Different Takes on the Causal Modelling of Spatio-Temporal Data

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

March 6, 2024

Machine Learning: It Should be Causal!

Machine learning is changing science and the society we live in thanks to our ability to learn models that can capture information effectively.

Article

Highly accurate protein structure prediction with AlphaFold

Nature | Vol 596 | 26 August 2021 | **583**

nature computational science

PERSPECTIVE

Scaling digital twins from the artisanal to the industrial

Steven A. Niederer, Michael S. Sacks, Mark Girolami and Karen Willcox

NATURE COMPUTATIONAL SCIENCE | VOL 1 | MAY 2021 | 313–320

The Economist

Science & technology | Generative AI

Large, creative AI models will transform lives and labour markets

They bring enormous promise and peril. In the first of three special articles we explain how they work

It creates black boxes that use probabilistic associations for prediction.

It is not equipped to understand the causality that drives reality.

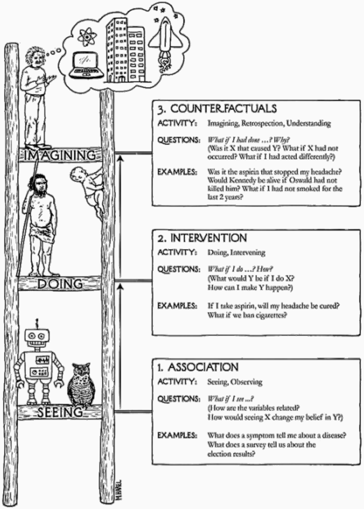
Scientific questions are inherently causal. Causation is central to how we think as human beings and how we understand the world.

Causality is a Network

Judea Pearl [3] has worked out a rigorous theory of causality that uses directed (acyclic) graphs to represent causes and effects. With it, we can reason about

- what we see,
- what happens if we affect change,
- what would have happened.

How can we learn them?



Causality and Statistical Learning

Research on causality is limited because

1. it concentrates on inference,
2. models are given or learned from toy data sets.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–16: Review
https://doi.org/10.1080/01621459.2021.1874961

Graphical Models for Processing Missing Data

Karthika Mohan[a] and Judea Pearl[b]

DAGs with NO TEARS: Continuous Optimization for Structure Learning

Xun Zheng[1], Bryon Aragam[1], Pradeep Ravikumar[1], Eric P. Xing[1,2]

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

We must be able to learn causal networks to get to do inference!

nature COMMUNICATIONS

https://doi.org/10.1038/s41467-021-25743-9 OPEN

Causality in digital medicine

nature COMMUNICATIONS

https://doi.org/10.1038/s41467-020-15195-y OPEN

Causal networks for climate model evaluation and constrained projections

nature

OUTLOOK | 24 February 2023

Why artificial intelligence needs to understand consequences

A machine with a grasp of cause and effect could learn more like a human, through imagination and regret.

Learning a causal network means

$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}},$$

with its structure $\mathcal{G}$

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G})\,\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \mathrm{P}(\mathcal{G}) \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\,\mathrm{P}(\Theta \mid \mathcal{G})\,d\Theta$$

and its parameters Θ

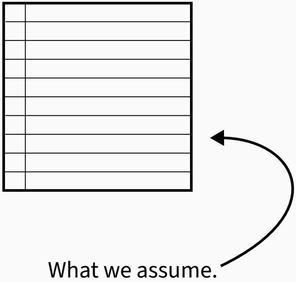
$$\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D}) = \mathrm{P}(\Theta \mid \mathcal{G})\,\mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta).$$

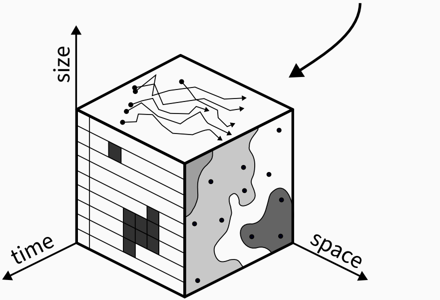
We used to ask domain experts for information [2, 1]; now we rely increasingly on the data $\mathcal{D}$.

What assumptions do we make on the data?

- Observations are independent and there are no missing values;
- We observe all variables, that is, no latent variables introducing confounding into the model.

They are too restrictive!





Distributions for Complex Data

Complete independence	$\mathcal{D} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{I})$
Temporal dependence	$\mathcal{D} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}_t)$
Spatial dependence	$\mathcal{D} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}_s)$
Spatio-temporal dependence	$\mathcal{D} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s)$

And we are not even considering interactions between space and time parameters, much less missing values and latent confounders.

Q1 Can we afford it? (Hint: no.)

Q2 Do we need it? (Hint: not necessarily.)

Nuisance Parameters, Parameters of Interest

A causal network has two components: the graph $\mathcal{G}$ and the parameters Θ. Causal inference defines queries using $\mathcal{G}$:

- Conditional independence, via d-separation.
- Intervention, via mutilation.
- Counterfactual, via the twin network.

Our ability to answer scientific questions using the causal network rests on having the right nodes in the network. If we don't, we cannot even formulate our question.

- The dimensions we use in the queries (interest) should be represented as nodes.
- The dimensions we do not (nuisance) should be represented as parameters in the local distributions.

The Local Distributions in a Causal Networks

A causal network defines the local distributions

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i}) \quad \text{where} \quad \Pi_{X_i} = \{\text{parents of } X_i\}.$$

where $\bigcup_{\mathbf{X}} \Theta_{X_i} = \Theta$. If X_i and Π_{X_i}, we commonly assume

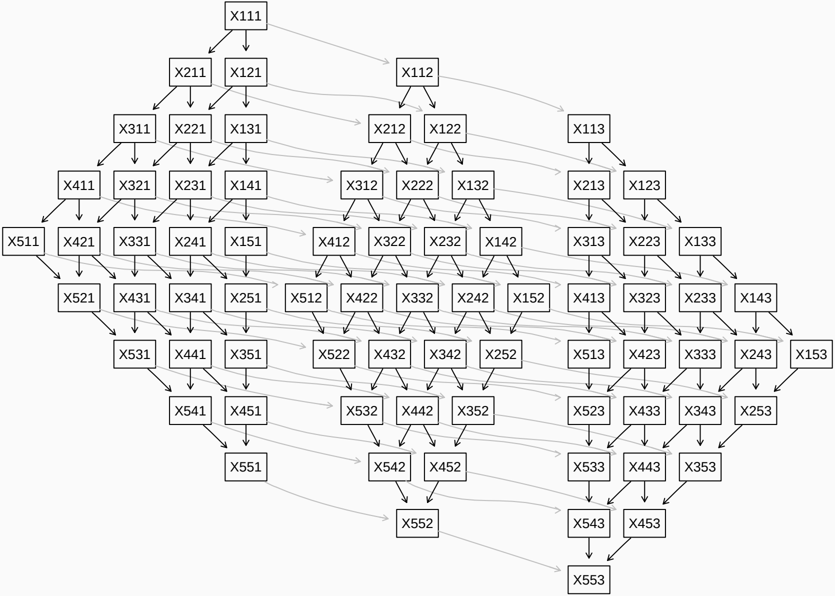
$$X_i = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i} + \boldsymbol{\varepsilon}_{X_i}, \qquad \boldsymbol{\varepsilon}_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I}).$$

It is quite all right to embed space, time or both in $\boldsymbol{\varepsilon}_{X_i}$ as

$$\boldsymbol{\varepsilon}_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I} \otimes \Sigma_t \otimes \Sigma_s)$$

instead of keeping $s \times t$ location-time combinations as separate nodes.

Time and Space as Nodes, No Nuisance Parameters



Time and Space as Nodes, No Nuisance Parameters

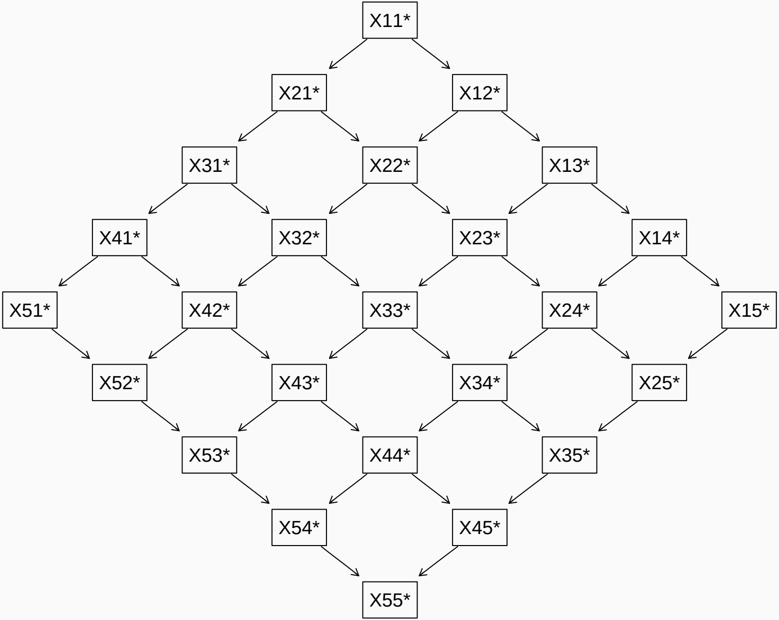
PROS:

- We can formulate queries across time points.
- We can formulate queries across locations.
- We can capture any state-space dependence structure across data.
- Local distributions are straightforward since $\varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I})$.

CONS:

- Sample size requirements are impractical.
- Visual inspection is impossible.
- Causal inference has exponential computational complexity.
- Not all arcs have a causal interpretation?

Space as Nodes, Time as a Nuisance Parameter



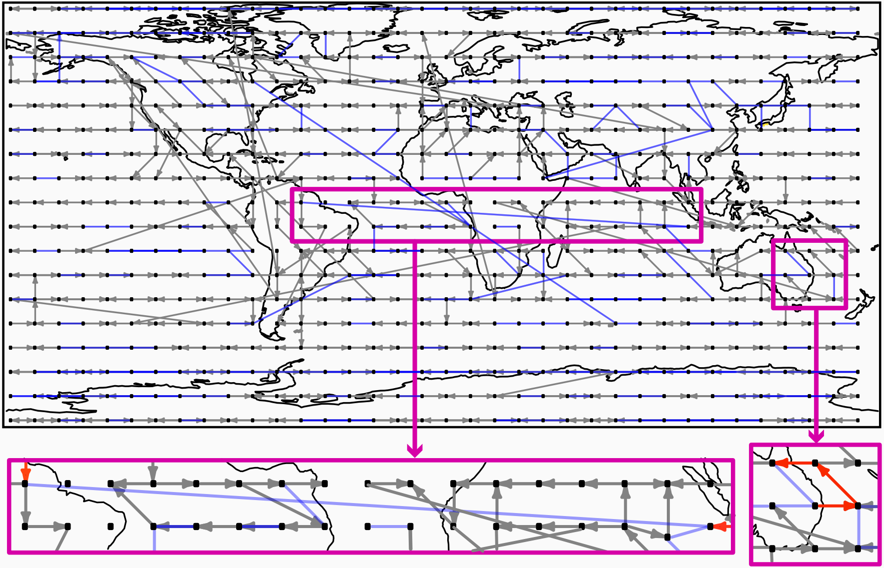
PROS:

- We can formulate queries across locations.
- We can capture any dependence structure across space.
- Local distributions are straightforward since $\varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I} \otimes \Sigma_t)$ where Σ_t can be the AR(1) correlation matrix.
- We can use thinning!

CONS:

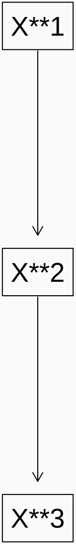
- We cannot formulate queries across time points.
- Sample size requirements are still difficult to meet.
- Visual inspection is next to impossible.
- Causal inference has exponential computational complexity.
- We can use Granger causality to allow for causal interpretation.

Space as Nodes: an Example



M. Scutari and C. E. Graafland and J. M. Gutiérrez (2019). [4]

Time as Nodes, Space as a Nuisance Parameter



Time as Nodes, Space as a Nuisance Parameter

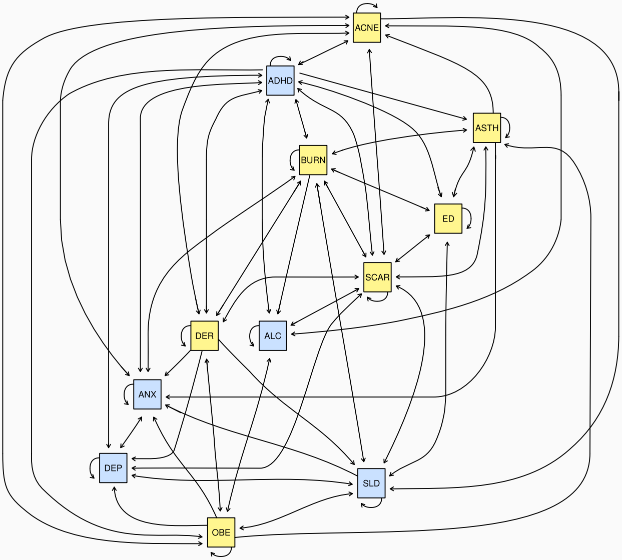
PROS:

- We can formulate queries across time points.
- We can capture any dependence structure across time: autocorrelation, seasonality, drift, etc.
- Sample size requirements are easier to meet.
- Visual inspection and inference are straightforward.

CONS:

- We cannot formulate queries across locations.
- Local distributions are quite complicated: $\boldsymbol{\varepsilon}_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I} \otimes \Sigma_s)$ and it is much more difficult to parameterise Σ_t than Σ_s.

Time as Nodes: an Example



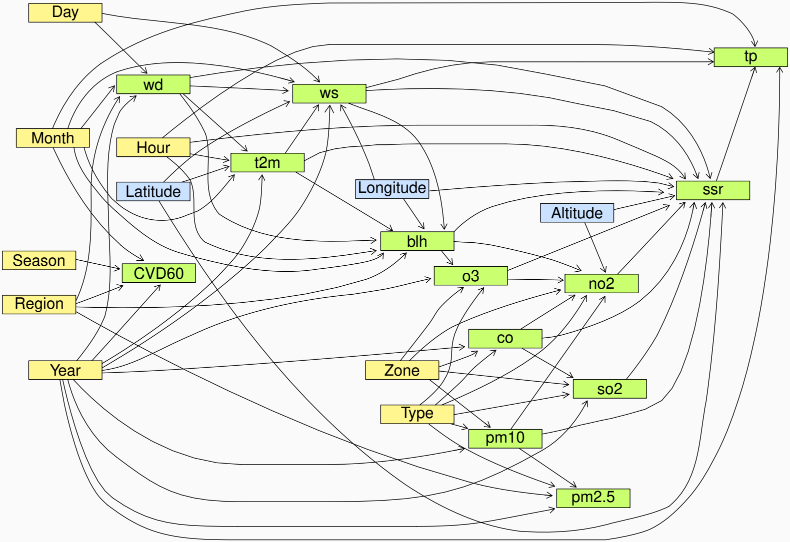
M. Scutari and D. Kerob and S. Salah (2024). [5]

PROS:

- Sample size requirements are easy to meet since we aggregate data across both time and space.
- Visual inspection and inference are straightforward.

CONS:

- We cannot formulate any query across time points or locations.
- We cannot model any kind of feedback loop.
- Local distributions are extremely complicated:
  $\boldsymbol{\varepsilon}_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I} \otimes \Sigma_t \otimes \Sigma_s)$.



C. Vitolo, M. Scutari, M. Ghalaieny, A. Tucker and A. Russell (2018). [6]

Few Conclusions, Many Open Questions

A1 Which approach we can afford to use depends on how many variables, locations and time points we have and on how regular their pattern is.

A2 Which approach we need depends entirely on which questions we want to answer with causal inference.

Q3 Computational complexity trade-offs?

Q4 Software availability?

Q5 Prior distributions to keep statistical learning on the reasonable side?

Q6 Implications on causal inference?

Acknowledgements

Catharina Elisabeth Graafland, José Manuel Gutiérrez
Institute of Physics of Cantabria (CSIC-UC)

Allan Tucker, Andrew Russell, Mohamed Ghalaieny
Brunel University London

Claudia Vitolo
European Space Agency

Samir Salah, Delphine Kerob
L'Oréal, La Roche-Posay

That's all!

Happy to discuss in more detail.

- M. J. Druzdel and L. C. van der Gaag.
  Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information.
  In *UAI*, pages 141–148, 1995.
- M. J. Druzdel and L. C. van der Gaag.
  Building Probabilistic Networks: "Where Do the Numbers Come From?".
  IEEE Transactions on Knowledge and Data Engineering, 12(4):481–486, 200.
- J. Pearl and D. Mackenzie.
  The Book of Why: the New Science of Cause and Effect.
  Basic Books, 2018.
- M. Scutari, C. E. Graafland, and J. M. Gutiérrez.
  Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms.
  International Journal of Approximate Reasoning, 115:235–253, 2019.
- M. Scutari, D. Kerob, and S. Salah.
  Inferring Skin-Brain-Skin Connections from Infodemiology Data Using Dynamic Bayesian Networks.
  Scientific Reports, 2023.

References II

- C. Vitolo, M. Scutari, A. Tucker, and A. Russell.
  Modelling Air Pollution, Climate and Health Data Using Bayesian Networks: a Case Study of the English Regions.
  Earth and Space Science, 5(4):76–88, 2018.