Causality in Biomedicine Lecture Series: Lecture 1

Ava Khamseh

22 Jan, 2020

Overview of the course

- Estimating causal effects
- Randomised trial vs observational data
- **Causal inference (of effects)** [DoWhy and others]
    - Rubin: Potential outcomes framework (observed confounders)
    - Rubin (unobserved confounders)
    - Simulations
    - Pearl: Structural causal models framework (observed and unobserved confounders)
    - Simulations
- **Causal discovery**
    - Constraint-based algorithms
    - Score-based algorithms
    - ML techniques

For biologists, computational biologist, XDFs, …

References

- For biologists:
  - https://www.youtube.com/watch?v=W1C5lFLEG84&list=PL_onPhFCkVQimvhuSAFrC8VWLEyNygQR5
  - The Book of Why by Judea Pearl

- For researchers with quantitative backgrounds:
  - R. Guo et al., A Survey of Learning Causality with Data
  - Review of Causal Discovery methods Based on Graphical Models, Glymour et al.
  - Causality book by Judea Pearl
  - Elements of Causal Inference book by Jonas Peters et al.

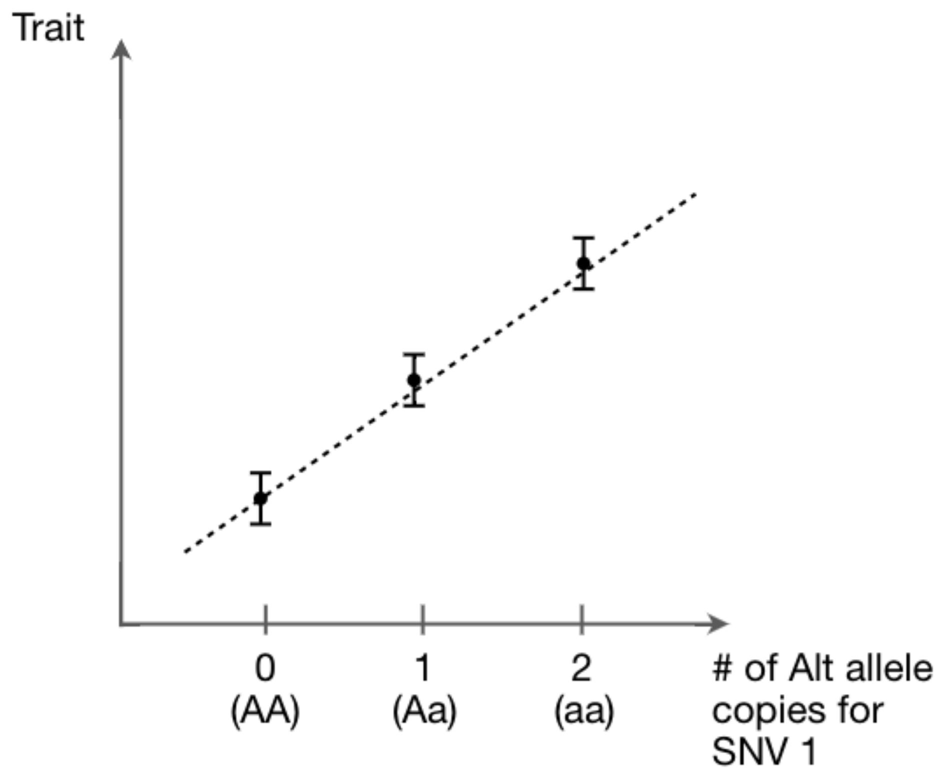
Outcomes of the course

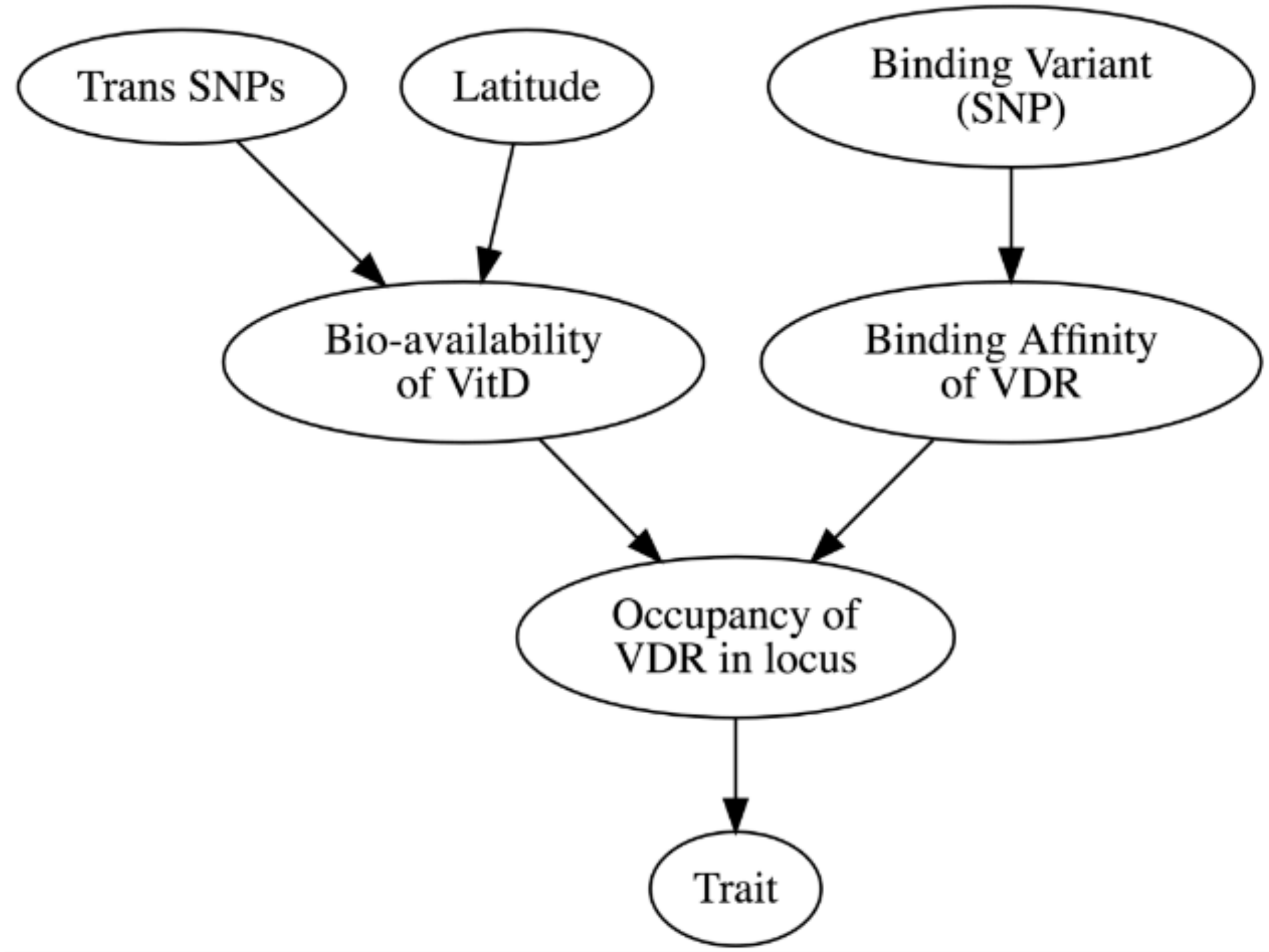
- Be able to find and follow papers that have developed causal techniques
- Understand which area of causal analysis the papers apply to
- Be able to apply causal techniques to a particular problem of interest
- Use causal analysis packages in R and Python (Microsoft DoWhy, CausalGraphicalModels)
- Be able to modify a current technique in such a way that applies to a particular problem of interest
- A foundation to start developing techniques in causal inference and causal discovery

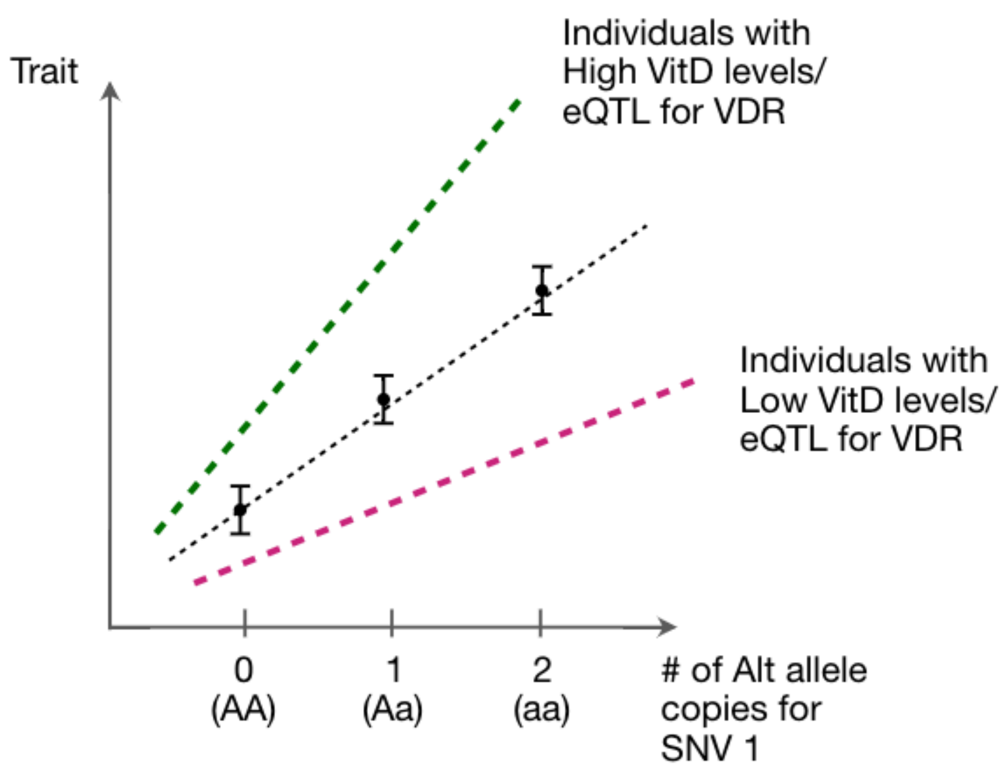
Biological Motivation I: Personalised Medicine

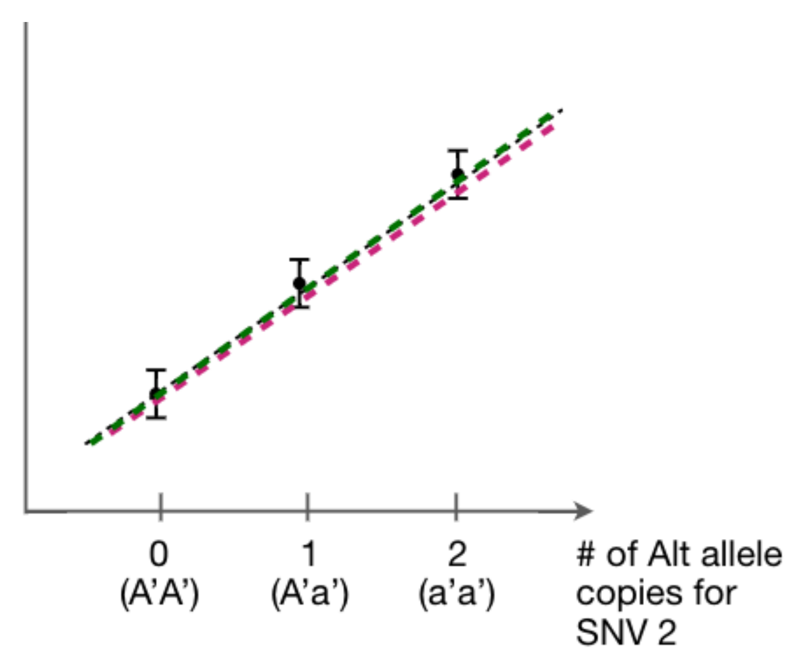
- Patient diagnosed with a particular **disease**
- Certain **baseline covariates** are known, e.g. age, weight, BMI, blood sugar, …
- **Question:** Should **treatment A** or **treatment B** be given
  - What is the causal effect of A vs B
  - Design a **policy**: Features —> {A,B}
  - i.e. best treatment for **a given individual**
- Source: **Electronic Health Records**

Biological Motivation II: Identifying causal SNPs



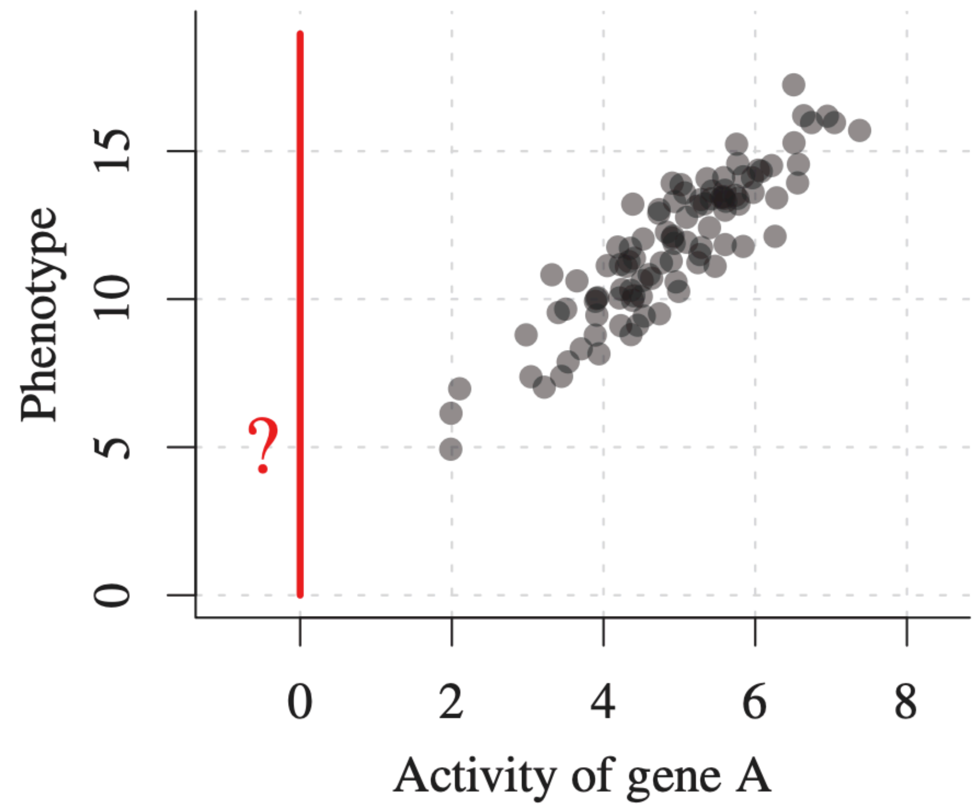


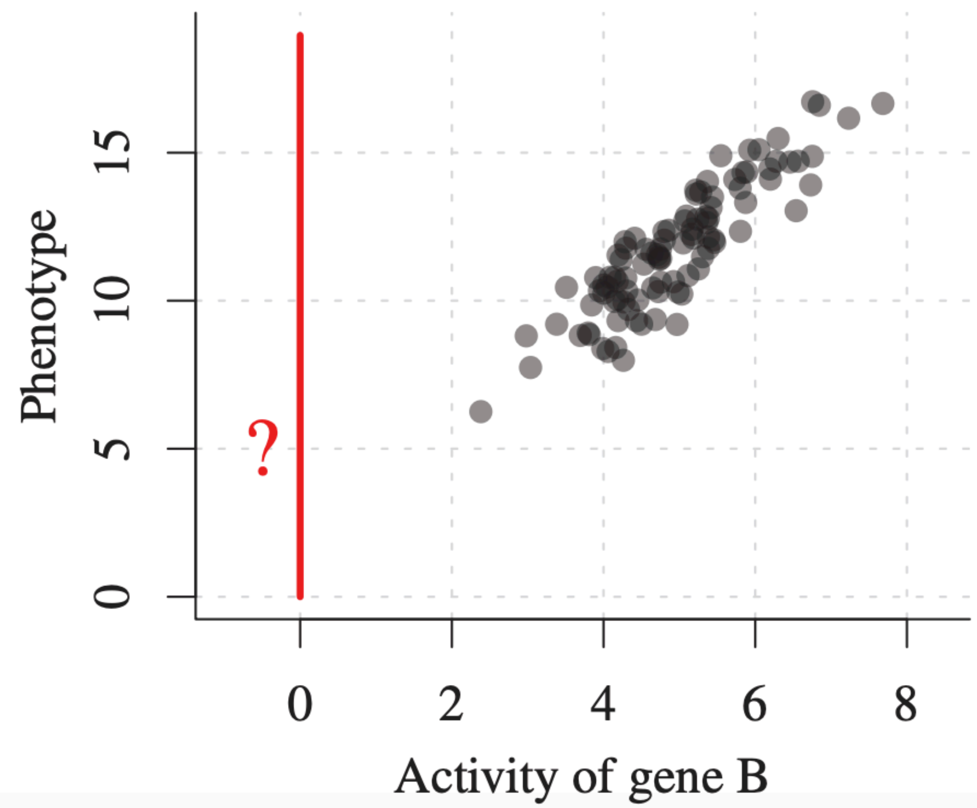




Neil Clark,
Chris Ponting

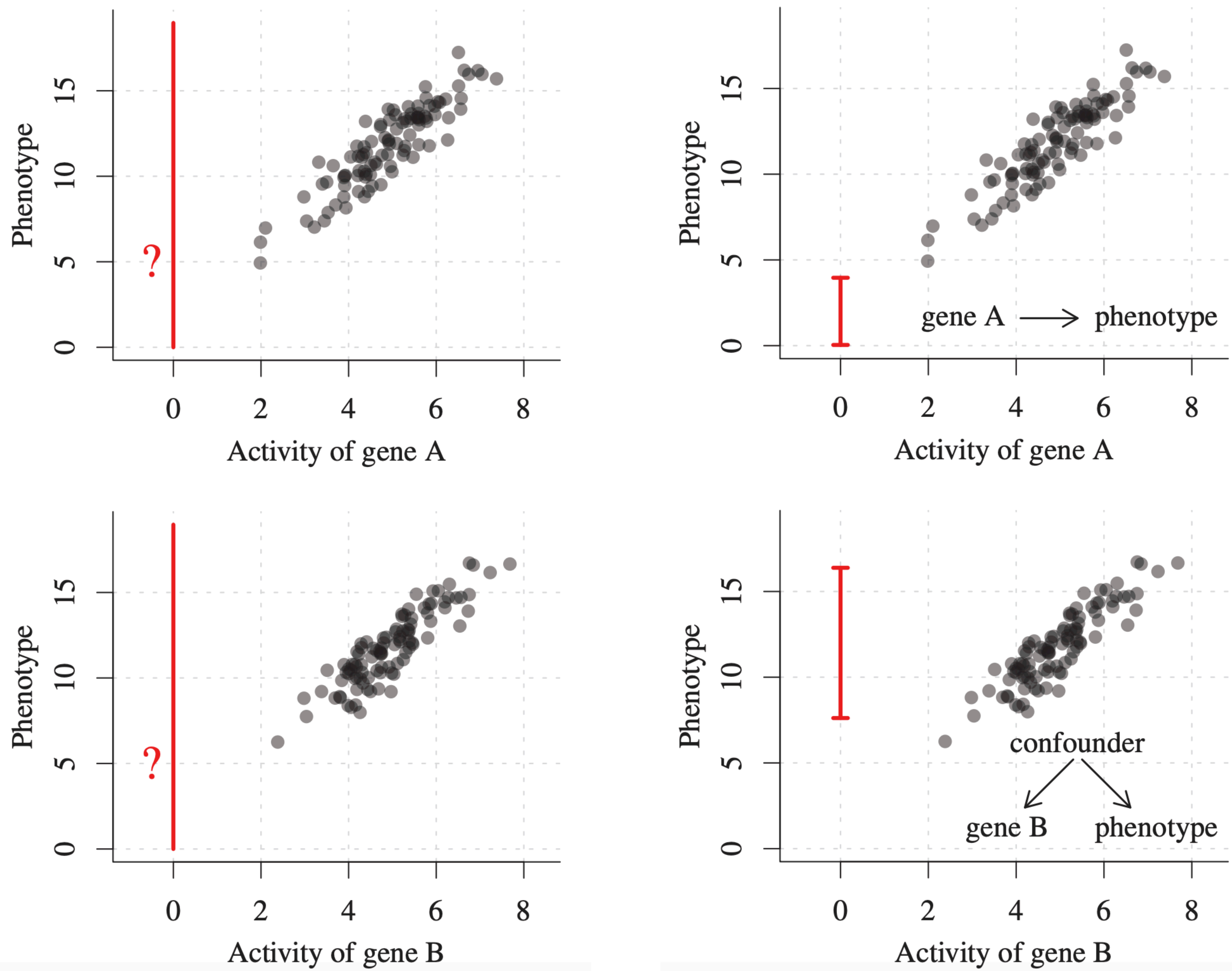
Biological Motivation III: Gene Perturbation





Elements of Causal Inference, Peters et al.

Biological Motivation III: Gene Perturbation

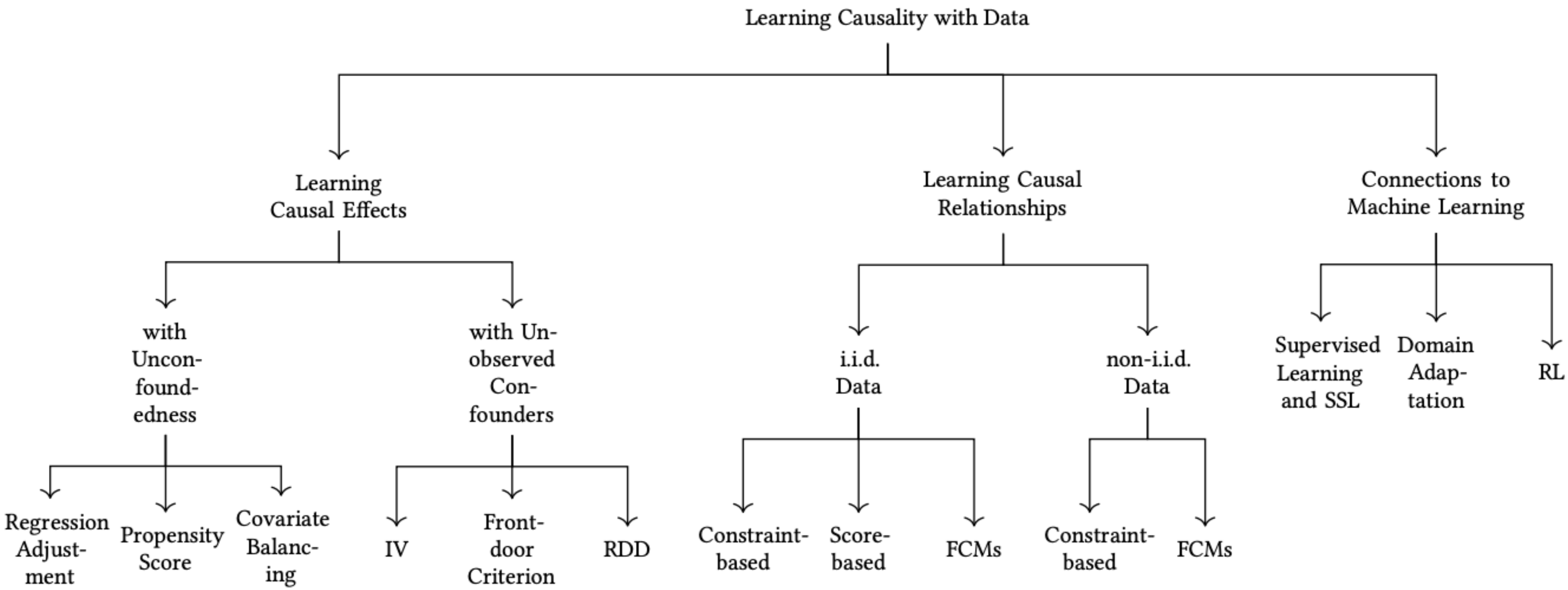


Elements of Causal Inference, Peters et al.

Causal Inference (of effects) vs Causal Discovery

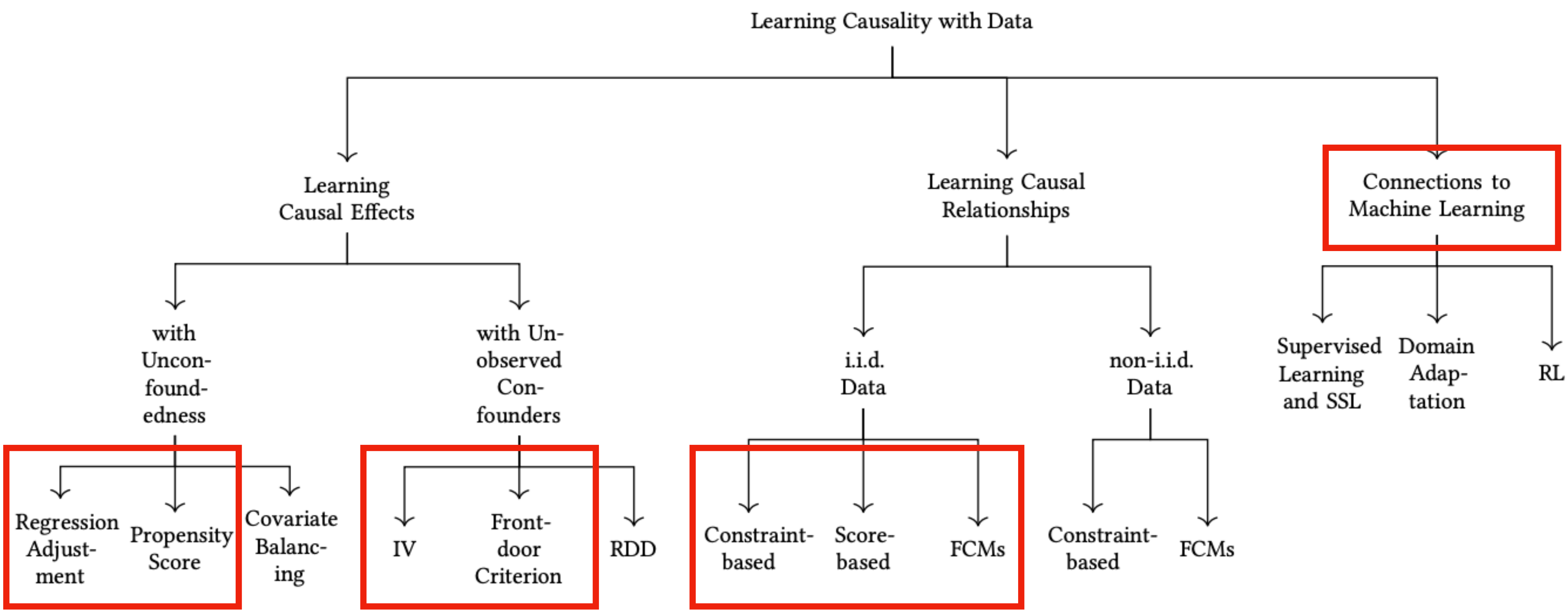
- **How much would some variables (features or labels) change if we manipulate the value of another variable?**
    - Have a prior causal knowledge (may be incomplete)
    - Wish to estimate degrees of causal dependencies
- **By modifying the value of which variables could we change the value of another variable?**
    - Wish to discover the causal graph
    - Apply causal inference

Overview of the field



R. Guo et al., A Survey of Learning Causality with Data

Overview of the field

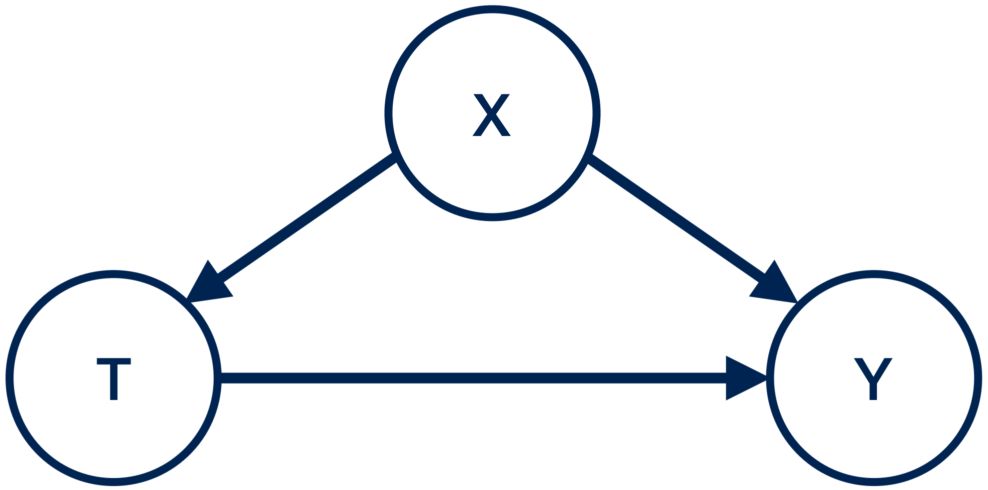


+ applications in biology

R. Guo et al., A Survey of Learning Causality with Data

Conventions

- Variable to be manipulated: **treatment (T)**, e.g. drug
- Variable we observe as response: **outcome (Y)**, e.g. success/ failure of drug
- Other observable variables that can affect treatment and outcome causally and we wish to correct for: **confounders (X)**, e.g. age, gender, …
- Unobservable confounder **(U)**



Causal Inference (of Effects)

- Have a prior causal knowledge (may be incomplete) and know the treatment/outcome pair, c.e., weight gain, hours online

- Interested in estimating the **effect size**:

$$\mathbb{E}\big[y_{t=1}(x) - y_{t=0}(x)\big] = \int \big(y_1(x) - y_0(x)\big)p(x)dx$$

Note: The features/confounders x for both treatment and control groups are drawn from the **<u>same</u>** distribution p(x)

- Goal: Find an **unbiased estimator**, e.g. signal/noise ratio

Randomised experiments: Already in causal framework

- In a **randomised experiment**, p(x) is designed to be the same for both treatment groups (t=0 or t=1), typically uniform
- Paired '**clones**' in treatment and outcome groups
- Simply take the difference of the averages:

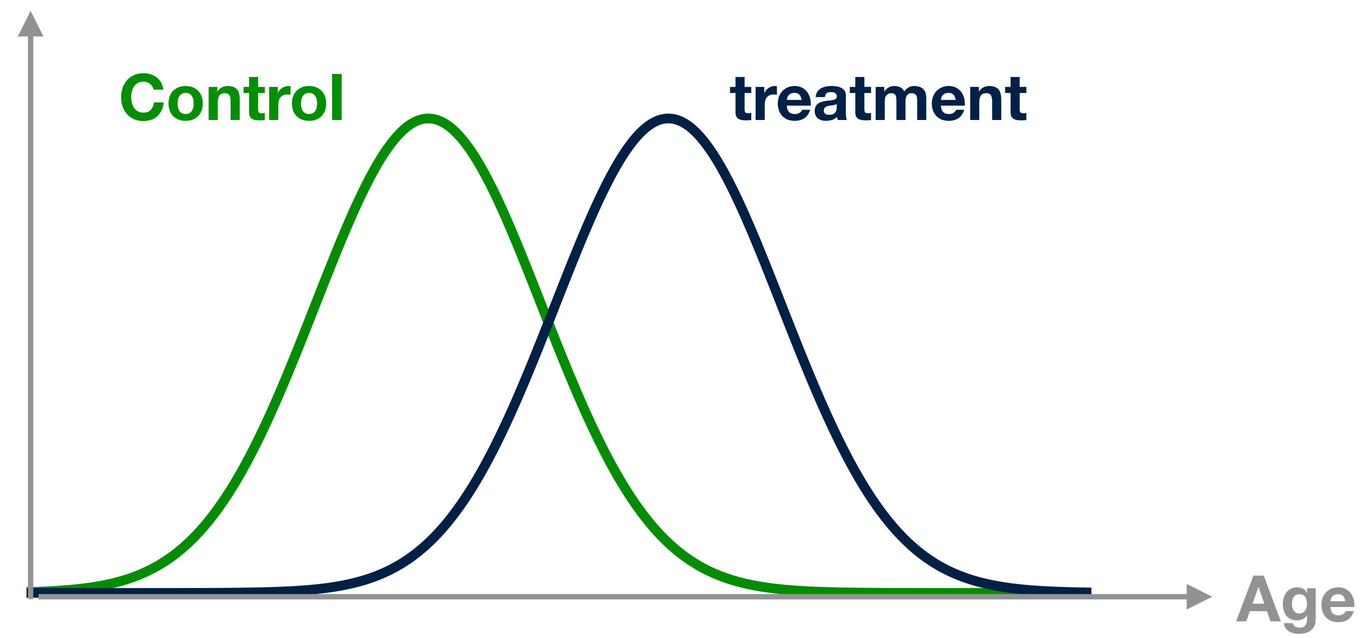
$$\Delta\hat{\mu} = \hat{\mathbb{E}}\big[y_{t=1}(x) - y_{t=0}(x)\big] = \frac{1}{N}\sum_{i=1}^{N}\left(y_1^{(i)}(x) - y_0^{(i)}(x)\right)$$

- Statistical test: e.g. T-test and p-values …

$$\frac{\Delta\hat{\mu}}{\sqrt{\frac{(\hat{\sigma_1})^2+(\hat{\sigma_0})^2}{N}}} > t^*$$

Observational data: What goes wrong?

$$p(x|t=1) \neq p(x|t=0)$$



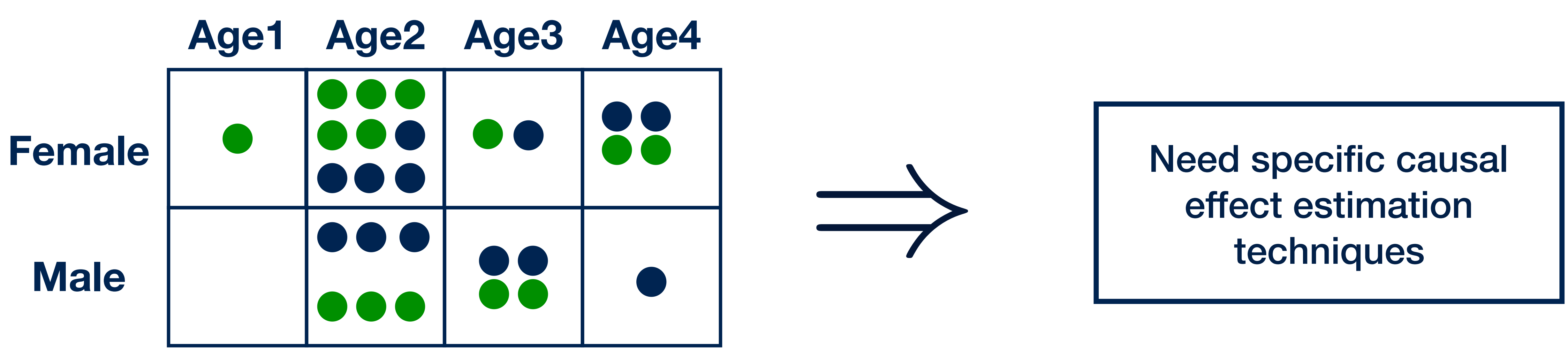
$$\left(\int y_1(x) p(x|t=1) dx \; - \int y_0(x) p(x|t=0) dx \right) \neq \int \big(y_1(x) - y_0(x)\big) p(x) dx$$

Observational data: Stratification

- Measure outcome (success/failure), **within** each of the young/old groups **separately**
- Take weighted average by the probability of being young/old

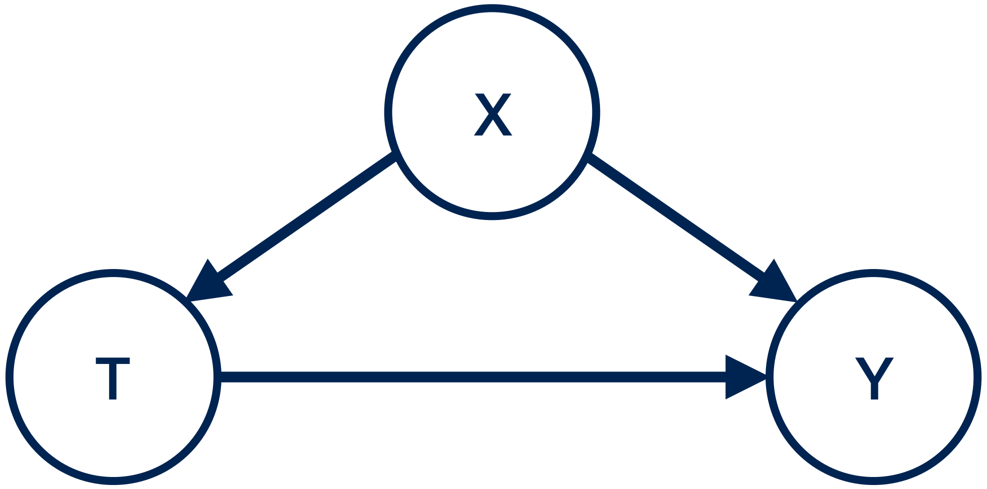
$$\mathbb{E}(\text{Healed}|t=1) = \mathbb{E}(\text{Healed}|t=1, \text{young})p(\text{young}) + \mathbb{E}(\text{Healed}|t=1, \text{old})p(\text{old})$$

- Disadvantages:
  - All possible confounders need to be observed
  - Assumes overlap between the two distributions (if there is no overlap, sample is not representative, e.g. performing the experiment only for old people)
  - Bad estimates as confounder dimensionality increases



Two main Frameworks for causal inference/discovery

- **Potential outcomes (Rubin):**
    - Requires a given treatment-outcome pair (known directionality)
    - Mainly applies to causal inference (learning effects)
    - More familiar to biologists



- **Structural causal models (Pearl):**
    - Causal graph
    - Structural equations
    - Algorithmic: Causal Discovery

Extend the language of probability theory: do-calculus

$$x = f_x(\epsilon_x),\ t = f_t(x, \epsilon_t),\ y = f_y(x, t, \epsilon_y)$$

Assumption: Independent noise terms: $\epsilon_x \perp\!\!\!\perp \epsilon_t \perp\!\!\!\perp \epsilon_y$

Potential Outcomes Framework (Rubin)

- **Definition:** Given treatment, t, and outcome, y, the **potential outcome** of instance/individual (i) is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual (i) had been under treatment t.

- $y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes
- $t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1
- **Observed** outcomes: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)}$$

- **Individual treatment effect**: $\tau^{(i)} = y_1^{(i)} - y_0^{(i)}$

- **Average treatment effect:**

$$\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N} \sum_{i=0}^{N} \left(y_1^{(i)} - y_0^{(i)} \right)$$

Potential Outcomes Assumptions (Rubin)

- **SUTVA:** Stable Unit Treatment Value Assumption
    - Well-defined treatment (no different versions)
    - No interference: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

- **Consistency:** The observed outcome is independent of how the treatment is assigned

- **Unconfoundedness (ignorability)**

Potential Outcomes Framework (Rubin)

- **Unconfoundedness**: Treatment assignment is random, given X:

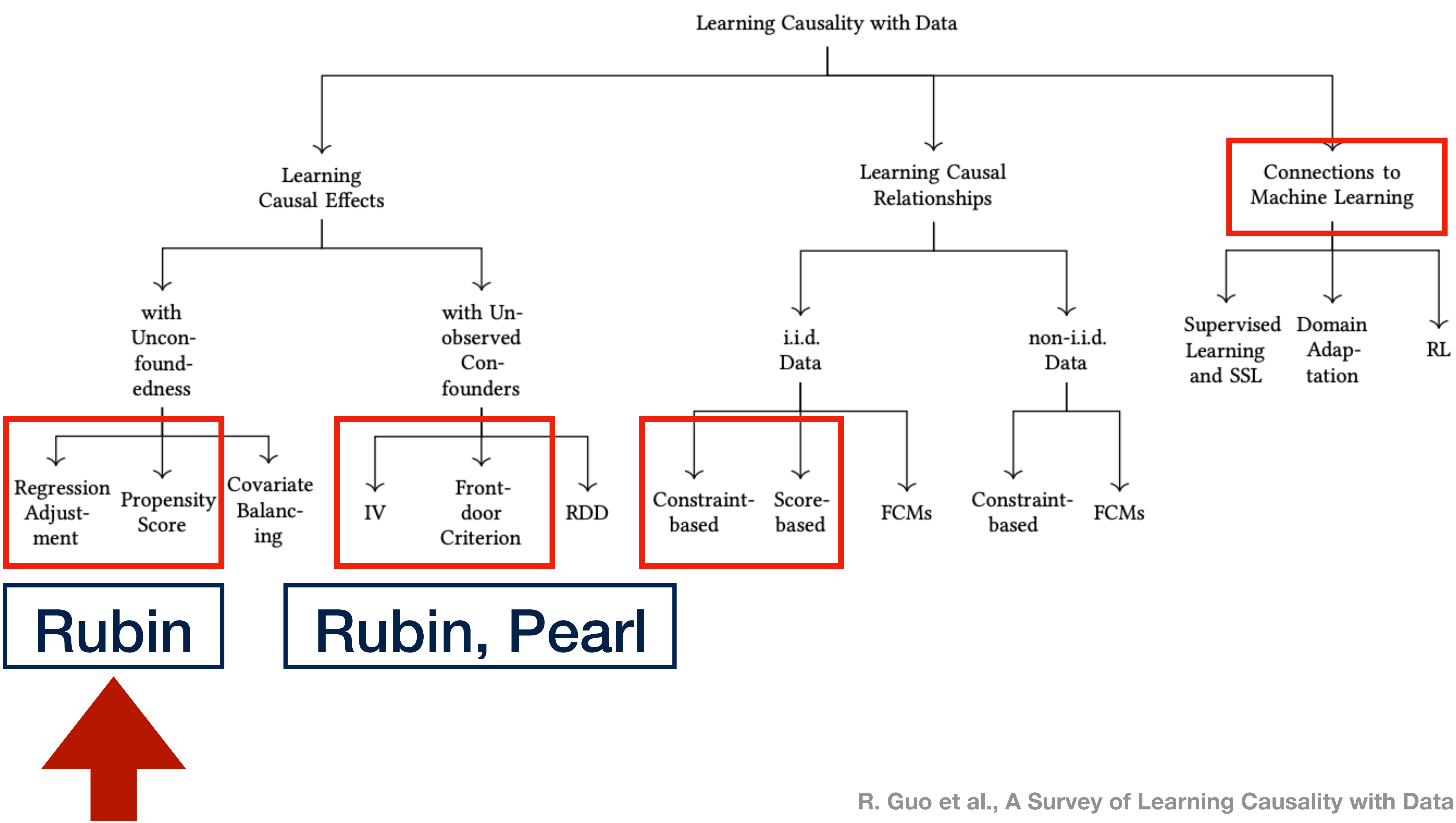
$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

- i.e. given X, individual (i) has no preference to get assigned to either of experiment or control groups
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person B
- There may be difference in power
- However, if we do not restrict to the old group, there is a clear preference: older individuals are more likely to receive the drug
- **No unobserved confounders** (see later: unverifiable in observational data)

- **Strong ignorability:** Every individual has a non-zero chance of receiving the treatment/control:

$$p(t=1|x) \in (0,1) \text{ if } P(x) > 0$$

Overview of the field



R. Guo et al., A Survey of Learning Causality with Data

Causality in Biomedicine Lecture Series: Lecture 1

Ava Khamseh

22 Jan, 2020