CO902
Probabilistic and statistical inference

Lecture 1

Tom Nichols
Department of Statistics &
Warwick Manufacturing Group

t.e.nichols@warwick.ac.uk

Preliminaries

Contact

Email: t.e.nichols@warwick.ac.uk
Office: D0.03 (Statistics)
Office Hours: 15:00-16:00 Tuesdays

Format

Lecture 2h/week: Motivate, guide reading
Labs 2h/week: Build Matlab, data experience
Problem Sets: Set biweekly, not graded

Survey

Probability? Statistics? Real Data Analysis?
Matlab? (R?)

Evaluation

Week	Assessment	Issued	Deadline	How assessed	%credit
5	Written assignment	Mon 28 Feb	Mon 4 Mar, 9am	Written report	25
8 & 9	Critical reading assignment	Mon 11 Feb	25 Feb & 4 Mar, 9-11am	In-class presentation	25
10	Oral Examination	11 & 12 Mar		Oral examination	50

Written assignment =
Report of practical data analysis

Critical reading assignment =
In class presentation of paper review

Oral Examination
'viva' ~20-30 min

Schedule

Class Calendar

Week	Lecture		Lab	
1	Mon 7 Jan	Lecture 1	Tue 8 Jan	Lab 1
2	Mon 14 Jan	Lecture 2	Tue 15 Jan	Lab 2
3	**Thr** 24 Jan	Lecture 3	**Fri** 25 Jan	Lab 3
4	Mon 28 Jan	Lecture 4	Tue 29 Jan	Lab 4
5	Mon 4 Feb	Lecture 5	Tue 5 Feb	Lab 5
6	Mon 11 Feb	Lecture 6	Tue 12 Feb	Lab 6
7	Mon 18 Feb	Lecture 7	Tue 19 Feb	Lab 7
8	Mon 25 Feb	Presentations	Tue 26 Feb	Lab 8
9	Mon 4 Mar	Presentations	Tue 5 Mar	Lab 9
10	Mon 11 Mar	Lecture 8		
Oral Exams	Tue 12 Mar; also Mon 11 Mar PM, if needed			

Homework for Tuesday! Install Matlab on your laptop

http://www2.warwick.ac.uk/services/its/servicessupport/software/matlab

(or search for "Matlab" on Warwick site)

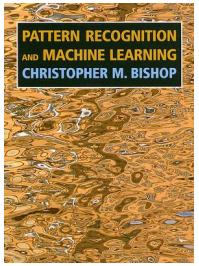
Books

- Web Schedule has references for each class' lectures

- Main textbook:

Bishop

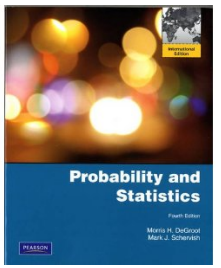
"Pattern Recognition and Machine Learning", Springer, 2006



- Others:

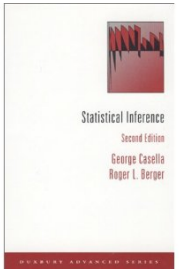
De Groot & Schervish

"Probability and Statistics (4th Ed)", Pearson 2011



Casella & Berger

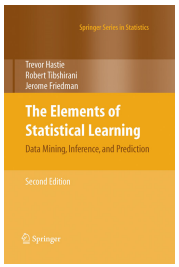
"Statistical Inference (2nd Ed)", Duxbury Press, 2001



Hastie, Tibshirani, Friedman

"The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed)", Springer, 2009

Free PDF! http://www-stat.stanford.edu/~tibs/ElemStatLearn/



Outline of course

A. Basics: Probability, random variables (RVs), pmfs and pdfs, introduction to statistical inference

B. Supervised learning: Regression, classification, including high-dimensional issues and intro to Bayesian approaches

C. Unsupervised learning: Dimensionality reduction, clustering and mixture models

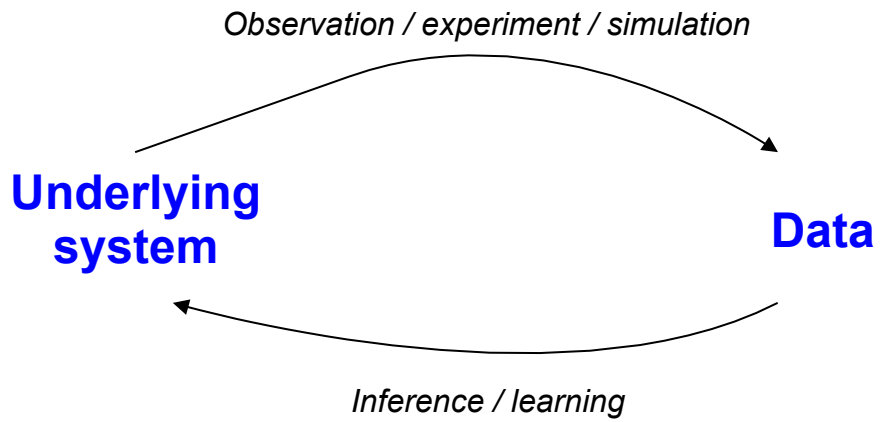
D. Networks: Probabilistic graphical models, learning in graphical models, inferring network structure

Overview & Motivation

- What it's all about**:**
  Machine learning and statistical inference.
  - We'll highlight open problems, current research areas as we go along
  - Light on mathematical/statistical details,
    focus on intuition, application of methods
  - Will make use of molecular biology & other areas to *illustrate* issues: (modern) bio is interesting for quantitative scientists, offers many opportunities for research
  - But equally, ML and stat inf are *highly* general approaches

Inference: from data to prediction and understanding

- Inference is about "reasoning backwards" – going from noisy data to saying something about the underlying system or process



- Concepts and methods in **machine learning** and **statistical inference** are highly general, and find application in biology, linguistics, AI, CS, engineering...
- Three examples: biology, social networks, images

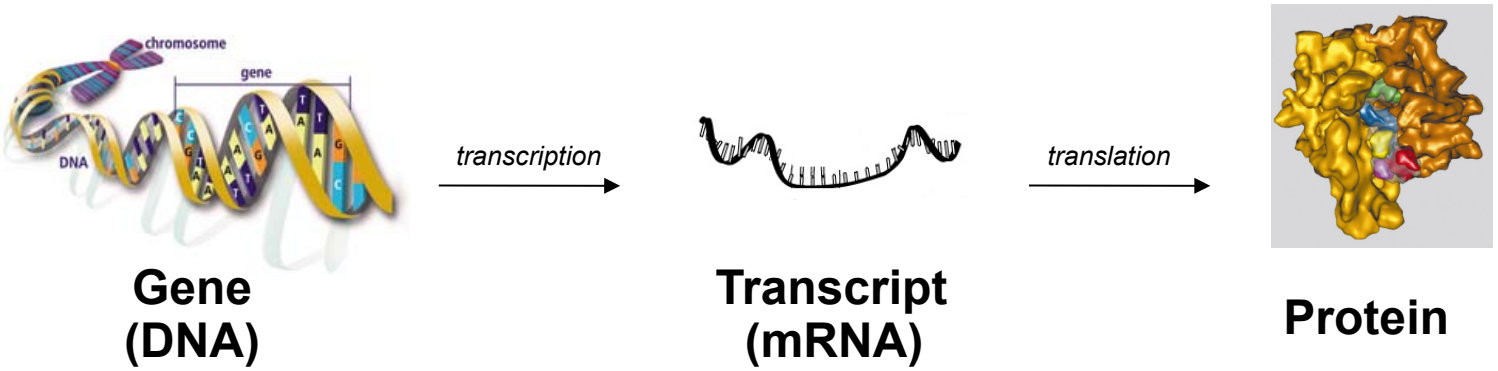
Example I: biology

?

- Biological systems
  - Extraordinarily complex (~10^{13} cells, each with ~10^{6} components, of potentially ~10^{5} distinct types!)
  - Poorly understood – we don't have a really good understanding at any level, from cell to organ to organism
  - But biology is undergoing a rapid transformation into a quantitative science
  - Many see this as arguably the key research challenge of the 21st century

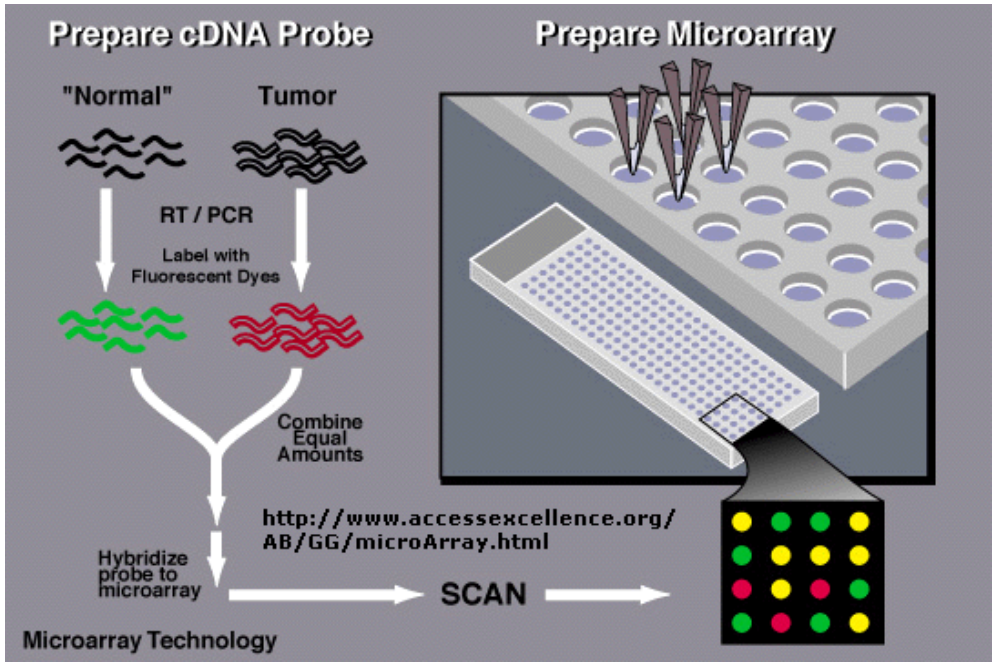
Molecular biology in one slide

- Last half-century has seen remarkable progress in understanding basic mechanisms of living systems



- Picture is a huge simplfication, but a useful framework
- Not long ago, biology was all about painstakingly measuring one gene, or one protein, and writing a paper about it: not much role for number crunching!
- What's changed in recent years is the ability to measure lots of things at once in an automated fashion

Example: gene expression microarrays

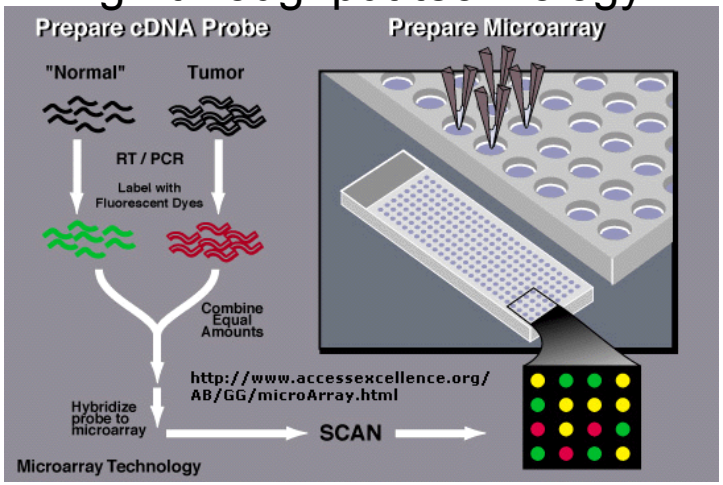


- Roughly speaking, gene expression is the "activity level" of a gene
- Microarrays can measure all 23,000 genes in one go!
- That is, you get a vector in R^{23k} under each condition, or across a range of conditions, through time etc...
- Now widely used in all areas of biomedical discover
- Have triggered off an enormous amount of research in quantitative biology: long-term aim is to "reverse-engineer" living systems

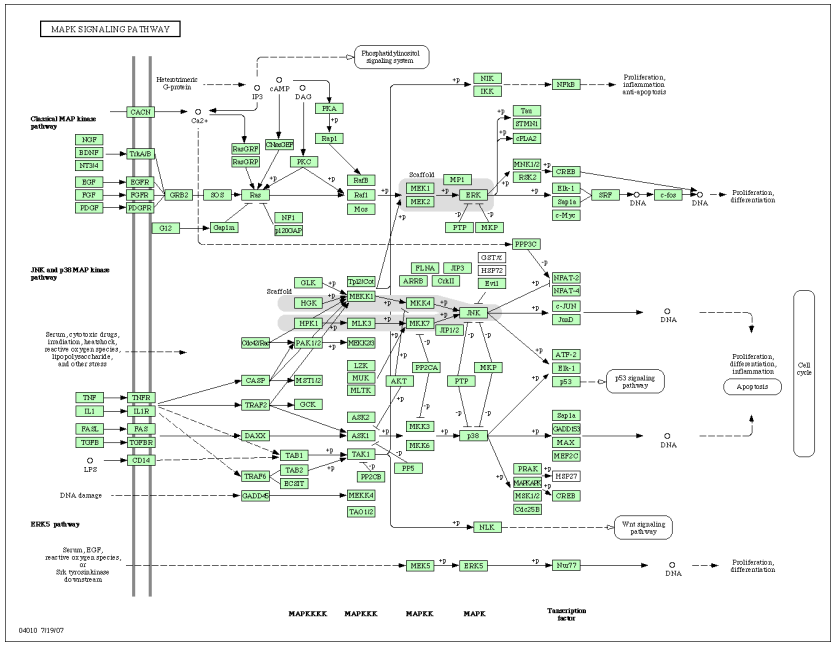
“Reverse-engineering” living systems

- Rapid progress in making system-wide measurements on biological systems
- We now know many of the many thousands of components - genes, transcripts, proteins, small molecules - whose interplay governs biological function
- We can measure many of these “players” on a large-scale, thousands at a time
- Inference problem is then using these data to (i) say something about how the system works, and/or (ii) make predictions about what is likely to happen under certain conditions, e.g. if a gene is present, or a certain drug is used

High-throughput technology

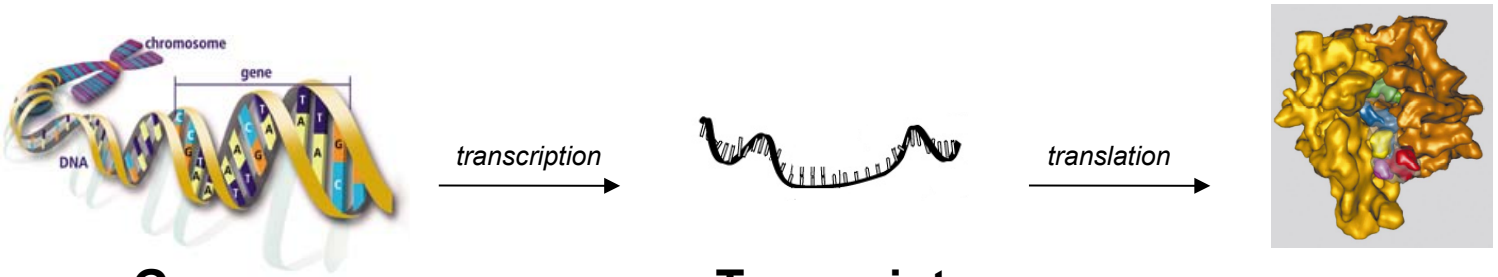


Detailed models of cell function



High-throughput data

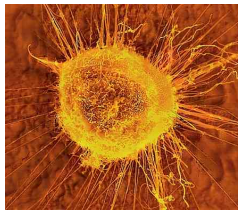
- Various technologies allow us to measure the state of the system at various levels

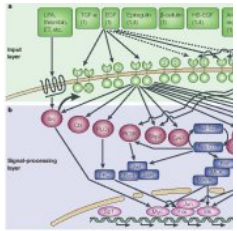


- One example from current research: breast cancer

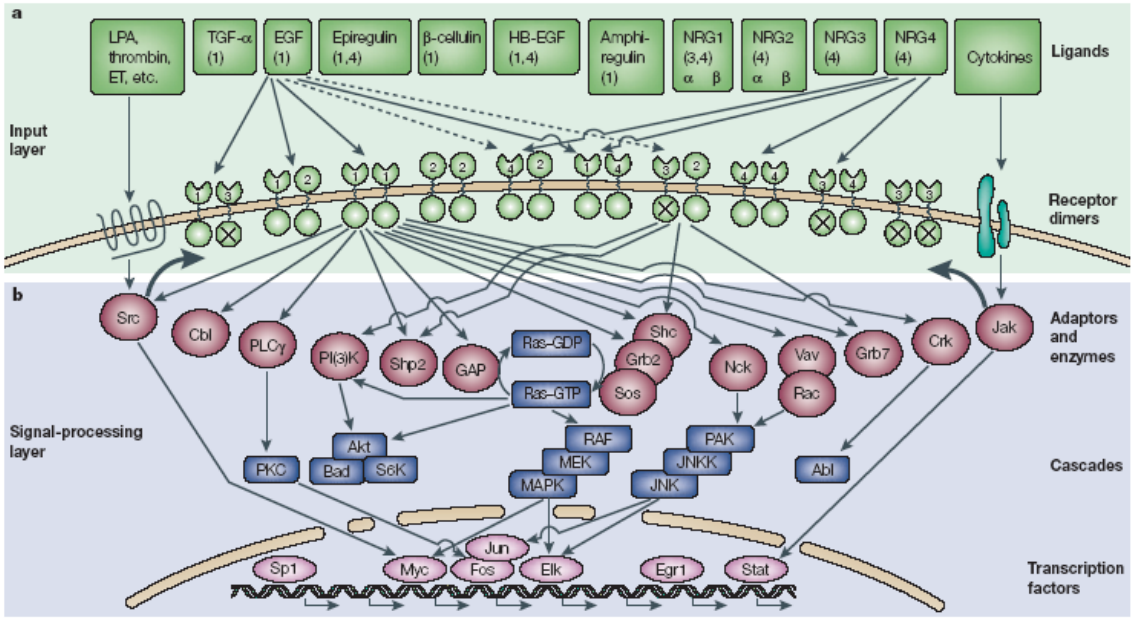
Breast cancer

- Most common cancer among women
  - >40k new cases/year in UK
  - ~1 in 9 women get BC during lifetime
- Aberrant functioning in a networks of molecules called **signalling networks** is a key molecular cause of BC
- **Signalling networks** are biological information-carrying systems
- Take "messages" from outside the cell through layers of "circuitry" to bring about major changes like cell division, cell-death etc.
- Drugs like **Herceptin "**target" systems like this
- EGFR – a signalling system



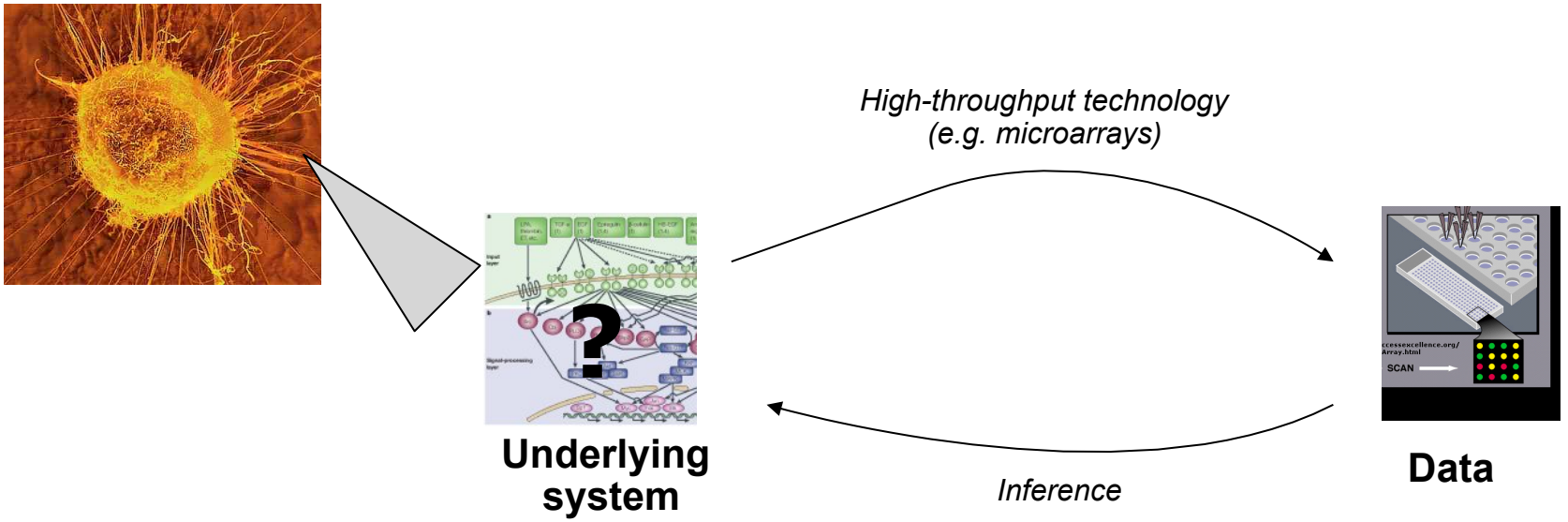


EGFR: a biological network



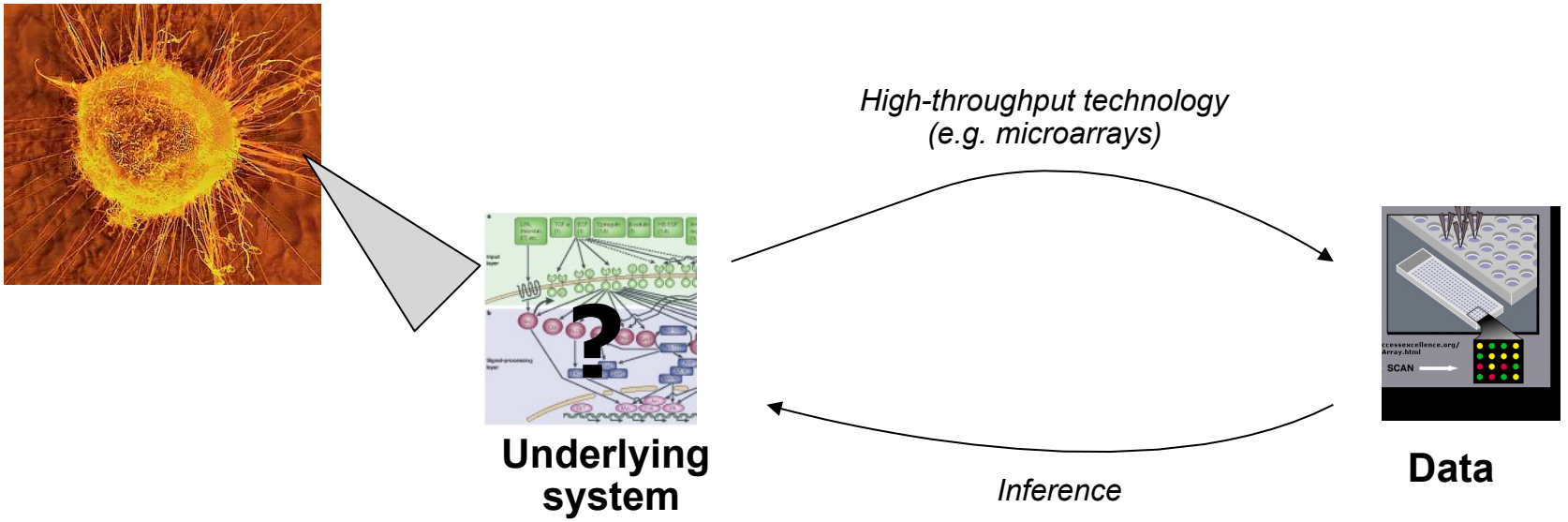
- Heavily involved in BC
- Microarrays and other data can shed light on system (which is actually not very well understood)

Prediction and modelling using high-throughput data



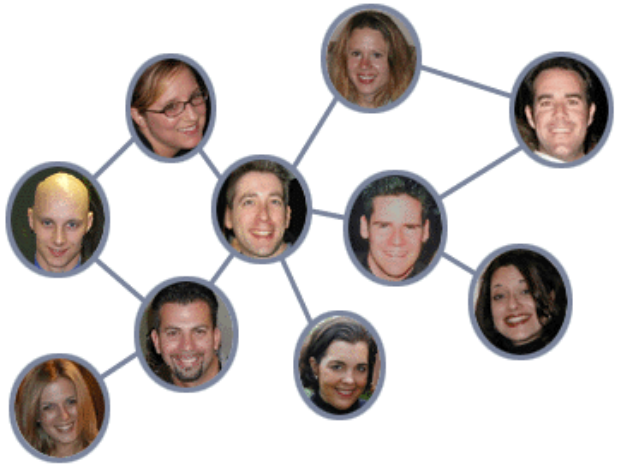
- **Prediction**
  - "if molecule A and B are present at high levels, drug X is likely to work"
  - "the molecular signature of this tumour tells us it's subtype Z"
- **Modelling**
  - "are there new subtypes of cancer with special molecular signatures?"
  - Complex Qs about network connectivity or dynamics

Prediction and modelling using high-throughput data



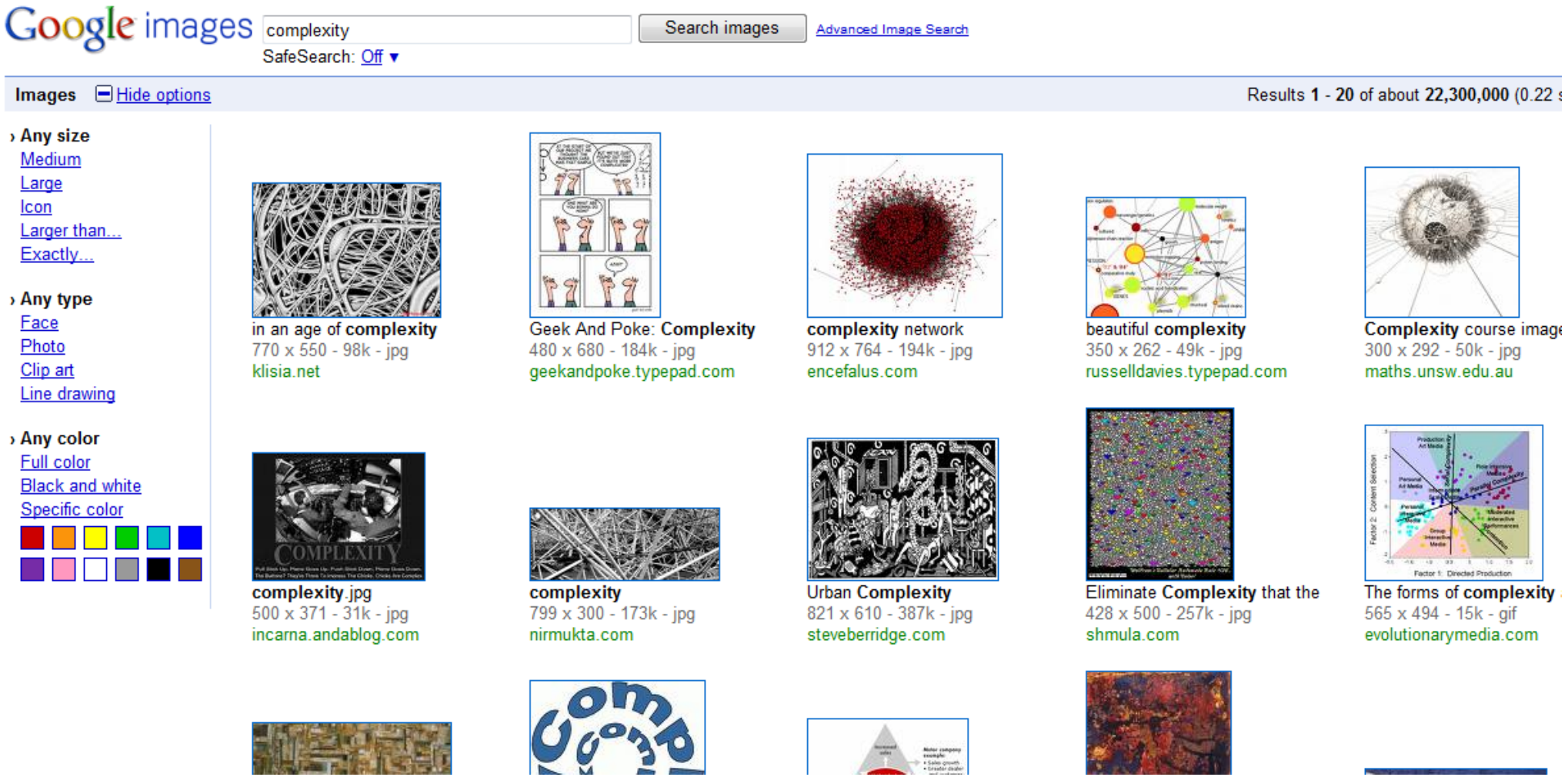
- Task of understanding all of this has only just begun
- Quite simply tonnes to do, open problems almost wherever you look

Example II: social networks



- **Social networking** sites like Facebook are ubiquitous
- Lots of data available on who is friends to whom, shared interests etc.
- Question: does knowing what books X's friends are reading help make a prediction about what books X will want to buy? How good is the prediction? What can we say about the spread of ideas? *Markov models* will have a role to play in this area
- Equally: related questions about how ideas - even diseases - spread, understanding social dynamics etc.
- Netflix prize

Example III: object categorization & image search



Research opportunities

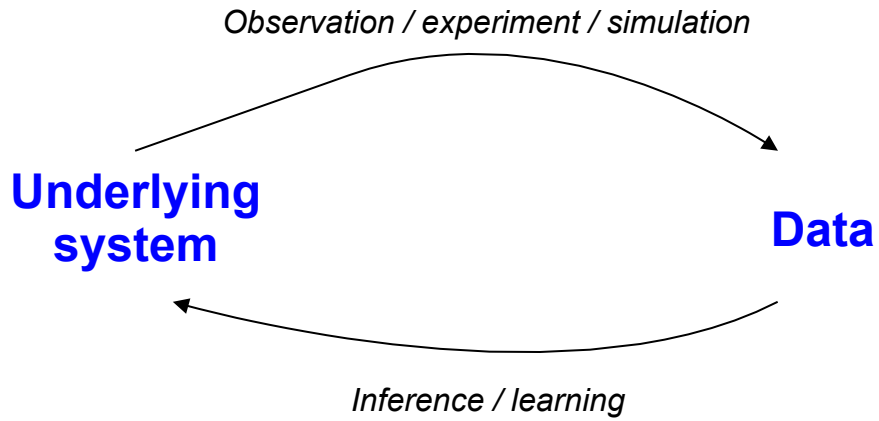
- All of these areas are **rich in research opportunities**
- Far from being "done and dusted", these areas are still on the frontier
- Any substantive advance makes a big difference: smart people are needed!

Supervised vs unsupervised learning

- **Supervised learning**, or "learning with a teacher", where the dataset has "answers", and you want to learn a general rule. E.g. Cancer classification, Netflix

- **Unsupervised learning**, is simply looking for structure in data, e.g. learning object categories from dataset of images, discovering cancer subtypes by "clustering" data from cancers

Data, predictions, understanding



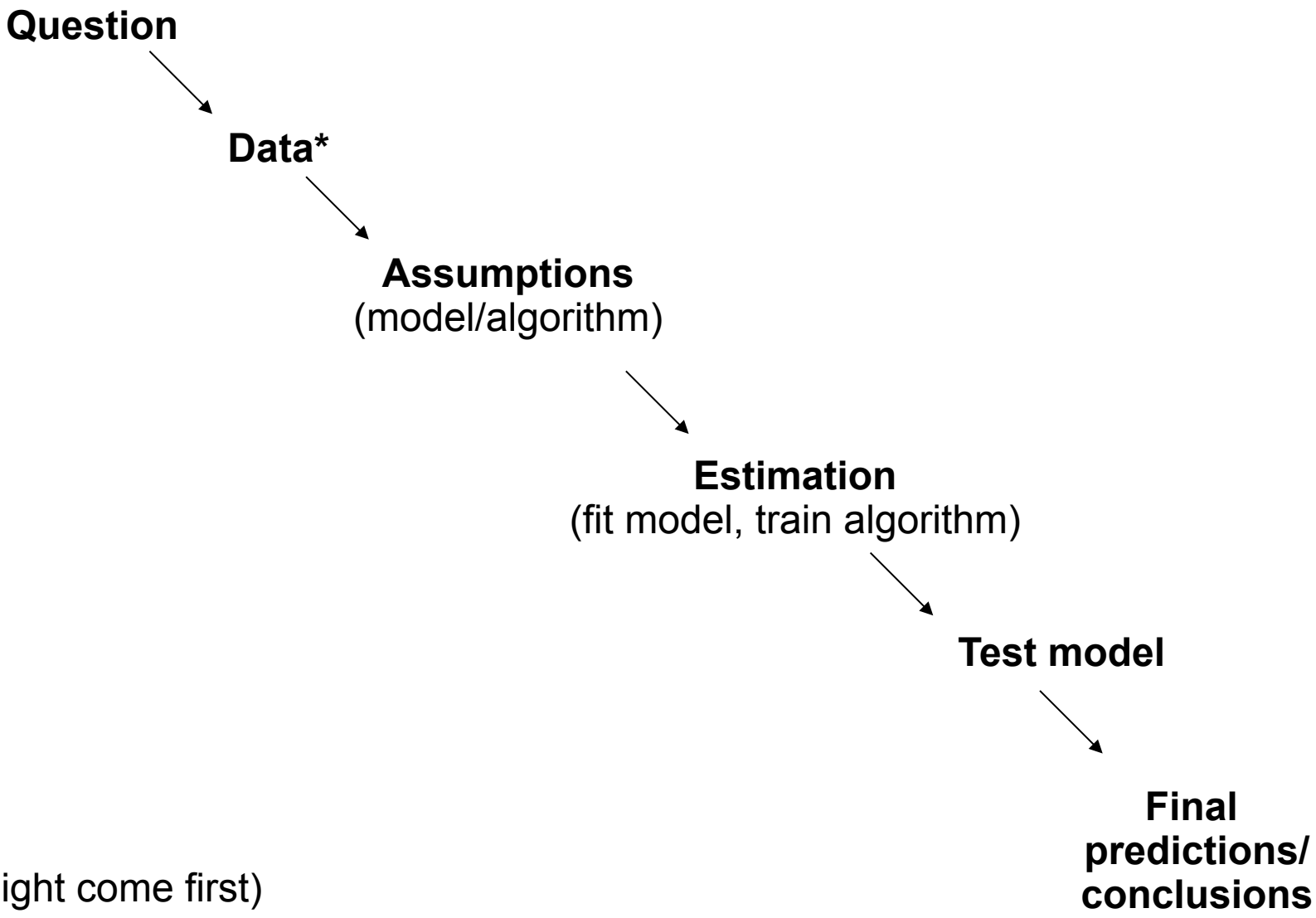
- All these problems characterized by **variability**, in underlying system, in data, or on account of gaps in our understanding
- Rarely enough data to be absolutely certain of any conclusions drawn
- **Probabilistic models** and **machine learning** have become very popular in these and other fields because they allow us to

(i) take account of **variability** in a principled manner and

(ii) **quantify our uncertainty** about conclusions.

- This course is about **machine learning** and **statistical inference**

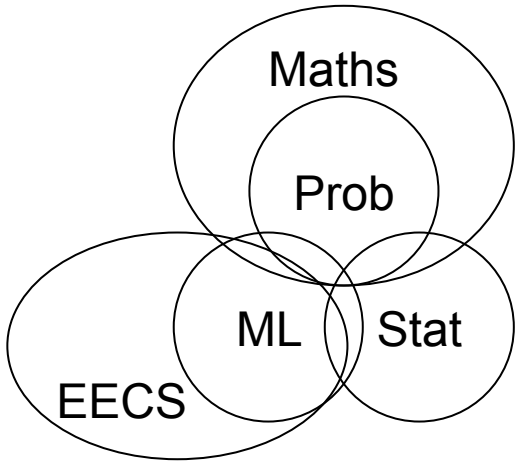
Stages in data modelling



(*might come first)

Probability, statistics, machine learning

- **Probability**: mathematics of chance, framework for inference
- **Statistics**: founded largely by Fisher, huge influence on how science is done, applications in analyzing data, experimental design, agriculture, clinical trials etc.
- **Machine learning**: emerged out of AI, now a distinct discipline, especially successful in developing statistical approaches for broad range of problems not always obviously statistical, and in advancing associated computer methodology. Current application areas include biology, finance, engineering, AI, etc.



Outline of course

A. Basics: Probability, random variables (RVs), pmfs and pdfs, introduction to statistical inference

B. Supervised learning: Regression, classification, including high-dimensional issues and intro to Bayesian approaches

C. Unsupervised learning: Dimensionality reduction, clustering and mixture models

D. Networks: Probabilistic graphical models, learning in graphical models, inferring network structure

Probability

- Probability: mathematics of chance events
- Much of this course will utilize probabilistic ideas, so we have to learn some probability - but this is *not* a mathematics module, emphasis is on intuition and not rigour

Sample space, outcomes, events

- Sample space: set of everything that can happen in setting of interest
- Outcomes: elements of sample space
- Events: subsets of sample space
- Example: toss a coin twice
  - SS = {HH, HT, TH, TT}
  - Four outcomes
  - Events:

"First toss is heads"

"Both tosses are tails"

- Q: When you actually toss a coin all sorts of things "happen" – it spins, lands, settles down etc. Why are the outcomes just H and T ?

Probability

- To every event A, assign a number P(A). To qualify as probabilities the P(A)s must satisfy:

$$
\begin{aligned}
P(A) &\geq 0 \\
P(\mathcal{S}) &= 1 \\
P\left(\bigcup_i A_i\right) &= \sum_i P(A_i) \quad \text{for disjoint } A_i
\end{aligned}
$$

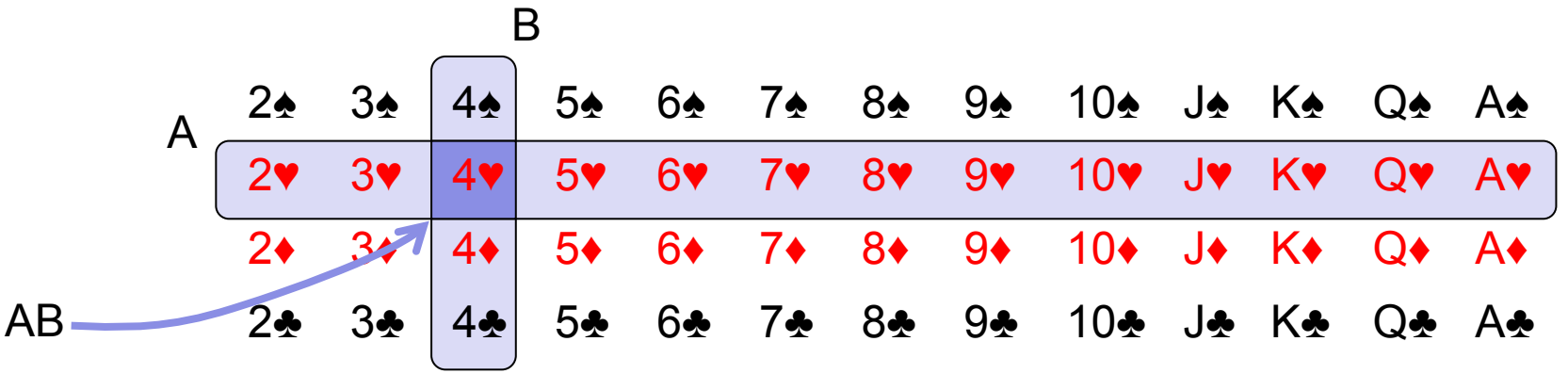
- Many useful properties follow

Interpretation of probabilities

- Intuitively P(A) represents how likely the event A is.
- Two views:
  - Limiting frequency: e.g. coin tosses
  - Measure of uncertainty (possibly subjective): e.g. climate change
- Distinction doesn't make *too* much difference for practical problem solving, but we'll run into it again when we look into inference

Joint probability

- Consider drawing a card from a deck
- We can think of events
  - A: "getting hearts"
  - B: "getting 4"
  - "getting hearts AND getting 4", that is the *intersection* AB
- There are four suits and 13 numbers
- Think of a 13x4 table, with the probability of each joint event stored in a cell
- What would you *do with the table* to get P(B), i.e. P(getting 4)
- Generalizing, we get the **law of total probability or "sum rule"…**



Sum rule

- If *A*'s *partition* the sample space, *P(B)* can be obtained simply by "summing out" the *A*'s...

$$P(B) = \sum_i P(A_i B)$$

- Simple rule, but profoundly important, e.g.:
  - Calculating the probability of a particular "network model", we may be interested in the connectivity of the network, *not* the specific parameters. But the model is easy to specify with both network AND parameters. So we "sum out" the parameters to get the probability we need
- Also know as "Law of Total Probability"

Data

- Practical problems have **numbers**: data of one kind or another
- The concept of a **random variable (RVs)** links our set theoretic story to numbers

- Once introduced, we'll mostly deal with RVs, but the sample space is always in the background

Random variables

$$\mathcal{S} = \{H, T\}$$

Outcome
Sample space
Event

$$X(s),\ s \in \mathcal{S}$$

Random variable
function *from* sample space to reals
e.g. X(H) = 1, X(T) = 0

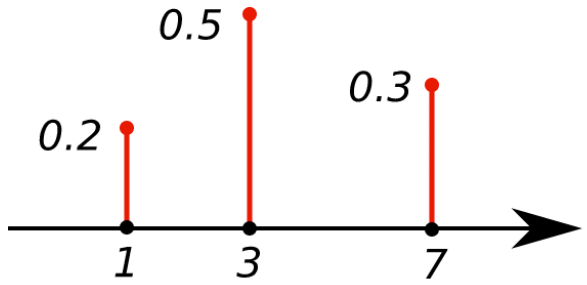
Probability mass function

- Let X be a Discrete RV (i.e. the range of X is finite or countably infinite)
- Then we can write the probability that X takes on a specific value x in terms of the underlying sample space:

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

- This is a **probability mass function or pmf**
- It's a function (from range(X) to unit interval [0,1]) of the possible value x
- Can think of an array of numbers, one column for each possible value of the random quantity

x	1	3	7
P(X=x)	0.2	0.5	0.3



Probability mass function

- Pmf must sum to one, because the RV covers the whole sample space. Simply put: something has to occur, e.g. a coin can't be neither H nor T!

$$\sum_{x \in \mathcal{X}} P(X = x) = 1$$

- **What is the pmf for an RV *X* representing the toss of a fair coin?**
- **What is the pmf for an RV *Y = (1+X)* ? What about *Z= X_1 + X_2*, where *X_1, X_2* are two tosses of the coin ?**

Expectation

- For (discrete) RV *X*

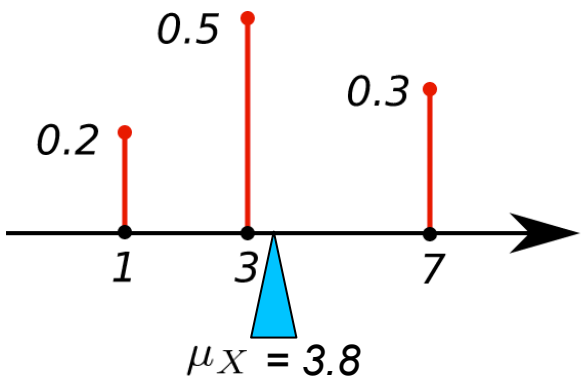
$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

is the **expectation** or **expected value** or **mean** of *X*

- This is simply a weighted sum of all possible values, weighted by how likely it is that we get each such value
- The mean is often written as μ_X

x	1	3	7
P(X=x)	0.2	0.5	0.3

1×0.2 + 3×0.5 + 7×0.3 = 3.8



- More generally, if *g(X)* is a function of RV *X*, *g(X)* is also an RV, with expected value:

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) P(X = x)$$

Variance

- The **variance** is the expected value of $g(X) = (X - E[X])^2$, i.e. the mean squared deviation from the expected value:

$$VAR(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- This gives an indication of how much the RV varies, hence the name
- Often denoted by σ_X^2

- The (positive) square root of *VAR(X)* is the **standard deviation** of *X*, or *STD(X)* (this has the advantage of being on the same scale as *X*)
- Often denoted by σ_X

- Note, other measures of 'spread' but more annoying analytically
  - E.g. Mean Absolute Deviation, Interquartile Range

Joint distribution

- Consider again drawing cards from a deck
- RV *X* represents the *suit* of the card (hearts, diamonds etc.)
- RV *Y* represents the *rank* of the card (Ace, 2, 3 etc.)
- Then we can write down a pmf for both RVs together:

$$P(X = x, Y = y) \quad = \quad P(\{s \in \mathcal{S} : X(s) = x \wedge Y(s) = y\})$$

this is the **joint distribution** of X and Y.

- **Q: The joint is a function. What are its domain and range?**

		Y											
		2	3	4	5	6	7	8	9	10	J	K	Q	A
X	♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	K♠	Q♠	A♠
	♥	2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	K♥	Q♥	A♥
	♦	2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	K♦	Q♦	A♦
	♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	K♣	Q♣	A♣

Multi-dimensional joint

- More generally, for *p* RVs $X_1 \ldots X_p$, the joint distribution is written:

$$P(X_1, X_2 \ldots X_p)$$

- As always, we can write any pmf as an table of probabilities.
- **Q: If each X can take on two possible values, how many columns does the joint table contain?**

Joint distributions are **BIG**

- This is big – joint distributions get unwieldy very quickly!
- But real-world problems are rich in settings with many RVs, where the "joint" information is important: e.g. genes, words
- Much of ML and comp stats is about addressing this problem by:
  - Making use of structure, i.e. how RVs are related, e.g. Biological networks
  - Seeking parsimonious models: e.g. markov models for words

Marginal distribution and sum rule

- Going back to cards:
  - The 13x4 table is simply the joint distribution *P(X,Y)*
  - It's a table of probabilities for all possible pairs *(suit, rank)*
- How can we use this table to get a pmf for suit alone?
- That is, given a 13x4 array for the joint, what must we do to get P (queen) ?

2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	K♠	Q♠	A♠
2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	K♥	Q♥	A♥
2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	K♦	Q♦	A♦
2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	K♣	Q♣	A♣

Marginal distribution and sum rule

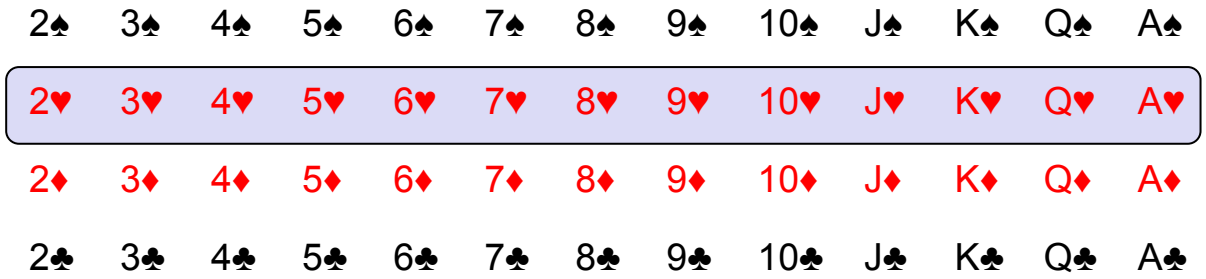
- This is the sum rule for RVs:

$$P(Y = y) \quad = \quad \sum_{x \in \mathcal{X}} P(X = x, Y = y)$$

- The sum is over values of *X*, gives us a function only of *y*
- Note that the sum has to be over the entire range of the RV *X*
- The word "marginal" is used because we're summing out over the RV we're not interested in to get to the **margin** of the table.
- (This sort of "summing out" is also called "marginalizing")

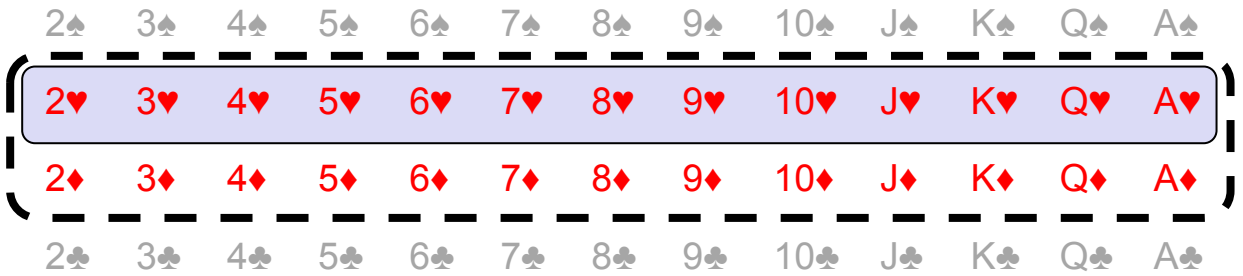
Conditional probability

- If I choose randomly, what's the probability *P(Y = heart)* ?



- Let's introduce a third RV *Z* for the *colour* of the suit
- If I tell you that I'm only going to choose from amongst the *red* cards, what is the probability of getting a heart?
- This is the probability of *Y* being heart, **conditioned on** Z being red
- Conditional probability is defined in the following way:

$$P(Y \mid Z) = \frac{P(Y, Z)}{P(Z)}$$



- Note that *P(Y | Z=z)* is itself a pmf for *Y*
- But, in general, *P(Y = y | Z)* is ***not*** a pmf for *Z !*
  - E.g. Consider P(Y = hearts | Z = red) + P(Y = hearts | Z = black)

Product rule

- Conditional probability (for *P(Y)>0*):

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

- For any *P(Y)*:

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

this is the **product rule** of probability

- These two, intuitive rules will crop up over and again, and, carefully applied, will enable us to proceed in quite complicated situations

"Conditional" sum and product rules

$$P(Y \mid Z) = \sum_{x \in \mathcal{X}} P(X, Y \mid Z)$$

$$P(X, Y \mid Z) = P(X \mid Y, Z) P(Y \mid Z)$$

- Intuitively, think of the variable Z as "background information", so it simply appears everywhere, after the "given".

Independence

- Two RVs X and Y are independent if and only if:

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

- **Q: If X, Y are independent, what is P(X | Y) ?**

Independence

- Two RVs X and Y are independent if and only if:

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

- **Q: If X, Y are independent, what is P(X | Y) ?**

- In other words, knowing Y doesn't give us any additional information about X

- **Q: Are successive coin tosses independent?**
- **Q: Suppose two football teams play three times, with results X_1, X_2, X_3. Are these RVs independent?**
- **Q: Let X be a RV representing #years education, Y that of person's partner. Are X, Y independent?**

Bayes theorem

- Joint *P(X,Y)* :

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

- We can equally well write:

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

- This gives:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

this is called **Bayes theorem**.

- **Q: show that (sum form of Bayes):**

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{\sum_{x \in \mathcal{X}} P(Y \mid X)P(X)}$$

- Bayes is an immensely useful expression, because it allows us to "turn conditionals around", getting X|Y in terms of Y|X

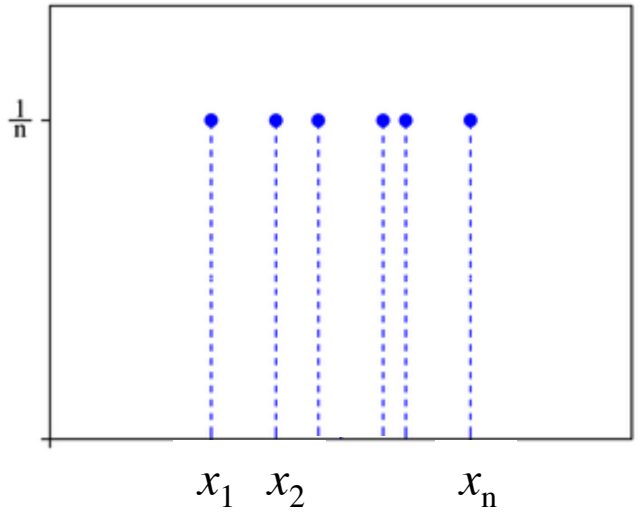
Bayes rule: example

- Q: You take a test for a disease which is 99% reliable in the following sense: if a person has the disease, there is a probability of 0.99 the test will be positive; if a person does *not* have the disease, there is a probability of only 0.01 that it comes back positive. The disease is known to affect 1 in 100,000 people. Your test comes back positive. What is the probability that you actually have the disease?

- This is a classic application of Bayes rule
- Arises so often in criminal cases it's called the *prosecutor's fallacy*

- As we shall see, Bayes is incredibly general.

Uniform pmf



- Gives equal probability mass for each possible value of *X*:

$$P(X = x) \quad = \quad \frac{1}{|\mathcal{X}|}$$

- Defined by sample space

Parameterized distributions

- In many cases, we work with pmfs which rather than being just any old table of numbers are members of families parameterized by a tunable parameter:

$$P(X = x \mid \theta) \quad = \quad f(x; \theta)$$

- Once we know the parameter, we have a fully specified pmf
- A pmf of this kind is a first example of a **statistical model**: it's a convenient functional form which we aim to use to describe some real-world observations, and thence make predictions about as yet unobserved events
- Some common pmfs...

Bernoulli distribution

- X has two possible outcomes, one is "success" (X=1) other "failure" (X=0):

- PMF:

$$P(X = x \mid \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

$$X \in \{0, 1\}$$

$$\theta \in [0, 1]$$

- Only one adjustable parameter ("success parameter")

Independent and identically distributed (i.i.d.)

$$P(X = x \mid \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

- When we used the **Bernoulli pmf** to describe coin tosses, made two implicit assumptions:

(i) that each toss is **independent** of the others

(ii) that the success parameter is the same for each toss, such that the pmfs are **identical**

- A set of RVs having these two properties are said to be "**independent and identically distributed**" or **i.i.d.**
- This is a very common assumption in inference

Models

- This approach – of choosing a sensible sounding probability function for a system of interest - is a key step in practical applications

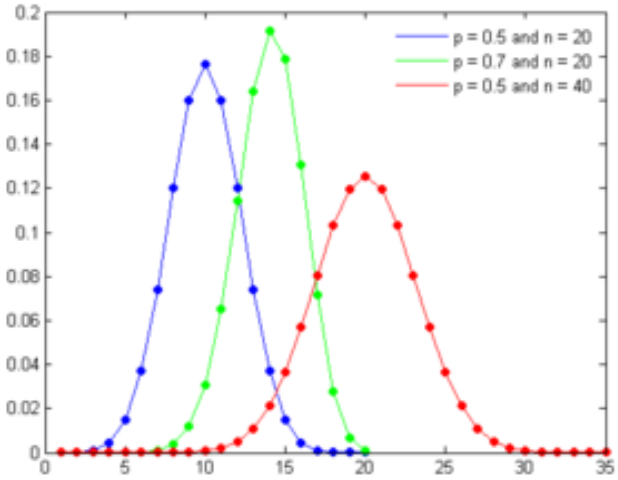
- Box famously said **"all models are wrong, some are useful"**
- That is, there's wrong and egregiously wrong
- E.g. an i.i.d. Bernoulli model would be quite wrong for the tennis match example – what sort of dependence might make more sense?

- In practice, we very often have to use models which are mathematically convenient. What we must do is check that they're actually "useful".

Binomial distribution

- Let X_1, X_2 ... X_n be i.i.d. Bernoulli RVs.
- What is the pmf of $X = X_1 + X_2 + ... + X_n$?

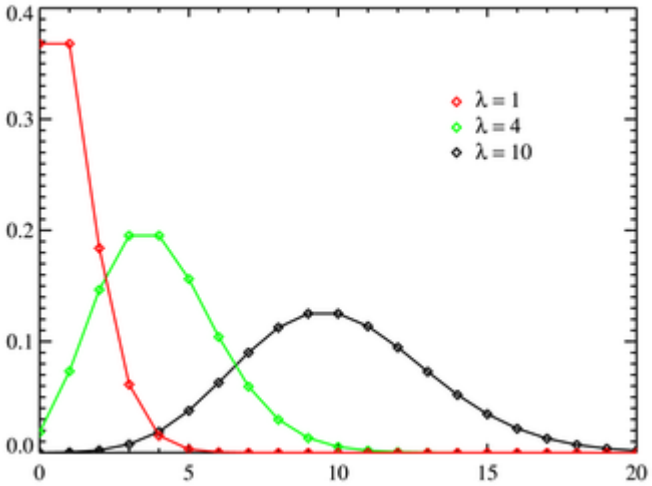
$$
\begin{aligned}
P(X = x|\theta, n) &= \binom{n}{x}\theta^x(1-\theta)^{(n-x)} \\
X &\in \{0, 1 \ldots n\} \\
\theta &\in [0, 1]
\end{aligned}
$$



- This is the **Binomial pmf**
- It's the distribution over the **number of successes** in *n* Bernoulli trials

Poisson pmf

$$
\begin{aligned}
P(X = x \mid \lambda) &= \frac{1}{x!} e^{-\lambda} \lambda^x \\
X &\in \{0, 1, 2 \ldots\} \\
\lambda &\in \mathbb{R}_+
\end{aligned}
$$



- Arises in modelling number of events, when the probability of the event occurring is constant in time
- For example, calls arriving at a telephone exchange: if the rate is *r* calls/minute, the distribution over the number of calls arriving in *T* minutes can be modelled as *Poisson(rT)*