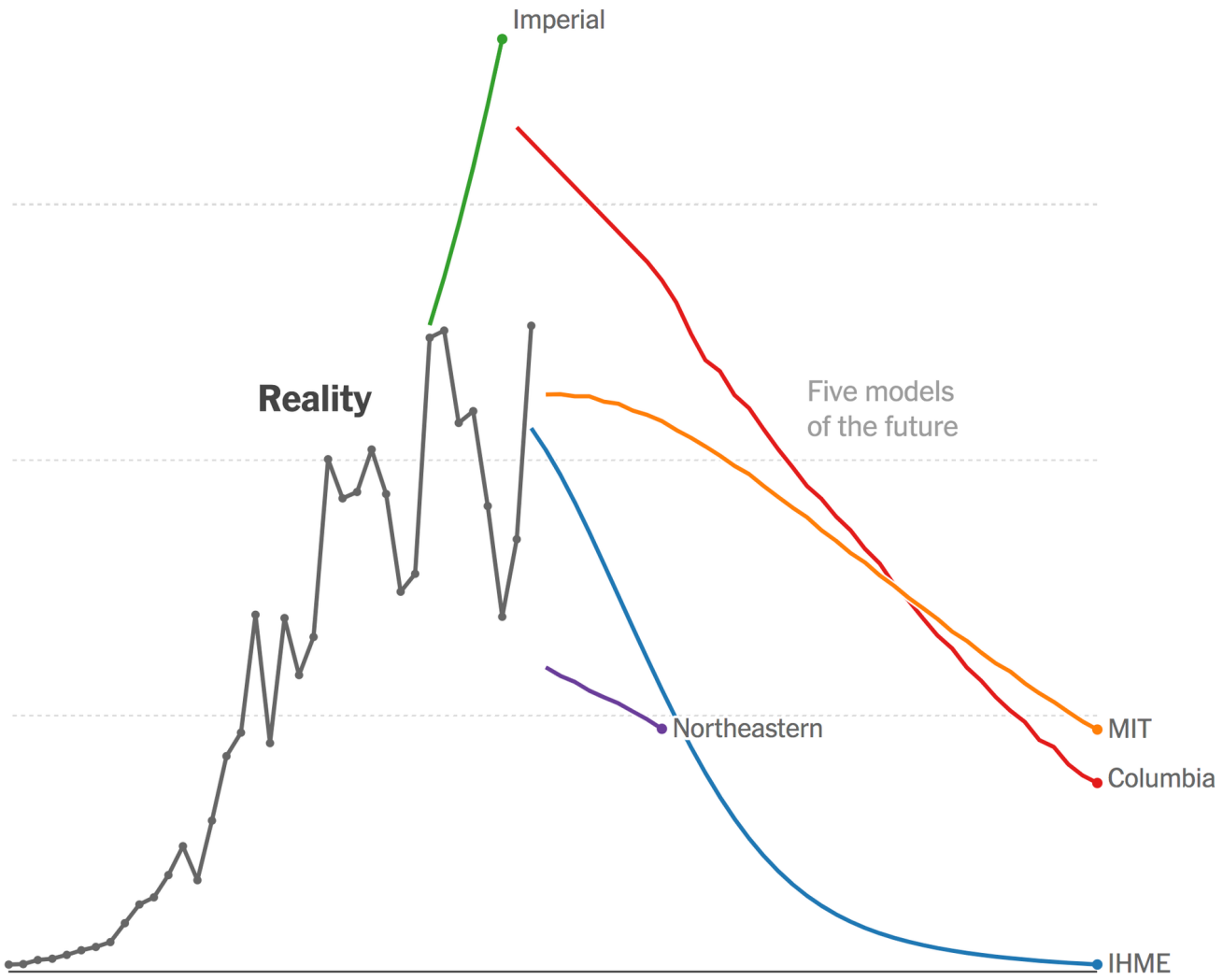
CSC696H: Probabilistic Methods in Machine Learning

Introduction and Course Overview

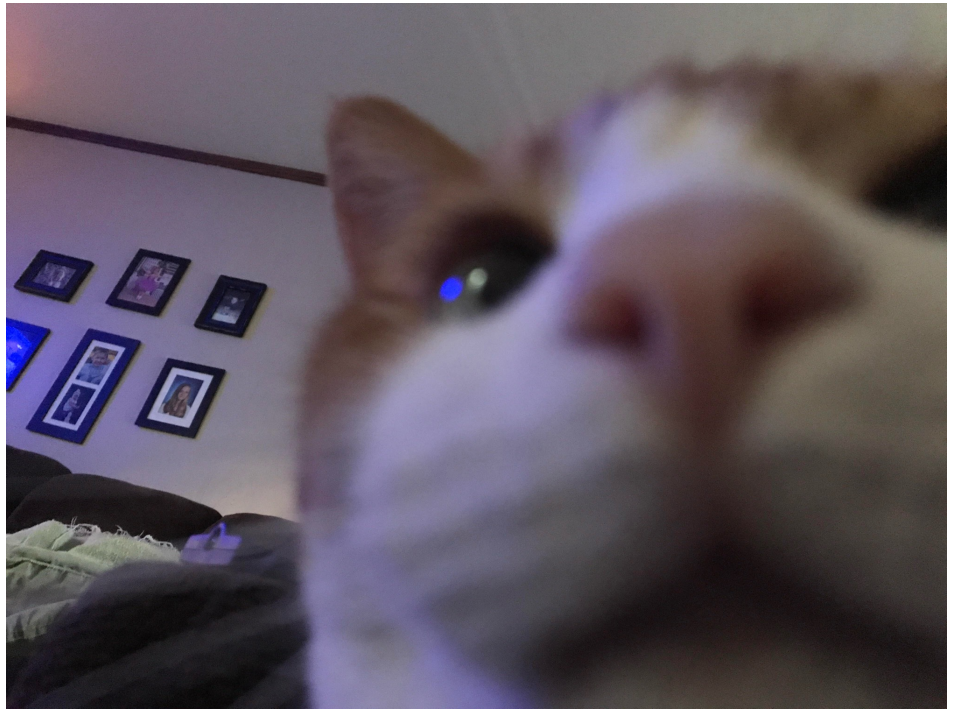
Prof. Jason Pacheco

Hurricane Prediction

Disease Prediction

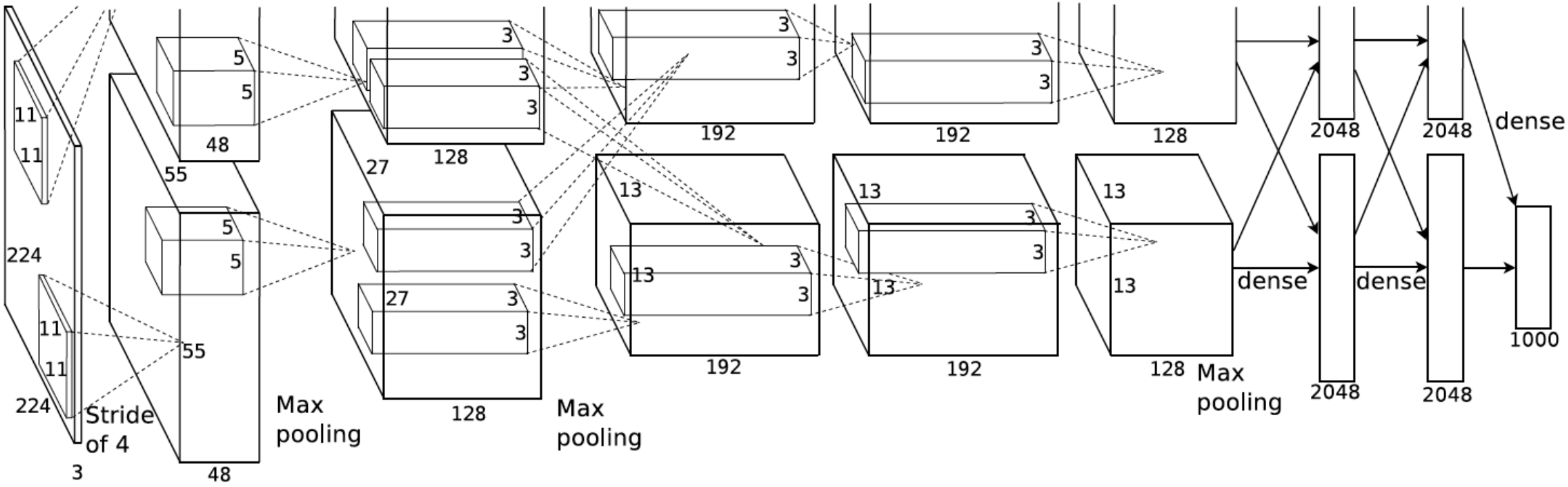


Object Detection



Modern Neural Networks

Modern *Deep Neural networks* add many hidden layers



…and some have many billions of parameters to learn

[Source: Krizhevsky et al. (NIPS 2012)]

Brittleness : Discontinuities in Predictions

Nearly imperceptible changes to input change prediction

All images in right column predicted as “ostrich”

Szegedy et al. “Intriguing properties of neural networks.” ICLR 2014

Safety Concerns

Variety of black-box *physical attacks* left-to-right:

- Artistic graffiti
- Subtle graffiti
- Poster

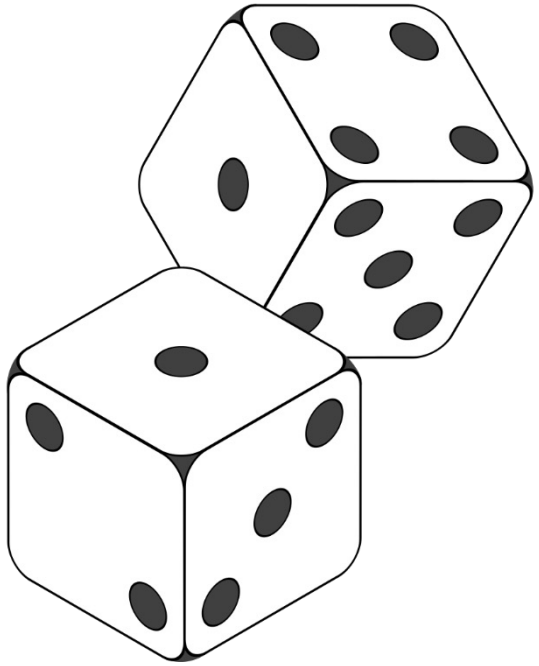
Can reliably cause ANN to misclassify as intended target (e.g. speed limit 45mph)

Does not require knowledge of network internals

Evtimov et al. "Robust Physical-World Attacks on ML Models." 2017

Probability Theory

Suppose we roll <u>two fair dice</u>…

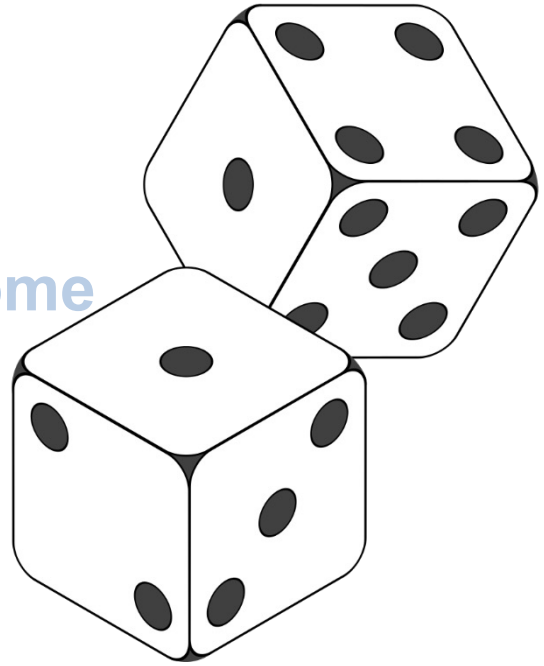


- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?
- What is the *probability* of rolling **odd** numbers?

…probability theory *gives a mathematical formalism to address such questions*

Random Events and Probability

Definition An **outcome** is a possible result of an experiment or trial, and the collection of all possible outcomes is the **sample space** of the experiment,



Example (1,1), (1,2), …, (6,1), (6,2), …, (6,6)

Outcome

Sample Space

Definition An **event** is a *set* of outcomes (a subset of the sample space),

Example Event Roll at least a single 1

{(1,1), (1,2), (1,3), …, (1,6), …, (6,1)}

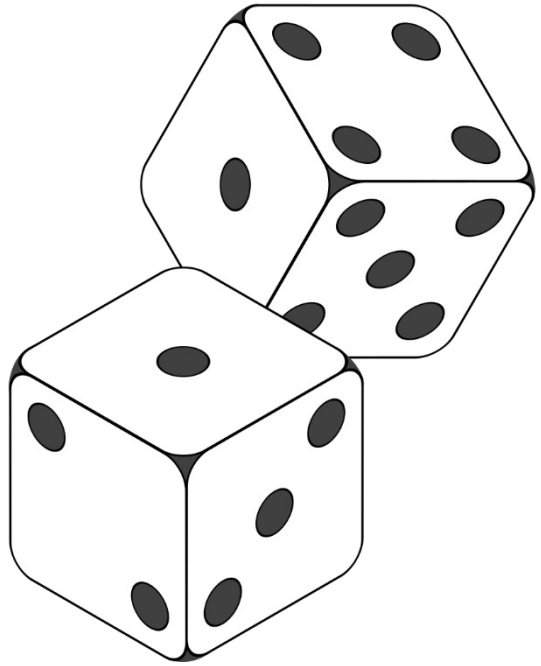
Random Events and Probability

Assume each outcome is equally likely, and sample space is finite, then the probability of event is:

$$P(E) = \frac{|E|}{|\Omega|}$$

Number of outcomes in event set

Number of possible outcomes in sample space



This is the **uniform probability distribution**

Example Probability that we roll *only* even numbers,

$$E^{\text{even}} = \{(2,2),(2,4),\ldots,(6,4),(6,6)\}$$

$$P(E^{\text{even}}) = \frac{|E^{\text{even}}|}{|\Omega|} = \frac{9}{36}$$

Probabilistic Inference

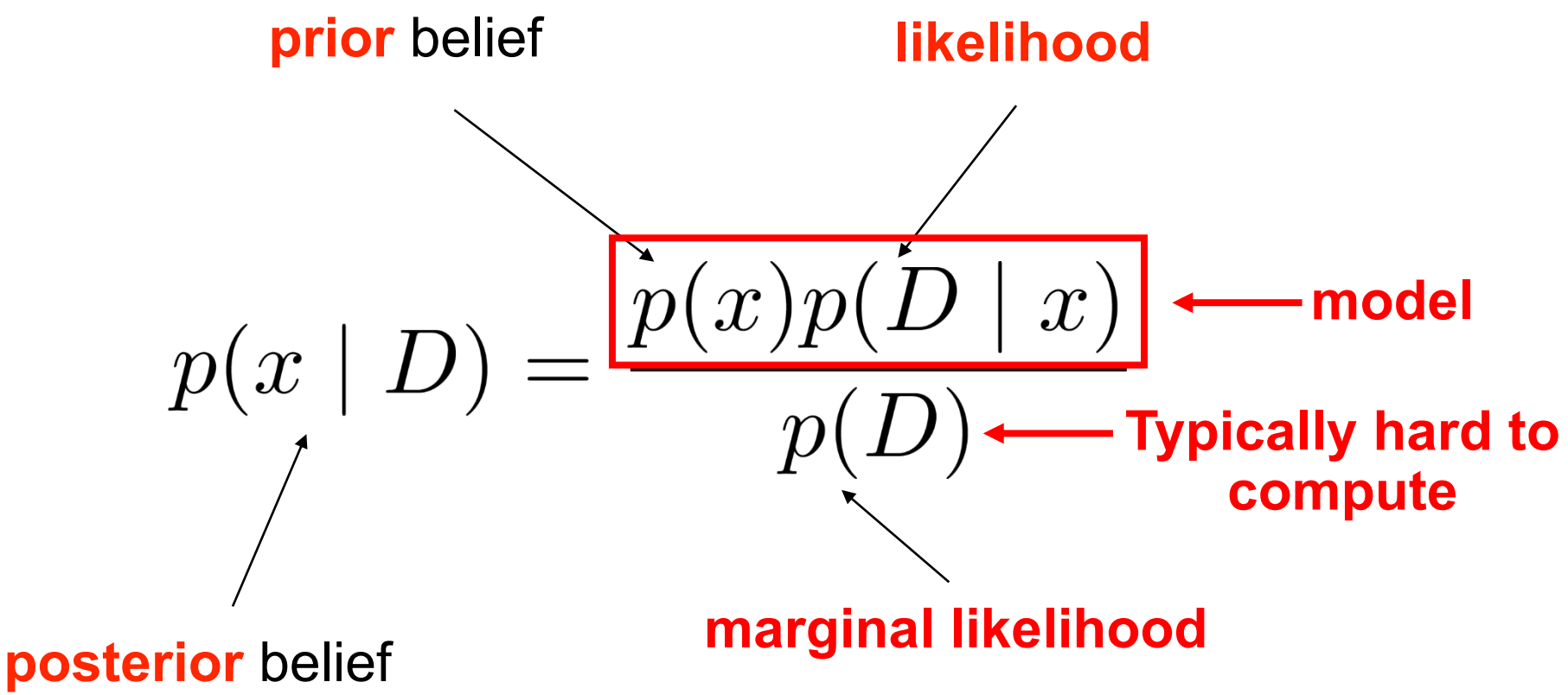
$$P(\text{what we don't know} \mid \text{what we do know})$$

Read as "given" or "conditioned on"

Bayesian Inference

X **Quantity to be inferred**

D **Observed Data**



Posterior encodes our *belief* about unknowns *given* data

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

A recent home test states that you have high BP. Should you start medication?

An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

Suppose we get a positive measurement, then posterior is:

$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

What conclusions can be drawn from this calculation?

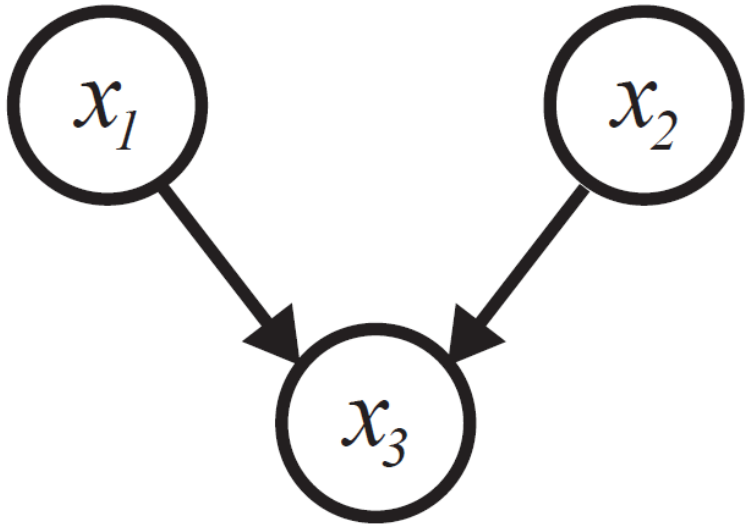
What is a Probabilistic Graphical Model?

*A probabilistic graphical model allows us to pictorially represent a probability distribution**

Probability Model:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$

Graphical Model:

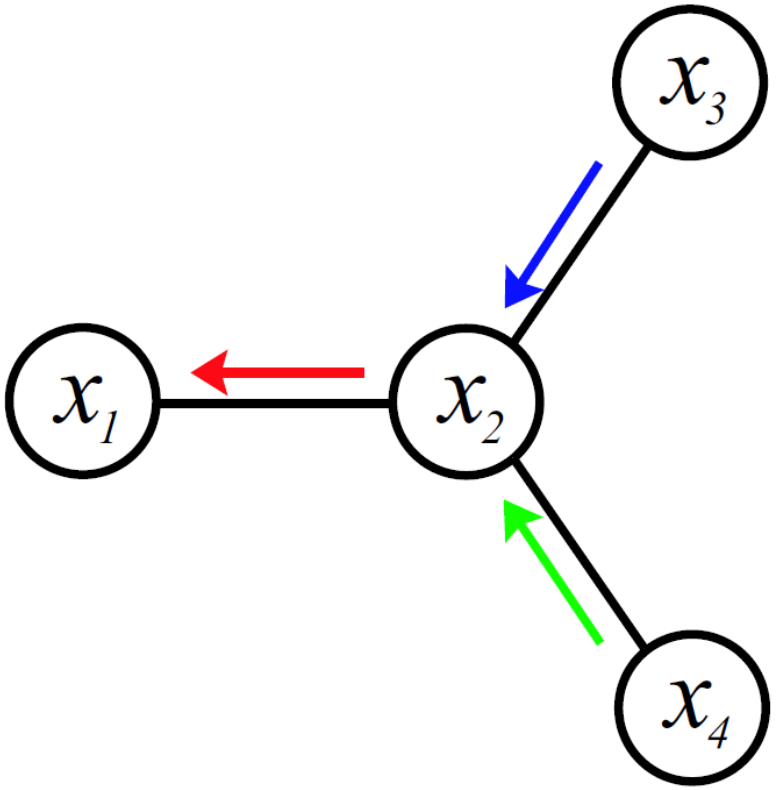


The graphical model structure *obeys* the factorization of the probability function in a sense we will formalize later

*** We will use the term "distribution" loosely to refer to a CDF / PDF / PMF**

Why Graphical Models?

Structure simplifies both **representation** and **computation**



Representation

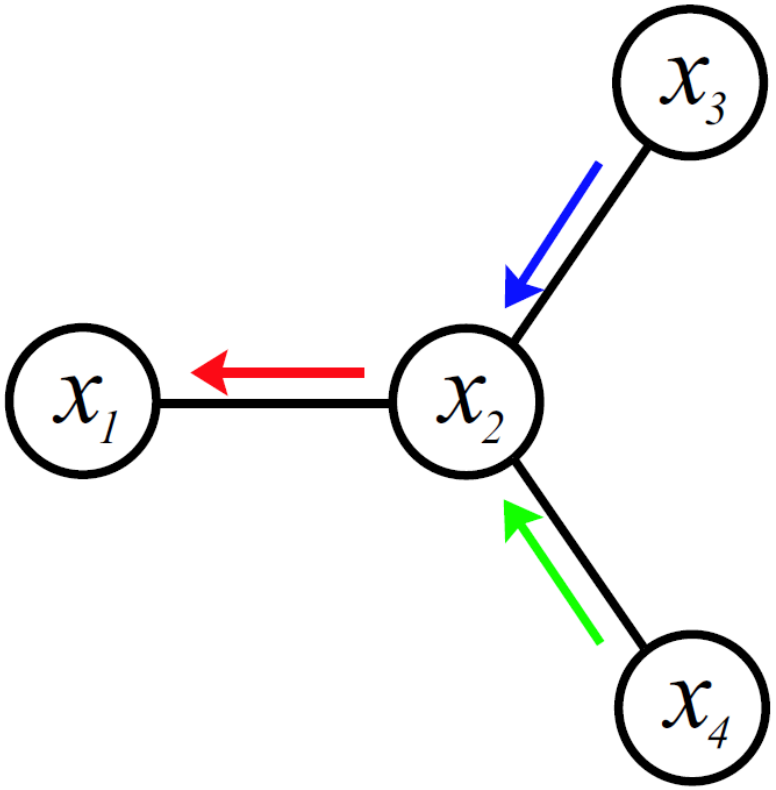
Complex global phenomena arise by simpler-to-specify local interactions

Computation

Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)

Why Graphical Models?

Structure simplifies both **representation** and **computation**



Representation

Complex global phenomena arise by simpler-to-specify local interactions

Computation

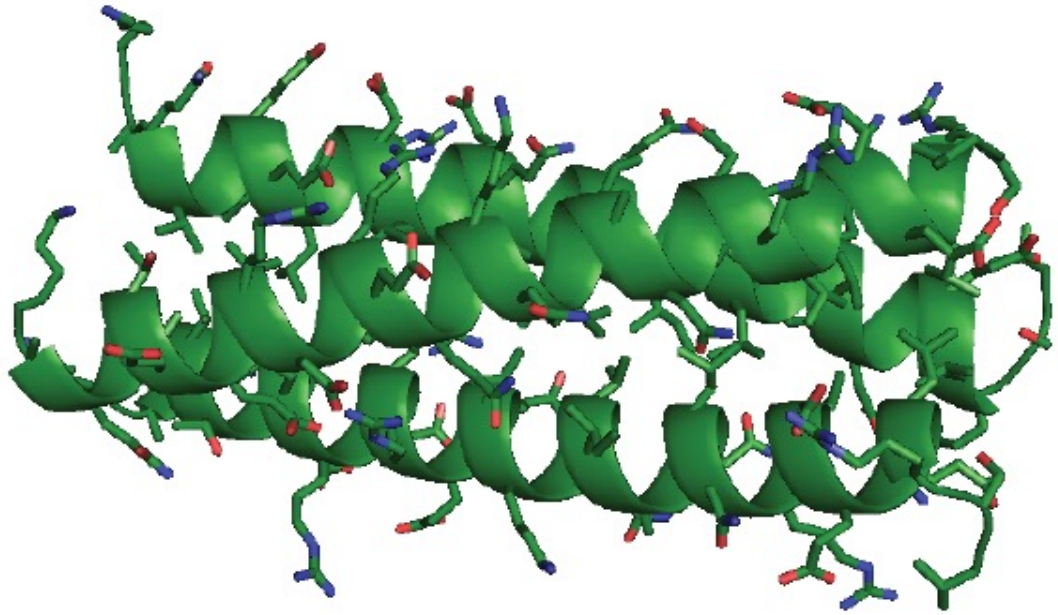
Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)

We will discuss inference later, but let's focus on representation…

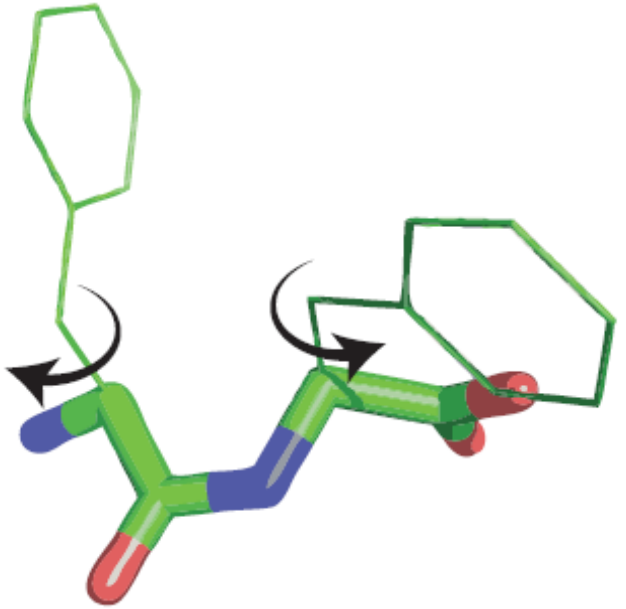
Protein Side Chain Prediction

Problem: Given 3D protein backbone structure, estimate orientation of every side chain molecule.

Backbone + Side Chains



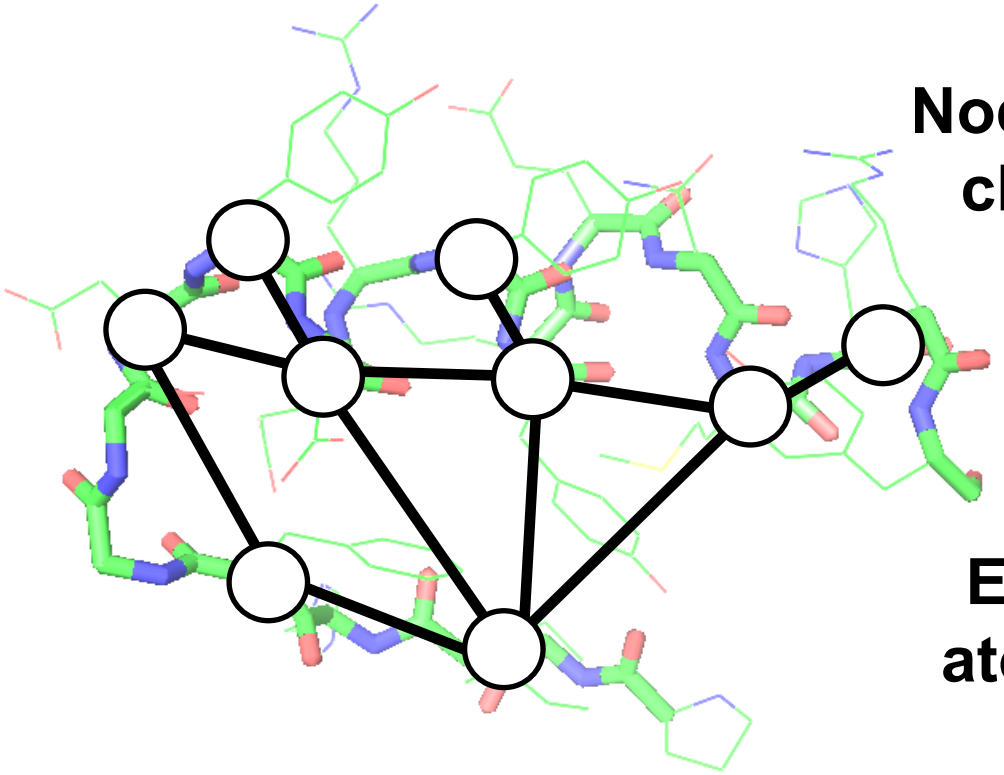
Side Chain Rotation



Solution: Just physics of atomic interaction. Easy, right!?

Protein Side Chain Prediction

Graphical Model

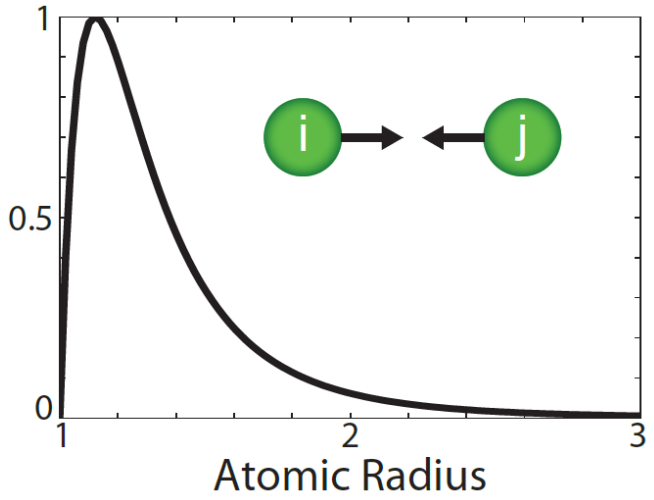


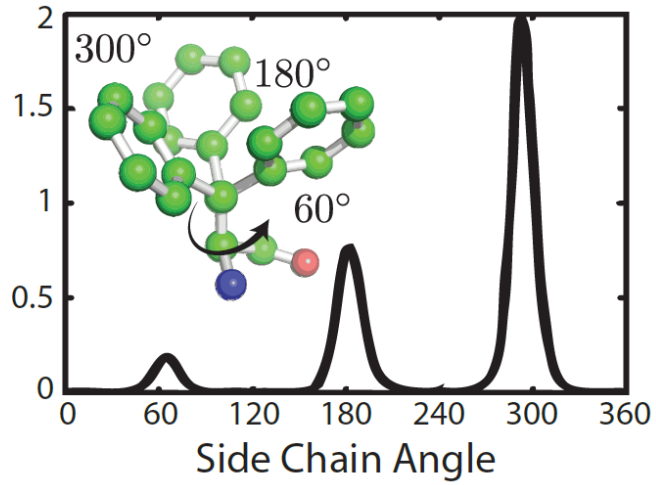
Nodes represent side chain orientations

Edges represent atomic interaction

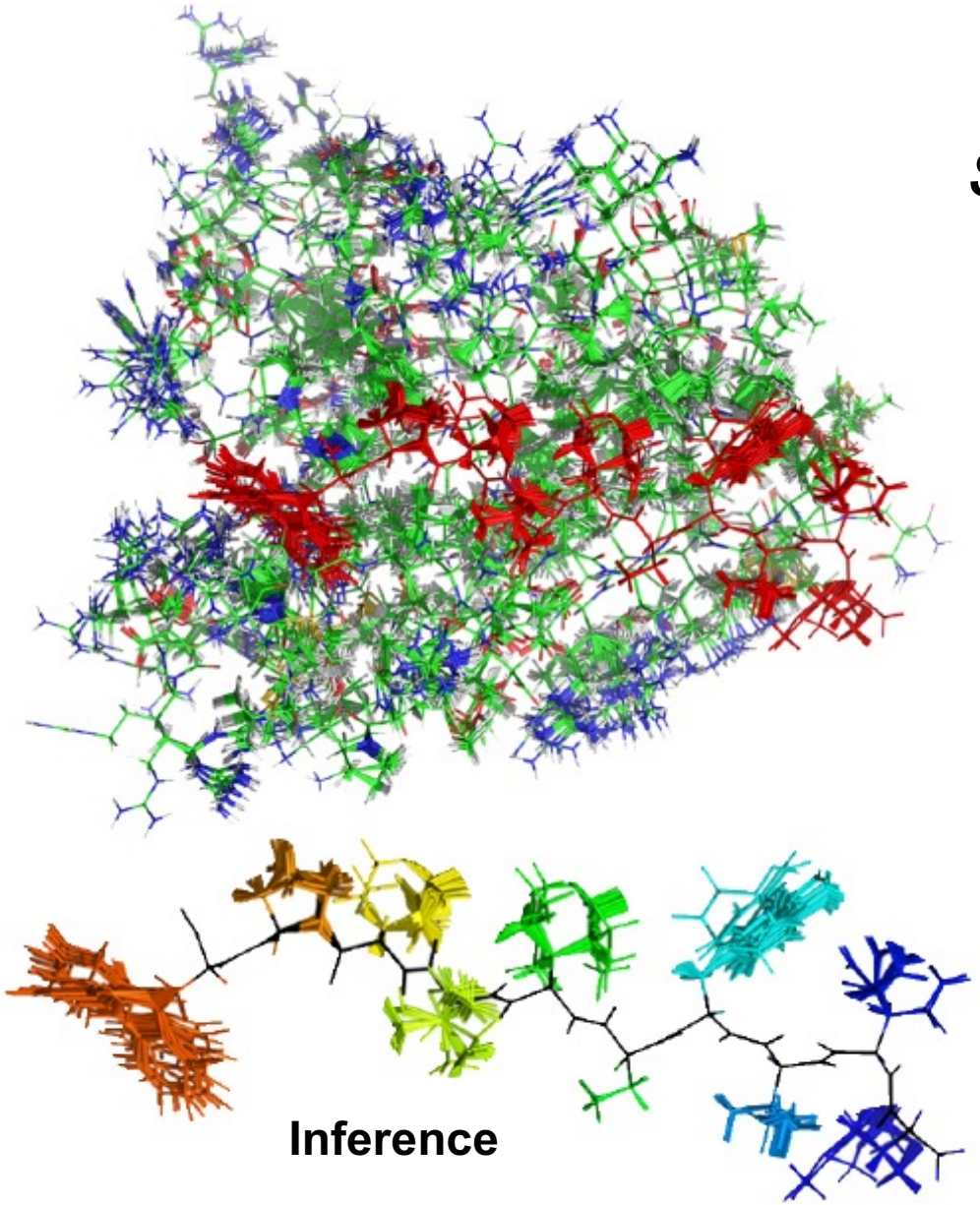
Complex phenomena specified by simpler atomic interactions

Configuration Likelihoods

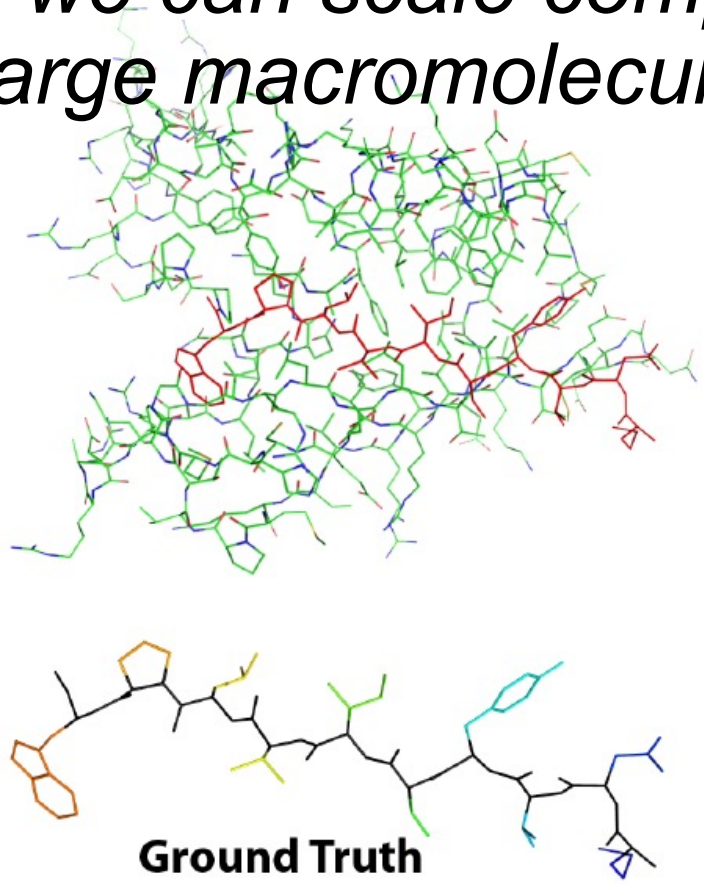




Protein Side Chain Prediction



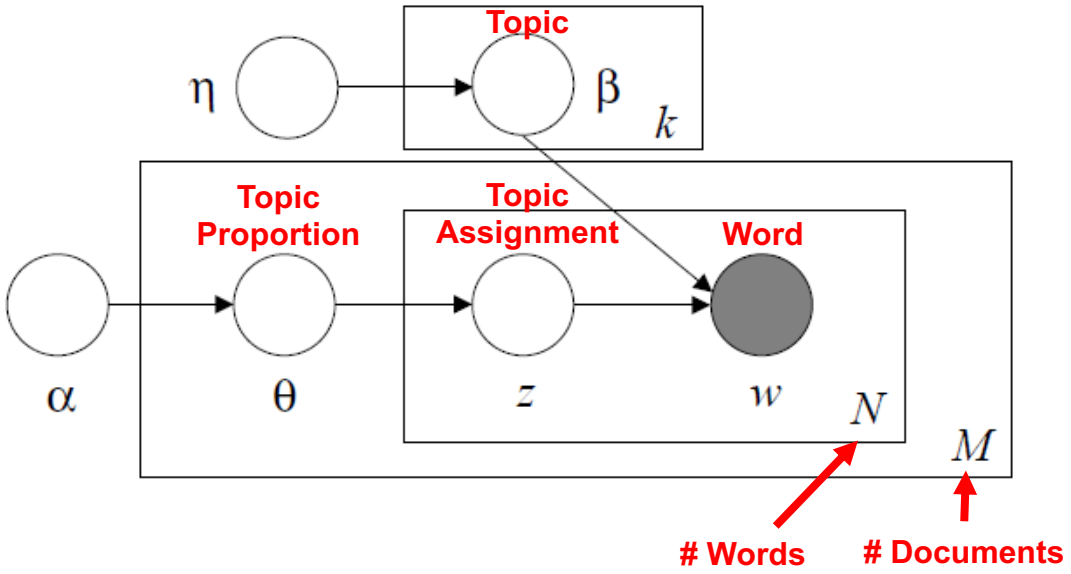
By exploiting graphical model structure we can scale computation to large macromolecules



Pacheco and Sudderth, ICML 2015

Topic Models

Latent Dirichlet Allocation (LDA)



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

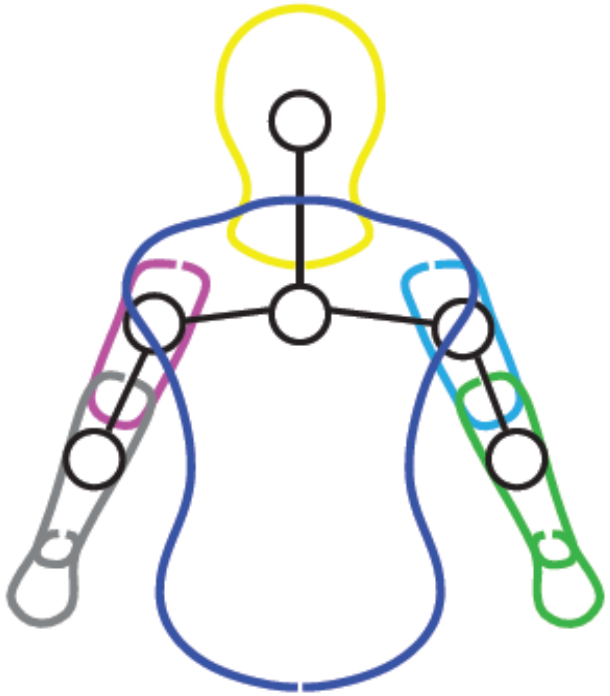
Allows *unsupervised learning* of document corpus via mixture modeling

Blei, D. et al., JMLR 2003

Pose Estimation

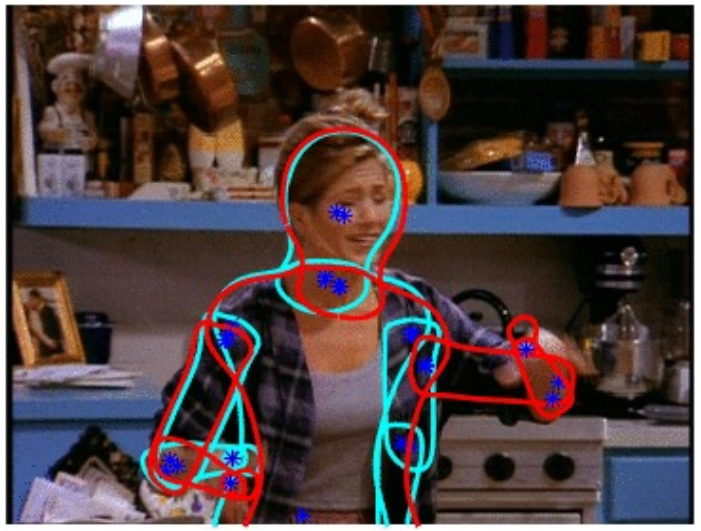
Estimate orientation / shape / pose of human figure from an image

Graphical Model

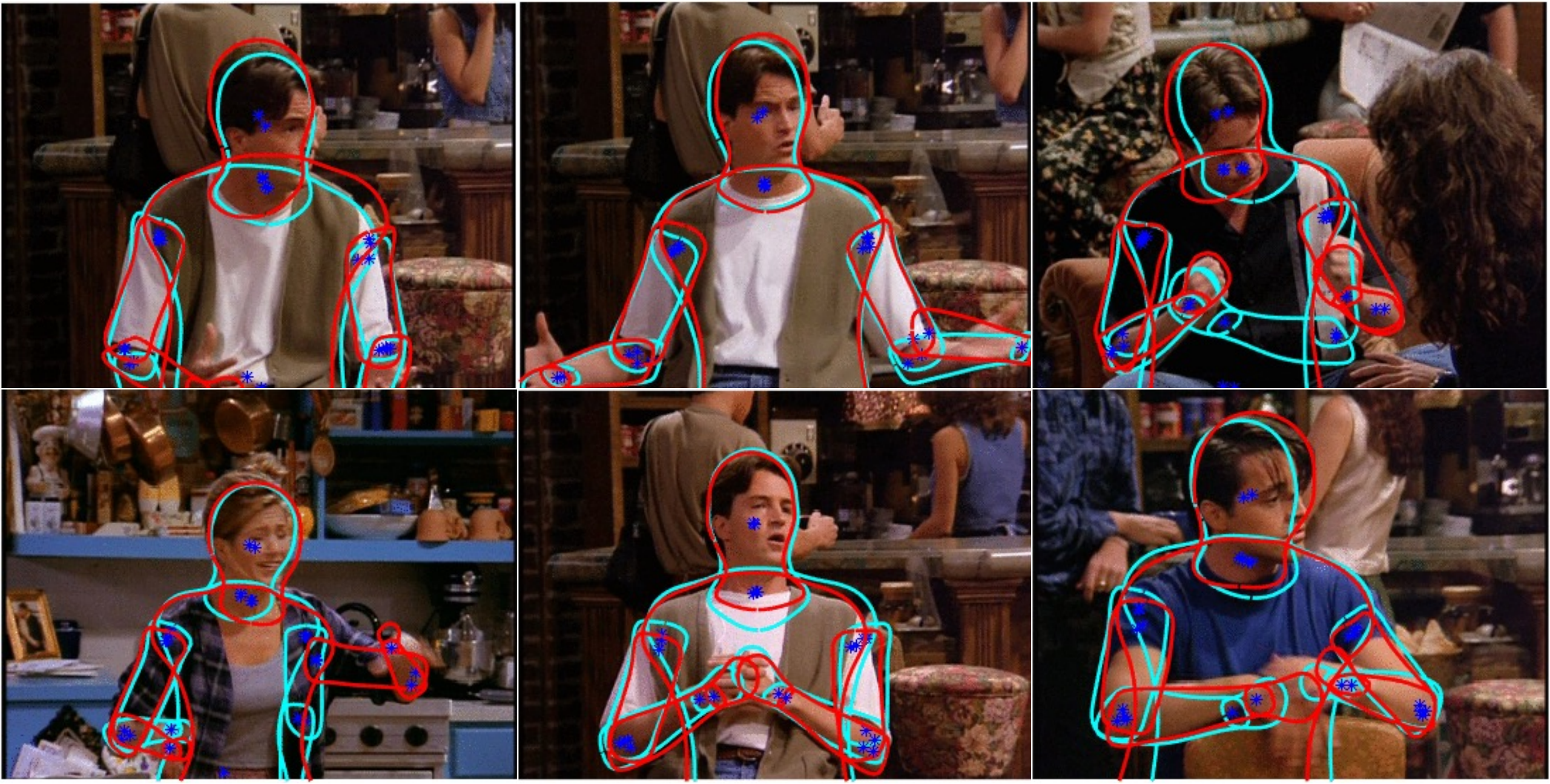


Data

Estimates

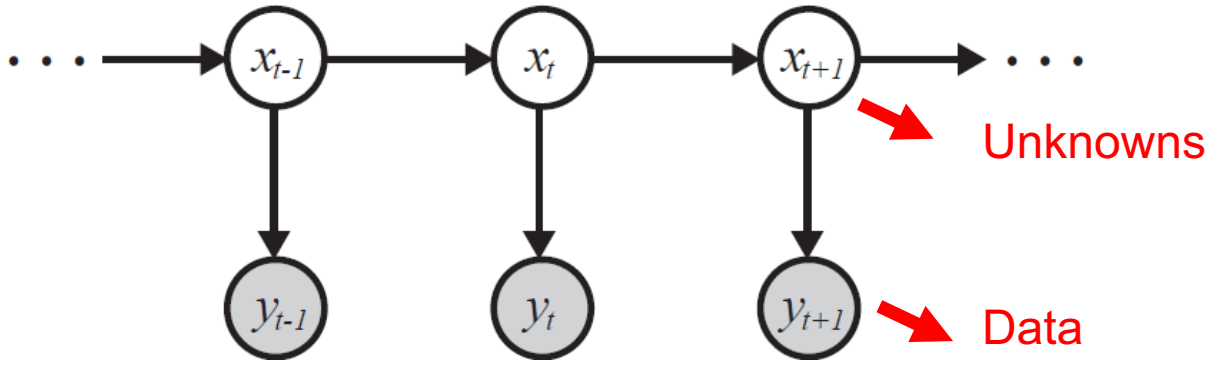


Pacheco, et al., ICML 2014



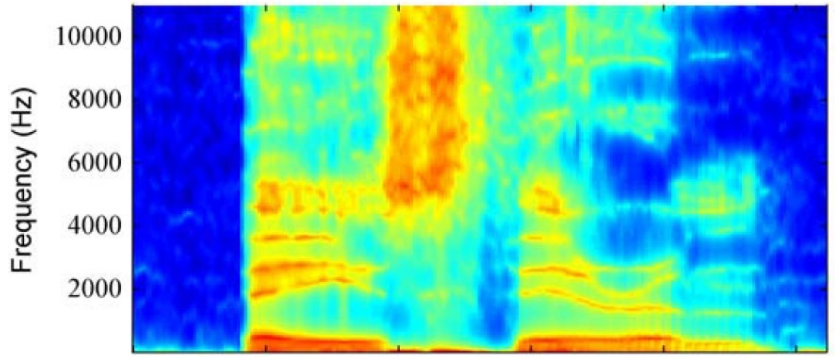
Hidden Markov Models

Sequential models *of* *discrete* *quantities of interest*



Example: Speech Recognition

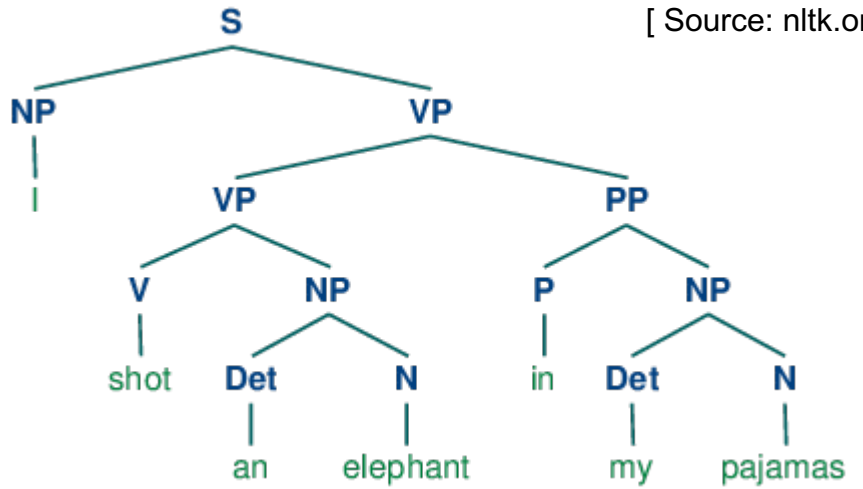
$y =$



x = b-ey-z-th-ih-er-em → Bayes' Theorem

[Source: Bishop, PRML]

[Source: nltk.org]



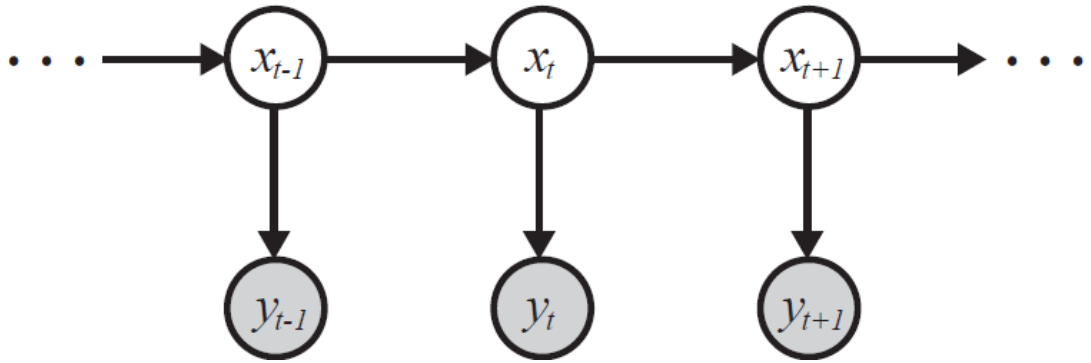
Example: Part-of-speech Tagging:

y = "I shot an elephant in my pajamas."

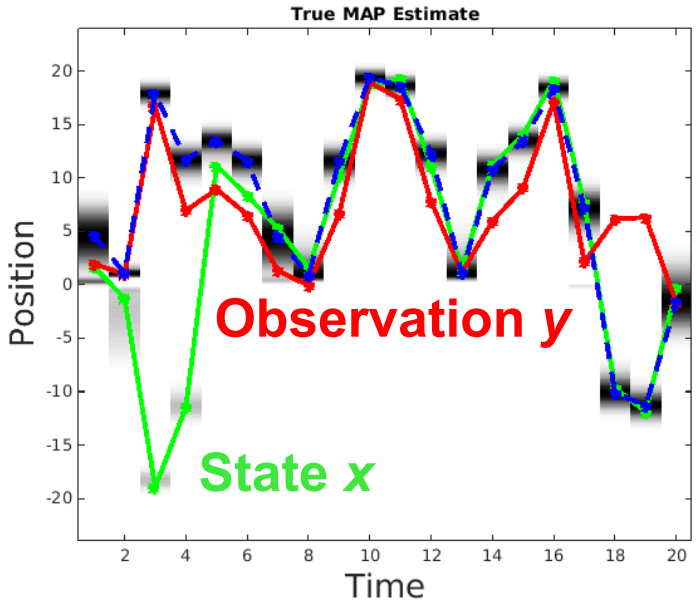
x = NP-V-Det-N-P-Det-N

Dynamical Models

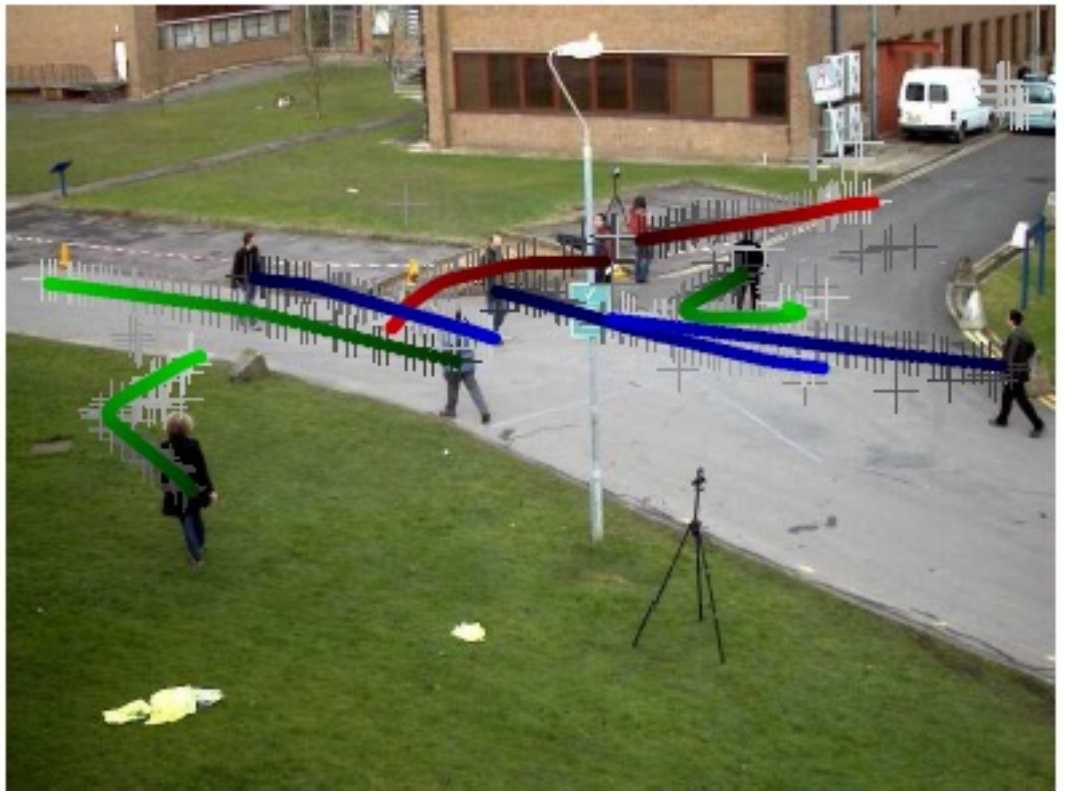
Sequential models *of* *continuous* *quantities of interest*



Example: Nonlinear Time Series

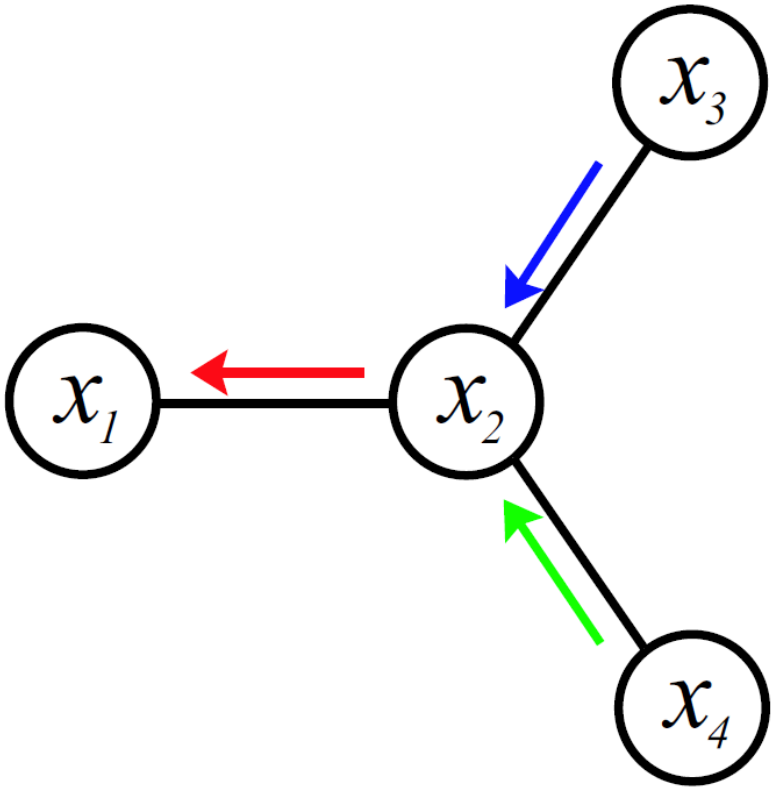


Example: Multitarget Tracking



Why Graphical Models?

Structure simplifies both **representation** and **computation**



Representation

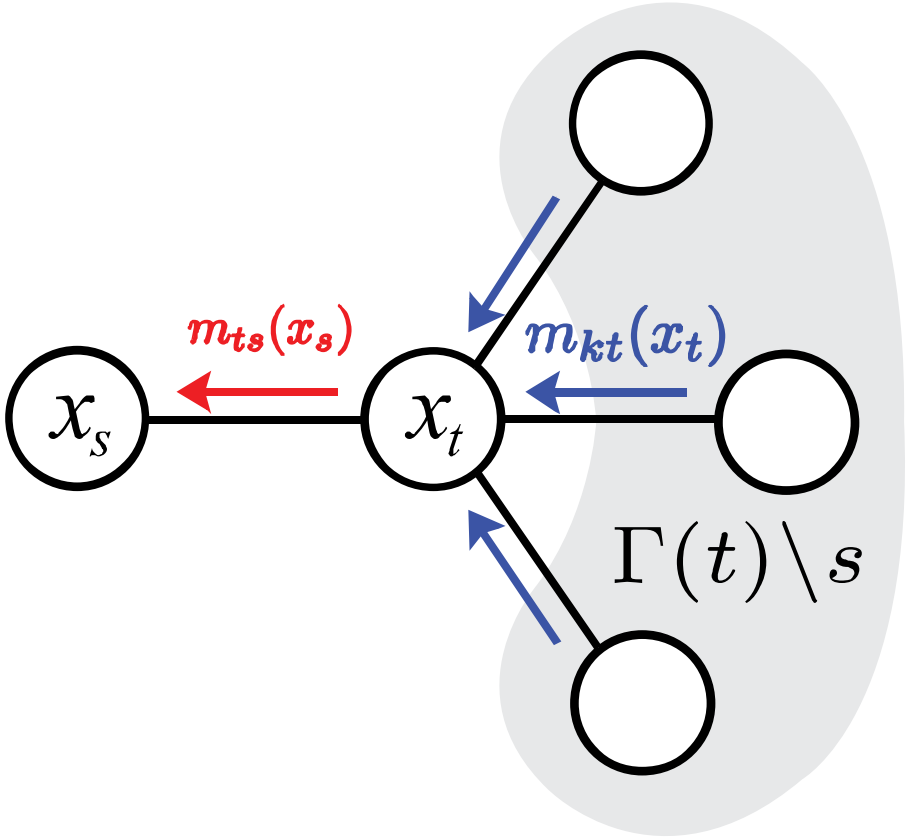
Complex global phenomena arise by simpler-to-specify local interactions

Computation

Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)

Dynamic Programming (DP)

Breaks difficult global computations into simpler local updates

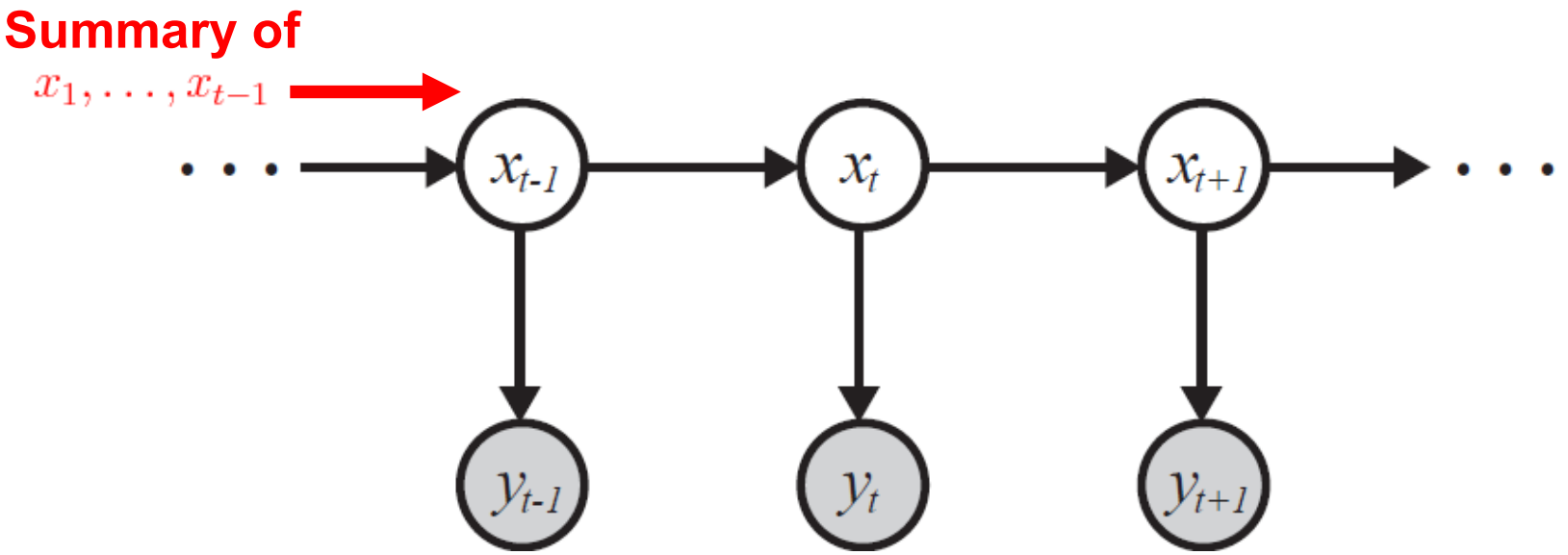


Many algorithms use some form of DP

- Belief propagation
- Gibbs sampling
- Particle filtering
- Viterbi decoder for HMMs
- Kalman filter (marginal inference)

Key Idea: Local computations only depend on the statistics of the current node and neighboring interactions

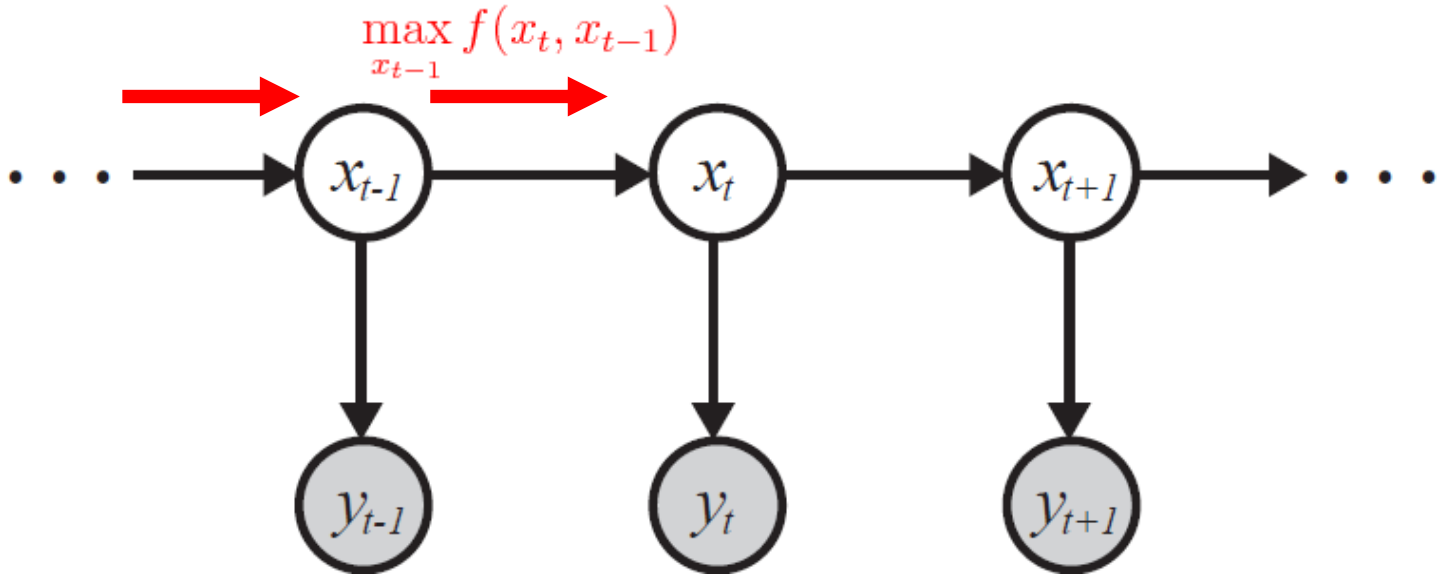
Viterbi Decoder



$$x^* = \underset{x}{\operatorname{argmax}}\, p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

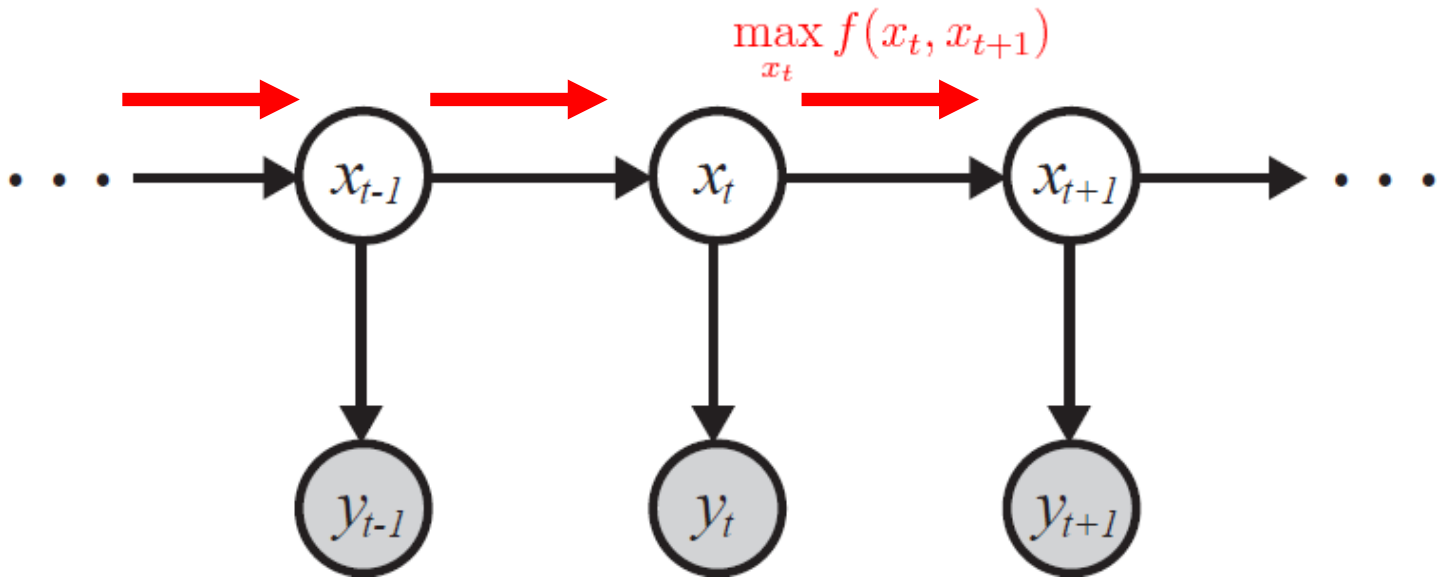
Viterbi Decoder



$$x^* = \operatorname*{argmax}_x p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

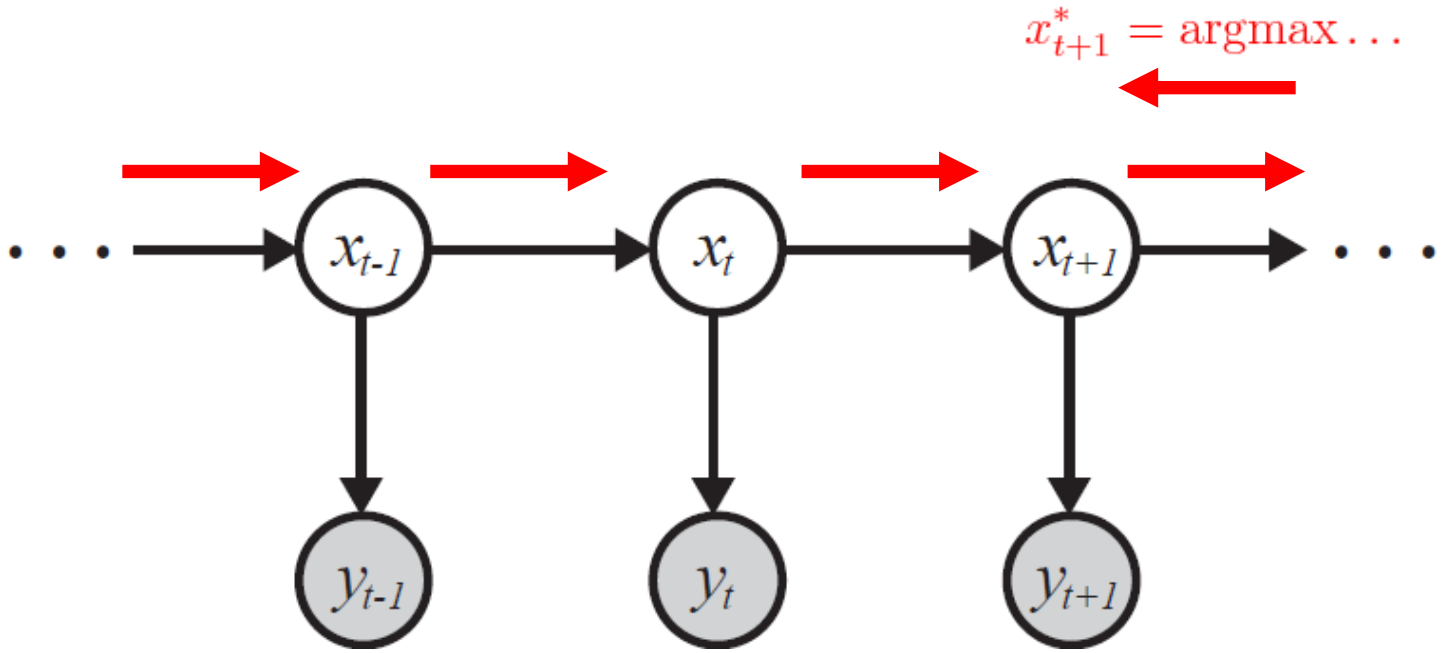
Viterbi Decoder



$$x^* = \operatorname*{argmax}_x p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

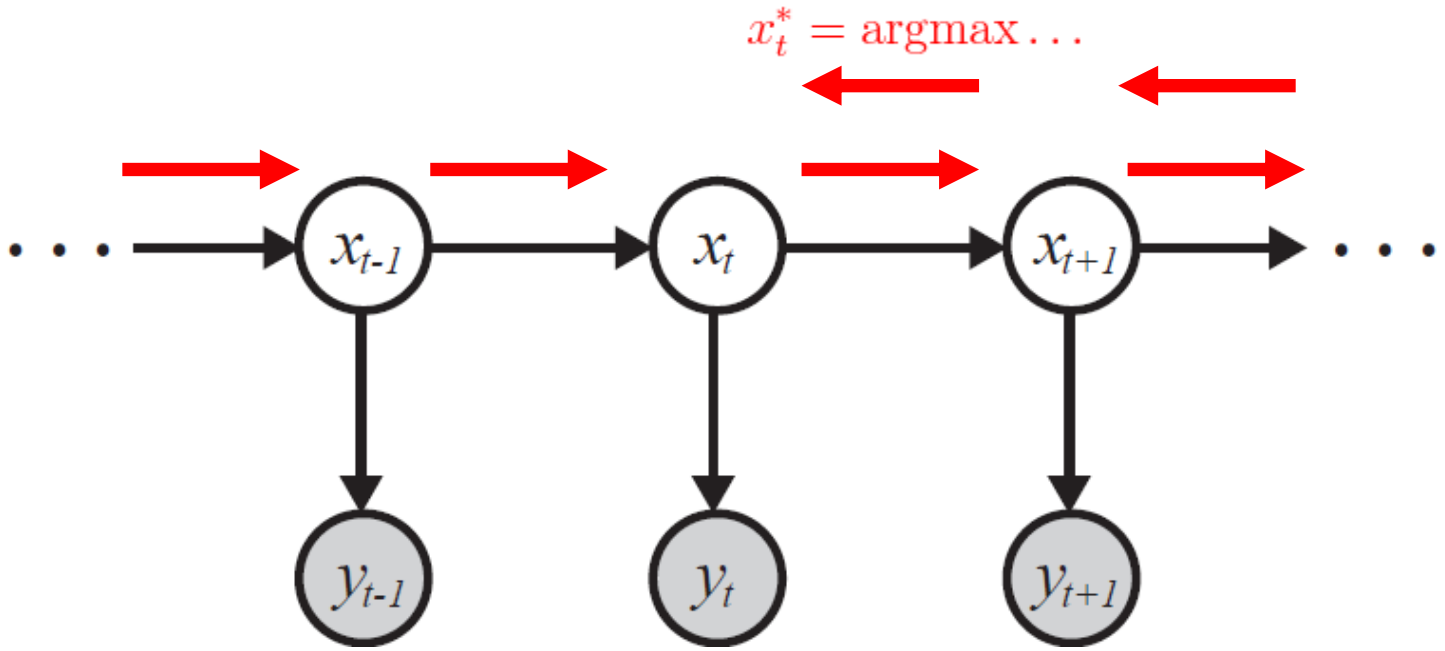
Viterbi Decoder



$$x^* = \underset{x}{\text{argmax}}\, p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

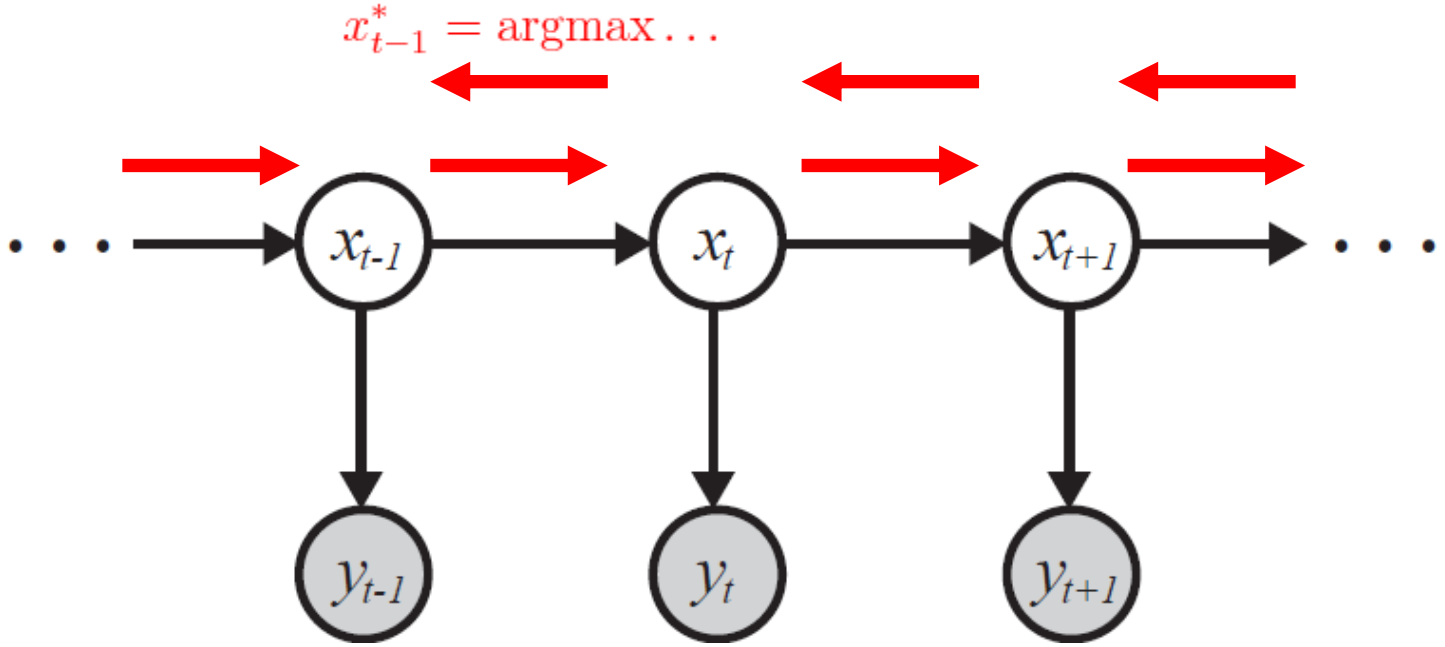
Viterbi Decoder



$$x^* = \underset{x}{\text{argmax}}\, p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

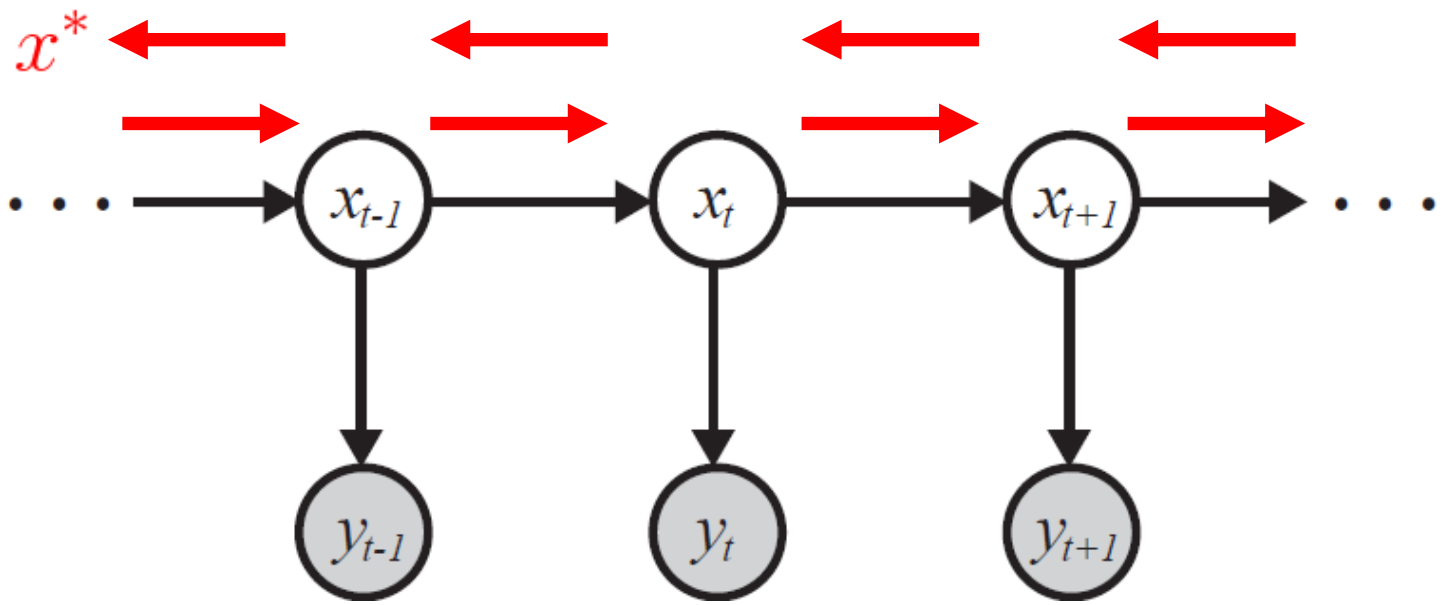
Viterbi Decoder



$$x^* = \underset{x}{\text{argmax}}\, p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

Viterbi Decoder



$$x^* = \underset{x}{\operatorname{argmax}}\, p(x \mid y)$$

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

Course Overview

Course is broken down into **five** primary topics…

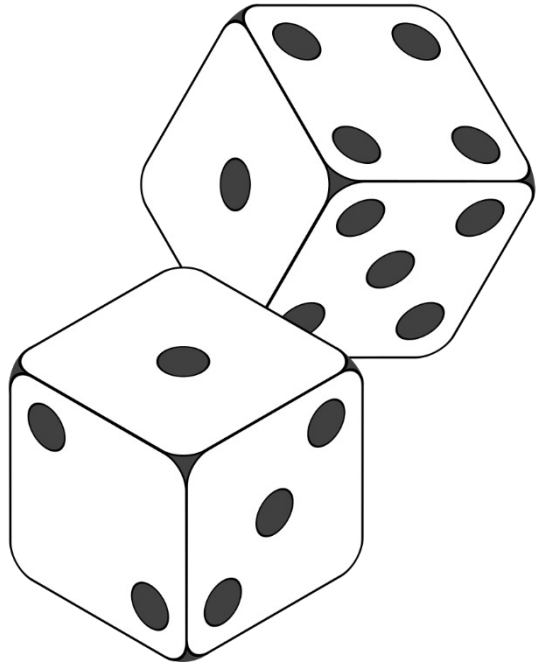
Probability and Statistics	Inference	Bayesian Deep Learning	Representation Learning	Uncertainty Quantification
Probability primer, Bayesian statistics, PGMs, Exponential families	Monte Carlo Methods, Variational Inference, Implicit Inference	Bayesian Neural Networks, Monte Carlo / Variational Dropout, Variational Autoencoder	Information Bottleneck, Information Dropout	Variational Information Dropout, Mutual Information Neural Estimation, Contrastive Predictive Coding

Probability and Statistics

Suppose we roll <u>two fair dice</u>…

- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?

*… this is an **experiment** or **random process**.*

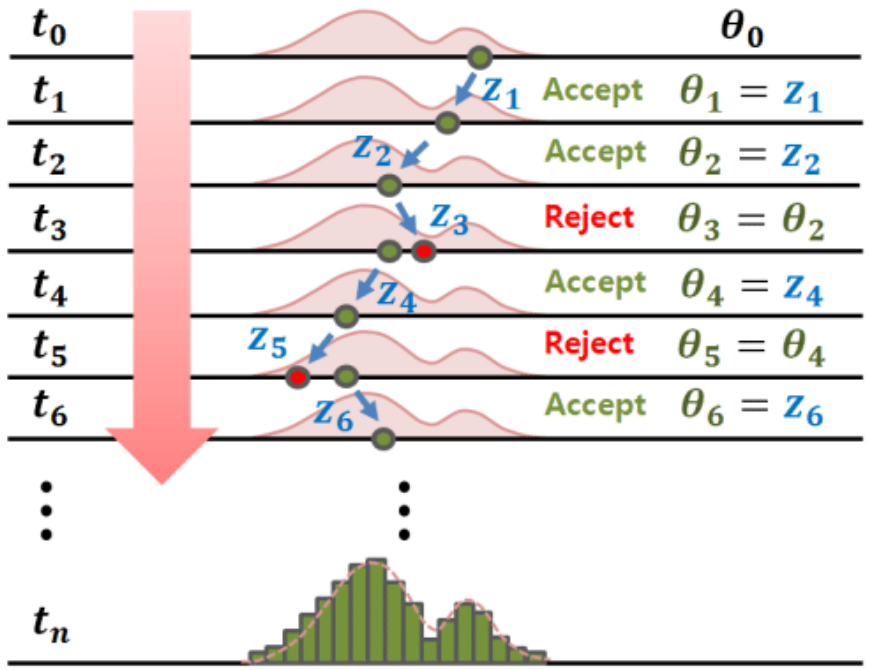


We will learn how to…

- Mathematically formulate outcomes and their probabilities?
- Describe characteristics of random processes
- Estimate unknown quantities (e.g. are the dice actually fair?)
- Characterize the uncertainty in random outcomes
- Identify and measure dependence among random quantities

Monte Carlo Methods

Sample-based methods that simulate realizations from the model to perform inference



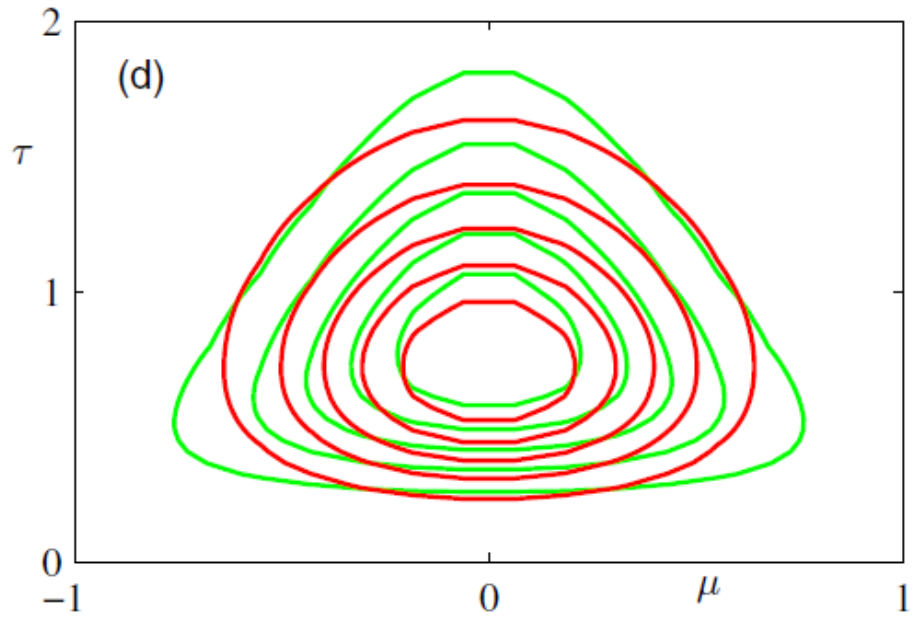
We will learn how to perform sample-based inference using:

- Rejection sampling
- Importance sampling
- Sequential importance sampling (particle filter)
- Markov chain Monte Carlo (MCMC) : Metropolis-Hastings
- MCMC : Gibbs Sampling

[Image: Jin, Seung-Seop et al. (2019)]

Variational Inference

Recasts statistical inference as the solution to an optimization problem



$$min_q \, \mathrm{KL}(q(x) \| p(x))$$

We will learn how to conduct inference via,

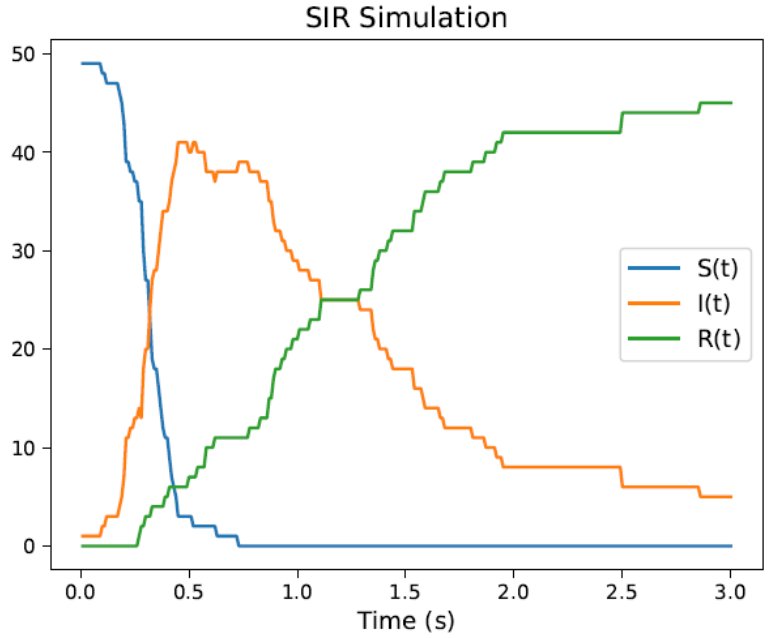
- *Mean field and variational Bayes*
- *Stochastic variational*
- *Inference for implicit likelihood models*

Implicit Models

Some observation models are naturally defined via simulation processes…

SIR Model Example : Epidemiological model of disease among **S**usceptible, **I**nfected, **R**ecovered,

$$
\begin{aligned}
S(t+\Delta_t) &= S(t) - \Delta I(t) \\
I(t+\Delta_t) &= I(t) + \Delta I(t) - \Delta R(t) \\
R(t+\Delta_t) &= R(t) + \Delta R(t)
\end{aligned}
$$



With random additive noise,

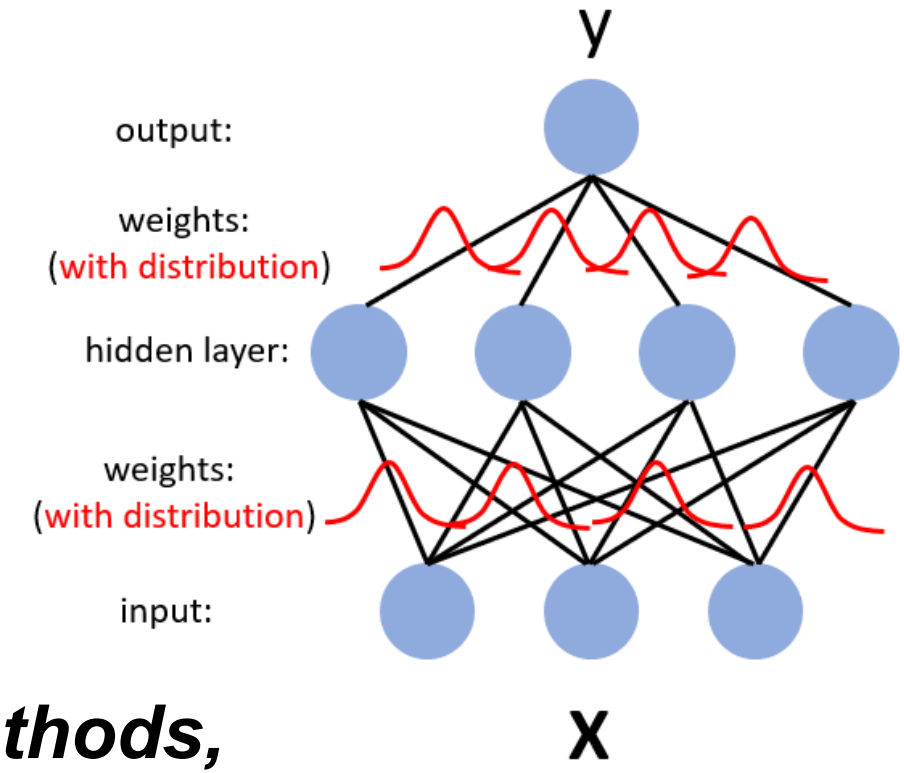
$$
\Delta I(t) \sim \text{Binomial}\left(S(t), \frac{\beta I(t)}{N}\right), \quad \Delta R(t) \sim \text{Binomial}(I(t), \gamma)
$$

For some time interval Δ_t we can simulate S/I/R from initial conditions, but have no closed-form likelihood,

$$
p(S, I, R \mid \beta, \gamma) = ???
$$

Bayesian Deep Learning

Combine probabilistic reasoning with Deep Learning models



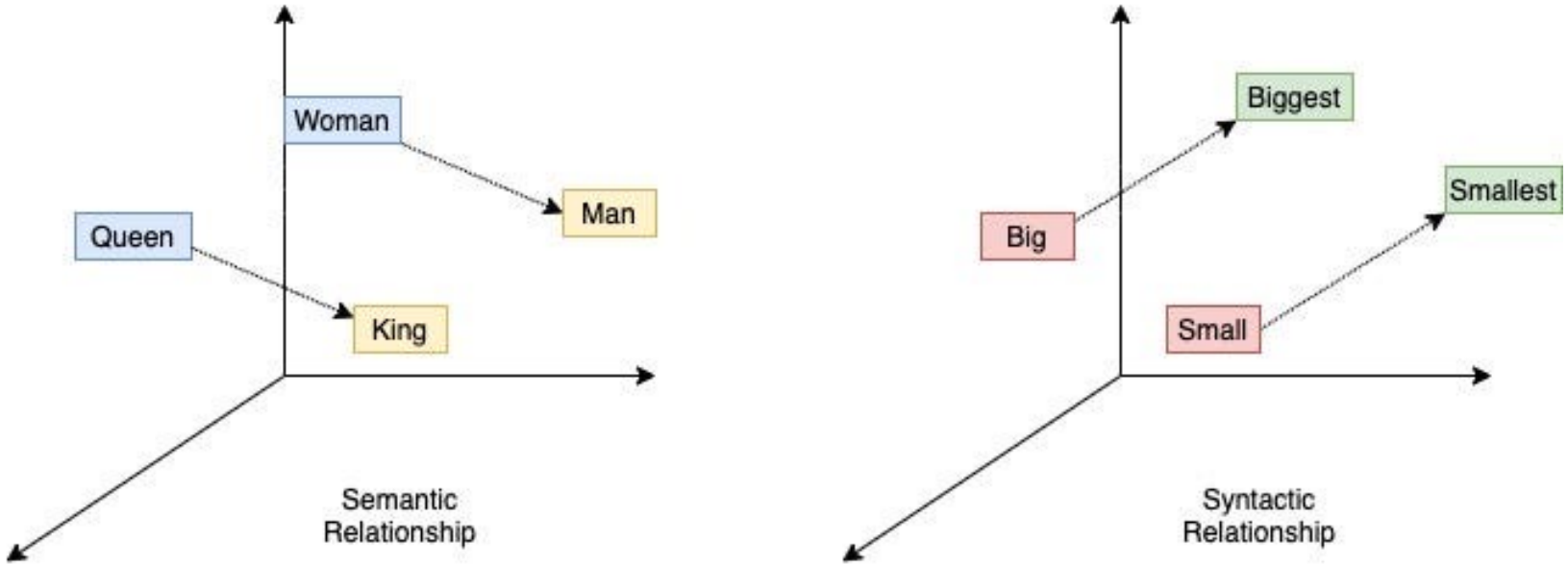
We will learn the following models and methods,

- *Variational autoencoder*
- *Bayesian Neural Network*
- *Structured Variational Autoencoders*
- *Dropout predictions*

[Image: Lin, Xi et al. (2018)]

Representation Learning

Capture underlying structure or patterns in the data



We will learn the fundamentals of representation learning,

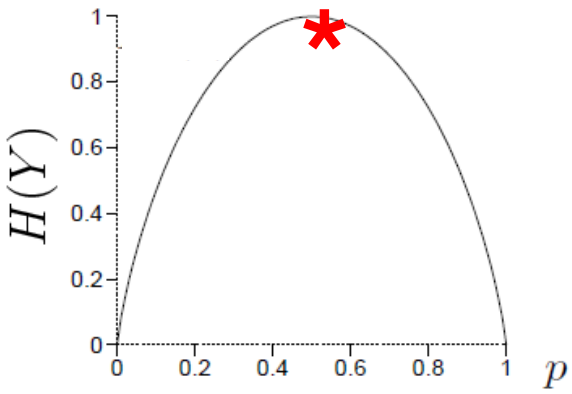
- *Information bottleneck*
- *InfoGAN*
- *Information Dropout*

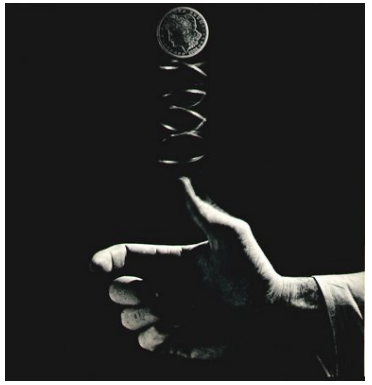
https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30

Uncertainty Quantification

How certain is your model about its predictions? How much information does the data contain about an unknown?

Coin Flip Example: $Y \sim \text{Bernoulli}(p)$





Maximum uncertainty when coin is fair.

We will learn the following methods for uncertainty quantification,

- *Variational information bounds*
- *Mutual information neural estimation*
- *Contrastive predictive coding*

Online Resources

All material (lectures / HWs / readings) are available on the **course webpage**:

http://pachecoj.com/courses/csc696h_spring24/

We will use **D2L** for grades:

https://d2l.arizona.edu/d2l/home/1415589

We will use **Piazza** for discussion:

https://piazza.com/arizona/spring2024/csc696h1

Grading Breakdown

- 20% - Paper presentations
- 10% - Critical reading summaries
- 30% - Term project proposal
- 40% - Term project (presentation and writeup)

Grading Questions

- I will announce in class and/or Piazza when grading of each item is complete
- Officially, you have **1 week** to raise any grading concerns (from the completion of grading)
- If you don't receive a grade, but should have, you must tell me **within 1 week**

Necessary Coding and Math Background

Basic Probability and Statistics
- Marginal / Conditional Probability
- Bayes' Rule

Familiarity with Probabilistic Graphical Models
- Bayes Nets / Markov Random Fields / Factor Graphs
- Conditional Independence

Basic understanding of optimization
- Nonlinear vs. linear programming
- Gradient ascent
- Dynamic programming (we will cover the basics)

Basic Linear Algebra
- Basic vector / Matrix algebra
- Matrix inversion / rank / condition

Basic coding and data structures (for semester project)
- Ideally familiarity with Python / Numpy / Scipy
- Should be able to write / manage code on the order of 1,000 lines

Critical Reading Summaries

- Will be maintained on Github and graded in batches
- A brief paragraph with at least one sentence on each:
  - What is the strength of this paper?
  - What is the primary weakness of this paper?
  - What would you have done differently to improve the paper?
- With each of the above:
  - **Be specific:** Reference figures / equations when possible
  - **Be original:** Consider aspects that are not obvious on a first reading
  - **Be concise:** Make your points in 3 to 6 sentences

Paper Presentations

- Each student will choose N assigned papers (N=1 or 2)
- Choose early:
  - Typically the “easy” papers are selected early
  - I will assign one to you if you don’t choose
- Plan for 45min presentations to leave room for discussion
- Board presentations are acceptable if it is appropriate for the material (e.g. lots of math)
  - Don’t use a board presentation for papers that don’t require them
  - If you’re unsure then check with me ahead of time

Term Project

1) Select a project topic
2) Write a proposal
3) Execute the project
4) Present during final exam period

- Joint projects limited to 2 people (individual projects allowed)
- A component of your graduate research can be used but must be proposed and executed as a standalone sub-problem
- I will be asking that you share Github repos for each project

Late Policy

Please complete reading summaries by the **night before lecture** on that topic

- I will usually grade these at the end of the week, if yours isn't submitted by then then it will not be graded

The **project proposal** and **project report** need to be submitted on time!

- If you are having issues then notify me ahead of time and I will deal with it on a case-by-case basis

Academic Integrity Continued

- All reading summaries **must** be done **independently**
- Project proposals and reports **must** be done **independently**
- You may discuss project details with others, but **do the work yourself**
- **Cite any and all resources** (that includes your own work)
- **Do not** submit your previous work for the term project—it must be novel work (I will look at Github commits)

Good Rule Cite any external resource you use that may be considered plagiarism without citation.

Lectures and Attendance

In-Person Attendance

- I ask that students attend in-person
- I have Zoom meetings scheduled on D2L, but this **should be used sparingly**
- If you might have COVID-19 symptoms then use Zoom
- Attendance and participation will be graded (10% of grade)

Lecture Recordings

- All lectures will be recorded and posted online after-the-fact
- Recordings are accessible via D2L
- Recorded lectures should supplement lecture, **they should not** be used in place of lecture

Office Hours

Use scheduled office hours for

- Specific homework questions
- Clarification on lecture / reading topics
- General course-related questions

Details

- 2 hours per week
- Office hours will be held on Zoom, but schedule is TBD
- Message me on Piazza if you have a conflict with hours and I will try to schedule something for you
- I still need to set a time for hours – ignore the current time in the Syllabus / course webpage

Piazza

- Use Piazza for **all course communication**
- If you email me directly I may not see it (I get a lot of email)
- You can ask / answer questions related to the course
- Also post course-related material (e.g. if you find something on the internet that is interesting / useful to the course)

Mental Well-Being

Some level of stress / depression / anxiety is normal, but sometimes you may need extra help

- Non-emergency UA resources at Counseling & Psych Services Mon-Fri
  - Phone: 520-621-3334
  - Web: https://health.arizona.edu/counseling-psych-services
- Emergency resources in Tucson in this Google Doc

I am happy to point you in the right direction, but keep in mind that I am not a mental health professional

Inclusivity

I want to foster a comfortable and inclusive classroom experience

Please let me know if you feel excluded in any way, e.g.

- “Alice-and-Bob” style examples of material
- Improper use of pronouns
- Microagressions
- Miscellaneous statements / interactions

You can message me anonymously on Piazza

Questions? Comments? Thoughts?