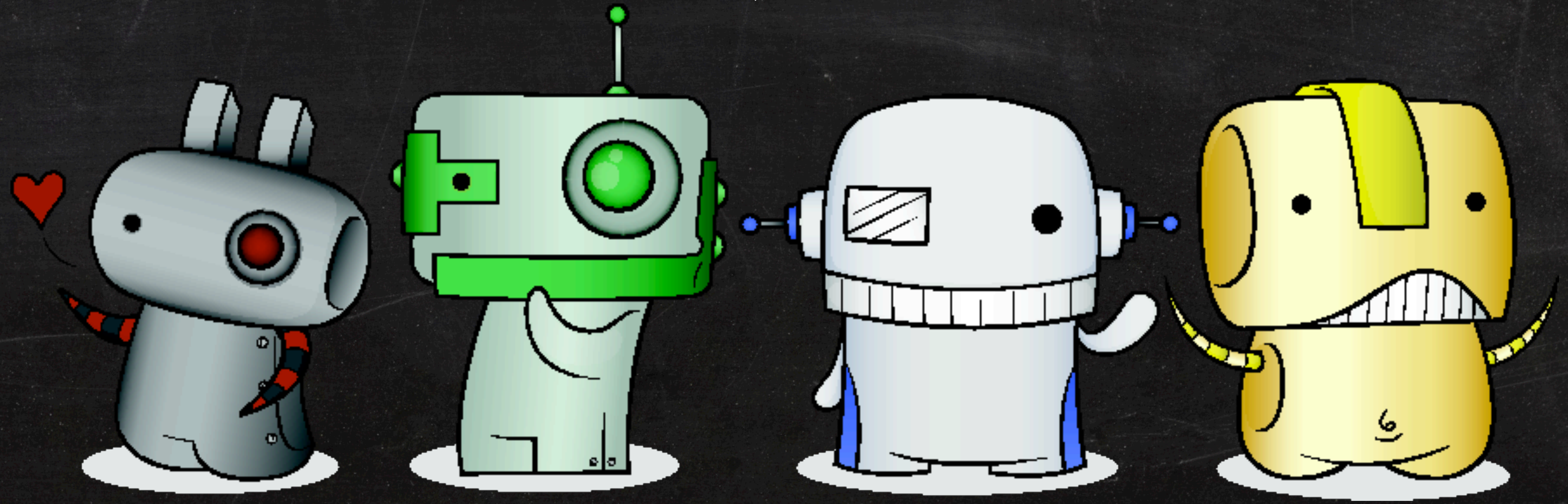
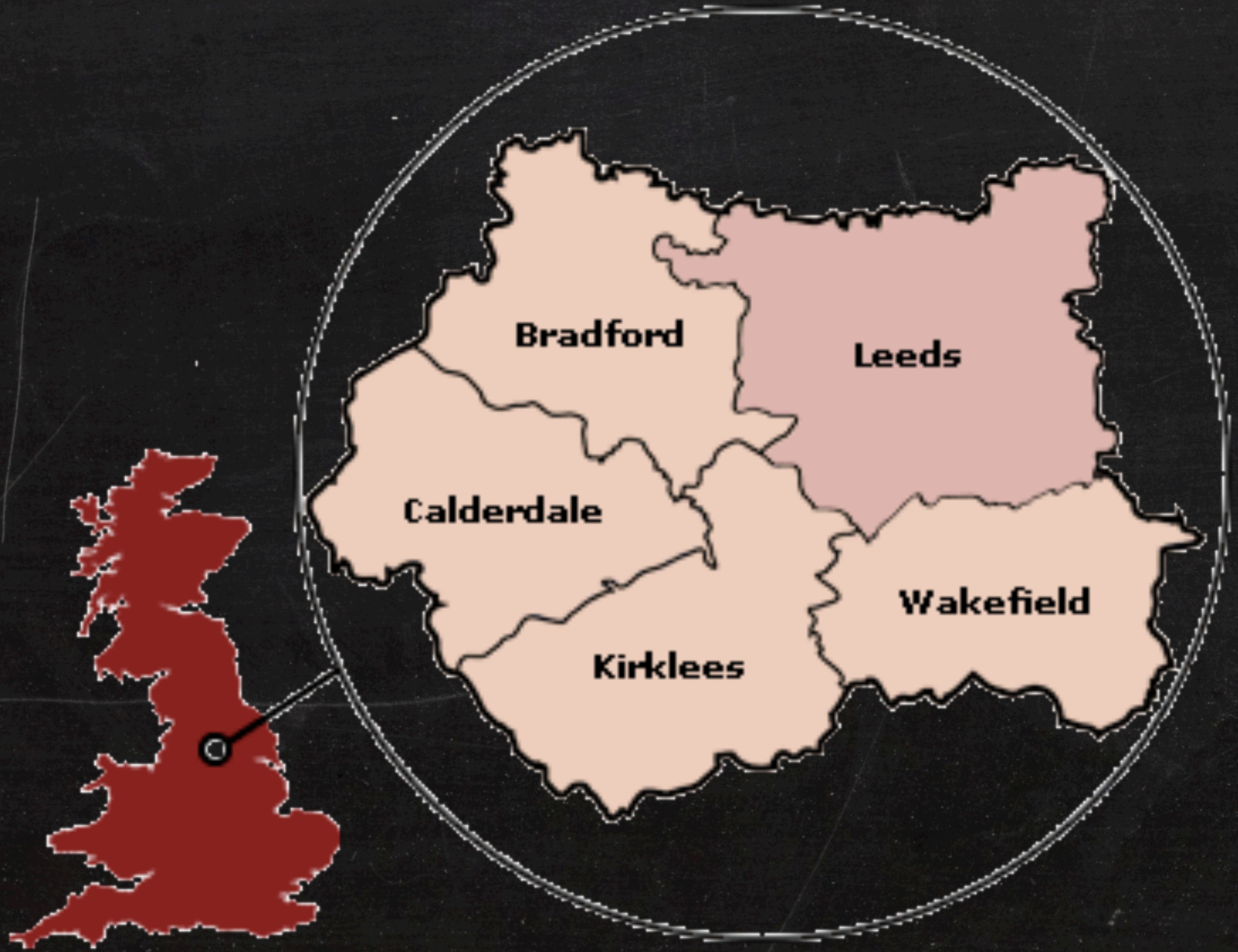
James McGivern

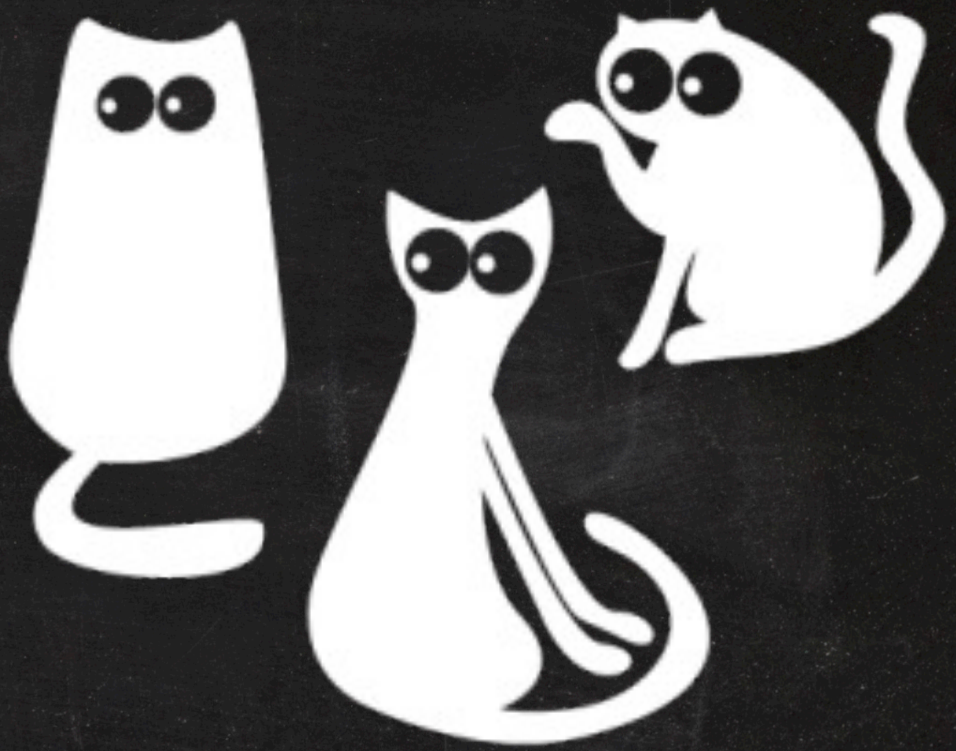
Probably, Definitely, Maybe



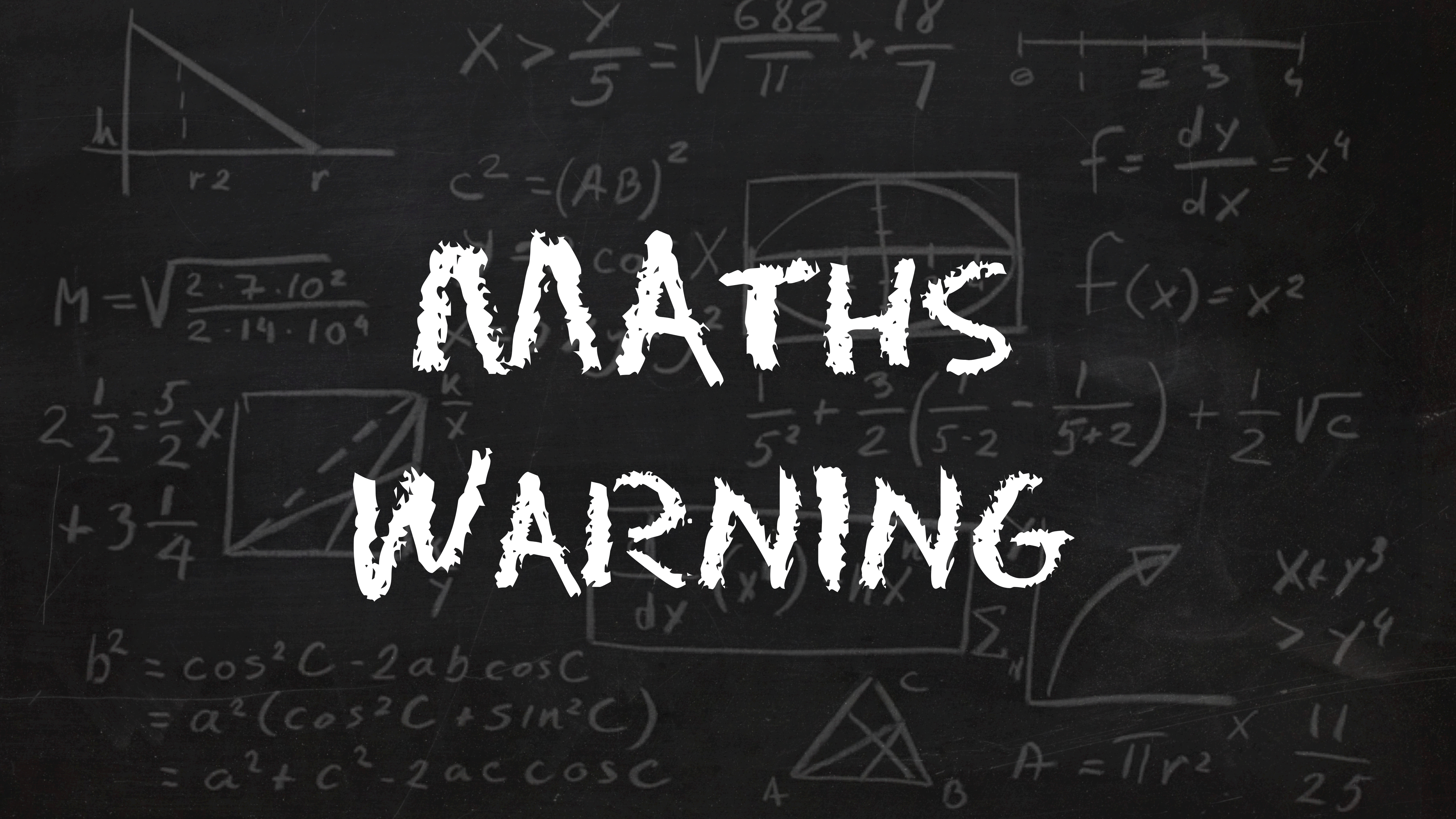
About Me

rockshore

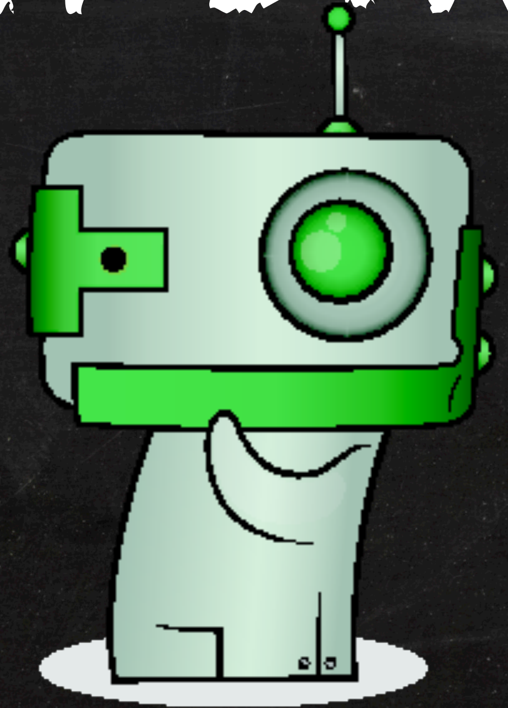




I Talk Fast!


MATHS
WARNING

- Chapter 1 -
Bayesian Probability
&
Bayesian Statistics



The Mean

Given a set of variables

$$X = \{x_1, \ldots, x_n\}$$

we define the mean (average)

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=0}^{n} x_i$$

Example: Amazon

- Users can rate a product by voting 1-5 stars
- product rating is the mean of the user votes

Q: how can we rank products with different number of votes?

Simple "Bayesian Ranking"

$$\overline{rank} = \frac{Cm + Rv}{m + v}$$

- C - the mean vote across all items
- v - number of votes for a given item
- R - the mean rating of the item
- m - number of votes required to be in top n percentile

Book	Number of votes (v)	Average Rating (R)	Bayesian Rank
A	100	5	4.333333...
B	70	5	4.17
C	50	4	3.5
D	30	4	3.375
E	20	3.5	3.14
F	30	3	3
G	5	2	2.91

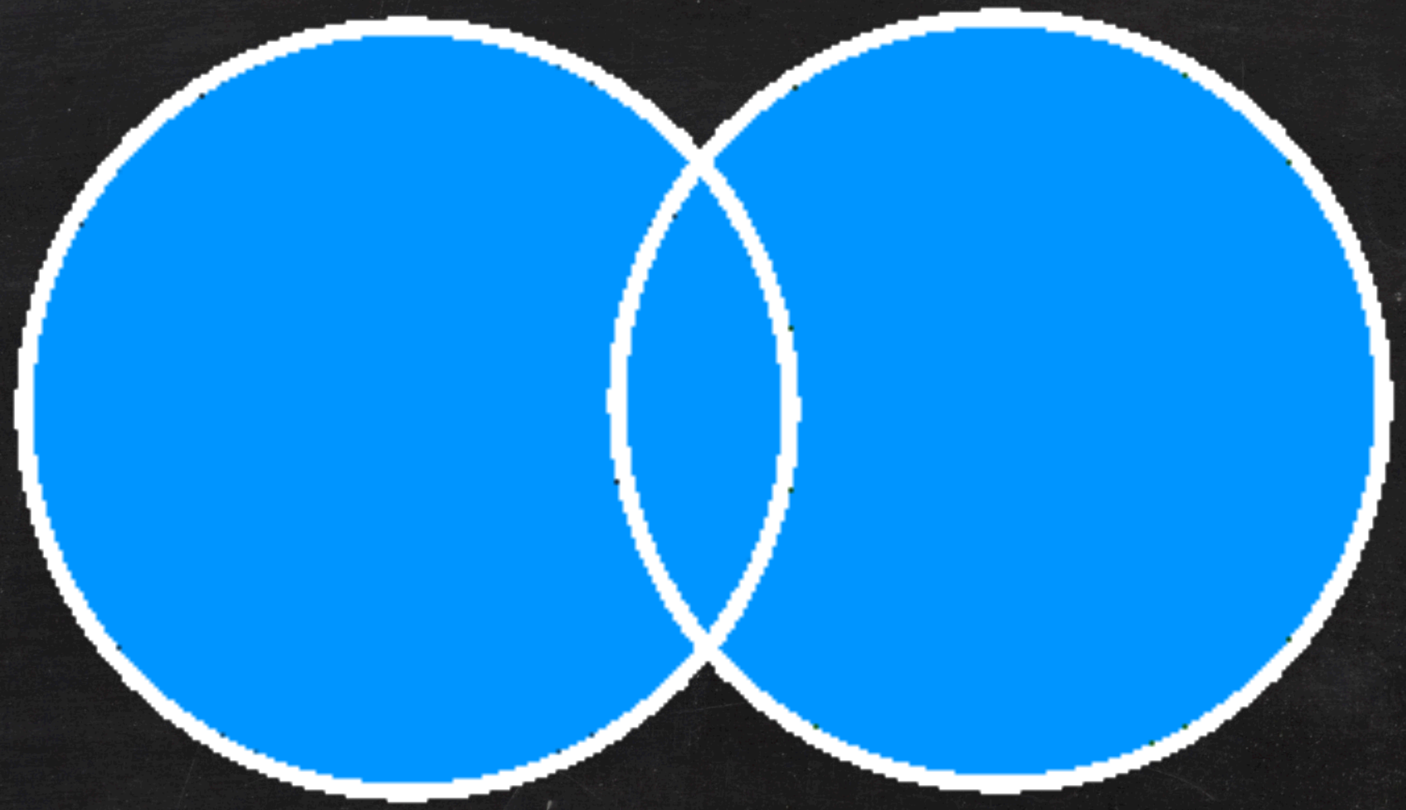
$C = 3$ $m = 50$

A Detour In To Probability Basics

Events

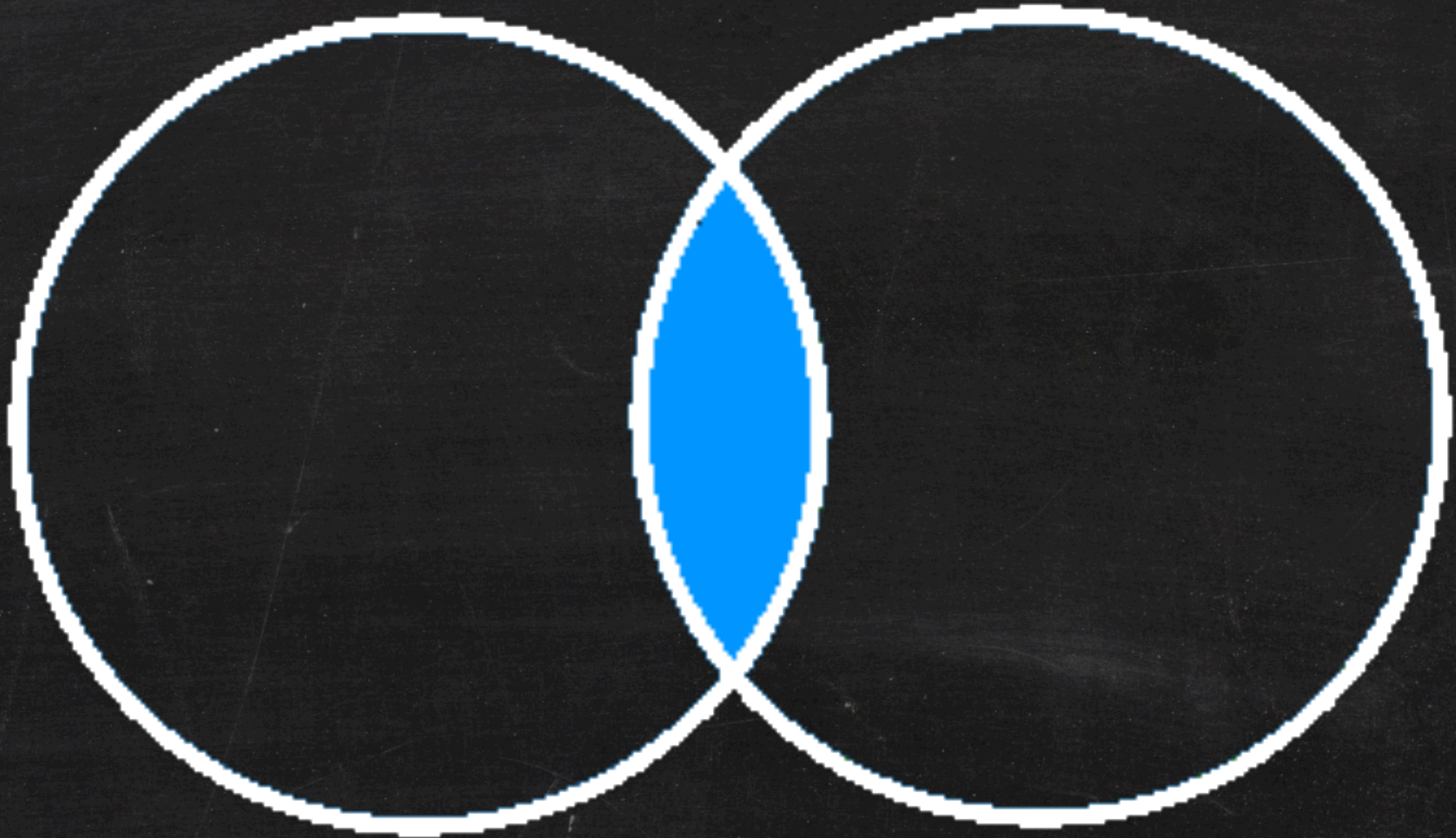
- Consider an experiment whose set of all possible outcomes Ω, called the sample space, is $\{x_1, \ldots, x_n\}$
- We define an event E as a subset of Ω and say that E occurs iff the experiment outcomes equal E

Union



$$E_1 \cup E_2 \cup \ldots \cup E_n = \bigcup_{i=1}^{n} E_i$$

Intersection



$$E_1 \cap E_2 \cap \ldots \cap E_n = \bigcap_{i=1}^{n} E_i$$

Probability Axioms: I

- We denote the probability of an event A by P(A)
- For any event A, $0 \leq P(A) \leq 1$
- The certain event, Ω, always occurs and $P(\Omega)=1$
- The impossible event $\emptyset$ never occurs and $P(\emptyset)=0$

Probability Axioms: 2

- We say that events A and B are disjoint if $A \cap B = \emptyset$
- if A and B are disjoint then $P(A \cup B) = P(A) + P(B)$
- for a set of disjoint events, the addition law gives us:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$$

Probability Lemmas

- For any event E, $P(E^c) = P(\neg E) = 1 - P(E)$
- $P(A-B) = P(A) - P(A \cap B)$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - \ldots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} E_i\right)$$

Random Variables

- Consider a random variable X, then $\{X \leq x\}$ is the event that X has a value less than or equal to the real number x. Hence the probability that this event occurs is $P(X \leq x)$
- If we allow x to vary we can define the distribution function

$$F(x) = P(X \leq x) \qquad -\infty < x < \infty$$

- Note that:
  - $P(X>x) = 1 - F(x)$
  - $P(a<X<b) = F(b) - F(a)$

Probability Mass Function

The probability mass function (PMF) of X

$$f_x(x) = P(X = x) = P(\{s \in S : X(s) = x\})$$

is a probability measure of the possible values for the random variable. Of course

$$\sum_{x \in A} f_x(x) = 1$$

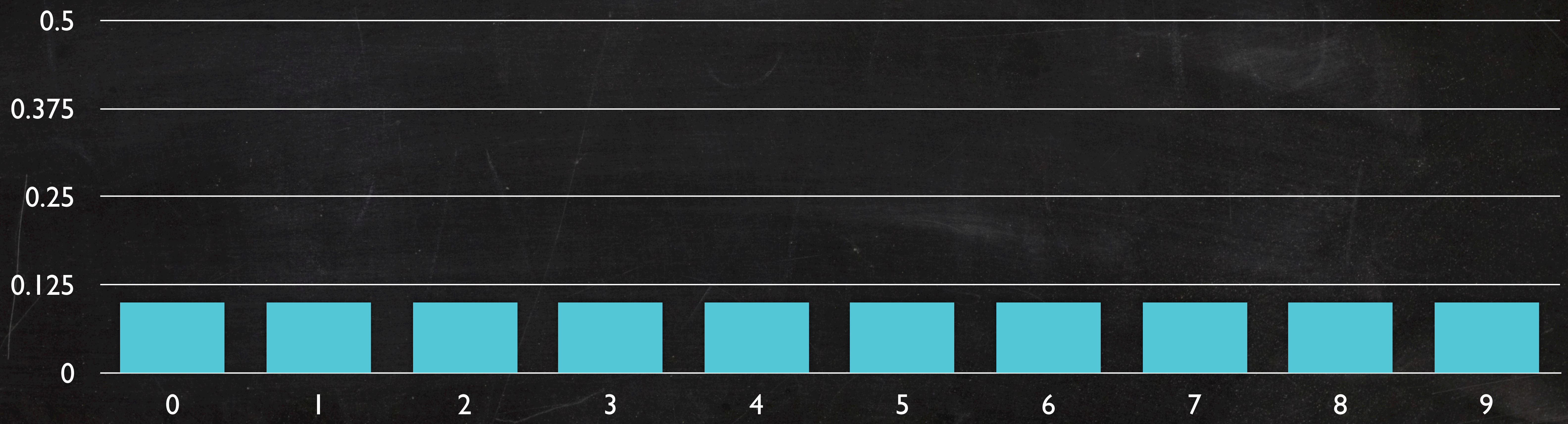
PMF for a fair 6 sided dice

$$f_x = \begin{cases} 1/6 & x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases}$$

Example: PMF of 2 Fair Die

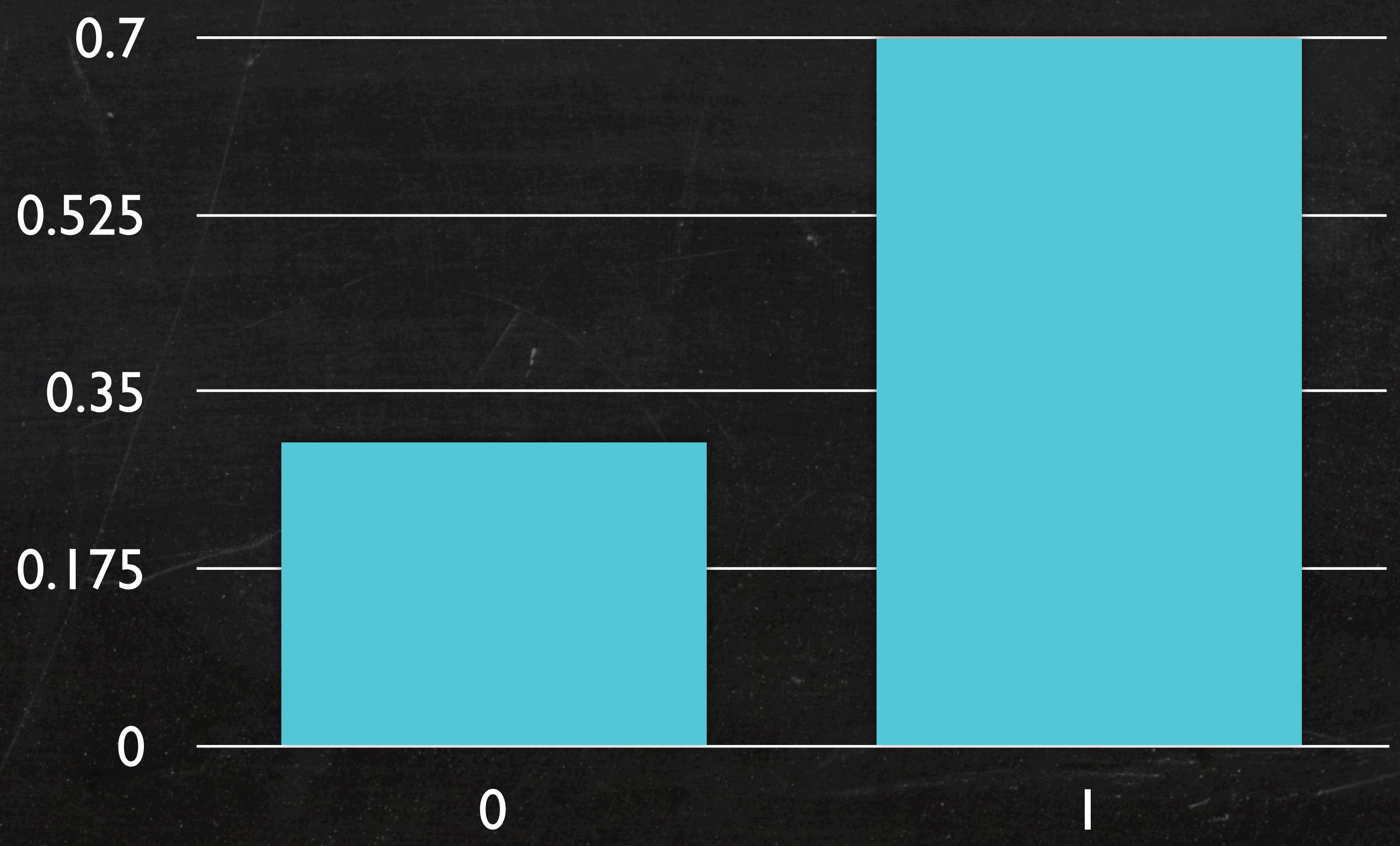
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
					1,6					
				1,5	2,5	2,6				
			1,4	2,4	3,4	3,5	3,6			
		1,3	2,3	3,3	4,3	4,4	4,5	4,6		
	1,2	2,2	3,2	4,2	5,2	5,3	5,4	5,5	5,6	
1,1	2,1	3,1	4,1	5,1	6,1	6,2	6,3	6,4	6,5	6,6
2	3	4	5	6	7	8	9	10	11	12

Uniform Distribution



Bernoulli Distribution

$$F(k,p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$



Binomial Distribution

$$F(k,n,p) = P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability Axioms: 3

- Given two events A and B we define the conditional probability P(A | B) by:

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

- Given two events A and B we say that they are independent iff:

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

Prior & Posterior Distributions

- $P(\theta)$ the prior probability distribution of θ
- $P(\theta|X)$ is the posterior probability of θ given X
- The posterior probability can be written in the memorable form as:

  posterior probability * likelihood * prior probability

- If the posterior distributions $P(\theta|X)$ are in the same family as the prior probability distribution $P(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior

Bayes' Theorem

From the definition of conditional probability we know:

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

Hence

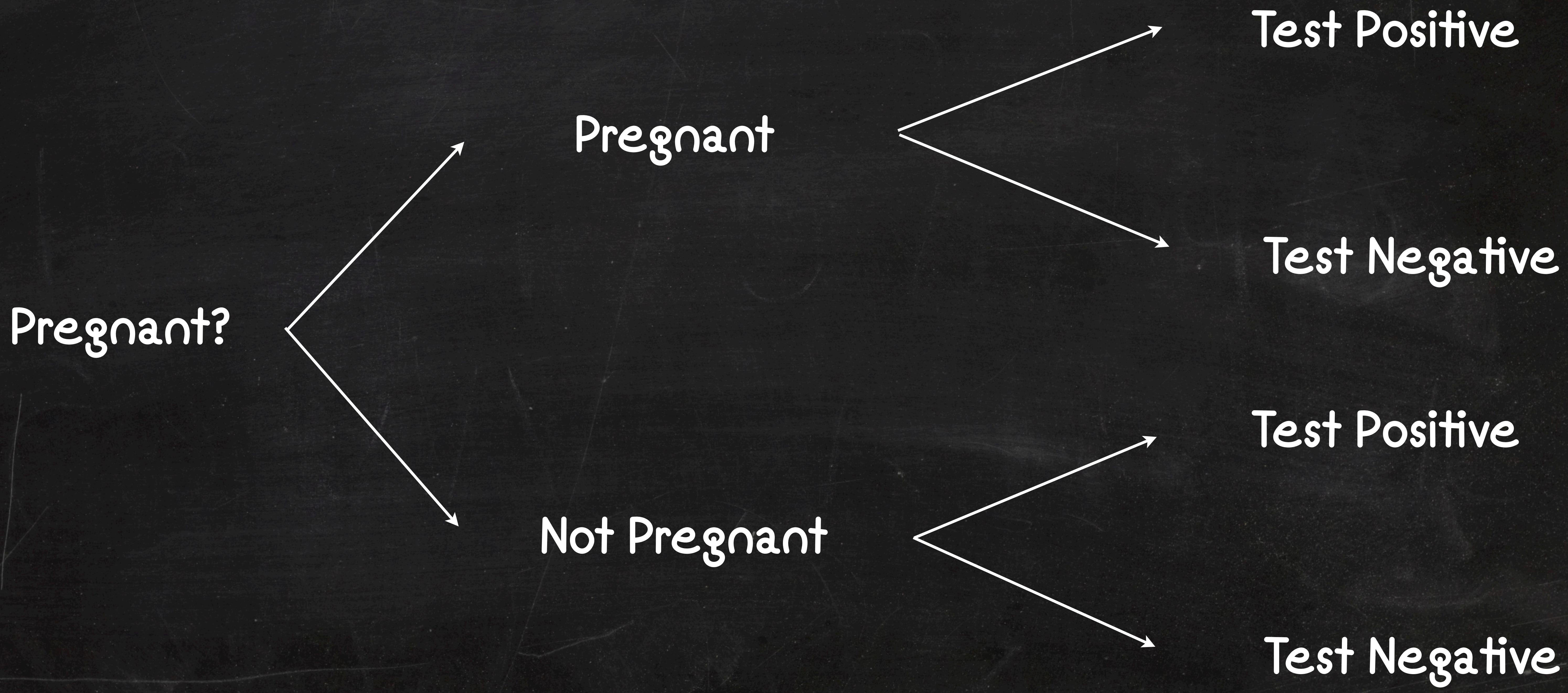
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

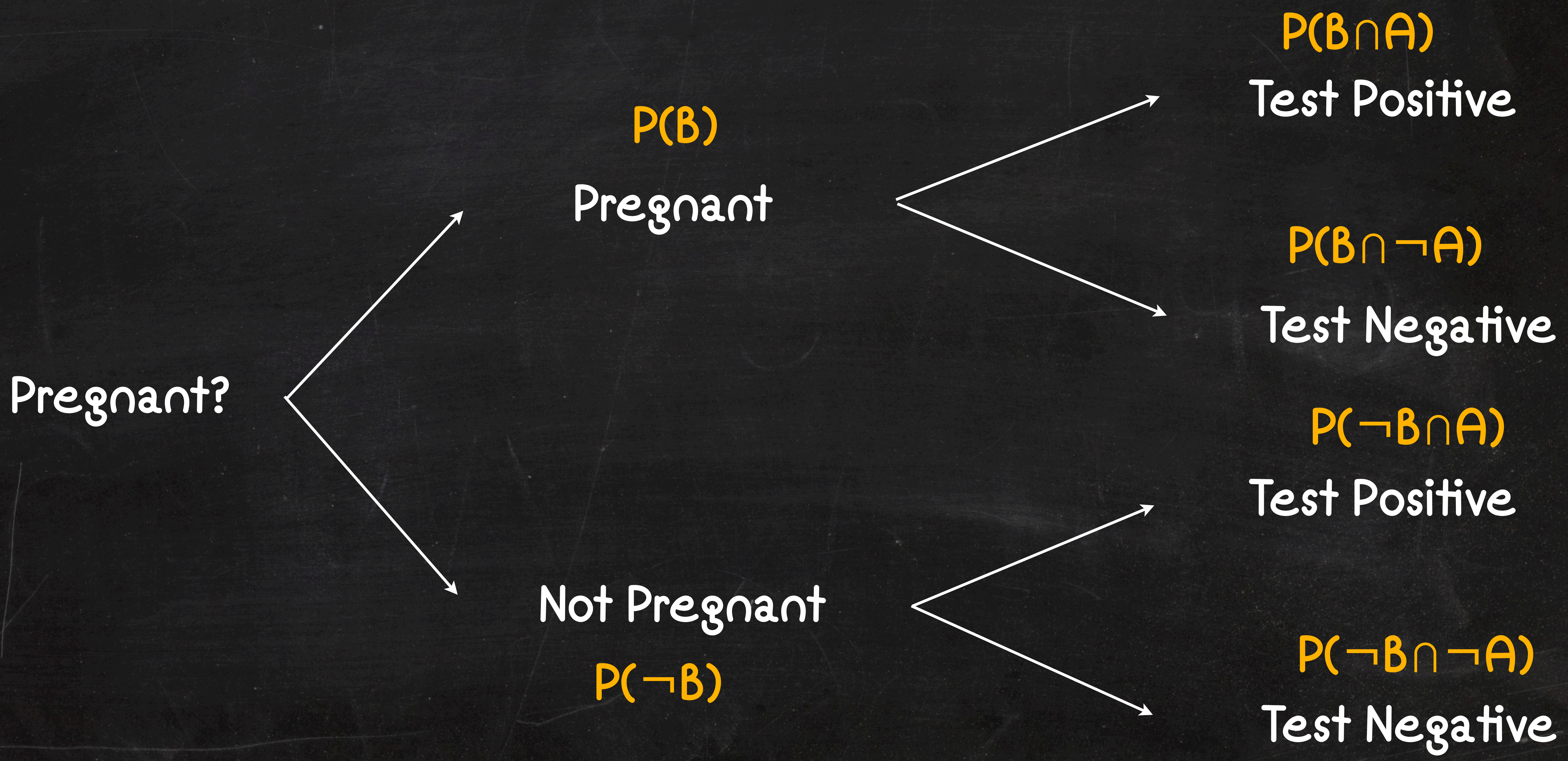
Or if $B_1,\ldots,B_n$ form a partition of the sample space

$$P(B_n|A) = \frac{P(A|B_n)P(B_n)}{\sum_i P(A|B_i)P(B_i)}$$

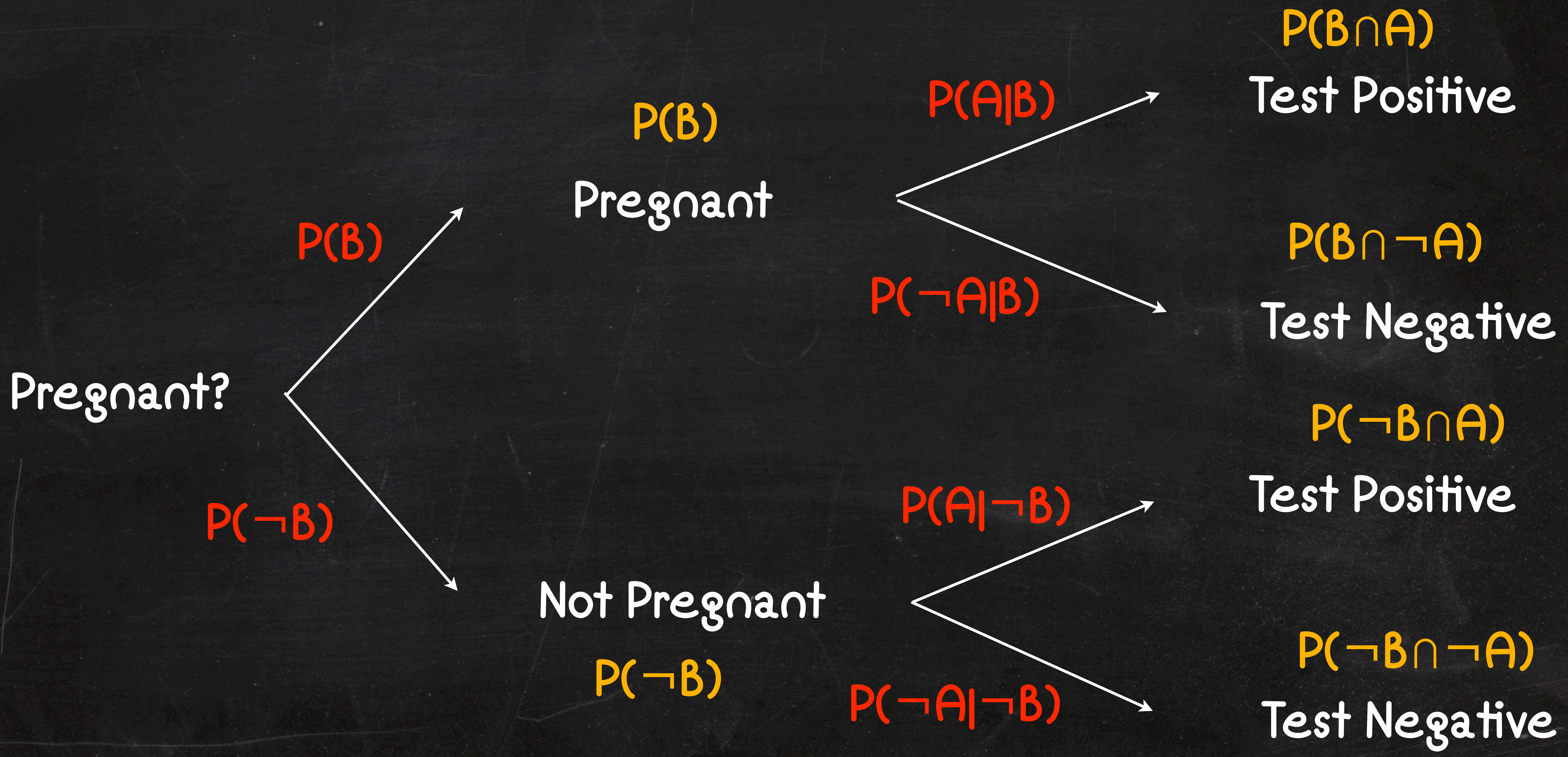
Example: Pregnancy Tests

- Pregnancy tests detect the presence of hCG, or human chorionic gonadotropin, in the blood or urine
- A “false positive” is when the test incorrectly returns a positive result, and “false negative” when it incorrectly returns a false one.
- False positives in the hcg test include:
  - non-pregnant production of the hCG molecule
  - use of drugs containing the hCG molecule
  - Some medications cause a positive reaction in the tests
- The actual probability of being pregnant depends on many messy biological factors

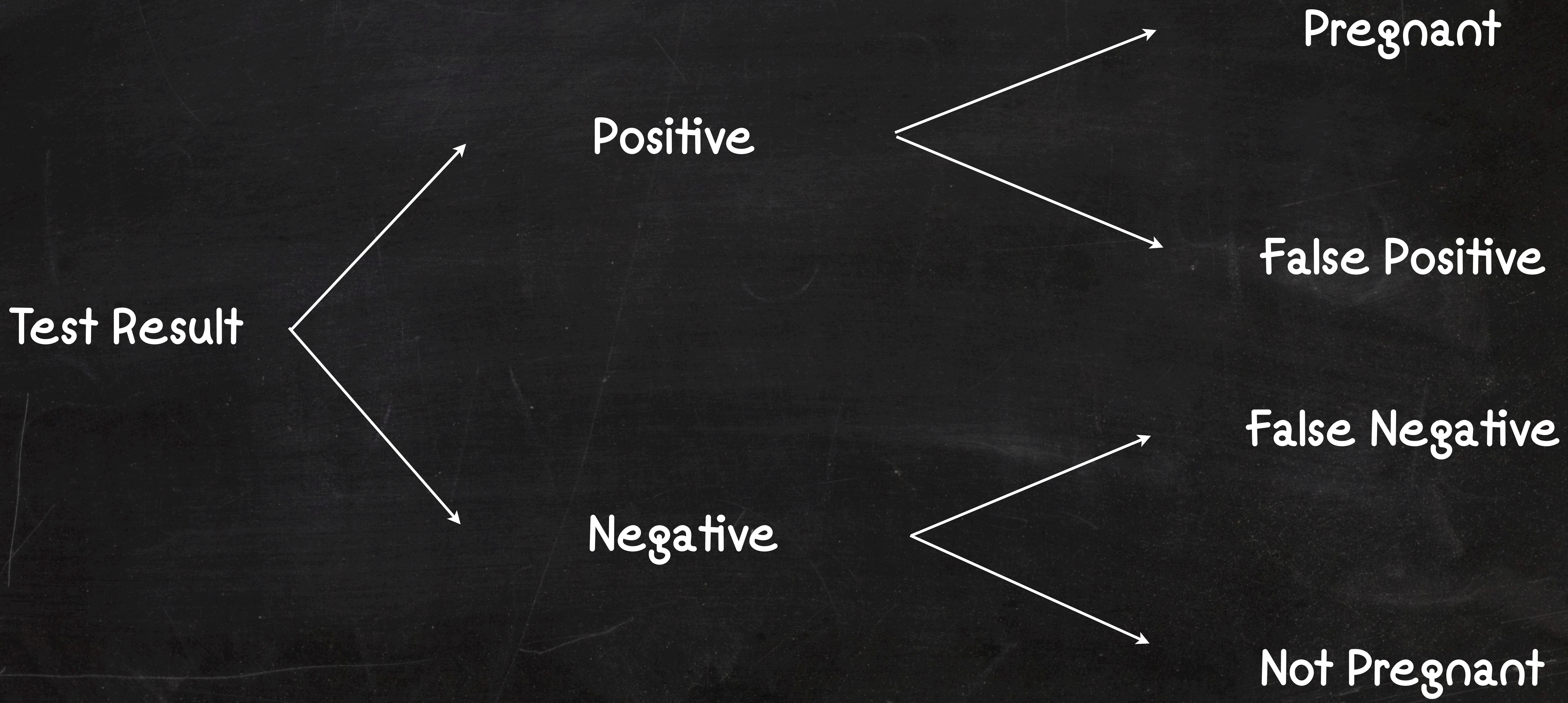

Pregnant?
Pregnant
Not Pregnant
Test Positive
Test Negative
Test Positive
Test Negative

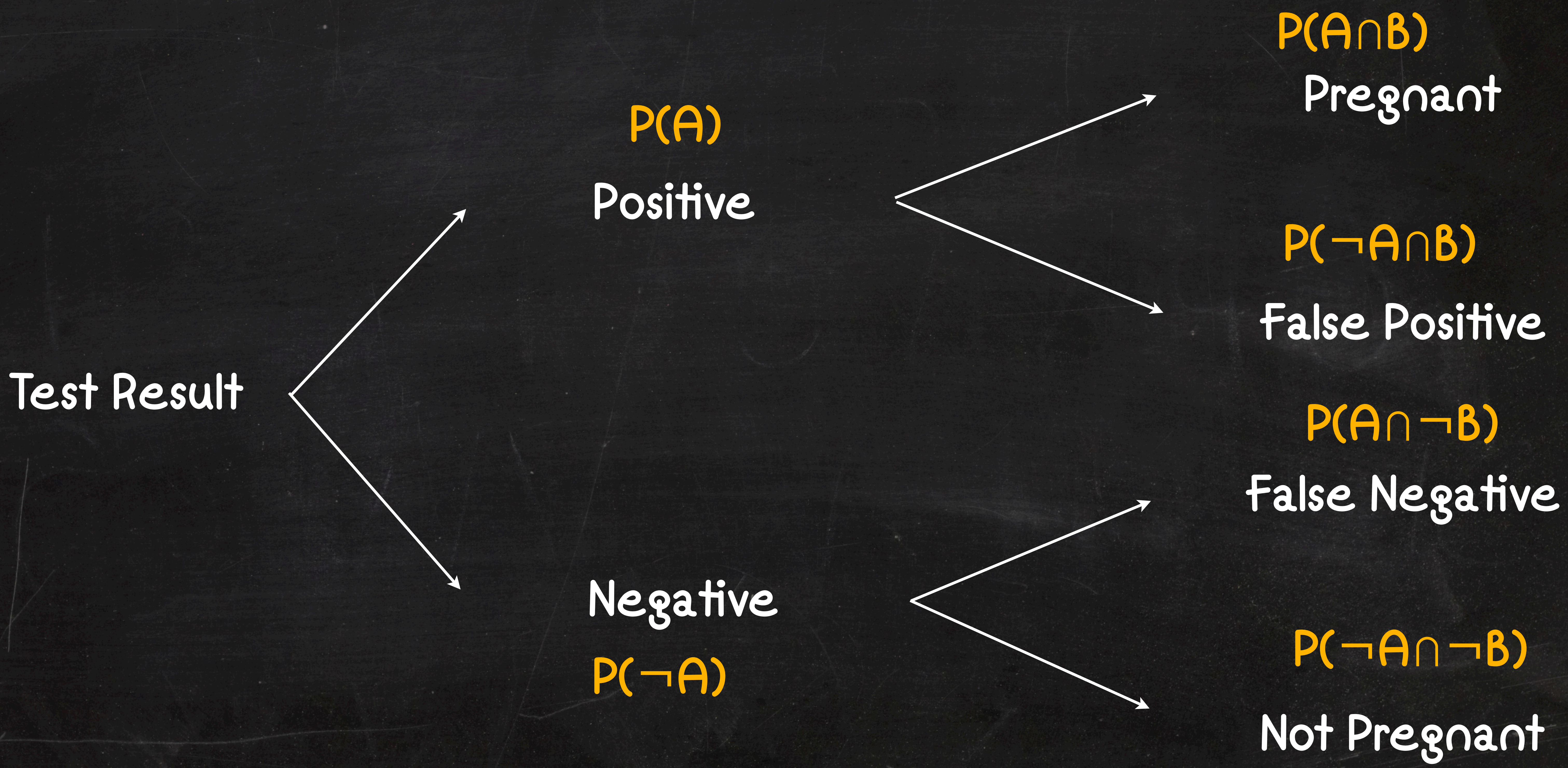


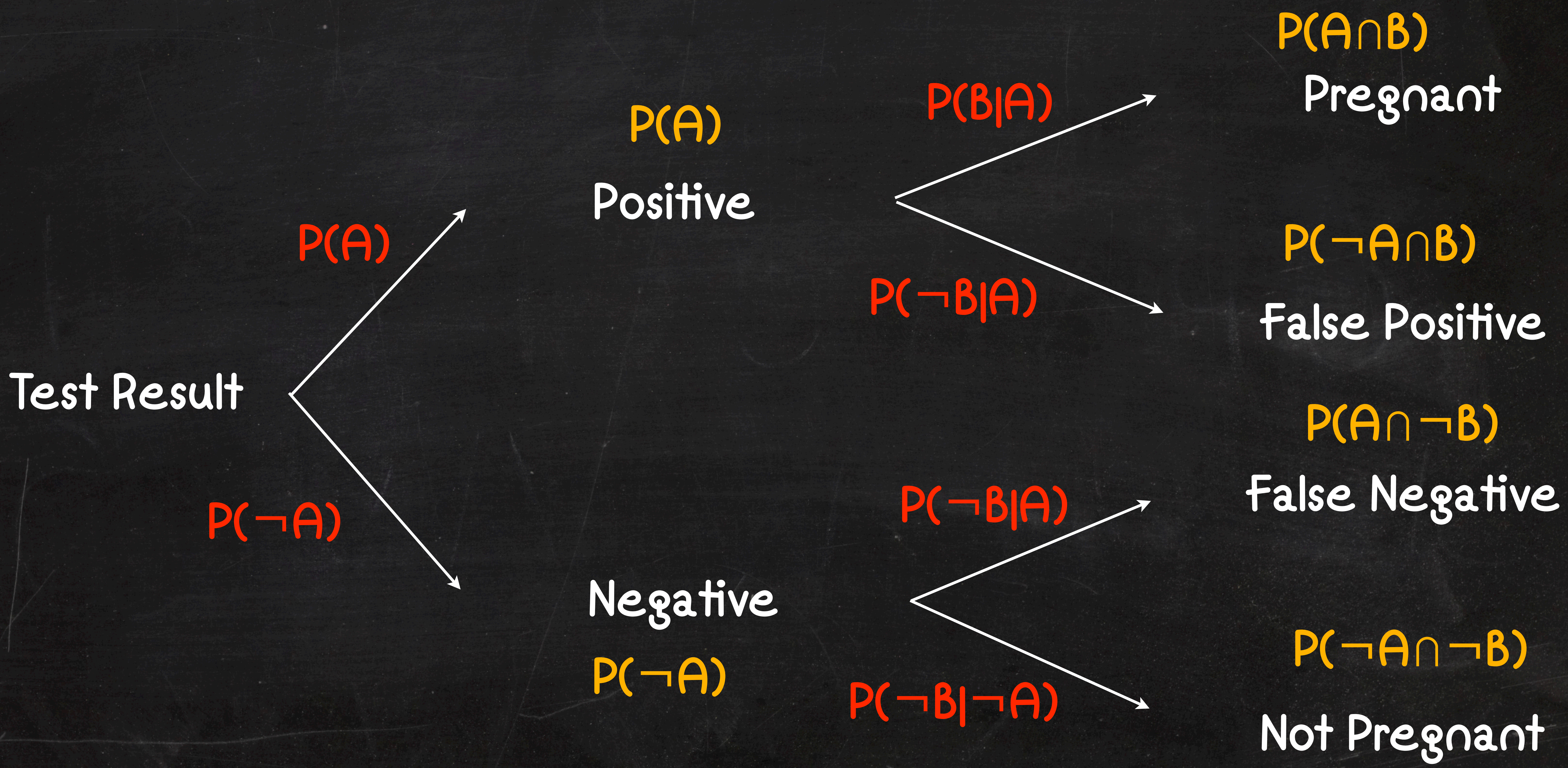
Pregnant?
$P(B)$
Pregnant
Not Pregnant
$P(\neg B)$
$P(B \cap A)$
Test Positive
$P(B \cap \neg A)$
Test Negative
$P(\neg B \cap A)$
Test Positive
$P(\neg B \cap \neg A)$
Test Negative


Pregnant?
P(B)
P(¬B)
P(B)
Pregnant
Not Pregnant
P(¬B)
P(A|B)
P(¬A|B)
P(A|¬B)
P(¬A|¬B)
P(B∩A)
Test Positive
P(B∩¬A)
Test Negative
P(¬B∩A)
Test Positive
P(¬B∩¬A)
Test Negative

Q: Given the test is positive what is the probability that the subject is pregnant?


Test Result
Positive
Negative
Pregnant
False Positive
False Negative
Not Pregnant


Test Result
P(A)
Positive
Negative
P(¬A)
P(A∩B)
Pregnant
P(¬A∩B)
False Positive
P(A∩¬B)
False Negative
P(¬A∩¬B)
Not Pregnant


Test Result
P(A)
P(¬A)
P(A)
Positive
Negative
P(¬A)
P(B|A)
P(¬B|A)
P(¬B|A)
P(¬B|¬A)
P(A∩B)
Pregnant
P(¬A∩B)
False Positive
P(A∩¬B)
False Negative
P(¬A∩¬B)
Not Pregnant

Example: Disease Diagnosis

Q: Consider the set $S=\{s_i : i = 0...N\}$ of all disease symptoms, and $D_{s+}=\{s_i : s_i \text{ in } S\}$ are the diagnostically inclusive symptoms of Ebola, and $D_{s-}=\{s_i : s_i \text{ in } S\}$ the exclusionary symptoms. Given a patient has some combination of symptoms $P_s=\{s_i : s_i \text{ in } S\}$, what is the probability they have Ebola?

- The presence or absence of some symptoms can completely rule out the diagnosis
- By updating the model based on real outcomes it is possible to provide more and more accurate predictions

Expected Values & Moments

- Suppose random variable X can take value x_1 with probability p_1, value x_2 with probability p_2, and so on. Then the expectation of this random variable X is defined as:

$$E[X] = p_1x_1 + p_2x_2 + \ldots + p_kx_k$$

- The variance of a random variable X is its second central moment, the expected value of the squared deviation from the mean $\mu = E[X]$:

$$Var(X) = E[(X-\mu)^2]$$

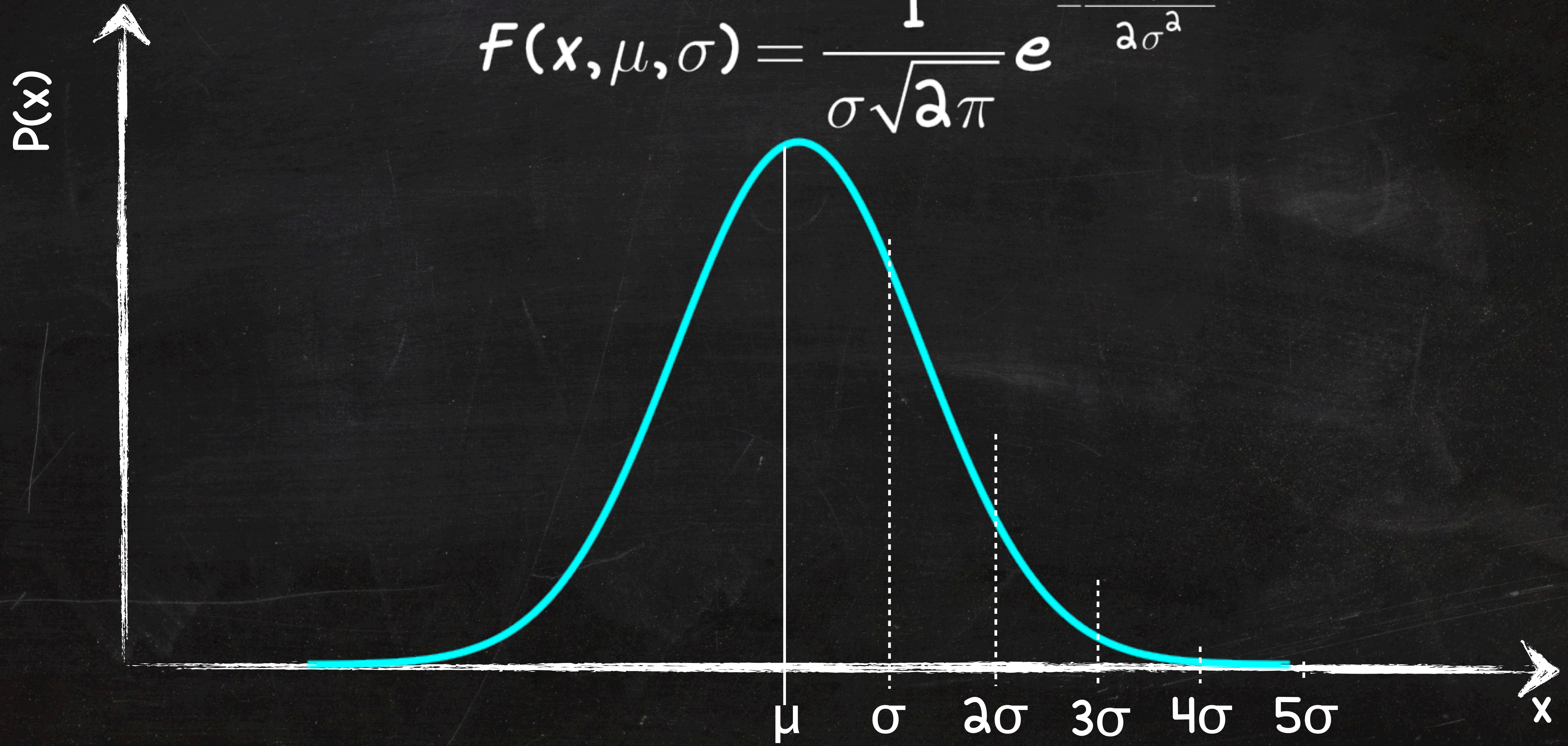
Variance & Covariance

- Variance is a measure of how far a set of numbers differs from the mean of those numbers. The square root of the variance is the standard deviation σ
  - CERN uses the 5-sigma rule to rule out statistical anomalies in sensor readings, i.e. is the value NOT the expected value of noise
- The covariance between two jointly distributed random variables X and Y with finite second moments is defined as:

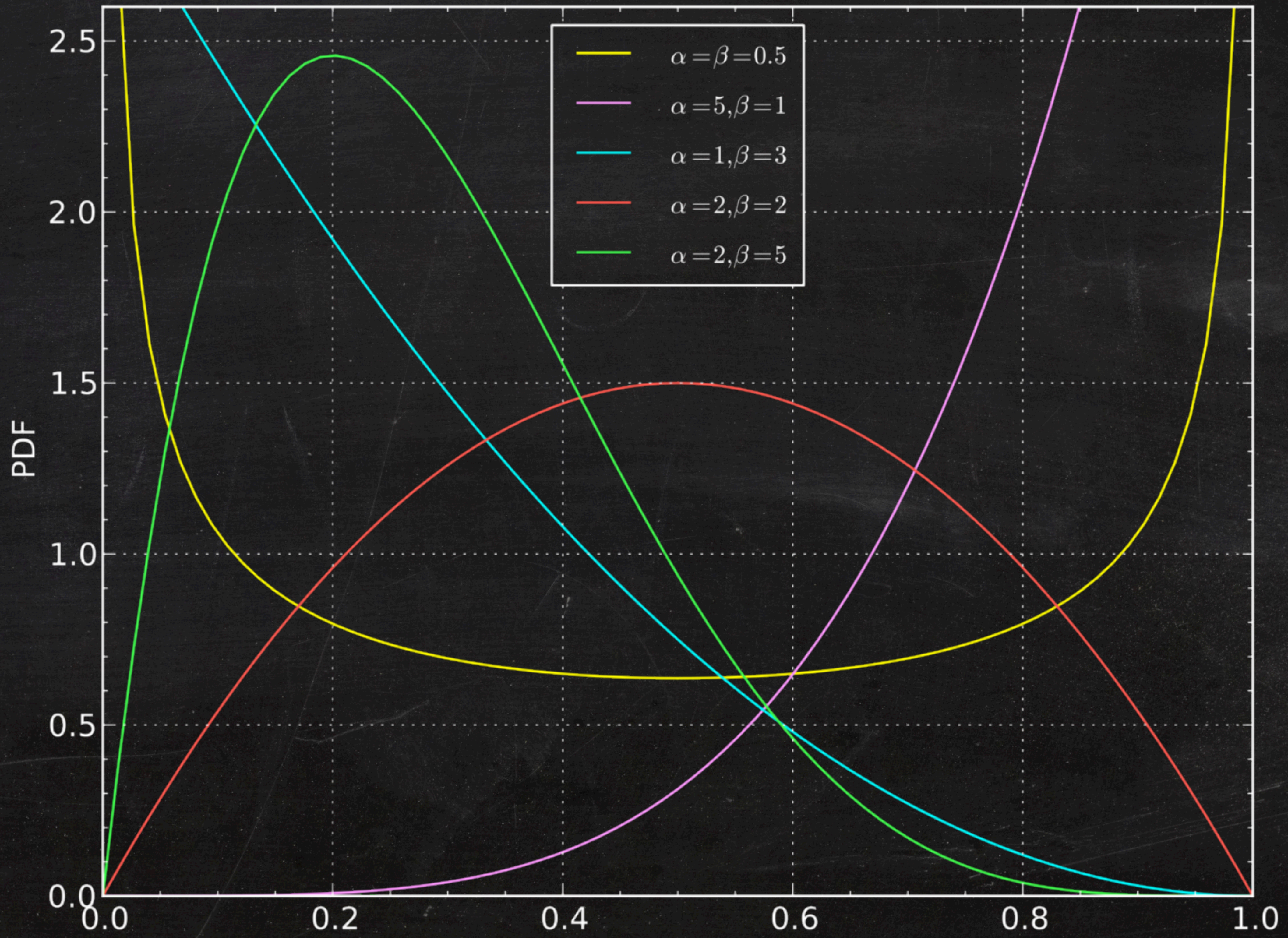
$$\sigma(X,Y) = E[(X - E[X]) \bullet (Y - E[Y])]$$

Gaussian Distribution

$$F(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Beta Distribution



End of Detour

Example: Amazon (Revisited)

- Users can rate a product by voting 1-5 stars
- product rating is the mean of the user votes

Q: how can we rank products with different number of votes?

Simple "Bayesian Ranking"

$$\overline{rank} = \frac{Cm + Rv}{m + v}$$

Assume the vote posterior distribution is a Normal, then the prior is also a Normal*, with mean

prior mean

$$\frac{\tau_0 \mu_0 + \tau \sum_{i=1}^{n} x_i}{\tau_0 + n\tau}$$

precision of vote distribution

prior precision

(*) http://en.wikipedia.org/wiki/Conjugate_prior

Example: YouTube

- Users can rate a clip by voting +/-1
- clip rating is the mean of the user votes
- clips also record the number of views

Q: how can we compare clip ratings with different number of votes and views?

Q: how can we make results more relevant? e.g. take in to account how old ratings are, author provenance, cost of incorrect ranking promotion

“Bayesian Rank”

- Since this is a +/- Bernoulli Trial we can model the prior belief distribution by a Beta function*

$$f(x,\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- Let α = upvote bias + number of up votes
- Let β = downvote bias + number of down votes + 1
- Every time we receive a new vote we just recalculate the distribution

- To map between a belief and a sorting criterion we make a decision using a loss function L

- Since the value of L depends on the value of a random variable we use instead the expected value of L

- Consider a multilinear loss function:

$$L_k(x, X) = \begin{cases} k(X - x) & : x < X \\ x - X & : x \geq X \end{cases}$$

since we want to minimise the loss we have:

$$\min(E[L_k(x, X)]) = I_x(U+1, D+1) = \frac{1}{1+k}$$

Extending The Loss Function

- Suppose in addition to the vote counts we also record the timestamp of the votes
- Items becomes less relevant in the rank the longer it is since the last vote following a pattern of exponential decay
- Hence the current up or down vote count is now determined by
- Thus we derive an updated rank function from

$$v' = v \times a^{-t/\lambda} + 1$$

$$I_x(U \times a^{-t/\lambda}, D \times a^{-t/\lambda}) = \frac{1}{1+k}$$

The House That Skynet Built

- SkyNet SmartHome™ is a system designed to manage the state of various household resources, e.g. heating, lighting, media-centre, etc
- it communicates with users smart phones to identify their location
- its aims to are to maximise the comfort (e.g. room temperature, hot water) of the users and minimise on waste (e.g. power consumption)
- users can provide feedback to correct inappropriate behaviour, e.g turning the heating up too high

Predicting Behaviour

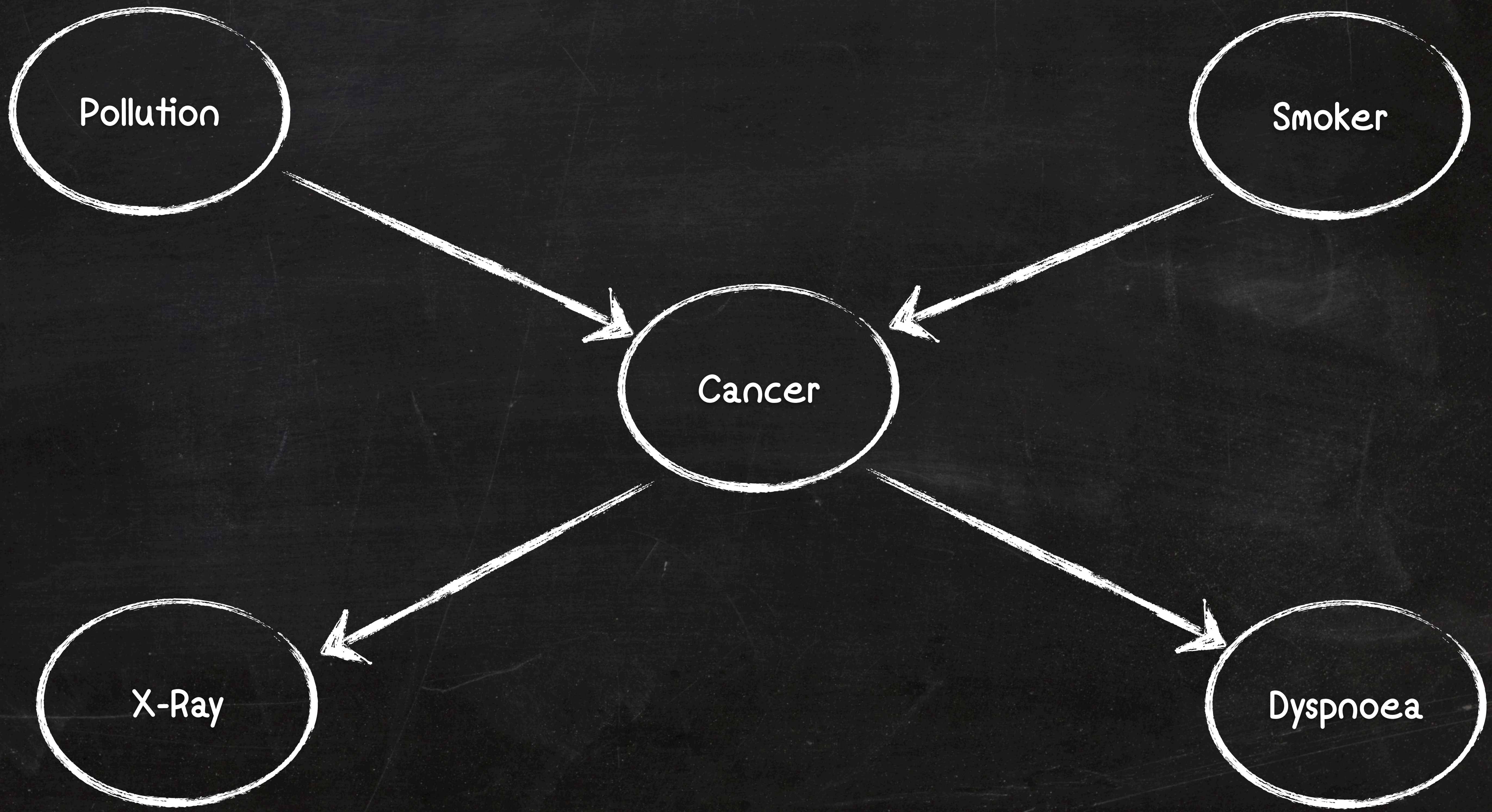
Q: Given the user is leaving the office what actions should Skynet smarthome take?

- the variables could include:
  - day of the week
  - work day or holiday
  - weather
  - season
  - calendar events

Bayesian Networks

- A BN is a probabilistic directed acyclic graphical model that represents a set of random variables and their conditional dependencies
- Vertices may be observable quantities, latent variables, unknown parameters or hypotheses
- Vertices that are not connected represent variables that are conditionally independent of each other

Example: Cancer



Example: Cancer

P	P(P)
Low	0.9
High	0.1

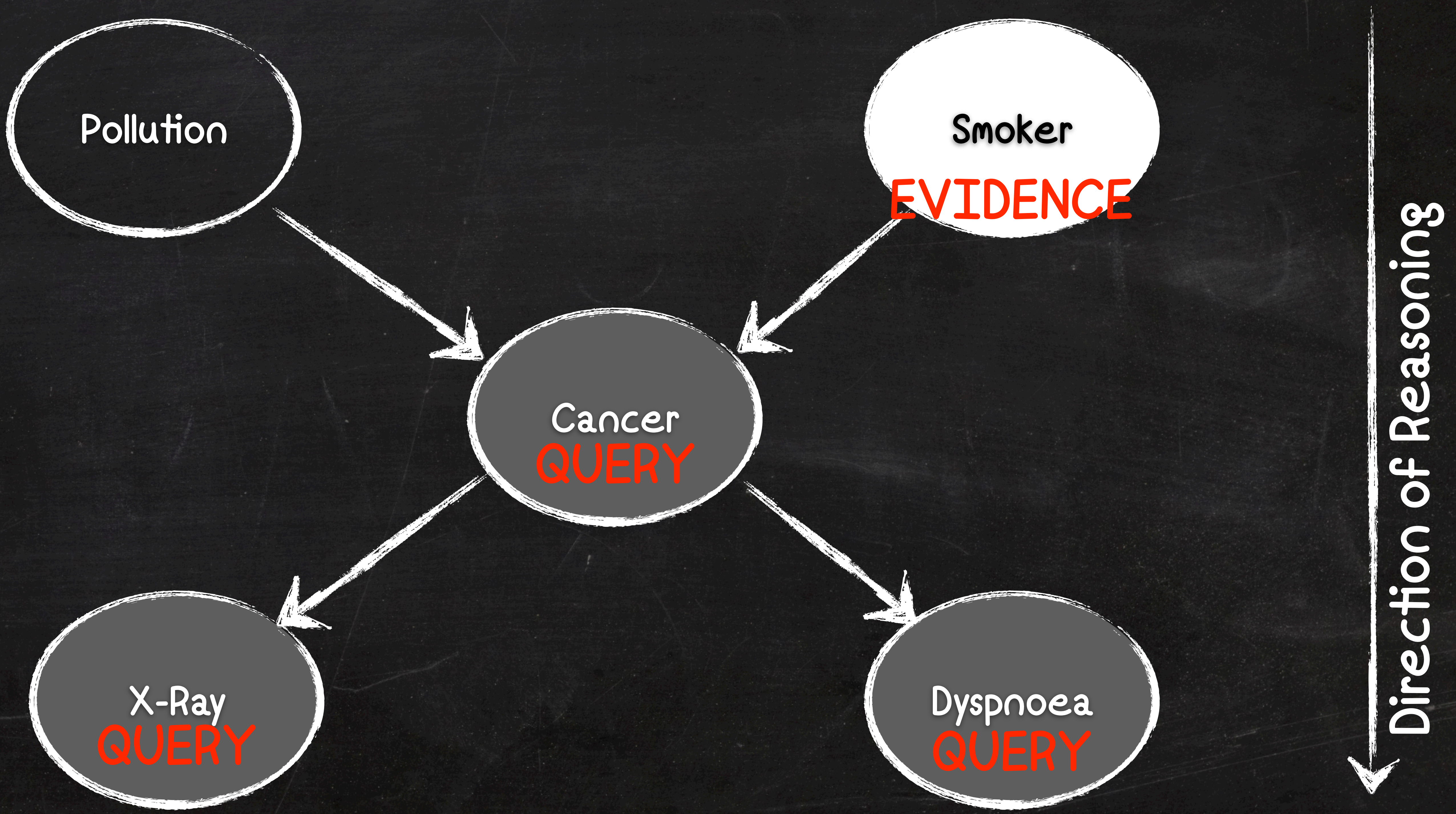
S	P(S)
TRUE	0.3
FALSE	0.7

P	S	$P(C \mid P \cap S)$
Low	T	0.03
Low	F	0.001
High	T	0.05
High	F	0.02

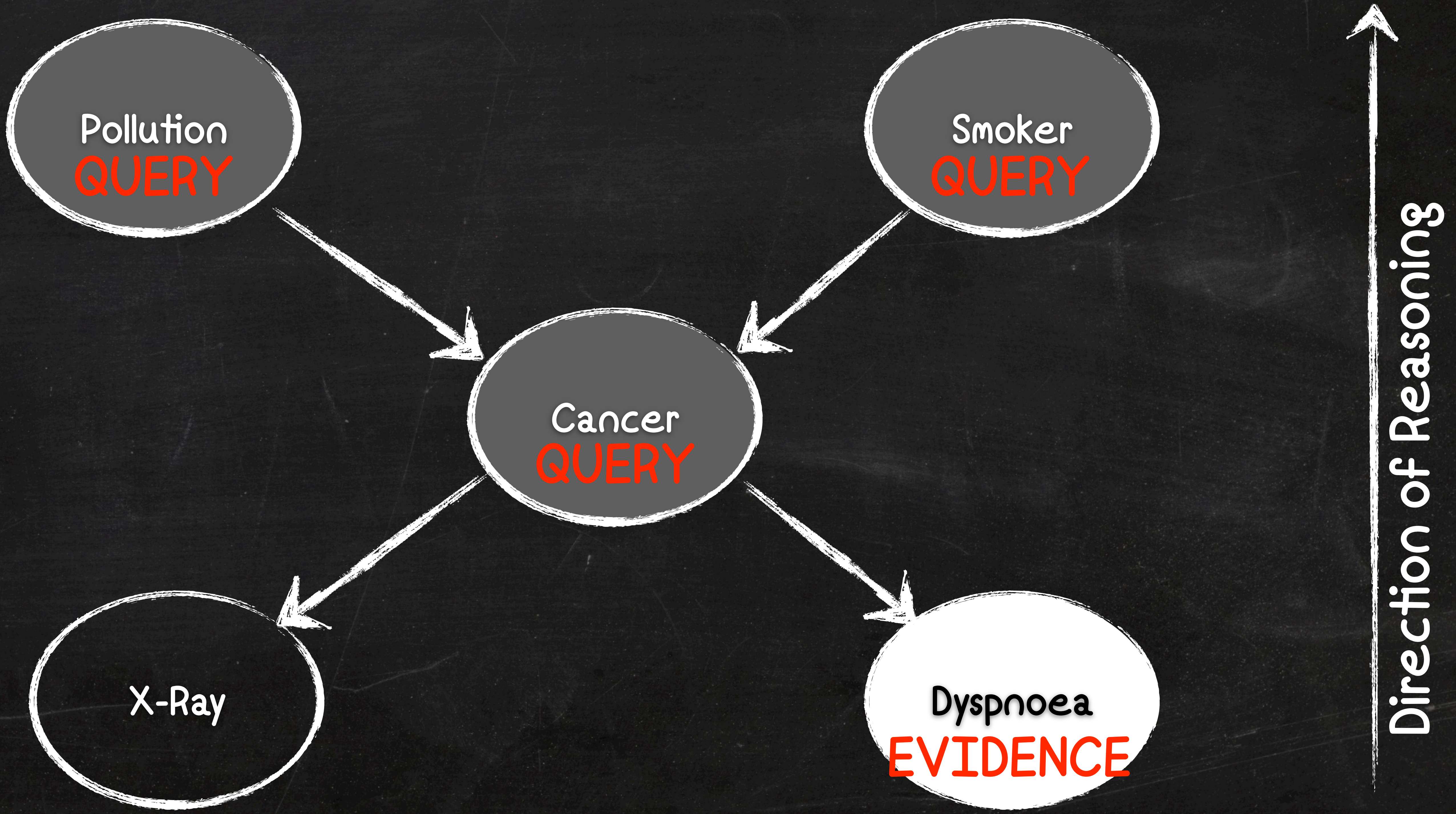
C	P(XRay+|C)
T	0.9
F	0.2

C	P(D+|C)
T	0.65
F	0.3

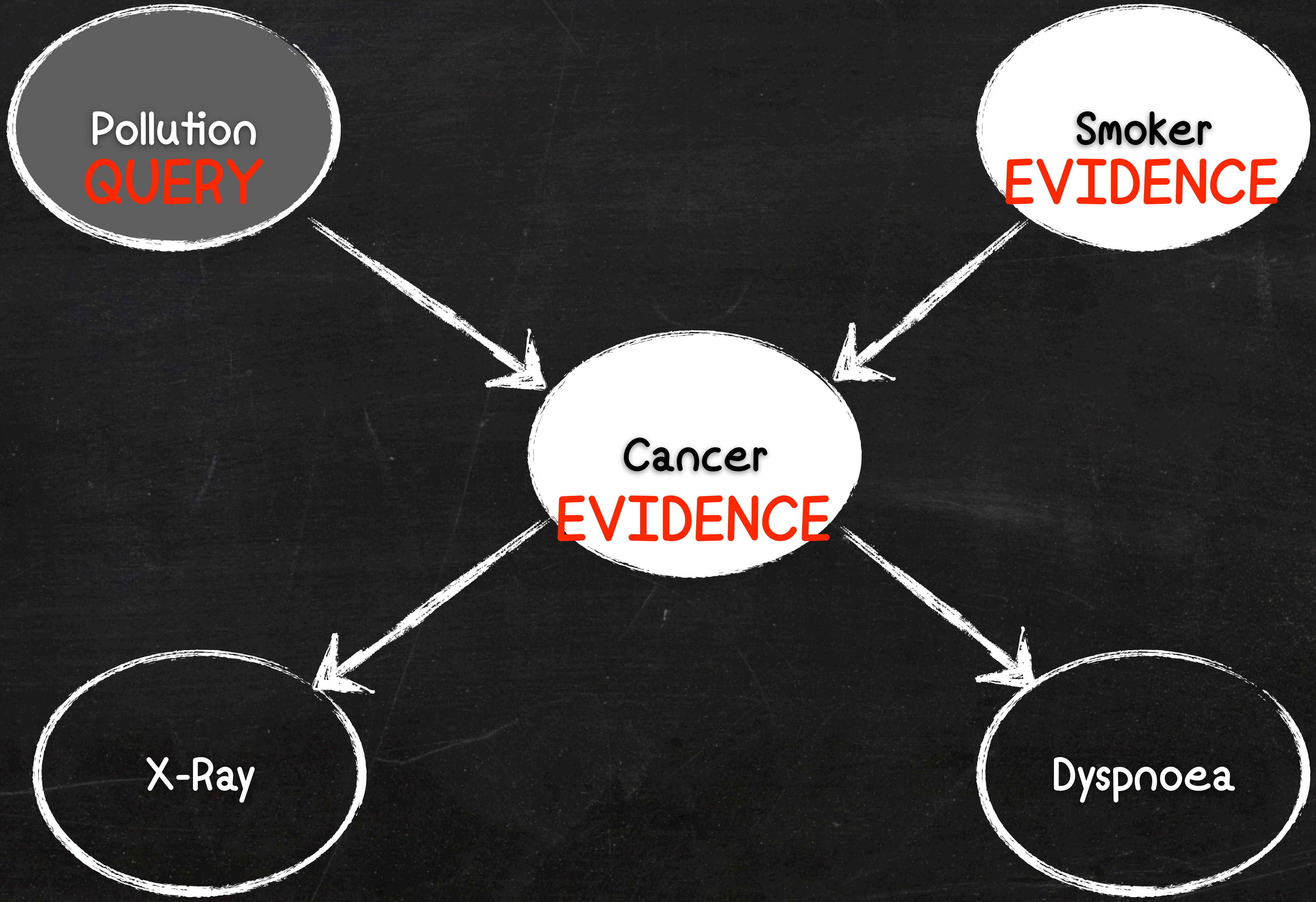
Predictive Reasoning



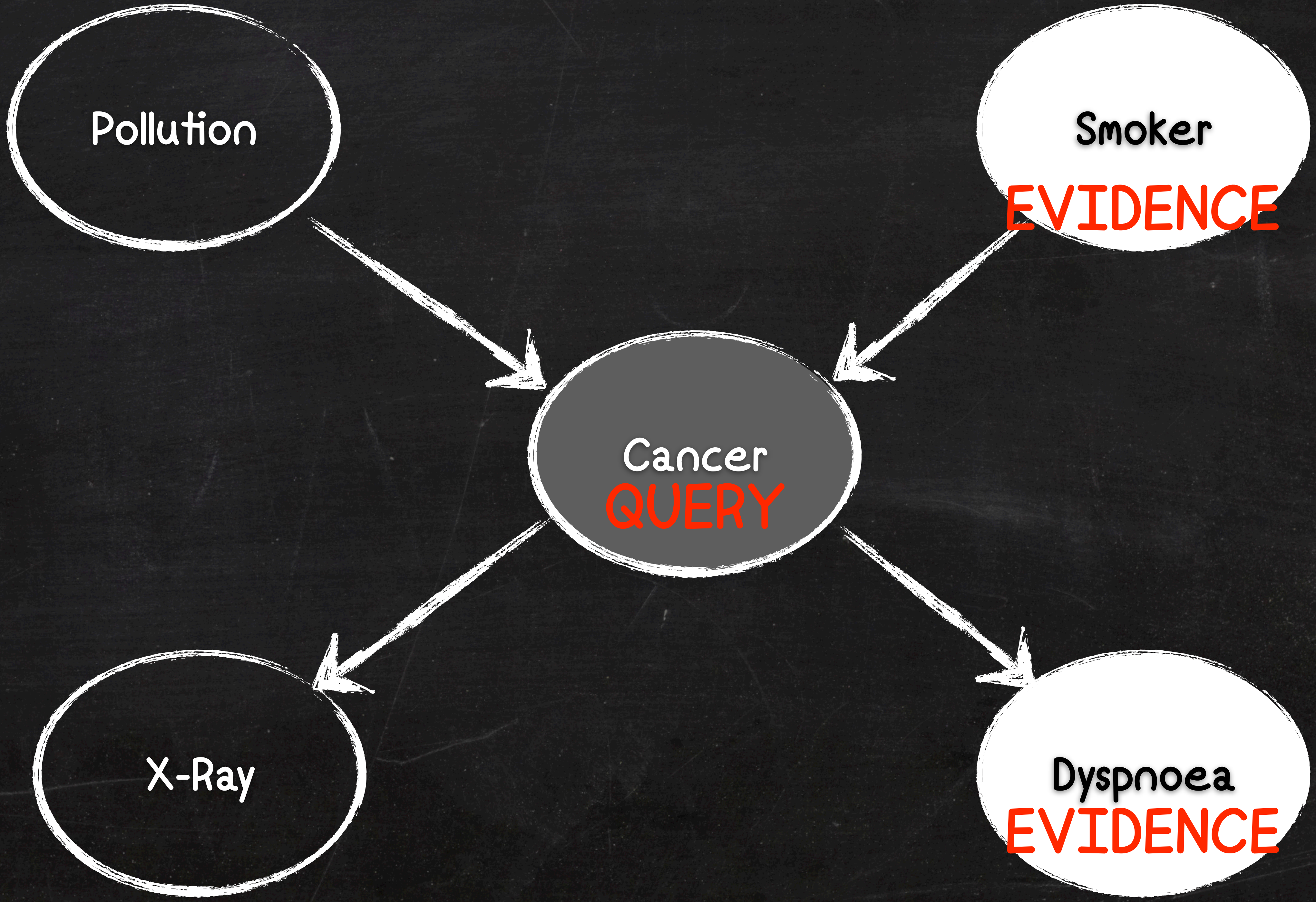
Diagnostic Reasoning



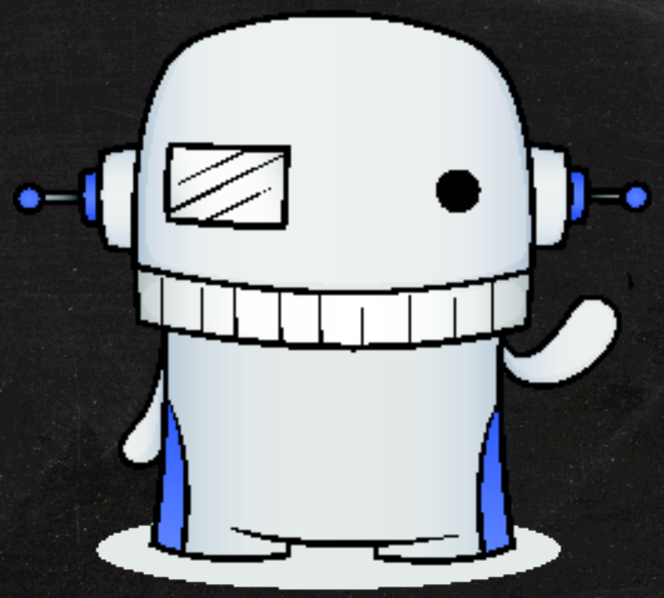
Intercausal Reasoning



Combined Reasoning



- Chapter II -
Markov Models



Stochastic Process

- A stochastic process, or random process, is a collection of random variables representing the evolution of some system over time
- Examples: stock market value and exchange rate fluctuations, audio and video signals, EKG & EEG readings
- They can be classified as:
  - Discrete time & discrete space
  - Discrete time & continuous space
  - Continuous time & discrete space
  - Continuous time & continuous space

A Detour In To Matrix Algebra

Vectors & Matrices

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad B = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \end{bmatrix}$$

1x3 matrix
or vector

3x2 matrix

Matrix Addition

If A and B are two m by n matrices then addition is defined by:

$$A + B$$
$$= C, \text{ where } C_{ij} = A_{ij} + B_{ij}$$
$$= \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & b_{ij} & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & a_{ij} + b_{ij} & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

Matrix Multiplication

If A is n*m matrix and B is an m*p matrix then multiplication is defined by:

$$A \bullet B = AB = \begin{bmatrix} (AB)_{11} & \cdots & (AB)_{1p} \\ \vdots & \ddots & \vdots \\ (AB)_{n1} & \cdots & (AB)_{np} \end{bmatrix}$$

where

$$(AB)_{i,j} = \sum_{k=1}^{m} A_{i,k} B_{k,j}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \bullet \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \bullet \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \bullet \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \bullet \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

The Identity Matrix

If A is n*m matrix and then the identity matrix I is an m*n matrix such that A I=A.
I is defined by:

$$A_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

So the 3*3 identity matrix is:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Inverse Matrix

If A is n*n matrix and then the inverse of A , A^{-1}, is an n*n matrix such that A A^{-1}=I. Non-square matrices do not have inverses

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$a_{11}b_{11}+a_{12}b_{21}=1, \quad a_{11}b_{12}+a_{12}b_{22}=0$$

$$a_{21}b_{11}+a_{22}b_{21}=0, \quad a_{21}b_{12}+a_{22}b_{22}=1$$

Matrix Transposition

If A is n*m matrix and then the transpose of A , A^T, is an m*n matrix defined by:

$$A^T_{ij} = A_{ji}$$

For example, consider a 2*2 matrix:

$$\begin{pmatrix} a & b \\ d & c \end{pmatrix}^T = \begin{pmatrix} a & d \\ b & c \end{pmatrix}$$

End of Detour

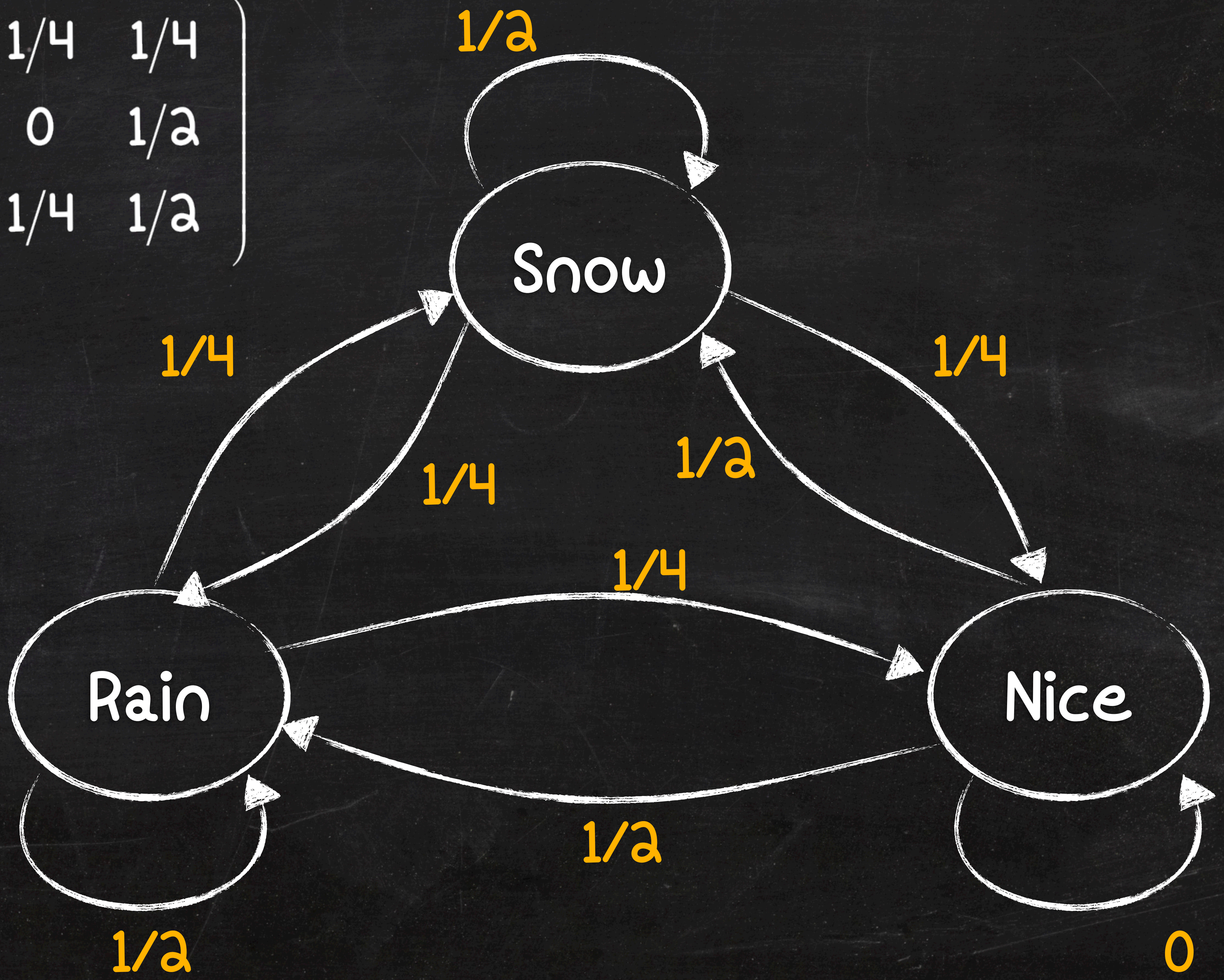
- A Markov chain is a directed graph whose vertices represent states and the edges the probability of transition between the two states
- Frequently we use an adjacency matrix representation of the graph T called the transition matrix
- Hence for a chain with N vertices the transition matrix is NxN
- The initial state of the system, S_0, is also an NxN matrix
- The state evolves according to $S_{n+1} = S_nT = ST^n$
- The value of $S_{n+1(i,j)}$ is the probability of being at that state in step n+1

Example: British weather

- England is the land of rain. We never have two nice days in a row. In fact if a nice day is always followed by either rain or snow. If there is a change from rain or snow only half the time is this a change to sunny weather

$$T = \begin{array}{c} \\ \text{Rain} \\ \text{Nice} \\ \text{Snow} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Rain} & \text{Nice} & \text{Snow} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}$$

$$
T = \begin{array}{c} \\ R \\ N \\ S \end{array}
\begin{array}{c}
\begin{array}{ccc} R & N & S \end{array} \\
\begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}
\end{array}
$$



- We can determine the long-term state of the process by calculating T^n as n increases towards infinity. The system will converge on a stationary value.
- For our weather example we find:

	Rain	Nice	Snow
Rain	4/10	2/10	4/10
Nice	4/10	2/10	4/10
Snow	4/10	2/10	4/10

Absorbing Markov Chains

- A state s_i of a Markov chain is called absorbing if it is impossible to leave it (i.e., $T_{i,i} = 1$)
- A Markov chain is absorbing if:
  - it has at least one absorbing state
  - every state is connected to at least one absorbing state
- In an absorbing Markov chain, the probability that the process will be absorbed is 1 (i.e., $Q^n \to 0$ as $n \to \infty$)

Example: The Wandering Drunk

- Consider a city divided up in some some grid, e.g. square blocks.
- The drunk can move 1 block per turn, each direction has equal probability
- If the drunk reaches Home or the Bar they will stay there
- Questions we can answer:
    - What is the expect time until an absorbing state is reached?
    - How many times does the drunk visit each intersection?

Example: Predictive Texting

Hidden Markov models

- The name is misleading! Nothing is unknown...
- In a basic Markov model the states of the system are visible. E.g. we can see if it is snowing
- In a hidden Markov Model the (entire) state is not directly visible but some outputs dependent on the state are observable
- However we still need to know all the transition probability values!

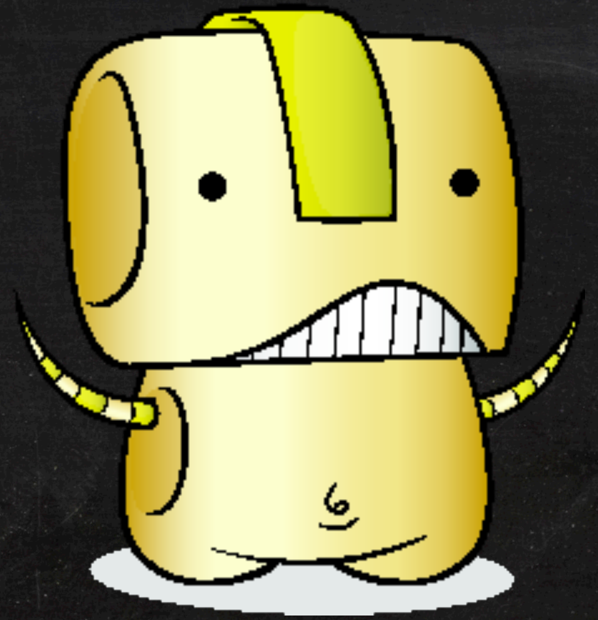
Markov decision processes

- Markov decision processes are an extension of Markov chains; the difference is the addition of actions (allowing choice) and rewards (giving motivation)
- A Markov decision process is a 5-tuple (S, A, P_A, R_A, L)
- The core problem is to choose an action π that will maximise some cumulative function, e.g.

$$\sum_{t=0}^{\infty} L_t R_{A_t}(s_t, s_{t+1}), \text{ where } A_t = \pi(s)$$

- MDPs can be easily solved by linear (e.g. Simplex method) or dynamic programming (e.g. Map-Reduce)

- Chapter III -
Kalman Filters



A Linear Dynamic System

- Continuous time definition:

$$\frac{\partial}{\partial t} x(t) = A \cdot x(t)$$

- Discrete time definition:

$$x_{n+1} = A \cdot x_n$$

- The systems are called linear since given any two solutions x(t) & y(t) then any linear combination of these solutions is also a solution

$$z(t) = \alpha x(t) + \beta y(t)$$

Example: Climate Control

- SkyNet SmartHome™ is able to monitor the temperature of rooms in the house and effect heating/AC to regulate the temperature
- The temperature sensors contain noise
- heating and AC are either ON or OFF
- similar systems exist for humidity

Consider...

$$X_n = AX_{n-1} + Bu_k + \omega_{k-1}$$

Kalman Filters

- Developed ~1960 by Rudolf E. Kálmán, Peter Swerling, and Richard S. Bucy.
- First implemented in NASA as part of the Apollo navigation computer
- Still used in many aeronautic and military applications, e.g. submarines, cruise missiles, NASA Space Shuttle, ISS
- There are generalisation of the basic Kalman filters for continuous time systems as well as non-linear systems

- Kalman Filters work by making a prediction of the future, getting a measurement from reality, comparing the two, moderating this difference, and adjusting its estimate with this moderated value.
- Kalman filters are:
  - discrete
  - recursive
  - Extremely accurate if you have a good model

Overview

current estimate

measured value

$$\hat{X}_k = K_k \cdot Z_k + (1 - K_k) \cdot \hat{X}_{k-1}$$

Kalman gain

previous estimate

Basic Kalman Filters

- Modelled on a Markov chain built on linear operators perturbed by errors that may include Gaussian noise
- The state of the system is represented by a vector of real numbers
- The filter is recursive, i.e. only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state
- typically we describe the algorithm in two phases: predict & update

1: Predict: State Estimation

State transition matrix

Control matrix

$$\hat{x}_{\hat{n}} = A \cdot \hat{x}_{n-1} + B \cdot u_{n-1}$$

Predicted state

Previous estimate of state

Control vector

2: Predict: Error Estimation

State transition matrix

$$\hat{P}_{\hat{n}} = A \cdot \hat{P}_{n-1} \cdot A^{T} + Q$$

Covariance prediction

Previous covariance estimate

Estimated process error covariance

3: Update: Innovation Covariance

Observation Matrix

Measurement Vector

$$\tilde{y} = z_n - H \cdot \hat{X}_{\hat{n}}$$

Covariance Innovation

Predicted state (step 1)

4: Update: Innovation Covariance

Observation Matrix

$$S = H \cdot \hat{P}_{\hat{n}} \cdot H^{T} + R$$

Covariance Innovation

Covariance prediction
(step 2)

Estimated measurement erro
covariance

5: Update: Kalman Gain

Covariance prediction
(step 2)

Observation Matrix

$$K = \hat{P}_{\hat{n}} \cdot H^{T} \cdot S^{-1}$$

Kalman Gain

Covariance Innovation
(step 3)

6: Update: State

Predicted state (step 1)

Covariance Innovation (step 3)

$$\hat{x}_n = \hat{x}_{\hat{n}} + K\tilde{y}$$

New state estimate

Kalman Gain (step 5)

7: Update: Covariance

Kalman Gain (step 5)

$$\hat{P}_n = (I - K \cdot H)\hat{P}_{\hat{n}}$$

New estimate of error

Observation Matrix

Covariance prediction (step 2)

Example: Voltmeter

- Consider a voltmeter measuring a constant DC voltage via a sensor with noise.
- The system can be described by:

$$V_n = V_{n-1} + w_n$$

- Since the voltage is constant using a Kalman filter allows us to filter out the noise w_n
- Also since this is a single state example all matrices are of size 1*1

- A: State transition - since the previous state should equal the current state A=1
- H: Observation transform - since we're taking direct measurements from the sensor H=1
- B: Control matrix - we have no controls so B=0
- Q: Process covariance - since we know the model very accurately Q=0.00001
- R: Measurement covariance - we don't trust the sensor too much so R=0.1
- $\hat{X}$: Initial state estimate = any number
- $\hat{P}$: Inital covariance estimate = 1 (because)

$$\widehat{X}_{\hat{n}} = \widehat{X}_{n-1}$$

$$\widehat{P}_{\hat{n}} = \widehat{P}_{n-1} + 0.00001$$

$$\tilde{y} = z_n - \widehat{X}_{\hat{n}}$$

$$S = \widehat{P}_{\hat{n}} + 0.1$$

$$K = \widehat{P}_{\hat{n}} \cdot S^{-1}$$

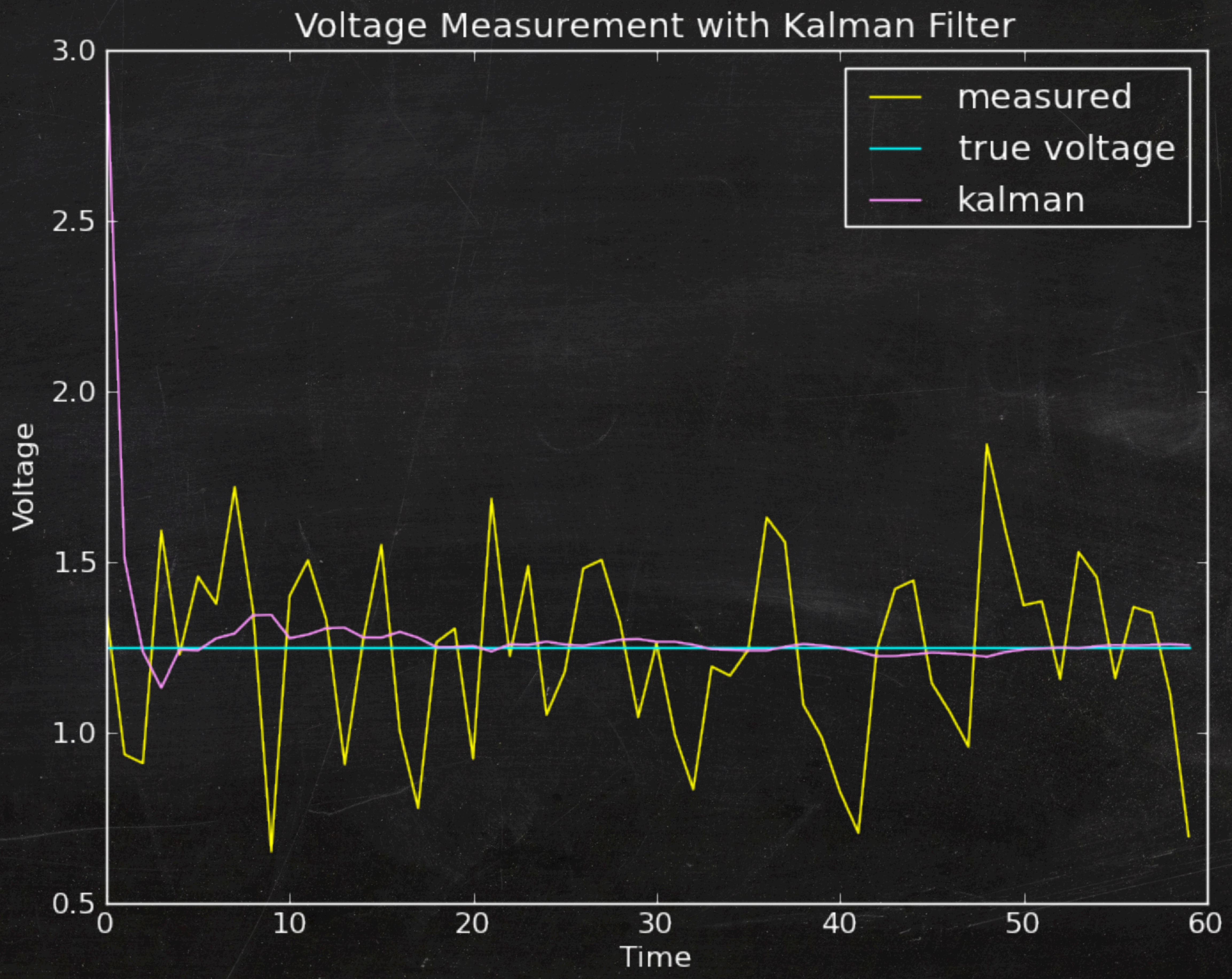
$$\widehat{X}_n = \widehat{X}_{\hat{n}} + K\tilde{y}$$

$$\widehat{P}_n = (I - K)\widehat{P}_{\hat{n}}$$

$$K = \frac{\widehat{P}_{n-1} + 0.00001}{(\widehat{P}_{n-1} + 0.00001) + 0.1}$$

$$\widehat{X}_n = Kz_n - (K - 1)\widehat{X}_{n-1}$$

$$\widehat{P}_n = (1 - K)(\widehat{P}_{n-1} + 0.00001)$$


Voltage Measurement with Kalman Filter
measured
true voltage
kalman
Voltage
Time
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60

As a programmer your challenge is to find the right filter model and determine the values of the matrices

Example: Robo-copter

Freezin Eskimo

XCell Tempest Helicopter

RL: Helicopter

- http://library.rl-community.org/wiki/Helicopter_(Java)
- Sensors to determine:
  - bearing
  - acceleration (velocity)
  - position (GPS)
  - rotational rates
  - inertial measurement unit
  - and more...

Summary

Events & Random Variables

Events & Random Variables

Conditional Probability

Events & Random Variables

Bayesian Networks

Conditional Probability

Events & Random Variables

Markov Chains

Bayesian Networks

Conditional Probability

Events & Random Variables

Bayesian Networks

Markov Chains

Random Walks

Conditional Probability

Events & Random Variables

Markov Chains

Bayesian Networks

Random Walks

Markov Decision Processes

Conditional Probability

Events & Random Variables

Markov Chains

Kalman Filters

Bayesian Networks

Random Walks

Markov Decision Processes

Conditional Probability

Events & Random Variables

Markov Chains

Kalman Filters

Bayesian Networks

Random Walks

Markov Decision Processes

Conditional Probability

Books

- Introduction to Probability - Grindstead & Snell

  http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html

- Bayesian Artificial Intelligence - Kevin B. Korb & Ann E. Nicholson
- An Introduction to Stochastic Modelling - Mark A Pinsky & Samuel Karlin
- Stochastic Processes and Filtering Theory - Andrew H. Jazwinski
- Artificial Intelligence: A Modern Approach - Stuart Russell and Peter Norvig

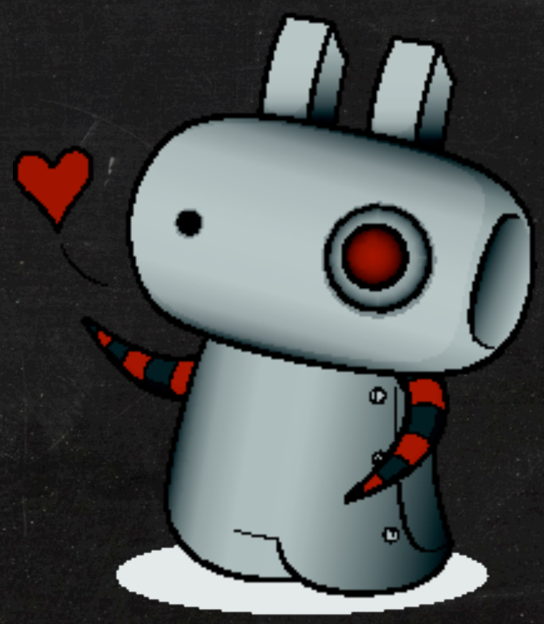
Java Libraries

- Apache Commons Math: http://commons.apache.org/proper/commons-math/
- Colt - high performance data structures and algorithms: http://dst.lbl.gov/ACSSoftware/colt/
- Parallel Colt: https://sites.google.com/site/piotrwendykier/software/parallelcolt
- JBlas - high performance Java API for native libraries LAPACK, BLAS, & ATLAS: http://mikiobraun.github.io/jblas/
- The rest... http://code.google.com/p/java-matrix-benchmark/
- Jayes - A Java framework for Bayesian Networks

Other Resources

- http://www.probabilitycourse.com
- http://masanjin.net/blog/bayesian-average - detailed derivation of bayesian averaging via normal distributions
- http://fulmicoton.com/posts/bayesian_rating/ - an alternative derivation of bayesian "averaging"
- http://www.tina-vision.net/docs/memos/1996-002.pdf - a beautifully simple derivation of Kalman filters
- http://www.intechopen.com/books/kalman-filter - articles on applications of Kalman filters

Thank You



James McGivern

Probably, Definitely, Maybe