BOS DATAFESTIVAL

Boston, Sep 19th, 2015

ACTING ON DATA

Rahul Dave (rahuldave@gmail.com)

@rahuldave

$\mathcal{L}x$Prior

IACS Harvard John A. Paulson School of Engineering and Applied Sciences
IACS Institute for Applied Computational Science

$\mathcal{L}_x$Prior

Data Science.

Simulation.

Software.

Social Good.

Interesting Problems.

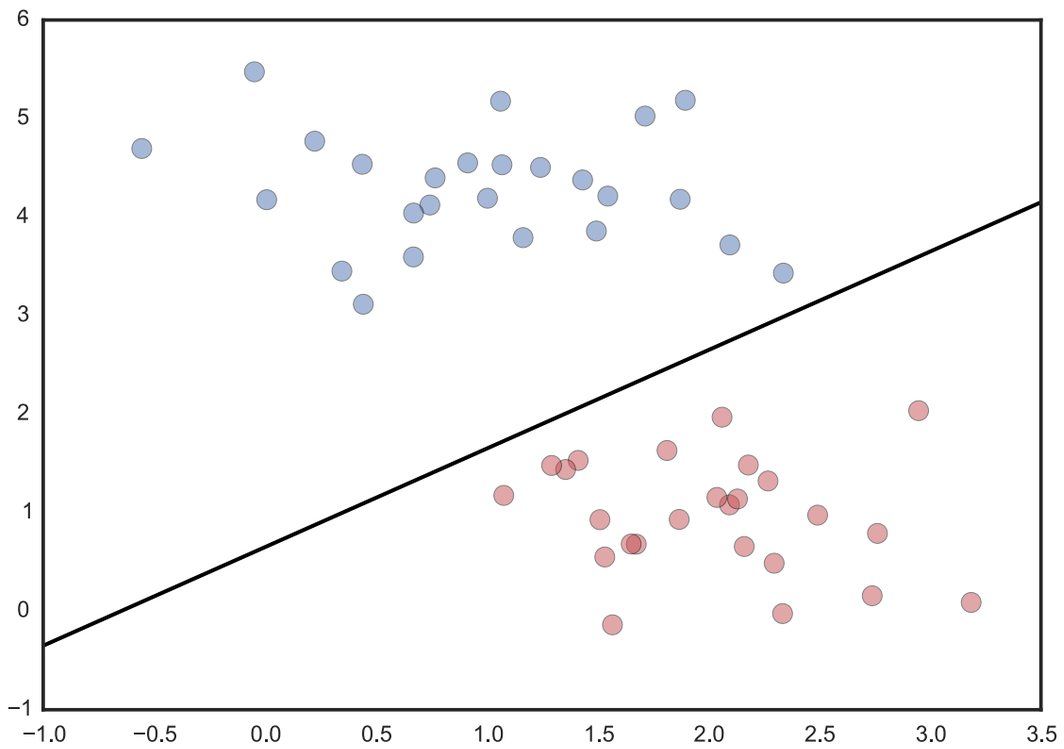
Harvard John A. Paulson School of Engineering and Applied Sciences
IACS Institute for Applied Computational Science

machine learning, complex systems, stochastic methods, viz, extreme computing

DEGREE PROGRAMS:

- Master's of science– one year
- Master's of engineering – two year with thesis/ research project

CLASSIFICATION

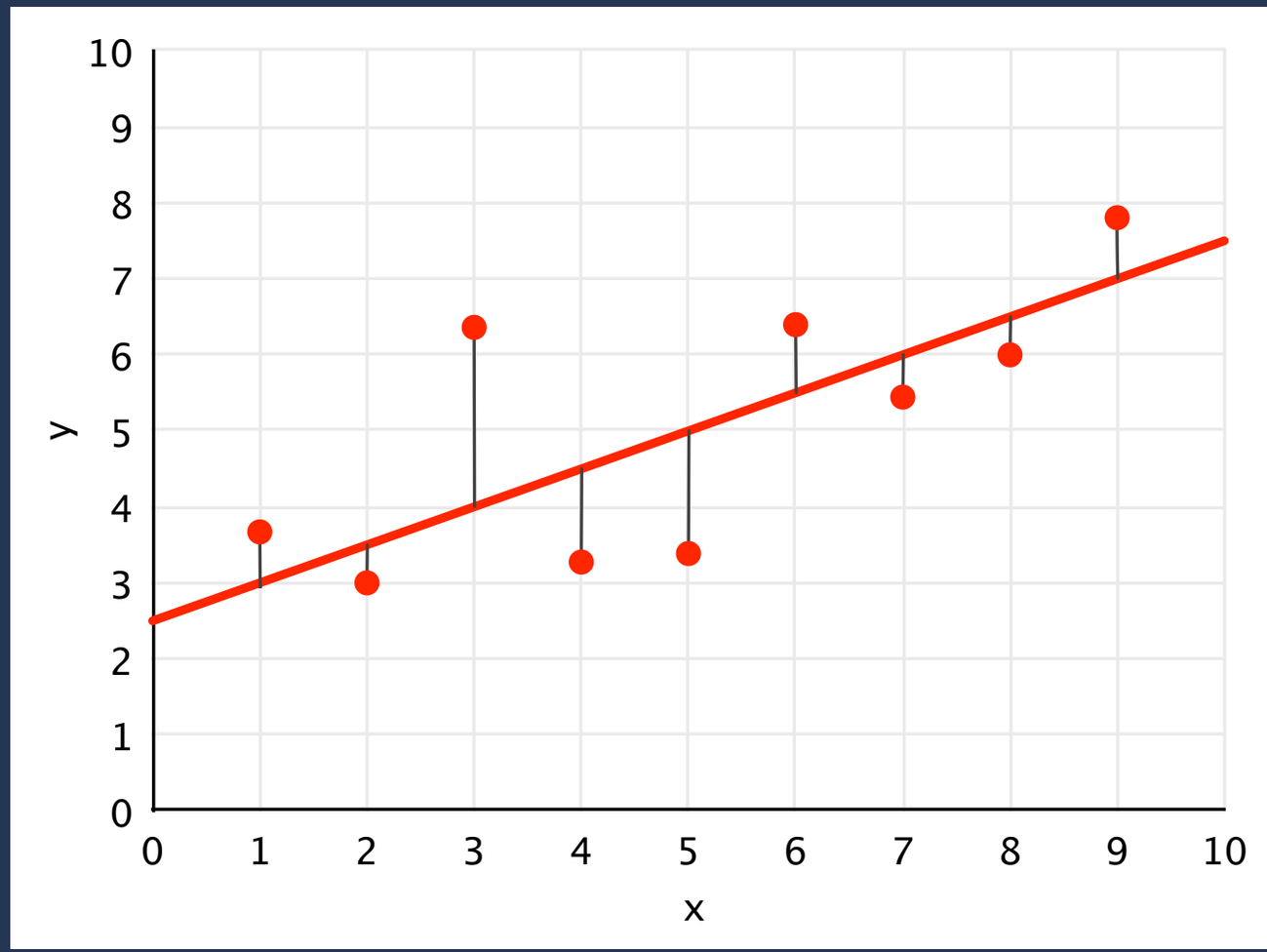


- will a customer churn?
- is this a check? For how much?
- a man or a woman?
- will this customer buy?
- do you have cancer?
- is this spam?
- whose picture is this?
- what is this text about?[j]

[j] image from code in http://bit.ly/1Azg29G

REGRESSION

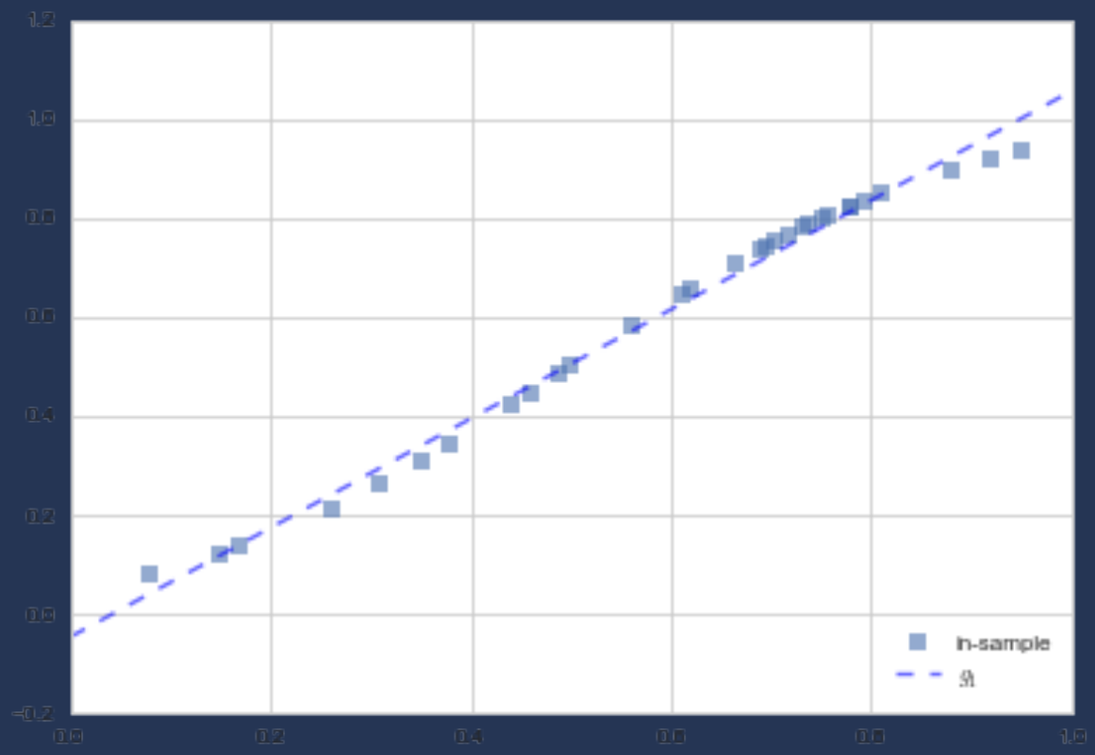
- how many dollars will you spend?
- what is your creditworthiness
- how many people will vote for Bernie t days before election
- use to predict probabilities for classification
- causal modeling in econometrics

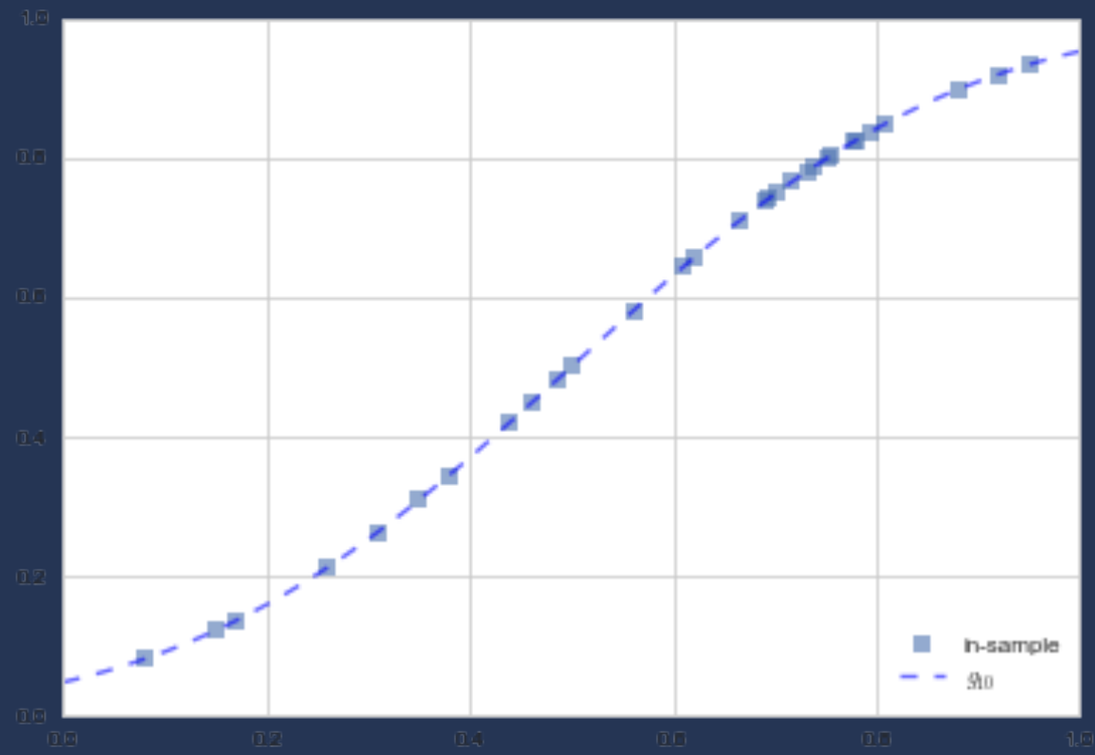


From Bayesian Reasoning and Machine Learning, David Barber:

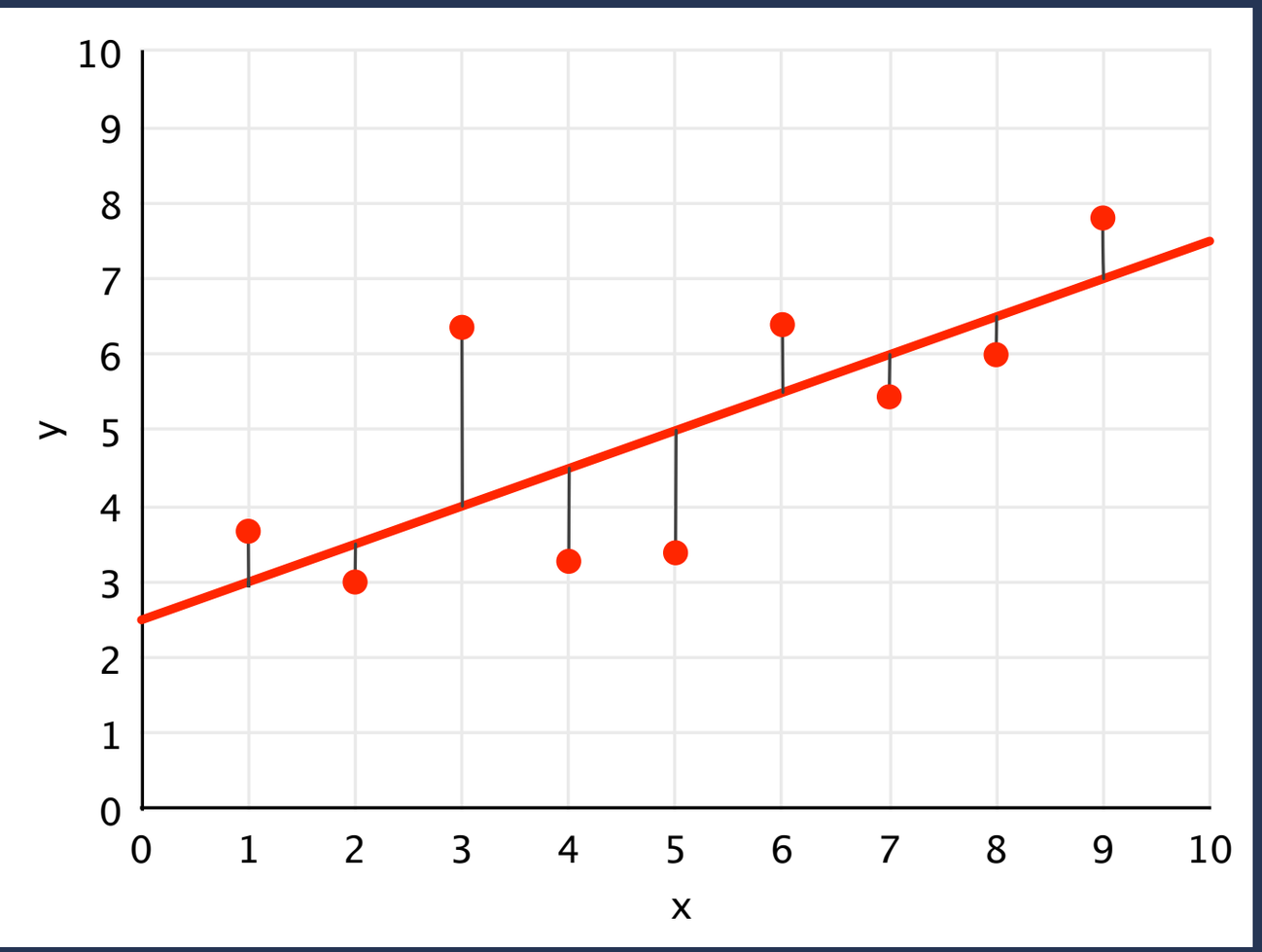
"A father decides to teach his young son what a sports car is. Finding it difficult to explain in words, he decides to give some examples. They stand on a motorway bridge and ... the father cries out ‘that’s a sports car!’ when a sports car passes by. After ten minutes, the father asks his son if he’s understood what a sports car is. The son says, ‘sure, it’s easy’. An old red VW Beetle passes by, and the son shouts – ‘that’s a sports car!’. Dejected, the father asks – ‘why do you say that?’. ‘Because all sports cars are red!’, replies the son."

30 points of data. Which fit is better? Line in $\mathcal{H}_1$ or curve in $\mathcal{H}_{20}$?





What does it mean to FIT?



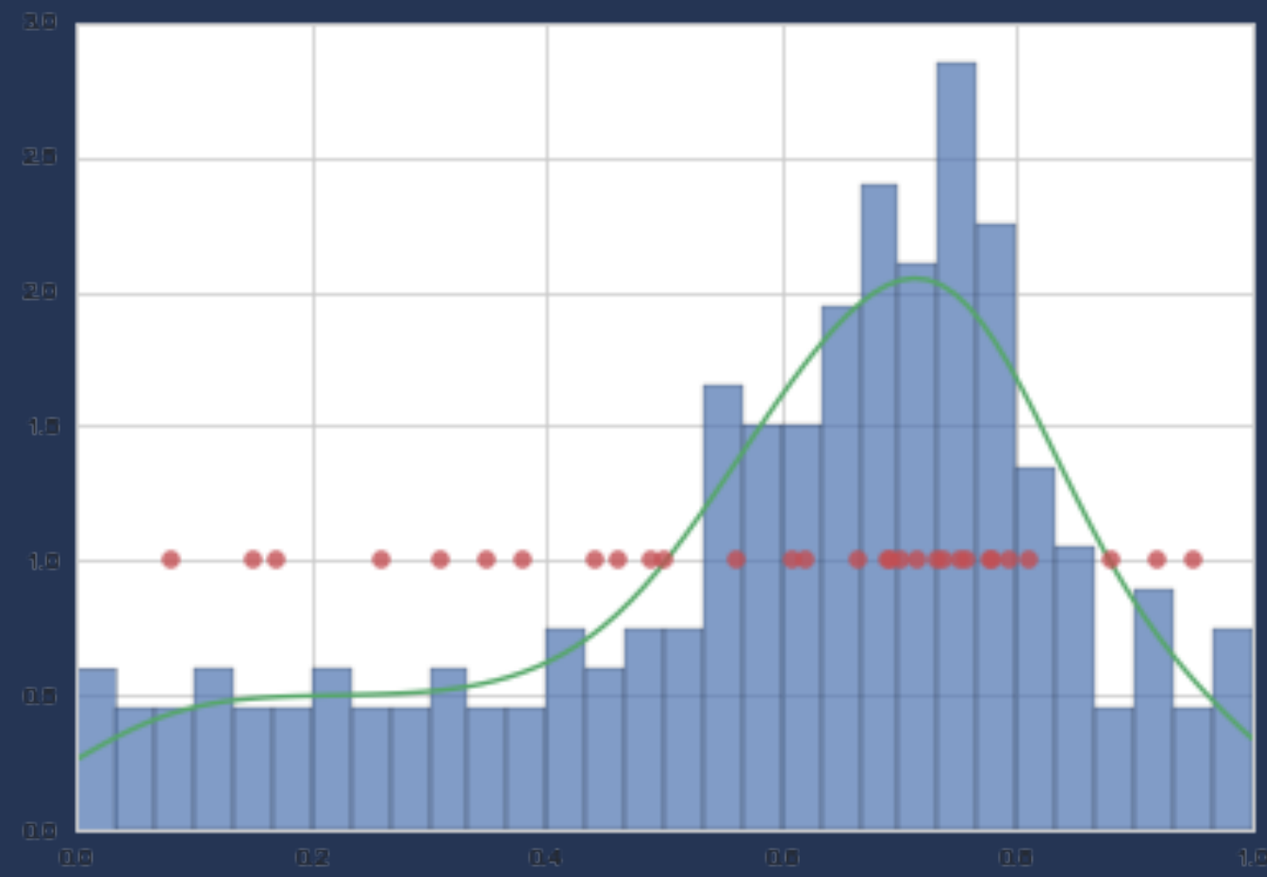
Minimize distance from the line?

$$R_{\mathcal{D}}(h_1(x)) = \frac{1}{N}\sum_{y_i \in \mathcal{D}}(y_i - h_1(x_i))^2$$

Minimize squared distance from the line.

$$g_1(x) = \arg\min_{h_1(x) \in \mathcal{H}} R_{\mathcal{D}}(h_1(x)).$$

Get intercept w_0 and slope w_1.



EMPIRICAL RISK MINIMIZATION

The sample must be representative of the population!

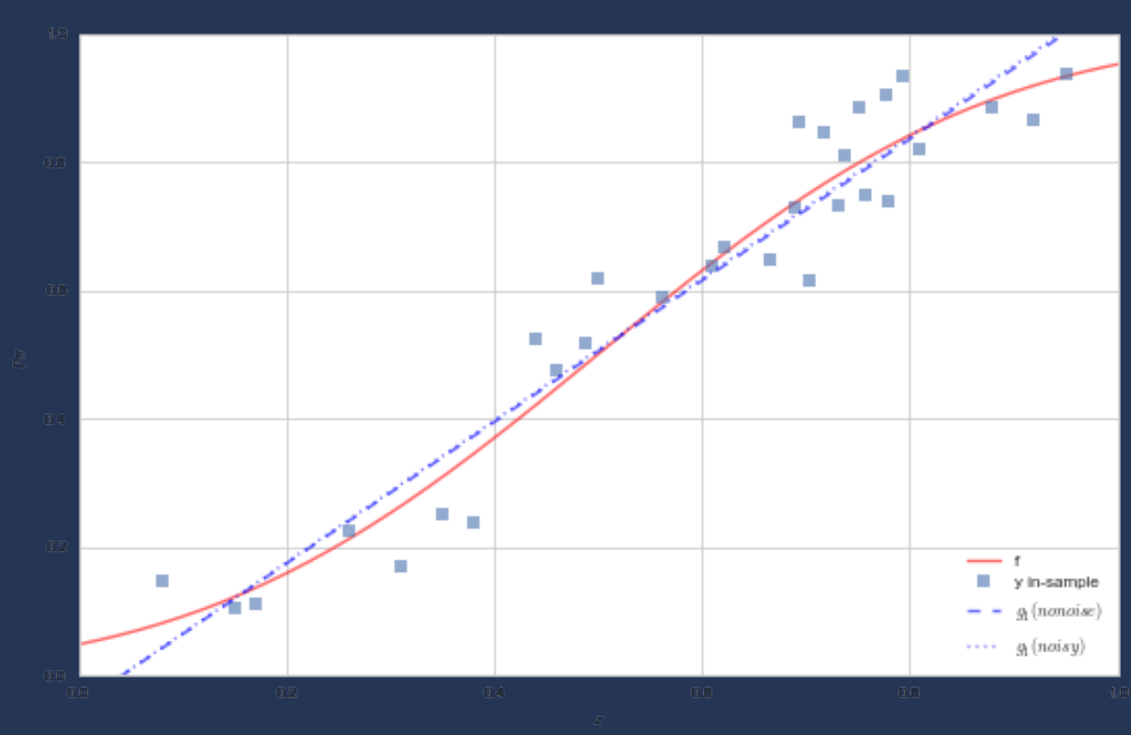
$$A: R_{\mathcal{D}}(g) \ \ smallest\ on\ \mathcal{H}$$
$$B: R_{out\ of\ sample}(g) \approx R_{\mathcal{D}}(g)$$

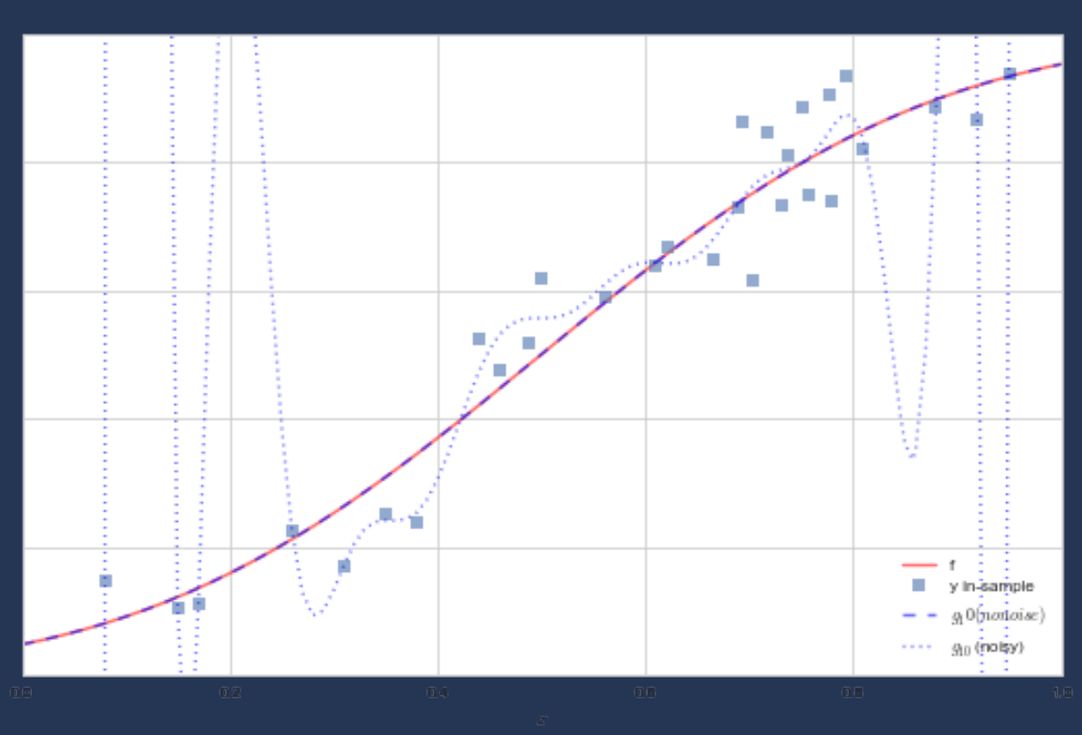
A: Empirical risk estimates out of sample risk.
B: Thus the out of sample risk is also small.

THE REAL WORLD HAS NOISE

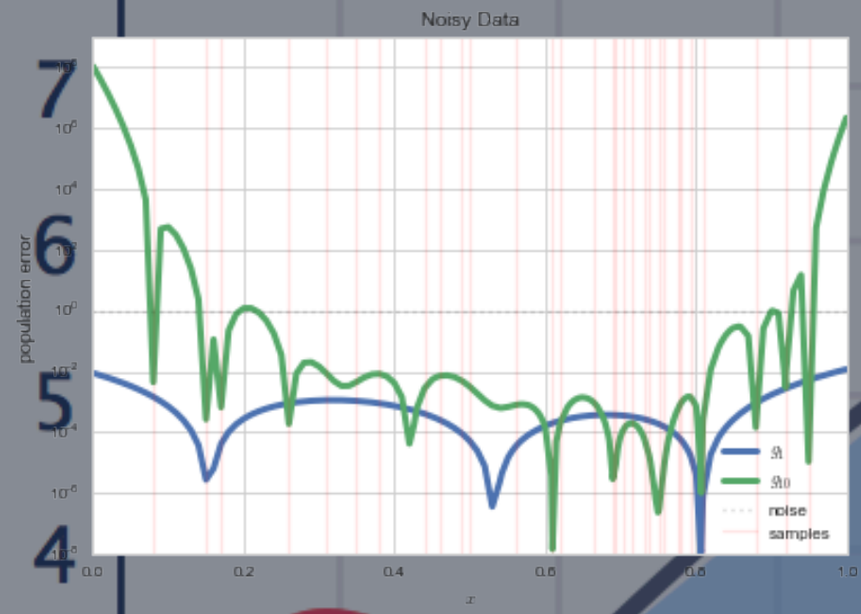
Which fit is better now?

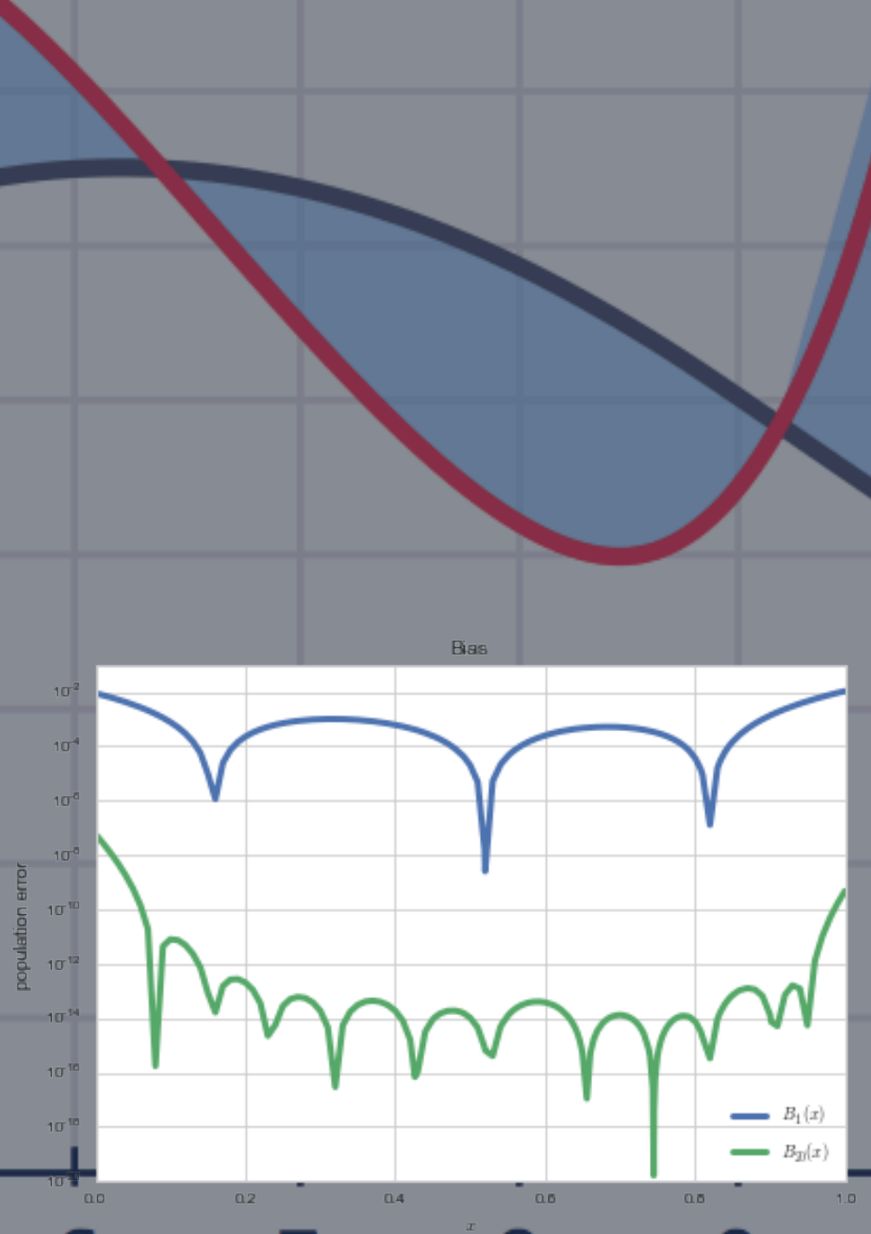
The line or the curve?



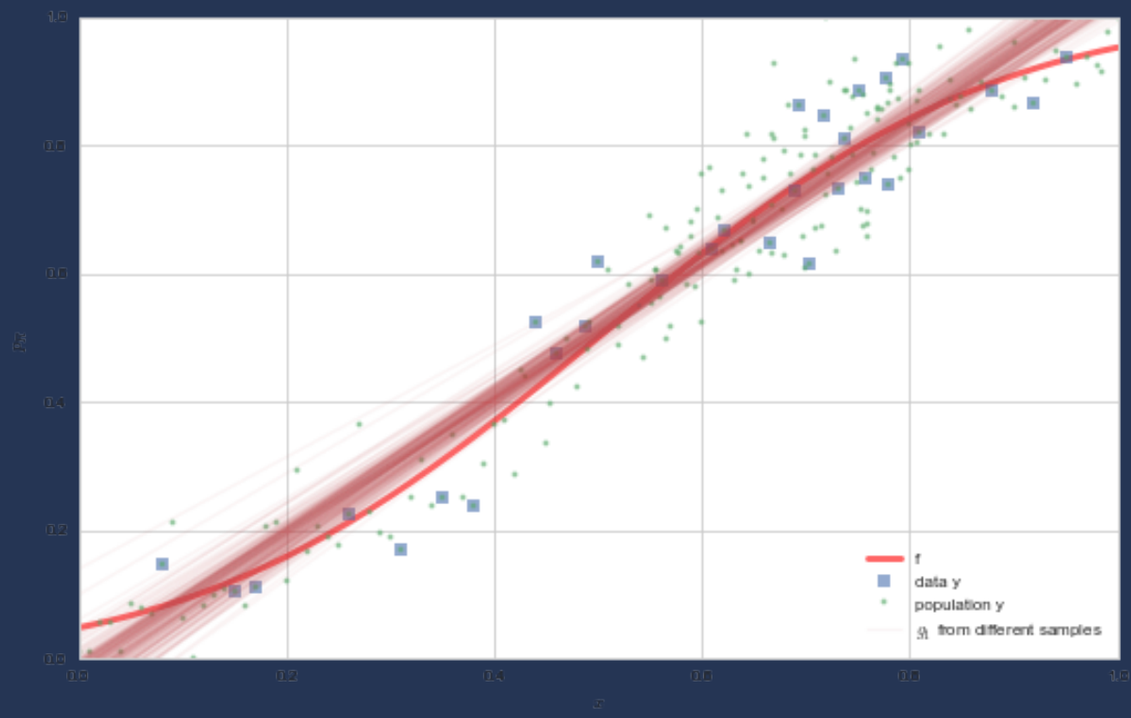


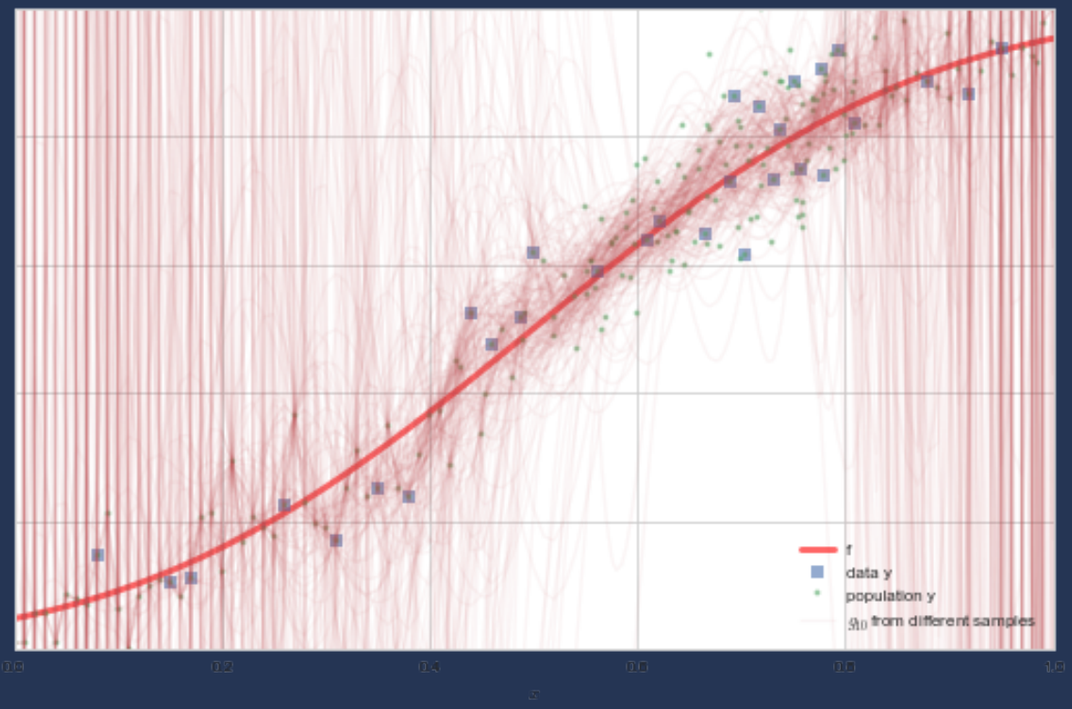
DETERMINISTIC NOISE (Bias) vs STOCHASTIC NOISE





UNDERFITTING (Bias) vs OVERFITTING (Variance)



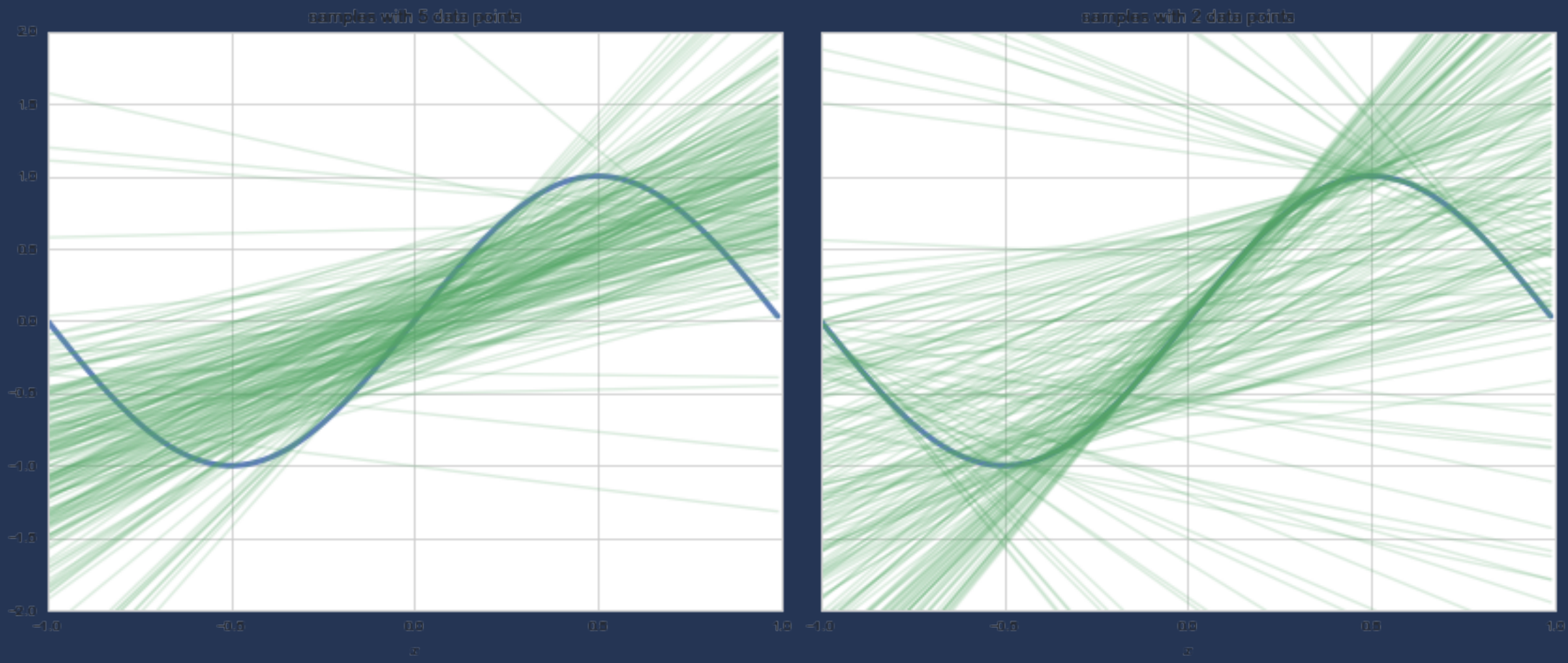


If you are having problems in machine learning the reason is almost always[1]

OVERFITTING

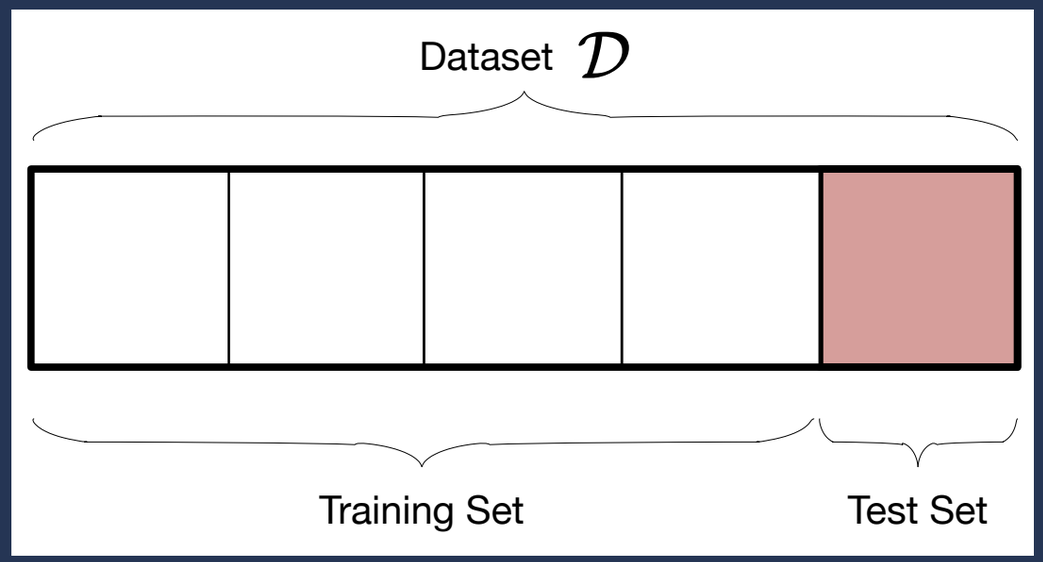
[1] Background from http://commons.wikimedia.org/wiki/File:Overfitting.svg

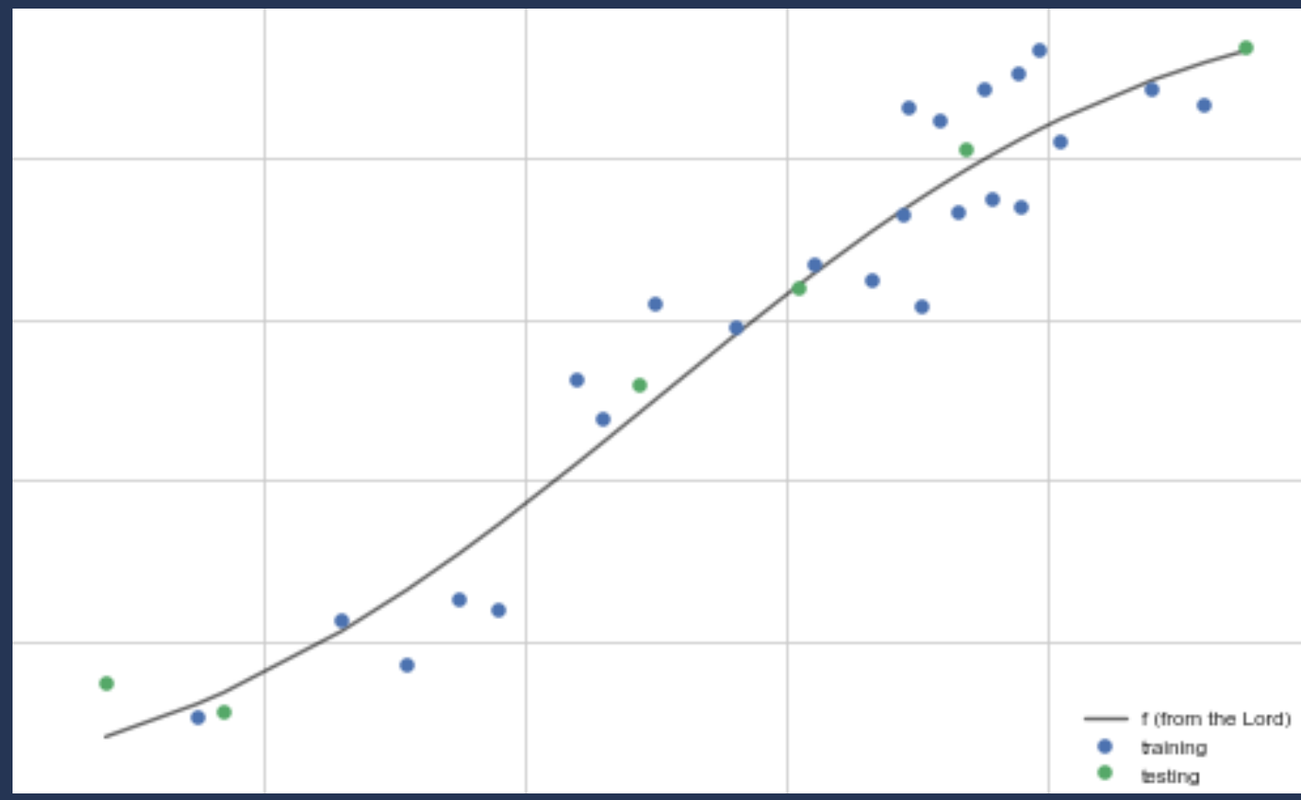
DATA SIZE MATTERS: straight line fits to a sine curve



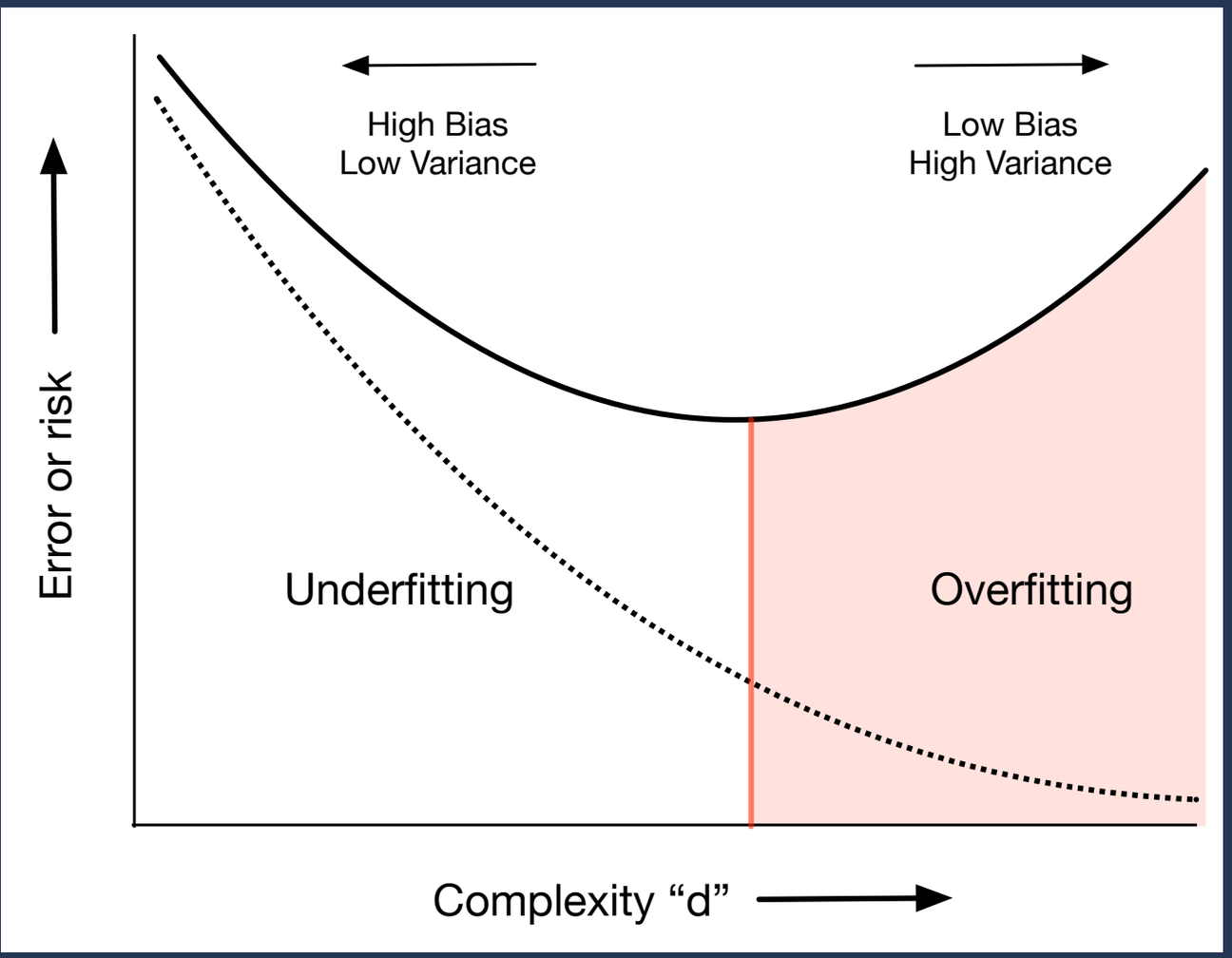
Corollary: Must fit simpler models to less data!

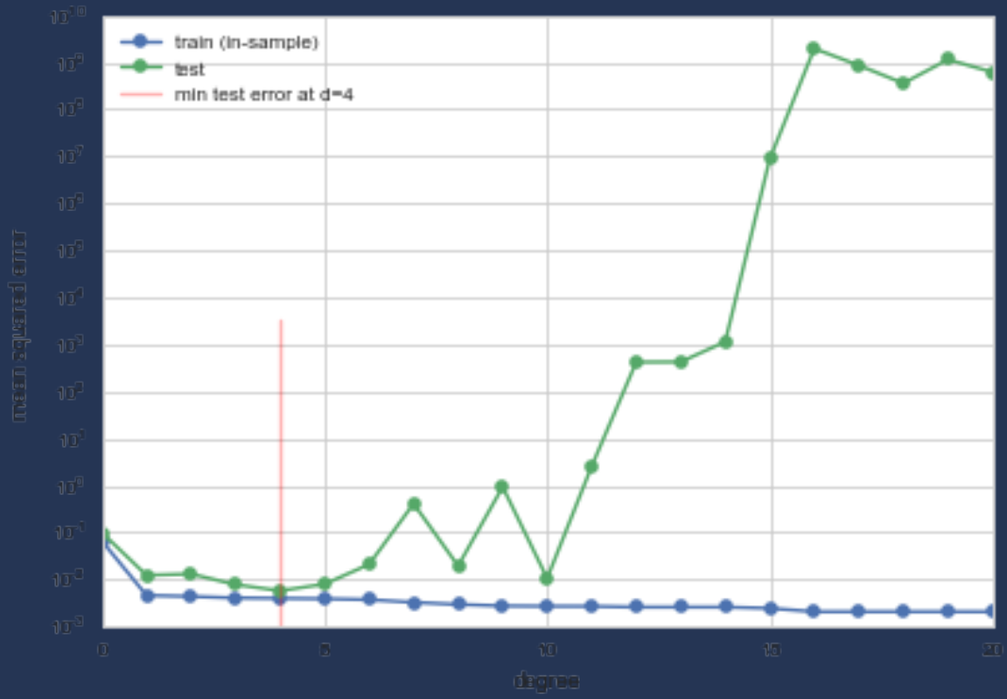
HOW DO WE LEARN?





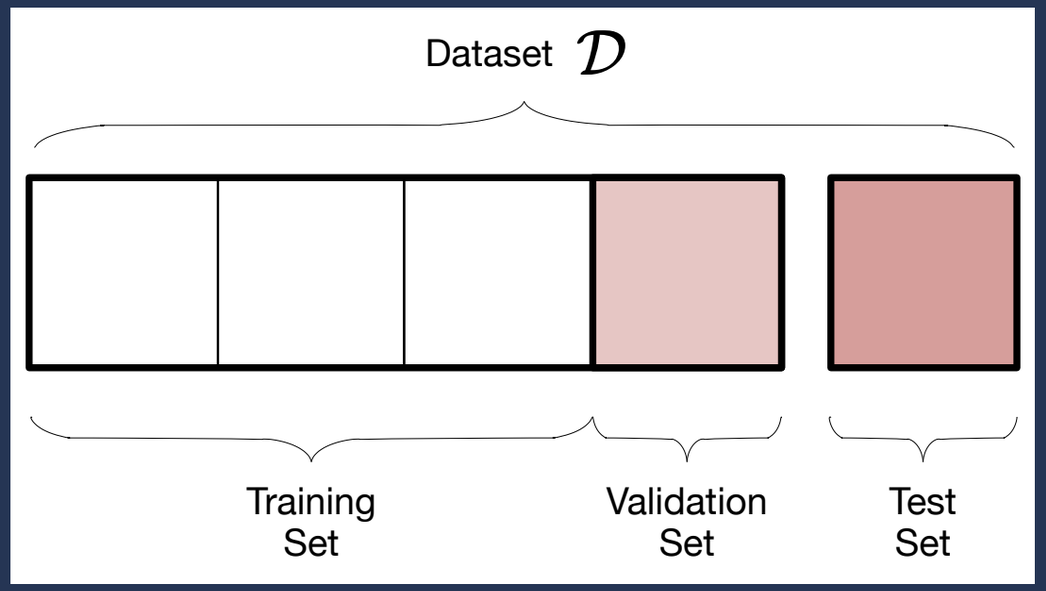
BALANCE THE COMPLEXITY

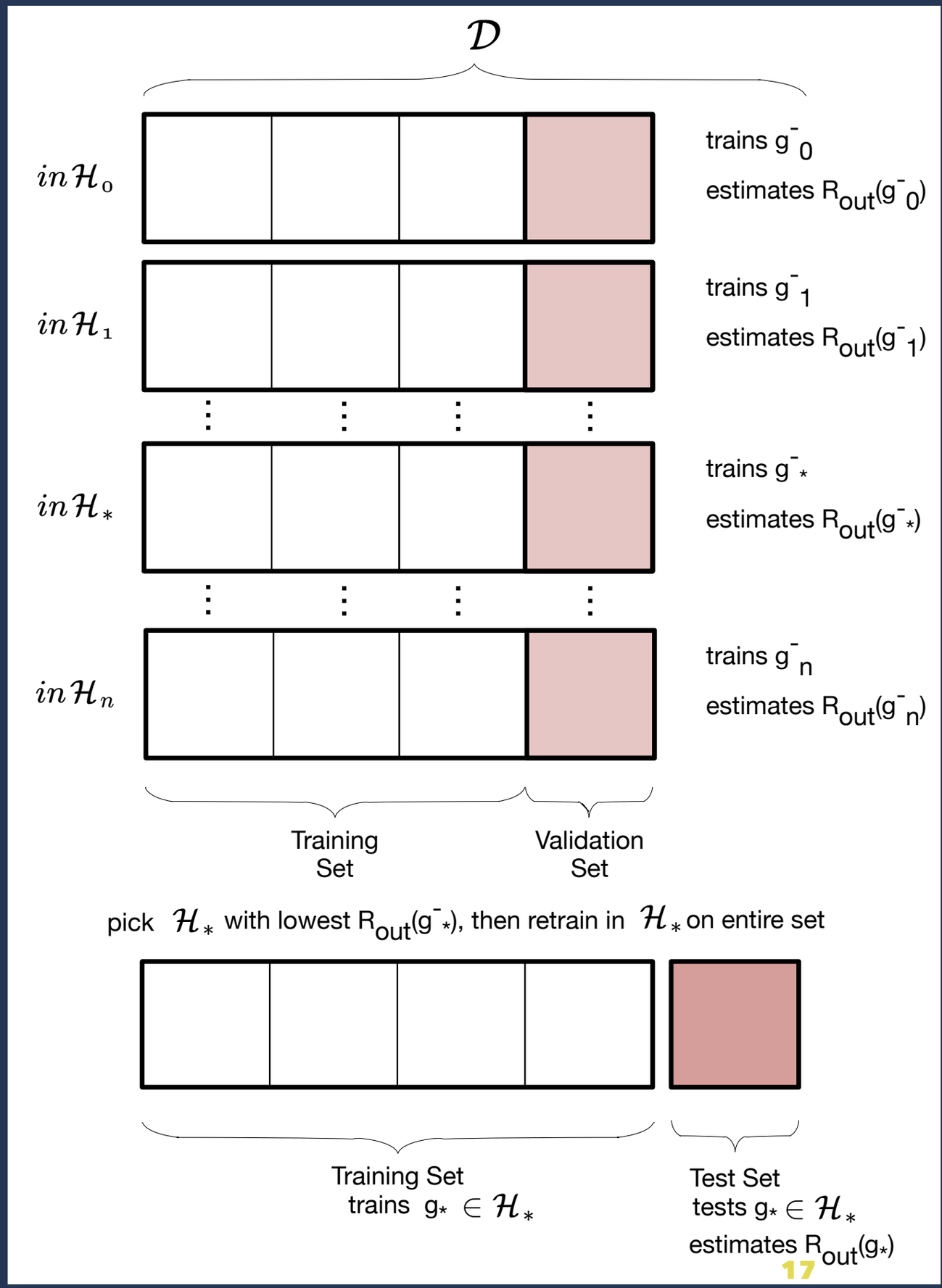




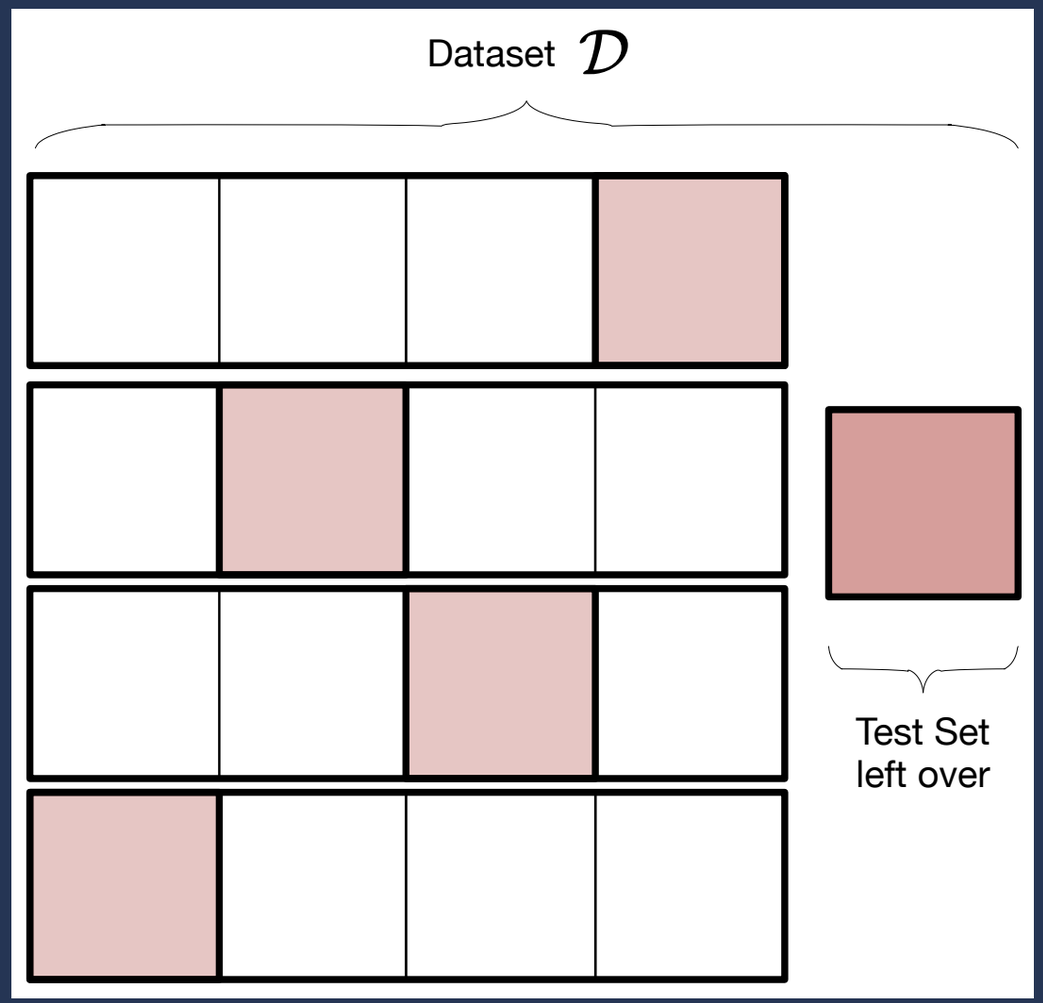
VALIDATION

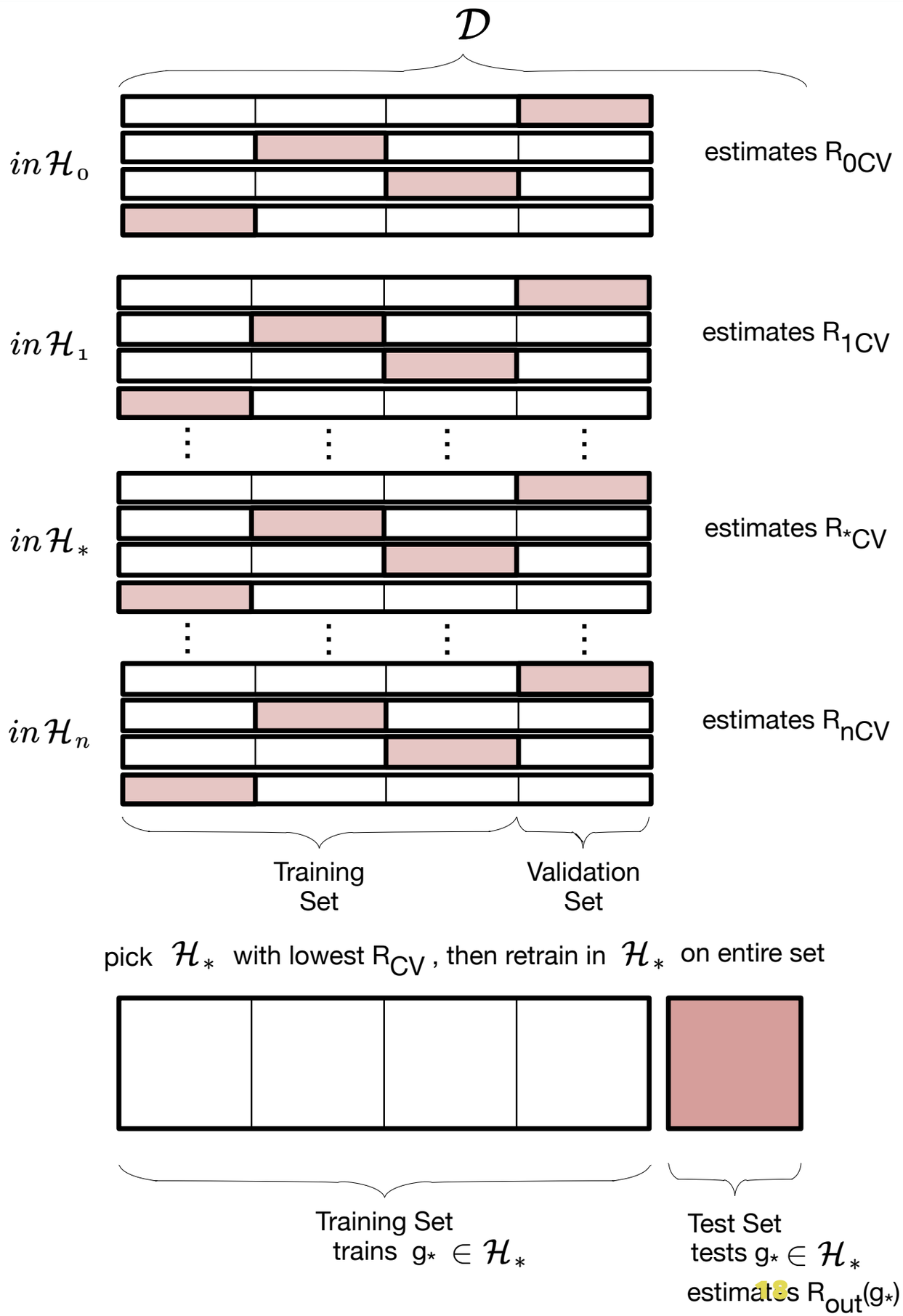
- train-test not enough as we *fit* for d on test set and contaminate it
- thus do train-validate-test





CROSS-VALIDATION





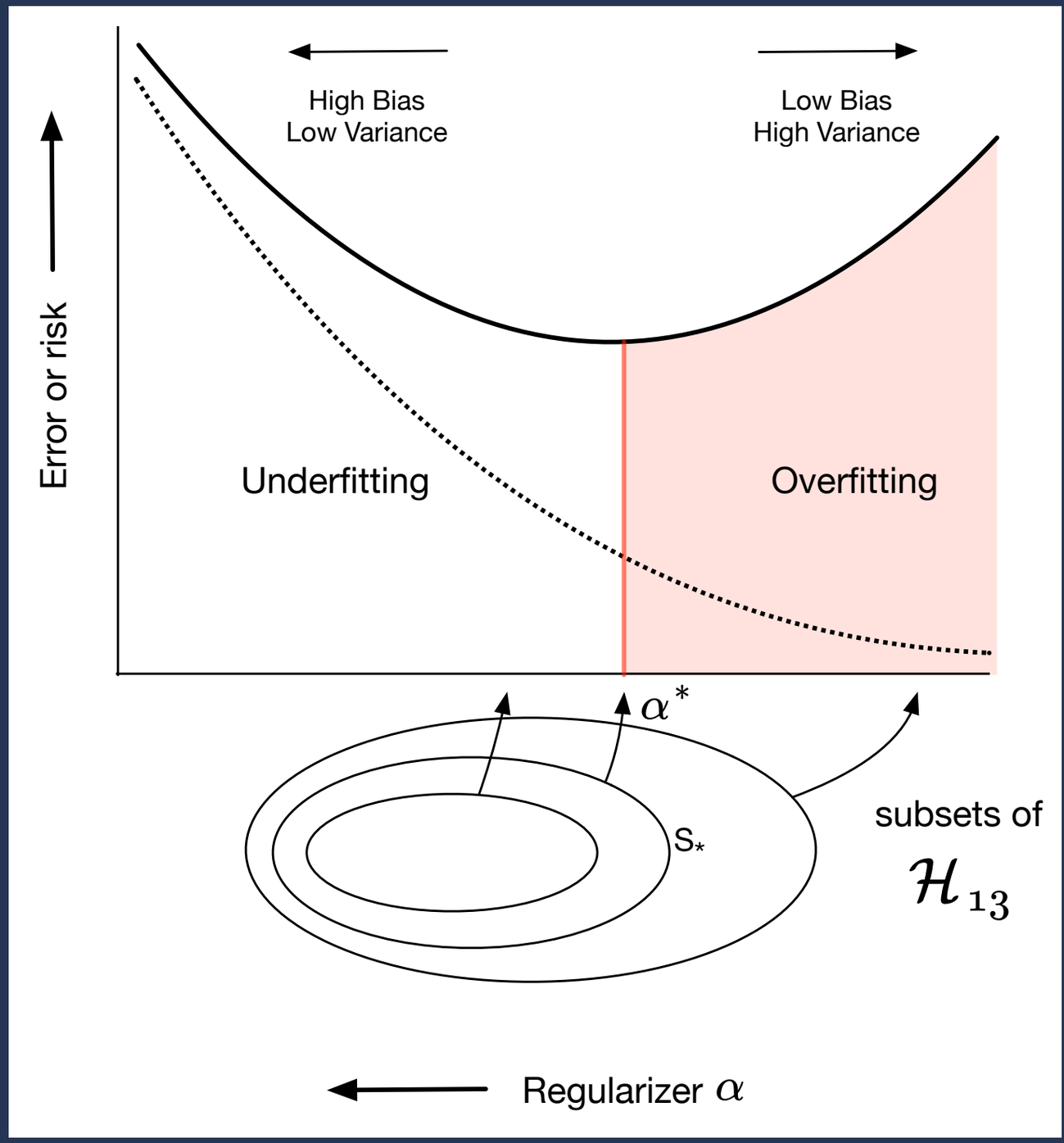
REGULARIZATION

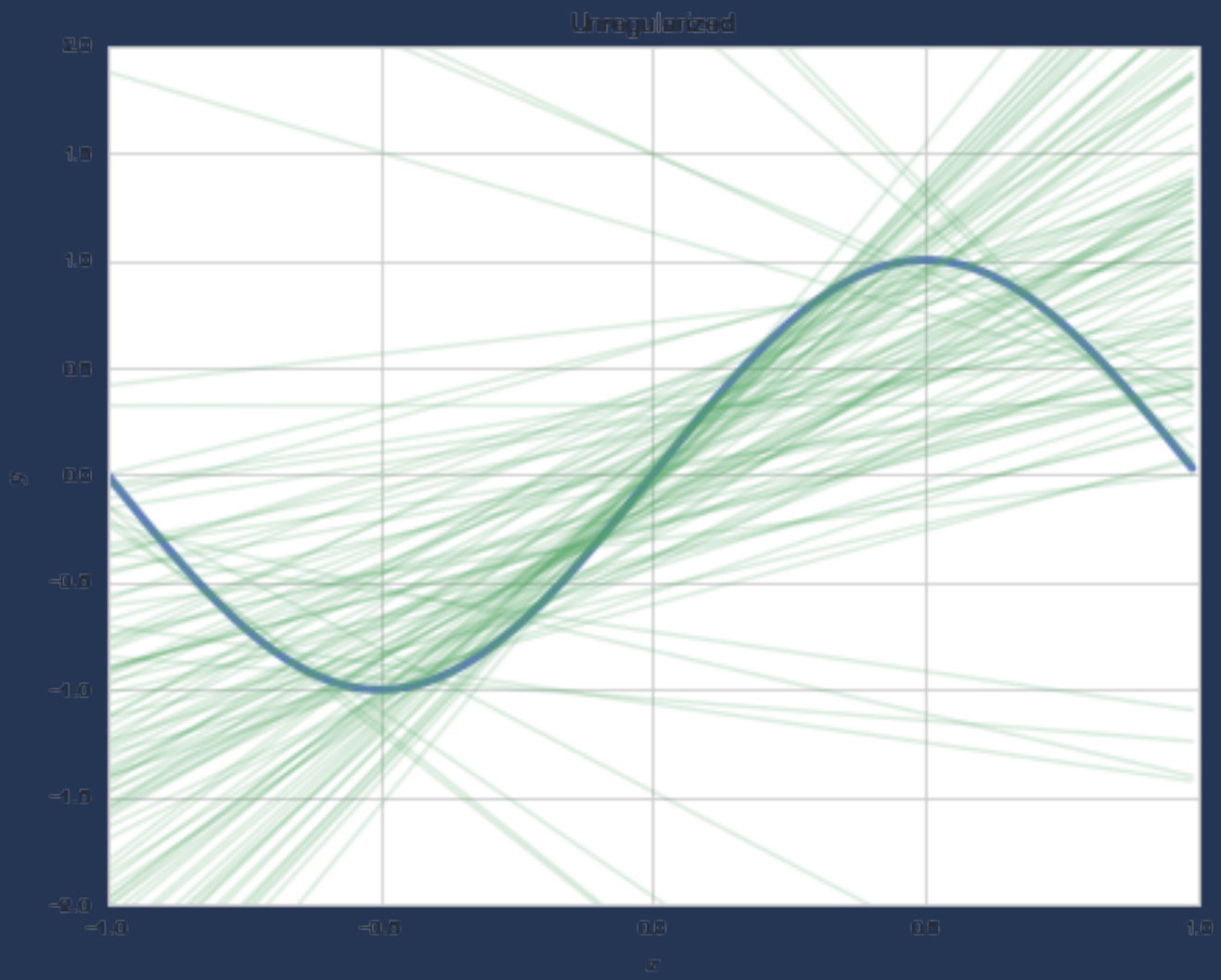
Keep higher a-priori complexity and impose a

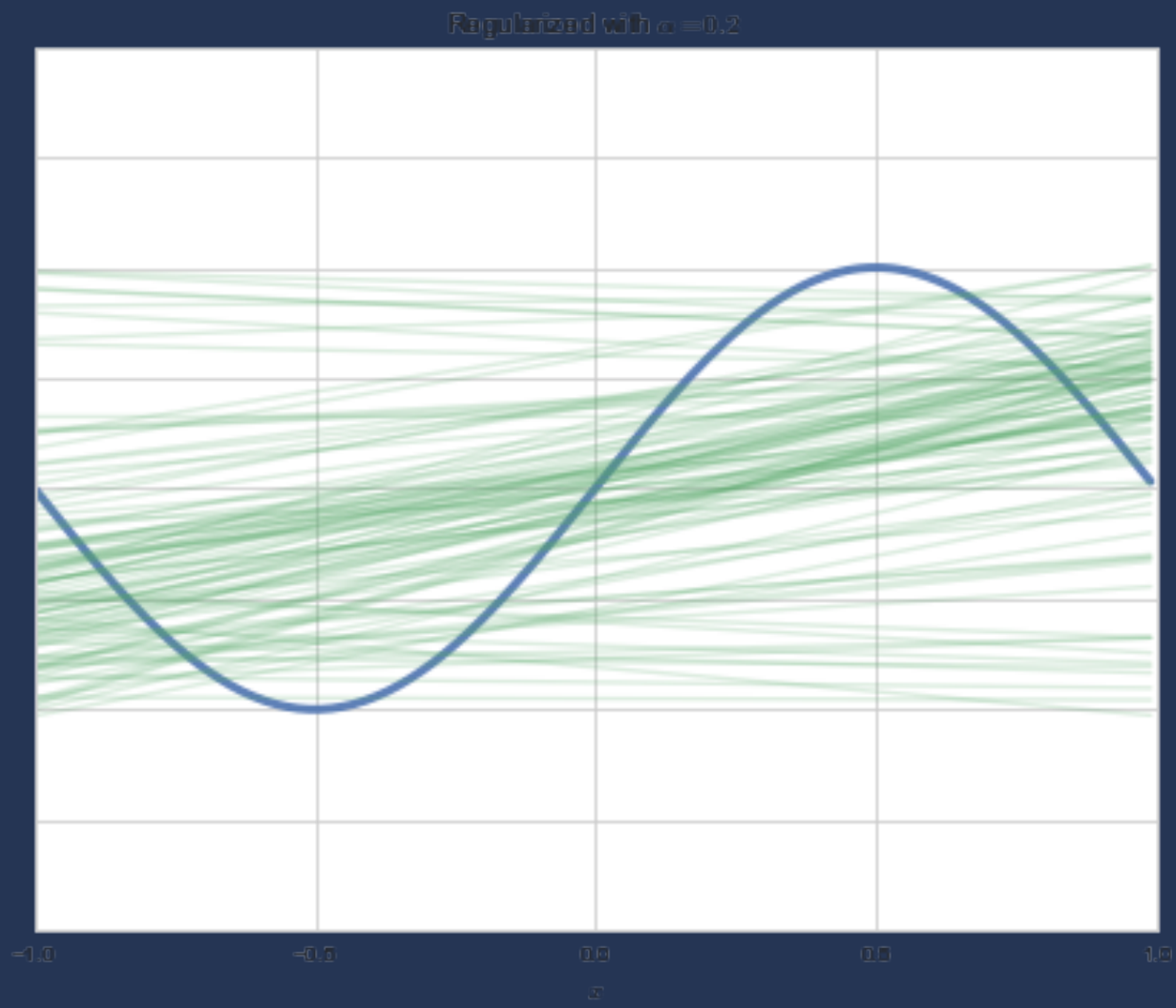
complexity penalty

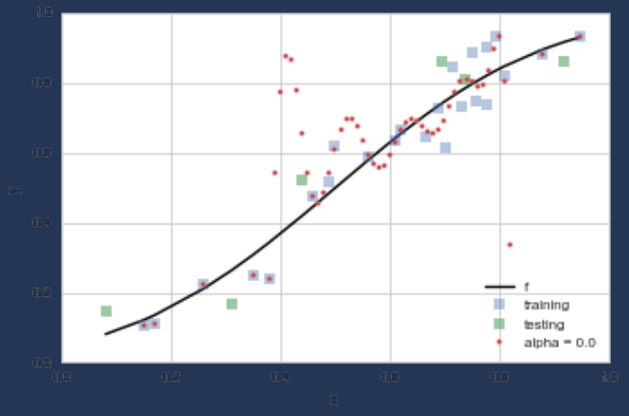
on risk instead, to choose a SUBSET of $\mathcal{H}_{big}$.
We'll make the coefficients small:

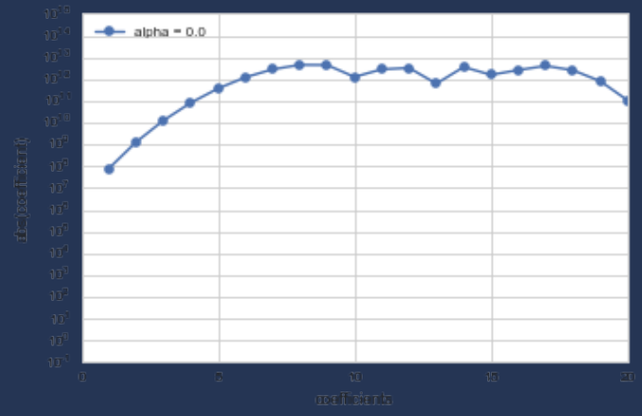
$$\sum_{i=0}^{j} a_i^2 < C.$$

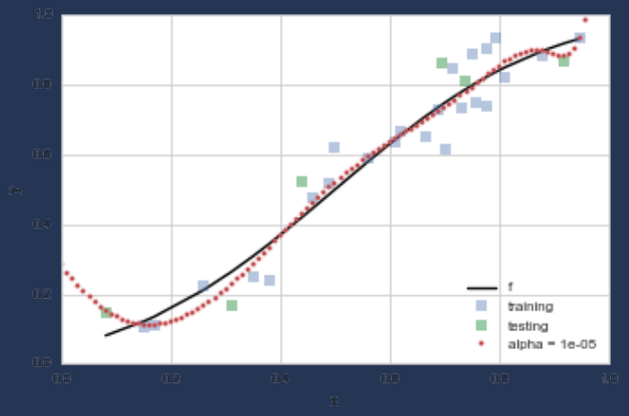


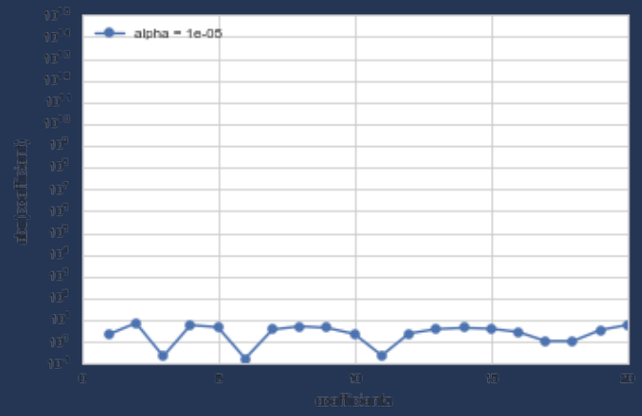

Unregularized
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.0
−0.5
0.0
0.5
1.0

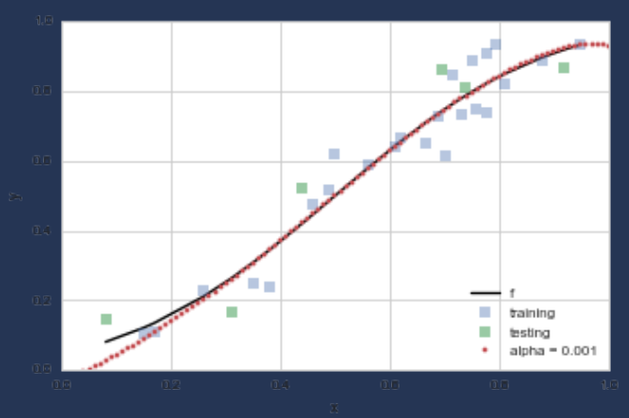

Regularized with α = 0.2
−1.0
−0.5
0.0
0.5
1.0

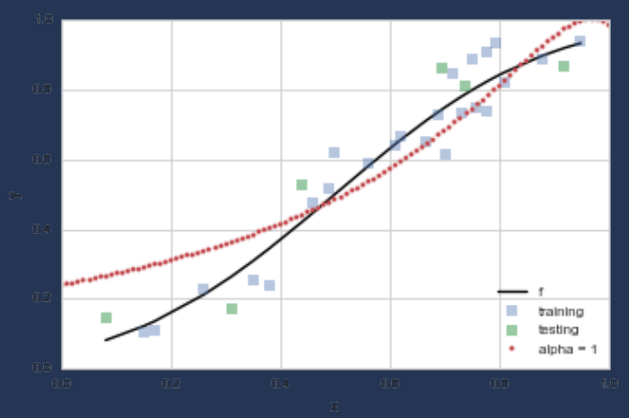


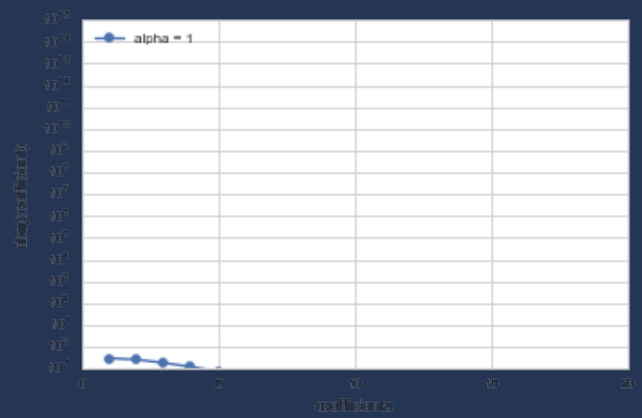












REGULARIZATION

$$\mathcal{R}(h_j) = \sum_{y_i \in \mathcal{D}} (y_i - h_j(x_i))^2 + \alpha \sum_{i=0}^{j} a_i^2.$$

As we increase α, coefficients go towards 0.

Lasso uses $\alpha \sum_{i=0}^{j} |a_i|$, sets coefficients to exactly 0.

Thus regularization automates:

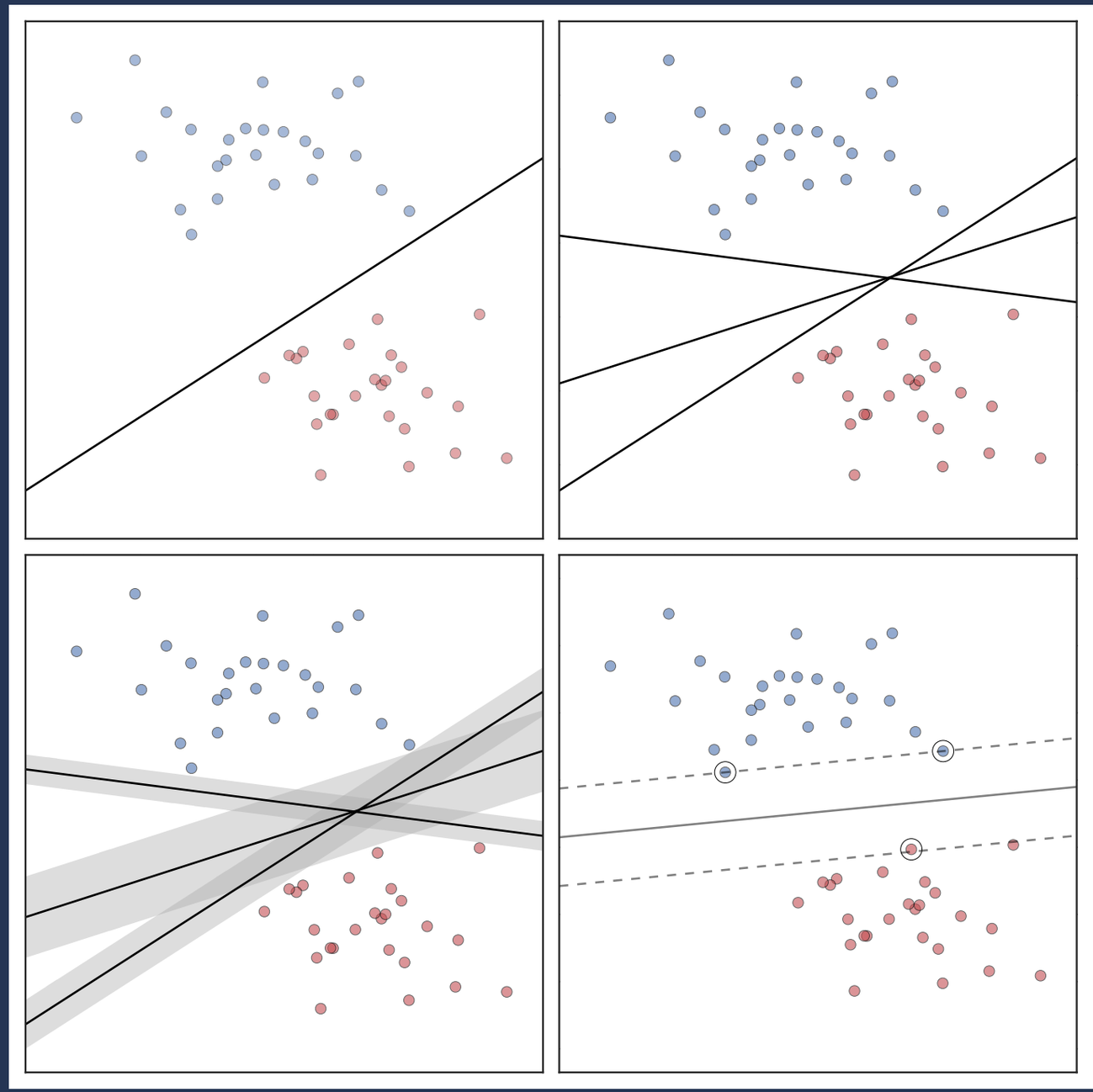
FEATURE ENGINEERING

CLASSIFICATION

BY LINEAR SEPARATION

Which line?

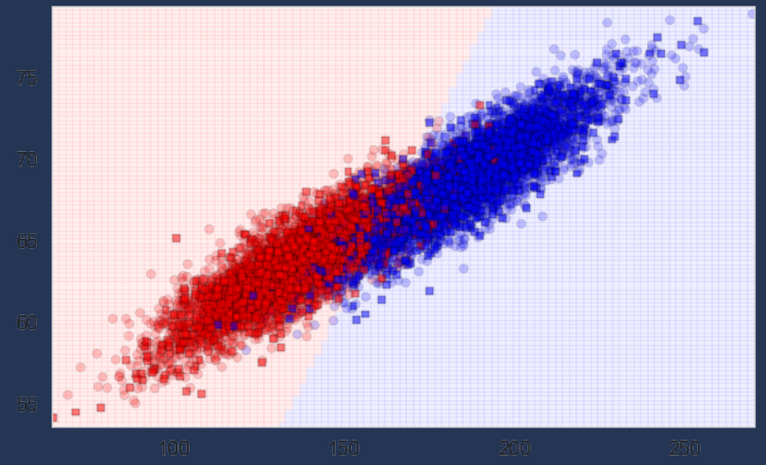
- Different Algorithms, different lines.
- SVM uses max-margin[j]

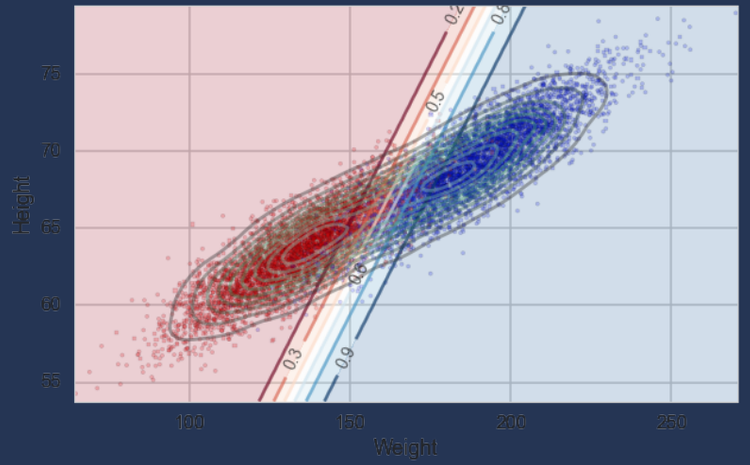


[j] image from code in http://bit.ly/1Azg29G

DISCRIMINATIVE CLASSIFIER

$$P(y|x) : P(male|height, weight)$$

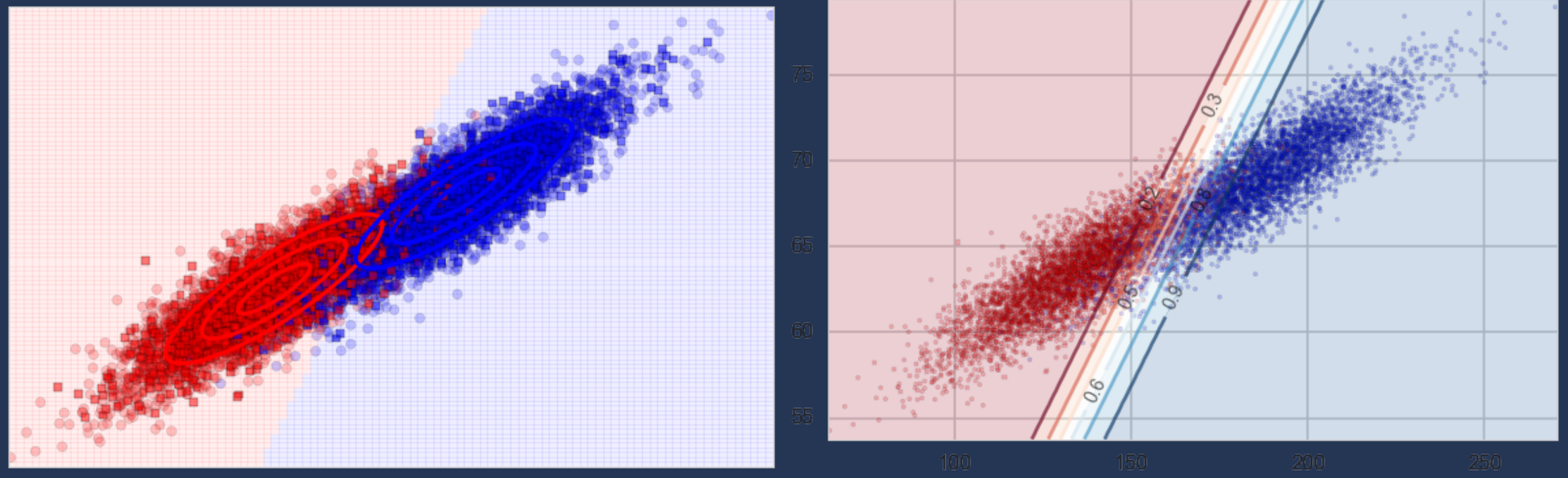




VS. DISCRIMINANT

GENERATIVE CLASSIFIER

$$P(y|x) \propto P(x|y)P(x) : P(height, weight|male) \times P(male)$$

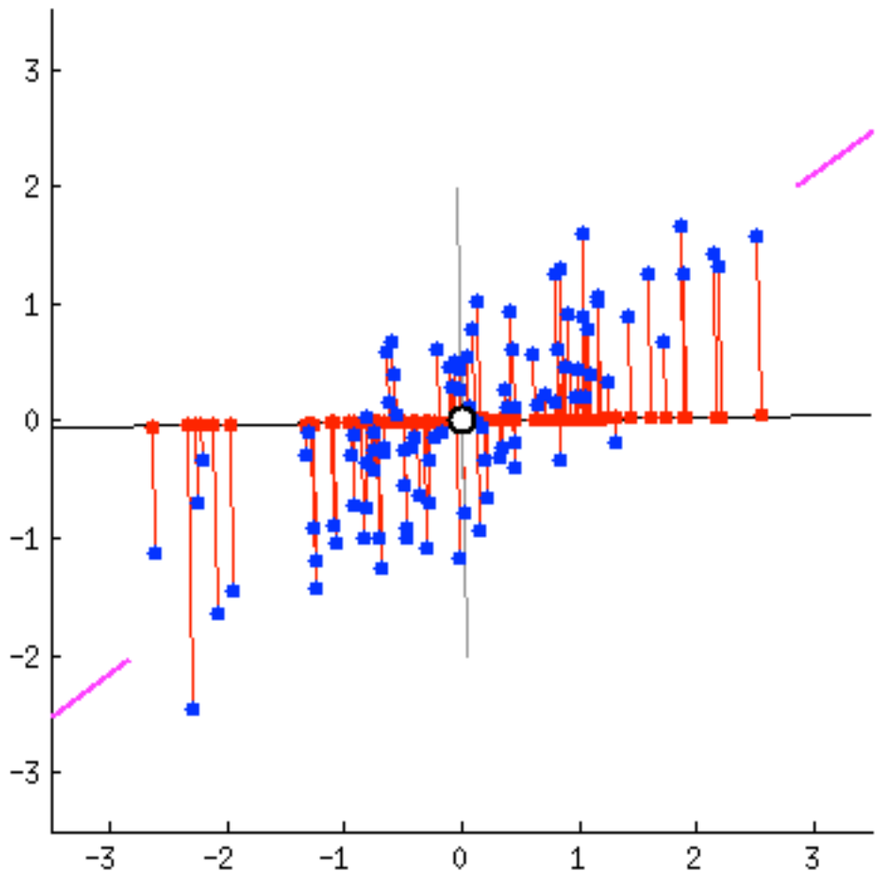


The virtual ATM[:]

[:] Inspired by http://blog.yhathq.com/posts/image-classification-in-Python.html

PCA

unsupervised dim reduction from 332x137x3 to 50



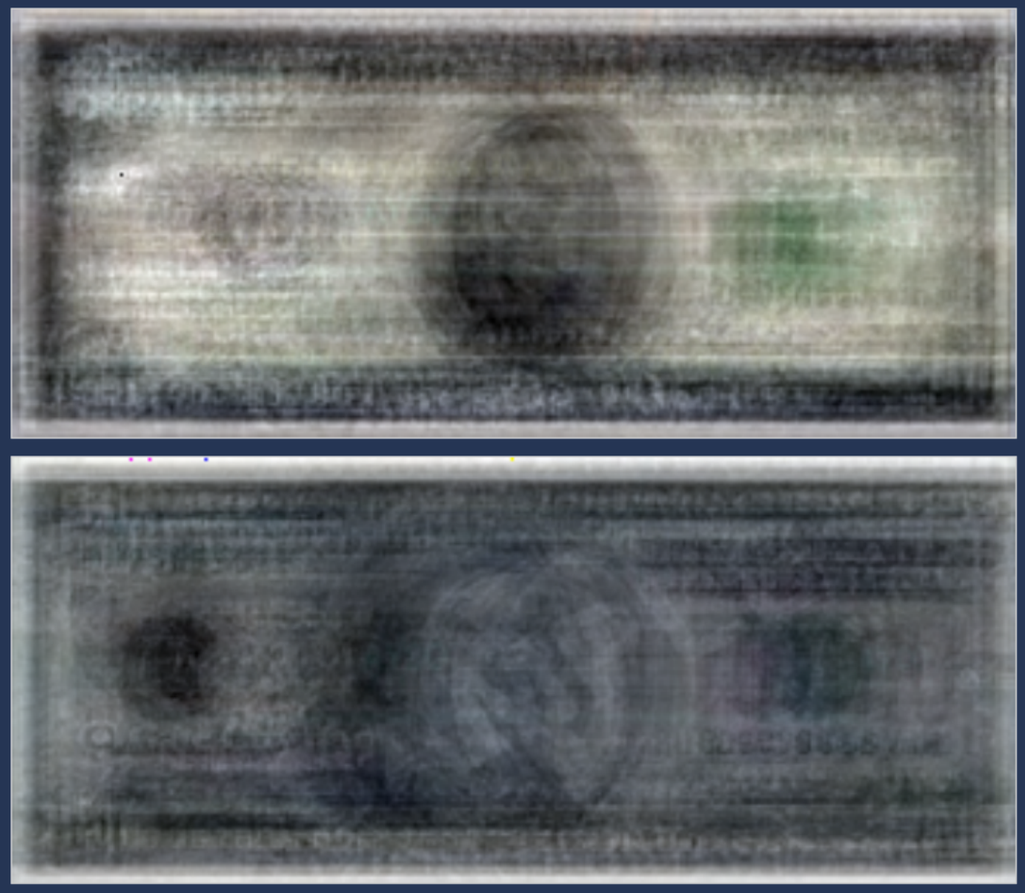
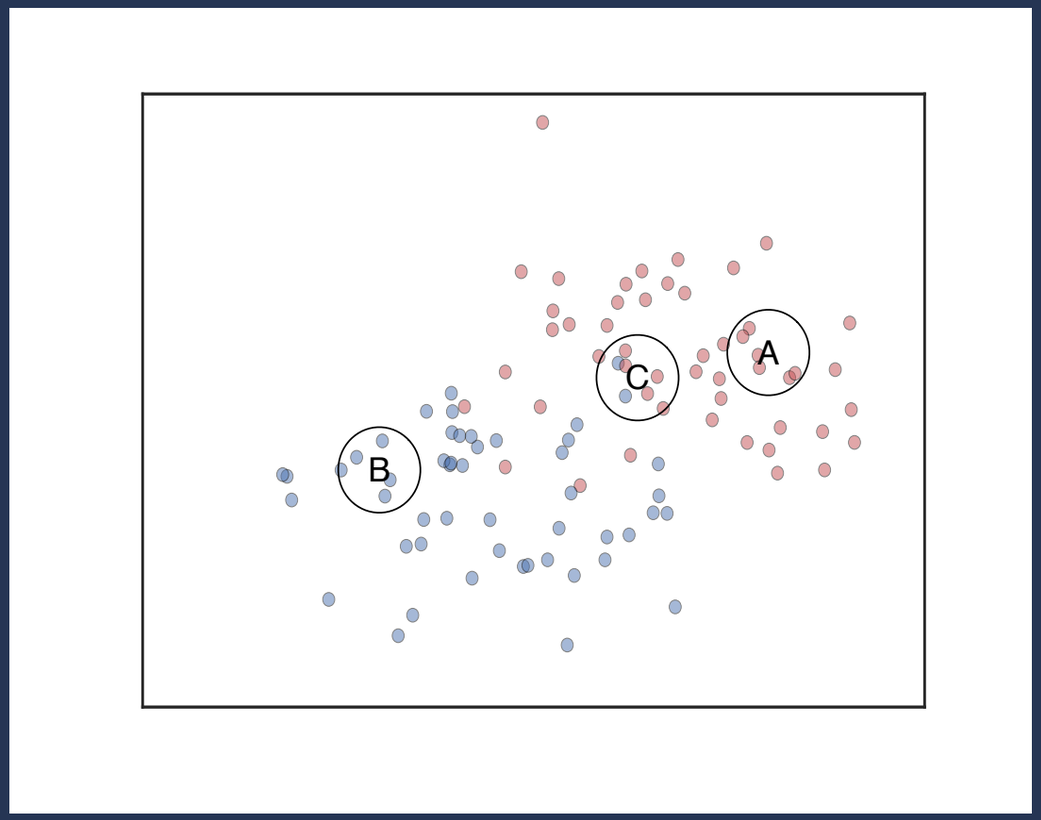
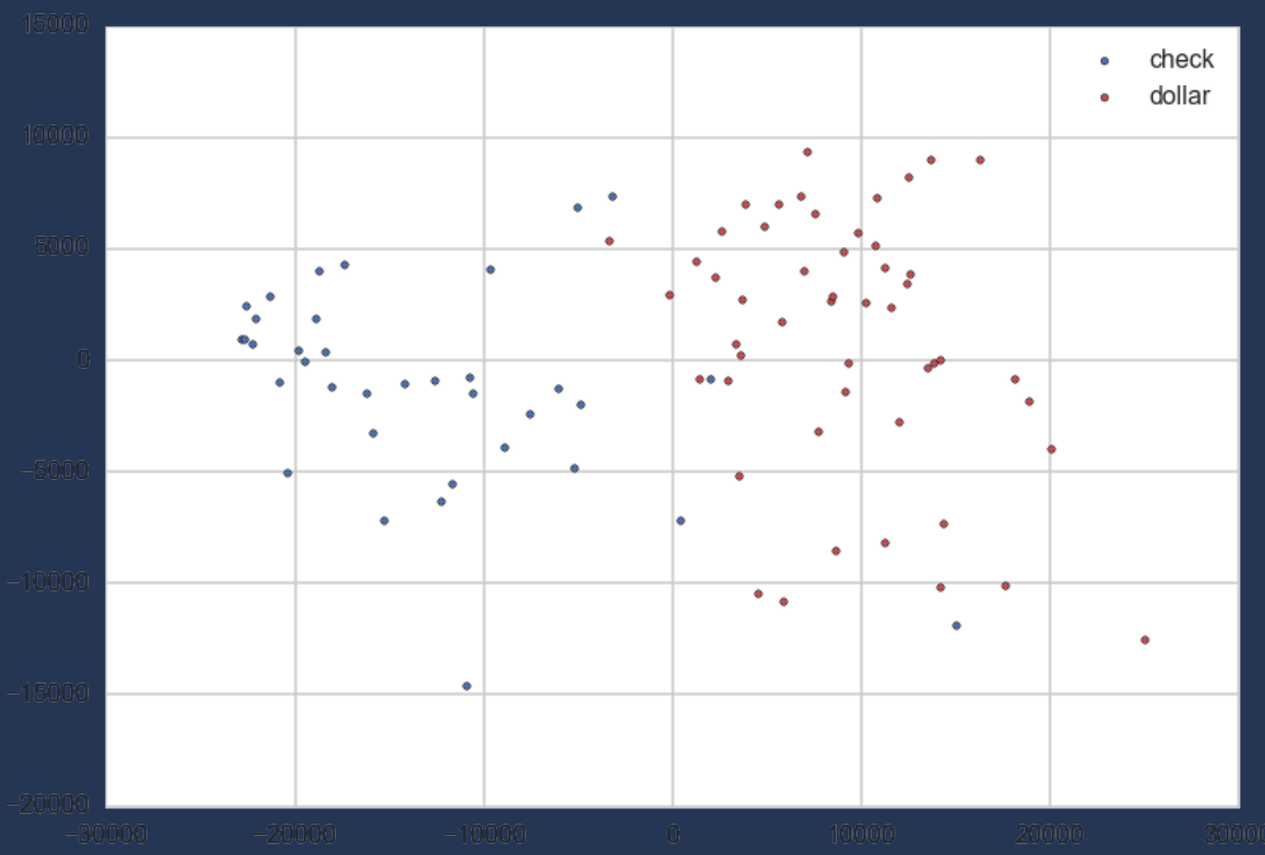


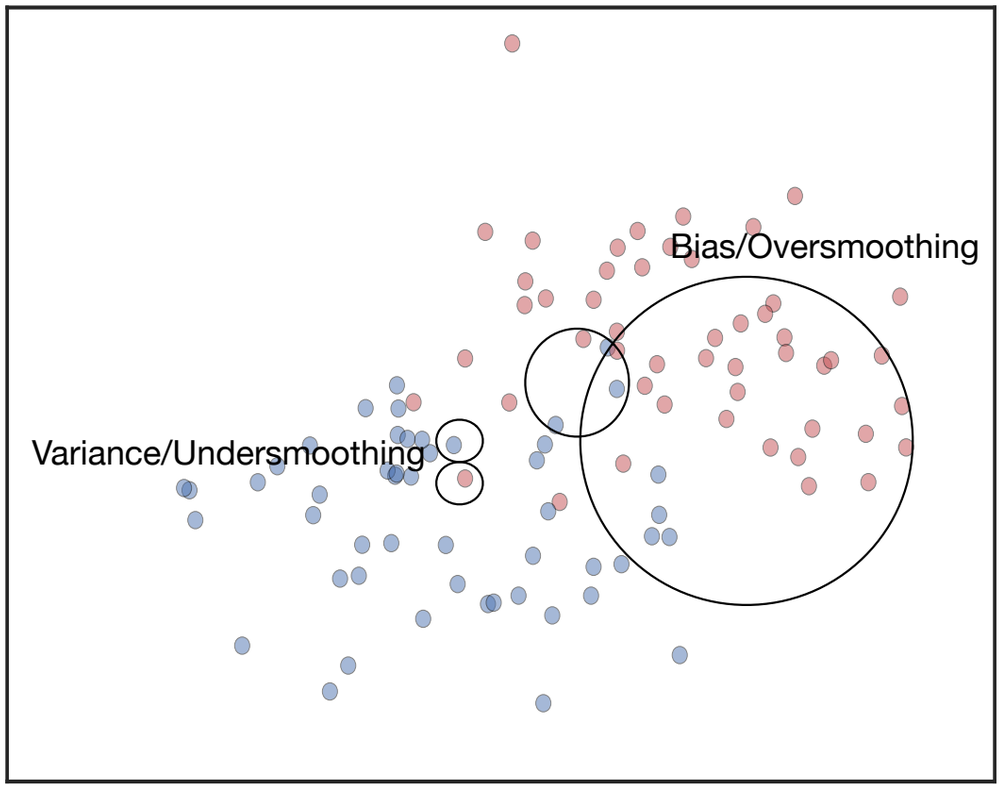
Diagram from http://stats.stackexchange.com/a/140579

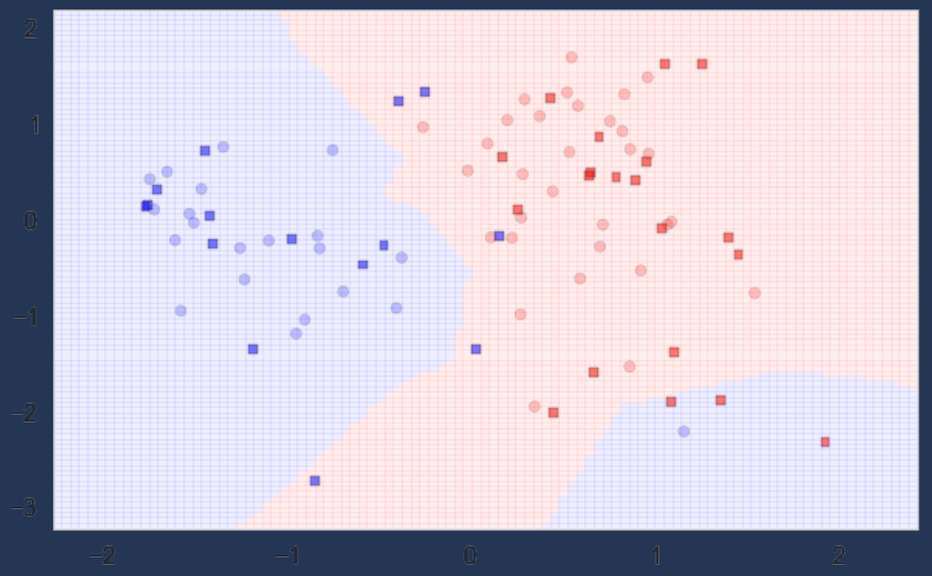
kNN

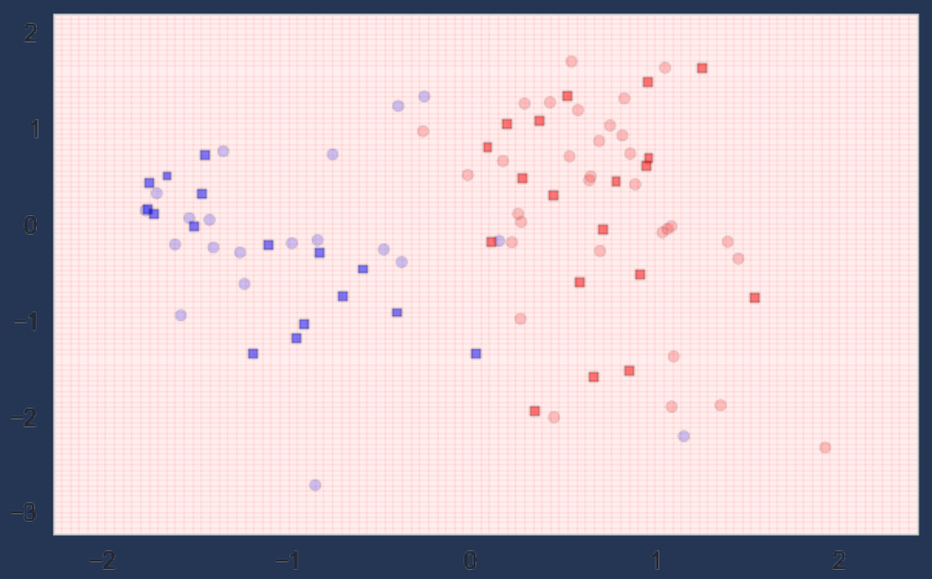

A
B
C


check
dollar
15000
10000
5000
0
−5000
−10000
−15000
−20000
−30000
−20000
−10000
0
10000
20000
30000

BIAS and VARIANCE in KNN

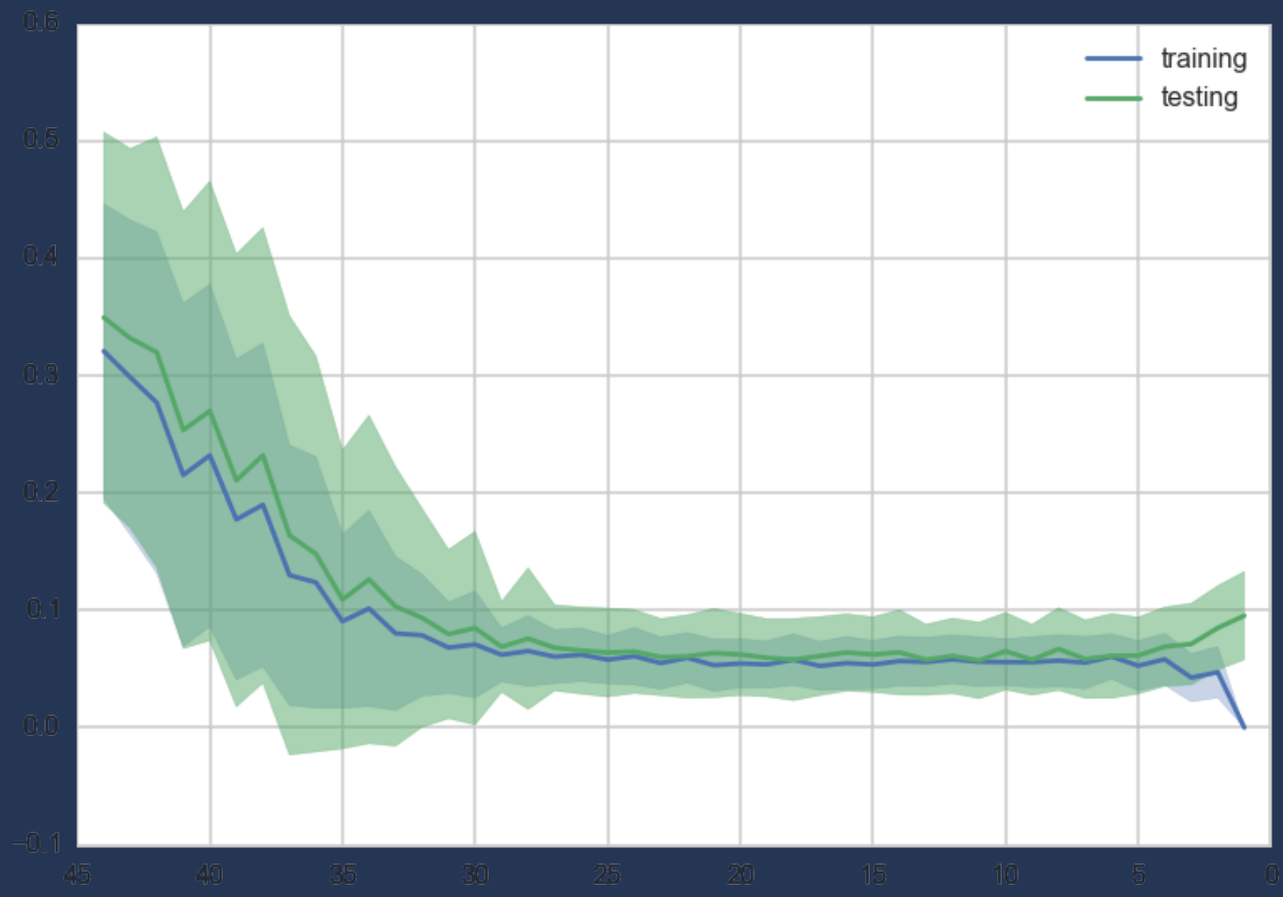


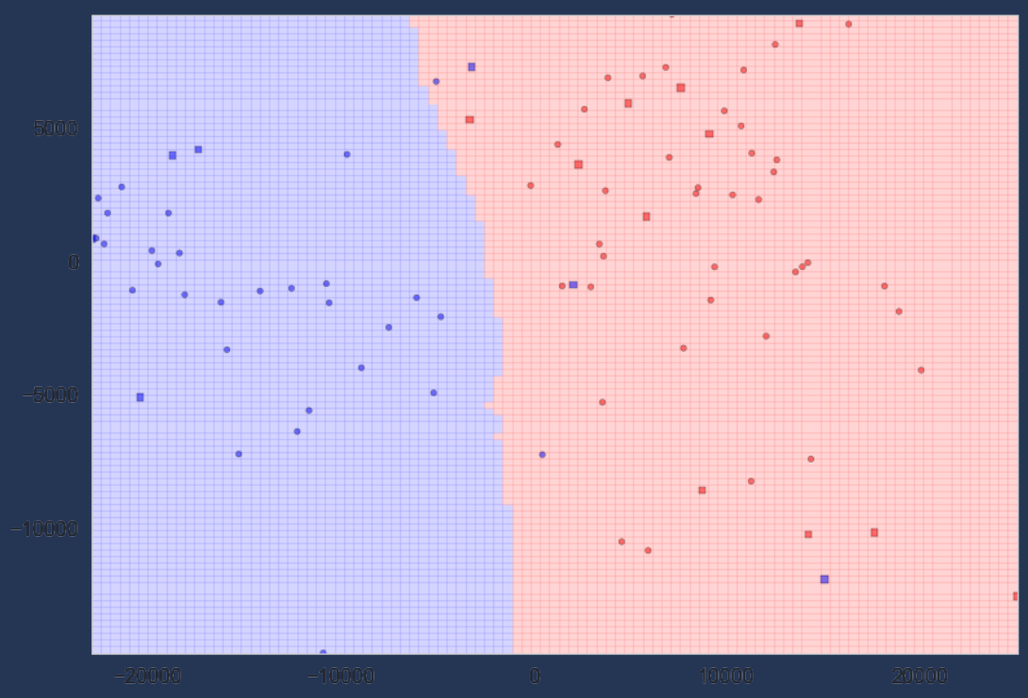




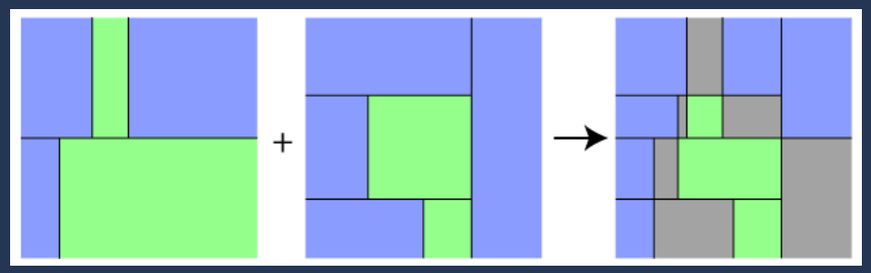
kNN

CROSS-VALIDATED





ENSEMBLE LEARNING

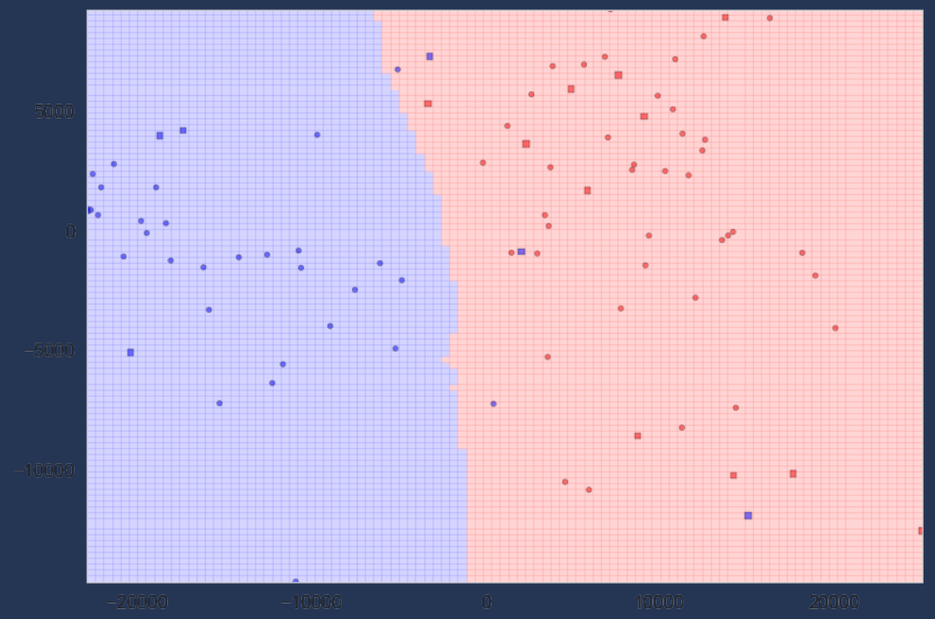


Combine multiple classifiers and vote.

Egs: Decision Trees → random forest, bagging, boosting

EVALUATING CLASSIFIERS

Observed \ Predicted	0	1	
0	TN True Negative	FP False Positive	ON Observed Negative
1	FN False Negative	TP True Positive	OP Observed Positive
	PN Predicted Negative	PP Predicted Positive	

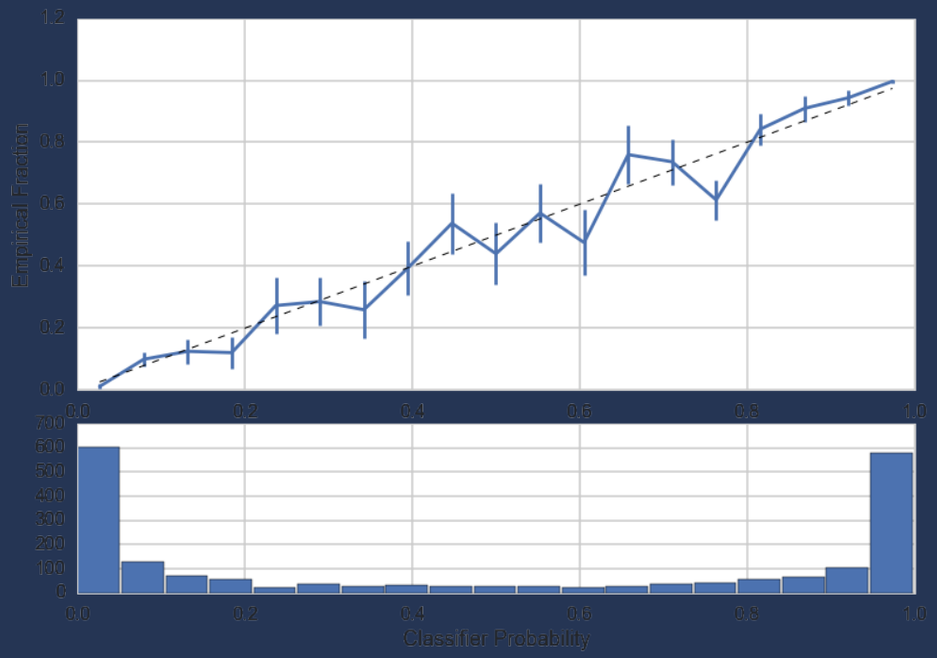


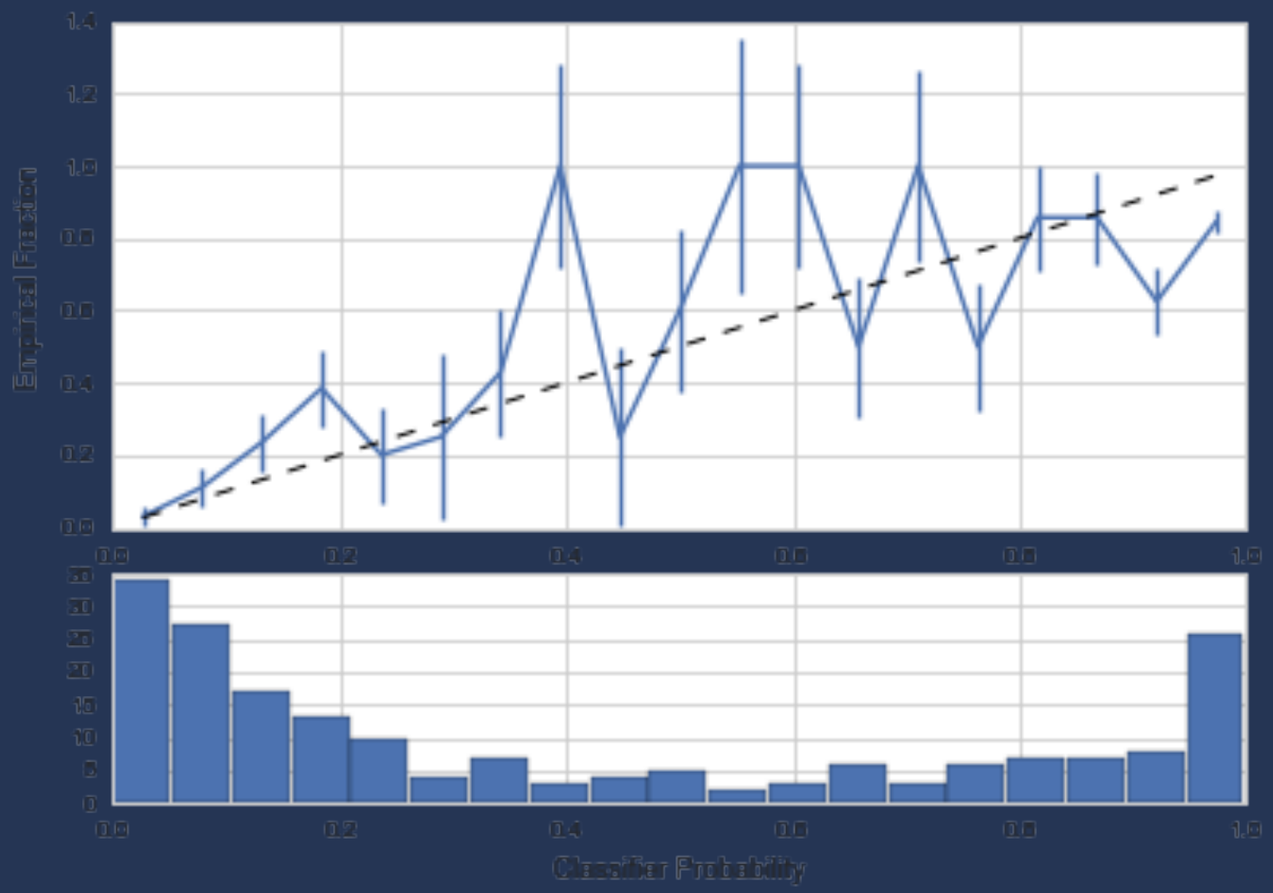

```
confusion_matrix(bestcv.predict(Xtest), ytest)

[[11, 0],
 [3, 4]]
```

checks=blue=1,
dollars=red=0

CLASSIFIER PROBABILITIES

- classifiers output rankings or probabilities
- ought to be well callibrated, or, atleast, similarly ordered





CLASSIFICATION RISK

- $R_{g,\mathcal{D}}(x) = P(y_1|x)\ell(g, y_1) + P(y_0|x)\ell(g, y_0)$
- The usual loss is the 1-0 loss $\ell = 1_{g \neq y}$.
- Thus, $R_{g=y_1}(x) = P(y_0|x)$ and $R_{g=y_0}(x) = P(y_1|x)$

CHOOSE CLASS WITH LOWEST RISK

1 if $R_1 \leq R_0 \implies$ 1 if $P(0|x) \leq P(1|x)$.

choose 1 if $P(1|x) \geq 0.5$! Intuitive!

ASYMMETRIC RISK^

want no checks misclassified as dollars, i.e., no false negatives: $\ell_{10} \neq \ell_{01}$.

Start with $R_g(x) = P(1|x)\ell(g,1) + P(0|x)\ell(g,0)$

$$R_1 = l_{11}p_1 + l_{10}p_0 = l_{10}p_0$$
$$R_0 = l_{01}p_1 + l_{00}p_0 = l_{01}p_1$$

choose 1 if $R_1 < R_0$

---

^ image of breast carcinoma from http://bit.ly/1QgqhBw

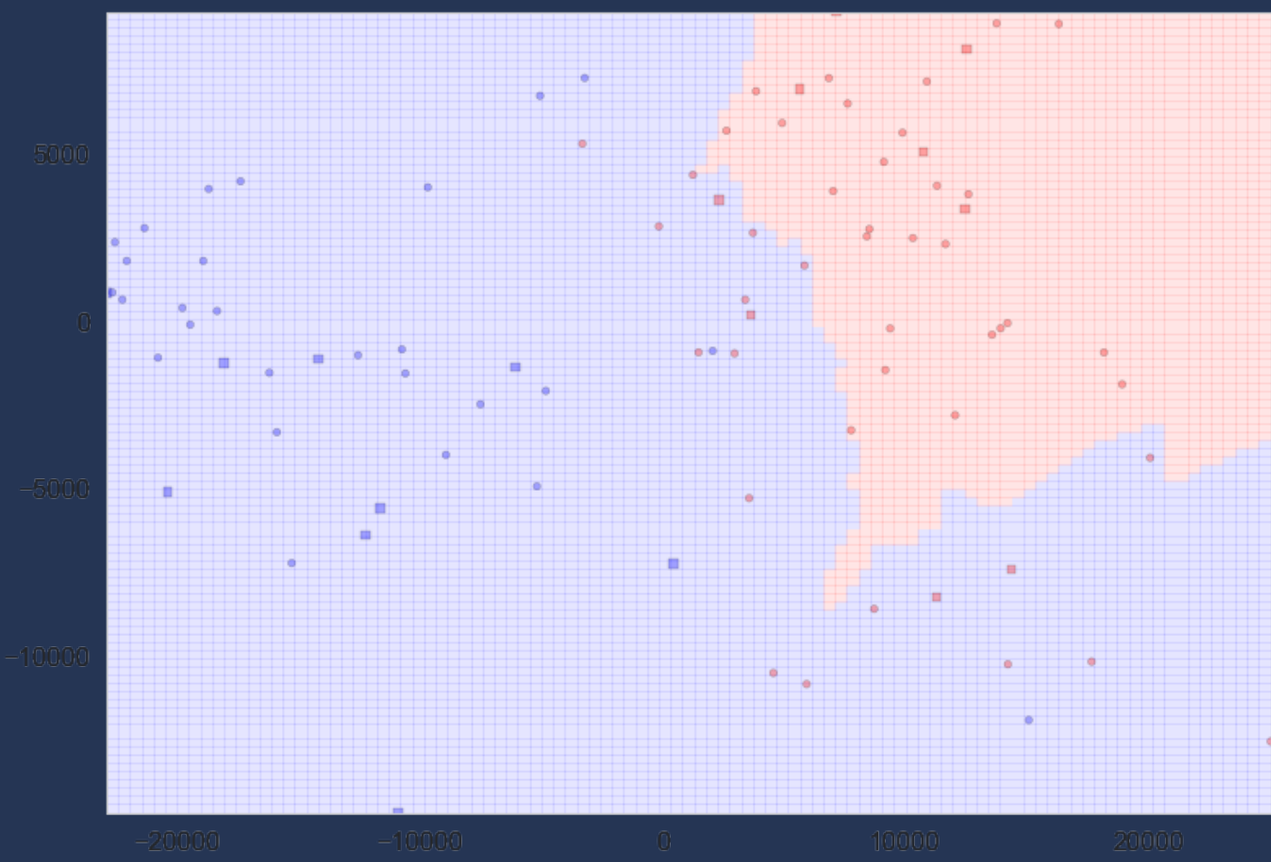
ASYMMETRIC RISK

i.e. $rp_0 < p_1$ where

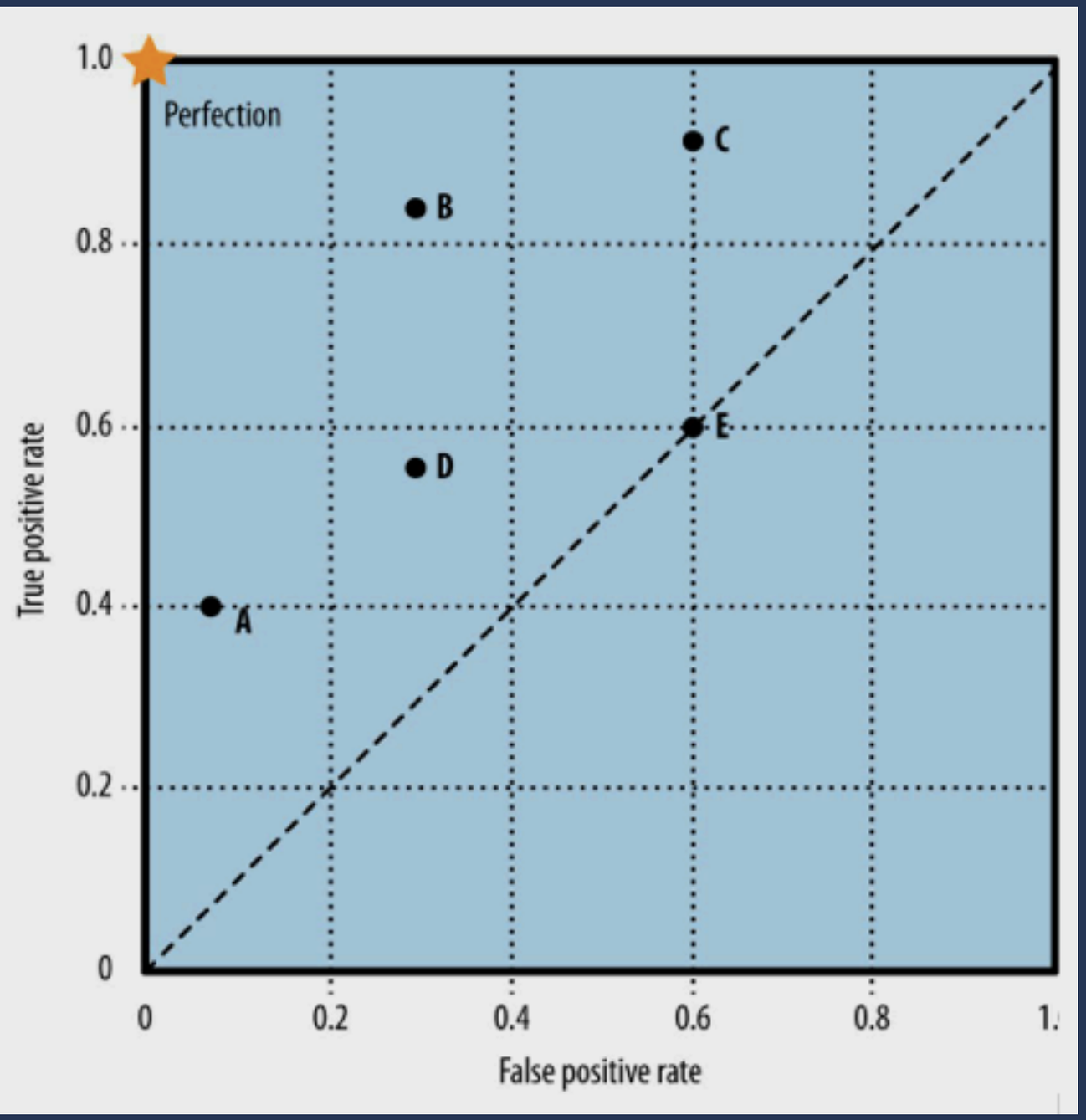
$$r = \frac{l_{10}}{l_{01}} = \frac{l_{FP}}{l_{FN}}$$

Or $P_1 > t$ where $t = \frac{r}{1+r}$

```
confusion_matrix(ytest, bestcv2, r=10, Xtest)
[[5, 4],
 [0, 9]]
```



ROC SPACE[+]



$$TPR = \frac{TP}{OP} = \frac{TP}{TP + FN}.$$

$$FPR = \frac{FP}{ON} = \frac{FP}{FP + TN}$$

Observed	Predicted 0	Predicted 1	
0	TN True Negative	FP False Positive	ON Observed Negative
1	FN False Negative	TP True Positive	OP Observed Positive
	PN Predicted Negative	PP Predicted Positive	

[+] this+next fig: Data Science for Business, Foster et. al.

matrix from matrix

Instance	Class	Score
.	.	.
.	.	.
.	.	.
.....	p	0.65
.....	p	0.71
.....	n	0.74
.....	p	0.75
.....	n	0.80
.....	p	0.84
.....	p	0.85
.....	p	0.87
.....	n	0.93
.....	p	0.95
.....	n	0.96
.....	p	0.98
.....	p	0.99

	p	n
Y	8	4
N	92	96

(0.04, 0.0

	p	n
Y	2	1
N	98	99

(0.01, 0.0

	p	n
Y	2	0
N	98	100

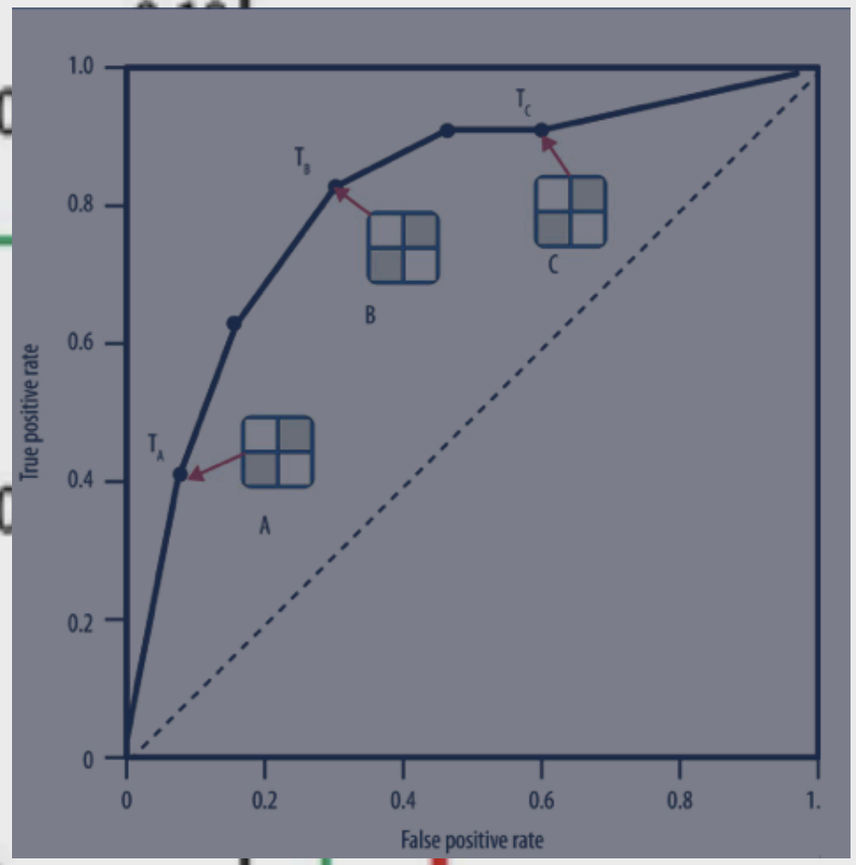
(0, 0.02)

	p	n
Y	0	
N	9	

(01)

	p	n
Y	0	0
N	100	100

(0, 0)



0.04

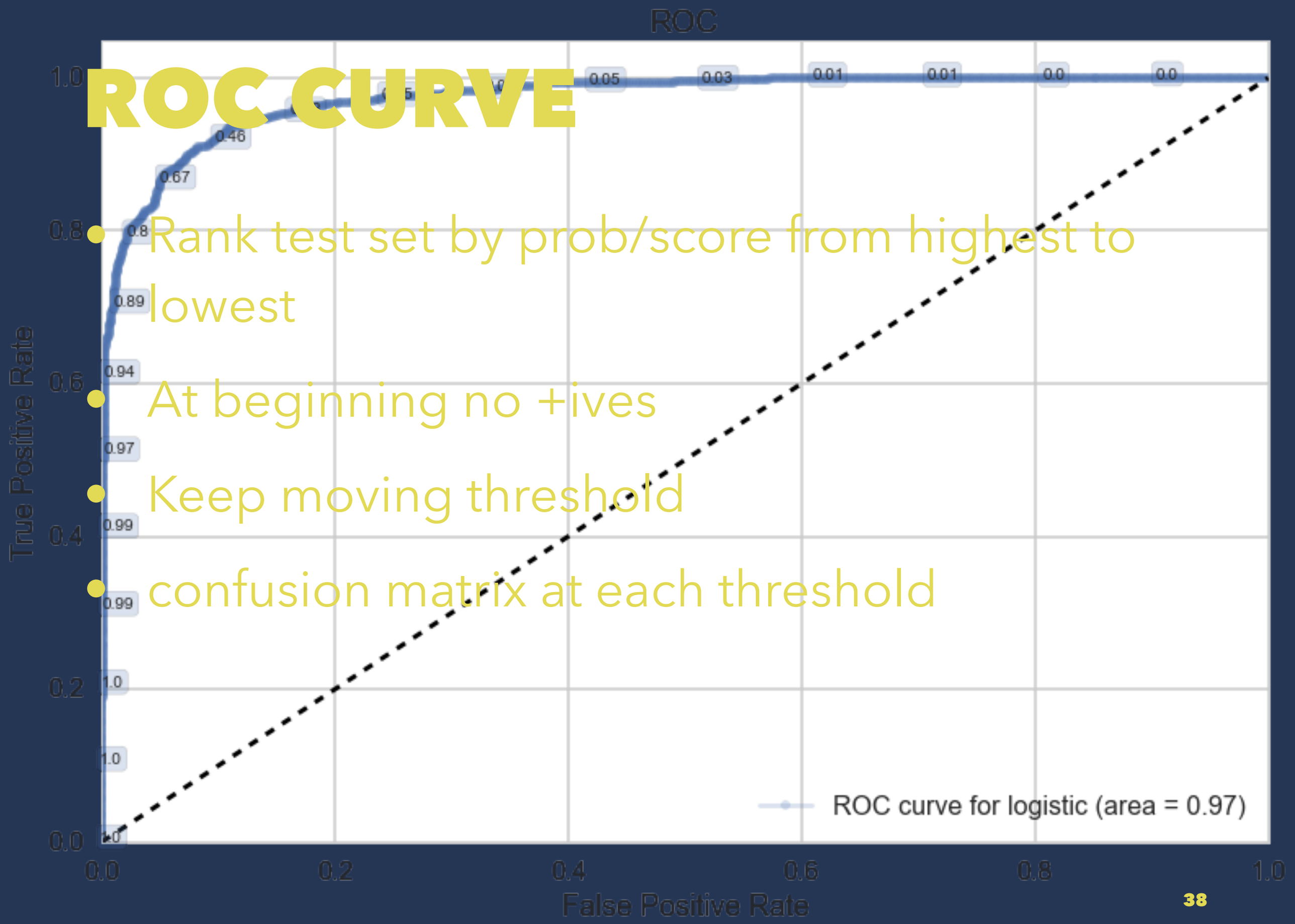
0.02

0

ROC CURVE

ROC CURVE

- Rank test set by prob/score from highest to lowest
- At beginning no +ives
- Keep moving threshold
- confusion matrix at each threshold



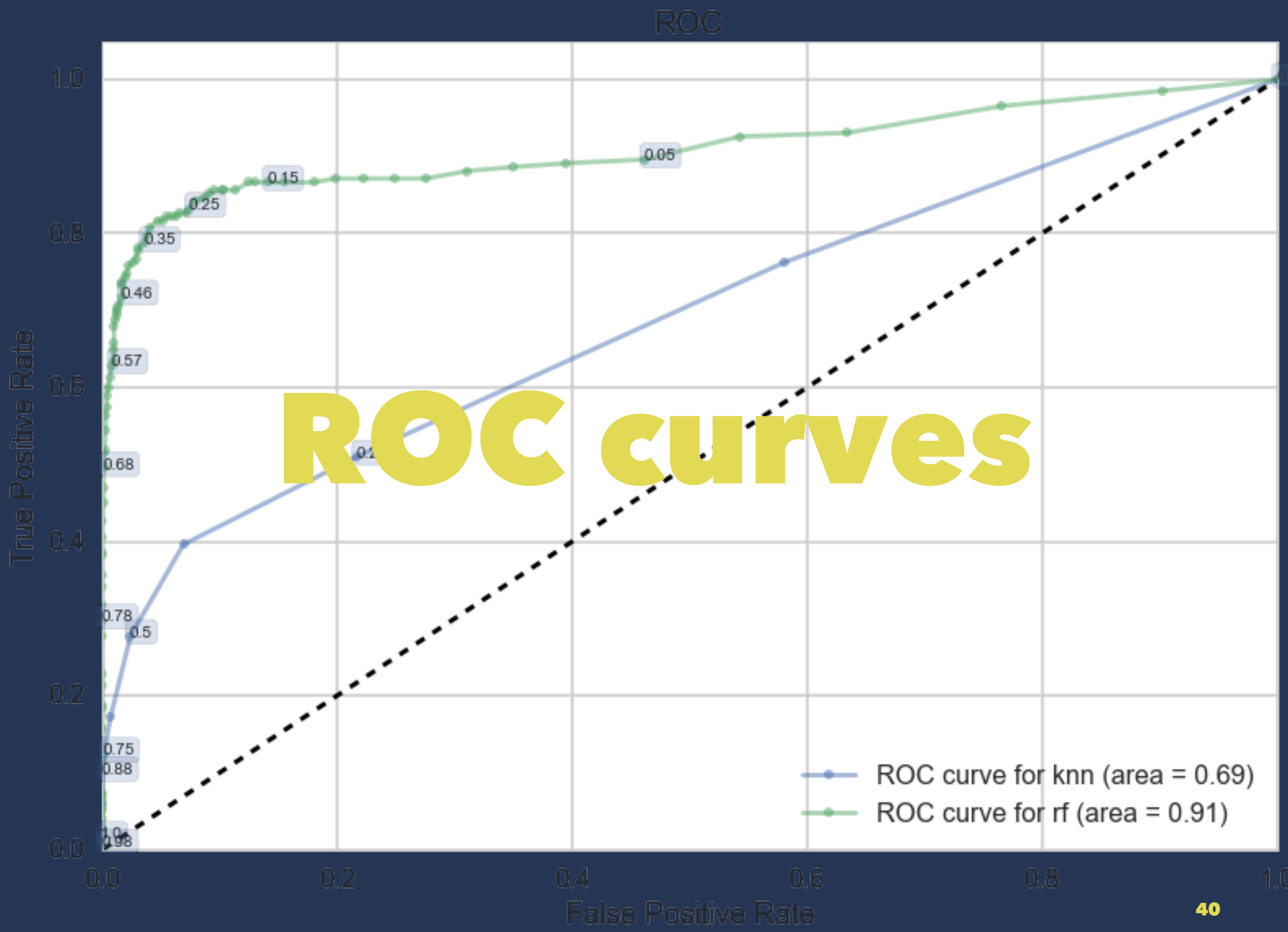
COMPARING CLASSIFERS

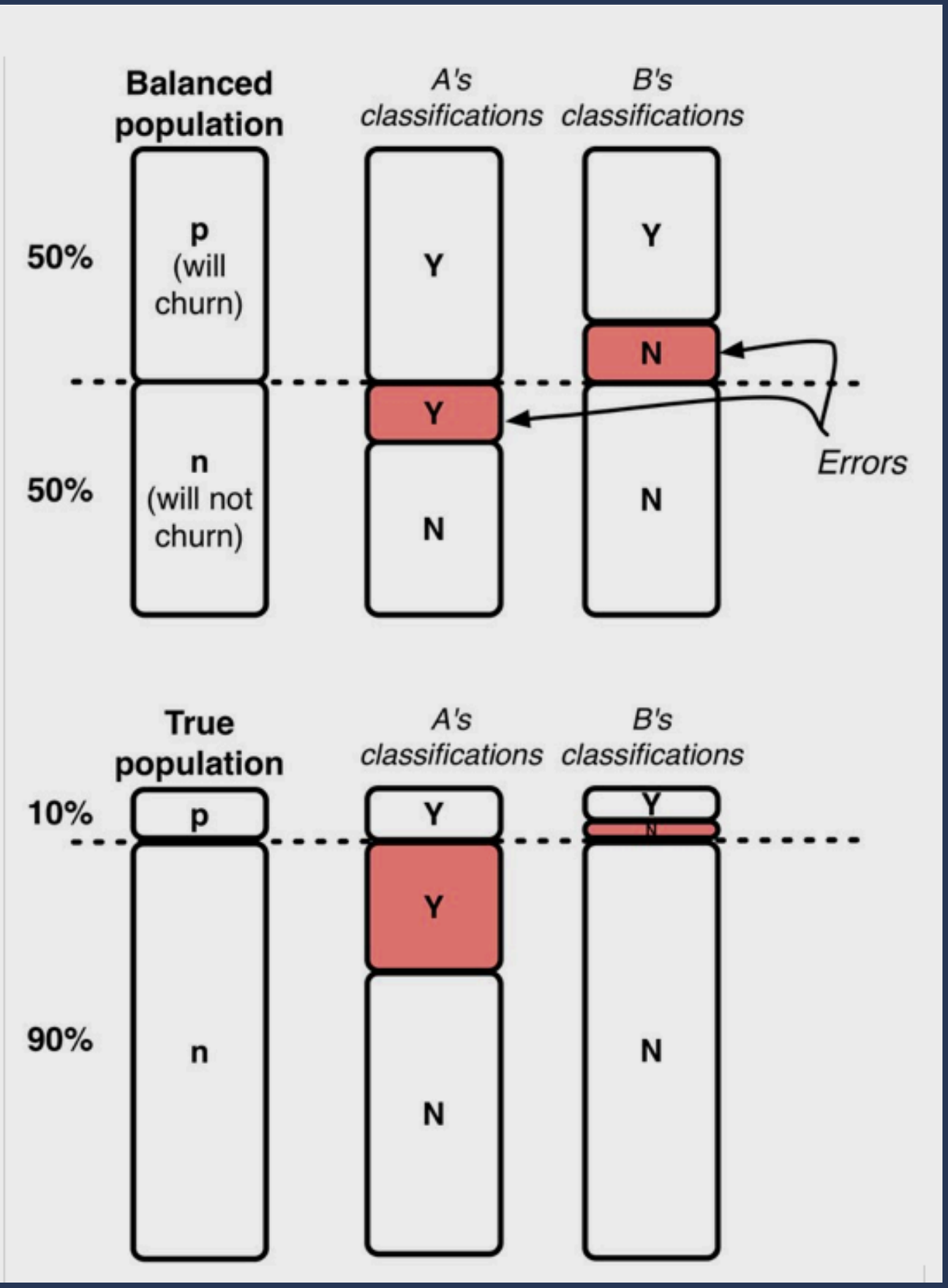
Telecom customer Churn data set from @YhatHQ[<]

VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False.
26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False.
0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False.
0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.
0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.

[<] http://blog.yhathq.com/posts/predicting-customer-churn-with-sklearn.html

ROC curves


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.05
0.15
0.25
0.35
0.46
0.57
0.68
0.78
0.5
0.75
0.88
ROC curve for knn (area = 0.69)
ROC curve for rf (area = 0.91)



ASYMMETRIC CLASSES

- A has large FP[#]
- B has large FN.
- On asymmetric data sets, A will do very bad.
- But is it so?

[#] figure from Data Science for Business, Foster et. al.

EXPECTED VALUE FORMALISM

Can be used for risk or profit/utility (negative risk)

$$EP = p(1,1)\ell_{11} + p(0,1)\ell_{10} + p(0,0)\ell_{00} + p(1,0)\ell_{01}$$

$$\begin{aligned} EP = p_a(1)[TPR\,\ell_{11} + (1-TPR)\ell_{10}] \\ + p_a(0)[(1-FPR)\ell_{00} + FPR\,\ell_{01}] \end{aligned}$$

Fraction of test set pred to be positive $x = PP/N$:

$$x = (TP + FP)/N = TPR\,p_o(1) + FPR\,p_o(0)$$

ASYMMETRIC CLASSES

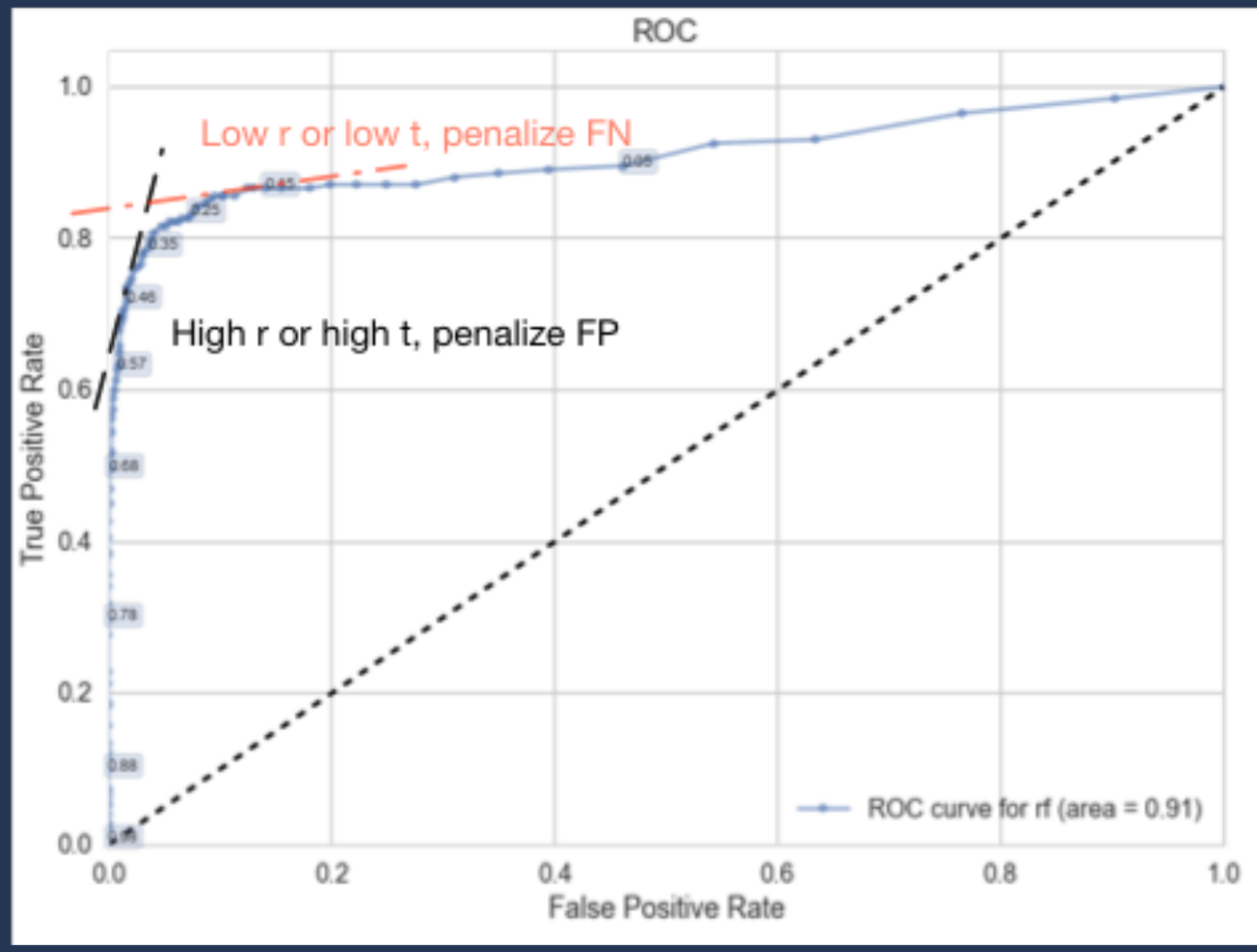
$$r = \frac{l_{FP}}{l_{FN}}$$

Equicost lines with slope

$$\frac{r\,p(0)}{p(1)} = \frac{r\,p(-)}{r\,p(+)}$$

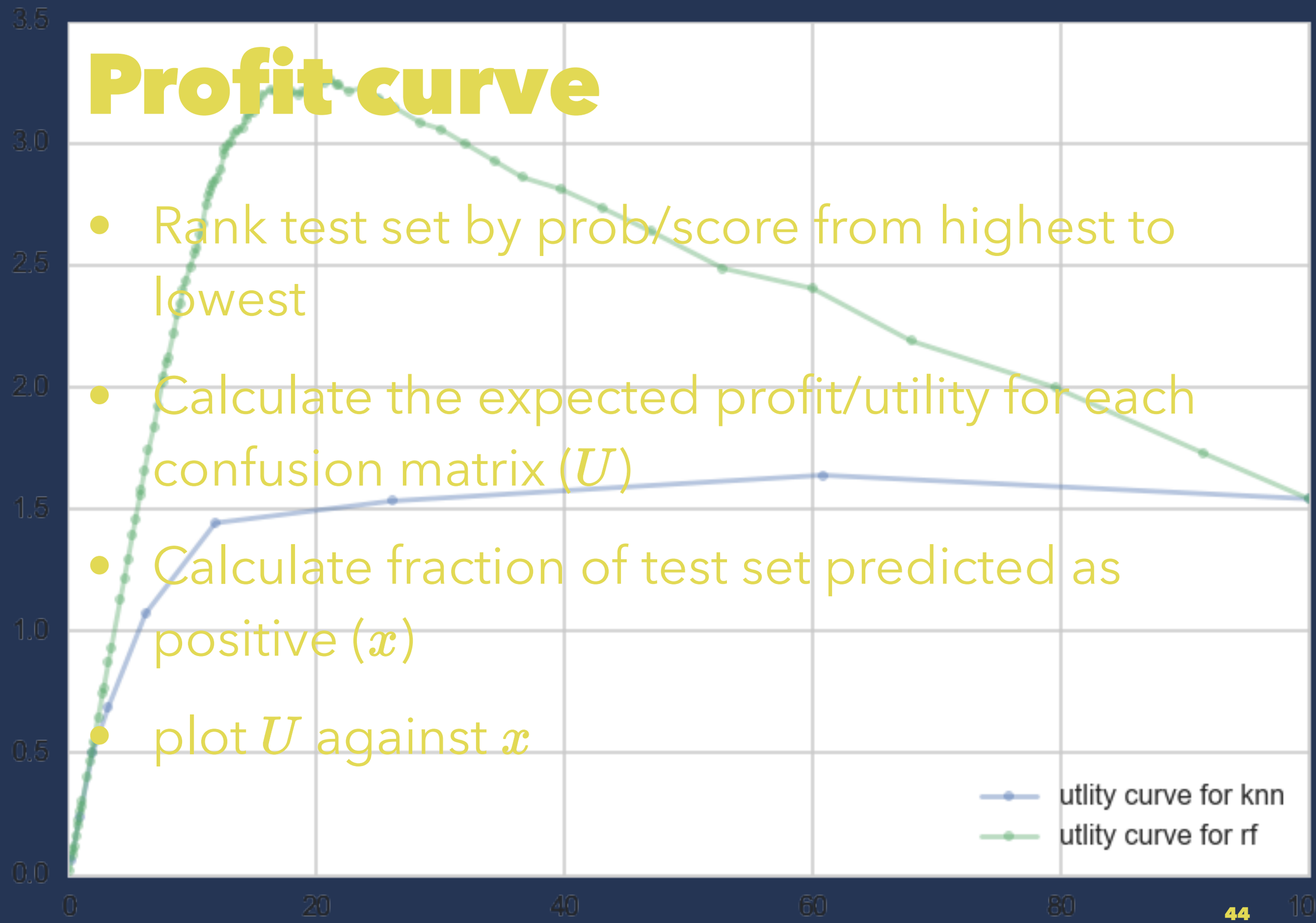
Small r penalizes FN.

Churn and Cancer u dont want FN: an uncaught churner or cancer patient (P=churn/cancer)



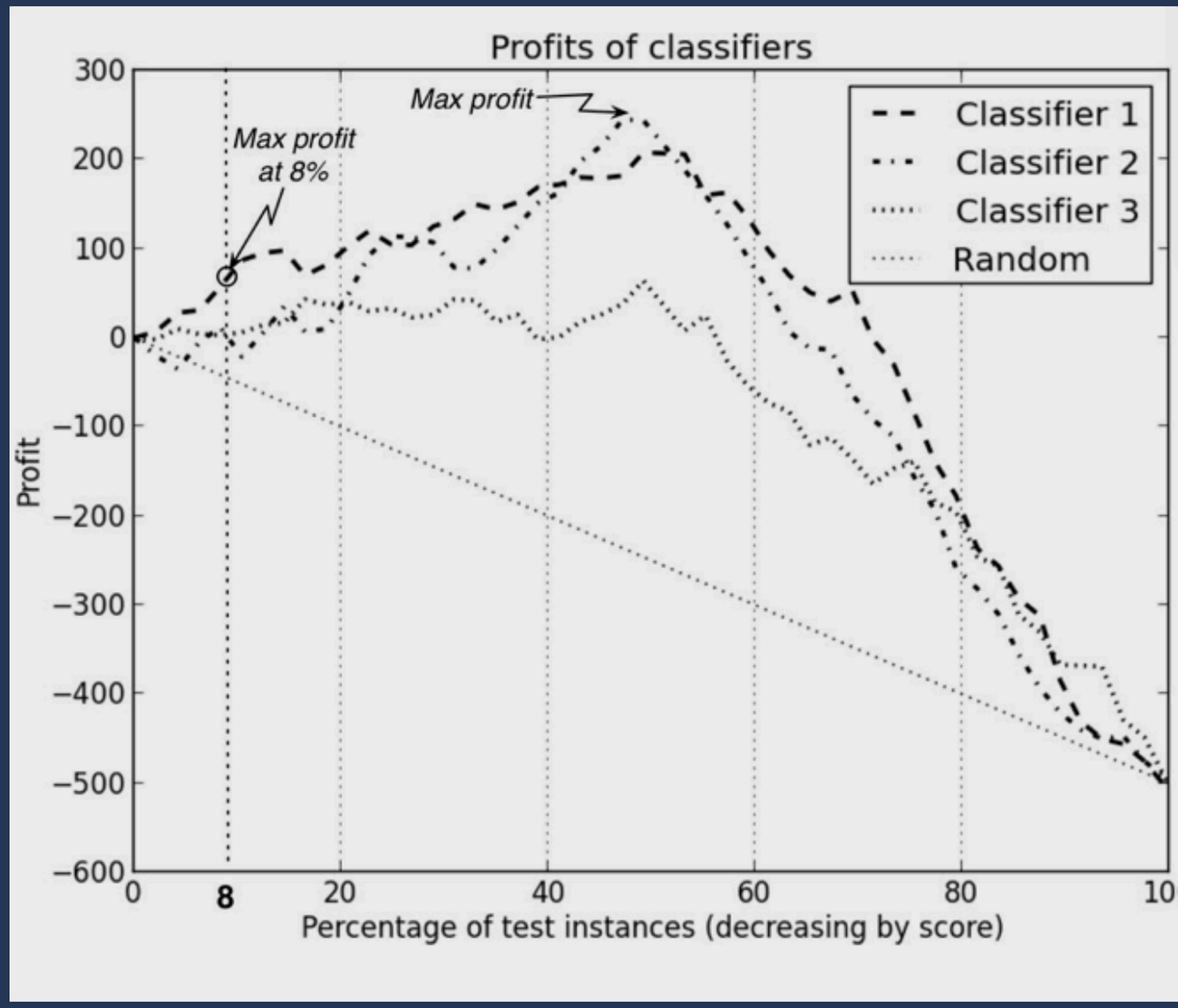
Profit curve

- Rank test set by prob/score from highest to lowest
- Calculate the expected profit/utility for each confusion matrix (U)
- Calculate fraction of test set predicted as positive (x)
- plot U against x



Finite budget[#]

- 100,000 customers, 40,000 budget, 5$ per customer
- we can target 8000 customers
- thus target top 8%
- classifier 1 does better there, even though classifier 2 makes max profit



[#] figure from Data Science for Business, Foster et. al.

WHERE TO GO FROM HERE?

- Follow @YhatHQ, @kdnuggets etc
- Read Provost, Foster; Fawcett, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
- check out Harvard's cs109: cs109.org
- Ask lots of questions of your data science team
- Follow your intuition

THANKS!!

@rahuldave