Modeling

Chris Piech
CS109, Stanford University

Midterm Location by Last Name

CEMEX:

A through K

Hewlett 200:

L through S

Hewlett 201:

T through Z

Look out for your assignment! Let us know if you are at risk

Stanford University

No Class on Monday

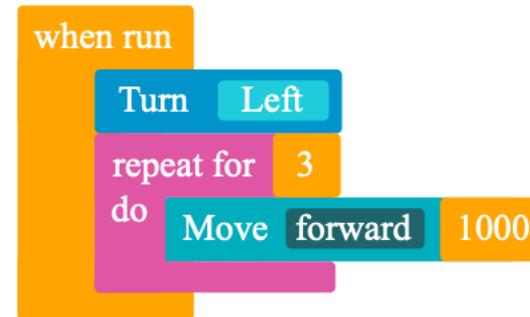This is how I imagine students are when they are studying ☺

(recall why we are here. Learning is a joy. You are growing)

Computers Couldn't Understand Student Code

60,000 students attempted this problem
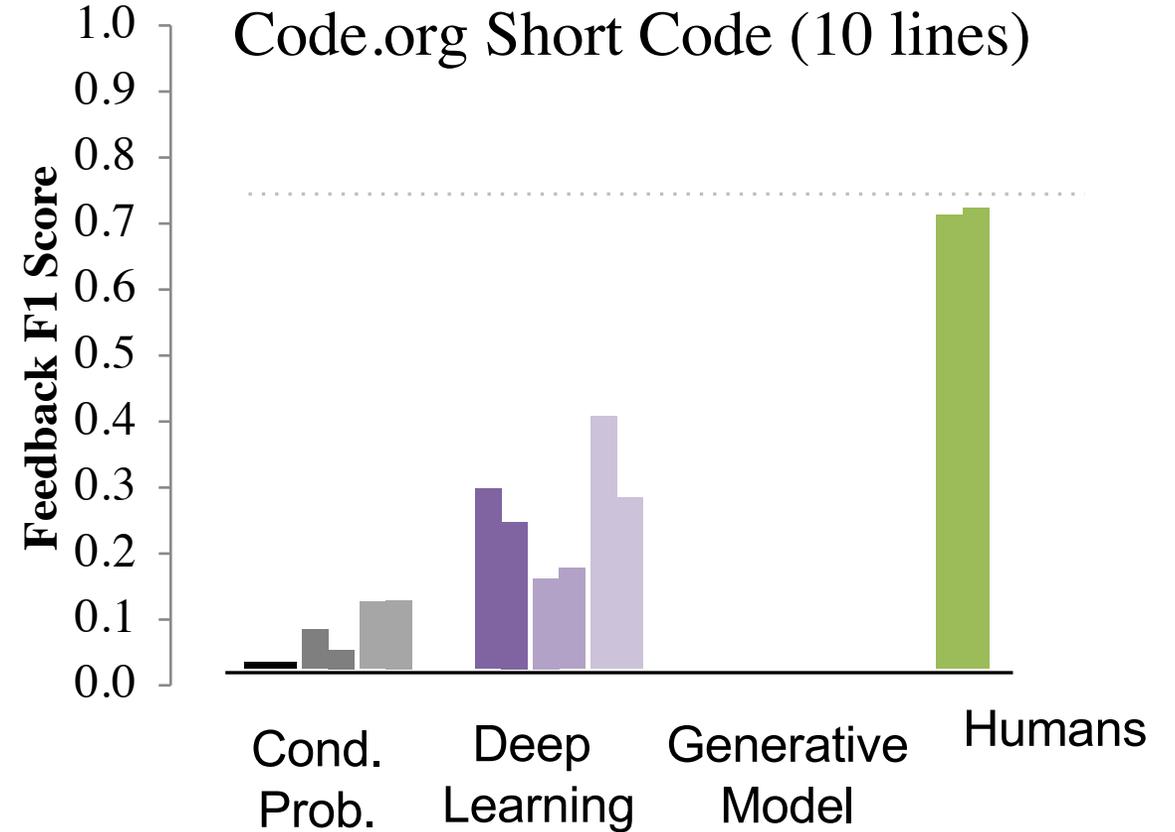37,000 unique solutions

Challenge

when run
Turn   Left
repeat for   3
do   Move   forward   1000

Student Code

You need to move and turn in your loop

Insight

Computers Couldn't Understand Student Code

Code.org Short Code (10 lines)

Generative Model of Characters

Lake et al, 2015
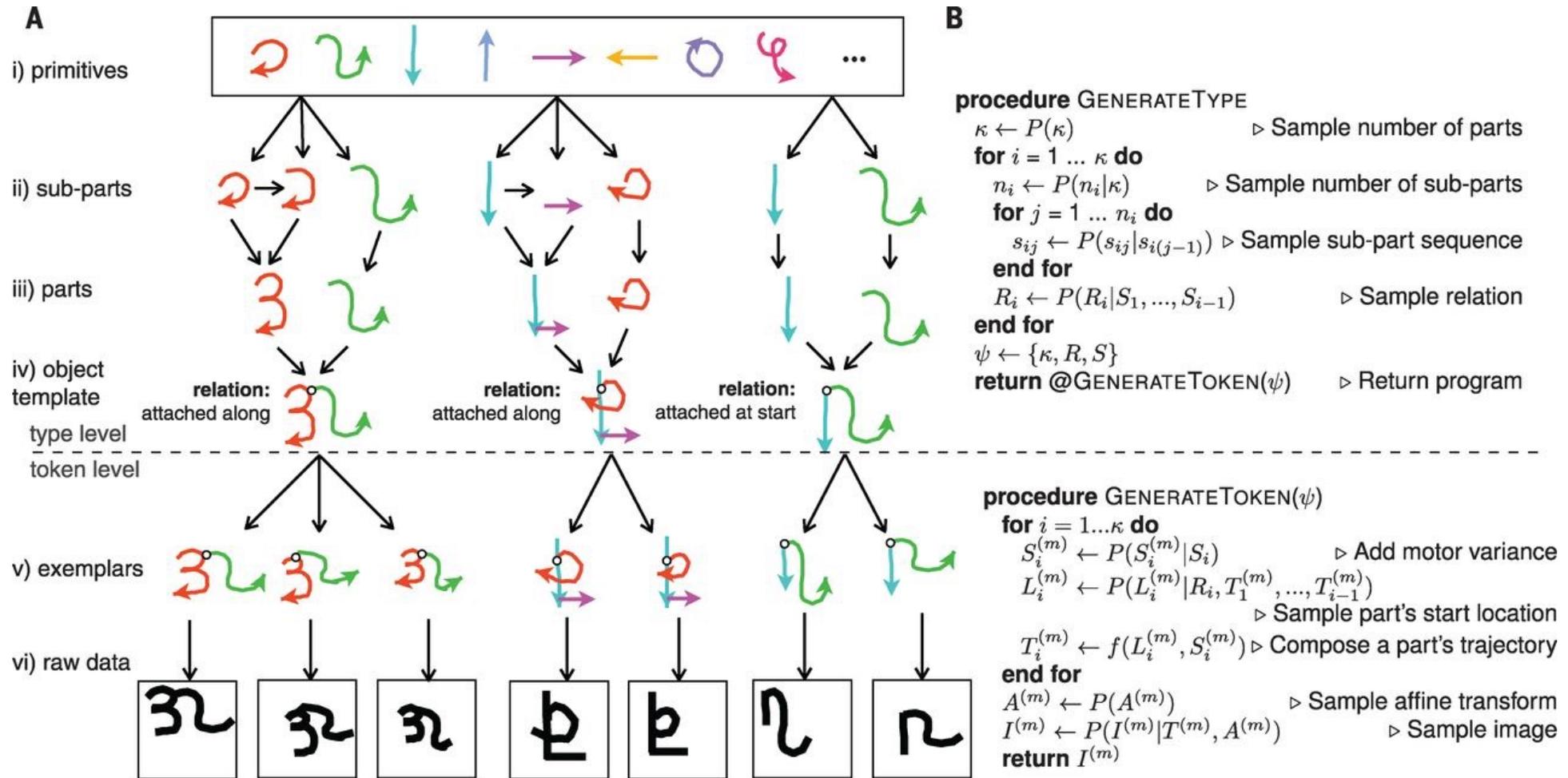
Computers Couldn't Understand Student Code

Code.org Short Code (10 lines)

Outstanding Student paper award, AAAI 2019
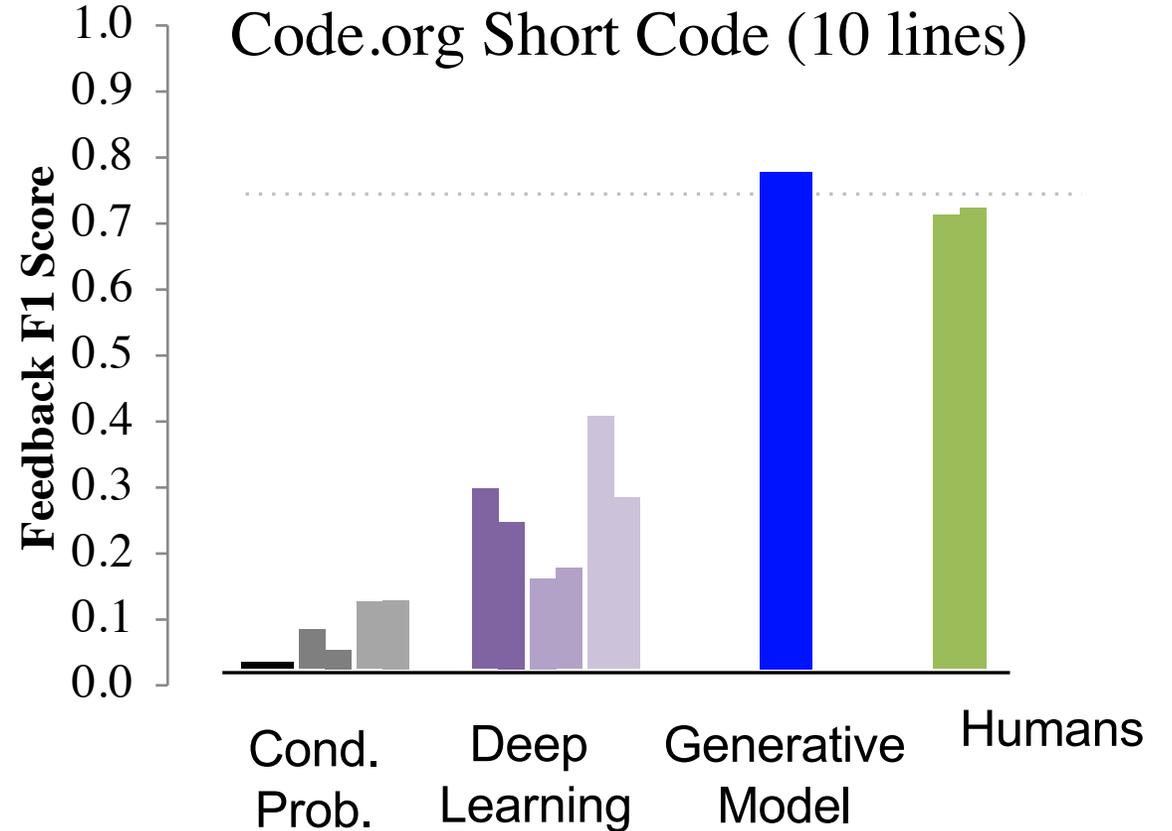
Review

Last Week: Joint Distributions

Joint Distribution *noun*

The probability of a simultaneous assignment to *__all__* the random variables in a probabilistic model.

Eg:

$$P(X = x, Y = y)$$
$$f(X = x, Y = y)$$
$$P(X = x, Y = y, \cdots, Z = z)$$

Monday, Wednesday: Inference

Inference *noun*

An updated belief about a random variable (or multiple) based on conditional knowledge regarding another random variable (or multiple) in a probabilistic model.

TLDR: conditional probability with random variables.
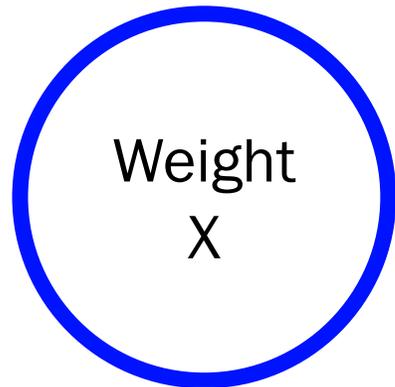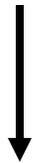
Inference with Continuous

Q: At birth, girl elephant weights are distributed as a Gaussian with mean = 160kg, std = 7kg. At birth, boy elephant weights are distributed as a Gaussian with mean = 165kg, std = 3kg. All you know about a newborn elephant is that it is 163kg. What is the probability that it is a girl?

Model Shown Graphically

IsFemale
G

$G = 1$ is Bern($p = 0.5$)

Weight
X
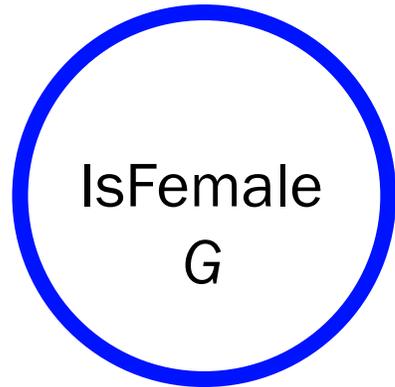
$X \mid G = 1$ is $N(\mu = 160, \sigma^{2\,=}\,7^2)$
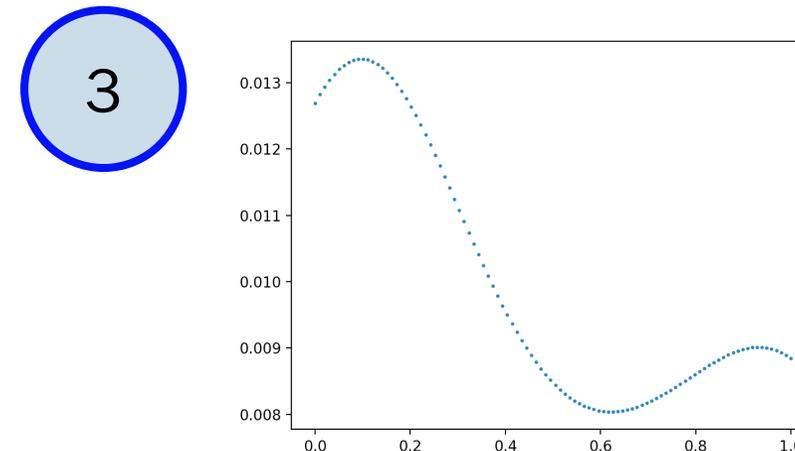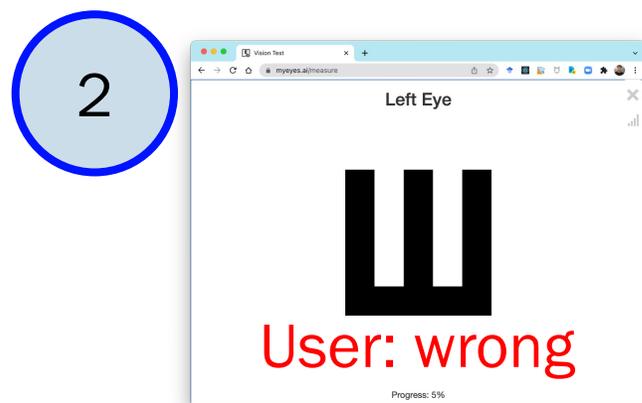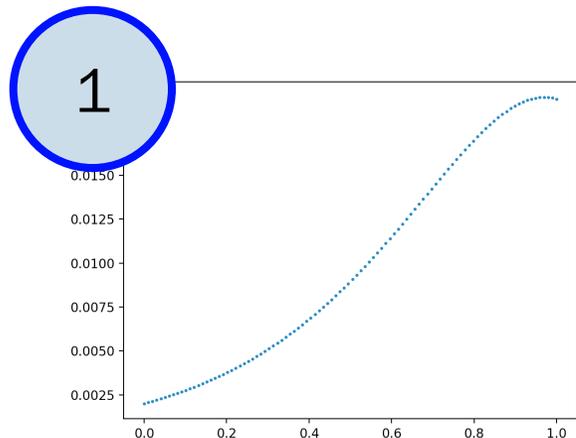$X \mid G = 0$ is $N(\mu = 165, \sigma^{2\,=}\,3^2)$

Does this define the joint?

$$f(G = g, X = x)$$
$$= f(X = x | G = g) P(G = g)$$

Q: What is P(G = 1 | X = 163)

Harder when inferred random variable
is not a Bernoulli

Inference on a non-bernoulli random variable

$$P(A = a)$$

Observation $Y = 0$

$$P(A = a|E)$$

Known function. Also needs font size (constant)

$$P(A = a|Y = 0) = \frac{P(Y = 0|A = a)P(A = a)}{P(Y = 0)}$$
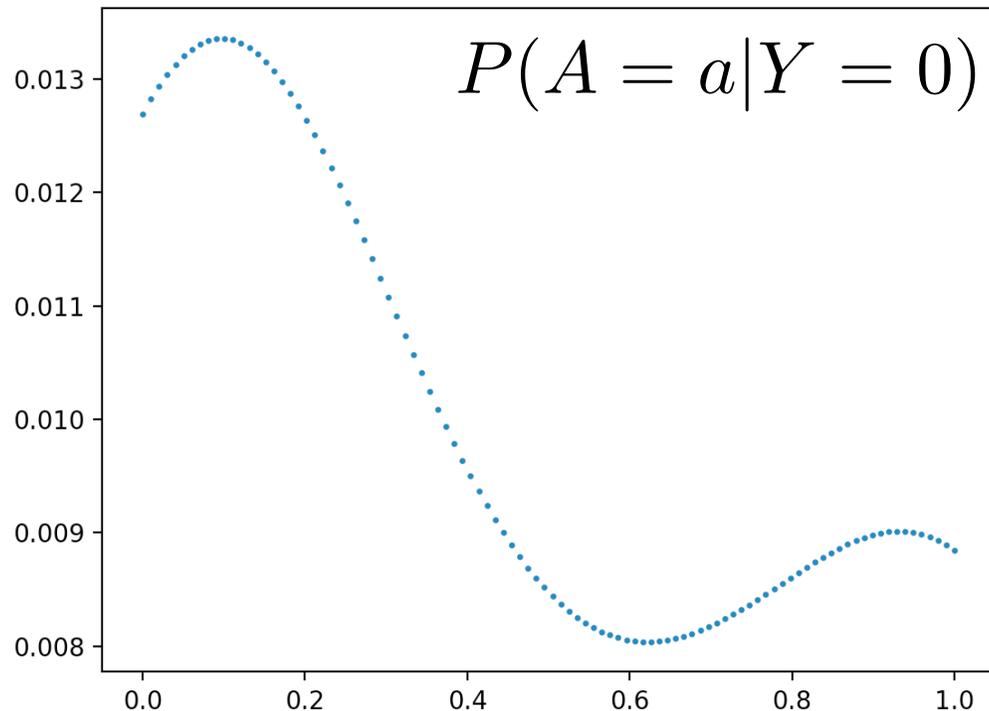
Number or Dictionary?

belief
$$P(A = a | Y = 0) = \frac{P(Y = 0 | A = a) P(A = a)}{P(Y = 0)}$$

belief[a] = 0.001

Inference on a non-bernoulli random variable

In plain English: run bayes for each value of a

$$P(A = a | Y = 0)$$

```
# RV bayes as code
def update(belief, obs):
    for a in support:
        prior_a = belief[a]
        likelihood = calc_likelihood(a, obs)
        belief[a] = prior_a * likelihood
    normalize(belief)
```

likelihood

$$P(A = a | Y = 0) = \frac{P(Y = 0 | A = a) P(A = a)}{P(Y = 0)}$$

Normalize???

```
# RV bayes as code
def update(belief, obs):
    for a in support:
      prior_a = belief[a]
      likelihood = calc_likelihood(a, obs)
      belief[a] = prior_a * likelihood
    normalize(belief)
```

In plain English: this is the numerator, summed over all values of A

$$P(A = a|Y = 0) = \frac{P(Y = 0|A = a)P(A = a)}{P(Y = 0)}$$

$$= \frac{P(Y = 0|A = a)P(A = a)}{\sum_{x \in A} P(Y = 0, A = x)}$$

$$= \frac{P(Y = 0|A = a)P(A = a)}{\sum_{x \in A} P(Y = 0|A = x)P(A = x)}$$
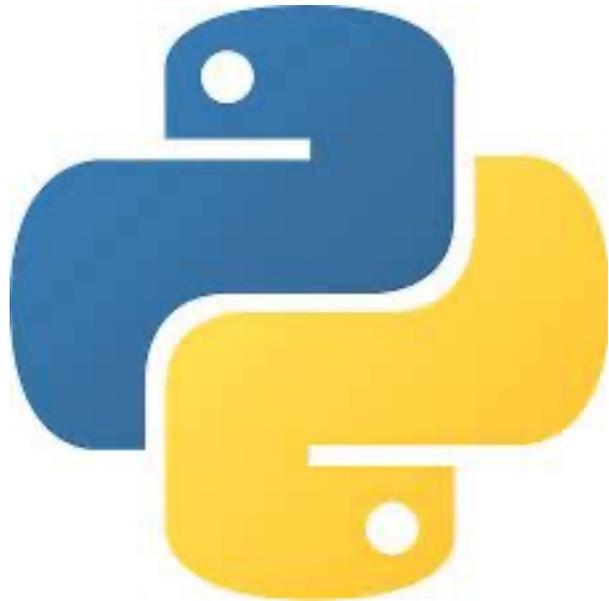
Learning Goals

1. Perspective on the artform of how to design probabilistic models
2. How to calculate Correlations
3. Use and verify Independence with Random Variables
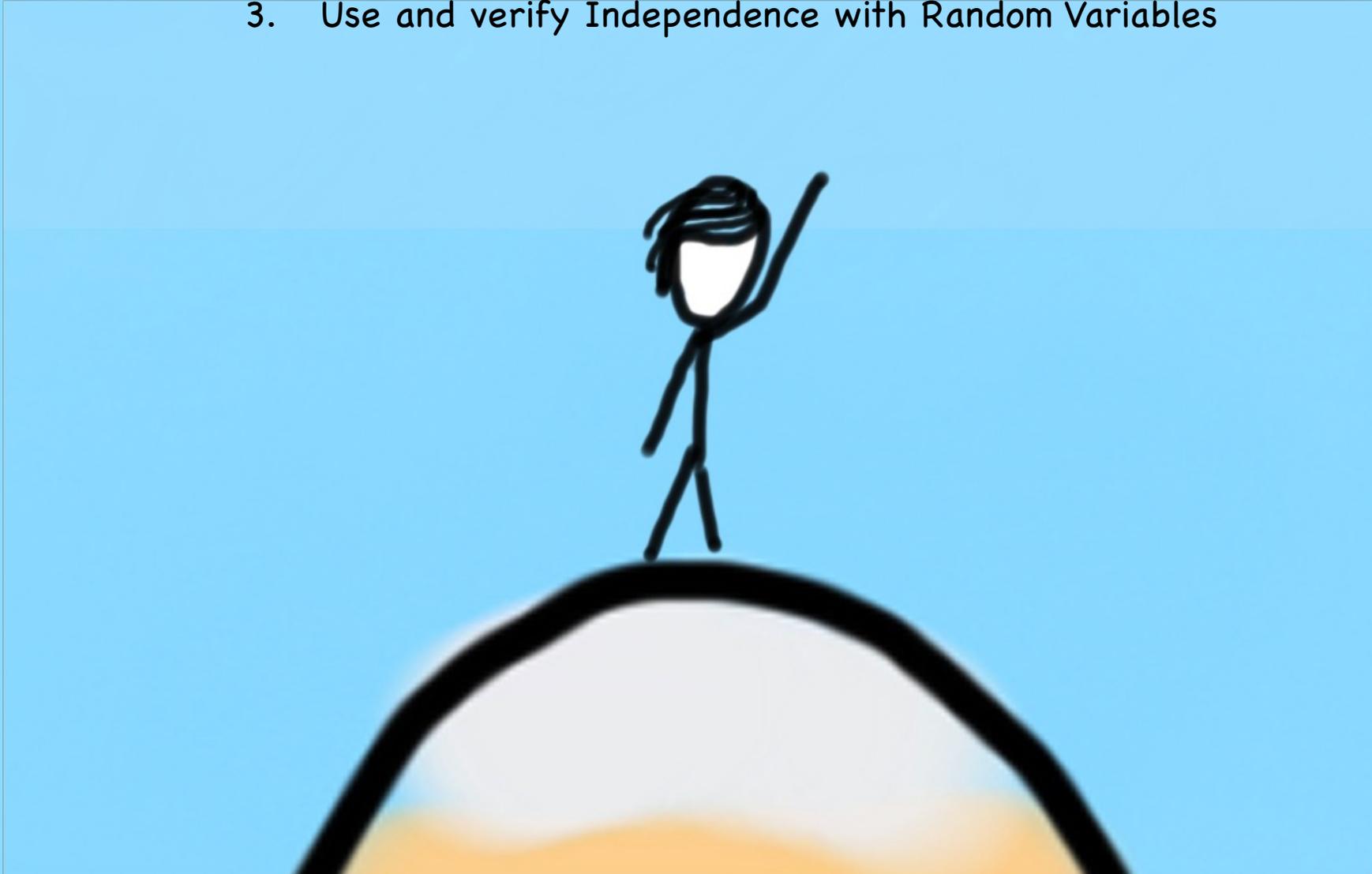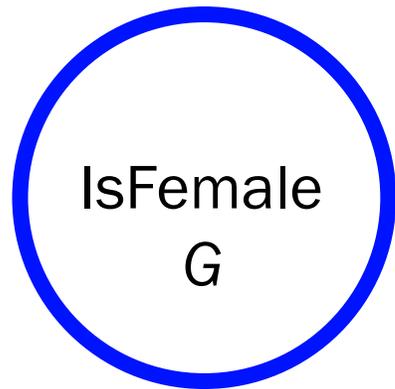
Lets talk about how to make a model

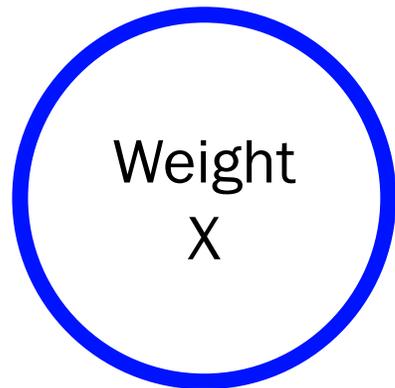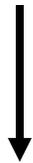# Model Version #1: Python That Outputs a **Joint** Sample

Model Version #2: Bayesian Network

IsFemale
G

G = 1 is Bern(p = 0.5)

Does this define the joint?

$$f(G = g, X = x)$$
$$= f(X = x | G = g) P(G = g)$$

Weight
X

$X \mid G = 1$ is $N(\mu = 160, \sigma^{2 = 7^2})$
$X \mid G = 0$ is $N(\mu = 165, \sigma^{2 = 3^2})$

Why You Need a Model

Inference

Inference

Cold

Strep Throat

Flu

Under-grad

Nausea

Chest pain

Fever

Tired

Sore Throat
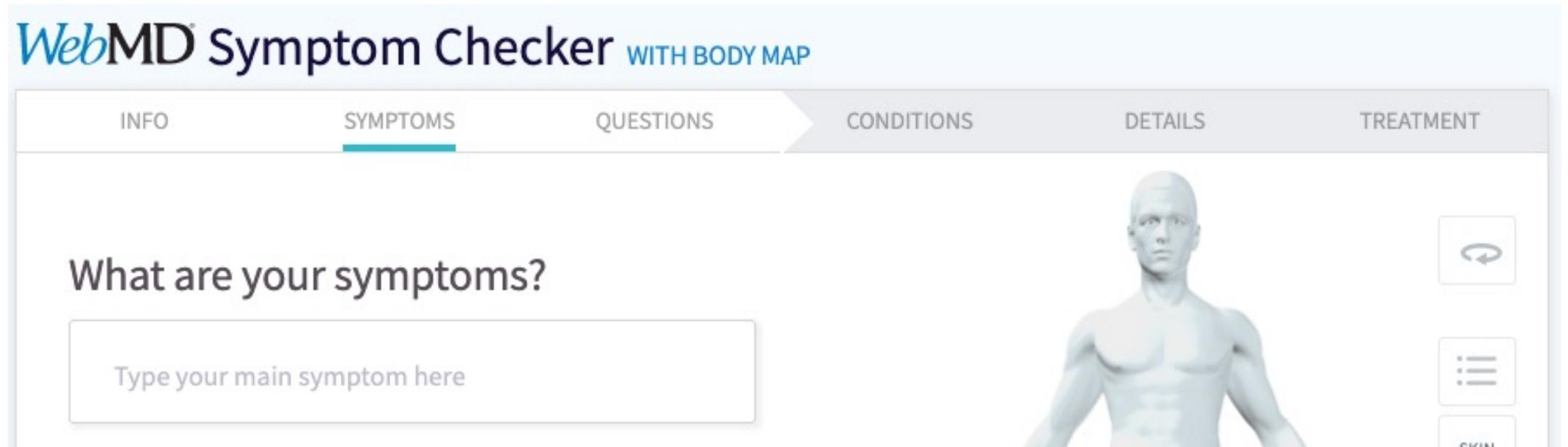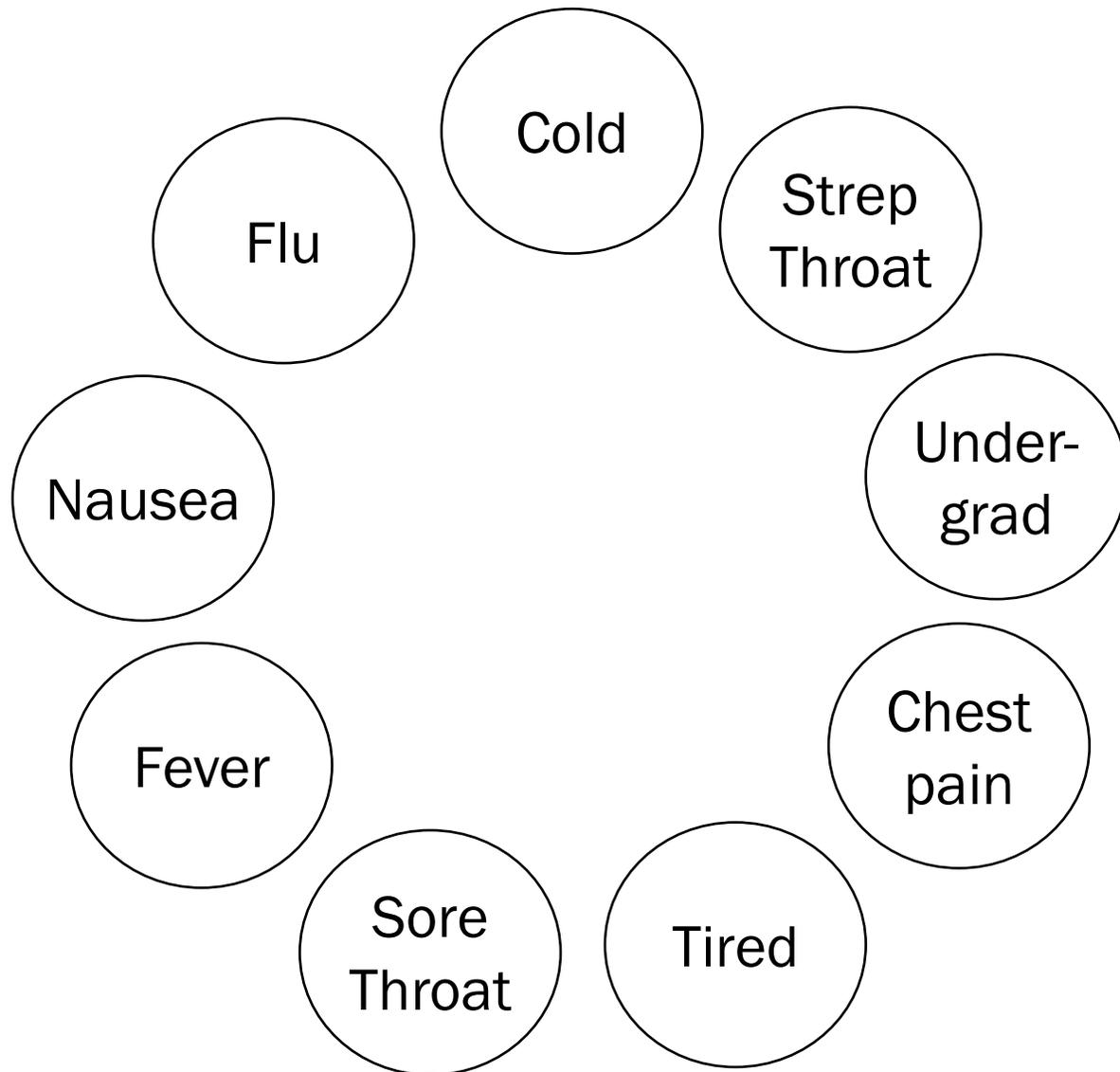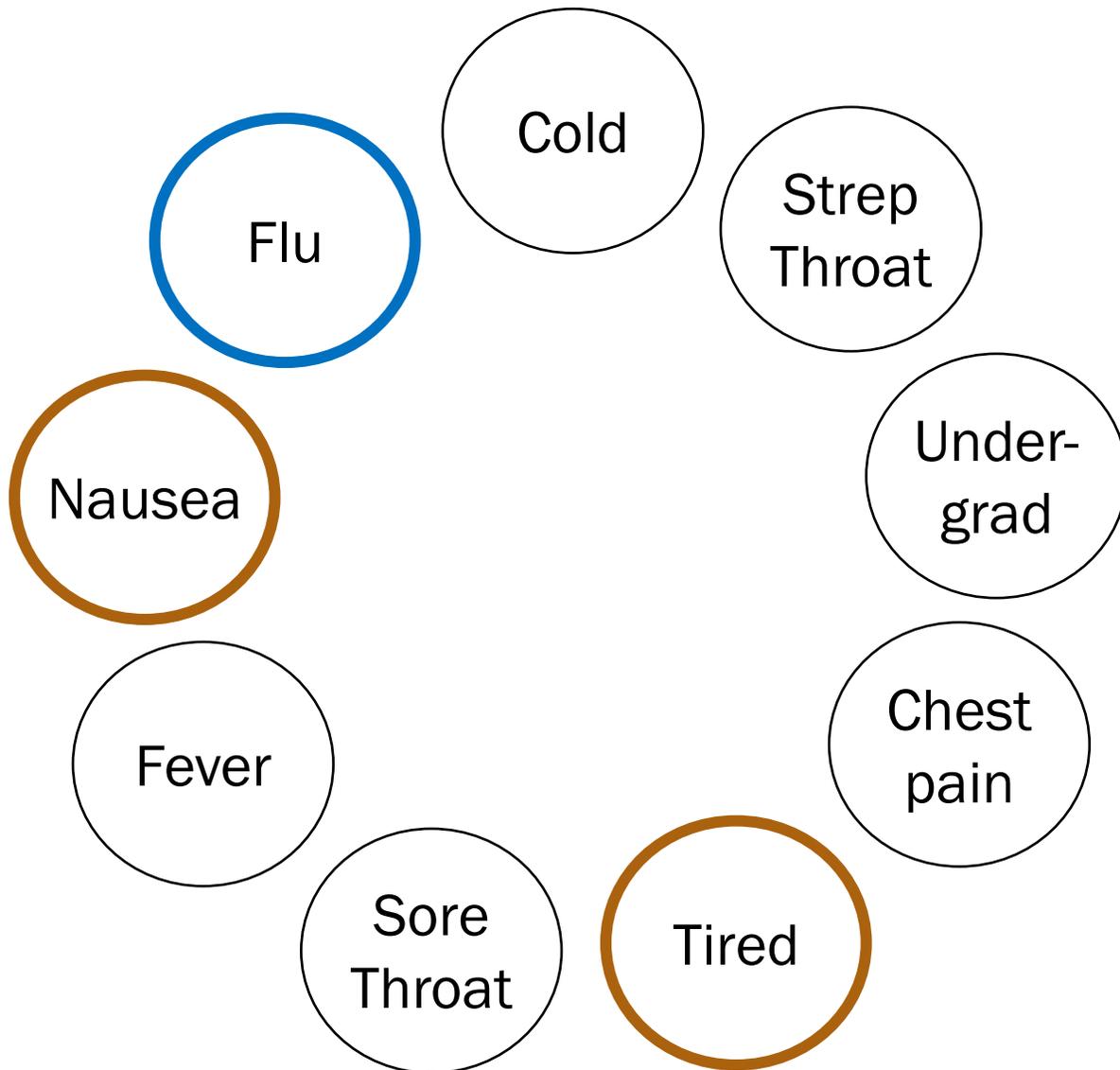
Inference question:

Given the values of some random variables, what are the conditional distributions of some other random variables?
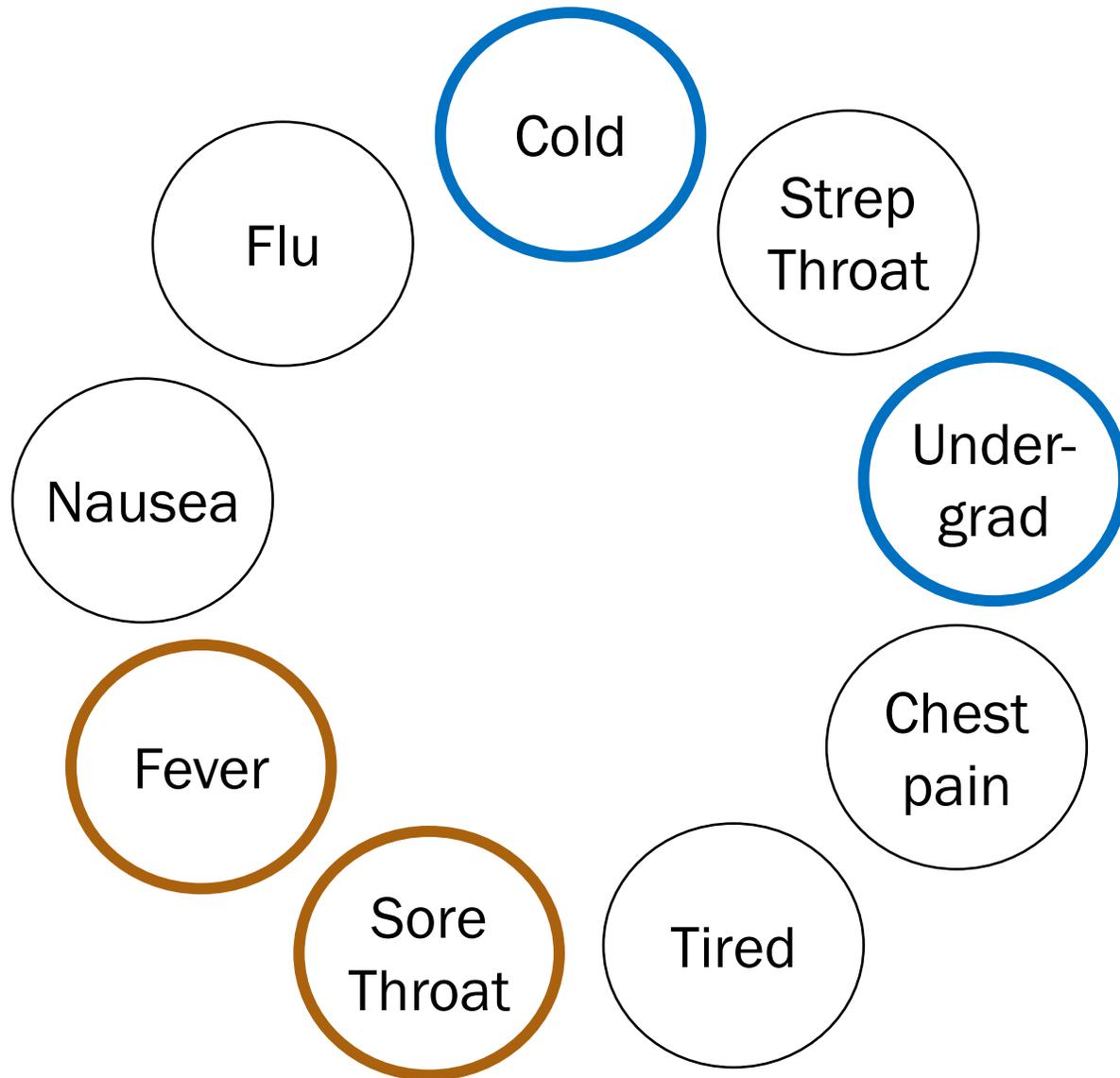
Inference

One inference question:

$$P(F = 1 | N = 1, T = 1)$$

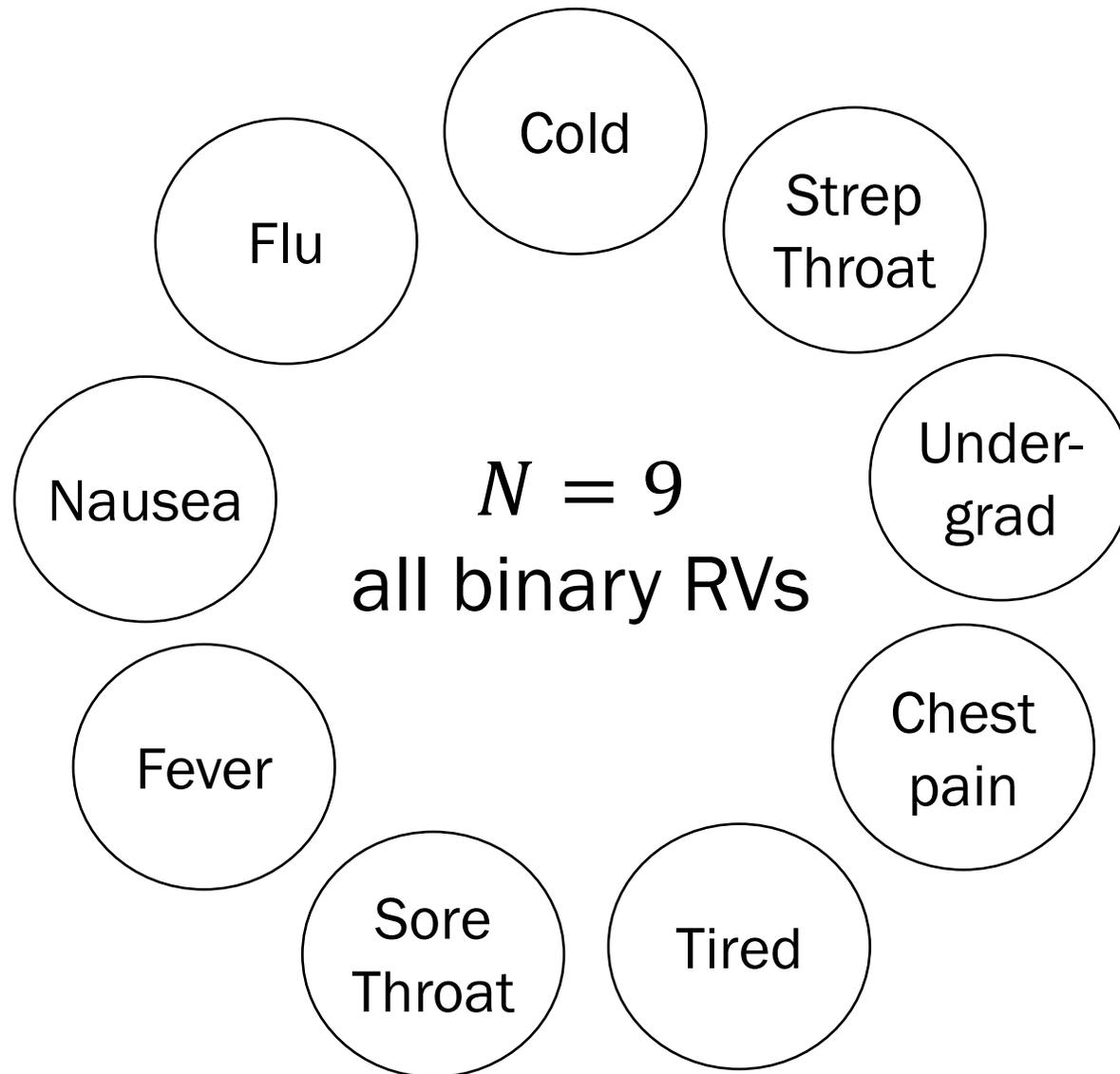$$= \frac{P(F = 1, N = 1, T = 1)}{P(N = 1, T = 1)}$$

Inference

Another inference question:

$$P(C_o = 1, U = 1 | S = 0, F_e = 0)$$

$$= \frac{P(C_o = 1, U = 1, S = 0, F_e = 0)}{P(S = 0, F_e = 0)}$$

Inference

Cold

Flu

Strep Throat

Nausea

$N = 9$

all binary RVs
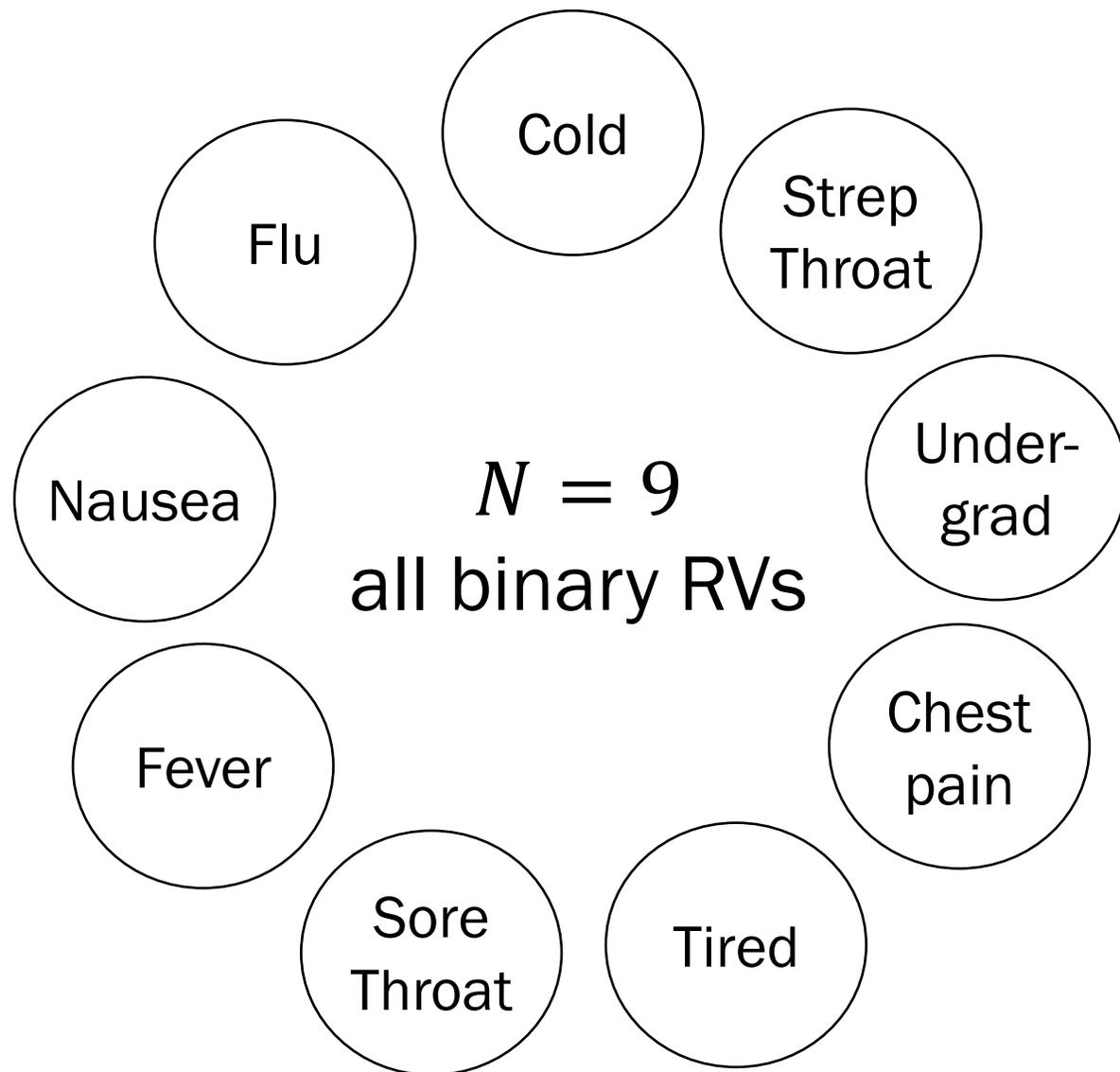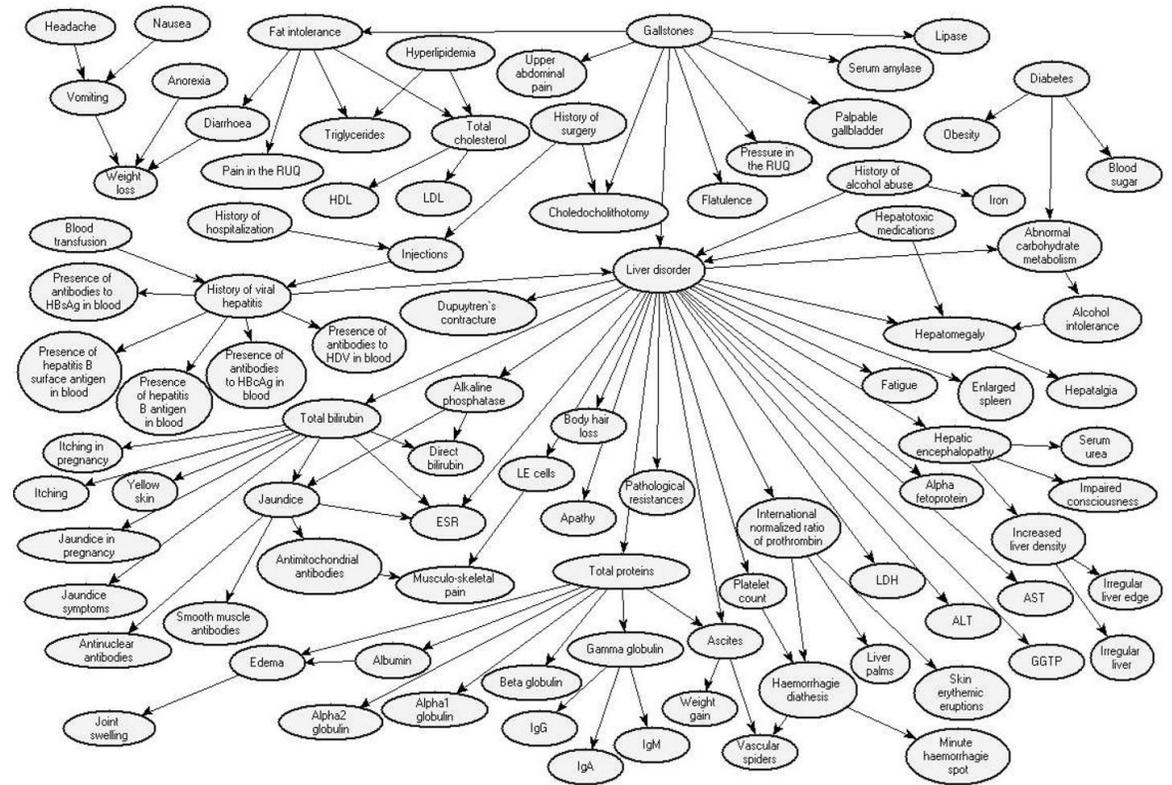
Under-grad

Fever

Chest pain

Sore Throat

Tired

If we knew the **joint distribution**, we can answer all probabilistic inference questions.

What is the size of the joint probability table?

A. 2^{N-1} entries
B. N^2 entries
C. 2^N entries
D. None/other/don't know

Inference

Cold

Flu

Strep Throat

Nausea

$N = 9$

all binary RVs
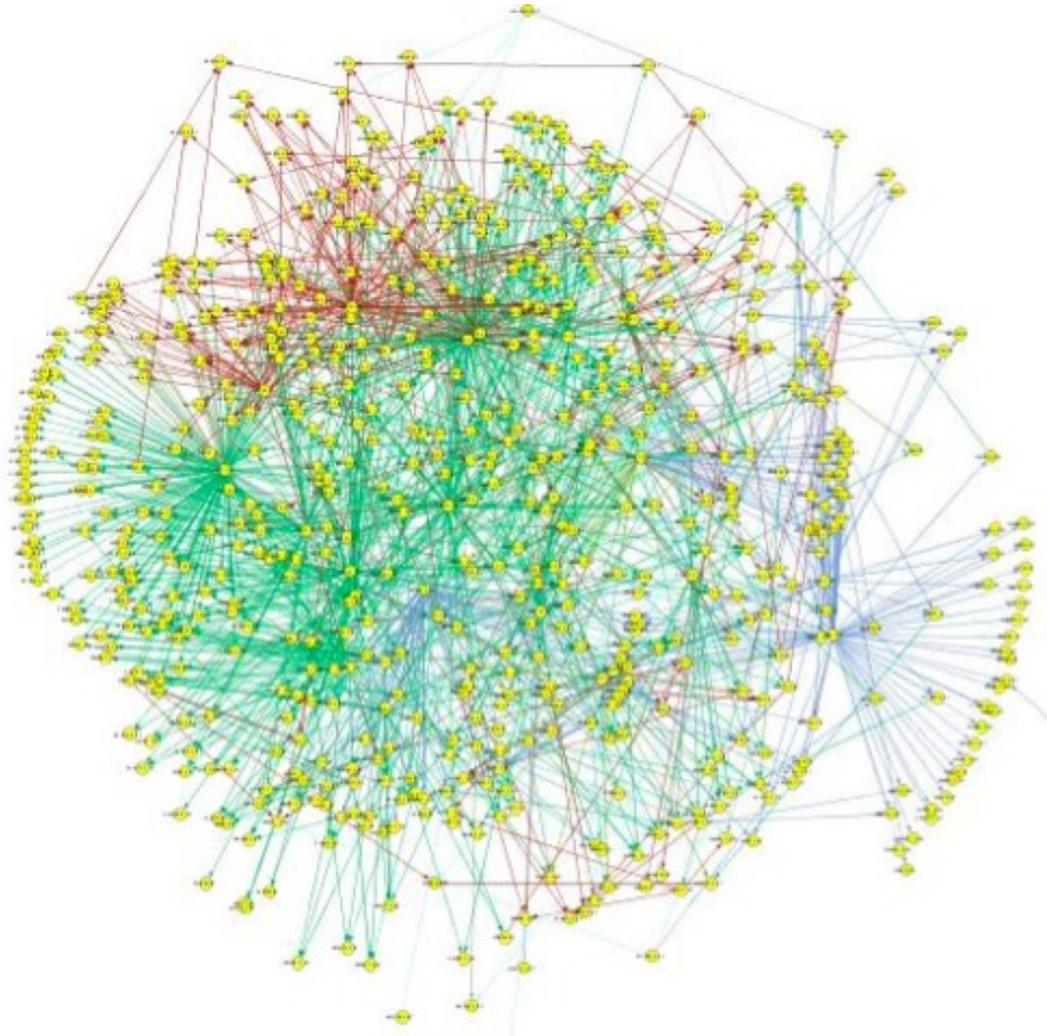
Under-grad

Fever

Sore Throat

Tired

Chest pain

If we knew the **joint distribution**, we can answer all probabilistic inference questions.

What is the size of the joint probability table?

A. 2^{N-1} entries

B. N^2 entries

C. 2^N entries

D. None/other/don't know

Naively specifying a joint distribution is, in general, intractable.

N can be large…

Bayesian Networks

A simpler WebMD
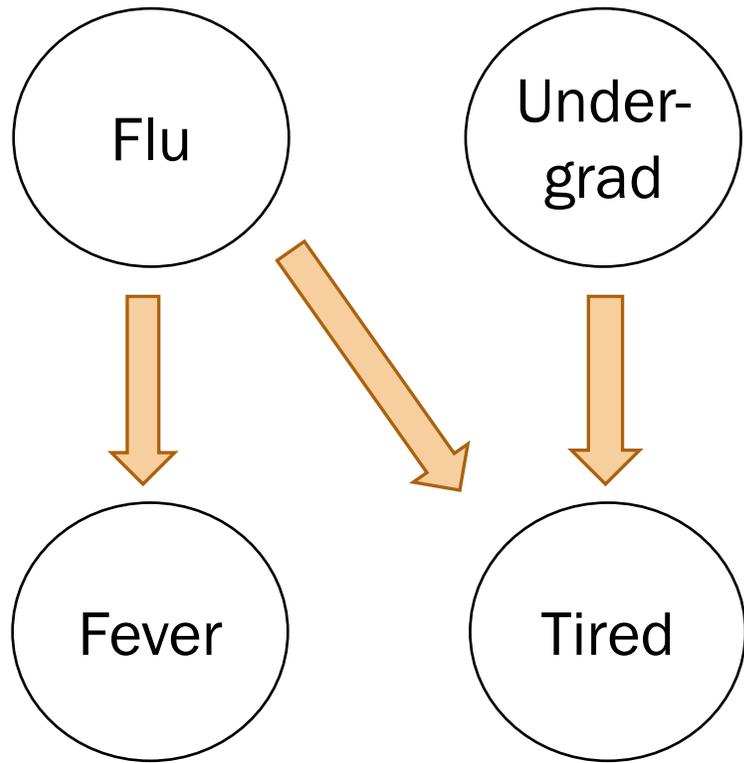
Flu

Under-grad

Fever

Tired

Great! Just specify $2^4 = 16$ joint probabilities...?

$$P(F_{lu} = a, F_{ev} = b, U = c, T = d)$$

What would a Stanford flu expert do?
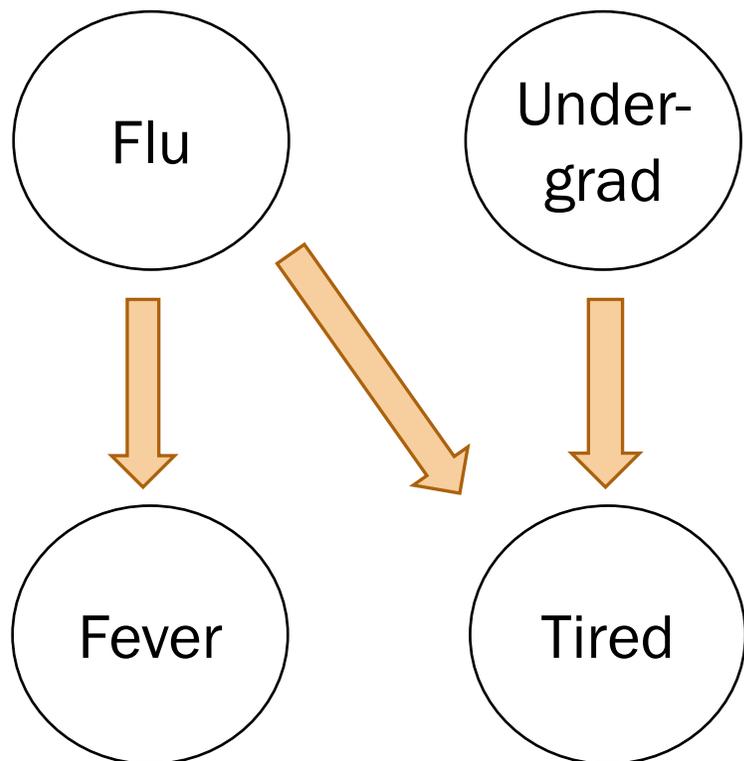
Describe the joint distribution using causality!

Constructing a Bayesian Network

What would a Stanford flu expert do?

✅ 1. Describe the joint distribution using causality.

2. Provide $P(\text{values}|\text{parents})$ for each random variable
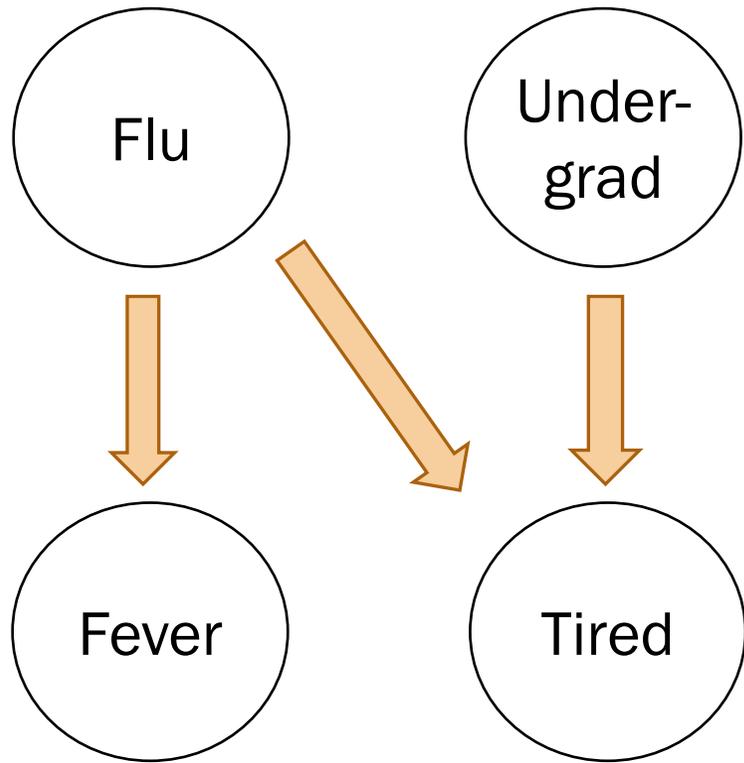
3. Implicitly assumes independences.

Constructing a Bayesian Network

In a Bayesian Network,
Each random variable is caused by its parents. Def P(node | parents)

- Node: random variable
- Directed edge: causality

Examples:
- $P(F_{lu} = 1)$
- $P(U = 0)$
- $P(F_{ev} = 1 | F_{lu} = 1), P(F_{ev} = 1 | F_{lu} = 0)$
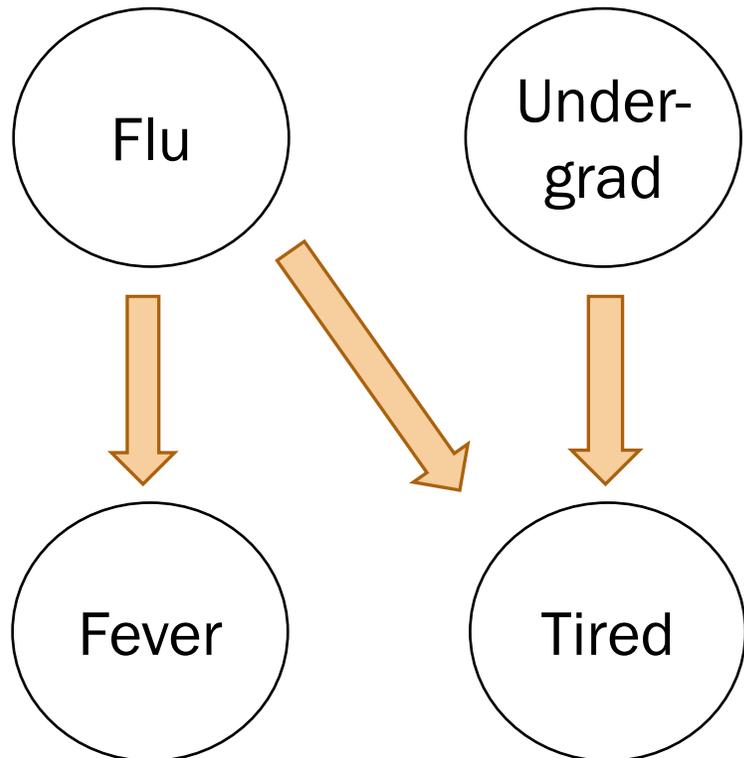- $P(T = 1 | F_{lu} = 0, U = 0) \dots$

$P(T = 1 | F_{lu} = 0, U = 0)$
$P(T = 1 | F_{lu} = 0, U = 1)$
$P(T = 1 | F_{lu} = 1, U = 0)$
$P(T = 1 | F_{lu} = 1, U = 1)$

Constructing a Bayesian Network

What would a Stanford flu expert do?

✅ 1. Describe the joint distribution using causality.

✅ 2. Provide $P(\text{values}|\text{parents})$ for each random variable

✅ 3. Implicitly assumes independences.
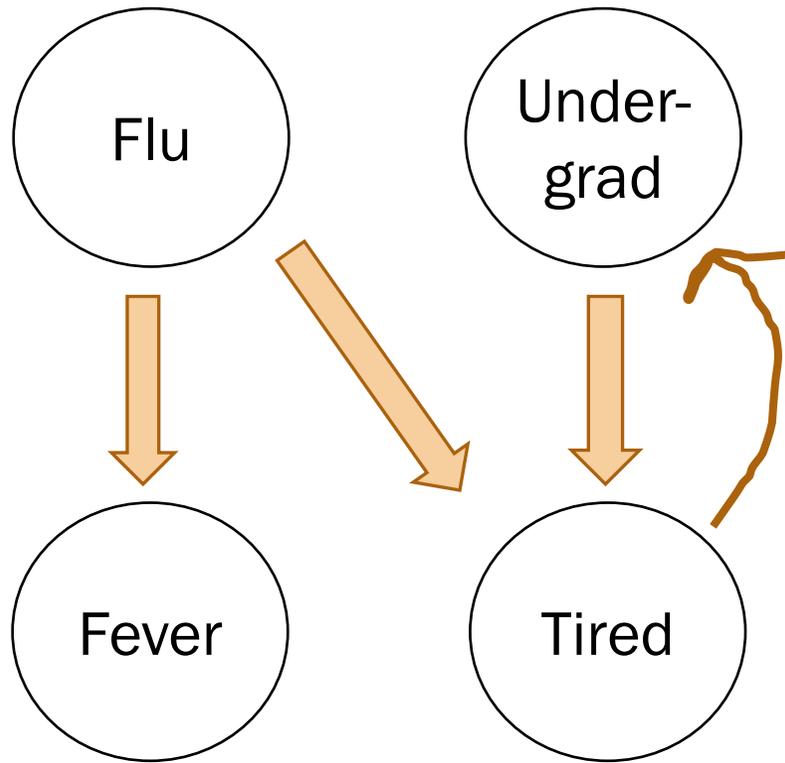
Constructing a Bayesian Network

This model assumes that Flu and being an Undergraduate are independent.

Neat trick: it also assumes that fever and tired are conditionally independent given Flu.
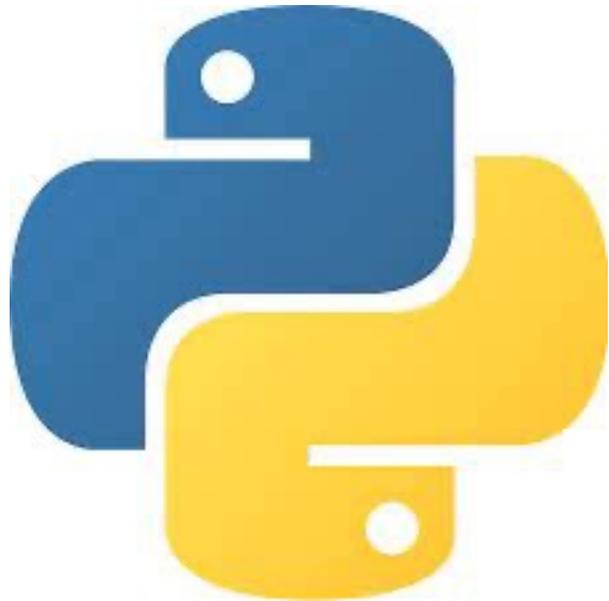
You need to tell a generative story. You do **not** need to be able to reason about all the implied independencies

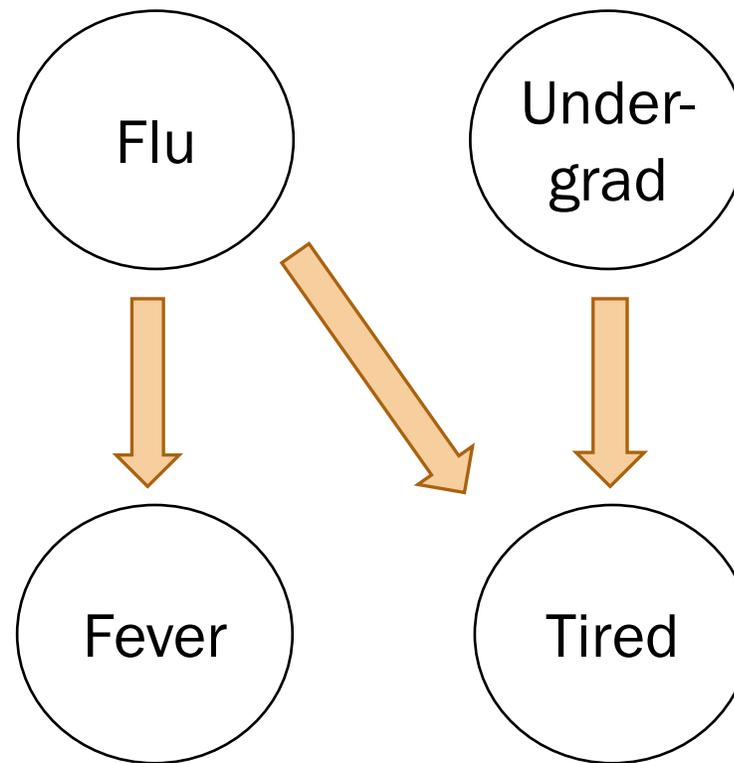# Bug: Constructing a Bayesian Network

Must by acyclic!

Model Version #1: Python That Outputs a **Joint** Sample

Sample Baby
Elephant

Sex: Female
Weight: 161kg

Bayesian Network Assumption:

$$P(\text{Joint}) = P(X_1 = x_1 \ldots X_n = x_n)$$

$$= \prod_i P(X_i = x_i | \text{parents of } X_i)$$

Bayes Nets tell a generative story.

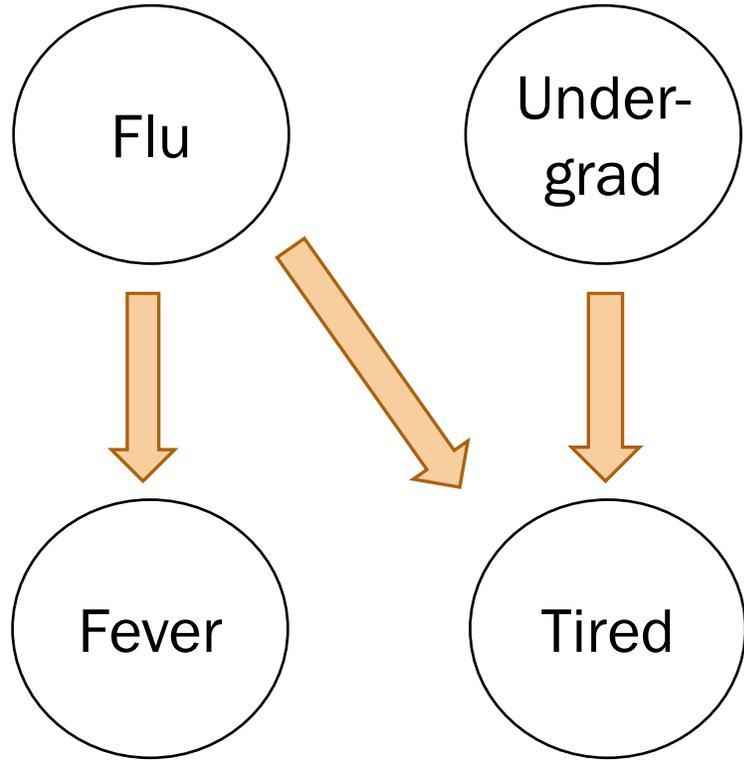This leads to many independence assumptions

Makes it **tractable** to represent the joint

Challenge: Exact Inference in a Bayes Net

$$P(Fl = x)$$

$$P(U = u)$$

(Flu)

(Under-grad)

(Fever)

(Tired)

$$P(F = f | Fl = x)$$

$$P(T = t | U = u, Fl = x)$$

$$P(Fl = 0 | Fe = 1)$$

$$= \frac{P(Fl = 0, Fe = 1)}{P(Fe = 1)}$$

$$= \frac{P(Fe = 1 | Fl = 0) P(Fl = 0)}{\sum_i P(Fe = 1 | Fl = i) P(Fl = i)}$$

Independence of RVs

Independent discrete RVs

Recall the definition of independent events E and F:

$$P(EF) = P(E)P(F)$$

Two discrete random variables X and Y are **independent** if:

for all x, y:

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

Different notation, same idea:

$$p_{\ }(x, y) = p_{\ }(x)p_{\ }(y)$$

- Intuitively: knowing value of X tells us nothing about the distribution of Y (and vice versa)
- If two variables are not independent, they are called **dependent**.

Dice (after all this time, still our friends)

Let: D_1 and D_2 be the outcomes of two rolls
$S = D_1 + D_2$, the sum of two rolls

- Each roll of a fair, 6-sided die is an independent trial.
- Random variables D_1 and D_2 are independent.

1. Are events $(D_1 = 1)$ and $(S = 7)$ independent?

2. Are events $(D_1 = 1)$ and $(S = 5)$ independent?

3. Are random variables D_1 and S independent?
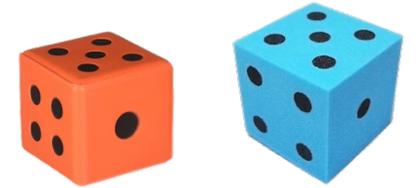
Dice (after all this time, still our friends)

Let: D_1 and D_2 be the outcomes of two rolls
$S = D_1 + D_2$, the sum of two rolls
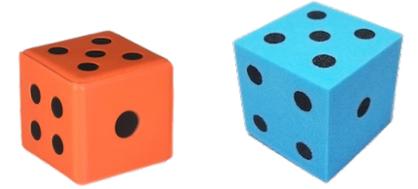
- Each roll of a 6-sided die is an independent trial.
- Random variables D_1 and D_2 are independent.

1. Are events $(D_1 = 1)$ and ✅
   $(S = 7)$ independent?

2. Are events $(D_1 = 1)$ and ❌
   $(S = 5)$ independent?

3. Are random variables D_1 and S independent? ❌

All events $(X = x, Y = y)$ must be independent for X, Y to be independent RVs.

Can I discover independence from data?

Music	Dance	Folk	Country	Classical music	Musical	Pop	Rock	Me...
5	2	1	2	2	1	5	5	
4	2	1	1	1	2	3	5	
5	2	2	3	4	5	3	5	
5	2	1	1	1	1	2	2	
5	4	3	2	4	3	5	3	
5	2	3	2	3	3	2	5	
5	5	3	1	2	2	5	3	
5	3	2	1	2	2	4	5	
5	3	1	1	2	4	3	5	
5	2	5	2	2	5	3	5	
5	3	2	1	2	3	4	3	
5	1	1	1	4	1	2	5	
5	1	2	1	4	3	3	4	
5	5	3	2	1	5	5	2	
5	2	1	1	2	3	4	5	
1	2	2	3	4	3	3	5	
5	3	1	1	1	2	4	4	
5	3	3	3	2	2	4	4	
5	5	4	3	4	5	5	4	
5	3	3	2	4	2	2	4	
5	3	2	3	4	3	2	5	
5	1	1	3	2	2	2	5	
5	3	2	3	3	3	4		
5	4	2	2	2	4	4	5	
5	3	1	1	4	3	3	5	
5	4	2	1	2	3	5	1	
5	5	5	4	5	3	4	4	
4	3	4	1	3	2	2	4	
5	5	1	1	1	1	3	4	
5	3	4	2	3	2	3	4	
4	4	3	3	3	3	4	4	
4	4	1	3	2	3	5	3	
5	3	1	3	2	3	3	4	
5	2	2	3	4	5	4	3	

Why is it harder to find independences here than for bat DNA expression?

Dance of the Covariance

Recall our Ebola Bats

Bat Data

Gene1	Gene2	Gene3	Gene4	Gene5	Trait
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
...					
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE

Stanford University

Expression Amount

Gene5	Trait
0.76	0.83
0.94	0.85
0.82	0.03
0.94	0.32
0.50	0.10
0.40	0.53
0.90	0.67
0.29	0.71
0.72	0.25
0.15	0.24
0.79	0.98
0.68	0.77
0.71	0.37
0.36	0.18
0.62	0.08
0.59	0.38
0.82	0.76

Spot The Difference

Spot The Difference

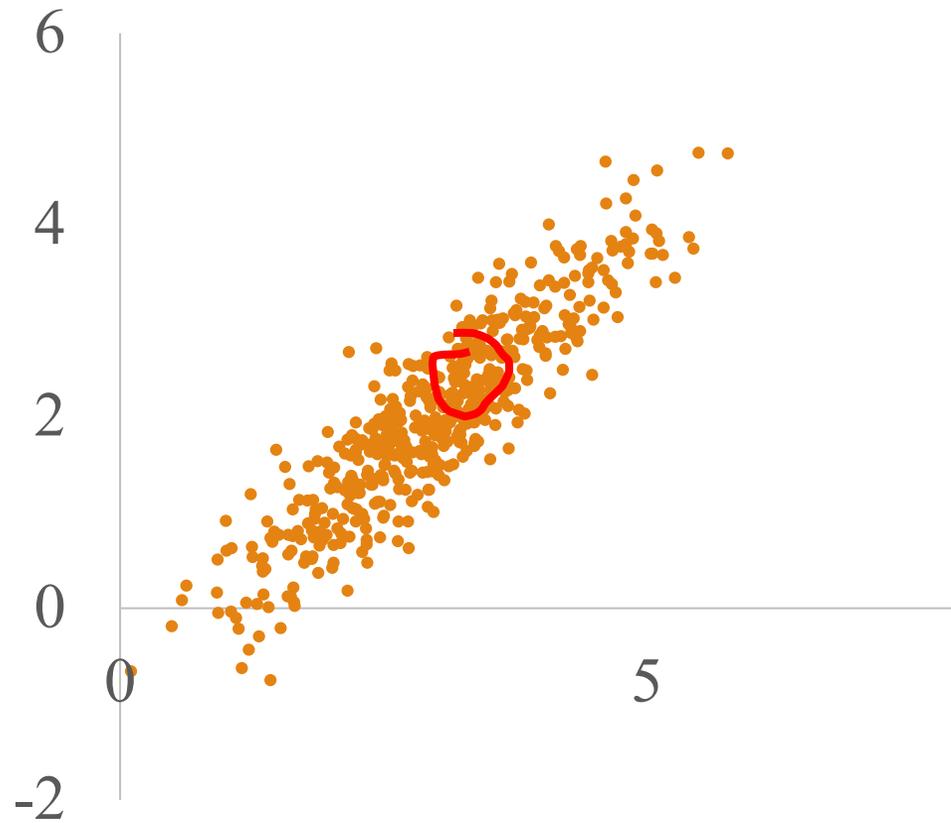**Stanford University**
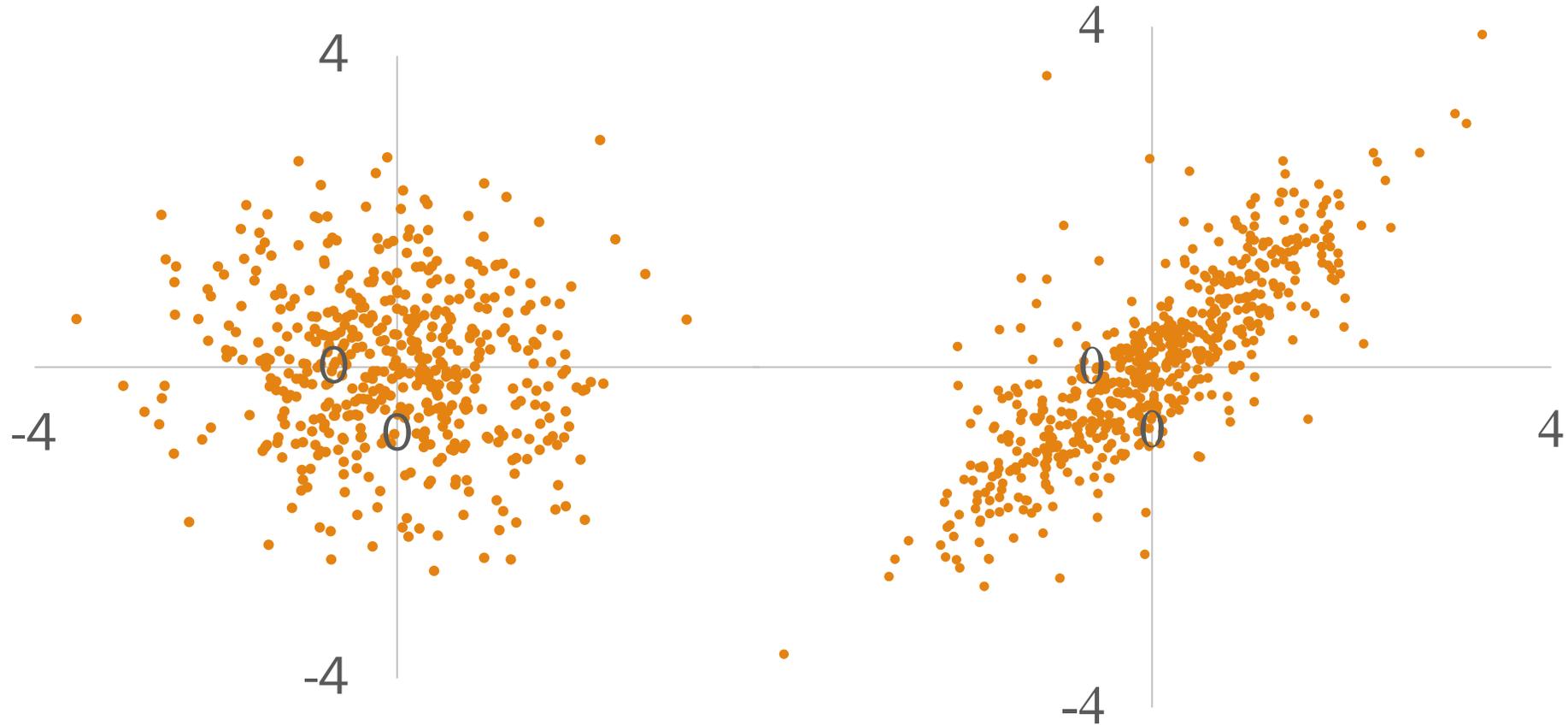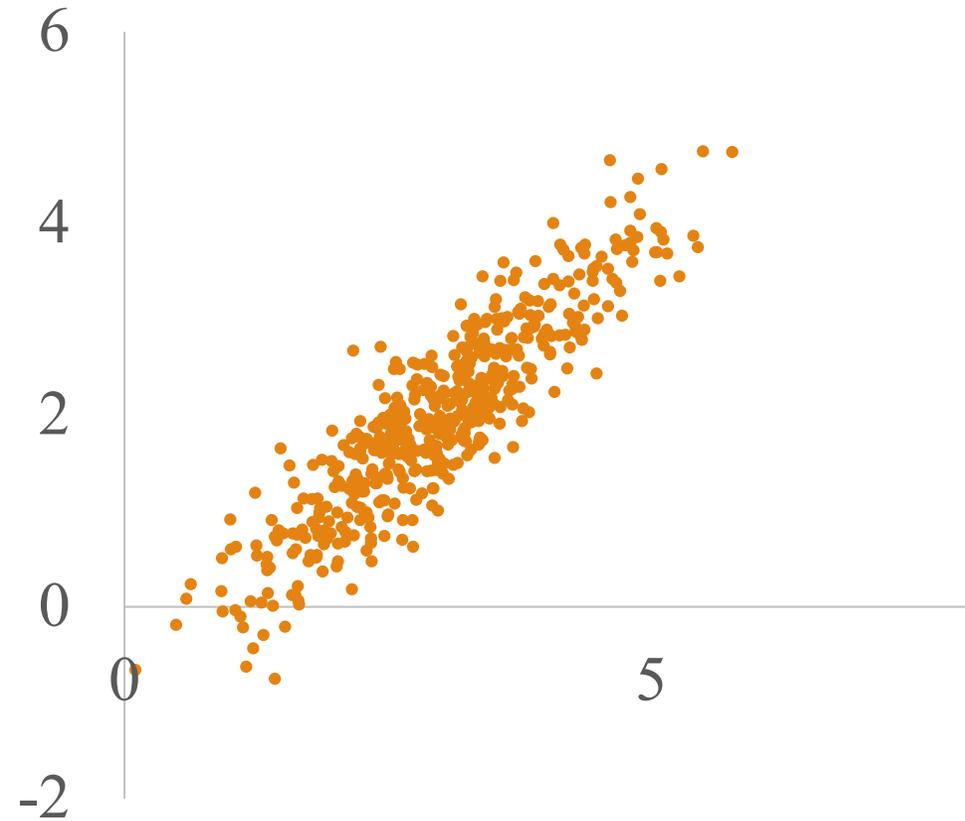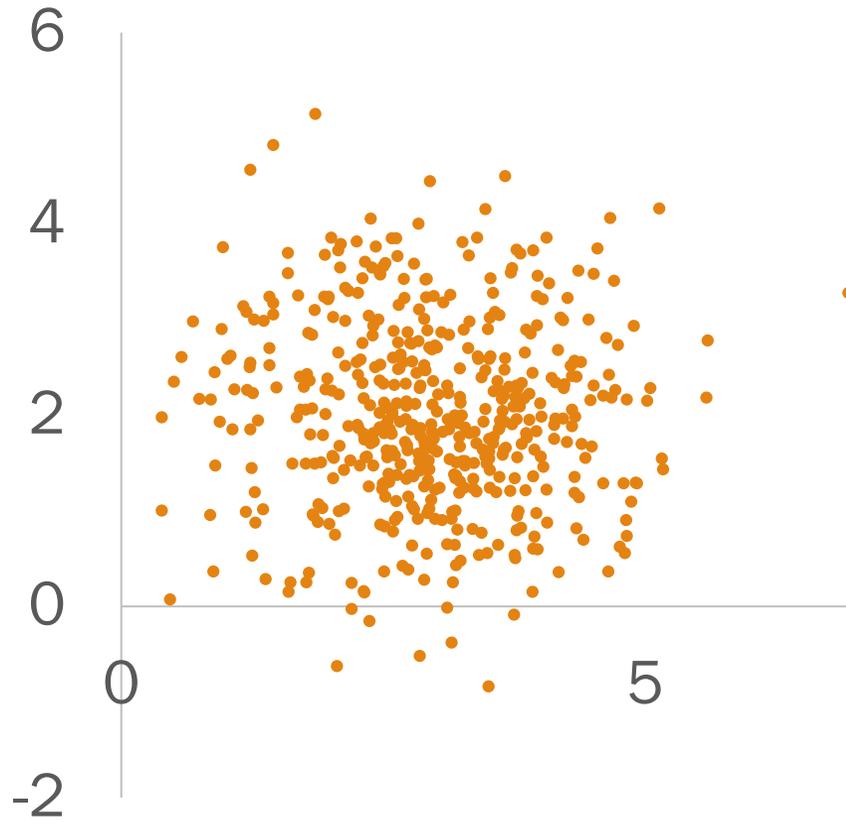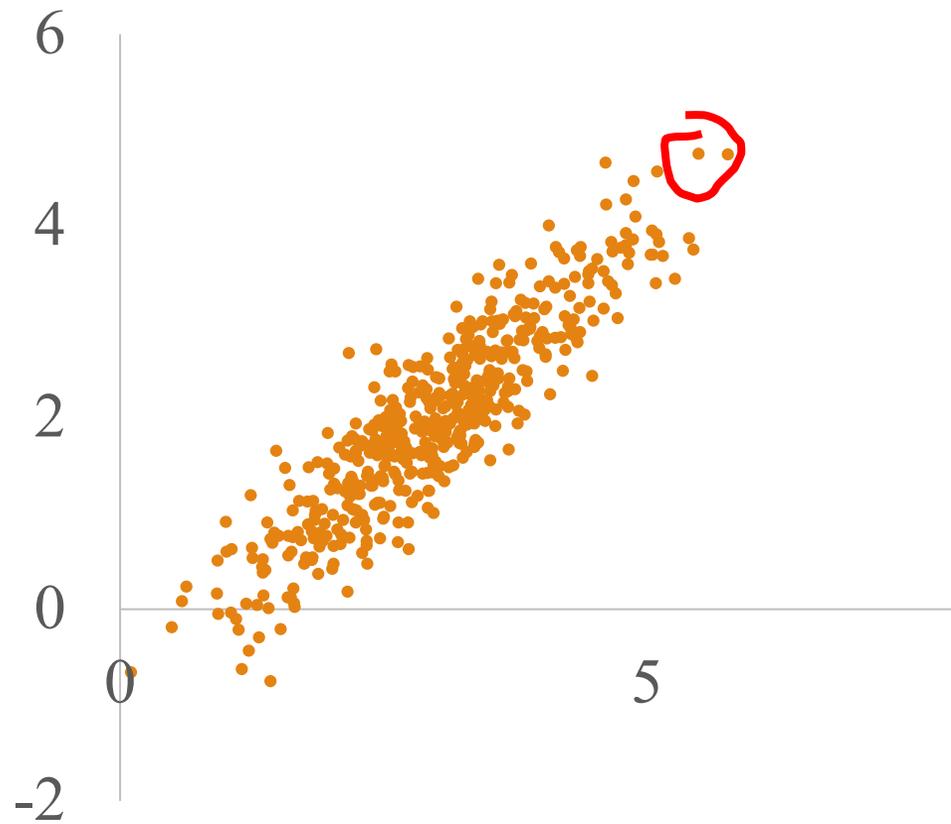
Vary Together

$$x - E[x] = 3$$

$$y - E[y] = 2.6$$

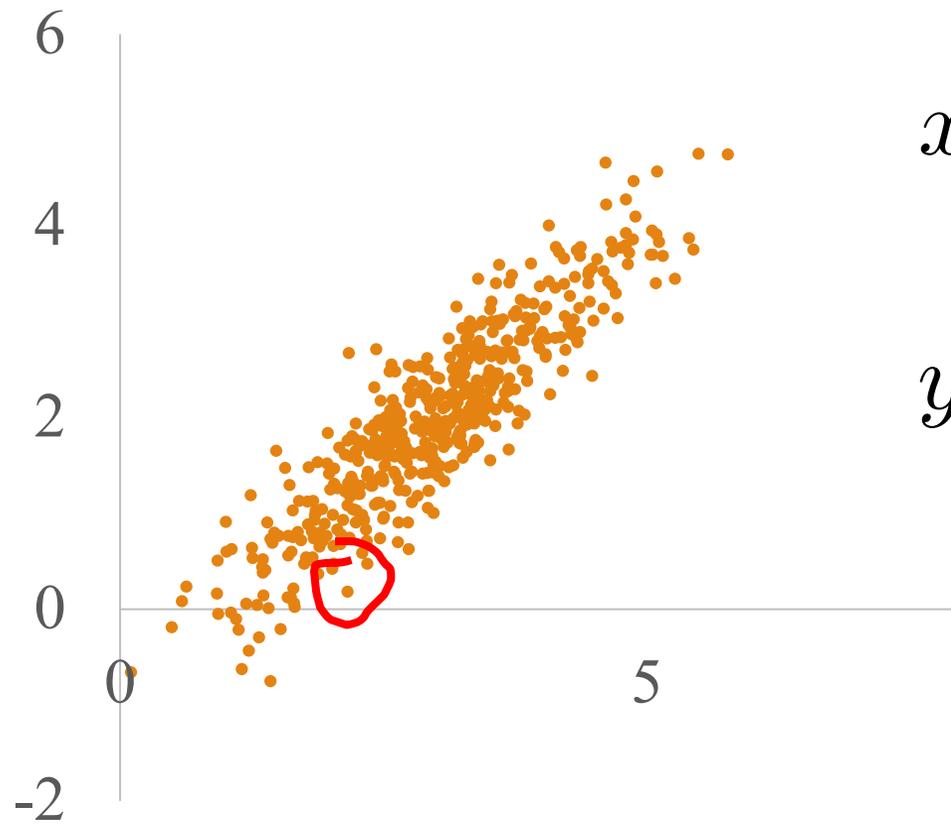$$(x - E[x])(y - E[y]) = 7.8$$

Vary Together

$$x - E[x] \approx 0$$

$$y - E[y] \approx 0$$
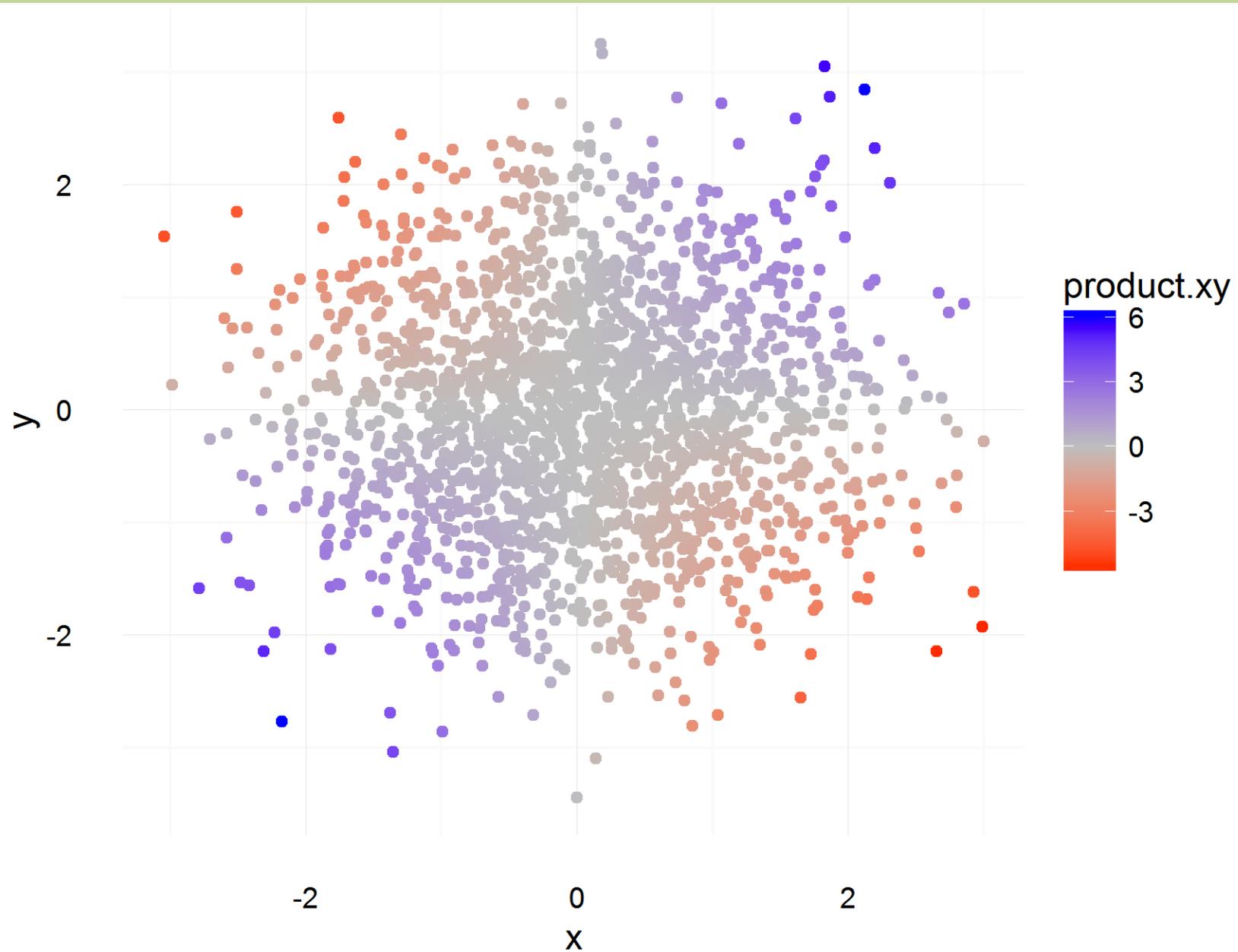
$$(x - E[x])(y - E[y]) = 0$$

Vary Together

$$x - E[x] = -1.1$$

$$y - E[y] = -2.8$$

$$(x - E[x])(y - E[y]) \approx 3.1$$

Understanding Covariance

The Dance of the Covariance

Say X and Y are arbitrary random variables

Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

x	y	(x – E[X])(y – E[Y])p(x,y)
Above mean	Above mean	Positive
Bellow mean	Bellow mean	Positive
Bellow mean	Above mean	Negative
Above mean	Bellow mean	Negative

The Dance of the Covariance

Say X and Y are arbitrary random variables

Covariance of X and Y:
$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Equivalently:

$$\text{Cov}(X,Y) = E[XY - E[X]Y - XE[Y] + E[Y]E[X]]$$
$$= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]$$
$$= E[XY] - E[X]E[Y]$$

- X and Y independent, E[XY] = E[X]E[Y] → Cov(X,Y) = 0
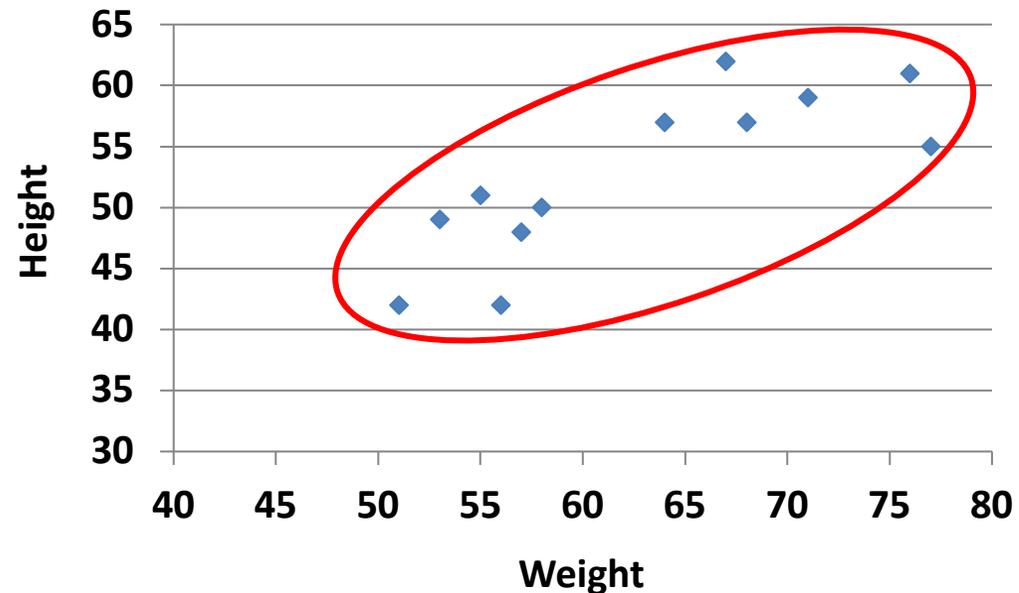- But Cov(X,Y) = 0 does **<u>not</u>** imply X and Y independent!

Covariance and Data

Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

E[W] = 62.75	E[H] = 52.75	E[W*H] = 3355.83

$$Cov(W, H) = E[W*H] - E[W]E[H]$$
$$= 3355.83 - (62.75)(52.75)$$
$$= 45.77$$

Covariance

Poll: (a) positive, (b) negative, (c) zero
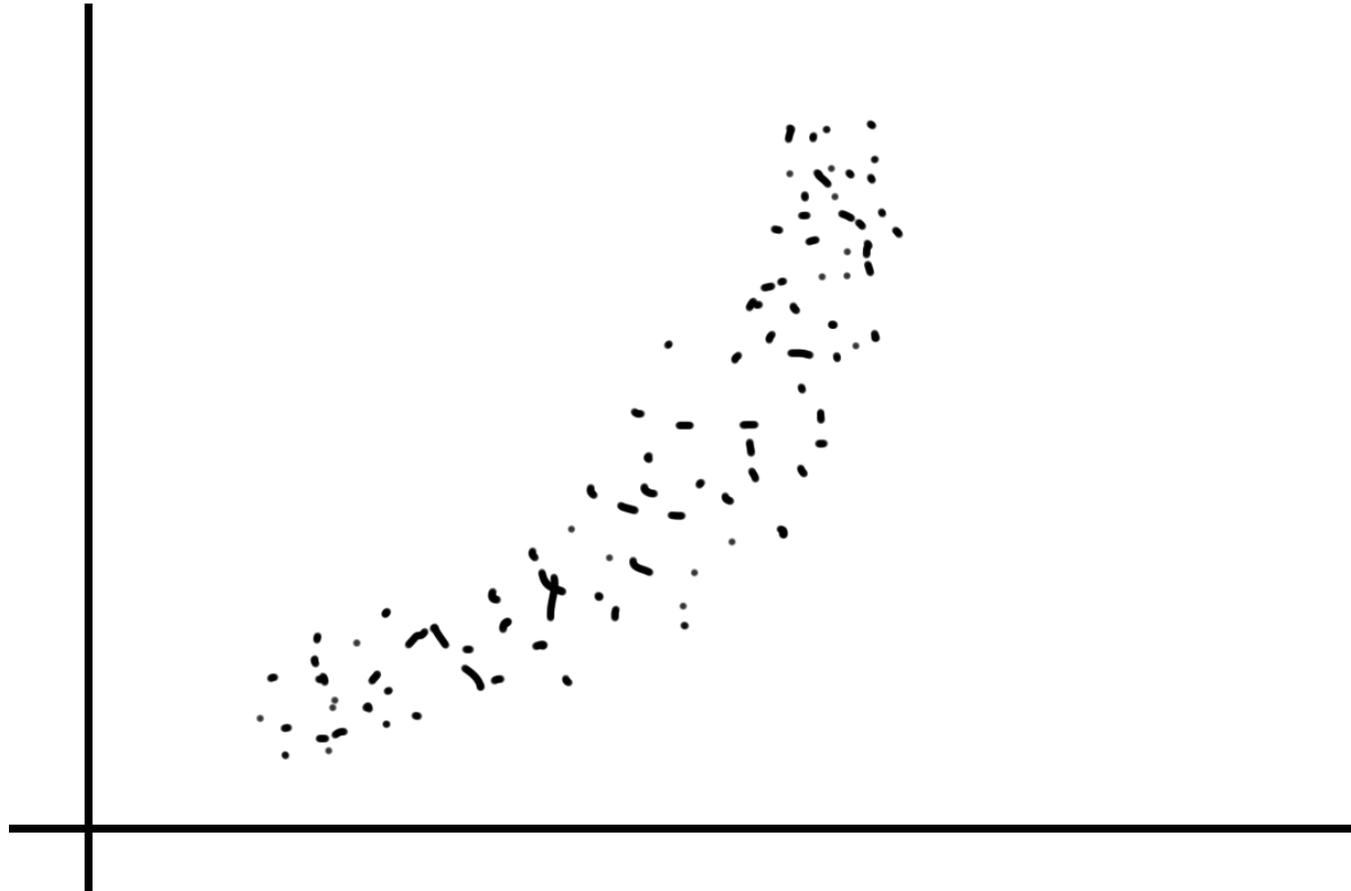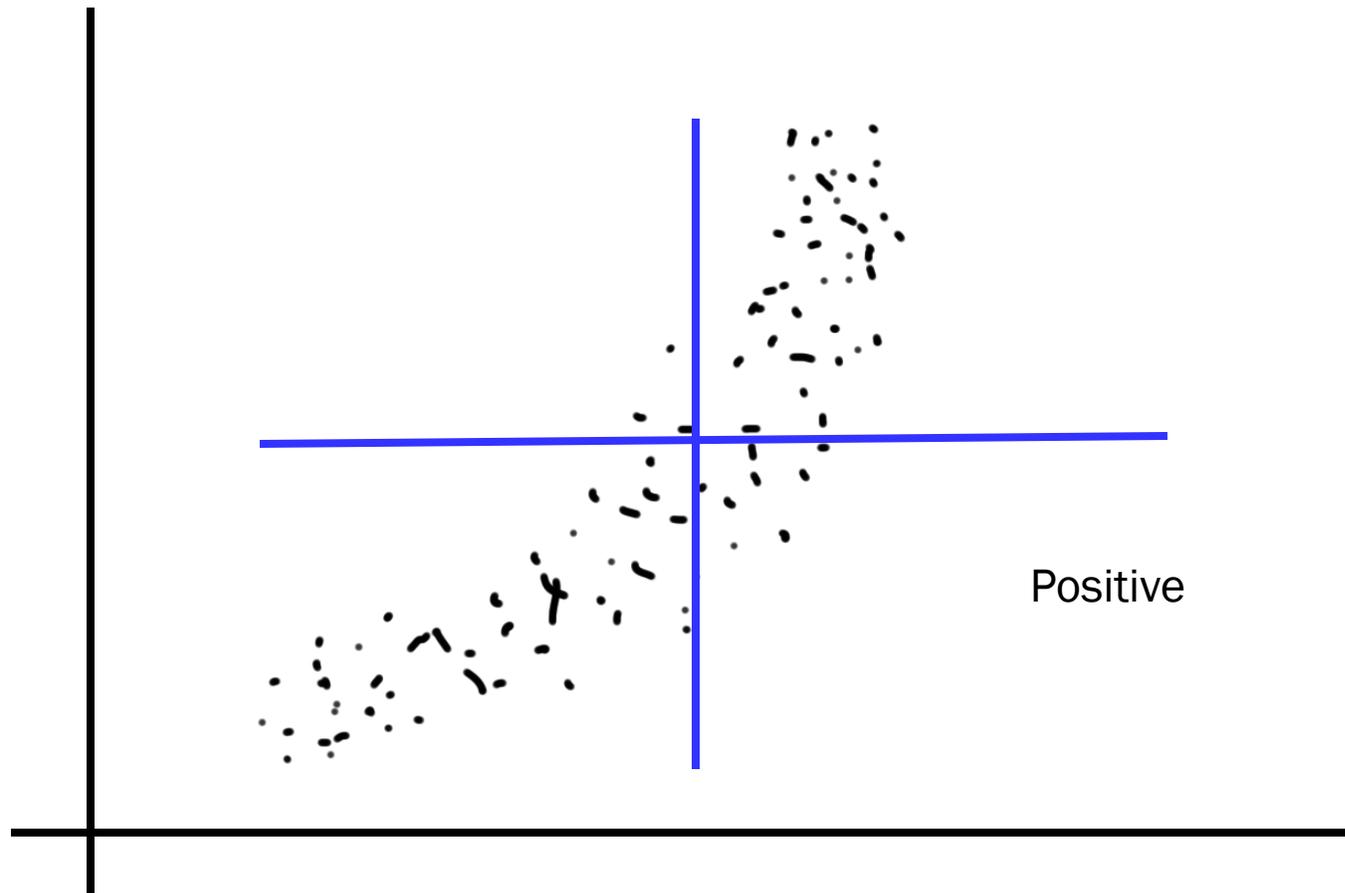
Covariance

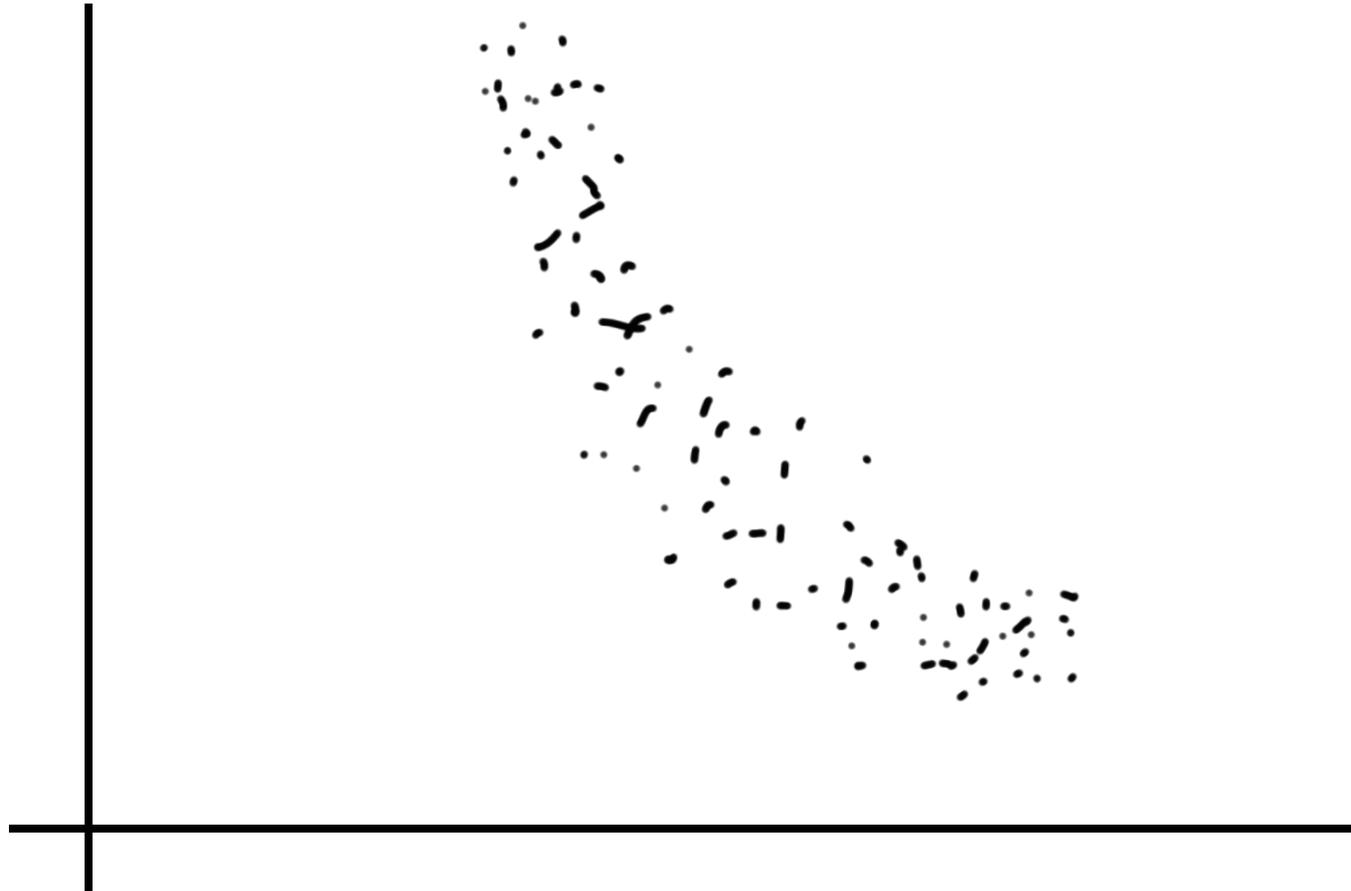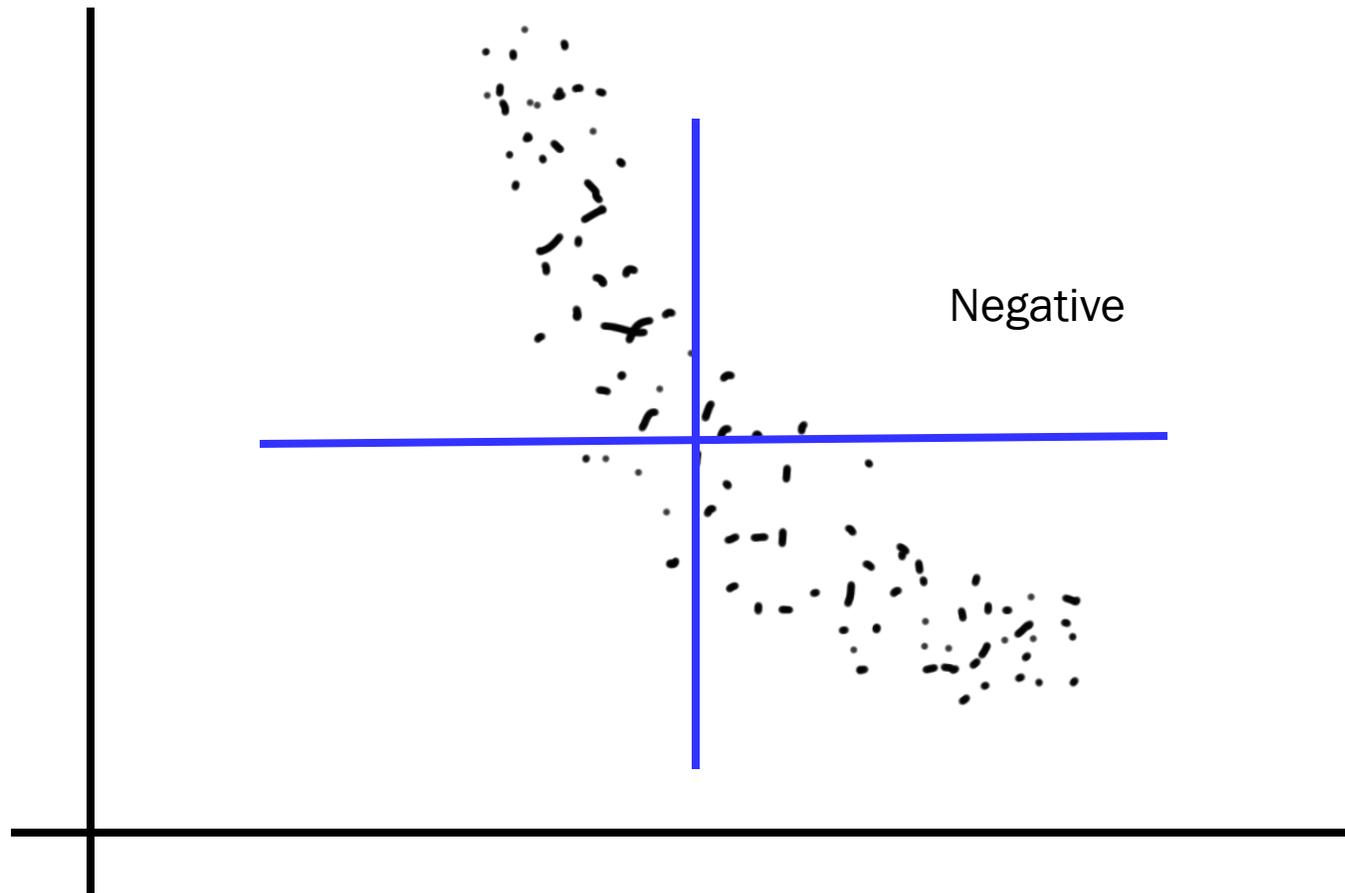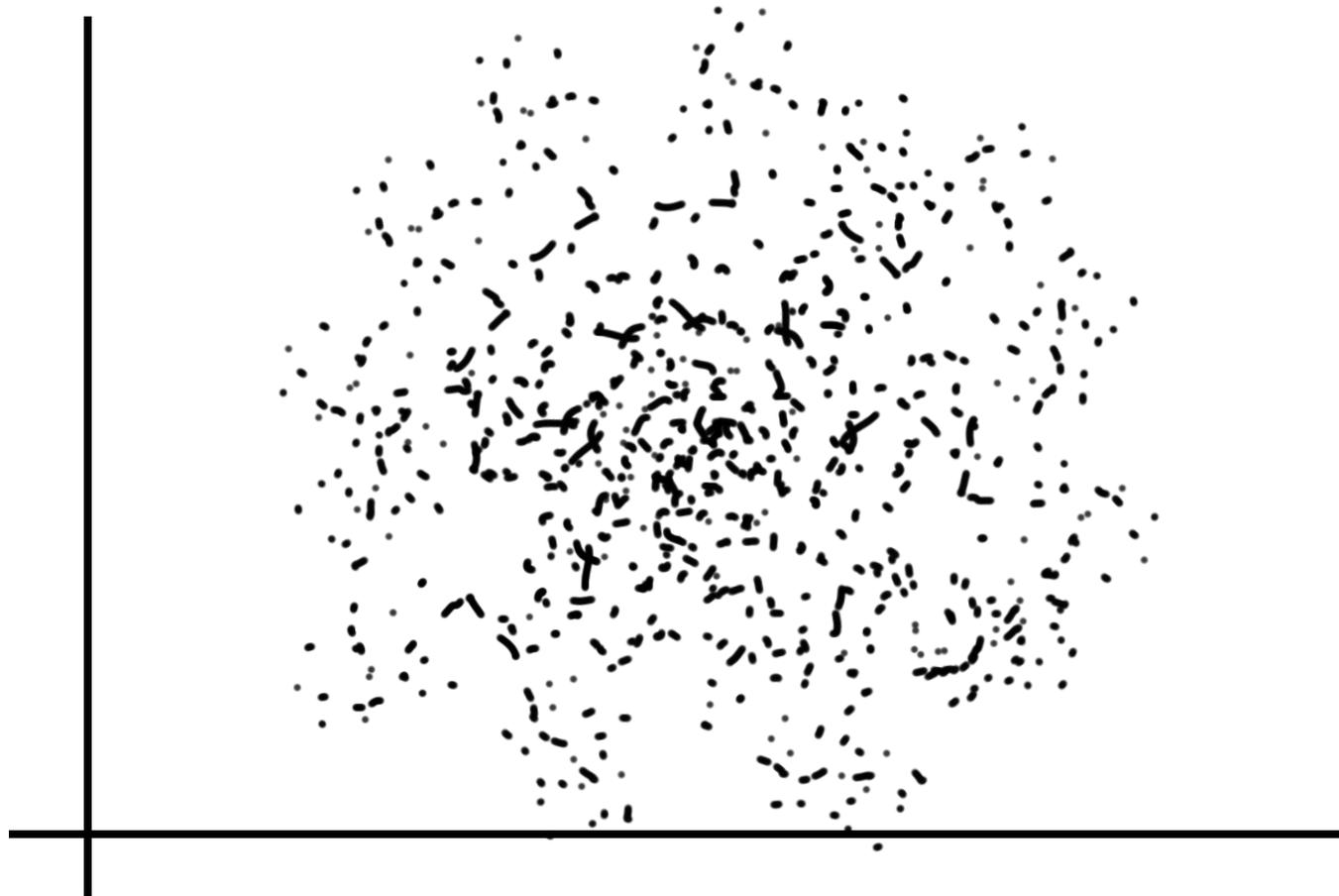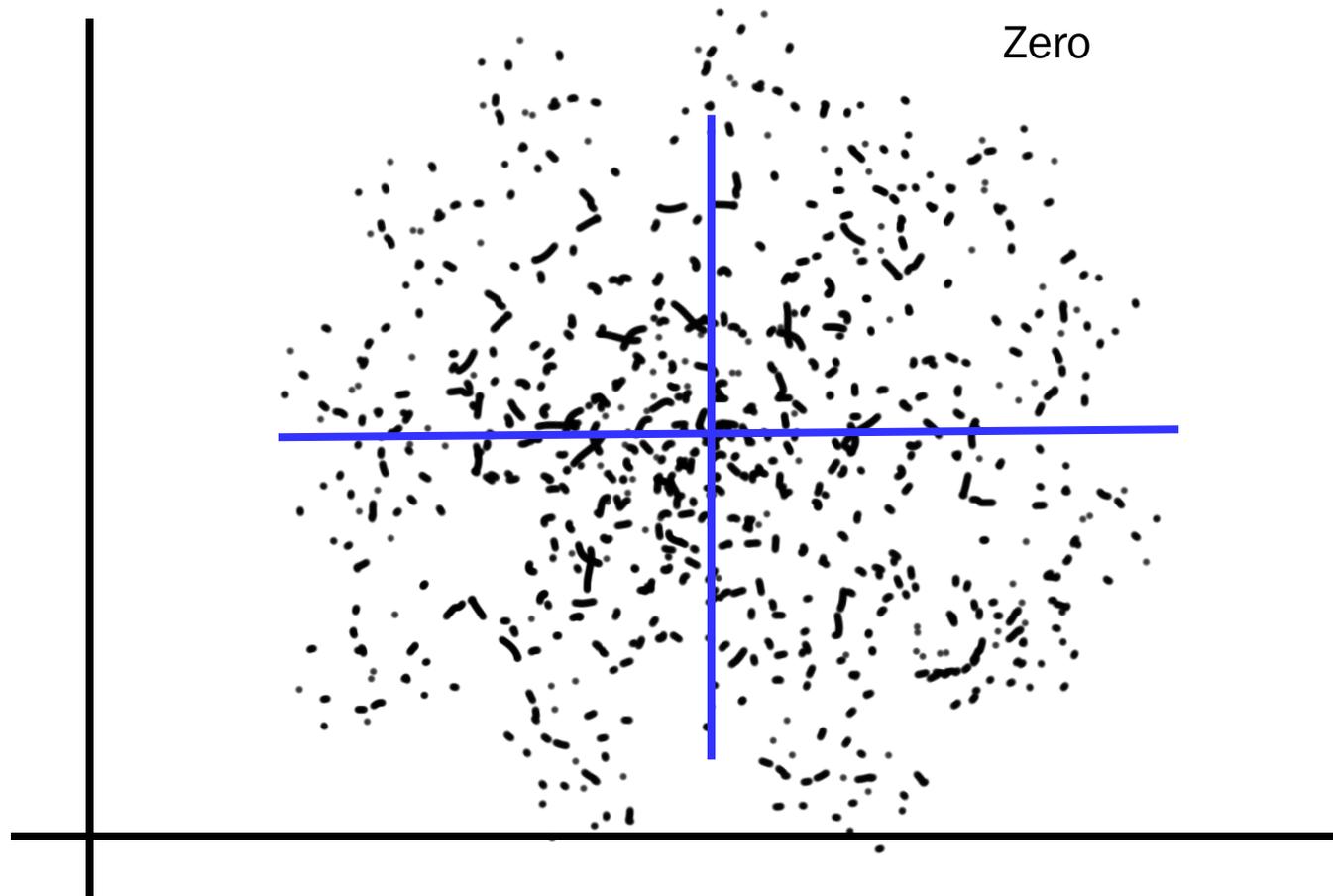Is the Covariance: (a) positive, (b) negative, (c) zero

Positive

Stanford University

Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero

Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero

Negative

Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero

Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero

Zero

Stanford University

Independence and Covariance

X and Y are random variables with PMF:

X \ Y	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	1

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

- E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0

- E[Y] = 0(2/3) + 1(1/3) = 1/3

- Since XY = 0, E[XY] = 0

- Cov(X, Y) = E[XY] – E[X]E[Y] = 0 – 0 = 0

But, X and Y are clearly dependent!

Properties of Covariance

Say X and Y are arbitrary random variables

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

- $\text{Cov}(X, X) = E[X^2] - E[X]E[X] = \text{Var}(X)$

- $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$

Correlation
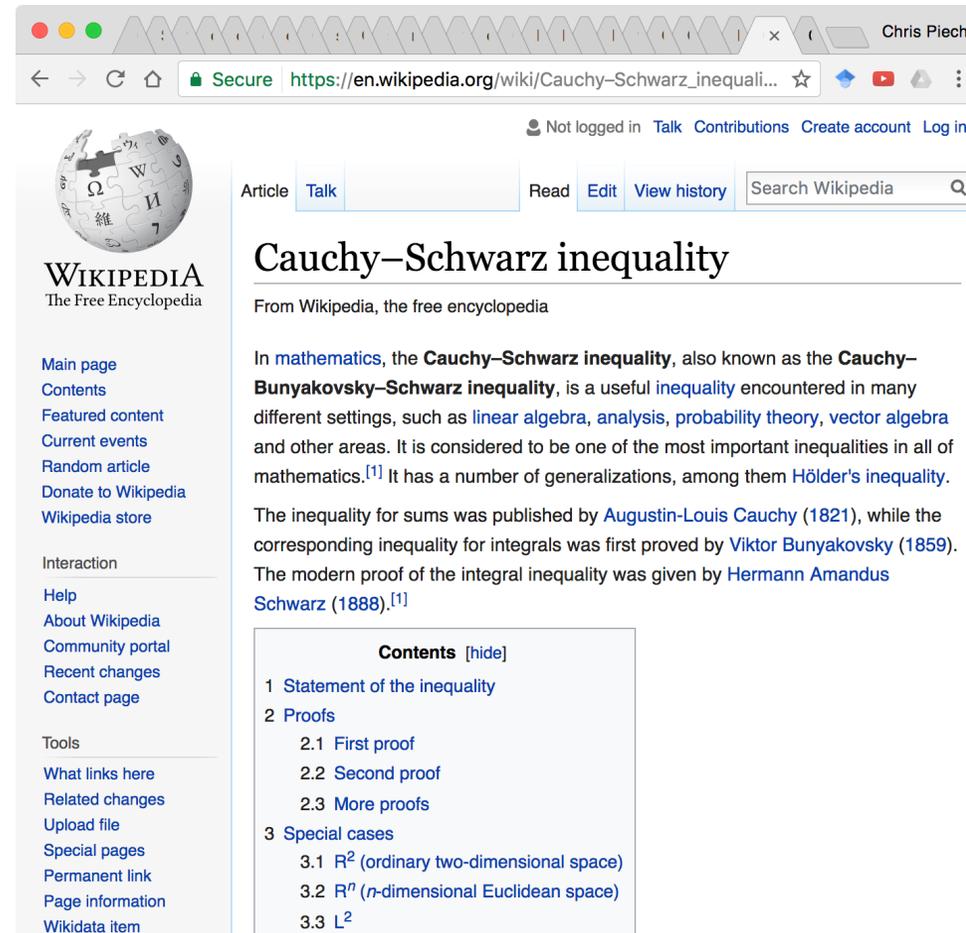
What is Wrong With This?

Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

E[W] = 62.75	E[H] = 52.75	E[W*H] = 3355.83
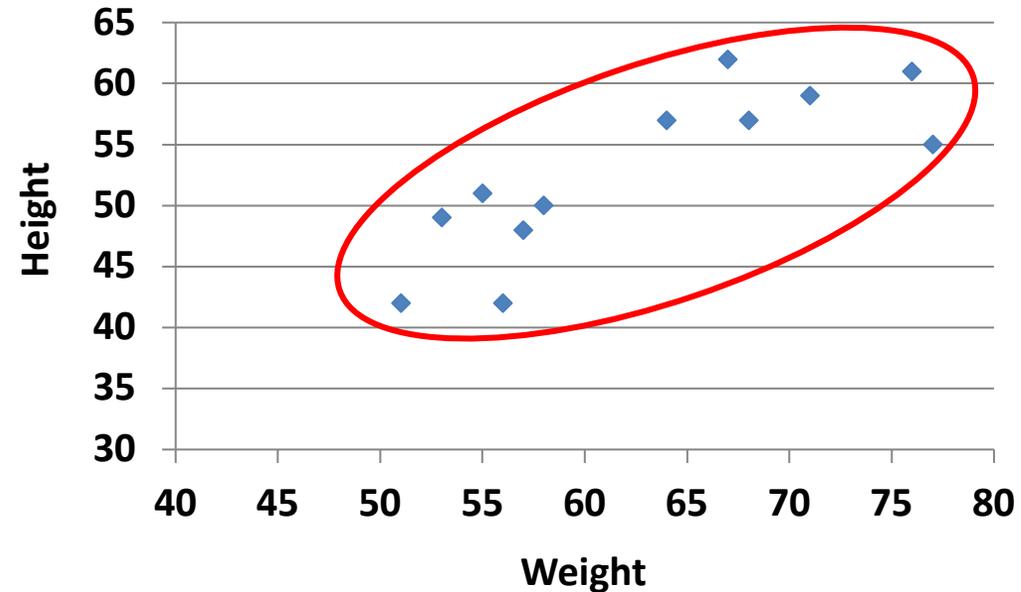
$$Cov(W, H) = E[W*H] - E[W]E[H]$$
$$= 3355.83 - (62.75)(52.75)$$
$$= 45.77$$

Cauchy Schwarz, a great way to normalize!

$$-\mathrm{Std}(X)\mathrm{Std}(Y) \leq \mathrm{Cov}(X,Y) \leq \mathrm{Std}(X)\mathrm{Std}(Y)$$

Viva La Correlatión

Say X and Y are arbitrary random variables

- Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Note: $-1 \leq \rho(X, Y) \leq 1$

- Correlation measures <u>linearity</u> between X and Y

- $\rho(X, Y) = 1 \qquad \Rightarrow \quad Y = aX + b \quad$ where $a = \sigma_y/\sigma_x$

- $\rho(X, Y) = -1 \qquad \Rightarrow \quad Y = aX + b \quad$ where $a = -\sigma_y/\sigma_x$

- $\rho(X, Y) = 0 \qquad \Rightarrow \quad$ absence of <u>linear</u> relationship
  - But, X and Y can still be related in some other way!

- If $\rho(X, Y) = 0$, we say X and Y are "uncorrelated"
  - Note: Independence implies uncorrelated, but <u>**not**</u> vice versa!

Viva La Correlatión

Say X and Y are arbitrary random variables

- Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Say Y = cX. Correlation should be 1.

| Music	Dance	Folk	Country	Classical music	Musical	Pop	Rock	Me...
5	2	1	2	2	1	5	5	
4	2	1	1	1	2	3	5	
5	2	2	3	4	5	3	5	
5	2	1	1	1	1	2	2	
5	4	3	2	4	3	5	3	
5	2	3	2	3	3	2	5	
5	5	3	1	2	2	5	3	
5	3	2	1	2	2	4	5	
5	3	1	1	2	4	3	5	
5	2	5	2	2	5	3	5	
5	3	2	1	2	3	4	3	
5	1	1	1	4	1	2	5	
5	1	2	1	4	3	3	4	
5	5	3	2	1	5	5	2	
5	2	1	1	2	3	4	5	
1	2	2	3	4	3	3	5	
5	3	1	1	1	2	4	4	
5	3	3	3	2	2	4	4	
5	5	4	3	4	5	5	4	
5	3	3	2	4	2	2	4	
5	3	2	3	4	3	2	5	
5	1	1	3	2	2	2	5	
5	3	2	3	3	3	4		
5	4	2	2	2	4	4	5	
5	3	1	1	4	3	3	5	
5	4	2	1	2	3	5	1	
5	5	5	4	5	3	4	4	
4	3	4	1	3	2	2	4	
5	5	1	1	1	1	3	4	
5	3	4	2	3	3	3	4	
4	4	3	3	3	3	4	4	
4	4	1	3	2	3	5	3	
5	3	1	3	2	3	3	4	
5	2	2	3	4	5	4	3	

Correlation of Music Tastes

Correlation of Music Tastes

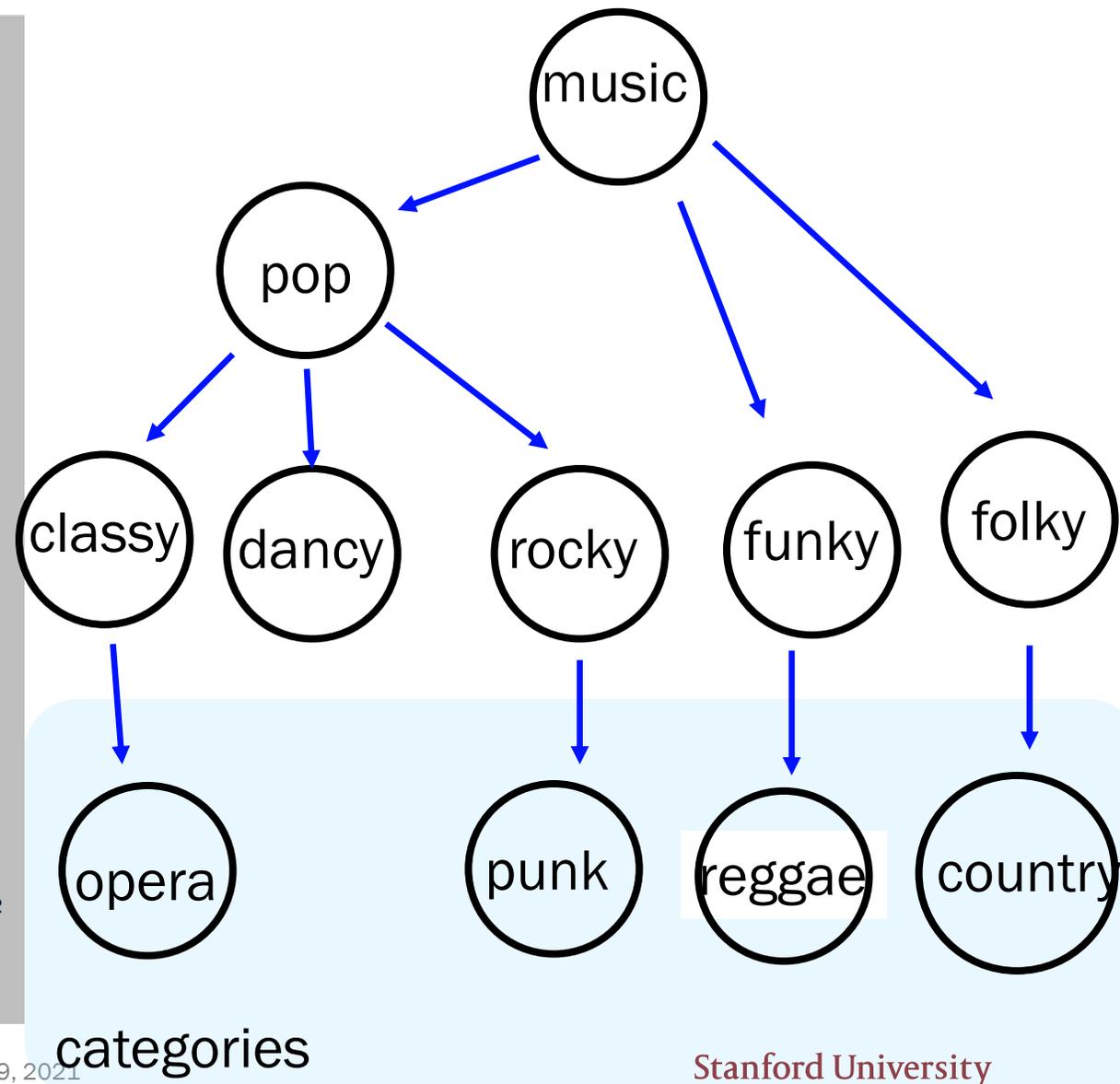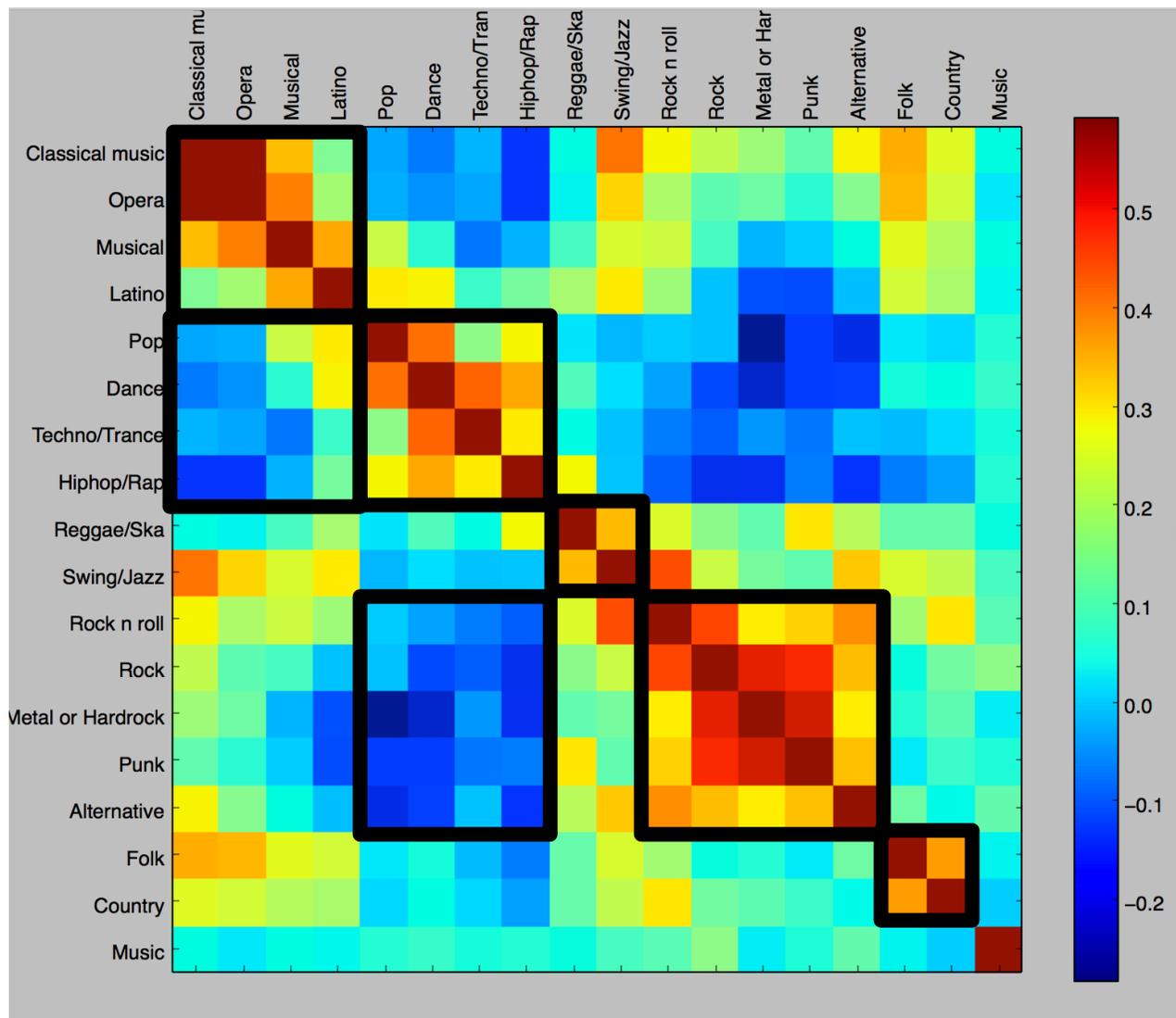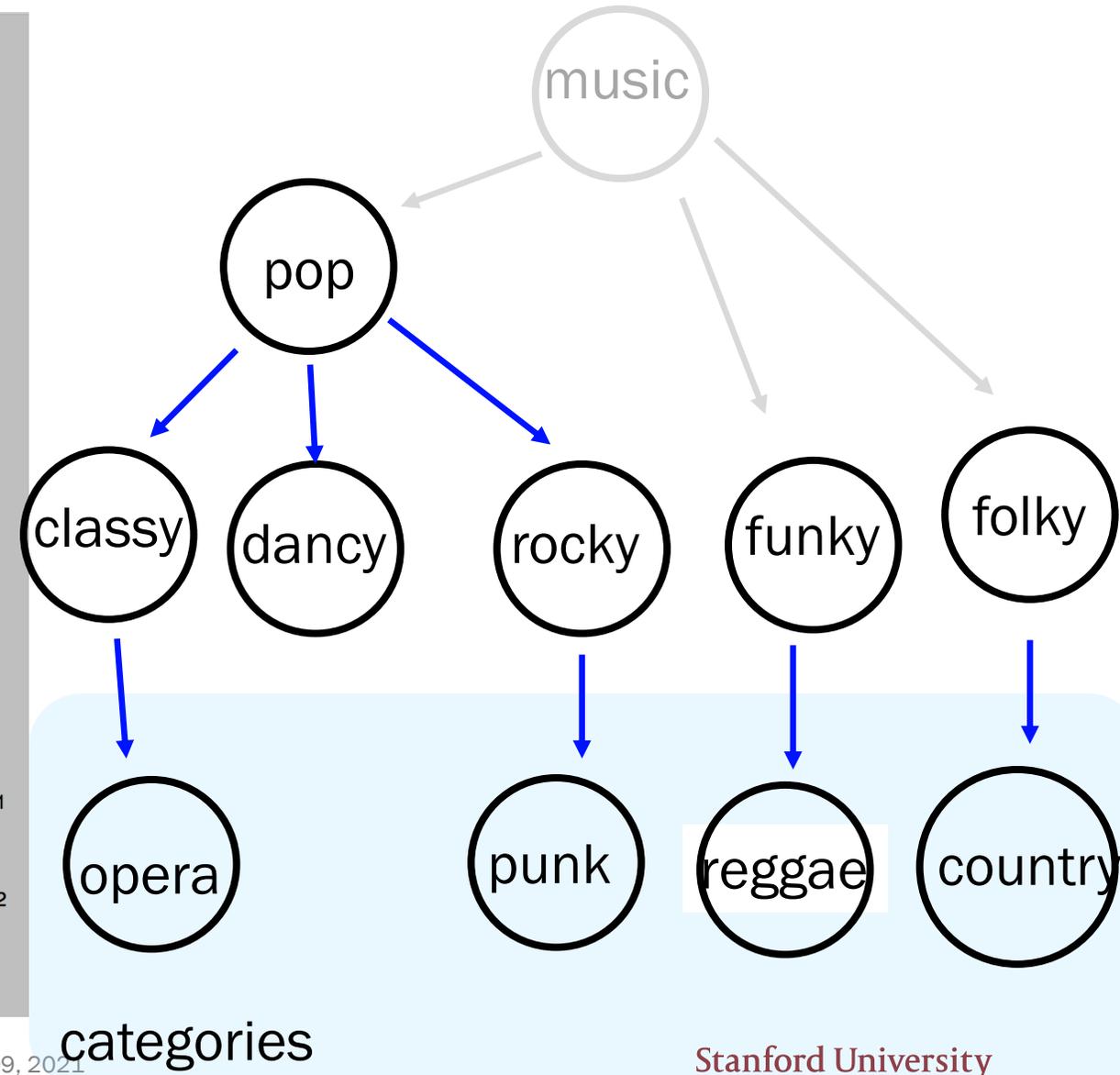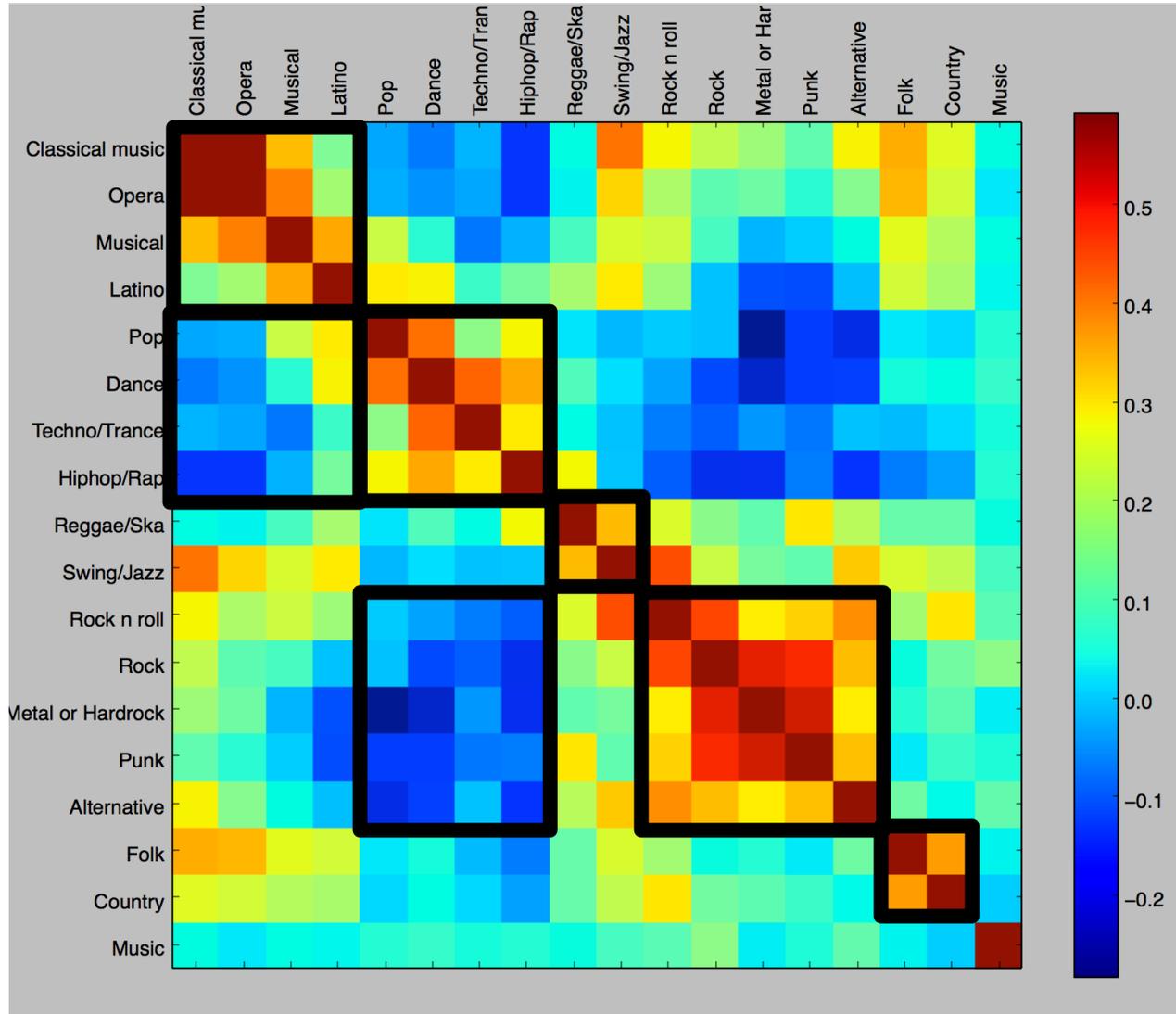# From Correlation to Bayes Net. Alternative!
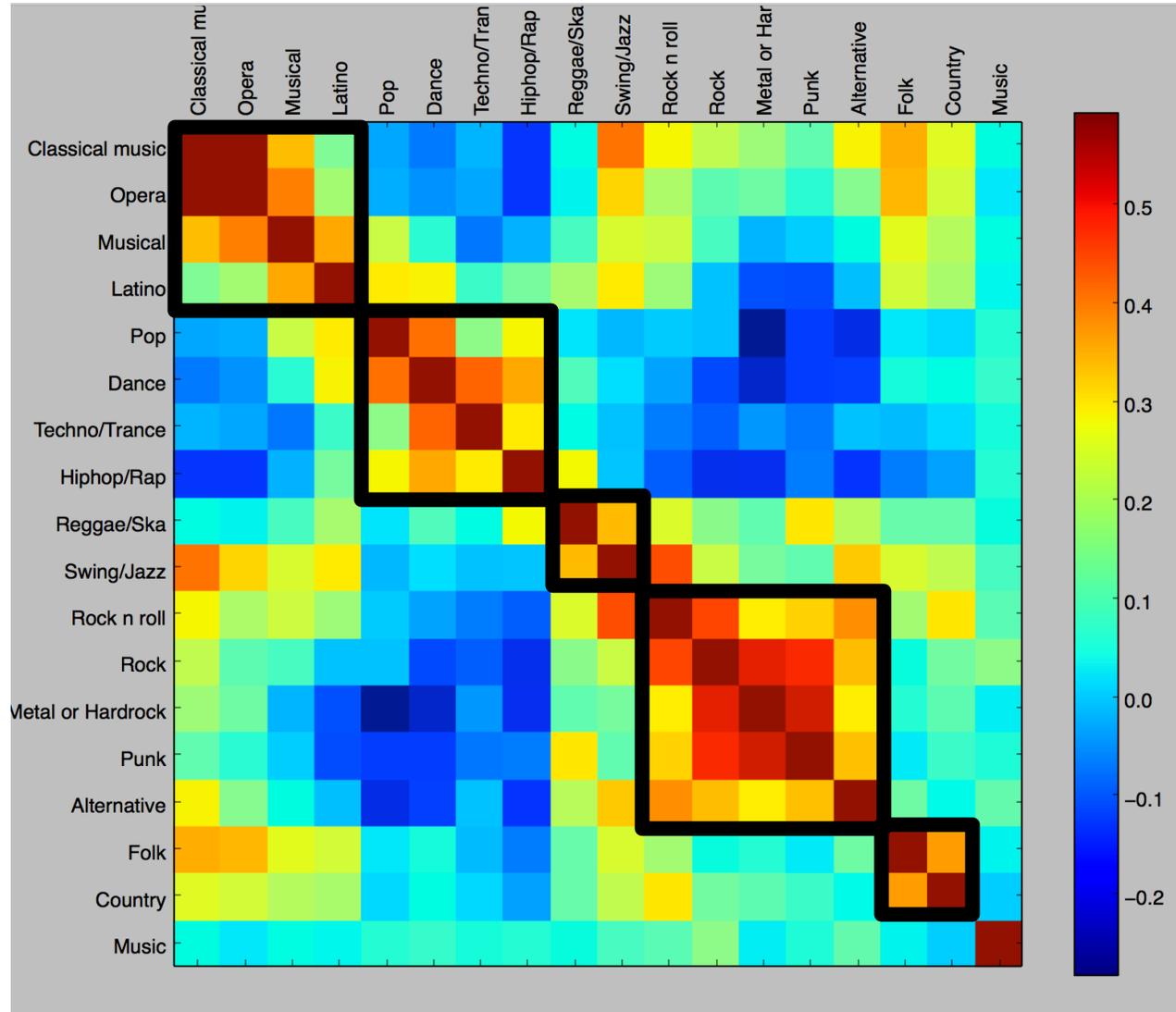
From Correlation to Bayes Net. Alternative!

From Correlation to Bayes Net. Alternative!

categories
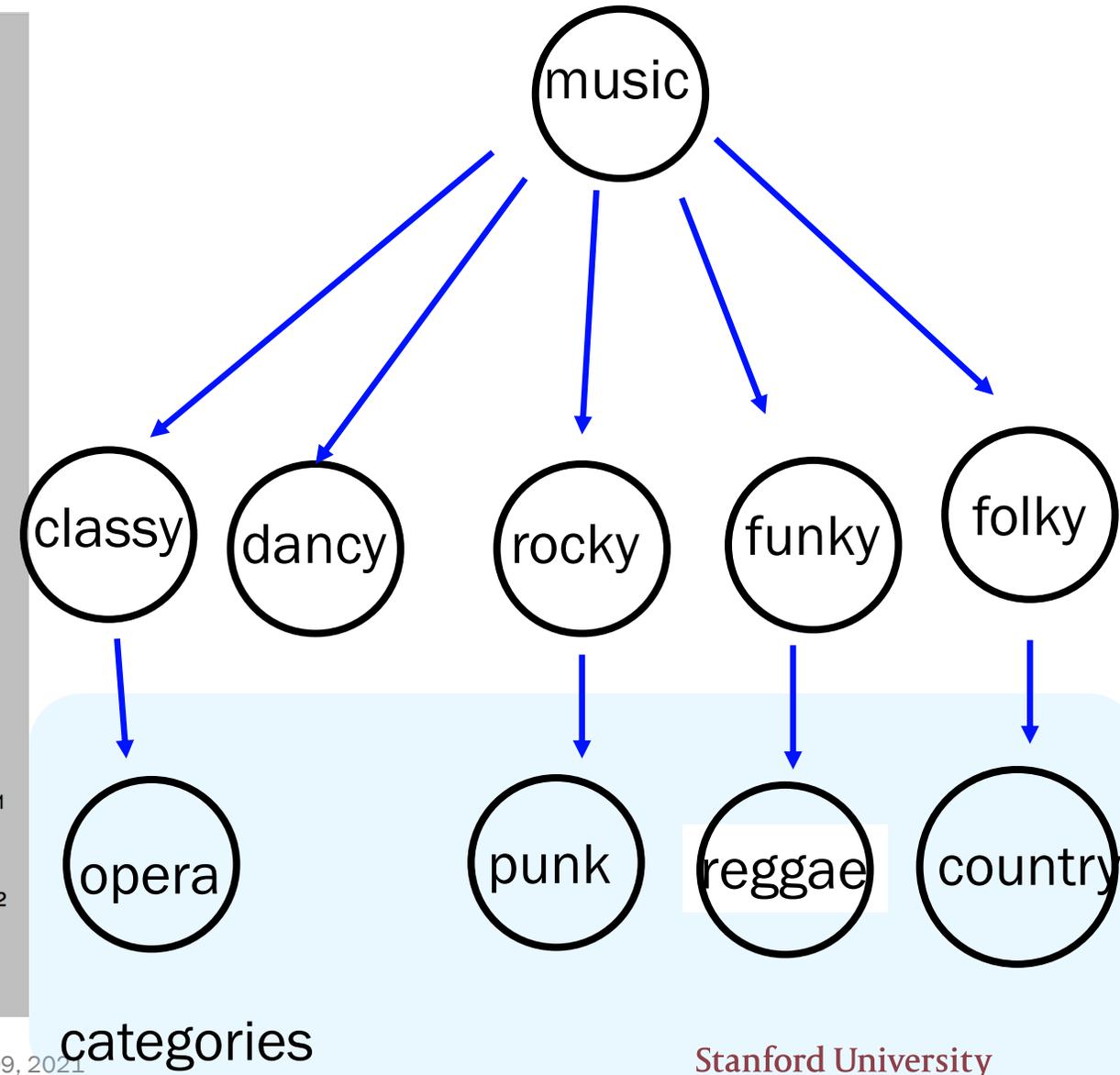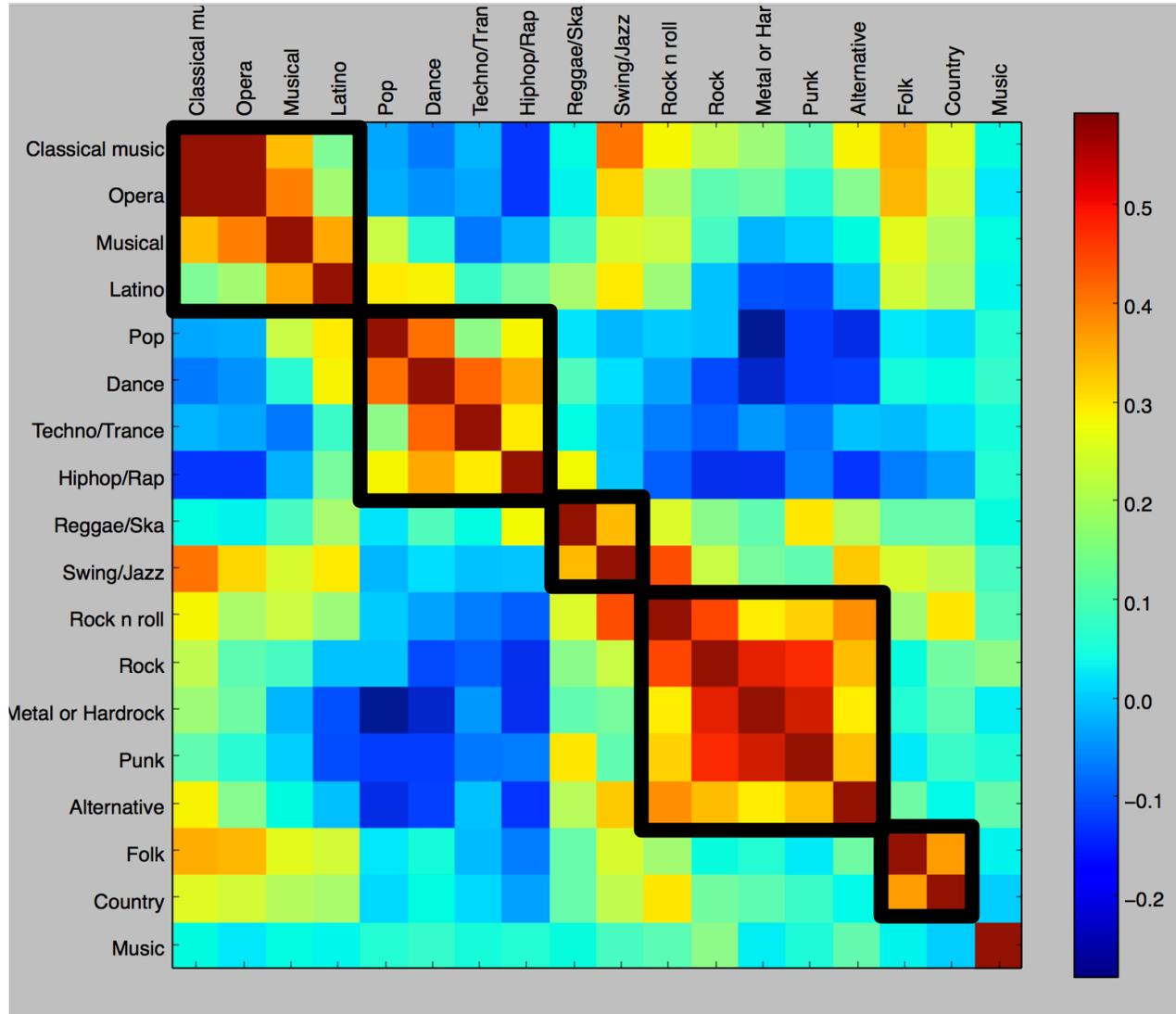
From Correlation to Bayes Net. Alternative!

How do you know if your model is good?

Answer: it is accurate at inference (especially tasks you care about)

Rock Music Vs Oil?

Hubbert Peak Theory

http://www.aei.org/publication/blog/
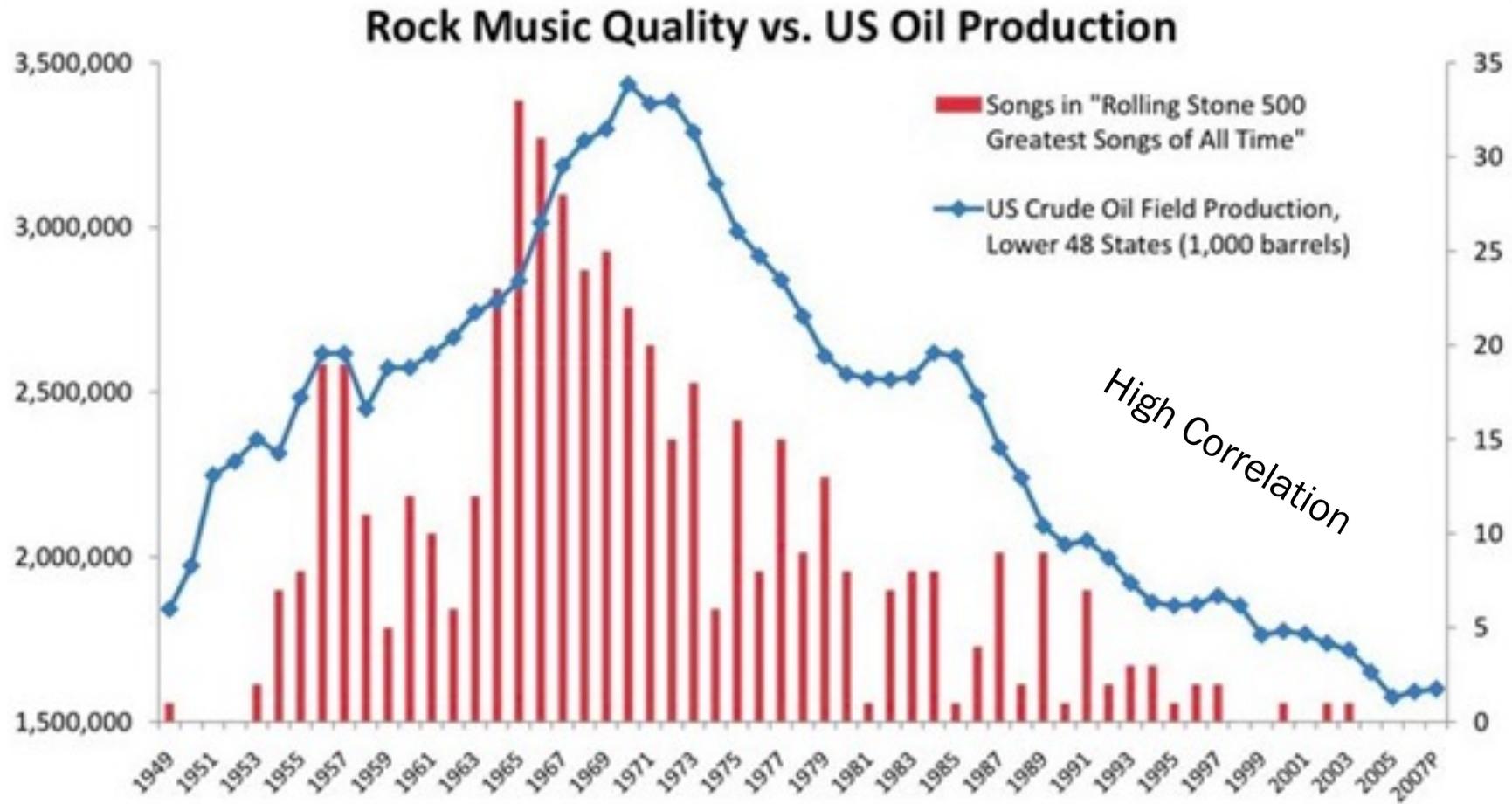
Tell your friends!
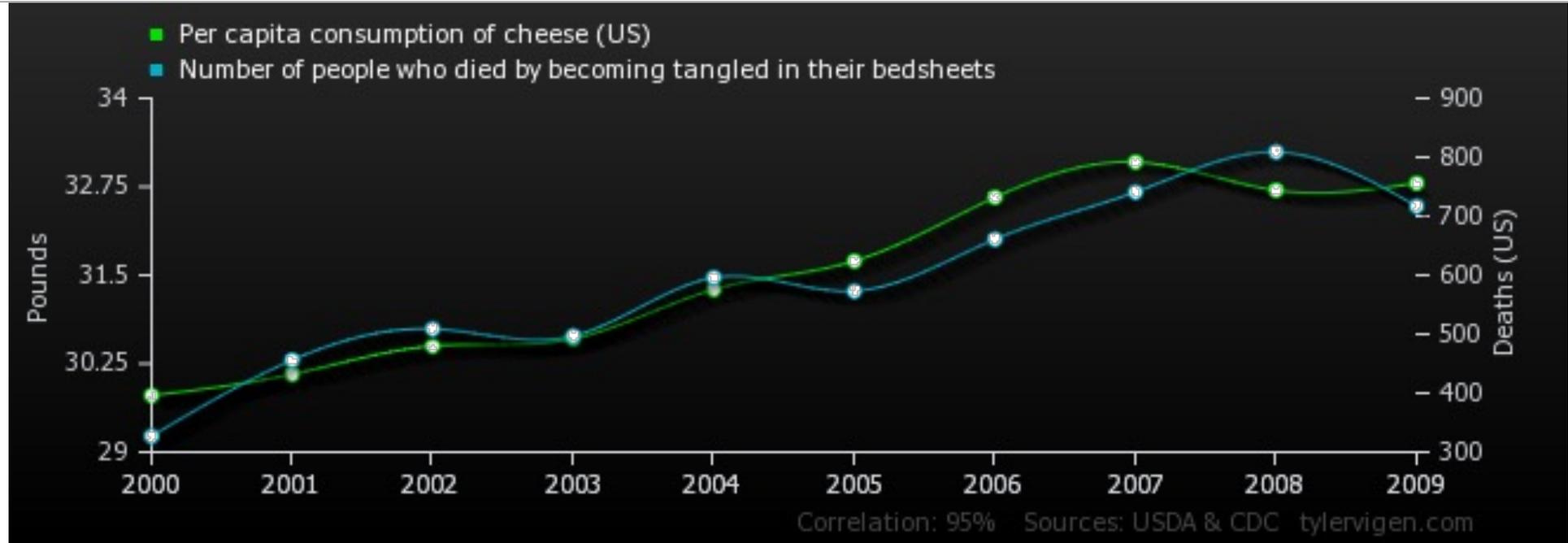
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

Stanford University

Divorce Vs Butter?
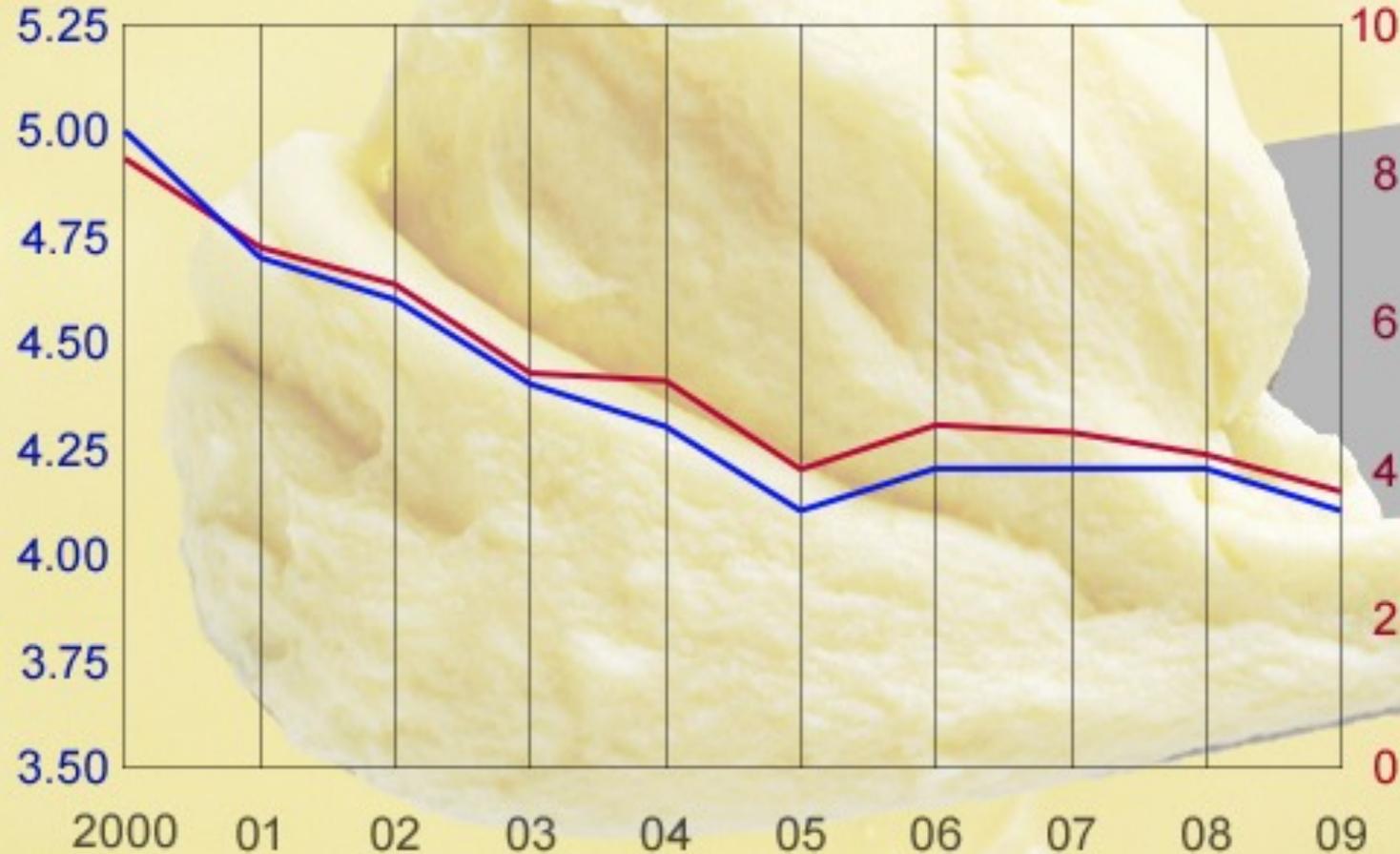
Source: US Census, USDA, tylervigen.com

http://www.bbc.com/news/magazine-27537142

What about Code.org?

Computers Couldn't Understand Code

60,000 students attempted this problem
37,000 unique solutions
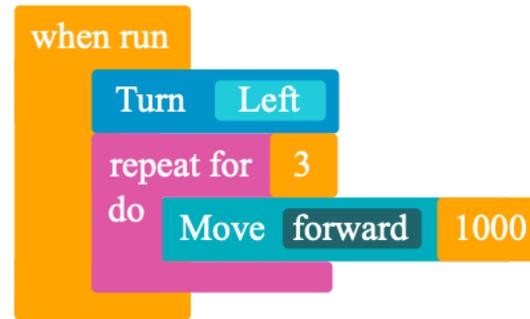
Challenge

Student Code

You need to move and turn in your loop

Insight

Computers Couldn't Understand Code

Code.org Short Code (10 lines)

Feedback F1 Score

Cond. Prob. Deep Learning Generative Model Humans

Stanford University
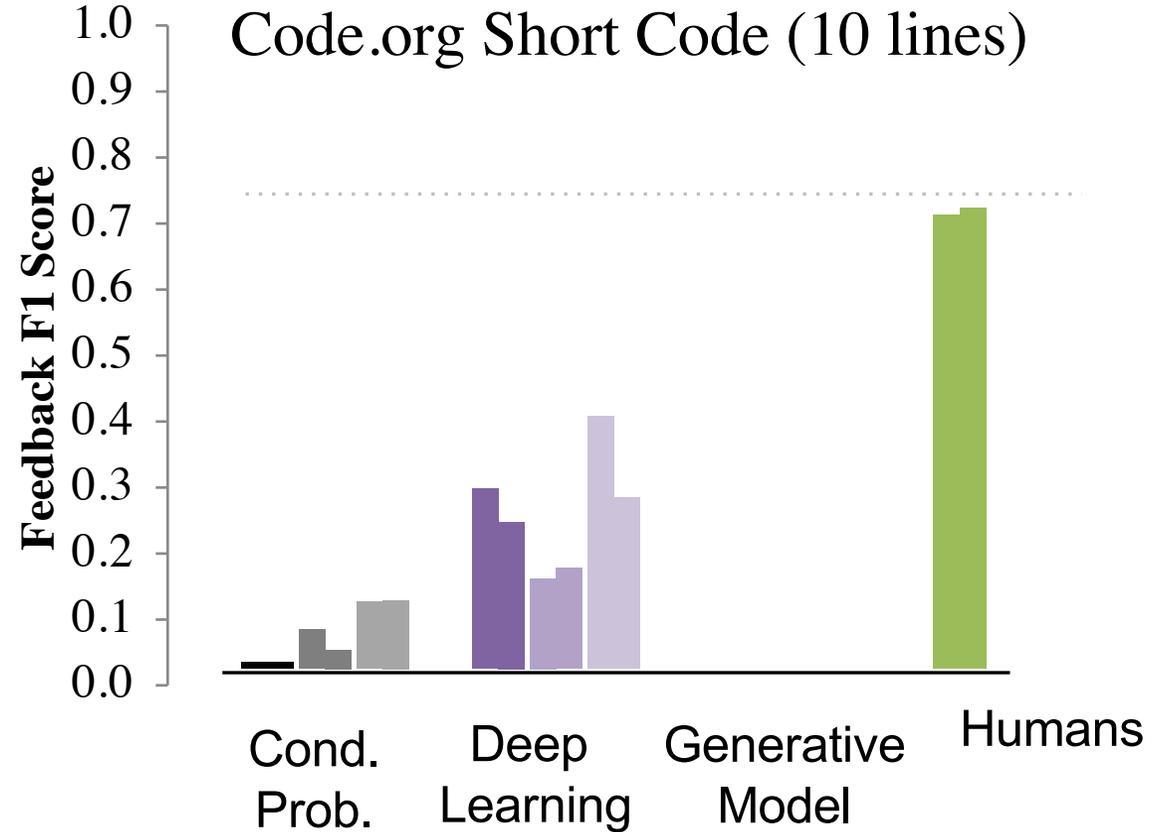
Code is hard to think about...

Code Zipf Plot

$$f(k) = \frac{1/k^s}{\sum_n^N (1/n^s)}$$

Probability Mass (log scale) vs Rank Order (log scale)
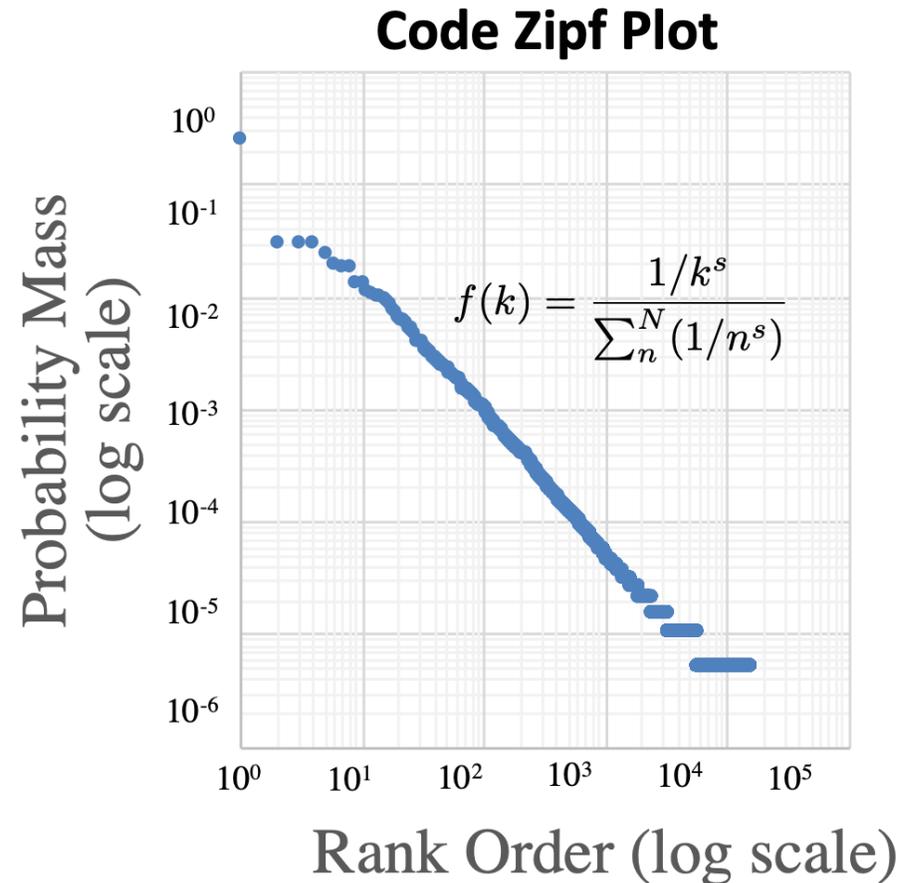
Exponential combination of decisions.
Super fat tailed. Everything looks unique

Generative Model is Easier than Straight Inference
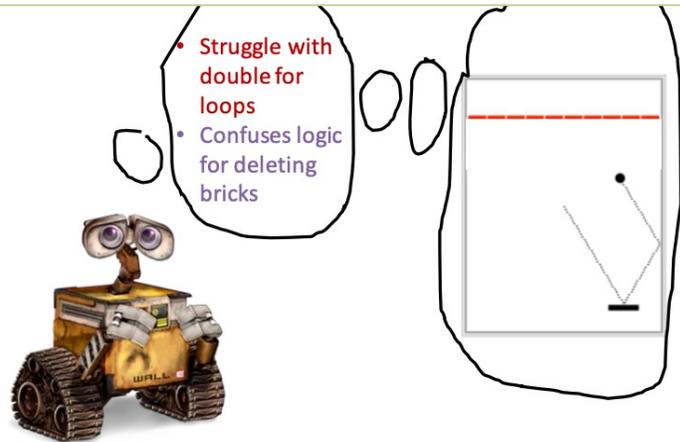
This is easy and exponential

This is hard and linear

A students "ability"
$$\Theta \sim \text{pythonSample}$$

A students "choices"
$$C \sim \text{pythonSample}|\Theta$$

The resulting code
$$\Pi \sim \text{pythonSample}|C$$

$$P(\Theta, C|\Pi)$$

Infer ability and choices from code

Made a probabilistic programming language:

ideaToText

Write a model of each **decision point**, given parents

```python
 7
 8    # This python class is a RubricSampling Decision
 9    # it generates programs that print the numbers 10 -> 1
10   class Countdown(Decision):
11
12     def registerChoices(self):
13        # these are the main strategies for printing out a
14        # countdown
15        self.addRubricChoices('loop-style', {
16           'for' : 100,
17           'while' : 40,
18           'none' : 1,
19           'empty' : 1
20        })
21
22     # we can make some grading choices based on which
23     # strategy they chose (did they actually use a loop?)
24     def processChoices(self):
25        style = self.getChoice('loop-style')
26        hasLoop = style != 'none' and style != 'empty'
27        self.addLabel('rubric-hasLoop', hasLoop)
28        self.addLabel('rubric-printsNums', style != 'empty')
29
30     # Based on their strategy render a different decision
31     def renderCode(self):
32        style = self.getChoice('loop-style')
33        if style == 'for': return '{ForSoln}'
34        if style == 'while': return '{WhileSoln}'
35        if style == 'none': return '{NoLoopSoln}'
36        if style == 'empty': return ''
37
```
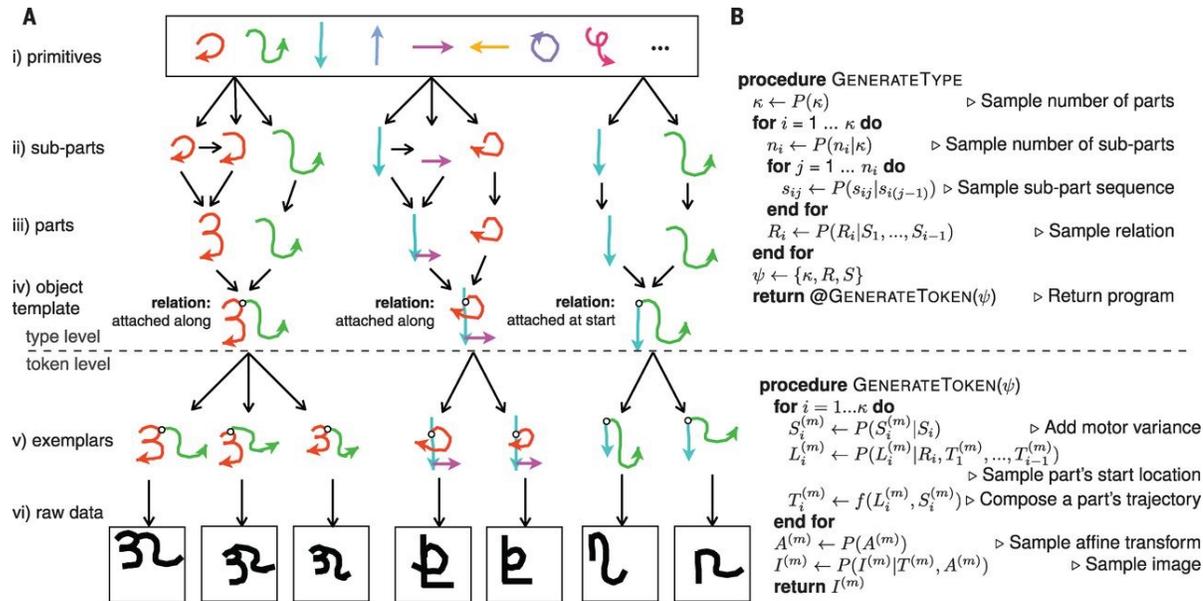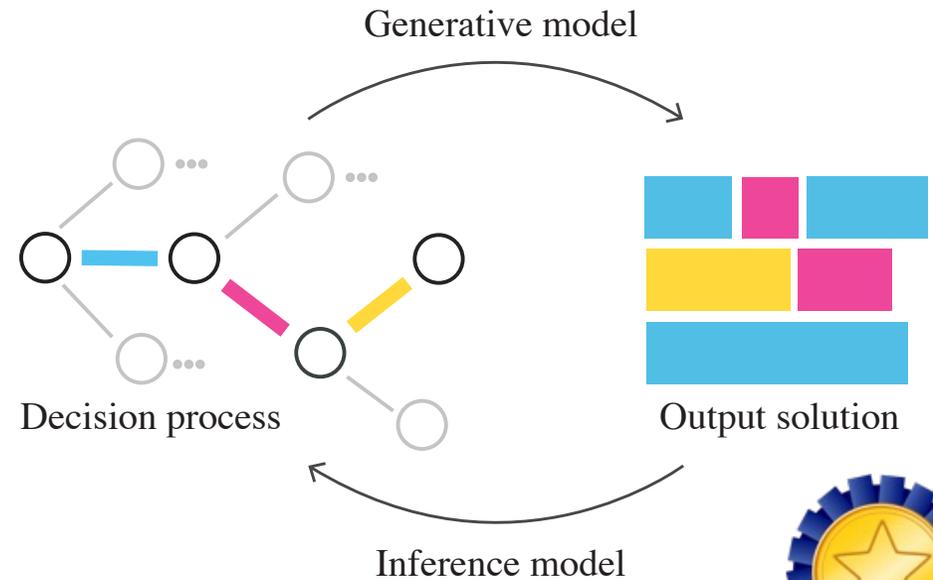
1. This is code for a single decision point

2. Give a name to the choice that the student is making

3. How do those choices translate into grades?

4. What does the code look like? Often evokes other decision points
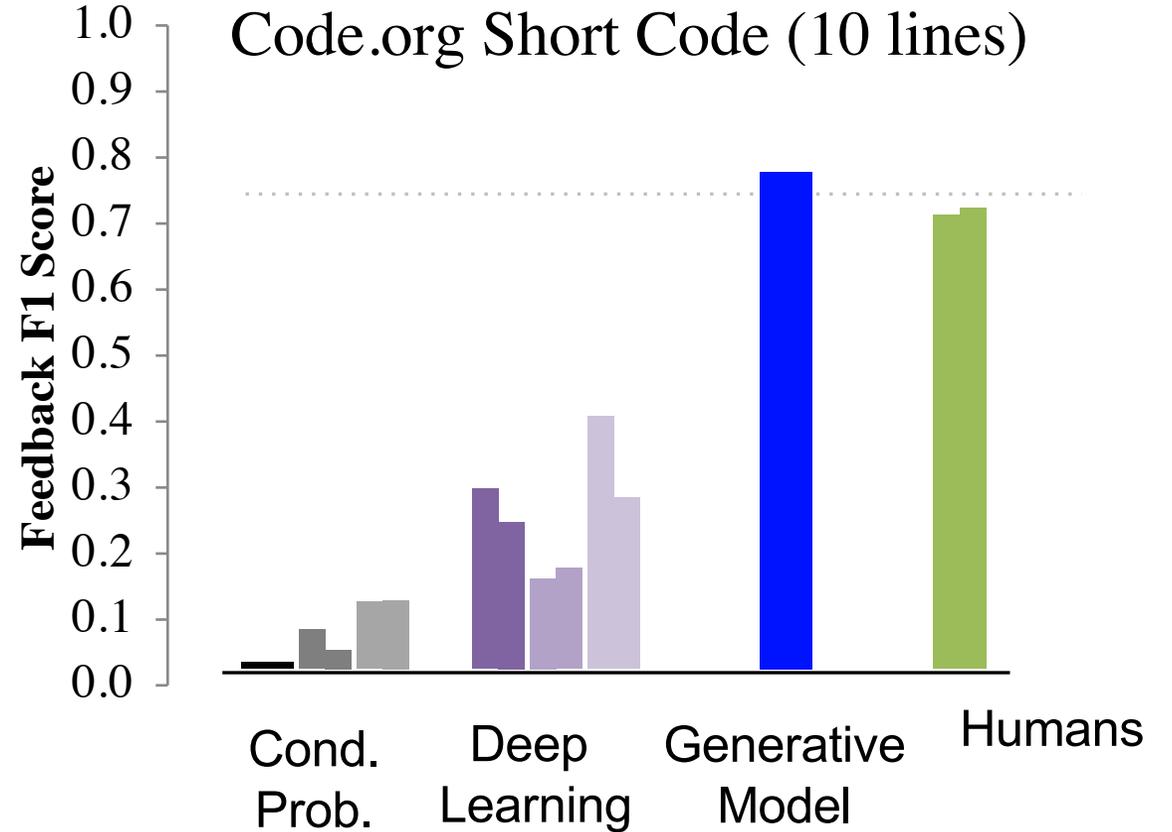
Generative Model of Grading

Lake et al, 2015

Muke Wu, Ali Malik, Noah Goodman, Chris Piech, 2019

Outstanding Student paper award, AAAI 2019

Stanford University
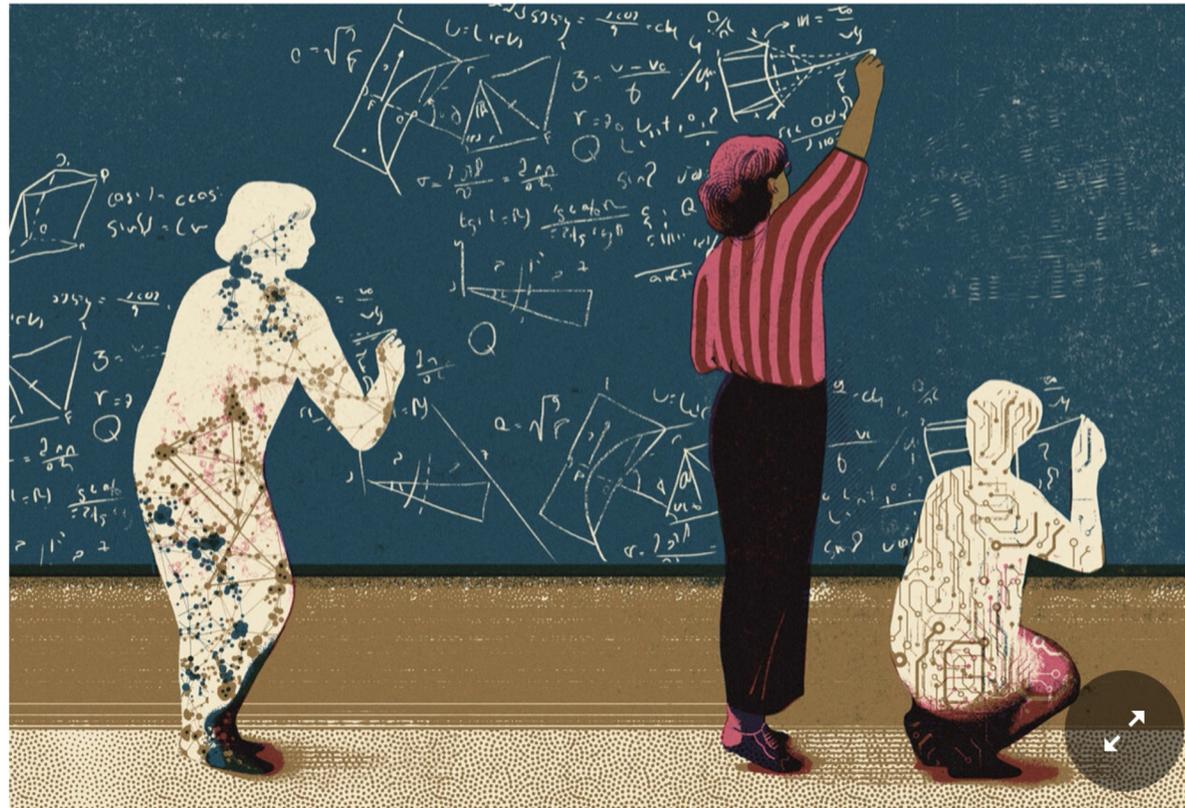
Computers Couldn't Understand Code

Code.org Short Code (10 lines)

Can A.I. Grade Your Next Test?

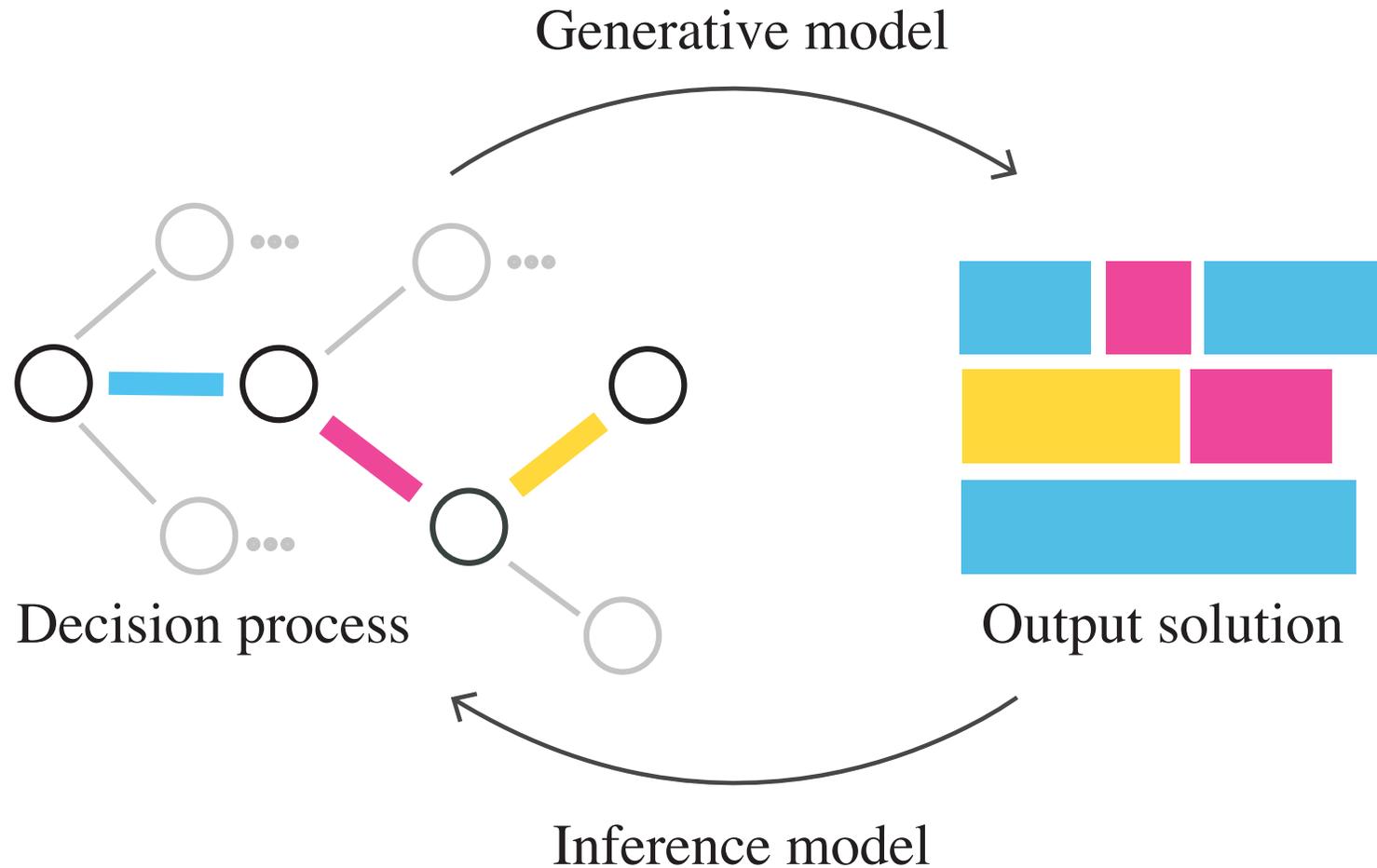Neural networks could give online education a boost by providing automated feedback to students.

Friday: General Inference

Generative model

Decision process

Output solution

Inference model

Que te vayas bien

No Class on Monday

This is how I imagine students are when they are studying ☺

(recall why we are here. Learning is a joy. You are growing)