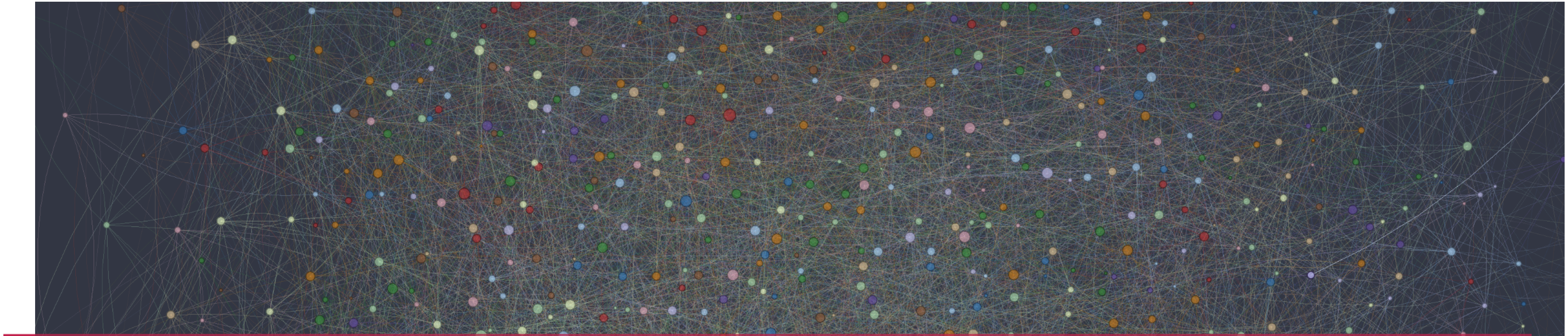
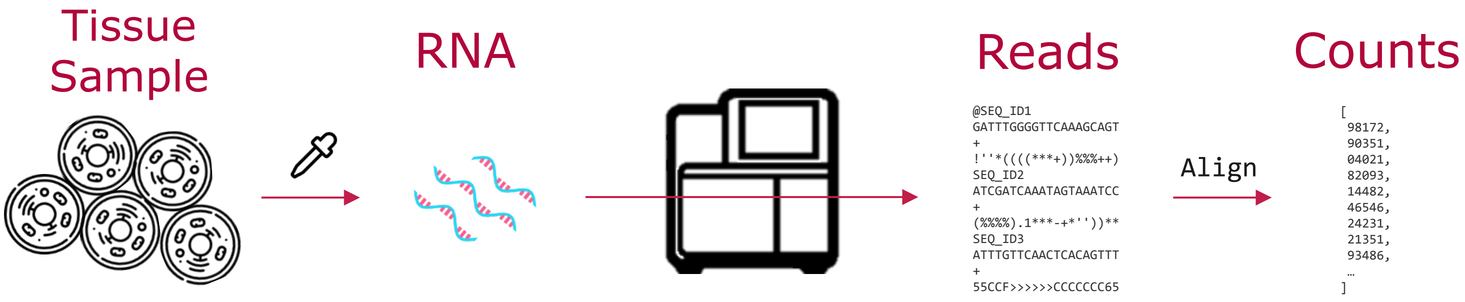


Trends in Bioinformatics: **Causal Inference on Gene Expression Data**

Philipp Bode

Causal Inference on **Gene Expression Data**: Snapshotting the Transcriptome[1]



Tissue Sample

RNA

Reads

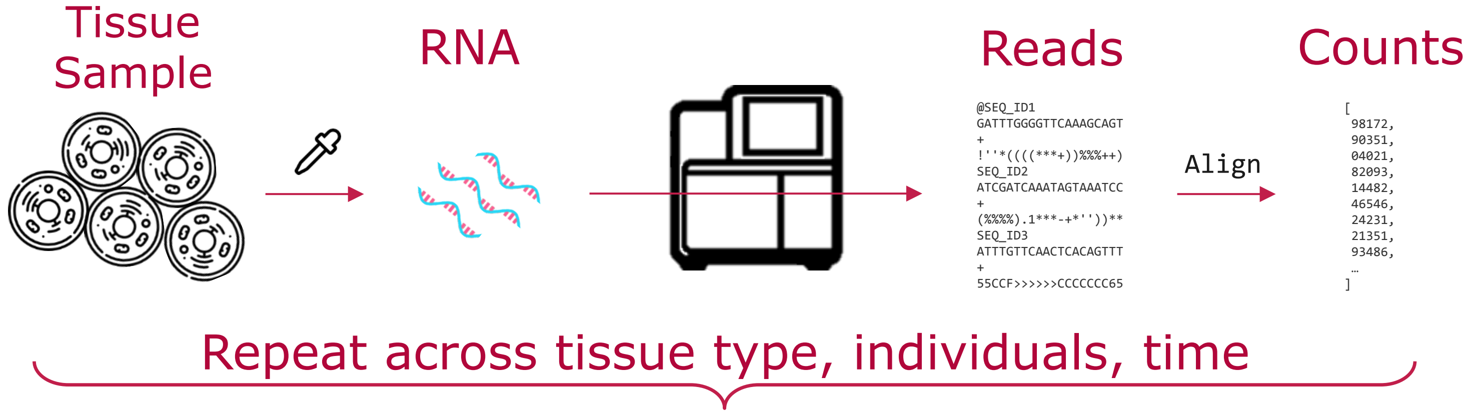
```
@SEQ_ID1
GATTTGGGGTTCAAAGCAGT
+
!''*((((***+))%%%++)
SEQ_ID2
ATCGATCAAATAGTAAATCC
+
(%%%%).1***-+*''))**
SEQ_ID3
ATTTGTTCAACTCACAGTTT
+
55CCF>>>>>>CCCCCCC65
```

Align

Counts

```
[
 98172,
 90351,
 04021,
 82093,
 14482,
 46546,
 24231,
 21351,
 93486,
 …
]
```

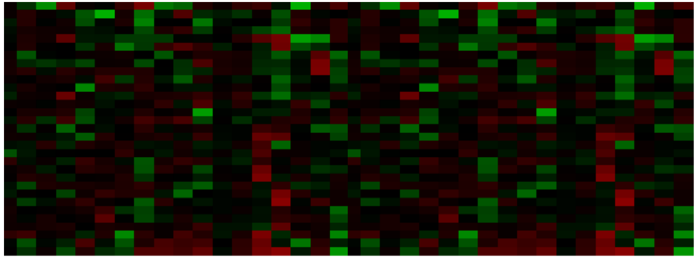
Causal Inference on **Gene Expression Data**: Snapshotting the Transcriptome[1]



```
@SEQ_ID1
GATTTGGGGTTCAAAGCAGT
+
!''*((((***+))%%%++)
SEQ_ID2
ATCGATCAAATAGTAAATCC
+
(%%%%).1***-+*''))**
SEQ_ID3
ATTTGTTCAACTCACAGTTT
+
55CCF>>>>>>CCCCCCC65
```

```
[
 98172,
 90351,
 04021,
 82093,
 14482,
 46546,
 24231,
 21351,
 93486,
 …
]
```

Differential Expressions



Differential Gene Expression Analysis

Motivation[2]

- Cellular functions heavily regulated by RNA expression
- Gain insights into processes in healthy and cancerous cells:
  - As biomarkers for prognostic or diagnostic evaluation
  - As potential drug targets

TCGA

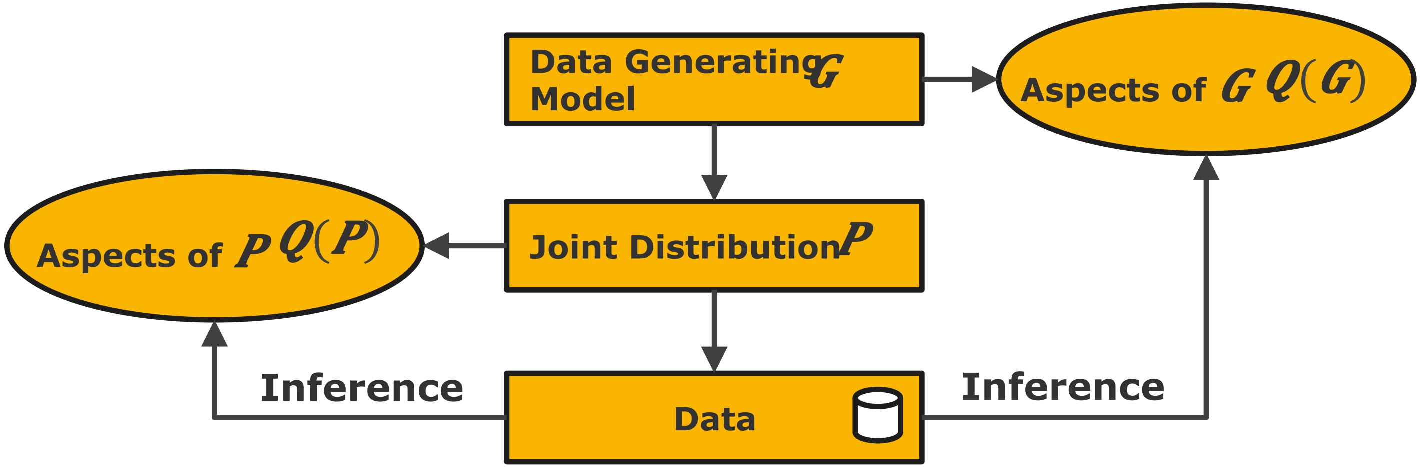
- Large steady-state observational RNA-seq data sets available

Causal Graphical Models

Motivation[3]

Traditional Statistical Inference Paradigm

Paradigm of Structural Causal Models



Data Generating Model G

Aspects of G $Q(G)$

Joint Distribution P

Aspects of P $Q(P)$

Data

Inference

Inference

E.g., is gene G2 higher expressed if **we see** that gene G1 is higher expressed?

$Q(P) = PExpression$

$G2$□$Expression\ G1$

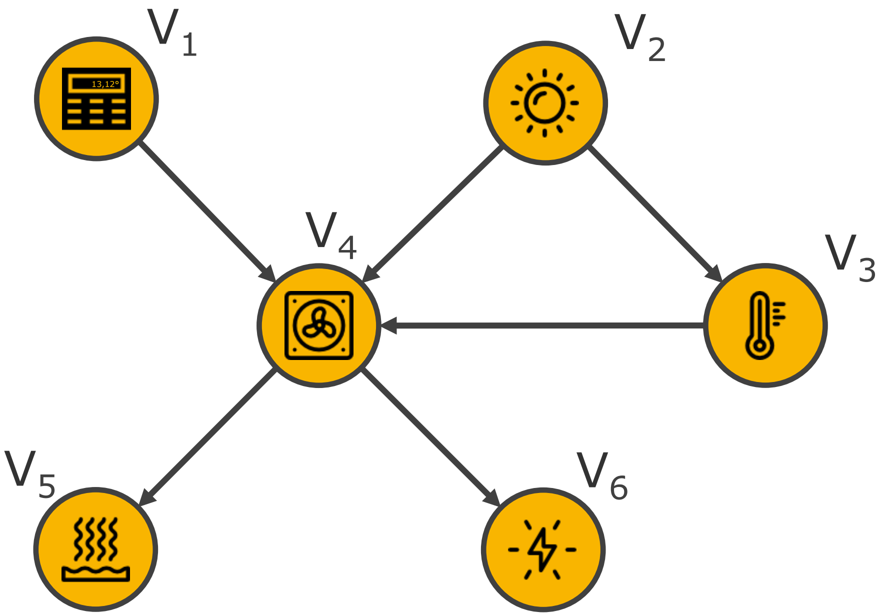
E.g., is gene G2 higher expressed if **we do** express gene G1 higher?

$Q(G) = PExpression$

$G2$□$do(Expression\ G1)$

Causal Inference on Gene Expression Data: Graphical Causal Models

Cooling House Example:



V_1: Target temp.

V_2: Sunlight level

V_3: Outside temp.

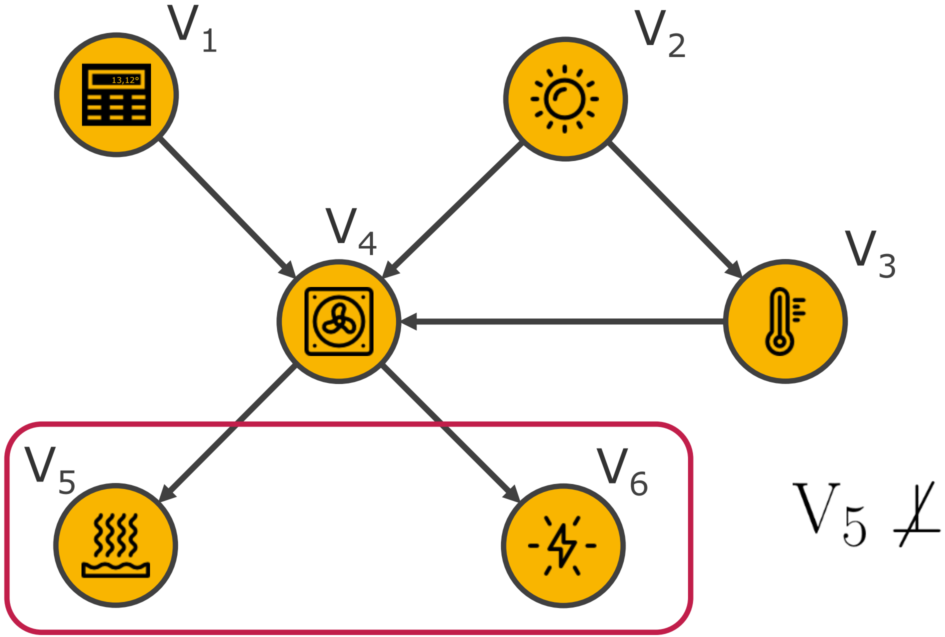
V_4: Cooling action

V_5: Thermal waste

V_6: Electricity consumption

Causal Inference on Gene Expression Data: Conditional Independence

Cooling House Example:



$$V_5 \not\perp V_6$$

V_1: Target temp.

V_2: Sunlight level

V_3: Outside temp.

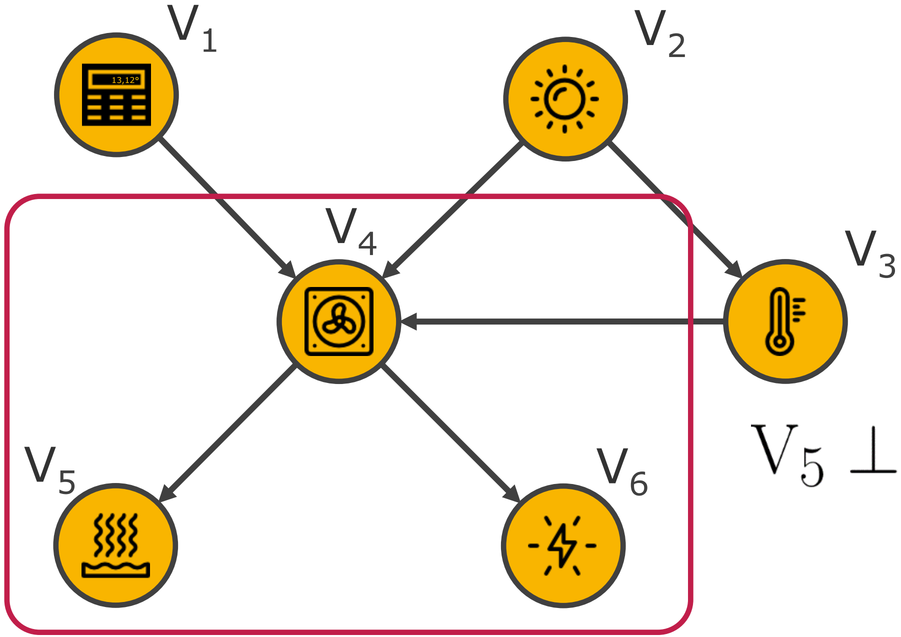
V_4: Cooling action

V_5: Thermal waste

V_6: Electricity consumption

Causal Inference on Gene Expression Data: Conditional Independence

Cooling House Example:



V_1: Target temp.

V_2: Sunlight level

V_3: Outside temp.

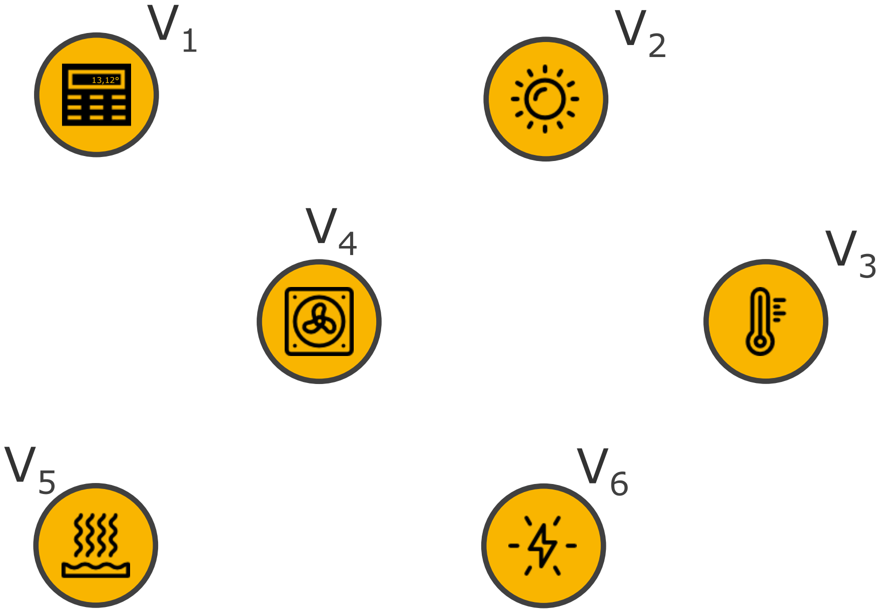
V_4: Cooling action

V_5: Thermal waste

V_6: Electricity consumption

$$V_5 \perp V_6 \mid V_4$$

Causal Inference on Gene Expression Data: The Peter-Clark algorithm



V_1: Target temp.

V_2: Sunlight level

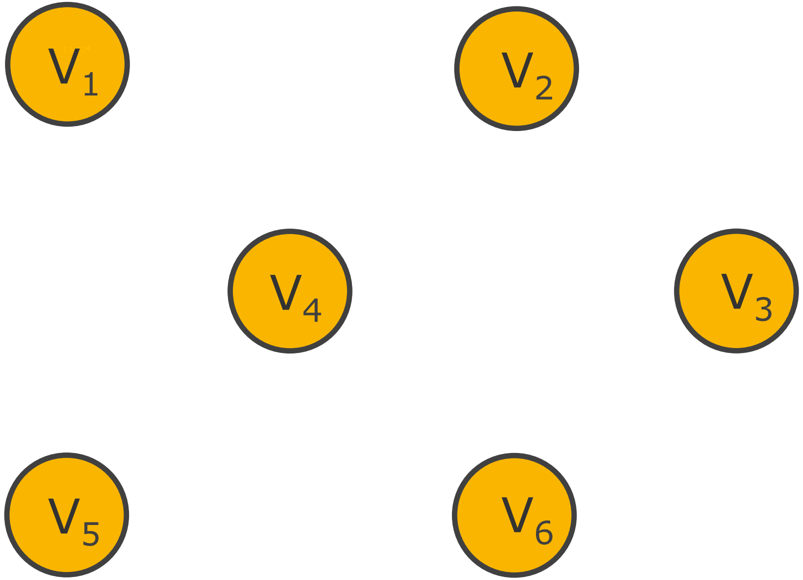
V_3: Outside temp.

V_4: Cooling action

V_5: Thermal waste

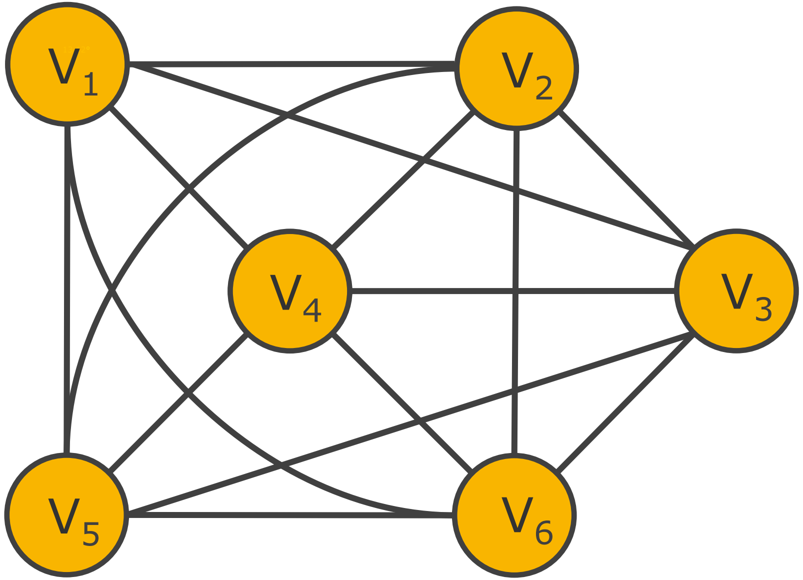
V_6: Electricity consumption

Causal Inference on Gene Expression Data: The Peter-Clark algorithm



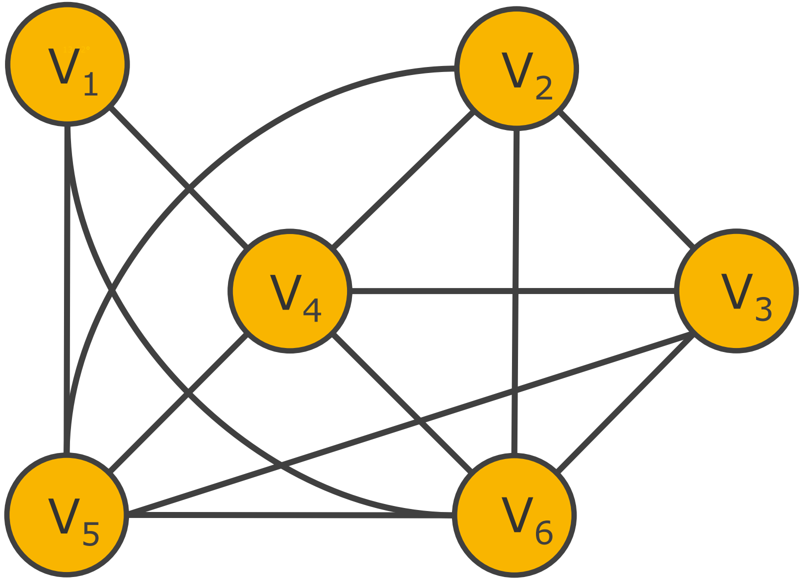
Causal Inference on Gene Expression Data: The Peter-Clark algorithm

- Fully connected graph



Causal Inference on Gene Expression Data: The Peter-Clark algorithm

First iteration: Remove edges without direct correlation

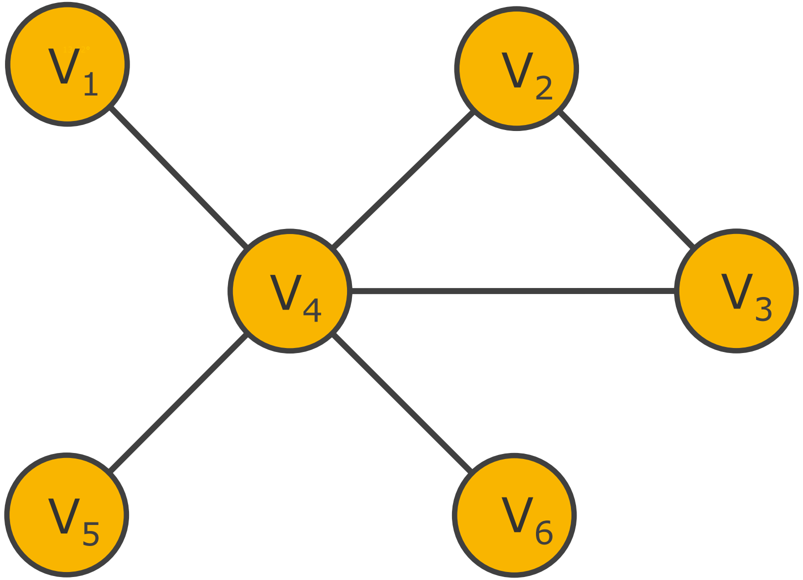


$V_1 \perp V_2$

$V_1 \perp V_3$

Causal Inference on Gene Expression Data: The Peter-Clark algorithm

- Second iteration: Remove conditionally independent edges



$V_1 \perp V_5 \mid V_4$

$V_1 \perp V_6 \mid V_4$

$V_2 \perp V_5 \mid V_4$

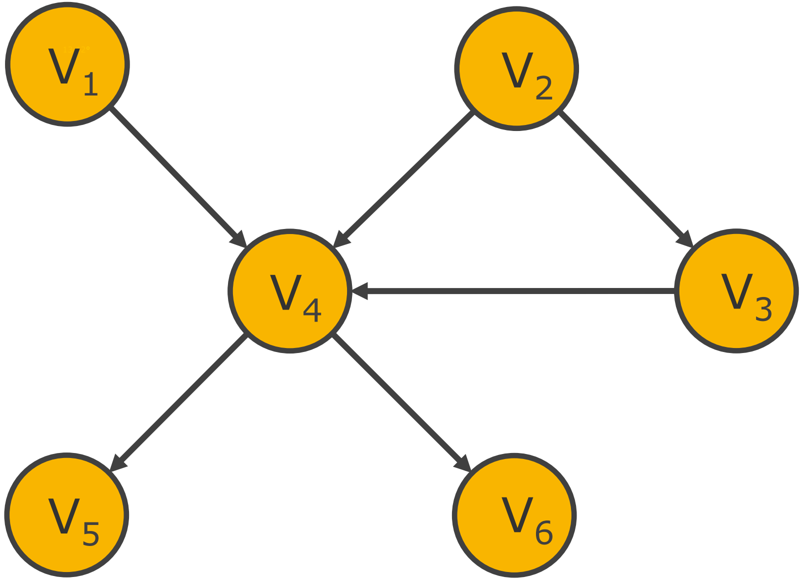
$V_2 \perp V_6 \mid V_4$

$V_3 \perp V_5 \mid V_4$

$V_3 \perp V_6 \mid V_4$

Causal Inference on Gene Expression Data: The Peter-Clark algorithm

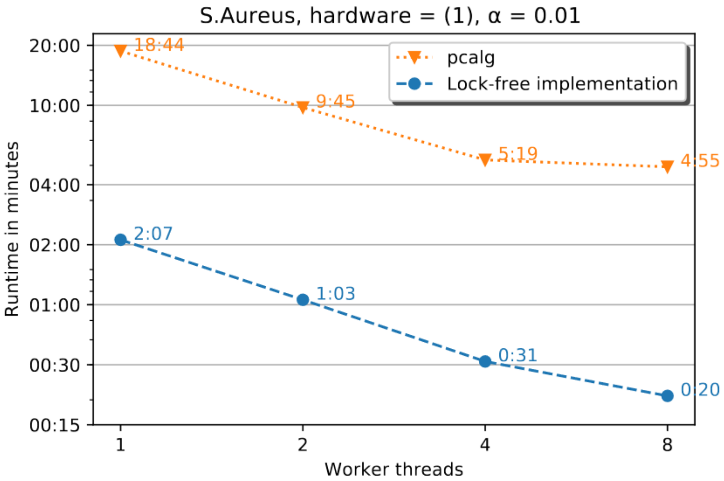
- Rule-based edge directing

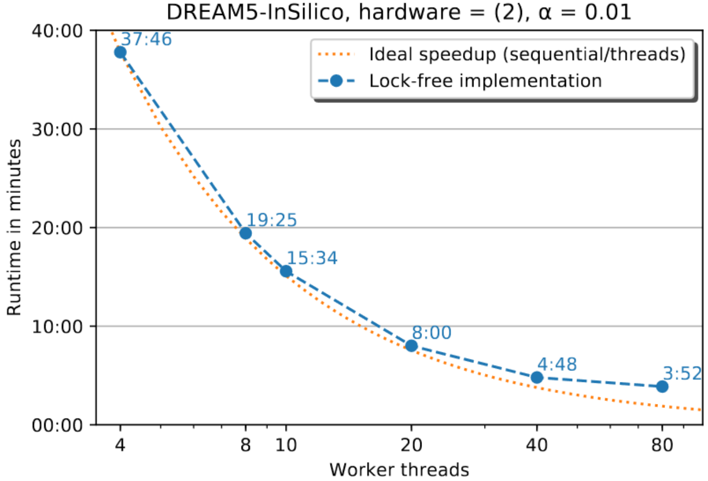


Causal Inference on Gene Expression Data:
Motivation

- Working with causal modeling instead of statistical approach:[4]
  - Approximate gene regulatory networks
  - Incorporate known effects of knock-out/down trials

- PC-algorithm: Limited preprocessing and massively parallelizable





Causal Inference on Gene Expression Data:
Challenges

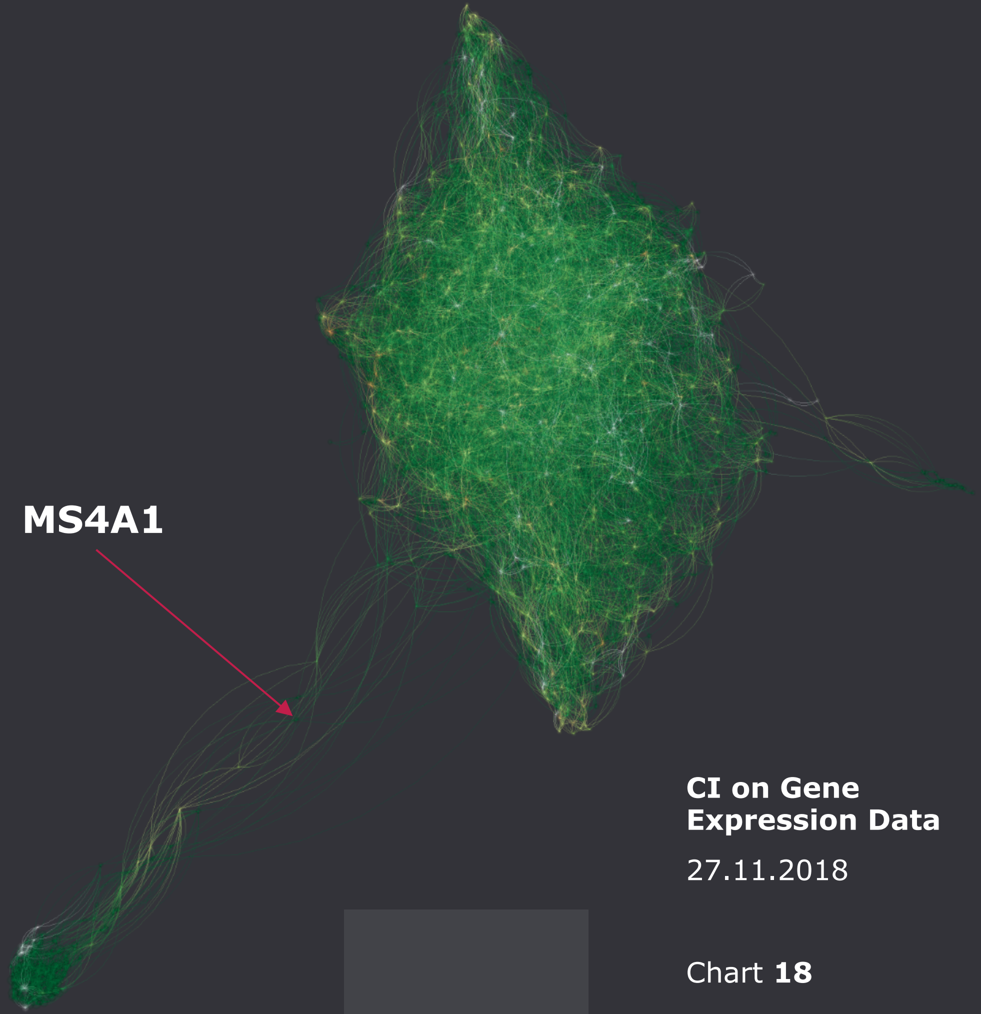
- Feasibility of constraint-based learning approach:
  - High dimensionality: 35K genes
  - Density of underlying causal graph
- (Most probably) many non-linear dependencies
  - Conditional independence tests computationally expensive[5]
- How to:
  - Interpret results?
  - Combine samples to form data sets?

Causal Inference on GED: Goals and Next Steps

- Next Steps:
  - Run on discretized values
  - Differential graph analysis on healthy/ cancerous tissue
  - (Integrate test for non-linear dependencies)

- Goals: Evaluate…
  - …feasibility of PC-algorithm
  - …resulting causal graphs:
    - With external knowledge bases
    - In comparison to other approaches

- Multi-cancer samples
- Top 2500 genes by variance
- 20.043.623.346 tests
- 5h on 64 cores



MS4A1

Aggretated OpenTargets neoplasm association

🟥 🟥 🟧 🟧 🟨 🟨 🟩 🟩 🟩 🟩

High Low

Sources

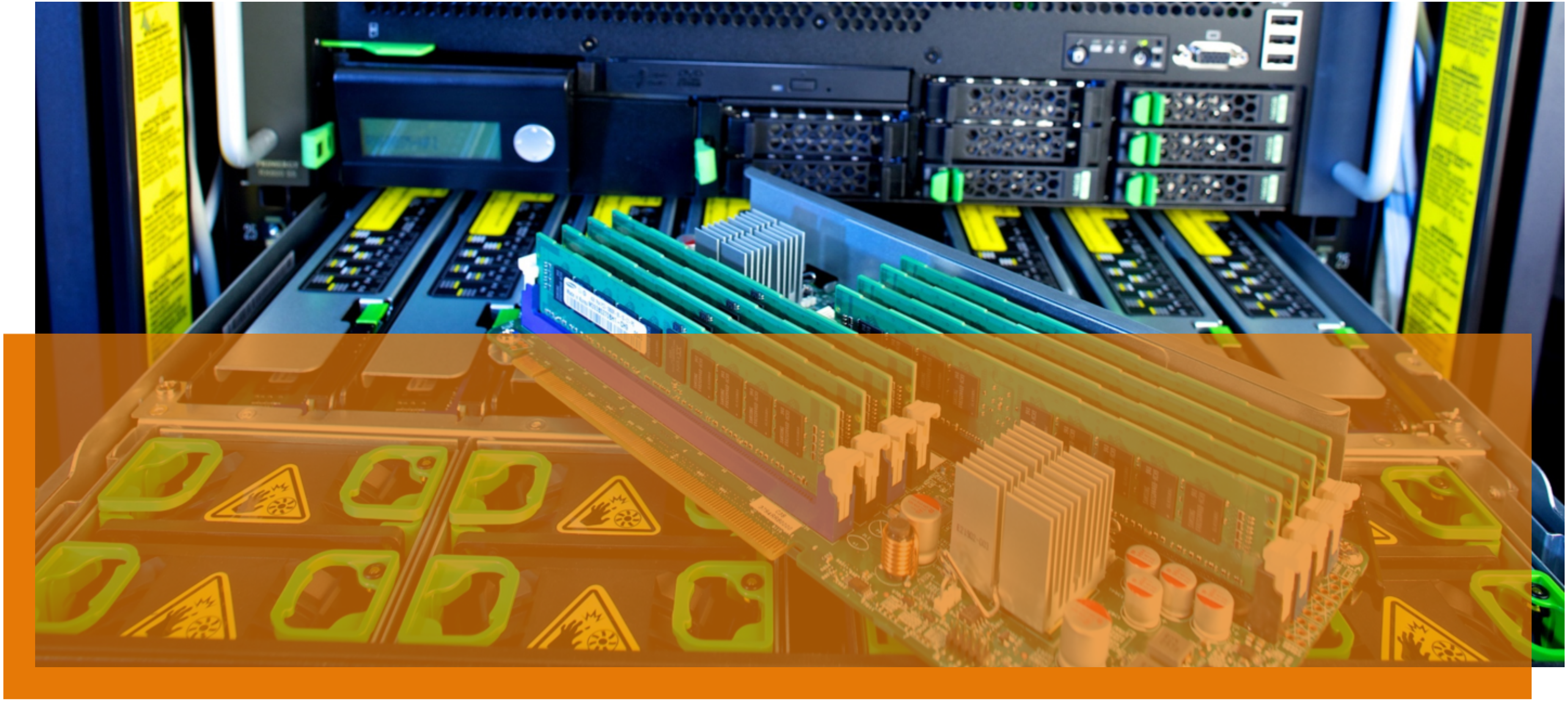
- [1] Oshlack, Alicia, Mark D. Robinson, and Matthew D. Young. "From RNA-seq reads to differential expression results." *Genome biology* 11, no. 12 (2010): 220.
- [2] ICGC-TCGA DREAM Somatic Mutation Calling - RNA Challenge https://www.synapse.org/#!Synapse:syn2813589/wiki/401435
- [3] Causal Inference – Theory and Applications: https://hpi.de/plattner/teaching/archive/summer-term-2018/causal-inference-theory-and-applications.html
- [4] Rau, Andrea, Florence Jaffrézic, and Grégory Nuel. "Joint estimation of causal effects from observational and intervention gene expression data." *BMC systems biology* 7, no. 1 (2013): 111.
- [5] Ramsey, Joseph D. "A scalable conditional independence test for nonlinear, non-Gaussian data." *arXiv preprint arXiv:1401.5031* (2014).

- Made by Freepik from www.flaticon.com :

Universität Potsdam

HPI Hasso Plattner Institut

Digital Engineering · Universität Potsdam



Universität Potsdam

HPI Hasso Plattner Institut

Digital Engineering · Universität Potsdam