UCB - CS189
Introduction to Machine Learning
Fall 2015
Lecture 9: Performance evaluation

Isabelle Guyon

ChaLearn

Come to my office hours…

Wed 2:30-4:30 Soda 329

Last time

Kernel machines

PARAMETRIC (Perceptons)

$f(\mathbf{x}) = \mathbf{w} \bullet \Phi(\mathbf{x})$

$\mathbf{w} = \sum_k \alpha_k \Phi(\mathbf{x}^k)$

(Large margin) Perceptron

$\Delta\mathbf{w} \sim y_k \Phi(\mathbf{x}^k)$ if $y_k f(\mathbf{x}^k) < 1$

$\sim \mathbf{1}(1-z_k)\, y_k \Phi(\mathbf{x}^k)$ $z_k = y_k f(\mathbf{x}^k)$

(Rosenblatt 1958)

Logistic regression

$\Delta\mathbf{w} \sim S(-z_k)\, y_k \Phi(\mathbf{x}^k)$

(Cox 1958)

LMS regression or classification

$\Delta\mathbf{w} \sim (y_k - f(\mathbf{x}^k))\, \Phi(\mathbf{x}^k) \sim (1 - z_k)\, y_k \Phi(\mathbf{x}^k)$

(Widrow-Hoff, 1960)

NON PARAMETRIC (Kernel machines)

$f(\mathbf{x}) = \sum_k \alpha_k k(\mathbf{x}^k, x)$

$k(\mathbf{x}^k, \mathbf{x}) = \Phi(\mathbf{x}^k).\Phi(\mathbf{x})$

Potential Function algorithm

$\Delta\alpha_k \sim y_k$ if $y_k f(\mathbf{x}^k) < 1$

$\sim \mathbf{1}(1-z_k)\, y_k$

(Aizerman et al 1964)

Dual logistic regression

$\Delta\alpha_k \sim S(-z_k)\, y_k$

Dual LMS

$\Delta\alpha_k \sim (y_k - f(\mathbf{x}^k)) \sim (1 - z_k)\, y_k$

4

Come to my office hours…
Wed 2:30-4:30 Soda 329

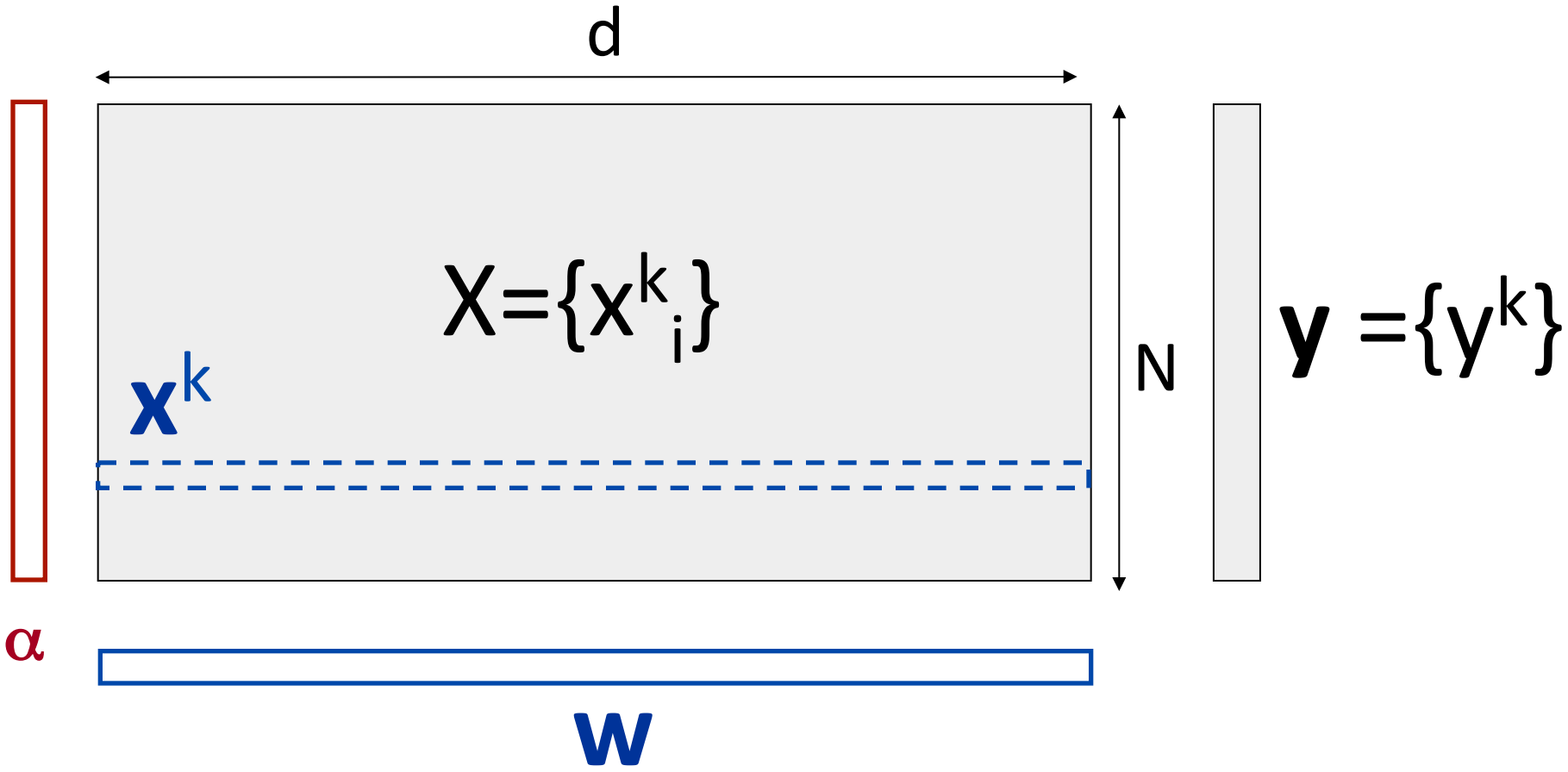
Today

Math prerequisites

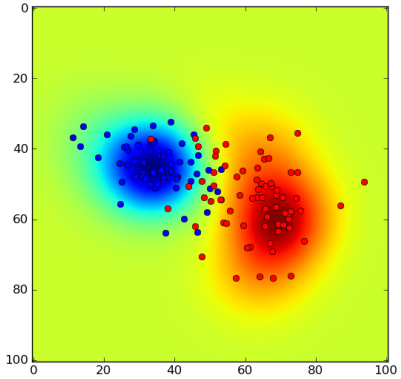
- Bernouilli distribution
- Binomial law
- Statistical tests

Changing roles

You have formulated your problem as a “machine Learning” problem



Main question:



How well do models generalize on new test data?

The true generalization error is the expected risk
$R[f] = \int L(f(\mathbf{x}, \mathbf{w}), y)\, dP(\mathbf{x}, y)$

But you do not know $P(\mathbf{x}, y)$, you can only get n TEST examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)$, distinct from the N training examples.

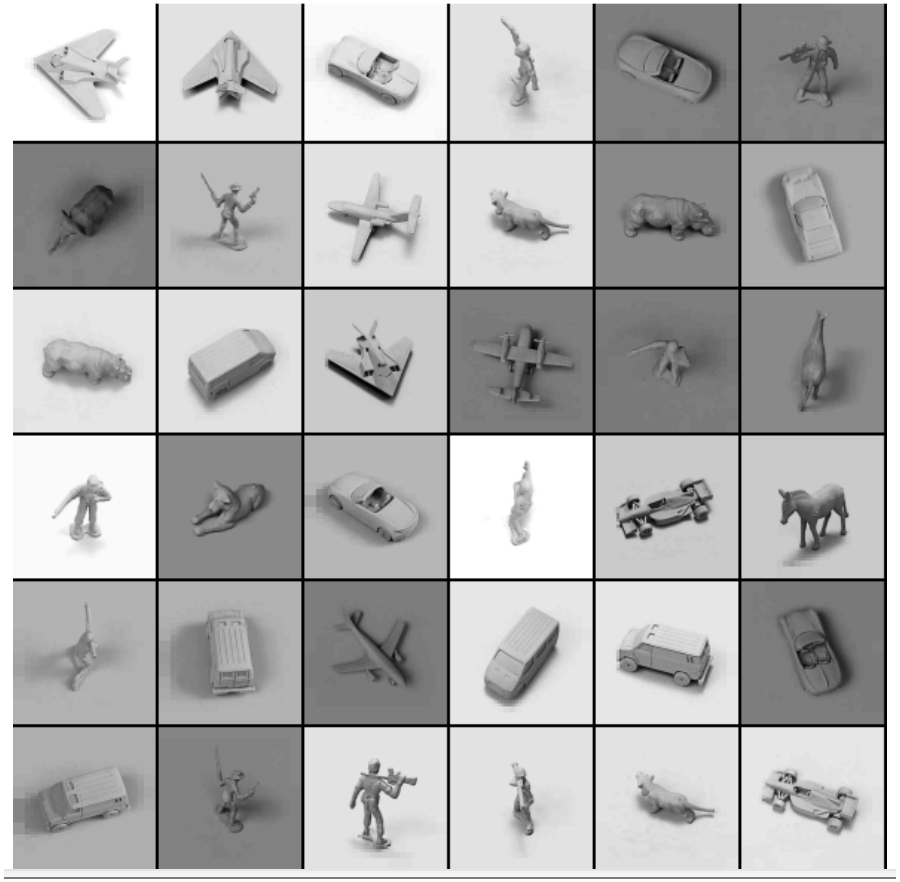
$R_{test}[f] = (1/n) \sum_{k=1:n} L(f(\mathbf{x}^k), y^k)$

Give them good data!

- Are your data I.I.D.?
- Split your data into training and test set.
- Do NOT release the target values y for the test data.
- Beware of data leakage
  - Sample IDs
  - Sample order
  - Confounding factors

Example of confounding

Have a good study design



- Control.
- Block.
- Randomize.

http://www.cs.nyu.edu/~ylclab/data/norb-v1.0/

Your "staff" has taken CS 189/289

- They know about **linear and kernel methods**:

$f(\mathbf{x}) = \mathbf{w} \bullet \Phi(\mathbf{x}) = \sum_k \alpha_k \, k(\mathbf{x}^k, \mathbf{x})$

- They know **how to "train" them**:

$f^*(\mathbf{x}) = \text{argmin}_f \, R_{train}[f] + \lambda \Omega[f]$ $\quad \lambda > 0$

with $R_{train}[f] = (1/N) \sum_{k=1:N} L(f(\mathbf{x}^k), y^k)$

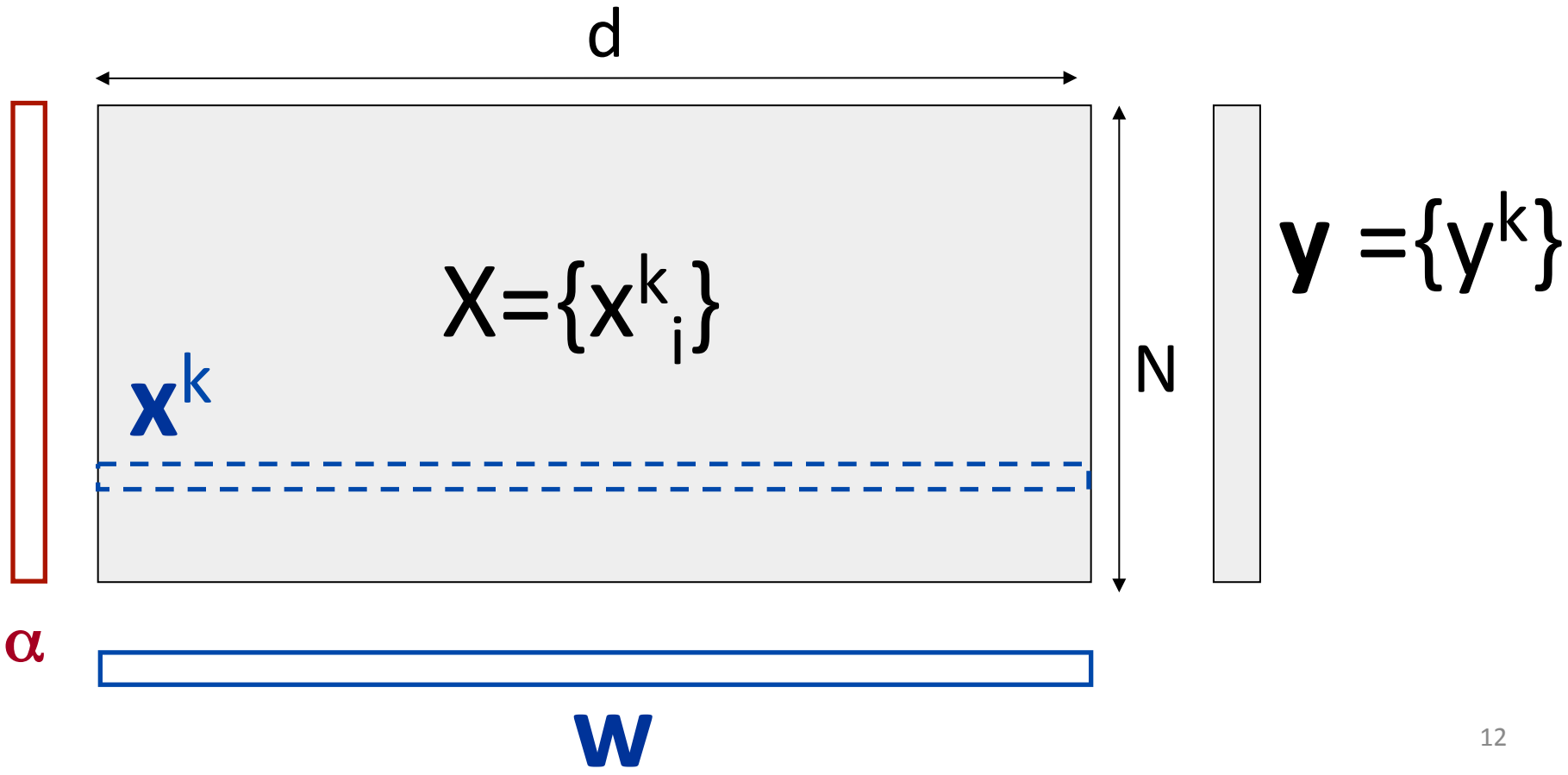
for a choice of loss functions L

(hinge loss, logistic loss, square loss)

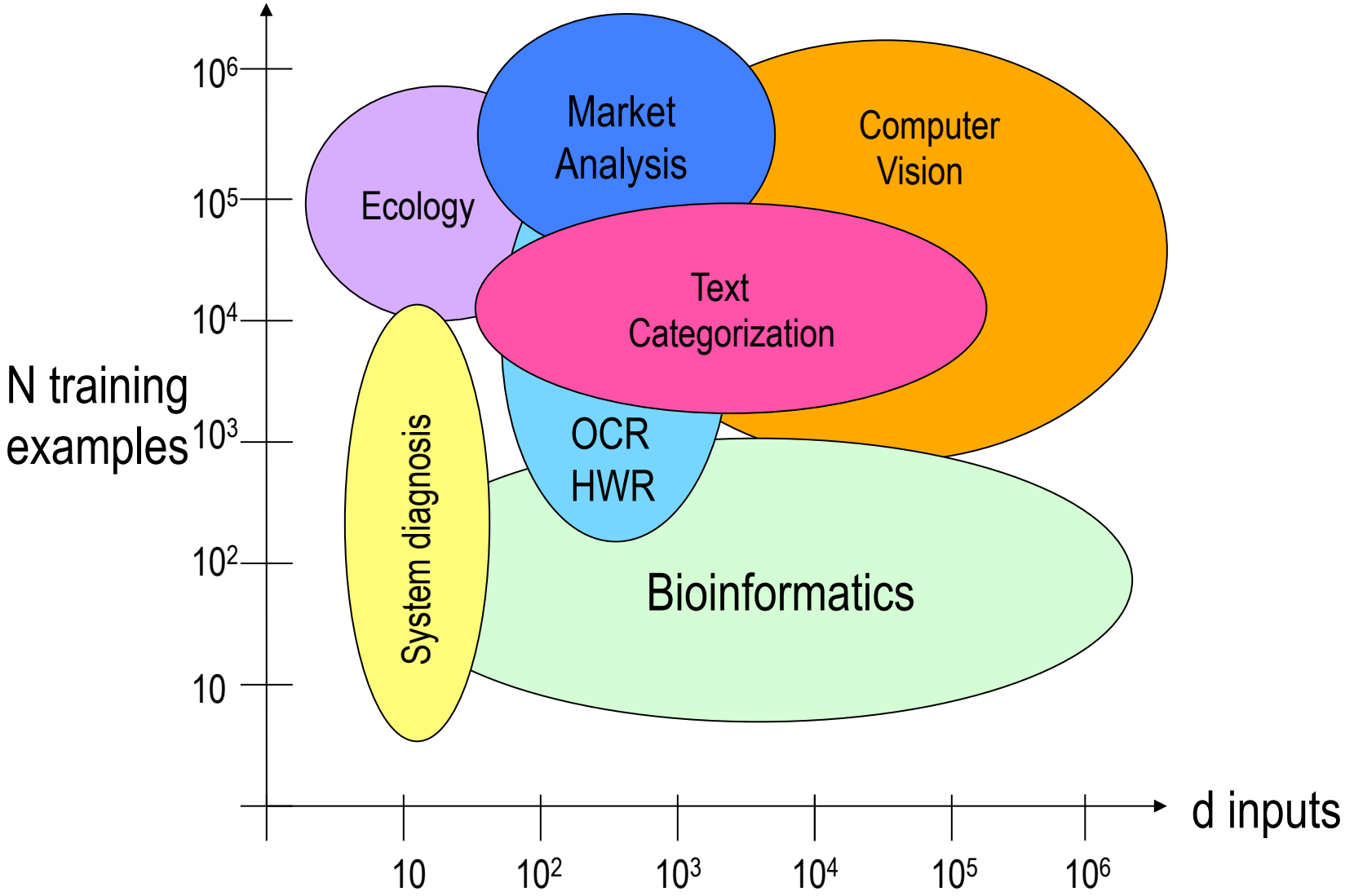
and a choice of regularizers Ω (such that $\|\mathbf{w}\|^2$)

Ask them to produce 3 models:

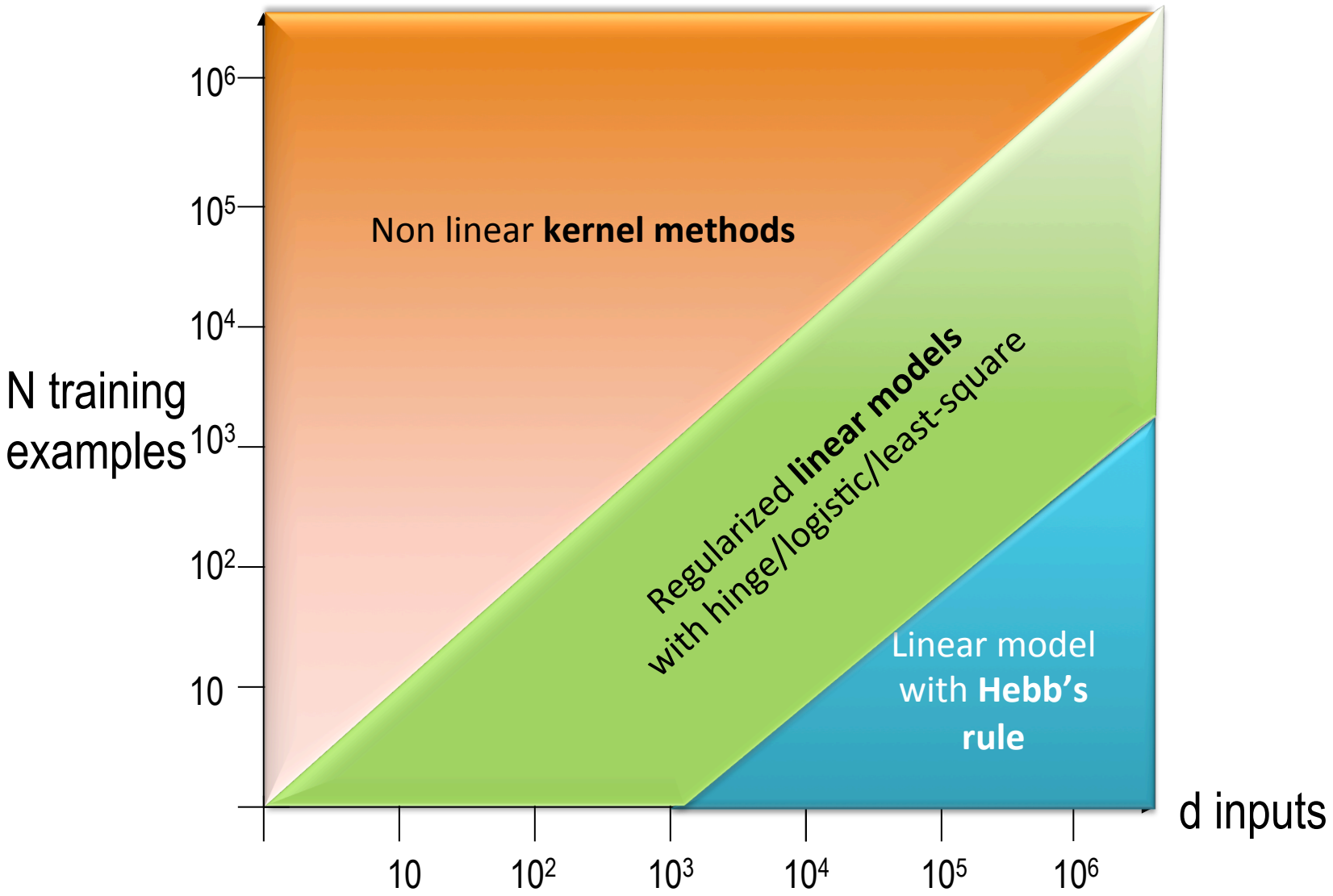
1) "Hebb's rule" linear model $f(\mathbf{x}) = \mathbf{w} \bullet \mathbf{x} = \sum_k y_k \mathbf{x}^k \bullet \mathbf{x}$
2) Regularized linear model $f(\mathbf{x}) = \mathbf{w} \bullet \mathbf{x} = \sum_k \alpha_k \mathbf{x}^k \bullet \mathbf{x}$
3) Non-linear model $f(\mathbf{x}) = \mathbf{w} \bullet \Phi(\mathbf{x}) = \sum_k \alpha_k k(\mathbf{x}^k, \mathbf{x})$



Can they always train 3 models?

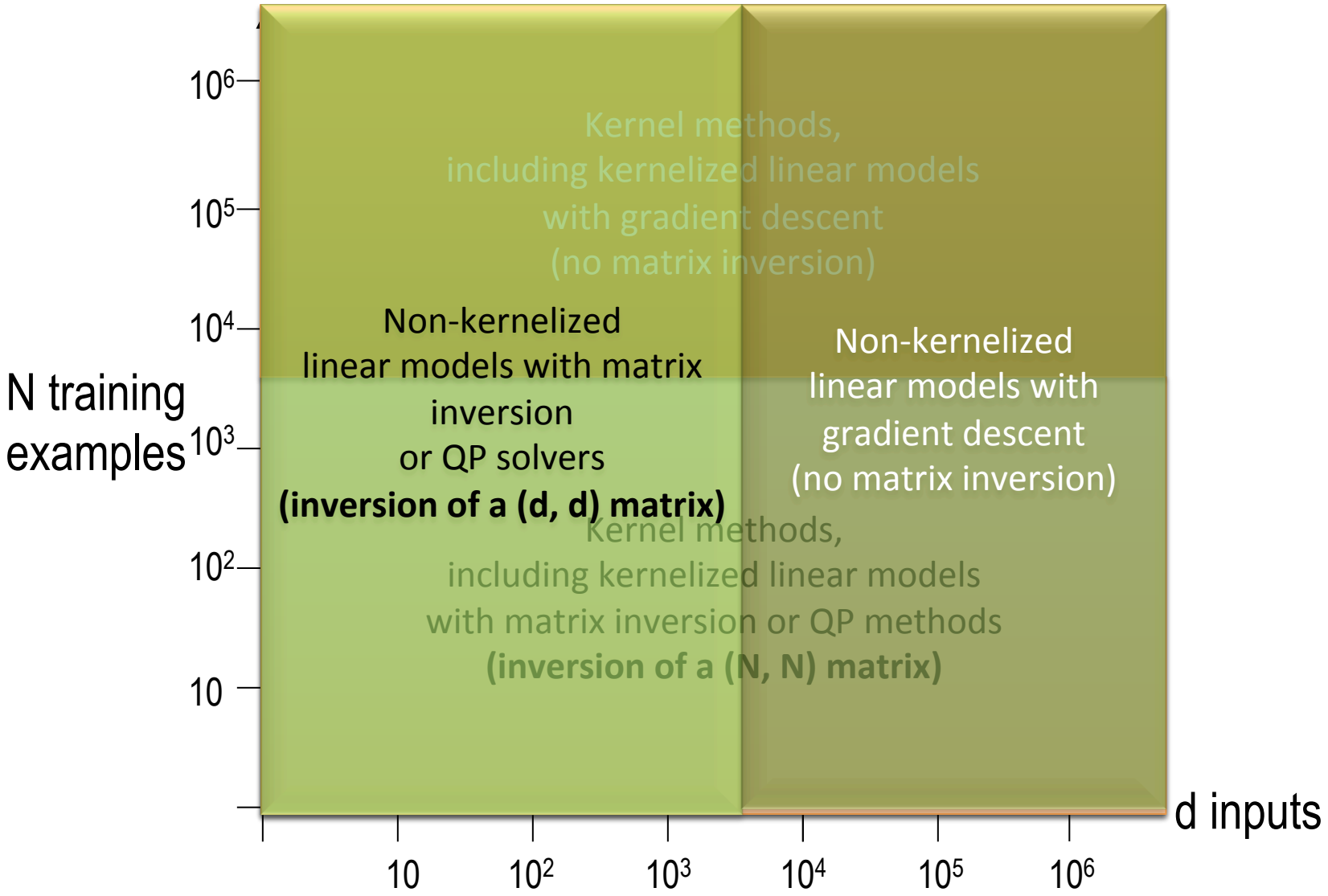


Statistical domains



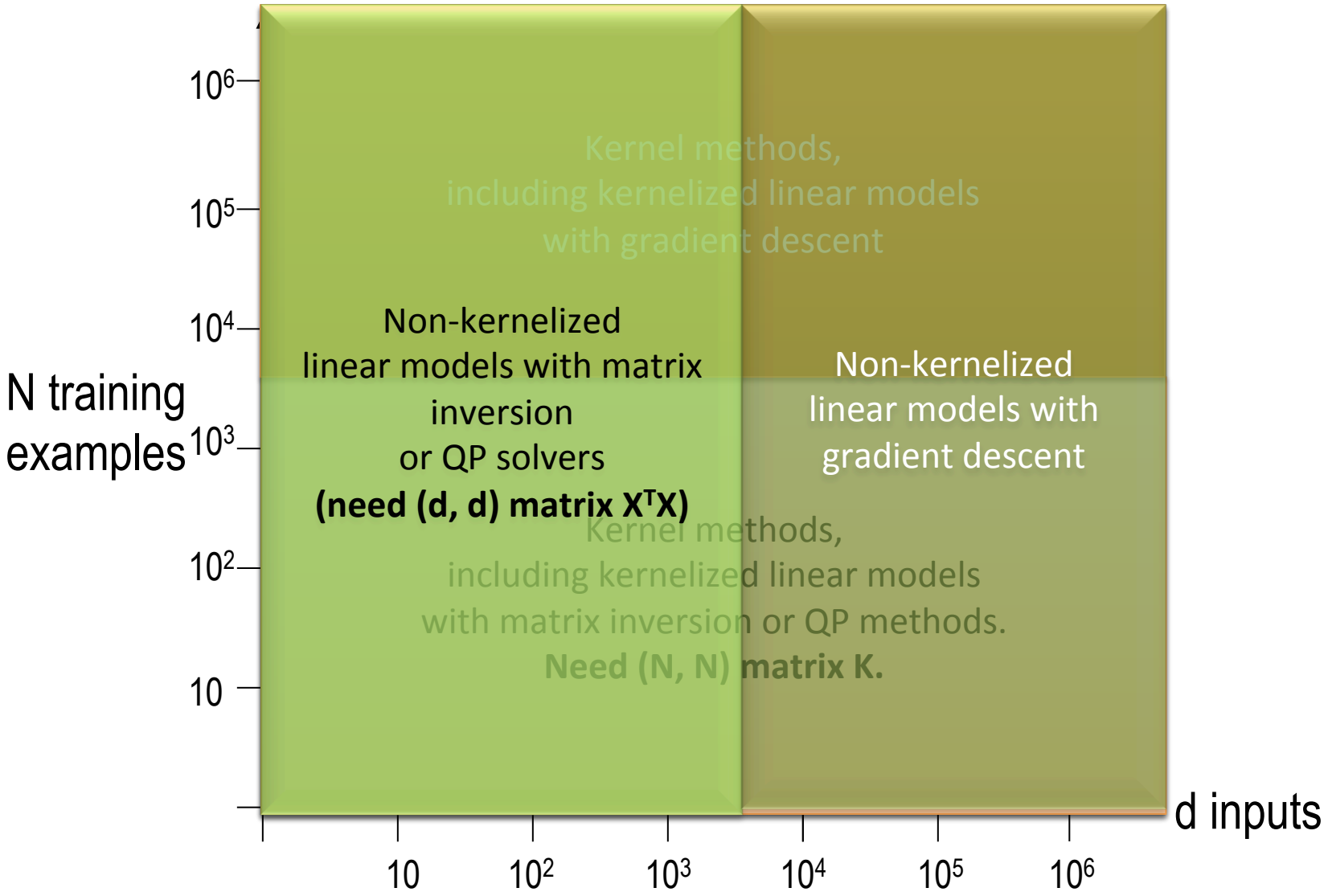
Computational domains

If you have to invert a matrix, either N or d must be small



Memory usage domains

Key: do we need to store the (N, N) matrix K or (d, d) matrix X^TX?



Other questions:

1) Should the loss function L be the same for training and testing?

2) How do I compute error bars?

3) How many test examples are needed to get “significant” results?

4) Can the test set be used to compare several models?

5) Should I continue the project or not?

6) Should I hire better data scientists or collect more data?

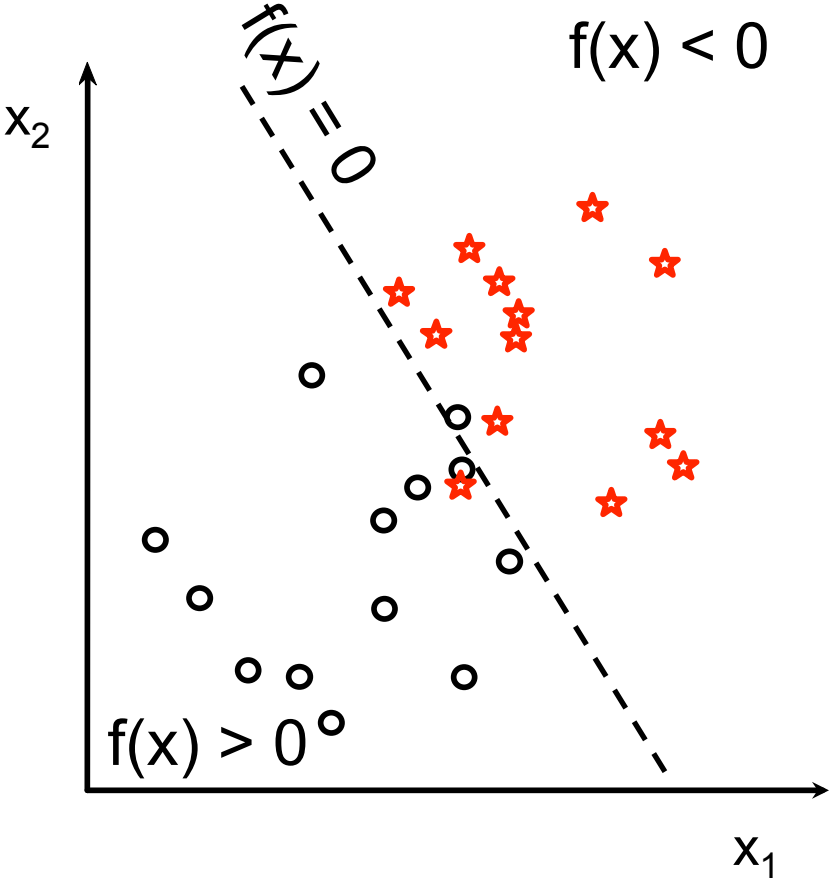
Cost functions to compute a "score"

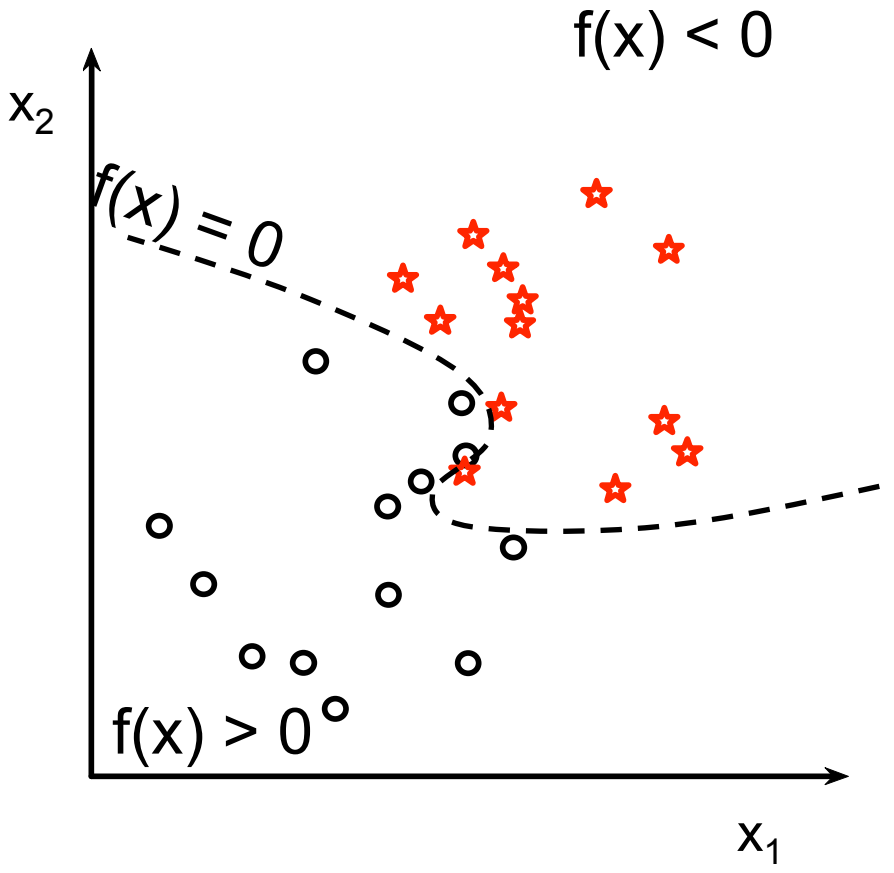
- Error rate E (or accuracy A = 1-E).
- Balanced Error Rate BER (or balanced accuracy BAC=1-BER).
- Cross-entropy H= $-\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right]$).
- A loss based on a cost matrix:

		Predicted		
		Class i	…	Class j
True	Class i	λ_{ii}	…	λ_{ij}
	⋮	⋮	⋱	⋮
	Class j	λ_{ji}	…	λ_{jj}

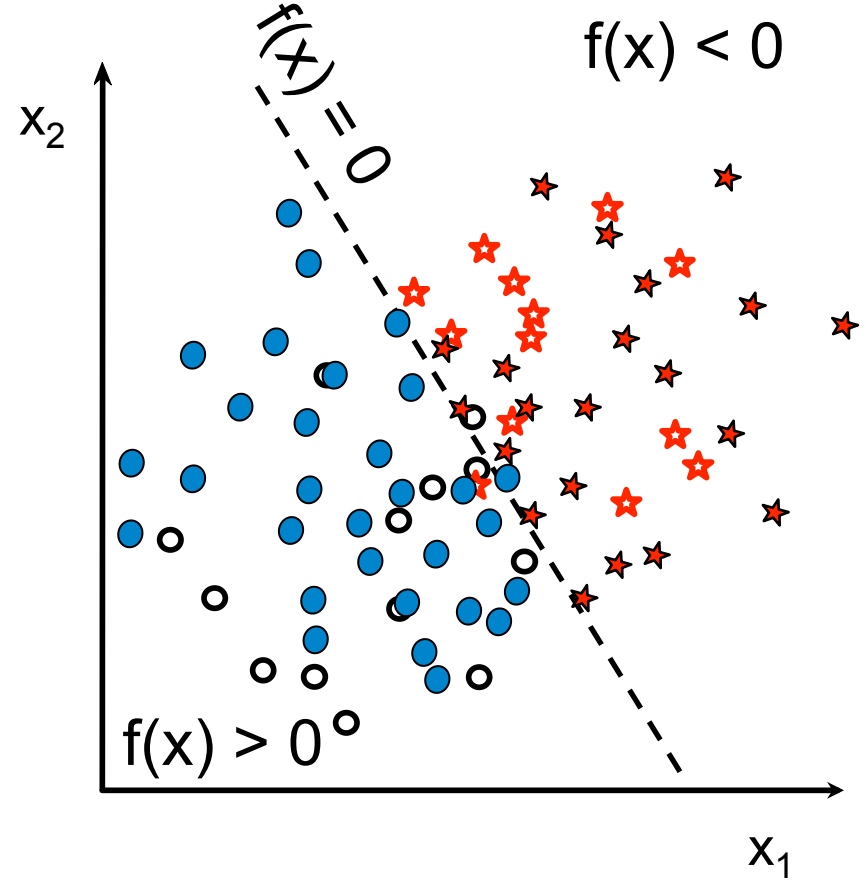
- Area under ROC curve.
- Mean square error or the R^2 = 1-MSE/VAR.

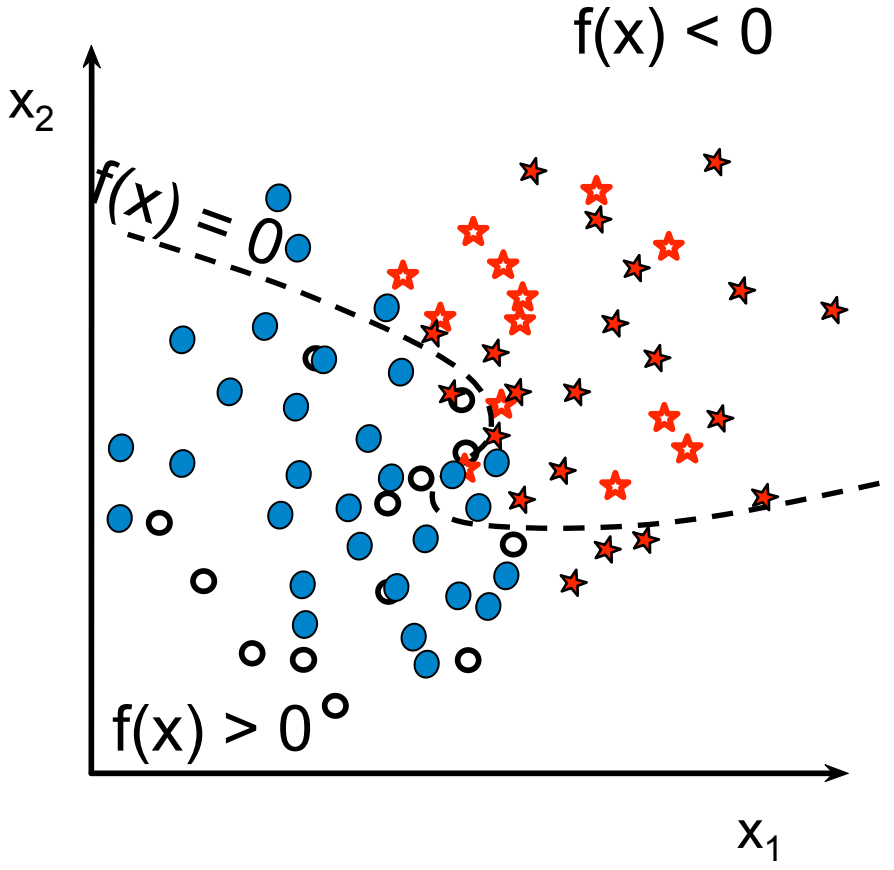
Original decision boundary



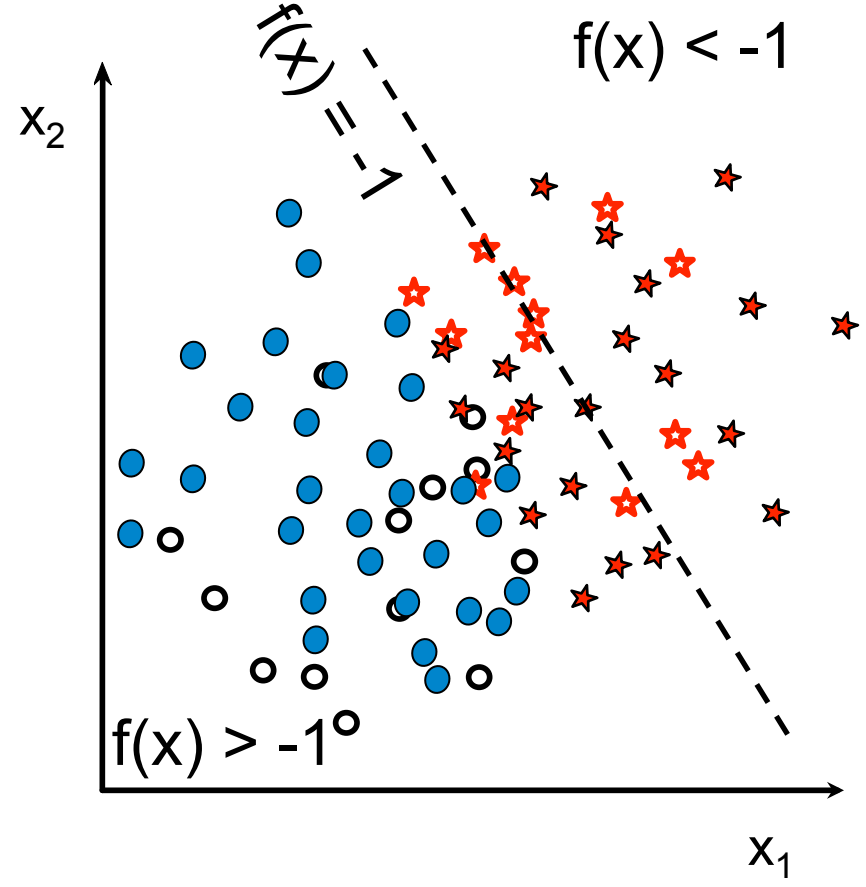


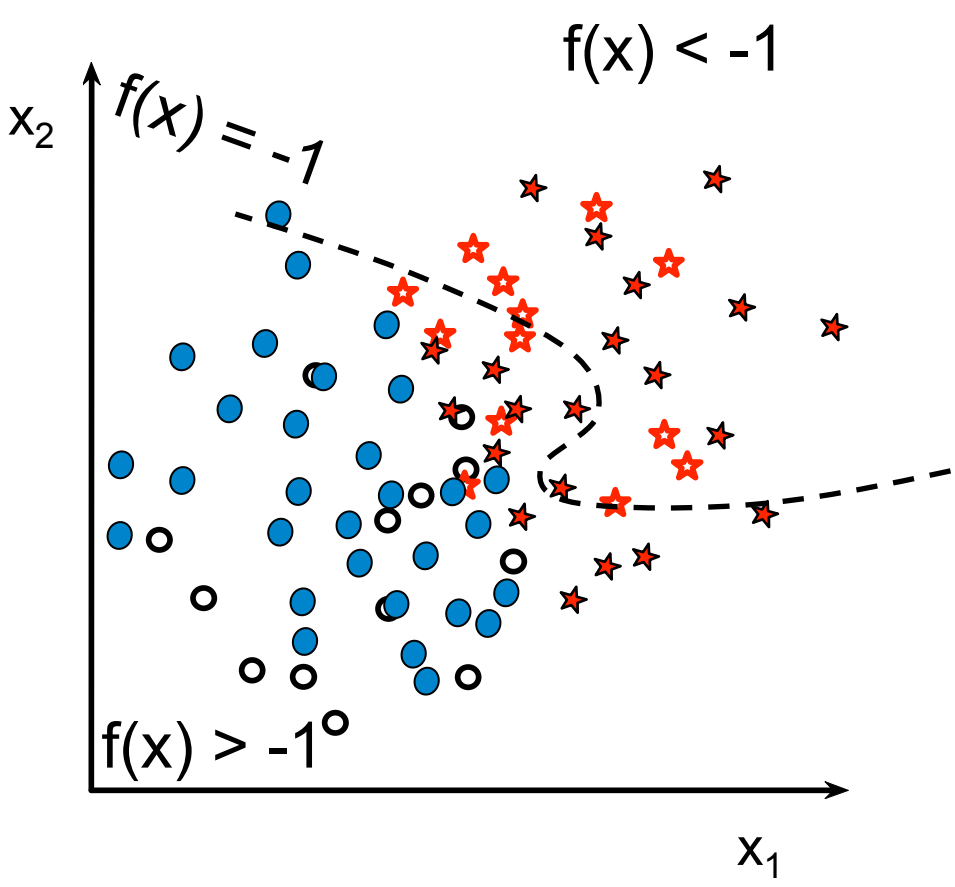
Overlay test examples, count FP and FN



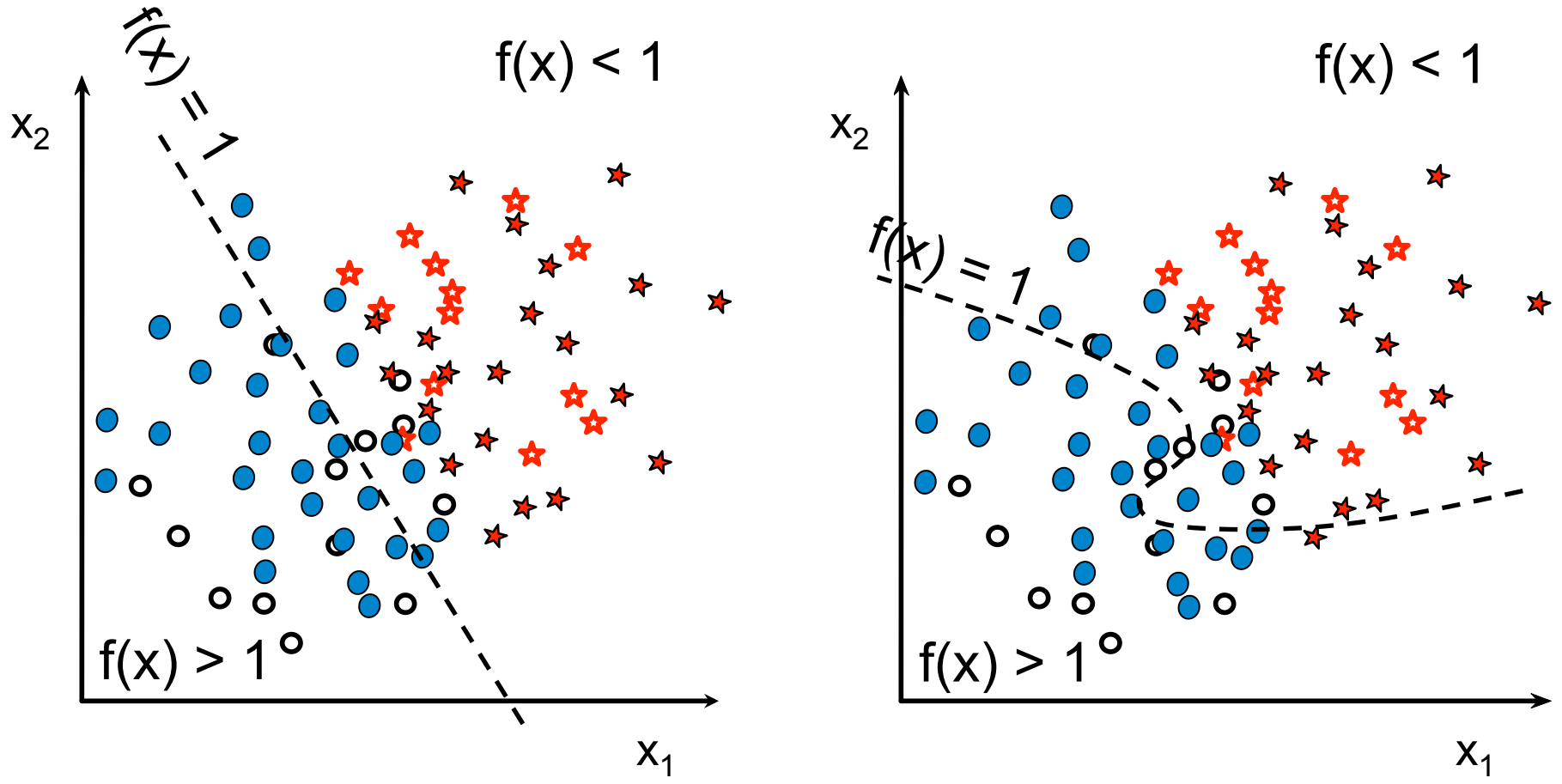


Move up the decision boundary: fewer False Negative



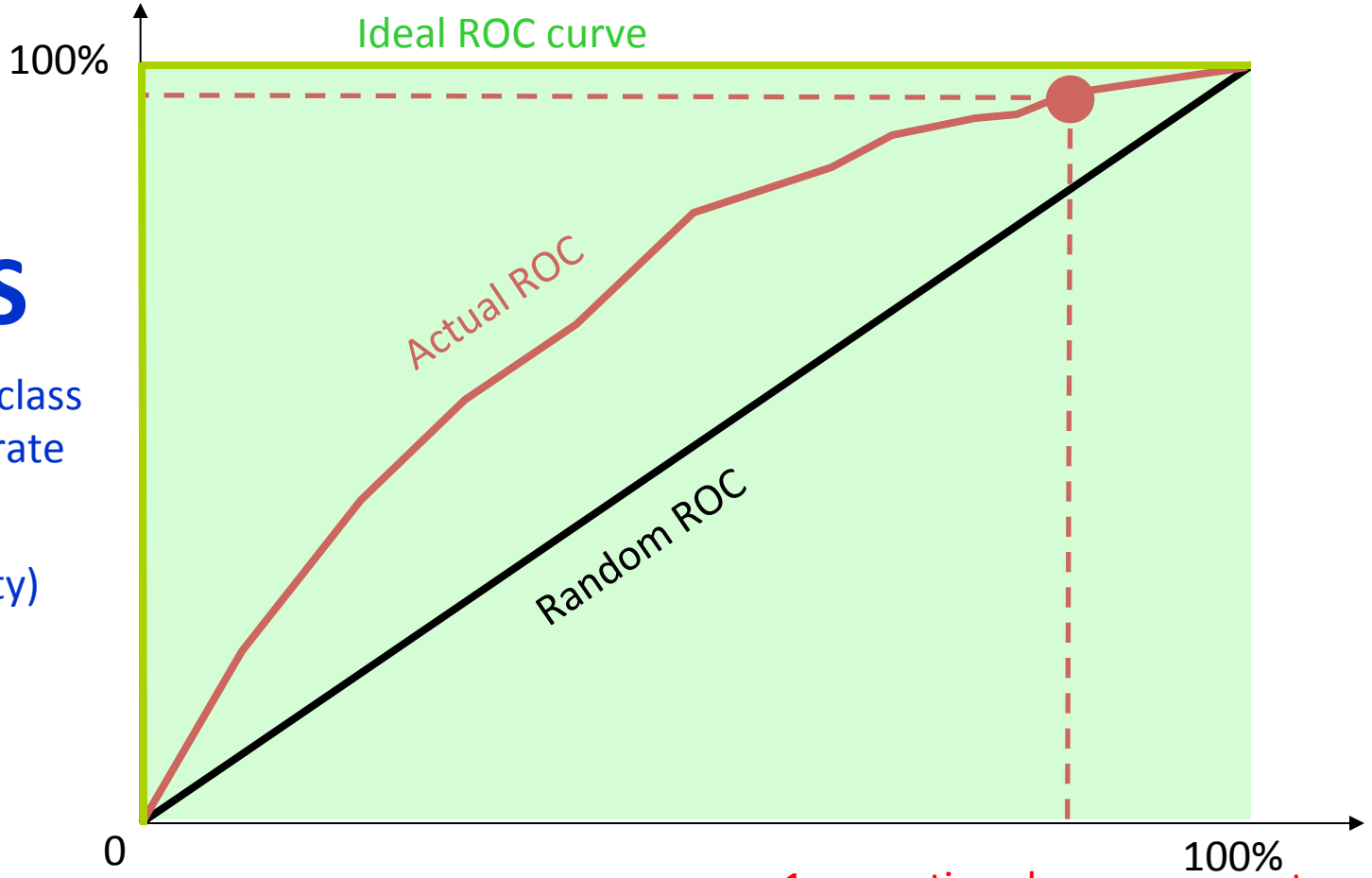


Move down the decision boundary: fewer False Positive



For a given threshold on f(x), you get a point on the ROC curve.

ROC Curve



HITS

Positive class success rate

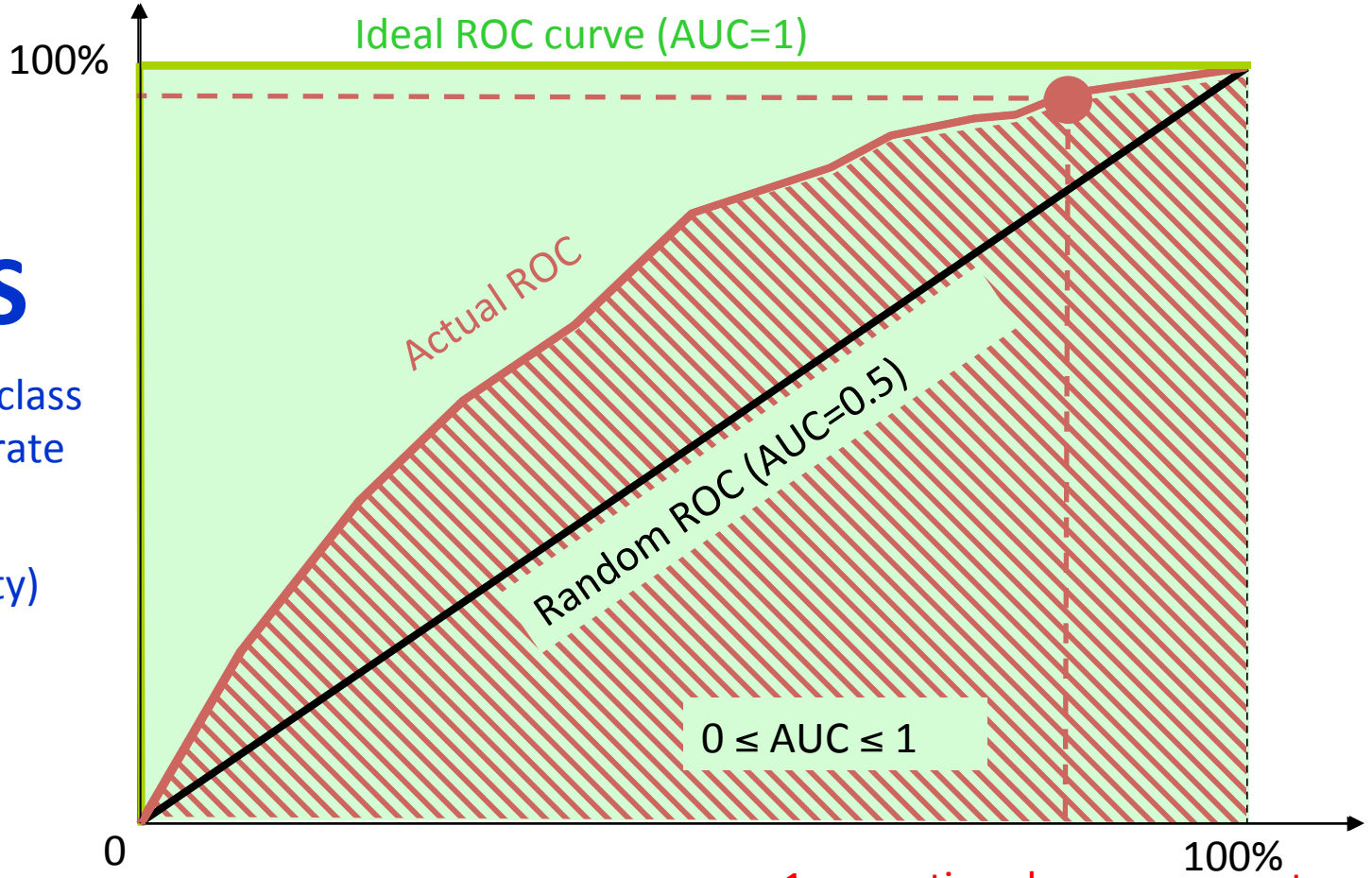
(hit rate, sensitivity)

FALSE ALARMS

1 - negative class success rate
(false alarm rate, 1-specificity)

For a given threshold on f(x), you get a point on the ROC curve.

Area under ROC Curve



HITS

Positive class success rate

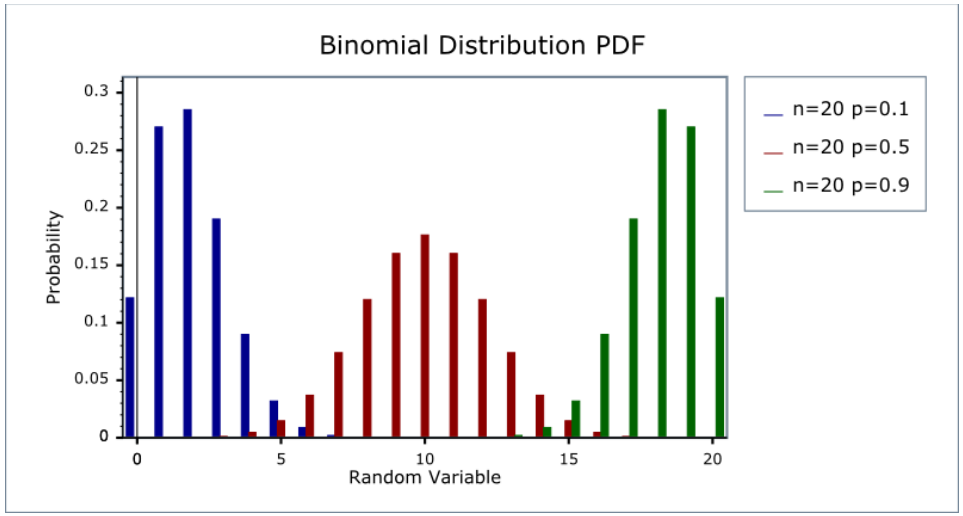
(hit rate, sensitivity)

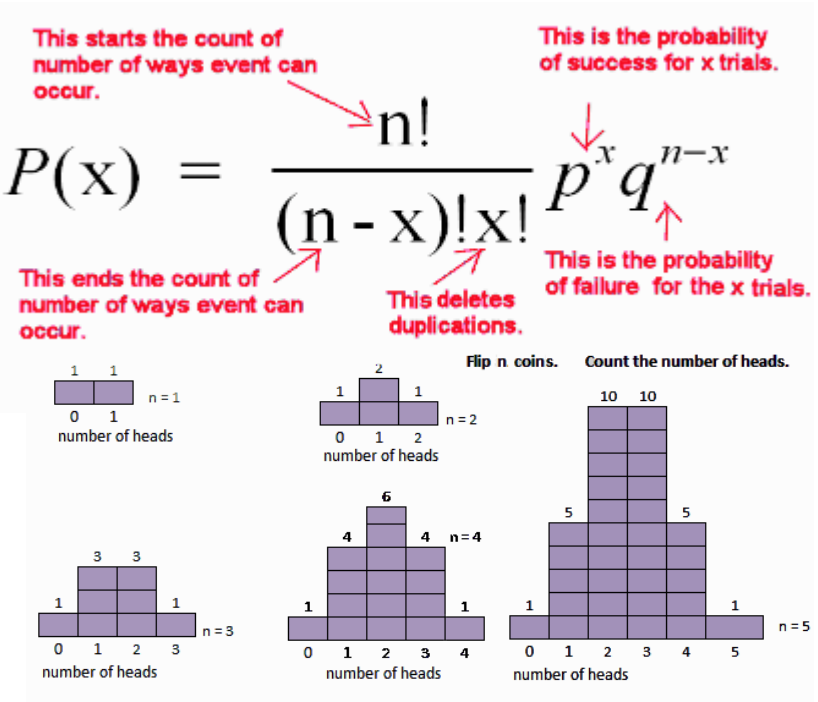
FALSE ALARMS

1 - negative class success rate (false alarm rate, 1-specificity)

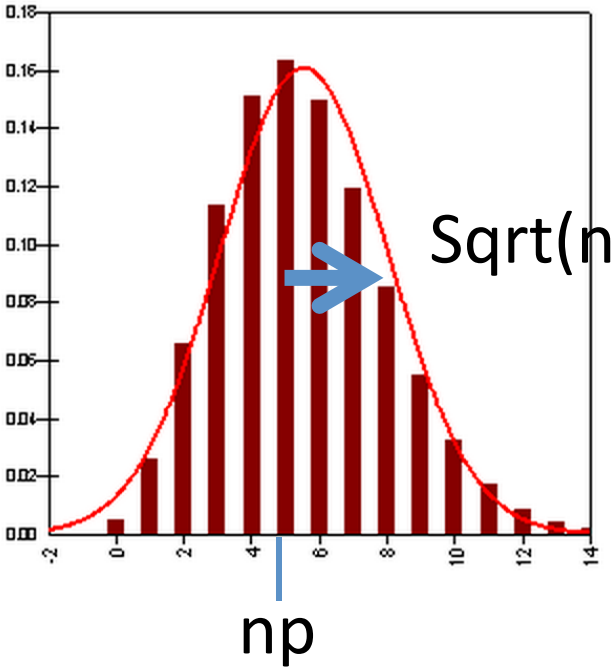
Error bars

- Counting errors is like counting heads when tossing a biased coin.
- Proba p of head and q=(1-p) of tail.
- After n tosses, proba of x errors:





Error bars (continued)



Density of getting x errors (heads) after n=50 tosses with proba p = 0.1.

Follows the binomial distribution.

Expected value of x: **np**

Variance of x: **np(1-p)**

Expected value of the error rate **E = x/n = p**

Variance of x/n: σ^2 **= var(x)/n² = p(1-p)/n**

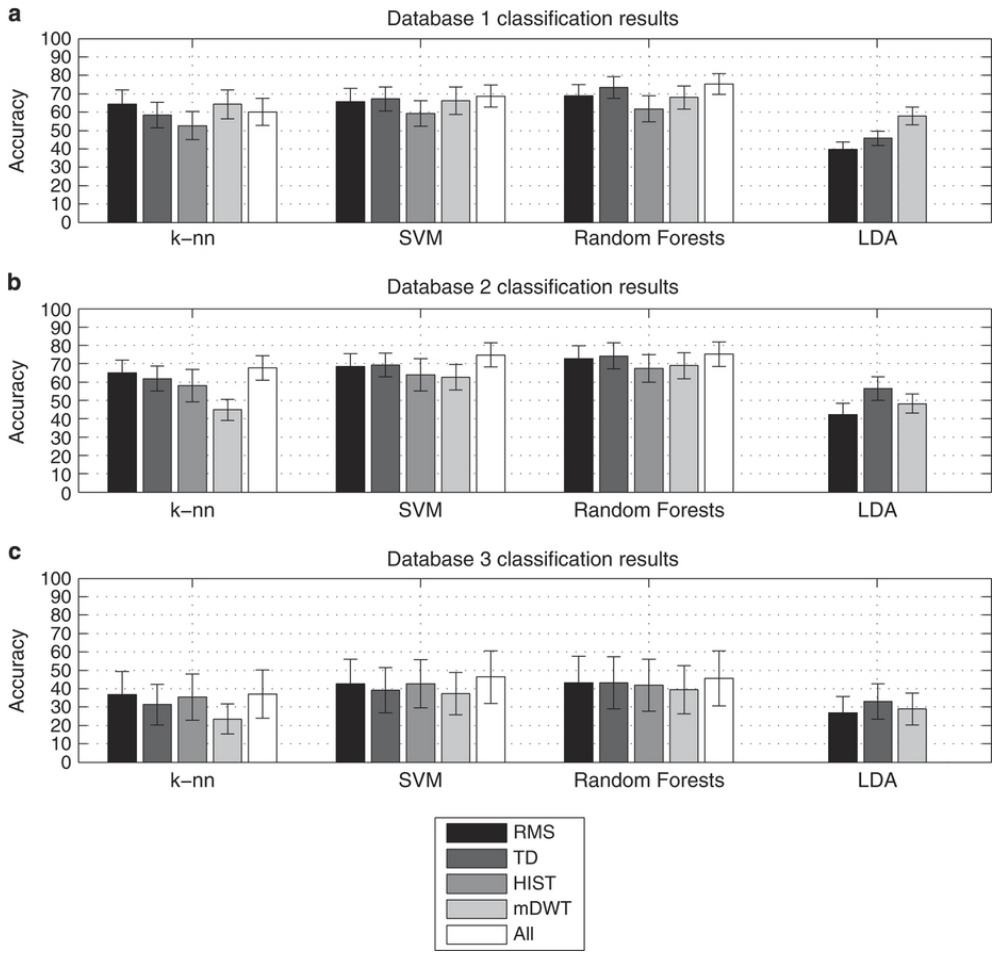
Use as error bar of the error rate E:

$$\sigma = \sqrt{E(1-E)/n}$$

n is the number of test examples.

More on error bars: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175406/

ML people like error bars…



http://www.nature.com/articles/sdata201453/figures/6

Statisticians like box plots…

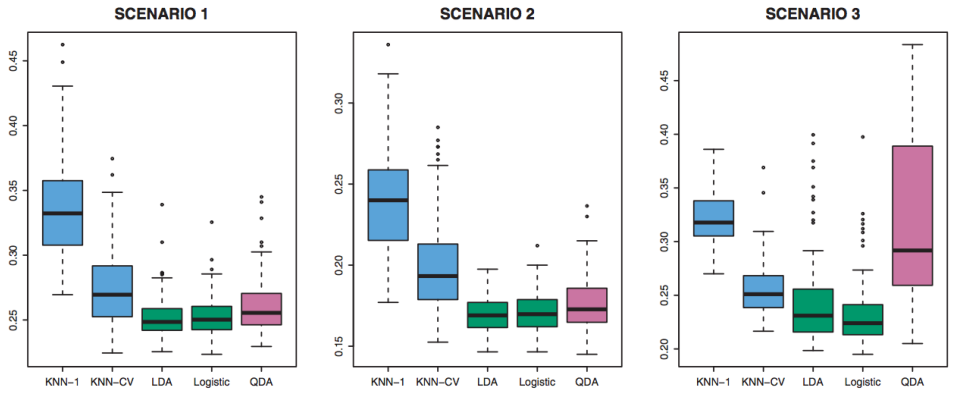


FIGURE 4.10. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

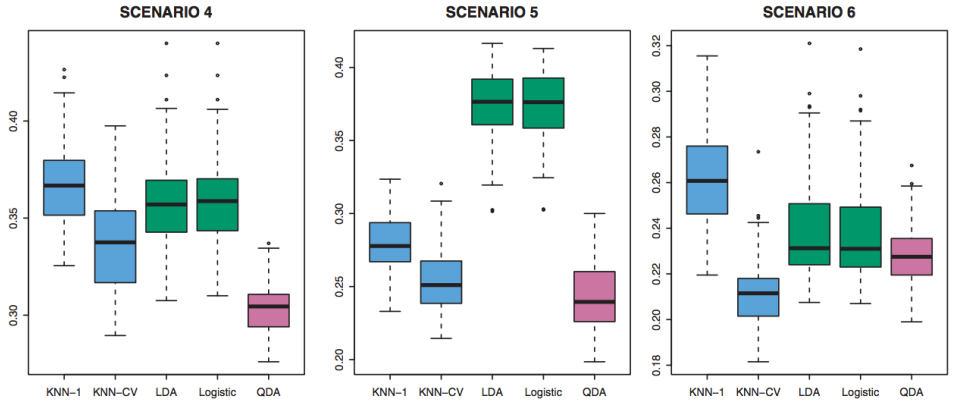


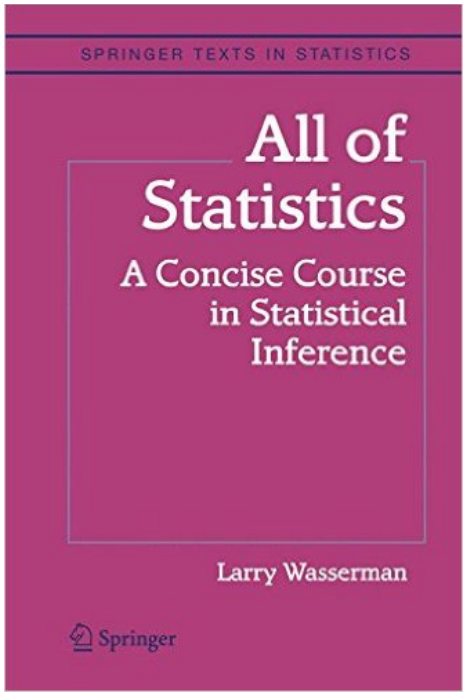
FIGURE 4.11. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf, p. 152

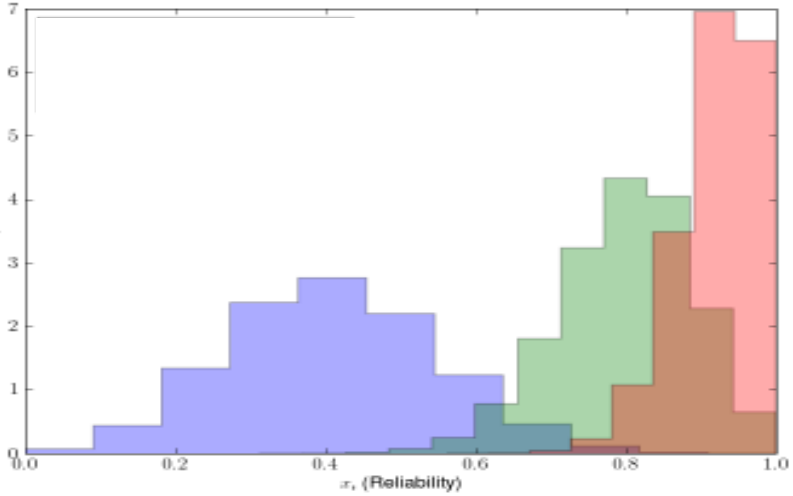
But where do we find the distributions of our favorite cost function???

OR use the bootstrap:

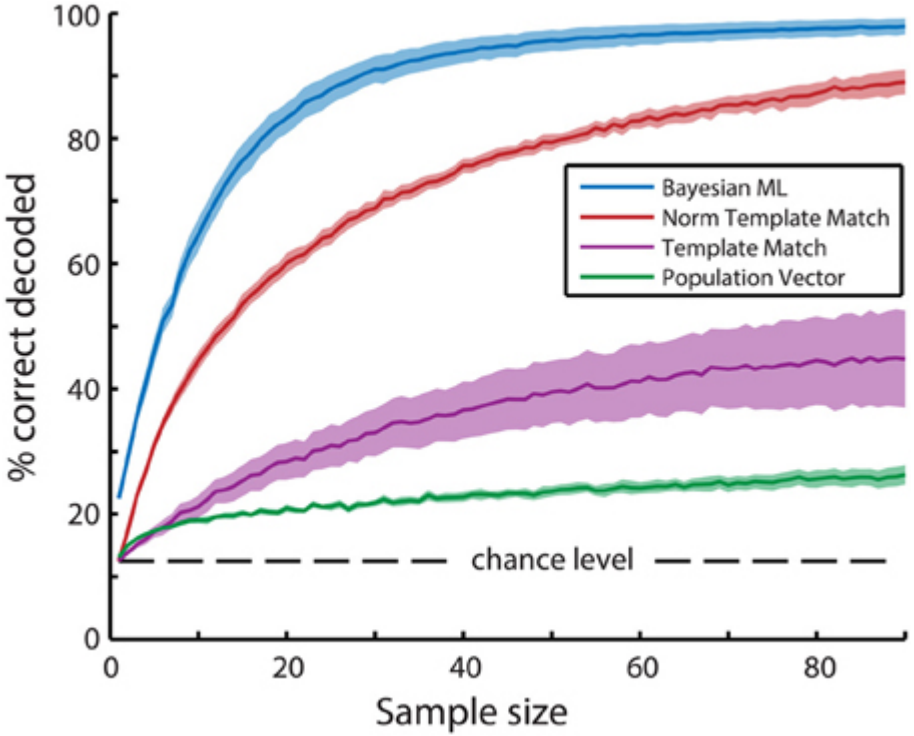
Read this book:

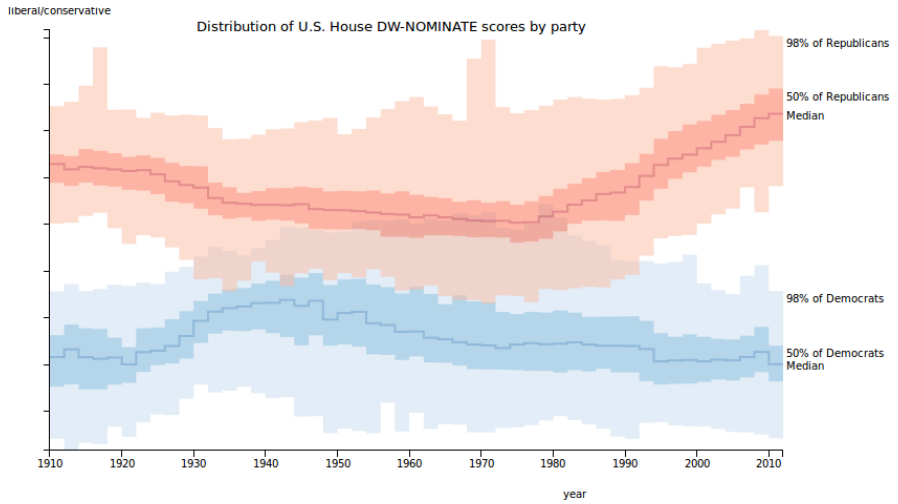


- No math required!
- You have n test samples; just **resample** (with replacements) subsets of n test samples many times (typically thousands) until the statistics you desire seem stable.
- The average number of distinct observations in each sample is about 0.632n.



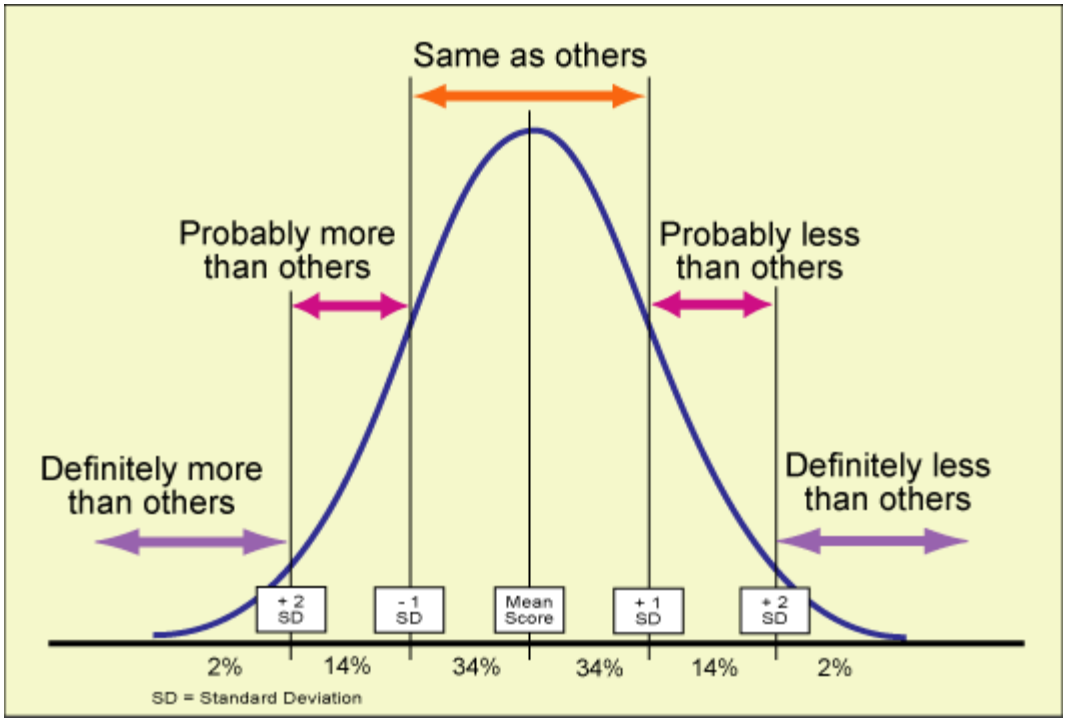
Bootstrap for error bars





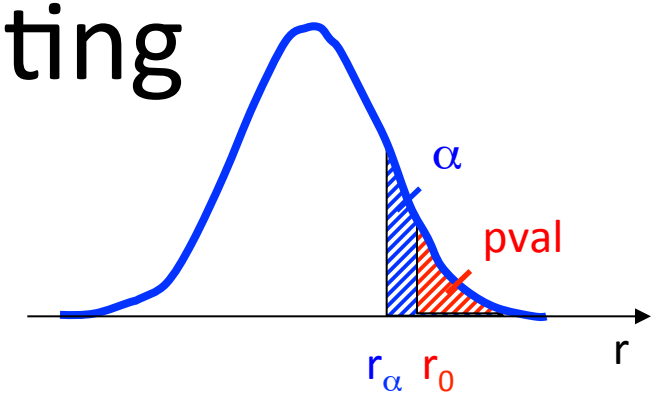
http://journal.frontiersin.org/article/10.3389/fncom.2014.00058/full

Result “significance”



http://brothersgrimmandgorey.blogspot.com/2010/03/using-statistical-significance-95.html

Hypothesis Testing



Ingredients:

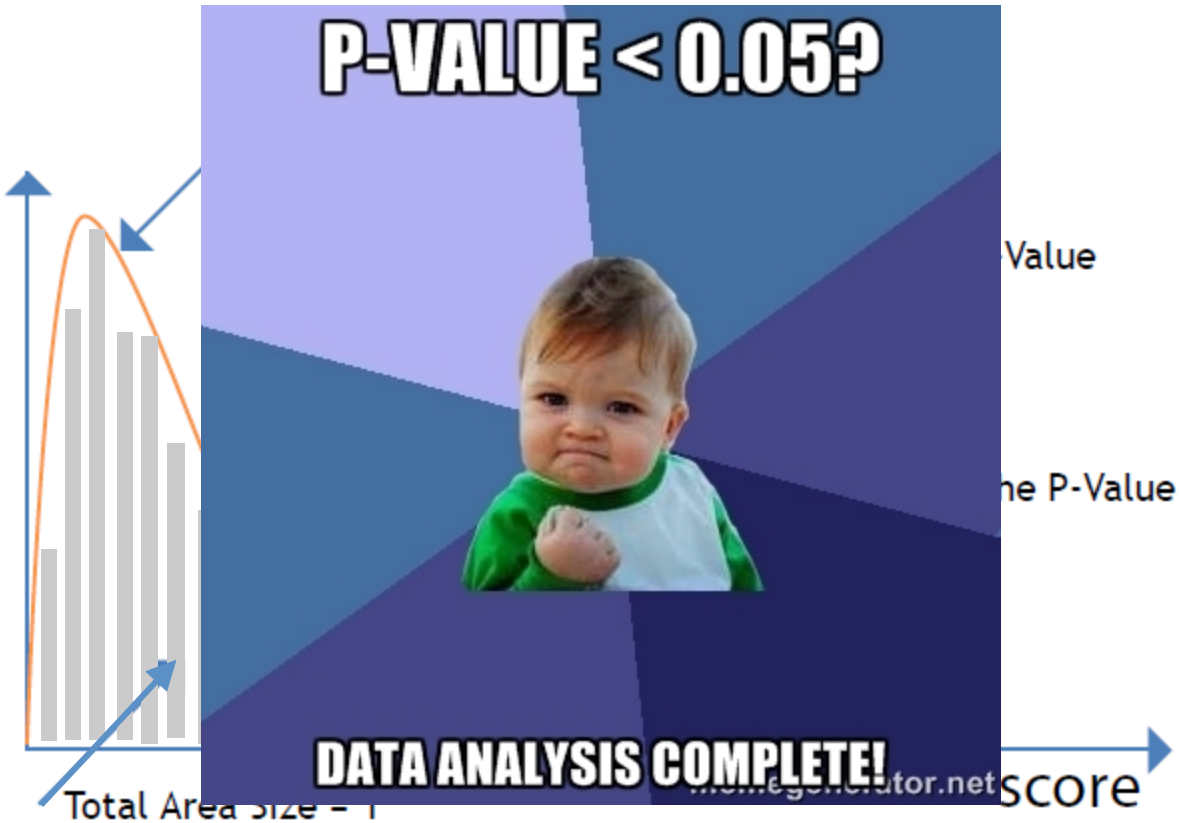
- A "null hypothesis" H_0.

"H_0: my score is no different than the baseline"

- A test statistic R *(the score).*
- A distribution of R if H_0 is true *(null distribution)*
- A risk value α and its corresponding threshold r_α, such that $\alpha = Proba(R > r_\alpha)$.
- A realization r_0 of R to be tested.

If $r_0 > r_\alpha$, reject H_0, with risk α of being wrong.

Pvalue



Pvalue = fraction of score values above the reference score.
SMALL pvalues are good: they shed doubt on the “null hypothesis”.

Number of test examples needed

Variance of test error rate σ^2= **E(1-E)/n**

For E<<1, $\sigma^2 \cong E/n$ (1)

Choose a given coefficient of variance $\sigma/E=0.1$,

that is $\sigma^2 = 0.01\ E^2$ (2)

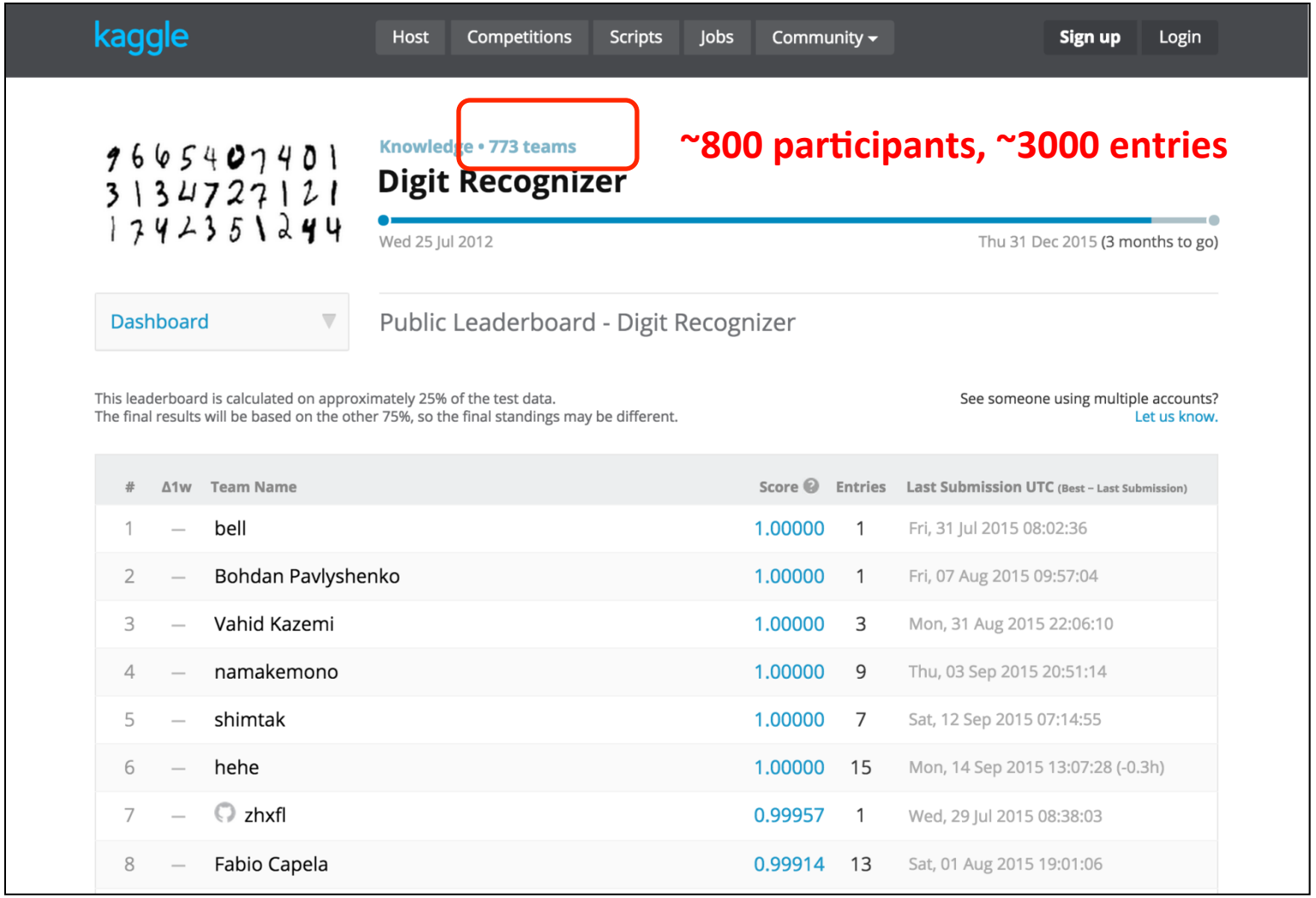
Combining (1) and (2):

$E/n = 0.01\ E^2$

$n = 100/E$

Guyon et al., 1996. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.3326

Can I use my test set to compare multiple methods?



#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	bell	1.00000	1	Fri, 31 Jul 2015 08:02:36
2	—	Bohdan Pavlyshenko	1.00000	1	Fri, 07 Aug 2015 09:57:04
3	—	Vahid Kazemi	1.00000	3	Mon, 31 Aug 2015 22:06:10
4	—	namakemono	1.00000	9	Thu, 03 Sep 2015 20:51:14
5	—	shimtak	1.00000	7	Sat, 12 Sep 2015 07:14:55
6	—	hehe	1.00000	15	Mon, 14 Sep 2015 13:07:28 (-0.3h)
7	—	zhxfl	0.99957	1	Wed, 29 Jul 2015 08:38:03
8	—	Fabio Capela	0.99914	13	Sat, 01 Aug 2015 19:01:06

Bonferroni correction

Pvalue = small proba of wrong conclusion.

Can I try m different ways of getting a small pvalue to increase my chances?

No! If m independent tests are performed, the fraction of correct decisions will be $(1\text{-pval})^m$.

Bonferroni correction: Replace pval by m pval. (or the risk α by α/m in a statistical test).

Paired test of significance

- What difference in error rate between 2 classifiers is statistically significant?
- McNemar paired test:

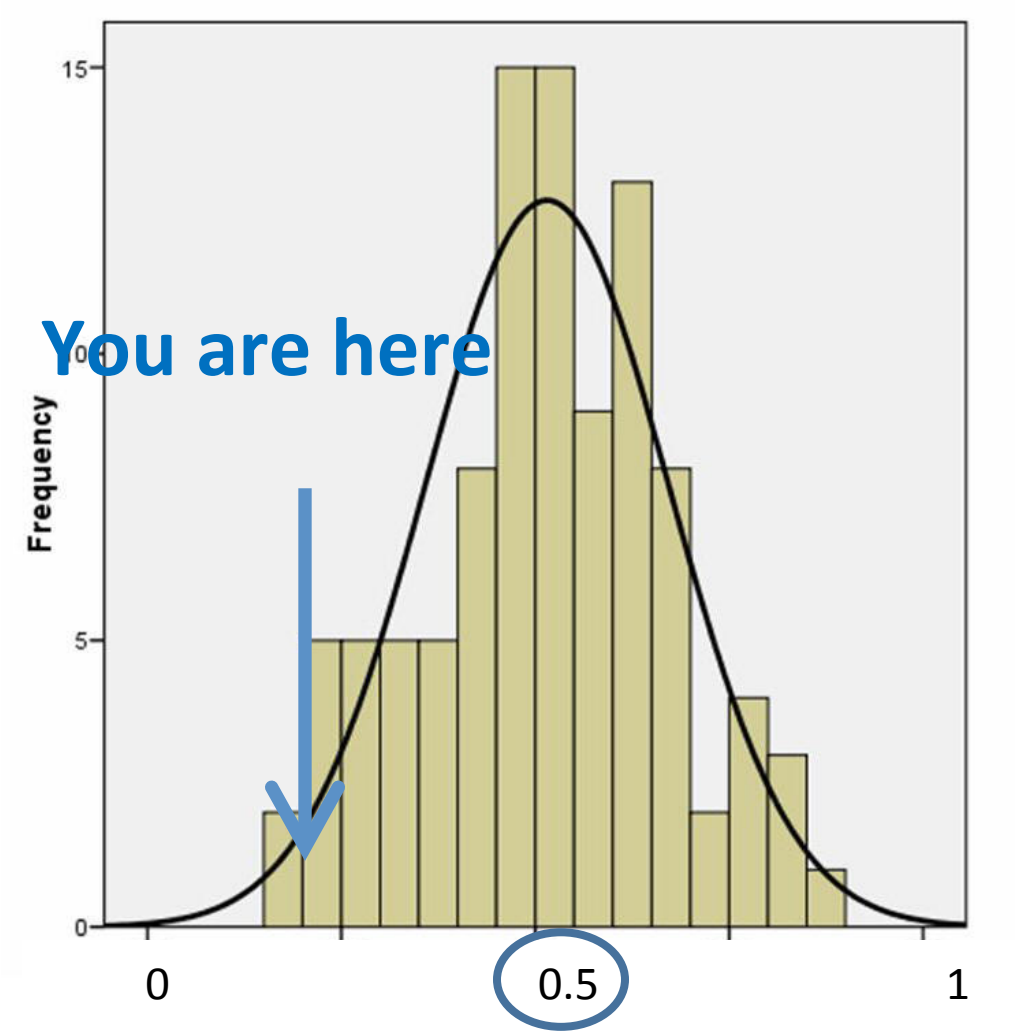
  - assume classifier 1 is better

  - ν_i=number of errors classifier i makes that the other classifier does not make.

  - if $E_2-E_1 \geq (z_\alpha/\nu)\text{sqrt}(\nu_1+\nu_2)$ reject H_0 of equality of error rates with risk α.

  - one sided risk α=0.01, z_α=2.33.

Should I continue the project?



You can also compare yourself to the **chance distribution**: permute at random many times the predicted values and re-compute the score.

Summary

- **Data is everything** (more always better).
- But: need "**good data**". Avoid confounding and data leakage. Control, block, randomize.
- Need to choose the **right metric** (AUC, pb. of imbalanced classes).
- Compute **error bars**: σ^2=E(1-E)n or bootstrap.
- **Case 1:** Really **crummy baseline results**: Are they better than random? **Permutation test**.
- **Case 2:** Really **great results**: Do we have enough errors to distinguish the baseline method from a better method? Increase **test set size to n ~ 100/E**.
- **Case 3: Middle perf:** where the gain is potentially biggest. We want to try lots of methods. Understand pvalues and multiple testing. **Bonferroni correction = pval * m.**

Come to my office hours…
Wed 2:30-4:30 Soda 329

Next time: model search