Intro to machine learning

Ed Morrissey

Morrissey Group: Quantitative biology of cell fate and tissue dynamics
Centre for Computational Biology - WIMM

Overview

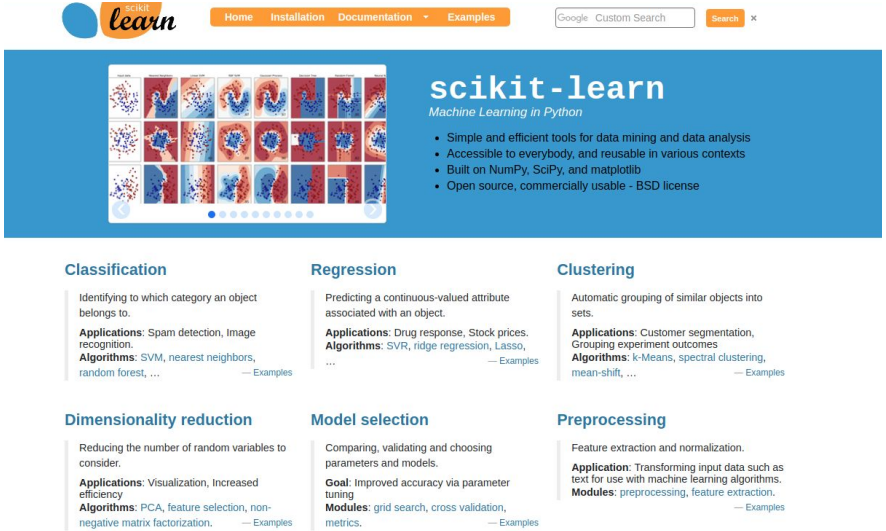
- Machine learning
  - Methods
  - Why it’s become popular
- Iris data set
- Decision trees
- Overfitting
- Random forest
- Dimensionality reduction methods

Kaggle

- Platform for predictive modelling and analytics competitions
  - https://www.kaggle.com/
- Companies post data and a challenge
  - Users compete to produce the best models for predicting and describing the datasets
  - Normally there is a cash prize
- Leaderboards
- After a competition has finished
  - Analysis scripts and results are then posted
  - Used as a learning resource

scikit-learn

- High quality python machine learning library
  - Very nicely structured
  - Well documented
  - Plenty of examples
- Has made machine learning more accessible
- http://scikit-learn.org/stable/



Machine learning overview

- Supervised
  - Regression
  - Classification
  - **Examples**
    - **Support Vector Machines (SVM)**
    - **Decision trees**
    - **Random forest**
- Unsupervised
  - Clustering
    - k-means
  - **Dimensionality reduction**
    - **PCA**
    - **t-SNE**
    - **Knn graphs**
  - Novelty detection
- Deep Neural Networks covers both (more on this in the next talk)

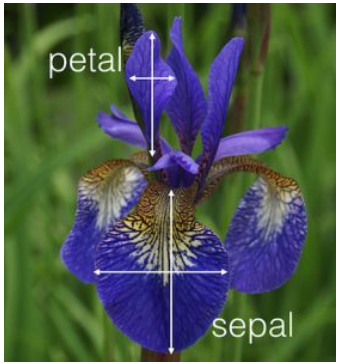
Classification: The aim

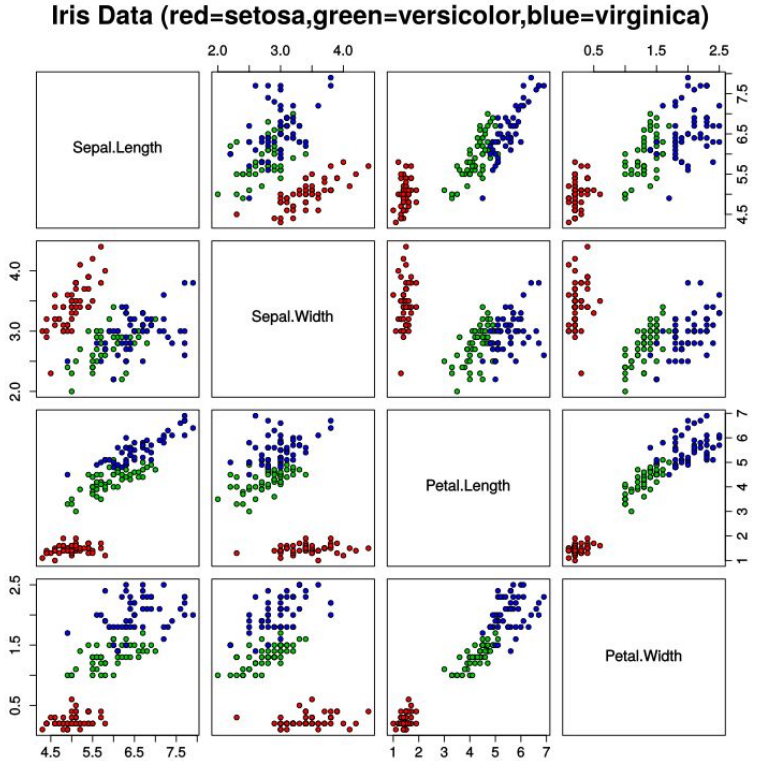
- Supervised so we have data set where we know the truth
- Derive a set of rules to classify unknown data
- Fisher’s Iris data: flower petal data
  - Classic data set
  - Used extensively in documents and tutorials
- Could use for example single cell data where you know the cell type

Data: Iris data set

- Fisher's data
  - Used a lot for tutorials etc (often in libraries)
  - Goto Code!
  - *1936 The use of multiple measurements in taxonomic problems*
  - 50 samples from three species of Iris
    - Iris setosa
    - Iris virginica
    - Iris versicolor
  - Measure 4 things
    - Petal width and length
    - Sepal width and length
- If someone gave you those 4 measurements
  - How would you predict the species?
  - Classifier!

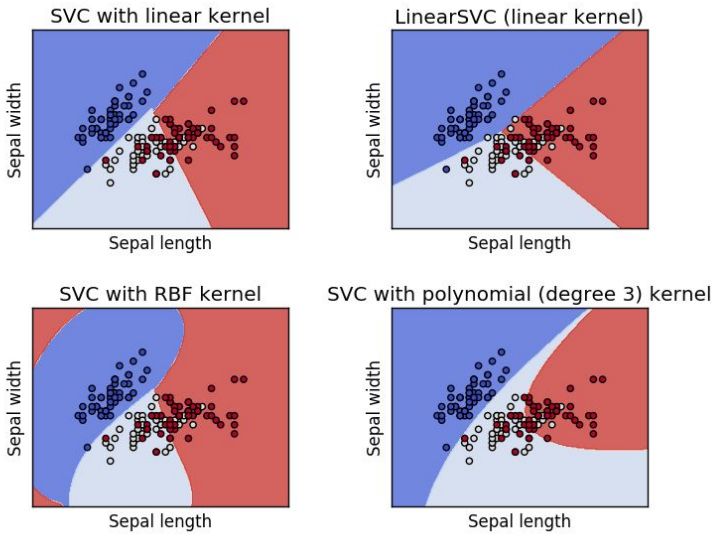
```
In [3]: iris.data
Out[3]:
array([[ 5.1,  3.5,  1.4,  0.2],
       [ 4.9,  3. ,  1.4,  0.2],
       [ 4.7,  3.2,  1.3,  0.2],
       [ 4.6,  3.1,  1.5,  0.2],
       [ 5. ,  3.6,  1.4,  0.2],
       [ 5.4,  3.9,  1.7,  0.4],
       [ 4.6,  3.4,  1.4,  0.3],
       [ 5. ,  3.4,  1.5,  0.2],
       [ 4.4,  2.9,  1.4,  0.2],
       [ 4.9,  3.1,  1.5,  0.1],
       [ 5.4,  3.7,  1.5,  0.2],
       [ 4.8,  3.4,  1.6,  0.2],
       [ 4.8,  3. ,  1.4,  0.1],
       [ 4.3,  3. ,  1.1,  0.1]
```

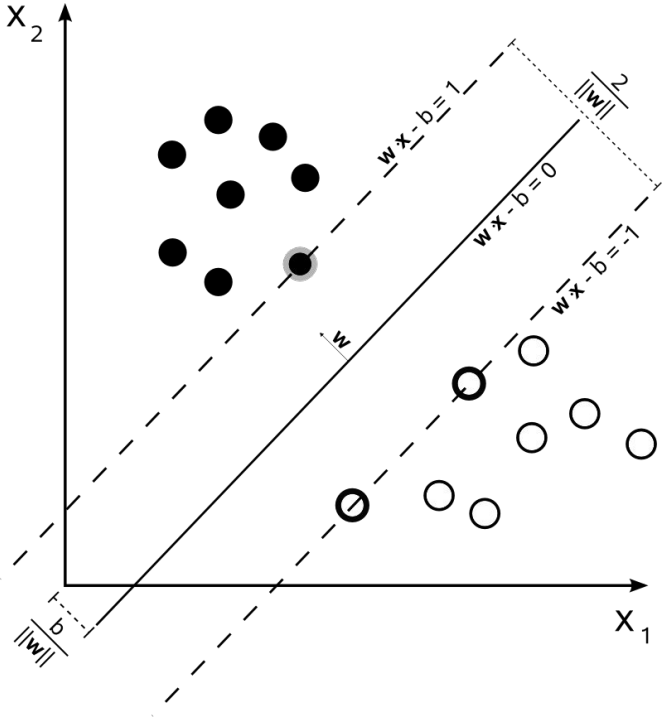




Classification: SVM

- Non probabilistic classifier
- Partition space to predict classes (hyperplanes)
  - Can use several functions for this
    - Example is linear

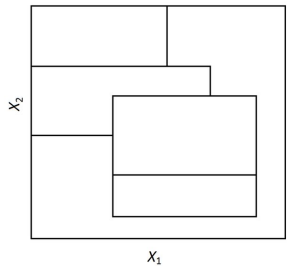




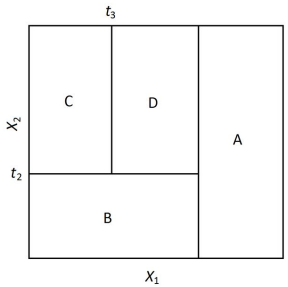
Classification: Decision trees

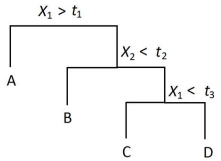
- Decision tree
  - Sequence of decisions
- Train with data to find tree
- Apply rules to new data to predict
- Tends to overfit!

This is not valid



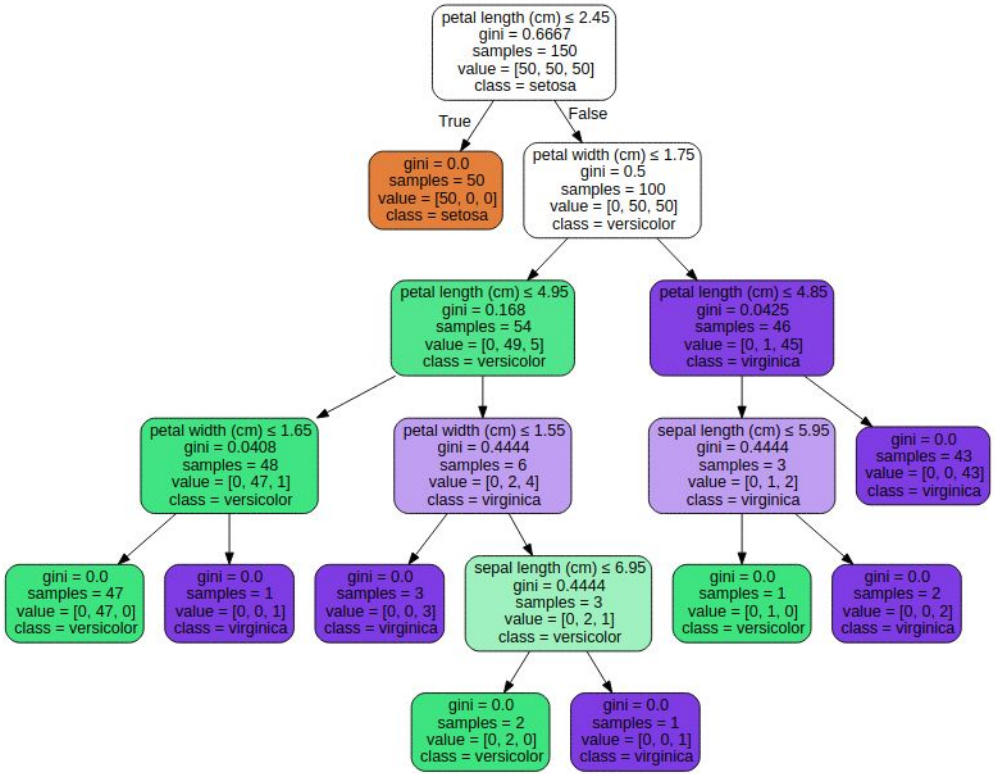
This is





Classification: Decision trees

- For the iris data set: Goto Code!

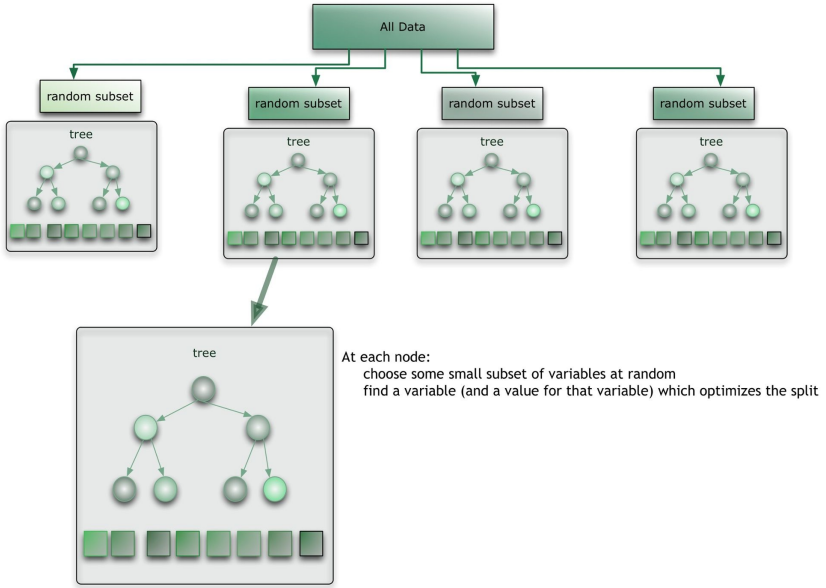


Overfitting

- A common problem in machine learning
- Makes prediction for new samples worse
- There are methods to try to minimise this
- For decision trees a very common variant to minimise overfitting
    - Random forest

Random forest

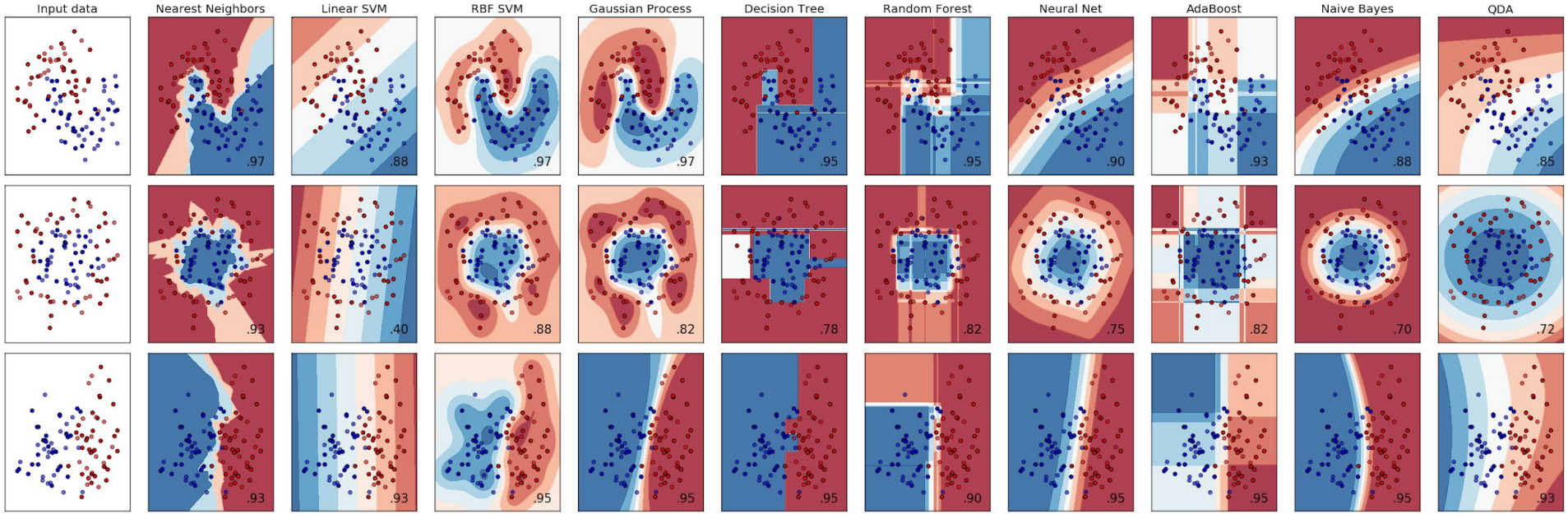
- Loads of decision trees run on random subsets of the data
  - Forest!
  - Also use subsets of the variables when splitting
- Consensus prediction from the forest of trees
- Reduces overfitting
- Finds variable importance



http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics

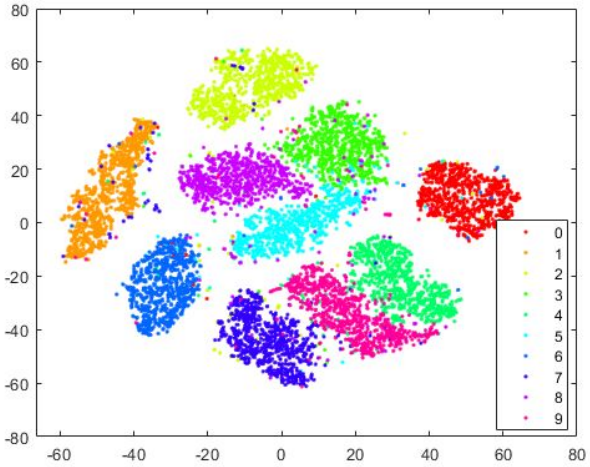
Classification: Other classifiers and performance

http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html



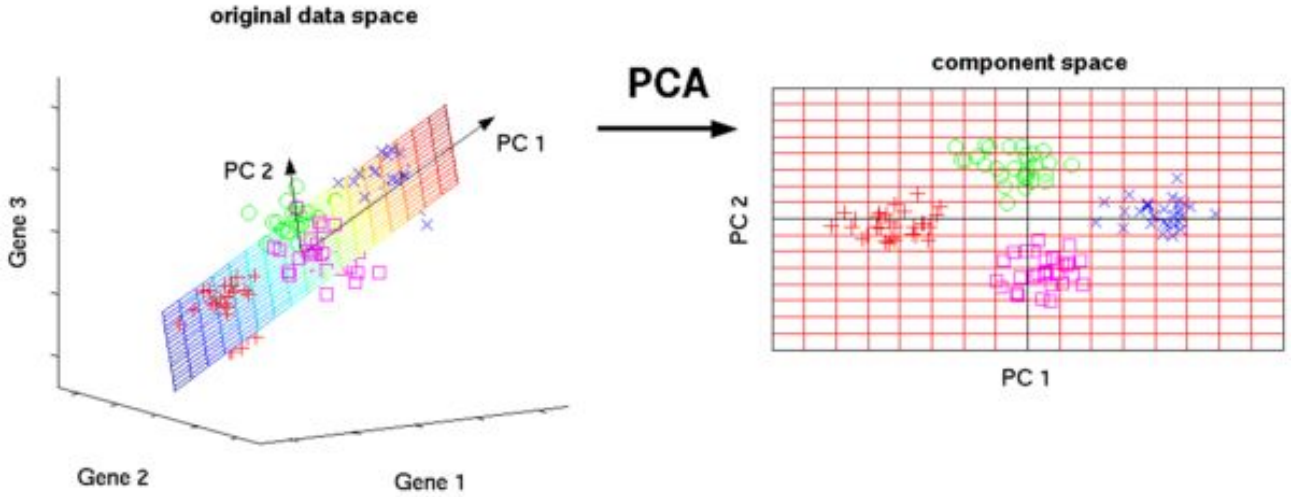
Dimensionality reduction

- Very useful in single cell techniques
  - scRNA-seq
    - 20,000 genes x 10,000 cells
- 20,000 => 2
  - Chucking out loads of information!
  - Not the truth, just a very useful tool
- Can be done in a number of ways
  - Right way?
- Examples
  - PCA - Linear
  - tSNE - non linear
  - Knn graphs - “non linear”

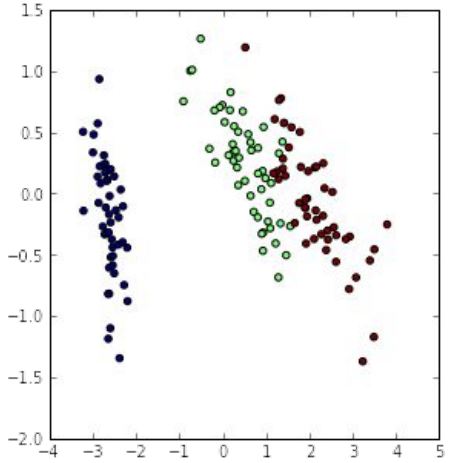


Dimensionality Reduction: PCA

- You get as many new dimensions (principal components) as you start with
- The new ones (PC) are ordered by “how much information” they have
- People normally do dimensionality reduction by discarding “low info” dimensions

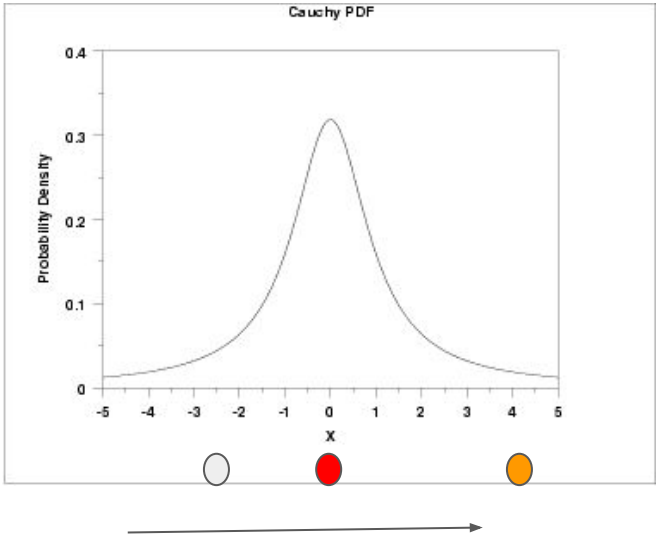


Iris data

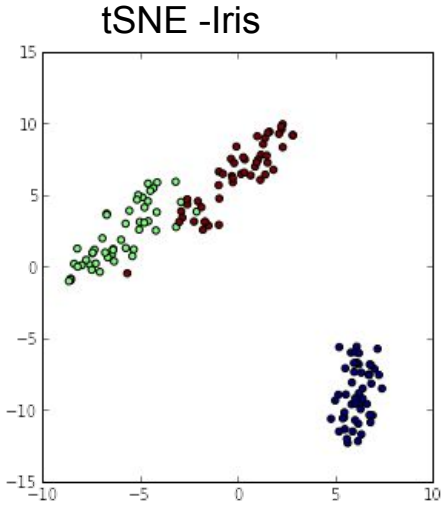


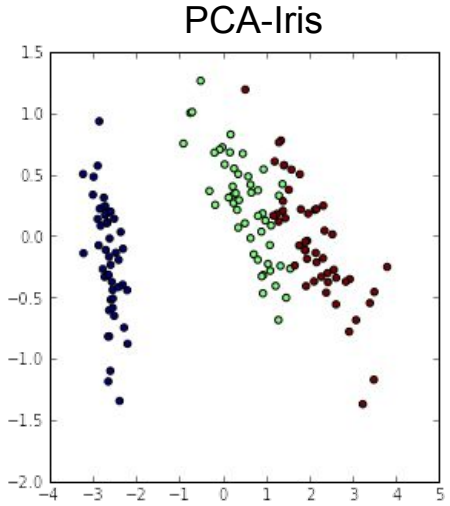
tSNE - nonlinear dimensionality reduction

- Calculate a distance between cells using a gaussian (and some tricks) in 20,000 dims
- Find a 2D arrangement that respects those distances (using t-distribution in 2d)
- https://distill.pub/2016/misread-tsne/



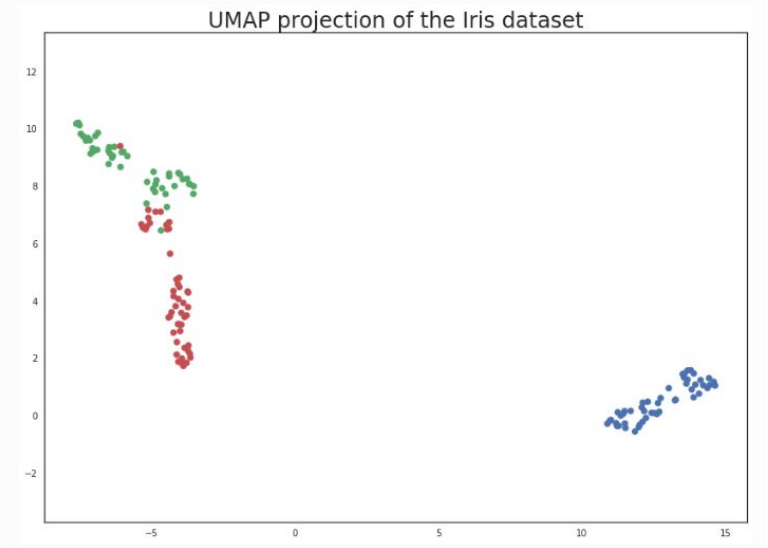
Expression of gene A





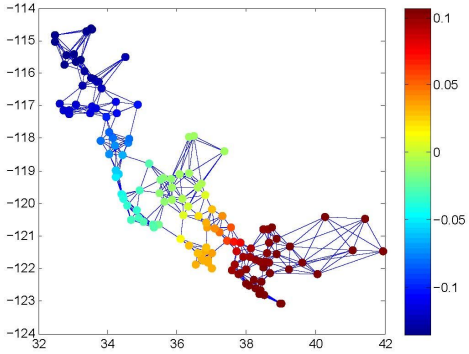
Umap:Uniform Manifold Approximation and Projection

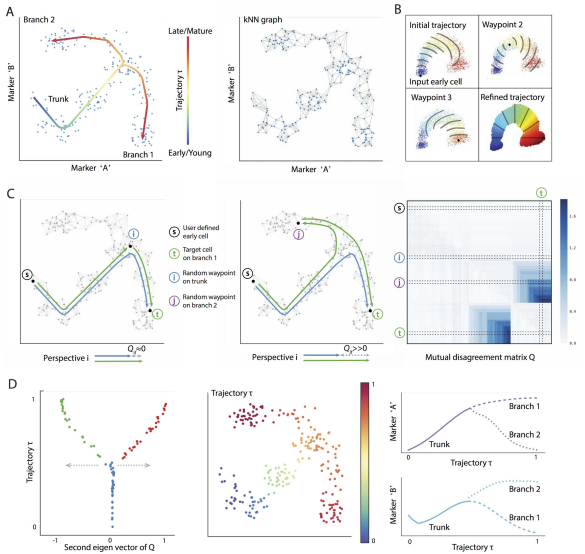
- Based on algebraic geometry
- Much faster than tSNE
- Can add new data
- Can also add labels
- Gives good results

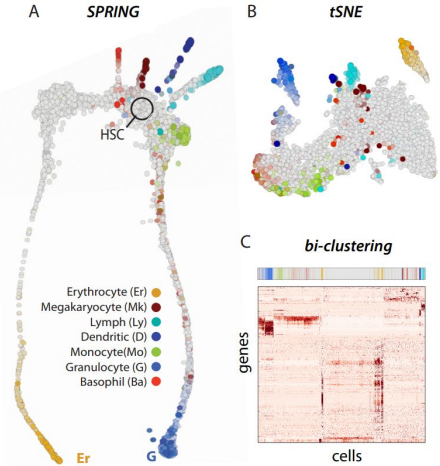


Knn

- K-nearest neighbour graph
- Can define trajectories or clusters
  - Phenograph uses community detection
- Quick
- A knn of k = 5
  - Calculate distance between cells ie d = $(gene\_A_{c1} - gene\_A_{c2})^2 + (gene\_A_{c2} - gene\_A_{c2})^2 + \ldots$
  - Join every cell to its 5 closest cells (most similar)
- Spring a web tool to calculate and view knn graphs

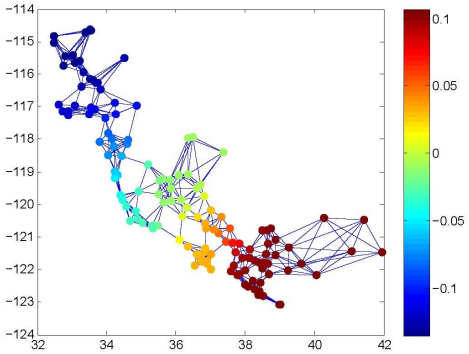


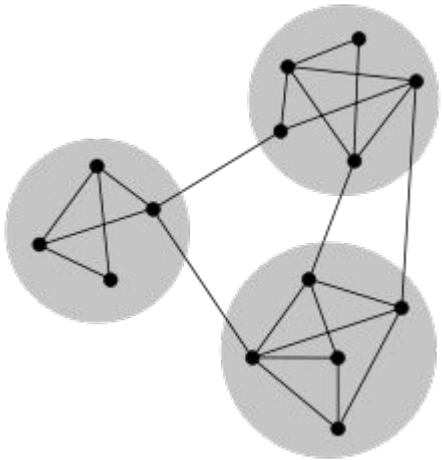




Knn for clustering

- K-nearest neighbour graph
- Find clusters
  - Nodes that are more connected between themselves than others
  - Used to find communities in social media analysis
  - Different names
    - Louvain clustering
    - Phenograph (used for cytof data)

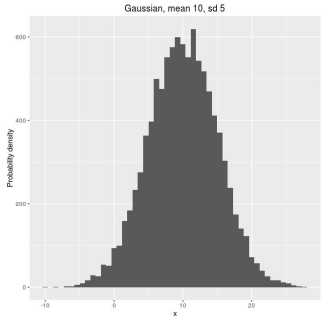


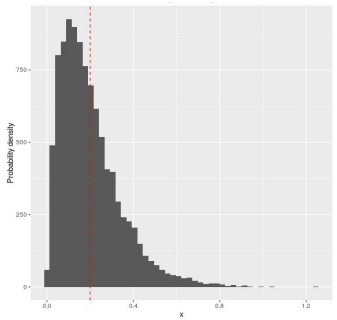


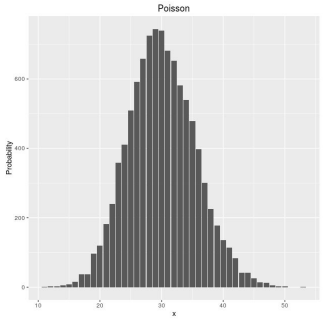
Machine learning vs Stats: Generally speaking

- Both
  - Very similar
  - Often use similar statistical tools to solve problems (sometimes with different names)
  - More and more they are blending into each other
- Stats
  - More interested in learning about the system
  - Make specific assumptions about the data (which you can normally check)
  - Quantify uncertainty
  - Typically moderate size data sets
- ML
  - Focused on accurate prediction and speed
  - Tends to deal with huge datasets
  - Less interested in understanding the underlying data and how it was generated
- Slightly provocative:
  - 'machine learning is statistics minus any checking of models and assumptions'. Prof Brian D. Ripley
  - ML tend to be seen as more hacky by statisticians. A bit black box like.
- Deep Neural Networks have made a huge impact and are drawing attention from statisticians

The End


Gaussian, mean 10, sd 5
Probability density
x


Probability density
x


Poisson
Probability
x