

Inference and Representation

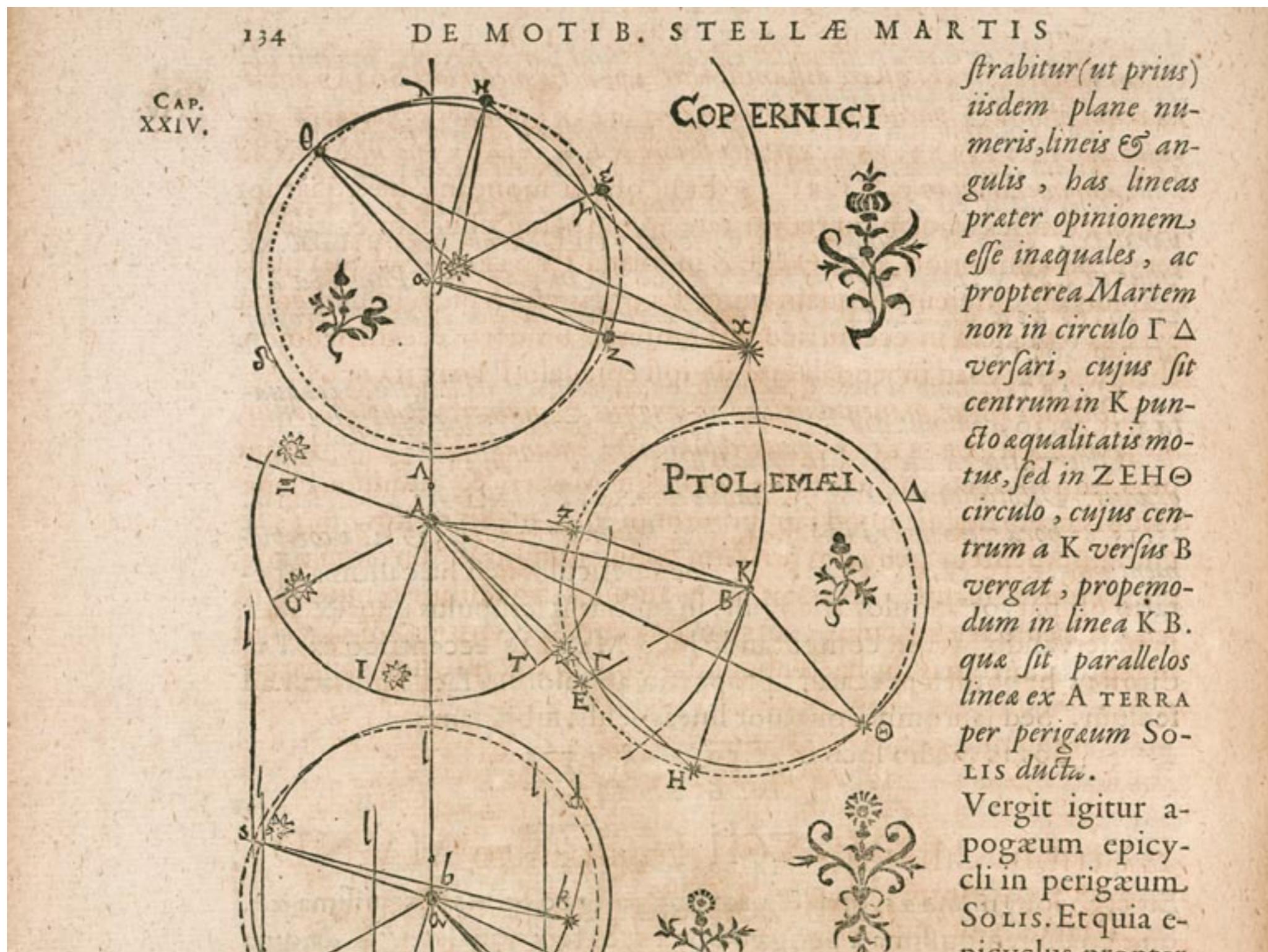
DS-GA-1005, CSCI-GA.2569

Joan Bruna

Courant Institute of Mathematical Sciences
Center for Data Science
NYU

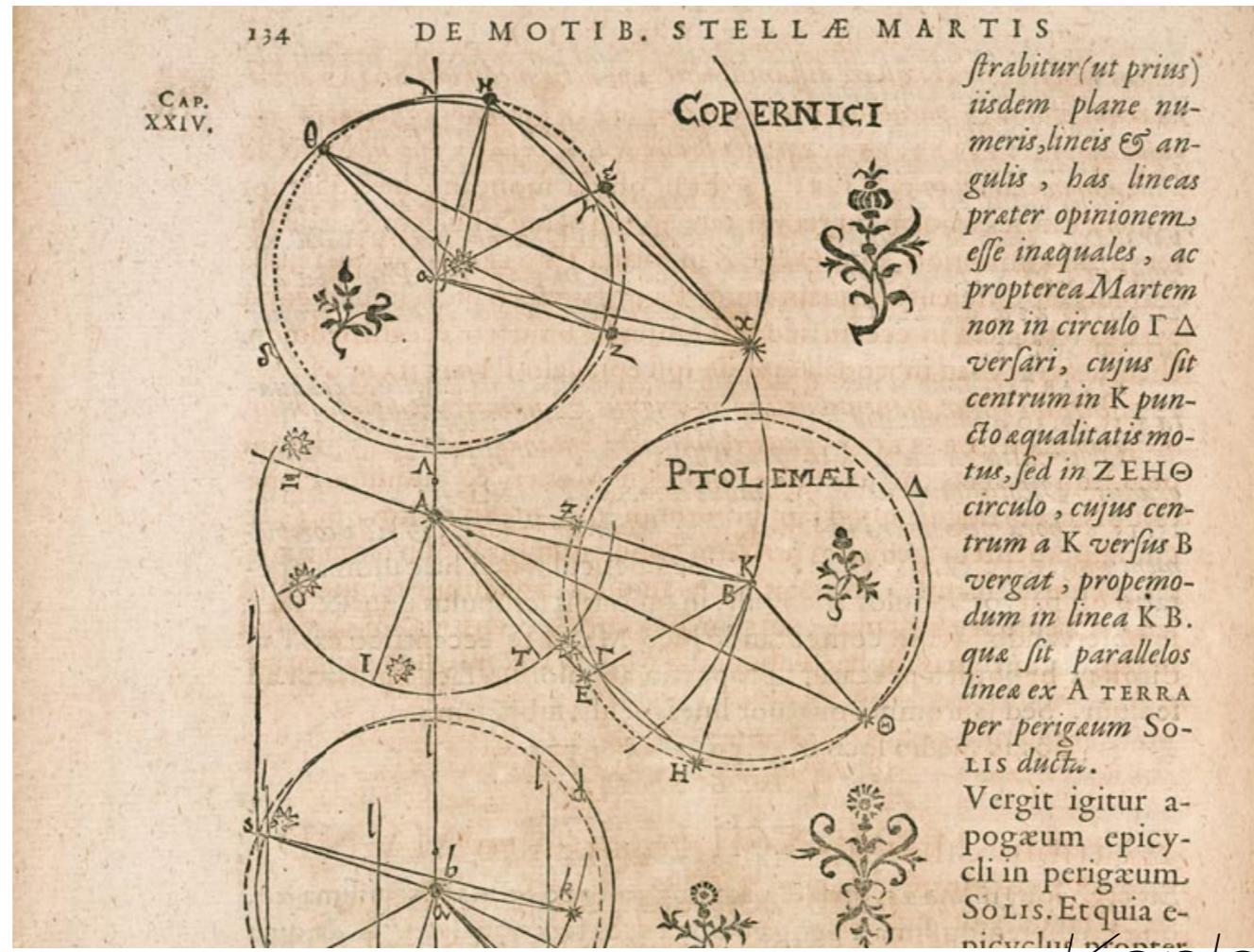


Celestial Mechanics



Kepler, 17th Century

Celestial Mechanics



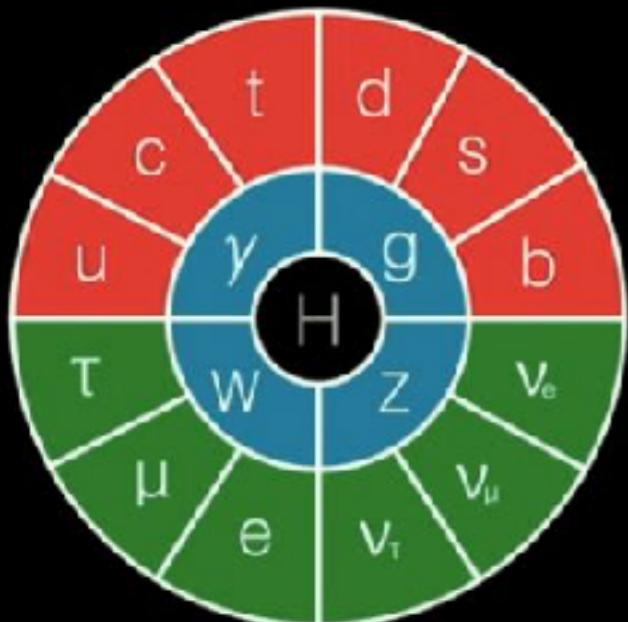
Kepler, 17th Century

- Determine the dynamics of planets, moons and stars given partial, noisy observations, and given partial, noisy mathematical model.
- Scientific Method: Build a model that can assign probabilities to observable events, and reject it whenever likelihood is too small.

Science = Model + Data

- Observations x_1, \dots, x_n .
 - E.g. x_1 : position of Jupiter at noon.
 x_2 : position of Jupiter at midnight.
 x_3 : size of the Moon at noon.
...
- Density models:
 - Model A: Joint probability distribution
 $p_A(x_1, \dots, x_n)$. (Solar system gravitates around Earth)
 - Model B: Joint probability distribution
 $p_B(x_1, \dots, x_n)$. (Solar system gravitates around Sun)
- Theories/Models are refuted from data by comparing likelihoods
 - Hypothesis testing, Likelihood ratios.

Higgs Boson



ONE OF THE THINGS PEOPLE PREDICT WILL COME OUT IS

THE
HIGGS
BOSON



THE HIGGS IS THE PARTICLE RESPONSIBLE FOR GIVING MASS TO OTHER PARTICLES.



WHAT IS MASS?

WHEN YOU THINK OF THINGS HAVING MASS, IT MEANS IT HAS "STUFF" TO IT, RIGHT?



IT'S NOT ACTUALLY "STUFF"

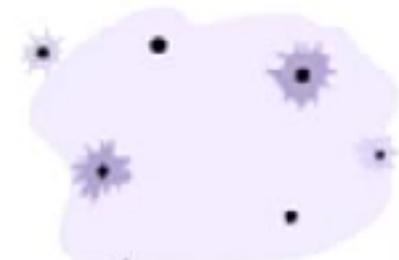
PARTICLES HAVE MASS BUT NO VOLUME.



(THEY'RE POINT PARTICLES)

MASS IS A CHARACTERISTIC OF A PARTICLE, LIKE CHARGE.

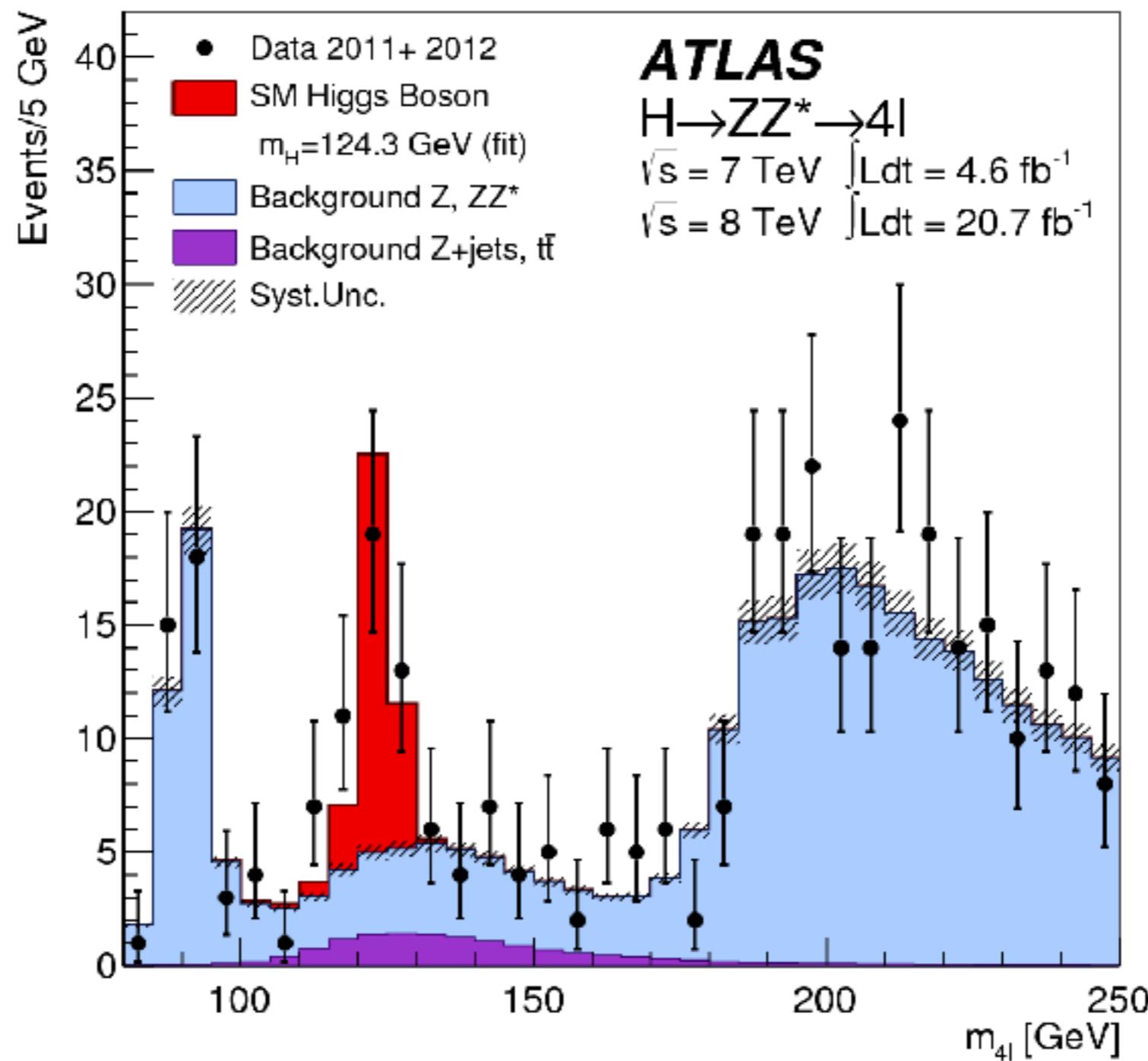
SOME HAVE IT, SOME DON'T



IT'S JUST A DIFFERENT KIND OF CHARGE...

Higgs Boson Discovery

- What does "discovery" mean in this context?



Fundamental Scientific Pipeline

- How to construct such models to explain large systems?
 - Representation Challenge: How to parametrize
 $p(x_1, \dots, x_n) = f_\theta(x_1, \dots, x_n)$?
- How to adjust the parameters of such models to best explain the data?
 - Learning Challenge: How to fit θ to the data?
- How to evaluate likelihoods under the model?
 - Inference Challenge
 - In most interesting cases, exact inference will be computationally intractable.
 - Major Topic of the course: develop computationally efficient approximate inference.

Applications: Modeling Text Data

- Topic Models allow us to infer, interpret large texts

Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Novie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (*Zeitschrift für physiologische chemie*, x,

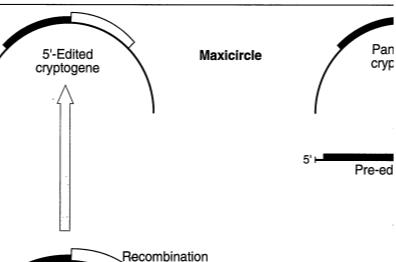
RNA Editing and the Evolution of Parasites

Larry Simpson and Dmitri A. Maslov

The kinetoplastid flagellates, together with their sister group of euglenoids, represent the earliest extant lineage of eukaryotic organisms containing mitochondria (1). Within the kinetoplastids, there are two major groups, the poorly studied bodonids-cryptobiids, which consist of both free-living and parasitic cells, and the better known trypanosomatids, which are obligate parasites (2).

Perhaps because of the antiquity of the trypanosomatid lineage, these cells possess several unique genetic features (see accompanying Perspective by Nilsen)—one of which is RNA editing of mitochondrial transcripts. This RNA editing function (3–7) creates open reading frames in “cryptogenes” by insertion (or occasional deletion) of uridine (U) residues at a few specific sites within the coding region of an mRNA (5'-editing) or at multiple specific sites throughout the mRNA (pan-editing). The

but there is disagreement on the nature of the primary parasitic host. The “invertebrate first” model (10, 11) states that the initial parasitism was in the gut of pre-Cambrian invertebrates. Coevolution of parasite and host would have led to a wide distribution of trypanosomatids in insects and leeches. In this theory, digenetic life cycles (alternating invertebrate and vertebrate hosts) evolved later as a result of the acquisition by some hemipterans and dipterans of the ability to feed on the blood



tion arth
netic woul
the al
In pothe
mitoc
quenc
Critic
cent nucle
as an
tical Tryp
the t by th
fish I tutes
trypa bran
separ

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (1) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (2) provide the most

convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

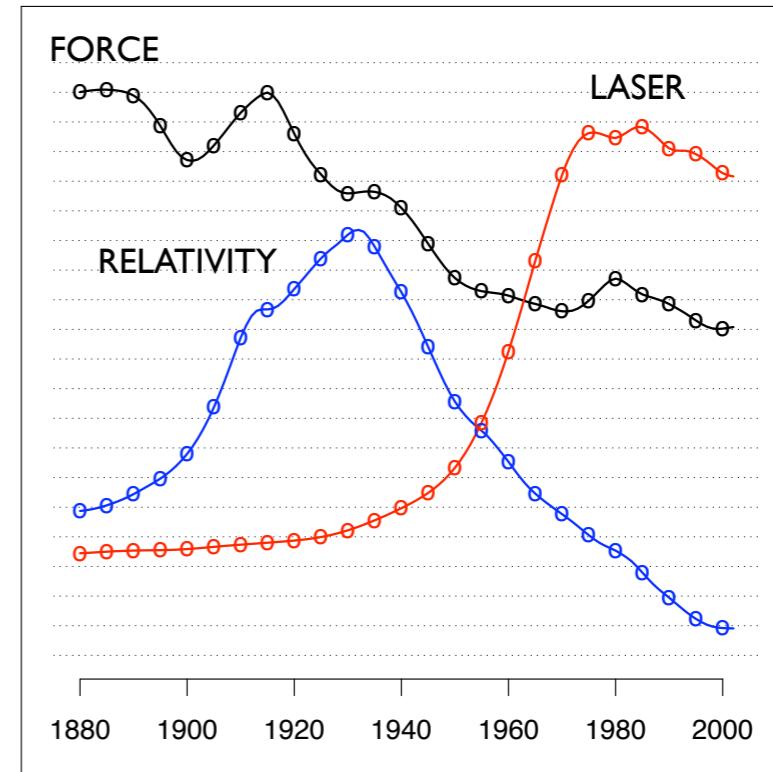
It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffered by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the tell-tale signatures of chaos. In phase space, chaotic trajectories come to lie on “strange attractors,” curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and, most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (3).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-



Cannibalism and chaos.
The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

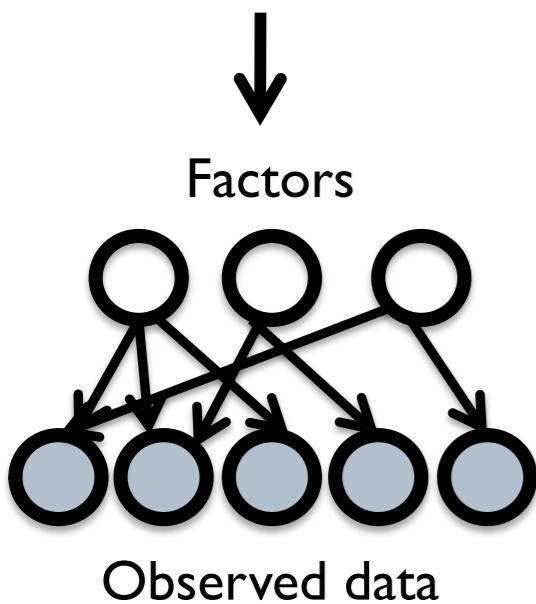
"Theoretical Physics"



human genome	evolution	disease	computer models
dna	evolutionary species	host bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Applications: Survey Data

```
1 0 0 1 1 0 0 0 1  
1 1 0 1 1 0 0 1 0  
0 1 0 1 0 0 1 0 0  
1 0 0 0 1 0 0 1 1  
0 1 0 1 0 0 1 1 0
```



- Social surveys with questions such as
 - Should tax rate be progressive?
 - Do you support the BLM movement?
 - Should affirmative action be used in college admissions?
- Goal is to automatically discover relevant factors, e.g.
 - Socioeconomic status
 - Health
 - Political values
- Factor Analysis models infer underlying beliefs/traits of survey data

Applications: Modeling Sequential Data

- Automatic Machine Translation

The screenshot shows the Google Translate interface. At the top, there's a navigation bar with the Google logo, followed by "Translate" in red, and dropdown menus for "From: English" and "To: Spanish". A blue "Translate" button is also present. Below this, there are three language tabs: "Spanish", "Chinese", and "English", with "English" being the active tab. On the left, the English input text is:

The top U.S. general, visiting Israel at a delicate and dangerous moment in the global standoff with Tehran, is expected to press for restraint amid fears that the Jewish state is nearing a decision to attack Iran's nuclear program.

A small "X" icon is located next to the input text. On the right, the Spanish output text is:

El máximo general de EE.UU., de visita en Israel en un momento delicado y peligroso en el enfrentamiento global con Teherán, se espera que presione a la moderación en medio de temores de que el estado judío se acerca a una decisión de atacar el programa nuclear de Irán.

A checkmark icon is located next to the output text. At the bottom, a note says: "New! Hold down the shift key, click, and drag the words above to reorder. Dismiss".

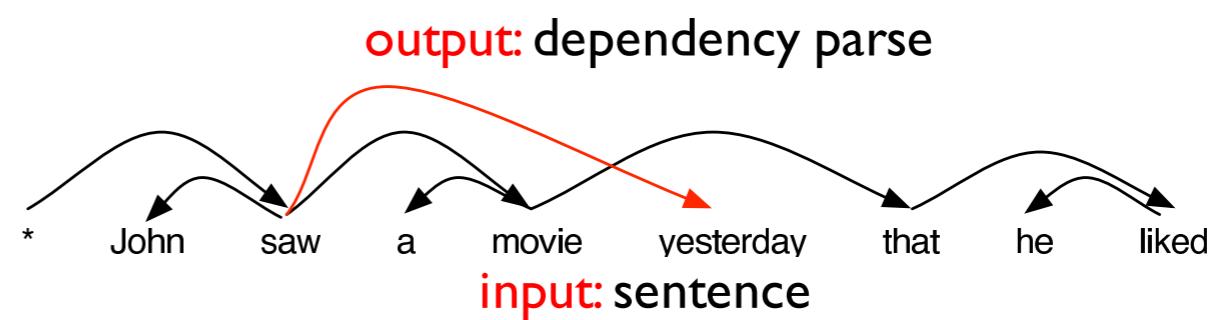
- nowadays mostly based on recurrent neural network models.

Applications: Structured Output Prediction

- Image Segmentation

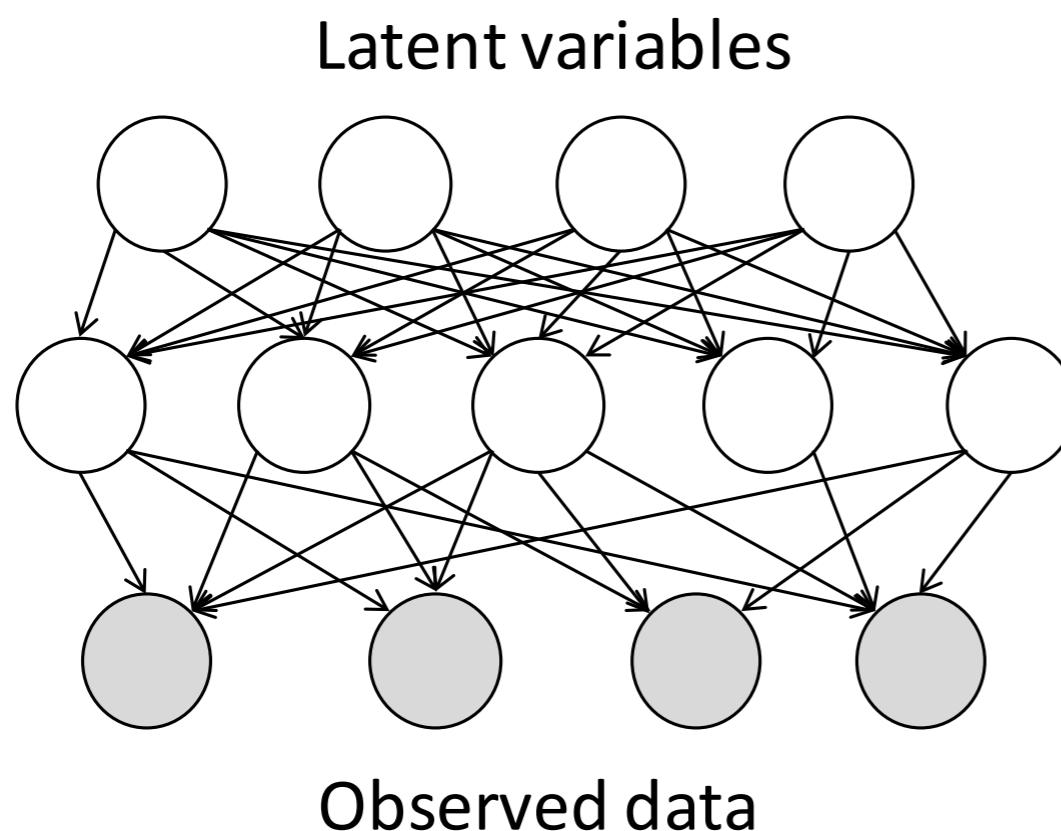


- Natural Language Processing



Applications: Modeling Images

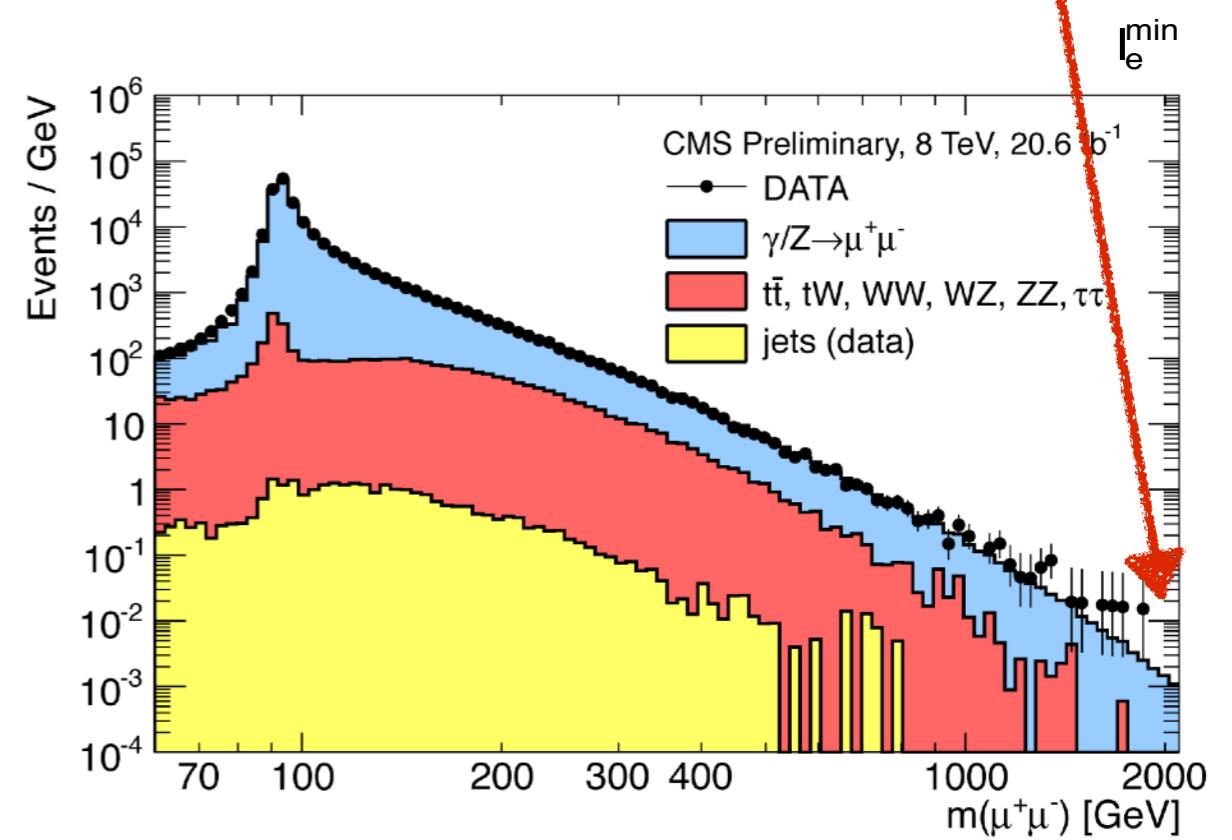
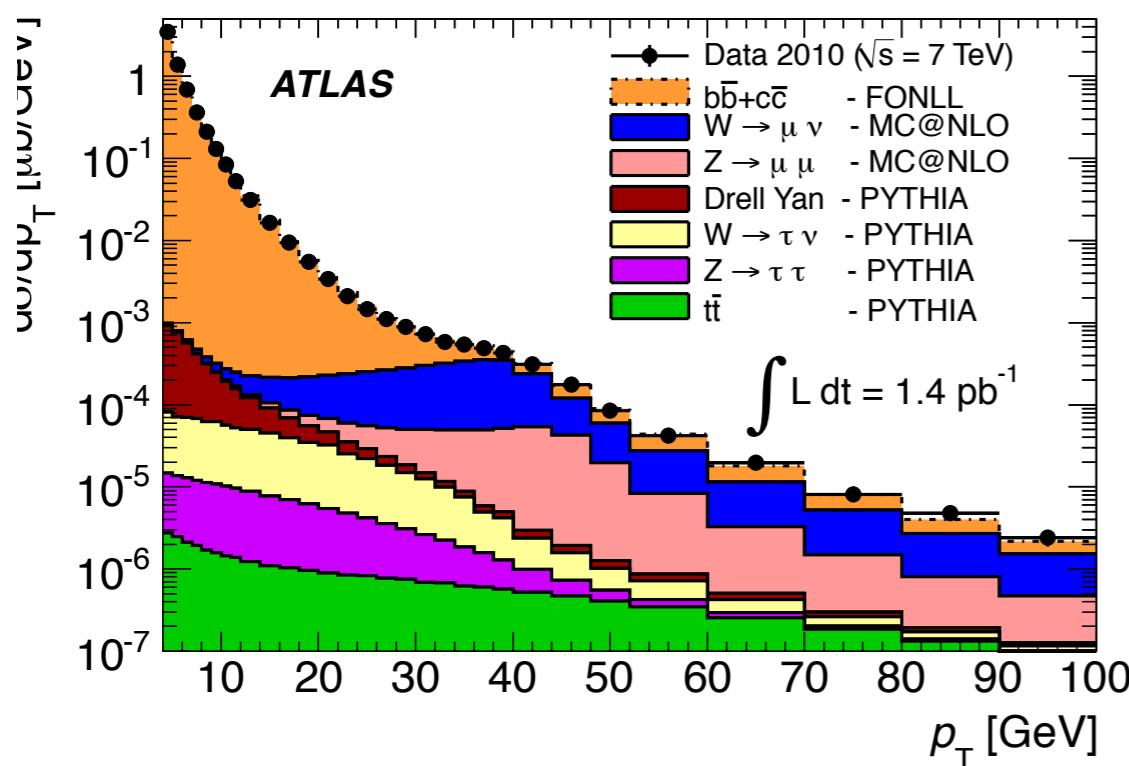
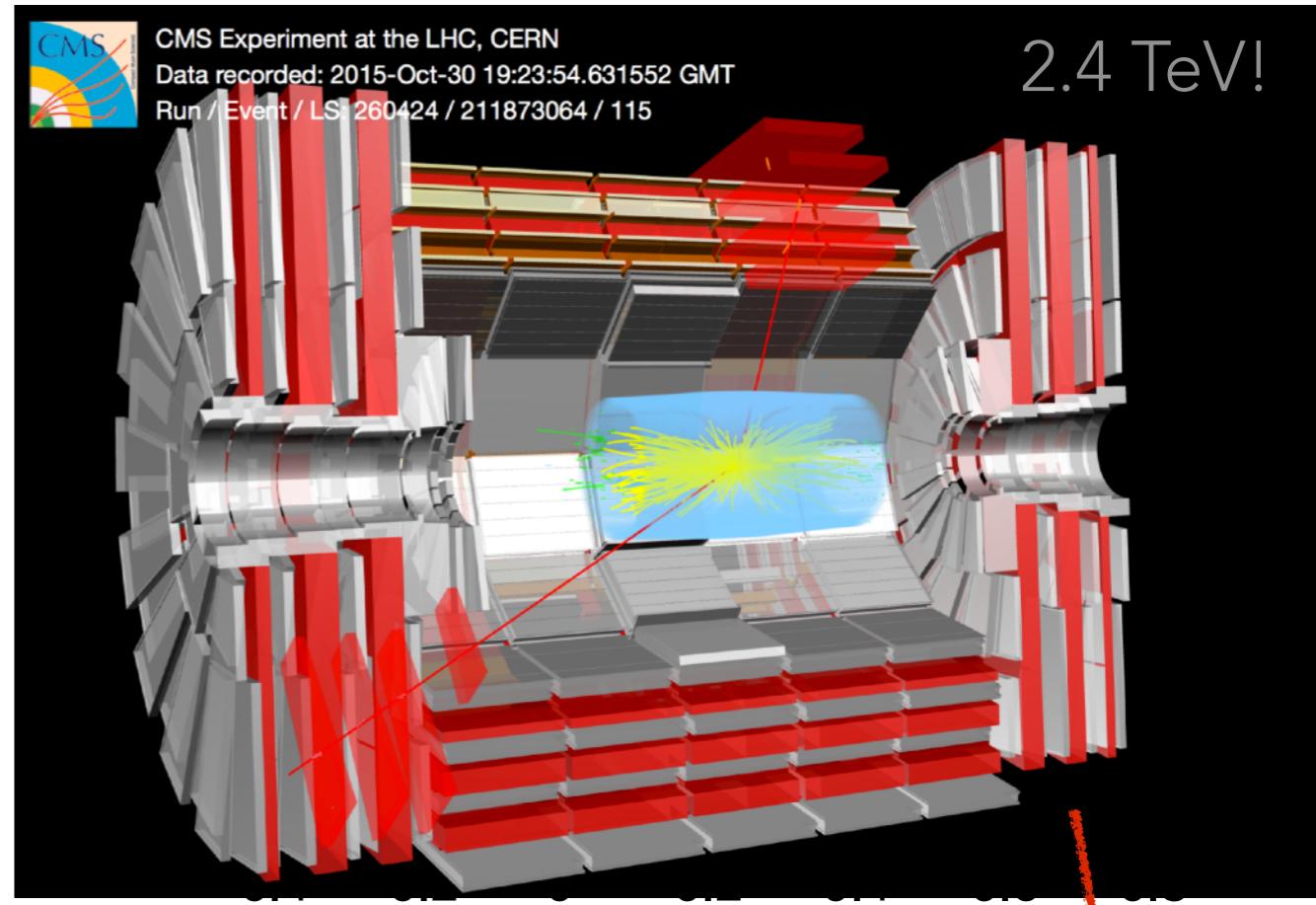
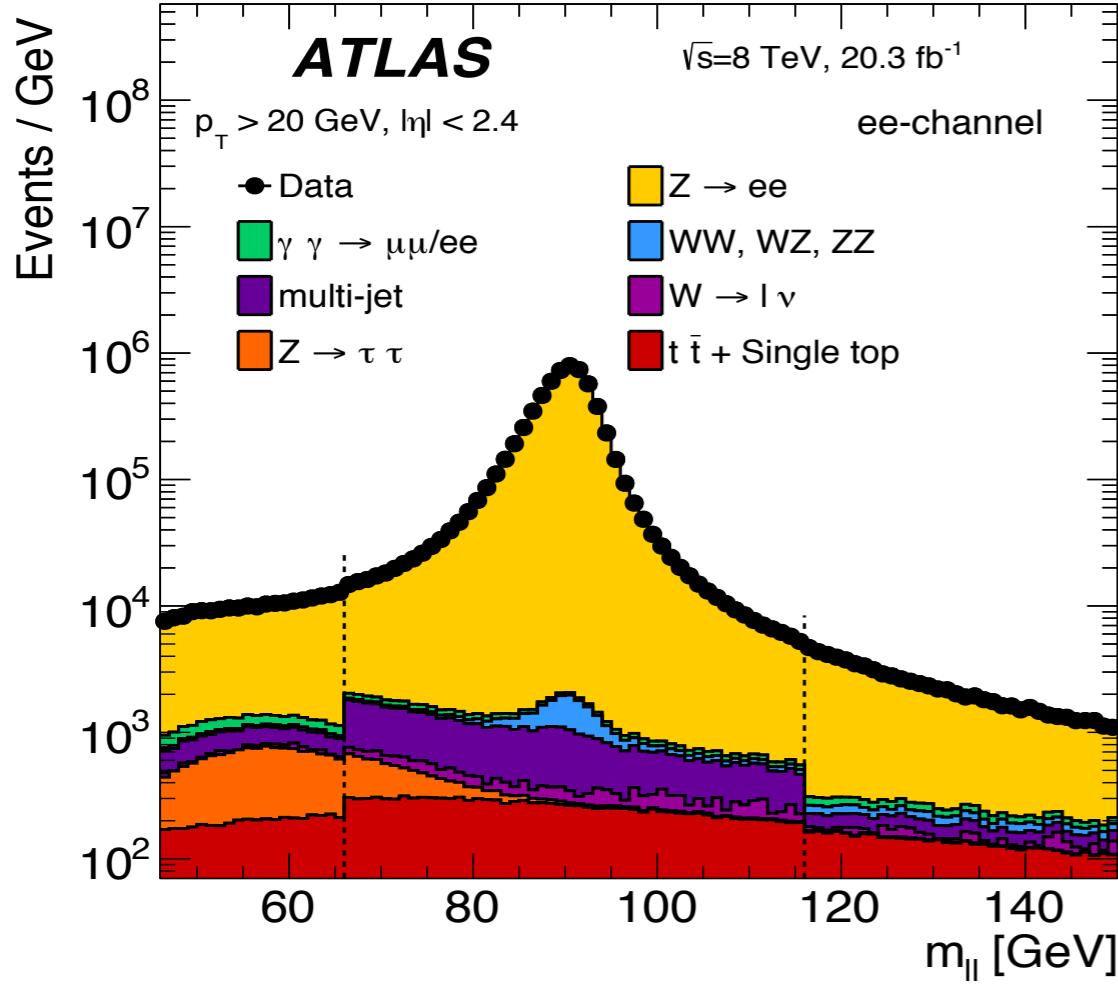
- Main challenge: high-dimensionality, complex, long-range dependencies.
 - Deep Generative Models, Autoencoders, Generative Adversarial Networks.



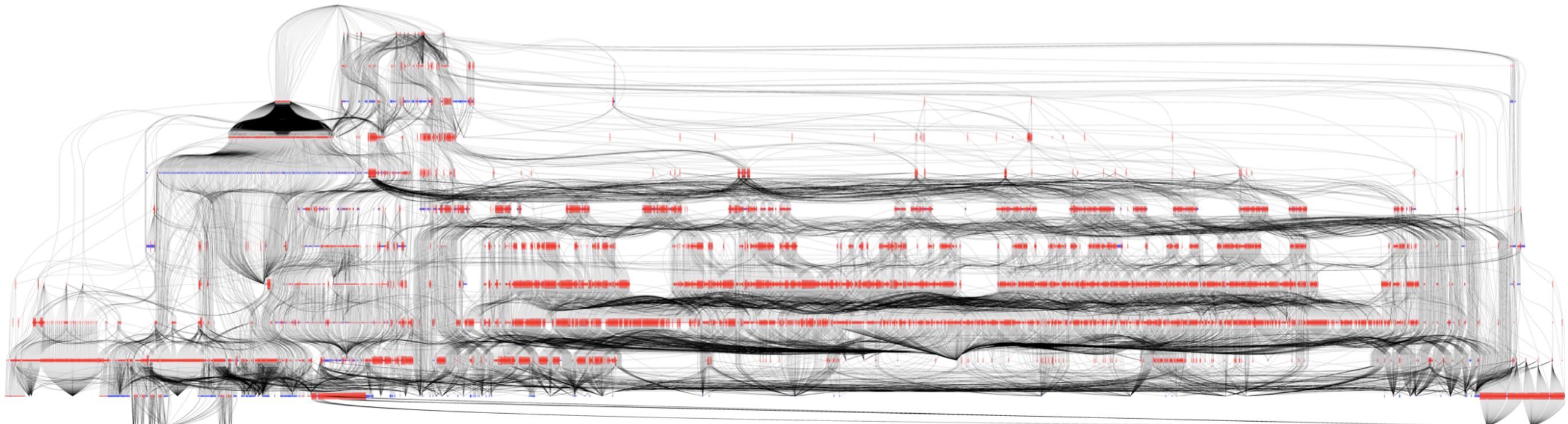
Samples from the model:

6	6	9	0	5	7	8	9	0	3
3	5	7	6	6	9	5	1	3	0
4	7	4	6	5	4	0	6	4	9
2	9	7	7	3	9	4	1	4	8
6	5	4	0	0	9	9	3	2	2
3	9	5	6	1	5	0	7	7	6
5	6	2	9	7	6	9	4	0	9
2	3	1	3	4	1	5	4	4	0
1	2	5	7	6	9	3	5	3	7
6	4	3	3	7	9	0	9	9	3

Applications: Particle Physics



Applications: Particle Physics



$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p | \alpha_p)$$

Syllabus Overview

- Definitions, Directed Graphical Models and Bayesian Networks (lect1)
- Undirected Graphical Models. Markov Random Fields (lect2)
- Topic Models and Bayesian Non-parametrics (lect 3)
- Principal Component Analysis with Applications (lect 4)
- Expectation-Maximization. MCMC (lect 5)
- Variational Inference (lect 6)
- Structured Output Prediction (lect 7)
- Sequential Models (lect 8)
- Boltzmann Machines, Variational Autoencoders (lect 9)
- Modeling images. GANs, Flows and Autoregressive (lect 10)
- Modern unsupervised learning with GANs and VAEs (lect 11)

Guest Lectures

- Kyunghyun Cho

- Kyle Cranmer

Logistics

- Website
- Piazza (you need to sign up)
- Required Readings for each lecture posted to course website.
- Books
 - Machine Learning: a probabilistic perspective, by Kevin Murphy (available online)
 - Mike Jordan's notes on Probabilistic Graphical Models.
- Office Hours
 - Mondays 10am.
- Lab/Recitation
- Graders
 - Sahana Athreya
 - Sreyas Mohan

Prerequisites and Grading

- Pre-requisite: DS-GA-1003/CSCI-GA 2567 Machine Learning and Computational Statistics)
- Grading
 - Problem Sets: 45%
 - Midterm Exam: 25%
 - Final Project: 25%
 - Class/Lab participation: 5%
- Final Project
 - Research project applying tools developed in class.
 - Proposal due by end of October
 - Feedback on proposal early November
 - Project writeup and Poster presentation: end of semester

Basic Definitions and Notation

- A central object of this class is a joint probability distribution over a high-dimensional space.
- To simplify, we start with a discrete space. Assume random variables X_1, \dots, X_n , each defined over a discrete space of q symbols: $\mathcal{X} = \{a_1, \dots, a_q\}$
- Joint probability distribution of random vector $\mathbf{X} = (X_1, \dots, X_n)$ is a mapping $p : \mathcal{X} \otimes \dots \otimes \mathcal{X} \rightarrow [0, 1]$ that computes
$$p(X_1 = a_{i_1}, \dots, X_n = a_{i_n}) \text{ for all } i_1, \dots, i_n \in \{1, q\}.$$
 - Normalized measure $\sum_{x \in \mathcal{X}^n} p(x) = 1$
- A *Probabilistic Model* is one such distribution, typically indexed by parameters $\theta \in \mathbb{R}^p$ that are fit to the data: p_θ
 - Two statistical paradigms
 - ❖ Parametric/ low-dimensional statistics: **p fixed as n increases.**
 - ❖ Non-parametric / high-dimensional statistics: **p increases with n .**

Curse of Dimensionality

- How large is the space of multivariate distributions as a function of n ?
 - Discrete case: the number of possible outcomes is q^n : exponential dependency.
 - Continuous case: under regularity assumptions, same phenomena.
- Estimation of a generic multivariate distribution is therefore intractable [PS1]: **statistical intractability**.
- Inference of conditional probabilities is **computationally intractable**

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- Therefore, we need to make assumptions about reality, and consider probabilistic models with smaller number of parameters.
- Q: How to introduce structure?

Bayes Rule

- Theorem-by-definition of conditional probability:

$$p(x_1, x_2) = p(x_1) p(x_2 \mid x_1)$$

- Iterating this rule, we obtain

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_1)p(x_2, \dots, x_n \mid x_1) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3, \dots, x_n \mid x_1, x_2) \\ &= \prod_{i=1}^n p(x_i \mid x_1 \dots x_{i-1}) . \end{aligned}$$

- the bulk of the terms on the rhs involve $\mathcal{O}(n)$ terms.
- what if we drop some conditioning variables?

$$p_G(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i \mid x_{i_1}, \dots, x_{i_{l_i}}) . \quad (i_j < i)$$

Example: Naive Bayes

- Consider n discrete features $X_1 \dots X_n$ and a categorical variable Y
- Suppose the joint distribution

$$p(X_1, \dots, X_n, Y) = P(Y) \prod_{i=1}^n P(X_i \mid Y, X_1, \dots, X_{i-1})$$

can be simplified to

$$p(X_1, \dots, X_n, Y) = P(Y) \prod_{i=1}^n P(X_i \mid Y)$$

- This means that X_i, X_j are independent given Y .
 - We will write $X_i \perp X_j \mid Y$, and say that they are *conditionally independent*.
- Q: Does $X_i \perp X_j \Rightarrow X_i \perp X_j \mid Y$?
- Q: Does $X_i \perp X_j \mid Y \Rightarrow X_i \perp X_j$?

Example: Classification with Naive Bayes

- Naive Bayes may be used to classify emails as spam ($Y = 1$) or not spam ($Y = 0$).
- By indexing the words on a given vocabulary $1 : n$, we construct

$X_i = 1$ if word i appears in email, 0 otherwise.

- Naive Bayes model

$$p(X_1, \dots, X_n, Y) = p(Y) \prod_{i=1}^n p(X_i \mid Y)$$

- The parameters of the model are estimated with maximum likelihood.
- Spam Prediction:

$$p(Y = 1 \mid X_1, \dots, X_n) = \frac{P(Y = 1, X_1 = x_1, \dots, X_n = x_n)}{P(X_1 = x_1, \dots, X_n = x_n)} .$$

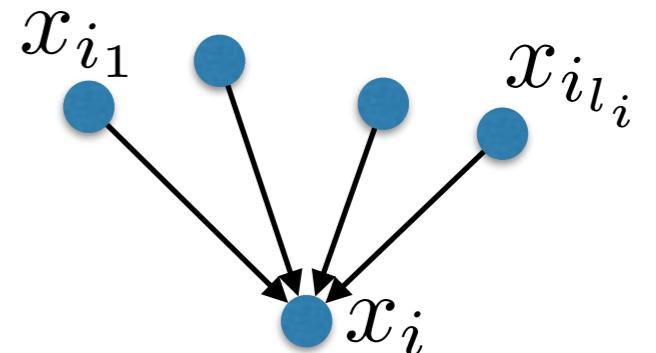
- Conditional Independence assumptions are strong, but resulting model is simple and useful.

Bayesian Networks

- We've seen that reducing the scope of conditional probability is critical in order to beat the curse of dimensionality.

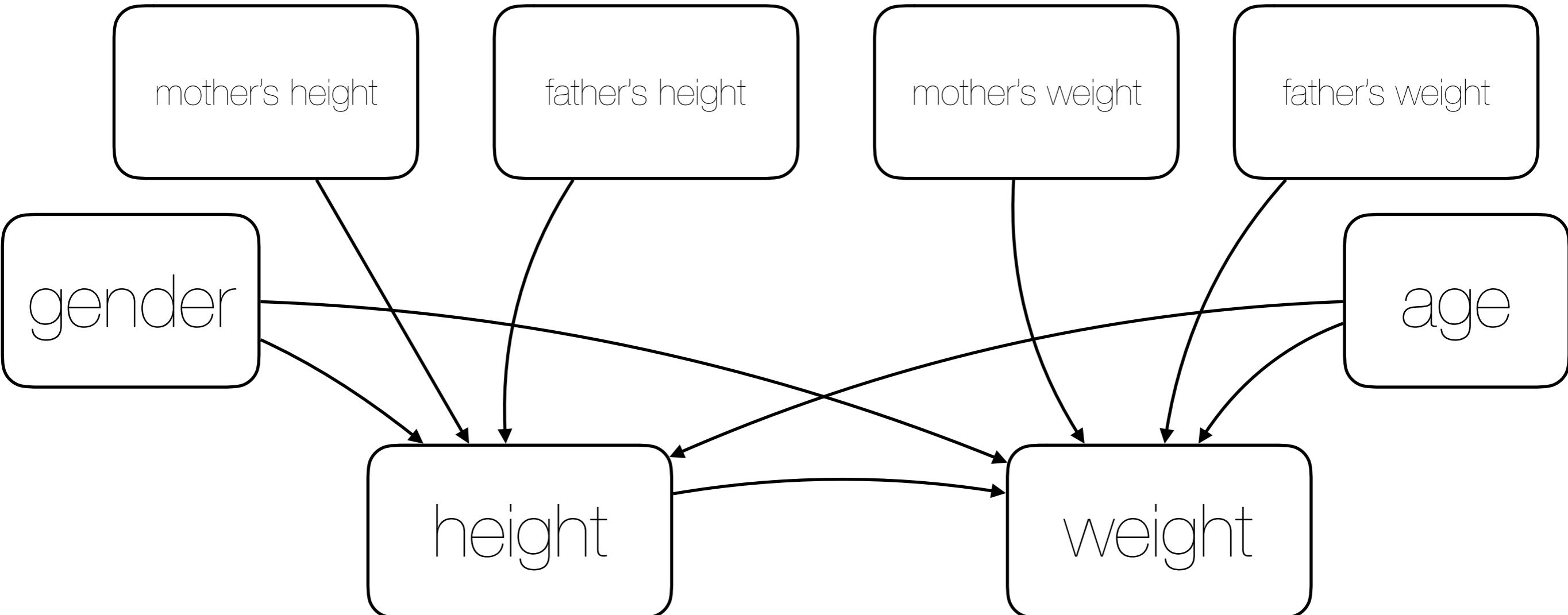
$$p_G(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i \mid x_{i_1}, \dots, x_{i_{l_i}}) .$$

- Q: How to describe those dependencies?
- Given a directed acyclic graph $G = (V, E)$ we encode a model as
 - One node $i \in V$ for each random variable.
 - One conditional probability distribution per node, where conditional factors become the parents in the graph:
- Properties of the joint distribution are now expressed in terms of the graph.
 - algorithms for inference
 - complexity questions.



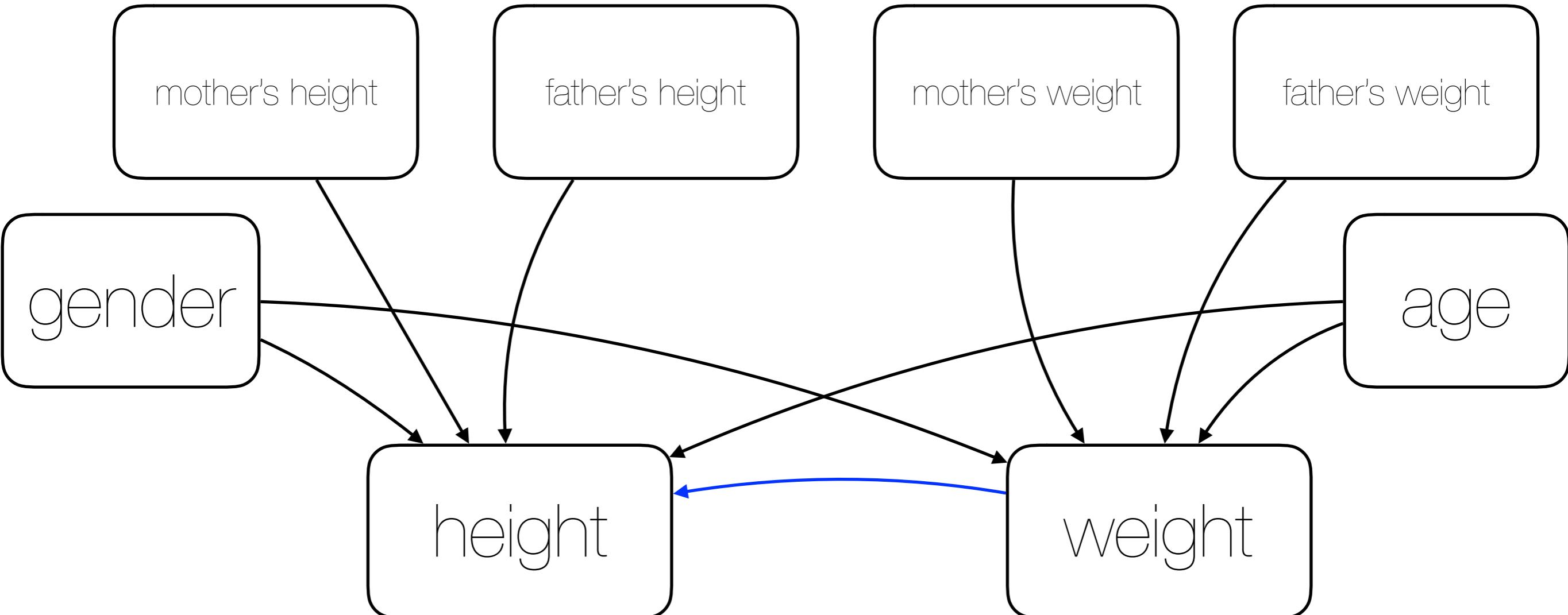
$$\{x_{i_1}, \dots, x_{i_{l_i}}\} := x_{Pa}(i)$$

Example



$$\begin{aligned} p(MH, FH, MW, FW, G, A, H, W) &= \\ &= p(G)p(A)p(MH)p(FH)p(MW)p(FW) \\ &\quad p(H \mid MH, FH, G, A)p(W \mid H, MW, FW, G, A) \end{aligned}$$

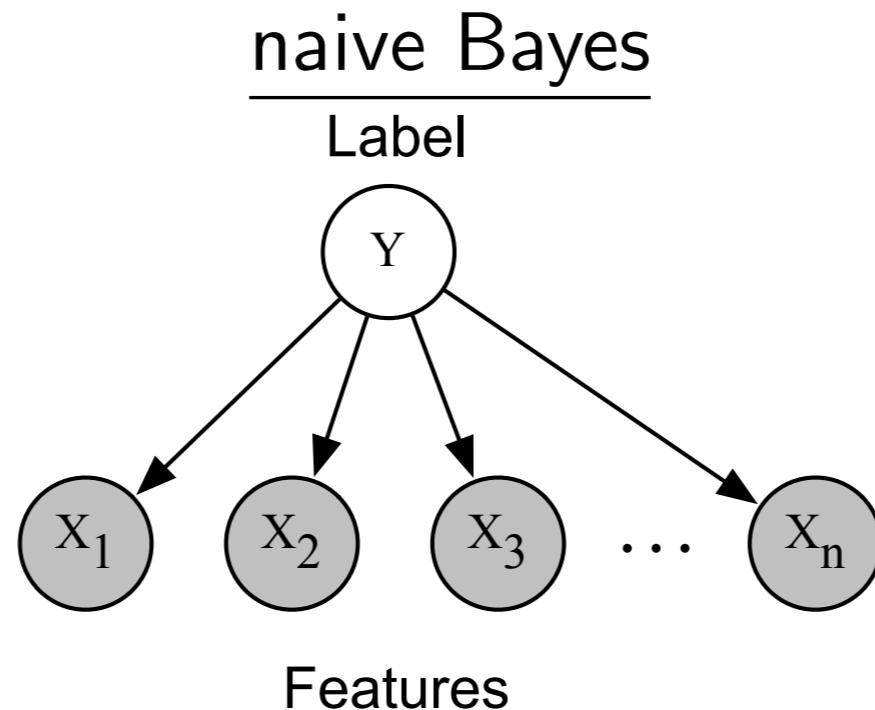
Example



$$\begin{aligned} p(MH, FH, MW, FW, G, A, H, W) &= \\ &= p(G)p(A)p(MH)p(FH)p(MW)p(FW) \\ &\quad p(W \mid MH, FH, G, A)p(H \mid W, MW, FW, G, A) \end{aligned}$$

same model!

Representing partially observed models



- Observing variables = shading nodes
- Posterior inference: $p(Y \mid X_1 = x_1, \dots, X_n = x_n)$
- Also, we can use a Bayesian network as a generative process:
 1. Sample $y \sim p(Y)$
 2. For each i , sample $x_i \sim p(X_i \mid Y = y)$.

Conditional Independence

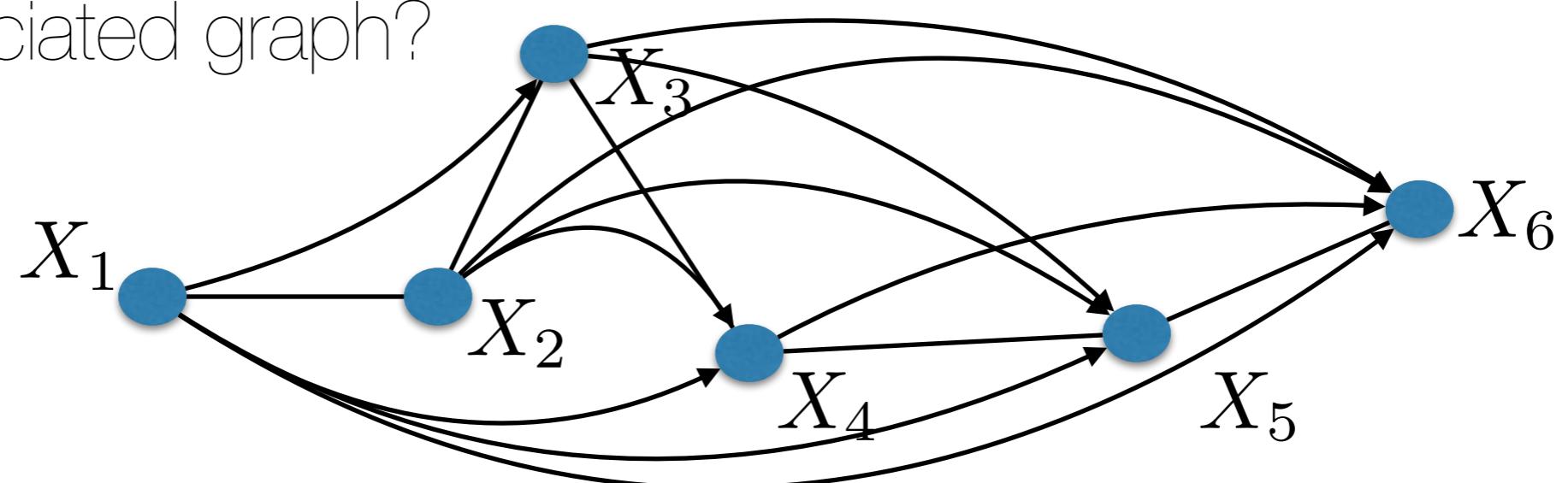
- Reminder:

$X_i \perp X_j \mid Y$ if $p(X_i, X_j \mid Y) = p(X_i \mid Y)p(X_j \mid Y)$

– equivalently, $p(X_i \mid X_j, Y) = p(X_i \mid X_j)$ whenever $p(X_j) > 0$.

- A generic joint distribution has no independence assumptions.

Which associated graph?



- Fully connected directed graph.

Conditional Independence

- In a specific graphical model, conditional probabilities have missing terms:

$$p(X_i \mid X_{i_1}, \dots, X_{i_{l_i}}) \text{ instead of } p(X_i \mid X_1, \dots, X_{i-1})$$

- Thus

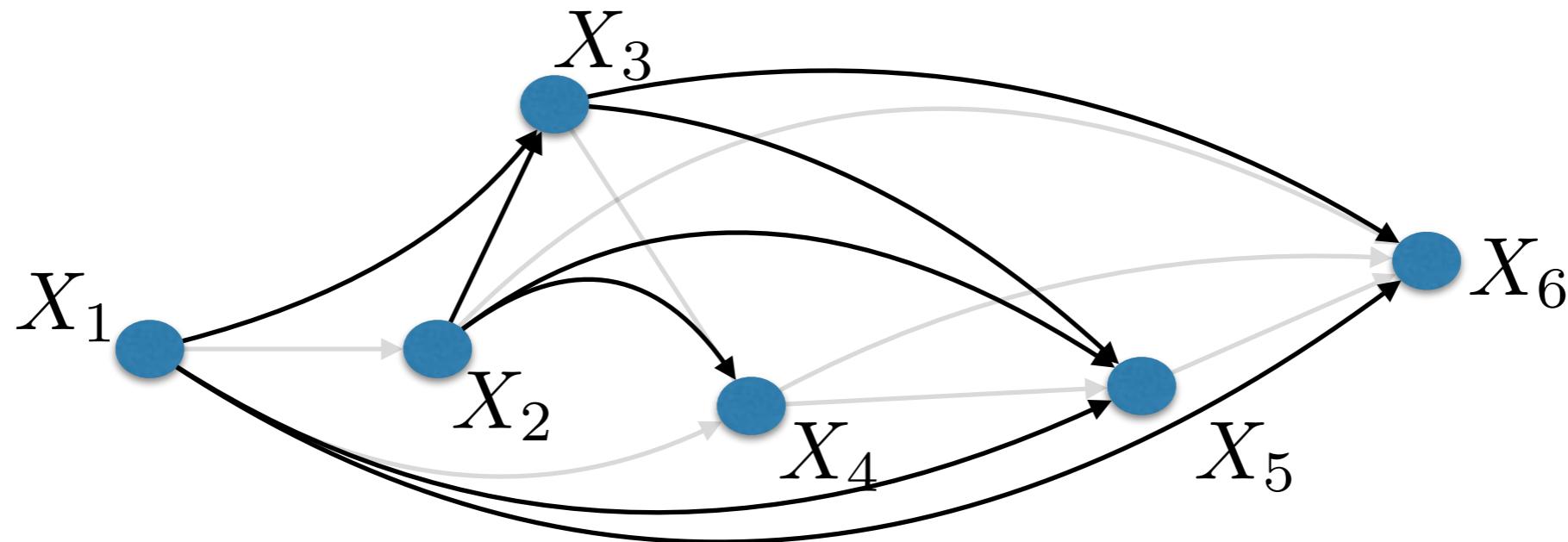
$$X_i \perp X_j \mid X_{Pa(i)} \text{ for } j < i, j \notin Pa(i).$$

- In fact, missing variables in the local conditional probabilities



missing edges in the corresponding graph

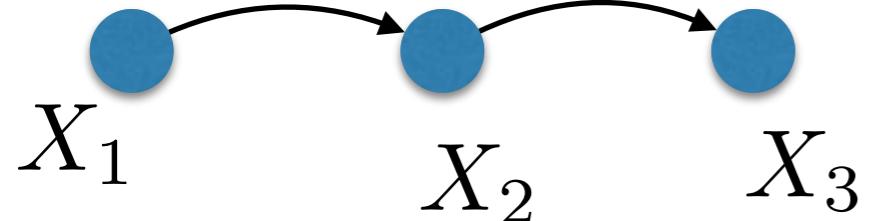
Conditional Independence



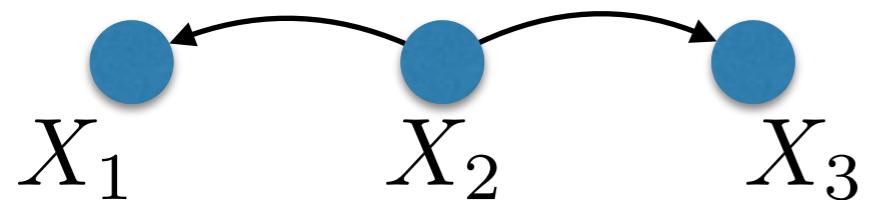
- So, removing edges creates conditional independences:
$$X_1 \perp X_2 \mid \emptyset$$
$$X_4 \perp \{X_1, X_3\} \mid X_2$$
$$X_5 \perp X_4 \mid \{X_1, X_2, X_3\}$$
$$X_6 \perp \{X_2, X_4, X_5\} \mid \{X_1, X_3\}$$
- A variable is independent from its non-descendants given its parents.
- Q: Are there other conditional independence relationships induced by the graphical model?

Independence in three-node graphs

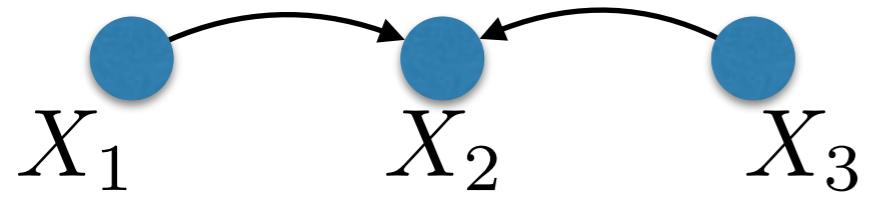
- Let's first focus on canonical, small graphical models.



"Cascade":
 $X_1 \perp X_3 \mid X_2$



"Common Parent"
 $X_1 \perp X_3 \mid X_2$



"V-structure"
Explaining away: knowing X_2 couples X_1 and X_3
 $X_1 \perp X_3$, but $X_1 \perp X_3 \mid X_2$ is false in general!

Example: Common Parent



We need to show that if $p(X_1, X_2, X_3)$ factors as

$$p(X_1, X_2, X_3) = p(X_2)p(X_1|X_2)p(X_3|X_2)$$

then

$$p(X_1, X_3 \mid X_2) = p(X_1 \mid X_2)p(X_3 \mid X_2)$$

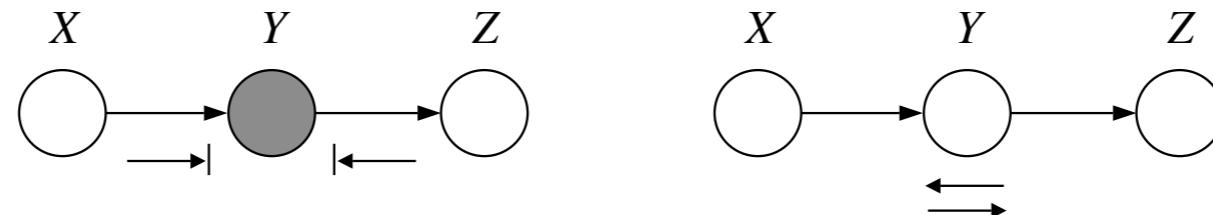
$$p(X_1, X_3 \mid X_2) = \frac{p(X_1, X_2, X_3)}{p(X_2)} = \frac{p(X_2)p(X_1 \mid X_2)p(X_3 \mid X_2)}{p(X_2)} \quad \square$$

D-separation

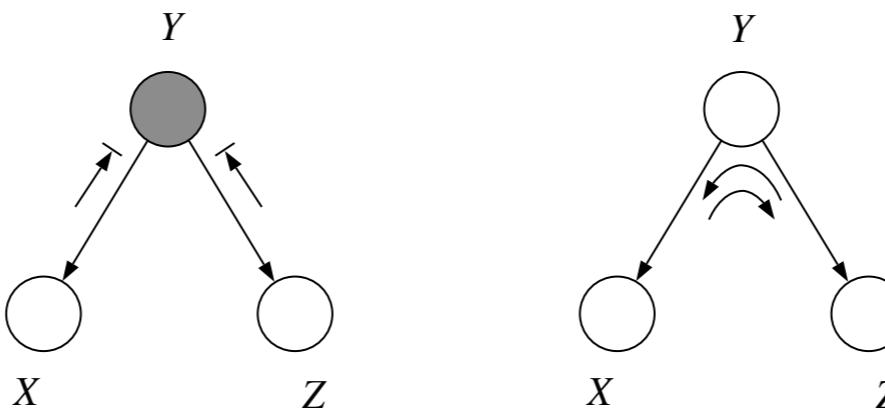
- In a general Bayesian Network, let us describe an algorithm to determine whether $X \perp Z | \mathbf{Y}$ by inspecting the graph.
- Idea: Determine an active path between X and Z when variables in \mathbf{Y} are observed.

- Use the canonical three-graphs as local rules:

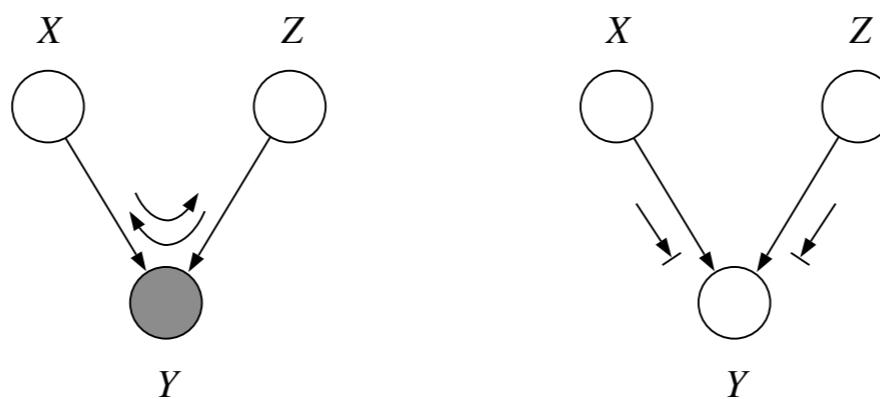
- Cascade:



- Common parent:



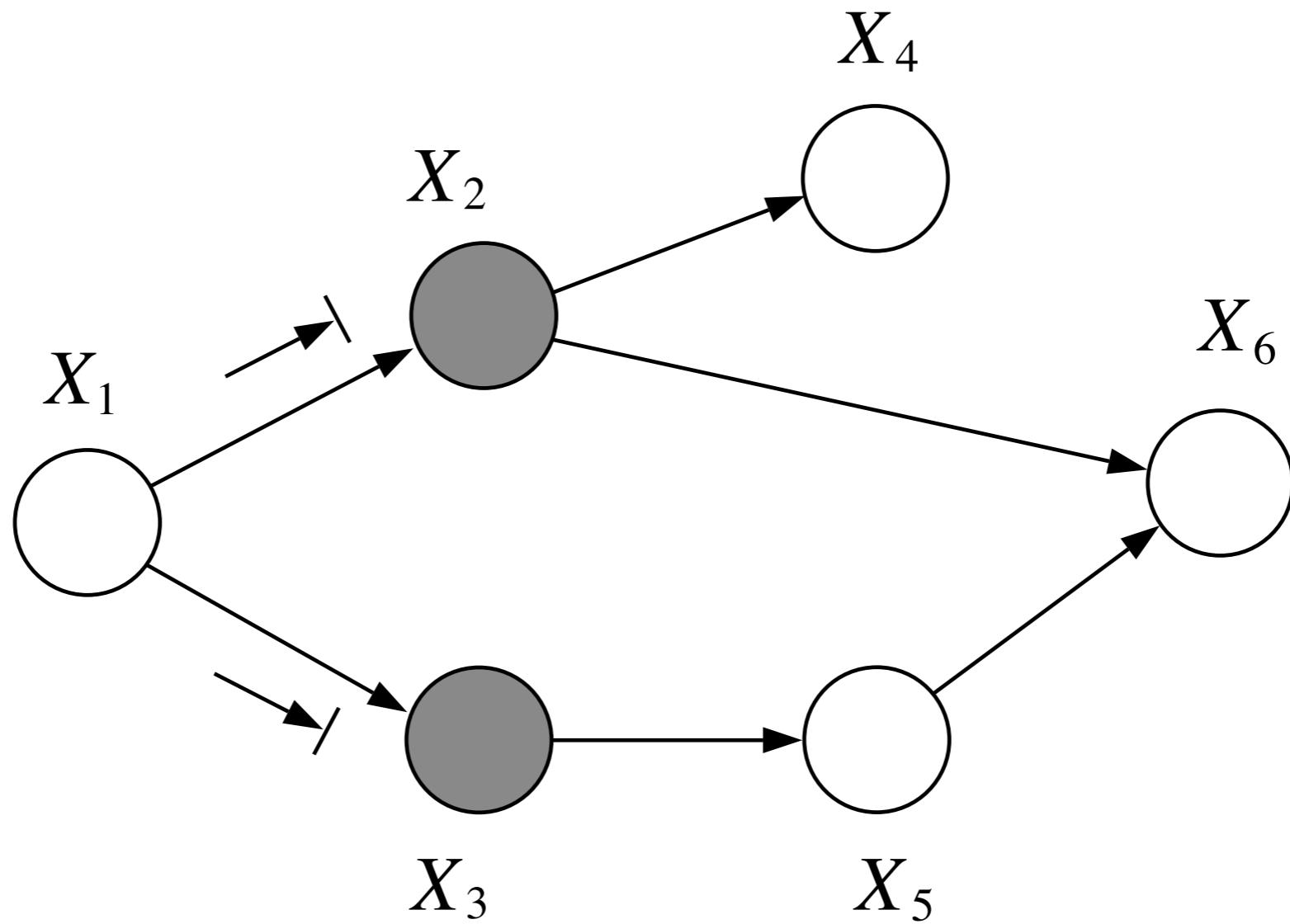
- V-structure:



D-separation

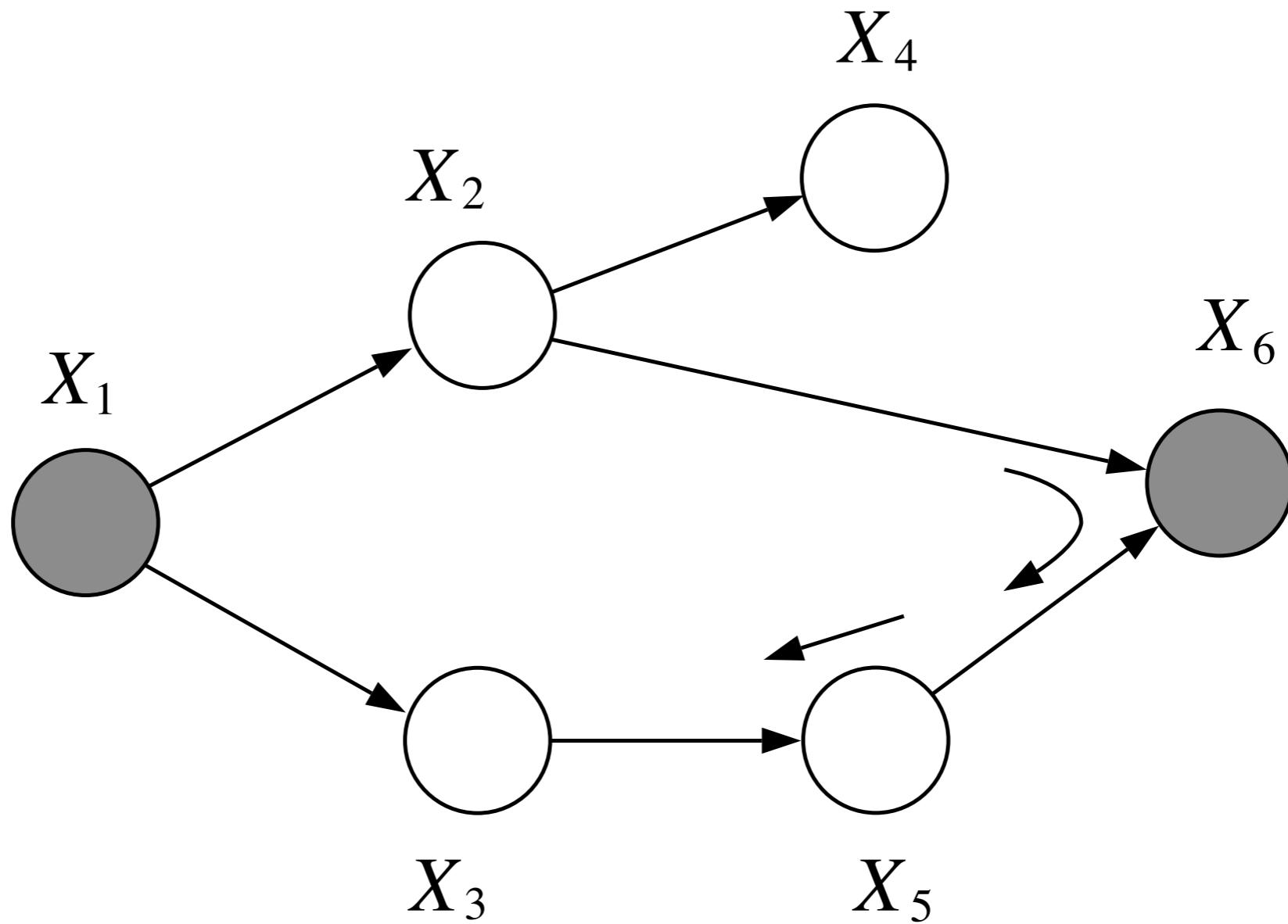
- If no such path exists, then we say that X and Z are d-separated wrt \mathbf{Y} , thus $X \perp Z \mid \mathbf{Y}$.
- Also referred as “Bayes Ball” algorithm
 - If we “throw” a ball from X , can it reach Z according to local rules?

D-separation example 1



Is $X_6 \perp X_5 | X_2, X_3?$ Is $X_4 \perp X_5 | X_2, X_3?$

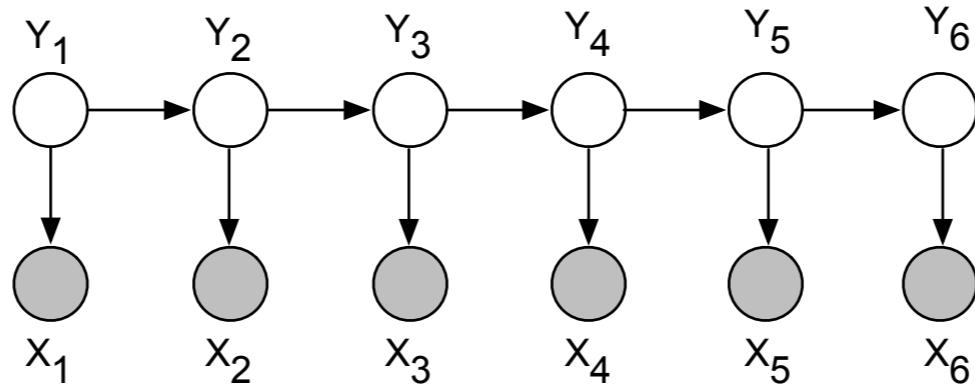
D-separation example 2



Is $X_4 \perp X_5 | X_1, X_6$?

What about if X_6 is not observed? I.e., is
 $X_4 \perp X_5 | X_1$?

Hidden Markov Models



- Very popular (today less so) in speech recognition and part-of-speech tagging.
- State-space model (more on that later in the course).
- Joint distribution

$$p(\mathbf{X}, \mathbf{Y}) = p(Y_1)p(X_1 \mid Y_1) \prod_{t=2}^T p(Y_t \mid Y_{t-1})P(X_t \mid Y_t) .$$

- $p(Y_t \mid Y_{t-1})$ is the *transition* probability kernel.
- $p(X_t \mid Y_t)$ is the output probability.
- Q: What are the conditional independencies?

past \perp future \mid present

Independence Maps

- Given a directed acyclic graph \mathbf{G} , we denote by $I(\mathbf{G})$ the set of conditional independencies implied by it.
- Given a joint distribution p , we denote by $I(p)$ the set of conditional independencies satisfied by it.
- \mathbf{G} is an *independence map* of p if $I(\mathbf{G}) \subseteq I(p)$.
 - Dense graphs are i-maps of any distribution (why?)
- \mathbf{G} is a *minimal independence map* if pruning \mathbf{G} (thus enlarging $I(\mathbf{G})$) makes it not an i-map.
 - Each node ordering may correspond to a different minimal i-map.
- \mathbf{G} is a *perfect independence map* for p if $I(\mathbf{G}) = I(p)$
 - Q: Does every distribution have a perfect map?

$$X, Y \sim \text{Ber}(0.5), Z = X \text{ XOR } Y.$$

Equivalent Graph Structures

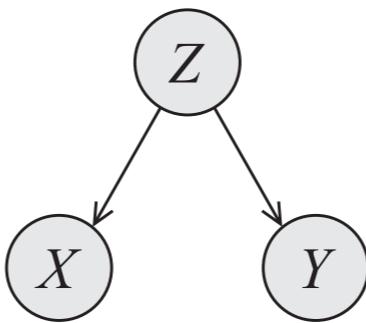
- Two different BNs can be equivalent: they encode the same conditional independence assumptions.
- Which of these are equivalent?



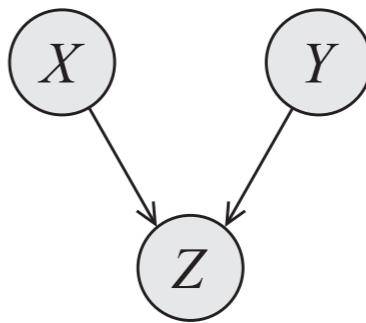
(a)



(b)



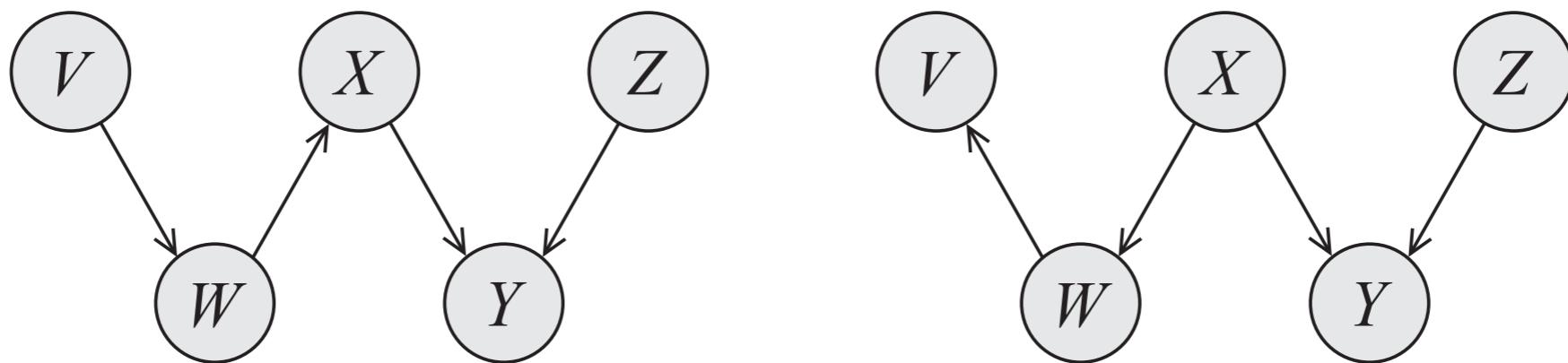
(c)



(d)

Equivalent Graph Structures

- Two different BNs can be equivalent: they encode the same conditional independence assumptions.
- Are these equivalent?



Lecture 1 Summary

- Bayesian Networks are constructed from a DAG G and a set of local conditional probability distributions.
- BN can be seen as a generative model specified by its topological order.
- Local and global conditional independence properties can be identified with the notion of d-separation.