Essential Probability and Statistics for Big Data

Jeff Chuang (with thanks to Joel Graber)

jeff.chuang@jax.org

Learning Objectives

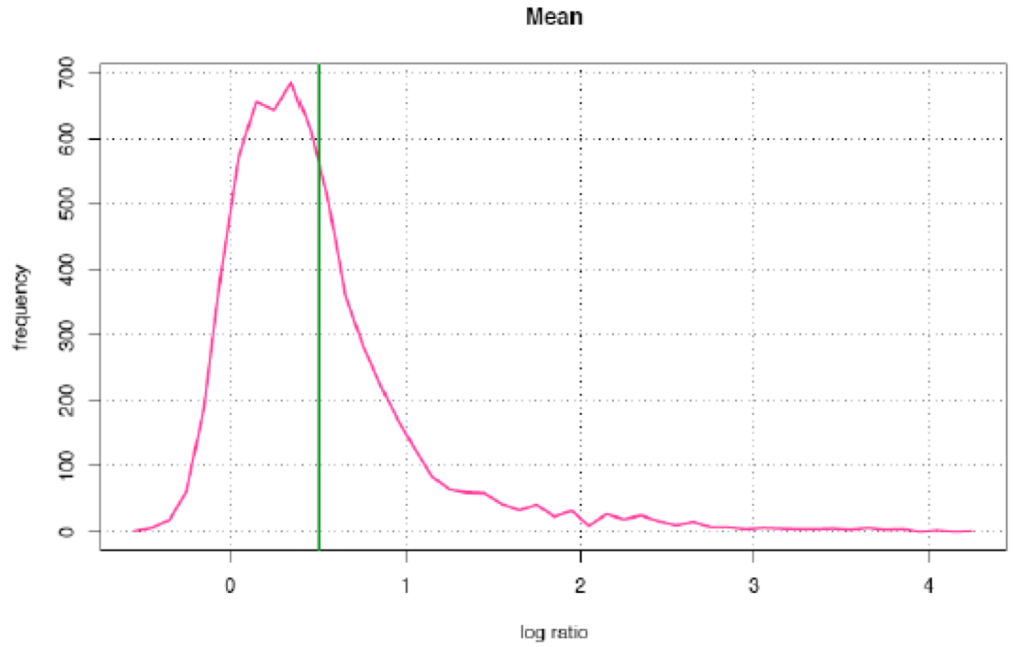
- Descriptive Statistics
- Distributions and p-values
- Bayesian Inference
- Multiple Hypothesis Testing
- Introduction to Data Mining

Descriptive Statistics

- With acknowledgement to Jacques van Helden for many of the images

Location parameters - Arithmetic mean

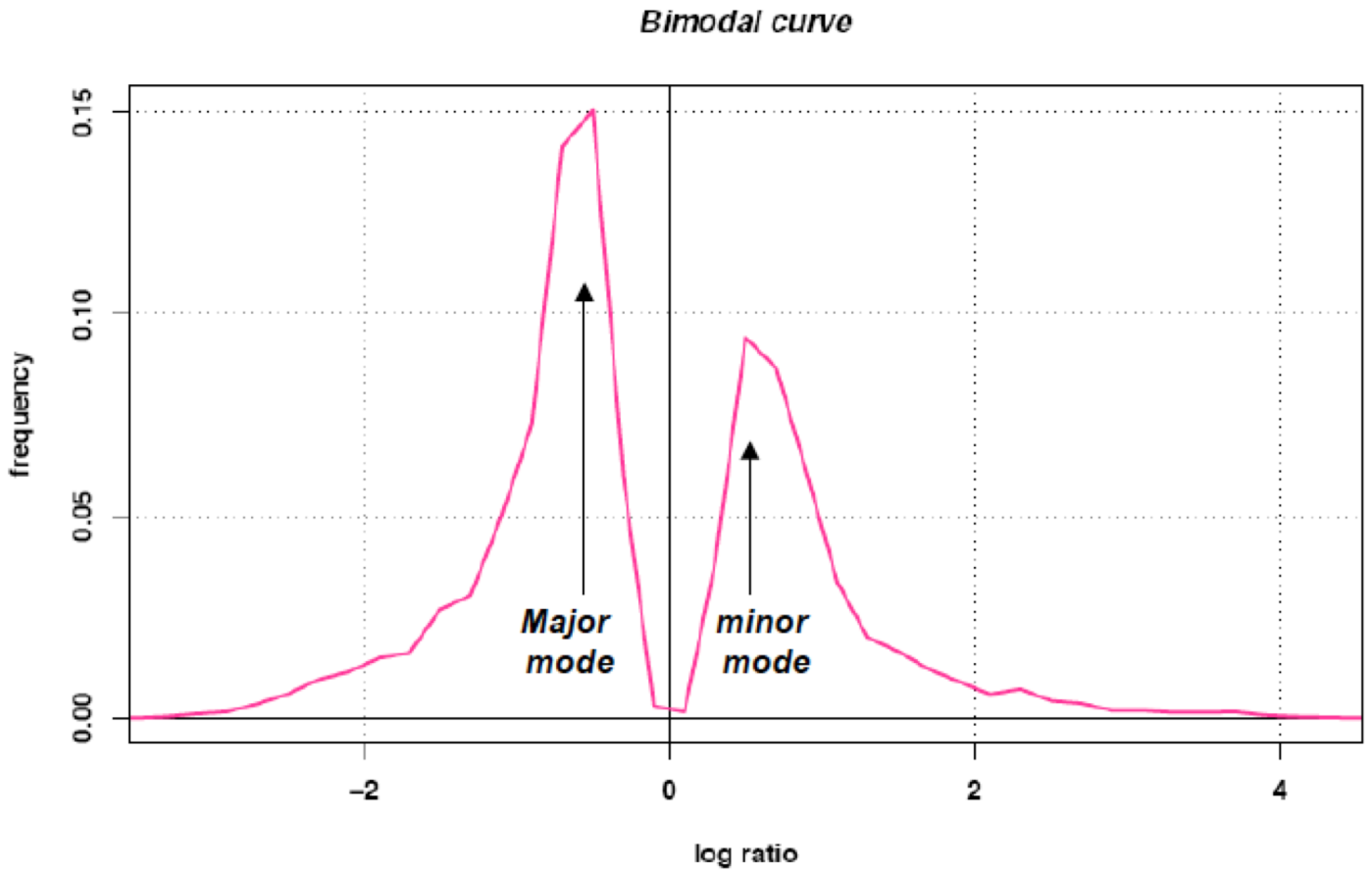
$$\bar{x} = a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- The mean is the gravity center of the distribution
- Beware: the mean is strongly influenced by outliers.
- Statistical "outliers" are generally biologically relevant objects (e.g. regulated genes).

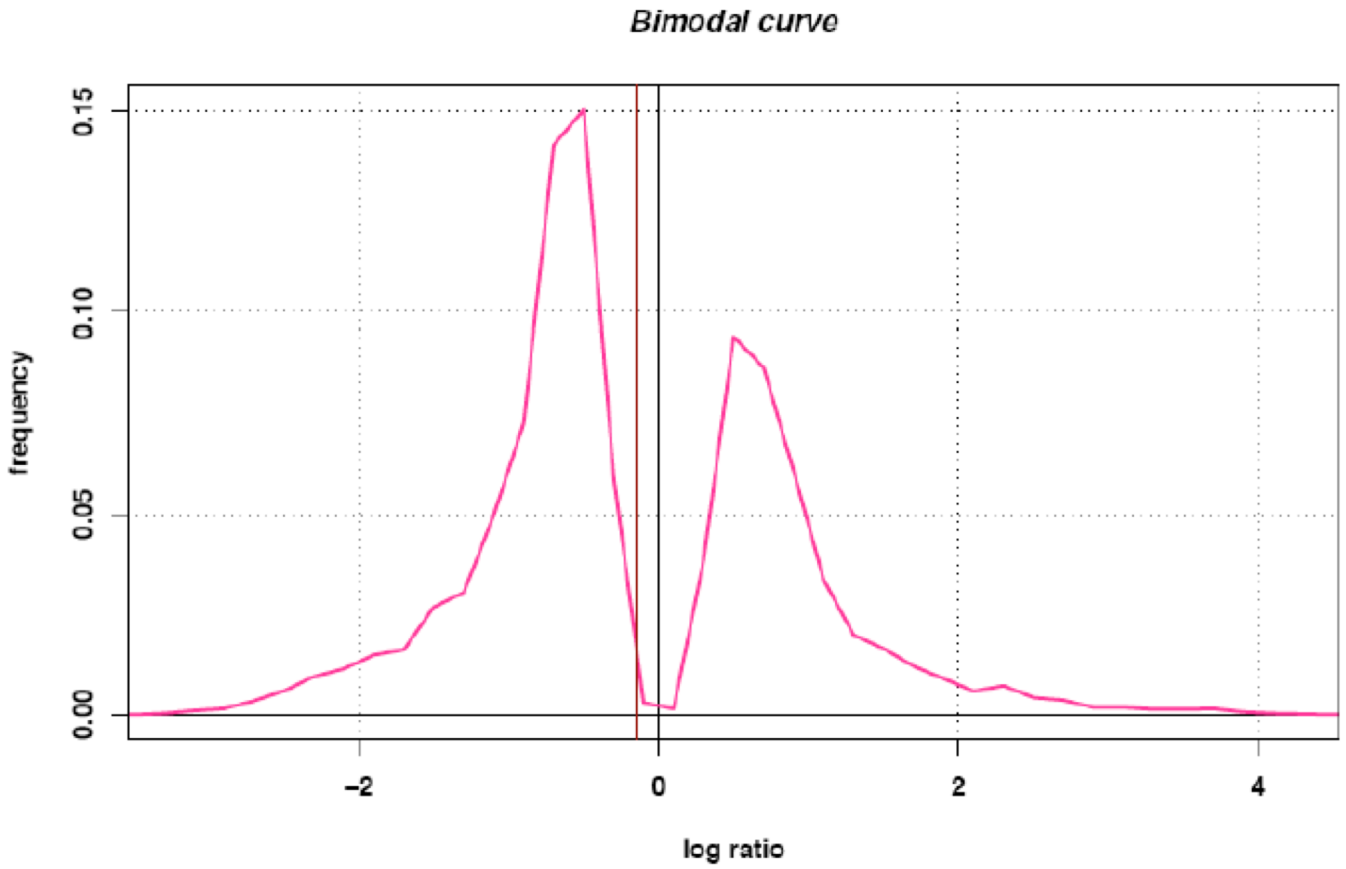
Multimodal curves

- E.g.: Extreme values in the gene expression data



Mean and bimodal curves

- For bimodal curves, the mean and the median poorly reflect the tendency of the population (almost no point has the mean value)



Dispersion parameters - Range

- $Range = max - min$
- The range only reflects 2 values: the min and max
- Strongly affected by outliers → poor representation of the general characteristics of the sample

Dispersion parameters - Variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2$$

- The variance is strongly affected by exceptional values

Dispersion parameters - interquartile range (IQR)

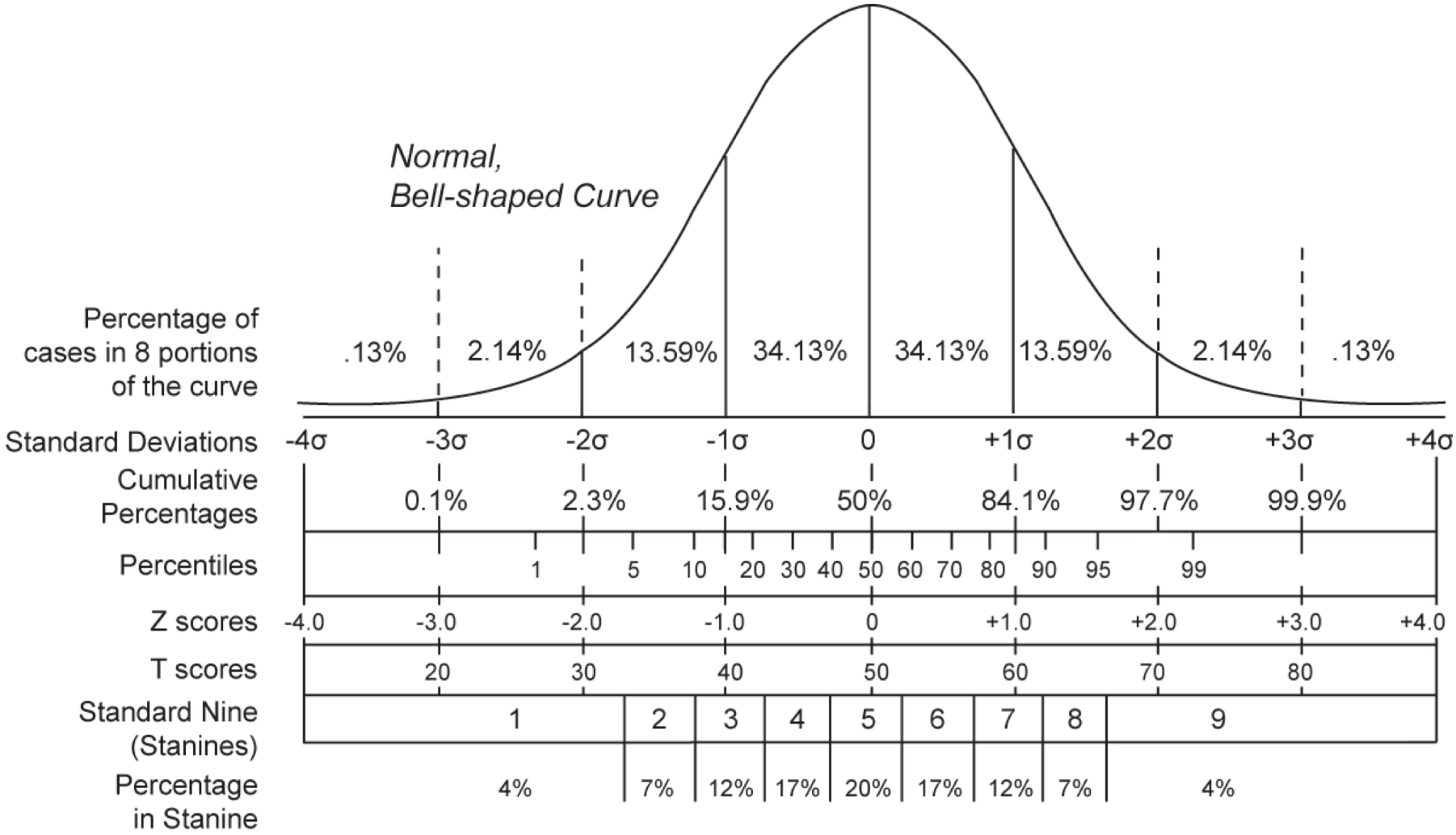
- The quartiles are an extension of the median
  - The first quartile ($Q1$) leaves 1/4 of the observations on its left.
  - The second quartile is the median.
  - The third quartile ($Q3$) leaves 3/4 of the observations on its left.
- The inter-quartile range (IQR=$Q3$-$Q1$) indicates the spread of the 50% central values.
- The inter-quartile range is robust to outliers, since it is is based on the ranks rather than the values themselves.

Normally distributed variates and Z-scores

- The Central Limit Theorem states that any sum of random variables, *regardless of their underlying distribution*, will tend towards a normal distribution
- A normal distribution can be described by two parameters, the mean (μ) and standard deviation (σ)
- Normally distributed scores are converted to *Z*-scores via:

$$Z = \frac{x - \mu}{\sigma}$$

Statistical significance: Z-scores and cumulative standard normal distributions



Outliers

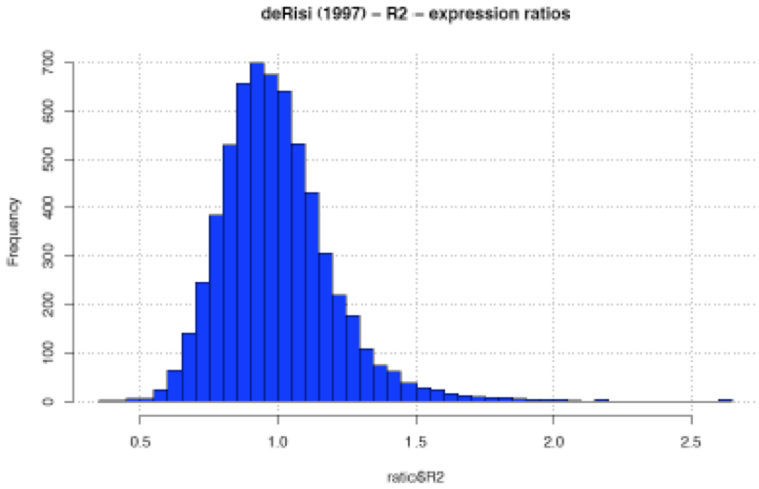
- An **outlier** is an observation point that is distant from other observations
- An outlier is *defined to be* the value in the sample that differs from the nearest quartile by more than $1.5IQR$
- Susceptible to outliers: mean, variance, standard deviation, range
- Not susceptible to outliers: quartiles, median, interquartile range

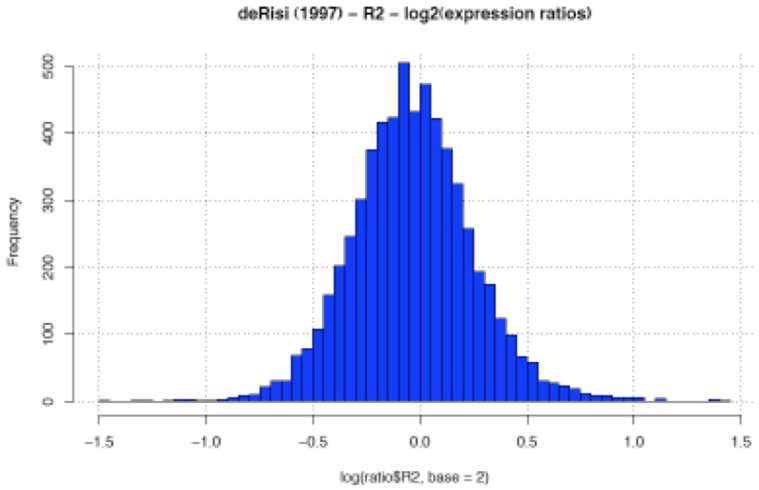
Things to remember

- Summary variables may not accurately reflect the distribution
- Biological systems are frequently noisy-- be wary of outliers and their impact on summary statistics
- Mean and variance are interpretable only if your data can be reasonably fit to a normal distribution

Distributions

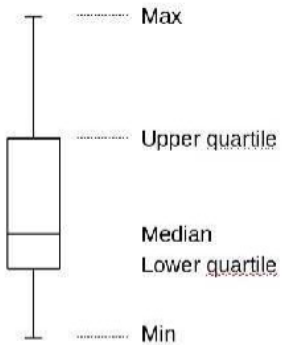
Histogram





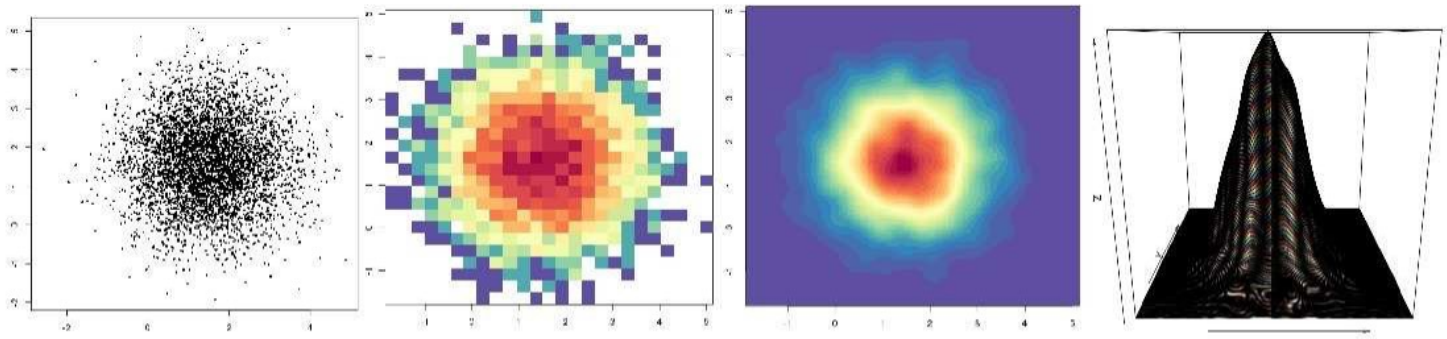
- The area above a given range is proportional to the frequency of this range
- Appropriate for absolute or relative frequencies
- Appropriate for representing class frequencies

Boxplot

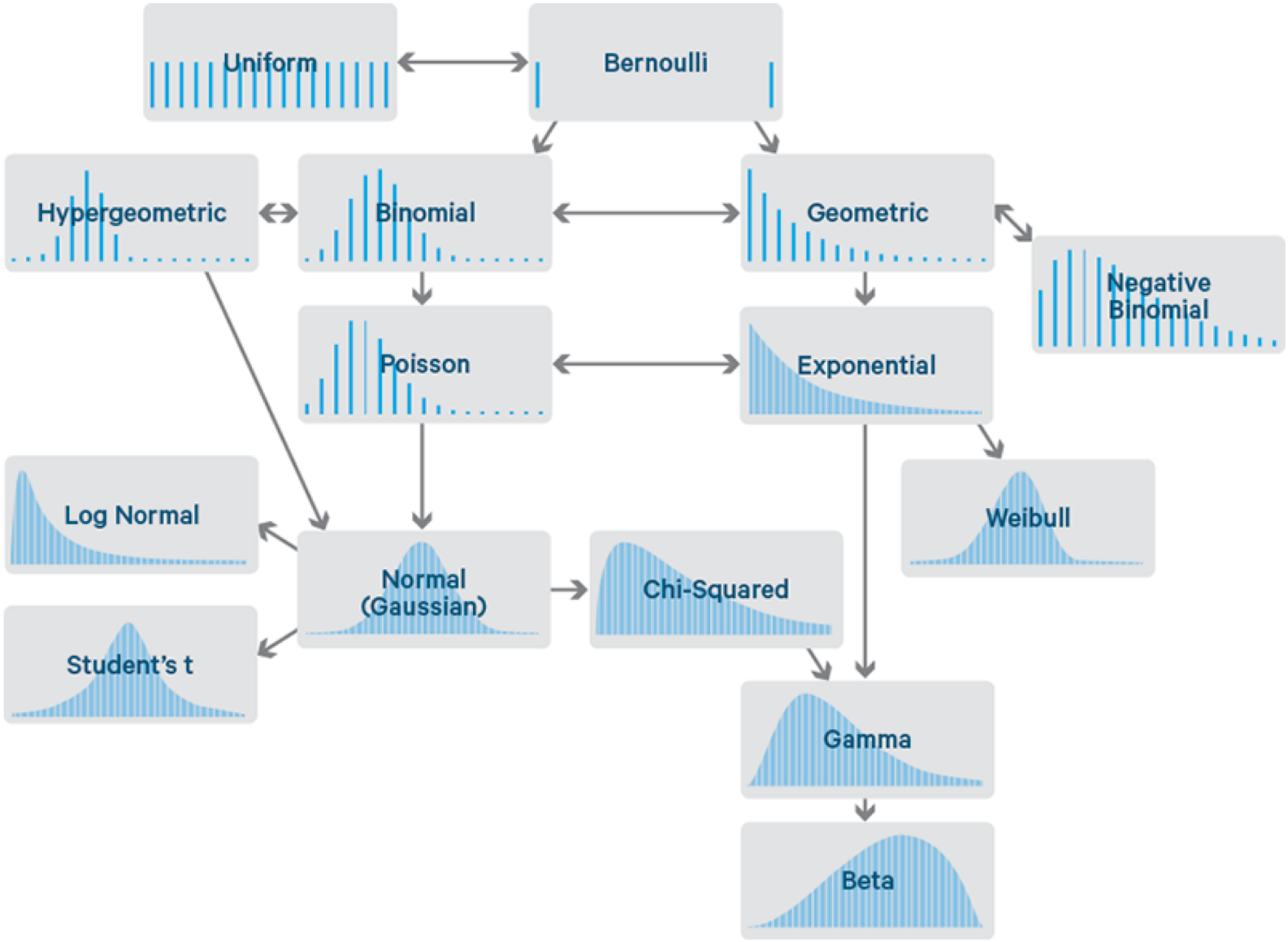


- A boxplot is a way of graphically depicting groups of numerical data through their quartiles
- Box plots also have lines extending the boxes (whiskers) indicating variability outside the upper and lower quartiles
- Outliers are plotted as individual points

2D-density



- A **heatmap** is a graphical representation of data where the values contained in a matrix are represented as colors
- Often the values to represent are frequencies

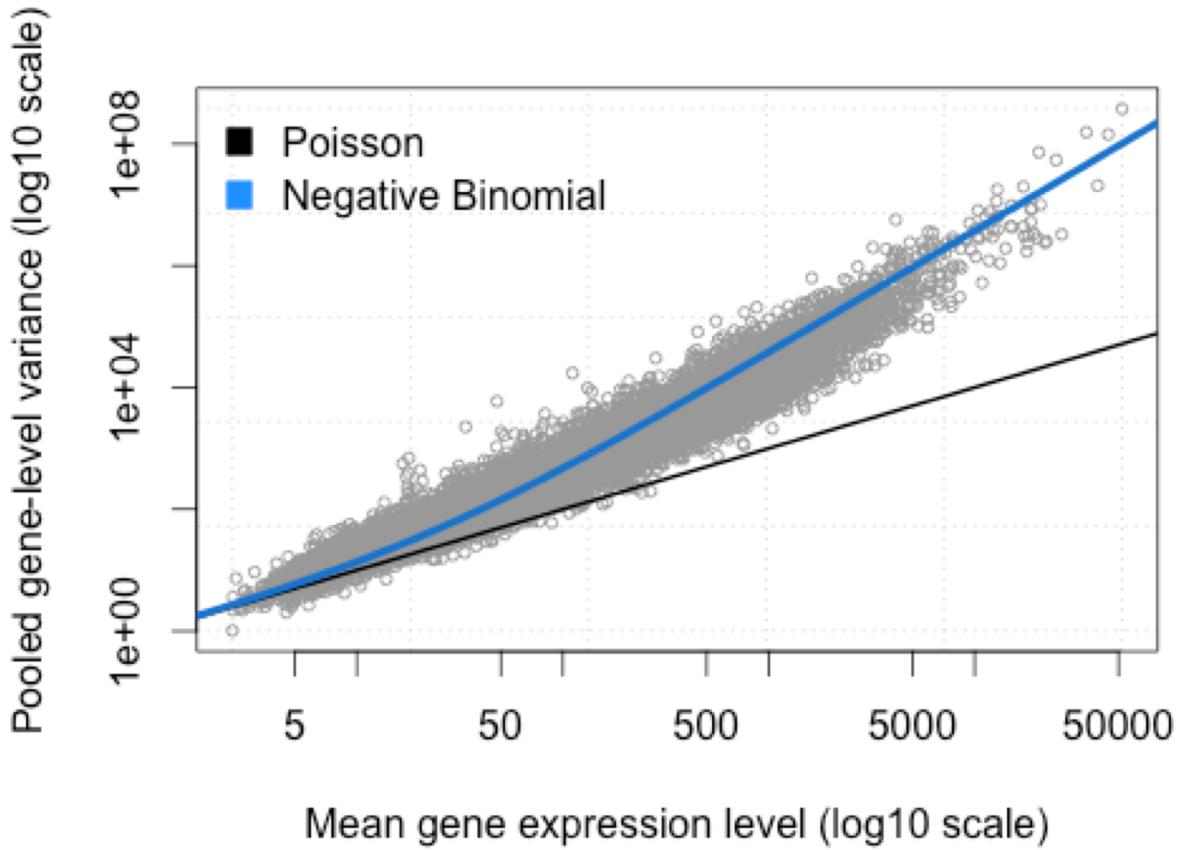


http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

Why do these distributions matter?

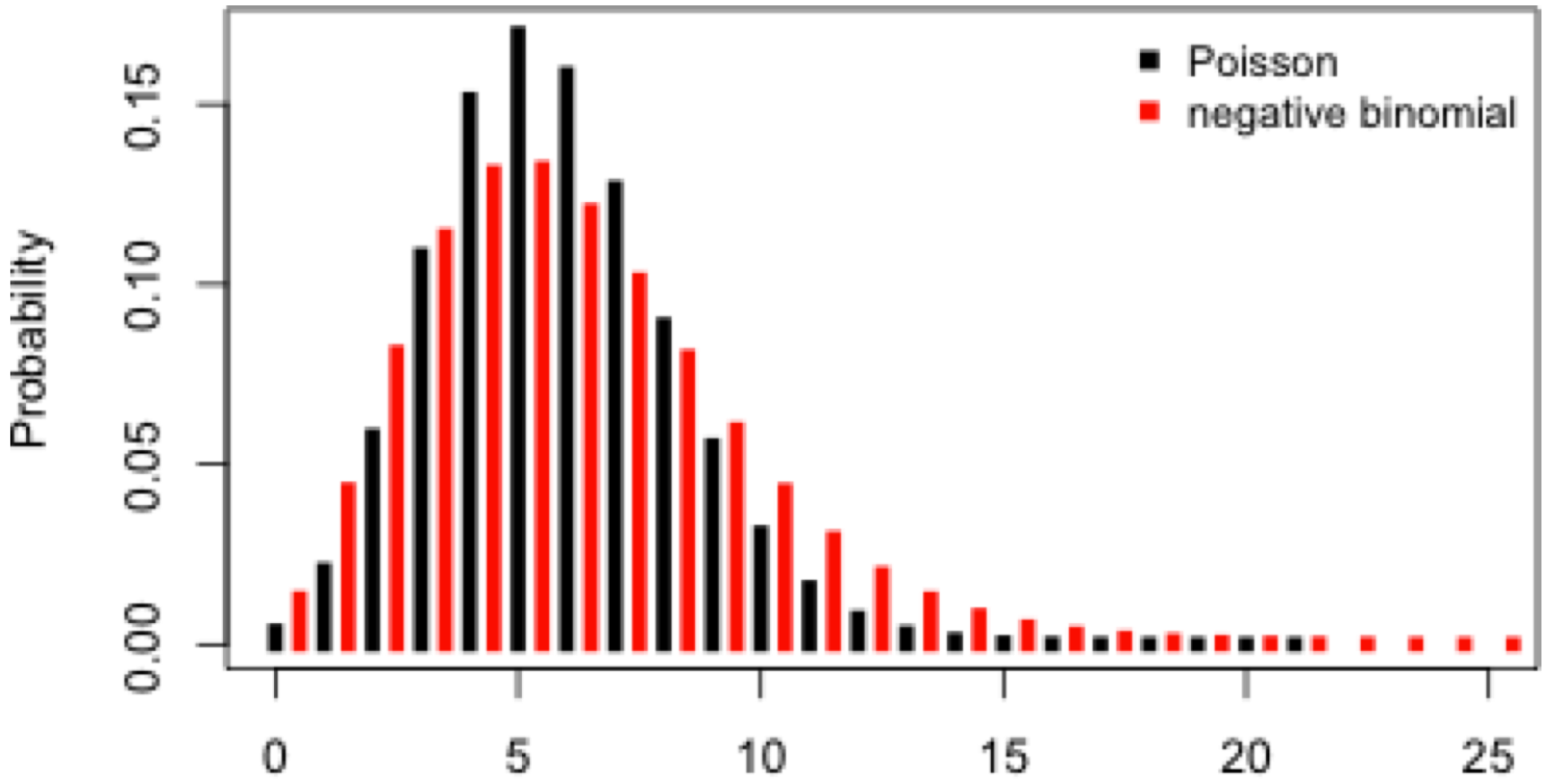
- We are frequently testing questions like:
  - Are these genes differentially expressed?
- Alternatively stated:
  - What is the likelihood of observing these counts under the assumption that they came from the same expression distribution?

Gene expression fits the negative binomial distribution



https://bioramble.wordpress.com/2016/01/30/why-sequencing-data-is-modeled-as-negative-binomial/

Poisson and Negative Binomial can result in very different “p-values”

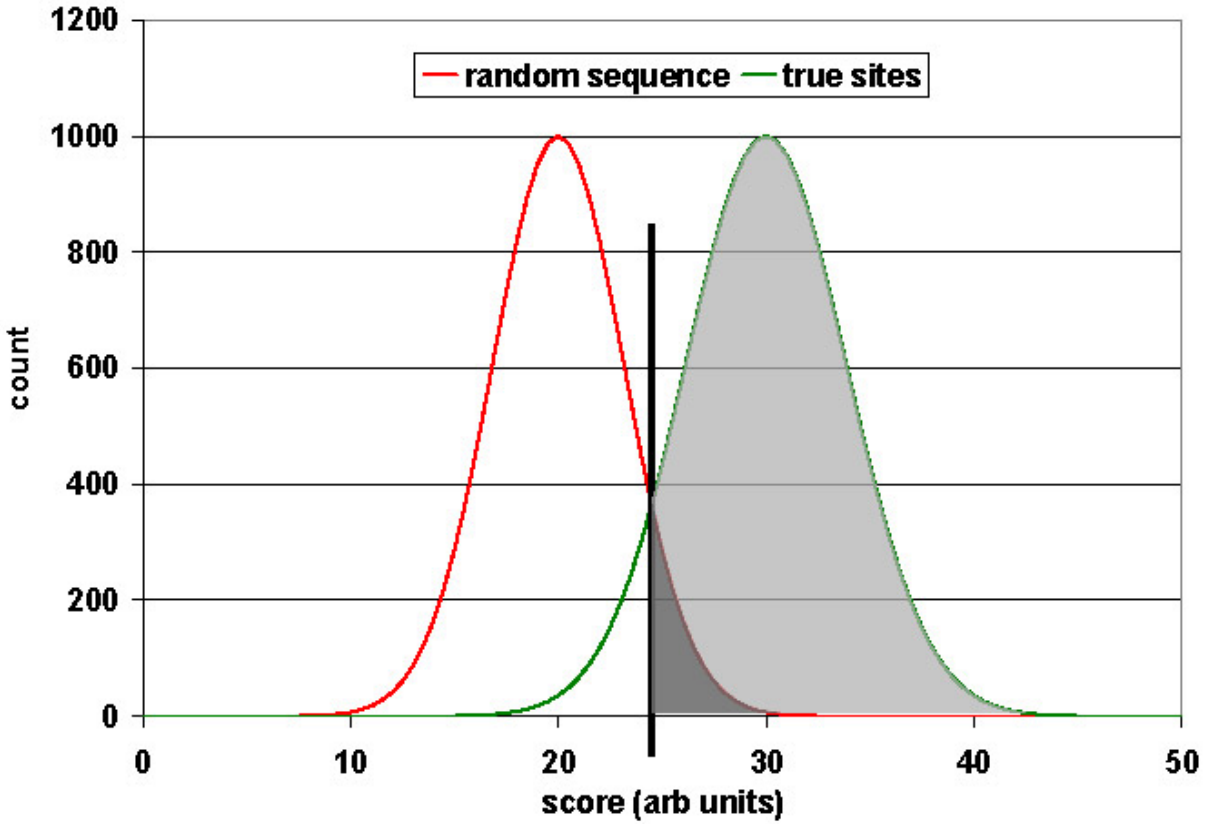


Model Evaluation

Assessing performance: Sensitivity and Specificity

- Testing of predictions is performed on sequences where the answer is known
- ***Sensitivity*** is the fraction of known elements correctly predicted
  - "Am I finding the things that I'm supposed to find?"
- ***Specificity*** is the fraction of predict elements that correspond to true elements
  - "What fraction of my predictions are true?"
- In general, increasing one decreases the other

Graphic View of Specificity and Sensitivity

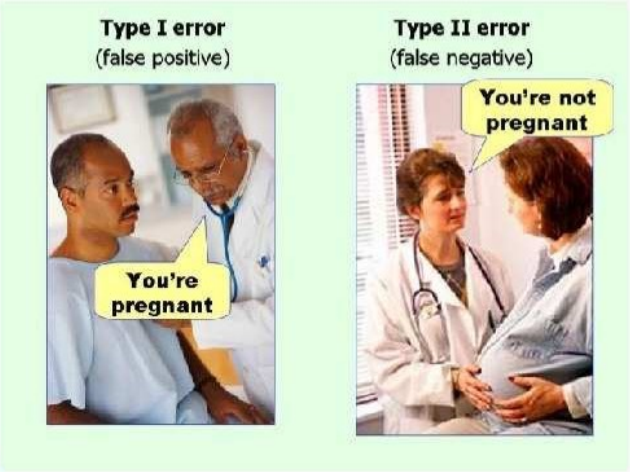


$$Sn = \frac{TruePositive}{AllTrue} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Sp = \frac{TruePositive}{AllPositive} = \frac{TruePositive}{TruePositive + FalsePositive}$$

False Positives and False Negatives

Actual condition	Test shows	
	"not pregnant"	"pregnant"
H_0 : Not pregnant	True Negative	False Positive **Type I error**
H_a : Pregnant	False Negative **Type II error**	True Positive

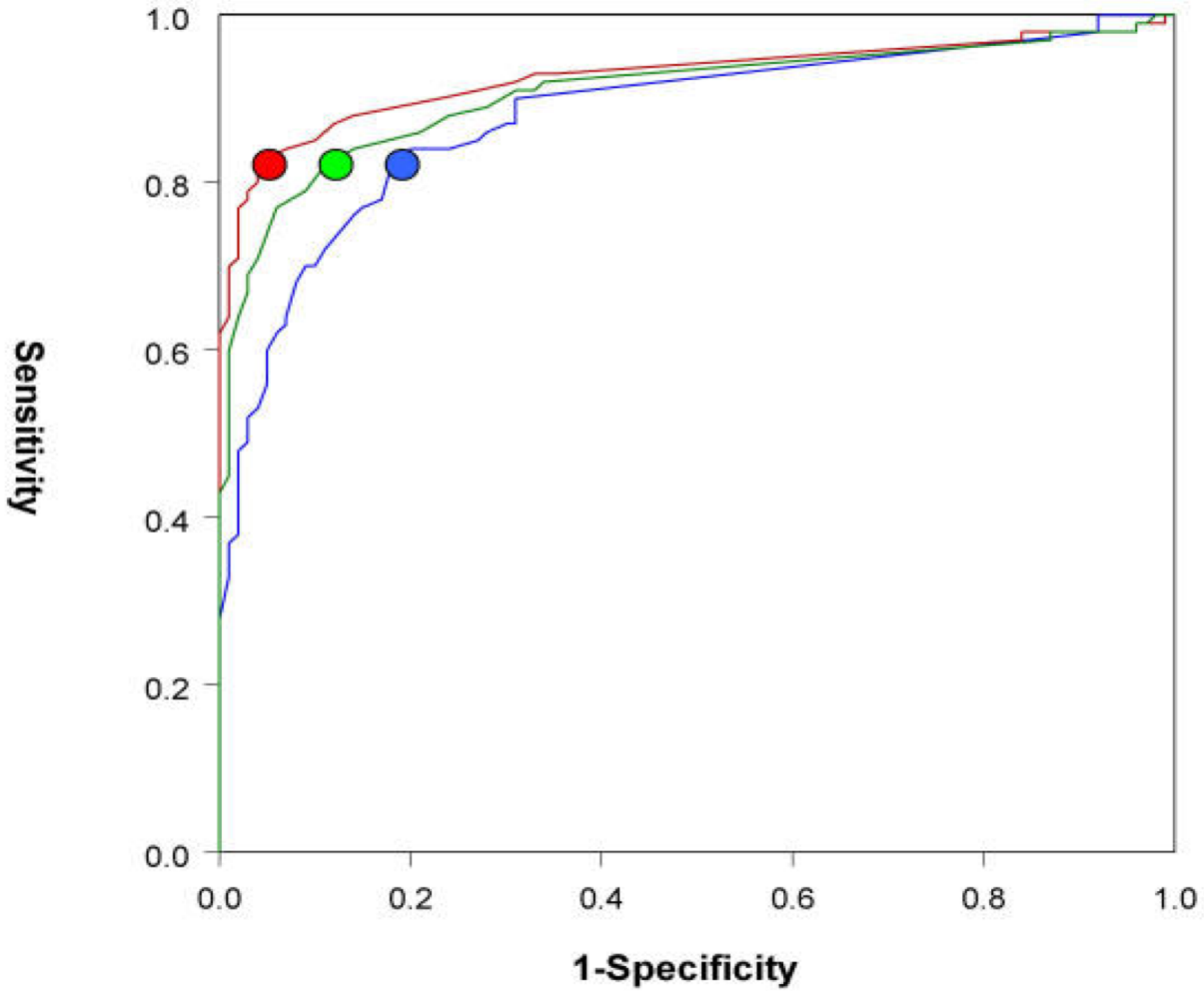


Confusion matrix

Actual condition	Test shows "not pregnant"	Test shows "pregnant"	Σ
H_0 : Not pregnant	TN	FP	TN + FP
H_a : Pregnant	FN	TP	FN + TP

Actual condition	Test shows "not pregnant"	Test shows "pregnant"	Σ
H_0 : Not pregnant	$\frac{TN}{TN+FP} =$ True Negative Rate = $1-\alpha =$ Specificity	$\frac{FP}{TN+FP} = \alpha =$ False Positive Rate	100%
H_a : Pregnant	$\frac{FN}{FN+TP} = \beta =$ False Negative Rate	$\frac{TP}{FN+TP} = 1-\beta =$ True Positive Rate = Sensitivity	100%

Quantifying the tradeoff: Receiver-Operator Curve analysis



Precision-recall curves provide similar evaluations while also adjusting for data imbalances

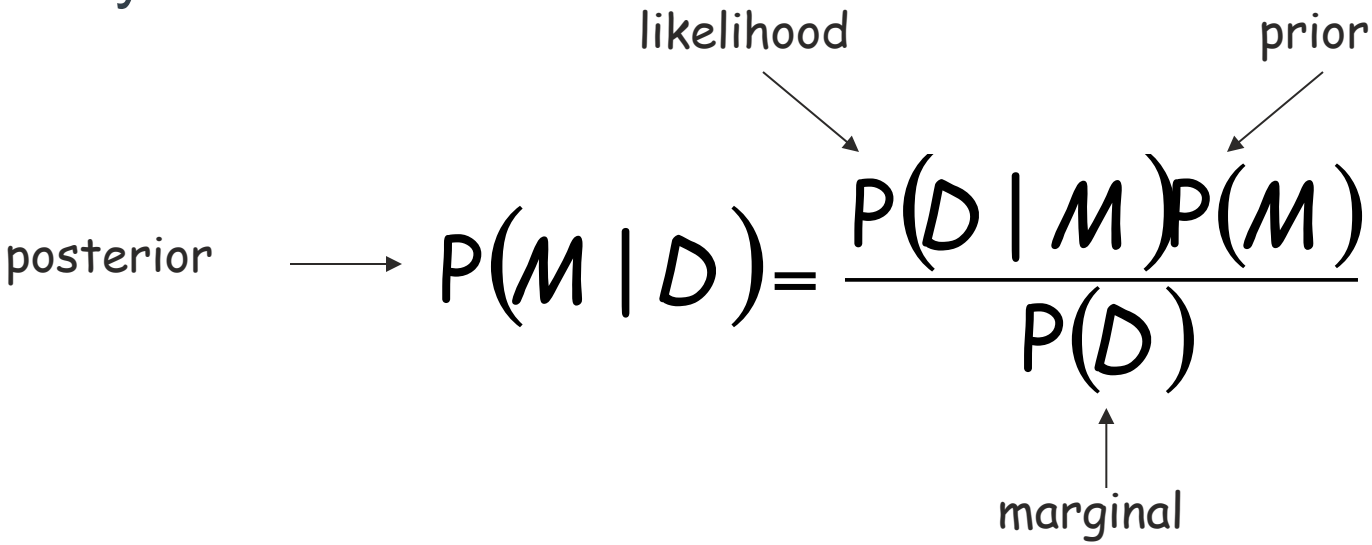
Basic Bayesian Statistics

Inverting the probability problem: statistical inference

- Standard analysis:
  - Given a model, compute the probability of observing one or more “events”
- Statistical inference:
  - Given two or more competing models, identify the model most probable to have generated the observed data

Bayesian Statistics: Statistical Inference

- Bayes' Rule



- *M*: the model, *D*: data or evidence

Basic Bayesian Statistics

- Bayes' Rule is at the heart of much predictive software
- Calculating marginal probabilities is complex and often computationally prohibitive
- In the simplest example, we can just compare a regulatory model vs. random and reduce it to a log-odds ratio

$$\log\frac{P(\text{model}|data)}{P(\text{random}|data)} = \log\frac{P(data|\text{model})}{P(data|\text{random})} + \log\frac{P(\text{model})}{P(\text{random})}$$

Multiple Hypothesis testing

Multiple Hypothesis testing

- Modern comp bio frequently involves simultaneous assessment of many hypotheses:
  - Are any regions of the genome amplified? (10^6-10^9)
  - Do any transcripts have unusual expression? (10^4-10^5)
  - Do any motifs recur? (4^4 - 4^{10})
- Low probability values appear simply because many things are tested
- We must correct for this in assessing significance

Types of multiple hypothesis correction

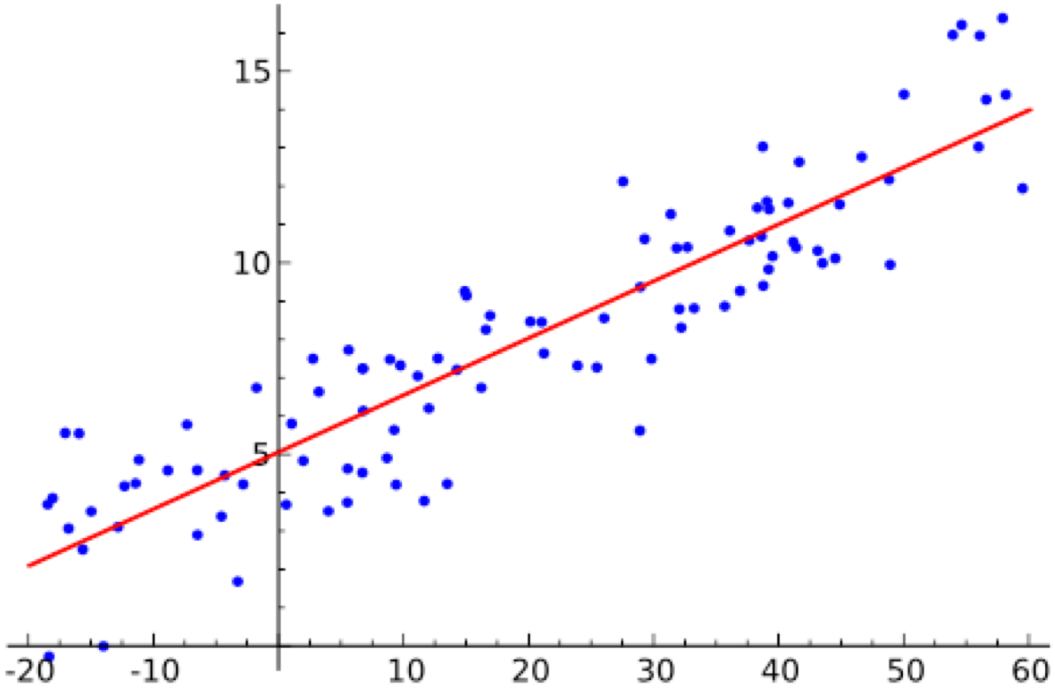
- Bonferroni correction
  - Multiply lowest probability obtained by the number of tests
  - Very conservative estimate
- Family-Wise Error Rate (FWER)
  - Determine the *best* probability obtained under a given model
  - Find all measurements better than this value in true set
  - Controls for *any* value <= p_{obs} under the null hypothesis
- False Discovery Rate (FDR, typically denoted *q*)
  - Essentially a scaled Bonferroni
    - Sort probabilities obtained
    - Multiply by N/rank(r)
    - Controls for the probability of obtaining the r^{th} value <= p_{obs} under the null

An example FDR correction

	sort →	Multiply each by N/rank →	Take minimum value for any greater rank →
0.340	0.005	0.104	0.104
0.039	0.014	0.138	0.134
0.005	0.020	0.134	0.134
0.912	0.027	0.137	0.137
0.405	0.039	0.157	0.157
0.425	0.172	0.573	0.573
0.241	0.241	0.688	0.650
0.487	0.329	0.822	0.650
0.027	0.340	0.757	0.650
0.429	0.405	0.811	0.650
0.699	0.425	0.772	0.650
0.020	0.429	0.716	0.650
0.172	0.441	0.679	0.650
0.483	0.483	0.691	0.650
0.755	0.487	0.650	0.650
0.521	0.521	0.651	0.651
0.441	0.699	0.823	0.823
0.014	0.755	0.839	0.838
0.796	0.796	0.838	0.838
0.329	0.912	0.912	0.912

Introductory Data Mining

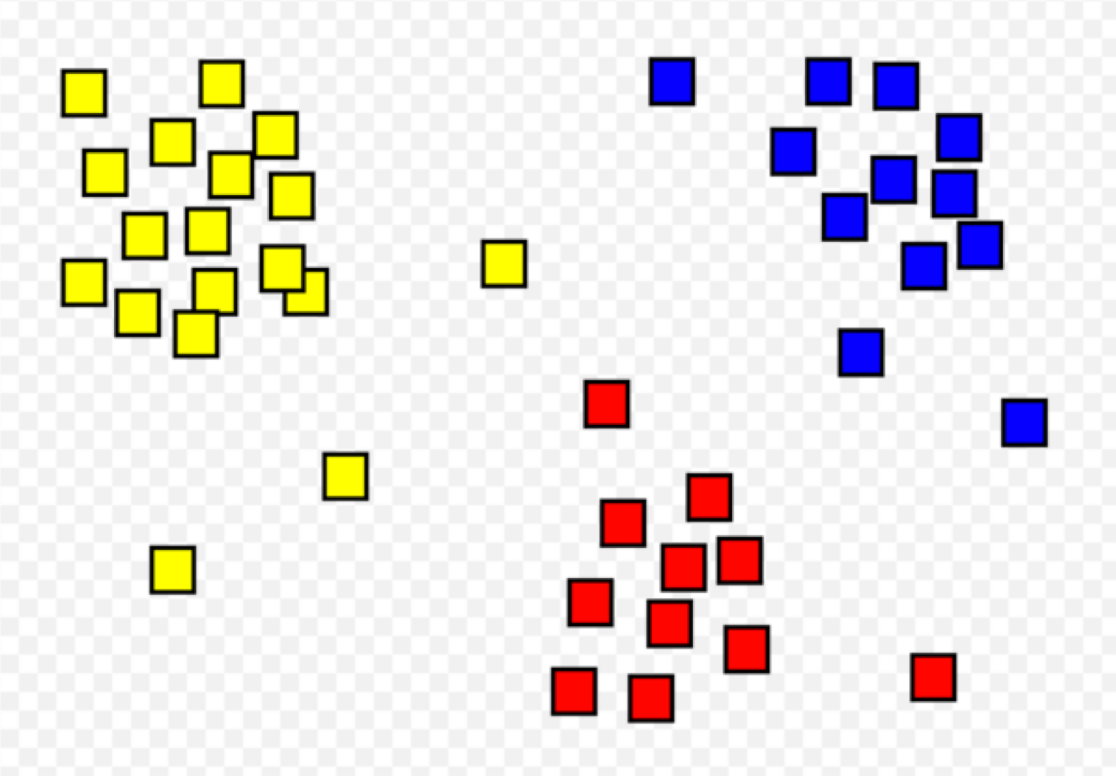
Regression



Modeling the relationship between a dependent variable y and one or more explanatory variables x

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Clustering



Clustering: grouping sets of objects by similarity.

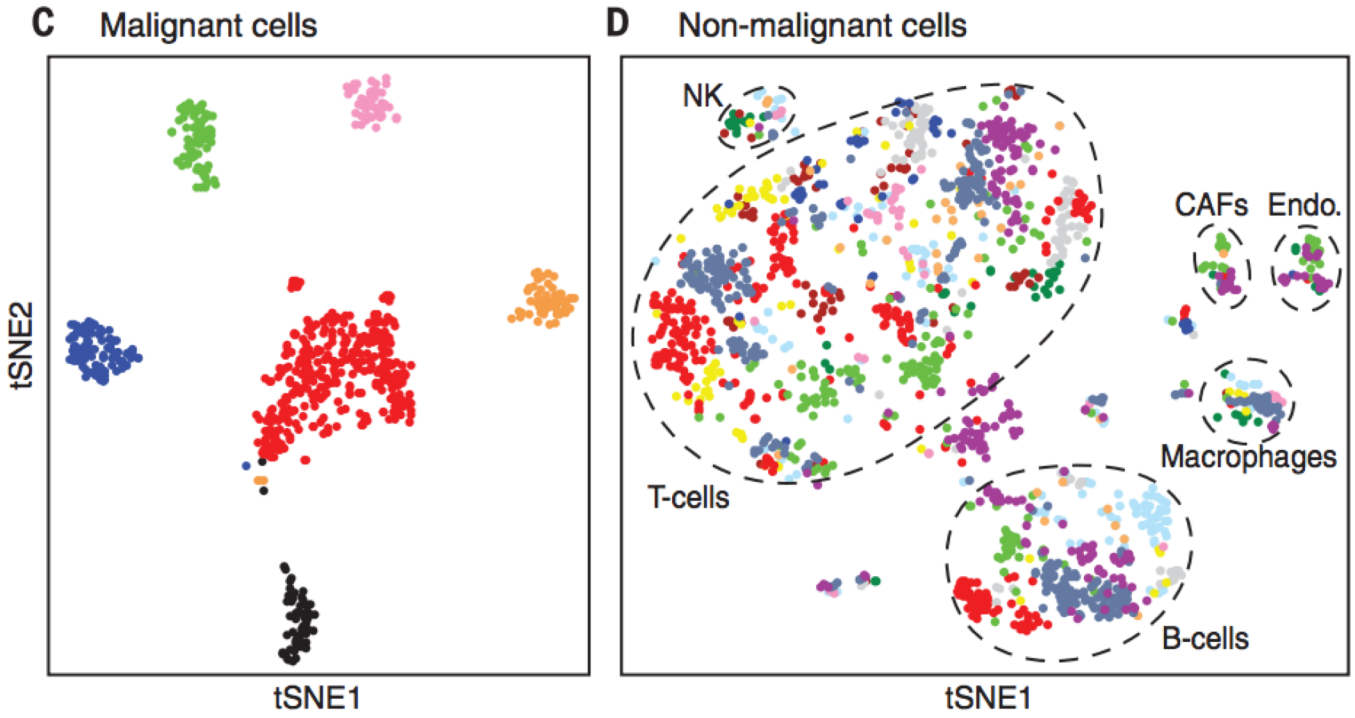
Common approaches include jierarchical clustering (e.g. UPGMA), K-means, and density-based clustering

CANCER GENOMICS

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

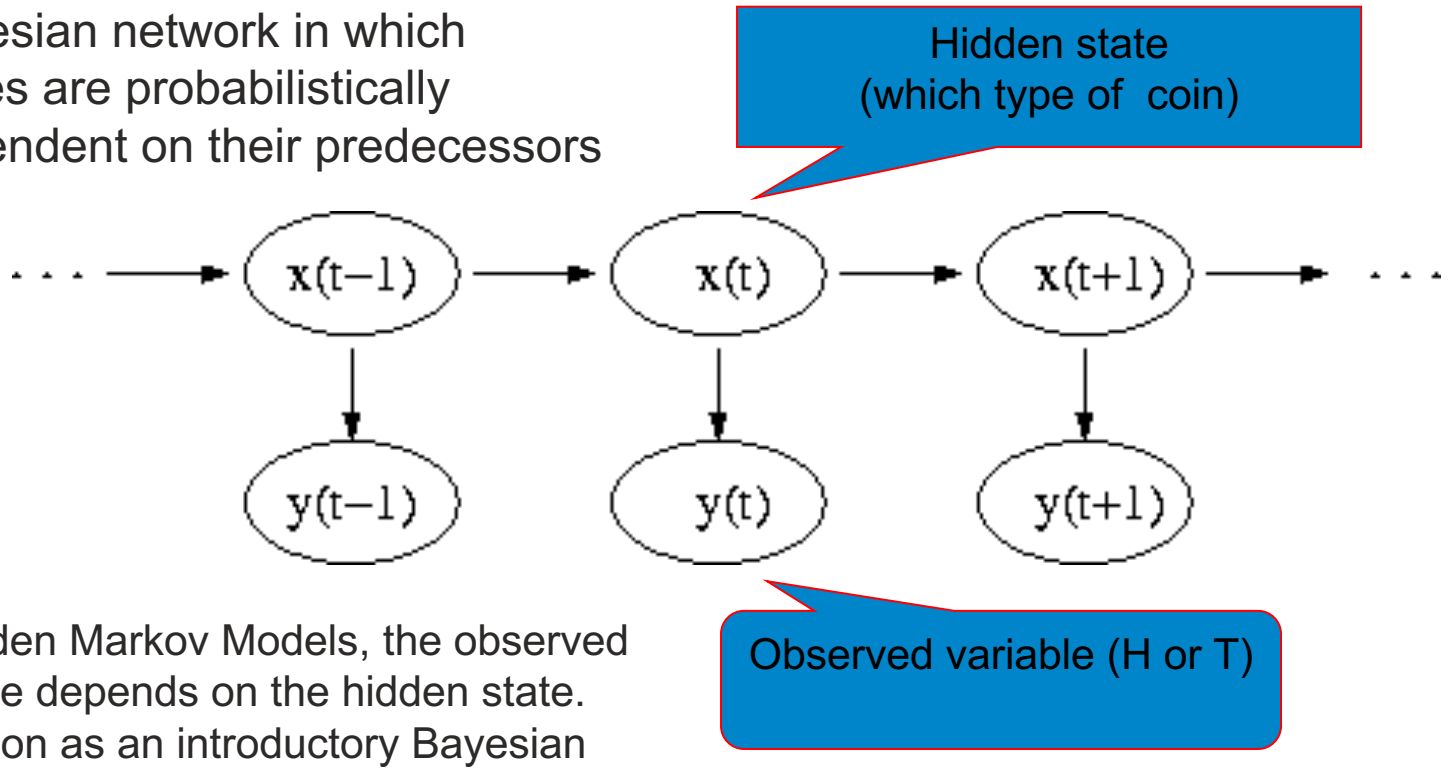
Itay Tirosh,[1*] Benjamin Izar,[1,2,3*†‡] Sanjay M. Prakadan,[1,4,5,6] Marc H. Wadsworth II,[1,4,5,6] Daniel Treacy,[1] John J. Trombetta,[1] Asaf Rotem,[1,2,3] Christopher Rodman,[1] Christine Lian,[7] George Murphy,[7] Mohammad Fallahi-Sichani,[8] Ken Dutton-Regester,[1,2,9] Jia-Ren Lin,[10] Ofir Cohen,[1] Parin Shah,[2] Diana Lu,[1] Alex S. Genshaft,[1,4,5,6] Travis K. Hughes,[1,4,6,11] Carly G. K. Ziegler,[1,4,6,11] Samuel W. Kazer,[1,4,5,6] Aleth Gaillard,[1,4,5,6] Kellie E. Kolb,[1,4,5,6] Alexandra-Chloé Villani,[1] Cory M. Johannessen,[1] Aleksandr Y. Andreev,[1] Eliezer M. Van Allen,[1,2,3] Monica Bertagnolli,[12,13] Peter K. Sorger,[8,10,14] Ryan J. Sullivan,[15] Keith T. Flaherty,[15] Dennie T. Frederick,[15] Judit Jané-Valbuena,[1] Charles H. Yoon,[12,13†] Orit Rozenblatt-Rosen,[1†] Alex K. Shalek,[1,4,5,6,11,16†] Aviv Regev,[1,17,18†‡] Levi A. Garraway[1,2,3,14†‡]

Example: t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.



Advanced: Hidden Markov Models

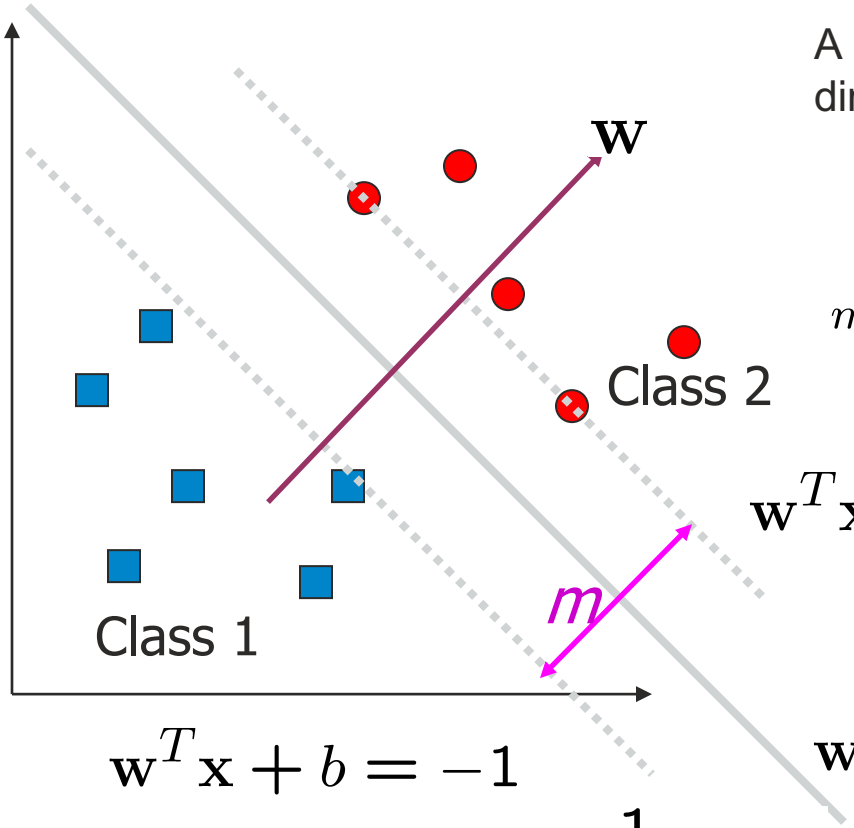
Markov models are a simple Bayesian network in which states are probabilistically dependent on their predecessors



In Hidden Markov Models, the observed variable depends on the hidden state. Common as an introductory Bayesian network in sequence analysis.

Advanced: Support Vector Machines

A binary classifier suitable for high-dimensional data



$$m = \frac{2}{||\mathbf{w}||}$$

$$\mathbf{w}^T\mathbf{x} + b = 1$$

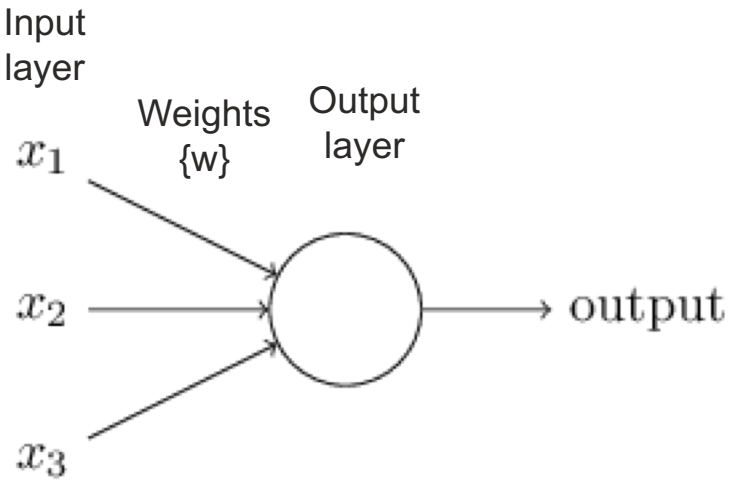
$$\mathbf{w}^T\mathbf{x} + b = -1$$

$$\mathbf{w}^T\mathbf{x} + b = 0$$

Minimize $\frac{1}{2}||\mathbf{w}||^2$

subject to $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \qquad \forall i$

Advanced: Neural Networks

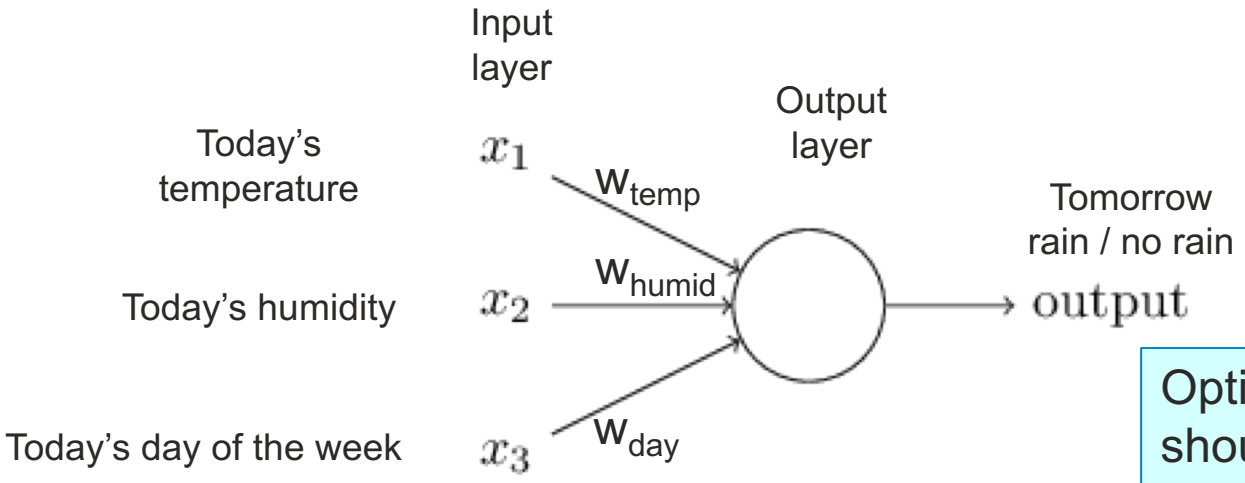


Training the neural network means finding the weights {w} which best fit the training data

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{threshold} \end{cases}$$

neuralnetworksanddeeplearning.com

Neural network example

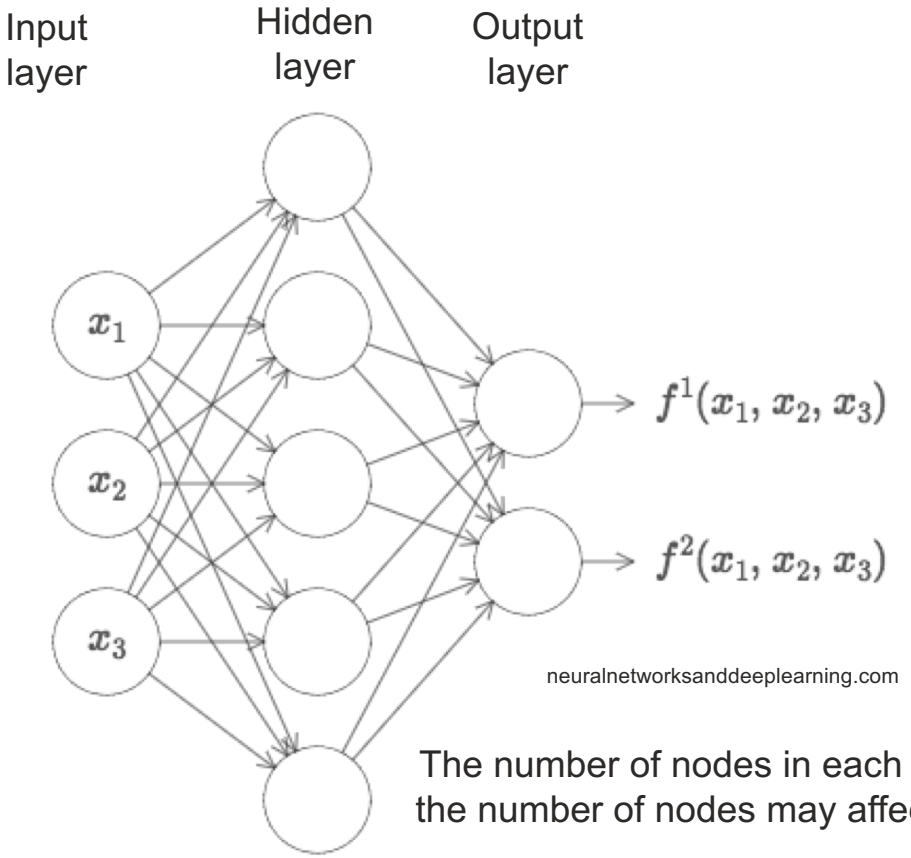


Optimal w_{temp} and w_{humid} should be non-zero.
Optimal w_{day} should be zero.

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{threshold} \end{cases}$$

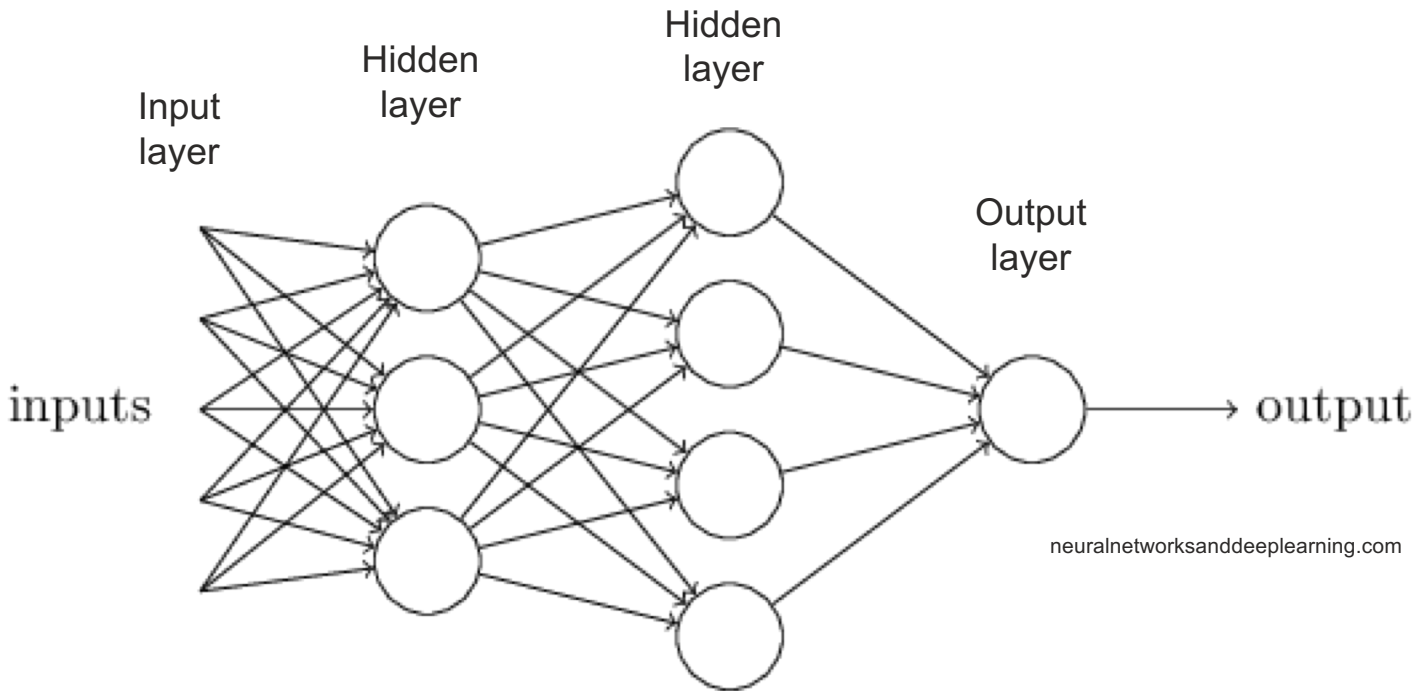
neuralnetworksanddeeplearning.com

Adding a hidden layer improves complexity of input-output fitting



The number of nodes in each layer is flexible. Choice of the number of nodes may affect accuracy of predictions.

Multilayer Neural Network



The number of hidden layers can be adjusted. ”Deep” networks may use 50-100 hidden layers.

Common ways to apply data mining methods

R, Python notebooks

Common software libraries: Inception, Keras, sklearn