Review: probability

- Monty Hall, weighted dice
- Frequentist v. Bayesian
- Independence
- Expectations, conditional expectations
  - Exp. & independence; linearity of exp.
- Estimator (RV computed from sample)
  - law of large #s, bias, variance, tradeoff

Covariance

- Suppose we want an approximate numeric measure of (in)dependence
- Let $E(X) = E(Y) = 0$ for simplicity
- Consider the random variable XY
  - if X, Y are typically both +ve or both -ve
  - if X, Y are independent

Covariance

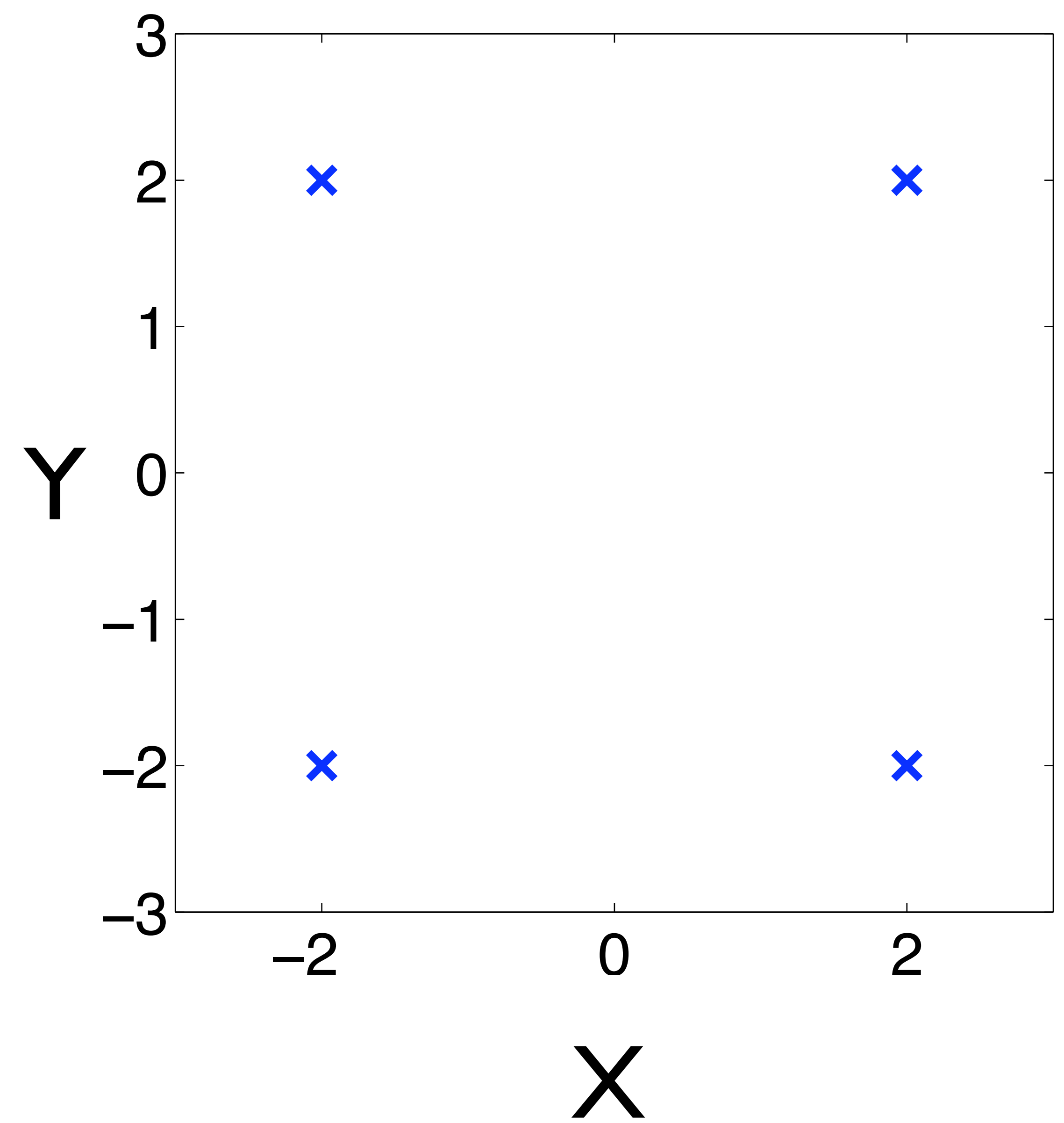
- cov(X, Y) =
- Is this a good measure of dependence?
  - Suppose we scale X by 10:

Correlation

- Like covariance, but controls for variance of individual r.v.s
- cor(X, Y) =
- cor(10X, Y) =

Correlation & independence

- Equal probability on each point
- Are X and Y independent?
- Are X and Y uncorrelated?



Correlation & independence

- Do you think that all independent pairs of RVs are uncorrelated?

- Do you think that all uncorrelated pairs of RVs are independent?

Proofs and counterexamples

- For a question $A \overset{?}{\Rightarrow} B$
  - e.g., X,Y uncorrelated $\overset{?}{\Rightarrow}$ X,Y independent
  - if true, usually need to provide a ***proof***
  - if false, usually only need to provide a ***counterexample***

Counterexamples

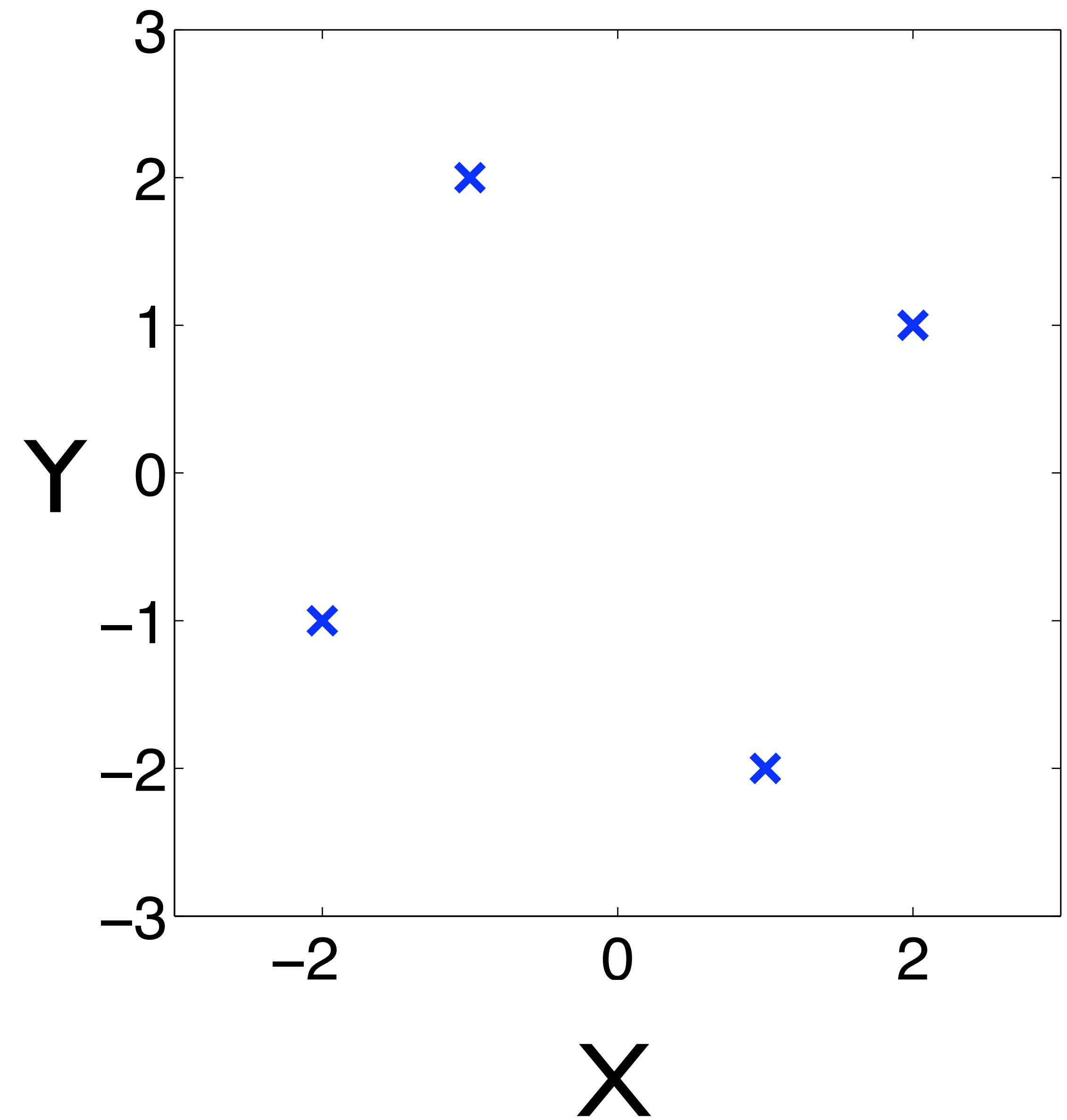
$$A \overset{?}{\Rightarrow} B$$

$$X, Y \text{ uncorrelated} \overset{?}{\Rightarrow} X, Y \text{ independent}$$

- Counterexample = example satisfying A but not B
- E.g., RVs X and Y that are ***not*** independent, but ***are*** correlated

Correlation & independence

- Equal probability on each point
- Are X and Y independent?
- Are X and Y uncorrelated?



Bayes Rule

Rev. Thomas Bayes
1702–1761

- For any X, Y, C
  - P(X | Y, C) P(Y | C) = P(Y | X, C) P(X | C)
- Simple version (without context)
  - P(X | Y) P(Y) = P(Y | X) P(X)
- Can be taken as definition of conditioning

Exercise

- You are tested for a rare disease, emacsitis—prevalence 3 in 100,000
- Your receive a test that is 99% ***sensitive*** and 99% ***specific***
  - sensitivity = P(yes | emacsitis)
  - specificity = P(no | ~emacsitis)
- The test comes out ***positive***
- Do you have emacsitis?

Revisit: weighted dice

- Fair dice: all 36 rolls equally likely
- Weighted: rolls summing to 7 more likely
- Data: 1-6 2-5

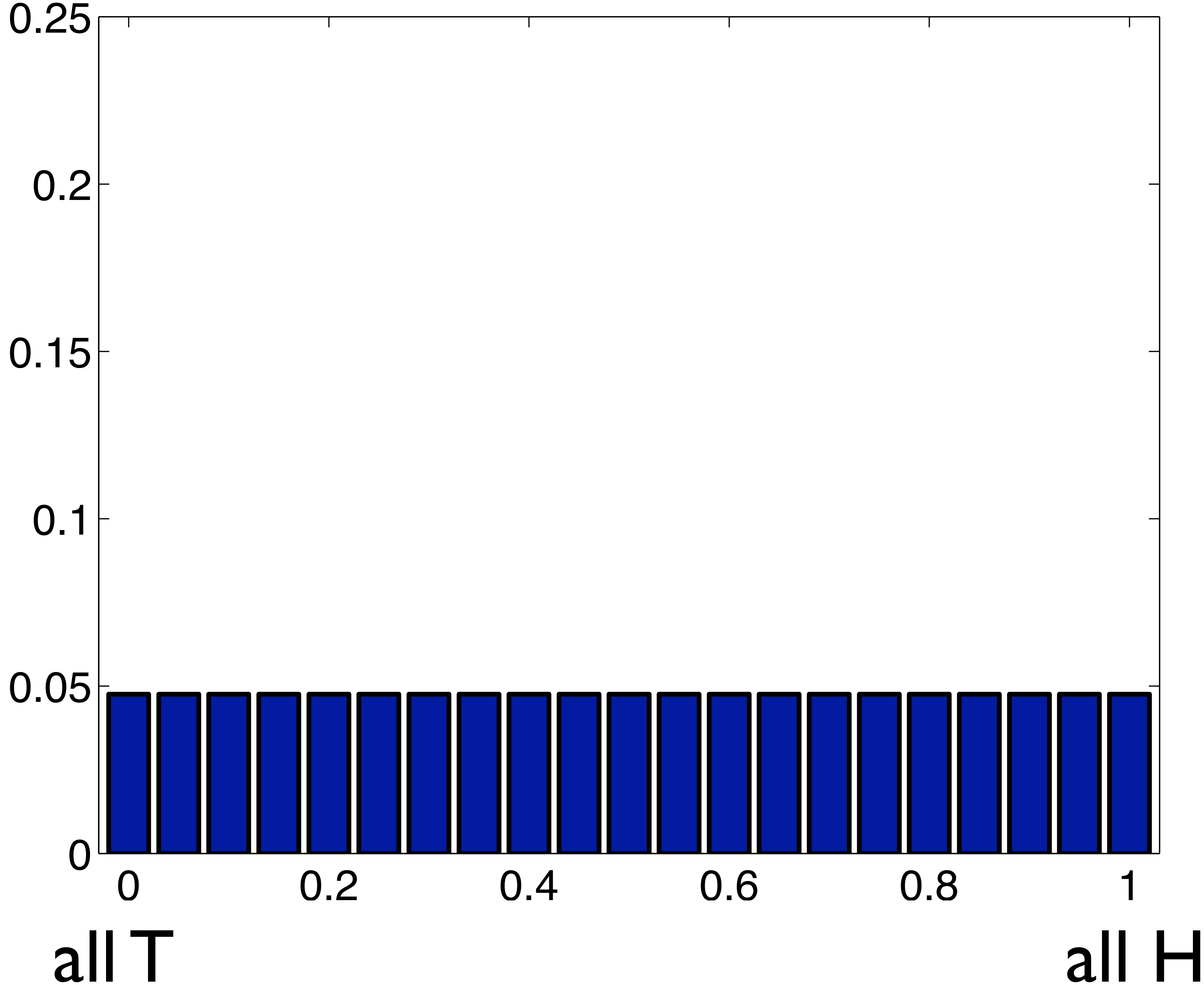
Learning from data

- Given a ***model class***
- And some data, sampled from a model in this class
- Decide which model best explains the sample

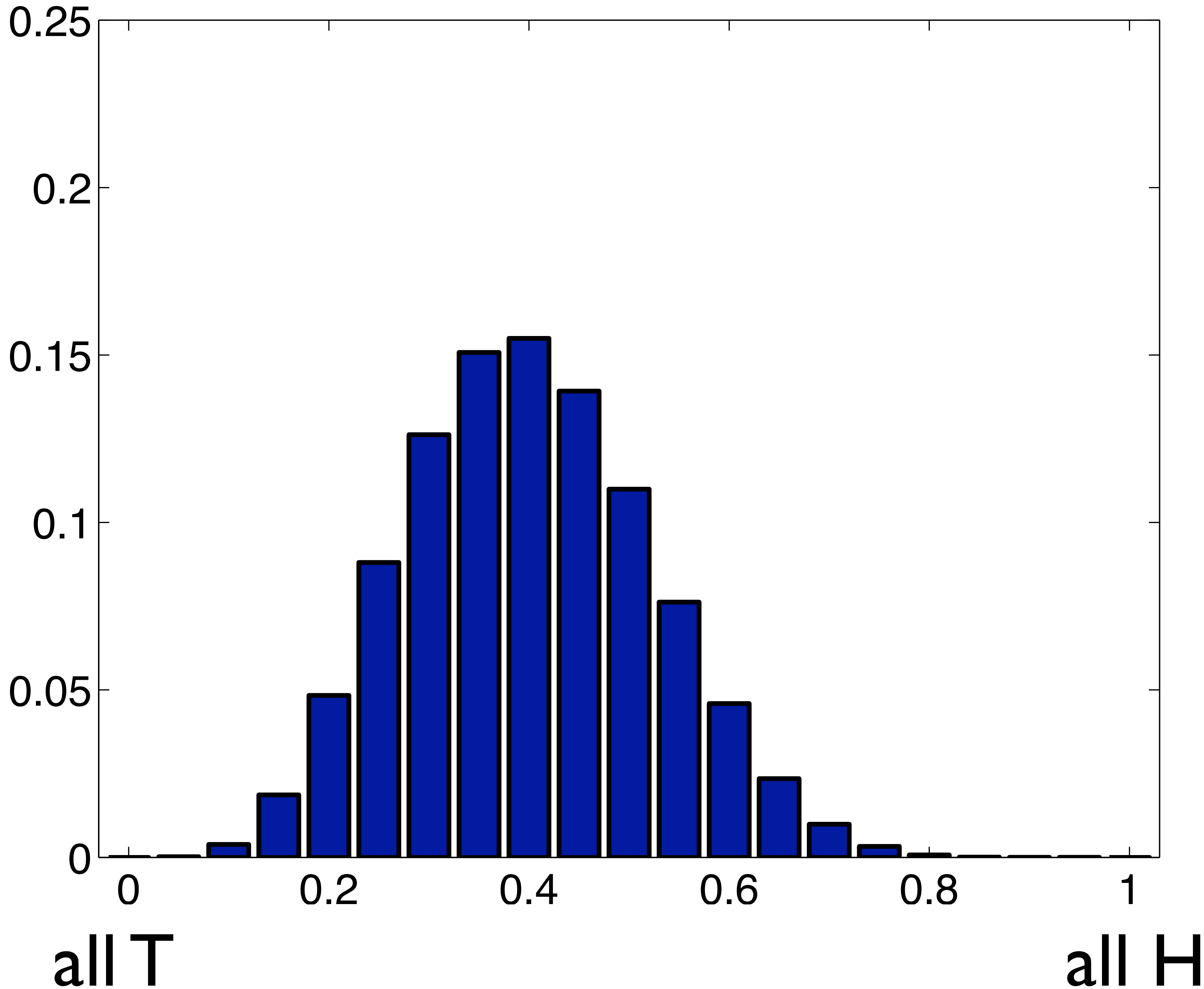
Bayesian model learning

- P(model | data) =
- Z =
- So, for each model, compute:
- Then:

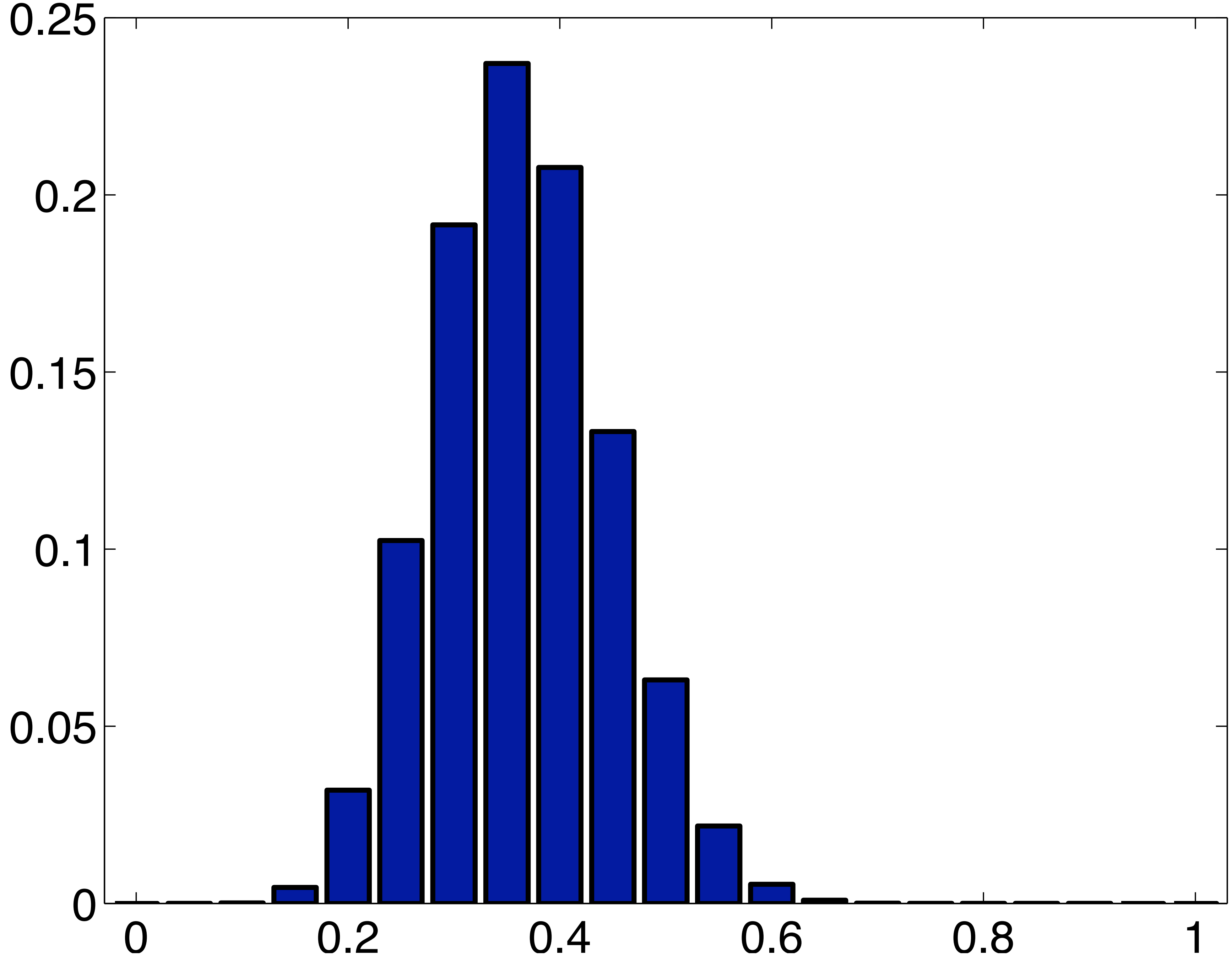
Prior: uniform



Posterior: after 5H, 8T



Posterior: 11H, 20T



all T all H

Graphical models

Why do we need graphical models?

- So far, only way we've seen to write down a distribution is as a big table
- Gets unwieldy fast!
    - E.g., 10 RVs, each w/ 10 settings
    - Table size =
- Graphical model: way to write distribution compactly using diagrams & numbers

Example ML problem

- US gov't inspects food packing plants
  - 27 tests of contamination of surfaces
  - 12-point ISO 9000 compliance checklist
  - are there food-borne illness incidents in 30 days after inspection? (15 types)
- Q:
- A:

Big graphical models

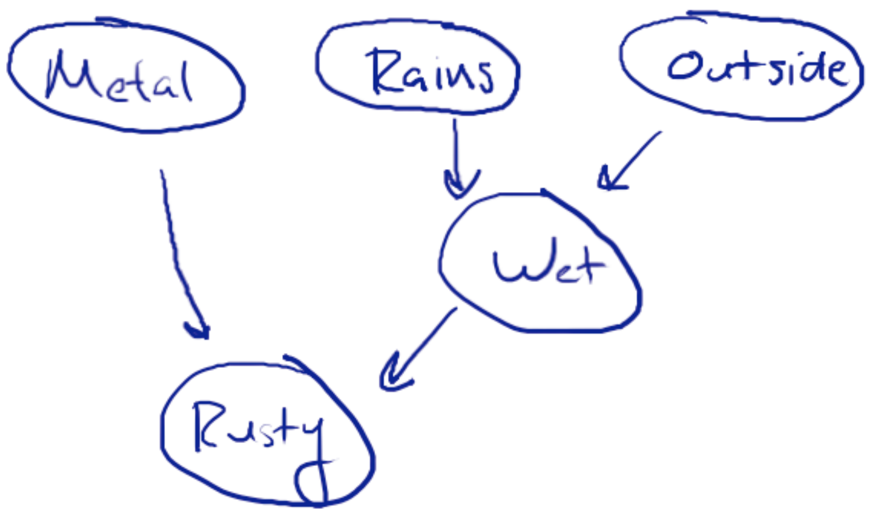
- Later in course, we'll use graphical models to express various ML algorithms
  - e.g., the one from the last slide
- These graphical models will be big!
- Please bear with some smaller examples for now so we can fit them on the slides and do the math in our heads…

Bayes nets

- Best-known type of graphical model
- Two parts: DAG and CPTs

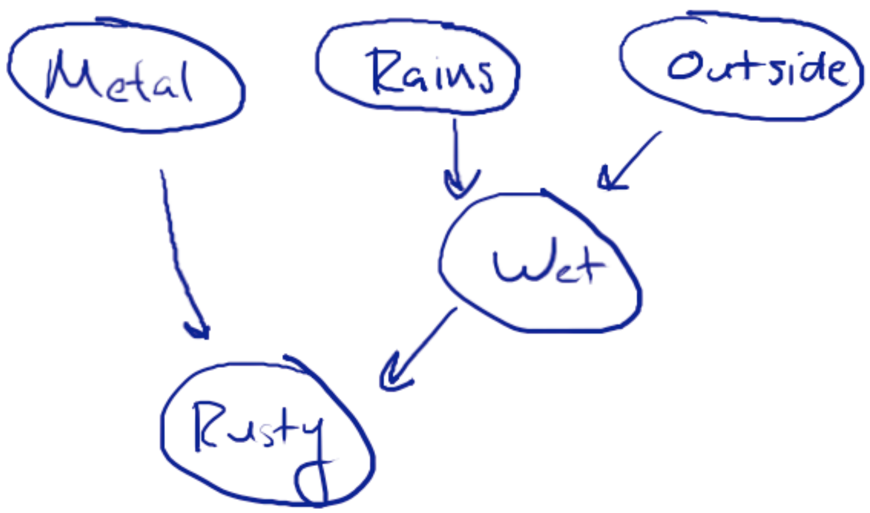
Rusty robot: the DAG

Rusty robot: the CPTs



- For each RV (say X), there is one CPT specifying P(X | pa(X))

Interpreting it



Benefits

- 11 v. 31 numbers
- Fewer parameters to learn
- Efficient ***inference*** = computation of marginals, conditionals $\Rightarrow$ posteriors

Inference example

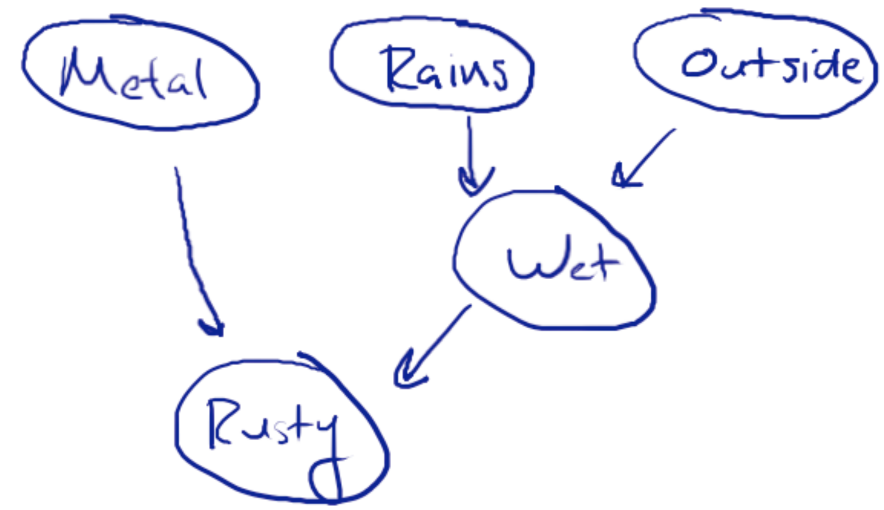
- $P(M, Ra, O, W, Ru) =$

  $P(M)\, P(Ra)\, P(O)\, P(W|Ra,O)\, P(Ru|M,W)$

- Find marginal of M, O

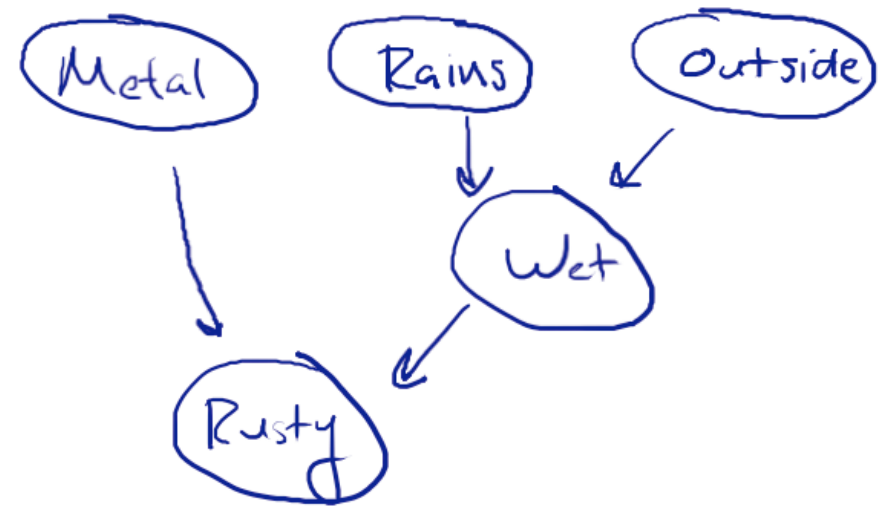
Independence

- Showed M $\perp$ O
- Any other independences?



- Didn't use
  - independences depend only on
- May also be "accidental" independences

Conditional independence



- How about O, Ru?   O   Ru
- Suppose we know we're not wet
- P(M, Ra, O, W, Ru) =

  P(M) P(Ra) P(O) P(W|Ra,O) P(Ru|M,W)
- Condition on W=F, find marginal of O, Ru

Conditional independence

- This is generally true
  - conditioning on evidence can make or break independences
  - many (conditional) independences can be derived from graph structure alone
  - “accidental” ones are considered less interesting

Graphical tests for independence

- We derived (conditional) independence by looking for factorizations
- It turns out there is a purely graphical test
  - this was one of the key contributions of Bayes nets
- Before we get there, a few more examples

Blocking

- Shaded = observed (by convention)

Explaining away

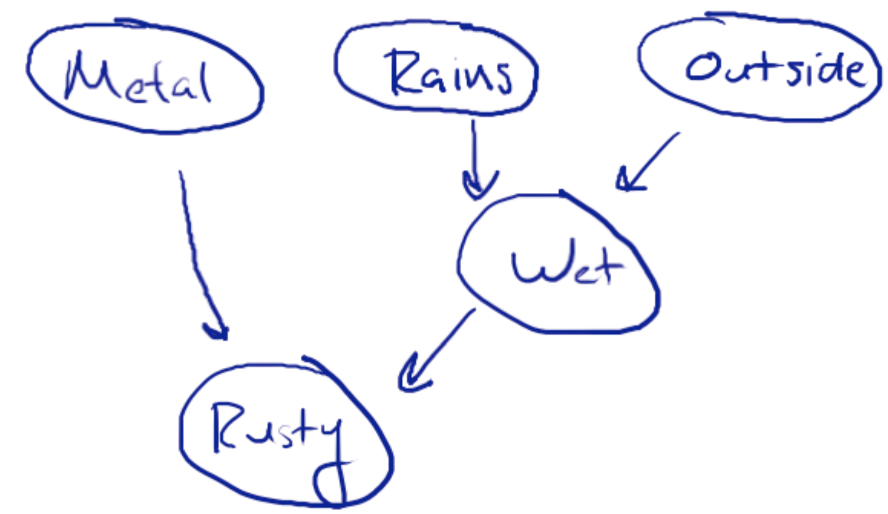
- Intuitively:

Son of explaining away

d-separation

- General graphical test: "d-separation"
  - d = dependence
- $X \perp Y \mid Z$ when there are no ***active paths*** between X and Y
- Active paths (W ***outside*** conditioning set):

Longer paths

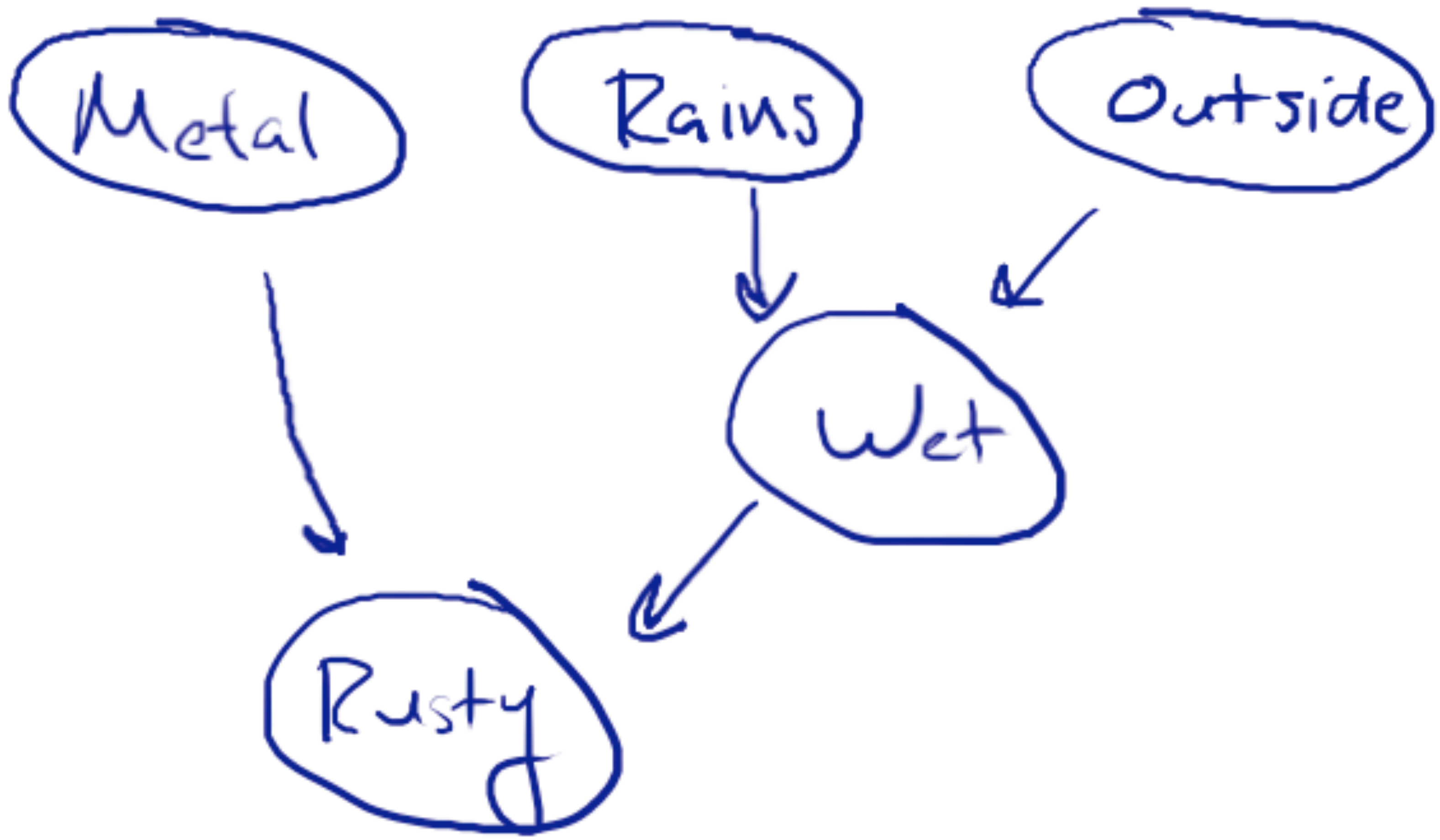


- Node is active if:

  and inactive o/w

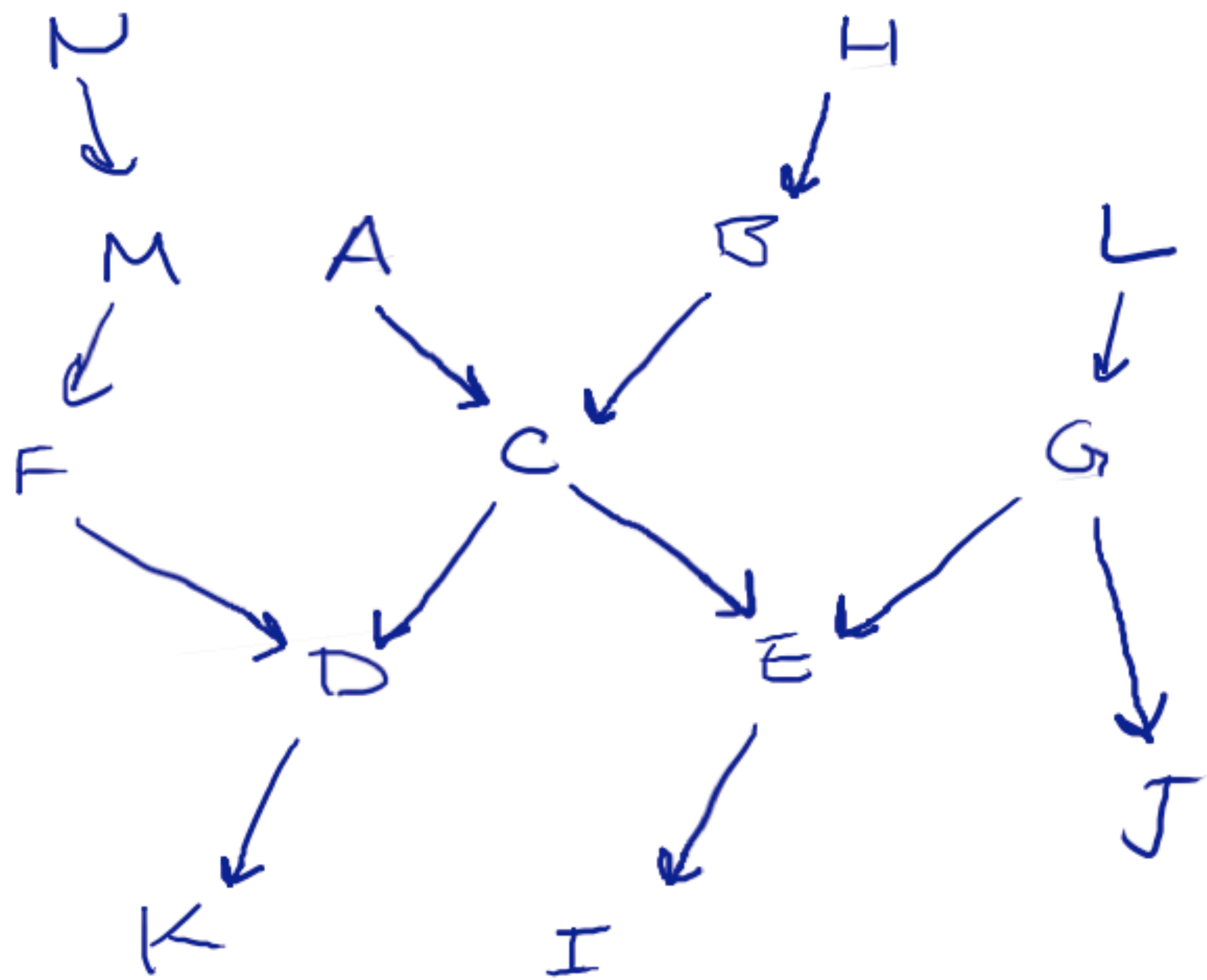
- Path is active if intermediate nodes are

Another example

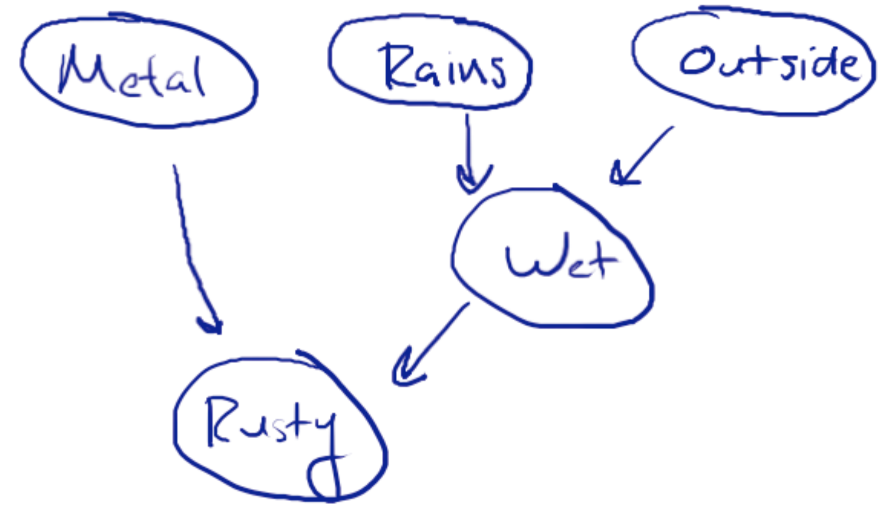


Markov blanket

Markov blanket of C = minimal set of observations to render C independent of rest of graph



Learning Bayes nets



P(M) =

P(Ra) =

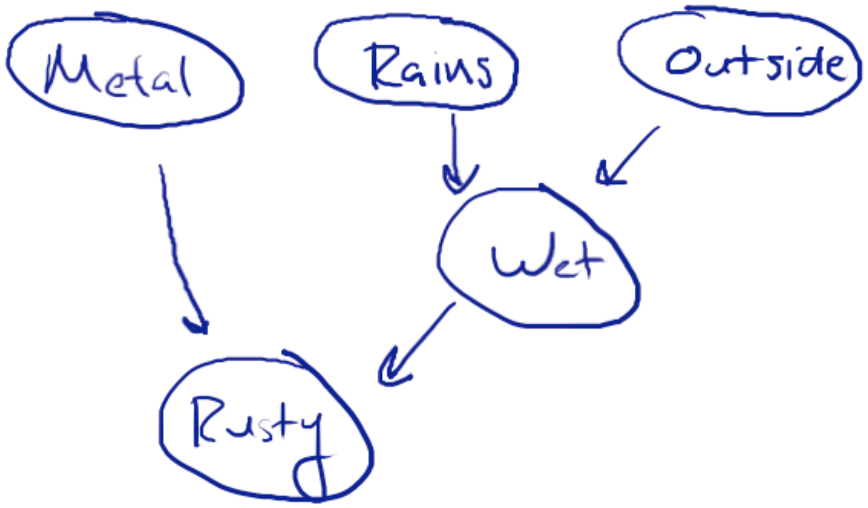
P(O) =

P(W | Ra, O) =

P(Ru | M, W) =

M	Ra	O	W	Ru
T	F	T	T	F
T	T	T	T	T
F	T	T	F	F
T	F	F	F	T
F	F	T	F	T

Laplace smoothing



P(M) =

P(Ra) =

P(O) =

P(W | Ra, O) =

P(Ru | M, W) =

M	Ra	O	W	Ru
T	F	T	T	F
T	T	T	T	T
F	T	T	F	F
T	F	F	F	T
F	F	T	F	T

Advantages of Laplace

- No division by zero
- No extreme probabilities
  - No near-extreme probabilities unless lots of evidence

Limitations of counting and Laplace smoothing

- Work ***only*** when all variables are observed in all examples
- If there are ***hidden*** or ***latent*** variables, more complicated algorithm—we'll cover a related method later in course
  - or just use a toolbox!

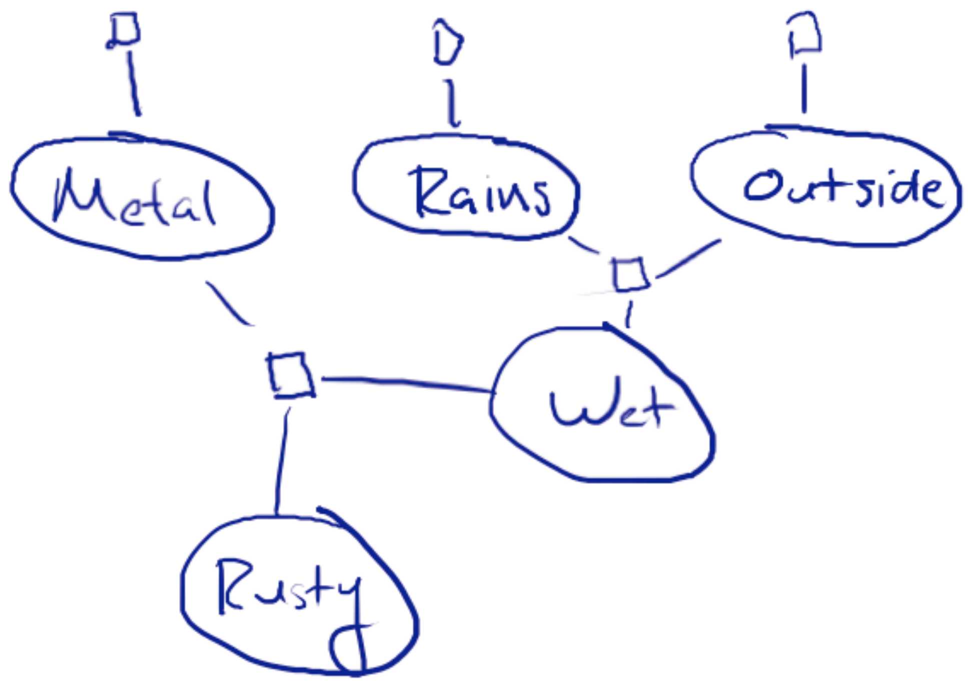
Factor graphs

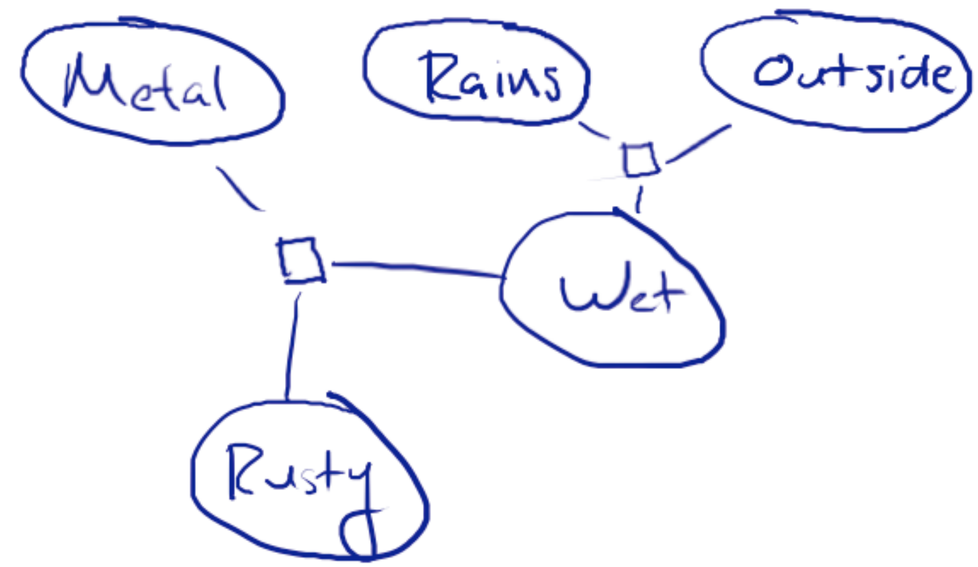
- Another common type of graphical model
- Uses ***undirected, bipartite*** graph instead of DAG

Rusty robot: factor graph

P(M) P(Ra) P(O) P(W|Ra,O) P(Ru|M,W)

Convention





- Don't need to show unary factors
- Why? They don't affect algorithms below.

Non-CPT factors

- Just saw: easy to convert Bayes net → factor graph
- In general, factors need not be CPTs: any nonnegative #s allowed
- In general, P(A, B, …) =
- Z =

Ex: image segmentation

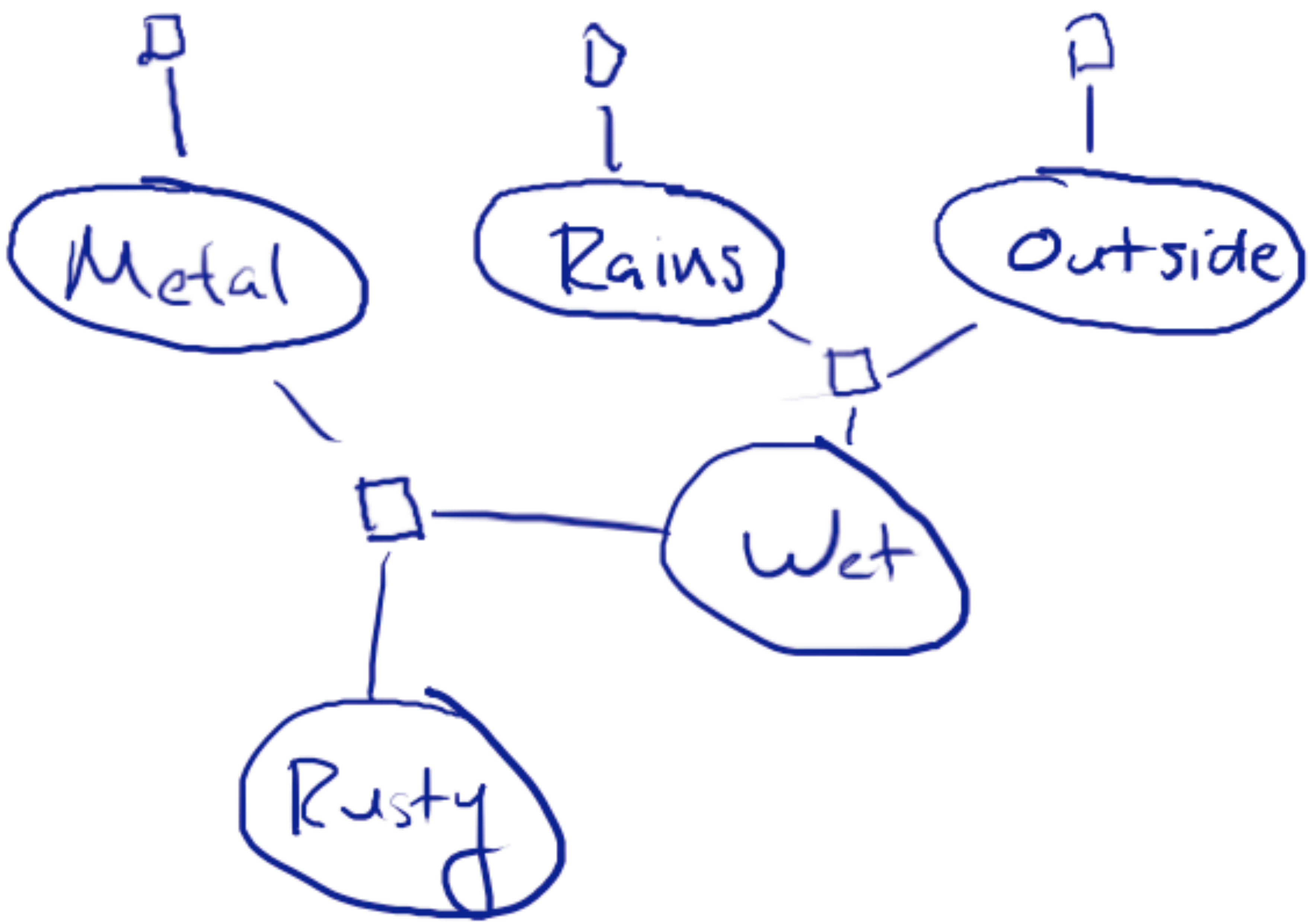
Factor graph → Bayes net

- Possible, but more involved
  - Each representation can handle ***any*** distribution
- Without adding nodes:
- Adding nodes:

Independence

- Just like Bayes nets, there are graphical tests for independence and conditional independence
- Simpler, though:
  - Cover up all observed nodes
  - Look for a path

Independence example



Modeling independence

- Take a Bayes net, list the (conditional) independences
- Convert to a factor graph, list the (conditional) independences
- Are they the same list?
- What happened?