

# Causal Inference and Invariance

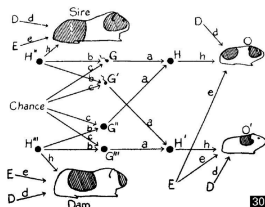
Charles Zheng and Qingyuan Zhao

Stanford University

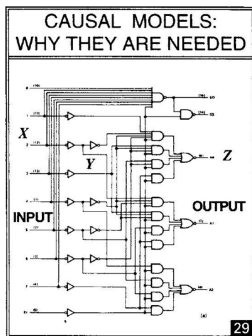
February 13, 2016

(Part 1/2)

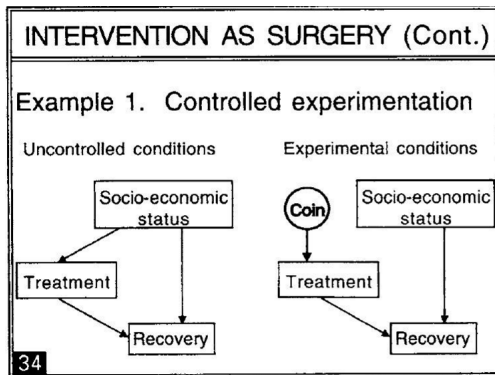
# Understanding = cause and effect



30

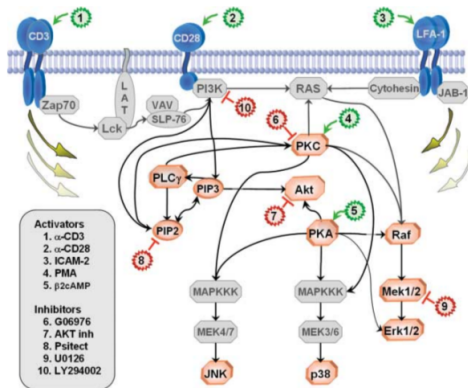


29



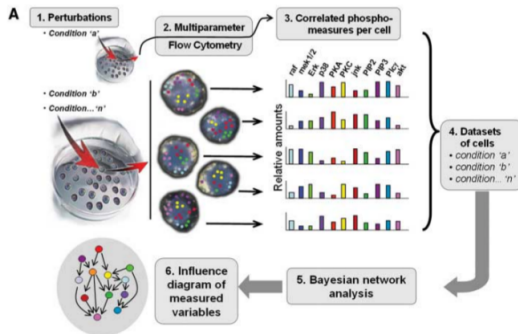
34

# A hot application: systems biology



- Causal relationships = *chemical interactions*.
- Experimenters *intervene* by injecting *activators* and *inhibitors*.

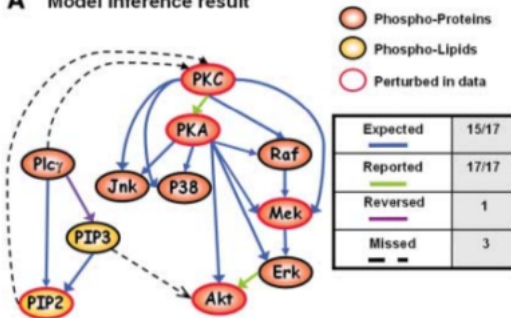
# Protein signalling data



- Flow cytometry data from Sachs et al. *Science*, 2005.
- 1 observational data set + 9 interventions.

# Putative causal model

**A** Model inference result



- Causal inference applied to observational + interventional data.
- Recovered most of the known interactions.

# The many facets of causality

- *Philosophy*. What is causality? How do we learn about cause and effect? *Aristotle, Hume*.
- *Computer science*. Can we build an artificial intelligence which reasons like humans? *Judea Pearl*.
- *Social science*. What influences an individual's life choices?
- *Law*. Whose “fault” is it??
- *Statistics*. Answering the above questions using data!

- *Estimating causal effects from data.* Can we predict a causal effect based on observational or experimental data? E.g. effect of a medical treatment based on clinical trial data? Motivation for potential outcomes approach developed by Rubin, etc.
- *Bayesian networks/structure learning from data.* Can we model multivariate relationships using a network structure? Networks *can be* given causal interpretation, but causal inference is not the only motivation. Motivation for graphical lasso.

## Principles of Causal Inference



# Principles of Causal Inference

Best framework: *Graphical approach* pioneered by Judea Pearl.

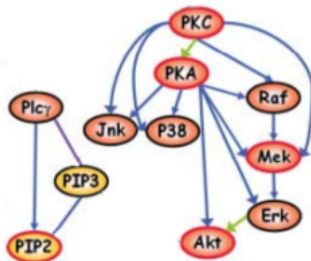
**TYPICAL DERIVATION IN CAUSAL CALCULUS**

```

graph LR
    S((Smoking)) --> T((Tar))
    T --> C((Cancer))
    S --> C
            
```

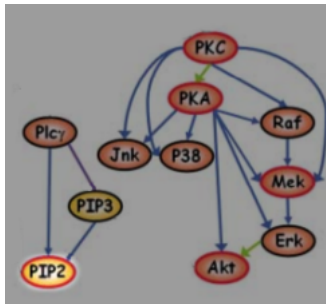
|   |                    |
|---|--------------------|
| $P(c \mid do\{s\}) = \sum_t P(c \mid do\{s\}, t) P(t \mid do\{s\})$       | Probability Axioms |
| $= \sum_t P(c \mid do\{s\}, do\{t\}) P(t \mid do\{s\})$                   | Rule 2             |
| $= \sum_t P(c \mid do\{s\}, do\{t\}) P(t \mid s)$                         | Rule 2             |
| $= \sum_t P(c \mid do\{t\}) P(t \mid s)$                                  | Rule 3             |
| $= \sum_{t'} \sum_t P(c \mid do\{t\}, s') P(s' \mid do\{t\}) P(t \mid s)$ | Probability Axioms |
| $= \sum_{t'} \sum_t P(c \mid t, s') P(s' \mid do\{t\}) P(t \mid s)$       | Rule 2             |
| <b>47</b> $= \sum_{t'} \sum_t P(c \mid t, s') P(s') P(t \mid s)$          | Rule 3             |

# Graphs: nodes and vertices



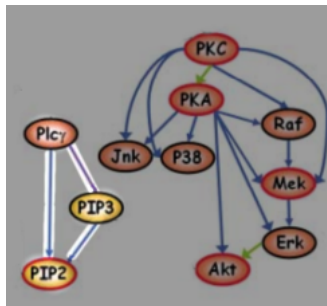
- Each variable in the dataset is given a *node*.
- Arrows indicate which variables *cause* which other variables. (Parents → children).
- Undirected or bidirected edges = correlation due to mutual causation or latent common causes.

# Causality and experiments



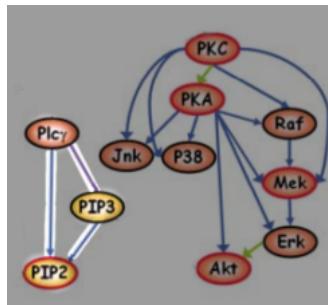
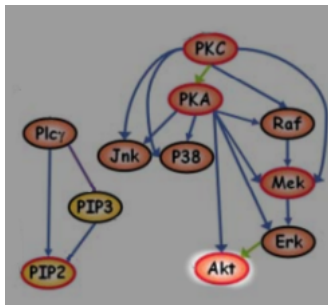
*Intervening* on variables in the system causes the distribution to change.

# Causality and experiments



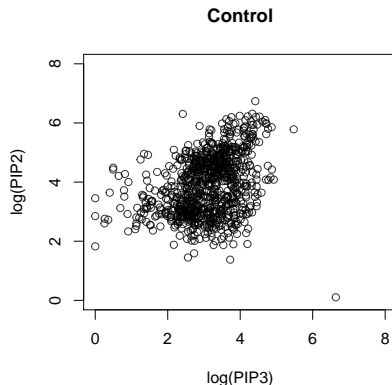
- Not every variable will be affected by the intervention!
- Following the arrows tells you *which* variables which are affected.

# Principle I: Which variables are affected.



- If we *inhibit* Akt, no other variables should be affected.
- If we *inhibit* PIP2, then we may not only change the distribution of PIP2, but also PIP3.

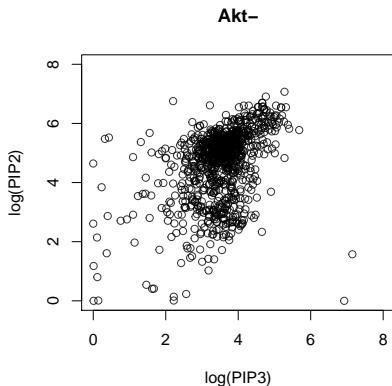
# Principle I: Which variables are affected.



Looking at Sachs data.

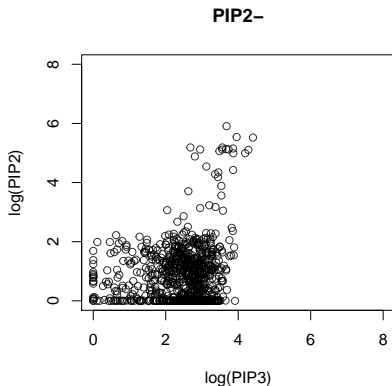
Joint distribution of PIP2 and PIP3 in the “control” case.

# Principle I: Which variables are affected.



Joint distribution of PIP2 and PIP3 when we intervene on Akt.

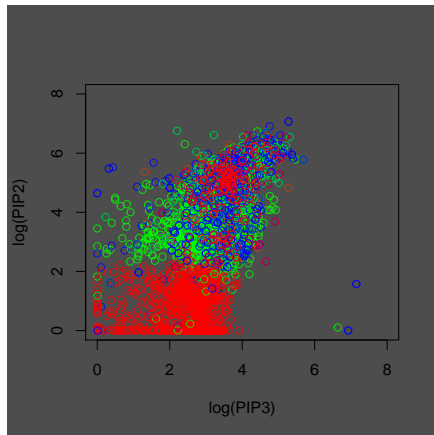
# Principle I: Which variables are affected.



Joint distribution of PIP2 and PIP3 when we intervene on PIP2.



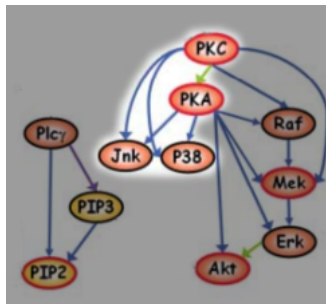
# Principle I: Which variables are affected.



**Control**, **PIP2-**, **Akt-**

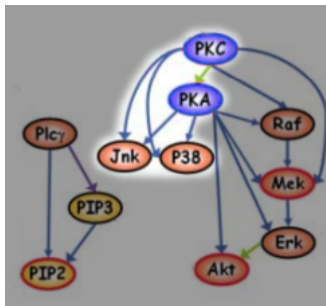
Intervening on PIP2 also affects the distribution of PIP3, while intervening on Akt does not (drastically) change the distribution.

# Principle II: Conditional independence.



- Surprisingly, the structure of the causal graph implies certain *conditional independence* relationships.
- This allows the potential to infer causal relationships from observational data.

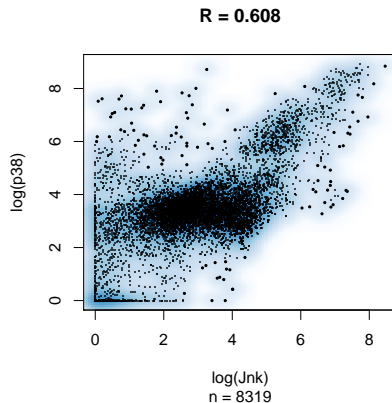
## Principle II: Conditional independence.



- Two variables are independent conditional on their common parents.
- Conditioning on PKC and PKA, Jnk and p38 should be independent.

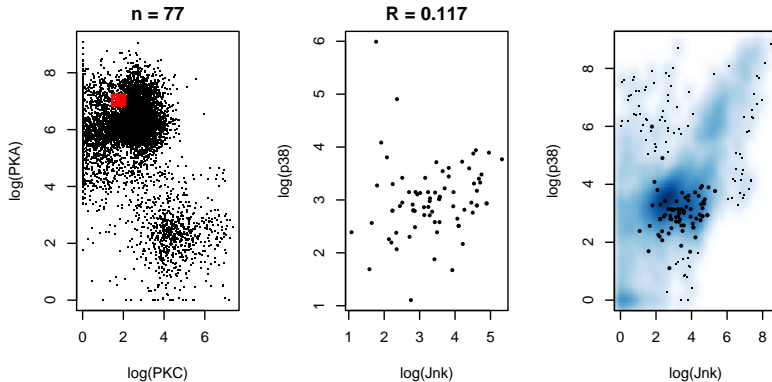


## Principle II: Conditional independence.



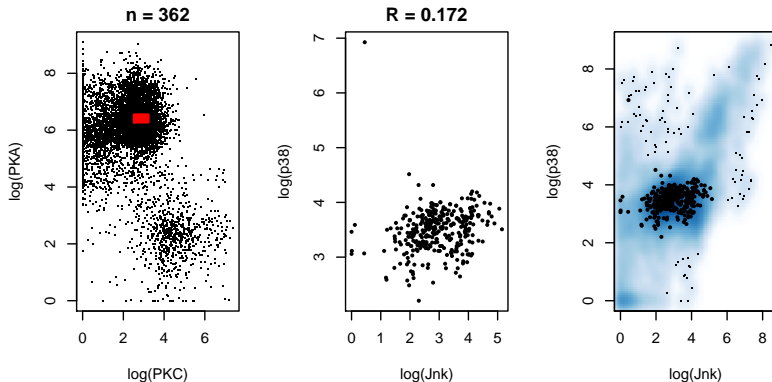
Marginally, p38 and Jnk are correlated.

# Principle II: Conditional independence.



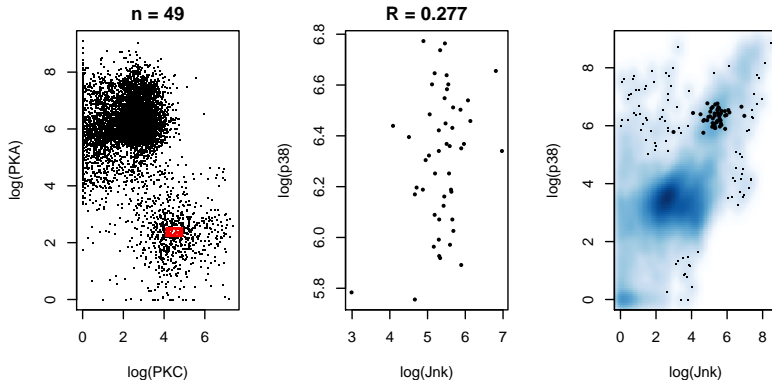
We can't condition on PKA and PKC since the data is continuous. But, conditioning on small windows seems to reduce association.

# Principle II: Conditional independence.



*Left:* We condition on (PKA, PKC) to lie within the indicated window.  
*Center:* Conditional joint distribution of (Jnk, p38). *Right:* Conditional joint distribution, overlaid on marginal distribution.

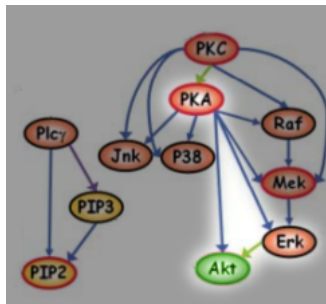
# Principle II: Conditional independence.



PKA and PKC *explain away* some (if not all) of the association between Jnk and p38. (Recall that  $R = 0.608$  marginally.)

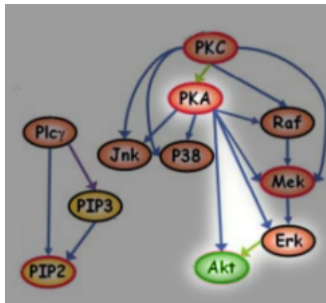


# Principle III: Predictive invariance



- The conditional distribution  $\Pr[Akt|PKA, Erk]$  is invariant to interventions applied to other variables.
- Therefore, the optimal rule for predicting  $\hat{Akt}(PKA, Erk)$  is invariant as well.

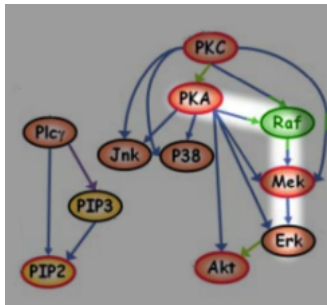
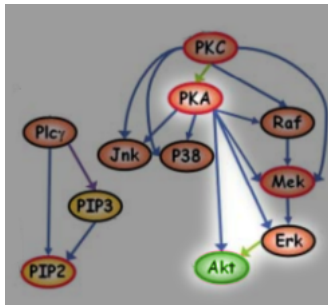
# Principle III: Predictive invariance



$\{PKA, Erk\}$  is an “invariant set” for  $Akt$  since:

- It includes all of the “direct” causes of  $Akt$  in the graph.
- It doesn’t include any variables caused by  $Akt$ .

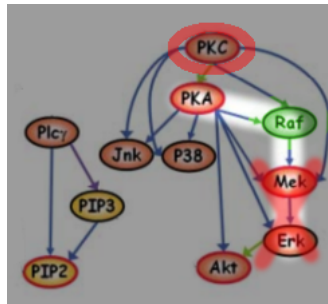
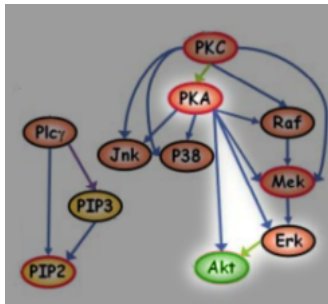
# Principle III: Predictive invariance



In contrast,  $\{PKA, Mek, Erk\}$  is *not* an invariant set for *Raf* since:

- .
- .

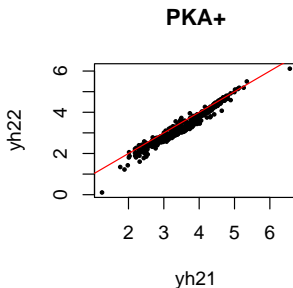
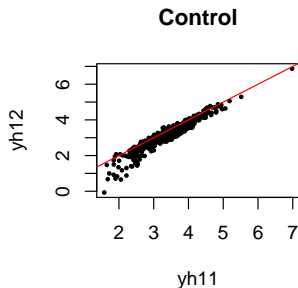
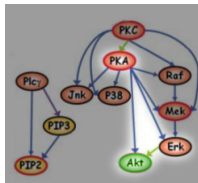
# Principle III: Predictive invariance



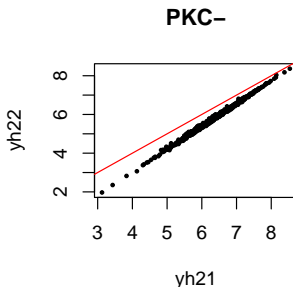
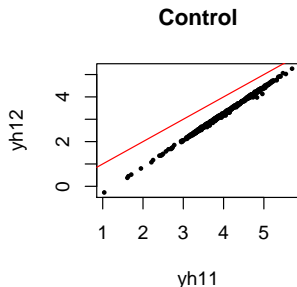
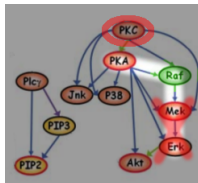
In contrast,  $\{PKA, Mek, Erk\}$  is *not* an invariant set for *Raf* since:

- It is missing a direct cause of *Raf*.
- It contains variables which are caused by *Raf*.

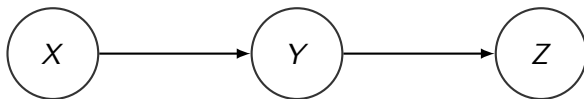
$\{PKA, Erk\}$  is an invariant set for *Akt*.



$\{PKA, Mek, Erf\}$  is not an invariant set for *Raf*.

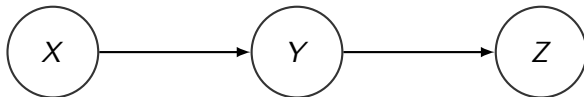


# Predictive Invariance: Example



- Suppose we are trying to predict  $Y$ .
- $X \sim N(0, a)$ .
- $Y|X \sim N(X, b)$ .
- $Z|Y \sim N(Y, c)$ .

# Predictive Invariance: Example



$$X \sim N(0, a), \quad Y|X \sim N(X, b), \quad Z|Y \sim N(Y, c).$$

- We can intervene by adding noise to  $X = \text{changing } a \rightarrow a'$ .
- Intervene by injecting noise to  $Z = \text{changing } c \rightarrow c'$ .
- Consider a linear model which predicts  $Y$  given  $X$  and  $Z$ .
- *Is the optimal prediction rule invariant under intervention?*



# Predictive Invariance: Example

$$X \rightarrow Y \rightarrow Z$$

$$X \sim N(0, a), \quad Y|X \sim N(X, b), \quad Z|Y \sim N(Y, c).$$

The joint distribution is

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & a & a \\ a & a+b & a+b \\ a & a+b & a+b+c \end{bmatrix} \right)$$

The optimal prediction rule is given by

$$\mathbf{E}[Y|X, Z] = \mu_Y + \Sigma_{Y,XZ} \Sigma_{XZ}^{-1} (X - \mu_X, Z - \mu_Z) = \frac{c}{b+c} X + \frac{b}{b+c} Z.$$

# Predictive Invariance: Example

$$X \sim N(0, a), \quad Y|X \sim N(X, b), \quad Z|Y \sim N(Y, c).$$

Optimal prediction rule:

$$\mathbf{E}[Y|X, Z] = \underbrace{\frac{c}{b+c}}_{\beta_X} X + \underbrace{\frac{b}{b+c}}_{\beta_Z} Z.$$

i.e.  $Y$  is a weighted average of  $X$  and  $Z$  ( $\beta_X + \beta_Z = 1$ ).

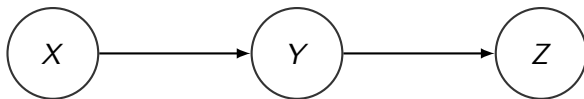
- Imagine  $c$  is very small, i.e.  $Z = Y + \text{tiny noise}$ . Then  $Z$  is a great predictor of  $Y$ !  $\beta_Z \approx 1$ .
- Conversely, if  $b$  is small, that means  $Y = X + \text{tiny noise}$ .  $\beta_X \approx 1$ .
- If  $b = c$ , then  $\beta_X = \beta_Z = 1/2$ .

# Predictive Invariance: Example

But is the OLS predictive rule invariant?

$$\mathbf{E}[Y|X, Z] = \frac{c}{b+c}X + \frac{b}{b+c}Z.$$

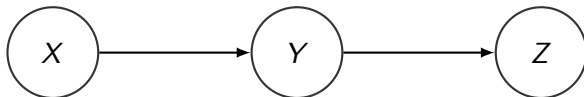
If we intervene on  $Z$ , changing  $c$  to  $c'$ , the OLS coefficients change too. The model is not invariant.



*“Real-life” example.*  $X$  = how much you weigh?  $Y$  = how many bagels you eat every day?  $Z$  = how many pull-ups you can do?

$Z$  is a good predictor of  $Y$ , unless you “intervene” by offering a \$100 prize for doing 10 pull-ups.

# Predictive Invariance: Example



- In contrast, consider predicting  $Y$  using *only*  $X$ .
- $\{X\}$  is an invariant set for  $Y$  because it contains all direct parents and no children of  $Y$ .
- Indeed,

$$\mathbf{E}[Y|X] = \frac{\text{Cov}(Y, X)}{\text{Cov}(X)} X = \frac{a}{a} X = X.$$

The OLS coefficient, 1, does not depend on  $a$  or  $c$ , and hence is *invariant* under interventions.

- *Exercise.* Is  $\{Z\}$  an invariant set for  $Y$ ?

# Overview: Principles of Causal Inference

Causal relationships in a system represented by a graph. The graph tells you:

- I. which variables are affected by an intervention.
- II. what conditional independence relationships exist in the joint distribution.
- III. which sets of predictors and responses will have “invariant” optimal predictive rules.

Section 3

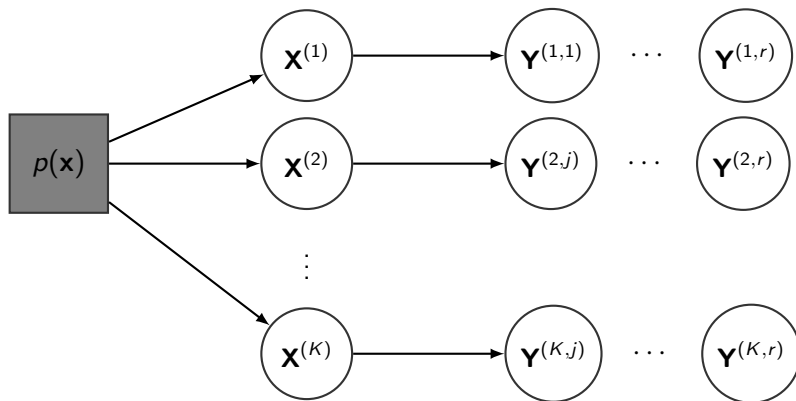
Statistical Methods

Section 4

Conclusions

# Look! A diagram!

Don't put this in the final presentation.



Legend:  $K = \{ \text{2}, \text{9}, \text{99}, \text{999} \}$