NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

BINARY CLASSIFICATION

- Positive and negative classes
- True positive (TP): correctly predicted as positive
- True negative (TN): correctly predicted as negative
- False positive (FP): incorrectly predicted as positive (type 1 error)
- False negative (FN): incorrectly predicted as negative (type 2 error)

ACCURACY

- Fraction of correct predictions
- # of correct decisions made / total # of decisions made
- (TP + TN) / (P + N) = (TP + TN) / (TP + FP + TN + FN)
- May be very misleading

ERROR RATE

- Fraction of incorrect predictions
- # of correct decisions made / total # of decisions made
- (FP + FN) / (P + N) = (TP + TN) / (TP + FP + TN + FN)
- 1 - accuracy

NO INFORMATION CLASSIFICATION

- Not using any information about input
- Always predict most frequent class

BASE RATE

- Accuracy of predicting most frequent class

NO INFORMATION REGRESSION

- Mean
- Median

SINGLE FEATURE PREDICTION

- Use only one feature
- Decision tree stump

CONFUSION MATRIX

- Summarize result of binary classification

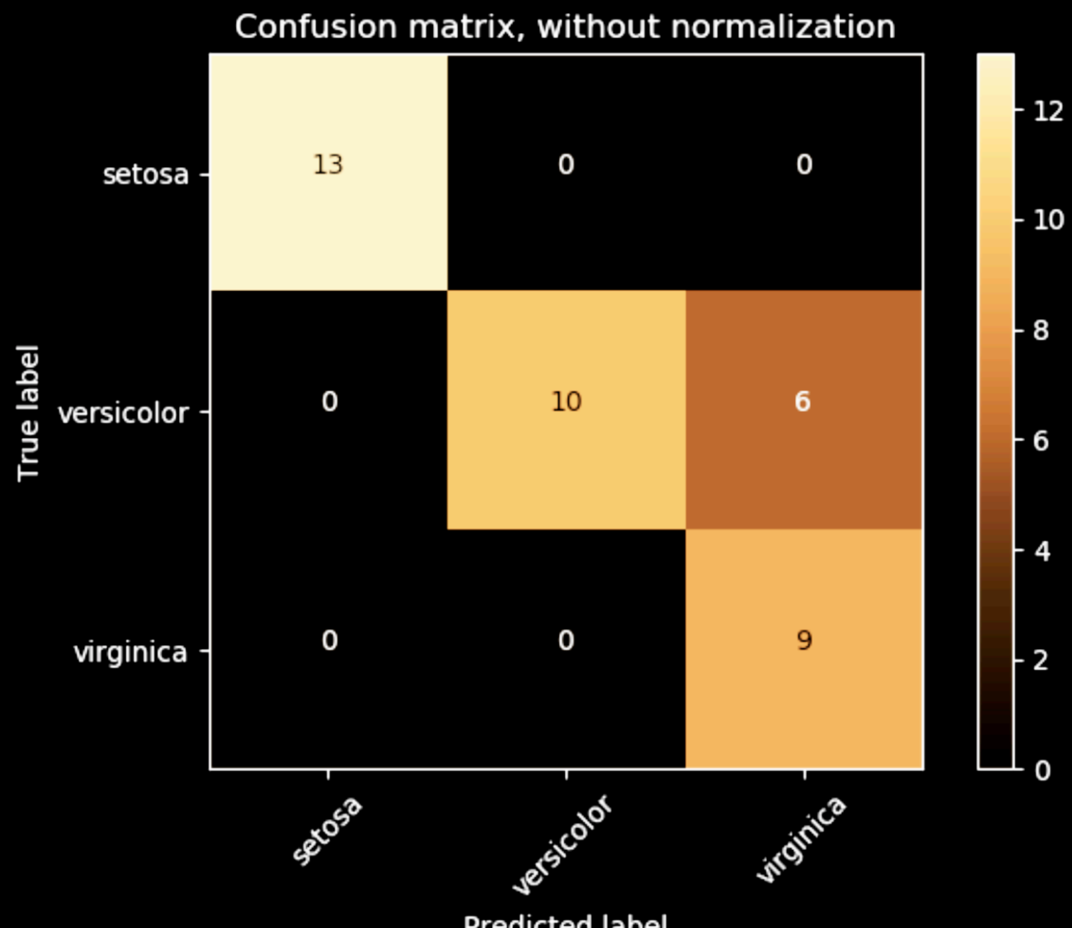
Actual

Predicted

	p	n
Y	True positives	False positives
N	False negatives	True negatives

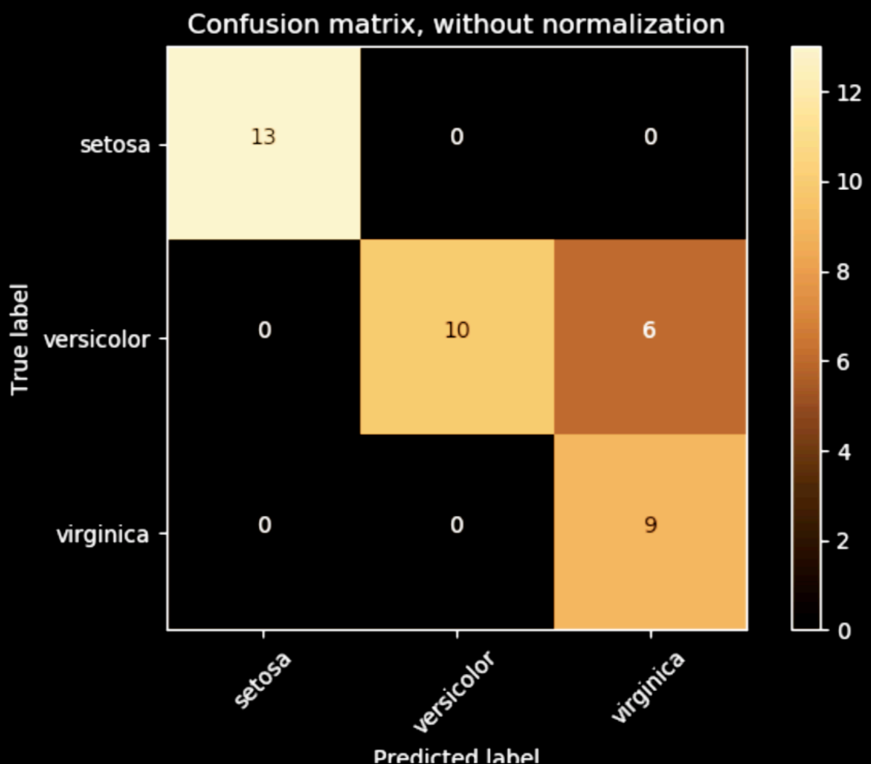
CONFUSION MATRIX

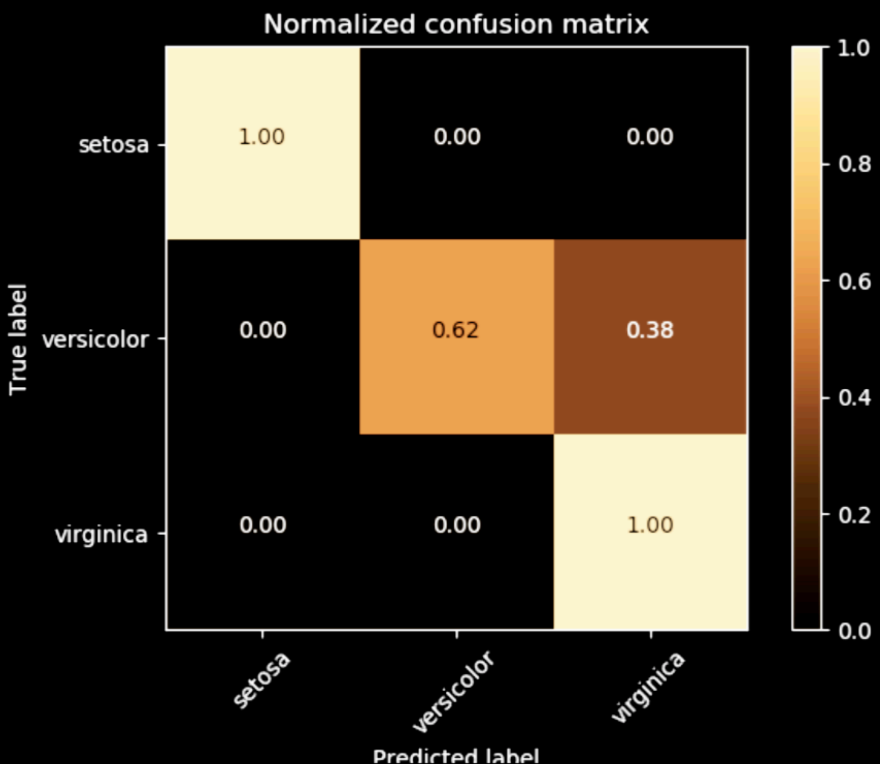
- Summarize result of multi-class classification



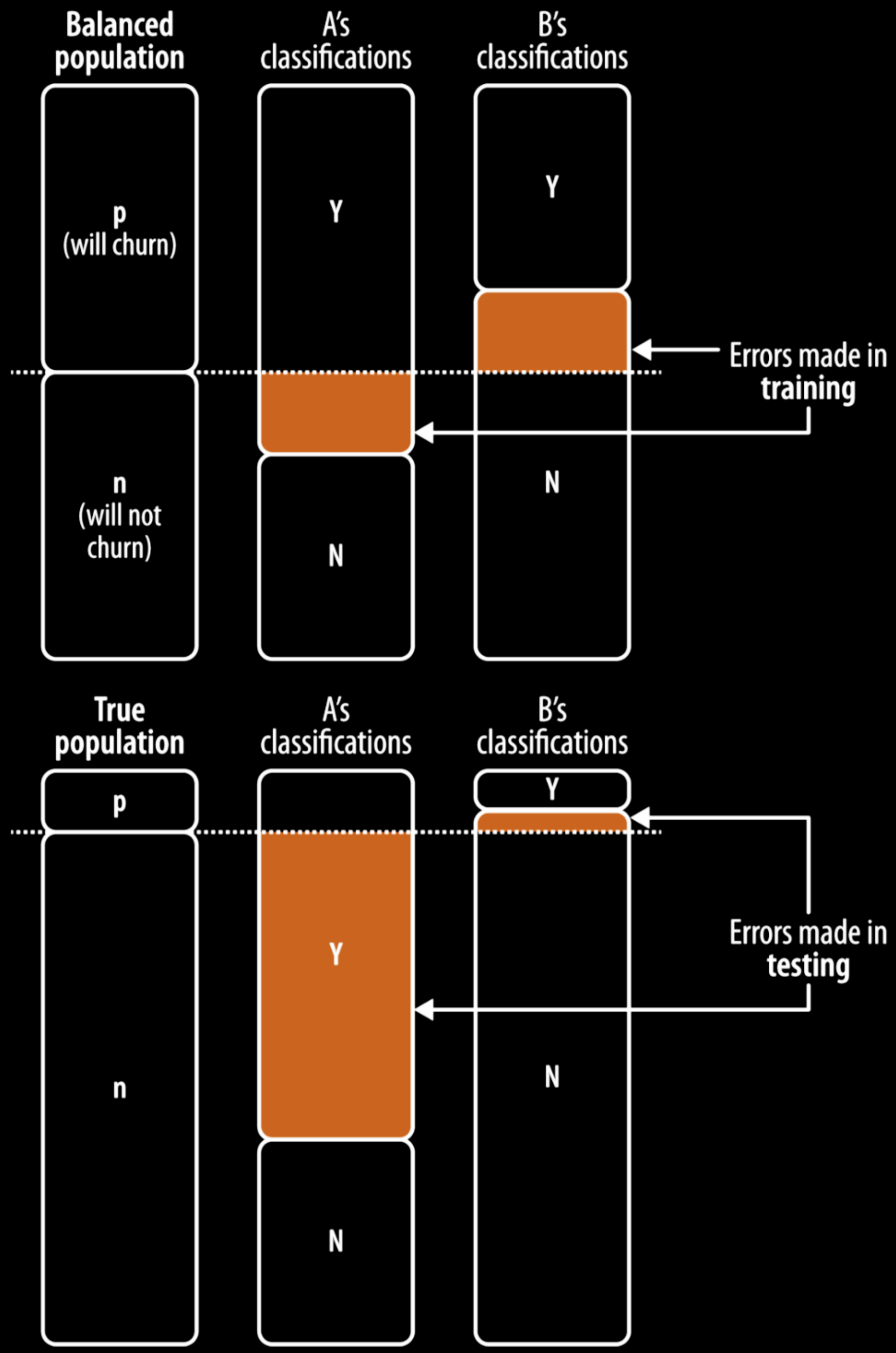
CONFUSION MATRIX

- Summarize result of multi-class classification





TRAINING AND TEST POPULATIONS



EXPECTED BENEFIT OF TARGETING

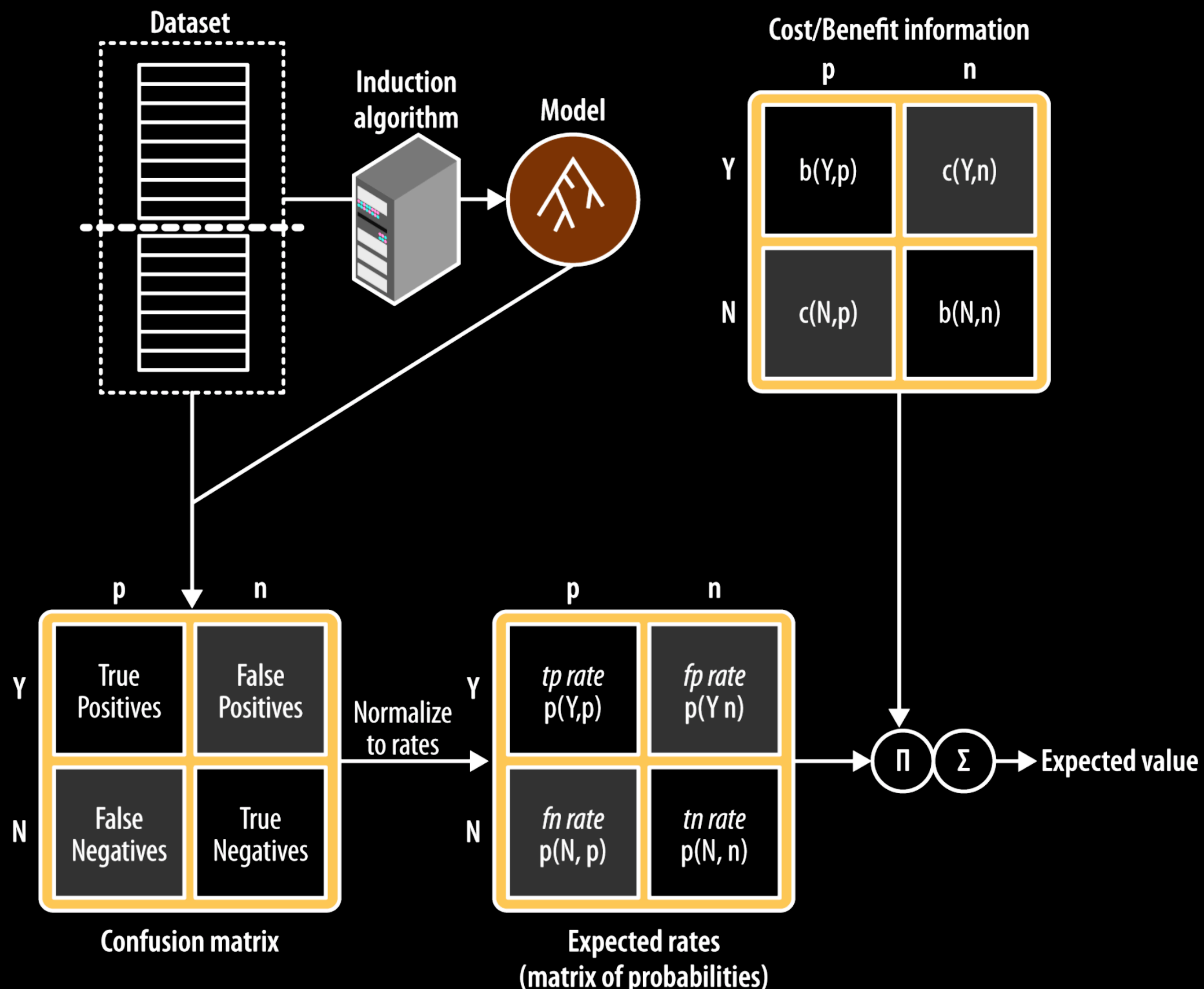
$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

$$p_R(\mathbf{x}) \cdot \$99 - [1 - p_R(\mathbf{x})] \cdot \$1 > 0$$

$$p_R(\mathbf{x}) \cdot \$99 > [1 - p_R(\mathbf{x})] \cdot \$1$$

$$p_R(\mathbf{x}) > 0.01$$

EXPECTED VALUE



CONFUSION PROBABILITIES

- p(hypothesis, actual) = count(hypothesis, actual)/T

	p	**n**
Y	56	7
N	5	42

$$T = 110$$

$$p(\mathbf{Y},\mathbf{p}) = 56/110 = 0.51 \quad p(\mathbf{Y},\mathbf{n}) = 7/110 = 0.06$$

$$p(\mathbf{N},\mathbf{p}) = 5/110 = 0.05 \quad p(\mathbf{N},\mathbf{n}) = 42/110 = 0.38$$

COST-BENEFIT MATRIX

$$
\begin{array}{cc}
 & \text{Actual} \\
\text{Predicted} & \begin{array}{c|cc} & \mathbf{p} & \mathbf{n} \\ \hline \mathbf{Y} & 99 & -1 \\ \mathbf{N} & 0 & 0 \end{array}
\end{array}
$$

		Actual	
		p	n
Predicted	Y	b(Y,p)	c(Y,n)
	N	c(N,p)	b(N,n)

EXPECTED PROFIT

$$\begin{aligned} Expected\ \ profit = {} & p(\mathbf{Y}, \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N}, \mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + \\ & p(\mathbf{N}, \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y}, \mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n}) \end{aligned}$$

$$p(x, y) = p(y) \cdot p(x \mid y)$$

$$\begin{aligned} \text{Expected profit} \quad = \quad & p(\mathbf{Y} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + \\ & p(\mathbf{N} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n}) \end{aligned}$$

$$\begin{aligned} \text{Expected profit} \quad = \quad & p(\mathbf{p}) \ \cdot \ [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\ & p(\mathbf{n}) \ \cdot \ [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot c(\mathbf{Y}, \mathbf{n})] \end{aligned}$$

EXPECTED PROFIT

	p	n
Y	56	7
N	5	42

$\mathrm{T} = 110$

$\mathrm{P} = 61$ $\quad$ $\mathrm{N} = 49$

$p(\mathbf{p}) = 0.55$ $\quad$ $p(\mathbf{n}) = 0.45$

$tp\ rate = 56/61 = 0.92$ $\quad$ $fp\ rate = 7/49 = 0.14$

$fn\ rate = 5/61 = 0.08$ $\quad$ $tn\ rate = 42/49 = 0.86$

$$
\begin{aligned}
\text{expected profit} \quad = \quad & p(\mathbf{p}) \cdot [\, p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] \, + \\
& p(\mathbf{n}) \cdot [\, p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{p}) \cdot c(\mathbf{Y}, \mathbf{n})] \\
= \quad & 0.55 \cdot [0.92 \cdot b(\mathbf{Y}, \mathbf{p}) + 0.08 \cdot b(\mathbf{N}, \mathbf{p})] \, + \\
& 0.45 \cdot [0.86 \cdot b(\mathbf{N}, \mathbf{n}) + 0.14 \cdot p(\mathbf{Y}, \mathbf{n})] \\
= \quad & 0.55 \cdot [0.92 \cdot 99 + 0.08 \cdot 0] \, + \\
& 0.45 \cdot [0.86 \cdot 0 + 0.14 \cdot \ \text{-}\ 1] \\
= \quad & 50.1 \text{ - } 0.063 \\
\approx \quad & \$\mathbf{50.04}
\end{aligned}
$$

PRECISION

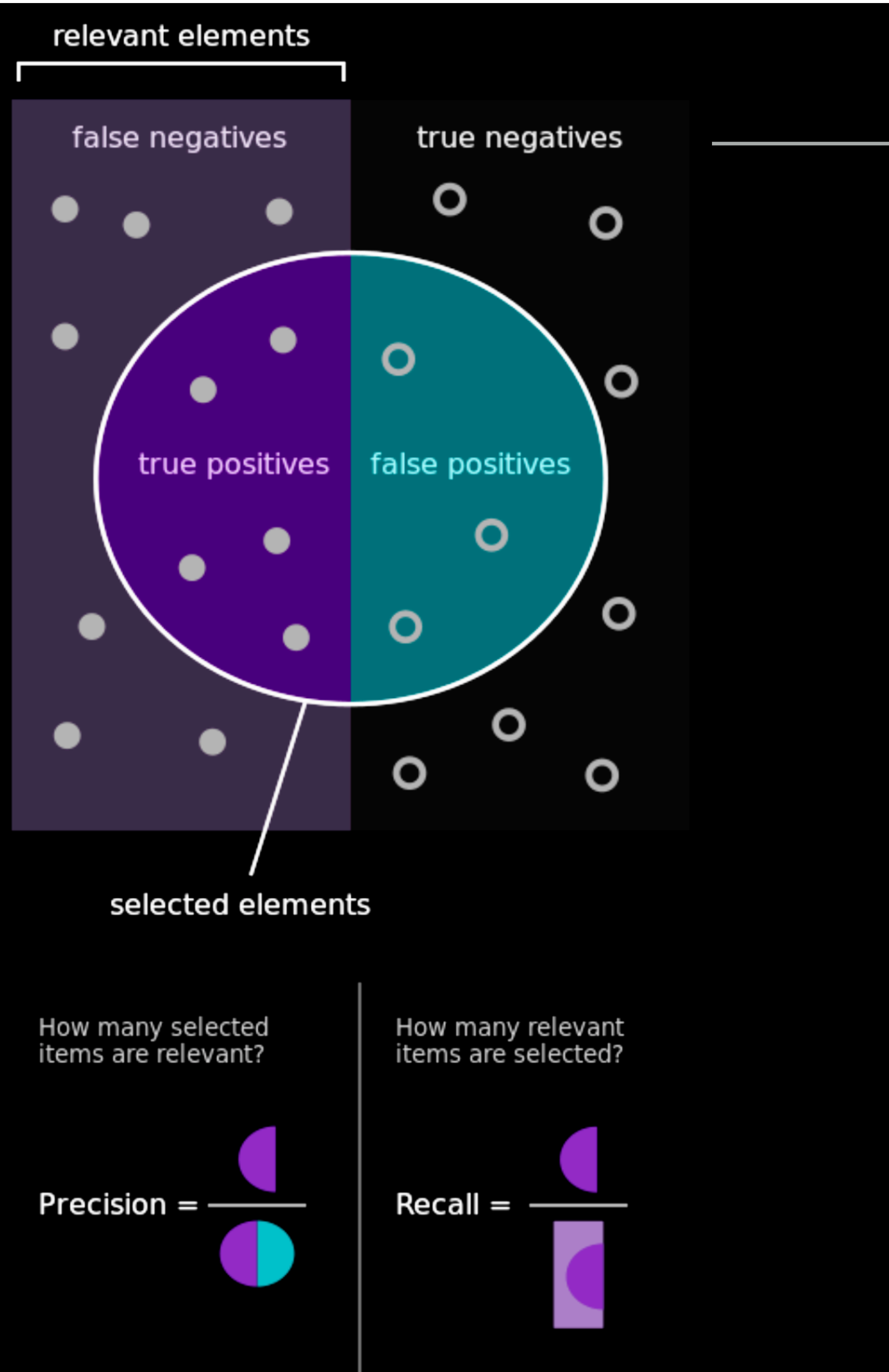
- Accuracy of positive predictions
- TP / (TP+FP)
- High precision: if you test positive you're probably positive
- % of documents offered as relevant that are actually relevant

RECALL = SENSITIVITY

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

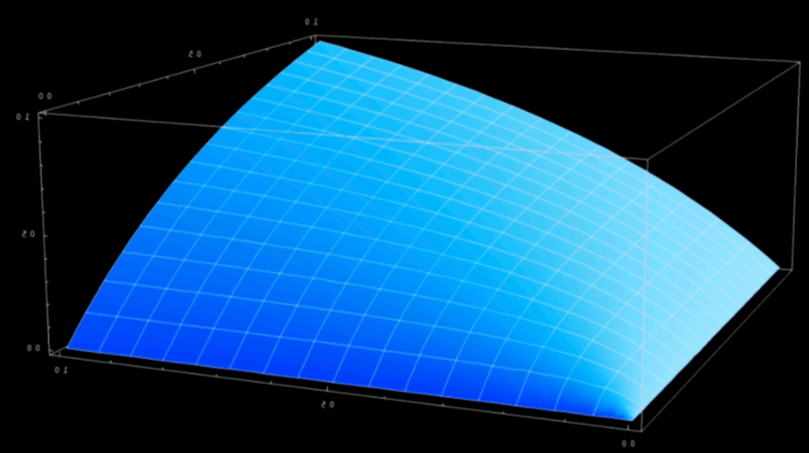
- True positive rate
- Accuracy of positive class
- TP / (TP + FN)
- High recall is not missing many positives
- % of relevant documents that were found
- What fraction of people with disease are identified
- How sensitive is the test to indicators of disease

PRECISION AND RECALL



F1 SCORE

- 2*(precision*recall)/(precision+recall)
- Harmonic mean of precision and recall in [0,1]
- Goal is high precision and high recall



F-BETA SCORE

- (1+beta^2)*(precision*recall)/(beta^2*precision+recall)
- Weigh precision and recall

SPECIFICITY

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

- True negative rate
- Accuracy on negative class
- TN / (FP + TN)
- What fraction of people without disease are identified
- High specificity: few false alarms

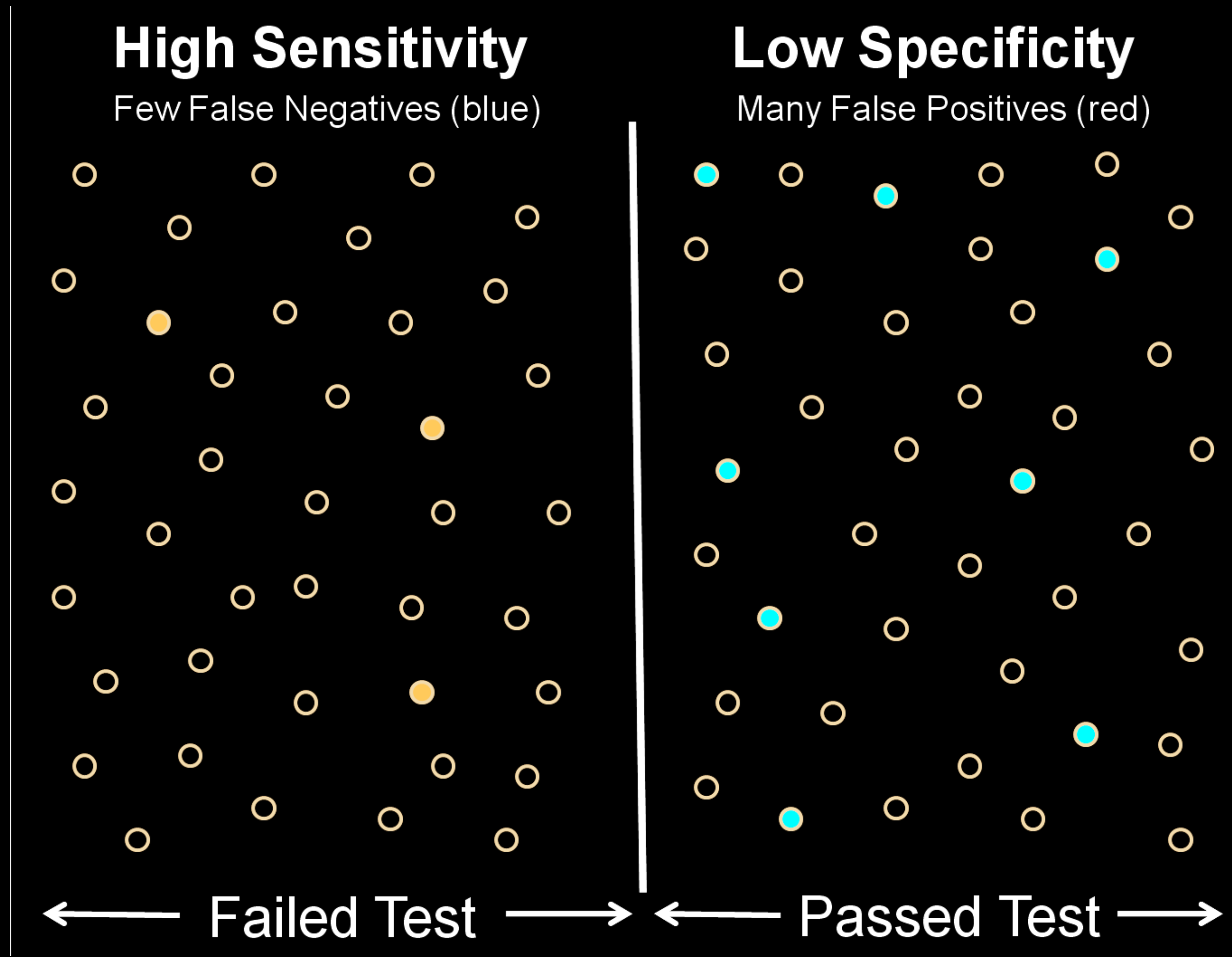
SENSITIVITY AND SPECIFICITY IN AIRPORT SECURITY

- Sensitivity: quantifies avoiding of false negatives
- Specificity: quantifies avoiding false positives
- For a test there is usually a trade-off between them.
- Low specificity and high sensitivity:
  - Testing passengers for potential safety threats
  - Scanners may be set to trigger alarms on low-risk items like buckles, keys (low specificity) to increase probability of identifying dangerous objects and minimize risk of missing objects that pose a threat (high sensitivity).

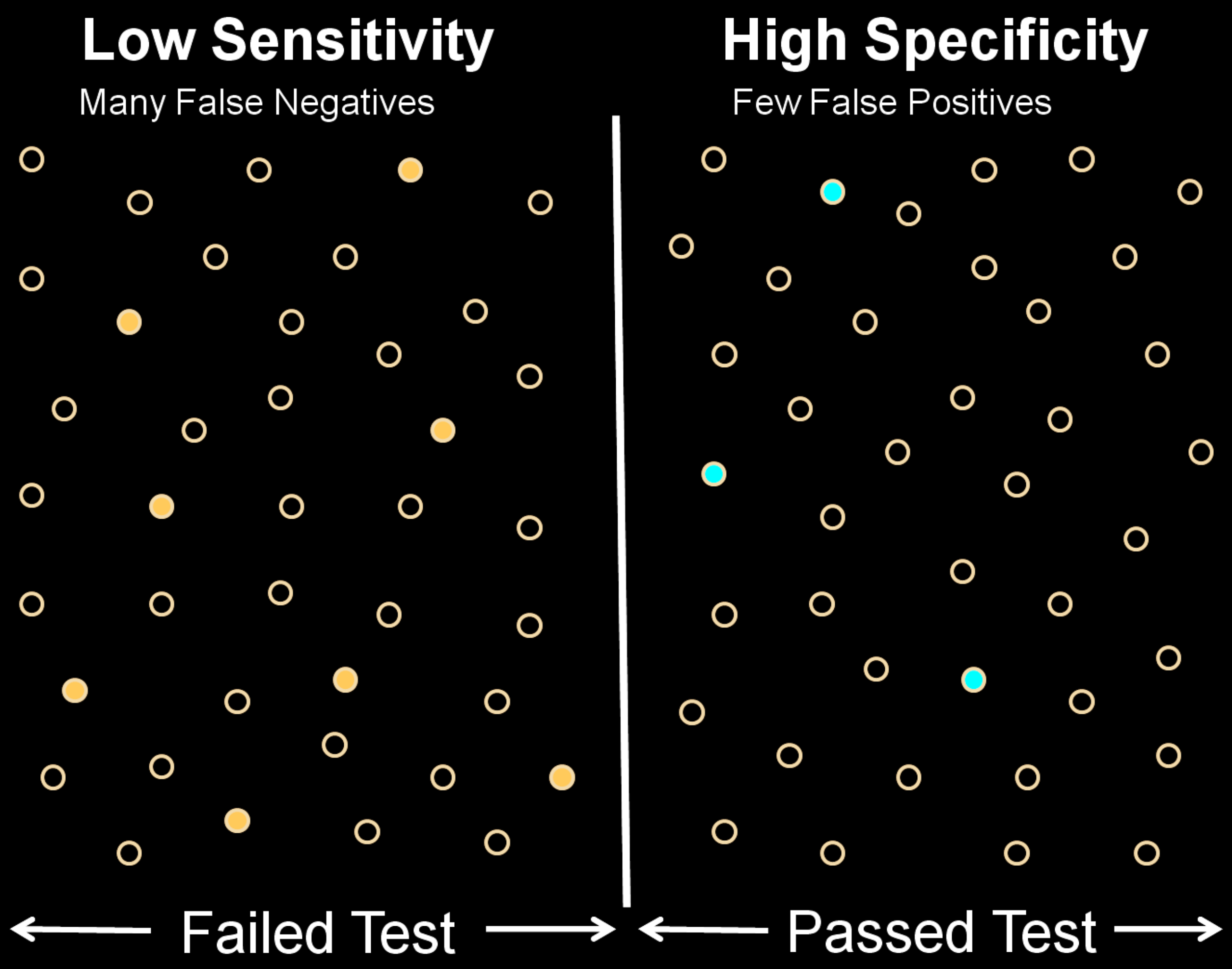
SENSITIVITY AND SPECIFICITY IN MEDICAL DIAGNOSIS

- Goal to have high sensitivity and high specificity
- Perfect predictor is both:
  - 100% sensitive: all sick people are correctly identified as sick
  - 100% specific: no healthy individuals are incorrectly identified as sick

SENSITIVITY AND SPECIFICITY TRADE-OFF



SENSITIVITY AND SPECIFICITY TRADE-OFF



FALSE POSITIVE RATE = FALL OUT = FALSE ALARM RATE

- Error rate on negative class
- FP / (FP + TN)

FALSE NEGATIVE RATE = MISS RATE

- FN / (FN + TP)

FALSE DISCOVERY RATE (FDR)

- FP / (FP + TP)

POSITIVE PREDICTIVE VALUE

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

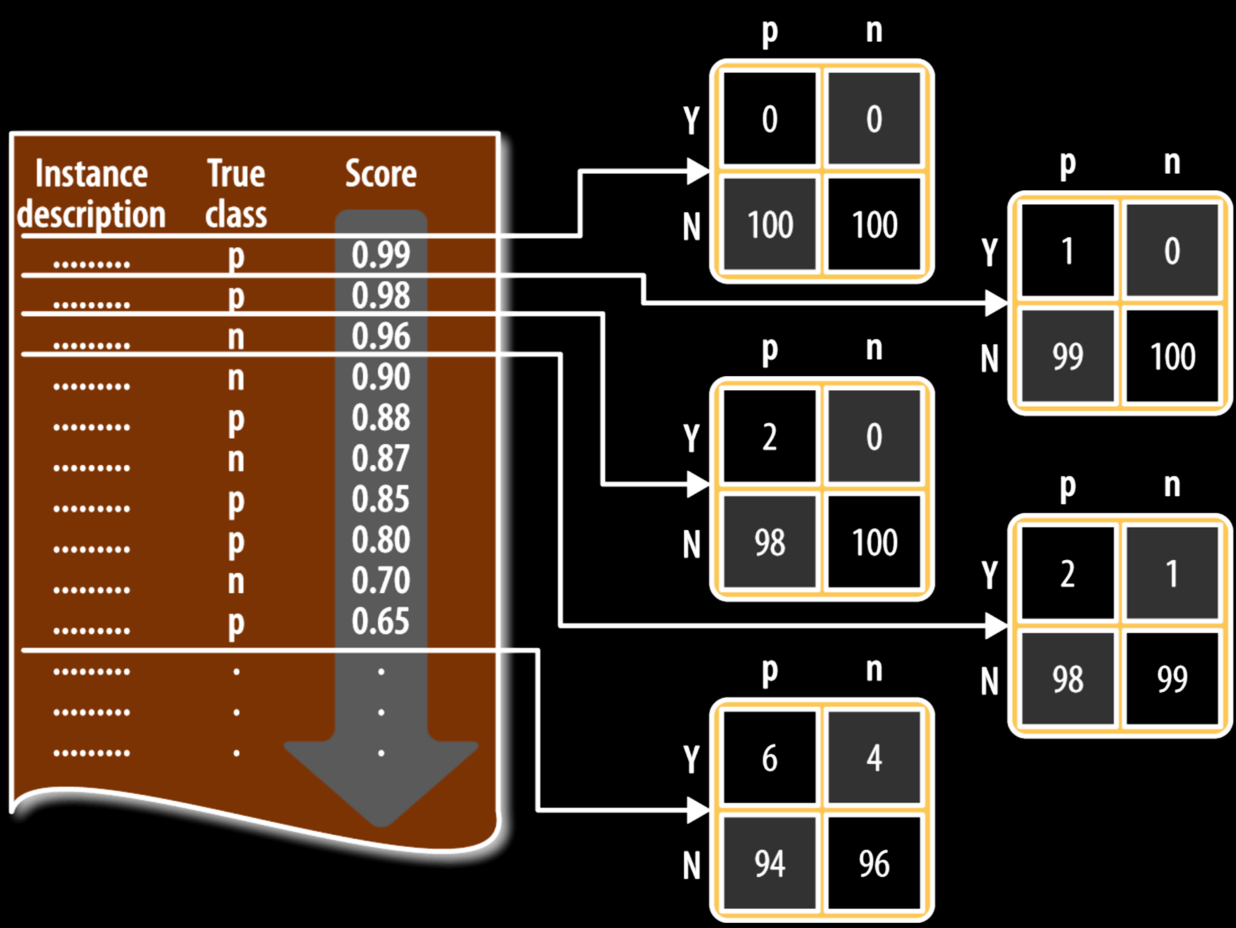
- TP / (TP + FP)
- PPV = 1 - FDR
- Small positive predictive value of 10% indicates many of positive results from testing procedure are false positives.
- Follow up any positive result with more reliable test to obtain more accurate assessment. Test may be useful if it is inexpensive and convenient.

SUMMARY OF MEASURES

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

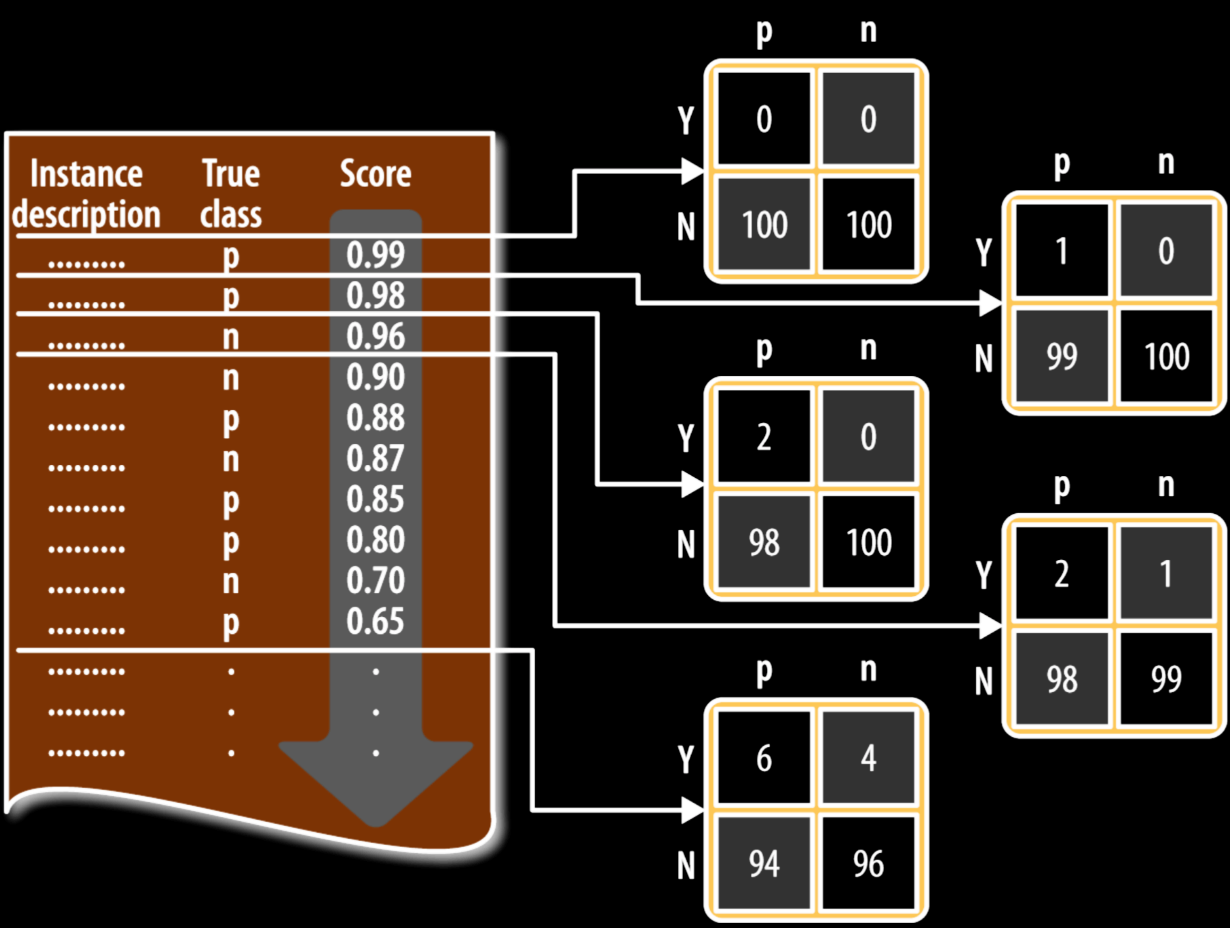
CONFUSION MATRIX AS A FUNCTION OF THRESHOLD

- Different thresholds results in different confusion matrices



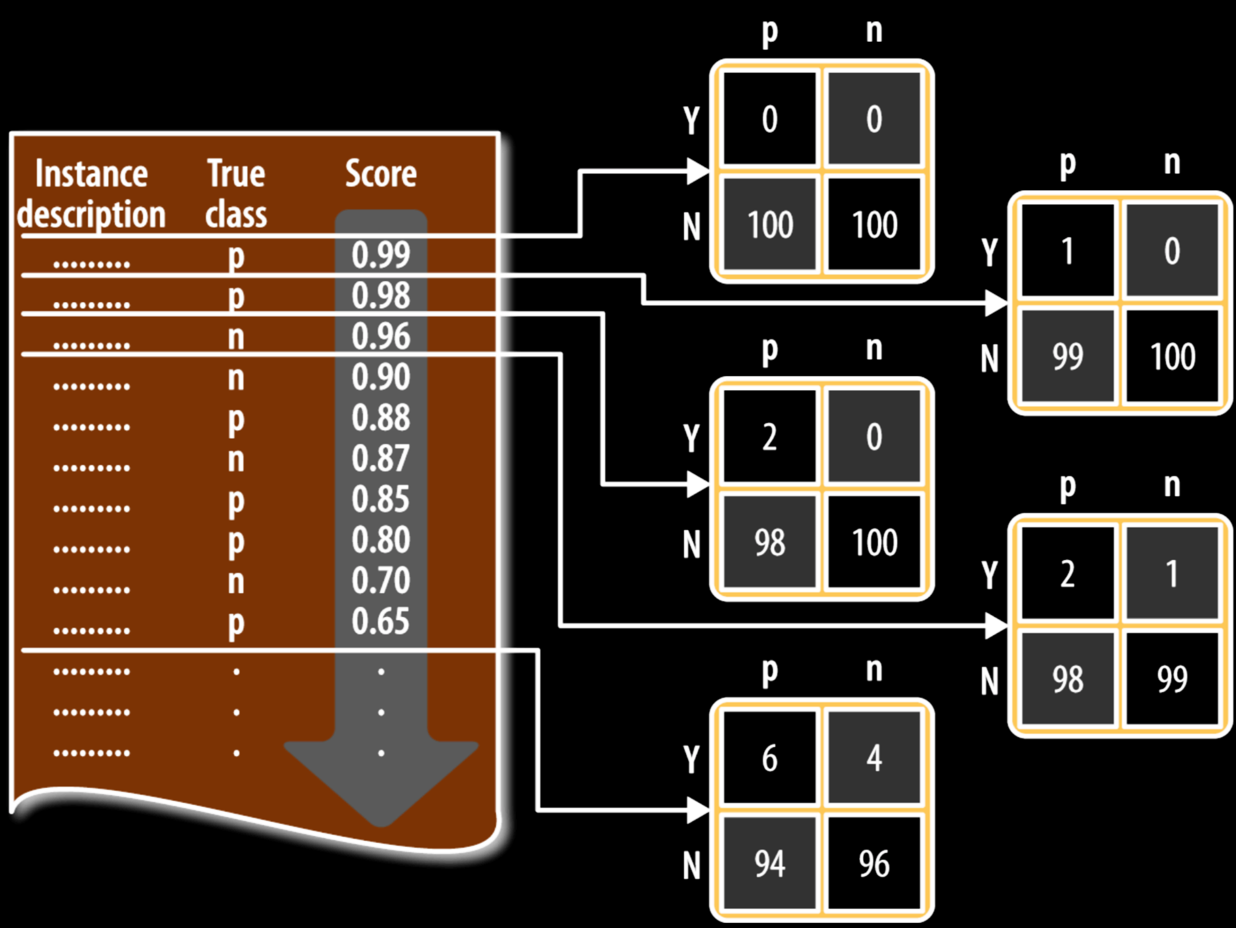
RECALL AS A FUNCTION OF THRESHOLD

- Decreasing threshold increases recall: TP / (TP + FN)



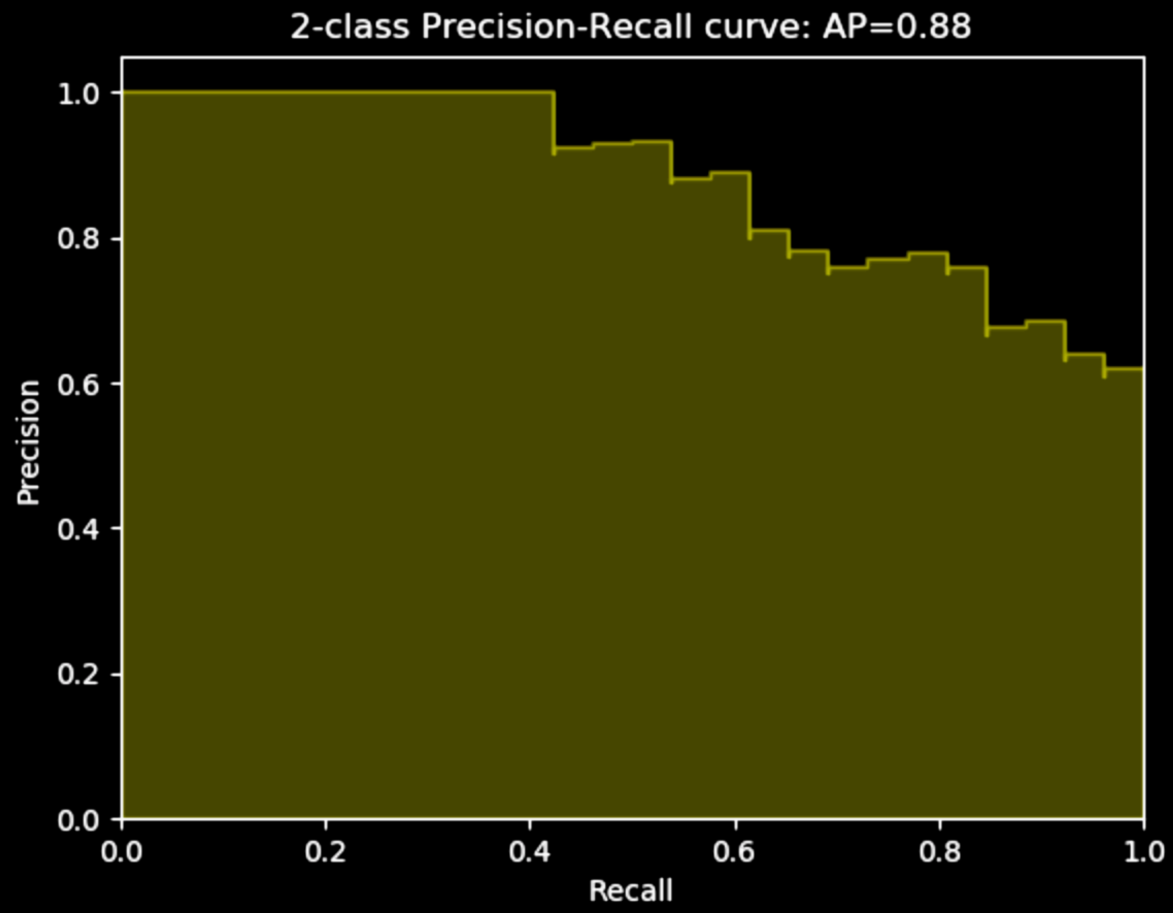
PRECISION AS A FUNCTION OF THRESHOLD

- Increasing threshold **may** increase precision: TP / (TP + FP)

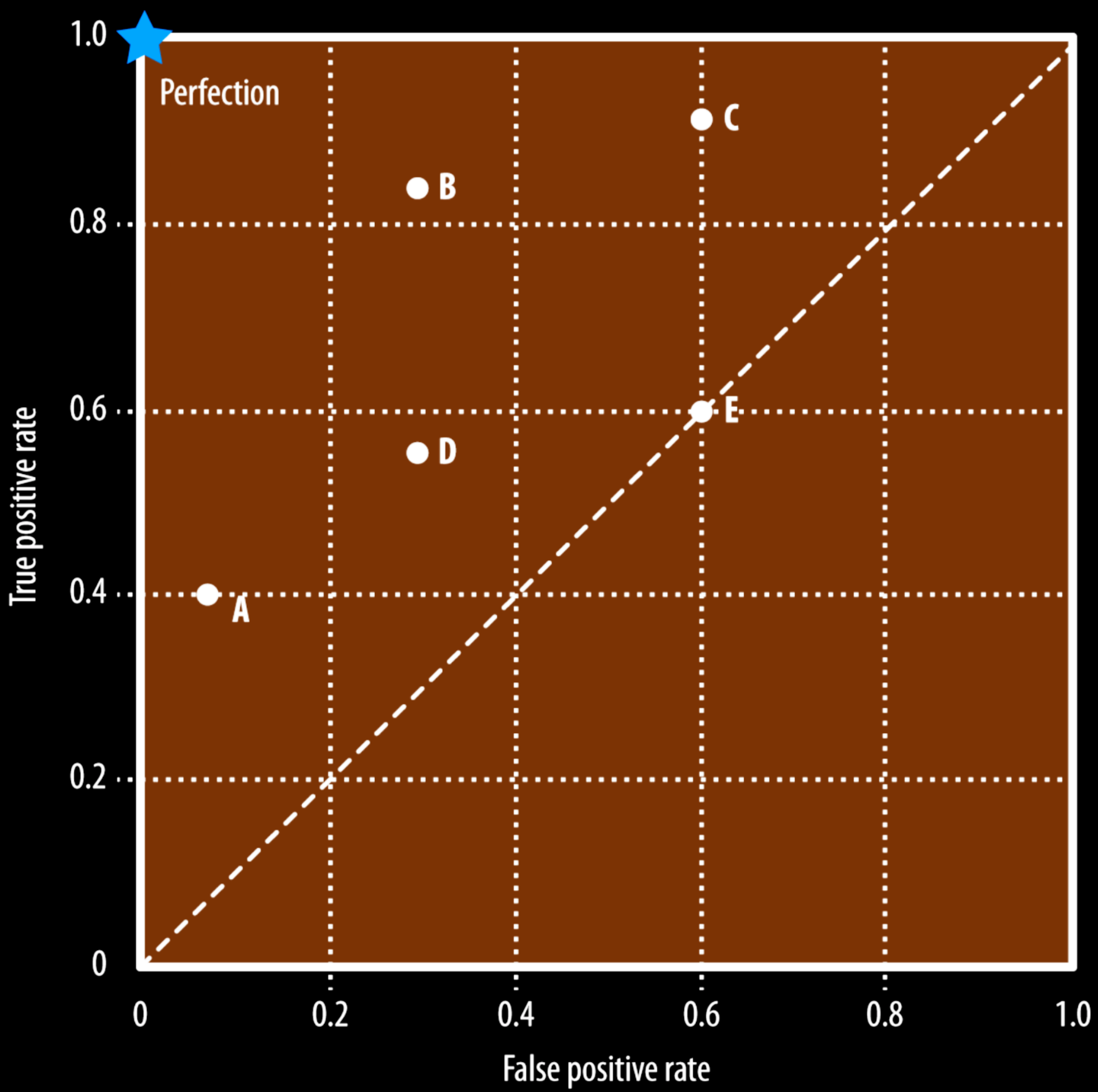


PRECISION VS. RECALL

- Not monotonic

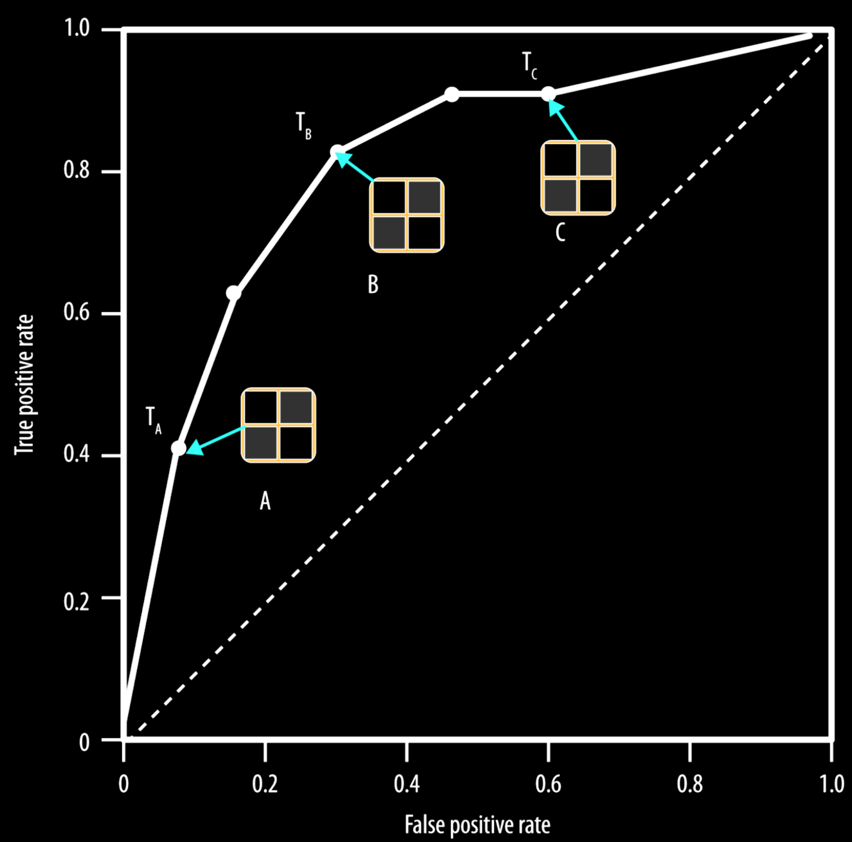


RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

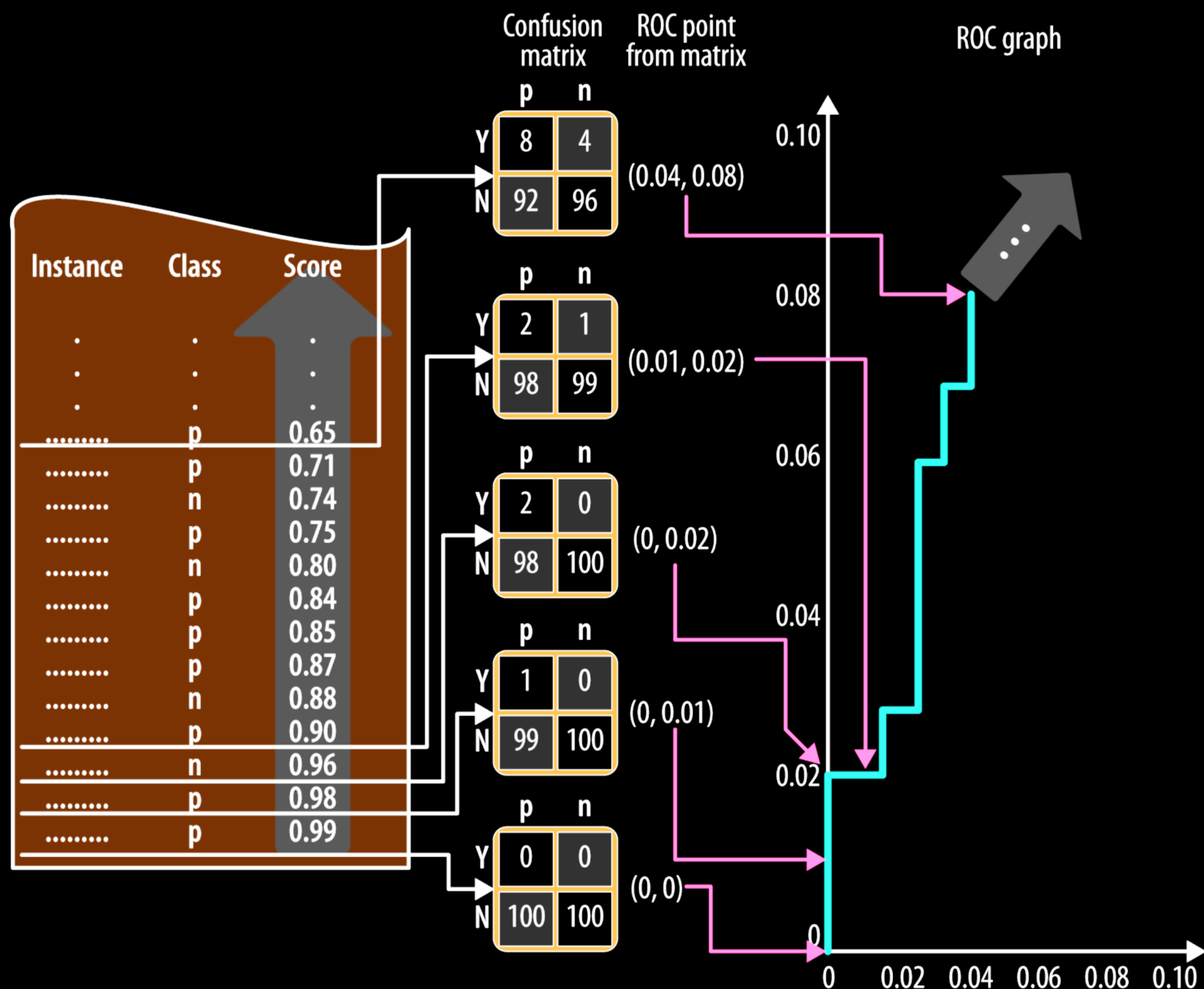


RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

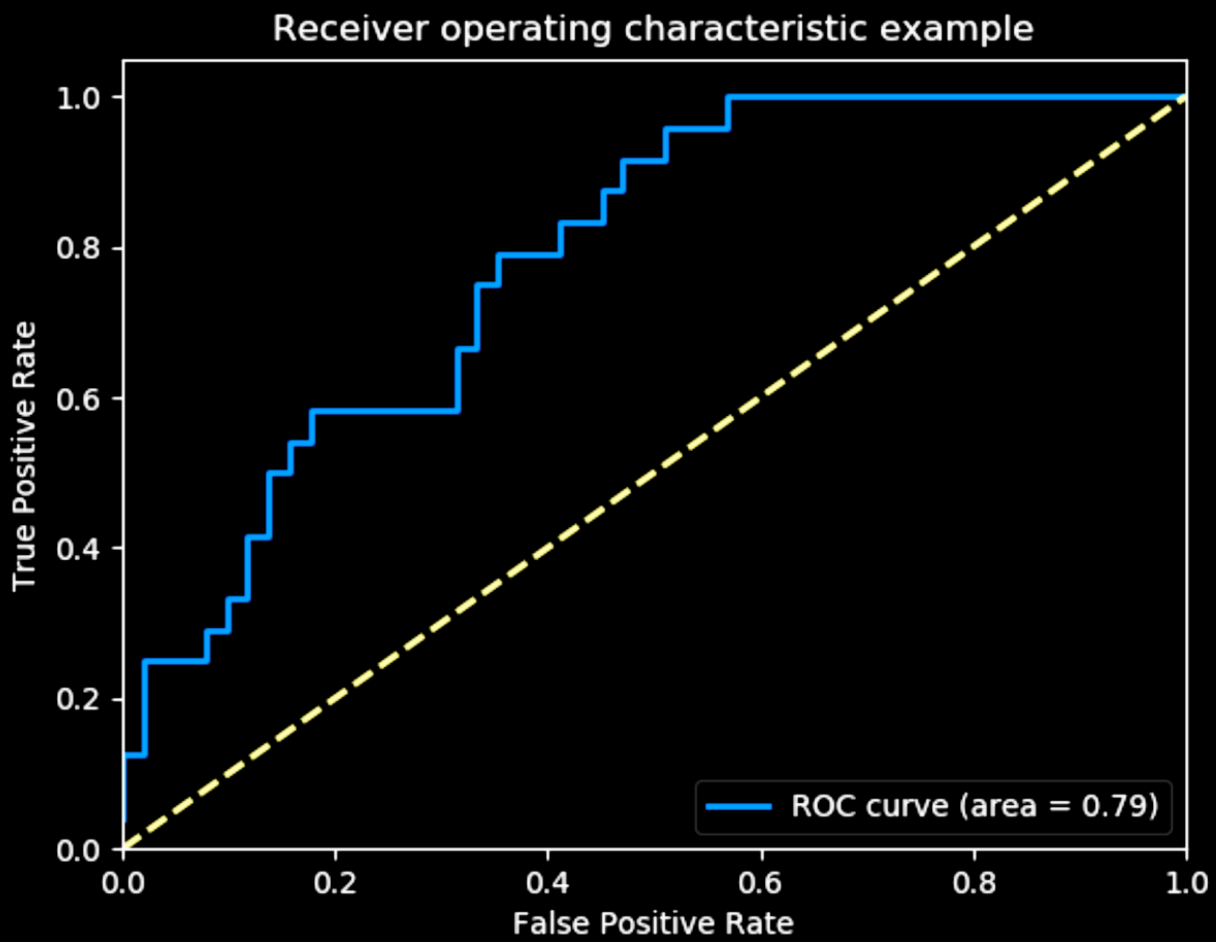
- Best curve would go straight up and left
- Non-increasing slope



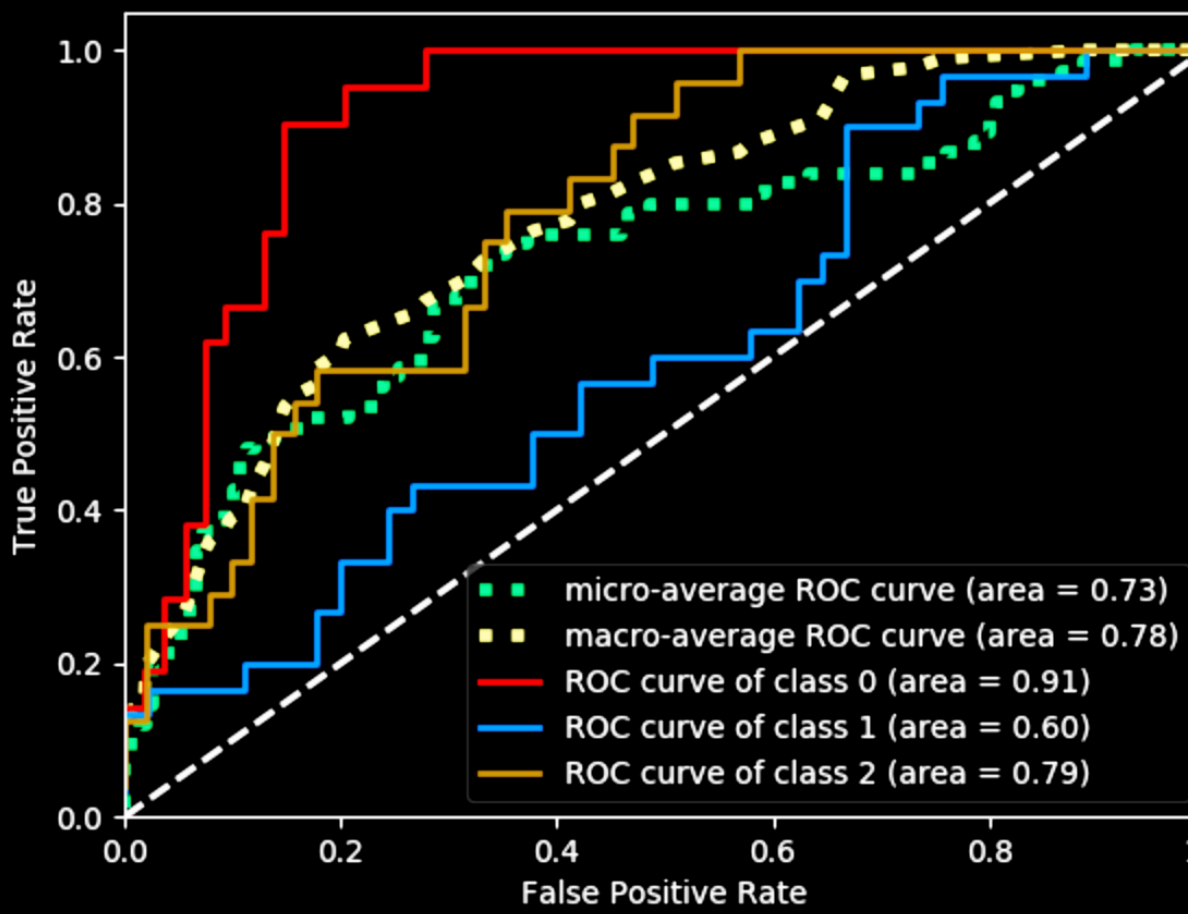
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



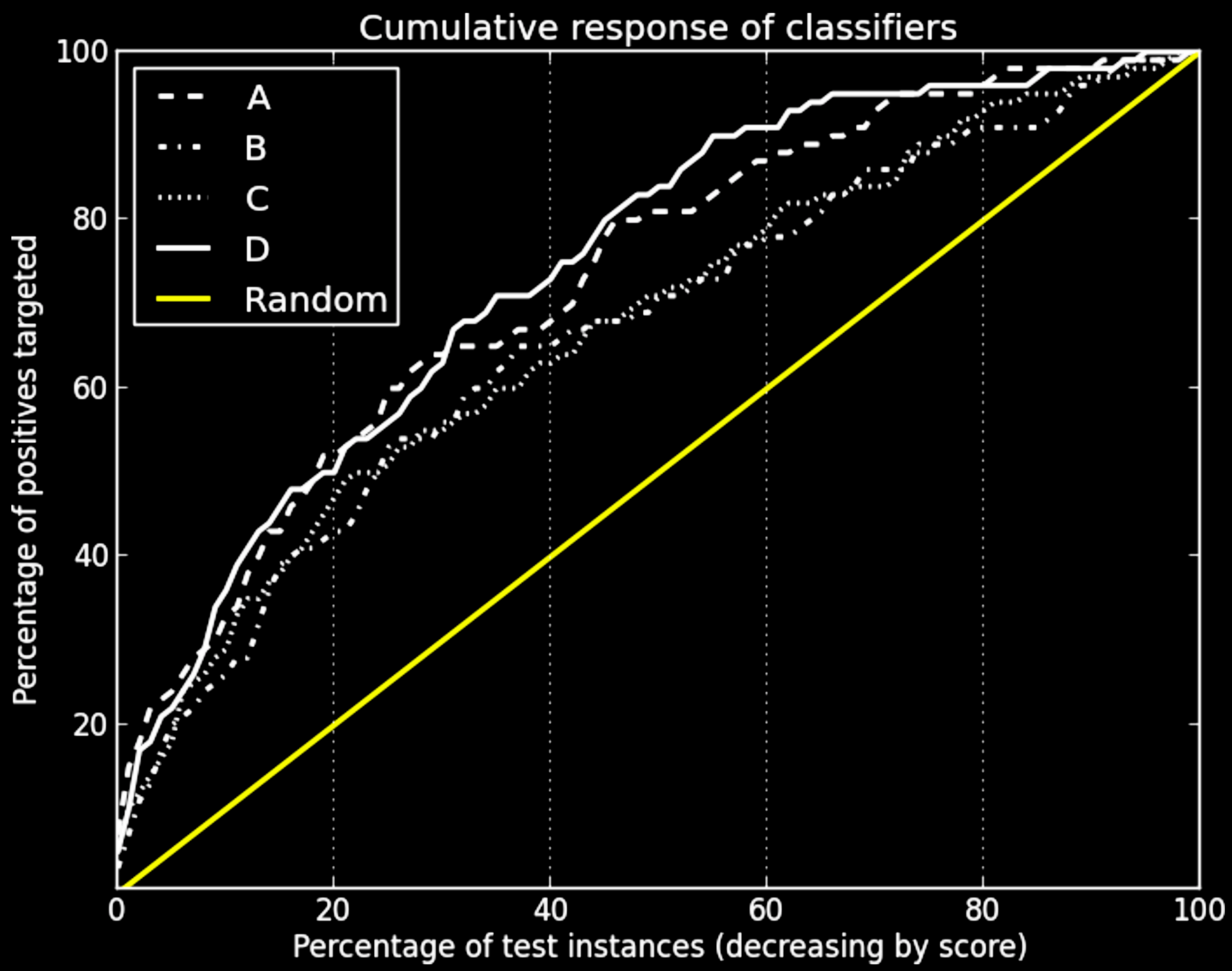
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



AREA UNDER (ROC) CURVE (AUC)

- Single number used to summarize classifier performance
- Perfect prediction is 1
- Random prediction is 1/2

CHURN EXAMPLE

- Customers switch carriers: churn
- Cheaper to retain customer than acquire new customer
- Retain customer by promotion or discount
- If you give a discount to someone who was about to churn you saved money. It you give a discount to someone who was not going to churn you wasted money.
- Trained classifier to predict churn. Choose score threshold for giving discount.

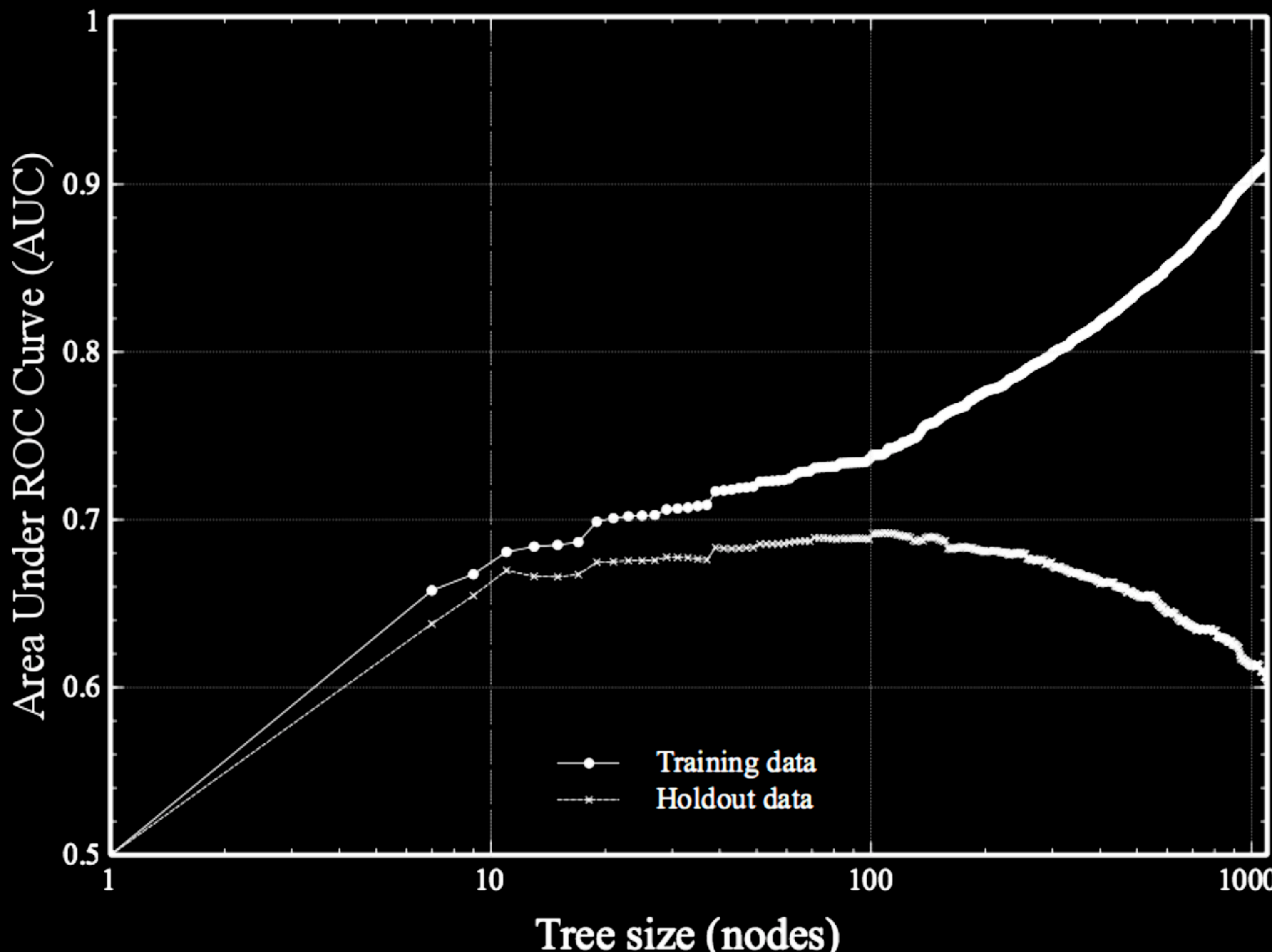
CHURN EXAMPLE

- Training accuracy

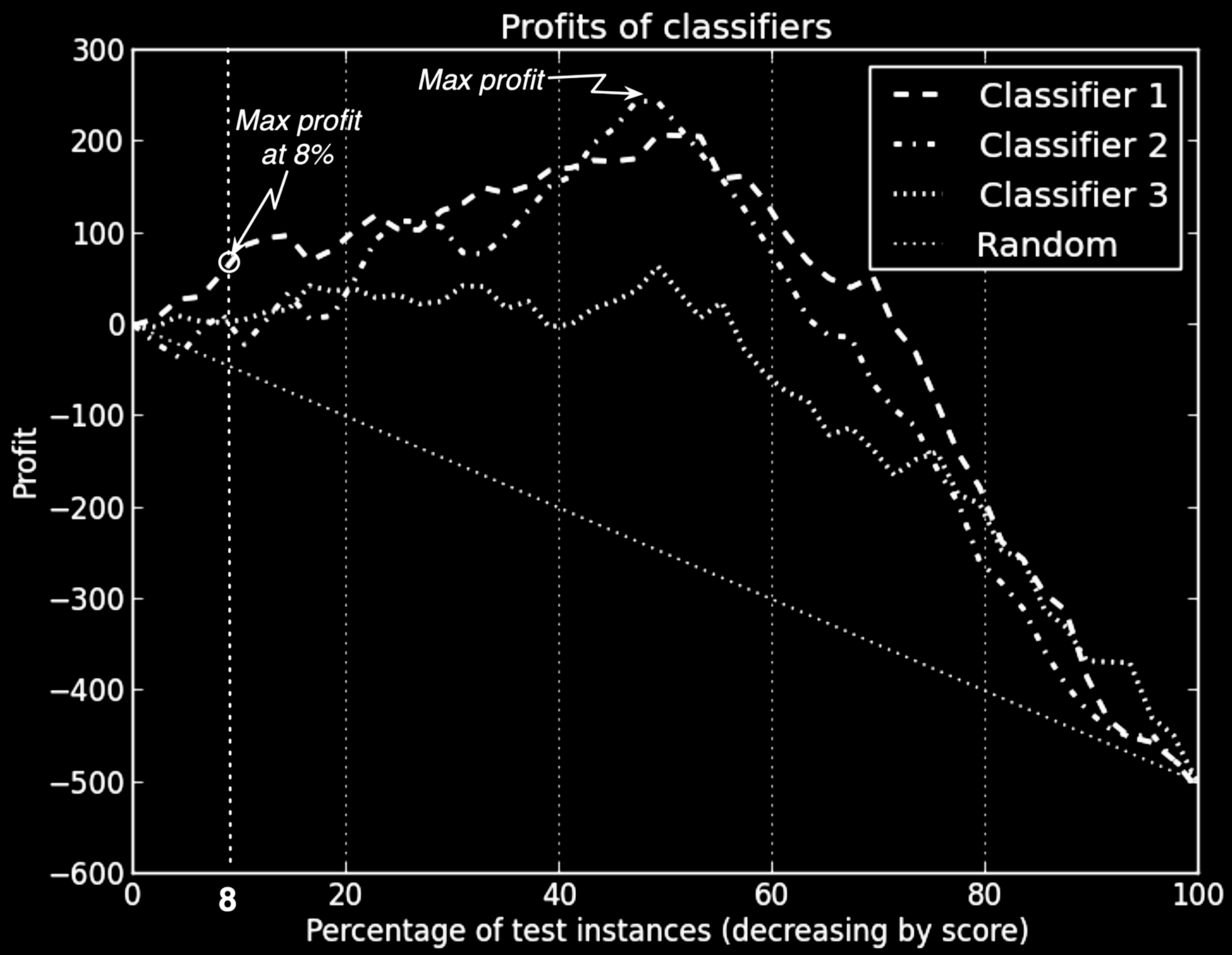
Model	Accuracy
Classification tree	95%
Logistic regression	93%
k-Nearest Neighbor	100%
Naive Bayes	76%

- Test accuracy and AUC

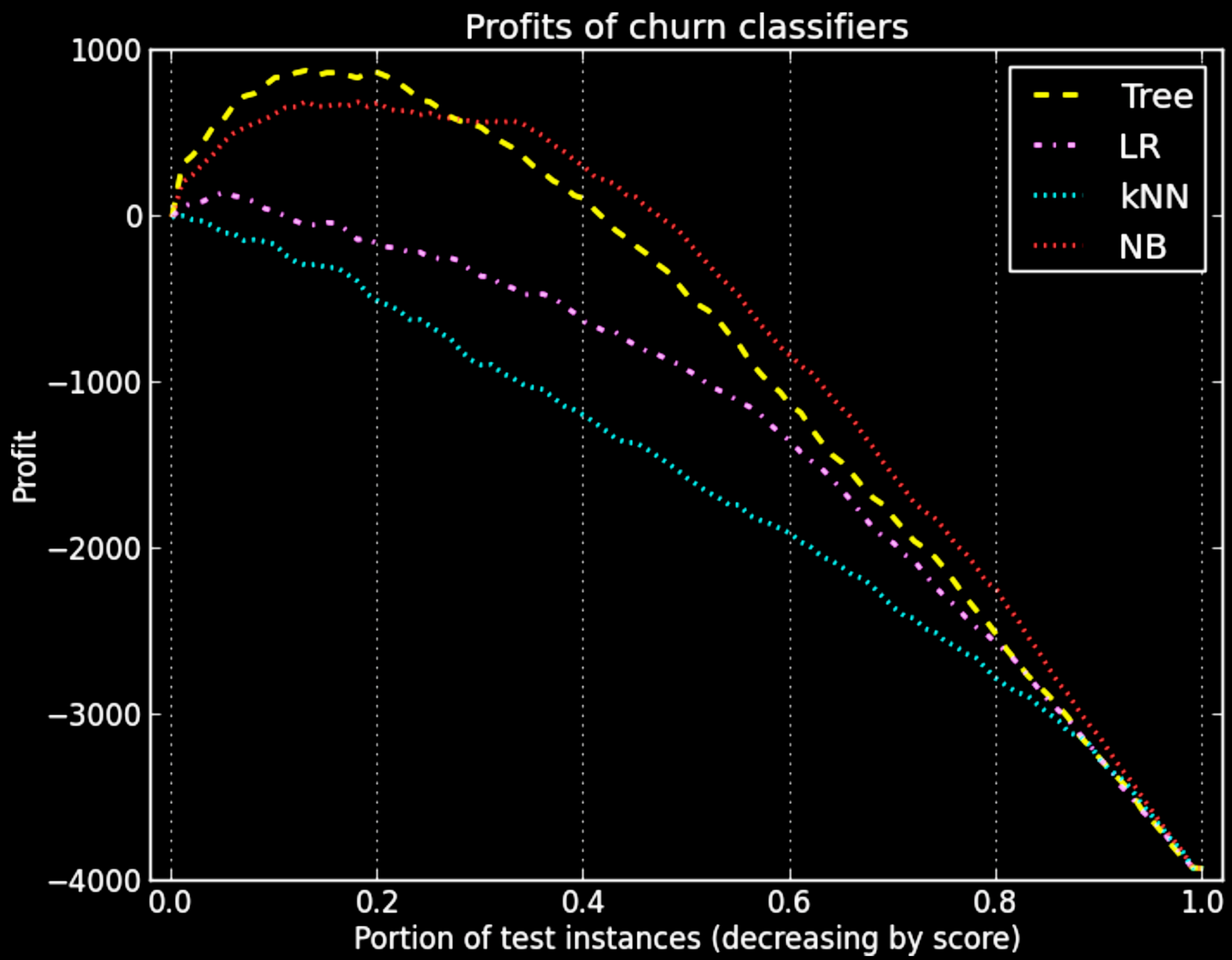
Model	Accuracy (%)	AUC
Classification Tree	91.8 ± 0.0	0.614 ± 0.014
Logistic Regression	93.0 ± 0.1	0.574 ± 0.023
k-Nearest Neighbor	93.0 ± 0.0	0.537 ± 0.015
Naive Bayes	76.5 ± 0.6	0.632 ± 0.019


Area Under ROC Curve (AUC)
Tree size (nodes)
Training data
Holdout data
1
0.9
0.8
0.7
0.6
0.5
1
10
100
1000

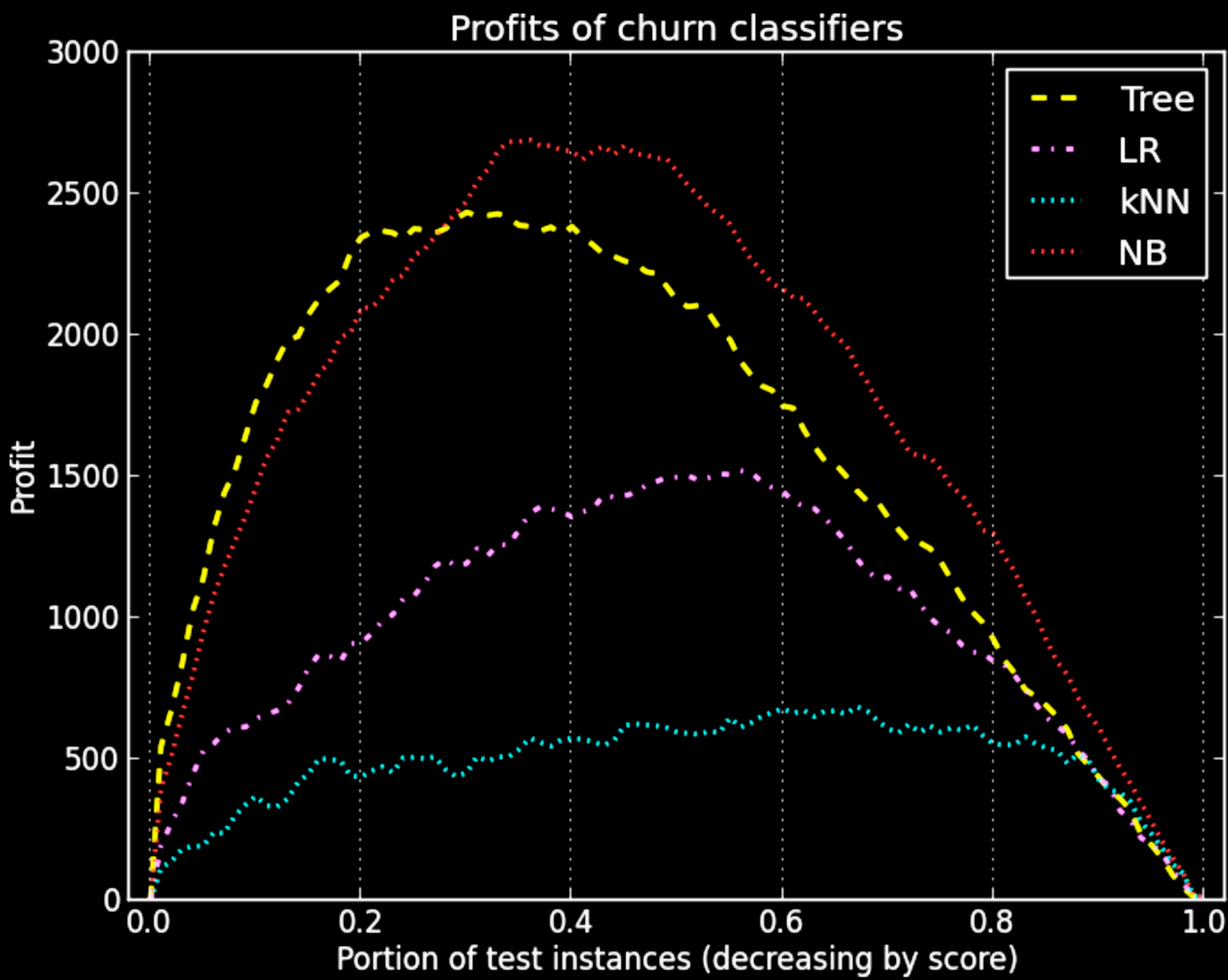
PROFIT CURVE



PROFIT CURVE



PROFIT CURVE



BAYES RULE FOR CLASSIFICATION

- p(C=c): prior probability of class
- p(E | C=c): likelihood of seeing evidence E when in class c
- p(E): likelihood of evidence
- p(C=c | E): posterior

$$p(C = c \mid \mathbf{E}) = \frac{p(\mathbf{E} \mid C = c) \cdot p(C = c)}{p(\mathbf{E})}$$

BAYES RULE FOR CLASSIFICATION

- Conditionally independent attributes

$$
\begin{aligned}
p(\mathbf{E} \mid c) &= p(e_1 \wedge e_2 \wedge \cdots \wedge e_k \mid c) \\
&= p(e_1 \mid c) \cdot p(e_2 \mid c) \cdots p(e_k \mid c)
\end{aligned}
$$

$$
p(c \mid \mathbf{E}) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdots p(e_k \mid c) \cdot p(c)}{p(\mathbf{E})}
$$

BAYES RULE FOR CLASSIFICATION

- Instances belonging to one class

$$
\begin{aligned}
p(\mathbf{E}) &= p(\mathbf{E} \wedge c_0) + p(\mathbf{E} \wedge c_1) \\
&= p(\mathbf{E} \mid c_0) \cdot p(c_0) + p(\mathbf{E} \mid c_1) \cdot p(c_1)
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{E}) = \; & p(e_1 \mid c_0) \cdot p(e_2 \mid c_0) \cdots p(e_k \mid c_0) \cdot p(c_0) \\
& + p(e_1 \mid c_1) \cdot p(e_2 \mid c_1) \cdots p(e_k \mid c_1) \cdot p(c_1)
\end{aligned}
$$

$$p(\mathbf{E}) = p(e_1 \mid c_0) \cdot p(e_2 \mid c_0) \cdots p(e_k \mid c_0) \cdot p(c_0) + p(e_1 \mid c_1) \cdot p(e_2 \mid c_1) \cdots p(e_k \mid c_1) \cdot p(c_1)$$

BAYES RULE FOR CLASSIFICATION

- p(ei | cj): proportion of examples in class cj in which feature ei appears.

$$p(c_0 \mid \mathbf{E}) = \frac{p(e_1 \mid c_0) \cdot p(e_2 \mid c_0) \cdots p(e_k \mid c_0) \cdot p(c_0)}{p(e_1 \mid c_0) \cdot p(e_2 \mid c_0) \cdots p(e_k \mid c_0) + p(e_1 \mid c_1) \cdot p(e_2 \mid c_1) \cdots p(e_k \mid c_1)}$$

LIFT

- Not interested in accuracy on entire data, only on %
- How much better than random prediction on the fraction of dataset
- lift(threshold) = % positives > threshold / % dataset > threshold

LIFT EXAMPLE

- If prevalence of hotel bookings in a randomly targeted set of consumers 0.01% and in our selected population it is 0.02%, then classifier gives us a lift of 2, ie selected population has double booking rate.
- If there is full feature independence:

$$p(c \mid \mathbf{E}) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdots p(e_k \mid c) \cdot p(c)}{p(e_1) \cdot p(e_2) \cdots p(e_k)}$$

LIFT EXAMPLE

- Each feature ei raises or lowers the probability of the class by a factor equal to that piece of evidence lift.

$$p(C = c \mid \mathbf{E}) = p(C = c) \cdot \text{lift}_c(e_1) \cdot \text{lift}_c(e_2) \cdots$$

$$\text{lift}_c(x) = \frac{p(x \mid c)}{p(x)}$$

LIFT EXAMPLE

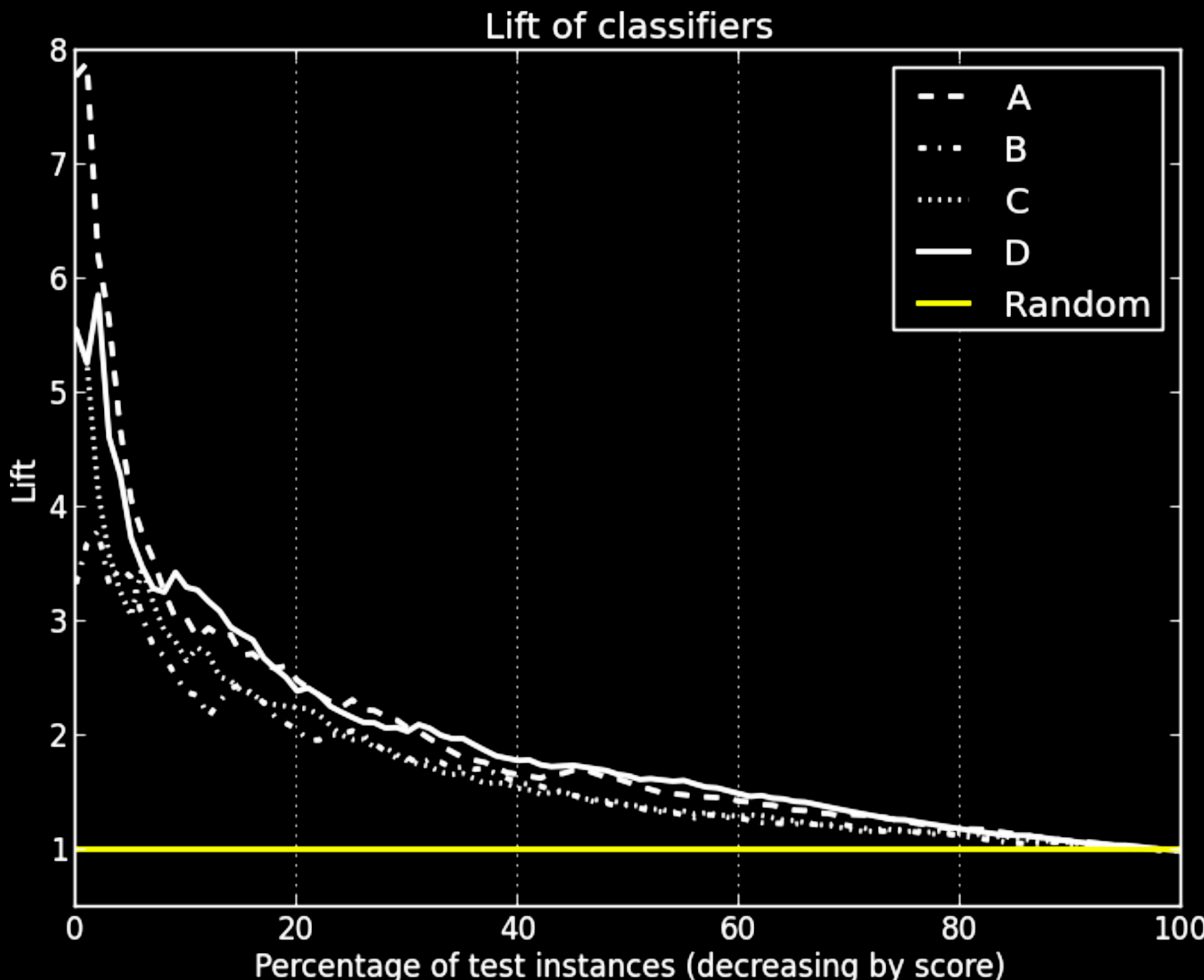
- Start with probability of booking, initialized to 0.0001 (prior probability, before seeing evidence, that a website visitor will book a room).
- Visited finance site? Multiply probability of booking by 2.
- Visit a truck-pull site? Multiply probability by 0.25.
And so on.
- After processing all ei evidence bits of E, resulting product is final probability (belief) that E is member of class c, ie that visitor E will book a room.

LIFT

- Target variable is IQ>130, 14% of sample is positive

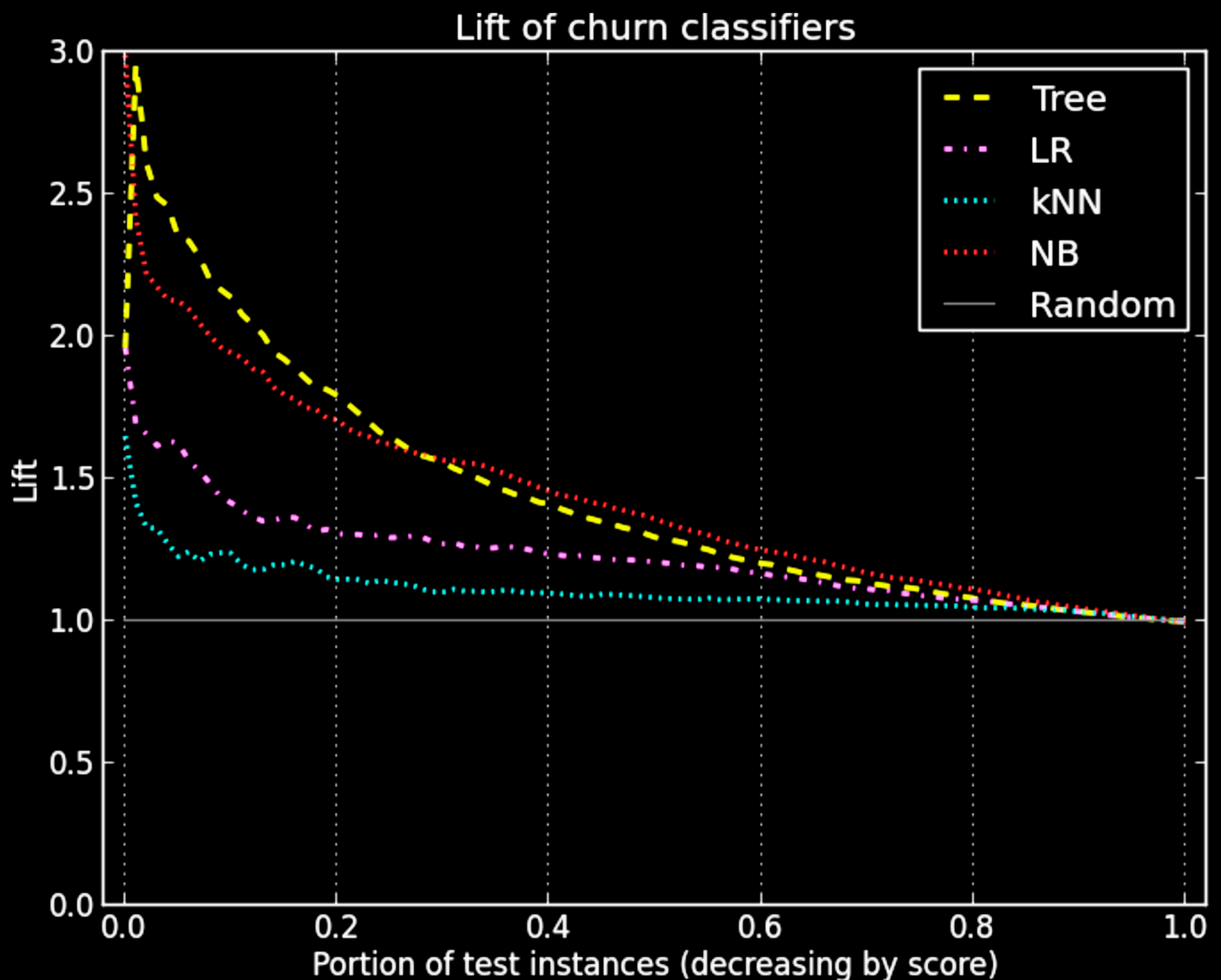
Like	Lift	Like	Lift
Lord Of The Rings	1.69	Wikileaks	1.59
One Manga	1.57	Beethoven	1.52
Science	1.49	NPR	1.48
Psychology	1.46	*Spirited Away*	1.45
The Big Bang Theory	1.43	Running	1.41
Paulo Coelho	1.41	Roger Federer	1.40
The Daily Show	1.40	*Star Trek*	1.39
Lost	1.39	Philosophy	1.38
Lie to Me	1.37	*The Onion*	1.37
How I Met Your Mother	1.35	*The Colbert Report*	1.35
Doctor Who	1.34	*Star Trek*	1.32
Howl's Moving Castle	1.31	Sheldon Cooper	1.30
Tron	1.28	*Fight Club*	1.26
Angry Birds	1.25	*Inception*	1.25
The Godfather	1.23	*Weeds*	1.22

LIFT



Lift of classifiers
Lift
Percentage of test instances (decreasing by score)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
A
B
C
D
Random

LIFT


Lift of churn classifiers
Tree
LR
kNN
NB
Random
Lift
Portion of test instances (decreasing by score)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Computer-based personality judgments are more accurate than those made by humans

Wu Youyou[a,1,2], Michal Kosinski[b,1], and David Stillwell[a]

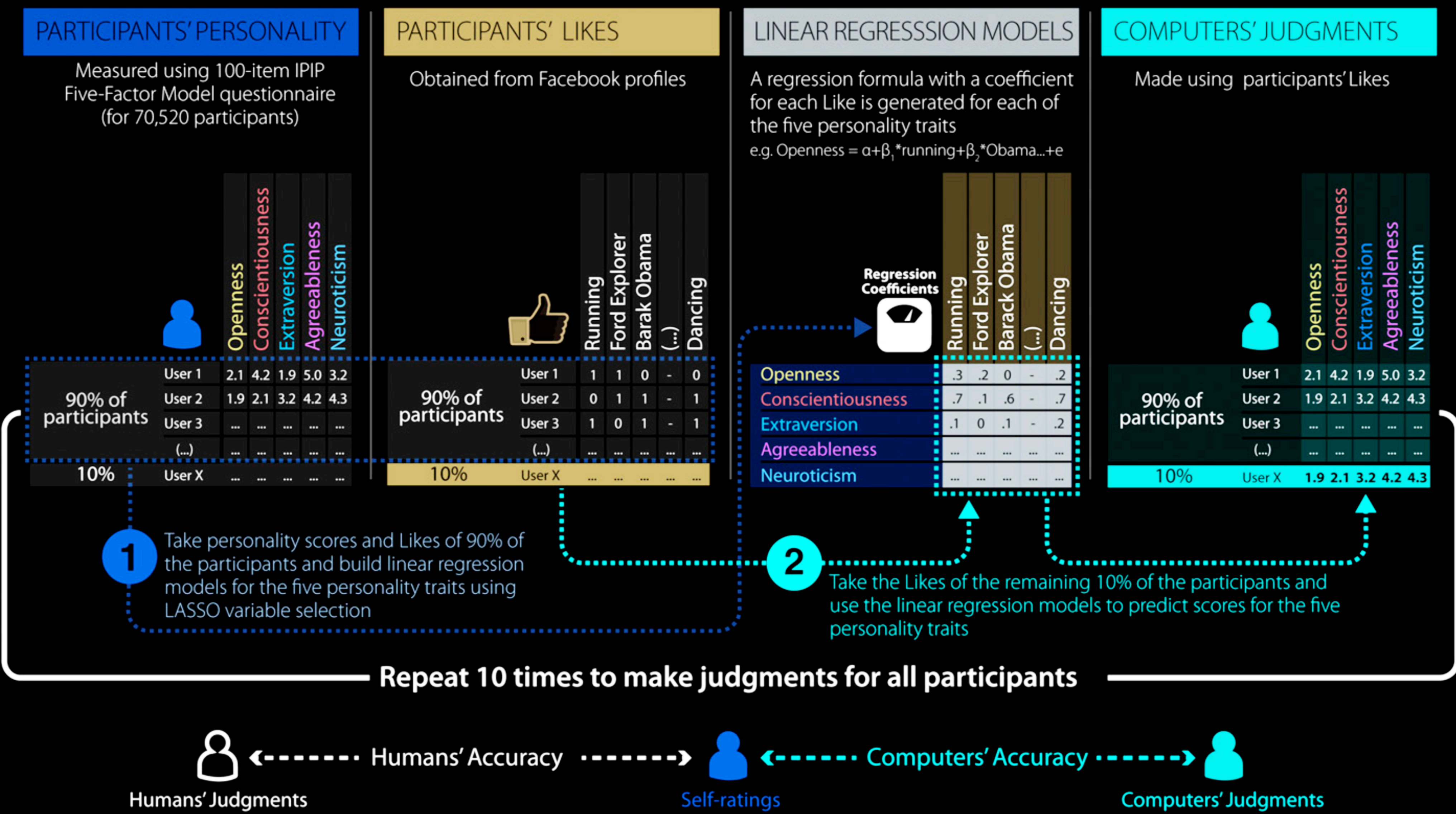
[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and [b]Department of Computer Science, Stanford University, Stanford, CA 94305

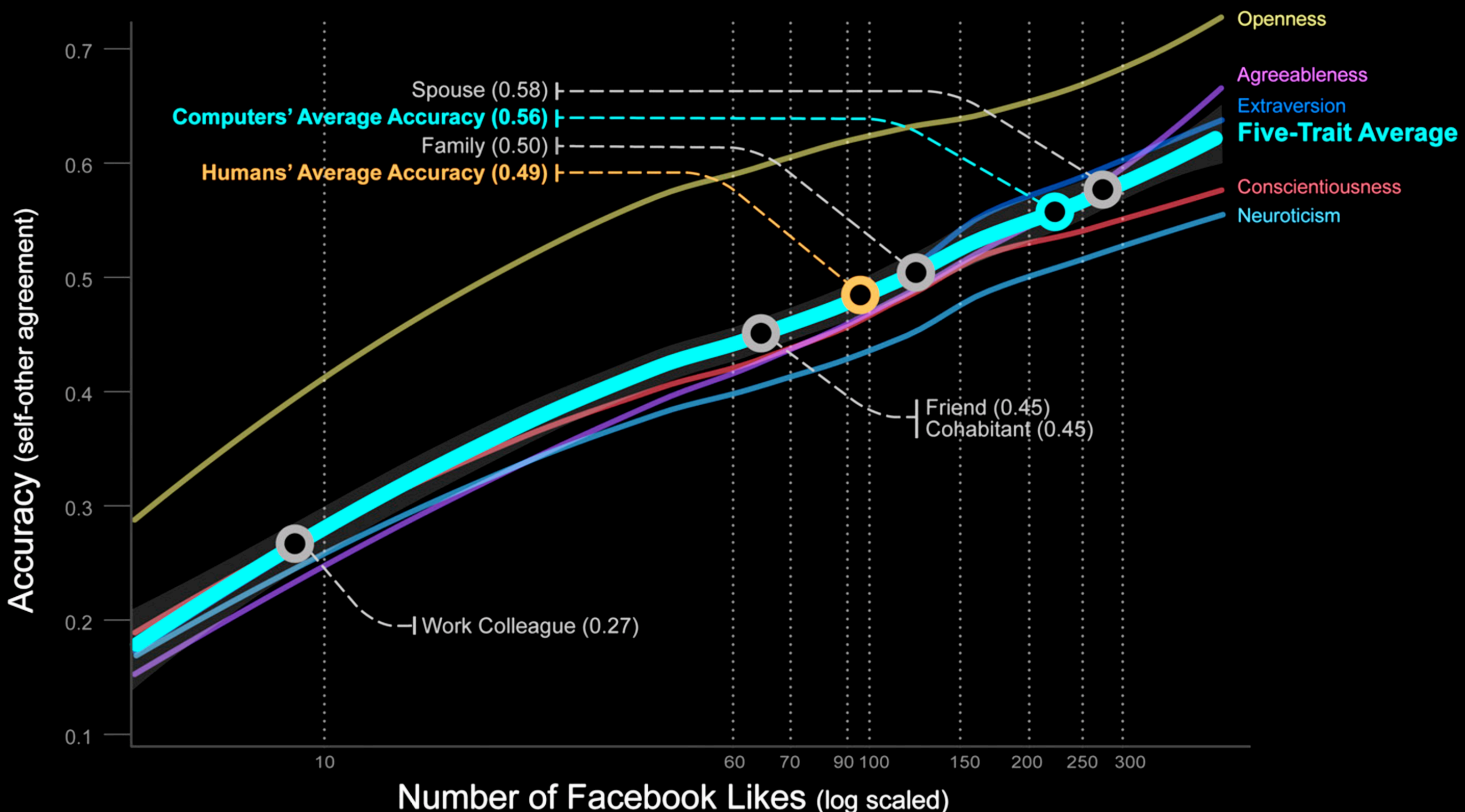
Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

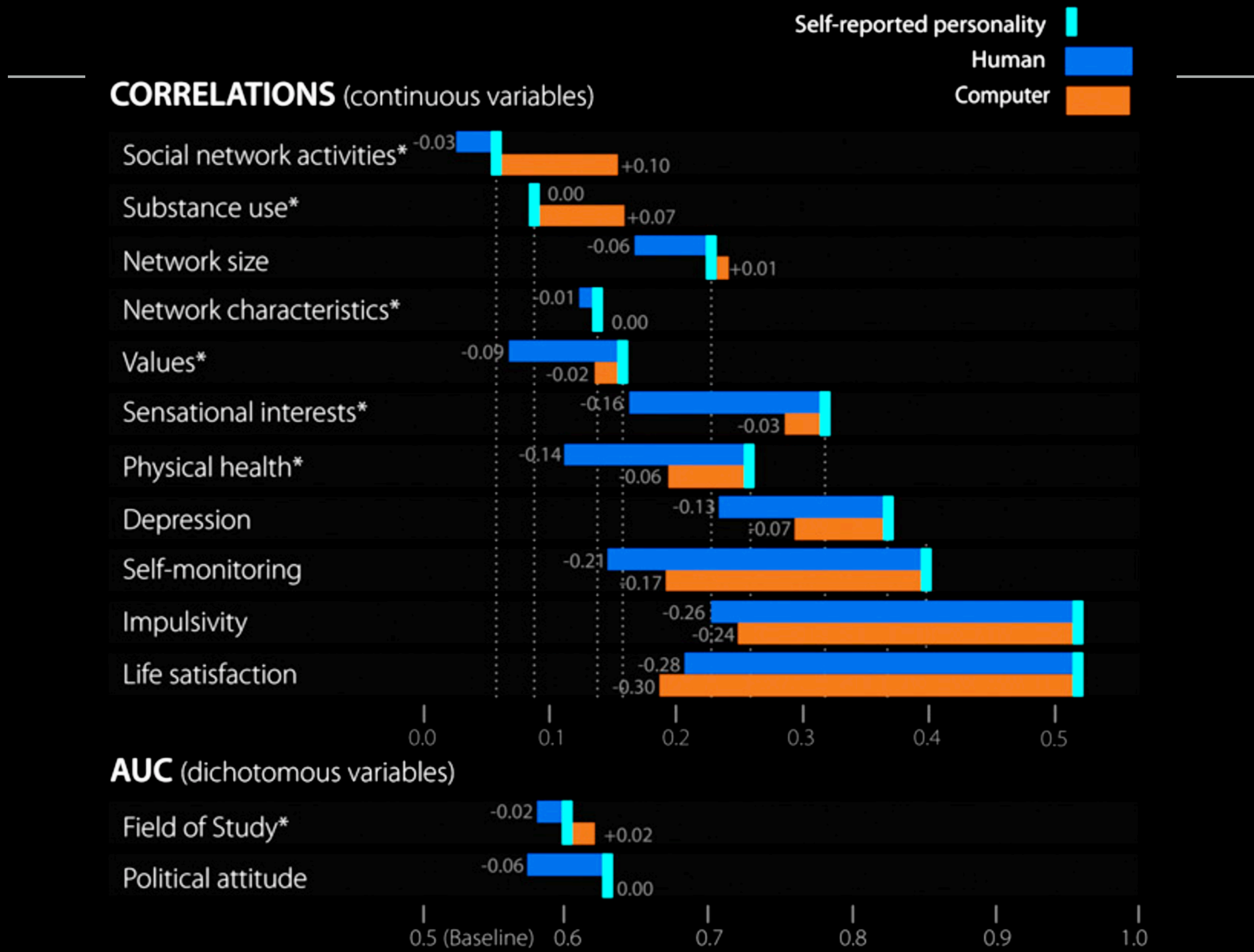
Judging others' personalities is an essential skill in successful social living, as personality is a key driver behind people's interactions, behaviors, and emotions. Although accurate personality judgments stem from social-cognitive skills, developments in machine learning show that computer models can also make valid judgments. This study compares the accuracy of human and computer-based personality judgments, using a sample of 86,220 volunteers who completed a 100-item personality questionnaire. We show that (*i*) computer predictions based on a generic digital footprint (Facebook Likes) are more accurate ($r = 0.56$) than those made by the participants' Facebook friends using a personality questionnaire ($r = 0.49$); (*ii*) computer models show higher interjudge agreement; and (*iii*) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores. Computers outpacing humans in personality judgment presents significant opportunities and challenges in the areas of psychological assessment, marketing, and privacy.

Significance

This study compares the accuracy of personality judgment—a ubiquitous and important social-cognitive activity—between computer models and humans. Using several criteria, we show that computers' judgments of people's personalities based on their digital footprints are more accurate and valid than judgments made by their close others or acquaintances (friends, family, spouse, colleagues, etc.). Our findings highlight that people's personalities can be predicted automatically and without involving human social-cognitive skills.


PARTICIPANTS' PERSONALITY
Measured using 100-item IPIP Five-Factor Model questionnaire (for 70,520 participants)
PARTICIPANTS' LIKES
Obtained from Facebook profiles
LINEAR REGRESSSION MODELS
A regression formula with a coefficient for each Like is generated for each of the five personality traits
e.g. Openness = α+β1*running+β2*Obama...+e
COMPUTERS' JUDGMENTS
Made using participants' Likes
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism
90% of participants
User 1 2.1 4.2 1.9 5.0 3.2
User 2 1.9 2.1 3.2 4.2 4.3
User 3 ... ...
(...) ... ...
10% User X ... ...
Running
Ford Explorer
Barak Obama
(...)
Dancing
90% of participants
User 1 1 1 0 - 0
User 2 0 1 1 - 1
User 3 1 0 1 - 1
(...) ... ...
10% User X ... ...
Regression Coefficients
Running
Ford Explorer
Barack Obama
(...)
Dancing
Openness .3 .2 0 - .2
Conscientiousness .7 .1 .6 - .7
Extraversion .1 0 .1 - .2
Agreeableness ... ...
Neuroticism ... ...
90% of participants
User 1 2.1 4.2 1.9 5.0 3.2
User 2 1.9 2.1 3.2 4.2 4.3
User 3 ... ...
(...) ... ...
10% User X 1.9 2.1 3.2 4.2 4.3
1
Take personality scores and Likes of 90% of the participants and build linear regression models for the five personality traits using LASSO variable selection
2
Take the Likes of the remaining 10% of the participants and use the linear regression models to predict scores for the five personality traits
Repeat 10 times to make judgments for all participants
Humans' Accuracy
Computers' Accuracy
Humans' Judgments
Self-ratings
Computers' Judgments


Accuracy (self-other agreement)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Number of Facebook Likes (log scaled)
10
60
70
90 100
150
200
250
300
Spouse (0.58)
Computers' Average Accuracy (0.56)
Family (0.50)
Humans' Average Accuracy (0.49)
Friend (0.45)
Cohabitant (0.45)
Work Colleague (0.27)
Openness
Agreeableness
Extraversion
Five-Trait Average
Conscientiousness
Neuroticism



Self-reported personality
Human
Computer
CORRELATIONS (continuous variables)
Social network activities* -0.03 +0.10
Substance use* 0.00 +0.07
Network size -0.06 +0.01
Network characteristics* -0.01 0.00
Values* -0.09 -0.02
Sensational interests* -0.16 -0.03
Physical health* -0.14 -0.06
Depression -0.13 -0.07
Self-monitoring -0.21 -0.17
Impulsivity -0.26 -0.24
Life satisfaction -0.28 -0.30
0.0
0.1
0.2
0.3
0.4
0.5
AUC (dichotomous variables)
Field of Study* -0.02 +0.02
Political attitude -0.06 0.00
0.5 (Baseline)
0.6
0.7
0.8
0.9
1.0

REFERENCES

- Reading: DSB Ch 7-9
- Optional: CASI Ch 12