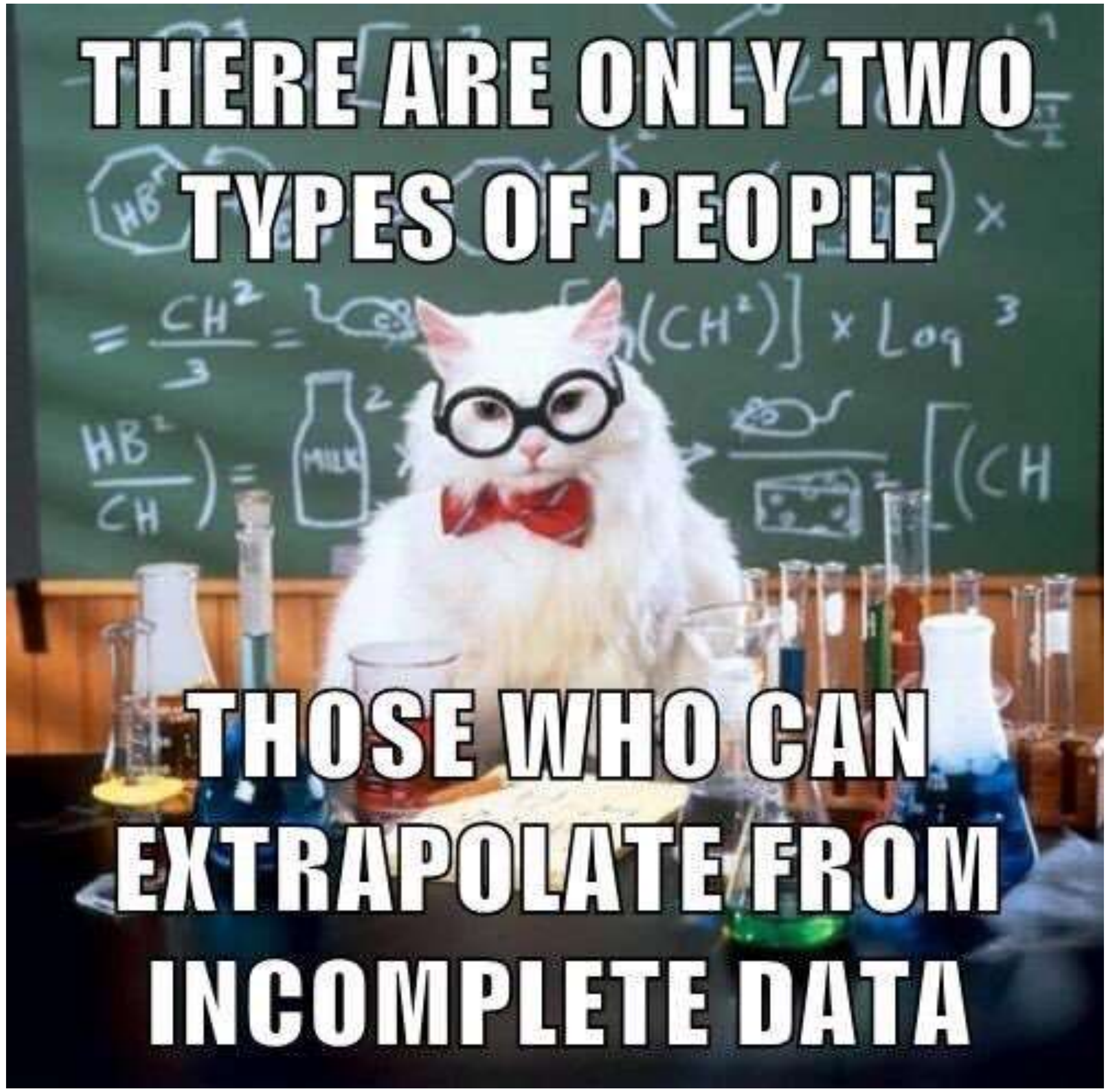
CORRELAID

GOOD CAUSES. BETTER EFFECTS.

CorrelAid MeetUp

30.11.2019 - Berlin



Missing Data & Causality

Questions to ask

Observational Questions *(What is?)*
„What if we see A“

$P(y|A)$

Action Questions *(What if?)*
„What if we do A?“

$P(y|do(A))$

Counterfactual Questions *(Why?)*
„What if we did things differently?“

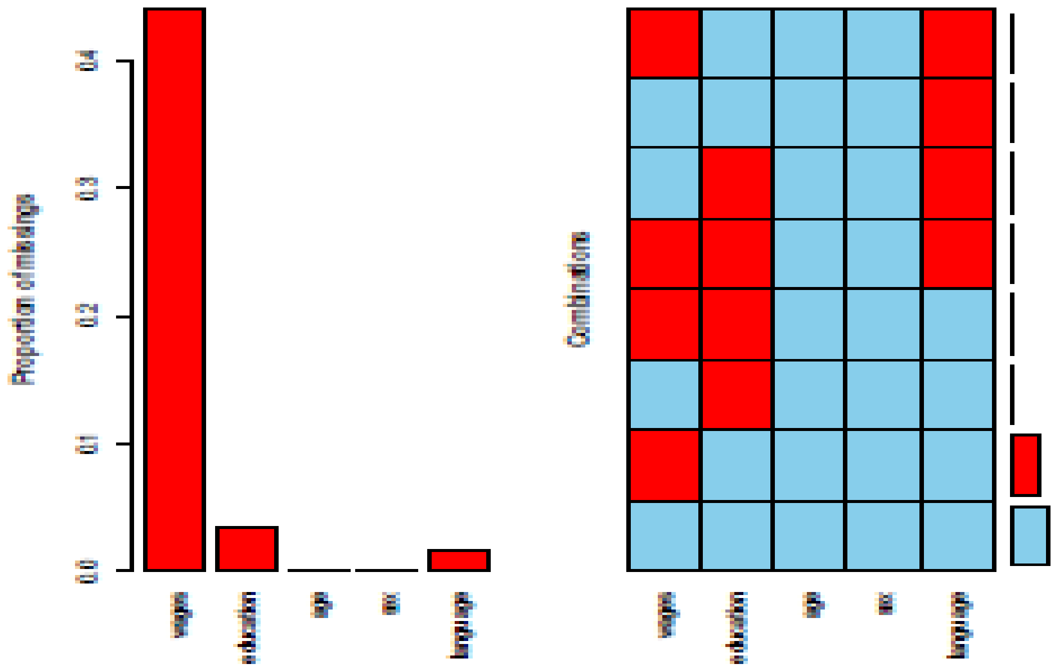
$P(y_{\neg A}|A)$

Options
„With what propability?“

Causal Hierarchy

Structure

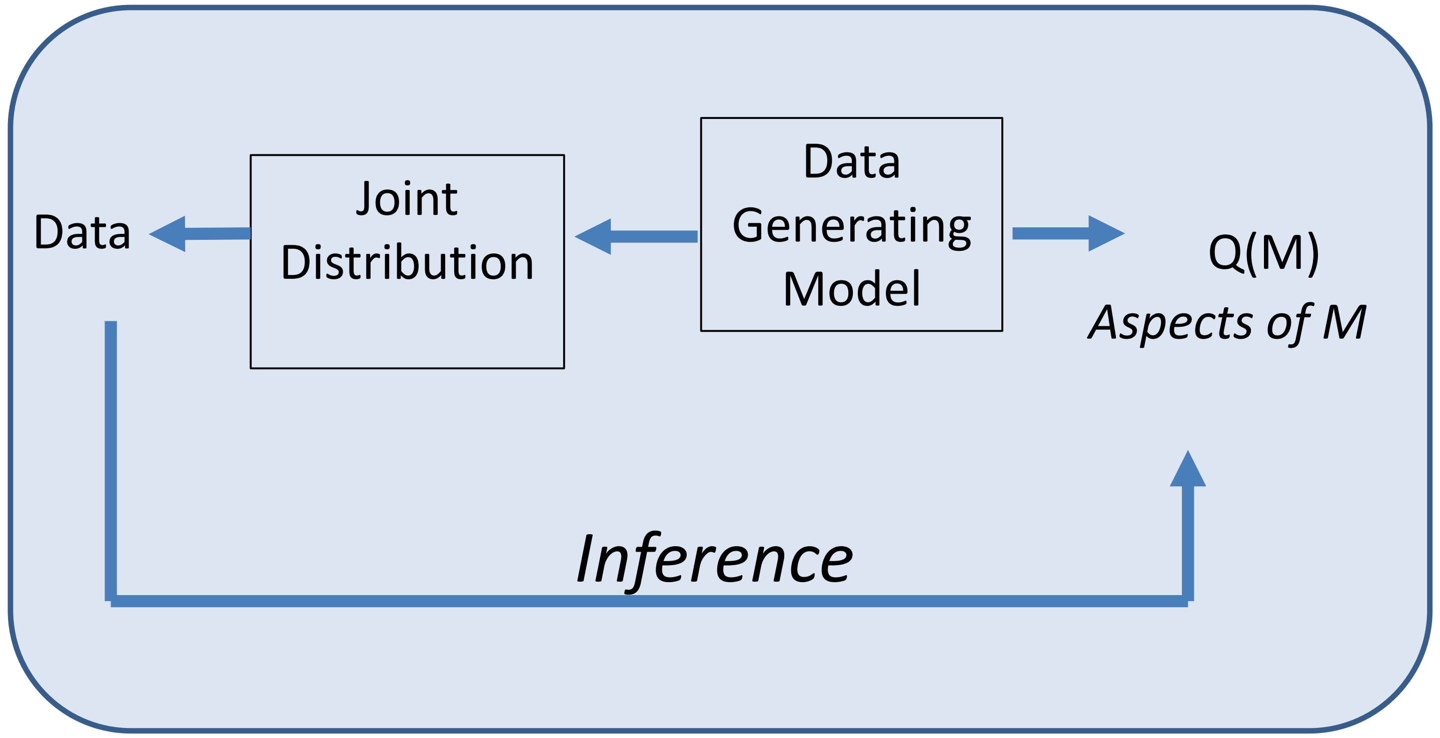
What is? / What if?



Missing Data:

- Missingness as a function
- Patterns vs. Mechanisms
- Naive Imputation
- Multiple Imputation

What if? / Why?



Causality:

- Data Generating Models
- Acyclic Graphical Models
- PC Algorithm

Data as a function

- Let Y be a population with the values $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ with $i \epsilon Y$
- The propability to draw a unit i, is $f(y_i; \theta)$ with θ indicating the parameter vector
- The observed part of Y can be denoted as Y_{obs} and the missing one as Y_{mis}, so that $Y=(Y_{obs,}\ Y_{mis})$ with the joint distribution $f(y_{obs}, i, y_{mis}; \theta)$ for each unit

Missingness as a function

- Let R be an **indicator variable** with unit i which is $f(r;\xi)$ distributed

$$R_{i,j} = \begin{cases} 1, \text{ if } Y \text{ observed with unit } i \\ 0 \text{ otherwise} \end{cases}$$

- The **joint distribution** of R and Y is

$$f(y, r, \theta, \xi) = f(y;\theta) f(r|y;\xi,\theta), (\theta,\xi) \epsilon \Omega_{\theta,\xi}$$

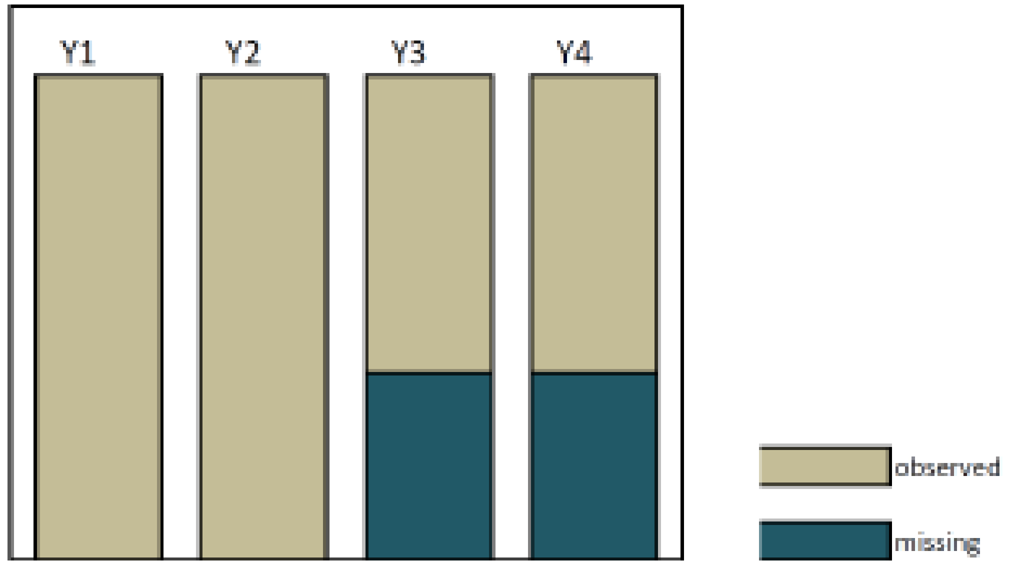
- The **likelihood** of a unit $\xi \epsilon \Omega$ is

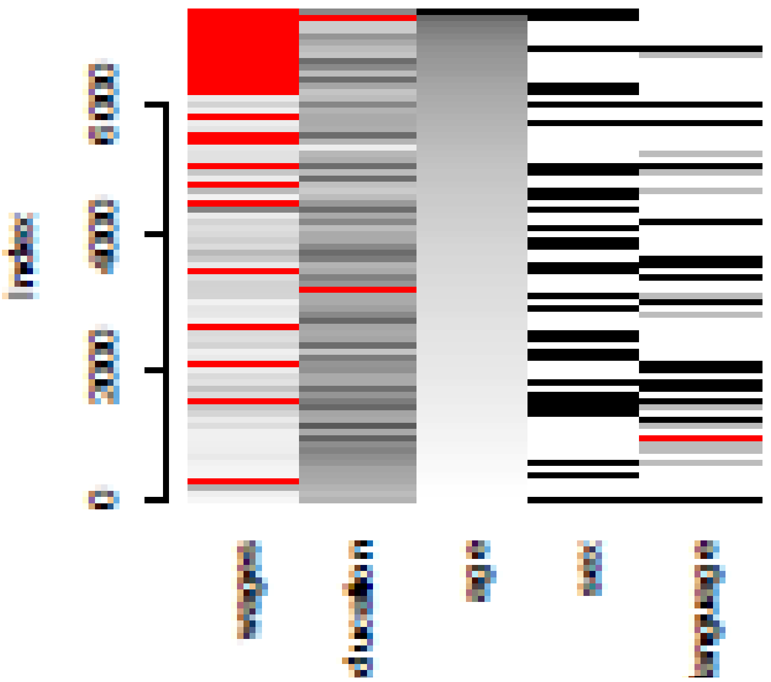
$$L(\theta,\xi; y_{obs}, r) = f(y_{obs}, y_{mis}, r; \theta, \xi) dy_{mis}$$

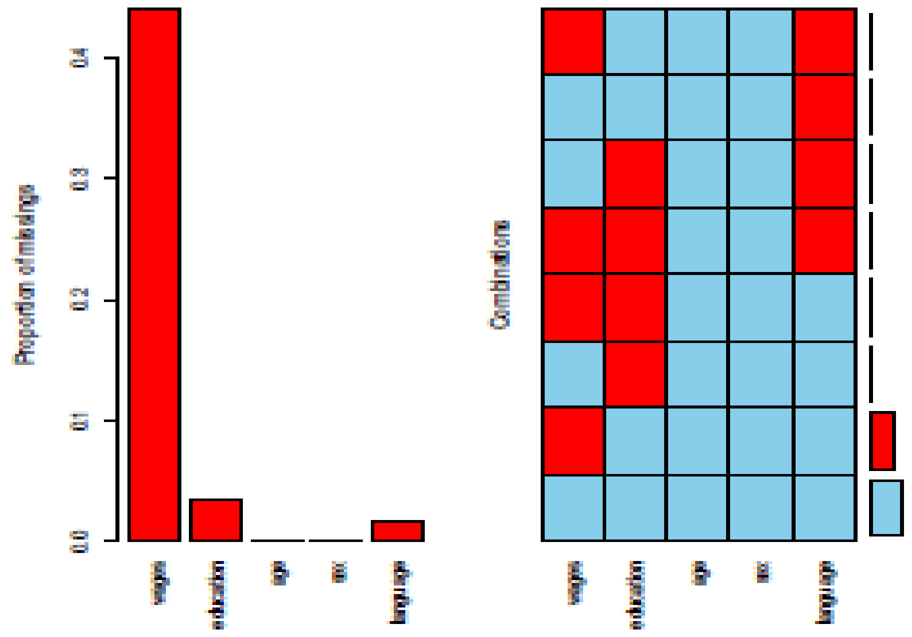
$$\int f(y_{obs}, y_{mis}; \theta) f(r|y_{obs}, y_{mis}; \theta) f(r|y_{obs}, y_{mis}; \xi, \theta) dy_{mis}$$

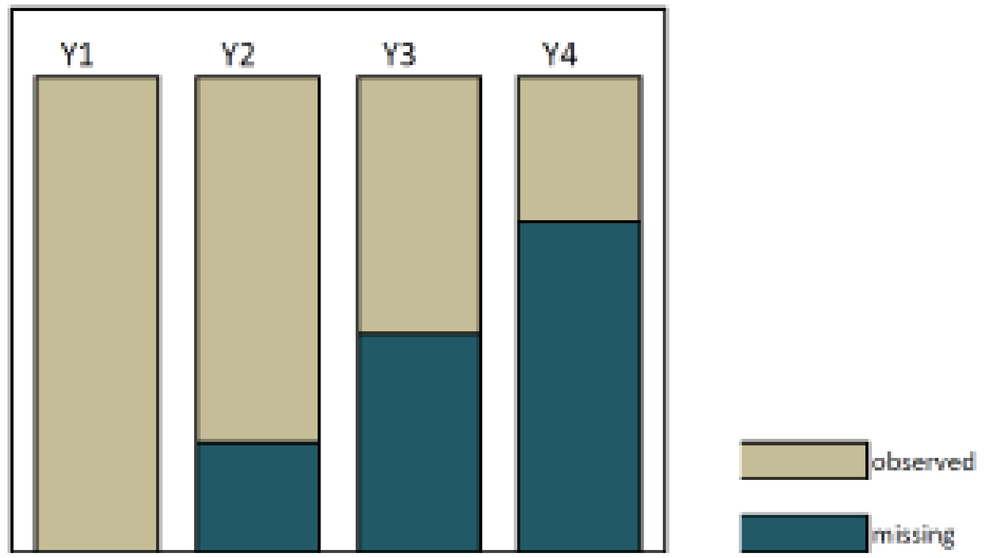
Index	Y1	Y2	R1	R2
1	y_{11}	y_{12}	1	1
2	y_{21}	y_{22}	1	1
3	y_{31}		1	0
4	y_{41}		1	0
5			0	0

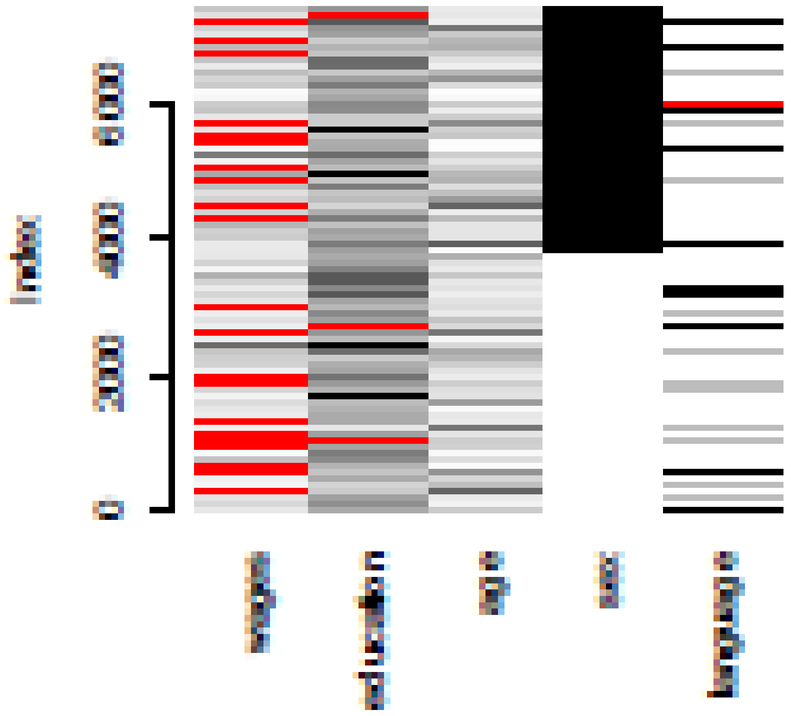
Patterns vs. Structures











A ***pattern*** *can be very non-random, but the* ***mechanism*** *might still be random!*

Notation for Missing Data mechanisms

Not Missing at Random (NMAR)

$f(r|y;\xi) = f(r|y_{obs}, y_{mis};\xi)$

Missing at Random (MAR)

$f(r|y;\xi) = f(r|y_{obs}, y_{mis};\xi) = f(r|y_{obs};\xi) \forall Y_{mis}$

Missing Completely at Random (MCAR)

$f(r|y;\xi) = f(r|y_{obs}, y_{mis};\xi) = f(r;\xi) \forall Y$

Assumption:

Variation freedom: θ and ξ are distinct/independent → $f(r|y;\xi,\theta) = f(r|y;\xi)$

Under **MAR** and variation freedom, the missing data mechanism is ignorable since the „observed-data“ likelihood of any unit resolves to:

$$L(\theta,\xi;y_{obs},r) = \int f(y_{obs},y_{mis};\theta) f(r|y_{obs},y_{mis};\theta,\xi) dy_{mis} = f(r|y_{obs};\xi) \int f(y;\theta) dy_{mis} \quad , \theta \epsilon \Omega_\theta, \xi \epsilon \Omega_\xi$$

Recognizing Mechanisms

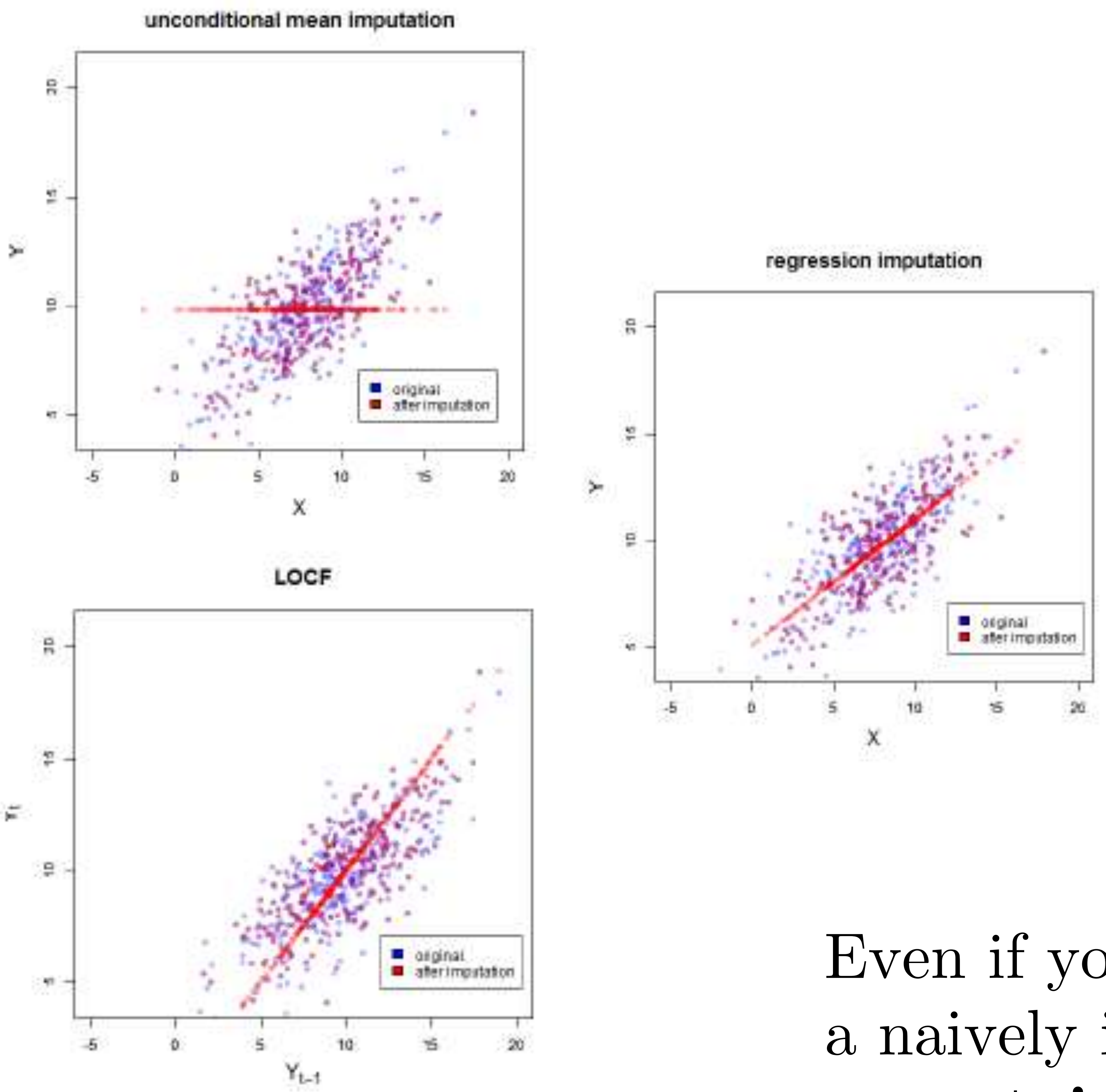
	Job performance ratings II			
IQ	Complete	MCAR	MAR	MNAR
78	9	–	–	9
84	13	13	–	13
84	10	–	–	10
85	8	8	–	–
87	7	7	–	–
91	7	7	7	–
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	–	7	–
99	7	7	7	–
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	–	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	–	12	12

MCAR:
Missingness neighter conditional on Y_{obs} nor Y_{mis} → random

MAR:
Missingness conditional on Y_{obs} but not Y_{mis}

MNAR:
Missingness conditional on Y_{mis}

To MI or not to MI...



When does it work?

- Crosstabs
- Correlations

→ Descriptive Statistics

Potential Pityfalls:

- Scewed Distributions
- Reduced variance / Confidence Intervals
- Hypothesis testing
- Model uncertainty

→ Inferential Statistics

Even if you‘re able to perform multivariate analysis with a naively imputed dataset you do not account for **model uncertainty.**

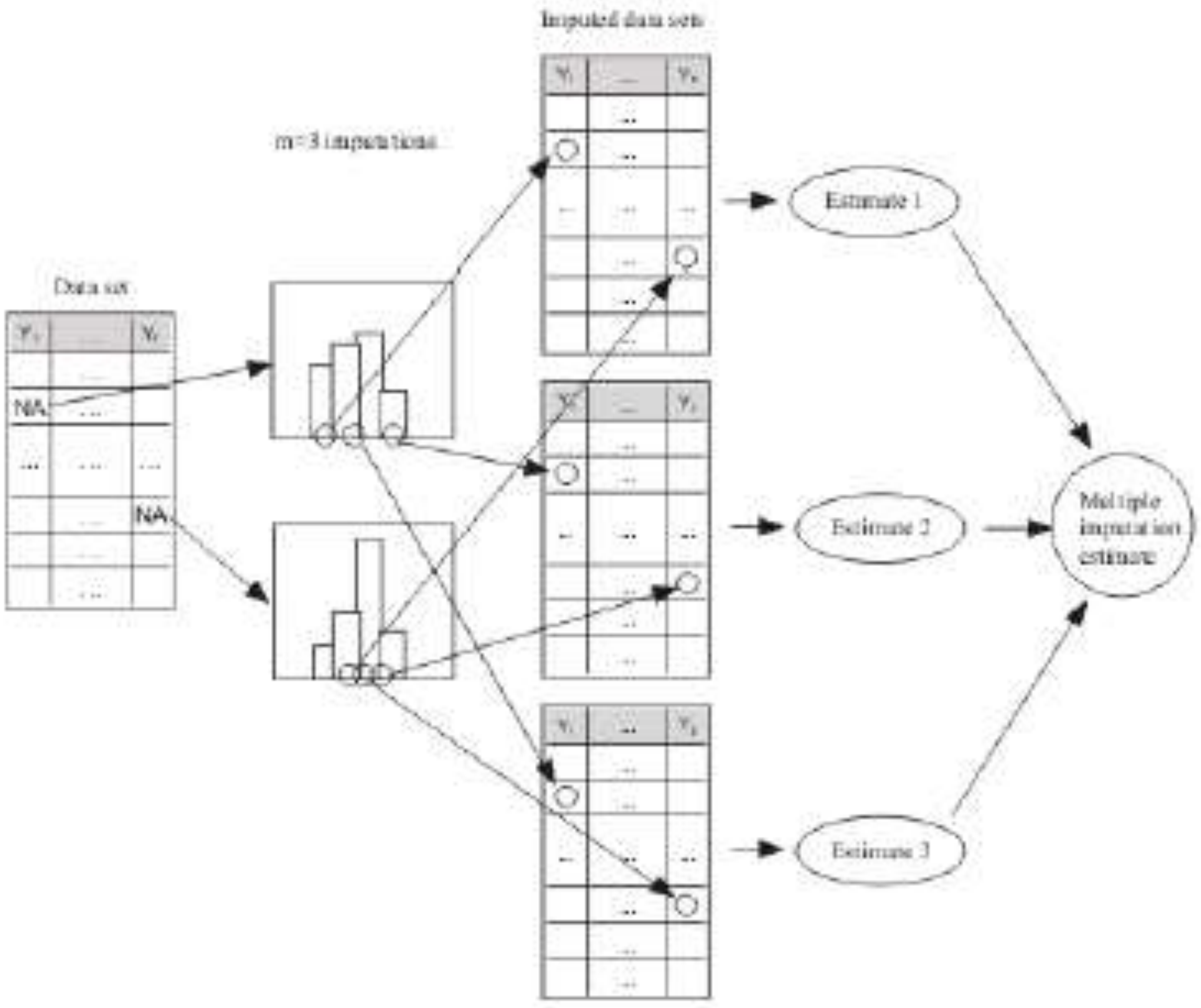
Ba(ye)sics of Multiple Imputation

Common to all approaches:

Produce several uniquely imputed datasets and calculate the mean.

Inferential MI:
→ Regression
→ Nearest Neighbour
→ Last Observation Carried Forward

Bayesian MI:
→ Draw from a Posterior Predictive Distribution



Ba(ye)sics of Multiple Imputation

$f(\theta)$ =: Prior distribution
$f(\theta|y)$ =: Posterior distribution
$f(y|\theta) = L(\theta;y)$ =: Likelihood
$f(y)$ =: Marginal distribution of the data

Posterior Predictive Distribution for MI:

$$f(y_{mis}|y_{obs}) = \int_{\Omega} f(\psi|y_{obs}) f(y_{mis}|\psi, y_{obs}) d\psi$$

Posterior-Step:

$$\tilde{\psi} \sim f(\psi|y_{obs})$$

Imputation-Step

$$y_{mis} \sim f(y_{mis}|\tilde{\psi}, y_{obs})$$

Interchanging draws from the observed-data posterior and the conditional predictive distribution yield random draws from $f(y_{mis}|y_{obs})$

Questions to ask

Observational Questions *(What is?)*
„What if we see A"

$P(y|A)$

Action Questions *(What if?)*
„What if we do A?"

$P(y|do(A))$

Counterfactual Questions *(Why?)*
„What if we did things differently?"

$P(y_{\neg A}|A)$

Options
„With what propability?"

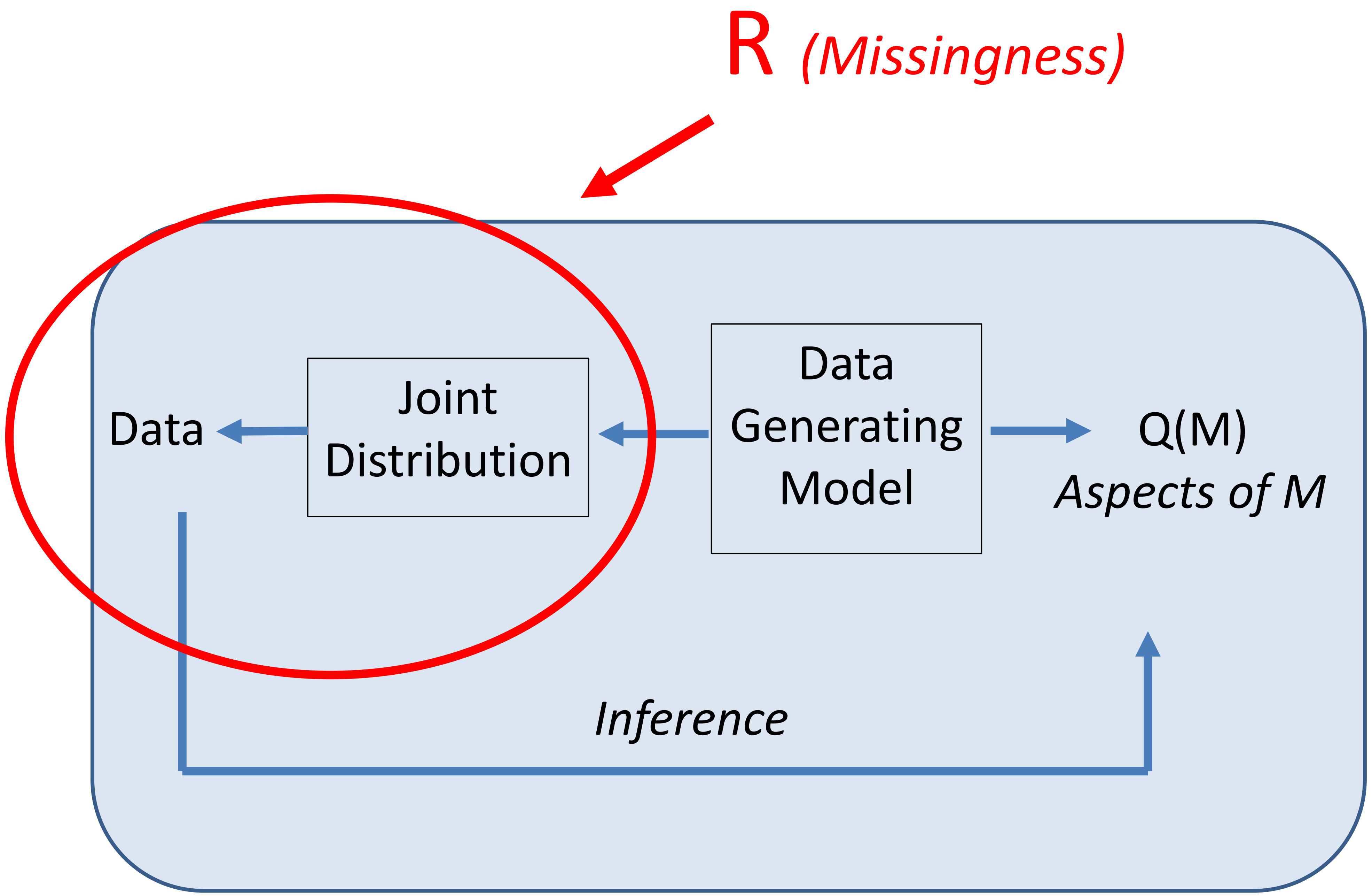
Causal Hierarchy

„Correlation does not imply causation.

–

But there is no correlation without causation.“

Basic Model of Causality



Fundamental laws of causal inference:

1. Law of counterfactuals

$Y_x(u) = Y_{M_x}(u)$

→ M generates and evaluates all counterfactuals

2. Law of Conditional Independence

$(X \perp\!\!\!\perp Y|Z)_{G(M)} \rightarrow (X \perp\!\!\!\perp Y|Z)_{P(V)}$

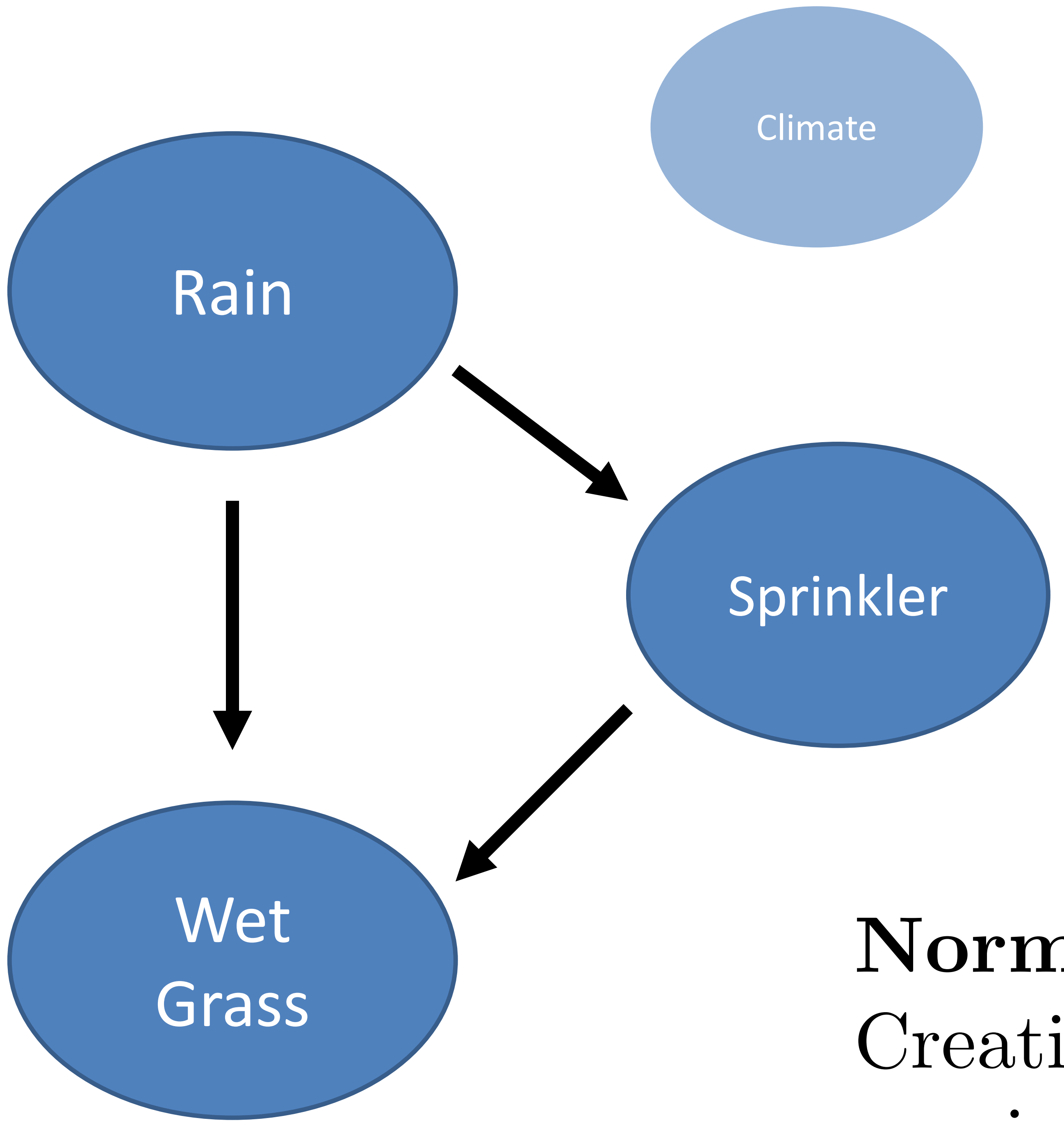
→ Separation in model entails independence in distribution

What do you think are good rules for choosing control variables?

Control für Z if...

- Z associated with Y (Outcome)
- Z associated with X (Covariate)
- Unaffected by X and associated with X and Y

Structural Causal Models



Acyclic Causal Graphs:

→ Show causal relationships between different variables
→ No circle-connections
→ Assuming you measured all relevant variables
→ There can be several markov-equivalent causal models

Normal Inference:
Creating one model and testing if it fits certain requirements
Causal Inference:
Searching for all plausible causal models

Constraint Based CI (PC-Algorithm)

Search for Markov Equivalent Classes (same d-seperations, nodes)

2 Main Steps:

1. Learn Skeleton Graph with undirected edges

→ Start with fully connected graph and test for conditional independence

2. Direct the edges to get causal model

→ Use of Additive Noisew Modelling *(lowest dependence between potential cause and ist residual when predicting the effect)*

Algorithm 1: Step 1 of the original-PC algorithm: learning the skeleton

Input: Dataset D with a set of variables $\mathbf{V}$, and significant level α
Output: The undirected graph G with a set of edges $\mathbf{E}$
Assume all nodes are connected innitially
Let depth $d = 0$
repeat
  for *each ordered pair of adjacent vertices X and Y in G* **do**
    if $(|adj(X,G)\backslash\{Y\}| >= d)$ **then**
      for *each subset $Z \subseteq adj(X,G)\backslash\{Y\}$ and $|Z| = d$* **do**
        Test $I(X,Y|Z)$
        if $I(X,Y|Z)$ **then**
          Remove edge between X and Y
          Save Z as the separating set of (X,Y)
          Update G and $\mathbf{E}$
          break
        end
      end
    end
  end
  Let $d = d + 1$
until $|adj(X,G)\backslash\{Y\}| < d$ *for every pair of adjacent vertices in G;*

CORRELAID

For instant updates on our work follow us on Twitter

@CorrelAid

Receive the latest infos on our projects via Facebook

@WeAreCorrelAid

Sign up for our mailing list on

www.correlaid.org