Causal Models:
Drawing on Category Theory

Ed Wike
MIT Category Theory Seminar
November 19, 2018

Agenda

- Project Overview
- Relevance and Interests
- Causal Models Overview
- Category Theory Framework
- Python Software Approach
- Demo – Pilot Version
- Next Phase - Programming
- Applicability and Benefits

Project Overview

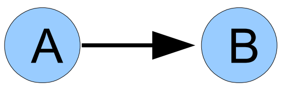
- In May, Brendan Fong gave a talk on "Causal theories: a categorical approach to Bayesian networks."
- Based on Brendan's masters thesis, "Causal Theories: A Categorical Perspective on Bayesian Networks".
- This thesis generalizes and expands upon Bayesian Networks using category theory.
- Causal inference is an important topic in analytics
- Today's talk is about a software project to automate the modeling methodology and see what it can do.
- We will see a demo of an early version of this program.
- Then we will discuss plans for a more robust version and some ideas to explore.

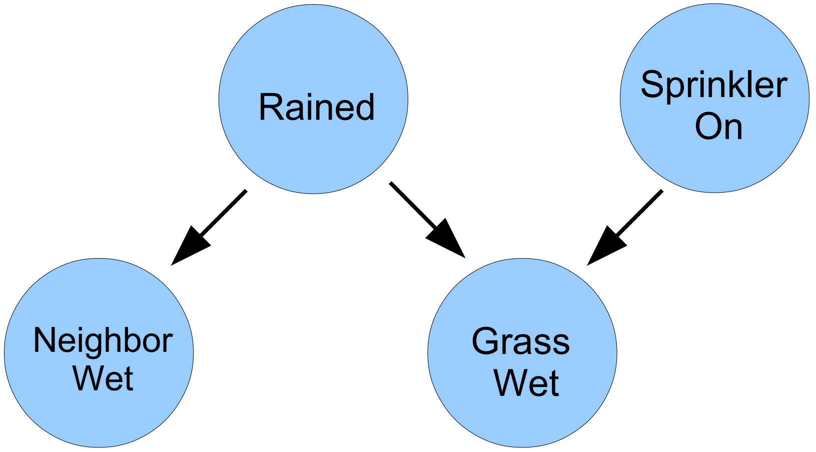
Relevance and Interests

- Causal modeling is increasingly important in the analytics world.
  - Ongoing debate between 'empiricism' and 'innate domain'.
  - Concerns about implicit bias and disparate impact of models
  - Lack of transparency and the need for explainability of models
- Personal interests and experience:
  - Math, operations research (MIT,1998), analytics, consulting
  - Category theory and applied category theory (+ analytics)
  - Modeling and data science experience
    - RMBS – prepayment, default, house price models
    - Mortgage lead generation - purchase, refinance, home equity
    - Customer retention models, marketing, fraud detection
    - Consumer credit, property, demographic, economic data
  - Network diagram programming for workflow scheduling models

Causal Models

- Example – Wet Grass
- P(AB) = P(A)P(B|A)





- Bayesian Network
  - Network (DAG) depiction of causal relationships
  - Nodes and arcs represent conditional probabilities
  - Intuitive interpretation of causal paths through network
  - Widely used for causal modeling and inference
  - Scalability – driven by conditional probability table
  - Key concept – conditional independence
  - More on conditional probability...

Conditional Probability

- Example: Random variables X (values 1, 2, 3) and Y with values (1,2)
- Joint probability table of X,Y and marginal probabilities X and Y:

Probability	X=1	X=2	X=3	Y Marginal
Y=1	0.10	0.40	0.15	0.65
Y=2	0.10	0.10	0.15	0.35
X Marginal	0.20	0.50	0.30	

- Dividing the joint probabilities of X,Y by the marginal probabilities for X:

Prob (Y | X)	X=1	X=2	X=3
Y=1	0.50	0.80	0.50
Y=2	0.50	0.20	0.50

- The columns of this [Y|X] matrix sum to 1 (stochastic matrix)
- If X is distributed as (0.8, 0.1, 0.1), then left multiplying by [Y|X] = (0.53,0.47)
- If X is deterministic – e.g., (0,1,0) then left multiplying by [Y|X] = column 2
- If all columns of [Y|X] are ~ equal then X and Y are ~ independent.

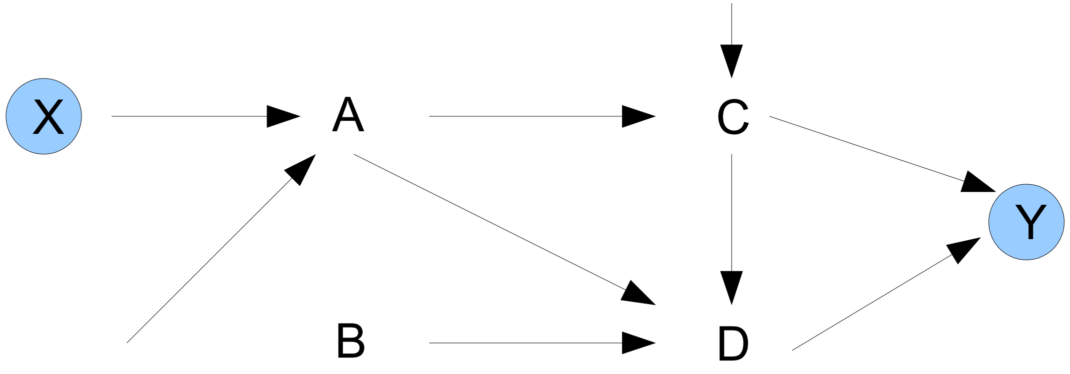
Conditional Independence

- A fundamental feature of Bayesian Networks – d-separation - is defined in terms of conditional independence.
- Conditional Indepence: $X \perp\!\!\!\perp Y \mid Z$ iff Prob(X|Y,Z) = Prob(X|Z)

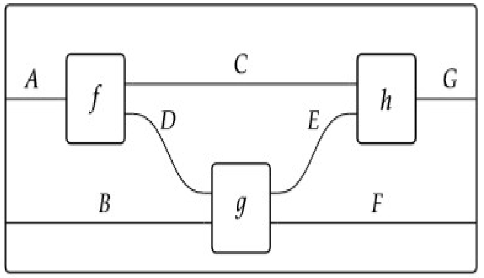
- Three important cases used in 'd-separation':
- <u>Fork</u>: $X \leftarrow Z \rightarrow Y$ Even if X, Y dep., then X,Y C.I. on Z
  - Example: X=Shoe Size, Z=Age of Child, Y=Reading Ability
- <u>Collider</u>: $X \rightarrow Z \leftarrow Y$ Even if X, Y ind., then X,Y C.D. on Z
  - Example: X=Tasty, Z=Choose to Eat, Y=Nutritious
- <u>Chain</u> <u>:</u> $X \rightarrow Z \rightarrow Y$ Even if X, Y dep., then X,Y C.I. on Z
  - Example: X=Fire, Z=Smoke, Y=Alarm

Causal Model Analysis

- Ladder of Causation (Pearl) – ... 'do-calculus'
  - Counterfactual – What if we had done X?
  - Intervention – What if we do X?
  - Association – What if we see X?
- Causal Effect Analysis [Y || X] (Causal Conditional)
  - What is the conditional effect on variable Y from an ancestor variable X which is on the causal path to Y?



Symmetric Monoidal Categories



- SMCs are categories* with a symmetric monoidal product
- Ideal for modeling serial and parallel processes
  - Serial processes using composition
  - Parallel processes using monoidal product
- For causal modeling we use
  - Serial for dependent causal relationships
  - Parallel for independent causal relationships
- Visualize as wiring (string) diagrams

* See the References slide.

Category Theory Framework

Causal Structure
DAG (V,A,s,t)

→

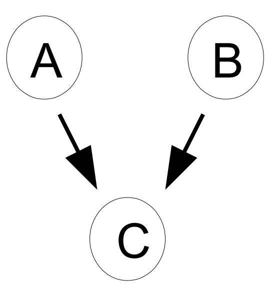
Causal Theory
(Hypothesis)
S.M.C.
(Syntax)

Model → Functor S.M.F.

Modeling
Codomain
S.M.C.
(Semantics)

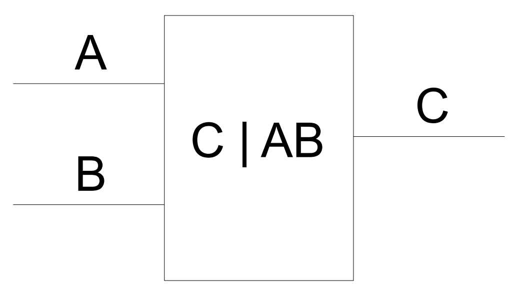
Nodes:
Symbols for variables

Arcs:
Causal relationships

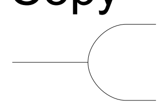


Objects: Strings in V

Morphisms:
Causal mechanisms
Comonoid maps



Copy



Discard

Stoch (Lawvere):

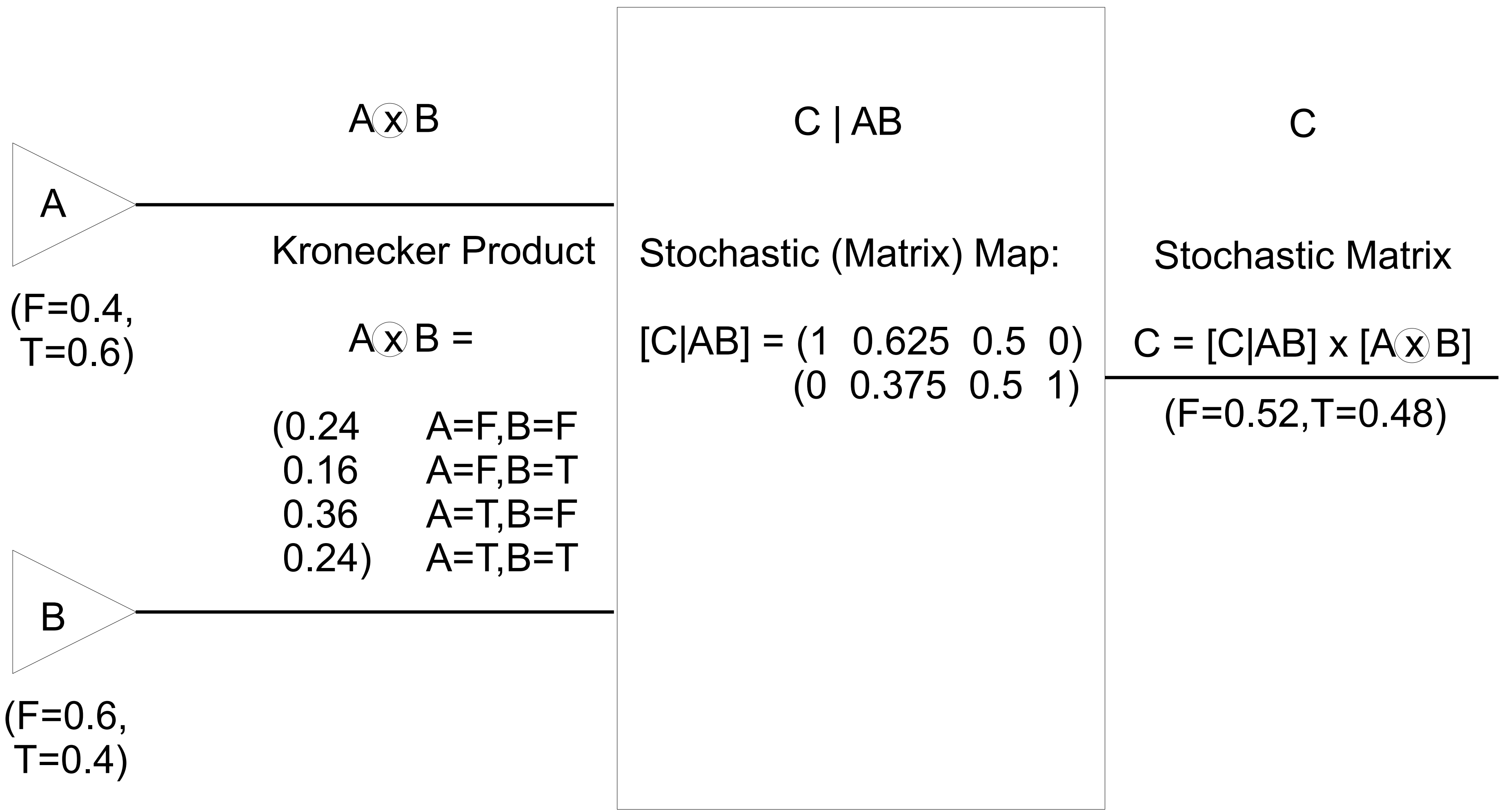
Objects:
Measurable spaces

Morphisms:
Stochastic maps

Subcategories:
FinStoch
CGStoch

Set, Rel, ...

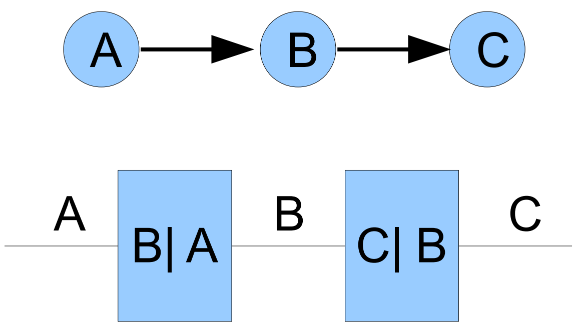
Causal Theory → FinStoch Example



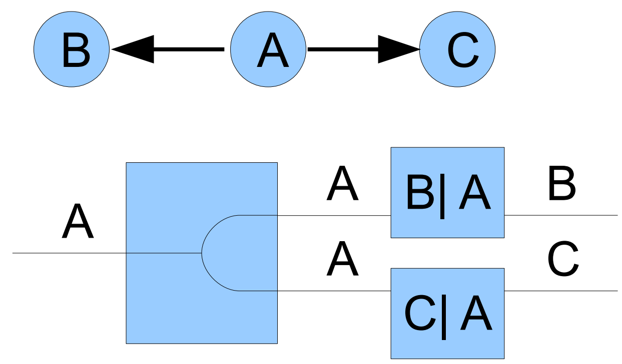
See slide 6 for more on stochastic matrices and conditional probability.

Causal Theory Models for Analysis

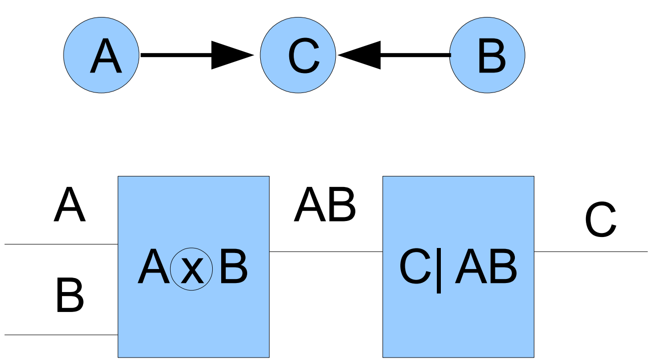
Chain



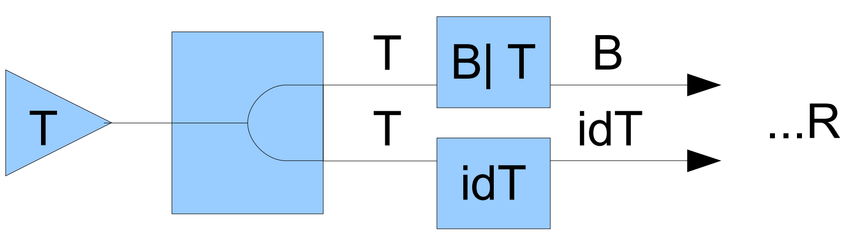
Fork



Collider



Causal Conditional



Causal Theory Models

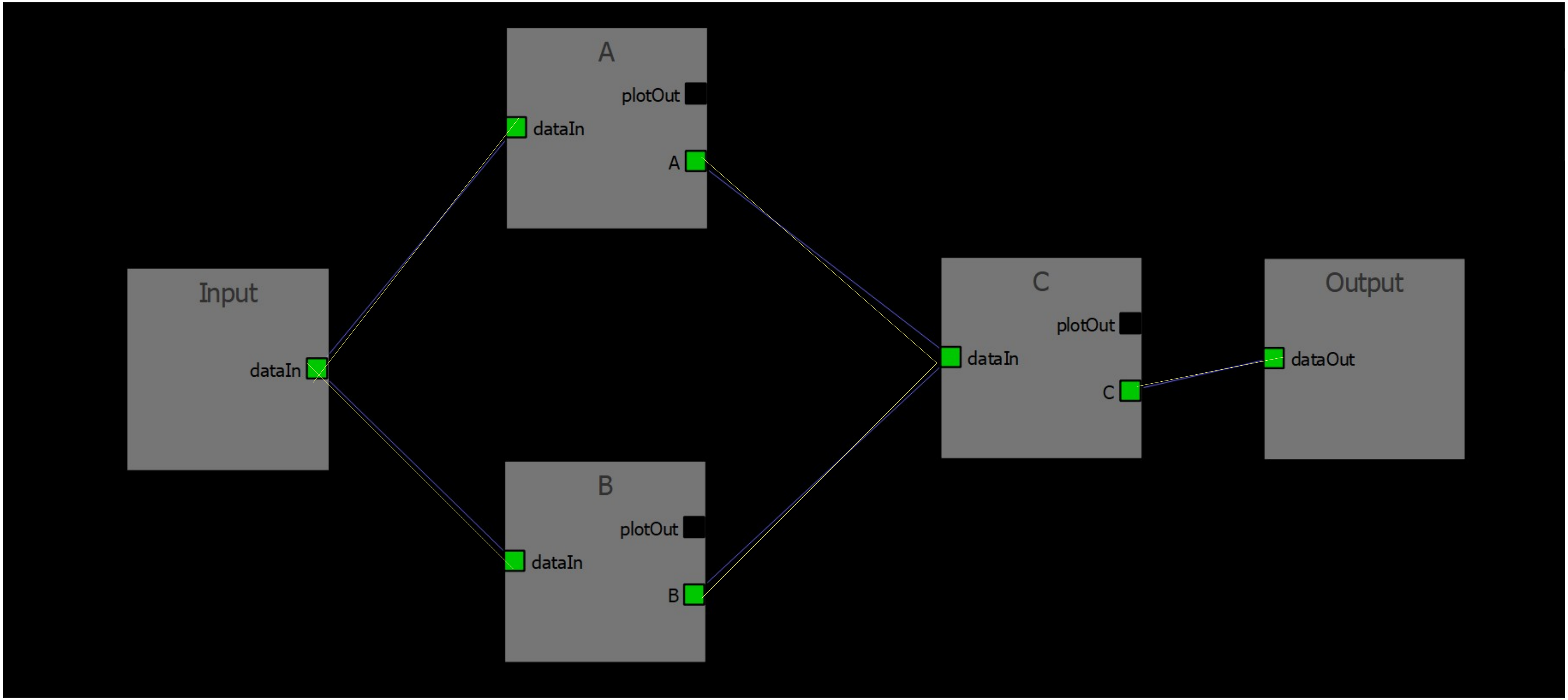
- Same functionality as Bayesian networks with...
- Visual representation of model variable relationships, but
- More functionality by being a monoidal category
- Causal graph generates factorization of joint probabiliity
- Causal model analysis capability – ladder, causal effects
- Generalized for continuous as well as discrete cases
- Can model with any symmetric monoidal category
- Model functors → natural transformations & comparisons
- Scalability – driven by size of causal graph
- Software implementation - next

Python Software Approach

- Desired functionality
  - User-drawn causal structure graph design
  - Load external data; generate maps automatically*
  - DAG-driven computing for probabilities and causal analysis
  - Generate symmetric monoidal category wiring diagram 'view'
  - Expandable to continuous case
- Libraries selected
  - Numpy, Scipy – ndarray, Kronecker product, A@B...
  - Pandas* – Data Frame, Groupby, Pivot
  - PyQt – multi-platform User Interface library
  - PyQtGraph – 'Flowchart' module on top of PyQt

* Tech note – We are using 'pandas' to generate stochastic maps and matrices.
These computations are coded to happen on the fly.

PyQtGraph:Flowchart Example



- Causal Graph flowchart view
- Nodes: Input, A, B, C, Output
- Terminals: dataIn, A, B, C, dataOut

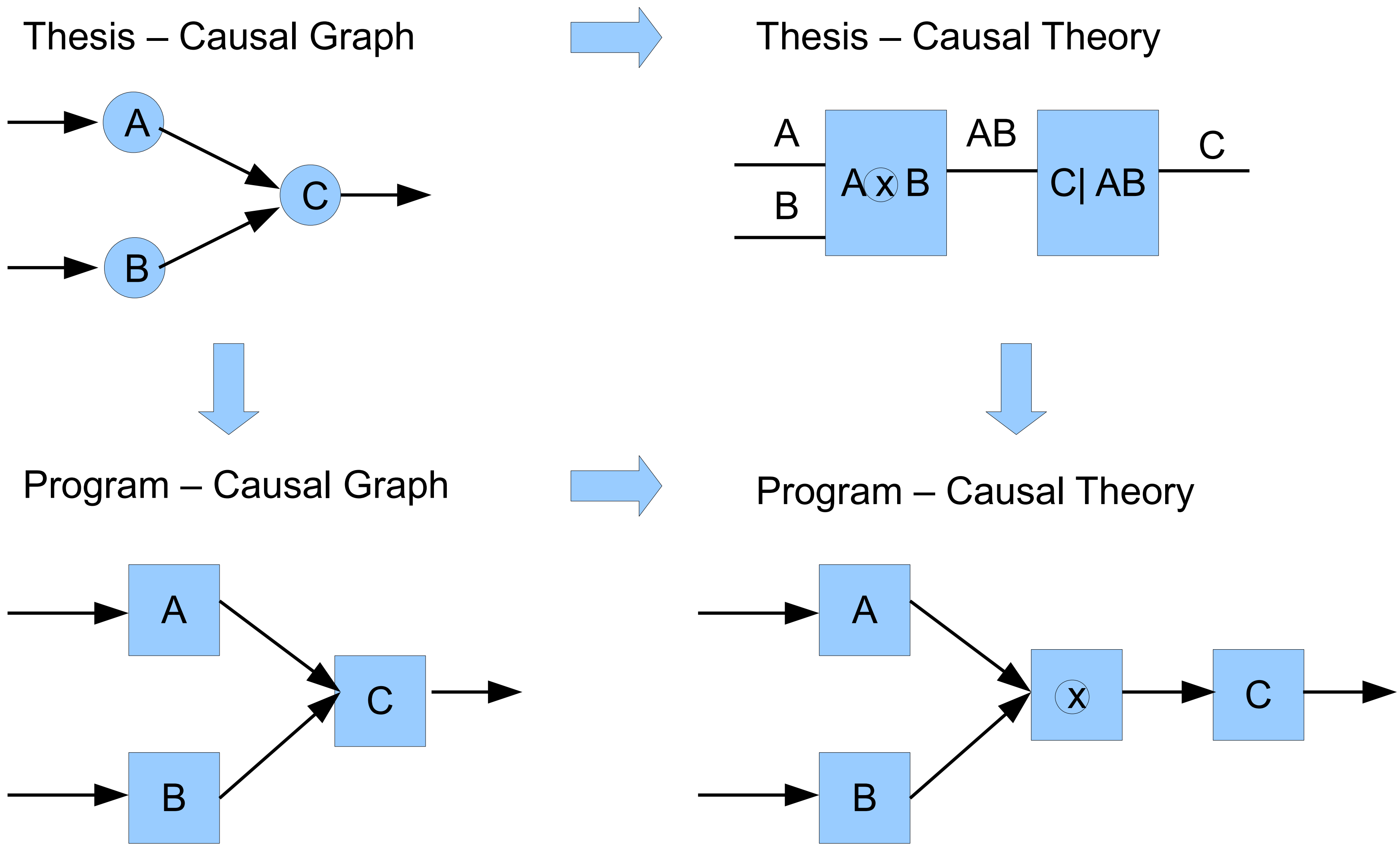
PyQtGraph:Flowchart Features

- Users can model and draw their own causal graphs
- Causal graphs can be generated programmatically
- Arcs connect nodes through 'out' and 'in' terminals
- 1-to-many and many-to-1 connections are allowed
- Nodes can have multiple terminals
- Programmers can create their own node types by making:
  - A 'process' function that runs for each node in order
  - Optional user interface node widget in the control panel
- Any object can be transmitted via terminals
- Flowchart process can run with UI refresh or without
- Flowcharts are nodes and can be embedded in flowcharts

Demo

Next Phase – Causal Theory View

Automatically make the Causal Theory (SMC) flowchart from the Causal Graph.



Next Phase - Programming

- Implement new node types for Causal Theory (SMC) view
- Generate the Causal Theory view automatically
- UI improvements including graph and table widget node views
- More robust exception handling and testing
- Volume testing for scalability
- Testing standard data sets and real world examples
- Database query capability – SQL to Data Frame
- Continuous node types
- In parallel – consider alternatives to PyQtGraph
- In parallel – drawing wiring diagrams for general SMCs

Applicability

- Causal modeling (by design)
- Supervised learning (to try out)
- Modeling analysis – pre/post modeling
- Modeling with aggregate data (e.g., BI, DW, cubes)
- Combining models for implementation
- Ad hoc models for hypothesis generation
- 80/20 data exploration for insights
- Causal modeling of model errors
  - Causes for poor predictability under certain conditions
  - Causes of false positives and false negatives
  - Causes of model drift over time

References

Brendan Fong, "Causal Theories: A Categorical Perspective on Bayesian Networks".

Brendan Fong and David I. Spivak, Seven Sketches in Compositionality

Tom Leinster, Basic Category Theory

Judea Pearl, Causality: Models, Reasoning and Inference

Emily Riehl, Category Theory in Context

David I. Spivak, Category Theory for Scientists