

CS4220 Machine Learning

Probabilistic Models

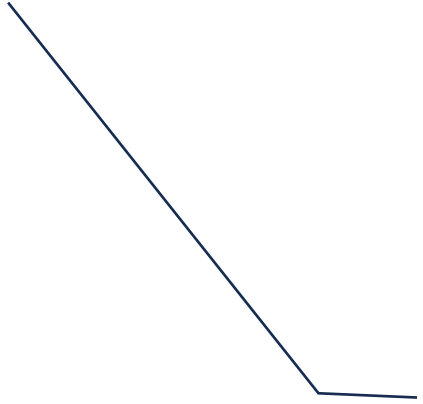
Monday, 4 December 2023

Merve Gürel



So far:

- Parametric/non-parametric classifiers
- Linear Regression, classification
- Nonlinear transformation
- Regularization
- Bias-variance tradeoff
- Cross-validation
- Curse of dimensionality



Related ones for today:
Week 2: Maximum
Likelihood, Maximum a
posteriori, regularization

This Week: Probabilistic Models

How do we incorporate prior knowledge?

- Bayesian Inference
- Bayesian Networks
(Probabilistic Graphical Models)

Credits. The remaining content is adapted from
Jesse Krijthe's slide deck with minor changes

MODIFIED BAYES' THEOREM:

$$P(H|X) = P(H) \times \left(1 + P(C) \times \left(\frac{P(X|H)}{P(X)} - 1 \right) \right)$$

H: HYPOTHESIS

X: OBSERVATION

P(H): PRIOR PROBABILITY THAT H IS TRUE

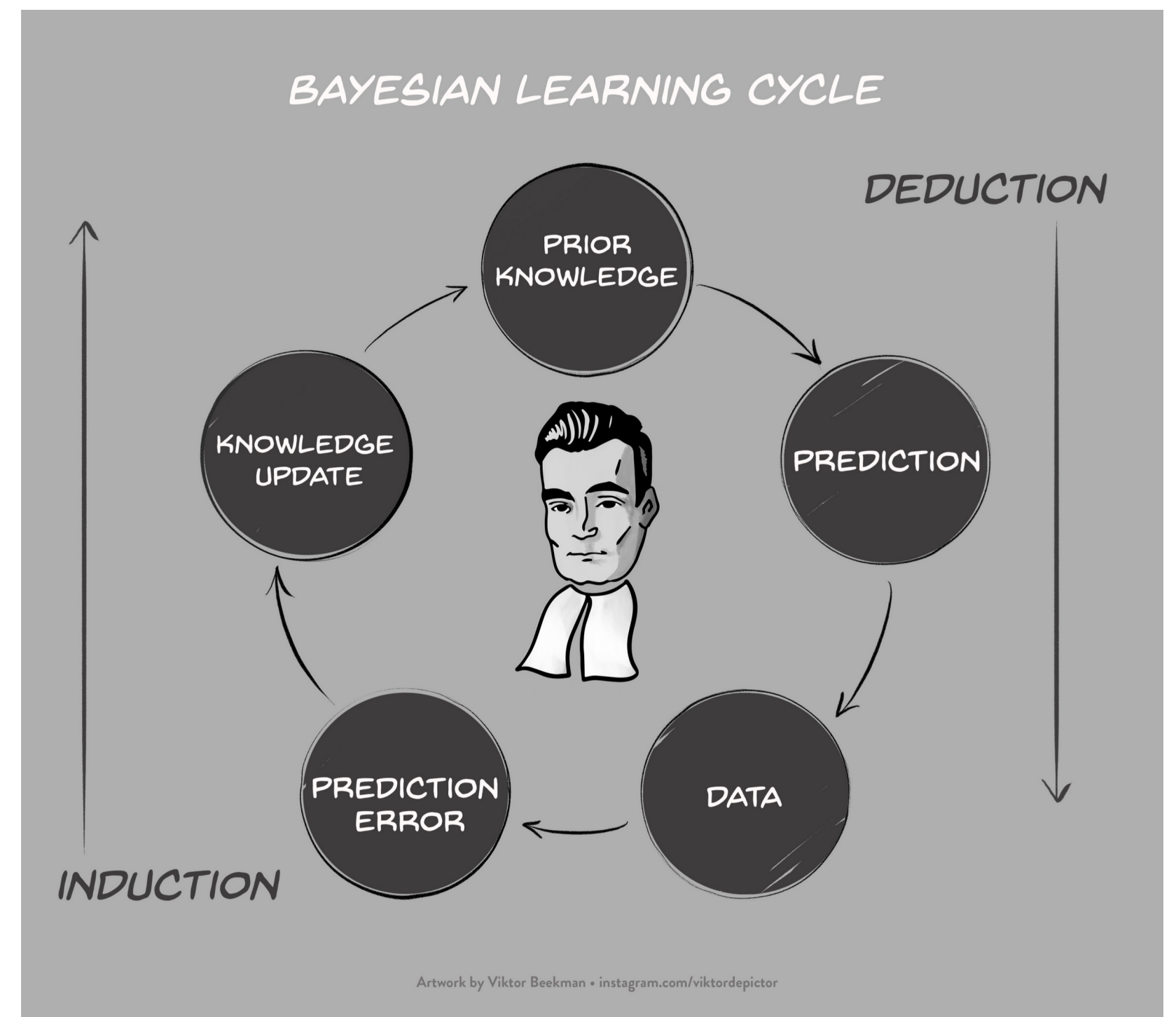
P(X): PRIOR PROBABILITY OF OBSERVING X

P(C): PROBABILITY THAT YOU'RE USING
BAYESIAN STATISTICS CORRECTLY

<https://xkcd.com/2059/>

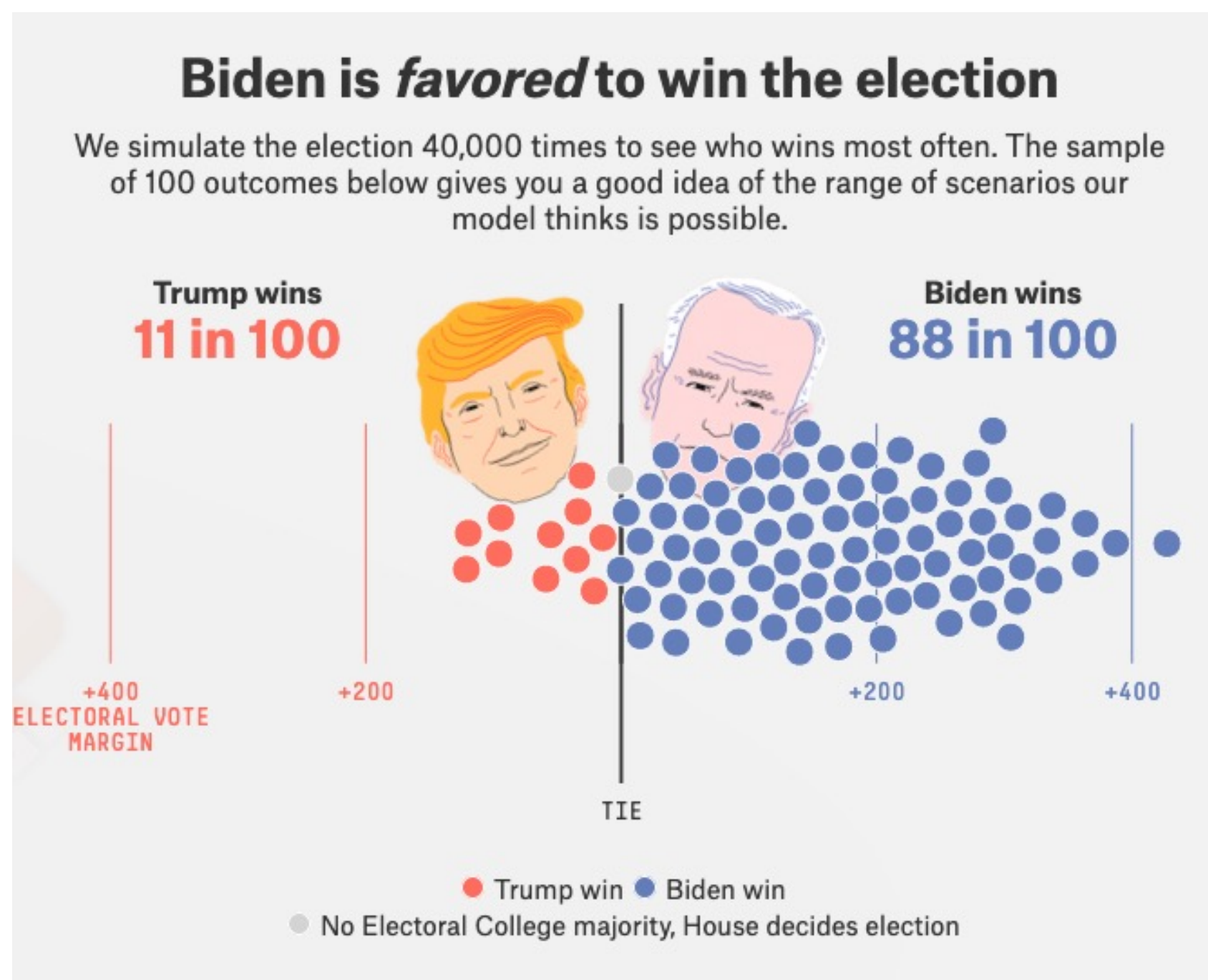
Why is Bayesian inference important?

- Principled way to incorporate new information, sequential updating
- Automatic complexity control
- Automatic uncertainty estimates

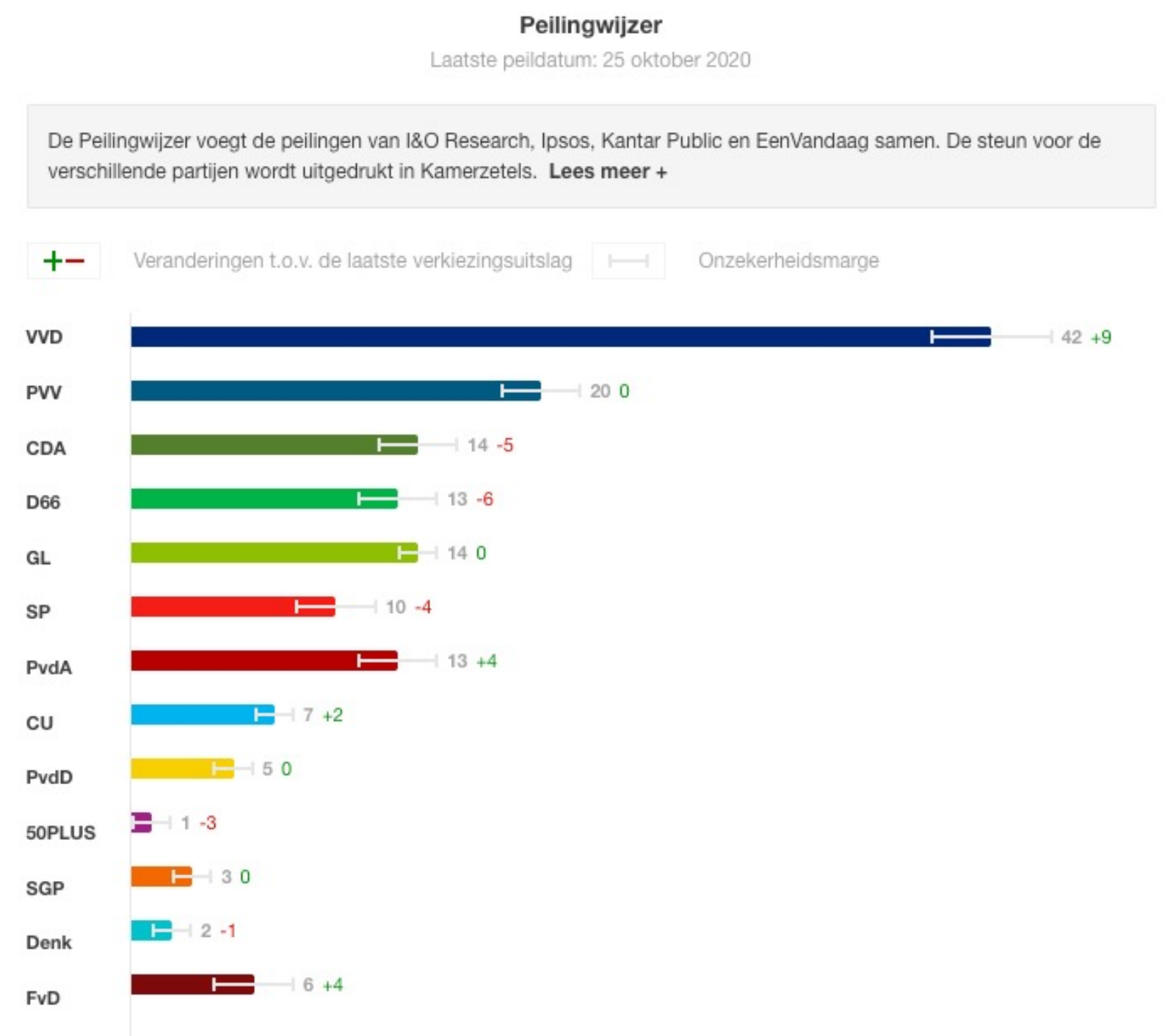


<https://www.bayesianspectacles.org/a-bayesian-perspective-on-the-proposed-fda-guidelines-for-adaptive-clinical-trials/>

Application: Election Forecasting



fivethirtyeight.com



peilingwijzer.nl

Application: BioNTech-Pfizer Trial

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	(N=18,198)		(N=18,325)			
	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	(N=19,965)		(N=20,172)			
	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.

† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

Polack, Fernando P., Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, et al. "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine." *New England Journal of Medicine* 383, no. 27 (December 31, 2020): 2603–15. <https://doi.org/10.1056/NEJMoa2034577>.

Bayesian Inference

$$\max_{\theta} p_{\theta}(D)$$

MAXIMUM LIKELIHOOD

Random Variable



$$\max_{\theta} p(\theta|D)$$

MAXIMUM A POSTERIORI

Not an estimator yet



$$p(\theta|D)$$

POSTERIOR

Bayesian Inference

Treat the parameters of the model as random variables and update their distributions when observing data

$$p(\theta|D)$$

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(\theta|D) = \frac{\overset{\text{Likelihood}}{p(D|\theta)} \overset{\text{Prior}}{p(\theta)}}{p(D)}$$

Linear Regression with Gaussian Error

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})$$

- \mathbf{X} be the $N \times M$ feature matrix
- \mathbf{w} be the $M \times 1$ weight vector
- Assume we know the measurement noise

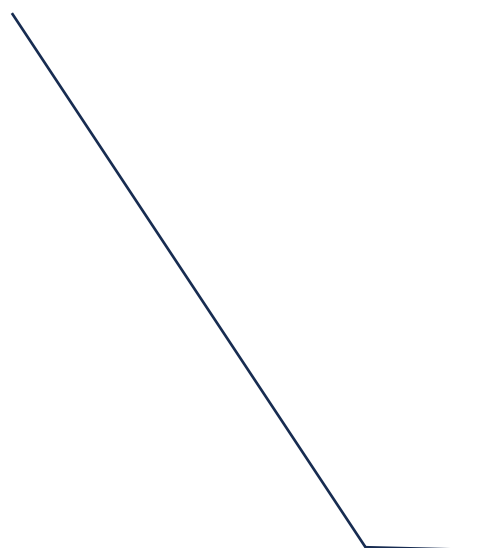
Bayesian Linear Regression

Likelihood
given weights

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbf{I})$$

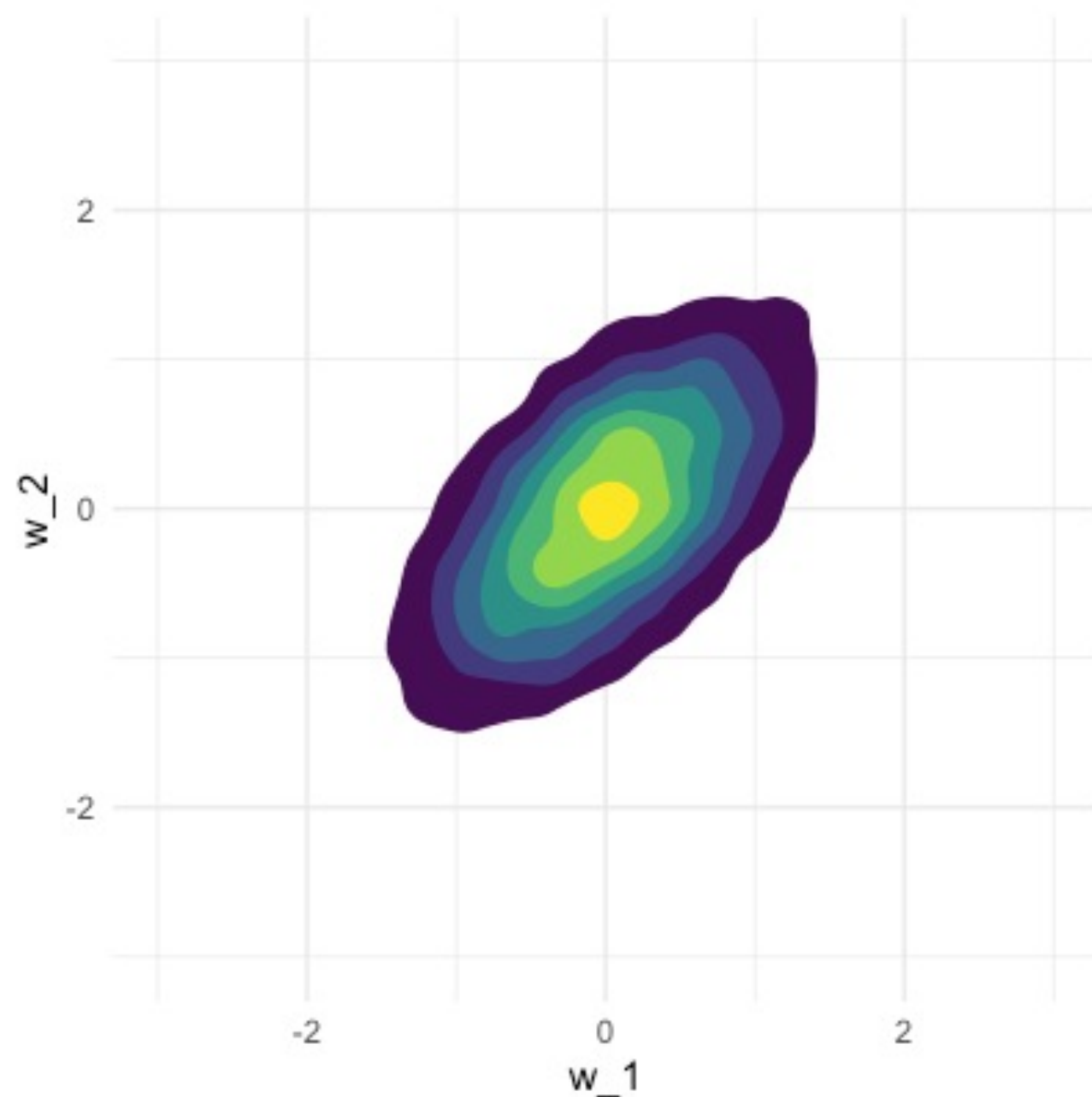


Terminology alert: Conjugate prior means that prior and posterior are from the same distribution family. If that's the case, the prior is called “conjugate prior” to the likelihood function

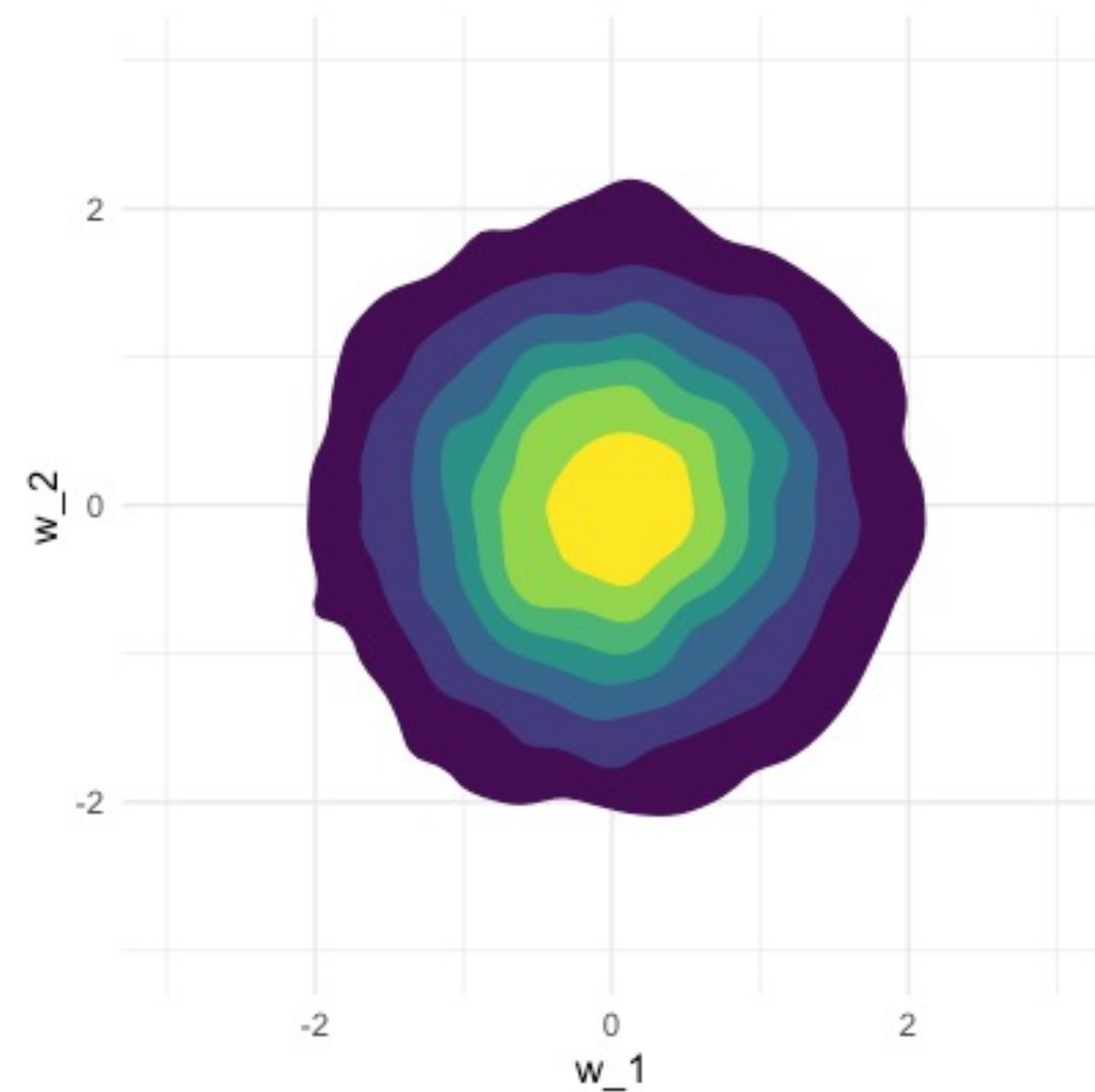
Check this out for Bayesian polynomial regression:
<https://jkrijthe.shinyapps.io/bayesian-poly/>

What does the prior probability look like?

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbf{I})$$



Plot A



Plot B

Getting to the Posterior

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} = \frac{\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})}{Z}$$
$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})$$

$$\begin{aligned}\log p(\mathbf{w}|\mathbf{y}) &= C_1 - \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + C_2 - \frac{1}{\alpha}\mathbf{w}^T\mathbf{w} - Z \\ &= \frac{2}{\sigma^2}\mathbf{y}^T\mathbf{X}\mathbf{w} - \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \frac{1}{\alpha}\mathbf{w}^T\mathbf{w} + C_1 + C_2 - Z - \frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y}\end{aligned}$$

Perhaps posterior is normal? Form: $-(\mathbf{w} - \mathbf{m})^T \mathbf{S}(\mathbf{w} - \mathbf{m})$

$$\mathbf{S} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha}\mathbb{I}$$

$$\mathbf{m} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha}\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{w} \mid \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha}\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}, \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha}\mathbb{I}\right)^{-1}\right)$$

Posterior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I}) \xrightarrow{\text{Observe Data}} p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}^{-1})$$
$$\mathbf{m} = \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \frac{1}{\alpha} \mathbb{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$\mathbf{S} = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \frac{1}{\alpha} \mathbb{I}$$

- What is the MAP estimate?
- Credible intervals vs. Confidence intervals

Posterior Predictive Distribution

$$\begin{aligned} p(y^{\text{new}}|\mathbf{y}) &= \int p(y^{\text{new}}|\mathbf{w})p(\mathbf{w}|\mathbf{y})d\mathbf{w} \\ &= \mathcal{N}(y_{\text{new}}|\mathbf{x}_{\text{new}}^{\top}\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^{\top}\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{X}^{\top}\mathbf{y}, \mathbf{x}_{\text{new}}^{\top}(\frac{1}{\sigma^2}\mathbf{X}^{\top}\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{x}_{\text{new}} + \sigma^2) \end{aligned}$$

- What about decisions?
 - Use decision theory: minimize the expected loss.
 - Why are all these estimates so similar here? (symmetry)
- Important use-case: model checking: posterior predictive checking

Arbitrary distributions and losses

- Not every distribution is conjugate and symmetric
- So the minimal expected loss will not always coincide with the MAP solution
- See examples in exercise session

Making Predictions

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Conjugacy (mainly considered here): closed form solutions to calculate posterior
- Sampling (Gibbs, Metropolis-Hastings, Hamiltonian Monte Carlo, Particle Filters): generate samples from the posterior
- (Variational) Approximations: approximate the posterior using a simpler distribution

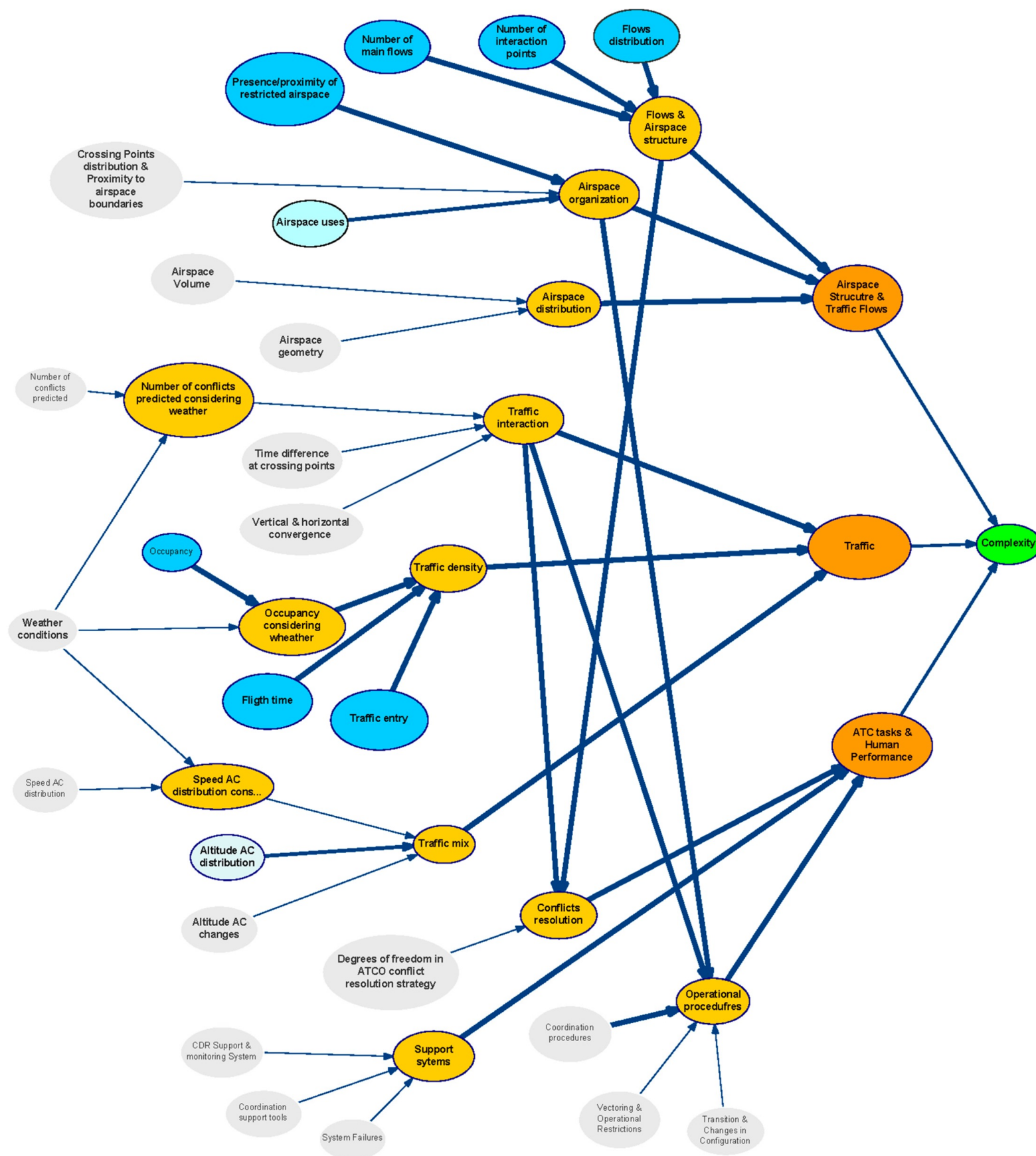
Interpretations of Bayesian Inference

- Subjective probability (prior knowledge)
- Non-Bayesian: way to construct models with good (frequentist) properties

Concluding Remarks

- Bayesian inference treats unknown parameters as random variables. We update these variables based on the observed data
- By taking into account different sources of information, and incorporating all uncertainties in the inference, we may be able to build good predictions/estimates

(Directed) Probabilistic Graphical Models *aka* Bayesian Networks



- Each node represents a random variable
- Edges express probabilistic relationship between these variables
- Good for expressing dependencies in data
- Visualization of structure
- Insights to the model: conditional independences
- Learning and inference in complex models with graphical manipulations

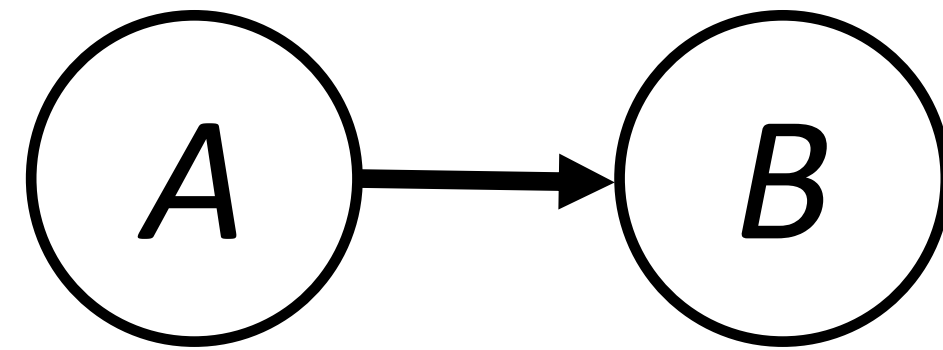
Modeling the Joint Distribution

$$P(X_1, X_2, \dots, X_M)$$

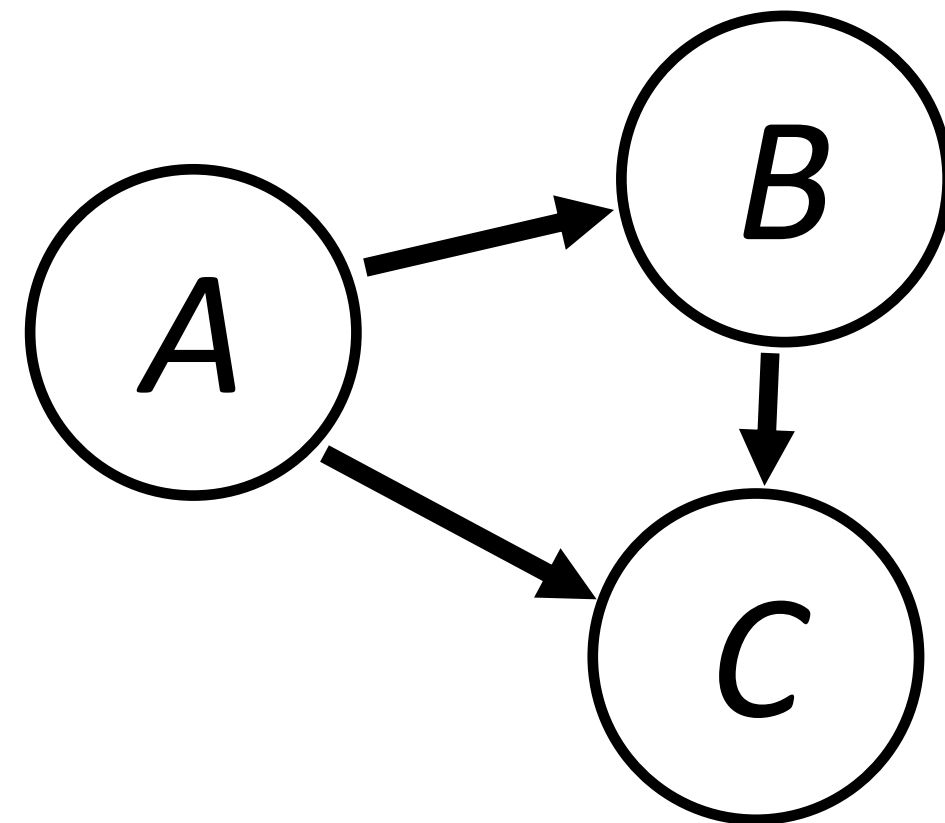
- Discrete: $D^M - 1$ values?
- Can we decompose it somehow to
 - ...represent it more efficiently?
 - ...reason about relations between variables?
 - ...do efficient inference?
 - ...incorporate our knowledge of P ?

Factorizing a PDF

$$P(A,B) = P(A)P(B|A)$$

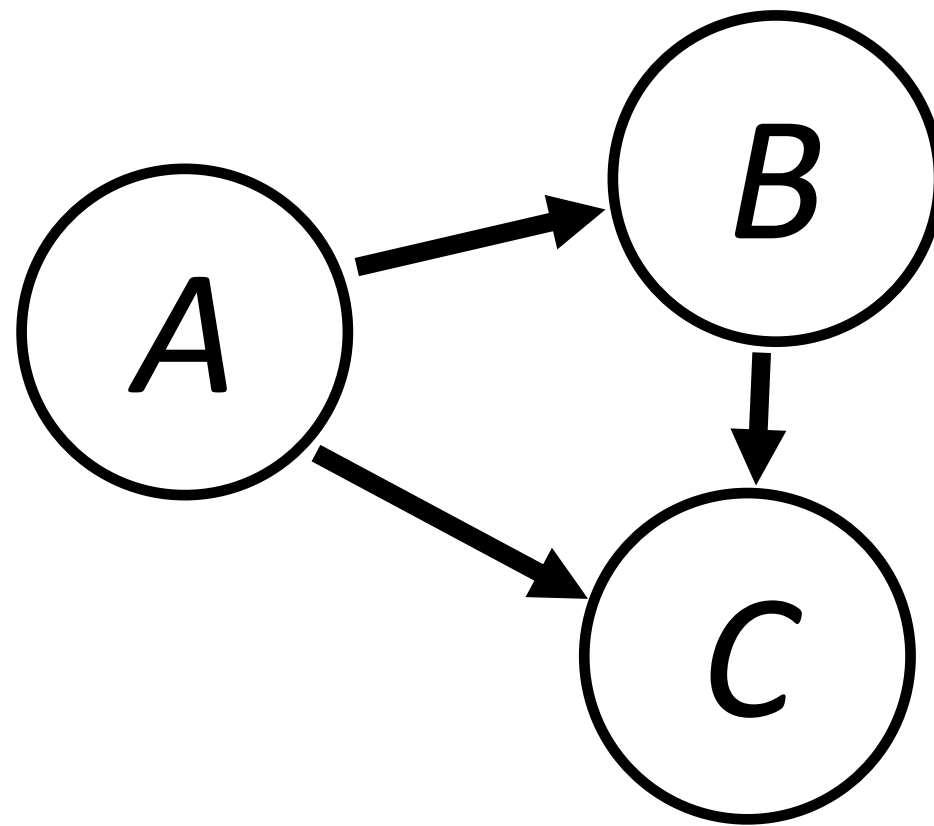


$$P(A,B,C) = P(A) P(B|A) P(C|A,B)$$



- Graphical representation: Directed Acyclic Graph (DAG)
- Graph captures the independence structures in the whole distribution
- Can we simplify the model by assuming some links are not there?

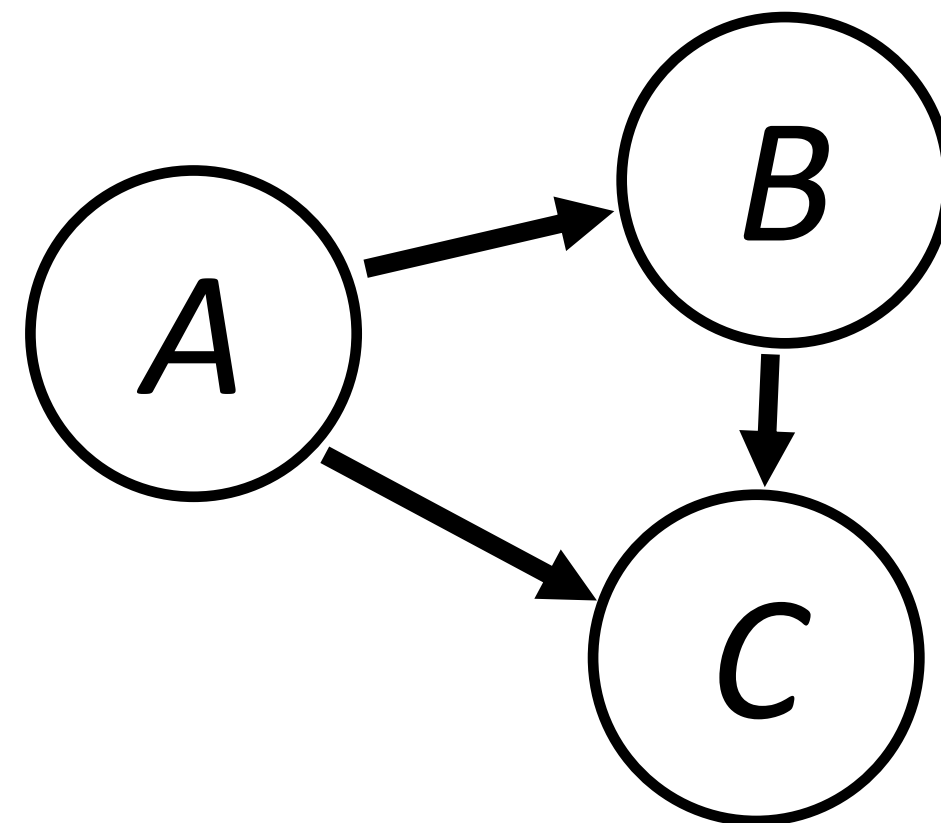
Directed Graphs: Bayesian Networks



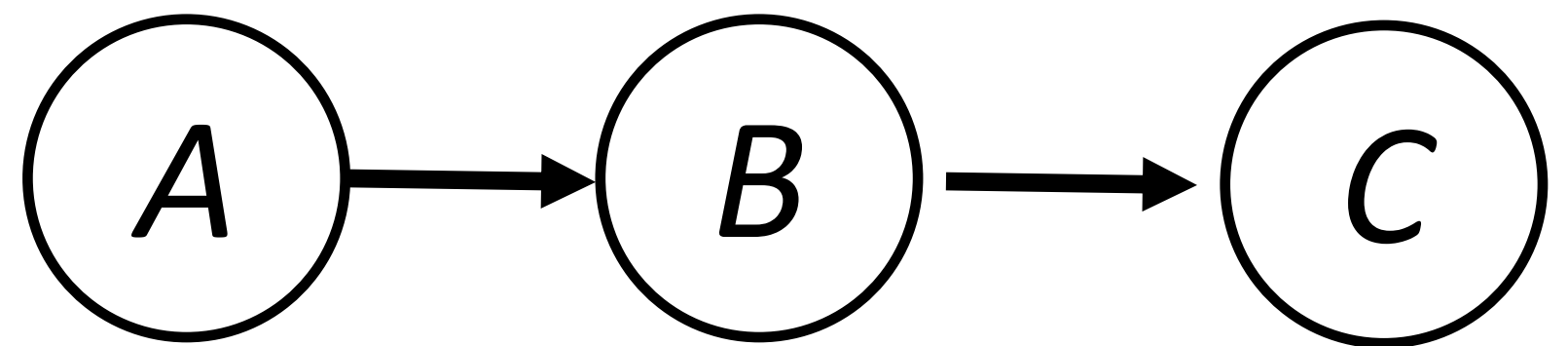
- Undirected vs. Directed
- Bayesian because they use Bayes rule, not (necessarily) because parameters are represented by random variables.

Exploiting Conditional Independence: Removing Links

Any $P(A,B,C)$ adheres to this:



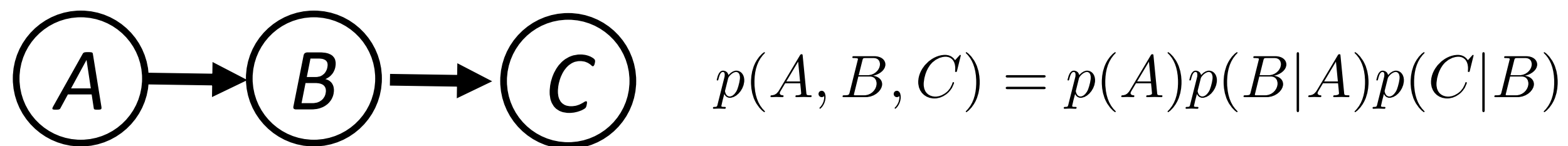
Not every $P(A,B,C)$ adheres to this:



Joint distribution in terms of Node Parents

You can express the joint distribution in terms of individual node distributions conditioned on the node's parents

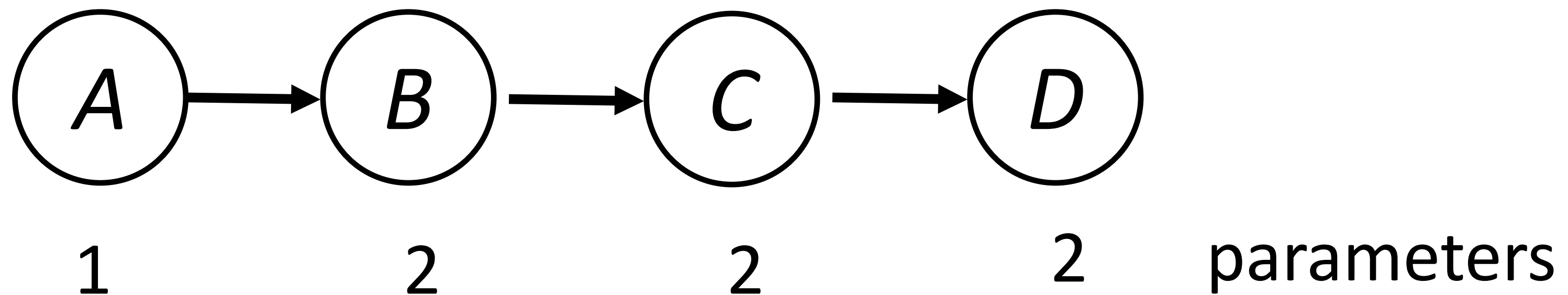
$$p(X_1, X_2, \dots, X_M) = \prod_{i=1}^M p(X_i | \text{parents}(X_i))$$



Example: Efficient Representation

M binary variables

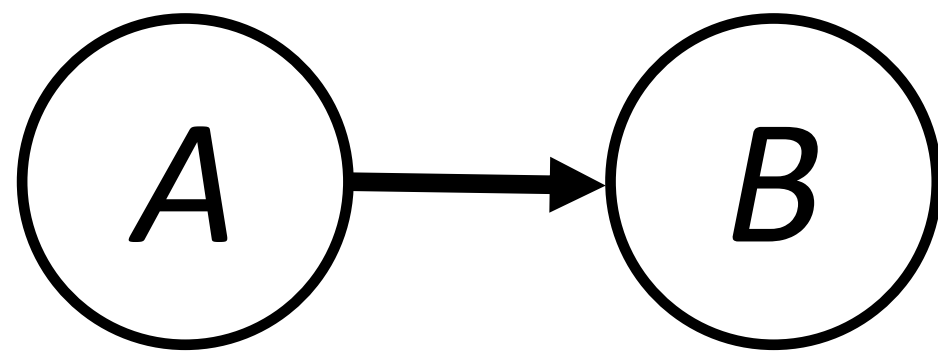
In general, we need $2^M - 1$ numbers to represent this distribution



When we can factorize as a chain, we only need $2(M-1)+1$ parameters

Check out Sesion 8.1.3 in Bishop's book

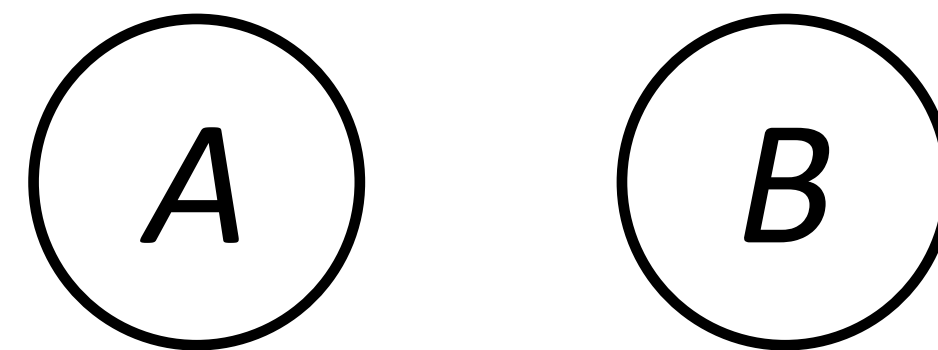
Independence: Two variables



$$p(B|A) \neq p(B)$$

(Or vice versa)

$$A \not\perp B$$



$$p(B) = p(B|A)$$

$$A \perp B$$

Only two variables, either independent or not

Conditional Independence

Independence

$$p(A|B) = p(A) \quad \text{or} \quad p(A,B) = p(A) p(B)$$

Observing B will not give me additional information about A

$$A \perp\!\!\!\perp B$$

Conditional Independence

$$p(A|B, C) = p(A|C) \quad \text{or} \quad p(A, B|C) = p(A|C) p(B|C)$$

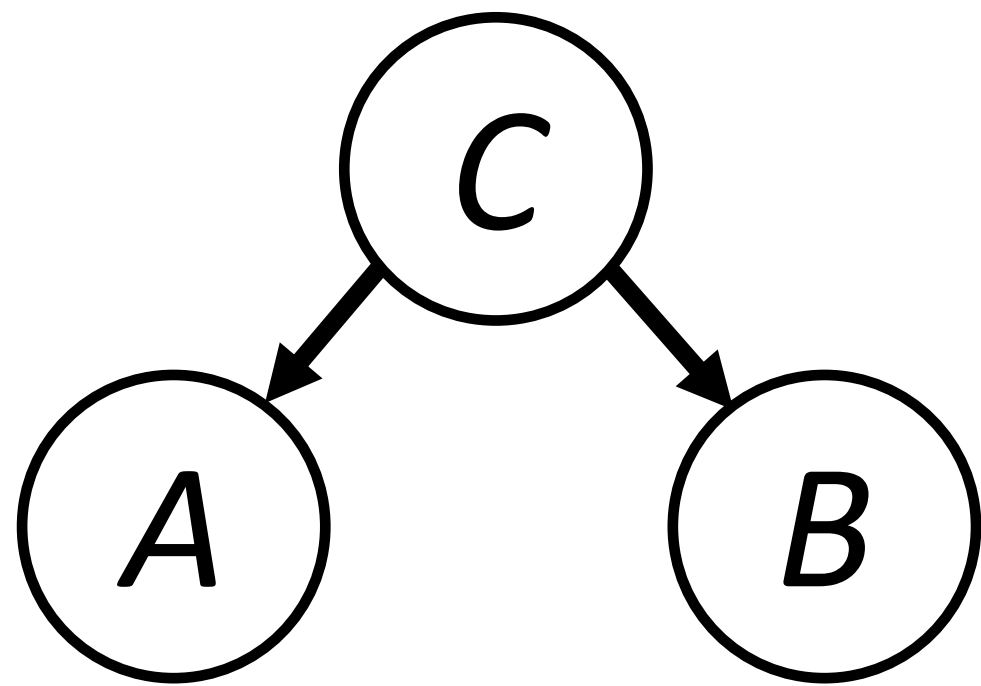
Given that I have observed C, observing B will not give me additional information about A

$$A \perp\!\!\!\perp B \mid C$$

Checking Independence

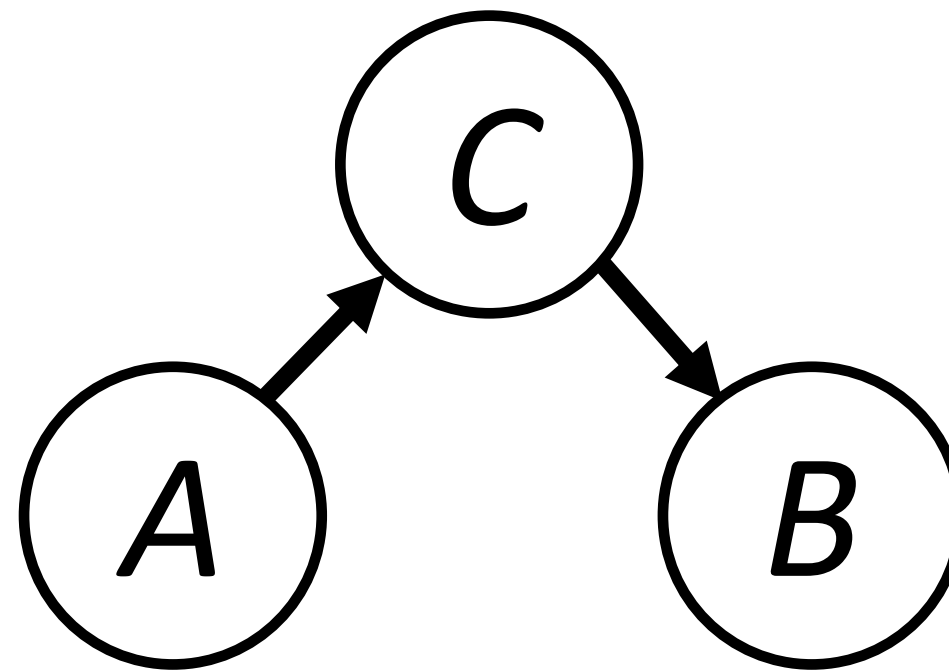
- We can check for independence by starting with $P(A,B,\dots,Z)$ and using algebraic manipulations to prove a particular independence holds.
- Cumbersome... can we find a way to check independences using the graph?

Conditional Independence: Three Variables



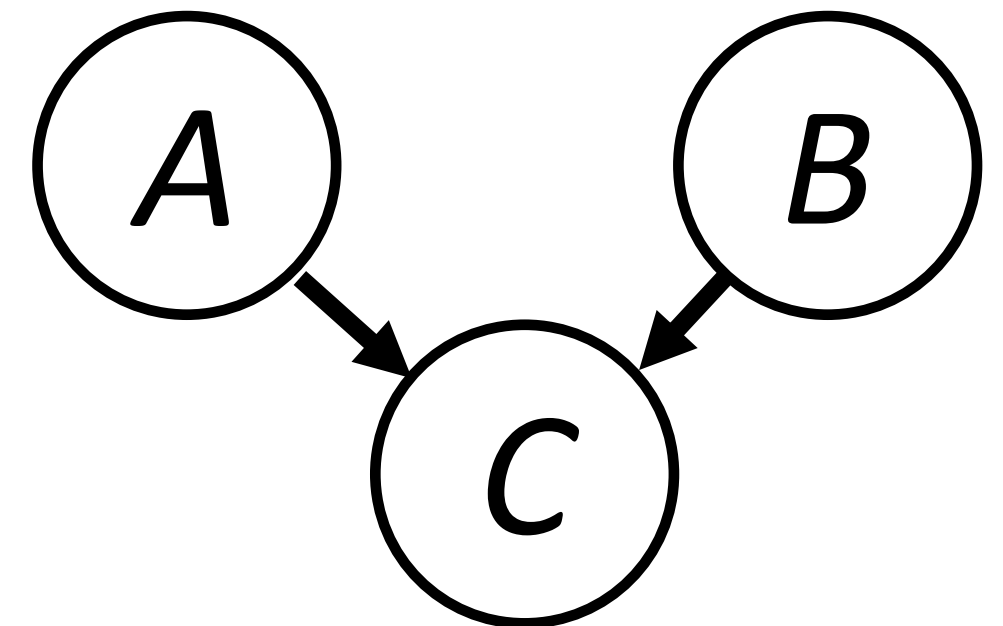
Fork

$$p(A|C) \ p(B|C) \ p(C)$$



Chain

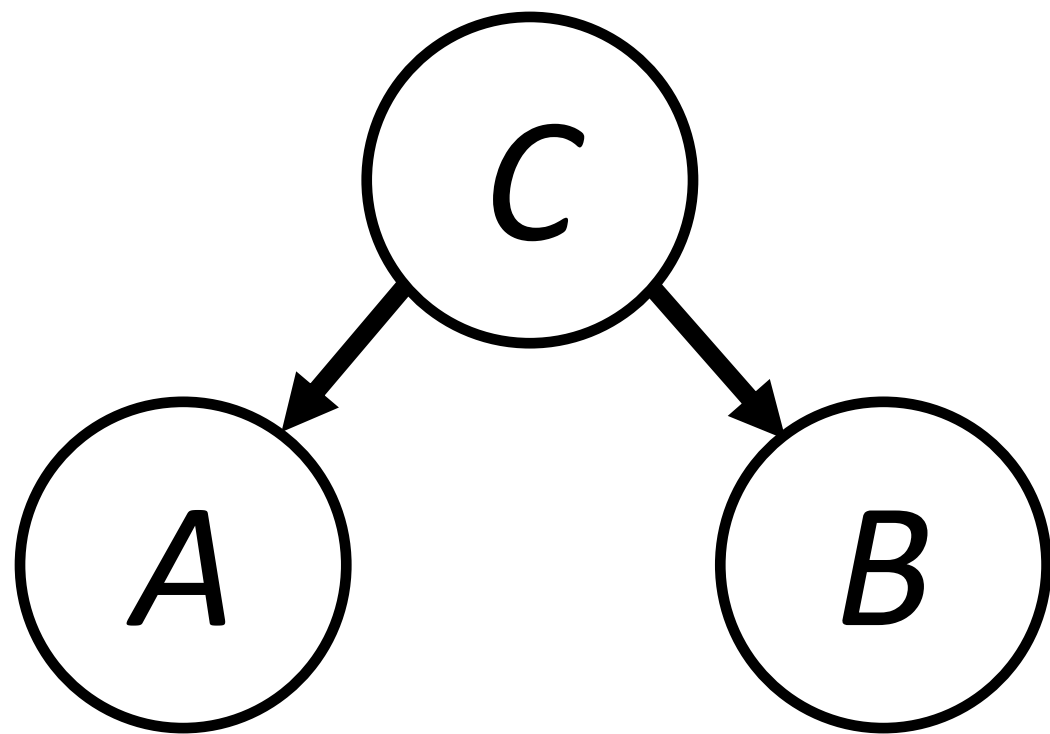
$$p(A) \ p(C|A) \ p(B|C)$$



Collider

$$p(A) \ p(B) \ p(C|A,B)$$

Fork/ Common Cause



Interpretation: if we don't observe C, observing A gives information on C, which gives information about B. If we do observe C, all the information about C is already present, and observing A adds nothing to our knowledge of B

$$p(A, B, C) = p(A|C) p(B|C) p(C)$$

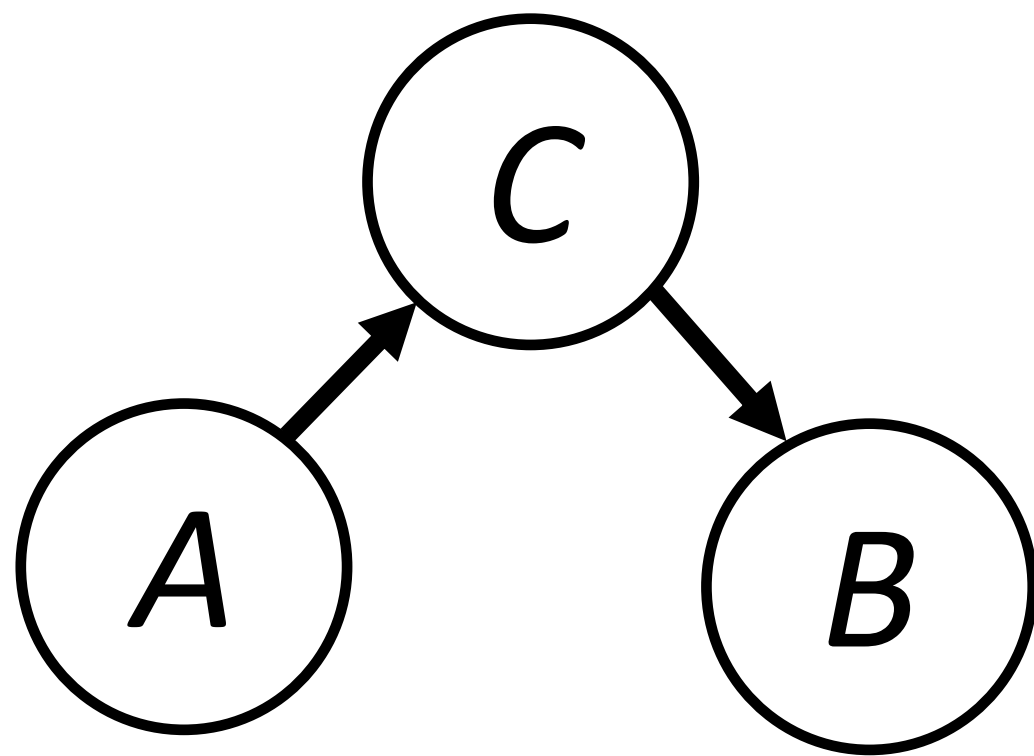
$$p(A, B) = \sum_C p(A|C) p(B|C) p(C) \neq p(A) p(B)$$

$$A \not\perp\!\!\!\perp B$$

$$p(A, B | C) = \frac{p(A, B, C)}{p(C)} = p(A | C) p(B | C)$$

$$A \perp\!\!\!\perp B | C$$

Chain



$$p(A) p(C|A) p(B|C) = p(C) p(A|C) p(B|C)$$

$$p(A, B) = p(A) \sum_C p(B|C) p(C|A) = p(A) p(B|A)$$

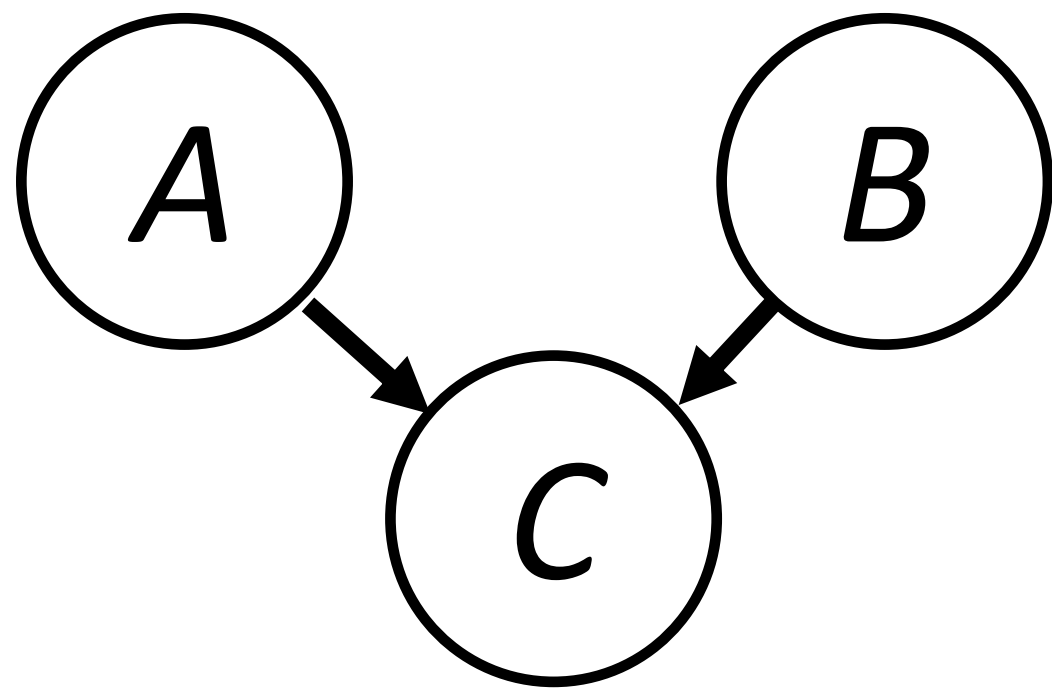
$$A \not\perp B$$

Interpretation: if we don't observe C, observing A gives information on C, which gives information about B. If we do observe C, all the information about B is already present.

$$p(A, B|C) = \frac{p(A, B, C)}{p(C)} = p(A|C) p(B|C)$$

$$A \perp\!\!\!\perp B \mid C$$

Collider



If we don't observe C, knowing A does not tell me what information B provided to C. If we do observe C, knowing A gives me information on what B must have been to explain the value of C

$$p(A, B, C) = p(A) p(B) p(C|A, B)$$

$$p(A, B) = p(A)p(B) \sum_C p(C|A, B) = p(A)p(B)$$

$$A \perp\!\!\!\perp B$$

$$p(A, B|C) = \frac{1}{p(C)} p(A)p(B)p(C|A, B) \neq p(A|C)p(B|C)$$

$$A \not\perp\!\!\!\perp B \mid C$$

Note: also true for descendants of C

Conditional Independence: N variables

Let's use the insights so far to construct a general criterion.

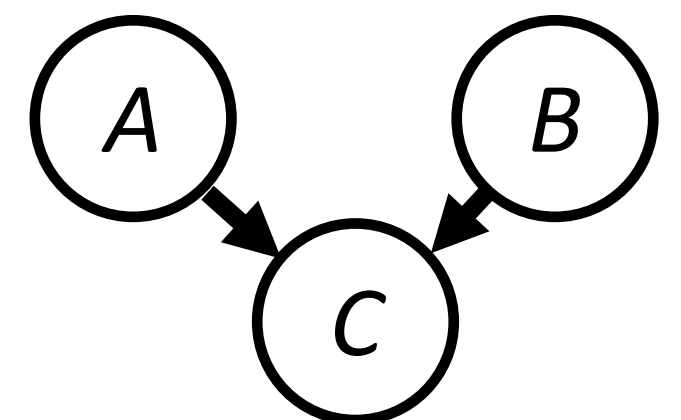
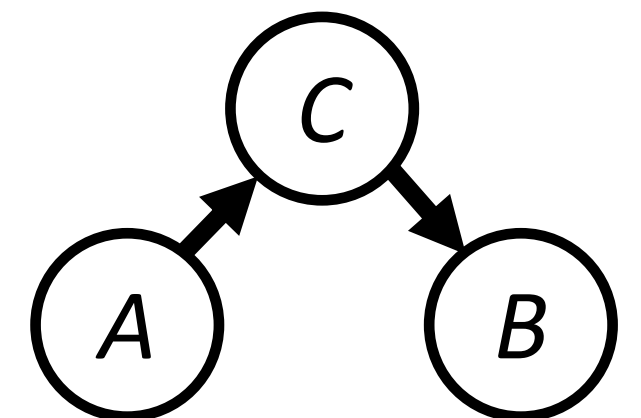
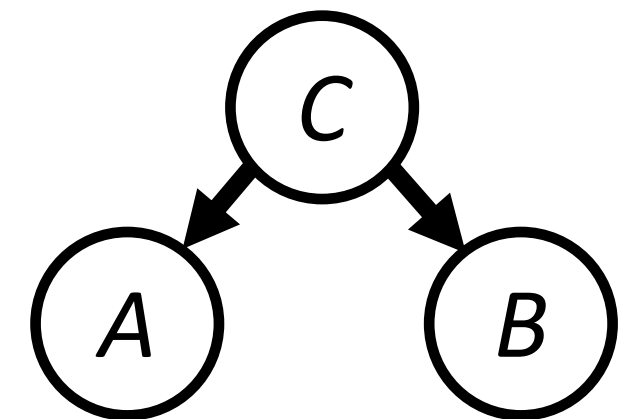
Here: A, B, C can be (sets of) variables in the graph

Blocking: We consider a path between A and B blocked given C if

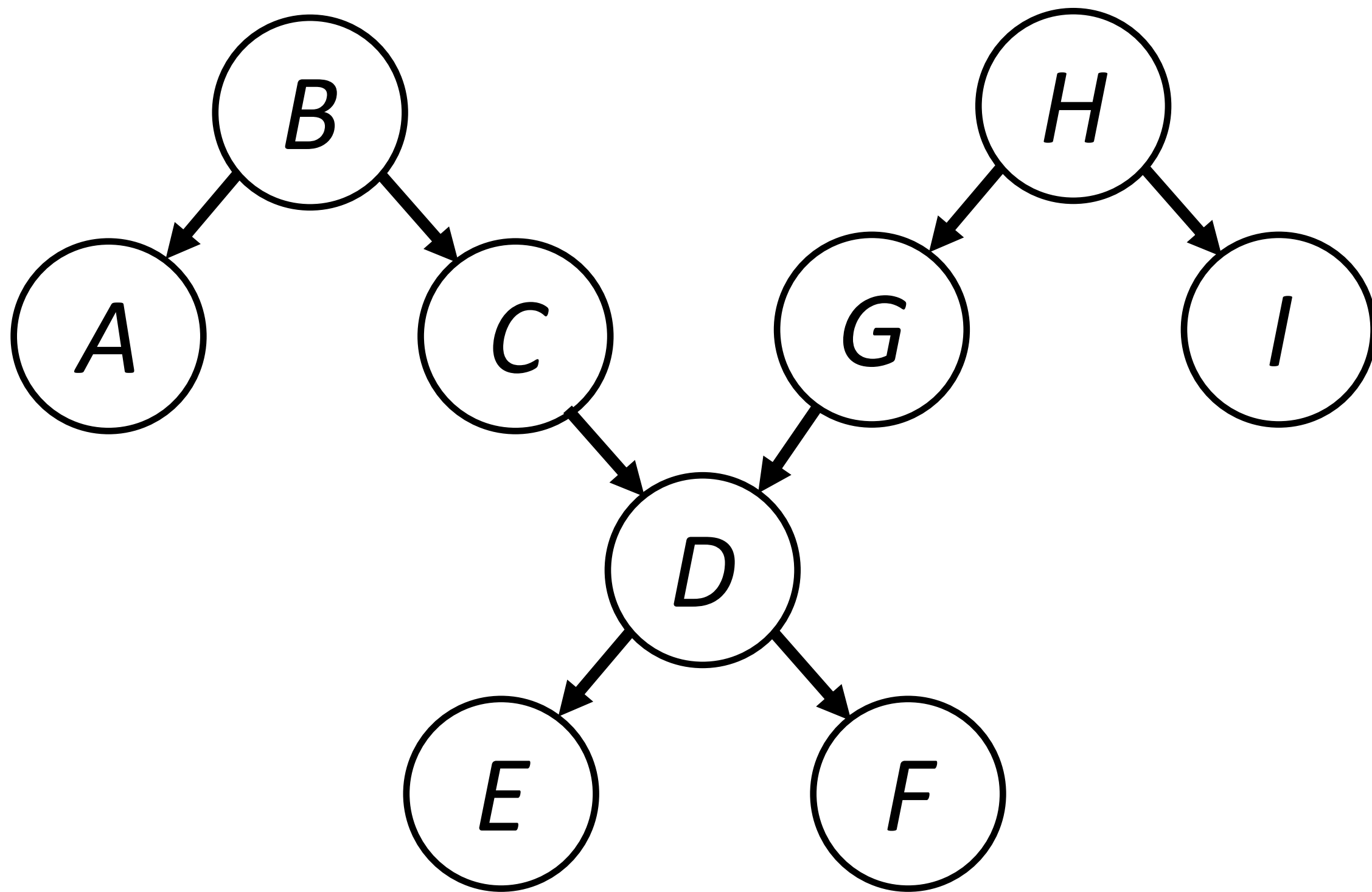
1. a non-collider node on the path is in C **OR**
2. there is a collider node on the path, and neither it or any of its descendants are in C

Directional-separation (d-separation)

If **every** path between A and B is blocked by C, then we say A is d-separated from B by C. That is $A \perp\!\!\!\perp B \mid C$



Practice Examples



Which of these are true:

$B \perp\!\!\!\perp E$ **FALSE**

$B \perp\!\!\!\perp E \mid C$ **TRUE**

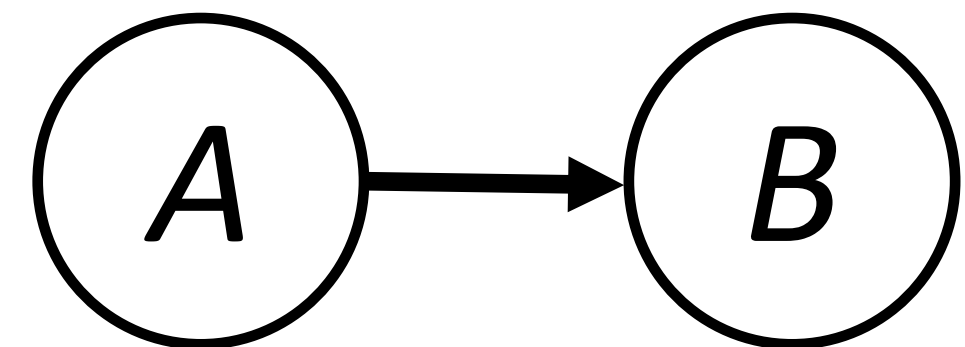
$B \perp\!\!\!\perp I$ **TRUE**

$B \perp\!\!\!\perp I \mid E$ **FALSE**

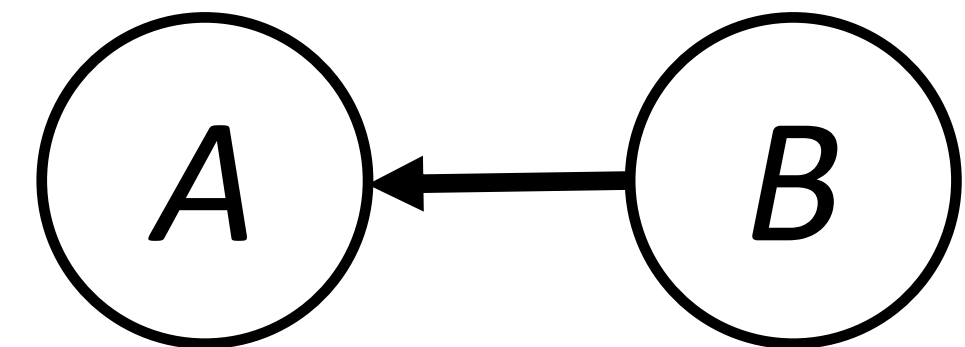
$B \perp\!\!\!\perp I \mid \{D, H\}$
TRUE

Note on Causality

- Often naturally follows from how we construct these graphs, but it is an additional assumption!
- Can have a correct probabilistic model, but not causal
- Important when we want to reason about the effect of actions

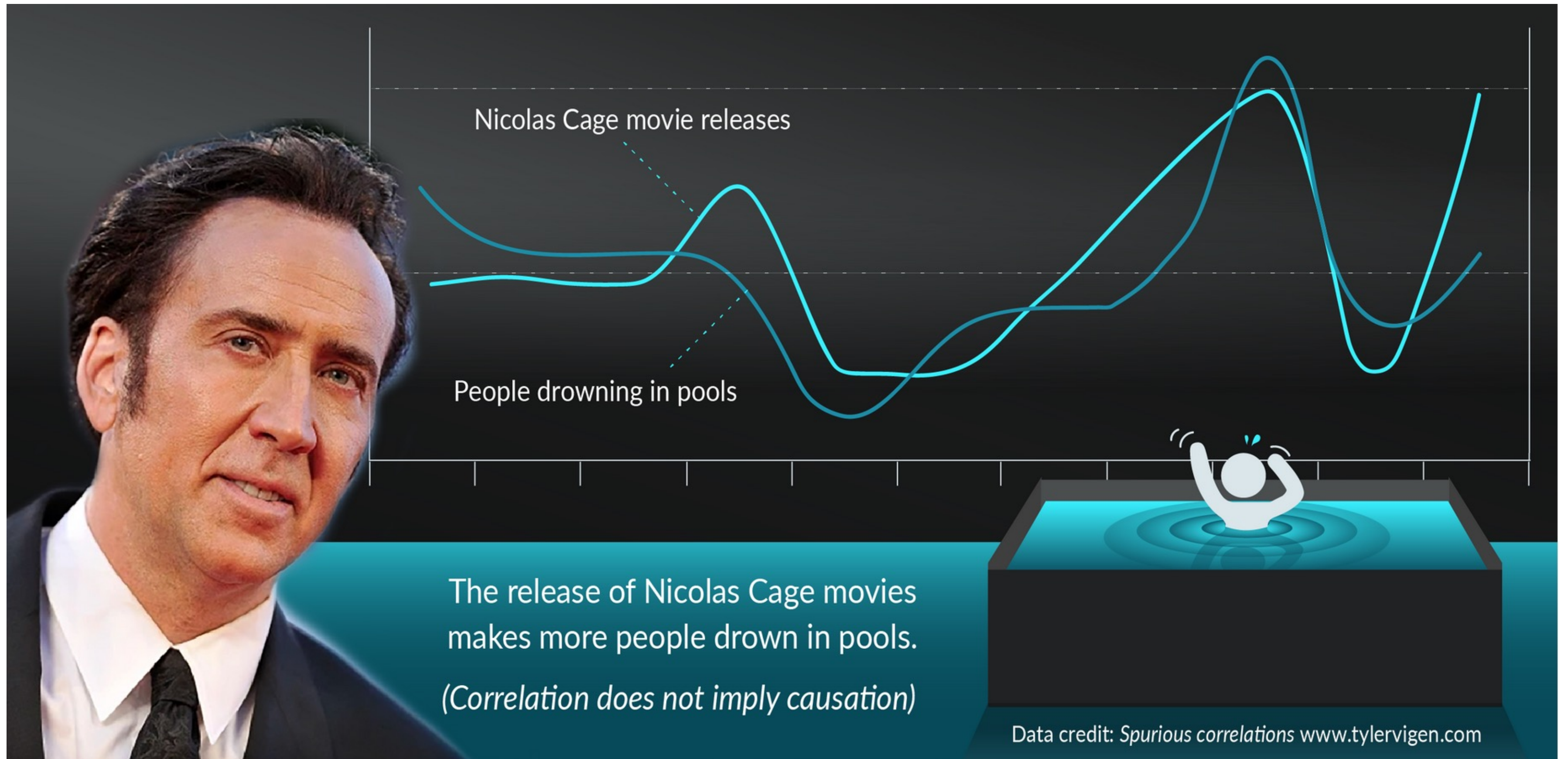


CAUSAL TRUTH



SAME DEPENDENCIES

Correlation vs. Causation

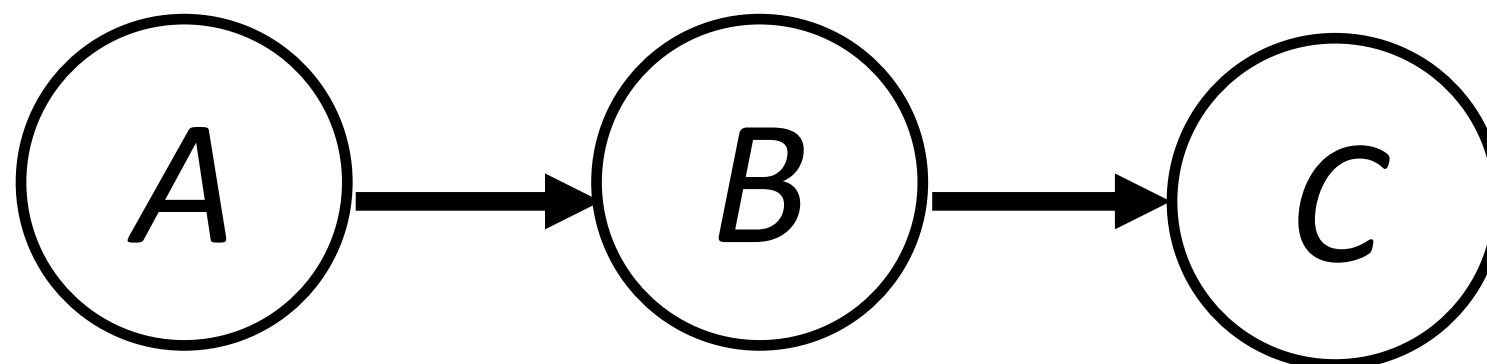


<https://modelthinkers.com/mental-model/correlation-vs-causation>

Correlation does not imply causation!

Inference in Bayesian Networks

- **Ancestral sampling:** Create many samples (then marginalize)
- **Message passing algorithms (belief propagation, a “sum-product” algorithm):** passing “messages” over the graph:
 - Suppose we want $p(A)$, $p(C)$
 - Naive: marginalizing the joint distribution
 - Use the structure to go from D^M to MD^2

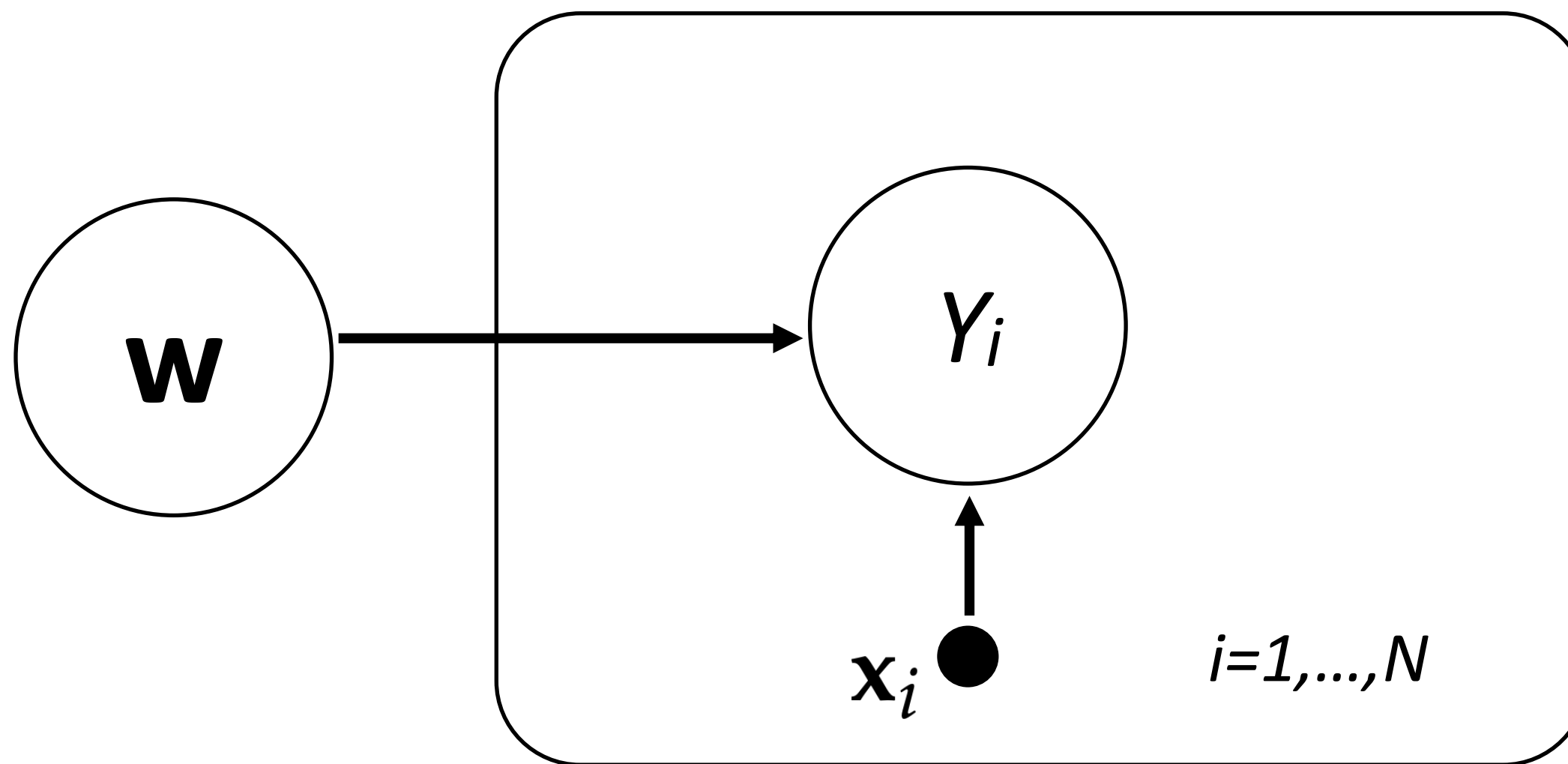


Learning the Network Structure

- We might not know the structure, but we can try to learn it
 - constraint-based methods
 - score and search methods
- Fewer links -> lower complexity

Machine Learning Models as PGMs

$$p(\mathbf{w}, \mathbf{Y}) = p(\mathbf{w}) \prod_{i=1}^N p(Y_i | \mathbf{w}, x_i, \dots)$$



See Section 8.1.1 in Bishop's book

Concluding Remarks

- Bayesian Inference treats (all) parameters in the model as random variables, and deals with consistently updating these random variables in the light of evidence.
- Bayesian networks are tools to represent, reason and do inference for joint probability distributions. d-separation is a graphical technique to reason about (in)dependencies among variables.

Next Class (video recording)

- Clustering
- Mixture Models

Reading

The reading material is from [Pattern Recognition and Machine Learning](#) by Bishop.

- Section 1.2.6
- Chapter 3 up to and including Section 3.3
- Chapter 8 up to and including Section 8.2
- Parts of similar material can be found in *The Elements of Statistical Learning* by Hastie *et al.* [also available on the web] and various other books.

Bishop, Christopher M., and Nasser M. Nasrabadi. “*Pattern recognition and machine learning*”. Vol. 4. No. 4. New York: springer, 2006.