Machine Learning for Computational Biologists

Boyko Kakaradov
UCSD Bioinformatics Bootcamp
20 September 2013

just a taste

- What is Machine Learning?
- How is it useful in Bioinformatics?
- Where to learn/do more?

What is Machine Learning?

Enabling computers to learn, decide, act without being explicitly programmed. ~ Arthur Samuel, 1959

What is Machine Learning?

Enabling computers to learn, decide, and act without being explicitly programmed. ~ Arthur Samuel, 1959

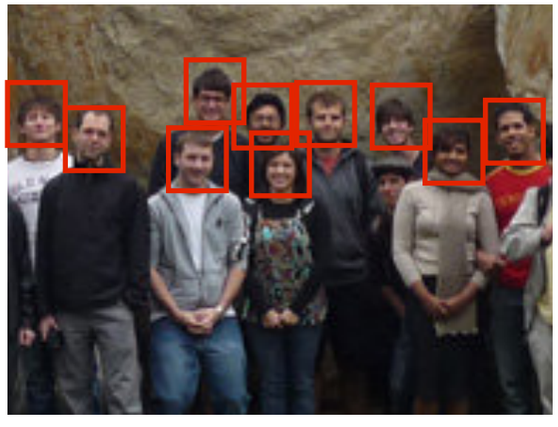


Any sufficiently advanced technology is indistinguishable from magic. ~ Arthur C. Clarke, 1973

What is Machine Learning?

Enabling computers to learn, decide, and act without being explicitly programmed. ~ Arthur Samuel, 1959

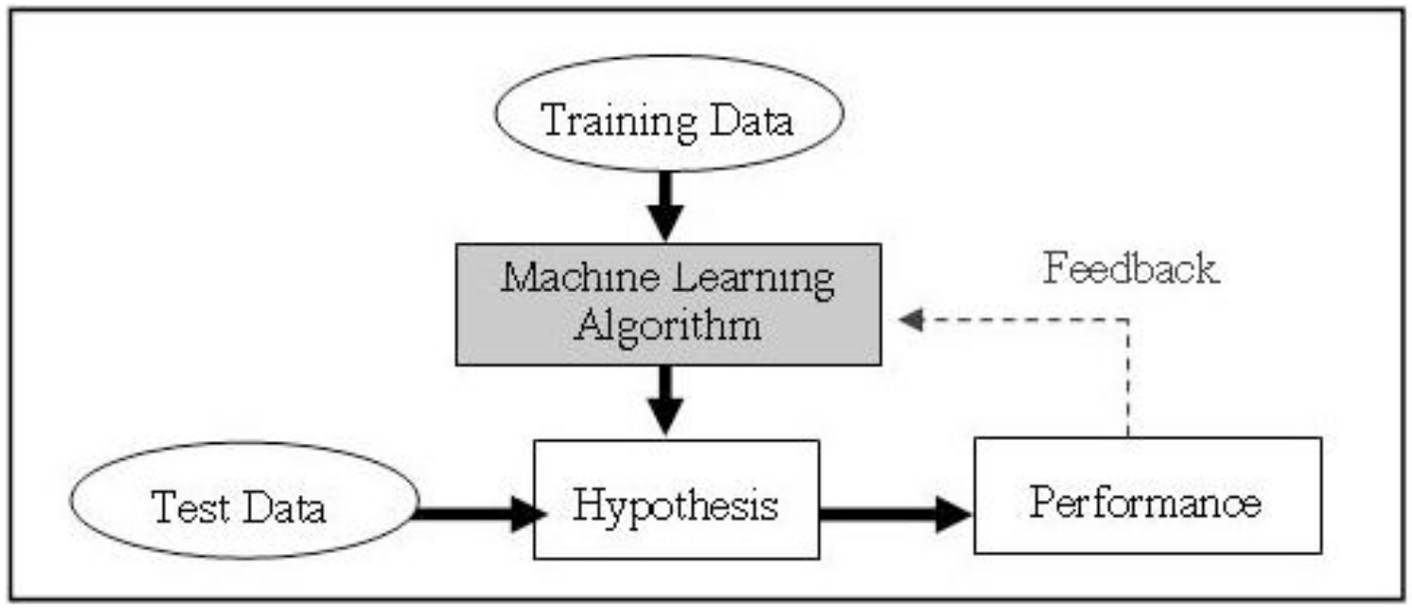
Learn



Decide

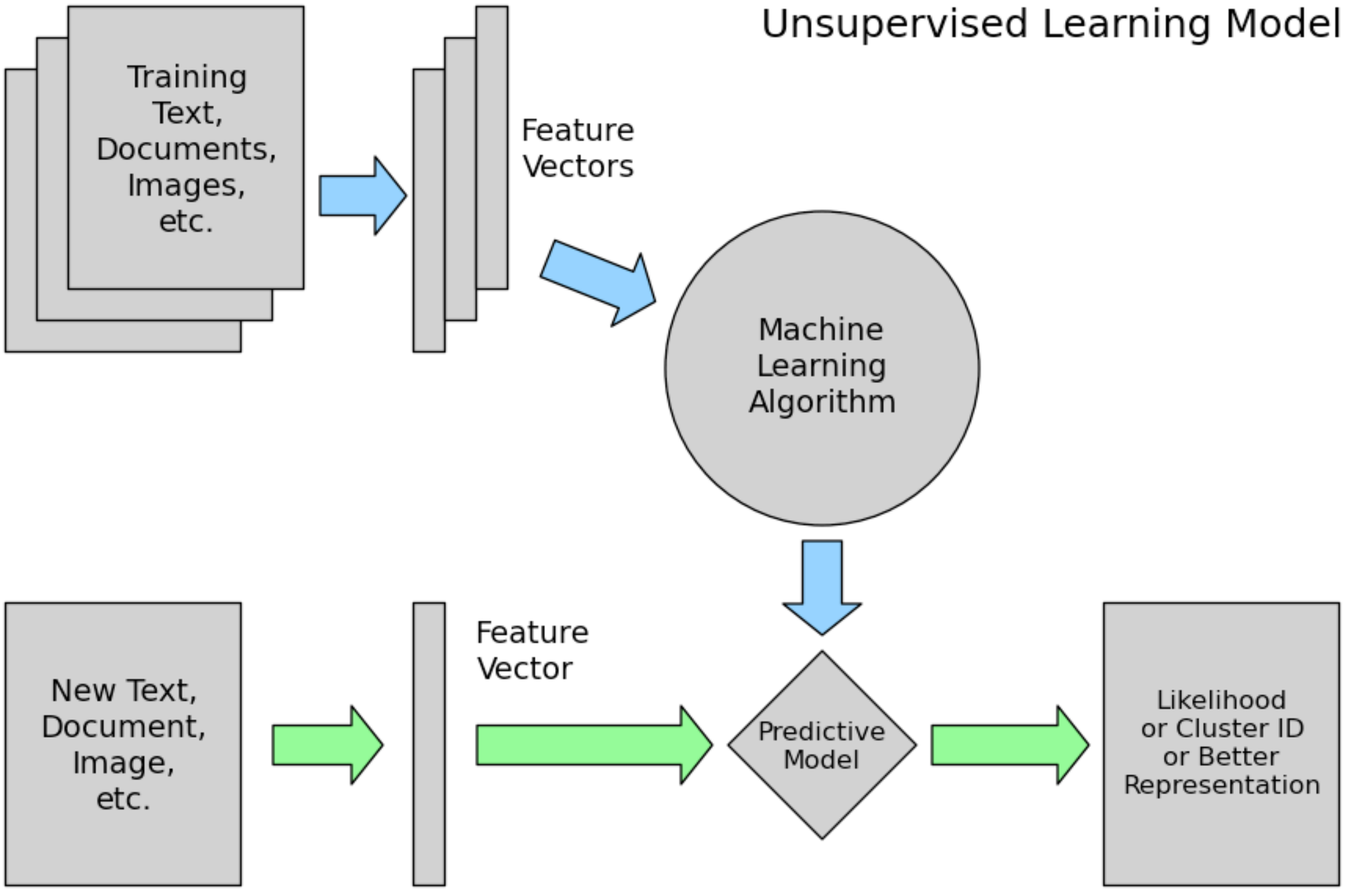
Act

ML in theory

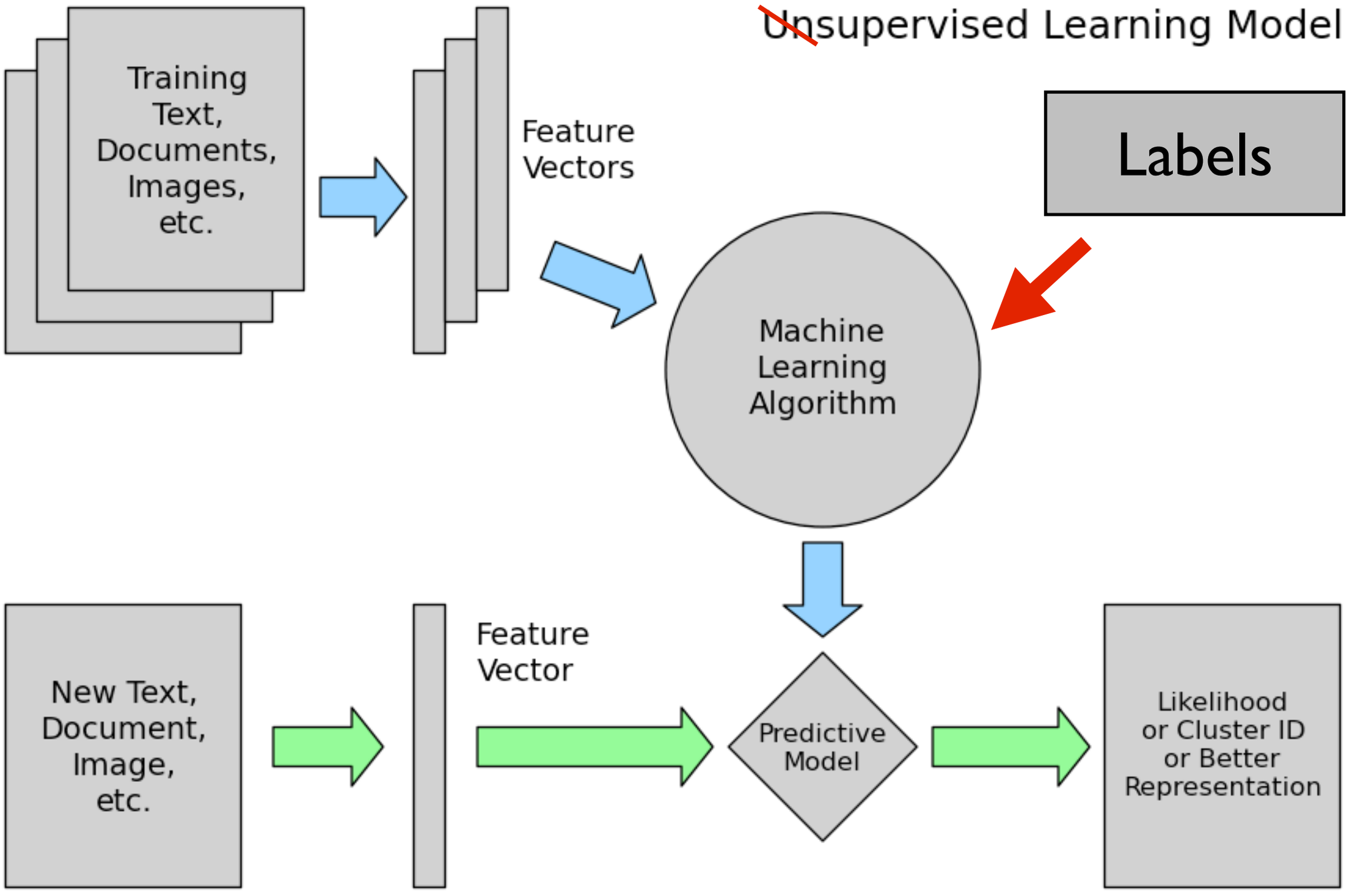


Inputs	Continuous, Categorical, Ordinal	Labels
Types	Unsupervised, Reinforcement	Supervised
Hypothesis	Clustering	Classification, Regression
Methods	optimization, sampling, Bayesian	active
Outputs	predictions and their confidence	

ML in practice



ML in practice



Example datasets

Samples:

texts
images
cells
genes
peptides

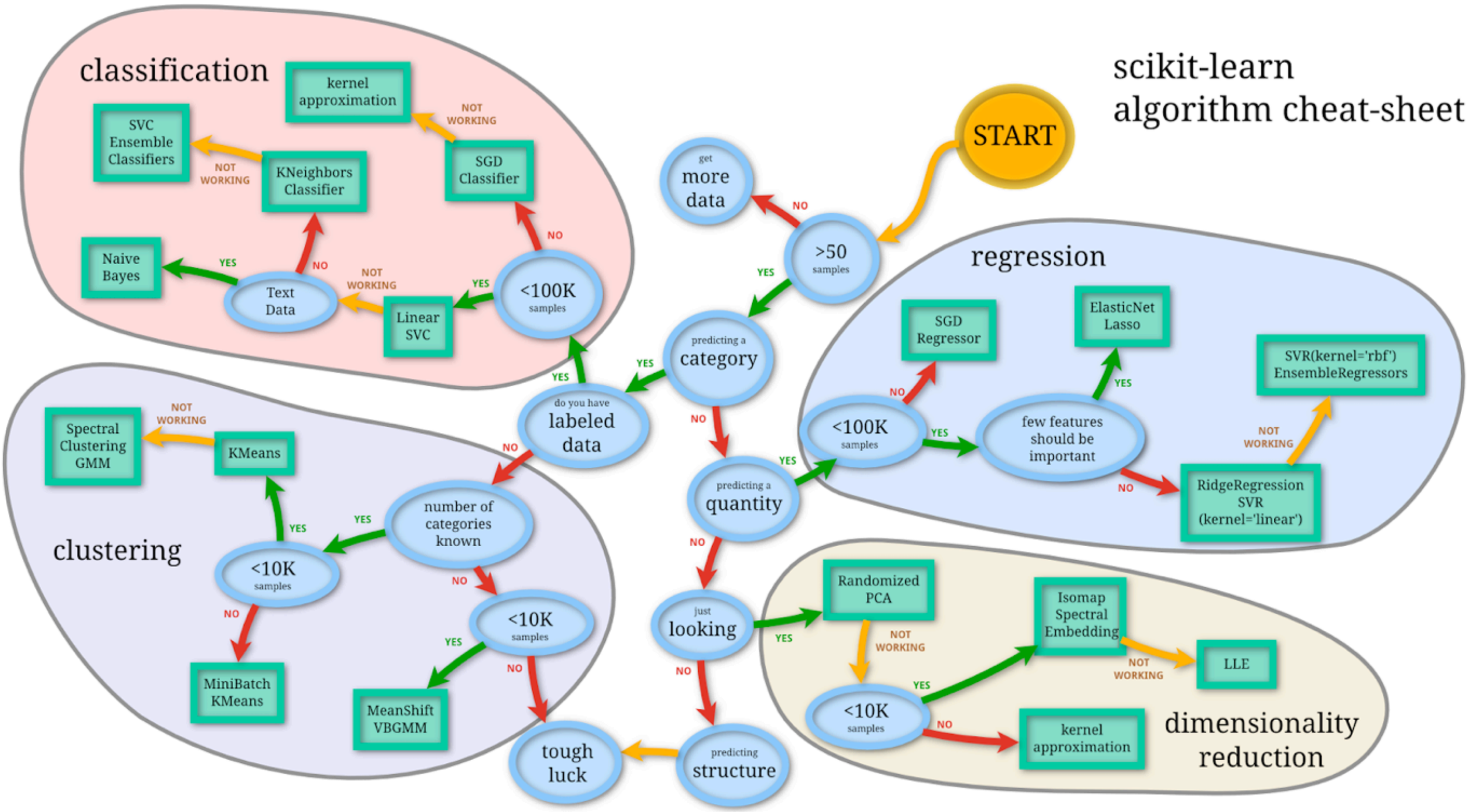
1	0.5	12.3	[152,138]	ACGT
0	0.3	32.8	[45,97,83]	GCGC
1	0.2	14.2	[84,92,...]	ATGG
1	0.7	3.7	...	AGGT
0	0.1	0.6	...	ACCG

Labels:

topic	ML
face_of	me
cancer	no
predispos.	-15%
protein ID	p53

Features: word count, pixel intensity, gene expression, protein mass spectra

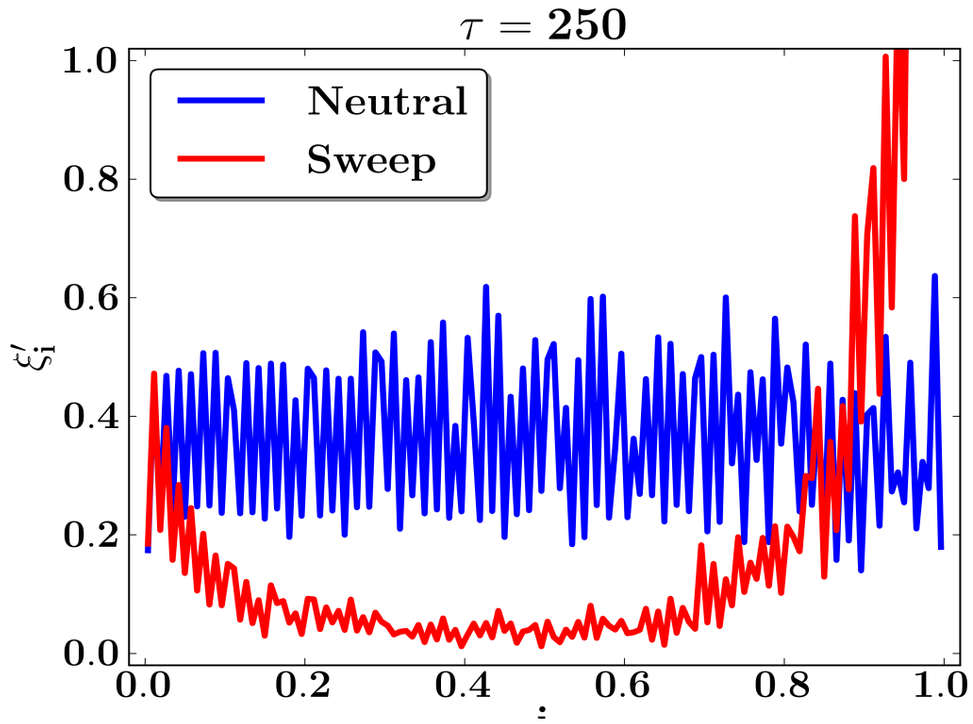
Example Algorithms



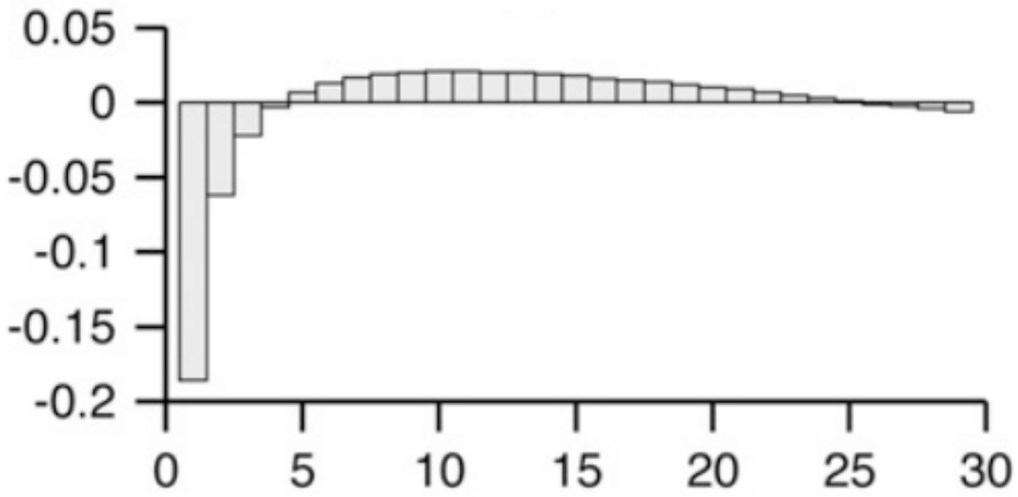
ML in Bioinformatics

- Gene and motif finding: HMMs, EM algo.
- Sequence alignment, homology prediction
- RNA & Protein structure prediction
- Gene/Protein network reconstruction
- Gene Expression module clustering

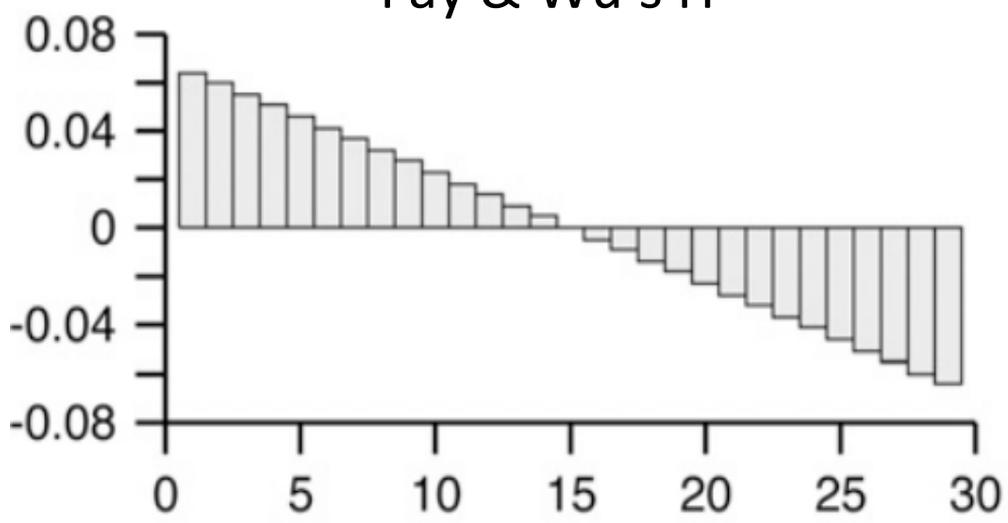
Learning Natural Selection from the Site Frequency Spectrum



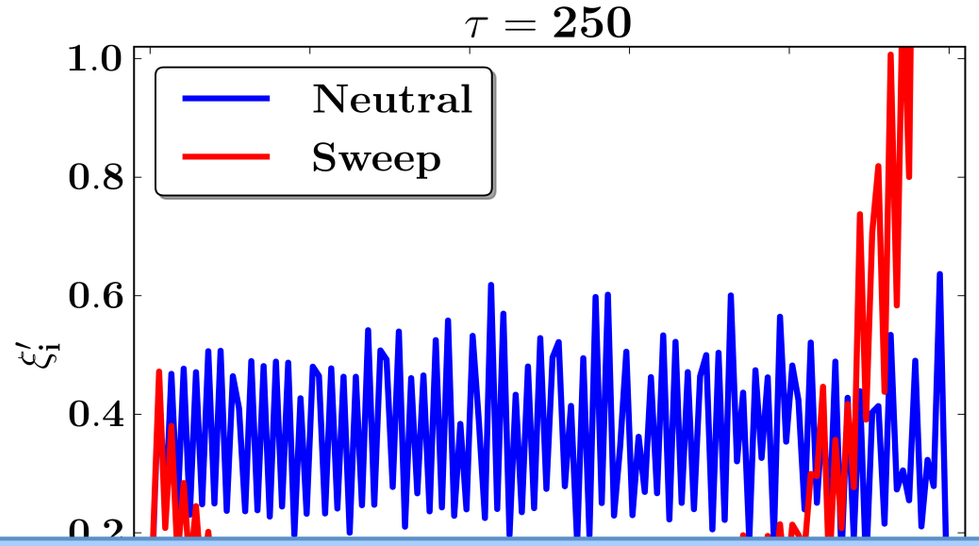
Tajima's D



Fay & Wu's H

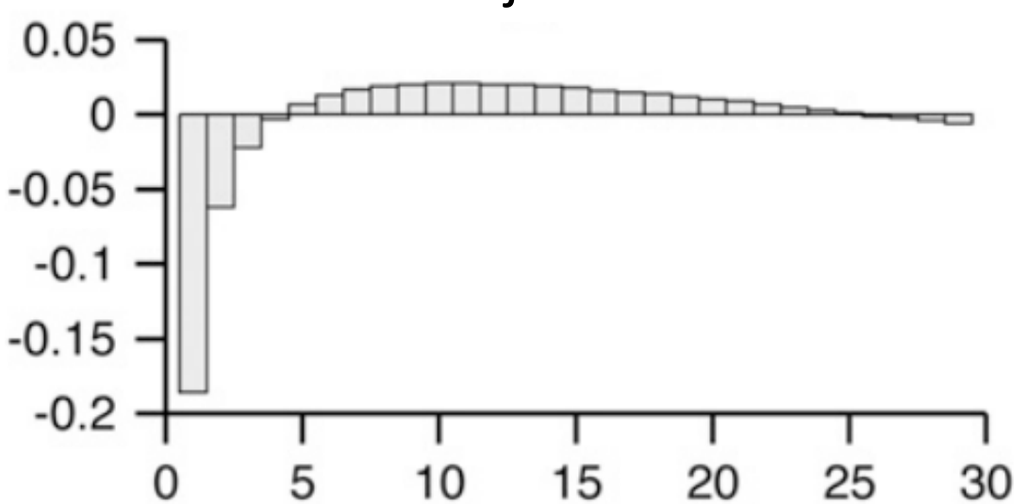


Learning Natural Selection from the Site Frequency Spectrum

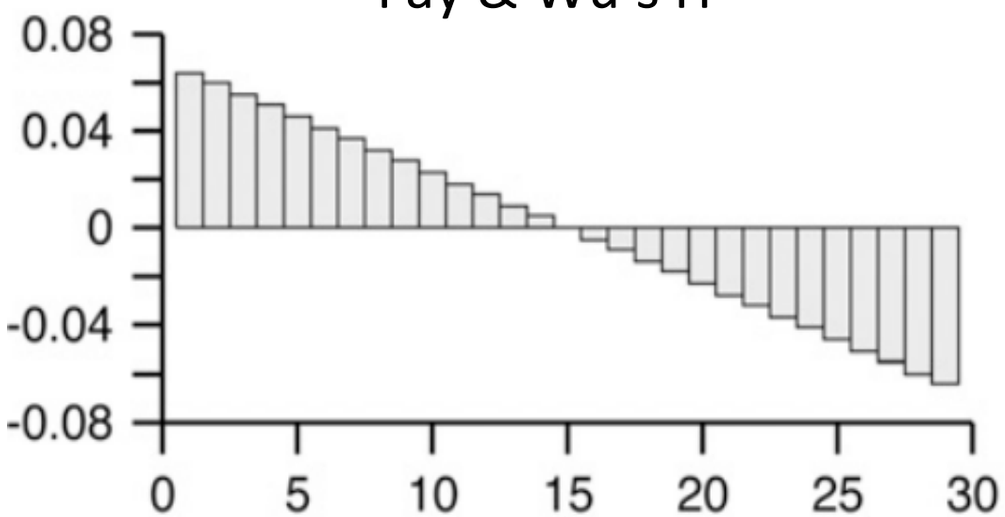


Classic: formulate weights from theoretical reasoning
Modern: learn weights from empirical data

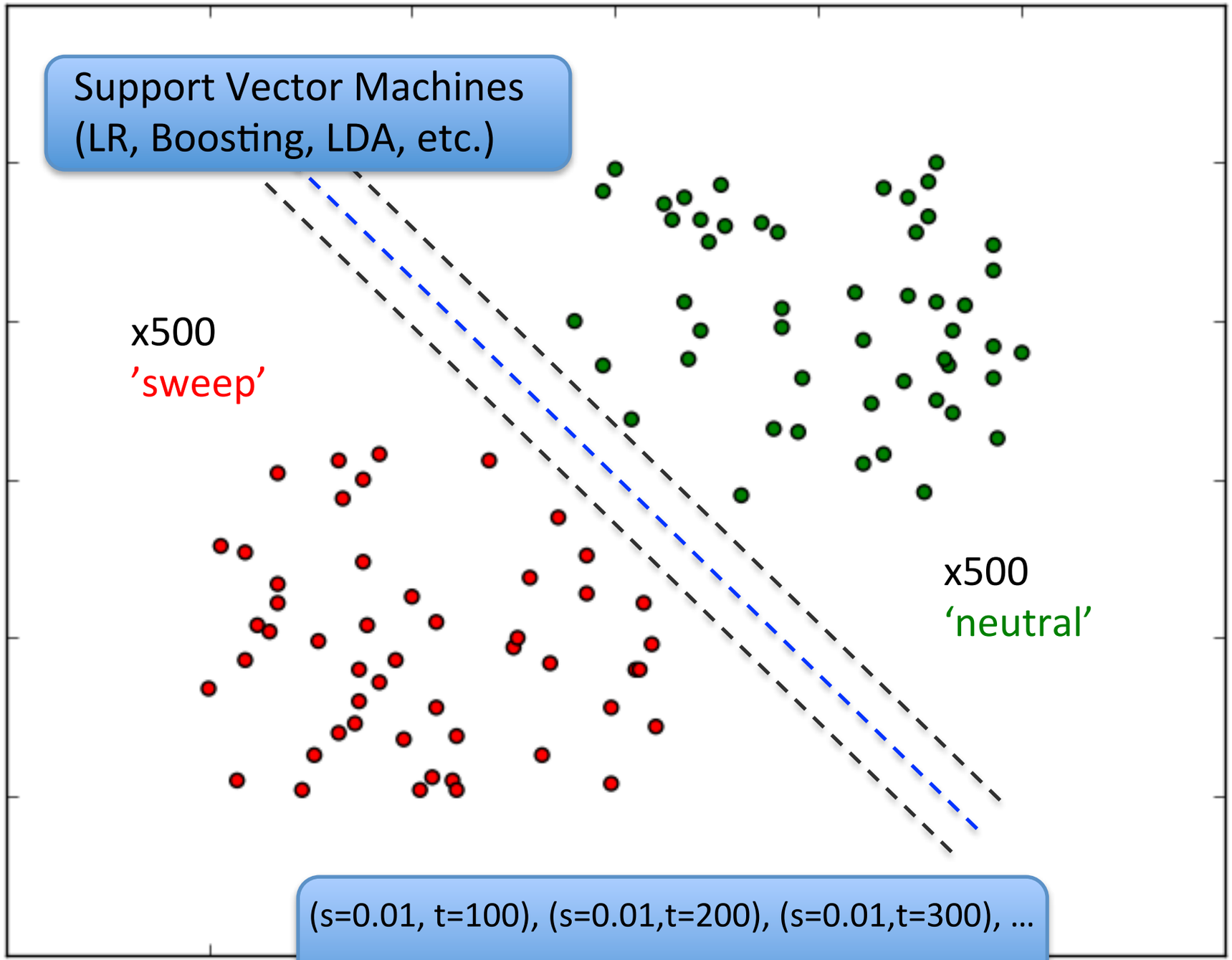
Tajima's D



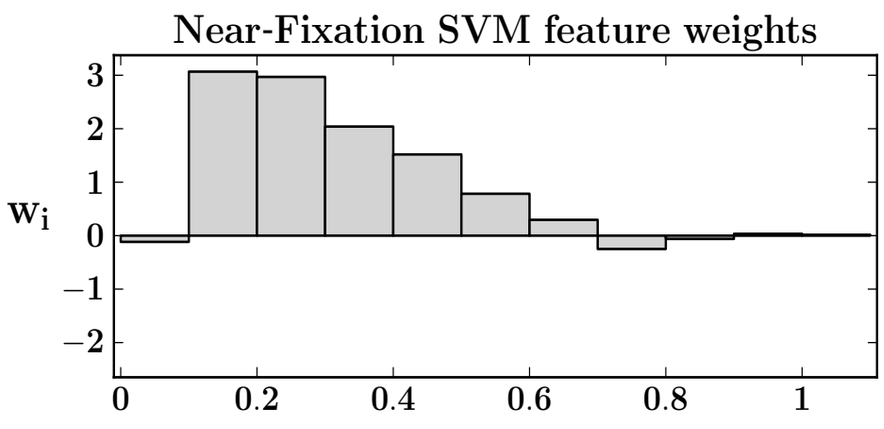
Fay & Wu's H

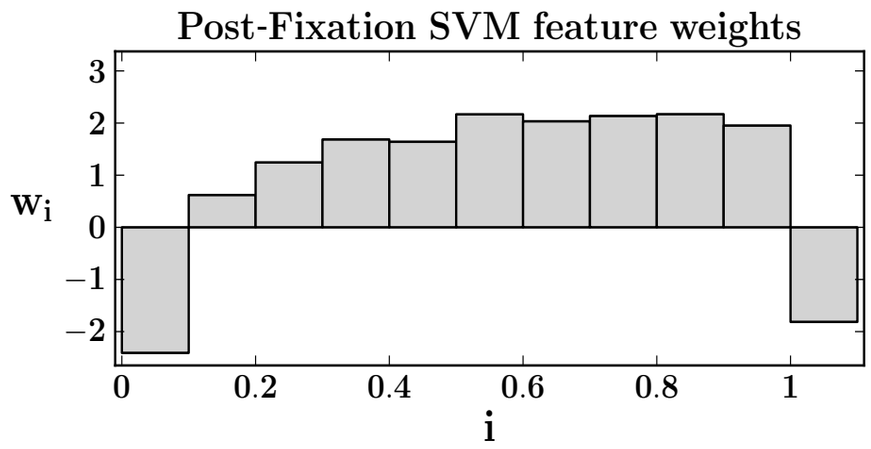


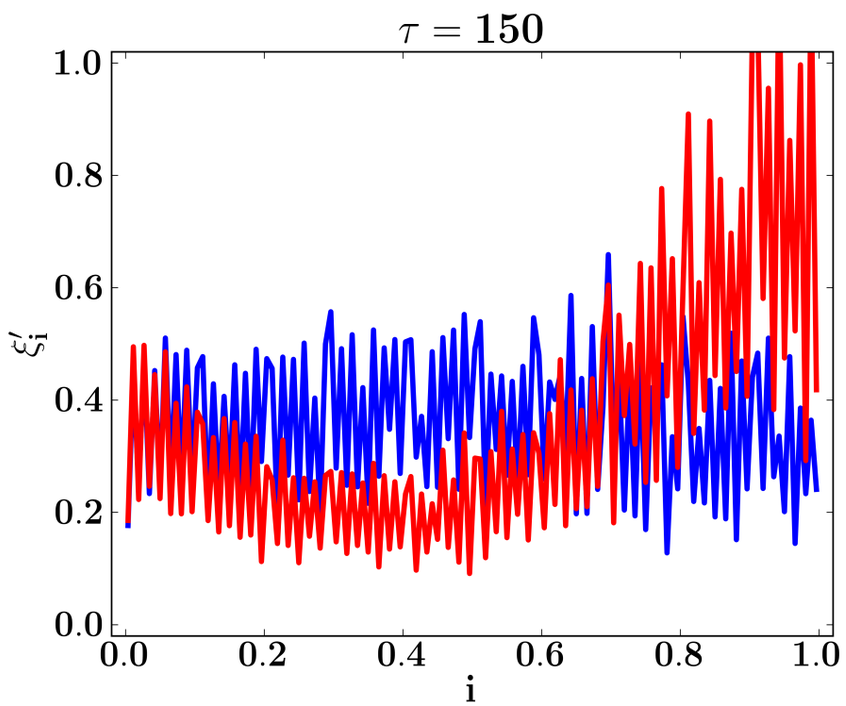
Supervised Learning

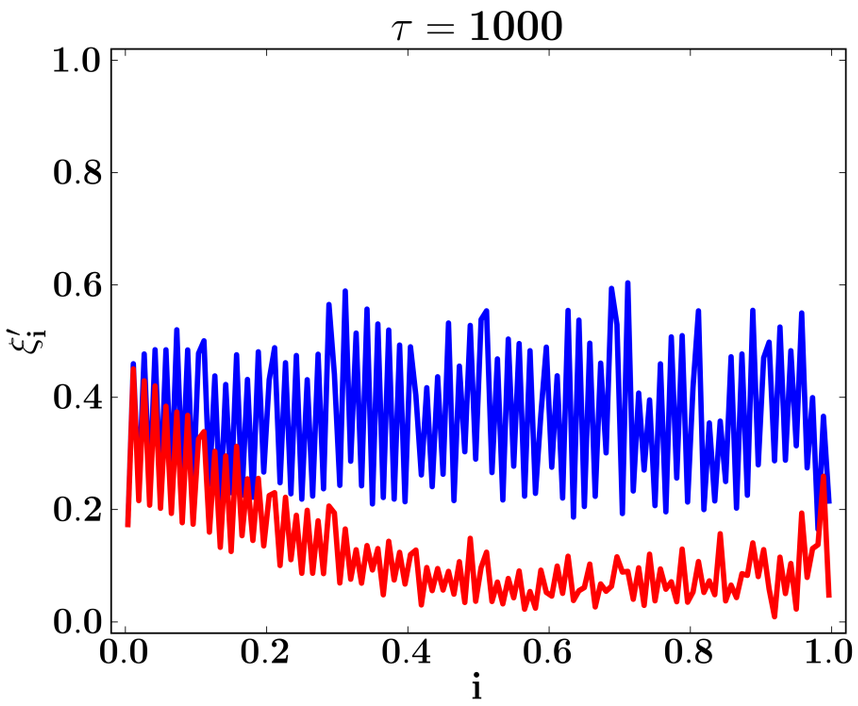


General Model Feature Weights









Recommended Classes

- On campus:

  - CSE151 Introduction to AI, Chaudhuri
  - CSE250 Basic foundations, Saul/Elkan
  - CSE291 Advanced topics, Saul/Dasgupta
  - CSE259 Weekly Seminar, various speakers
  - ECE273 Convex Optimization, Lanckriet
  - COSMAL mailing list, annual poster session

Recommended Courses

- MOOCs to supplement campus classes:
  - Machine Learning, A. Ng or P. Domingos

  - Probabilistic Graphical Models, D. Koller

  - Neural Networks for ML, G. Hinton

coursera

edX

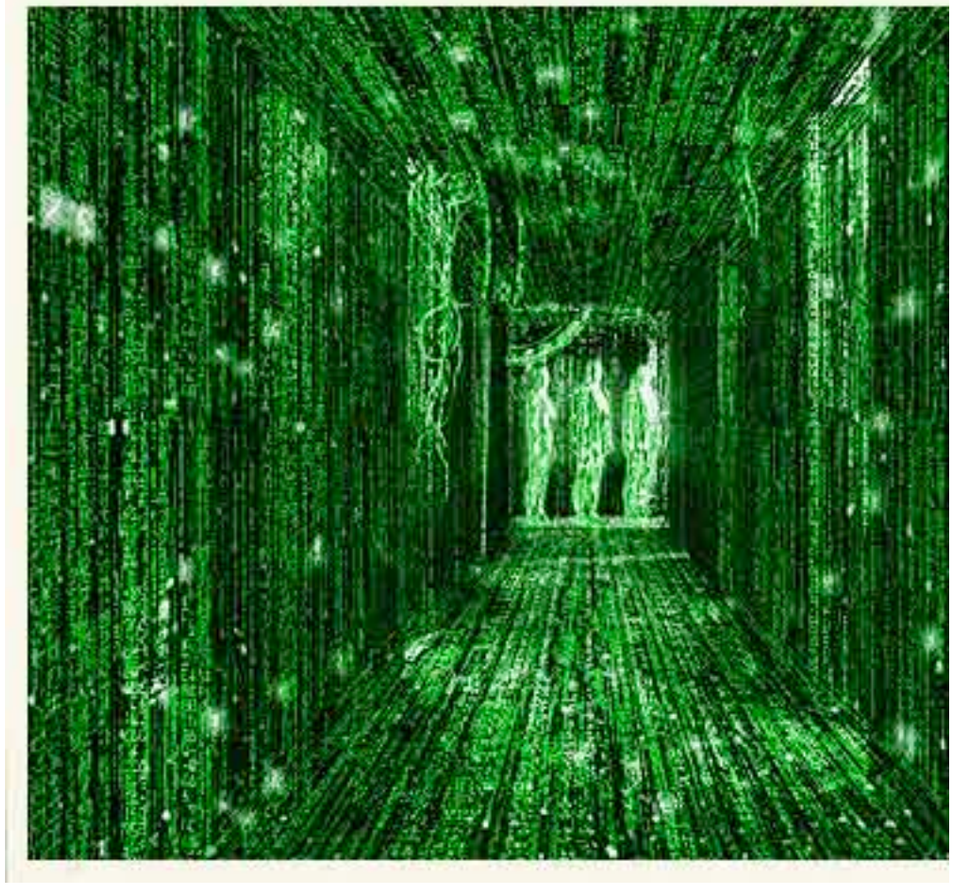
  - Network Analysis for Systems Biology,
  - Computer Vision, F.F. Lee and S. Savarese

Recommended Resources

- Tutorials: VideoLectures, Metacademy
- Discussion forums: MetaOptimize, CrossValidated
- Python packages: scipy, numpy, pandas, scikit-learn

Recommended Resources

- Conferences:
  - General ML: NIPS, ICML
  - Bioinformatics: ISMB, PSB
- Competitions:
  - DREAM challenges
  - Kaggle, InnoCentive, TopCoder



hope this helps