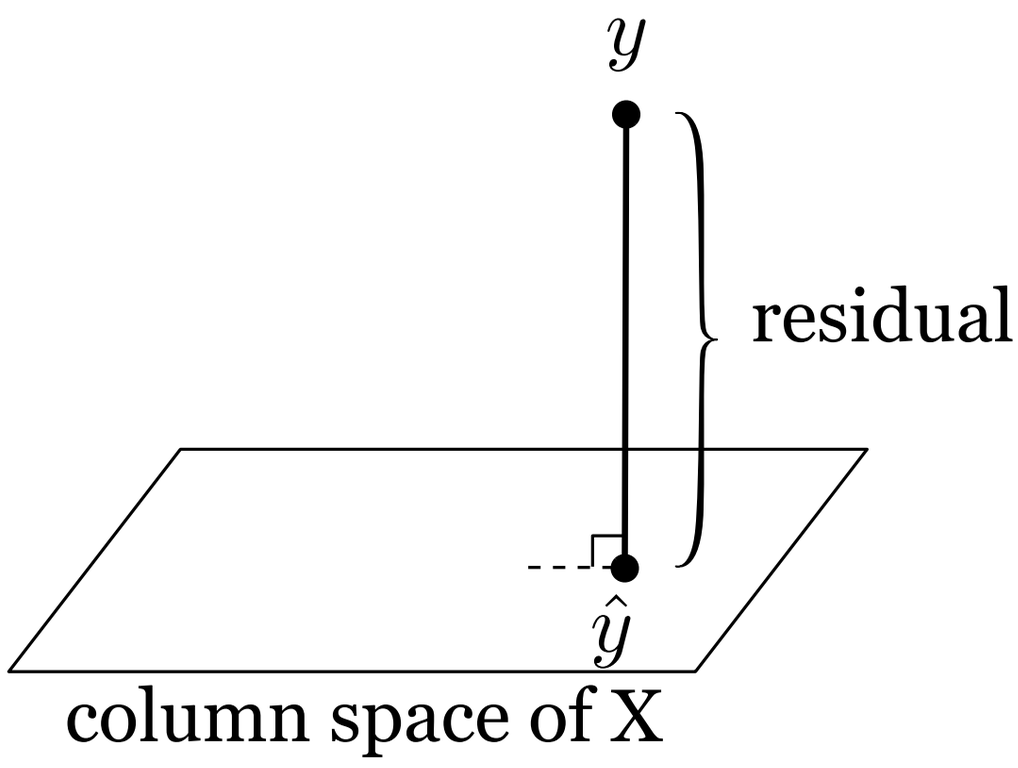
CS109/Stat121/AC209/E-109
Data Science
Classification and Clustering

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

- HW2 due 10/3 at 11:59 pm.
- Reminder: for this and all assignments, make sure to *check* through your submission, both before submitting and by re-downloading from the dropbox and then looking through the file. *No homeworks will be accepted more than 2 days late, since a maximum of 2 late days can be applied to an assignment.*
- Friday lab **10-11:30 am** in MD G115

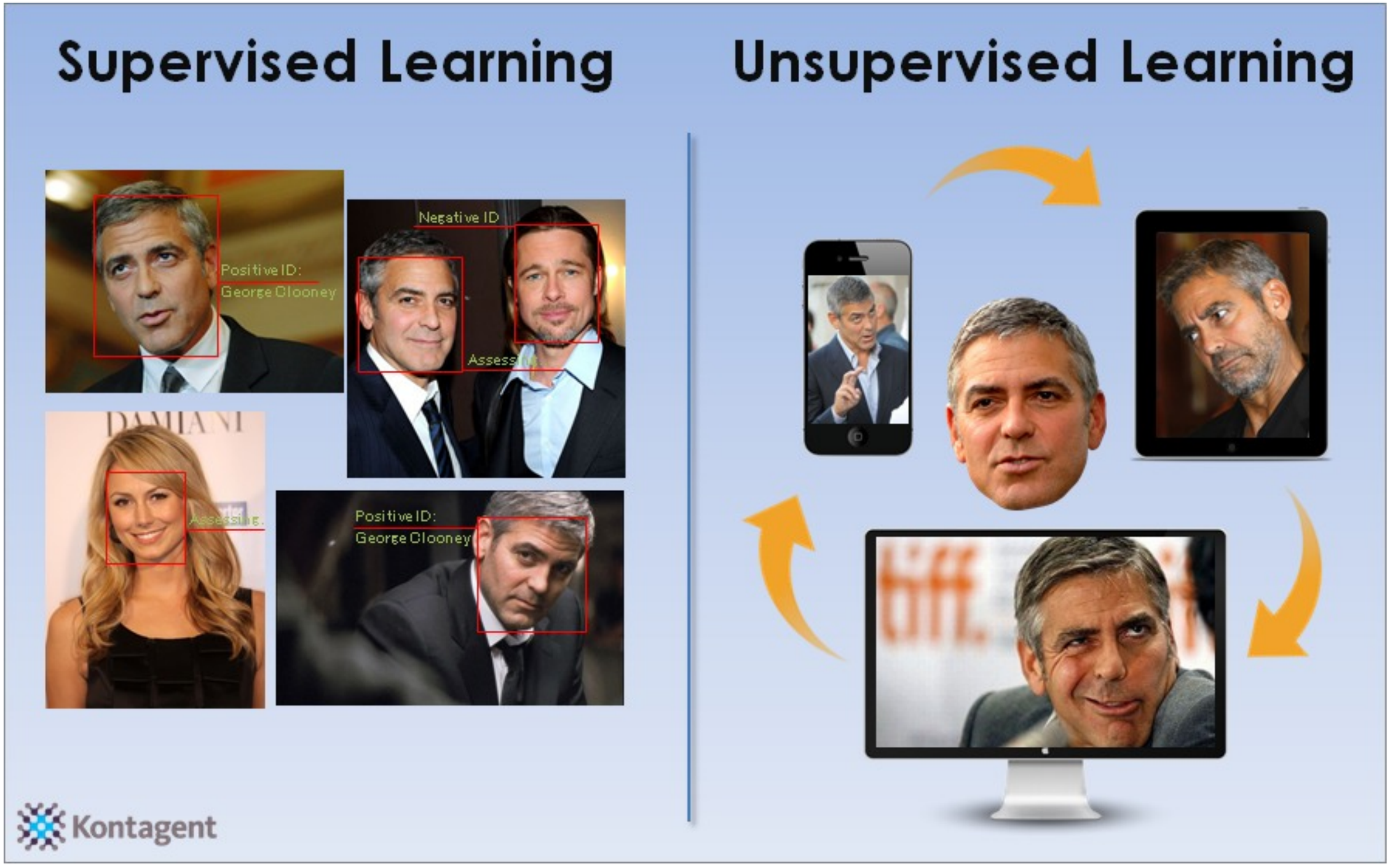
Classification vs. Clustering

Classification is *supervised learning.*
Clustering is *unsupervised learning.*

Classification has pre-defined classes, and training data with labels. Use x's to predict y's.

Clustering has no pre-defined classes. Group data points into "clusters", try to find structure in the data.

Classification vs. Clustering



http://www.kontagent.com/kaleidoscope/2013/01/09/kscope-profile-what-george-clooney-can-teach-you-about-ltv-and-machine-learning/

Discriminative vs. Generative Classifiers

What to model and what not to model?

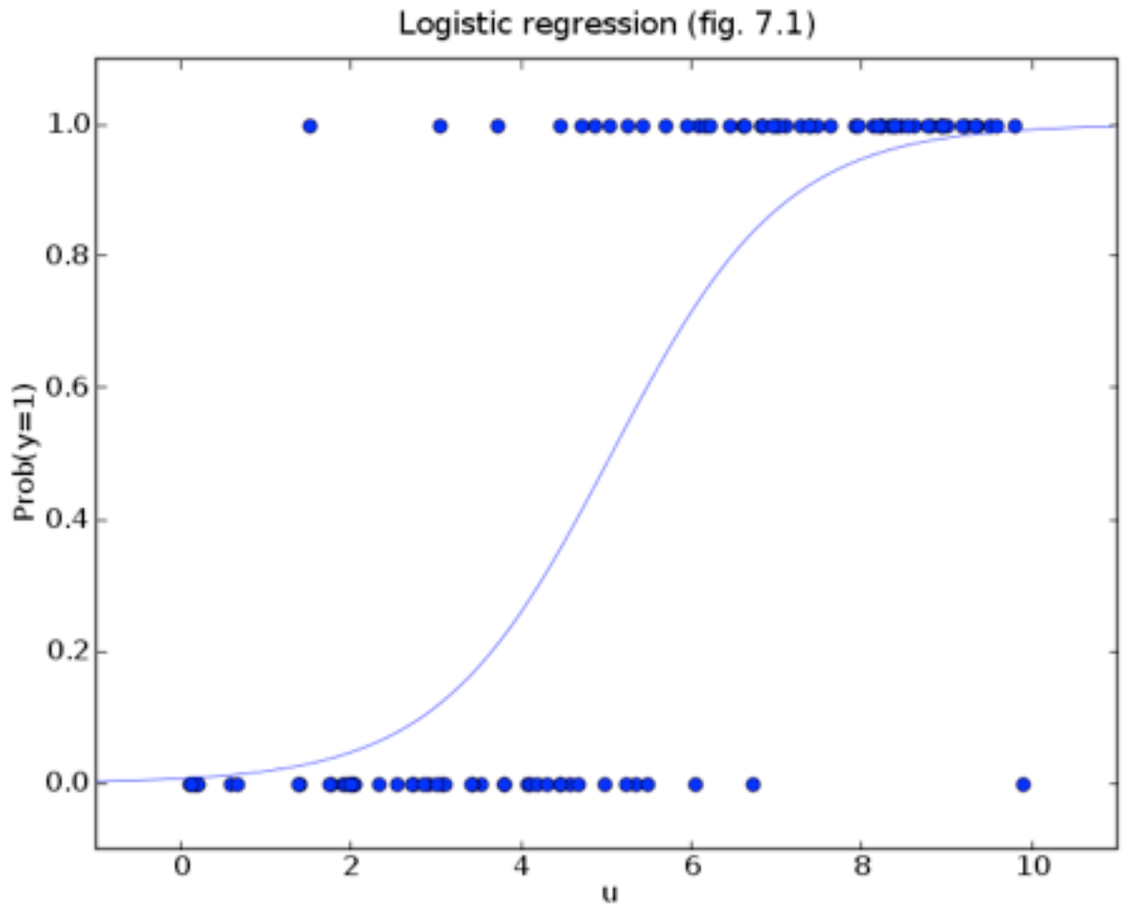
discriminative: directly model $p(y|x)$
generative: give a full model
$p(x,y)=p(x)p(y|x)=p(y)p(x|y)$

Classification via Logistic Regression

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$$

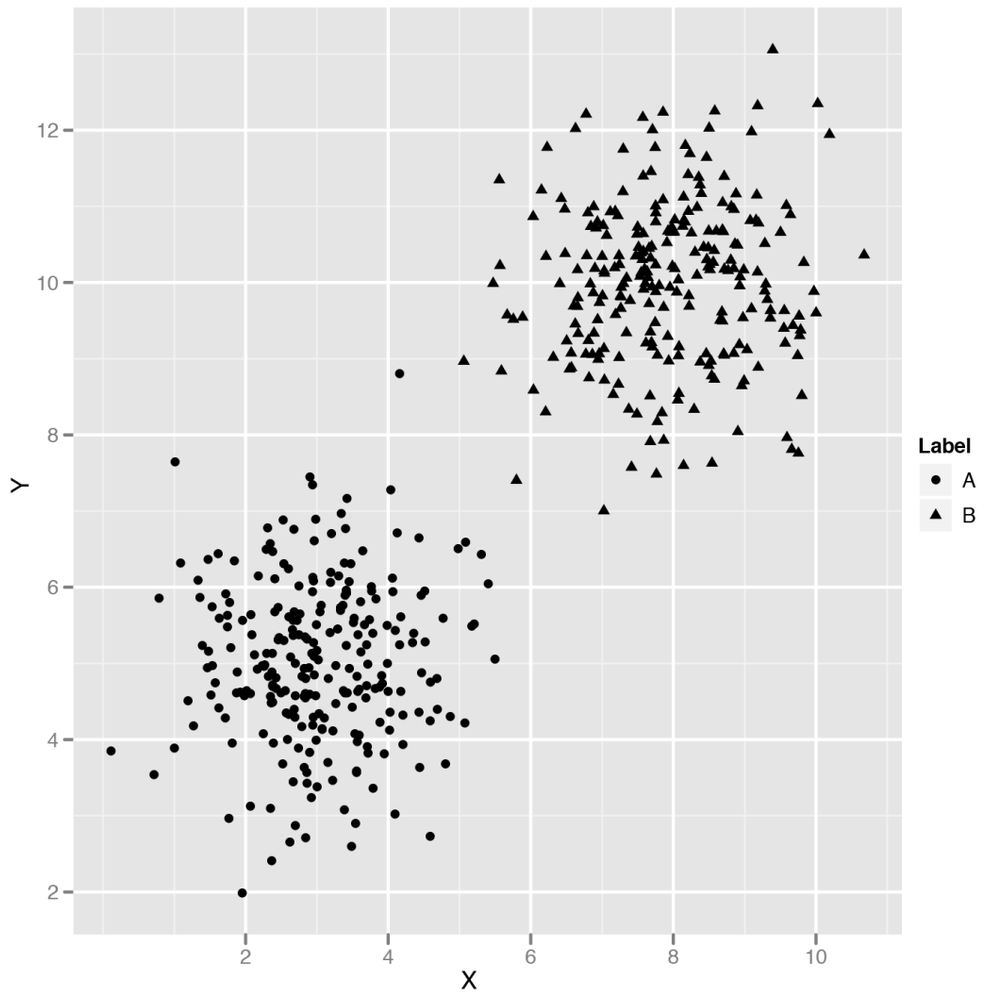
where p is the probability of being in group 1 (can also extend logistic regression to the case of more than 2 groups)

modeling approach: model p(y|x), don't model p(x)

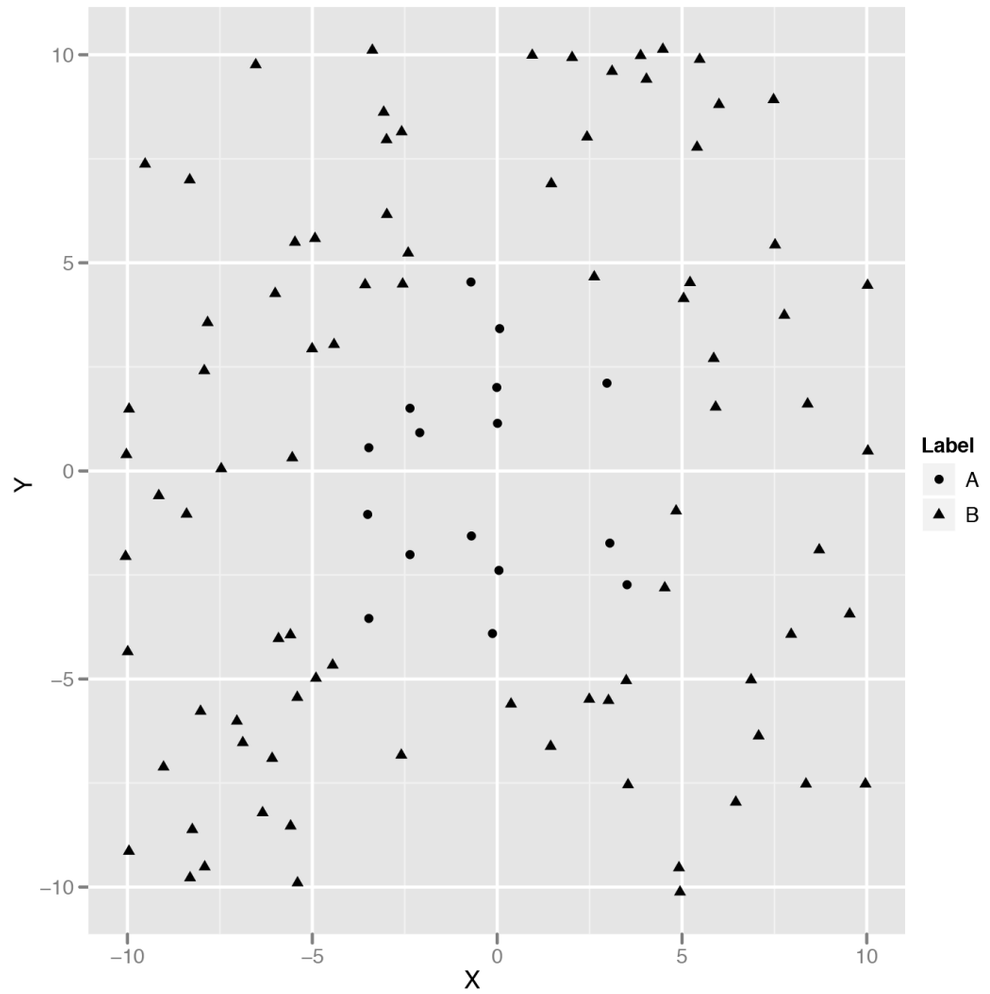


http://cvxopt.org/examples/book/logreg.html

linear decision boundary



nonlinear decision boundary



Conway and White, *Machine Learning for Hackers*

Generative Models

$$P(Y = 1|X = x) = \frac{f(x|Y = 1)P(Y = 1)}{f(x|Y = 1)P(Y = 1) + f(x|Y = 0)P(Y = 0)}$$

(by Bayes' Rule)

Then can model the densities f(x|Y=1), f(x|Y=0).

Gaussian and Linear Classifiers

Take the conditional X|Y to be Multivariate Normal.

This leads to a *quadratic decision boundary*. If the covariance matrices for the two groups are assumed equal, then get a *linear decision boundary*.

Naive Bayes

Naive conditional independence assumption:

$$f_j(x_1, \ldots, x_d) = f_{j1}(x_1) f_{j2}(x_2) \ldots f_{jd}(x_d)$$

Often unrealistic, but still may be *useful* esp. since it leads to a drastic reduction in the number of parameters to estimate.

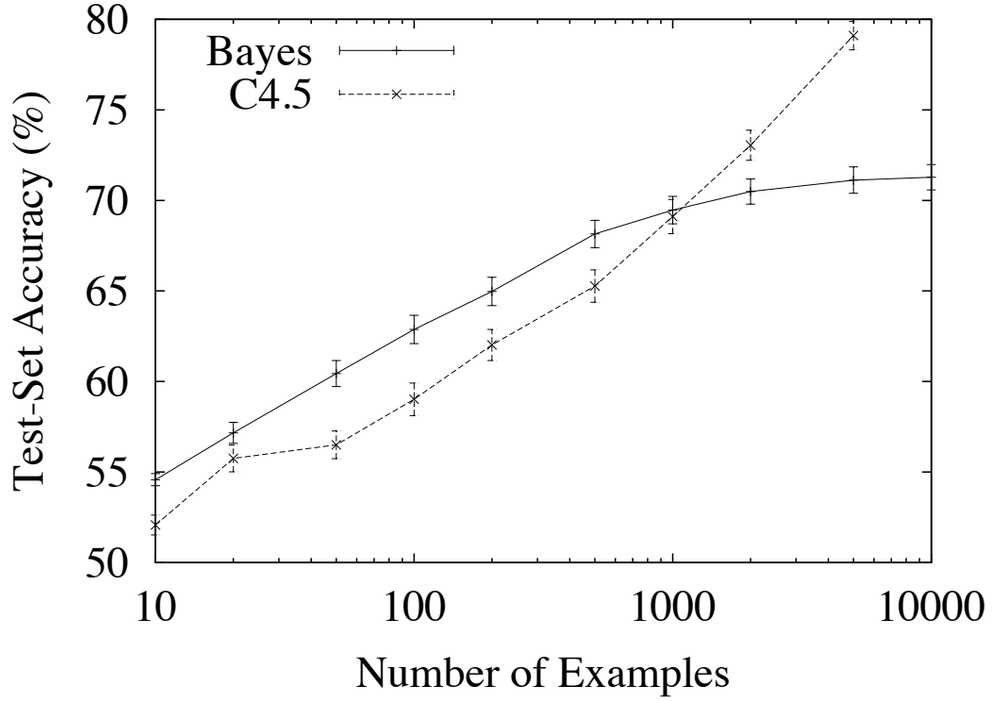
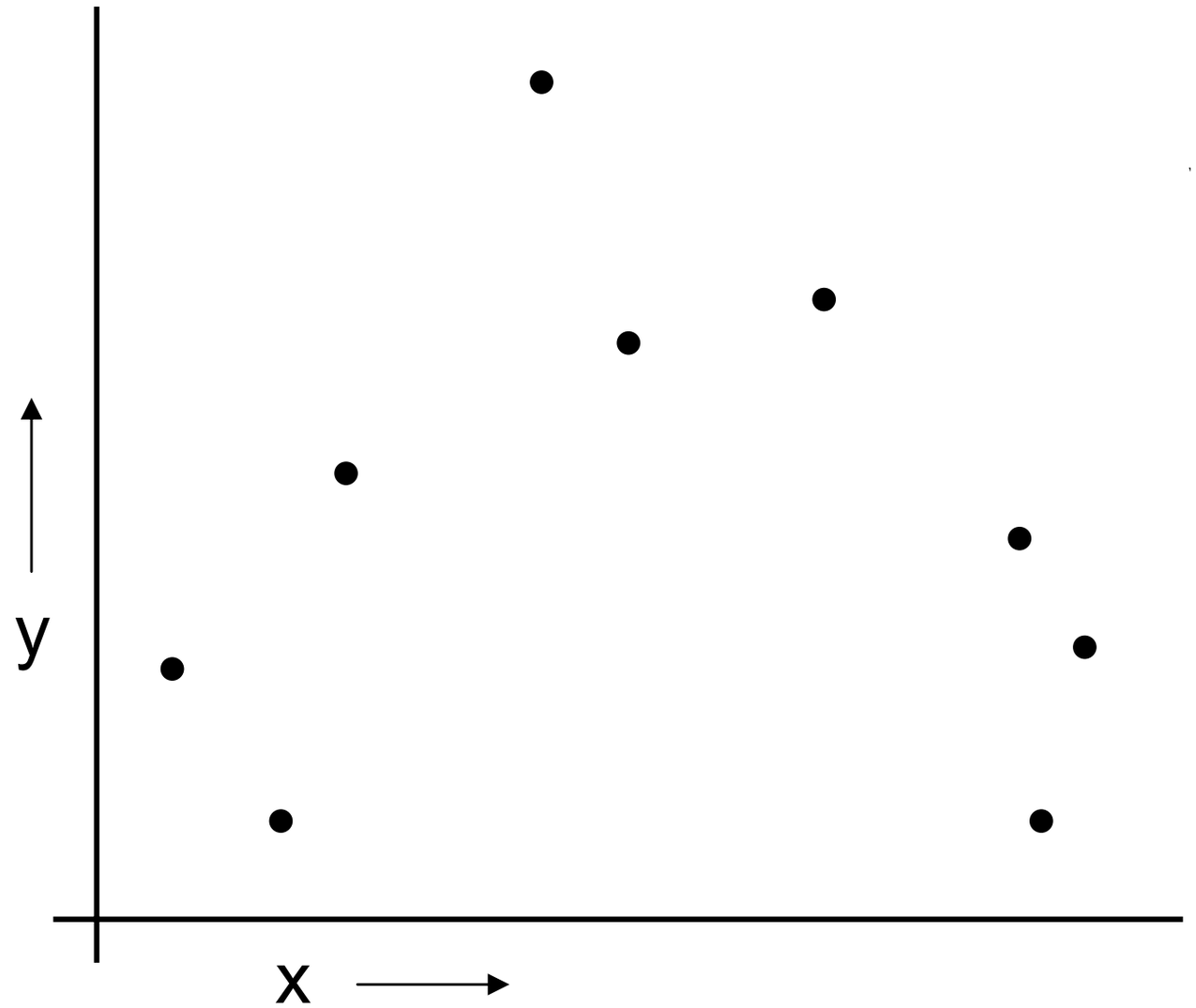


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules.

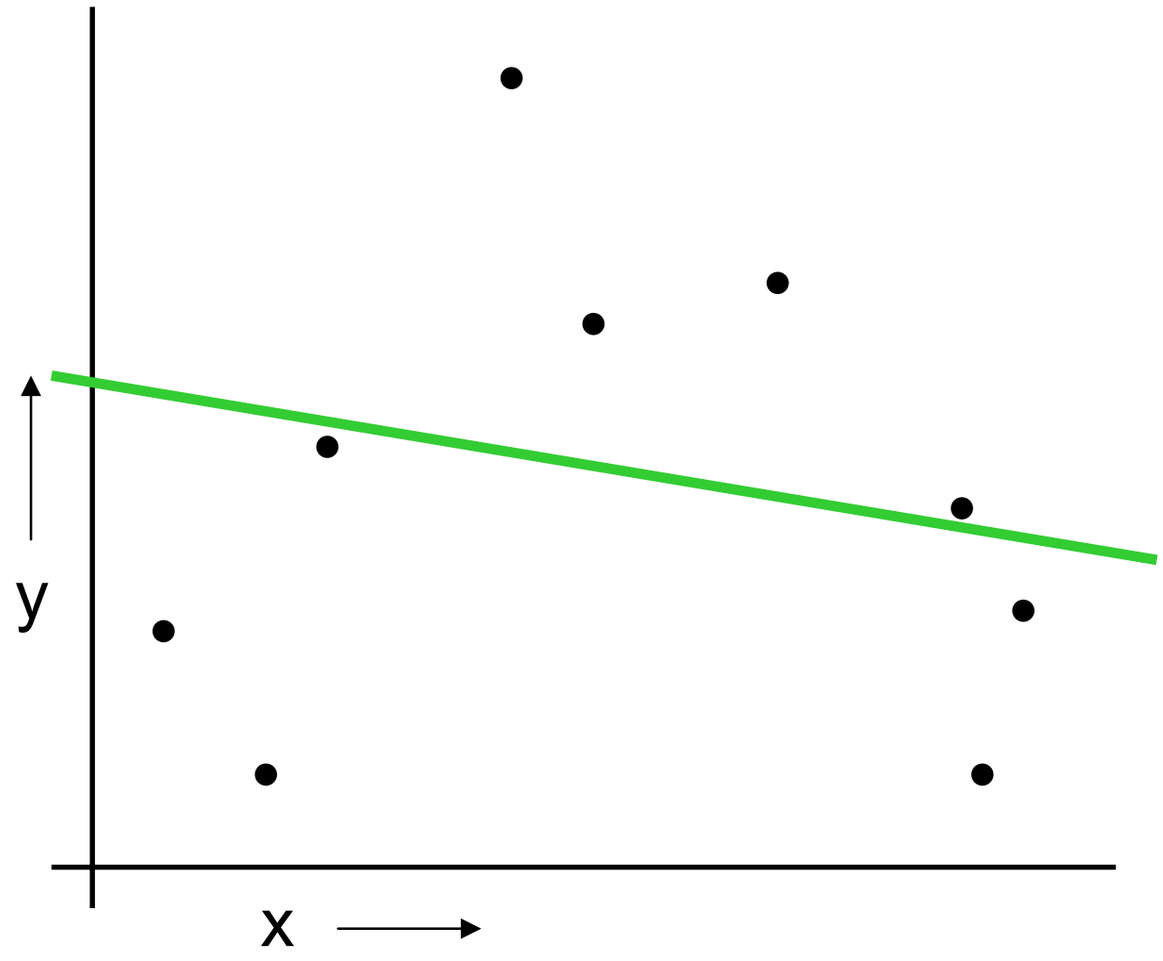
Domingos, http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

What kind of regression model should we use?



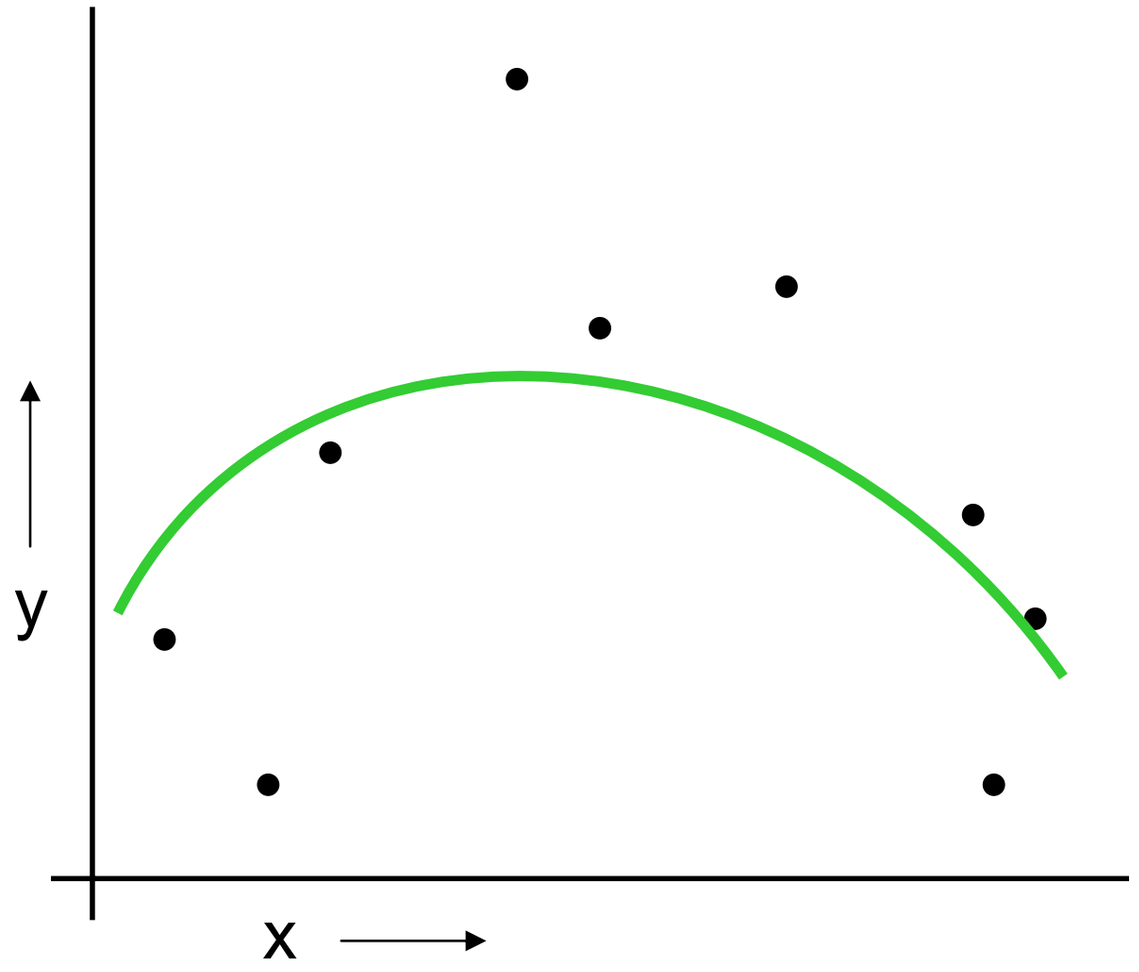
Moore, www.cs.cmu.edu/~awm/tutorials

Linear in x



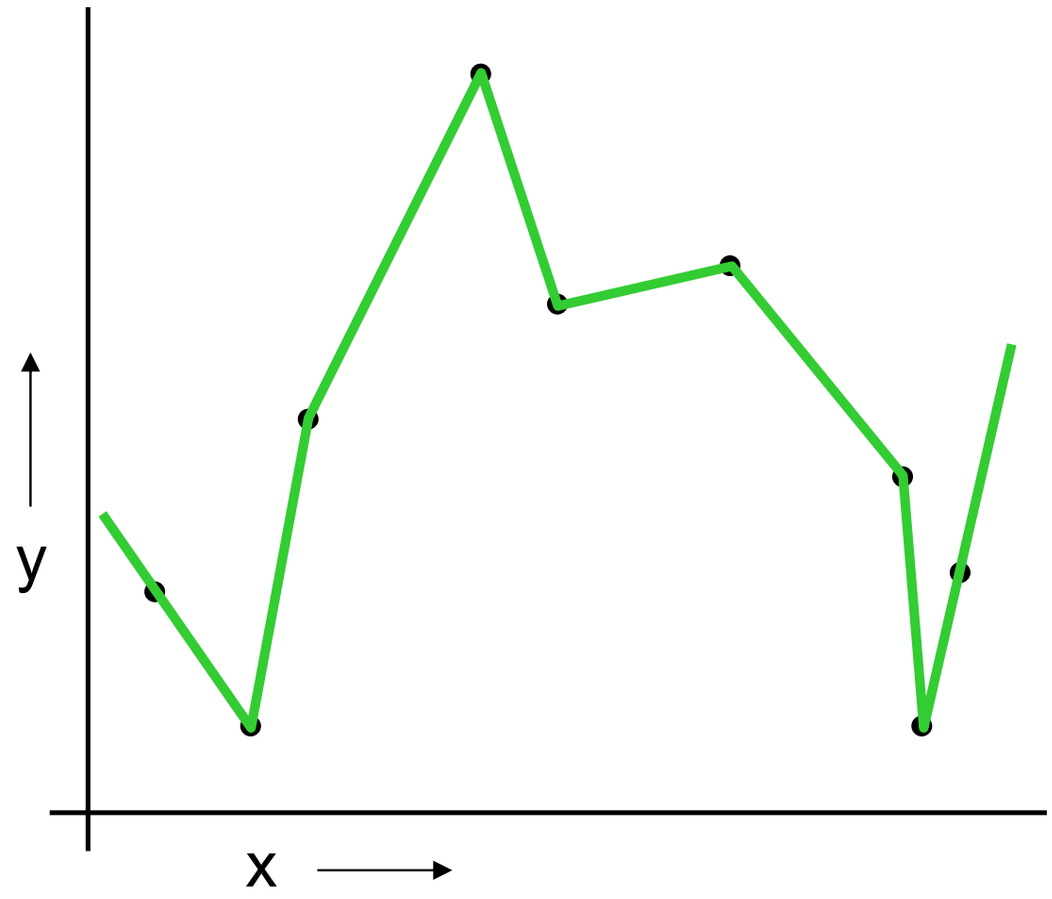
Moore, www.cs.cmu.edu/~awm/tutorials

Quadratic in x

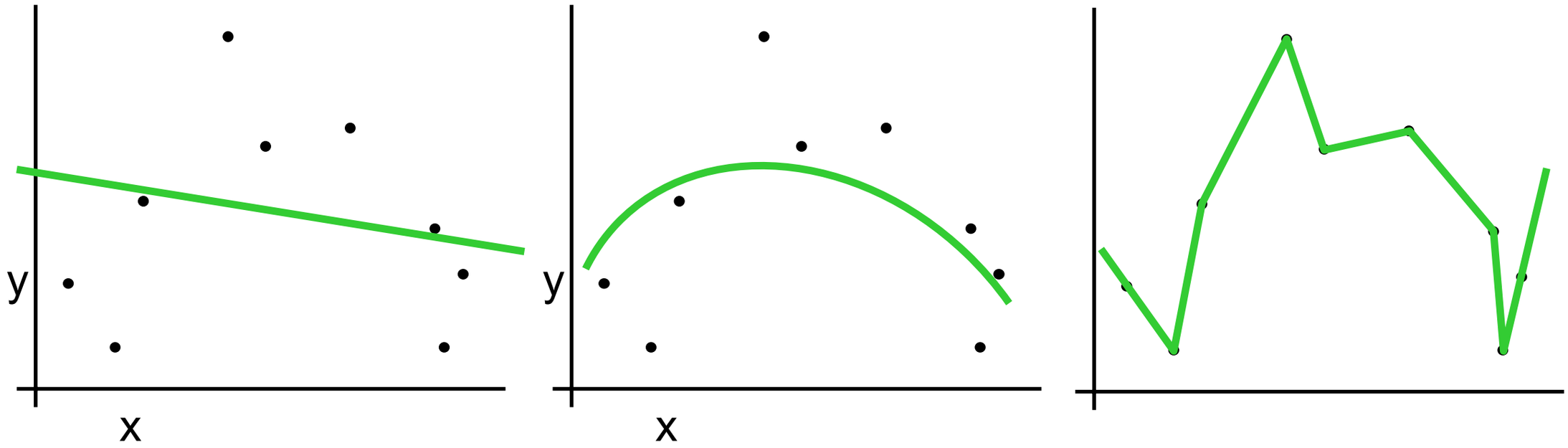


Moore, www.cs.cmu.edu/~awm/tutorials

Connect the dots

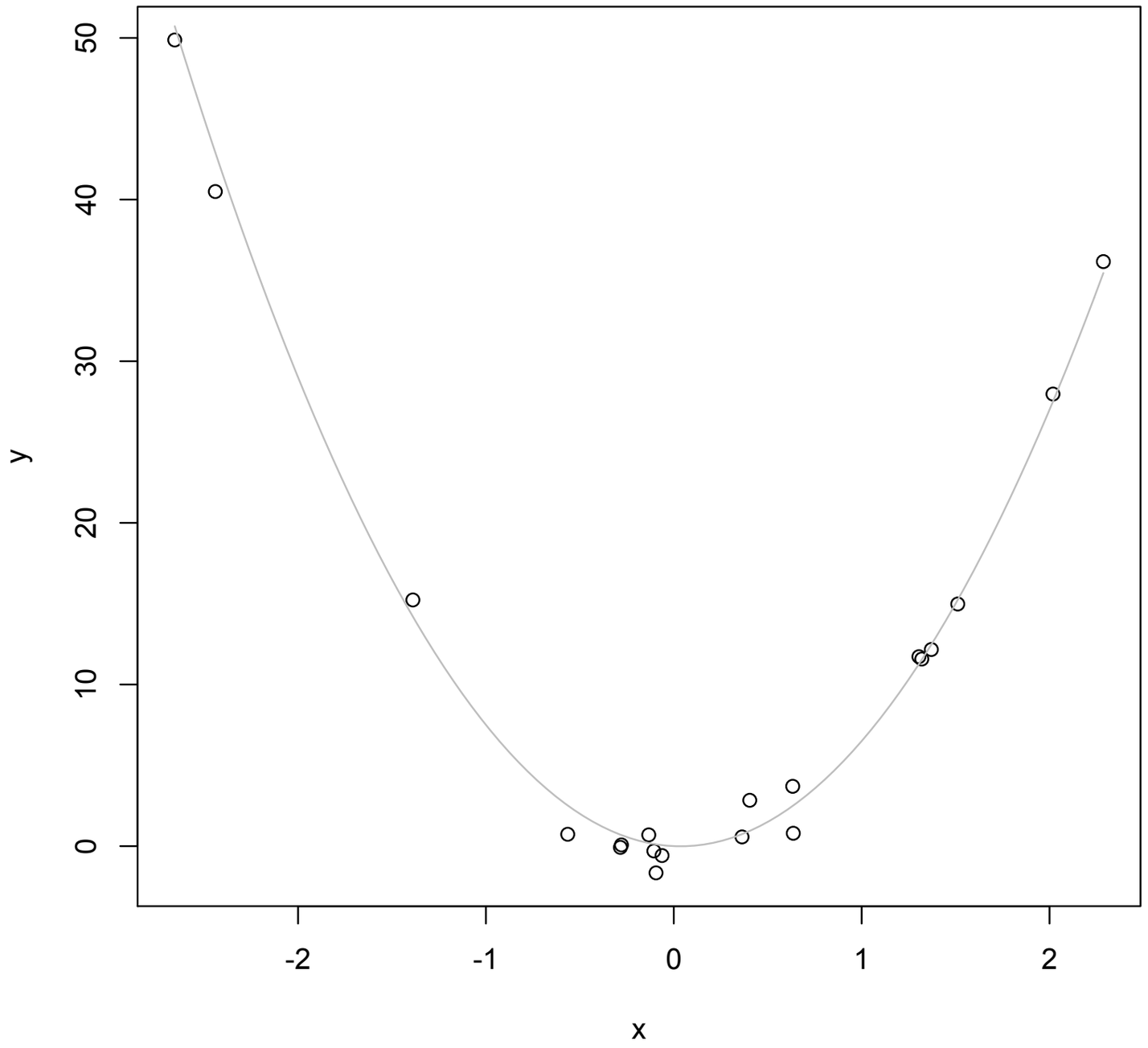


Moore, www.cs.cmu.edu/~awm/tutorials



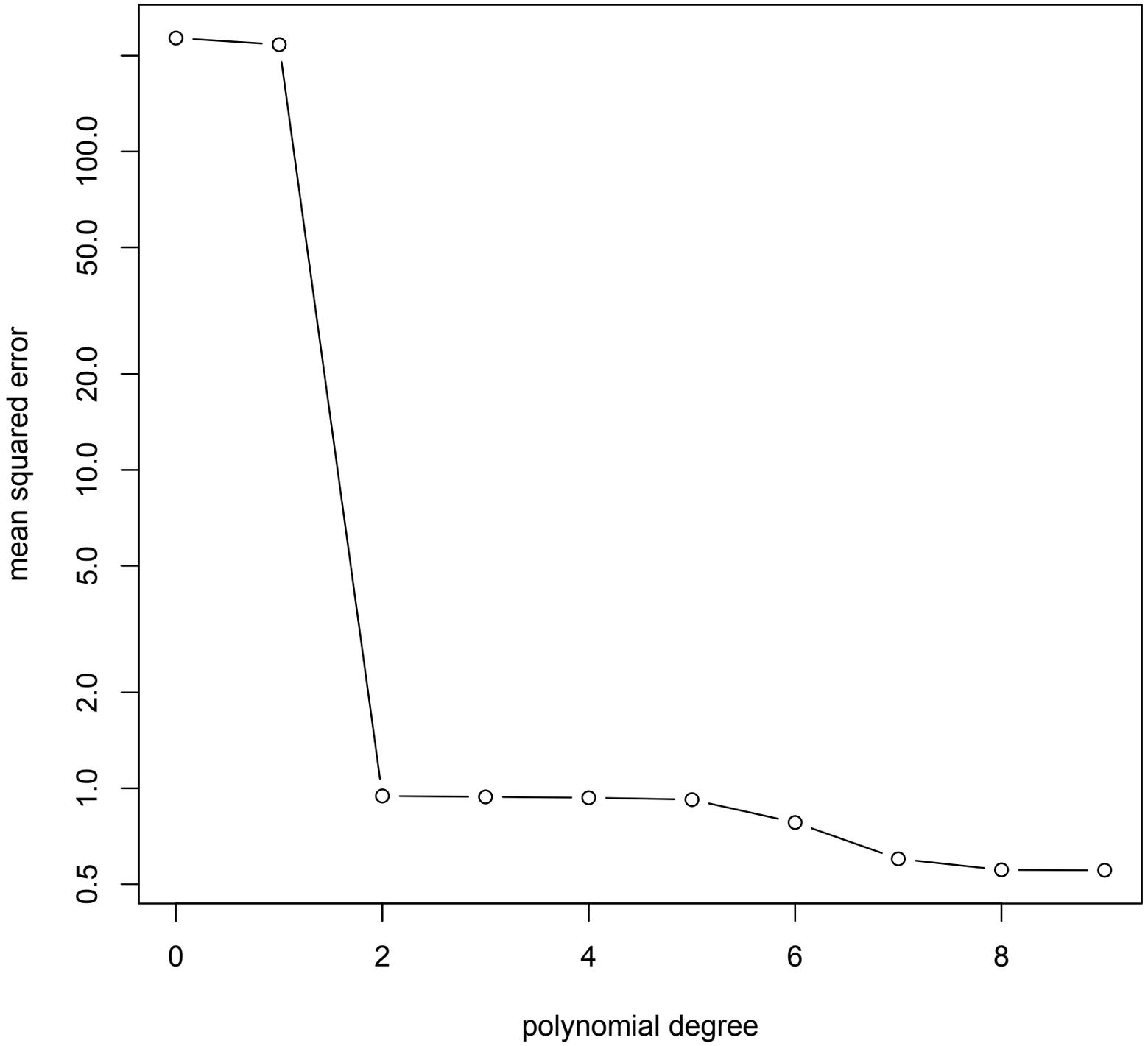
Moore, www.cs.cmu.edu/~awm/tutorials

Underfitting vs. Overfitting



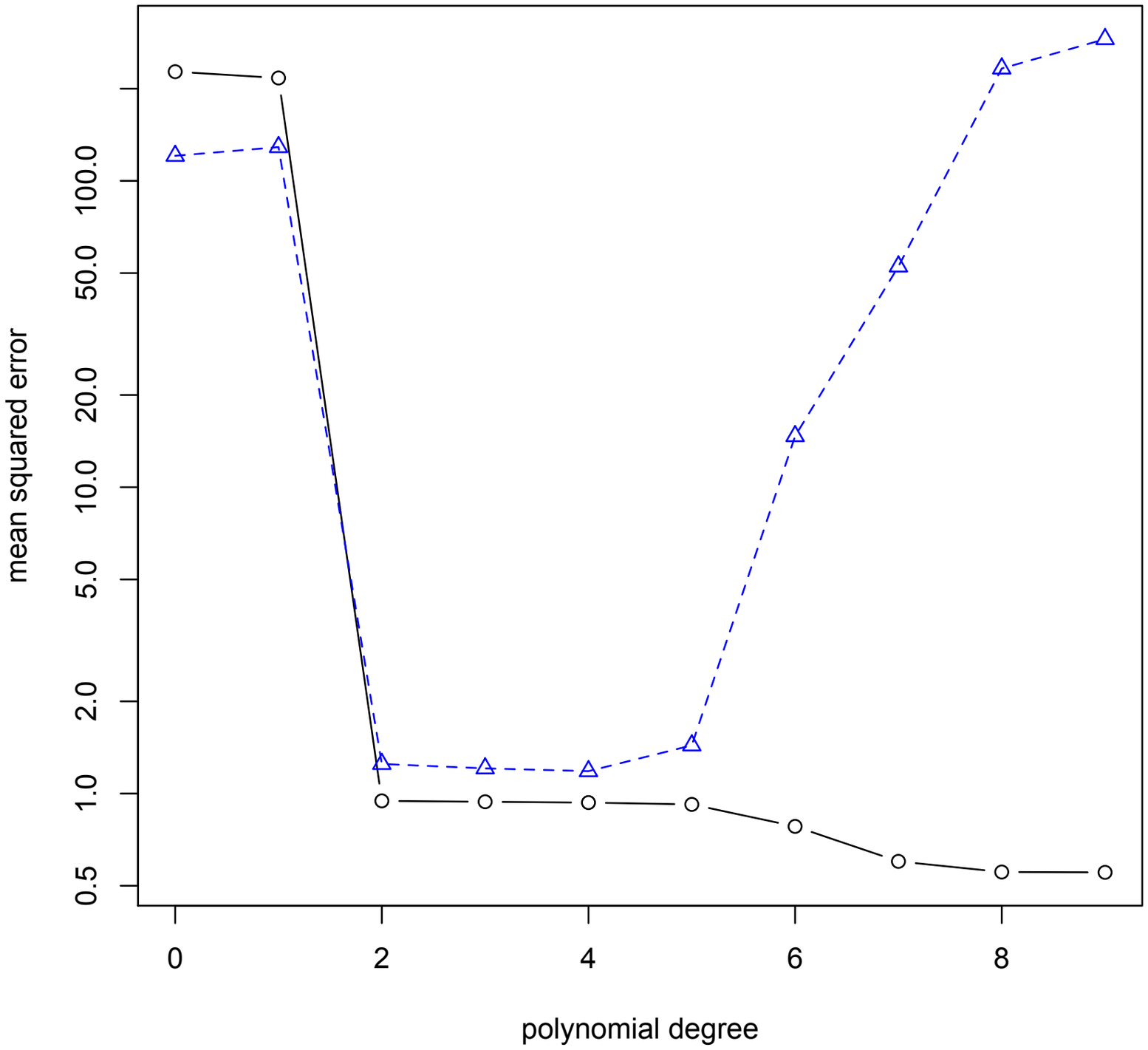
Shalizi, http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf

In-Sample MSE



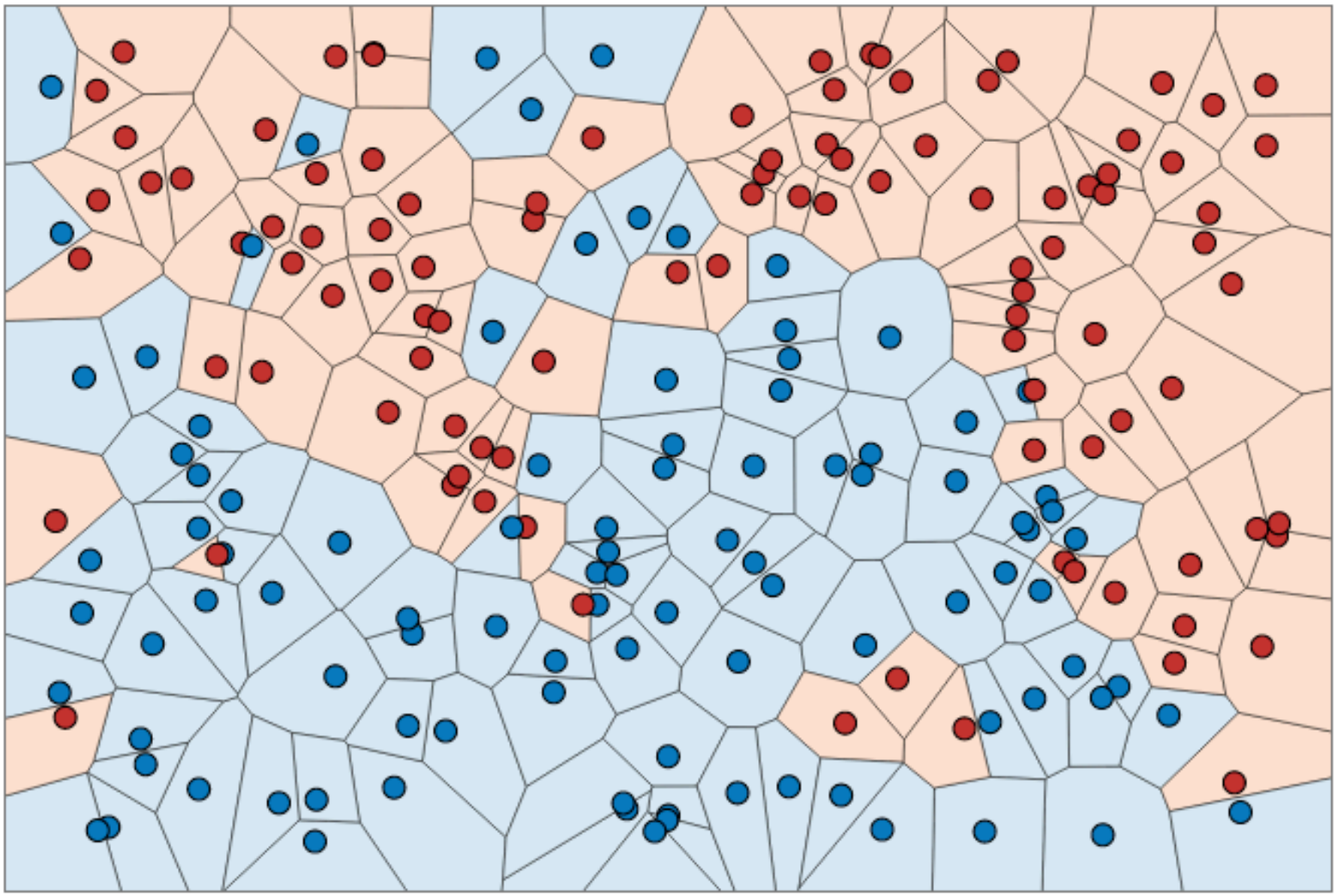
Shalizi, http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf

Out-Of-Sample MSE



Shalizi, http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf

kNN (k Nearest Neighbors)



http://scott.fortmann-roe.com/docs/BiasVariance.html

Choice of k is another bias-variance tradeoff.

kNN in Collaborative Filtering

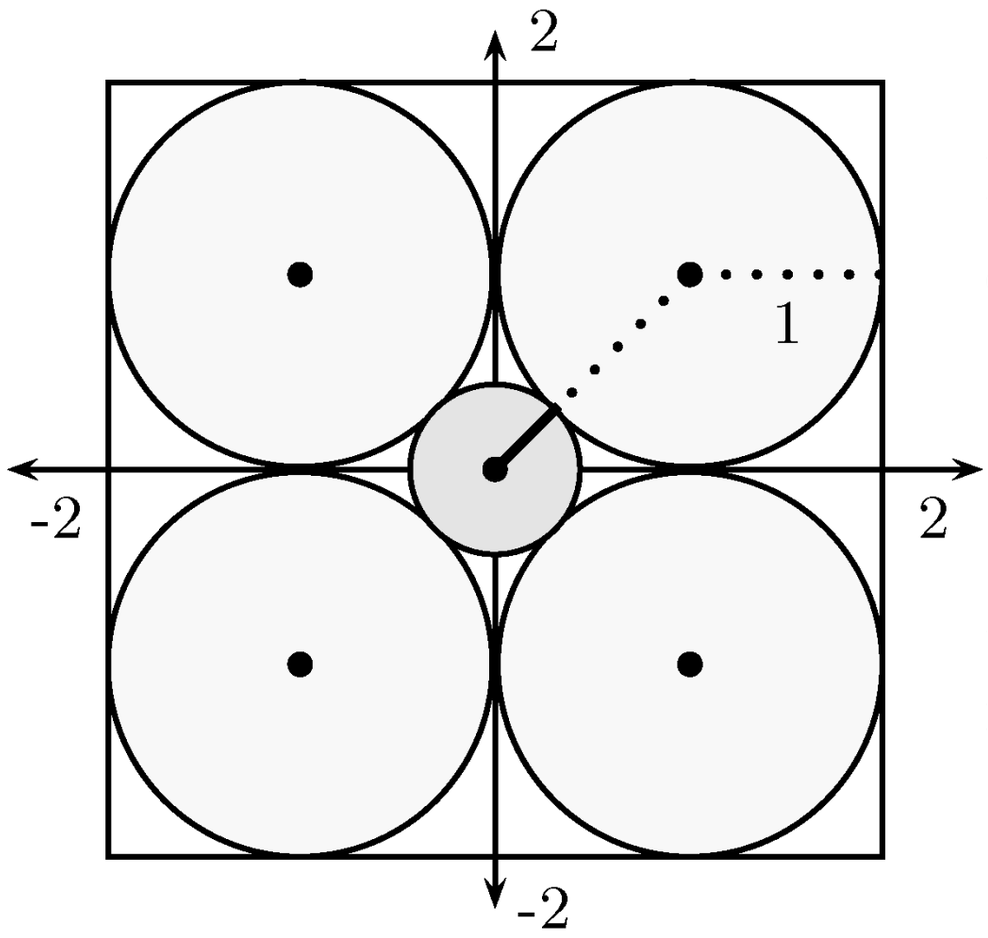
The most common tool for recommendation systems when the Netflix Prize began, and remained an integral tool of most of the successful teams.

$$\hat{r}_{ui} = \frac{\sum_{j \in N(i;u)} s_{ij} r_{uj}}{\sum_{j \in N(i;u)} s_{ij}}$$

Both user-oriented and item-oriented versions are useful.

How should k be chosen? How should the weights be chosen?

A Geometry Puzzle



In $\mathbb{R}^2$, one places a unit circle in each quadrant of the square $[-2,2]^2$.

A non-overlapping circle of maximal radius is then centered at the origin.

Fig. 4.1. This arrangement of $5 = 2^2 + 1$ circles in $[-2,2]^2$ has a natural generalization to an arrangement of $2^d + 1$ spheres in $[-2,2]^d$. This general arrangement then provokes a question which a practical person might find perplexing — or even silly. Does the central sphere stay inside the box $[-2,2]^d$ for all values of d?

Steele, *The Cauchy-Schwarz Master Class*

For 10 dimensions and higher, it extends outside the box. In fact, as dimension increases the % of its volume inside the box goes to 0.

Curse of Dimensionality

For n indep. Unif(-1,1) r.v.s, what is the probability that the random vector is in the unit ball?

n	probability
2	0.79
3	0.52
6	0.08
10	0.002
15	0.00001

In many high-dimensional settings, the vast majority of data will be near the boundaries, not in the center.

Blessing of Dimensionality

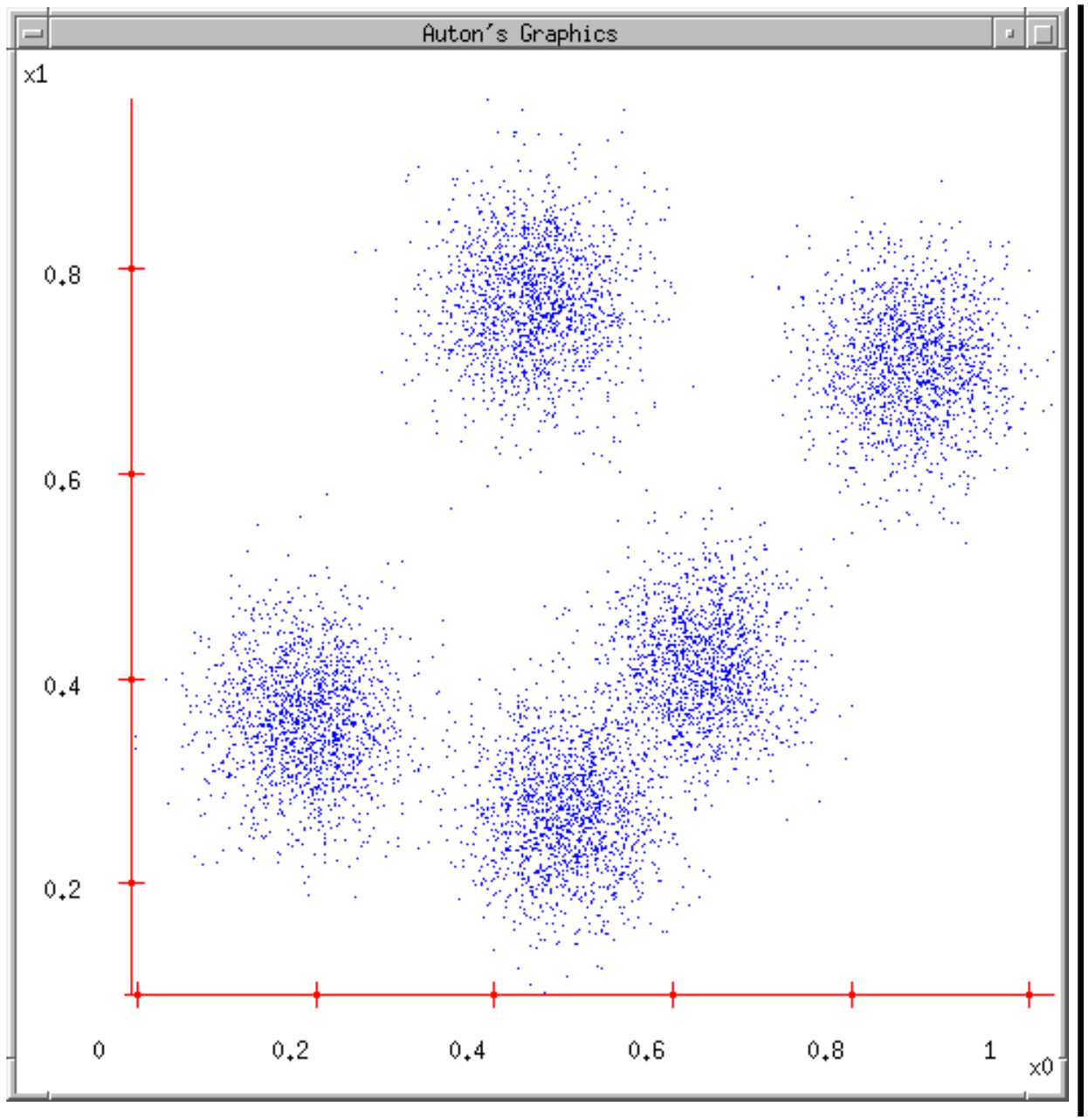
In statistics, “curse of dimensionality” is often used to refer to the difficulty of fitting a model when many possible predictors are available. But this expression bothers me, because more predictors is more data, and it should not be a “curse” to have more data....

With multilevel modeling, there is no curse of dimensionality. When many measurements are taken on each observation, these measurements can themselves be grouped. Having more measurements in a group gives us more data to estimate group-level parameters (such as the standard deviation of the group effects and also coefficients for group-level predictors, if available).

In all the realistic “curse of dimensionality” problems I’ve seen, the dimensions–the predictors–have a structure. The data don’t sit in an abstract K-dimensional space; they are units with K measurements that have names, orderings, etc.

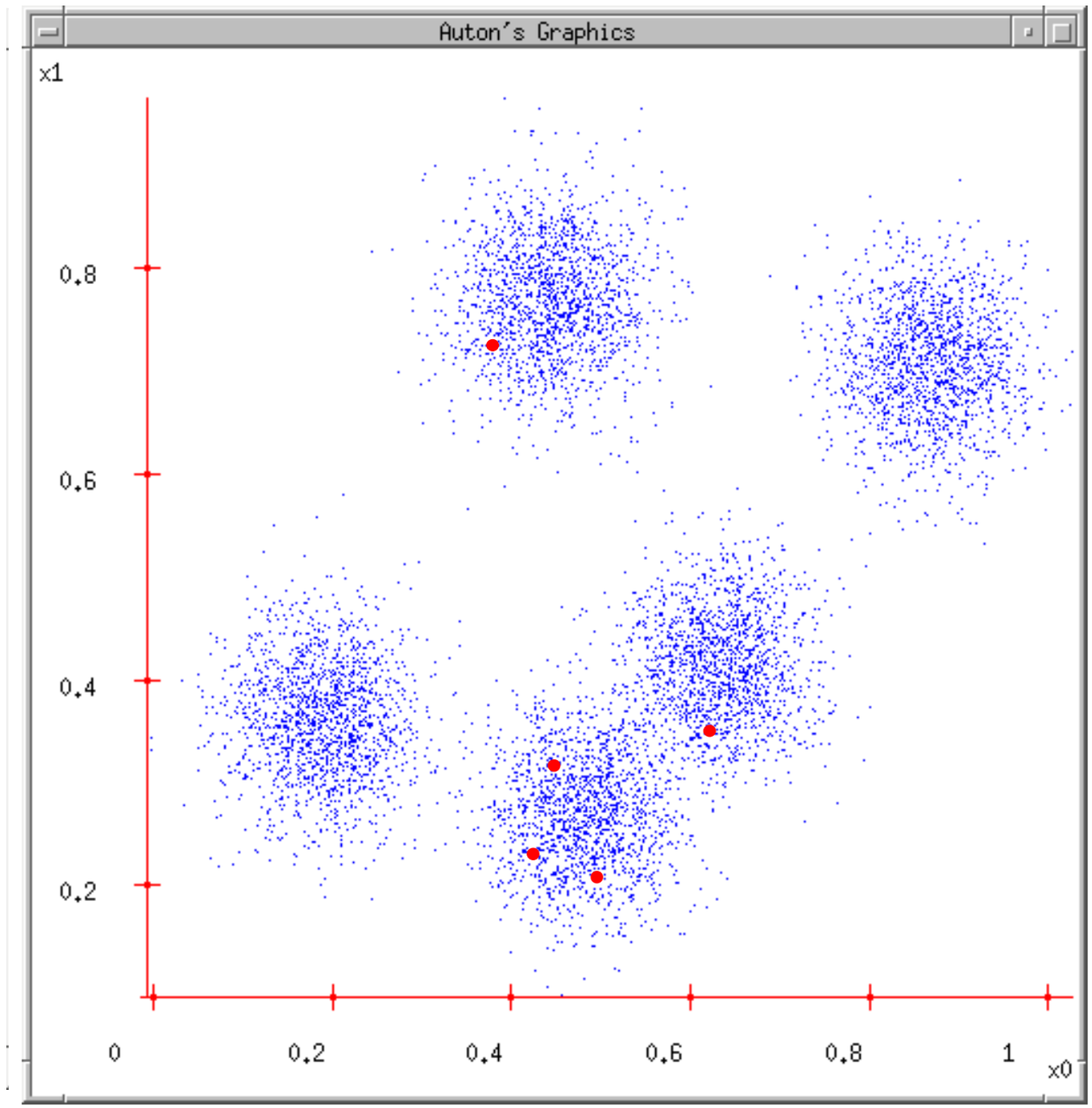
Gelman, http://andrewgelman.com/2004/10/27/the_blessing_of/

k-means Clustering



Moore, www.cs.cmu.edu/~awm/tutorials

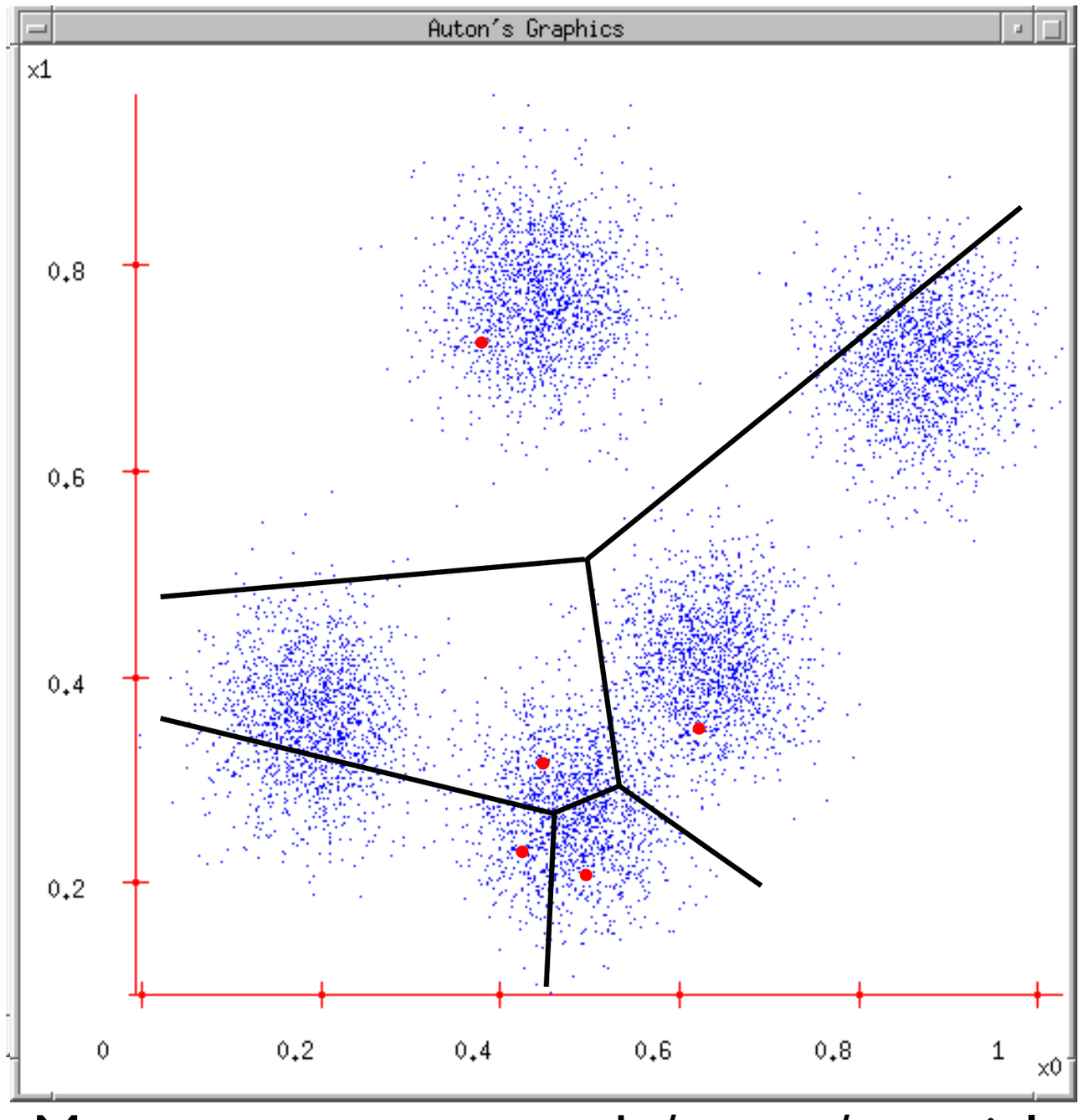
k-means Clustering



Moore, www.cs.cmu.edu/~awm/tutorials

guess cluster means

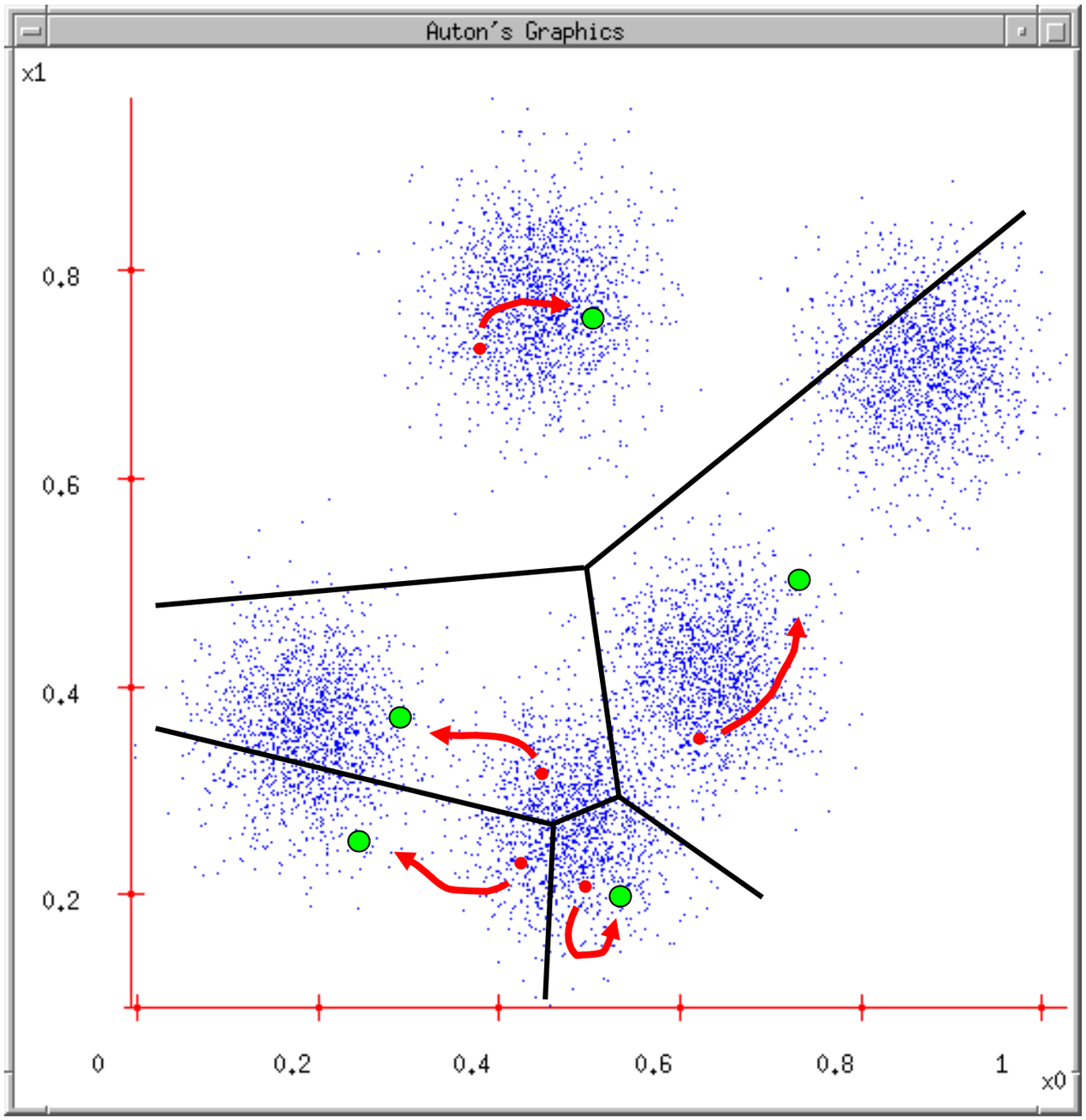
k-means Clustering



Moore, www.cs.cmu.edu/~awm/tutorials

each cluster mean takes responsibility for the data closest to it

k-means Clustering

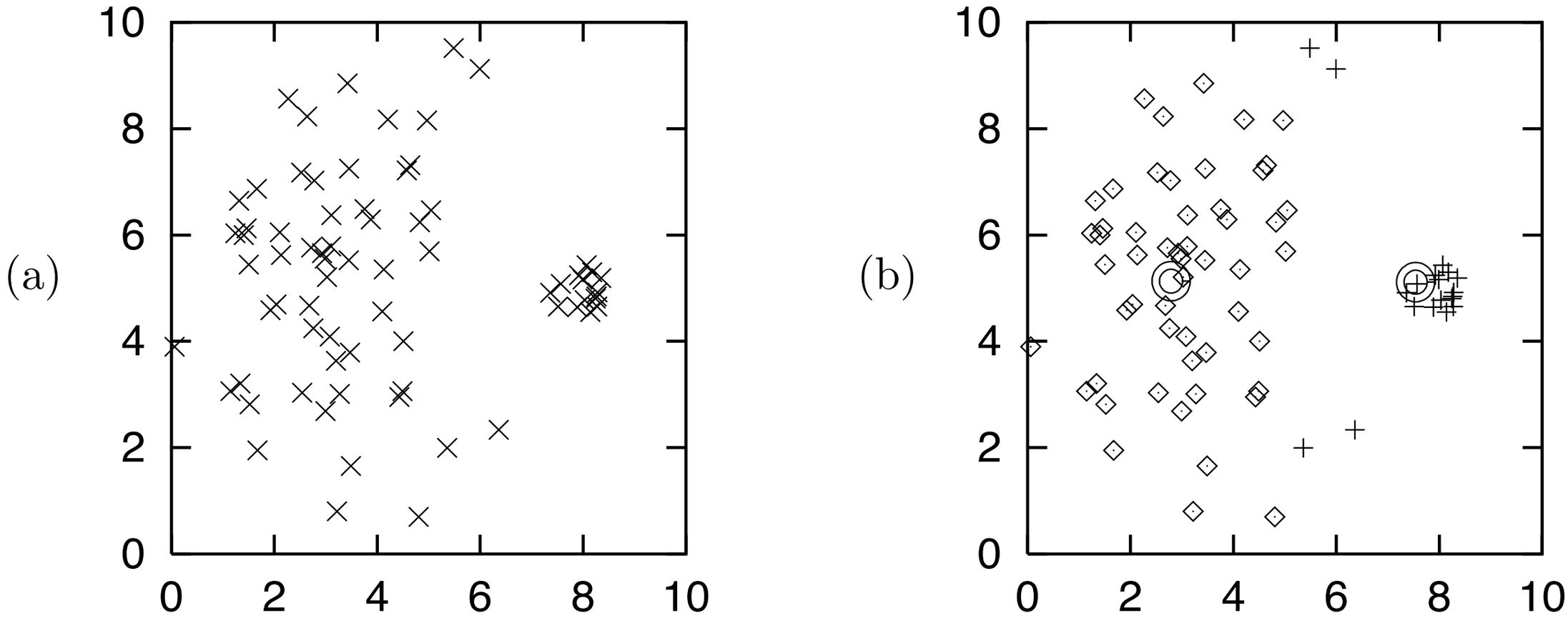


Moore, www.cs.cmu.edu/~awm/tutorials

recompute and iterate

K-means issues

number of clusters?
initial guess?
hard clustering vs. soft clustering?
non-Multivariate Normal looking shapes?



MacKay, http://www.inference.phy.cam.ac.uk/itila/Potter.html