bnlearn, Learning Bayesian Networks

15 Years Later

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

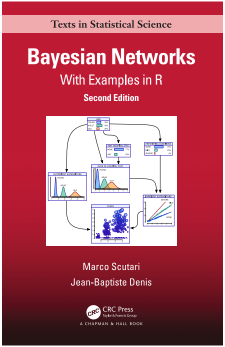
June 24, 2024

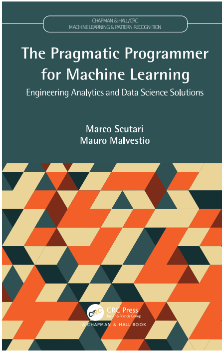
A Brief Intro

Sonia Shah

Kitty Lo
Scientist | Data scientist

David Balding





projects / bnlearn / commit git

summary | shortlog | log | commit | commitdiff | tree
(initial) | patch

commit ? search: re

Initial commit (v 0.1).

```
author     Marco Scutari <                 >
           Tue, 12 Jun 2007 18:53:43 +0000 (20:53 +0200)
committer  Marco Scutari <                 >
           Tue, 12 Jun 2007 18:53:43 +0000 (20:53 +0200)
commit     b8c24c841b6941fc631031ba061fbd3b0ac71de6
tree       48ad0bfcc78e0123df87bdc82b74d195ce46877b        tree | snapshot

Initial commit (v 0.1).

DESCRIPTION            [new file with mode: 0644]  blob
NAMESPACE              [new file with mode: 0644]  blob
R/cibn.R               [new file with mode: 0644]  blob
R/test.R               [new file with mode: 0644]  blob
R/utils.R              [new file with mode: 0644]  blob
man/bnlearn-package.Rd [new file with mode: 0644]  blob
man/gs.Rd              [new file with mode: 0644]  blob
```

bnlearn R package Atom RSS

Why I like Bayesian Networks

Machine learning creates black boxes that use probabilistic associations for prediction, but scientific questions are inherently causal. Causation is central to how we think and how we understand the world.

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

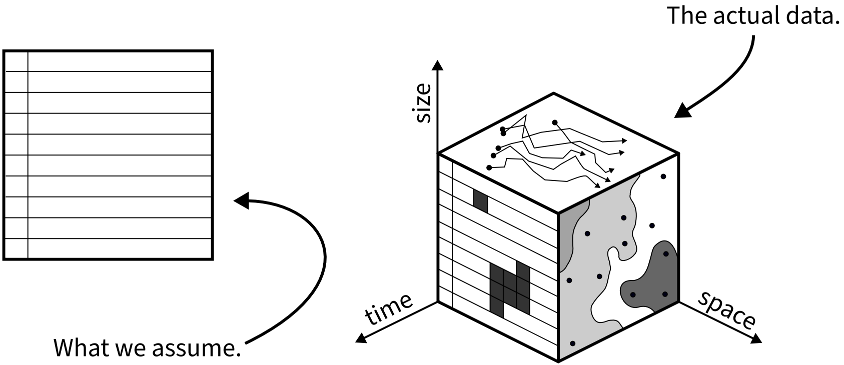
They bring enormous promise and peril. In the first of three special articles we explain how they work

Bayesian networks are the opposite: they promote understanding so that we act to improve the world, easily modelling how interventions will impact outcomes.

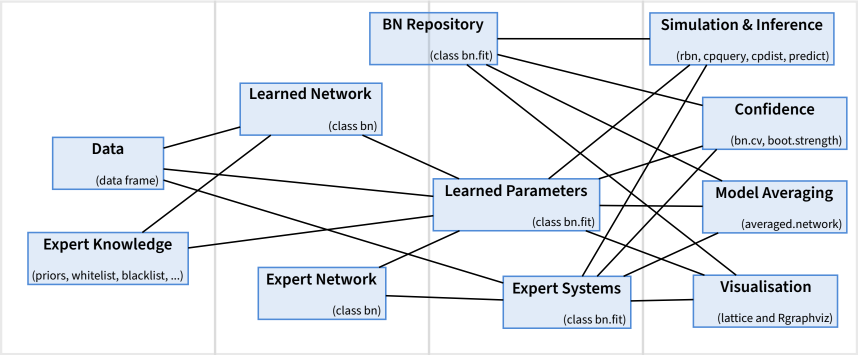
Bayesian networks have come a long way...

... but there is much we do not understand or we cannot do well:

- A theory of statistical learning for causal network models.
  - Handling latent confounders?
  - Model identifiability?
  - Relative contributions of parameter and structure learning?
- Scalable methods and software to empower applications.
- Model challenging data with complex structures.
  - Space-time data?
  - Missing data?
  - Heterogeneous populations?



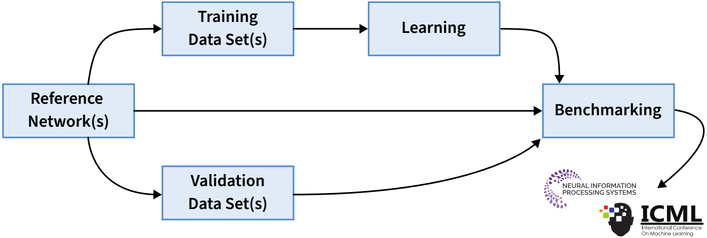
This why I develop bnlearn: I want a software developed with the best practices software engineering has to offer and that does not crash and burn when fed real-world data.

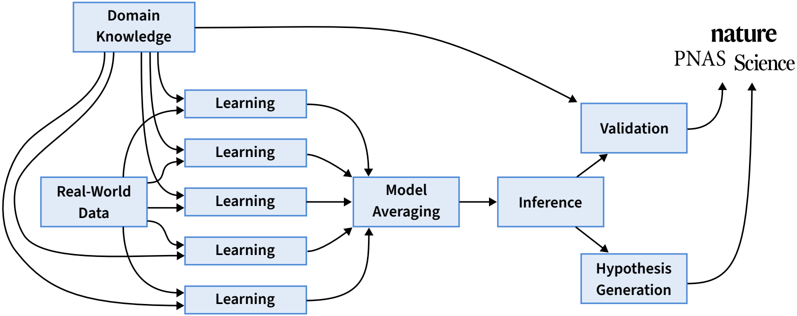


New and upcoming features: Don't delay, download today!

- Missing data (EM, PNAL).
- Custom tests and scores.
- Exact inference (DBN, GBN).
- Entropy and KL Divergence.
- More C code, for speed.
- Updating `bn.fit` objects (soon).

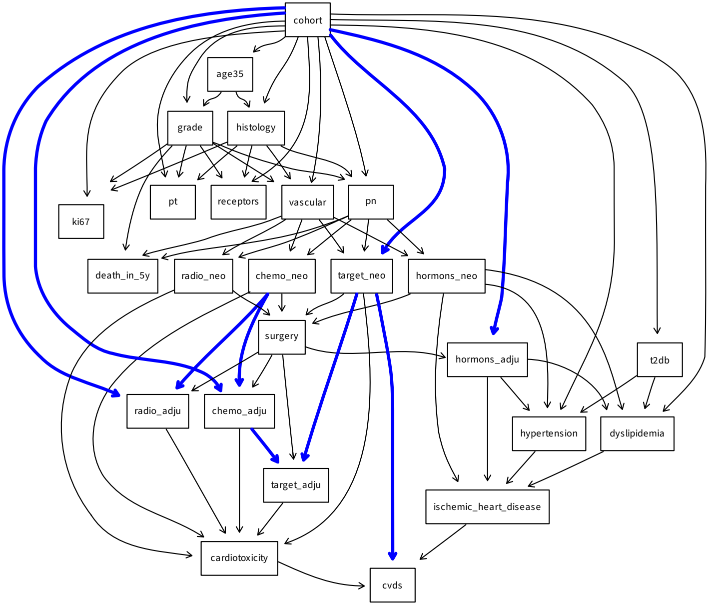
Two Main Workflows in bnlearn



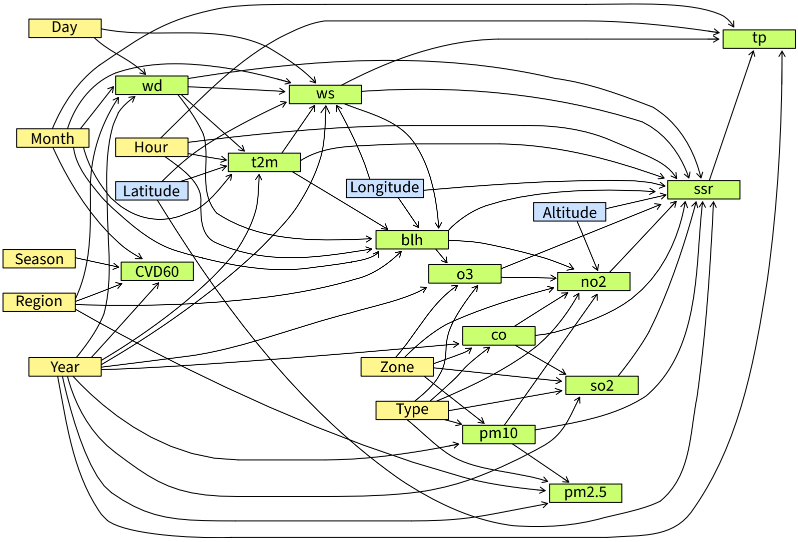


1. Choose some networks from `www.bnlearn.com/bnrepository`.
2. Possibly alter their structure and/or parameters to match the scientific question you want to answer.
3. Generate data sets with `rbn()`, with replicates across $n/p \in \{0.1, 0.2, 0.5, 1, 2, 5\}$.
4. Run the learning approach of your choice.
5. Benchmark its outputs.
   - Structural measures: `compare()` and `shd()`.
   - Information theoretic measures: `H()` and `KL()`.
   - Empirical measures, with a validation data set: `predict()` or `logLik()`.
   - Visual analysis: `graphviz.compare()` and `graphviz.chart()`.
6. Profit!

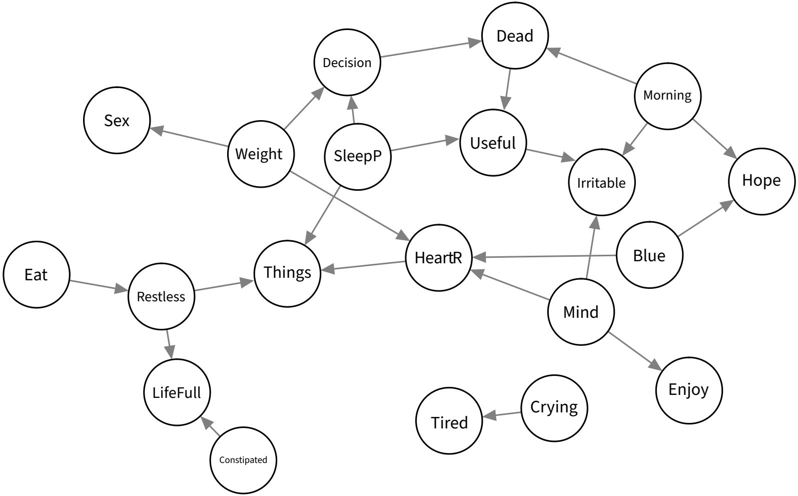
1. Preprocess the data.
   - Remove highly correlated variables: `dedup()`.
   - Possibly discretising variables: `discretize()`.
   - Possibly imputing missing values: `impute()`.
2. Structure learning with model averaging.
   - Establish a blacklist of arc directions that make no sense.
   - Learn multiple structures with `boot.strength()` or `custom.strength()`.
   - Average them with `averaged.network()`.
3. Parameter learning with `bn.fit()`.
4. Model validation and hypothesis generation.
   - Can you find literature supporting the existence of arcs and paths?
   - Do their answers (through `cpquery()`, `cpdist()`, `mutilated()`, `predict()`) to key questions agree with domain experts?



Combining population and clinical trial data for breast cancer survivors [1].



Linking pollution with respiratory and cardiovascular diseases in the UK [5].



Mapping causal paths between depression symptoms[2].

- At least one between GES and LiNGAM, which have become baselines in much of the causal discovery literature.
- Exact uniform sampling over DAGs from Kuipers & Moffa (2012).
- A query function that returns a conditional distribution.
- Better exact inference, including CGBNs.
- The Structural Intervention Distance from Peters & Bülmann (2015).
- Creating twin networks for counterfactuals.
- Support for either state-space or stratified data.
- Better support for incomplete data.

That's all!

Happy to discuss in more detail.

References I

- A. Bernasconi, A. Zanga, P. J. F. Lucas, M. Scutari, and F. Stella.
  Towards a Transportable Causal Network Model Based on Observational Healthcare Data.
  In *Proceedings of the 2nd Workshop on Artificial Intelligence for Healthcare, 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*, pages 67–82, 2023.
- G. Briganti, M. Scutari, and P. Linkowski.
  Network Structures of Symptoms from the Zung Depression Scale.
  Psychological Reports, 124(4):1897–1911, 2021.
- M. Scutari and J.-B. Denis.
  Bayesian Networks with Examples in R.
  Chapman & Hall, 2nd edition, 2021.
- M. Scutari and M. Malvestio.
  The Pragmatic Programmer for Machine Learning: Engineering Analytics and Data Science Solutions.
  Chapman & Hall, 2023.
- 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.