



School of
Computing

CS5340

Uncertainty Modeling in AI

Asst. Prof. Harold Soh
Dept of Computer Science
National University of Singapore

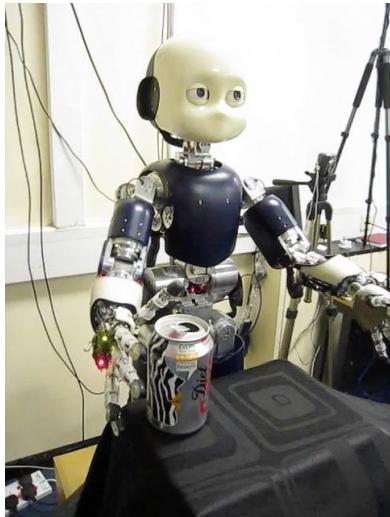
Poll Everywhere!

<https://pollev.com/haroldsohsoo986>





Introduction



Lecturer:

Harold Soh

Department of CS

Office: COM3-02-58

Email: harold@comp.nus.edu.sg

Web: <https://haroldsoh.github.io>

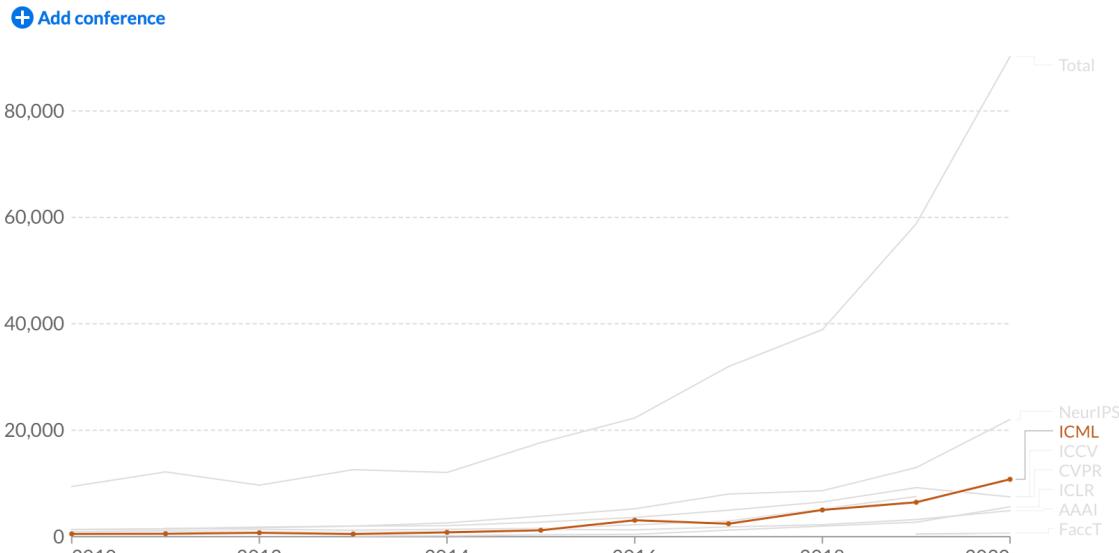
Research Interests:

Human AI/Robot Interaction,
Machine Learning, AI, Robotics

Artificial Intelligence is Cool (again)!

Annual attendance at major artificial intelligence conferences

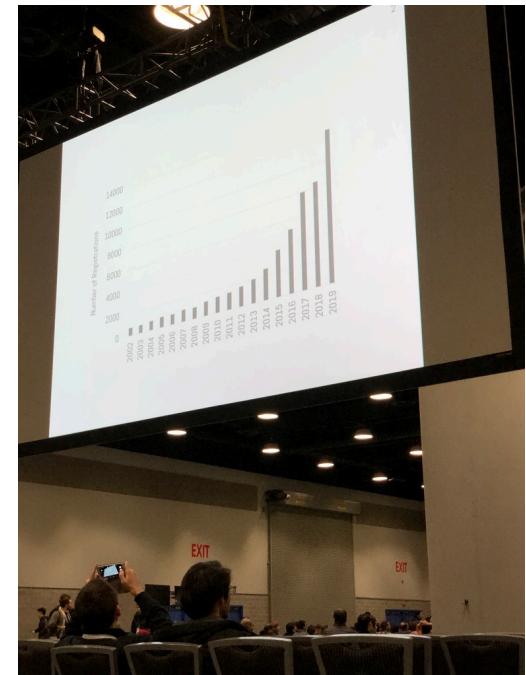
Sixteen major conferences are included.



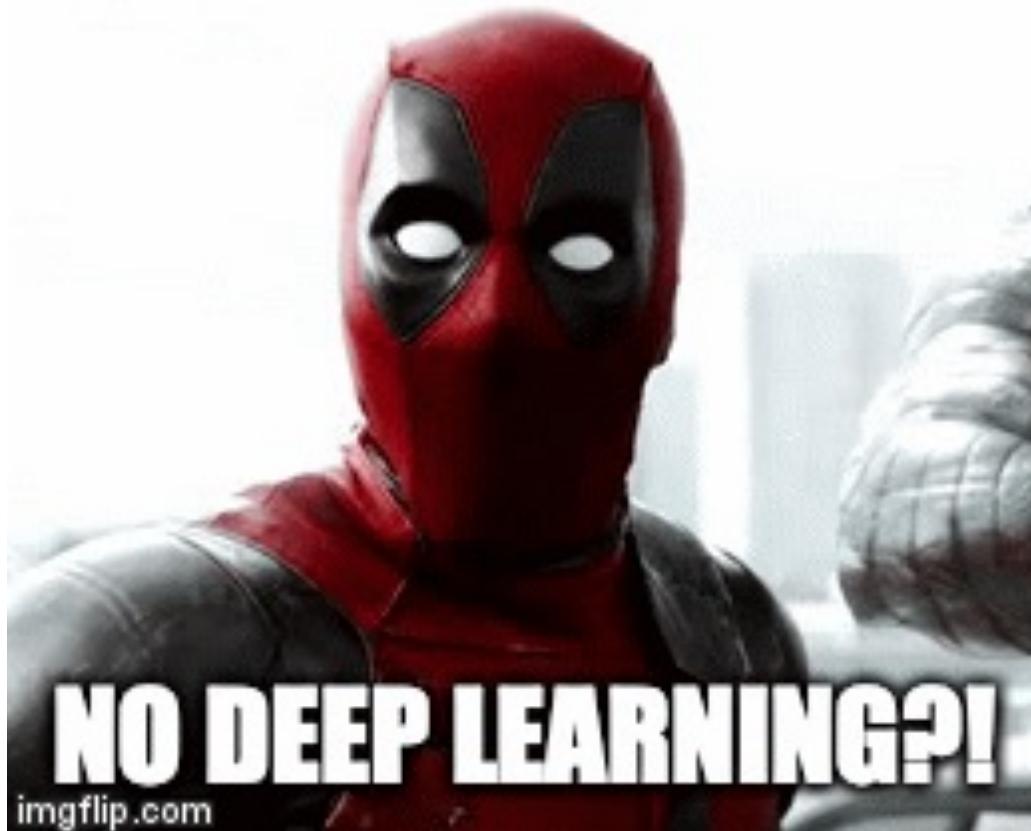
Source: AI Index Report (2022)

Note: Most conferences in 2020–2021 were held virtually due to the COVID-19 pandemic.

OurWorldInData.org/artificial-intelligence • CC BY



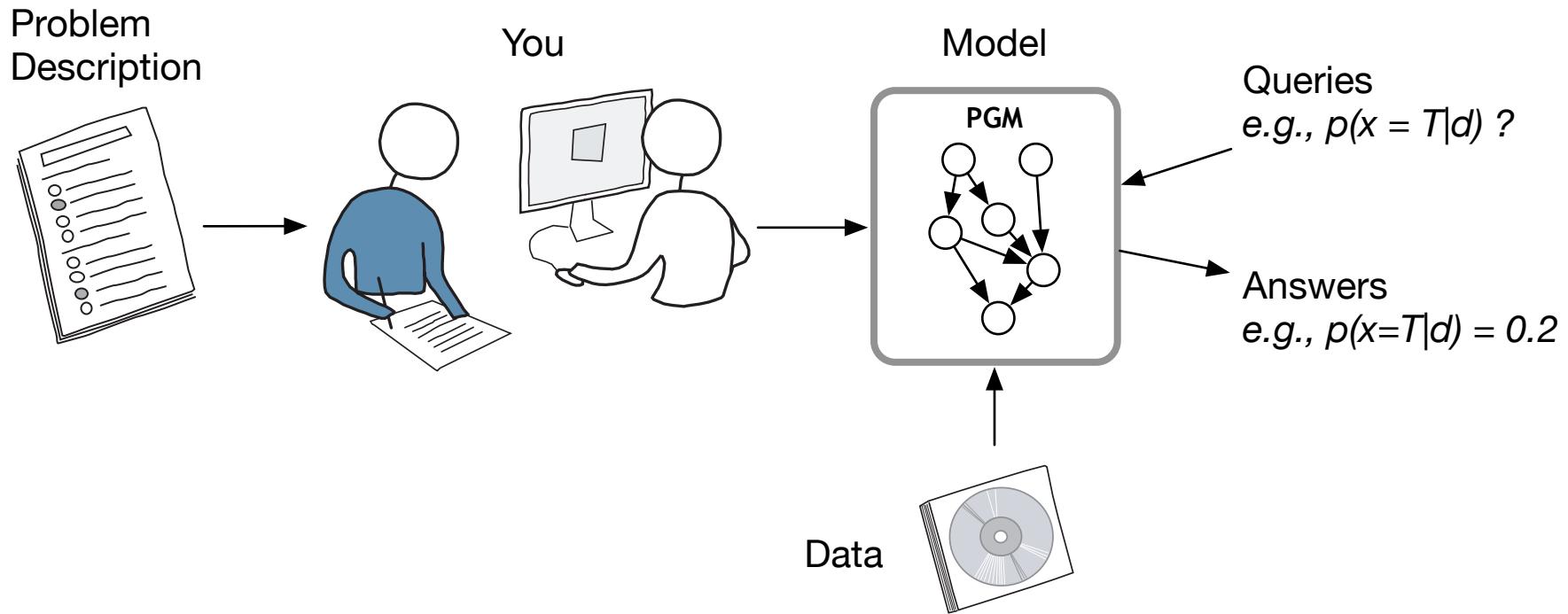
CS5340 is not a “deep learning” class



So, what is CS5340 about?

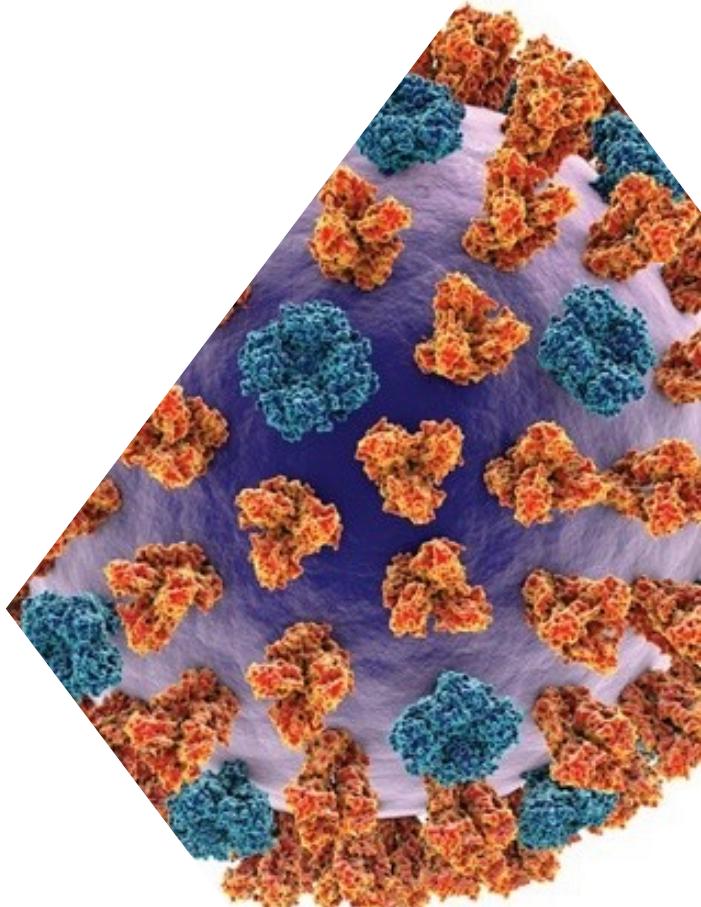
CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



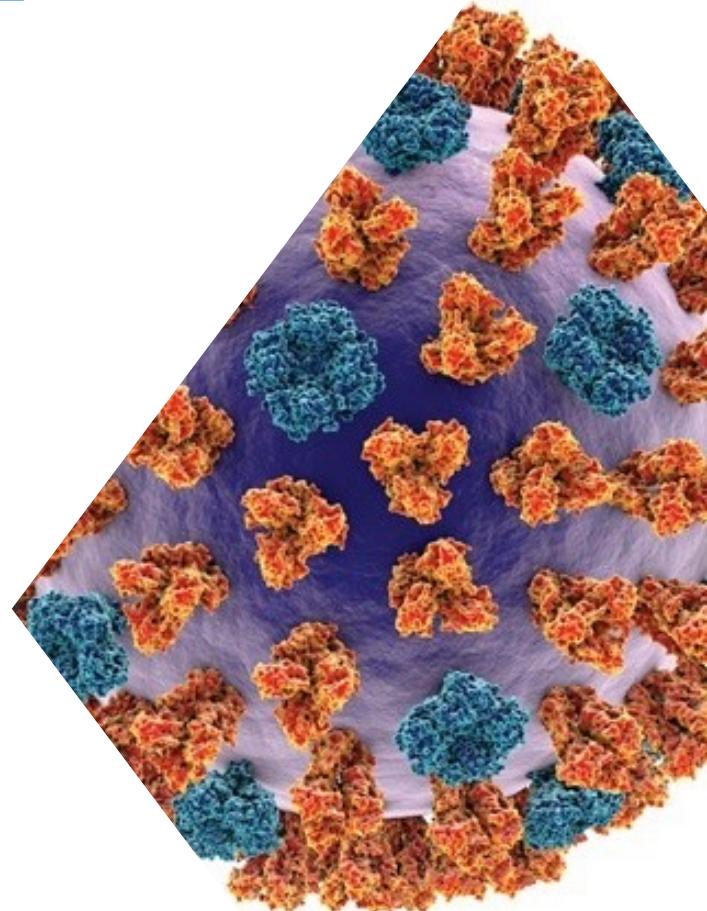
A Modeling Exercise: Influenza

- You're hired to design an AI Diagnostician.
- Scope: **Influenza** (Flu)
- **Problem:** *Given a patient X and information available, does X have the flu?*
- Modeling steps:
 - What are the *relevant variables*?
 - What are the *relationship* between the variables?



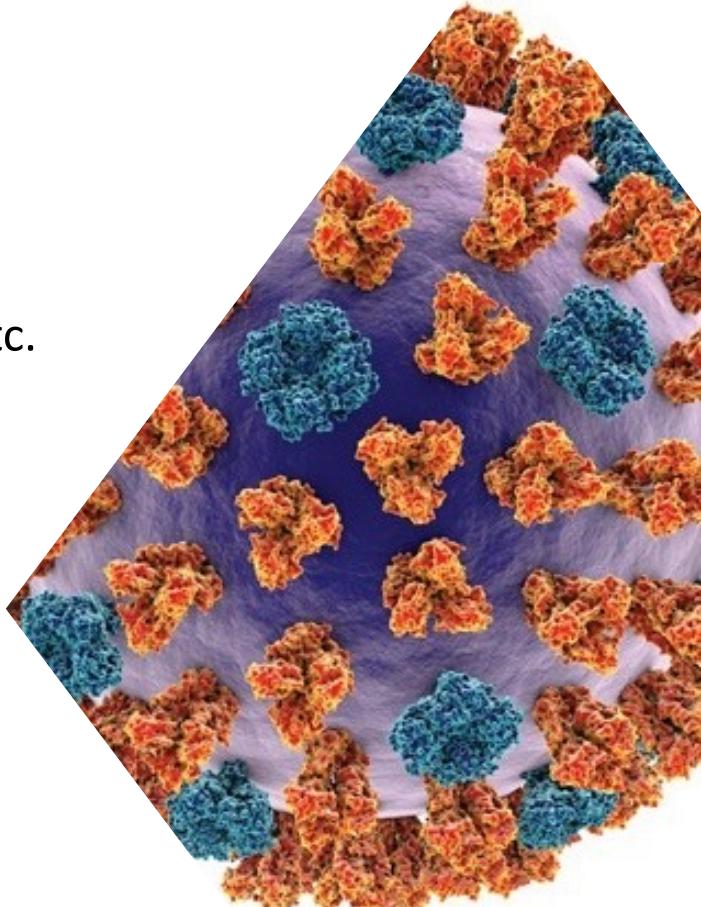
Possible Relevant Variables

<https://pollev.com/haroldsohsoo986>



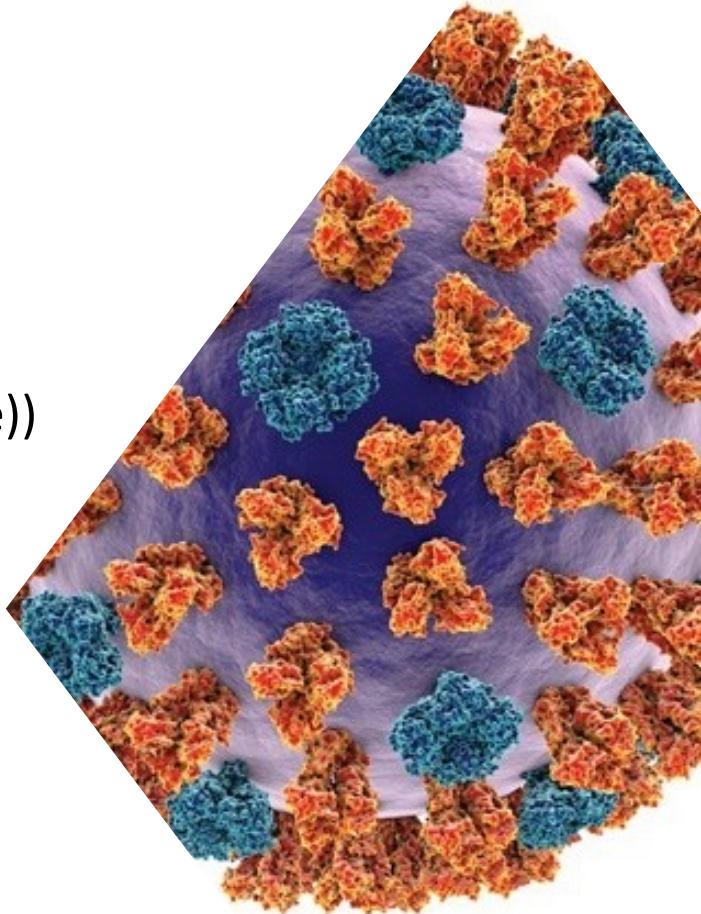
Possible Relevant Variables

- Whether X has the flu or not.
- Immunity to influenza
- Vaccinated
- Genetically immune
- Symptoms present or not:
 - Fever, Cough, Sore throat, Runny nose, Etc.
- Whether the person came into contact with a person with the flu recently.
- Current prevalence of flu in the population
- Structure of the population
- Flu test result
- and so on ...



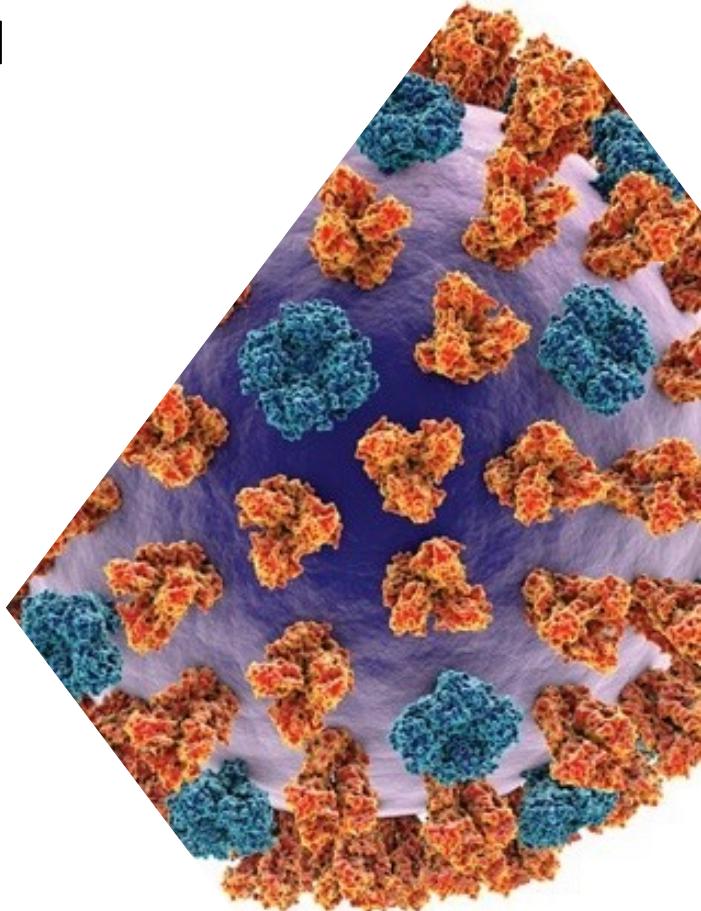
Modeling the Relationships

- Consider binary variables.
- How can we *represent* how the variables are related?
- One simple way:
 - Rules, e.g.,
 - **IF** (fever **AND** cough **AND** (**NOT** immune))
THEN flu=TRUE
ELSE flu=FALSE
- What are the problems?

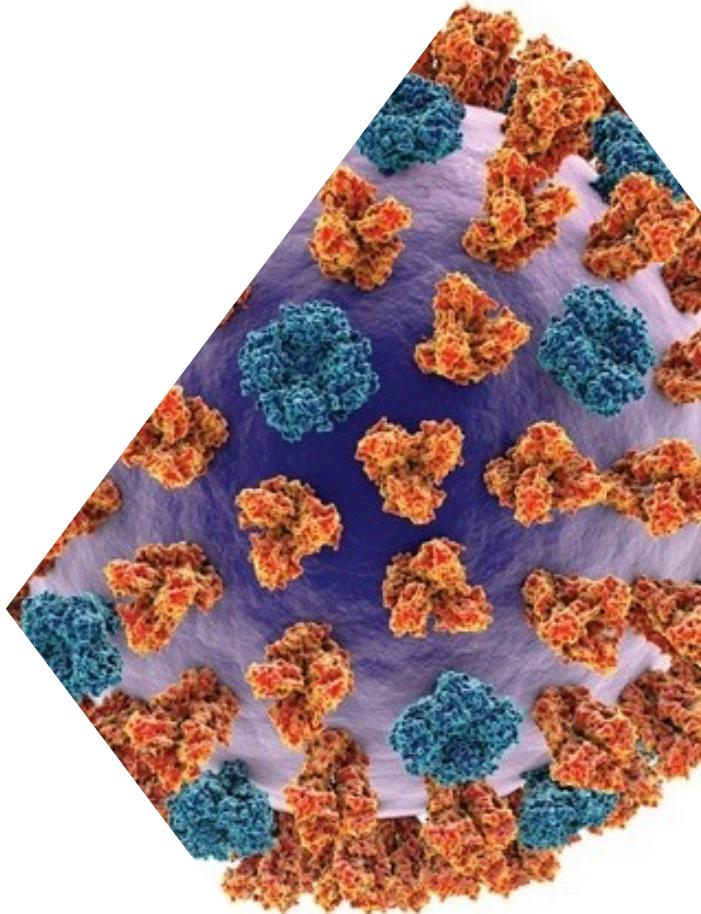
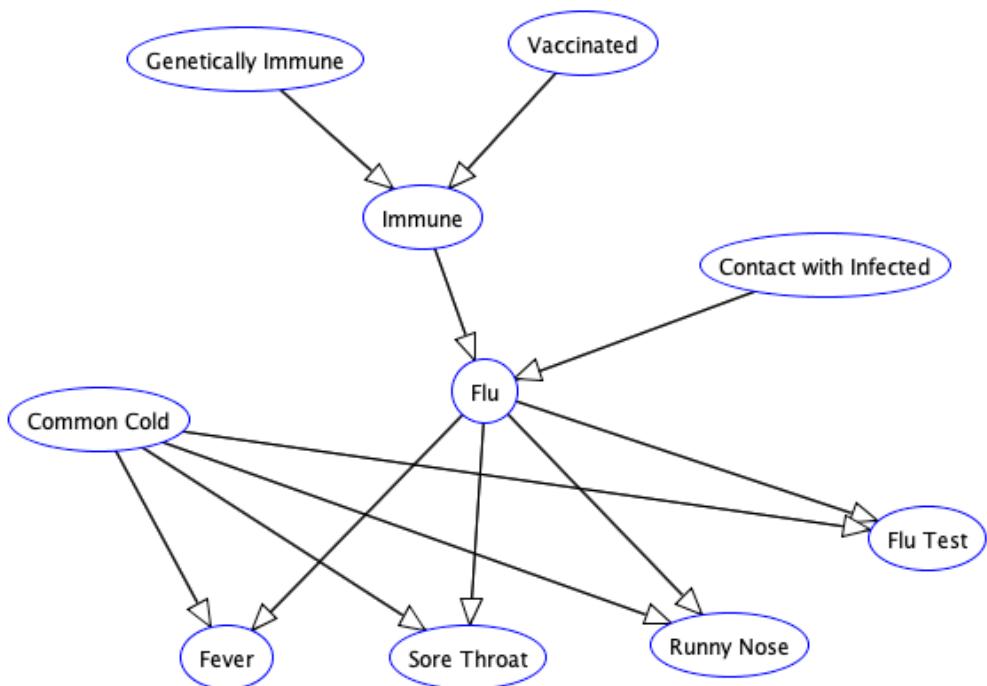


Generative (Causal) Modeling of Relationships between Variables

- A probabilistic approach.
- Capture the generative process to model the ***structure*** between variables and the ***conditional probabilities*** involved.
- Let's try this out.
- Belief and Decision Networks
 - <http://www.aispace.org/downloads.shtml>

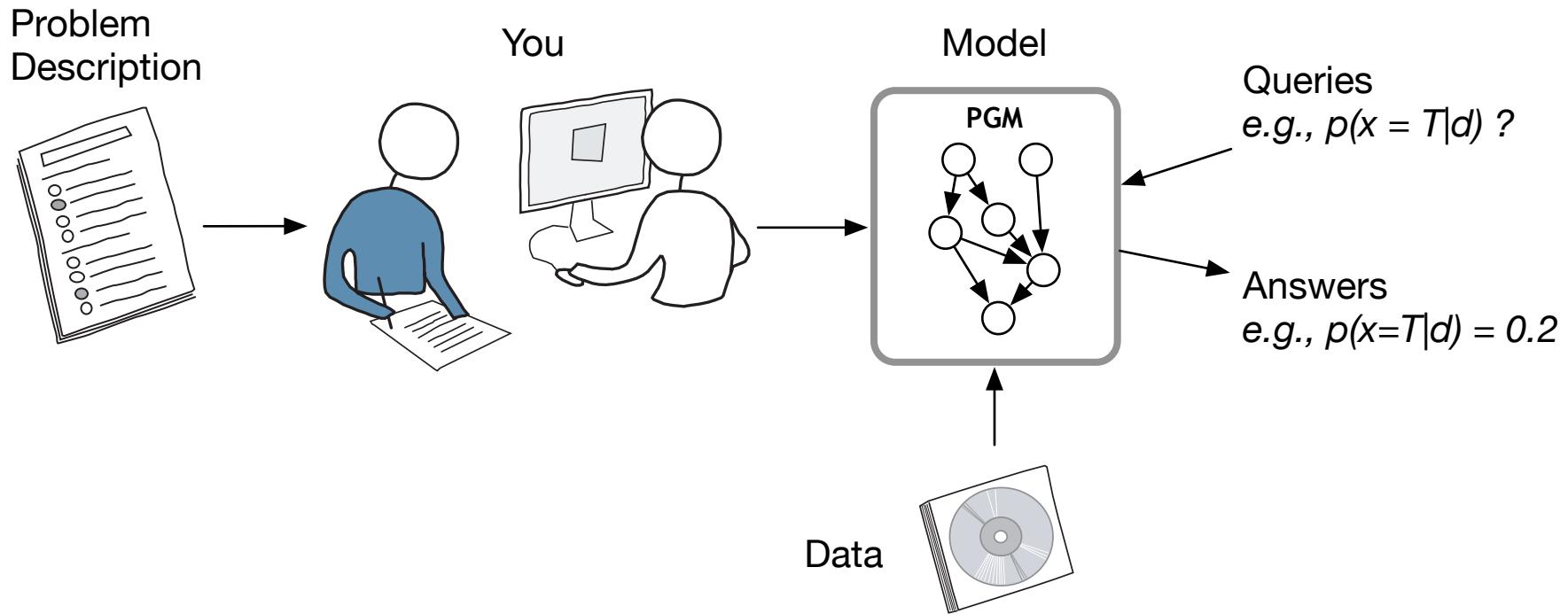


Generative (Causal) Modeling of Relationships between Variables



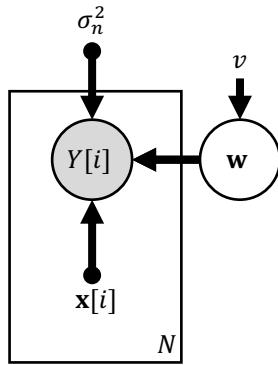
CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



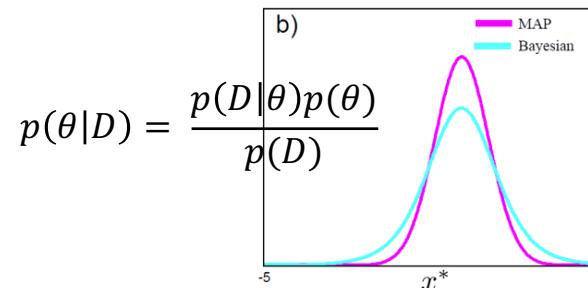
CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



Representation: The *language* is probability and **probabilistic graphical models** (PGM).

The language is used to **model problems**.



Reasoning: We use **learning** and **inference algorithms** to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

Types of Uncertainties

- **Aleatoric Uncertainty**
 - *Natural* randomness in the process
 - A parameter's value can change with each trial (the variability is random)
- **Epistemic Uncertainty**
 - Uncertainty in the *model*
 - A parameter takes on a certain value but we're not sure what that value is.
- **Exercise:** Consider the influenza diagnosis problem. What are the *aleatoric* and *epistemic* uncertainties present?

PGM: Applications

Robotics

Medical Diagnosis

Recommender Systems

Natural Language Processing

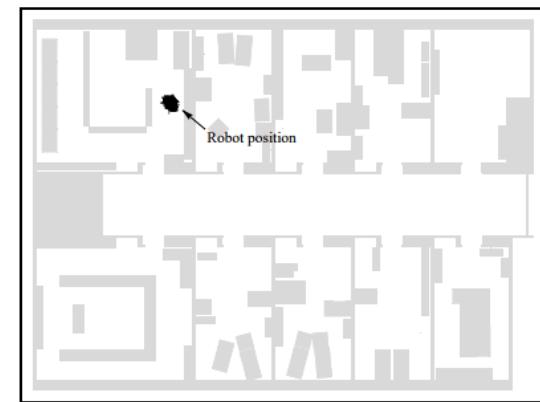
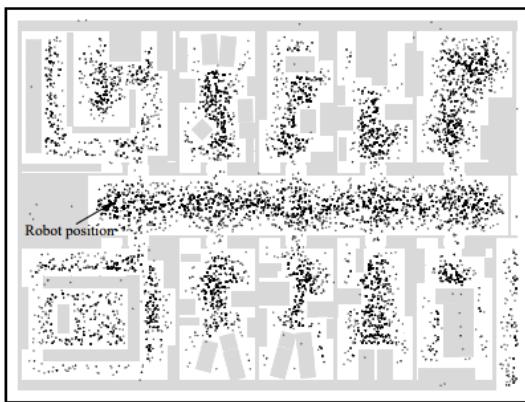
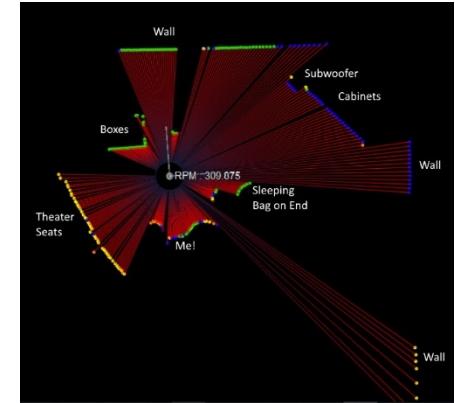
Computer Vision

Generative Modelling

And many more!

Applications

Markov Localization



“Monte Carlo Localization for Mobile Robots”, Frank Dellaert et. al., ICRA 1999

Applications

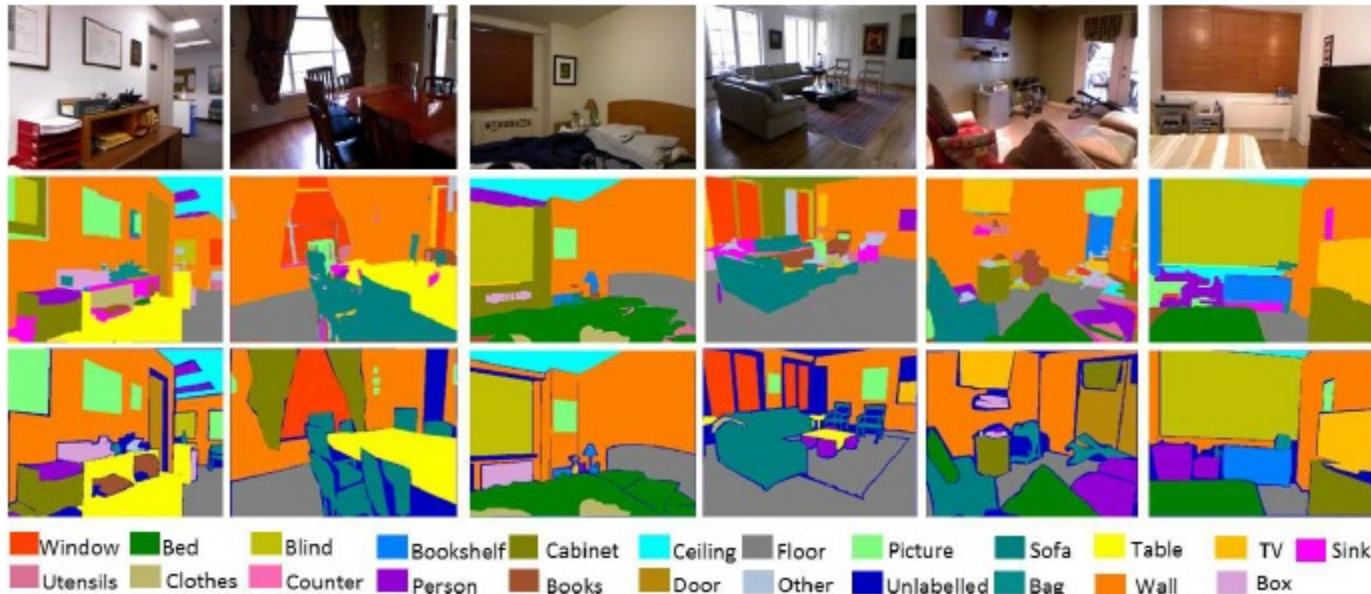
Markov Localization



<https://www.youtube.com/watch?v=P8wLQz9mpjg>

Applications

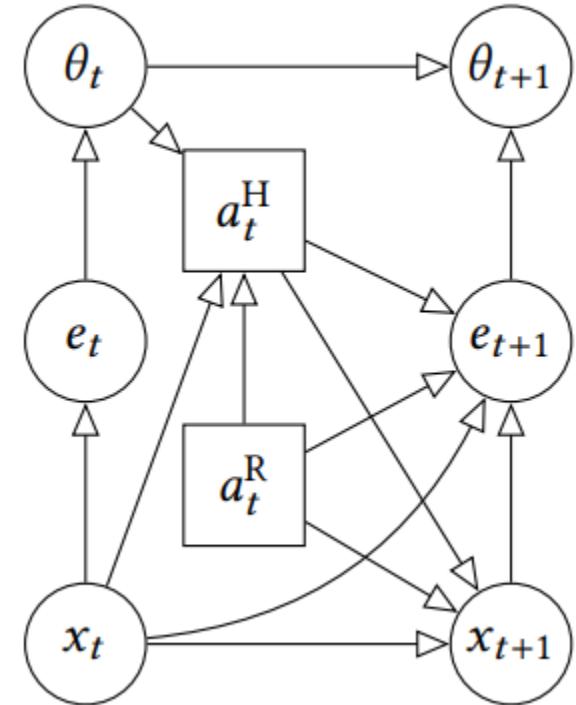
Scene Understanding



“Geometry Driven Semantic Labeling of Indoor Scenes”, Salman Hameed Khan et. Al. ECCV 2014

Applications

Human-Robot Collaboration

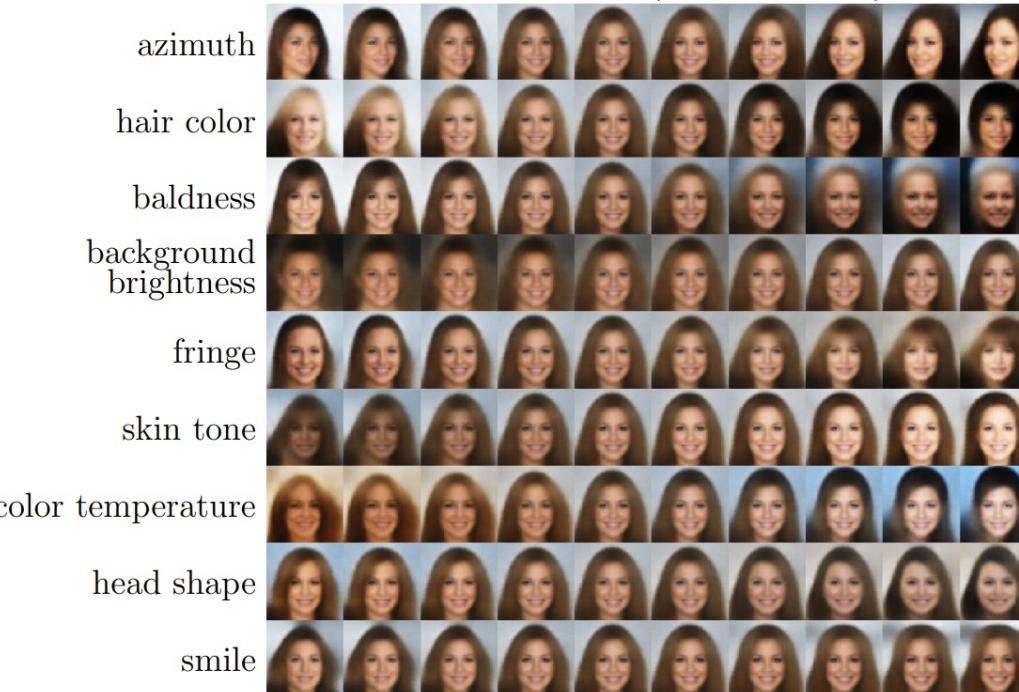


“Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning”, Chen et al., HRI 2018

Applications

Generative Modeling

CHyVAE ($\nu = 1000$)



Ansari, Abdul Fatir, and Harold Soh. "Hyperprior induced unsupervised disentanglement of latent representations." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019.

Applications

Generative Modeling via Diffusion Models

The screenshot shows a news article from arstechnica.com. The title is "With Stable Diffusion, you may never believe what you see online again". Below the title, it says "AI image synthesis goes open source, with big implications." and "BENJ EDWARDS - 9/6/2022, 9:30 PM". The article features three generated images: Abraham Lincoln as a cowboy, a woman with blonde hair, and a woman in a mechanical suit. At the bottom, there's a quote: "Al image generation is here in a big way. A newly released open source image synthesis model called [Stable Diffusion](#) allows anyone with a PC and a decent GPU to conjure up almost any visual reality they can imagine. It can imitate virtually any visual style, and if you feed it a descriptive phrase, the results appear on your screen like magic."



<https://arstechnica.com/information-technology/2022/09/with-stable-diffusion-you-may-never-believe-what-you-see-online-again/>

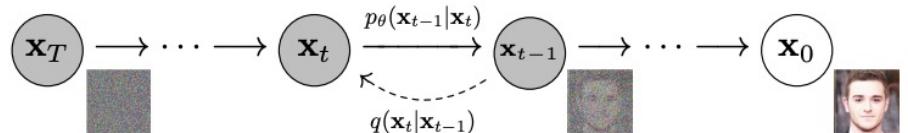
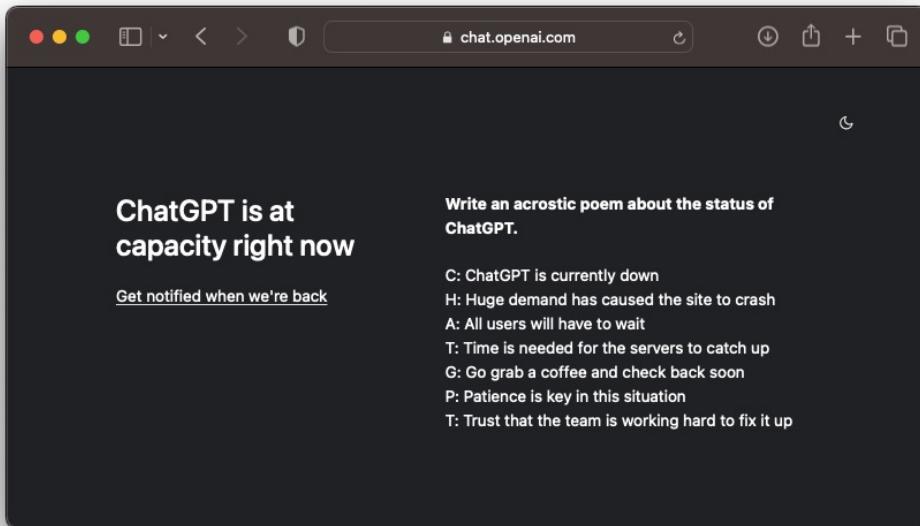


Figure 2: The directed graphical model considered in this work.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.

Applications

- Language Modeling



2. Approach

At the core of our approach is language modeling. Language modeling is usually framed as unsupervised distribution estimation from a set of examples (x_1, x_2, \dots, x_n) each composed of variable length sequences of symbols (s_1, s_2, \dots, s_n) . Since language has a natural sequential ordering, it is common to factorize the joint probabilities over symbols as the product of conditional probabilities (Jelinek & Mercer, 1980) (Bengio et al., 2003):

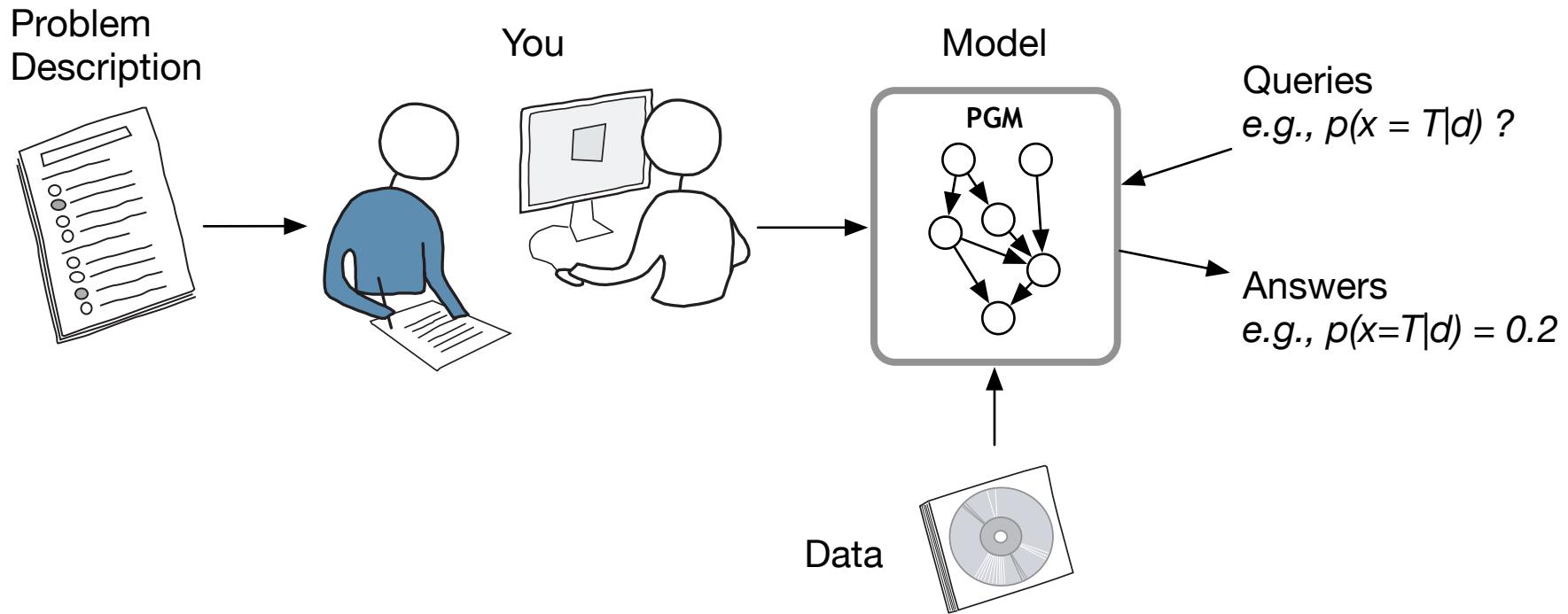
$$p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1}) \quad (1)$$

This approach allows for tractable sampling from and estimation of $p(x)$ as well as any conditionals of the form $p(s_{n-k}, \dots, s_n | s_1, \dots, s_{n-k-1})$. In recent years, there have been significant improvements in the expressiveness of models that can compute these conditional probabilities, such as self-attention architectures like the Transformer (Vaswani et al., 2017).

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



Questions?

<https://pollev.com/haroldsohsoo986>



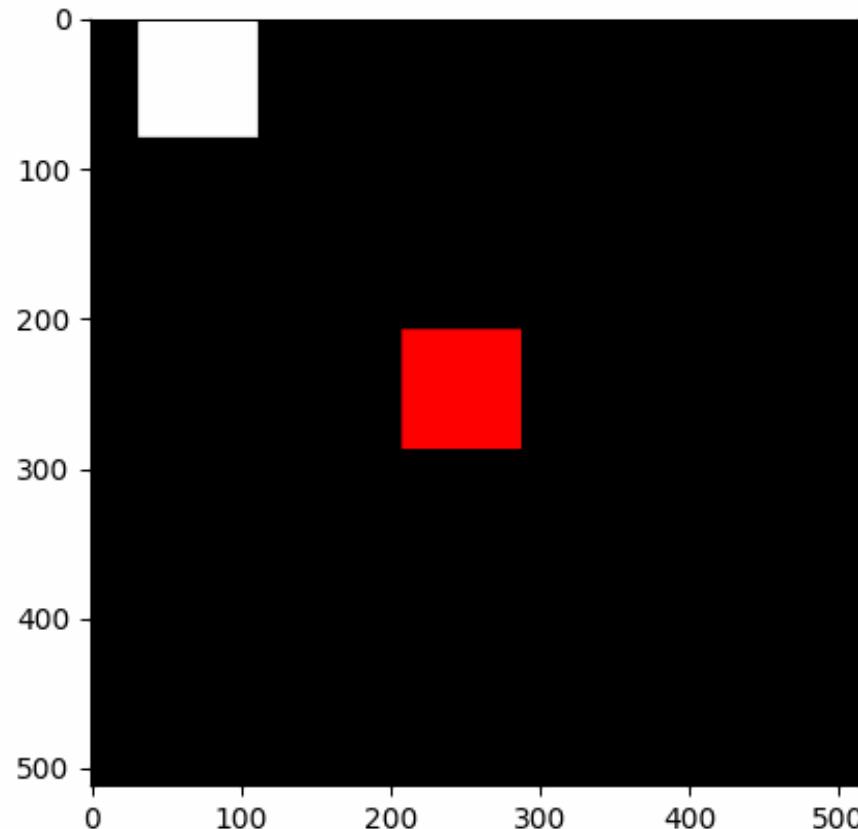
Course Administration

Course Info, Schedule, and Assessments

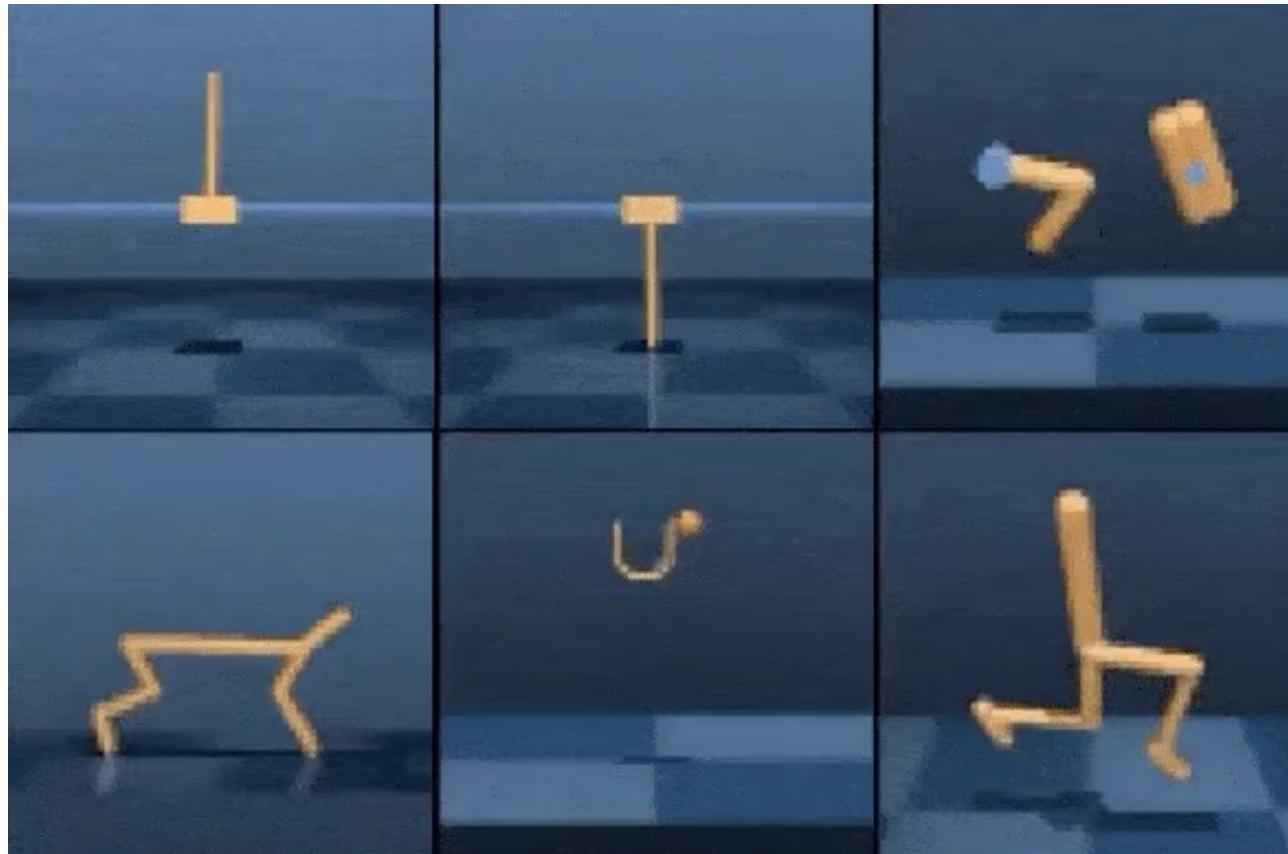
Course Schedule

Week	Date	Lecture Topic	Tutorial Topic
1	12 Jan	Introduction to Uncertainty Modeling + Probability Basics	Introduction
2	19 Jan	Simple Probabilistic Models	Probability Basics
3	26 Jan	Bayesian networks (Directed graphical models)	More Basic Probability
4	2 Feb	Markov random Fields (Undirected graphical models)	DGM modelling and d-separation
5	9 Feb	Variable elimination and belief propagation	MRF + Sum/Max Product
6	16 Feb	Factor graph and the junction tree algorithm	Quiz 1
-	-	RECESS WEEK	
7	2 Mar	Mixture Models and Expectation Maximization (EM)	Linear Gaussian Models
8	9 Mar	Hidden Markov Models (HMM)	Probabilistic PCA
9	16 Mar	Monte-Carlo Inference (Sampling)	Linear Gaussian Dynamical System
10	23 Mar	Variational Inference	MCMC + Sequential VAE
11	30 Mar	Inference and Decision-Making (Special Topic)	Quiz 2
12	6 Apr	Gaussian Processes (Special Topic)	Wellness Day
13	13 Apr	Project Presentations	Closing

From Zero to Control via Pixels



From Zero to Control via Pixels



<https://planetrl.github.io>

Course Information

Lecture and Tutorials: Blended Classroom style

- **Lecture Videos Online:** Posted on Canvas (Panopto) every Thursday
- **Tutorial:** Thursday 18:30 to 20:30 at LT15

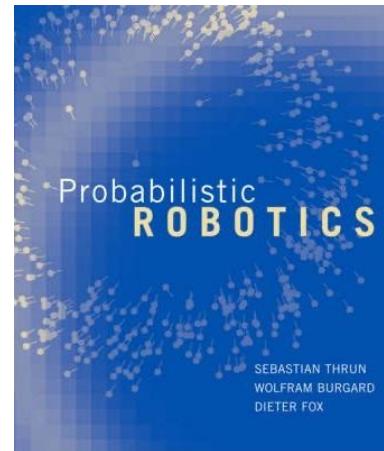
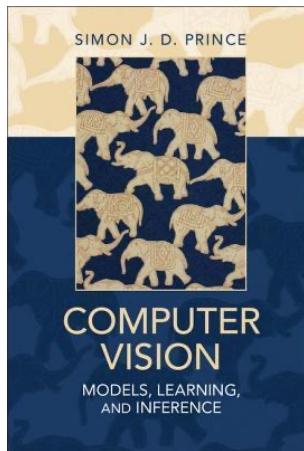
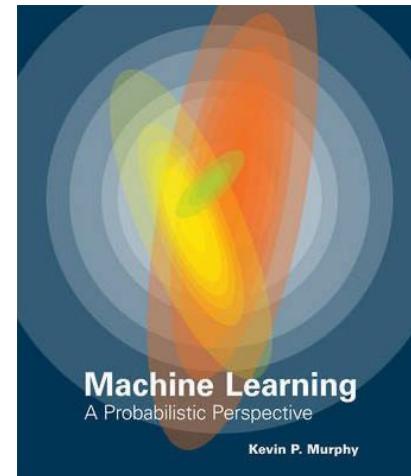
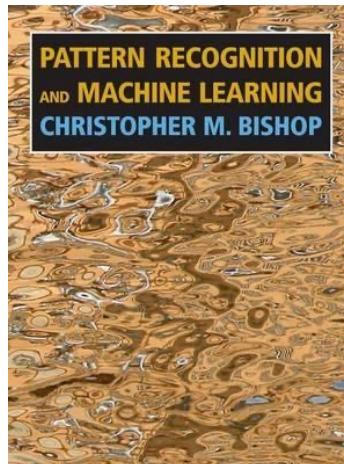
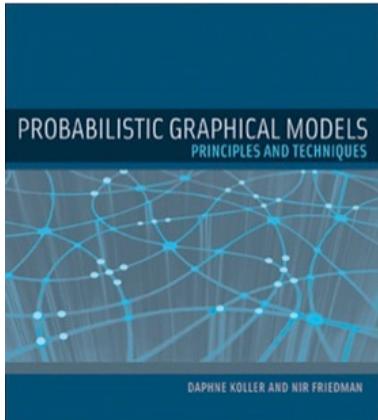
Prerequisites:

- A statistics/probability course: CS3243 and ((EE2012/A or ST2132 or ST2334 or ((MA2216 or ST2131) and (ST1131/A or ST1232 or DSC2008))))
 - **Please note:** we will spend very little time on basic probability.
- CS3243 Introduction to Artificial Intelligence

Drop Dates (Please verify):

- **'W' grade: 23 Jan 2022, 23:59hrs onwards.**
- **'F' grade: 27 Feb 2022, 23:59hrs onwards.**

Recommended Readings (Not Compulsory)



Teaching Assistant



Chen Kaiqi
Department of Computer Science
Email: kaiqi@comp.nus.edu.sg

Canvas

The screenshot shows the Canvas course interface for CS5340 Uncertainty Modelling. The left sidebar contains a vertical navigation menu with icons and labels: Account, Dashboard, Courses (selected), Calendar, Inbox, History, Studio, DYOM, Help, Home, Assignments, Discussions, Grades, People, Pages, Syllabus, Quizzes, Modules, Collaborations, New Analytics, Zoom, Videos/Panopto, Student Feedback, Course Readings, and Chat. The main content area displays the course modules. The first module is "Welcome" with sections: Welcome to CS5350!, Time Table, Pre-requisites, Materials/Textbook, Assessments, and Honor Code and Plagiarism. The second module is "Week 1: Introduction to Uncertainty Modeling" with a section: Week 1 Summary. A purple banner at the bottom left says "You are currently logged in to student view". Another banner at the bottom right says "Resetting the test student will clear all history for this student and allow you to view the course as a brand new student." Buttons for "Reset student" and "Leave student view" are also present.

CS5340 Uncertainty Modelling

[2220] 2022/2023 Semester 2

CS5340 > Modules

Home

Assignments

Discussions

Grades

People

Pages

Syllabus

Quizzes

Modules

Collaborations

New Analytics

Zoom

Videos/Panopto

Student Feedback

Course Readings

Chat

Welcome

- Welcome to CS5350!
- Time Table
- Pre-requisites
- Materials/Textbook
- Assessments
- Honor Code and Plagiarism

Week 1: Introduction to Uncertainty Modeling

- Week 1 Summary

You are currently logged in to student view

Resetting the test student will clear all history for this student and allow you to view the course as a brand new student.

Reset student

Leave student view

Discussion Forums

The screenshot shows the 'Discussions' page for the course CS5340 on the Canvas platform. The left sidebar includes links for Account, Dashboard, Courses (selected), Calendar, Inbox, History, Studio, DYOM, Help, and various course modules like Assignments, Grades, People, Pages, Syllabus, Quizzes, Modules, Collaborations, New Analytics, Zoom, Videos/Panopto, Student Feedback, Course Readings, and Chat.

The main content area displays 'Pinned discussions' and 'Discussions'. The pinned discussion is titled 'Happy New Year and Welcome to CS5340!' and was posted by 'All sections' on 'Last post at 5 Jan, 08:16'. Below this, there is a section for 'Discussions' which is currently empty, indicated by the message 'There are no discussions to show in this section' and 'Click here to add a discussion'. A small icon of two people talking is shown. At the bottom, there is a section for 'Closed for comments'.

At the bottom of the page, a purple banner states: 'You are currently logged in to student view' and 'Resetting the test student will clear all history for this student and allow you to view the course as a brand new student.' It includes buttons for 'Reset student' and 'Leave student view'.

Assessments

- **No Final Exam** 😊
- **Team Project:** 30%
- **Quizzes:** 40%
- **Tutorials:** 20%
- **Participation:** 10%
 - Forum Participation

Project (30%)

- **Team-based**
- 4 to 5 people per team
- Key Dates:
 - Form Teams: **3 Feb**
 - Abstract Due: **20 Feb**
 - Presentations: **13 April**
 - Report Due: **21 April**
- Deliverables:
 - Abstract
 - NeurIPS LaTeX Template
 - Up to 4 pages excl. references
 - Technical report
 - NeurIPS LaTeX Template
 - Up to 8 pages excl. references
 - 5-10 mins Presentation

Example Past Projects



- **Infectious Disease Modeling/Diagnosis**
 - Uncertainty in Patient Diagnosis
 - Modeling the Mortality Risk of COVID-19
 - Modeling COVID-19 Spread
- **Computer Vision / Image Processing**
 - Techniques for Image Denoising
 - Image Background Generation with VAEs and MRFs
- **Financial / Economic / Political Modeling**
 - Uncertainty Distribution of Sales
 - Modeling Political Behaviors with PGMs
 - Portfolio Risk and Return
- **Robotics Prediction / Control**
 - Robust Deep Reinforcement Learning
 - Drone Trajectory Estimation with HMMs
- **Natural Language Processing**
 - Repairing Multilingual Subtitles with Deep Markov Models

Abstract

- **Introduction**
 - What is the problem you want to solve?
 - Why do you want to solve it?
 - Why is it important/interesting?
- **Related Work**
 - What other work has been done in this area?
- **Methodology**
 - How do you propose to solve it?
 - Why do you want to solve it this way?
- **Preliminary Results (if any)**
 - What have you done so far?
- **Note:** Support your statements with evidence (references and/or results)

Quizzes



- **Quiz 1: 20%**
 - **16 Feb 2022**
 - Covers all material up to Week 5 (Variable Elimination and Belief Propagation)
- **Quiz 2: 20%**
 - **30 Mar 2022**
 - Covers all material up to Week 10 (Variational Inference)
- **Rules:**
 - Open book. Online Quiz on Canvas
 - No searching online for answers (or ChatGPT/GPT)

Academic Honesty

- Do ***NOT*** cheat
- **Strict** Plagiarism policy:
 - If you cheat, we will report you to the disciplinary board
 - If found guilty, you will get an F (University Policy)
- Please be academically honest.
 - “*Give credit where credit is due*”

Tutorials

- Every Thursday from 6:30pm to 8:30pm
- Graded Questions on Canvas “Quiz” Platform
- Each tutorial is worth 2.5% of your grade.
- Best 8 out of 10 tutorials to calculate your final score.
- **IMPORTANT:** The tutorial session is only open from 6:30pm to 9:30pm. No late entries will be allowed.
- Let’s do one now!

Course Schedule

Week	Date	Lecture Topic	Tutorial Topic
1	12 Jan	Introduction to Uncertainty Modeling + Probability Basics	Introduction
2	19 Jan	Simple Probabilistic Models	Probability Basics
3	26 Jan	Bayesian networks (Directed graphical models)	More Basic Probability
4	2 Feb	Markov random Fields (Undirected graphical models)	DGM modelling and d-separation
5	9 Feb	Variable elimination and belief propagation	MRF + Sum/Max Product
6	16 Feb	Factor graph and the junction tree algorithm	Quiz 1
-	-	RECESS WEEK	
7	2 Mar	Mixture Models and Expectation Maximization (EM)	Linear Gaussian Models
8	9 Mar	Hidden Markov Models (HMM)	Probabilistic PCA
9	16 Mar	Monte-Carlo Inference (Sampling)	Linear Gaussian Dynamical System
10	23 Mar	Variational Inference	MCMC + Sequential VAE
11	30 Mar	Inference and Decision-Making (Special Topic)	Quiz 2
12	6 Apr	Gaussian Processes (Special Topic)	Wellness Day
13	13 Apr	Project Presentations	Closing

Questions?

<https://pollev.com/haroldsohsoo986>



Please Prepare for Next Week

- Watch the videos.
- Do the tutorial.

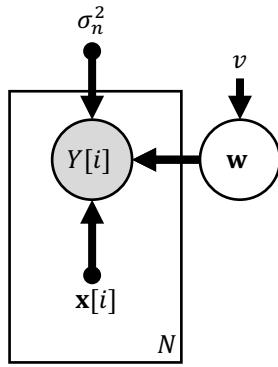
Week 1: Introduction to Uncertainty Modeling	✓	+	⋮
Week 1 Summary	✓	⋮	⋮
Tutorial	✓	⋮	⋮
Tutorial 0 12 Jan 2.5 pts	✓	⋮	⋮
Video Lectures	✓	⋮	⋮
L1 - Part 1 (Preliminaries)	✓	⋮	⋮
L1 - Part 2 (The Basics)	✓	⋮	⋮
L1 - Part 3 (Basic Operations)	✓	⋮	⋮
L1 - Part 4 (Probability Problems)	✓	⋮	⋮
L1 - Part 5 (Common Probability Distributions)	✓	⋮	⋮
Tutorial Sheet (for next week)	✓	⋮	⋮
Tutorial 1.pdf	✓	⋮	⋮

Preliminaries

Introduction

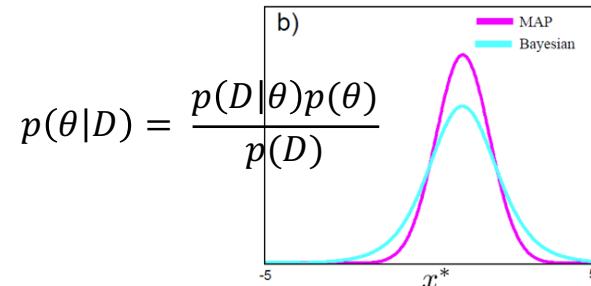
CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



Representation: The *language* is probability and **probabilistic graphical models** (PGM).

The language is used to **model problems**.



Reasoning: We use **learning** and **inference algorithms** to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

Probabilistic Graphical Modeling

Key Ideas:

- **Represent** the world as a collection of random variables X_1, \dots, X_N with joint distribution $p(X_1, \dots, X_N)$.
- **Learn** the distribution from data.
- Perform “**inference**” (compute conditional distributions $p(X_i | X_1 = x_1, \dots, X_N = x_N)$).

What does all this mean precisely?

Learning Outcomes

Students should be able to:

1. Describe uncertain quantities with **random variables** and **joint probabilities**.
2. Explain the basic rules of probability – **sum, product, Bayes', independence and expectation** rules.
3. Use the common probabilities distributions – **Bernoulli, categorical, univariate and multivariate normal** distributions.
4. Explain the use of **conjugate distributions**.

Acknowledgements

- A lot of slides and content of this lecture are adopted from:
 1. Simon Prince, “Computer Vision: Models, Learning, and Inference”, Chapter 1 and 2.
 2. Daphne Koller and Nir Friedman, "Probabilistic graphical models", Chapter 2.
 3. Christopher Bishop, “Pattern Recognition and Machine Learning”, Chapter 2.
 4. MIT Course: Mathematics for CS Readings, Chapter 14 and 18.
 5. Dr. Lee Gim Hee’s CS5340 slides.

The Basics

Events, Outcomes, and Probability

Problem: You're not feeling well...

- You're not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **positive** for *rare, fatal* disease.
- Should you:
 - A. Skip CS5340 and start planning your funeral?
 - B. Not worry. Be Happy.
 - C. Take the test again (and again) until it comes back negative.
 - D. Ask for more information.

Problem: You're not feeling well...

- You're not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **positive** for *rare, fatal* disease.
 - Disease affects 0.1% of the population.
 - Test correctly identifies 99% of the people who have the disease.
 - If you do not have the disease, test may come back positive 2% of the time.

What to do now?

Depends how much you believe whether you have the disease.

What do we **mean** when we say:

- “*the probability that I have the rare fatal disease is 90%*”
- “*the probability of getting an even number when rolling a die is $\frac{1}{2}$* ”

Probability Space

- A probability space (Ω, E, P) models a process consisting of outcomes that occur **randomly**.
- Consists of three parts:
 - Outcome or sample space Ω
 - Event space E
 - Probability function $P: E \rightarrow [0,1]$

Outcome and Event Spaces

- Outcome space is an agreed upon **space of possible outcomes**, denoted by Ω .

Example: Outcomes of a dice roll, $\Omega = \{1,2,3,4,5,6\}$.

- Event space $E \subseteq 2^\Omega$ is a **subset of the power set** of Ω , it is the set of **measurable events** to which we **assign probabilities**.

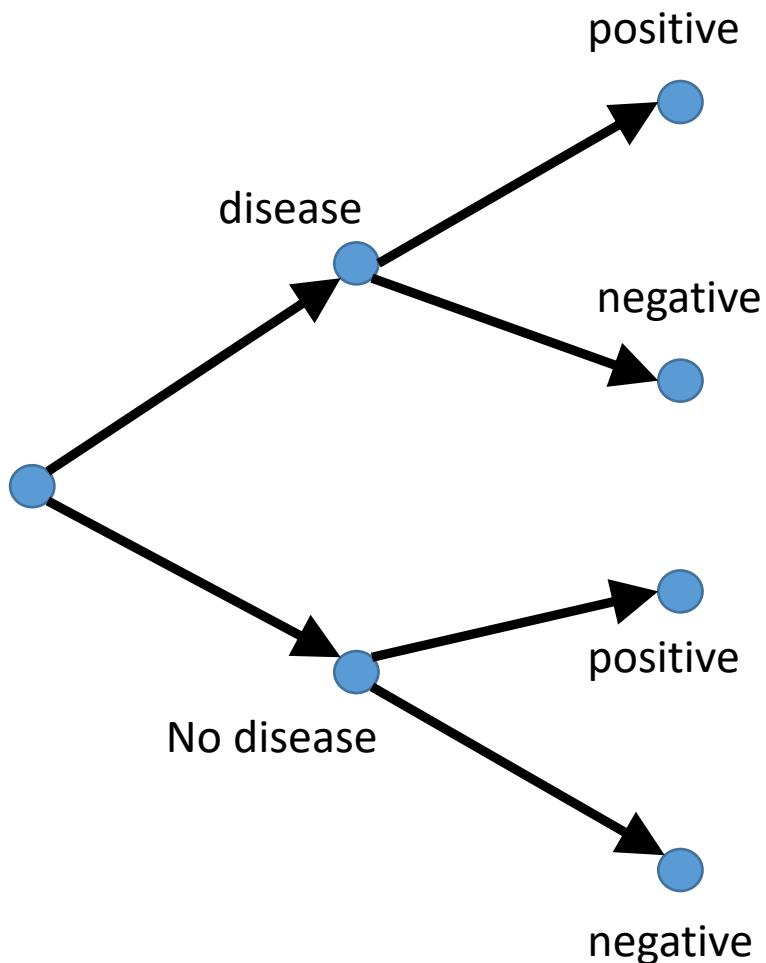
Example: The event space on whether a dice roll is odd or even, $E = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$.



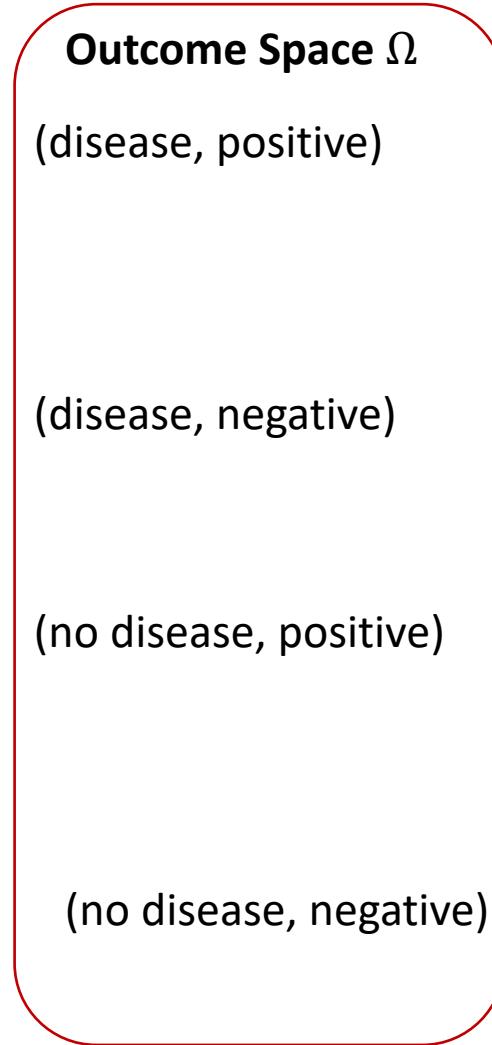
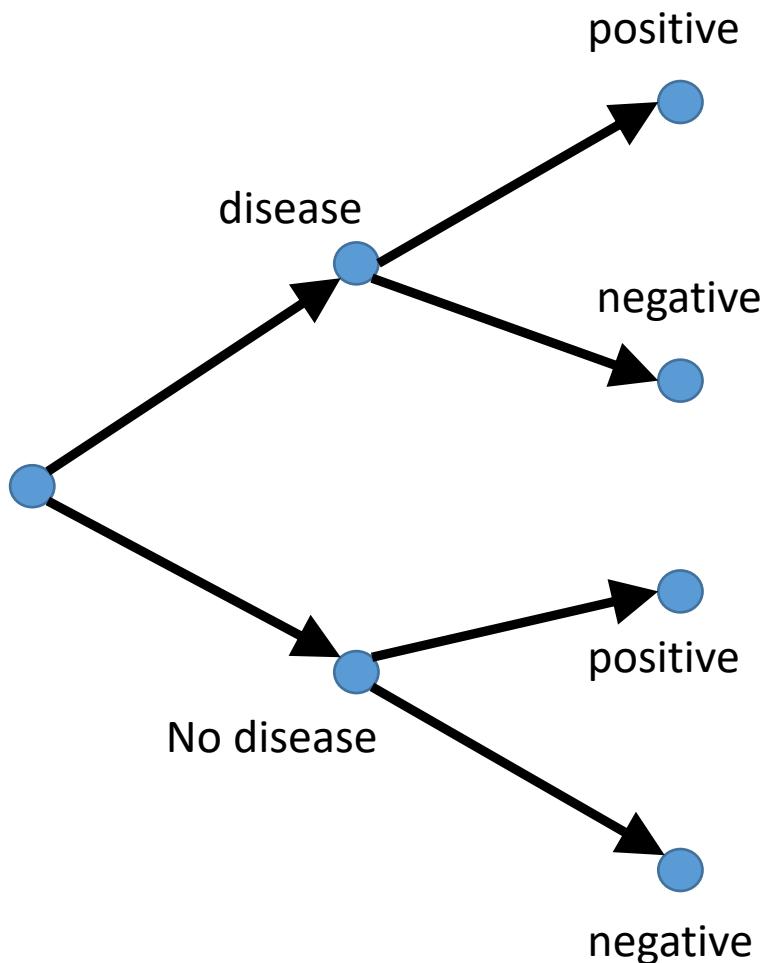
Outcome and Event Spaces

- Event space must satisfy **three basic properties**:
 1. It contains the **empty event** \emptyset , and the **trivial event** Ω .
 2. It is **closed under union**, i.e. if $\alpha, \beta \in E$, then so is $\alpha \cup \beta$.
 3. It is **closed under complement**, i.e. if $\alpha \in E$, then so is $\Omega - \alpha$.

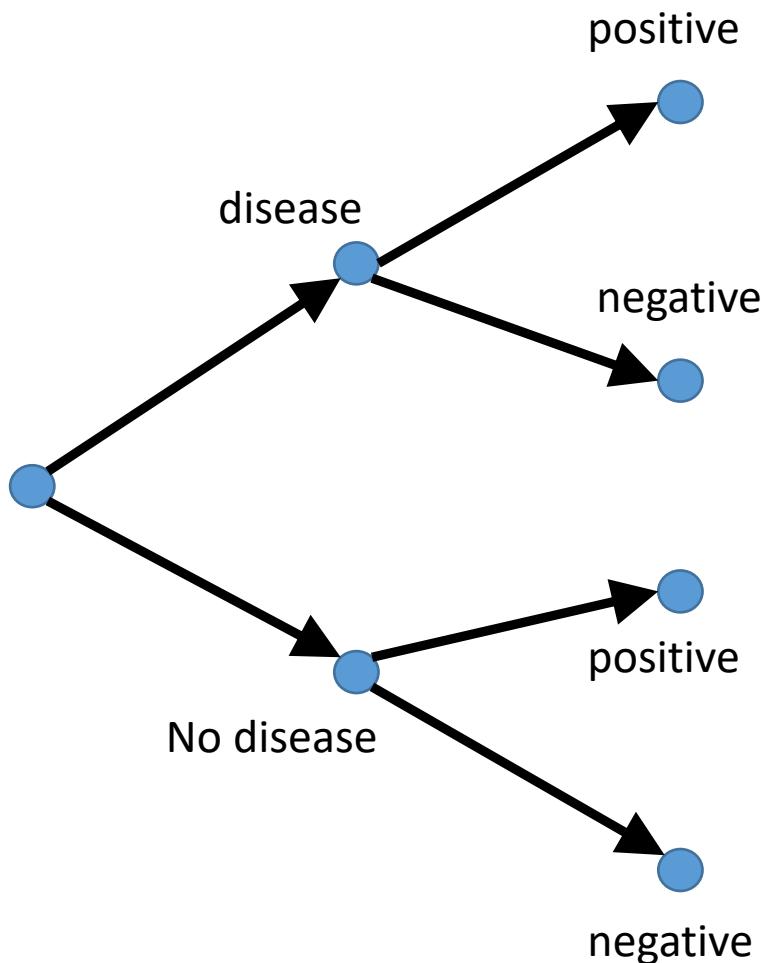
Tree Diagram Example



Tree Diagram Example



Tree Diagram Example



Outcome Space Ω

(disease, positive)

(disease, negative)

(no disease, positive)

(no disease, negative)

Example Event Spaces:

Test is positive or negative

Event Space E

\emptyset

$\{(disease, positive), (no disease, positive)\}$

$\{(disease, negative), (no disease, negative)\}$

Ω

Disease or no disease

Event Space E

\emptyset

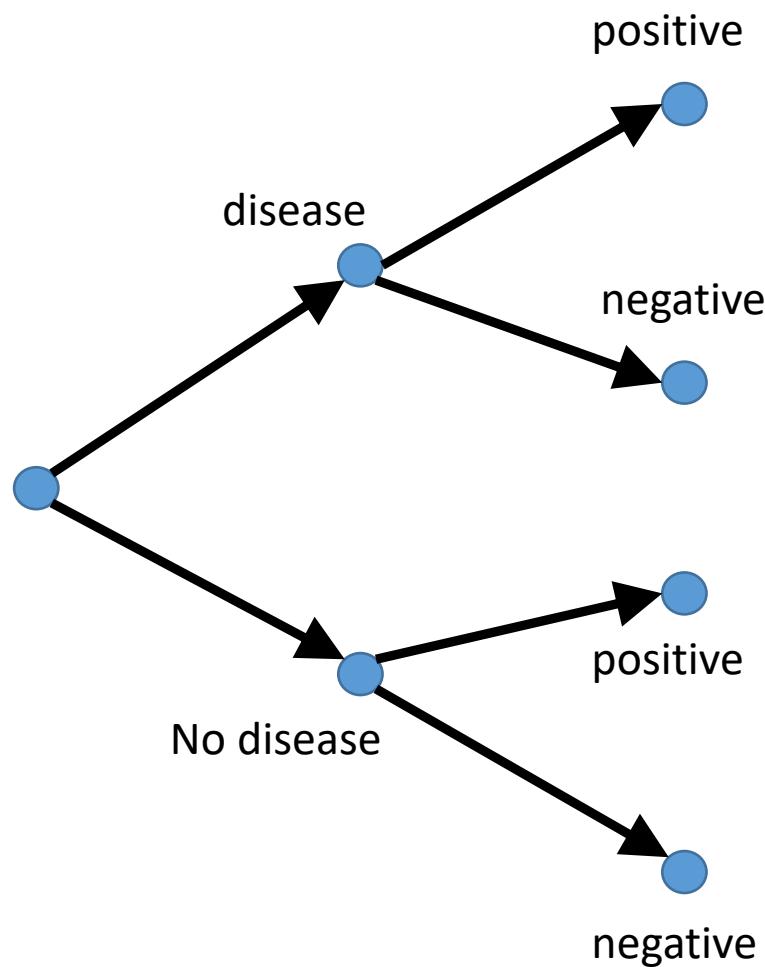
$\{(disease, positive), (disease, negative)\}$

$\{(no disease, positive), (no disease, negative)\}$

Ω

Tree Diagram Example

Example Event Spaces:



Outcome Space Ω

(disease, positive)

(disease, negative)

(no disease, positive)

(no disease, negative)

Event Space E

\emptyset

$\{(disease, positive)\}$

$\{(disease, negative)\}$

$\{(no\ disease, positive)\}$

$\{(no\ disease, negative)\}$

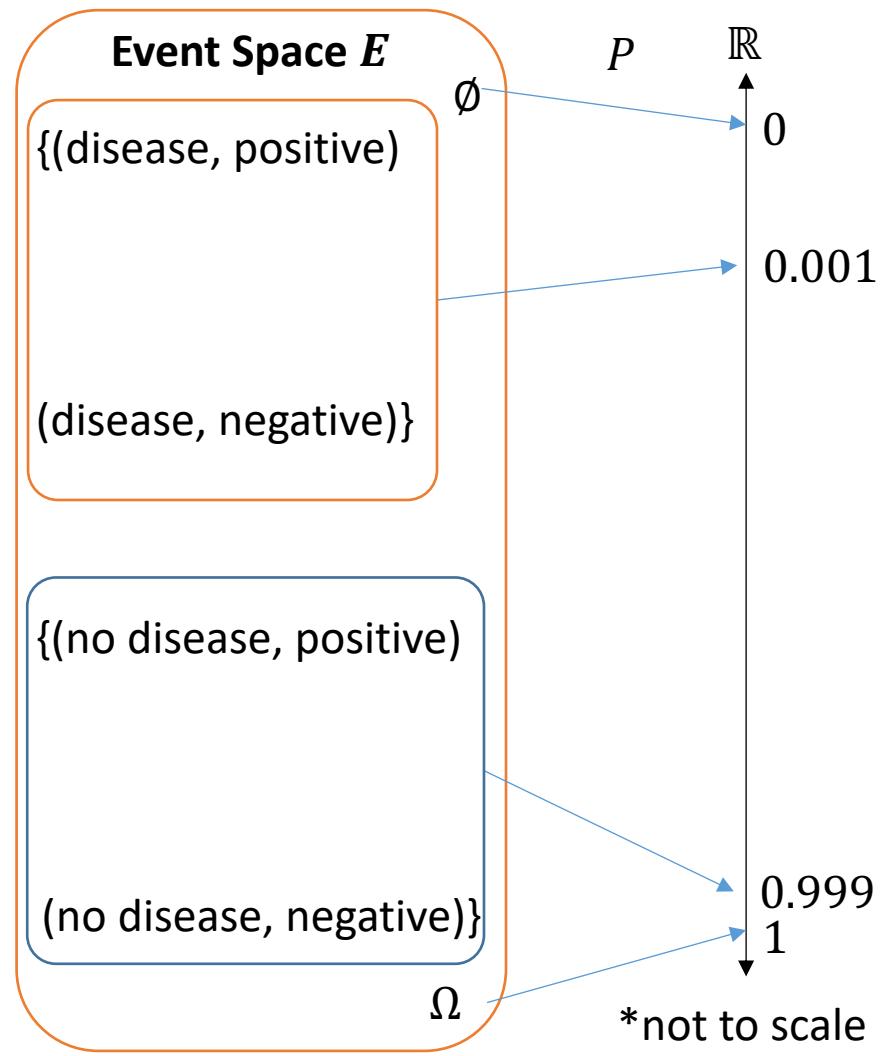
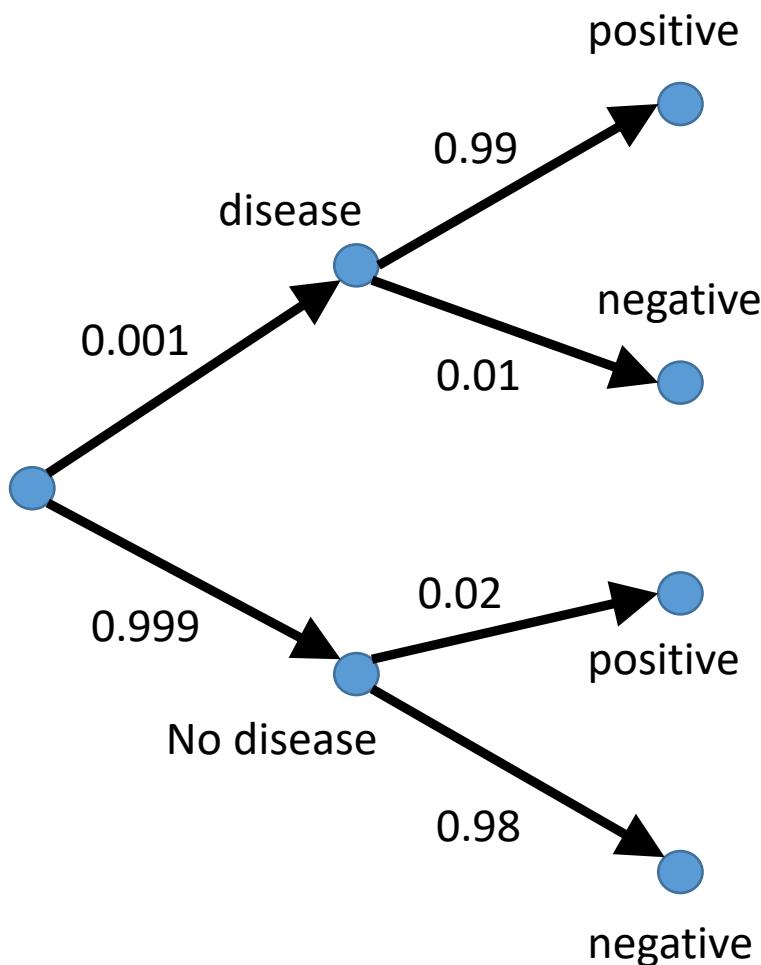
Question: This event space is incomplete.
What other events are missing?

Ω

Probability Distributions

- A probability distribution P over (Ω, E) is a **mapping from events in E to real values** that satisfies the following conditions, i.e. axioms of probability:
 1. **Non-negativity**, i.e. $P(\alpha) \geq 0$, $\forall \alpha \in E$.
 2. Probability of all outcomes **sums to 1**, i.e. $P(\Omega) = 1$.
 3. **Mutually disjoint events**: If $\alpha, \beta \in E$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

Tree Diagram Example



Random Variables

- A random variable, denoted as X (**upper case**), is the formal machinery for **discussing attributes** and their values in different outcomes.
- More formally, it is **a function** $X: \Omega \rightarrow S$ that maps a set of possible **outcomes** Ω to a space S
 - S usually a subset of \mathbb{R} , but can be other sets.

Random Variables: 2 Coin Flips

- Independent, Unbiased Coin Flips
- Possible outcomes $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$
- Random variable to indicate either both heads or both tails

$$X(\omega) = \begin{cases} 1 & \text{if HH or TT} \\ 0 & \text{otherwise} \end{cases}$$

Indicator Random Variables

- **Example:** 3 independent, unbiased coin flips.

$$\Omega = \{ \text{HHH}, \text{TTT}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH} \}$$

$$X(\omega) = \begin{cases} 1 & \text{if HHH or TTT} \\ 0 & \text{otherwise} \end{cases}$$

Exercise: Can you come up with a random variable which represents the *number* of heads? How does it relate to the event space?

Indicator Random Variables

- Indicator random variable maps every outcome to either 0 or 1.
- For example: whether you have the disease

$$\Omega = \{(d, \oplus), (\neg d, \oplus), (d, \ominus), (\neg d, \ominus)\}$$

$$X(\omega) = \begin{cases} 1 & \text{if } (d, \oplus) \\ 1 & \text{if } (d, \ominus) \\ 0 & \text{if } (\neg d, \oplus) \\ 0 & \text{if } (\neg d, \ominus) \end{cases}$$

Random Variables

- The **set of values** that a random variable X can take is denoted as $Val(X)$.
- A lower case letter, e.g. x , is a **generic value** of a random variable X , a.k.a. **realization** of the random variable.
 - E.g.: for an indicator random variable, $x = 1$ or $x = 0$
- The value of a random variable $Val(X)$ can be:
 - **Discrete**, i.e. takes values from a **predefined set**, or
 - **Continuous**, i.e. take values that are **real numbers**.

Random Variables

Examples:

Random variables with discrete values

- Rolling a six-faced die: $Val(X) = \{1, 2, \dots, 6\}$
- Weather conditions: $Val(X) = \{"rain", "cloud", "snow", "sun", "wind"\}$
- Number of people on the next train: $Val(X) = \mathbb{Z}_{\geq 0}$

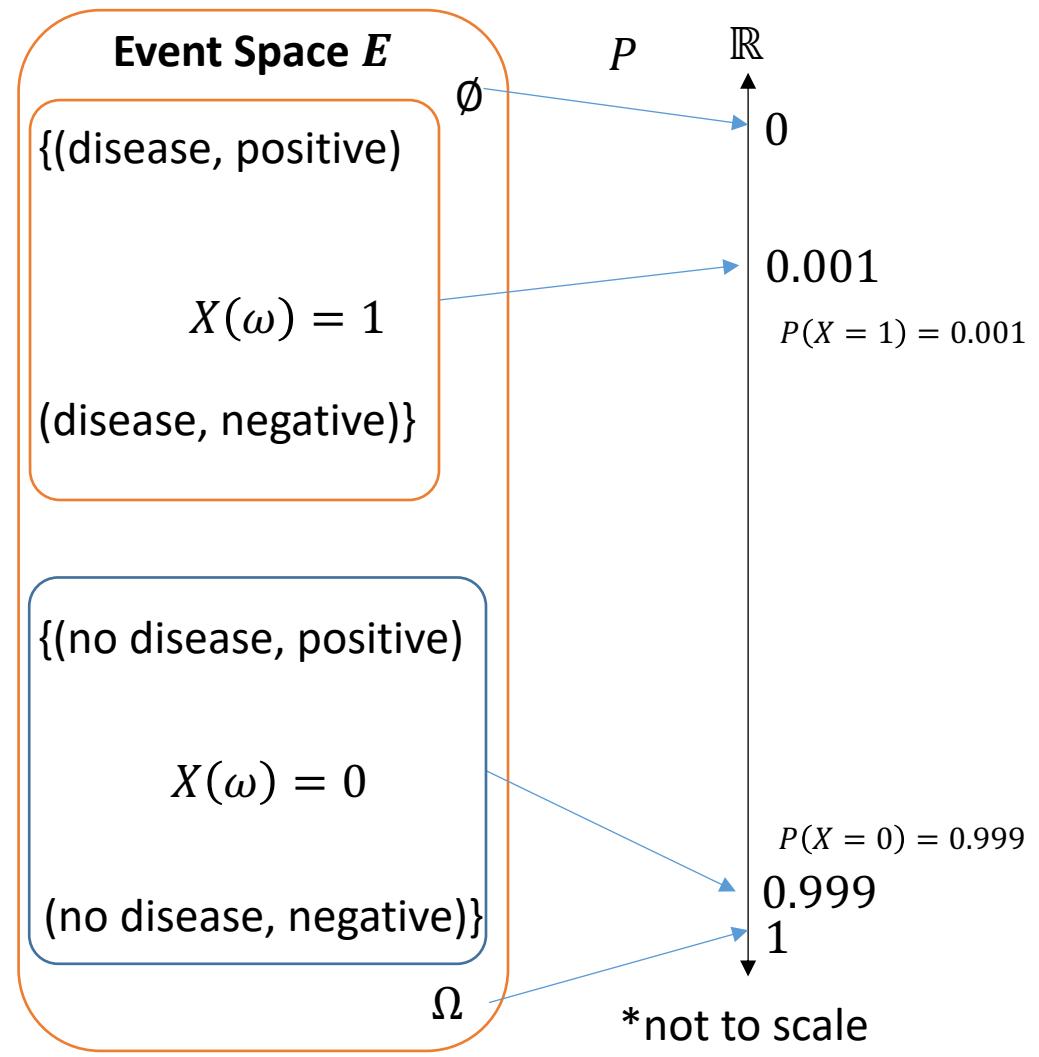
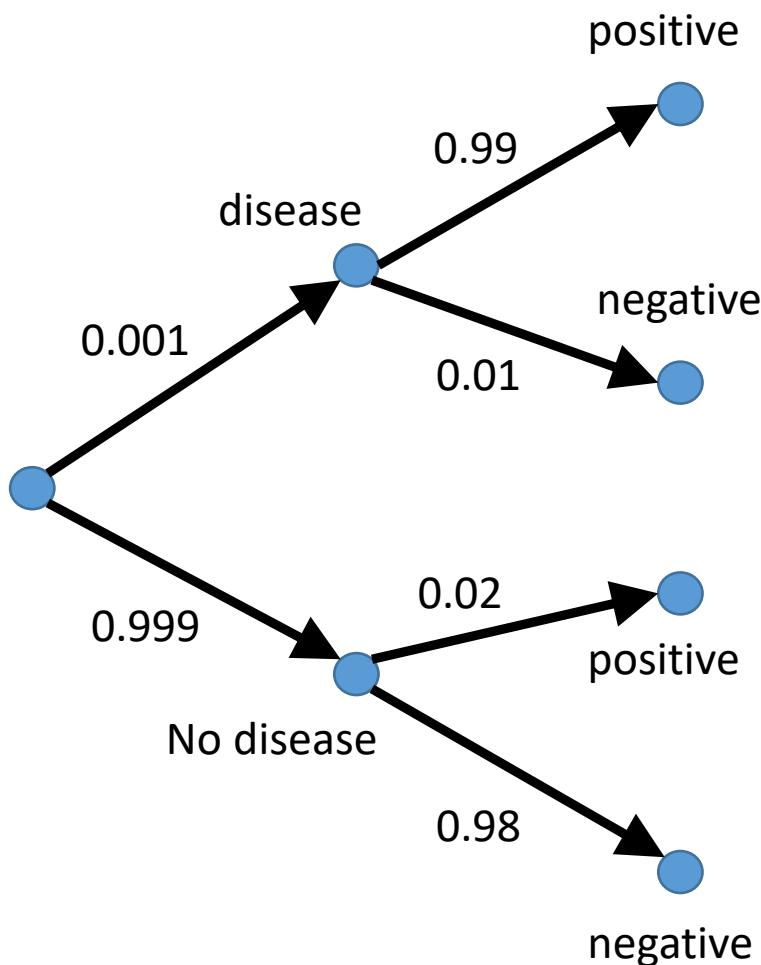
Continuous random variables

- Time taken to finish an exam: $Val(X) = [1, 2]$ hours
- Height of a tree: $Val(X) = \mathbb{R}_{>0}$
- Ambient Temperature: $Val(X) = \mathbb{R}$

Probabilities & Random Variables

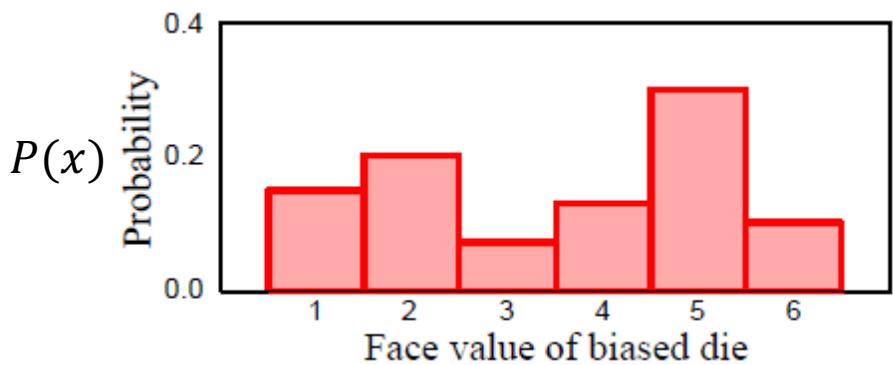
- Interested in the probability of random variables taking on certain values.
 - E.g: the probability of:
 - all heads/tails given 3 independent coin flips
 - the number new COVID-19 patients will be 0
- $P(x)$ is often used as a **shorthand notation** for $P(X = x)$.
 - Recall: x is a **generic value** of a random variable X
- We use the notation x^i to represent a **specific value** of X .

Tree Diagram Example



Probability Distributions: Discrete Vs Continuous

- Discrete: Probability mass function, $P(x)$



$$\sum_{i=1}^K P(X = x^i) = 1$$

$$0 \leq P(X = x^i) \leq 1, \forall i = 1, \dots, K$$

$$K = |Val(X)|$$

$$Val(X) = \{1, 2, 3, 4, 5, 6\}$$

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability Distributions: Discrete Vs Continuous

- For continuous distributions, we have a problem:

$$P(X = x^i) = 0, \forall x^i \in Val(X)$$



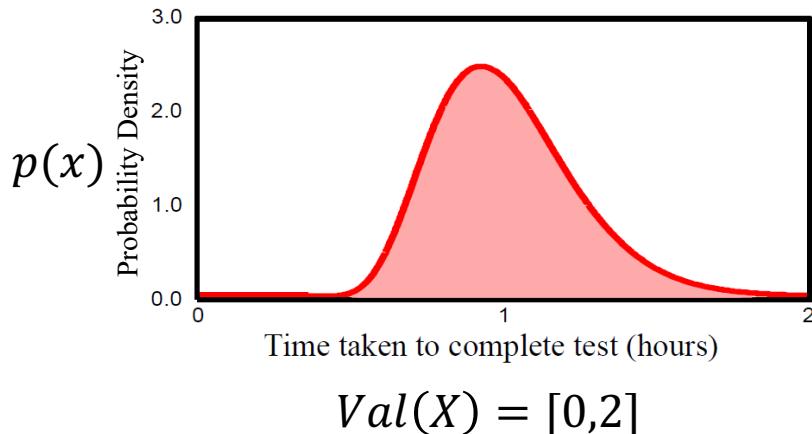
circumference = l
Consider interval
[a, b], what should be
its probability?

$$P([a, b]) \propto \text{length}[a, b]$$
$$P([a, b]) = (b - a)/l$$

Probability Distributions: Discrete Vs Continuous

- Continuous: **Probability density function** is a function (denoted by a lower case p) $p(x): \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$.

$$\int_{Val(X)} p(x)dx = 1 \quad p(X = x^i) \geq 0, \quad \forall x^i \in Val(X)$$



$P(X)$ is the **cumulative function** of X :

$$P(X \leq a) = \int_{-\infty}^a p(x) dx$$

$$P(a \leq X \leq b) = \int_a^b p(x) dx$$

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Back to our Wheel

- What is pdf $p(x)$ if we want uniform?
- Remember: the area under the curve must equal 1

$$\int_{-\infty}^{+\infty} p(x) dx = 1$$

$$\begin{aligned}\int_{-\infty}^{+\infty} p(x) dx &= \int_0^l p(x) dx \\ &= \int_0^l c dx \\ &= cl = 1\end{aligned}$$

So, we have $c = \frac{1}{l}$

circumference = l
What should the
probability density
function be?

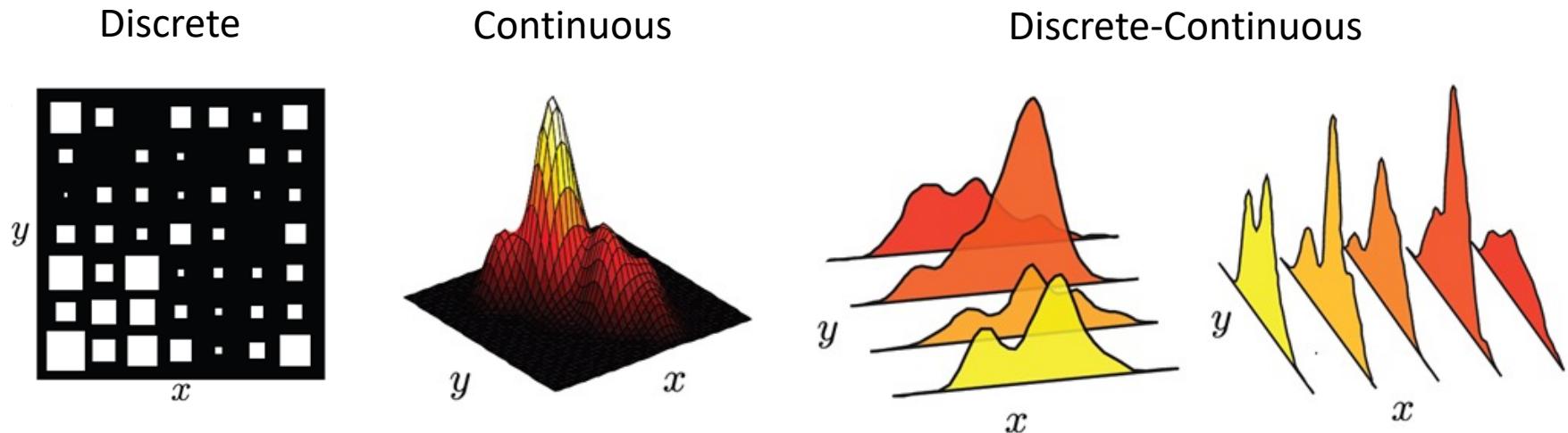
Probability Distributions: Discrete Vs Continuous

In this course, we **abuse notation** by denoting both the probability mass function and probability density function as the lower case $p(x)$

We silently note the property differences in $P(x)$ when X is **discrete or continuous**.

Probability: Joint Probability

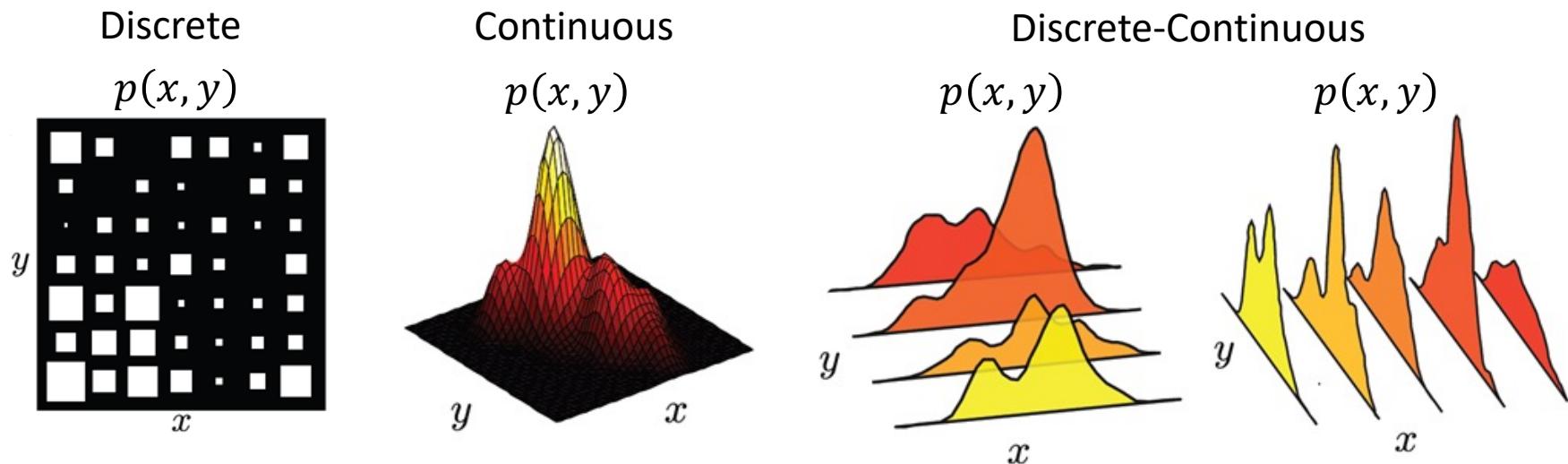
- Consider **all combination** of events of two random variables X and Y .
- Some combinations of outcomes are **more likely** than others.



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

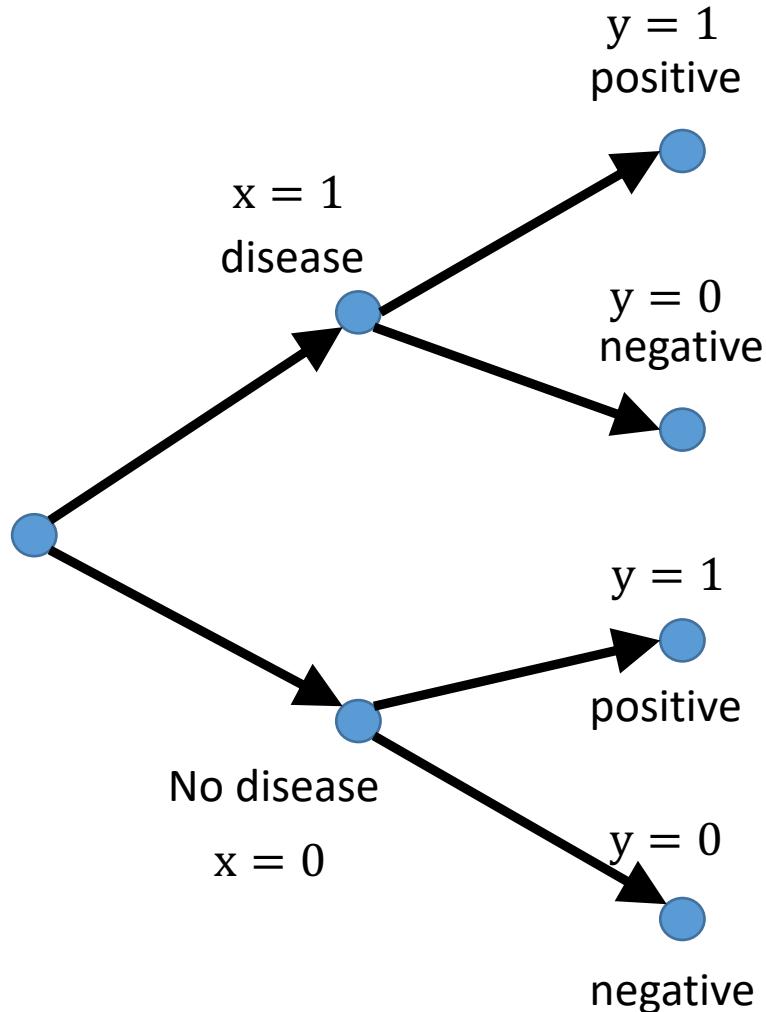
Probability: Joint Probability

- This is captured in the **joint probability** distribution $p(x, y)$.
- Read as “**probability of X and Y** ”.
- Can be **more than two** random variables, i.e. $p(a, b, c, \dots)$.



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Tree Diagram Example



We can now have:

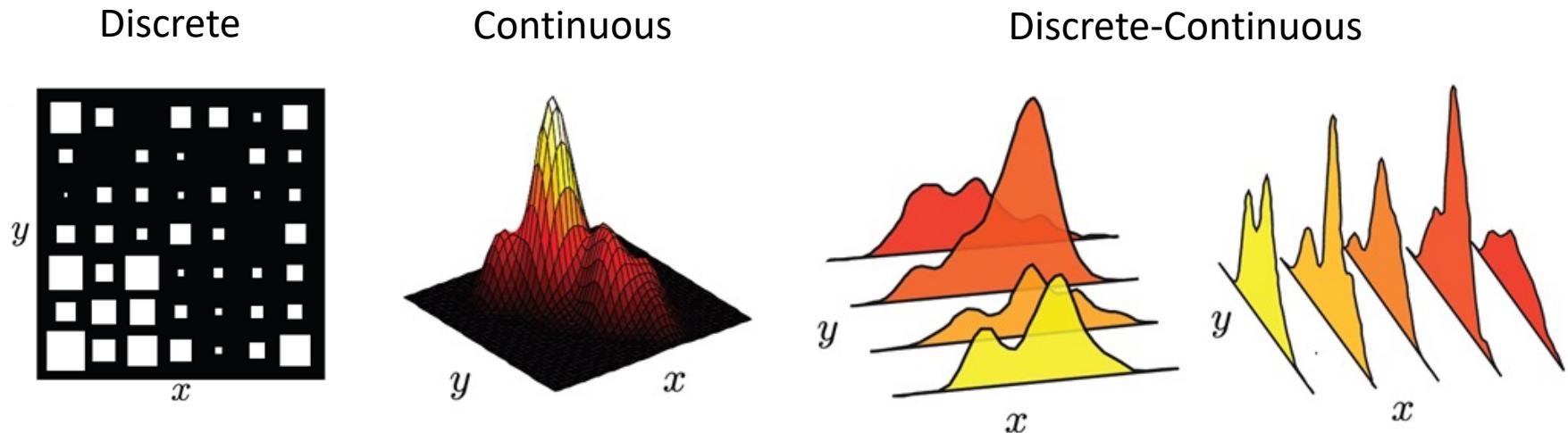
$$p(x, y) = P(X = x \text{ and } Y = y)$$

Basic Operations

*Marginalization, Conditioning, Bayes Rule and
Expectations*

Probability: Joint Probability

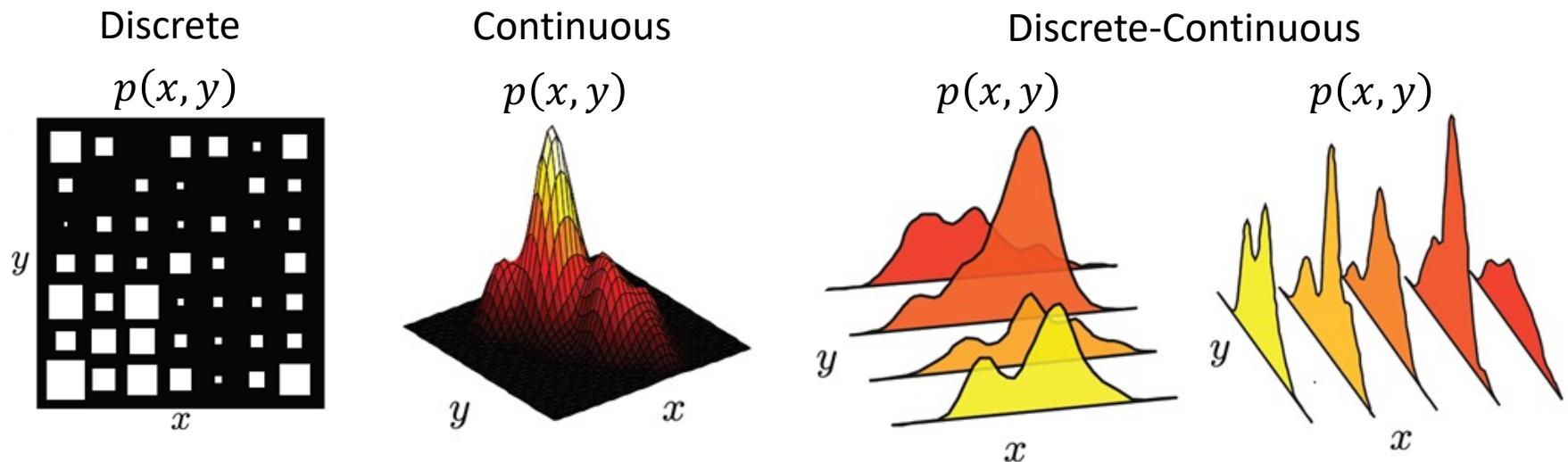
- Consider **all combination** of events of two random variables X and Y .
- Some combinations of outcomes are **more likely** than others.



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Joint Probability

- This is captured in the **joint probability** distribution $p(x, y)$.
- Read as “**probability of X and Y** ”.
- Can be **more than two** random variables, i.e. $p(a, b, c, \dots)$.



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Summary: Sum and Product Rules

- Sum rule:

$$p(x) = \int p(x, y) dy$$
$$p(x) = \sum_y p(x, y)$$

- Product/Chain rule:

$$p(x, y) = p(x|y)p(y)$$

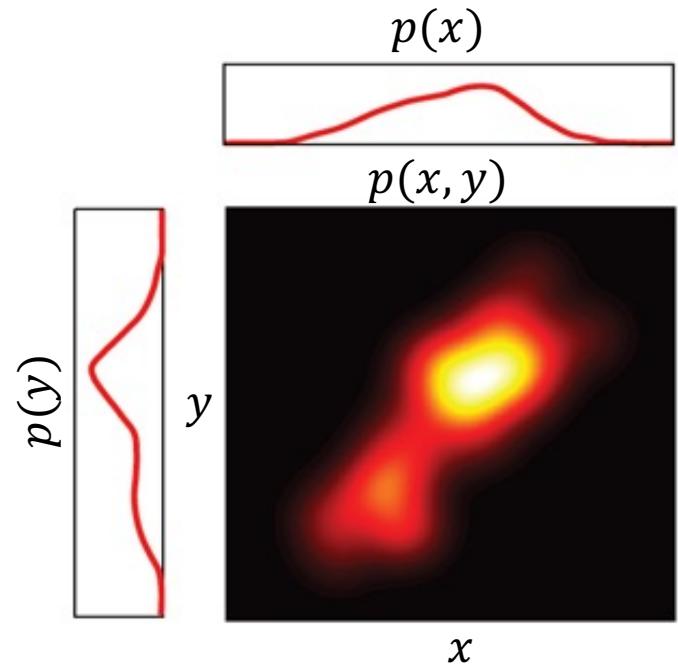
Probability: Marginalization

- Recover probability distribution of any variable in a joint distribution by **integrating (or summing)** over all other variables.
- Also known as the “**sum rule**” of probability.

Continuous:

$$p(x) = \int p(x, y) dy$$

$$p(y) = \int p(x, y) dx$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

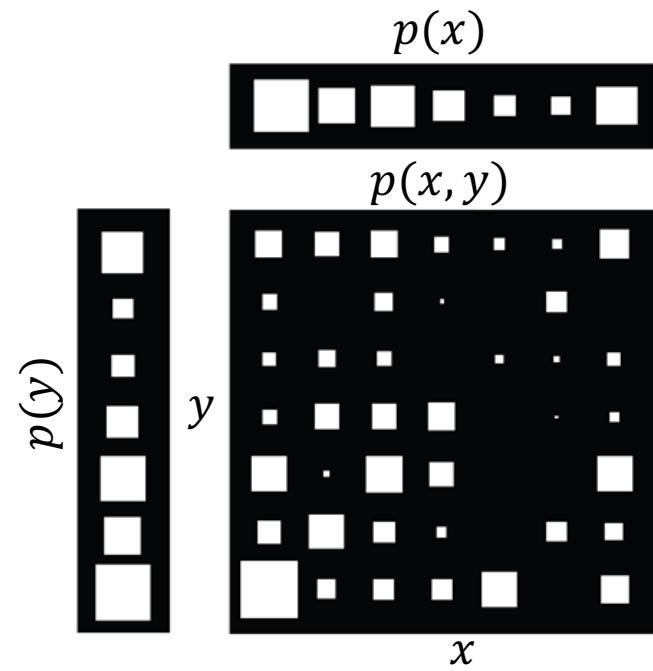
Probability: Marginalization

- Recover probability distribution of any variable in a joint distribution by **integrating (or summing)** over all other variables.
- Also known as the “**sum rule**” of probability.

Discrete:

$$p(x) = \sum_y p(x, y)$$

$$p(y) = \sum_x p(x, y)$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

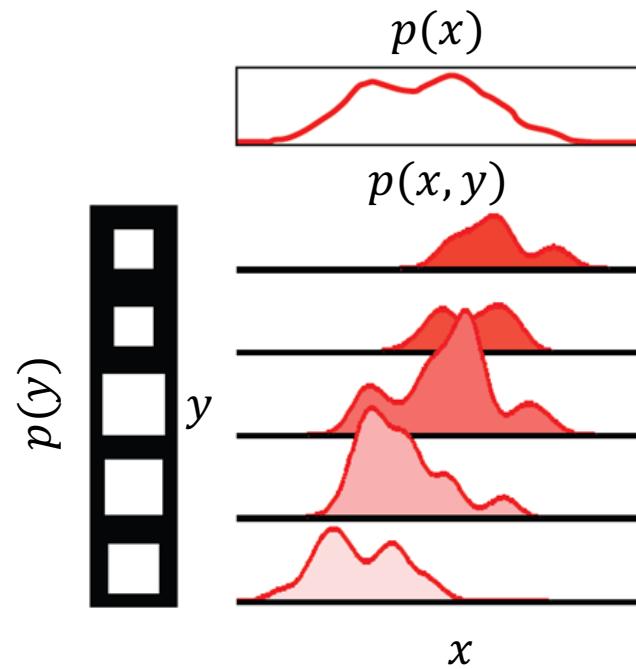
Probability: Marginalization

- Recover probability distribution of any variable in a joint distribution by **integrating (or summing)** over all other variables.
- Also known as the “**sum rule**” of probability.

Discrete-continuous:

$$p(x) = \sum_y p(x, y)$$

$$p(y) = \int p(x, y) dx$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Marginalization

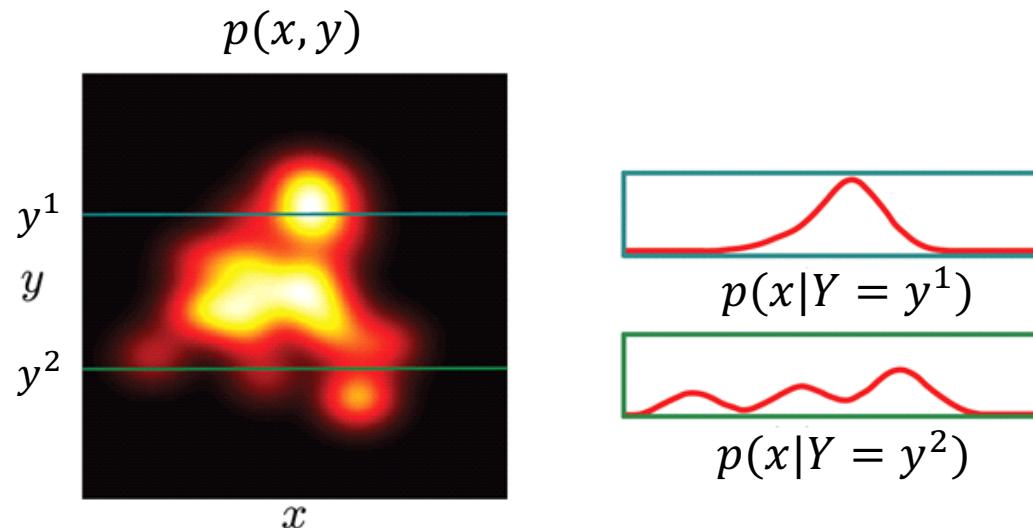
- Works in **higher dimensions** too!

Example:

$$p(x, y) = \sum_w \int p(w, x, y, z) dz$$

Probability: Conditional Probability

- $p(x|Y = y^*)$: “probability of X given $Y = y^*$ ”.
- **Relative propensity** of the random variable X to take different outcomes given that the random variable Y is fixed to value y^* .

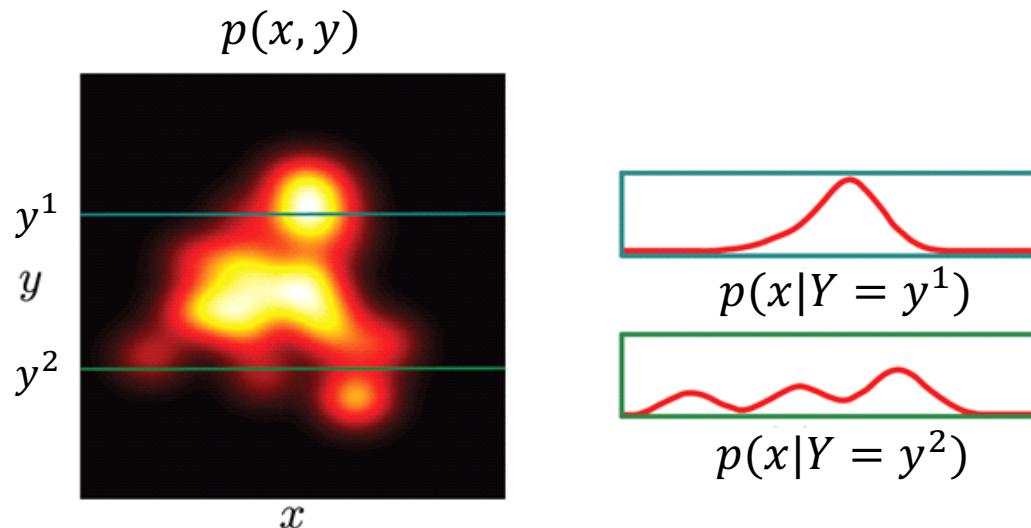


Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Conditional Probability

- Conditional probability can be **extracted from joint probability**.
- Extract appropriate slice and **normalize** (so that the area is 1):

$$P(x|Y = y^*) = \frac{p(x, Y = y^*)}{\int p(x, Y = y^*) dx} = \frac{p(x, Y = y^*)}{p(Y = y^*)}$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Conditional Probability

$$P(x|Y = y^*) = \frac{p(x, Y = y^*)}{\int p(x, Y = y^*) dx} = \frac{p(x, Y = y^*)}{p(Y = y^*)}$$

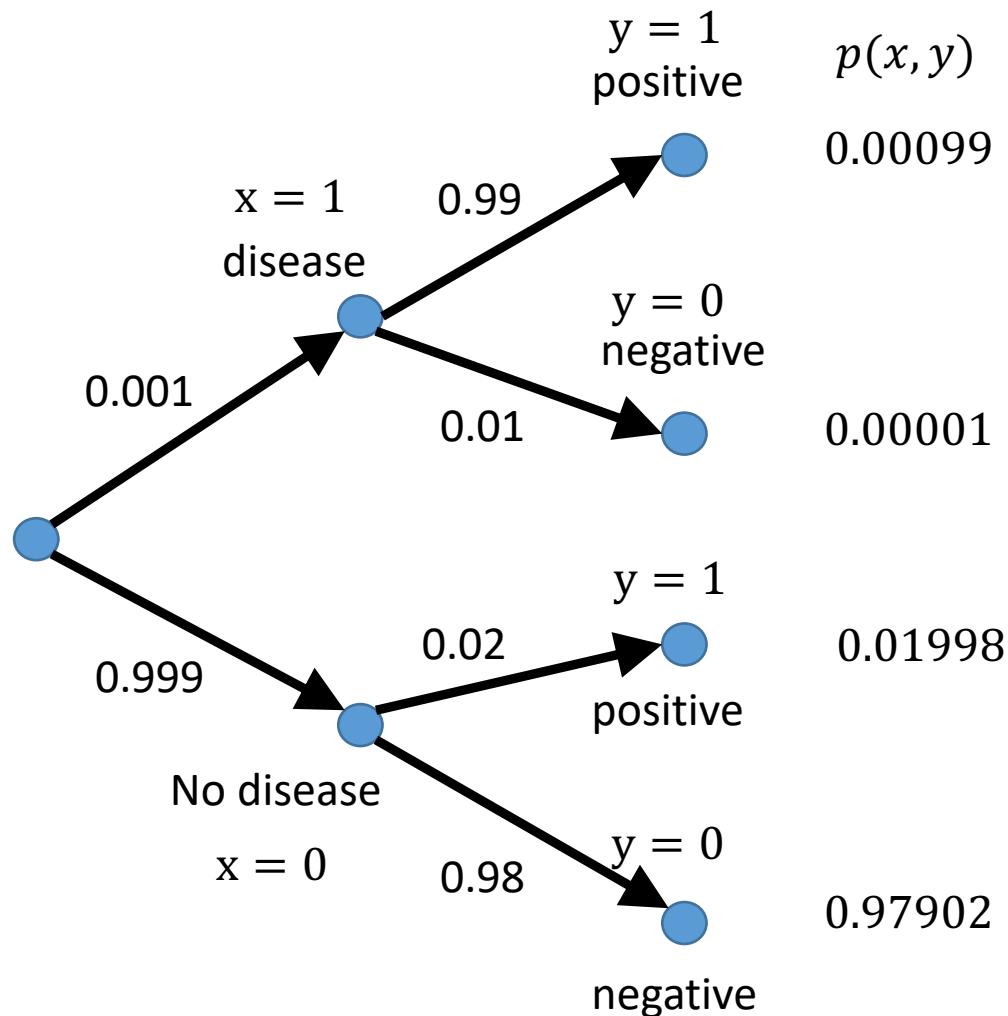
- Usually written in compact form:

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

- Which can be re-arranged to give:

$$\left. \begin{aligned} p(x, y) &= p(x|y)p(y) \\ p(x, y) &= p(y|x)p(x) \end{aligned} \right\} \text{known as "chain rule" or "product rule" of probability.}$$

What is the probability I have the disease given the test is positive?



We want: $p(X = 1 | Y = 1)$

$$\begin{aligned} p(x|Y = 1) &= \frac{p(x, Y = 1)}{p(Y = 1)} \\ &= \frac{p(x, Y = 1)}{\sum_x p(x, Y = 1)} \end{aligned}$$

$$\begin{aligned} p(Y = 1) &= 0.00099 + 0.01998 \\ &= 0.02097 \end{aligned}$$

$$\begin{aligned} p(X = 1|Y = 1) &= 0.00099/0.02097 \\ &= 0.0472 < 5\% \end{aligned}$$

Probability: Conditional Probability

$$p(x, y) = p(x|y)p(y)$$

- Works for **higher dimensions** too!

Example:

$$\begin{aligned} p(w, x, y, z) &= p(w, x, y|z)p(z) \\ &= p(w, x|y, z)p(y|z)p(z) \\ &= p(w|x, y, z)p(x|y, z)p(y|z)p(z) \end{aligned}$$

Summary: Sum and Product Rules

- Sum rule:

$$p(x) = \int p(x, y) dy$$

$$p(x) = \sum_y p(x, y)$$

- Product/Chain rule:

$$p(x, y) = p(x|y)p(y)$$

Probability: Bayes' Rule

- Formulated by Reverend Thomas Bayes in 1763 in “An Essay towards solving a Problem in the Doctrine of Chances”
- Further developed by Laplace and Jeffreys
- “[*Bayes Theorem*] is to the theory of probability what the Pythagorean theorem is to geometry”
 - Sir Harold Jeffreys



Thomas Bayes
1701–1761

Image source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Probability: Bayes' Rule

- Recall:

$$p(x, y) = p(x|y)p(y)$$
$$p(x, y) = p(y|x)p(x)$$



- Eliminating $p(x, y)$, we get:

$$p(y|x)p(x) = p(x|y)p(y)$$

Thomas Bayes
1701–1761

- Rearranging:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x,y)dy} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

Probability: Bayes' Rule

Terminology:

Likelihood – propensity for observing a certain value of X given a certain value of Y

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

Posterior – what we know about Y after observing X

Prior – what we know about Y before seeing X

Evidence – a constant to ensure that the left hand side is a valid distribution

Problem: You're sick!

- You're not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **positive** for *rare, fatal* disease.
 - Disease affects 0.1% of the population.
 - Test correctly identifies 99% of the people who have the disease.
 - If you do not have the disease, test may come back positive 2% of the time.

$$\begin{aligned} p(d|\oplus) &= \frac{p(\oplus|d)p(d)}{p(\oplus)} \\ &= \frac{p(\oplus|d)p(d)}{p(\oplus|d)p(d) + p(\oplus|\neg d)p(\neg d)} \end{aligned}$$

Problem: You're sick!

- Test comes back **positive** for *rare, fatal* disease.
 - Disease affects 0.1% of the population.
 - Test correctly identifies 99% of the people who have the disease.
 - If you do not have the disease, test may come back positive 2% of the time.

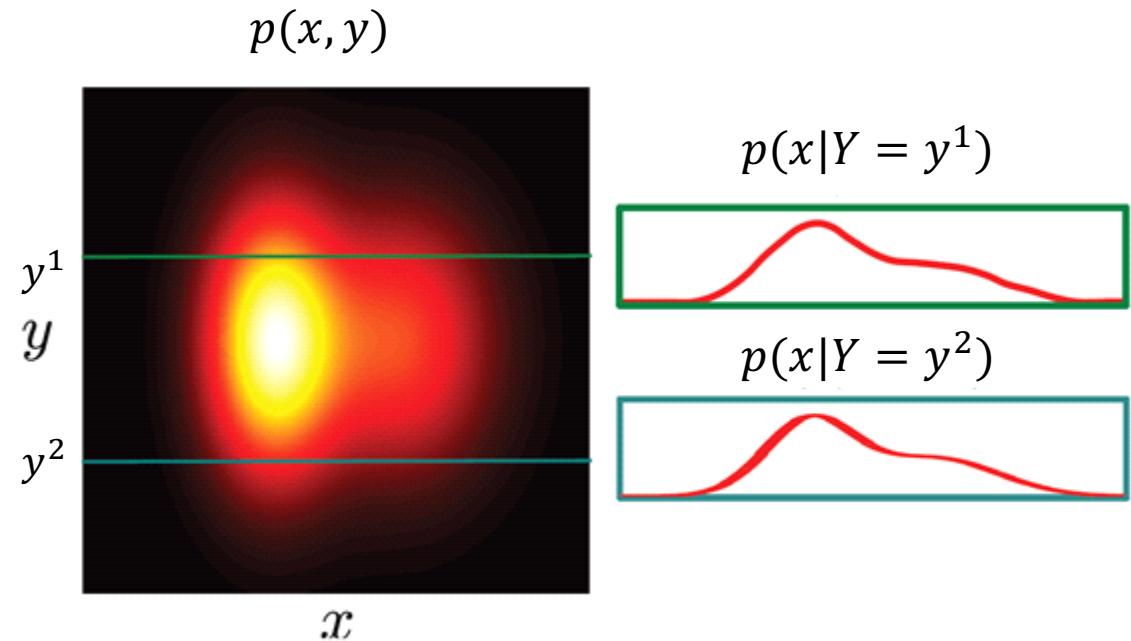
$$\begin{aligned} p(d|\oplus) &= \frac{p(\oplus|d)p(d)}{p(\oplus)} \\ &= \frac{p(\oplus|d)p(d)}{p(\oplus|d)p(d) + p(\oplus|\neg d)p(\neg d)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.02 \times 0.999} = 0.047 < 5\% \end{aligned}$$

Probability: Independence

- The independence of X and Y means that **every conditional distribution is the same**.
- The value of Y **tells us nothing** about X and vice-versa.

$$p(x|y) = p(x)$$

$$p(y|x) = p(y)$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

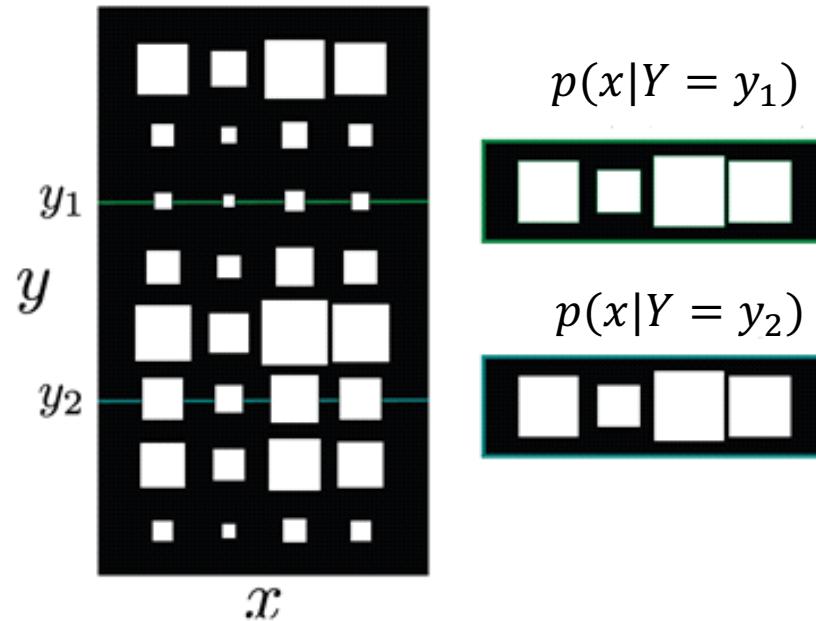
Probability: Independence

- The independence of X and Y means that **every conditional distribution is the same**.
- The value of Y **tells us nothing** about X and vice-versa.

$$p(x, y)$$

$$p(x|y) = p(x)$$

$$p(y|x) = p(y)$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Independence

- When variables are **independent**, the joint factorizes into a **product of the marginals**:

$$\begin{aligned} p(x, y) &= p(x|y)p(y) \\ &= p(x)p(y) \end{aligned}$$

Probability: Expectation

- The **expected or average value** of some function $f[x]$ taking into account the distribution of X .

Definition:

$$E[f[x]] = \sum_x f[x]p(x)$$

$$E[f[x]] = \int f[x]p(x)dx$$

Probability: Rules of Expectation

- **Rule 1:** Expected value of a **constant** is the constant.

$$E[\kappa] = \kappa$$

- **Rule 2:** Expected value of **constant times function** is constant times expected value of function.

$$E[\kappa f[x]] = \kappa E[f[x]]$$

Probability: Rules of Expectation

- **Rule 3:** Expectation of **sum of functions** is sum of expectation of functions.

$$E[f[x] + g[x]] = E[f[x]] + E[g[x]]$$

- **Rule 4:** Expectation of **product of functions in variables X and Y** is product of expectations of functions if X and Y are independent.

$$E[f[x]g[y]] = E[f[x]]E[g[y]],$$

if X and Y are independent

Probability Problems

Dastardly Decoys and Two Envelopes

Problem 1: Dastardly Decoys



A



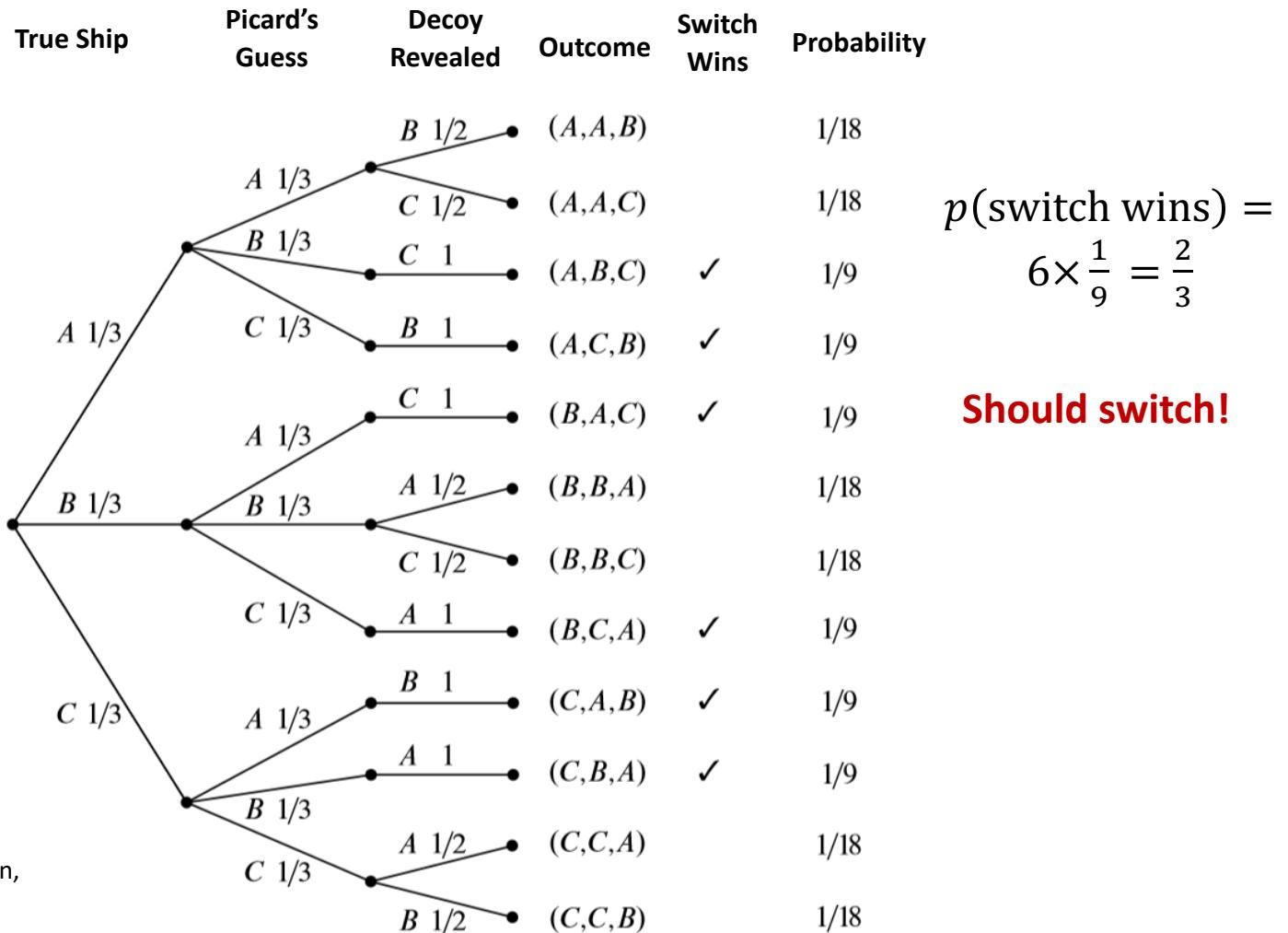
B

Should Picard switch targets?



C

Problem 1: Solution



Adapted from "Math for CS", Lehman, Leighton and Meyer



Monty-Hall Problem in Disguise

The Monty Hall Problem gets its name from the TV game show, *Let's Make A Deal*, hosted by Monty Hall¹. The scenario is such: you are given the opportunity to select one closed door of three, behind one of which there is a prize. The other two doors hide "goats" (or some other such "non-prize"), or nothing at all. Once you have made your selection, Monty Hall will open one of the remaining doors, revealing that it does not contain the prize². He then asks you if you would like to switch your selection to the other unopened door, or stay with your original choice. Here is the problem:

Does it matter if you switch?

This problem is quite interesting, because the answer is felt by most people — including mathematicians — to be counter-intuitive. For most, the “solution” is immediately obvious (they believe), and that is the end of it. But it’s not. Because most of the time, this “obvious” solution is incorrect. The correct solution is quite counterintuitive. Further, I’ve found that many persons have difficulty grasping the validity of the correct solution even

The Monty Hall problem is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show *Let's Make A Deal* and named after its original host, Monty Hall. The problem was originally posed (and solved) in a letter by Steve Selvin to the *American Statistician* in 1975 (Selvin 1975a), (Selvin 1975b). It became famous as a question from a reader's letter quoted in *Marilyn vos Savant's "Ask Marilyn"* column in *Parade* magazine in 1990 (vos Savant 1990a):

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Vos Savant's response was that the contestant should switch to the other door (vos Savant 1990a). Under the standard assumptions, contestants who switch have a $\frac{2}{3}$ chance of winning the car, while contestants who stick to their initial choice have only a $\frac{1}{3}$ chance.

2 Envelopes Game

- Team 1:
 - Pick 2 **different** numbers between 0 and 10.
 - Write each number on a piece of paper each.
 - Turn the papers face down.
- Team 2:
 - Objective is to pick the **larger number**.
 - Pick one of the pieces of paper.
 - Have a peek at the number.
 - **Decide:** do you keep this number or *switch*?
- **Question:** Can Team 2 win more than 50% of the time?

2 Envelopes Game

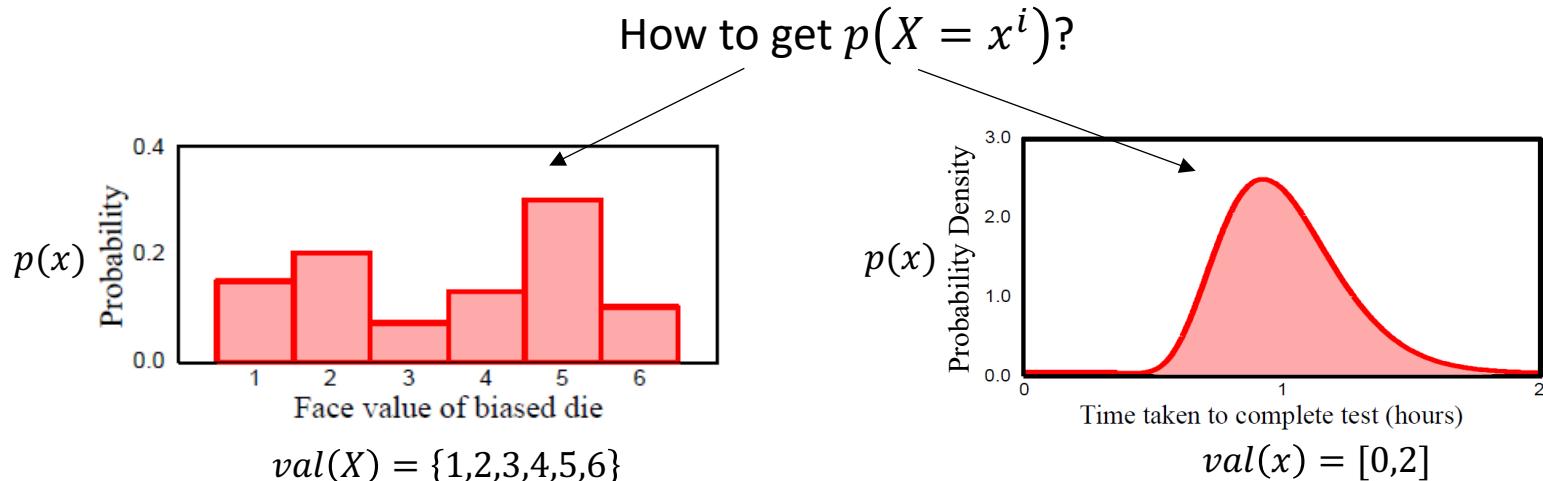
- Team 1:
 - Pick 2 **different** numbers between 0 and 10.
 - Write each number on a piece of paper each.
 - Turn the papers face down.
- Team 2:
 - Objective is to pick the **larger number**.
 - Pick one of the pieces of paper.
 - Have a peek at the number.
 - **Decide:** do you keep this number or *switch*?
- **Question:** Can Team 2 win more than 50% of the time? **Yes! But How and Why? Tutorial Next Week!**

Probability Distributions

Basic distributions and Conjugacy

Probability Distributions

- We have seen the definitions of random variables, probability, and rules for manipulating probabilities.
- Question: “How do we assign the values of $p(X = x^i)$?”



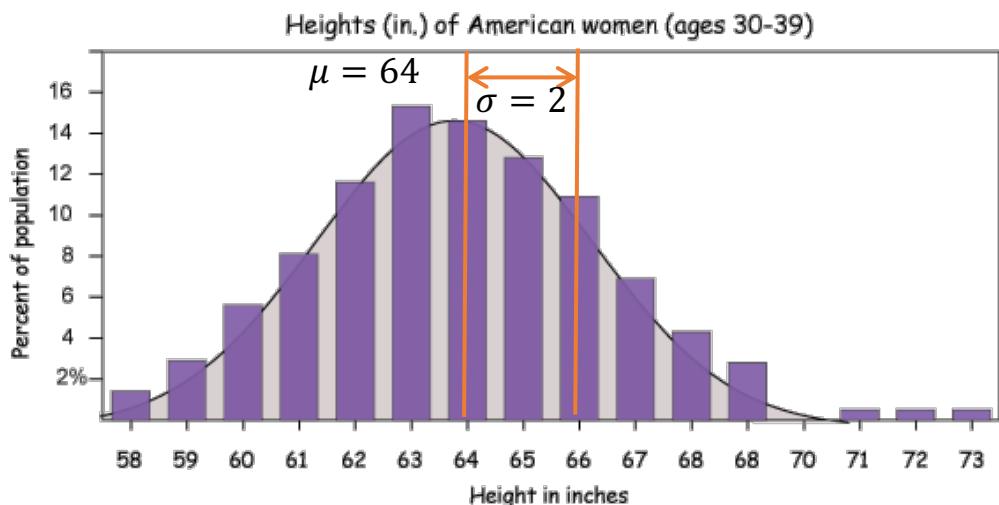
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

“Parametric” Probability Distributions

Q: “How do we assign the probability values?”

A: Use parametric **probability distributions** defined over some fixed set of **parameters**.

Example:



Fitting a Normal distribution to the heights of a population:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

Parameters: mean $\mu = 64$, variance $\sigma^2 = 4$ are learned from data.

Image source: http://www.drcruzan.com/ProbStat_Distributions.html

Common Probability Distributions

- The choice of distribution depends on the **type/domain** of data to be modeled.

Data Type	Domain	Distribution
univariate, discrete, binary	$x \in \{0, 1\}$	Bernoulli
univariate, discrete, multi-valued	$x \in \{1, 2, \dots, K\}$	categorical
univariate, continuous, unbounded	$x \in \mathbb{R}$	univariate normal
univariate, continuous, bounded	$x \in [0, 1]$	beta
multivariate, continuous, unbounded	$\mathbf{x} \in \mathbb{R}^K$	multivariate normal
multivariate, continuous, bounded, sums to one	$\mathbf{x} = [x_1, x_2, \dots, x_K]^T$ $x_k \in [0, 1], \sum_{k=1}^K x_k = 1$	Dirichlet
bivariate, continuous, x_1 unbounded, x_2 bounded below	$\mathbf{x} = [x_1, x_2]$ $x_1 \in \mathbb{R}$ $x_2 \in \mathbb{R}^+$	normal-scaled inverse gamma
multivariate vector \mathbf{x} and matrix \mathbf{X} , \mathbf{x} unbounded, \mathbf{X} square, positive definite	$\mathbf{x} \in \mathbb{R}^K$ $\mathbf{X} \in \mathbb{R}^{K \times K}$ $\mathbf{z}^T \mathbf{X} \mathbf{z} > 0 \quad \forall \mathbf{z} \in \mathbb{R}^K$	normal inverse Wishart

Problem: Infectious Agent

- Patient Zero is loose!
- When he comes into contact with someone, the chance he infects them is:
 $p(\text{Infected upon Contact}) = \lambda = 1/5$
- You know he came into contact with $n = 20$ people.
- You want to model the number of people Patient Zero infected.
- Which probability distribution applies to this scenario?
Assume each contact is *independent*.
 - Common probability distributions and their applications
 - https://en.wikipedia.org/wiki/Probability_distribution#Common_probability_distributions_and_their_applications

Bernoulli Distribution

- **Binary** random variable X , i.e. $x \in \{0,1\}$
- A **single parameter** $\lambda \in [0,1]$.

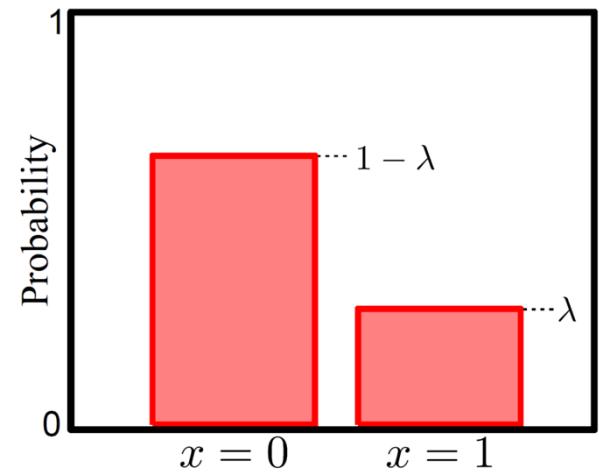
$$\begin{aligned} p(X = 0 | \lambda) &= 1 - \lambda \\ p(X = 1 | \lambda) &= \lambda \end{aligned}$$



Jacob Bernoulli
1654–1705

Or

$$\begin{aligned} p(x) &= \lambda^x(1 - \lambda)^{1-x}, \\ p(x) &= \text{Bern}_x[\lambda] \end{aligned}$$



Example:

X is the outcome of flipping a coin, $X = 1$ represents ‘heads’, and $X = 0$ represents ‘tails’.

Images source: “Pattern Recognition and Machine Learning”, Christopher Bishop
“Computer Vision: Models, Learning, and Inference”, Simon Prince

Binomial Distribution

- **Discrete** random variable X , i.e. $x \in \{0,1,2, \dots, n\}$
- **Two parameters** $n \in \{0,1,2, \dots\}$, $\lambda \in [0,1]$.

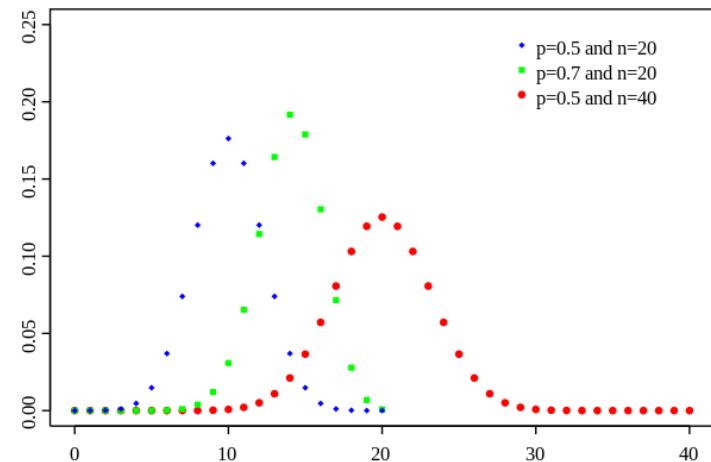
$$p(x) = \binom{n}{x} \lambda^x (1 - \lambda)^{n-x},$$
$$p(x) = \text{Bin}_x[n, \lambda]$$



Jacob Bernoulli
1654–1705

Example:

X is the number of heads when flipping a coin 10 times.



Images source: "https://en.wikipedia.org/wiki/Binomial_distribution", Wikipedia

Categorical Distribution

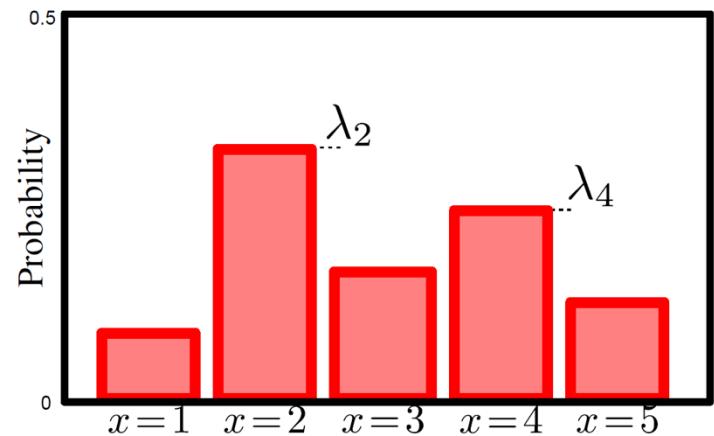
- Discrete variables X that take on **1-of- K possible mutually exclusive states**, e.g. a K -faced die.
- x is represented by a **K -dimensional vector \mathbf{e}_k** in which one of the elements $x_k = 1$, and $\sum_{k=1}^K x_k = 1$.
- e.g. $K = 5$, and $x = \mathbf{e}_3 = [0,0,1,0,0]^T$.
- **K parameters** $\lambda = [\lambda_1, \dots, \lambda_K]^T$, where $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$.

$$p(X = \mathbf{e}_k | \lambda) = \lambda_k$$

Or

$$p(x) = \prod_{k=1}^K \lambda_k^{x_k} = \lambda_k,$$

$$p(x) = \text{Cat}_x[\lambda]$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Univariate Normal Distribution

- Also known as the **Gaussian distribution**.
- Univariate normal distribution describes **single continuous variable** X , i.e. $x \in \mathbb{R}$.
- **Two parameters** $\mu \in \mathbb{R}$ (mean) and $\sigma^2 > 0$ (variance).



Carl Friedrich Gauss
1777–1855

$$p(X = a | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(a-\mu)^2}{2\sigma^2}, \quad a \in \mathbb{R}$$

Or

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}}$$
$$p(x) = \text{Norm}_x[\mu, \sigma^2]$$

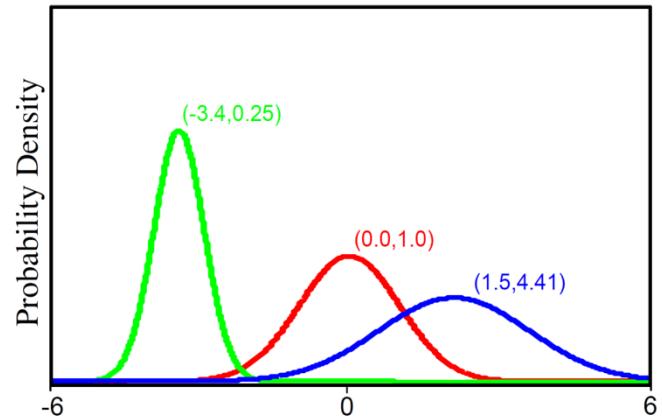


Image sources: "Pattern Recognition and Machine Learning", Christopher Bishop
"Computer Vision: Models, Learning, and Inference", Simon Prince

Multivariate Normal Distribution

- Multivariate normal distribution describes a D -dimensional continuous variable X , i.e. $\mathbf{x} \in \mathbb{R}^D$.
- D -dimensional mean $\boldsymbol{\mu} \in \mathbb{R}^D$, and $D \times D$ symmetrical positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}_+^{D \times D}$.

$$p(\mathbf{X} = \mathbf{a} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\{-0.5(\mathbf{a} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{a} - \boldsymbol{\mu})\}, \quad \mathbf{a} \in \mathbb{R}^D$$

Or

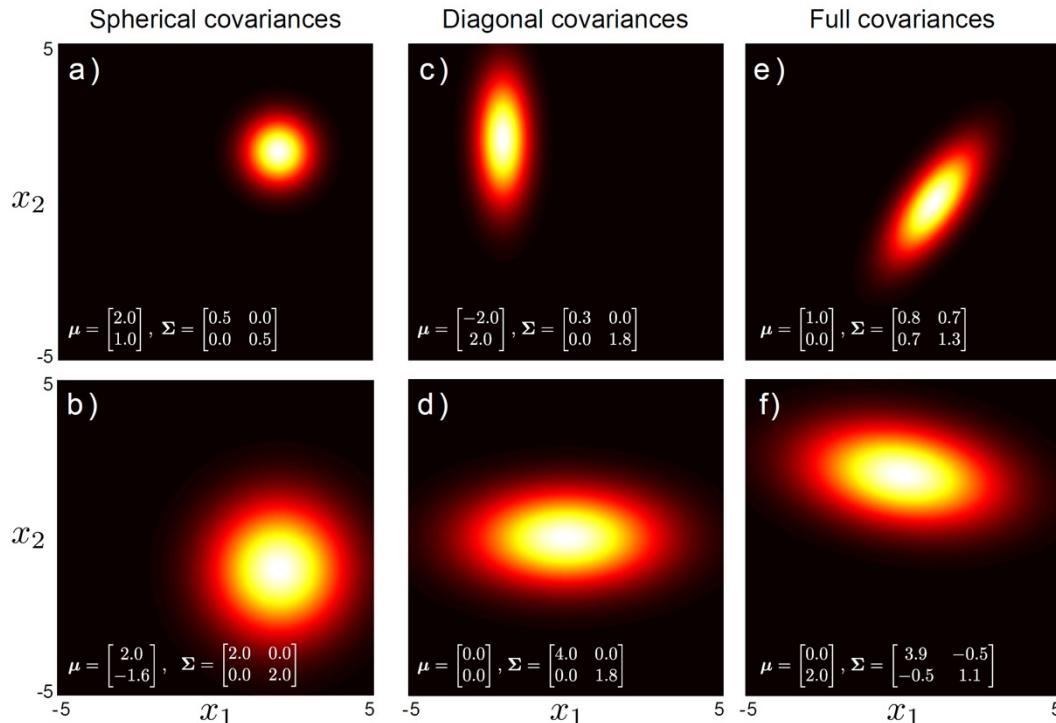
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\{-0.5(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}$$

$$p(\mathbf{x}) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

Types of Covariance

- Covariance matrix has three forms: **spherical**, **diagonal** and **full**.

$$\Sigma_{spher} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \Sigma_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \Sigma_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Problem: Infectious Agent

- Patient Zero is loose!
- When he comes into contact with someone, the chance he infects them is:
 $p(\text{Infected upon Contact}) = \lambda = 1/5$ 
- Where did this number come from?
- You know he came into contact with $n = 20$ people.
- You want to model the number of people Patient Zero infected.
- Which probability distribution applies to this scenario?
Assume each contact is *independent*.
 - Common probability distributions and their applications
 - https://en.wikipedia.org/wiki/Probability_distribution#Common_probability_distributions_and_their_applications

Learning with Bayes' Rule

Likelihood – propensity for observing the data given a certain parameter θ

Prior – what we know about θ before seeing D

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

Posterior – what we know about θ after observing the data D

Conjugate Distributions

- Conjugate distributions can be used to **model the parameters** of probability distributions.
- **Product** of a probability distribution and its conjugate has the **same form** as the conjugate **times a constant**.
- Parameters of conjugate distributions are known as **hyperparameters** because they control the parameter distributions.
- List of conjugate distributions: https://en.wikipedia.org/wiki/Conjugate_prior

Distribution	Domain	Parameters modeled by
Bernoulli	$x \in \{0, 1\}$	beta
categorical	$x \in \{1, 2, \dots, K\}$	Dirichlet
univariate normal	$x \in \mathbb{R}$	normal inverse gamma
multivariate normal	$\mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Importance of Conjugate Distributions

Learning the parameters θ of a parametric probability distribution:

Recall the Bayes' Rule:

1. Choose prior that is conjugate to likelihood

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

2. The posterior will have **same form** as conjugate prior distribution, i.e. **closed-form**.

Problem: Infectious Agent

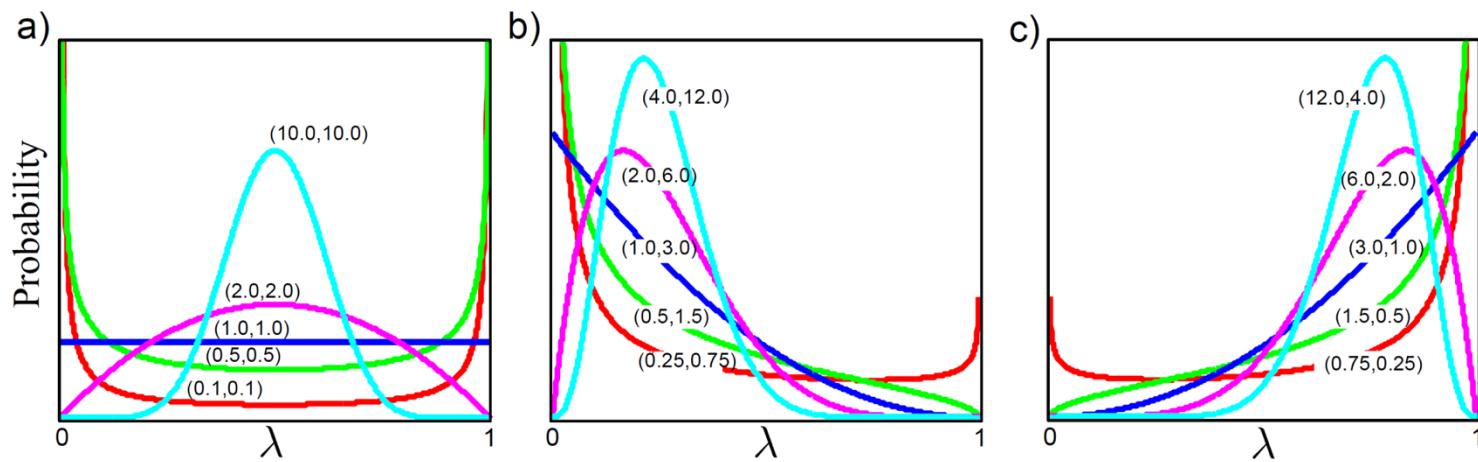
- Patient Zero is loose!
- You want to model the uncertainty over the binomial/Bernoulli parameter
 $\lambda = p(\text{Infected upon Contact})$
- What is the conjugate distribution for λ ?
Assume each contact is *independent*.
 - Conjugate Distributions:
https://en.wikipedia.org/wiki/Conjugate_prior

Conjugate Distribution: Beta Distribution

- Conjugate distribution of **Bernoulli distribution**.
- Defined over parameter of the Bernoulli distribution $\lambda \in [0,1]$.

$$p(\lambda) = \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1}$$

$$p(\lambda) = \text{Beta}_\lambda[\alpha, \beta]$$

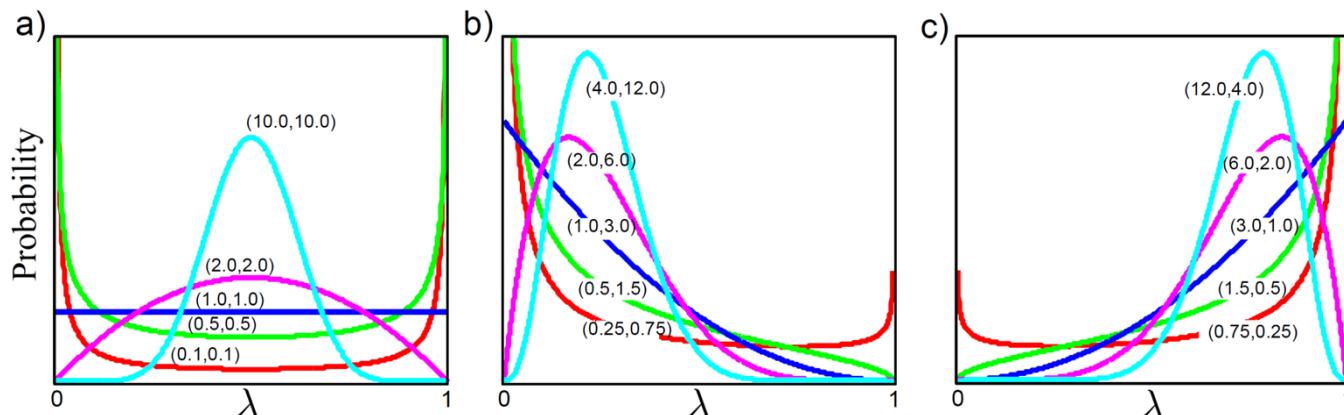


Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Conjugate Distribution: Beta Distribution

$$\begin{aligned} p(\lambda) &= \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1} \\ p(\lambda) &= \text{Beta}_\lambda[\alpha, \beta] \end{aligned}$$

- Two hyperparameters $\alpha, \beta > 0$.



Gamma Function:

$$\begin{aligned} \Gamma(z) &= \int_0^\infty t^{z-1} e^{-t} dt, \quad z \in \mathbb{C} \\ \Gamma(n) &= (n-1)! , \quad n \in \mathbb{Z}_{>0} \end{aligned}$$

$$\begin{aligned} B(\alpha, \beta) &= \frac{\Gamma[\alpha]\Gamma[\beta]}{\Gamma[\alpha + \beta]} \\ &= \int_{t=0}^1 t^{\alpha-1} (1-t)^{\beta-1} \end{aligned}$$

Tutorial: Beta-Binomial

- Show that the Beta distribution is conjugate to the Binomial distribution.

Bernoulli

$$p(x) = \lambda^x(1 - \lambda)^{1-x},$$

$$p(x) = \text{Bern}_x[\lambda]$$

Binomial

$$p(x) = \binom{n}{x} \lambda^x (1 - \lambda)^{n-x},$$

$$p(x) = \text{Bin}_x[n, \lambda]$$

Beta

$$\begin{aligned} p(\lambda) &= \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1} \end{aligned}$$

$$p(\lambda) = \text{Beta}_\lambda[\alpha, \beta]$$

Also:

$$\begin{aligned} B(\alpha, \beta) &= \frac{\Gamma[\alpha]\Gamma[\beta]}{\Gamma[\alpha + \beta]} \\ &= \int_{t=0}^1 t^{\alpha-1} (1 - t)^{\beta-1} \end{aligned}$$

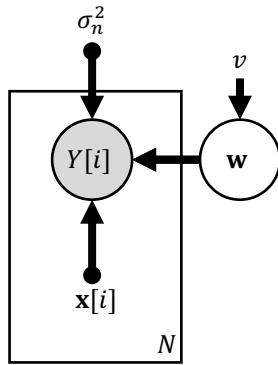
Learning Outcomes

Students should be able to:

1. Describe uncertain quantities with **random variables** and **joint probabilities**.
2. Explain the basic rules of probability – **sum, product, Bayes', independence and expectation** rules.
3. Use the common probabilities distributions – **Bernoulli, categorical, univariate and multivariate normal** distributions.
4. Explain the use of **conjugate distributions**.

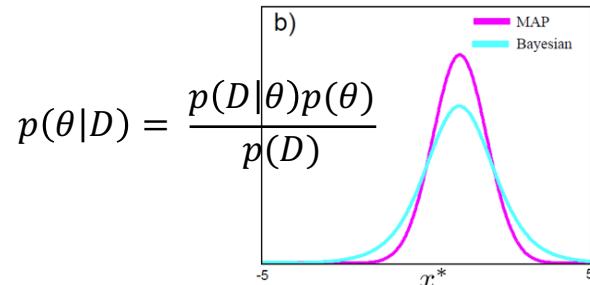
CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with **uncertainty** in a computer.



Representation: The *language* is probability and **probabilistic graphical models** (PGM).

The language is used to **model problems**.



Reasoning: We use **learning** and **inference algorithms** to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

Appendix

More Conjugate Distributions

Dirichlet Distribution

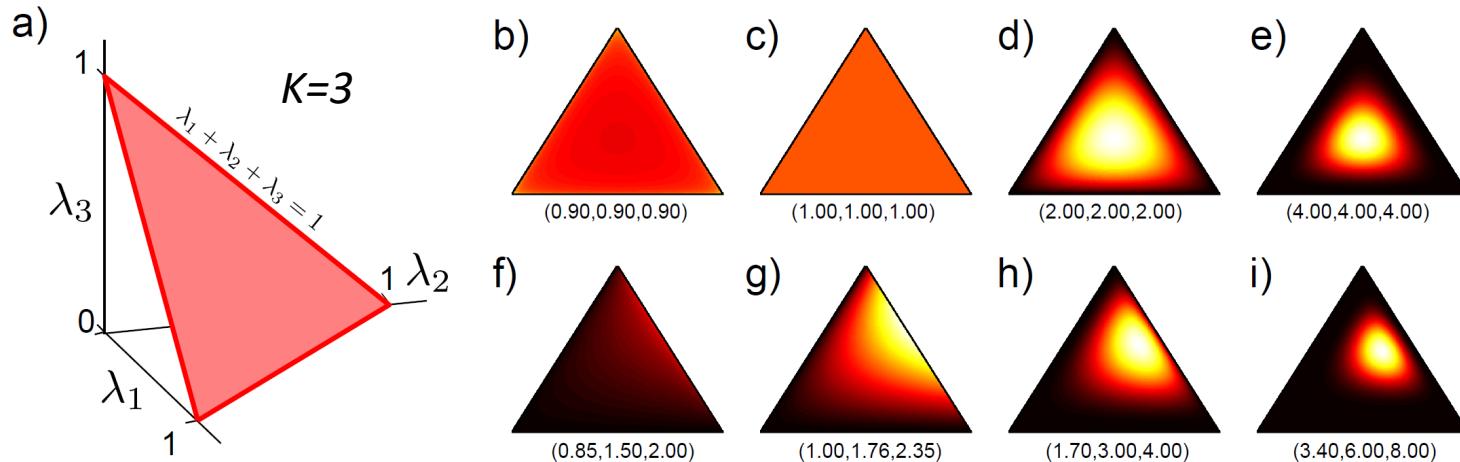
- Conjugate distribution of **categorical distribution**.
- Defined over K parameters of Categorical distribution, $\lambda_k \in [0,1]$, where $\sum_k \lambda_k = 1$.

$$p(\lambda_1, \dots, \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k-1},$$

$$p(\lambda_1, \dots, \lambda_K) = \text{Dir}_{\lambda_1 \dots K} [\alpha_1, \dots \alpha_K]$$



Peter Gustav Lejeune Dirichlet
(1805-1859)

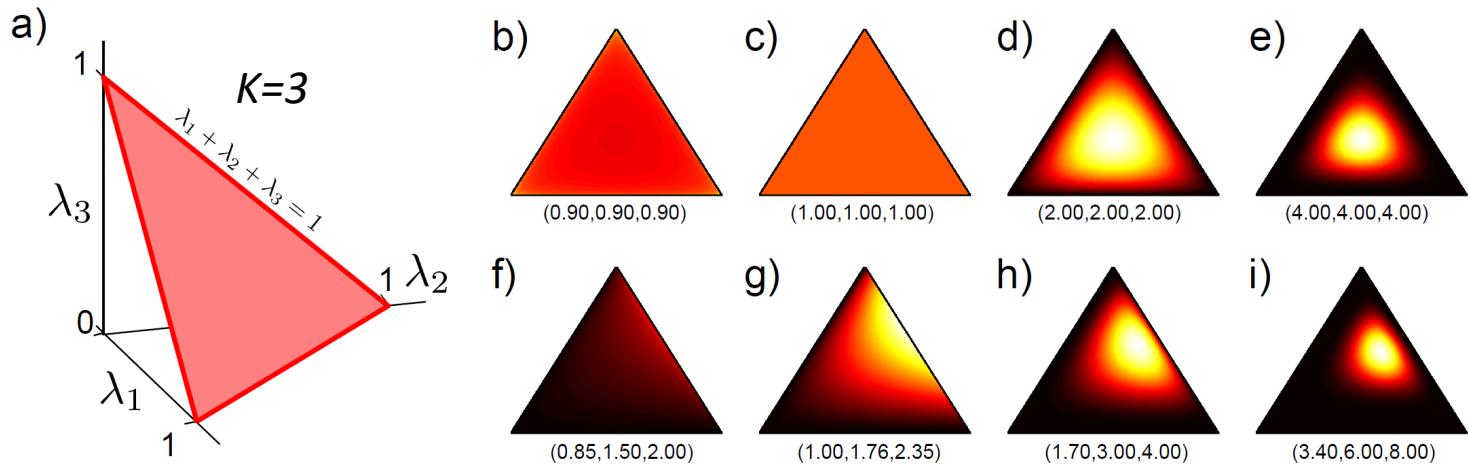


Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince
<http://www.amt.edu.au/biogdirichlet.html>

Dirichlet Distribution

$$p(\lambda_1, \dots, \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k-1},$$
$$p(\lambda_1, \dots, \lambda_K) = \text{Dir}_{\lambda_1 \dots K}[\alpha_1, \dots \alpha_K]$$

- K hyperparameters $\alpha_k > 0$.



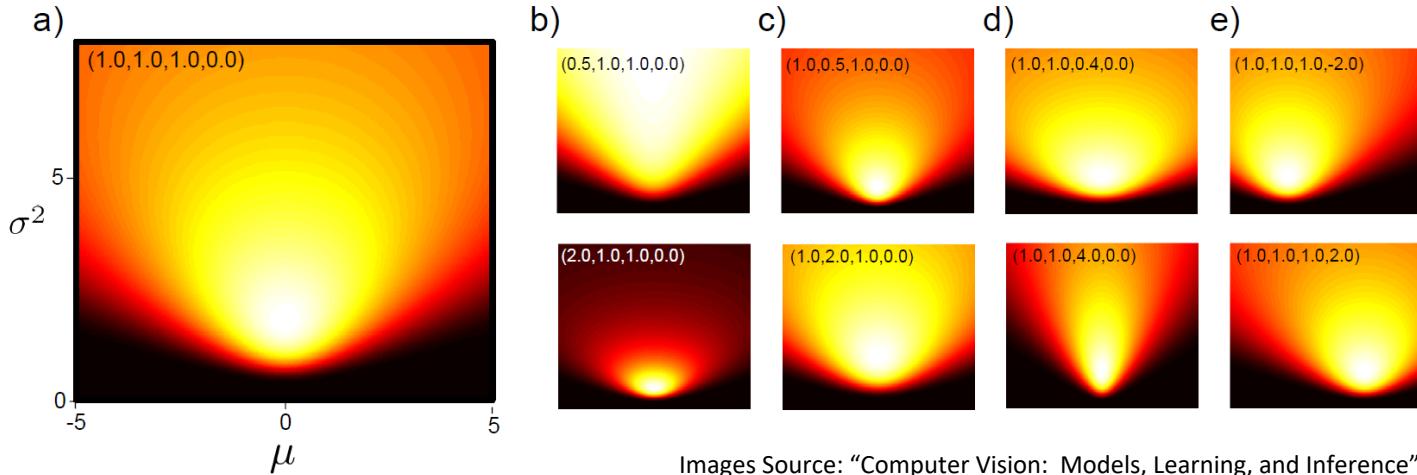
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Normal Inverse Gamma Distribution

- Conjugate distribution of **univariate normal distribution**.
- Defined on parameters $\mu, \sigma^2 > 0$ of univariate normal distribution.

$$p(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right]$$

$$p(\mu, \sigma^2) = \text{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta]$$



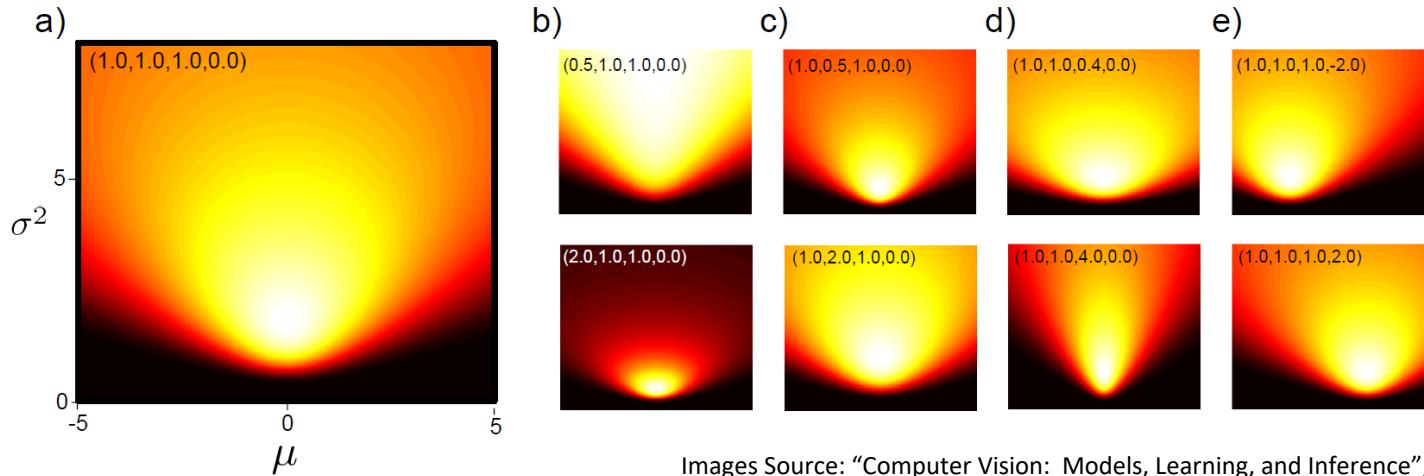
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Normal Inverse Gamma Distribution

$$p(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right]$$

$$p(\mu, \sigma^2) = \text{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta]$$

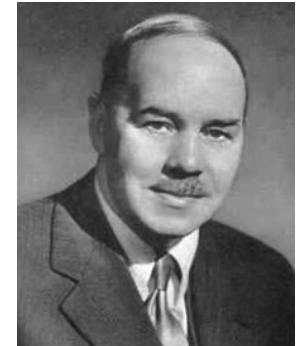
- Four hyperparameters $\alpha, \beta, \gamma > 0$ and $\delta \in \mathbb{R}$.



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Normal Inverse Wishart

- Conjugate distribution of **multivariate normal distribution**.
- Defined on parameters μ, Σ of multivariate normal distribution.



John Wishart
(1898-1956)

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\gamma^{D/2} |\boldsymbol{\Psi}|^{\alpha/2} \exp[-0.5 (\text{Tr} [\boldsymbol{\Psi} \boldsymbol{\Sigma}^{-1}] + \gamma(\boldsymbol{\mu} - \boldsymbol{\delta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\delta}))]}{2^{\alpha D/2} (2\pi)^{D/2} |\boldsymbol{\Sigma}|^{(\alpha+D+2)/2} \Gamma_D[\alpha/2]}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NorIWis}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}]$$

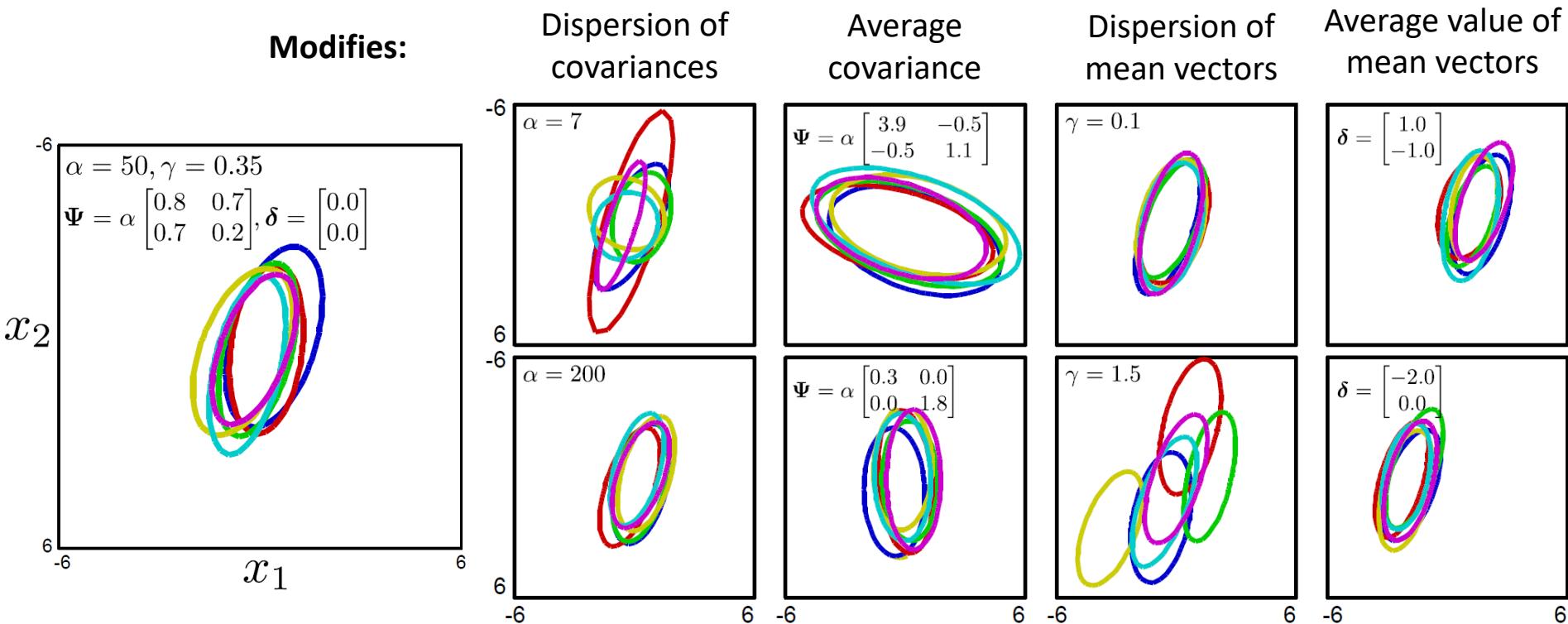
- **Four hyperparameters:** a positive scalar α , a positive definite matrix $\boldsymbol{\Psi} \in \mathbb{R}_+^{D \times D}$, a positive scalar γ , and a vector $\boldsymbol{\delta} \in \mathbb{R}^D$.

Multivariate gamma function:

$$\Gamma_D[a] = \pi^{a(a-1)/4} \prod_{j=1}^a \Gamma[a + (1-j)/2]$$

Normal Inverse Wishart

- Samples from Normal Inverse Wishart:



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince