# Probabilistic Graphical Models

## Lecture 1: Introduction, Probability Review

Matthew Walter

TTI-Chicago

April 7, 2020

Some slide content courtesy of David Sontag

How can we gain **global insight** based on **local observations**?

# What are Graphical Models?

In machine learning, we generally consider two things:

**Dataset** A set $\mathcal{D} = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(M)}\}$ of samples from $P^*$, which is unknown

**Model** $\mathcal{M}$ that "best" approximates $P^*$

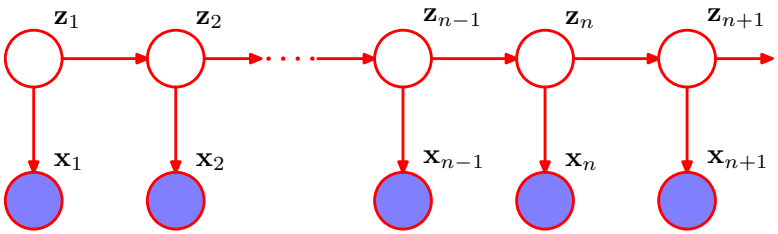
# What are Graphical Models?

In machine learning, we generally consider two things:

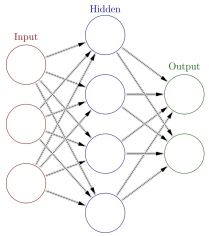
**Dataset** A set $\mathcal{D} = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(M)}\}$ of samples from $P^*$, which is unknown

**Model** $\mathcal{M}$ that "best" approximates $P^*$

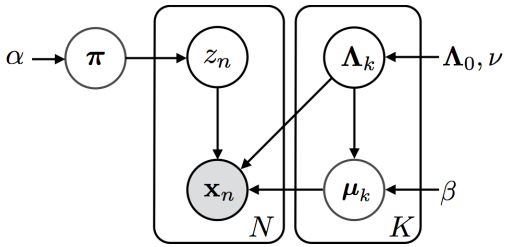
Essentially, graphical models are models that use a graph to represent the interactions between random variables



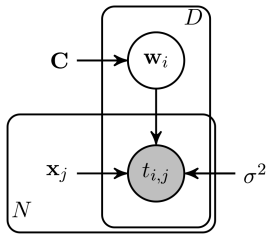

Hidden Markov Model




Neural Network




Mixture of Gaussians




Probabilistic PCA

# Key Ideas

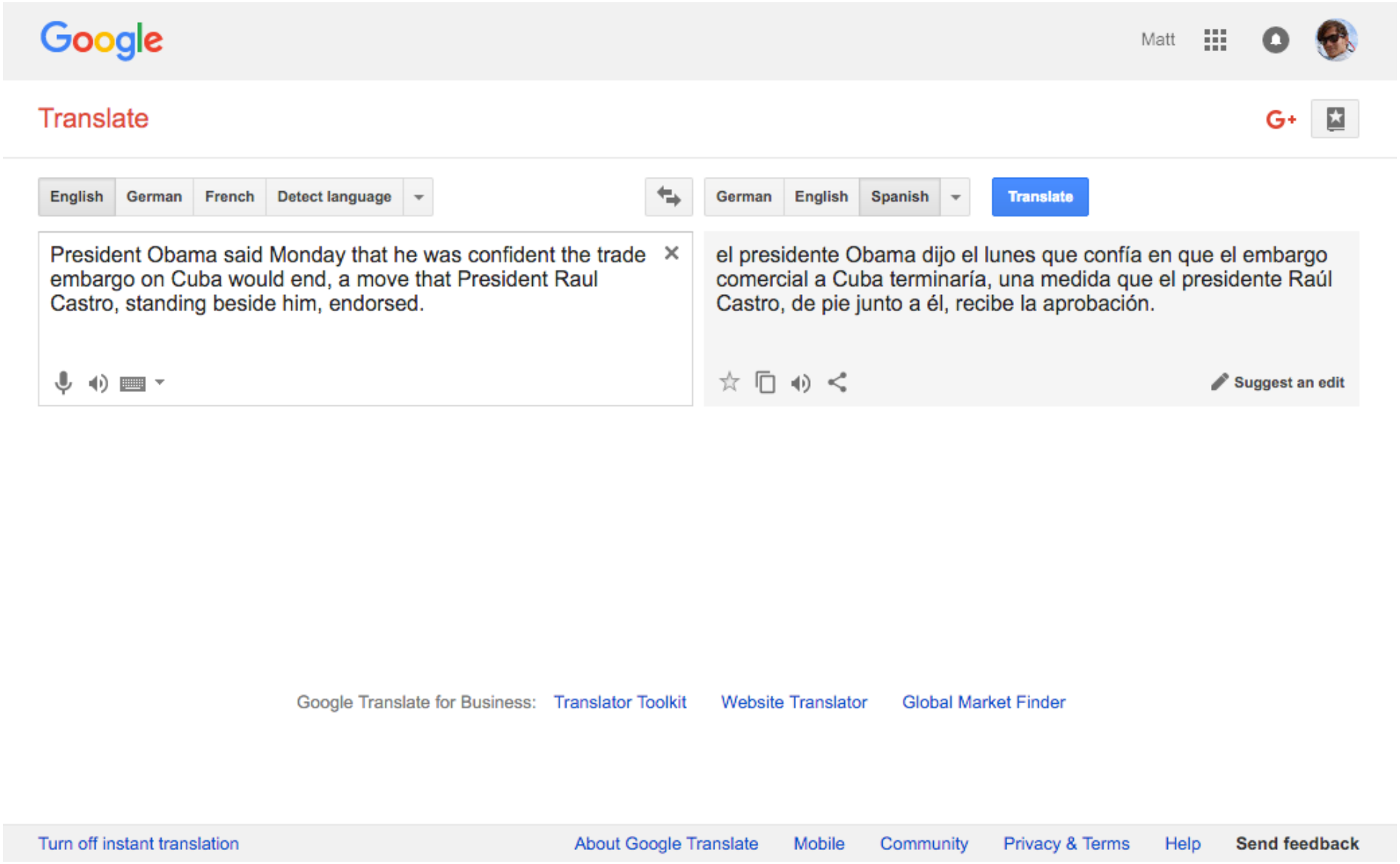
1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $P(X_1, \ldots, X_n)$
2. **Learn** the distribution from available data
3. **Infer** (predict) instances of random variables based upon evidence $P(X_i \,|\, X_1 = x_1, \ldots, X_m = x_m)$

# Reasoning Under Uncertainty

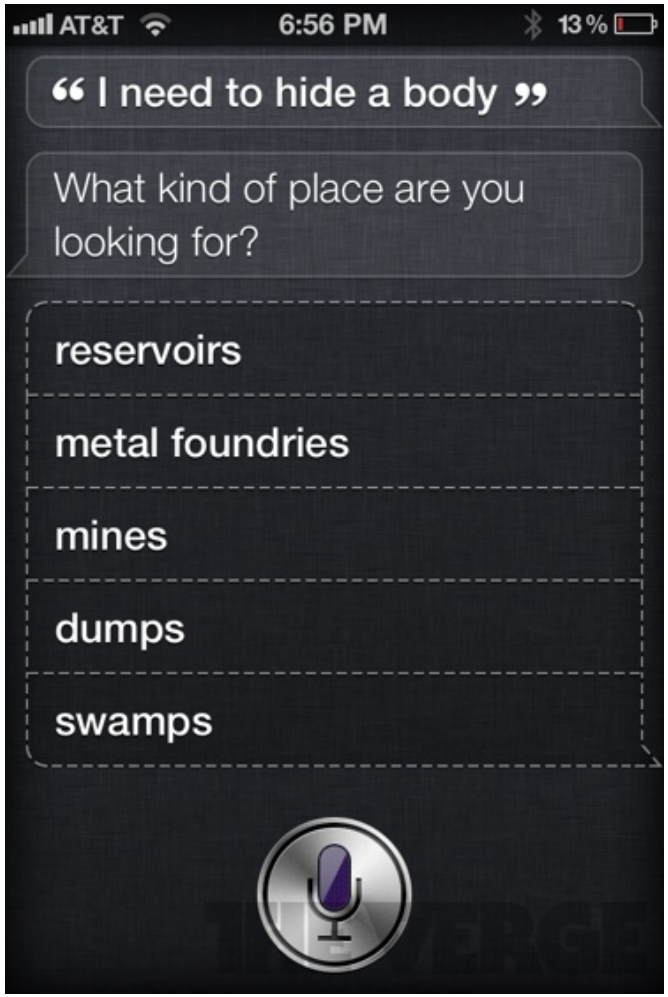
- Humans frequently make predictions in light of uncertainty
- Traditional (rule-based) approaches to AI ignore uncertainty
- The effectiveness of recent advances stems largely from the fact that they take a *probabilistic* approach
- Probabilistic graphical models provide a rich architecture for modeling complex domains with probability distributions
- By combining graph theory and probability theory, graphical models provide a flexible framework for representing intricate interactions among a large number of random variables

# Applications: Question Answering

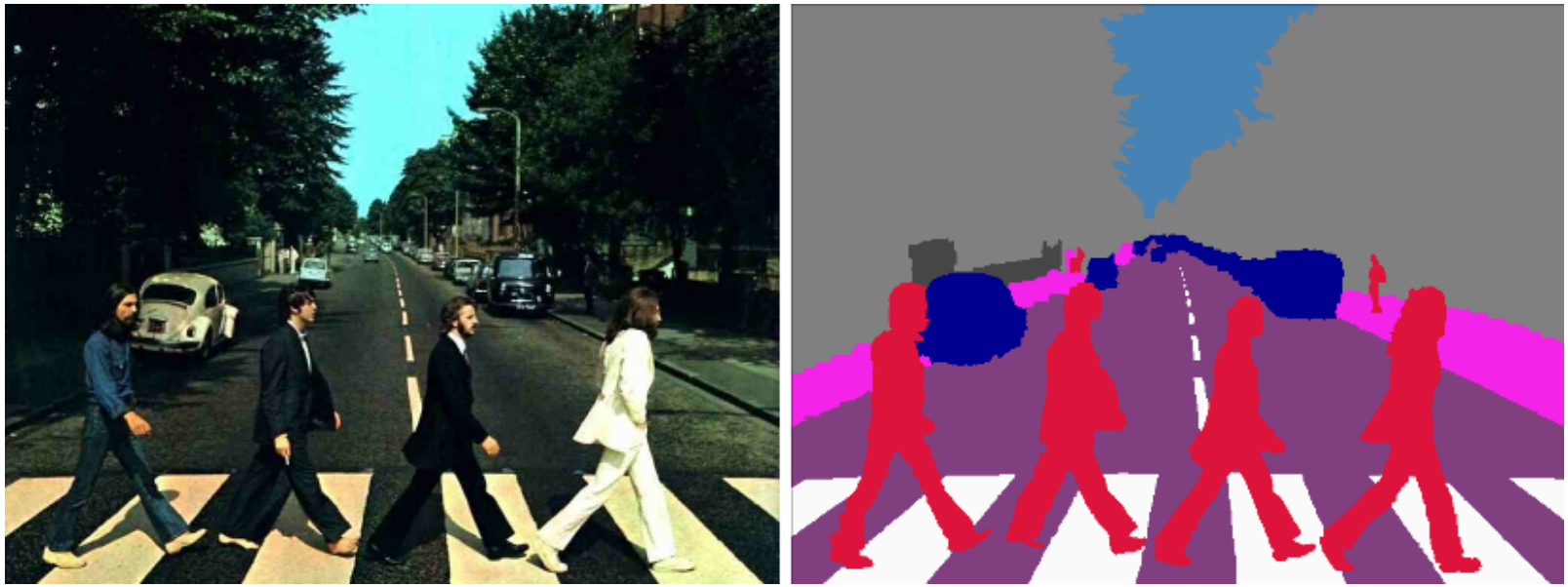
# Applications: Machine Translation



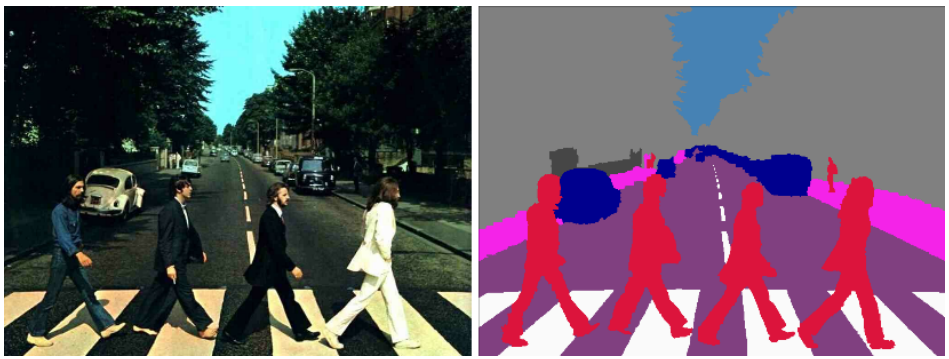
# Applications: Speech Recognition



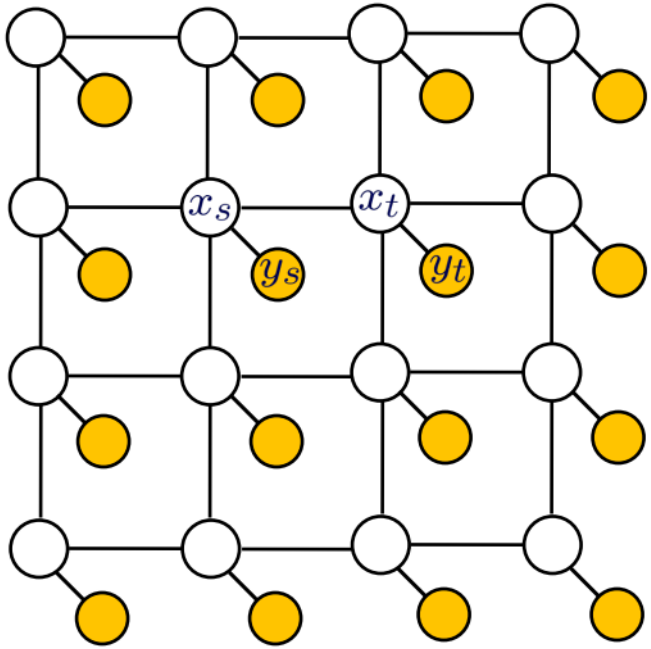
# Applications: Image Segmentation



# Applications: Image Segmentation



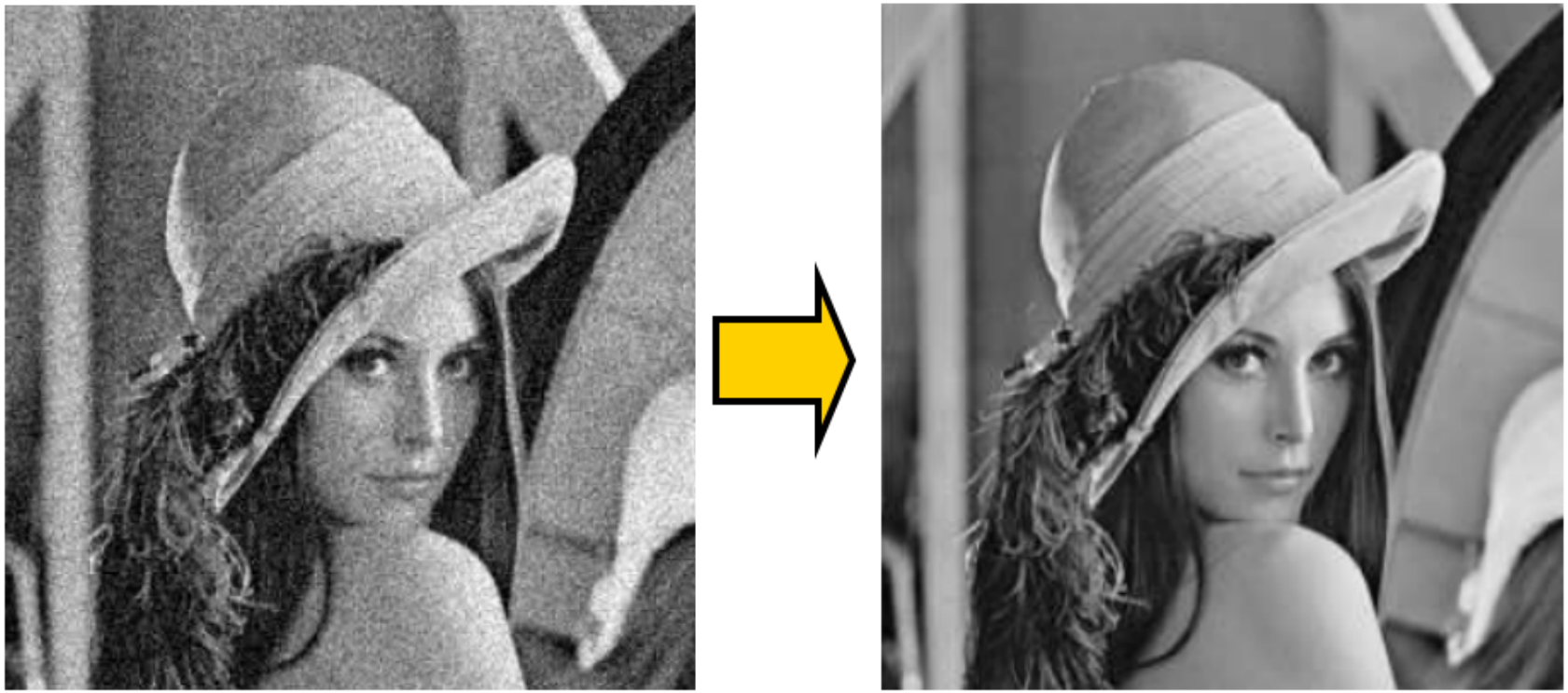
Markov Random Field



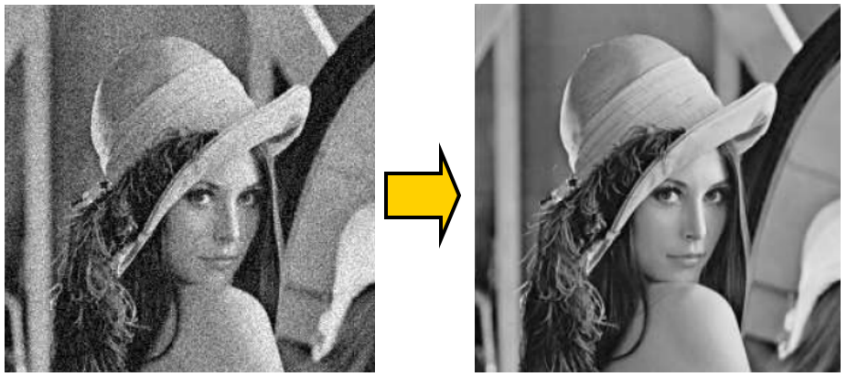

$x_i$: Unobserved (latent) label

$y_i$: Observed (super)pixel

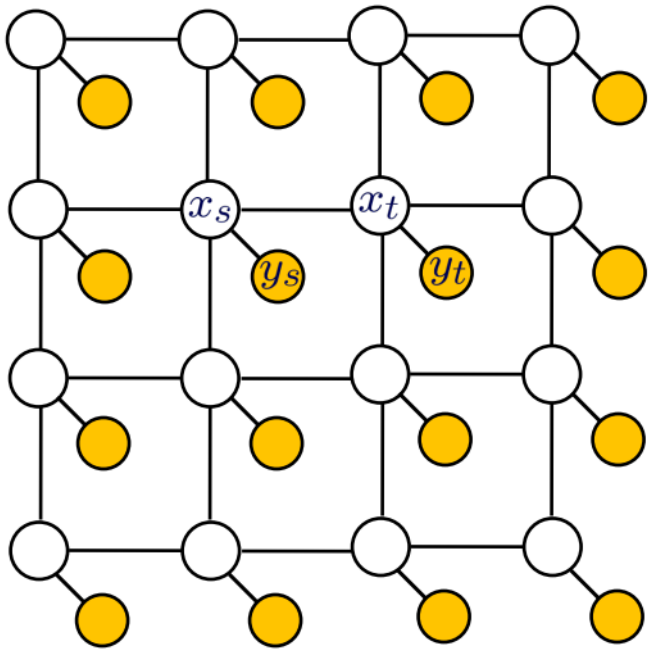
# Applications: Image Denoising



# Applications: Image Denoising



Markov Random Field



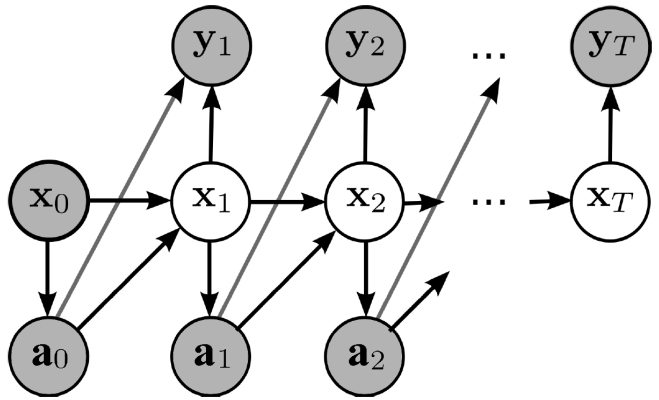

$x_i$: Unobserved "True" intensity

$y_i$: Observed (noisy) intensity

# Applications: Robot Localization

# Applications: Robot Localization


Dynamic Bayesian Network



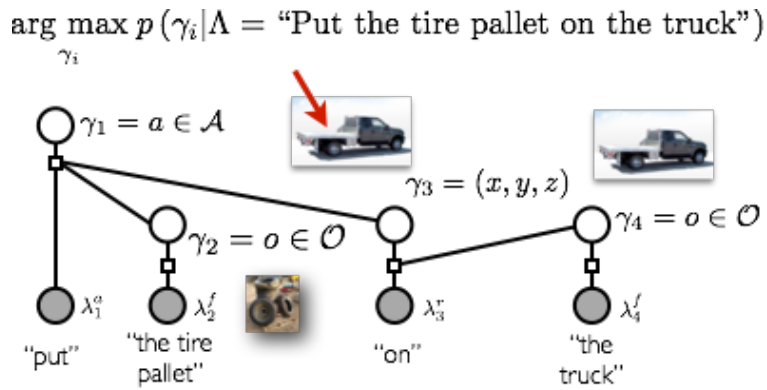

$x_i$: Robot's true pose

$a_i$: actions

$y_i$: Measurements

# Applications: Natural Language Understanding (Robotics)

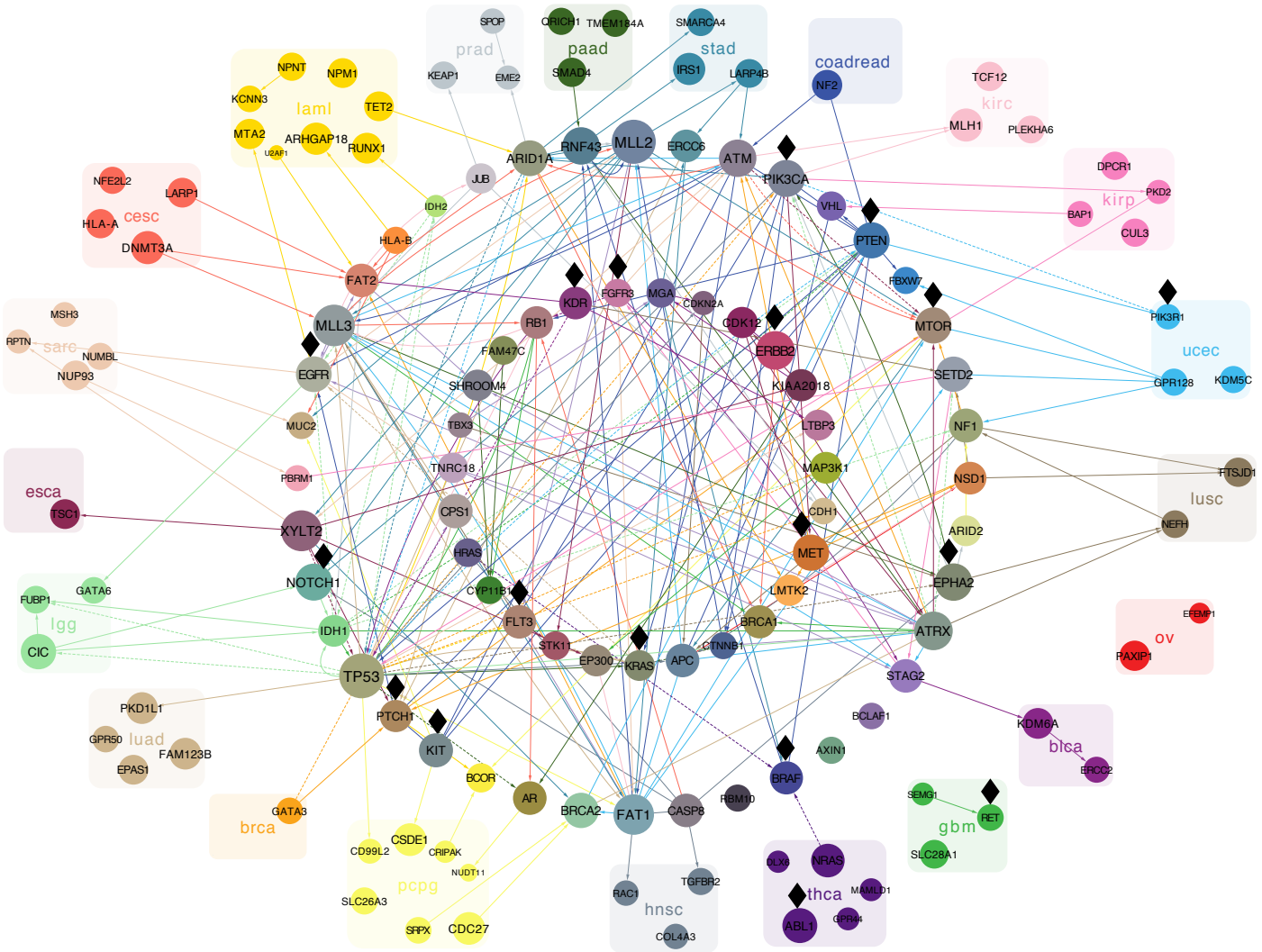

Factor Graph




$\gamma_i$: Groundings (objects, actions)

$\lambda_i$: Phrases in instruction

# Applications: Biology




Bayesian network model of interactions between gene mutations for different types of cancer (courtesy: J. Kuipers et al., 2017)

# Applications of Graphical Models

- Speech recognition (Hidden Markov Models)
- Computer vision (Markov Random Fields)
- Planning under uncertainty (dynamic Bayesian Networks)
- Natural language processing (probabilistic grammars)
- Robot localization (dynamic Bayesian Networks, Kalman Filters)
- ...

# Key Challenges

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $P(X_1, \ldots, X_n)$
   - How does one *compactly* describe this joint distribution?
   - Directed graphical models (Bayesian networks)
   - Undirected graphical models (Markov random fields, factor graphs)
2. Perform **inference** to predict instances of random variables based upon evidence $P(X_i \,|\, X_1 = x_1, \ldots, X_m = x_m)$
   - Exact vs. approximate
   - Variational
   - Sampling
3. **Learn** the distribution from available data
   - What do we maximize? Maximum likelihood?
   - How much data do we need?
   - How much computation does it take?

# Course Overview

- We will study Representation, Inference, & Learning
- Begin with simplest case
  - Only discrete variables
  - Fully observed models
  - Exact inference & learning
- Then generalize
  - Continuous variables
  - Learning with partially observed (hidden) variables
  - Approximate inference & learning
- Learn about theory, algorithms, & applications

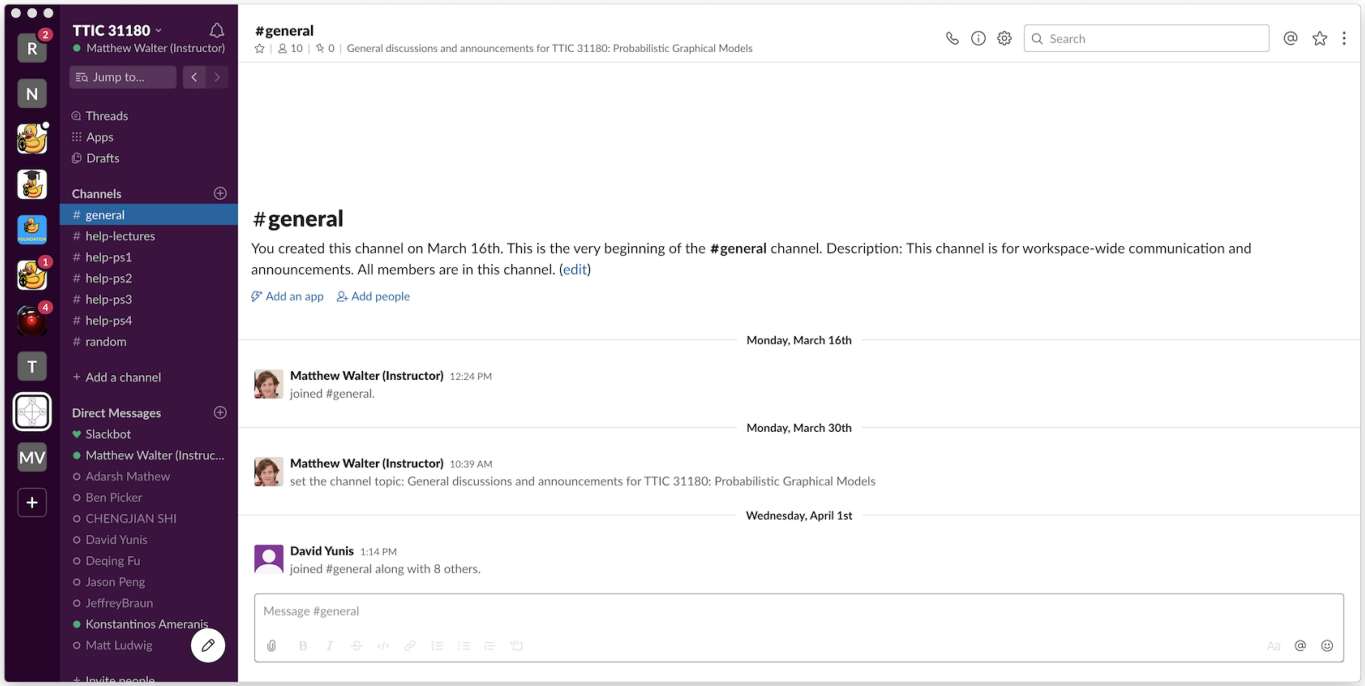
# Course Logistics

- **Instructor**: Matthew Walter
  - TTIC Room 429
  - `mwalter@ttic.edu`
- **Lectures**: Tuesdays & Thursdays, 09:30am–10:50am
  - The current plan is to live-stream lectures using Zoom
  - Lectures will be recorded and made available for offline viewing.
- **Office Hours**: TBD (via Zoom)
- **Webpage**: `http://canvas.uchicago.edu`

# Course Logistics: Zoom

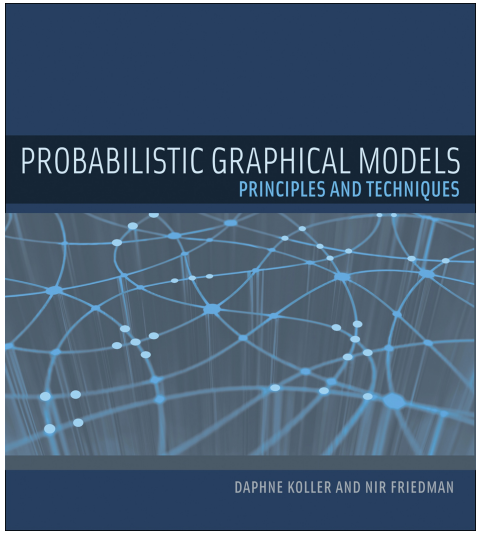
- We will use Zoom for lectures/discussions and office hours
  - **Meeting ID**: 558-526-524
  - **Password**: 31180
  - `https://uchicago.zoom.us/j/558526524`
  - See Canvas for more information on joining the meetings
- Zoom accounts are available via UChicago (see Canvas)
- I strongly encourage you to regularly interact with one another via Zoom (or similar)
- See `https://its.uchicago.edu/secure-zoom-meetings/` for recommendations on securing your Zoom meetings

# Course Logistics: Slack




- We will use Slack to as a regular discussion forum
- `http://ttic31180.slack.com/`
- See Canvas for information on how to get an account
- Dedicated channels to discuss lectures, problem sets, . . .
- I strongly encourage you to regularly interact with one another

# Course Logistics: Required Reading




Daphne Koller & Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press (2009)

Additional readings will be made available via Canvas

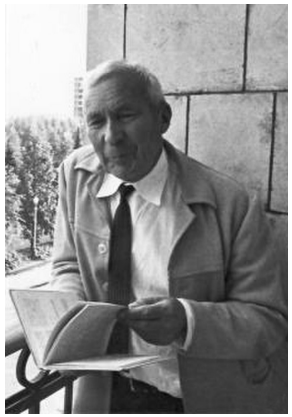
# Course Logistics: Grading

- **Required Reading**:
  Daphne Koller & Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press (2009)
- **Grading**:
  - Problem Sets: $50\%$
  - Midterm Exam: $25\%$
  - Final Exam (or final project): $25\%$
- **Problem Sets**:
  - Written (or typeset) and programming (Python) assignments
  - Submitted via Canvas
  - Late policy: $10\%$ off per day, no credit beyond $3$ days. Each student has a budget of $6$ late days
  - You are encouraged to discuss problems with one another, but please acknowledge your collaborator(s)
  - See syllabus for further policy details

# Probability Review: Probability Space

**Probability space** $(\Omega, \mathcal{S}, P)$ is a measure space that models a real-world process ("experiment") consisting of randomly occurring states

- $\Omega$: The **sample space** is the non-empty set of possible outcomes
- $\mathcal{S} \subseteq \Omega$: The **event space**, where each event is a measurable set containing zero or more outcomes. $\mathcal{S}$ is a $\sigma$-algebra (i.e., closed under countable unions and complement).
- $P : \mathcal{S} \rightarrow [0, \infty]$: A **probability measure** assigns probabilities to events

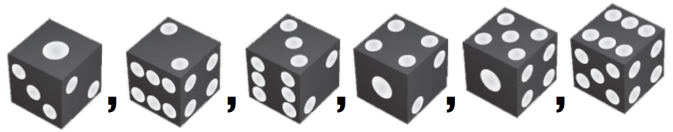


Andrey Kolmogorov (1903–1987)

# Probability Review: Probability Space

- The **sample space** (outcome space) $\Omega$ specifies the set of possible outcomes

$\Omega = \{$  ,  $\}$ Coin toss

$\Omega = \{$  $\}$ Die toss

# Probability Review: Probability Space

- The **sample space** (outcome space) $\Omega$ specifies the set of possible outcomes

$\Omega$ = {  ,  } Coin toss

$\Omega$ = {  } Die toss

- The **event space** $\mathcal{S} \subset \Omega$ is a subset of the outcome space ($\mathcal{S} \subseteq 2^{\Omega}$)

**E** = {  } Even die tosses

**O** = {  } Odd die tosses

# Probability Review: Probability Space

- The **sample space** (outcome space) $\Omega$ specifies the set of possible outcomes

$\Omega$ = {  ,  } Coin toss

$\Omega$ = {  } Die toss

- The **event space** $\mathcal{S} \subset \Omega$ is a subset of the outcome space ($\mathcal{S} \subseteq 2^{\Omega}$)

**E** = {  } Even die tosses

**O** = {  } Odd die tosses

- The **probability measure** is a function returning an event's probability

# Probability Review: Probability Distribution

The **probability** of an event is defined by the sum of the probabilities associated with the elements of the set

$$P(S) = \sum_{\omega \in S} P(\omega)$$

E.g., p(E) = p(  ) + p(  ) + p(  )

= 1/2, if fair die

# Probability Review: Axioms of Probability Theory

The Kolmogorov Axioms:

1. **Non-negativity**: $P(A) \geq 0 \;\; \forall A \in S$

2. **Normalization**: $P(\Omega) = 1$

3. **Countable Additivity**: For any *countable* sequence of *disjoint* sets $A_1, A_2, \ldots$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

# Exercise: Properties of Probability Measures

Let $(\Omega, \mathcal{S}, P)$ be a probability space and $A, B, A_i$ be events in $\mathcal{S}$

- Monotonicity: If $A \subseteq B$, then $P(A) \leq P(B)$
- Complement rule: $P(A^c) = P(\Omega \setminus A) = 1 - P(A)$
- Numeric bound: $0 \leq P(A) \leq 1$
- Addition law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

# Exercise: Monotonicity

If $A \subseteq B$, then $P(A) \leq P(B)$

**Proof.**

Let $A_1 = A$, $A_2 = B \setminus A$, and $A_i = \emptyset \; \forall i \in \{3, 4, \ldots, \infty\}$

$$P(B) = \sum_{i=1}^{\infty} P(A_i) \quad \text{(per Axiom 3)}$$

$$= P(A) + P(B \setminus A) + \sum_{i=3}^{\infty} P(A_i)$$

$$\therefore P(B) \geq P(A)$$

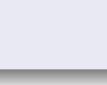
# Exercise: Complement rule

$$P(A^c) = P(\Omega \setminus A) = 1 - P(A)$$

## Proof.

Let $A_1 = A$, $A_2 = \Omega \setminus A$, and $A_i = \emptyset \; \forall i \in \{3, 4, \ldots, \infty\}$

$$\begin{aligned} P(\Omega) &= \sum_{i=1}^{\infty} P(A_i) \quad \text{(per Axiom 3)} \\ &= P(A) + P(\Omega \setminus A) + \sum_{i=3}^{\infty} P(A_i) \\ &= P(A) + P(A^c) \\ &= 1 \quad \text{per Axiom 2} \\ \therefore P(A^c) &= 1 - P(A) \end{aligned}$$



# Exercise: Numeric bound

$$0 \leq P(A) \leq 1$$

**Proof.**

$$\begin{aligned} P(A) + P(A^c) &= 1 \quad \text{(per complement rule)} \\ P(A) &= 1 - P(A^c) \\ &\leq 1 \quad \text{(per Axiom 1)} \end{aligned}$$

$$\therefore 0 \leq P(A) \leq 1$$

# Exercise: Addition law of probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

## Proof.

Let $A_1 = A \setminus (A \cap B)$, $A_2 = A \cap B$, $A_3 = B \setminus (A \cap B)$, and $A_i = \emptyset$ $\forall i \in \{4, 5, \ldots, \infty\}$

$$\begin{aligned} P(A \cup B) &= \sum_{i=1}^{\infty} P(A_i) \quad \text{(per Axiom 3)} \\ &= P(A \setminus (A \cap B)) + P(A \cap B) + P(B \setminus (A \cap B)) \\ &= \big(P(A \setminus (A \cap B)) + P(A \cap B)\big) + \\ &\qquad \big(P(B \setminus (A \cap B)) + P(A \cap B)\big) - P(A \cap B) \\ \therefore P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

# Chain Rule

Let $A_1, \ldots, A_n \in \mathcal{S}$ be events such that $P(A_i) > 0$

The joint distribution **factorizes** as

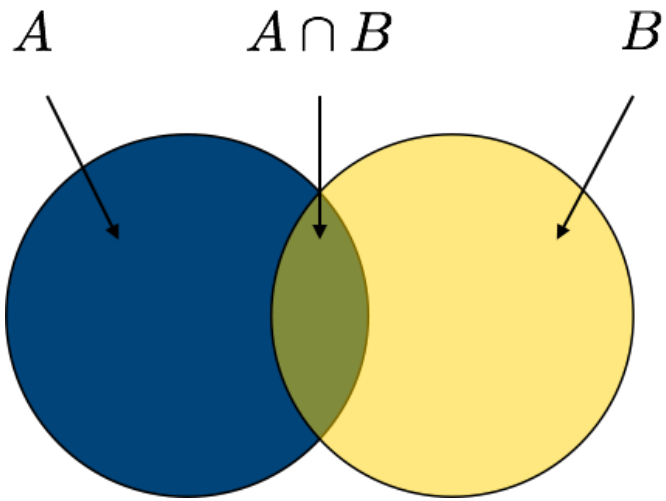
$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_{n-1} \ldots A_1)$$

# Bayes' Rule

## Definition (Bayes' Rule)

Let $A, B \in \mathcal{S}$ be events such that $P(A), P(B) > 0$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

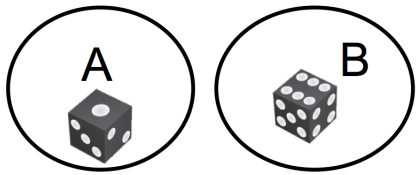


# Independence of Events

- Two events $A$ and $B$ are **independent** if

$$P(A \cap B) = P(A)P(B)$$

- Are the following events independent?



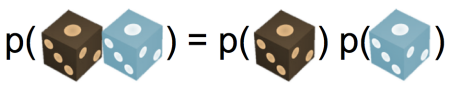

**No!** $p(A \cap B) = 0, \quad p(A)p(B) = \left(\frac{1}{6}\right)^2$

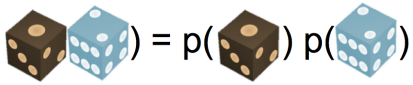
- Suppose that the outcome space consisted of two different die

$\Omega$ = {  } 2 die tosses

$6^2 = 36$ outcomes

and the probability distribution is such that each die is independent



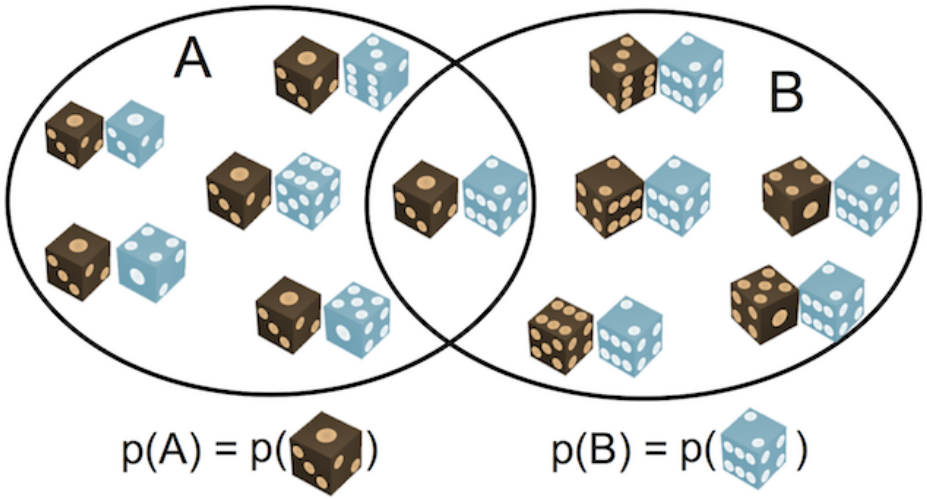



# Independence of Events

- Two events $A$ and $B$ are **independent** if

$$P(A \cap B) = P(A)P(B)$$

- Suppose $A = \{\text{black die is } 1\}$ & $B = \{\text{blue die is } 2\}$. Are $A$ and $B$ independent?

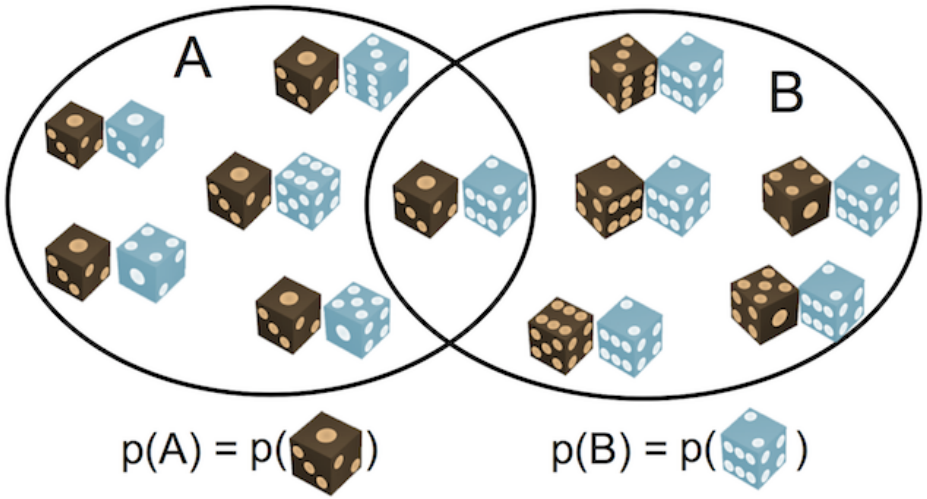


# Independence of Events

- Two events $A$ and $B$ are **independent** if

$$P(A \cap B) = P(A)P(B)$$

- Suppose $A = \{\text{black die is } 1\}$ & $B = \{\text{blue die is } 2\}$. Are $A$ and $B$ independent?




**Yes!**

$$p(A \cap B) = \text{p}( \quad )$$

$$p(A)p(B) = \text{p}( \quad ) \; \text{p}( \quad )$$

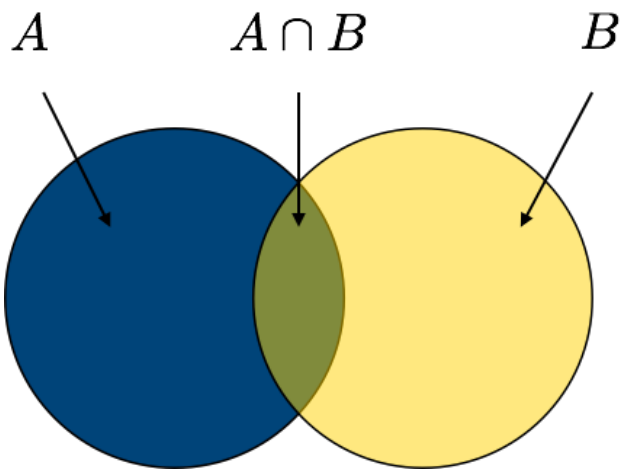
# Conditional Probability

- Let $A, B$ be events such that $P(B) > 0$
- Suppose we know that $B$ occurred. What does that tell us about $A$?

# Conditional Probability

- Let $A, B$ be events such that $P(B) > 0$
- Suppose we know that $B$ occurred. What does that tell us about $A$?
- $P(A \mid B) =$ probability of $A$ given that $B$ occurred ($B$ is the new universe)

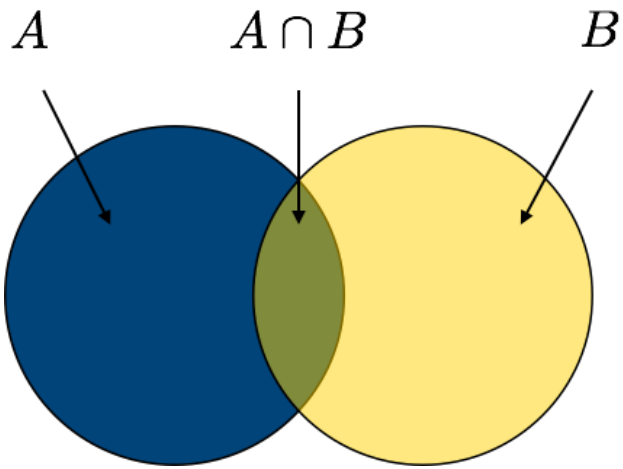
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



# Conditional Probability

- Let $A, B$ be events such that $P(B) > 0$
- $P(A \mid B) =$ probability of $A$ given that $B$ occurred ($B$ is the new universe)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$




- Conditional probability satisfies axioms of probability
- If $A$ and $B$ are independent, $P \vDash (A \perp B) \rightarrow P(A \mid B) = P(A)$

# Conditional Independence

- Let $A, B, C$ be events
- $A$ and $B$ are **conditionally independent** given $C$

$$P \models (A \perp B \,|\, C)$$

- **Proposition**: $P \models (A \perp B \,|\, C)$ iff

$$\begin{aligned} P(A, B \,|\, C) &= P(A \,|\, B, C) P(B \,|\, C) \\ &= P(A \,|\, C) P(B \,|\, C) \end{aligned}$$

- If $P(A \perp B \,|\, \emptyset)$, then $A$ and $B$ are **marginally independent**

# Random Variables

- Events can be difficult to work with
  - We often group events by **attributes** (e.g., Person $\rightarrow$ "age," "height," and "weight"), particularly numerical values
- A **Random variable** $X : \Omega \rightarrow D$ is a measurable function mapping an outcome to a measurable space $D$
  - $D$ may be discrete or continuous (e.g., $\mathbb{R}$)
  - Induces a partition on $\Omega$ of all outcomes: $\{\omega \in \Omega : X(\omega) = x\}$
- For a given $x \in D$ we define the "probability that $X$ instantiates as $x$"

$$P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

- $\text{Val}(X)$ is the set of values that $X$ can take
- $\sum_{x \in \text{Val}(X)} P(X = x) = 1$

# Random Variables: Examples

- Bernoulli distribution (i.e., "biased coin flips")
  - $D = \{\text{Heads}, \text{Tails}\}$
  - Specify $P(X = \text{Heads}) = p$, and thus, $P(X = \text{Tails}) = 1 - p$
  - $X \sim \text{Ber}(p)$

- Multinomial distribution (i.e., "biased m-sided dice")
  - $D = \{1, 2, \ldots, m\}$
  - Specify $P(X = i) = p_i$ s.t. $\sum_i p_i = 1$
  - $X \sim \text{Mult}(p_1, \ldots, p_m)$

# Random Variables: Example (Discrete)

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \text{Val}(X_i) \in \{0, 1\}$$

- Let $\Omega = \text{Val}(X_1) \times \text{Val}(X_2) \times \text{Val}(X_3)$

$$\Omega = \{(0,0,0), (1,0,0), \ldots, (1,1,1)\}$$

- $X_i(\omega)$ is the value of $X_i$ in the assignment $\omega \in \Omega$
- Specify $P(w)$ for each outcome $\omega \in \Omega$ via a look-up table

| $x_1$ | $x_2$ | $x_3$ | $p(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | .11 |
| 0 | 0 | 1 | .02 |
| | ⋮ | | |
| 1 | 1 | 1 | .05 |

- How many parameters do we need to define this model?

# Random Variables: Example (Discrete)

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \text{Val}(X_i) \in \{0, 1\}$$

- Let $\Omega = \text{Val}(X_1) \times \text{Val}(X_2) \times \text{Val}(X_3)$

$$\Omega = \{(0,0,0), (1,0,0), \ldots, (1,1,1)\}$$

- $X_i(\omega)$ is the value of $X_i$ in the assignment $\omega \in \Omega$
- Specify $P(w)$ for each outcome $\omega \in \Omega$ via a look-up table

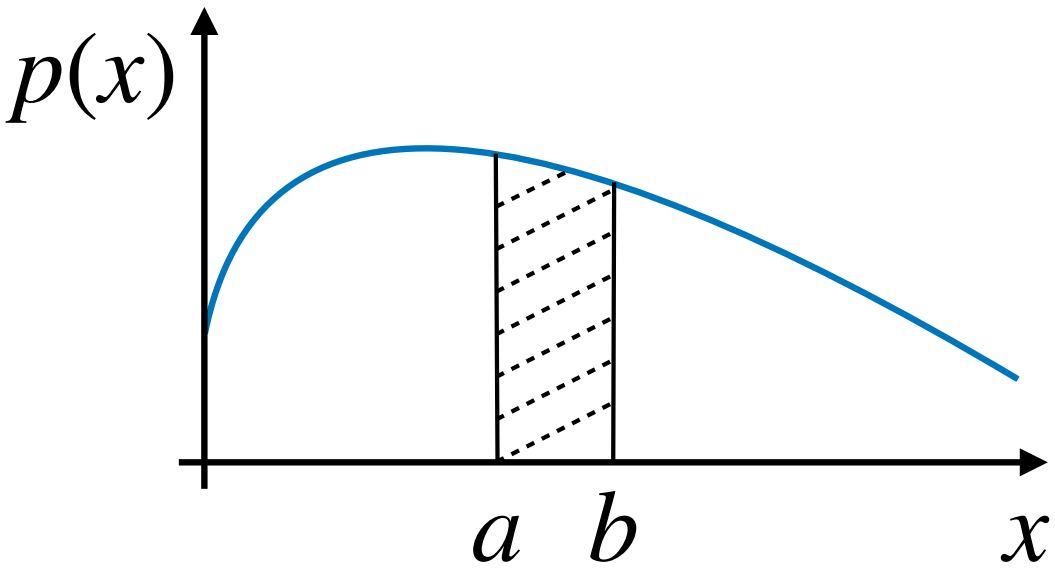
| $x_1$ | $x_2$ | $x_3$ | $p(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | .11 |
| 0 | 0 | 1 | .02 |
| | ⋮ | | |
| 1 | 1 | 1 | .05 |

- How many parameters do we need to define this model? $2^3 - 1$

# Continuous Random Variables

- $X$ takes on values in the continuum
- $p(X = x) = p(x)$ (lower-case $p$) is the *probability density function*

$$P(x \in (a, b)) = \int_a^b p(x)dx$$



# Multivariate Random Variables

- We have a random *vector* rather than a random variable

$$\mathbf{X}(\omega) = \begin{bmatrix} X_1(\omega) & X_2(\omega) & \dots & X_n(\omega) \end{bmatrix}$$

- $X_i = x_i$ is an event and the joint distribution is defined as

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1 \cap \dots \cap X_n = x_n)$$

- It is common to use shorthand

$$P(X_1 = x_1, \dots, X_n = x_n) = P(x_1, \dots, x_n)$$

# Joint Probability

- Model joint probability over three binary r.v. via look-up table

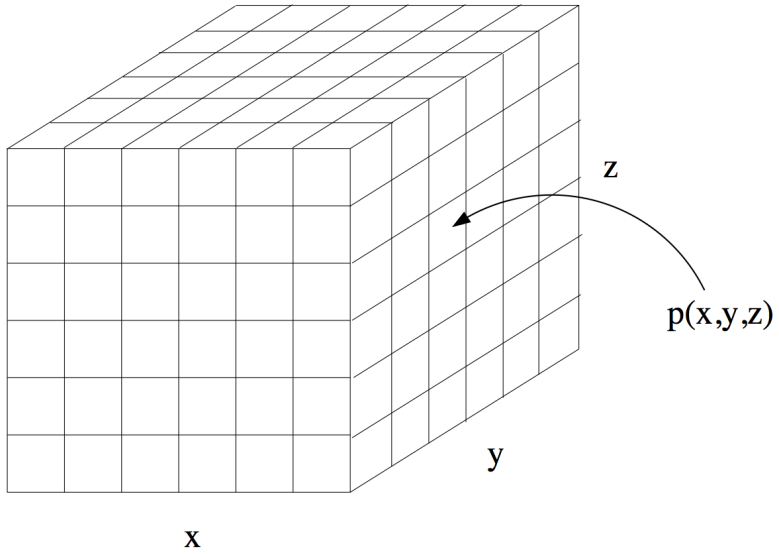
| $x_1$ | $x_2$ | $x_3$ | $p(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | .11 |
| 0 | 0 | 1 | .02 |
| | ⋮ | | |
| 1 | 1 | 1 | .05 |

# Joint Probability

- Model joint probability over three binary r.v. via look-up table

| $x_1$ | $x_2$ | $x_3$ | $p(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | .11 |
| 0 | 0 | 1 | .02 |
| | ⋮ | | |
| 1 | 1 | 1 | .05 |

- Alternatively, we can view this as a $2 \times 2 \times 2$ matrix



# Rules for Random Variables

- Chain Rule

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2 \,|\, x_1) \ldots P(x_n \,|\, x_1, \ldots, x_{n-1})$$

- Bayes' Rule

$$P(x \,|\, y) = \frac{P(y \,|\, x)P(x)}{P(x)}$$

# Marginalization

- Suppose that $X$ & $Y$ are random variables w/ distribution $P(X, Y)$
  $X$ : Intelligence, $\mathrm{Val}(X) = \{\text{“Very High,” “High”}\}$
  $Y$ : Grade, $\mathrm{Val}(Y) = \{\text{“A,” “B”}\}$
- Joint distribution specified by

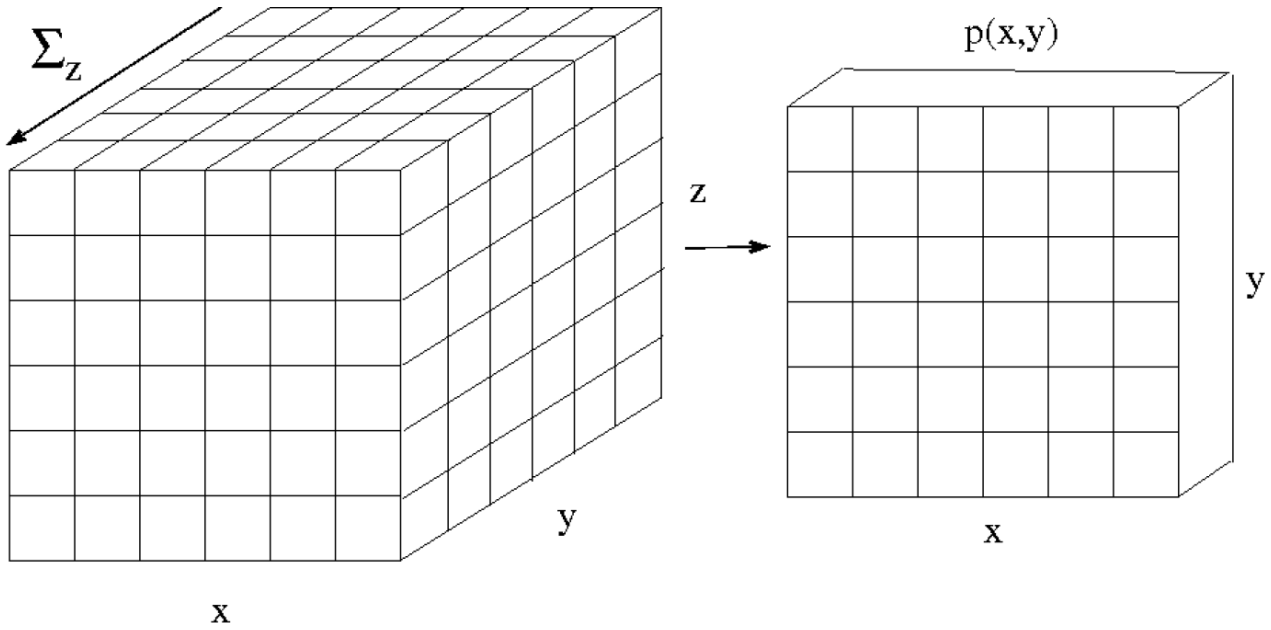
| | | X | |
|---|---|---|---|
| | | vh | h |
| Y | a | 0.7 | 0.15 |
| | b | 0.1 | 0.05 |

- $P(X = \text{vh}) = P(Y = \text{a}, X = \text{vh}) + P(Y = \text{b}, X = \text{vh})$
- More generally for a joint distribution $P(X_1, \ldots, X_n)$

$$P(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} P(x_1, \ldots, x_n)$$

# Marginalization

Analogous to adding slices of the table together




$$p(x,y) = \sum_{z \in \mathcal{Z}} p(x,y,z) \qquad p(x) = \sum_{y \in \mathcal{Y}} p(x,y)$$

# Conditioning

- Suppose that $X$ & $Y$ are random variables w/ distribution $P(X, Y)$
  $X$ : Intelligence, $\mathrm{Val}(X) = \{\text{“Very High,” “High”}\}$
  $Y$ : Grade, $\mathrm{Val}(Y) = \{\text{“A,” “B”}\}$

| | | X | |
|---|---|---|---|
| | | vh | h |
| Y | a | 0.7 | 0.15 |
| | b | 0.1 | 0.05 |

- We can compute the conditional probability

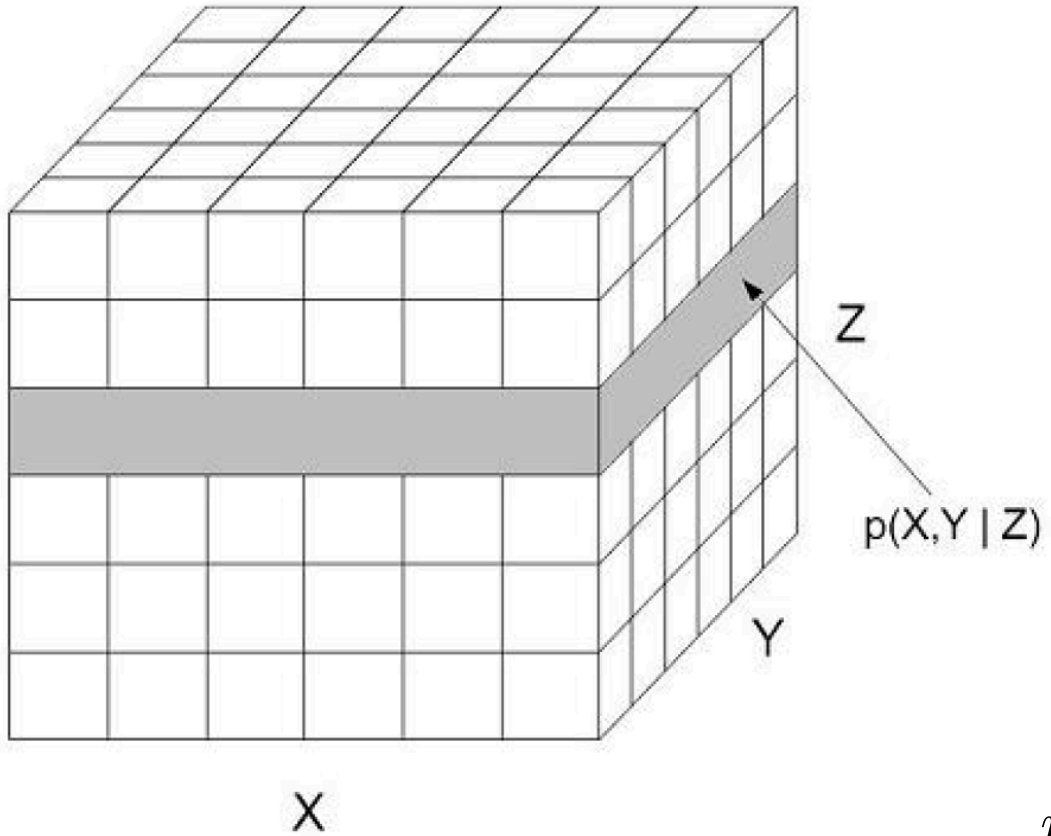
$$P(Y = \mathsf{a} \,|\, X = \mathsf{vh}) = \frac{P(Y = \mathsf{a}, X = \mathsf{vh})}{P(X = \mathsf{vh})}$$

$$= \frac{P(Y = \mathsf{a}, X = \mathsf{vh})}{P(Y = \mathsf{a}, X = \mathsf{vh}) + P(Y = \mathsf{b}, X = \mathsf{vh})}$$

$$= \frac{0.7}{0.7 + 0.1} = 0.875$$

# Conditioning

Analogous to taking a slice of the table (and making this the new universe)

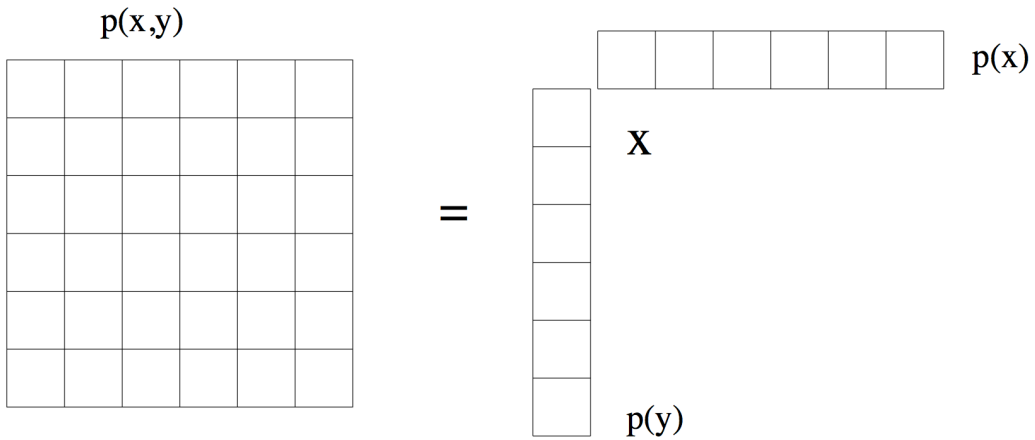



$$p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)}$$

# Independence & Conditional Independence

- Two variables $X$ and $Y$ are independent iff their joint factors as

$$P(X, Y) = P(X)P(Y)$$




- $X$ & $Y$ are **conditionally independent** given $Z = z$ iff slice factors

$$P(X, Y \mid Z = z) = P(X \mid Z = z)P(Y \mid Z = z) \quad \forall z$$

- As we will see, independence is critical to tractability

# Example: Medical Diagnosis

- Variable for each **symptom** (e.g., "fever," "cough," "nausea," etc.)
- Variable for each **disease** (e.g., "peumonia," "cold," "flu," etc.)
- Diagnosis is performed by doing **inference** in the model:

$$P(\text{flu} = 1 \,|\, \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- A famous model is the Quick Medical Reference (QMR-DT), which has $600$ diseases and $4000$ symptoms

# Representing the Distribution

- A naive representation of multivariate distributions is to use a table of probabilities for each outcome
- How many outcomes would there be for QMR-DT?

# Representing the Distribution

- A naive representation of multivariate distributions is to use a table of probabilities for each outcome
- How many outcomes would there be for QMR-DT? $2^{4600} - 1$

# Representing the Distribution

- A naive representation of multivariate distributions is to use a table of probabilities for each outcome
- How many outcomes would there be for QMR-DT? $2^{4600} - 1$
- Learning the joint distribution would require a *huge* amount of data
- Inference of conditional probabilities

$$P(\text{flu} = 1 \,|\, \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

would require summing over exponentially $(2^{4596})$ many values

# Representing the Distribution

- A naive representation of multivariate distributions is to use a table of probabilities for each outcome
- How many outcomes would there be for QMR-DT? $2^{4600} - 1$
- Learning the joint distribution would require a *huge* amount of data
- Inference of conditional probabilities

$$P(\text{flu} = 1 \,|\, \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

  would require summing over exponentially ($2^{4596}$) many values
- Tractable learning and inference requires exploiting independencies

# Exploiting Independence Structure

- If $X_1, \ldots, X_n$ are independent

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2) \cdots P(x_n)$$

- Only $n$ parameters are necessary vs. $2^n - 1$ (exponentially less)!
- But . . . lack of dependencies negates value of evidence

# Exploiting Independence Structure

- Suppose $P \vDash \{(X_1 \dots X_{i-1} \perp X_{i+1} \dots X_n \,|\, X_i) \; \forall i\}$

$$P(x_1, \dots, x_n) =$$

# Exploiting Independence Structure

- Suppose $P \models \{(X_1 \dots X_{i-1} \perp X_{i+1} \dots X_n \,|\, X_i) \; \forall i\}$

$$P(x_1, \dots, x_n) = P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2) \dots P(x_n \,|\, x_{n-1})$$

# Exploiting Independence Structure

- Suppose $P \models \{(X_1 \dots X_{i-1} \perp X_{i+1} \dots X_n \,|\, X_i) \; \forall i\}$

$$P(x_1, \dots, x_n) = P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2) \dots P(x_n \,|\, x_{n-1})$$

- How many parameters?

# Exploiting Independence Structure

- Suppose $P \vDash \{(X_1 \dots X_{i-1} \perp X_{i+1} \dots X_n \,|\, X_i) \; \forall i\}$

$$P(x_1, \dots, x_n) = P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2) \dots P(x_n \,|\, x_{n-1})$$

- How many parameters? $2n - 1$ (vs. $2^n - 1$ for the full joint)

# Naive Bayes Model

- Suppose that $X_1 \dots X_n$ are conditionally independent given $Y$, i.e., $P : (X_i \perp \boldsymbol{X}_{-i} \,|\, Y)$

$$P(y, x_1, \dots, x_n) = P(y)P(x_1 \,|\, y) \prod_{i=2}^{n} P(x_i \,|\, x_1, \dots, x_{i-1}, y)$$

$$= P(y) \prod_{i=1}^{n} P(x_i \,|\, y)$$

- How many parameters?

# Naive Bayes Model

- Suppose that $X_1 \dots X_n$ are conditionally independent given $Y$, i.e., $P : (X_i \perp \boldsymbol{X}_{-i} \,|\, Y)$

$$P(y, x_1, \dots, x_n) = P(y)P(x_1 \,|\, y) \prod_{i=2}^{n} P(x_i \,|\, x_1, \dots, x_{i-1}, y)$$

$$= P(y) \prod_{i=1}^{n} P(x_i \,|\, y)$$

- How many parameters? $2n + 1$

# Naive Bayes Model

- Suppose that $X_1 \dots X_n$ are conditionally independent given $Y$, i.e., $P : (X_i \perp \boldsymbol{X}_{-i} \,|\, Y)$

$$P(y, x_1, \dots, x_n) = P(y)P(x_1 \,|\, y) \prod_{i=2}^{n} P(x_i \,|\, x_1, \dots, x_{i-1}, y)$$

$$= P(y) \prod_{i=1}^{n} P(x_i \,|\, y)$$

- Simple, yet expressive

# Naive Bayes Model: Example

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index words in a dictionary
  - $X_i = 1$ if word $i$ appears in an e-mail
  - E-mails are drawn according to distribution $P(Y, X_1, \ldots X_n)$
- Suppose words are conditionally independent given $Y$

$$P(y, x_1, \ldots, x_n) = p(y) \prod_{i=1}^{n} p(x_i \,|\, y)$$

- Learn (estimate) the model via maximum likelihood
- Infer (predict) whether an e-mail is spam

$$P(Y = 1 \,|\, x_1, \ldots, x_n) = \frac{P(Y = 1) \prod_{i=1}^{n} p(x_i \,|\, Y = 1)}{\sum_{y \in \{0,1\}} P(Y = y) \prod_{i=1}^{n} p(x_i \,|\, Y = y)}$$

# Key Questions

- How do we specify distributions that satisfy particular independence properties?
  $\Rightarrow$ **Representation**
- How do we exploit independencies for efficient computation?
  $\Rightarrow$ **Inference**
- How do we identify independence properties present in data?
  $\Rightarrow$ **Learning**

**Next Lecture: Bayesian Networks**