subset selection was already discussed in the context of the linear regression, but is also relevant to classification in general
  - more in the context of dimensionality reduction and regularization
- remove redundant and/or irrelevant attributes
  - rule of thumb: keep top 50 attributes
- automated procedures:
  - best subset (among all exponentially many, computationally expensive)
  - backward elimination (top down approach)
  - forward selection (bottom up approach)
  - stepwise regression (combines forward/backward)

Number of Variables in Final Models

- About one-third of data miners typically build final models with 10 or fewer variables, while about 28% generally construct models with more than 45 variables.

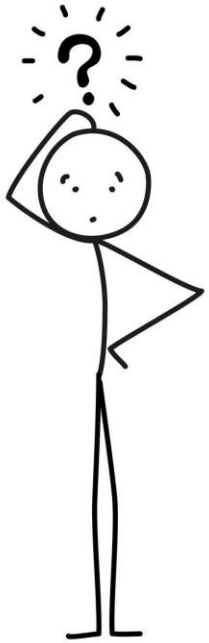


There were no significant differences in number of variables by company size, computing platform, or employment type.

Question: How many variables are typically incorporated into your final model?

Vendors were excluded from this analysis.

How would you define causality?



Outline for today

- Pruning of decision trees
- CRISP-DM process model
- Data preparation
  - Data cleaning
  - Balanced training data
  - Discretization
  - Feature selection
- **Causal inference**

Example: **True Advertising Lift**

1) You analyze the user journey of customers and show a display ad on a web site to likely customers (treatment).

2) Actually, 2.8% of those who have seen the ad buy a car as compared to 0.1% in the group who has not seen a display ad.

3) Success!?

Causal Inference

In many cases, we want to draw conclusions about the impact of a treatment (higher price) on an outcome (churn).

- Y outcome (dependent variable)
- T treatment indicator
- X covariate (pretreatment)

What would have happened to those who, in fact, received treatment, if they have not received treatment (or vice versa)?

Correlation does not imply causation!

Causal Inference

- Y_{1i} denotes the outcome of individual i given i was treated
- Y_{0i} denotes the outcome of individual i given i was in the control group
- $\Delta_i = Y_{1i} - Y_{0i}$ is the treatment effect on i

Sub.	Y_1	Y_0	Δ
A	15		
B	13		
C		8	
D		4	

Causal Inference

In a perfect world, we could observe both Y_{1i} and Y_{0i}.

- Individual treatment effect: $Y_{1i} - Y_{0i}$
- Average treatment effect: $E\ (Y_{1i} - Y_{0i})$
- Subgroup treatment effect: $E\ (Y_{1i} - Y_{0i}|X)$

Sub.	X	Y_1	Y_0	Δ
A	40	15	10	5
B	30	13	8	5
C	30	13	8	5
D	20	9	4	5

Fundamental problem of causal inference: Cannot observe both Y_{1i} and Y_{0i}.

We do not know the counterfactual. The best we can do is to find an approximation for the potential outcome (see Rubin's causal model).

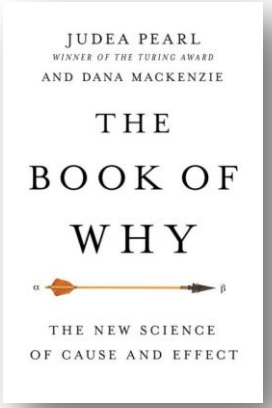
Judea Pearl on Causality

Pearl contends that until algorithms can reason about cause and effect, their utility and versatility will never approach that of humans:

"As much as I look into what's being done with deep learning, I see ***they're all stuck there on the level of associations. Curve fitting****."*

"The key is to replace reasoning by association with causal reasoning. Instead of the mere ability to correlate fever and malaria, machines need the capacity to reason that malaria causes fever."

Judea Pearl (ACM Fellow, Turing Award winner)



https://www.theatlantic.com/technology/archive/2018/05/machine-learning-is-stuck-on-asking-why/560675/

Internal and External Validity of a Study

A statistical study has **external validity** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

- Choice of training samples is important.

A statistical analysis has **internal validity** if the statistical inferences about causal effects are valid for the population being studied.

Selected problems:

- misspecification of the functional form of the model (already discussed)
- measurement errors in the independent variables
- simultaneous causality ($x \Rightarrow y$ and $y \Rightarrow x$)
- **sample selection bias**
  - The sample selection process influences the data and is related to the dependent variable.
  - Frequently an issue if data is not collected via randomized controlled trials.

Main Study Designs

Experimental studies

- *The researcher intervenes and observes what happens. Examples are:*
  - **Randomized controlled trials (RCTs)** describes randomized experiments, where each subject is randomly assigned to a treated group or a control group in order to control for extraneous factors.
    - Randomization mitigates the sample selection bias and the different comparison groups allow the researchers to determine any effects of the treatment when compared with the no treatment (control) group.
  - **Quasi-experiments** compare groups and measure effects without randomization of the subjects.
    - The independent variable (e.g., price change) is controlled, but the assignment of subjects is not random (e.g., age, ethnicity). The selection bias is an issue.

Observational studies

- *The researcher studies what occurs but doesn't change the subjects*.
- The independent variable is not under the control of the researcher.
- Also here, the selection bias is a concern.

Examples of Observational Studies

Cross-sectional study:

- Involves data collection from a population, or a representative subset, at one specific point in time.
- For example, take a sample of commuters on a given morning and study their modes of transport.

Panel (or cohort) study:

- A particular form of longitudinal study where a group of subjects is monitored over a span of time.
- For example, a set of households whose purchases are analyzed every month.

Case-control study:

- Study in which two existing groups differing in outcome are identified and compared on the basis of some supposed causal attribute.
- For example, compare the histories of cancer patients to non-cancer patients.

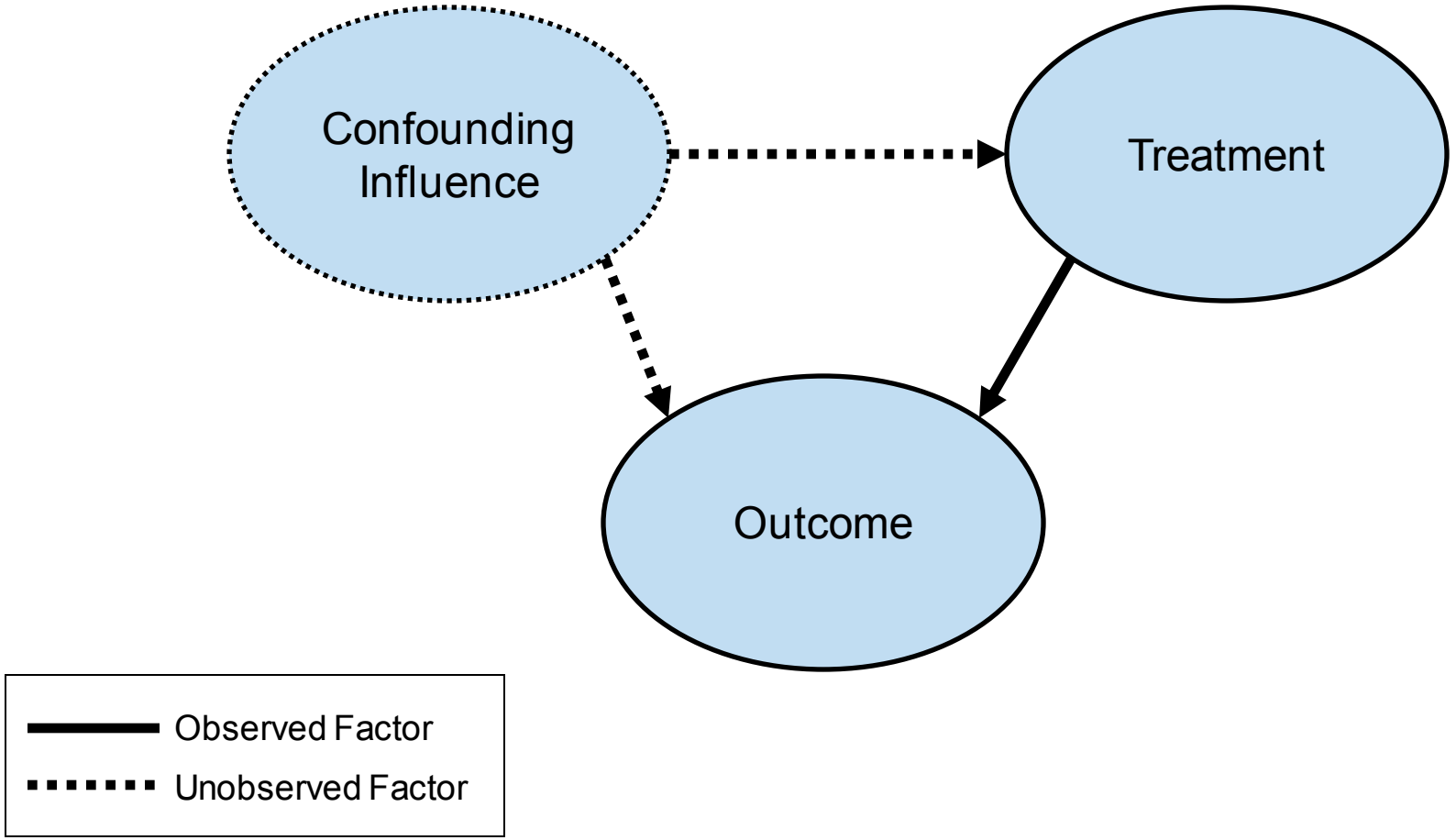
Confounding Variables

Often, we have data from **observational studies** or **quasi-experiments**.
Still, we want to identify causal relationships. However, there might be **confounding variables**:

- **Confounding variable** (also confounding factor, a confound, a lurking variable or a confounder) is an extraneous variable in a statistical model that correlates (directly or inversely) with both the dependent variable and the independent variable, in a way that "explains away" some or all of the correlation between these two variables.

- Remember, exogeneity was one of the Gauss-Markov assumptions in linear regression analysis.

The Problem in Causal Inference

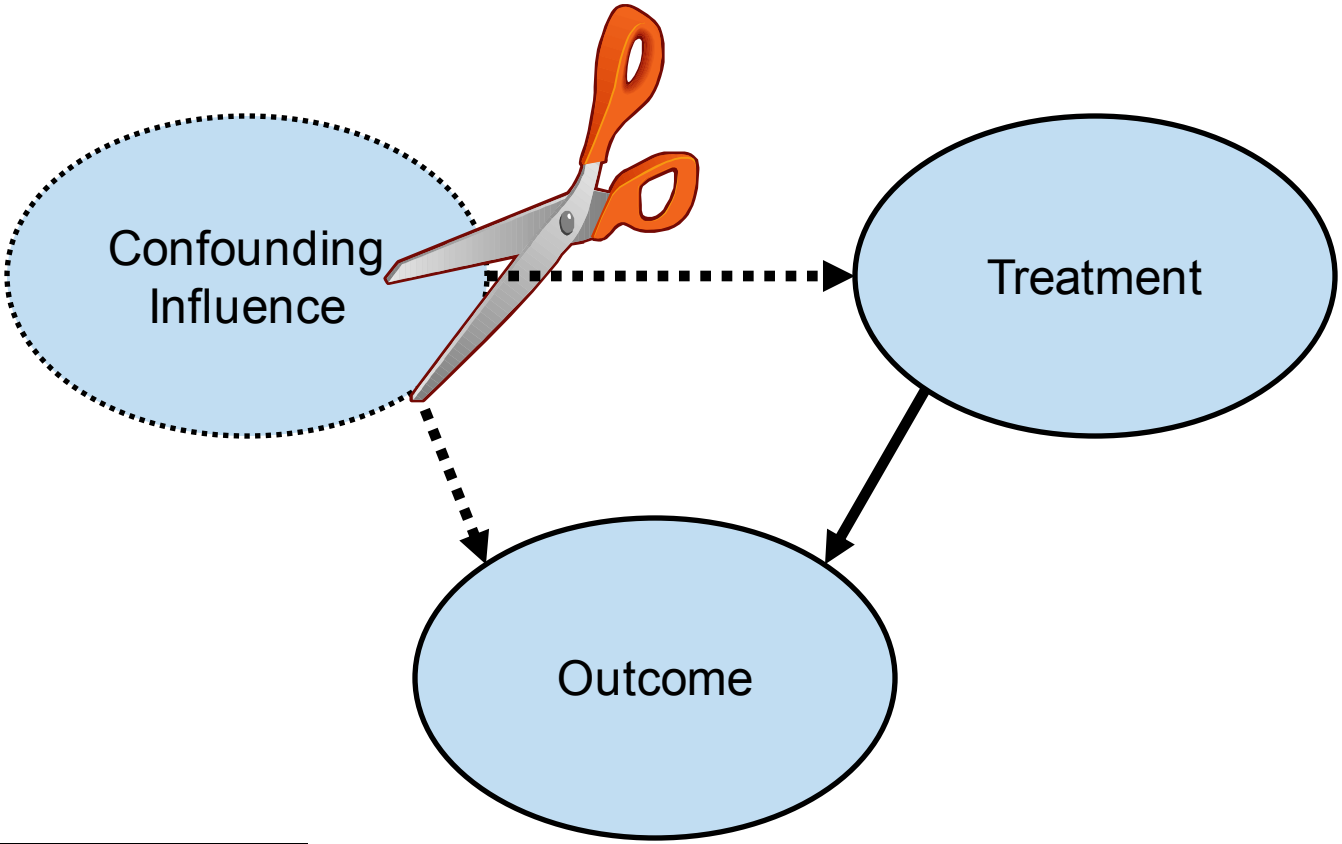


Identification Strategies

Different techniques deal with confounding variables, i.e. to **identify** causal effects:

1. Randomized controlled trials (gold standard, but often not feasible)
   - can be organized in the lab or in the field
2. Fixed or random effects models (for *panel data*)
3. Difference-in-differences (e.g., for *quasi-experimental data*)
4. Propensity score matching (e.g. for *cross-sectional data*)
5. Instrument variables *[not discussed in this course]*
6. Regression discontinuity analysis *[not discussed in this course]*
7. Double Machine Learning *[not discussed in this course]*
8. *…*

1. Randomized Controlled Experiments



1. Lab versus Field Experiments

Lab experiment

- create a situation with desired conditions
- manipulate some variables while controlling others
- examine the dependent variable

Field experiment

- research study in a natural setting
- manipulate some variables
- examine the dependent variable

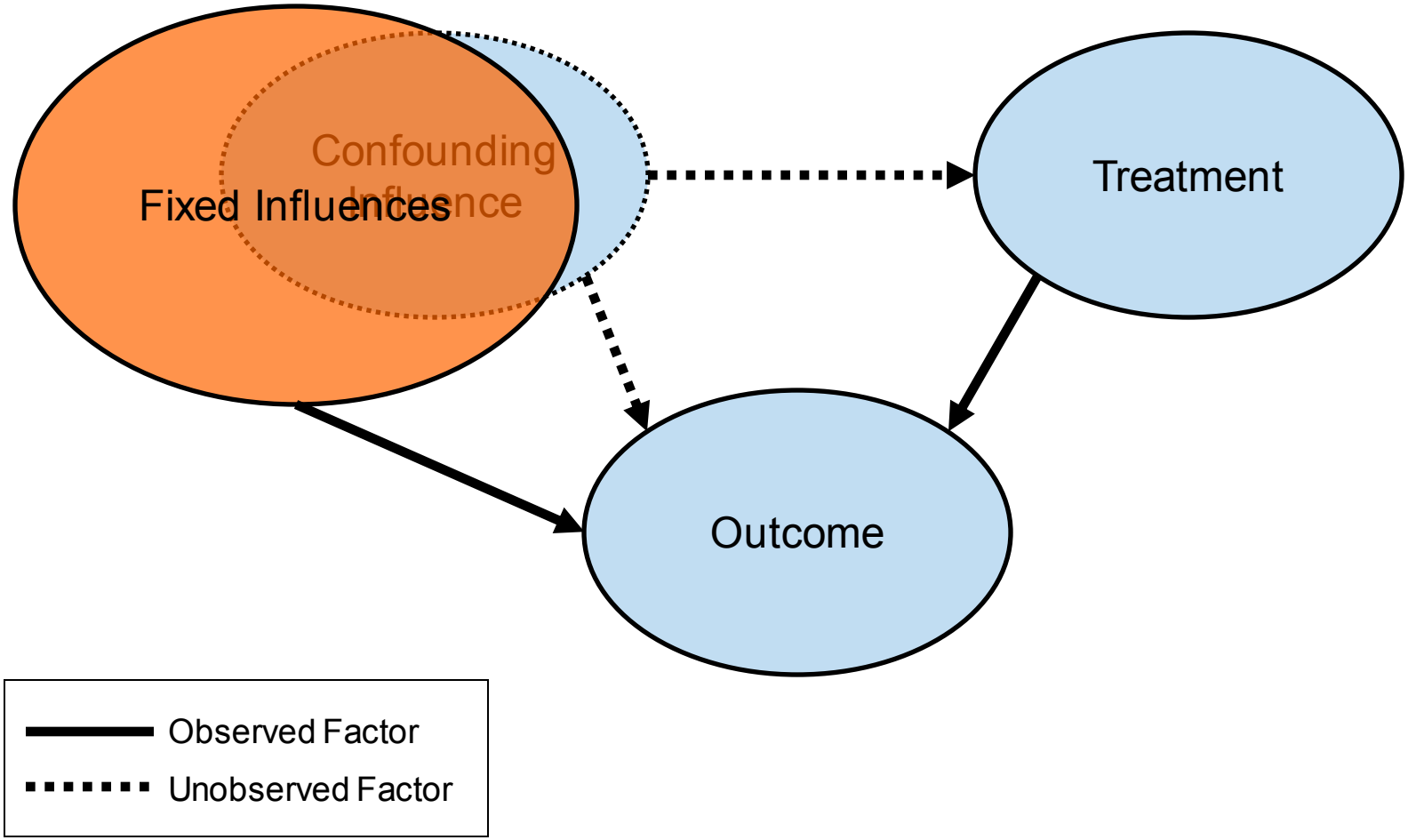
Nowadays online services allow for large-scale field experiments! with randomized assignment of subjects to treatments (RCTs)! (Often called *A/B-Tests*.)

Randomized controlled trials in the field address the selection bias!

	Randomized experiment	Quasi-experiment
Field	**High internal validity/ High external validity**	Low internal validity / High external validity
Lab	High internal validity/ Low external validity	Low internal validity/ Low external validity

↳ because of the "desired conditions"

2. Fixed Effects Models



2. Fixed Effects

Idea

- Fixed effect models assume that the explanatory variable has a fixed or constant relationship with the response variable across all observations.

Example

- You want to understand the probability of a vaccinated person falling sick due to Covid-19.
- You train a linear model with on blood pressure (continuous), the **fixed effects** diabetic (y/n), and country.

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_{bloodp}(BloodPressure) + \beta_{diabetic}(Diabetic) + \beta_{country}(Country)$$

- One could also model the country as a **random effect**, which will incorporate the variability due to picking a subset of all the countries (if only a limited number of countries are analyzed).
  - One can use a Hausman test to decide (see slides on regression diagnostics).
- In other models, there can be person or family fixed-effects that you want to control for to soak up the fixed effect of these factors.

3. Difference in Difference Models

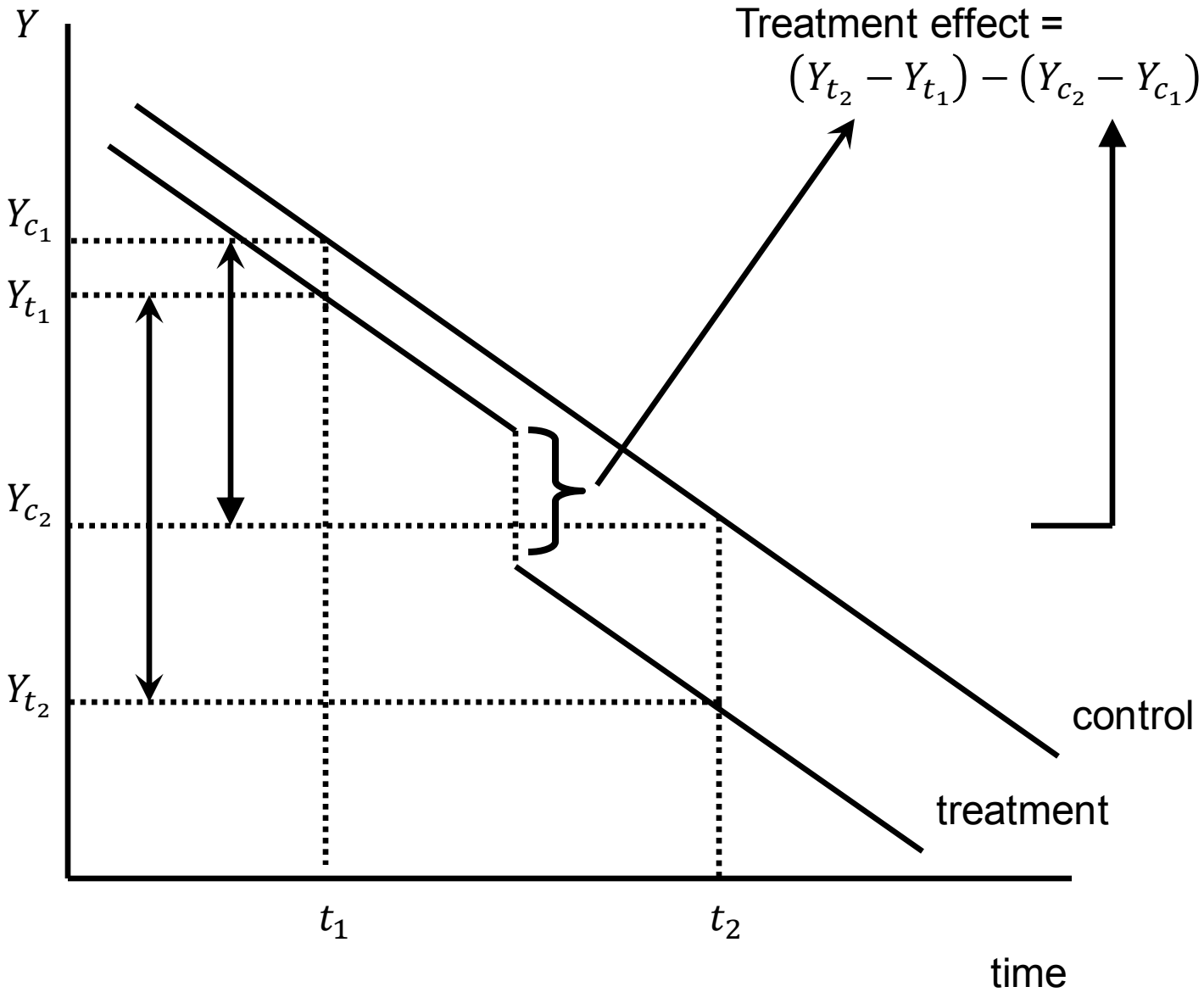
If the treatment in a quasi-experiment is **as if** subjects were randomly assigned, we can use a **differences regression**. However, there might still be differences among the groups (i.e. confounding variables).

We can take a **common trends assumption**, i.e. the treatment and control group would have the same overall trend.

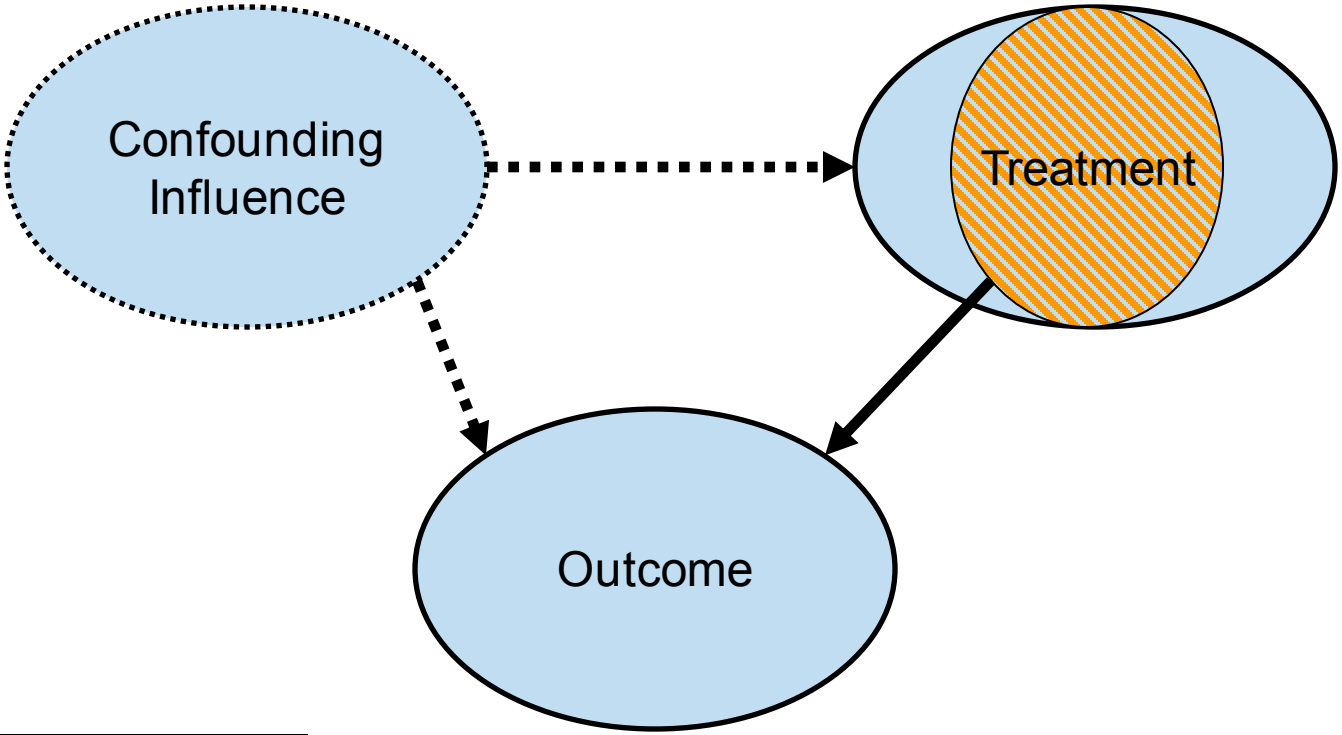
Example: Causal effect of a new law on rent control on rents for apartments per m^2. Cities w/o the new law (control) and with the new law (treatment).

Year	Treatment	Control
2020	9	10.2
2023	6.9	8.7

$$(Y_{t2} - Y_{t1}) - (Y_{c2} - Y_{c1}) = (6.9 - 9) - (8.7 - 10.2) = -0.6$$


Y
Treatment effect =
$(Y_{t_2} - Y_{t_1}) - (Y_{c_2} - Y_{c_1})$
Y_{c_1}
Y_{t_1}
Y_{c_2}
Y_{t_2}
control
treatment
t_1
t_2
time

4. Propensity Score Matching (PSM)



Observed Factor

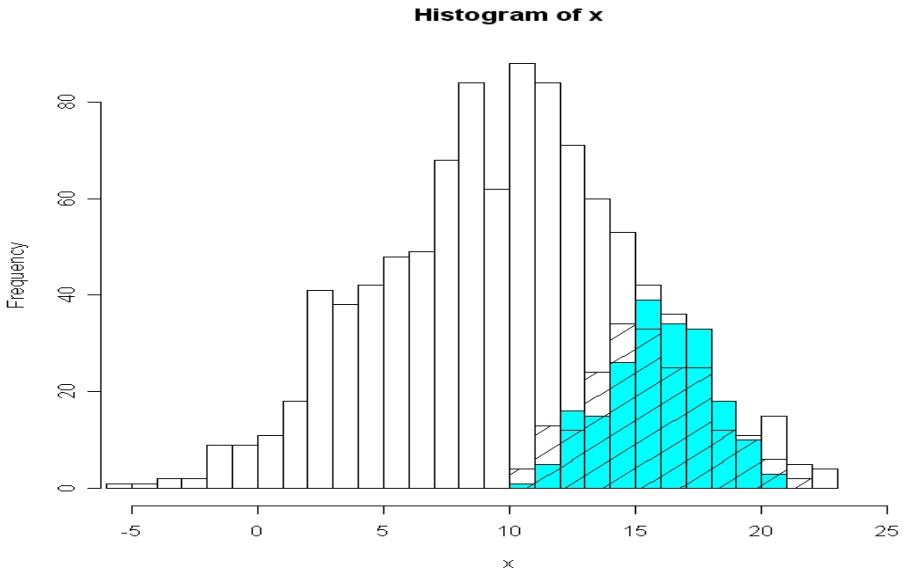
Unobserved Factor

Compare outcomes of similar subjects where the only difference is treatment; discards the rest of the subjects.

4. PSM Example: Matching

Matching to balance the covariate distribution
- to make the treated and control subject look alike before treatment
- to produce a regime which resembles a randomized experiment most, in terms of the observed covariates

Select 200 subjects in the control group, which resemble the treated most
- covariate X of subjects in treatment and non-treatment group have similar means
- comparison only made with matched subgroup



4. PSM Example: Standard Regression

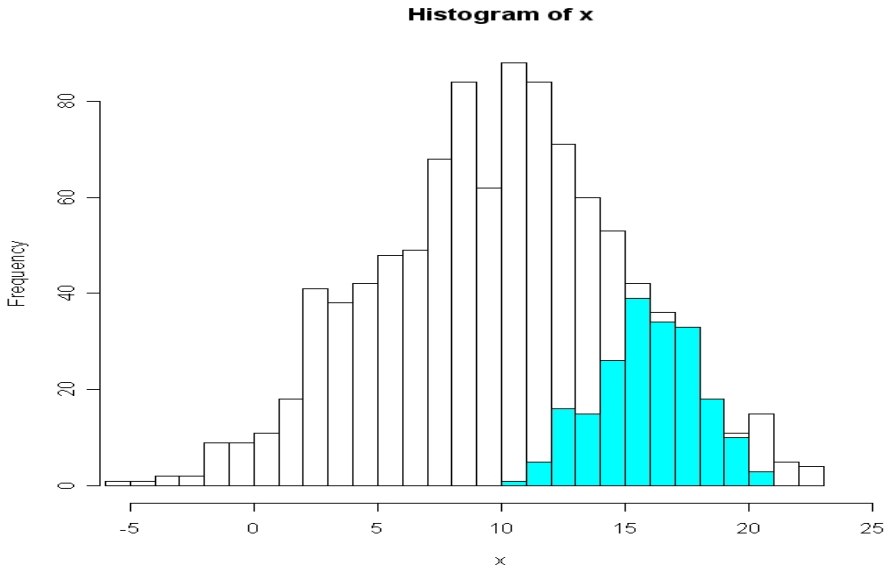
T: treatment indicator (1,0)

X: covariate, normally distributed

T=1, N(16,4), n1=200

T=0, N(10,25), n0=1000

Y: outcome



4. PSM Example: Standard Regression

Assumes subjects randomly selected into treatment!

$$Y_1 = \beta_0 + \beta_1 \times T + \beta_2 \times X + \varepsilon$$

	Estimate	Std. Error	t-value	Pr(>|t|)
T	11.1549	0.4294	25.98	<2e-16 ***

Overestimates the treatment effect: Subjects with high covariate X tend to select treatment

```
> ttest(x1,x0)

t = 27.6809, df = 745.829, p-value < 2.2e-16

95 percent confidence interval:
 5.515863 6.357964

sample estimates:
mean of x mean of y
15.885911  9.948997
```

4. Propensity Score Matching

1. Estimate propensity score
   - the likelihood/propensity of an individual being selected in the treatment
   - often done via logistic regression

2. Match subjects with similar propensity score
   - matching algorithm (e.g., nearest neighbour) iteratively finds the pair of subjects with the shortest distance. The goal is to balance the pretreatment covariates distribution

3. Evaluate quality of matching
   - check if the treatment and comparison group are similar across observable characteristics

4. Evaluate outcomes
   - based on the treatment and the matched comparison group

Empirical Evidence: A Study on Ad Lift

Gordon, Brett, et al. "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook." (Marketing Science, 2019)

We examine how common techniques used to *measure the causal impact* of ad exposures on users' conversion outcomes compare to the "gold standard" of a true experiment (*randomized controlled trial*). Using data from 12 US advertising lift studies at Facebook comprising 435 million user-study observations and 1.4 billion total impressions we *contrast the experimental results* to those obtained from *observational methods*, such as comparing exposed to unexposed users, matching methods, model-based adjustments, synthetic matched-markets tests, and before-after tests.

- We show that observational methods often fail to produce the same results as true experiments even after conditioning on information from thousands of behavioral variables and using non-linear models.
- Our findings suggest that common approaches used to measure advertising effectiveness in industry fail to measure accurately the true effect of ads.