Benchmarks, wikis, and open-source causal discovery

Patrik O. Hoyer

Univ. of Helsinki
Finland

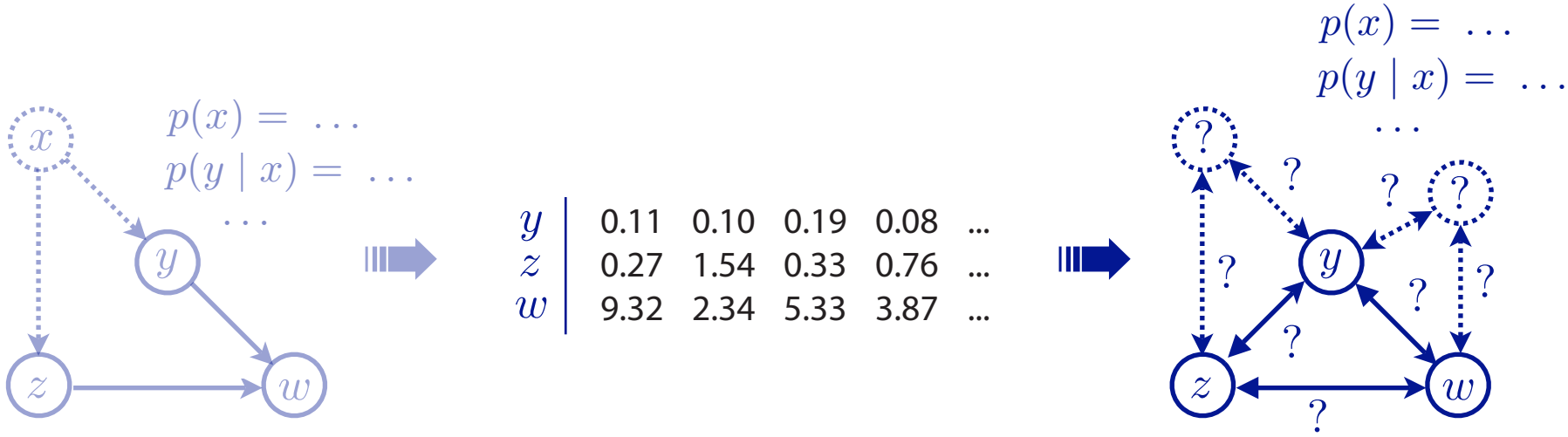
NIPS*08 workshop on causality
Dec 12, 2008

Beware! Not a technical talk!

The causal discovery problem

- Unknown data-generating ('causal') system
- We have some non-experimental and/or experimental data, from which we seek to infer the causal system...

  ... this is an extremely difficult 'inverse' problem which requires good assumptions/priors to succeed!

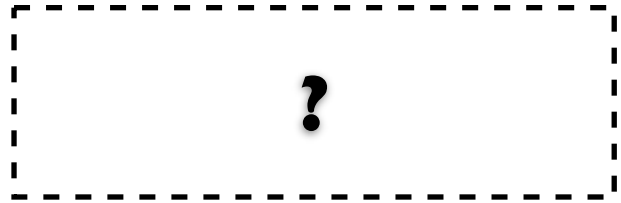


How well can we actually do it?

- Lots of different methods proposed
- Testing these methods on real data requires...
  - ...a set of non-experimental / experimental data
  - ...auxiliary knowledge of the true causal system!

...so testing causal discovery methods is quite more complicated than testing methods for regression, classification, or density estimation!

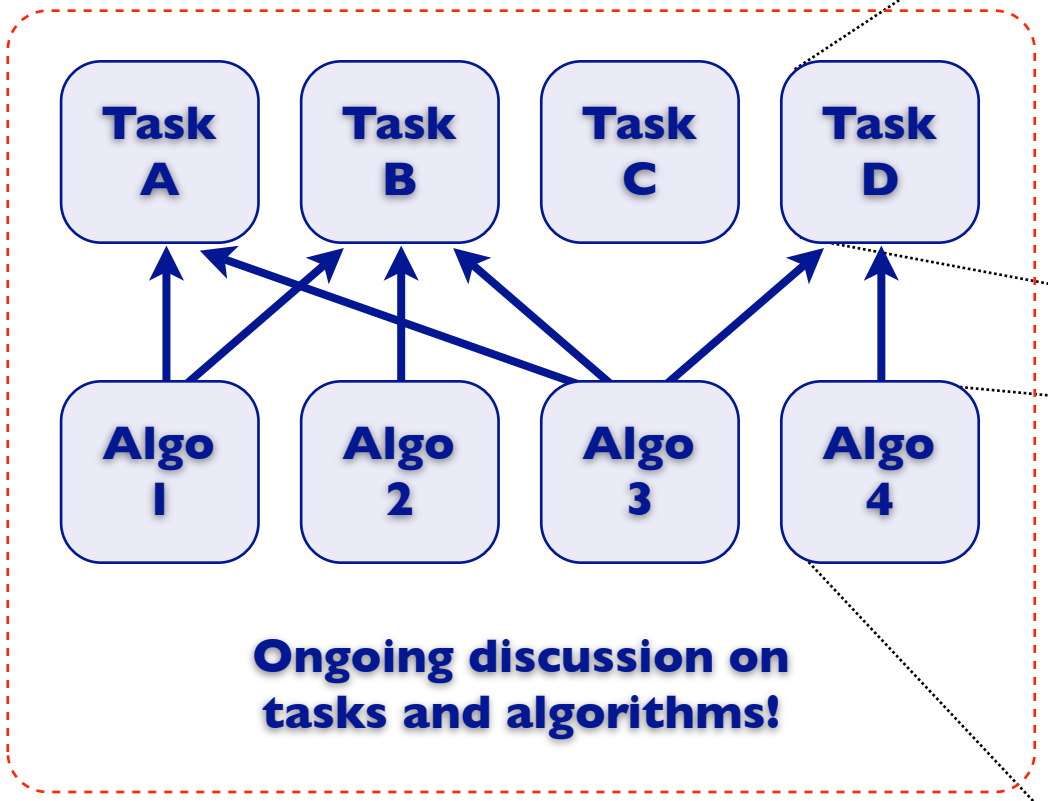
density estimation
classification
regression



causal discovery
causal inference

Causal discovery repository?

- Both problems and solutions...



...and anyone can evaluate how any algorithm performs on any task.

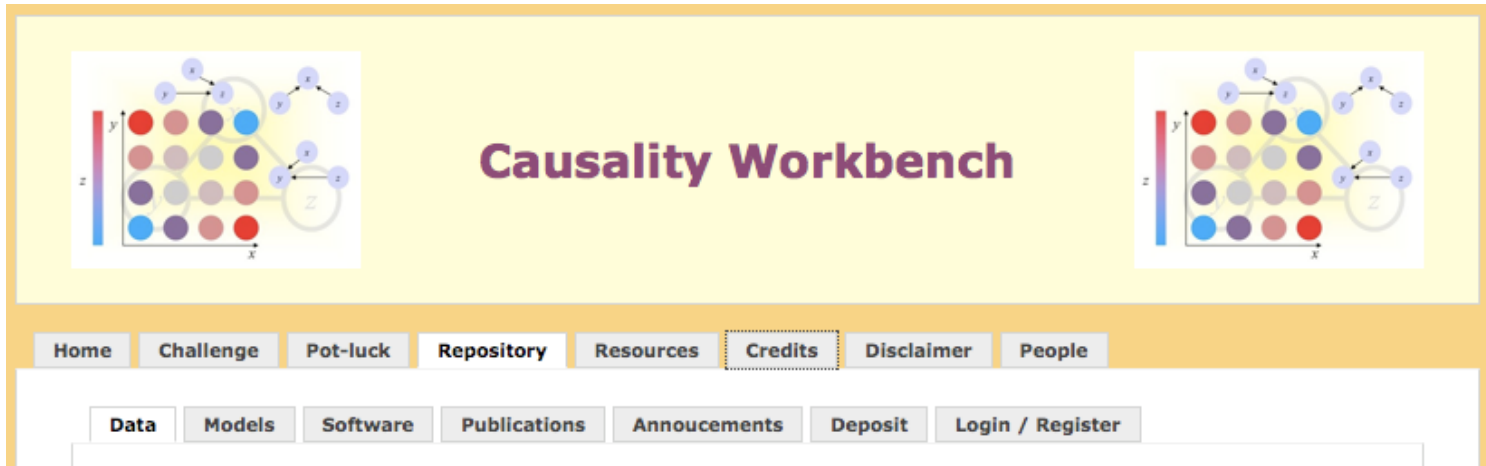
Task:

- Real or simulated data
- Precisely defined:
  - what should algo do?
  - exact scoring procedure
  - reasonable assumptions about the data

Algorithm:

- What does it do?
- Input-output well defined
- Open-source (if possible) or executable available (so anyone can run it on any dataset)

Causality workbench! (Isabelle Guyon et al.)



- 15 datasets (≈ 'tasks') already collected!
- The two challenges have spawned numerous approaches to solutions (≈ 'algorithms'), although these are not (at least yet) collected on-line for anyone to evaluate
- ...all-in-all, an excellent effort and a great start!

The nature of the project

- Volunteer-based collaborative effort
- Need for repository to become self-sustaining
- Look at successful examples for good ideas and principles...
  - Open-source software development (Linux, GNU)
  - Wikipedia

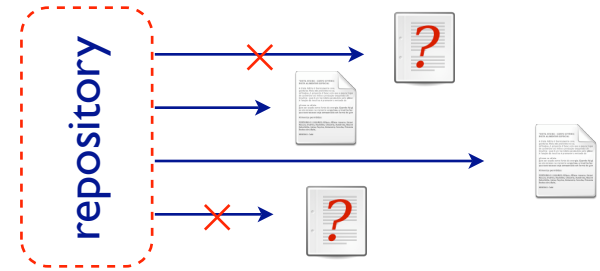


(obviously, slightly different scale of projects, but the basic principles are still pretty much the same...)

- What follows are my humble thoughts on what some of these principles are

1. All material – all versions – permanently available

- Store everything on-site rather than just as links
  - Eliminates broken links -problems (e.g. 'ICA Central' repository, quick test: 5/10 broken links)



  - Enables full versioning of all material (particularly important in a scientific context where priority is often an issue)

  - Easy full downloading of all material (ensuring that all material is permanently available)

(Thus, need storage capacity and bandwidth, and need to consider licensing issues)

2. As few technical restrictions as possible

- With full versioning (and easy reverting) there is no need for technical restrictions, rather one can rely on socially agreed-upon rules (may still require registration to make changes/updates)

  [assuming the collaborators are trying to help the project, technical restrictions are more annoying than helpful]

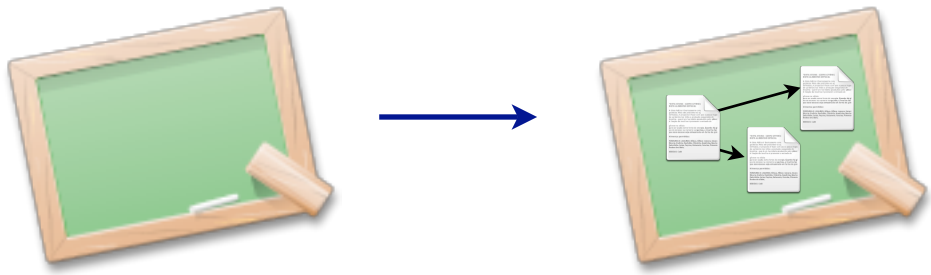
- Anyone can help with any part of the project
  - Contributing new tasks and solutions, and developing improved solutions based on earlier ones
  - Writing documentation
  - Clarifying the rules and conventions of the repository
  - Graphics and design

3. Emergence of structure

- It may be difficult to predict and impose the appropriate structure from the outset...

  ...so often useful to allow the structure to emerge as the project develops (and allow anyone to help in constructing the structure that best fits the changing needs)

- 'Wiki' software: Free-form and/or structured collaborative webpages, file uploads, categories, templates, full versioning with easy reverts, complete downloads, all text and structure is collaboratively edited by all users



MediaWiki
Because ideas want to be free.

+ lots of other options

Summary

- Might be worthwhile to consider...
  - ...storing all material on-site rather than as links
  - ...using full versioning of all material
  - ...relying on social rules rather than technical restrictions to keep the repository in order
  - ...making it possible for everyone to work on the structure as well as the content
- These features/aspects may be easiest to implement using freely available wiki software
- I hope (and believe) that together we can make the repository an extremely useful tool for benchmarking existing causal discovery methods and developing new ones.