QCB/MOL/COS 455/551 Introduction to genomics & computational biology

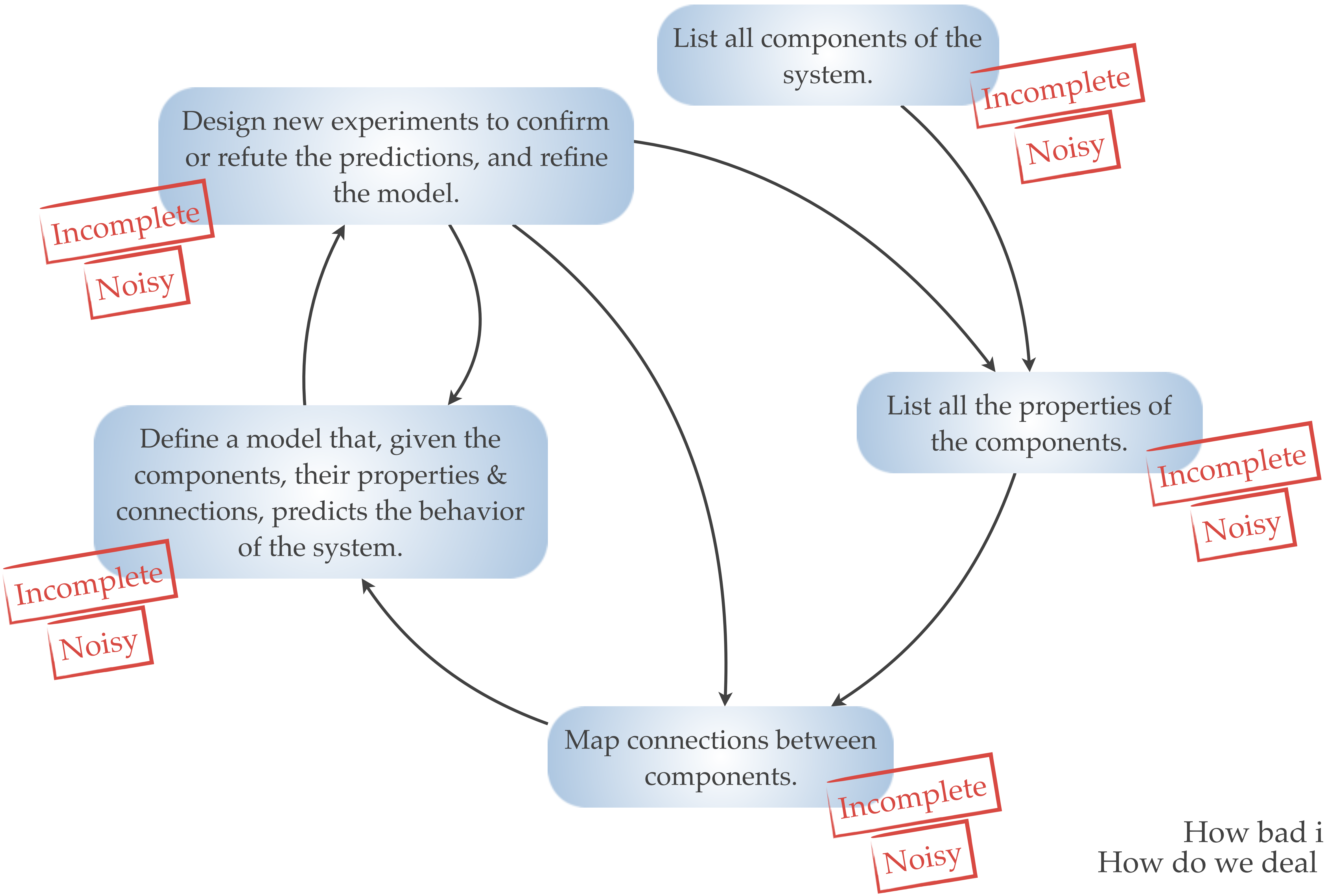
Basic statistical concepts in “omics”

October 11, 2016

Overview

1. True and false positives, precision/recall, ROC curves
2. Gene Ontology and other functional standards
3. P-values, multiple testing correction
4. Data exploratory analysis, heatmaps, clustering, visualization

The workflow of systems biology



List all components of the system.

Incomplete

Noisy

List all the properties of the components.

Incomplete

Noisy

Map connections between components.

Incomplete

Noisy

Define a model that, given the components, their properties & connections, predicts the behavior of the system.

Incomplete

Noisy

Design new experiments to confirm or refute the predictions, and refine the model.

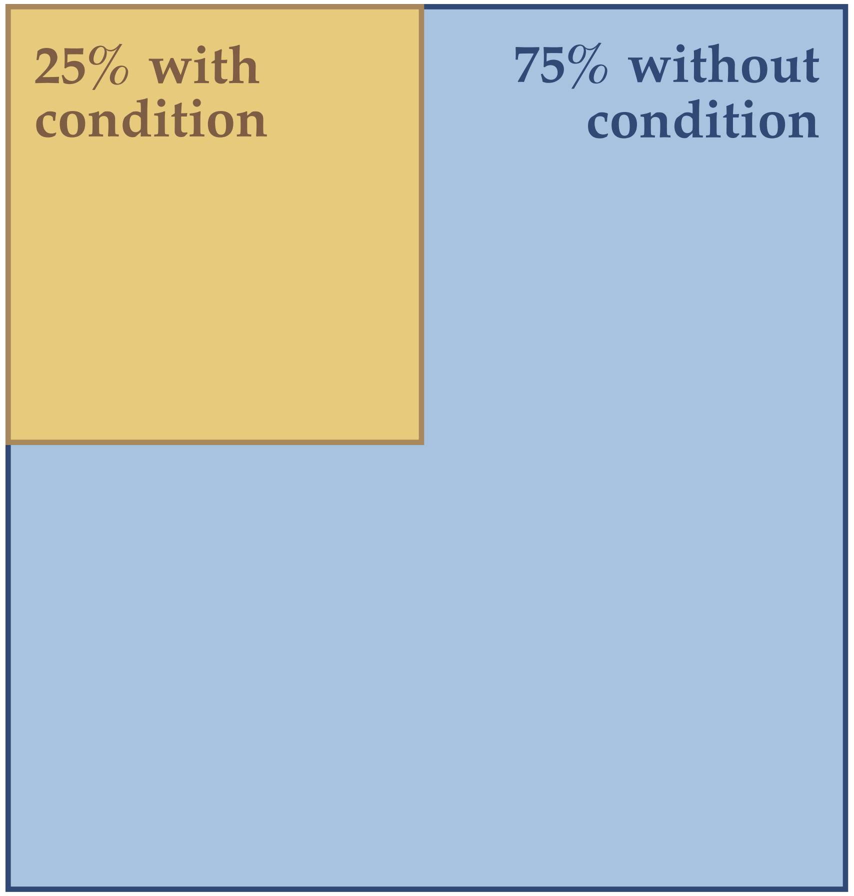
Incomplete

Noisy

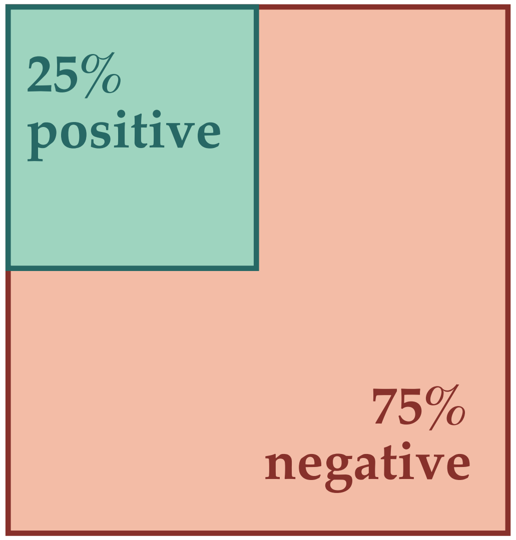
How bad is it?
How do we deal with that?

Reality vs Test

Population



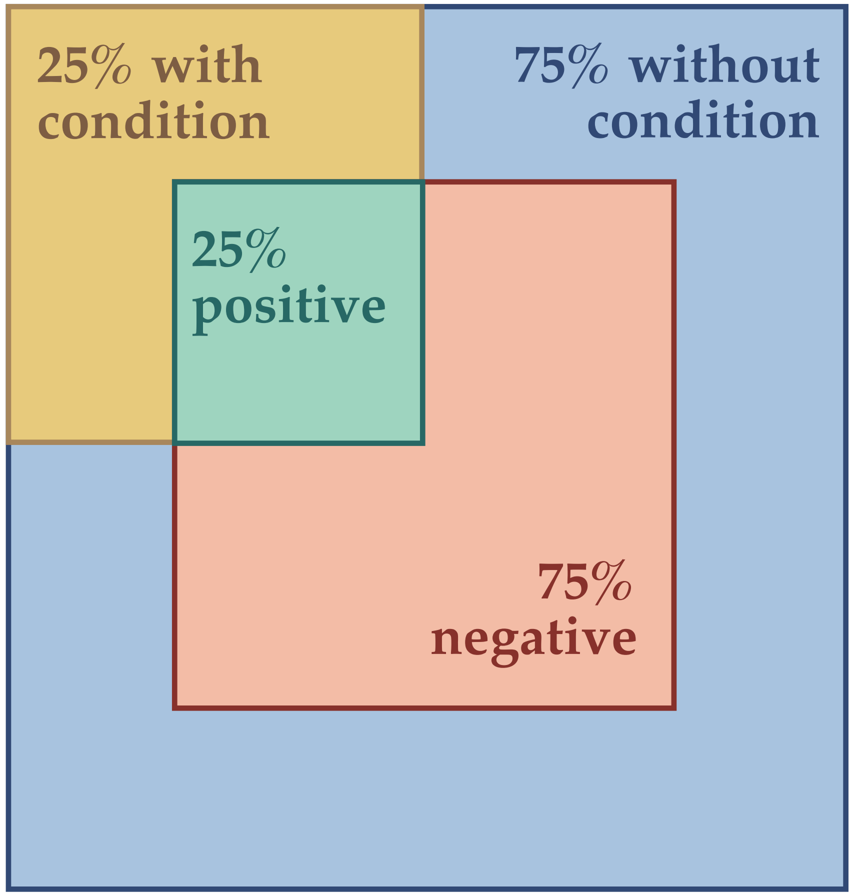
Test for the condition



Is this a good test?
It depends on how the positives & the negatives align with individuals with & without condition.

Reality vs Test

Best case scenario



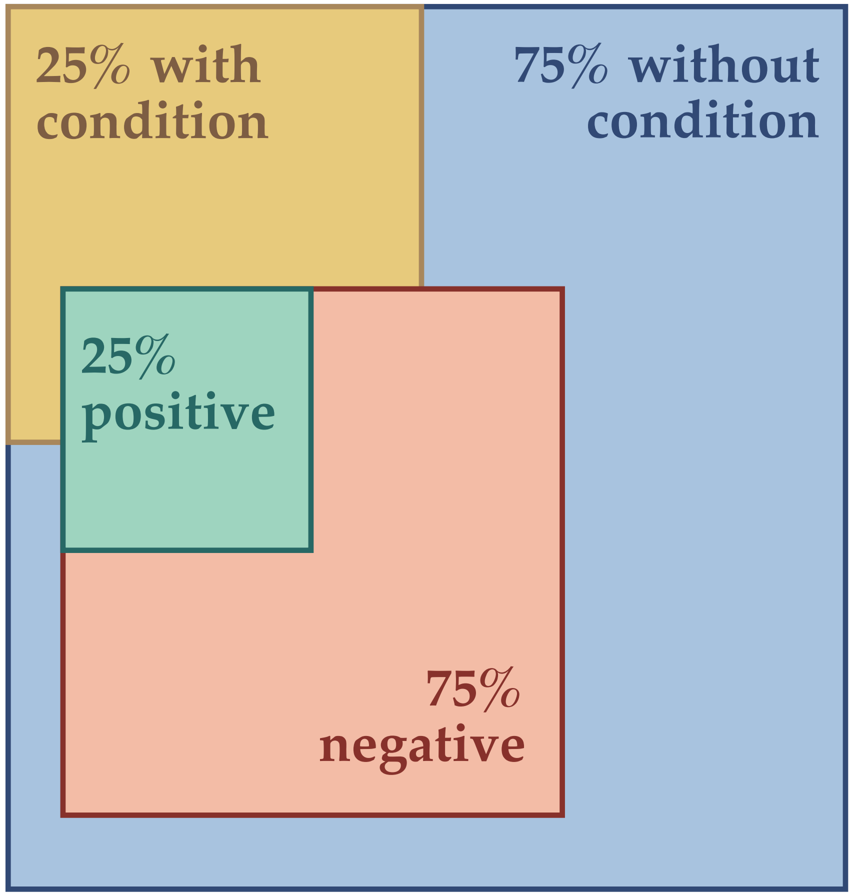
25% with condition

75% without condition

25% positive

75% negative

Most likely scenario



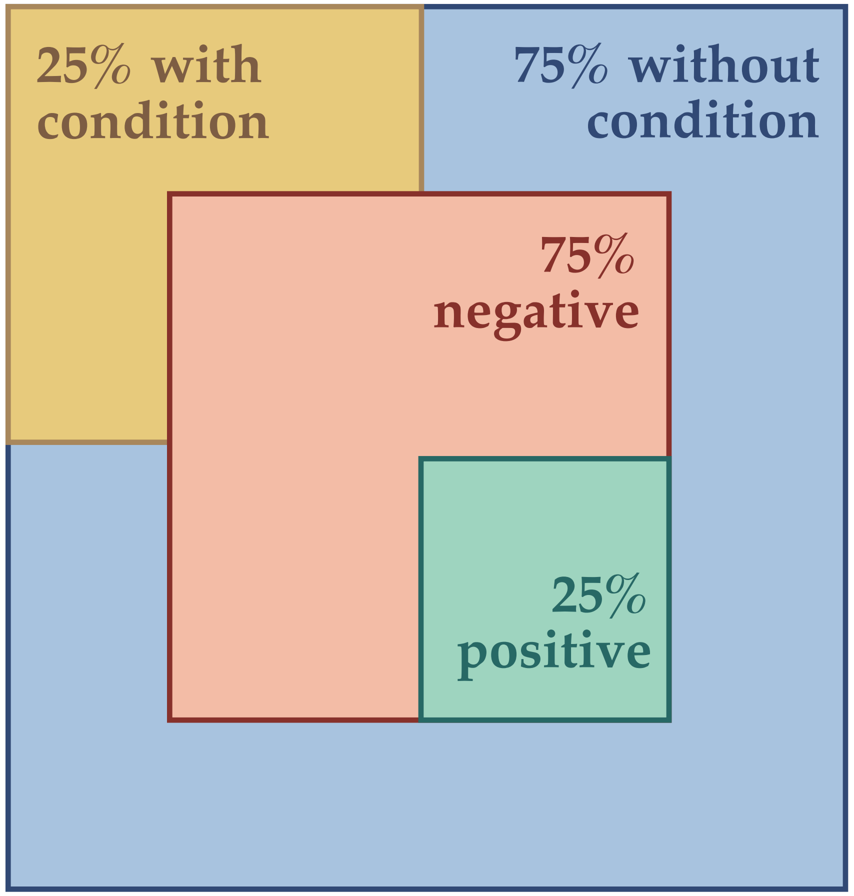
25% with condition

75% without condition

25% positive

75% negative

Worst case scenario



25% with condition

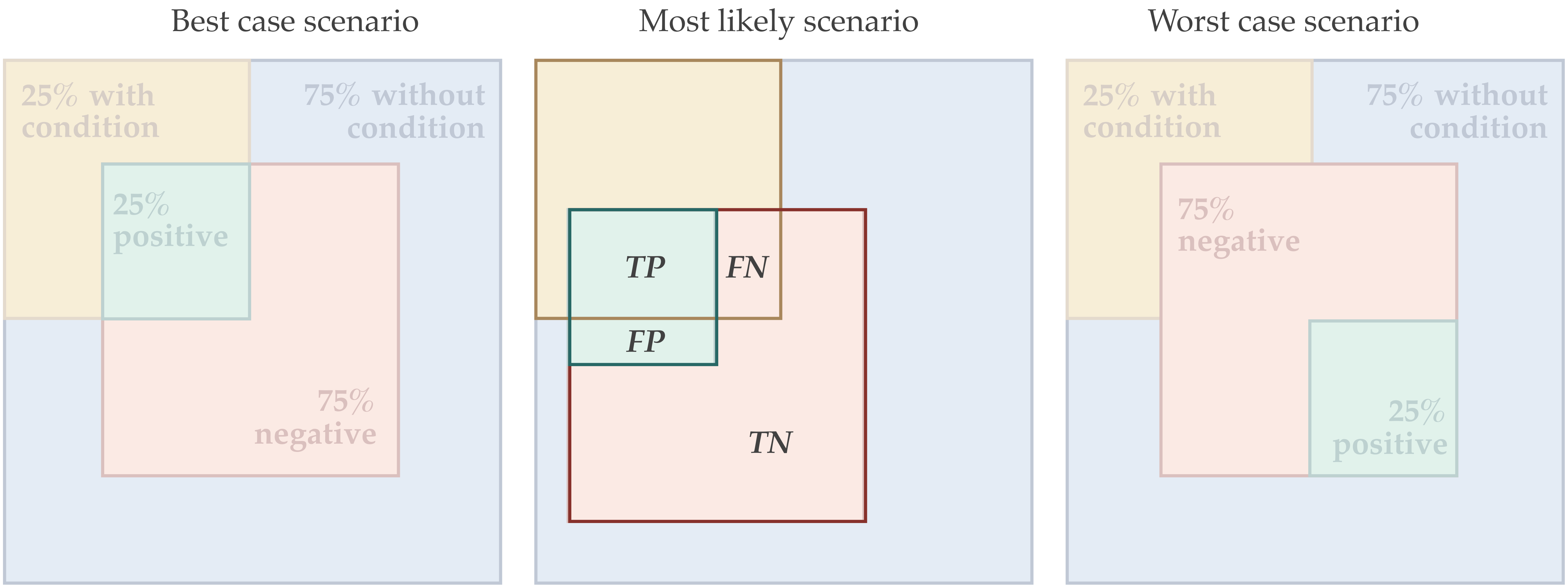
75% without condition

75% negative

25% positive

Can I quantify the quality of this test?

Some terminology (1)



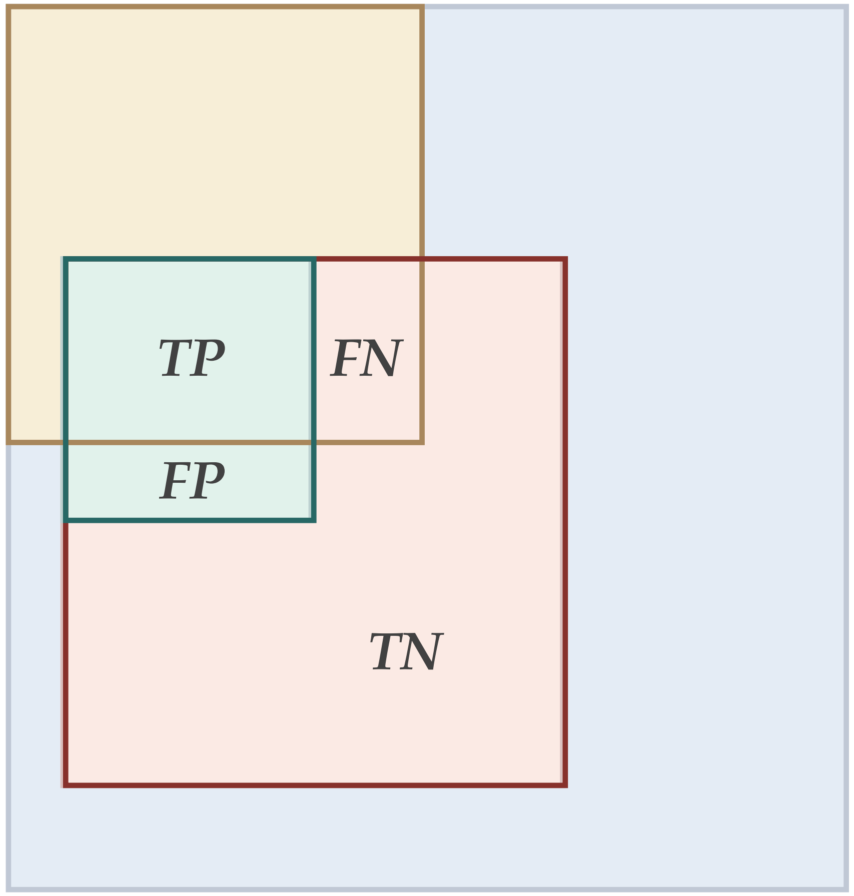
TP = true positives
FP = false positives
TN = true negatives
FN = false negatives

TP + *FP* = all positives in my test
TN + *FN* = all negatives in my test

TP + *FN* = all individuals with condition (among tested)
FP + *TN* = all individuals without condition (among tested)

Some terminology (2)

Most likely scenario



1. How many individuals **with condition** will be labelled as **positive**?

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

True Positive Rate
= Sensitivity
= Recall

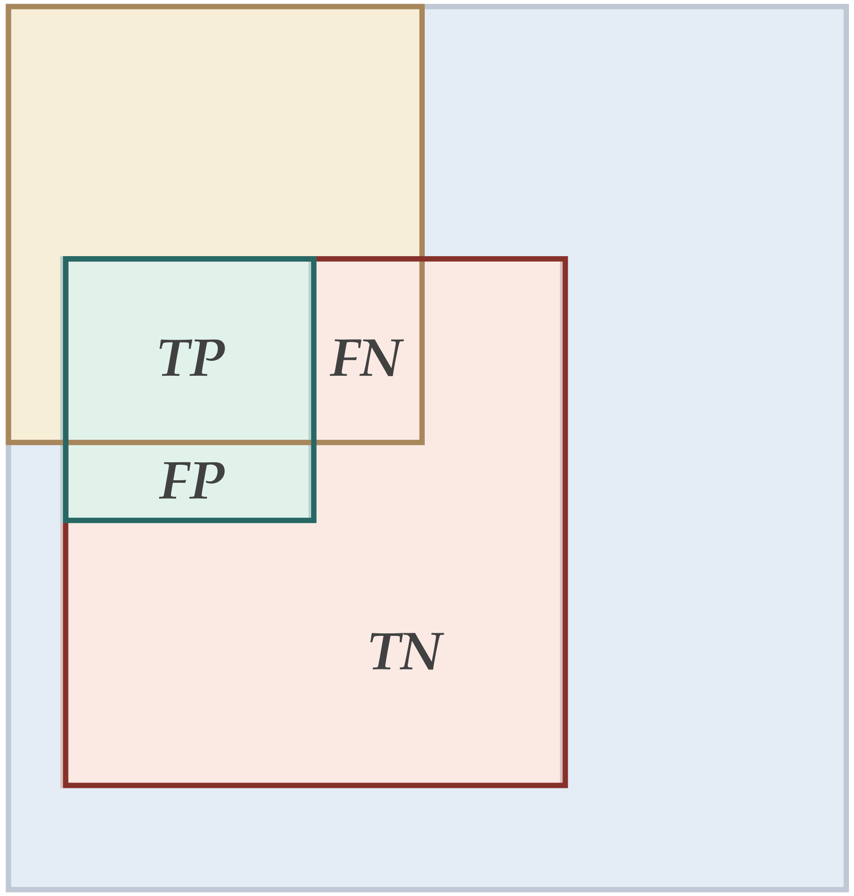
2. How many individuals **without the condition** will be labelled as **negative**?

$$TNR = \frac{TN}{TN + FP}$$

True Negative Rate
= Specificity

Some terminology (3)

Most likely scenario



1. How many individuals **without the condition** will be labelled as **positive**?

$$FPR = 1 - TNR = \frac{FP}{TN + FP}$$

False Positive Rate

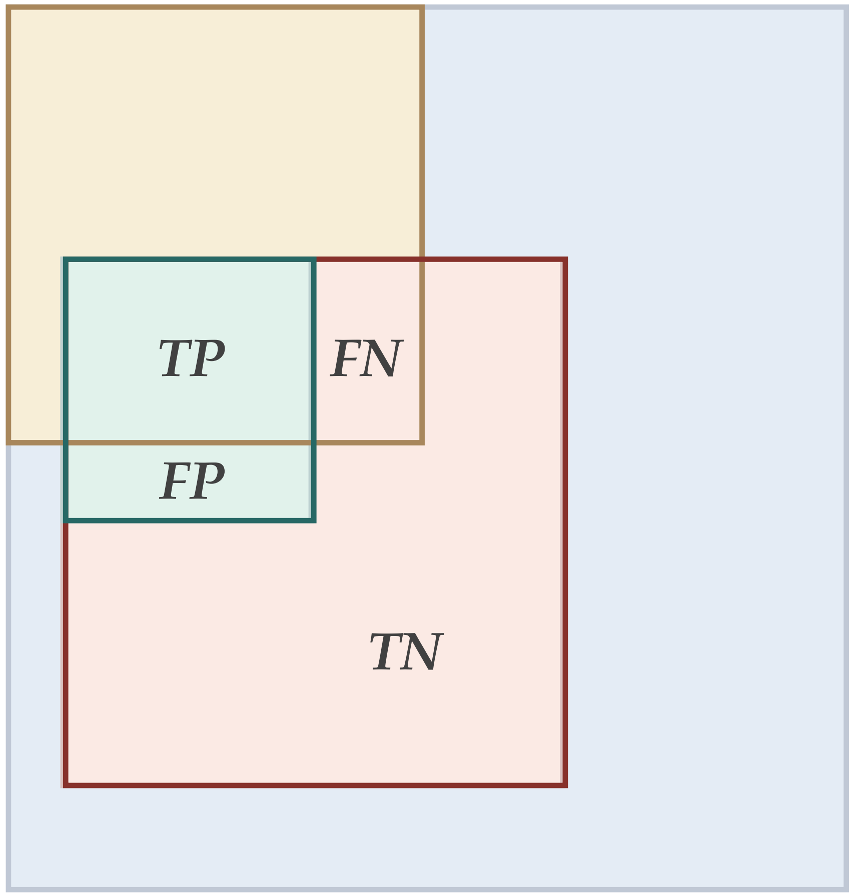
2. How many individuals **with the condition** will be labelled as **negative**?

$$FNR = 1 - TPR = \frac{FN}{FN + TP}$$

False Negative Rate

The lesser of two evils: high FPR or high FNR?

Most likely scenario



Not all of these measurements are equally informative in different contexts.

A) Development of a clinical test for early cancer diagnosis:

High *false positive rate (FPR)*	Many healthy individuals labelled as positives.	OK*
High *false negative rate* (FNR)	Many sick individuals labeled as negative.	Not OK

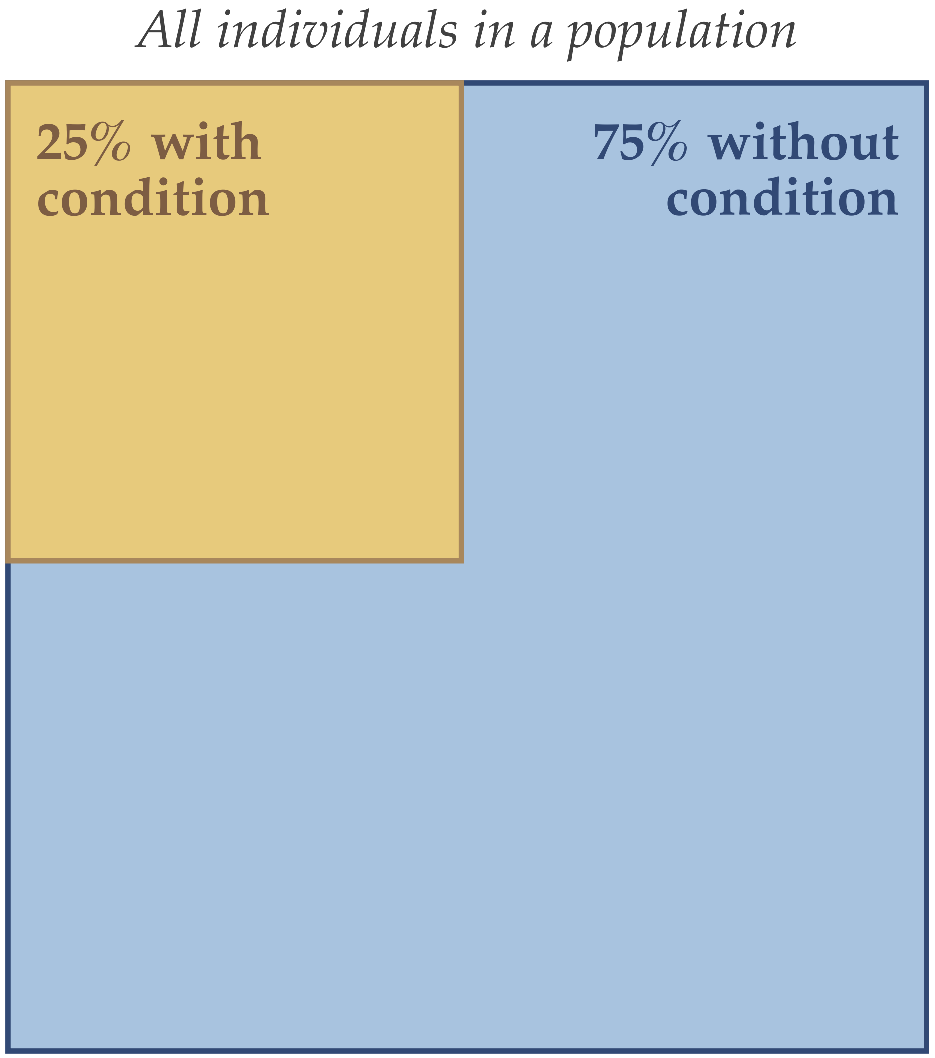
B) Development of a computational method for predicting physical interactions between proteins:

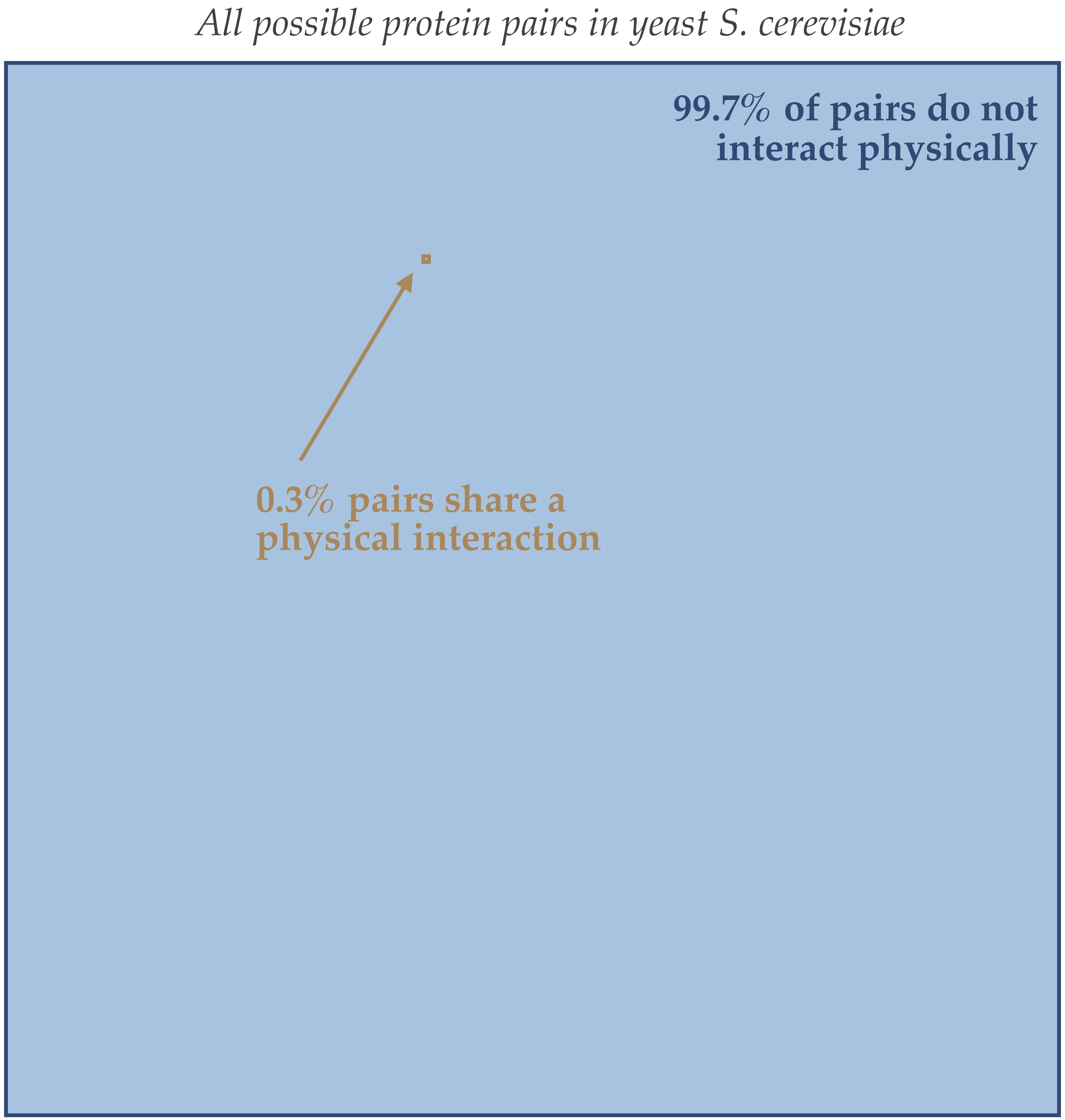
High *false positive rate (FPR)*	Many random proteins labelled as interacting.	Not OK
High *false negative rate* (FNR)	Many interactions missed.	OK

* Steven Salzberg's (@StevenSalzberg1) post on Forbes about genetic testing for breast cancer: http://goo.gl/xfj6wb (Sept. 15, 2014)

* Elizabeth Holmes & Theranos (Kidd et al., "Evaluation of direct-to-consumer low-volume lab tests in healthy adults", J Clin Invest, 2016)

The lesser of two evils: high FPR or high FNR?




All possible protein pairs in yeast S. cerevisiae
99.7% of pairs do not interact physically
0.3% pairs share a physical interaction

High *false positive rate (FPR)*
= Many random protein pairs are labelled as interacting.
= Huge contamination of a tiny dataset.

High *false negative rate (FNR)*
= Many interactions missed.
= Not great, but at least the ones I have are real.

So, should I aim for low *FPR*?

Low FPR is not informative for rare events

$$FPR = 1 - TNR = \frac{FP}{TN + FP}$$

Reality:

Total number of protein pairs = 18 000 000
Interacting = 54 000 = 0.3%
Non-interacting = 17 946 000 = 99.7%

New method:

Predicted to interact = 54 000
Predicted to not interact = 17 946 000

The true interactions & the predicted ones only overlap by half.

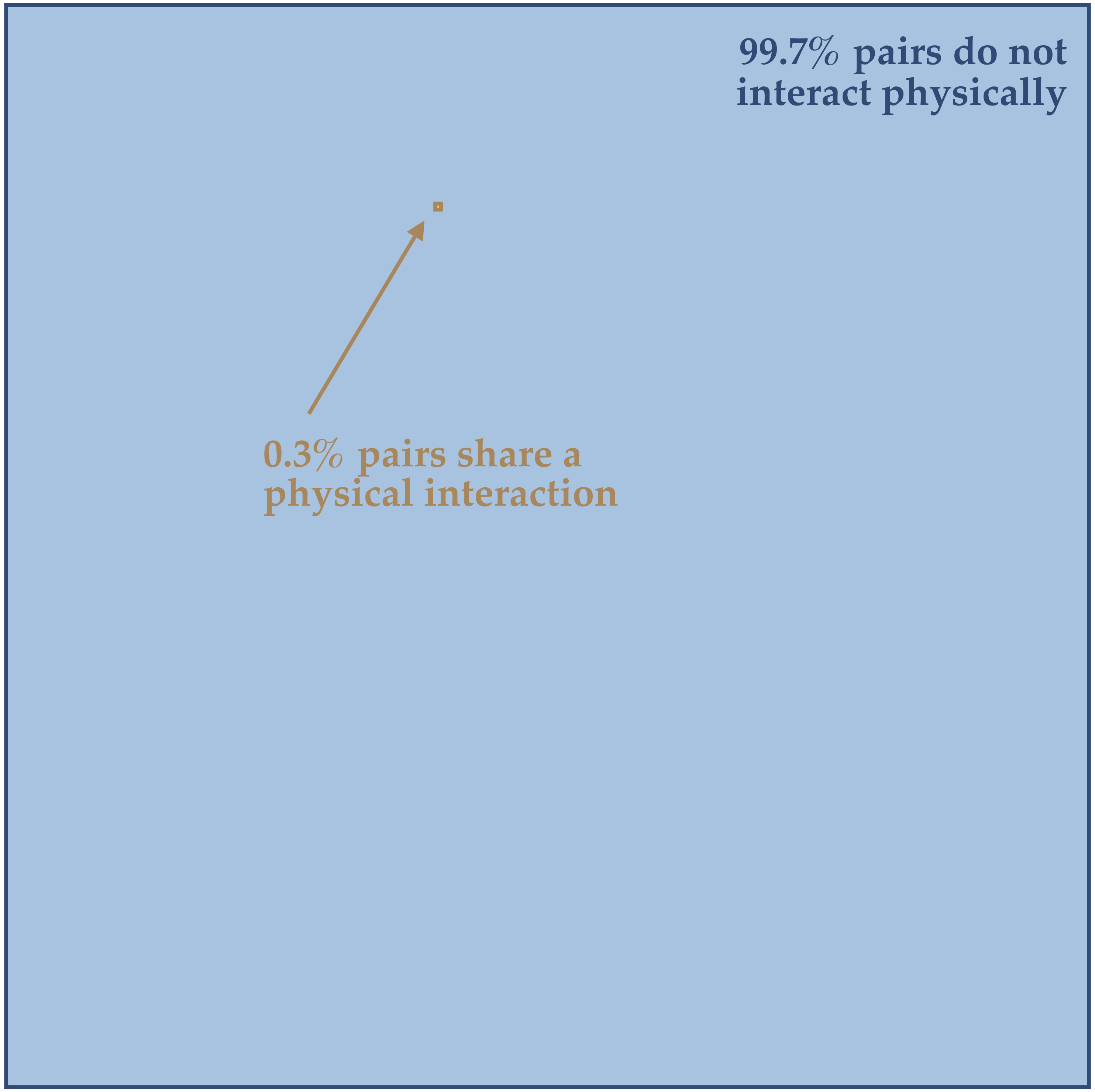
$FPR = 27\,000 \;/\; 17\,946\,000 = \;0.15\%$

FPR very low but half the data is false.

General rule:

The more specific the phenomenon (i.e., the less frequently it is observed), the less informative *FPR* is.

All possible protein pairs in yeast S. cerevisiae


99.7% pairs do not interact physically
0.3% pairs share a physical interaction

Better estimate of false positives: FDR

$$FDR = \frac{FP}{FP + TP}$$ False Discovery Rate

Total number of protein pairs = 18 000 000
Interacting = 54 000 = 0.3%
Non-interacting = 17 946 000 = 99.7%

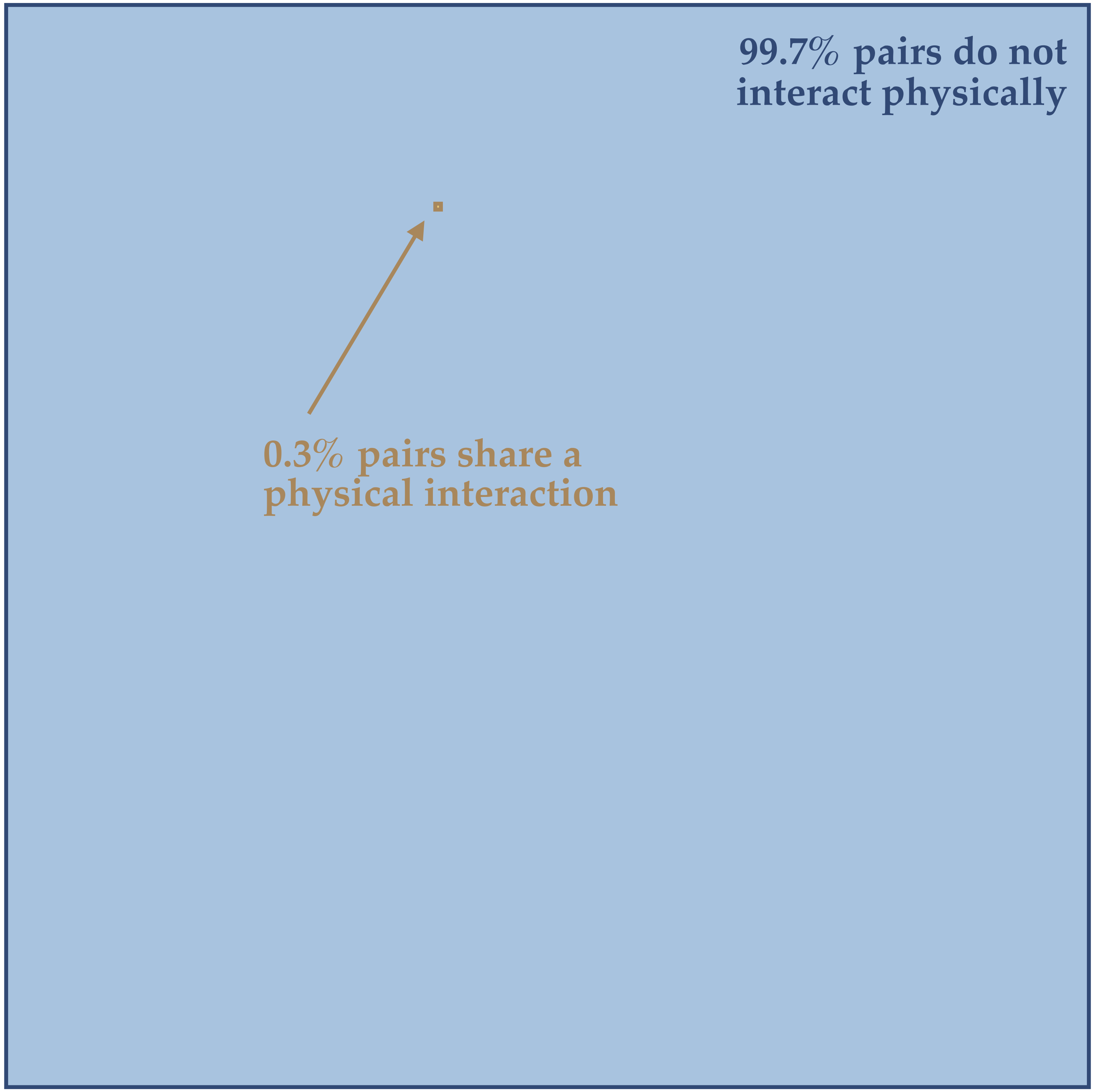
Predicted to interact = 54 000
Predicted to not interact = 17 946 000

The true interactions & the predicted ones only overlap by half.

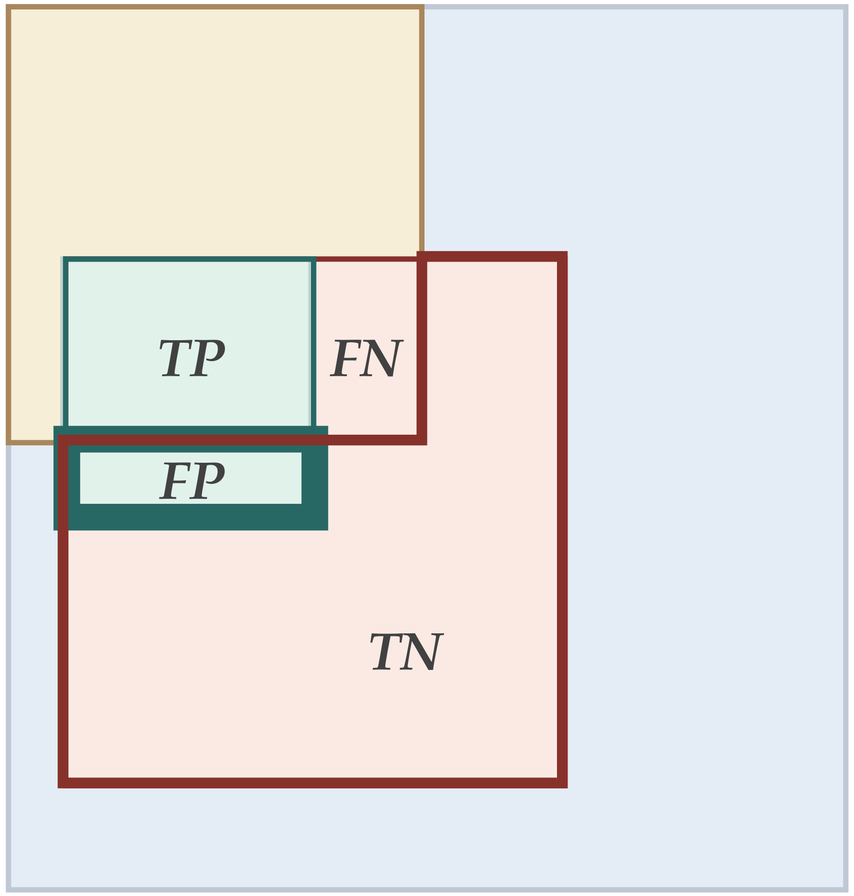
FDR = 27 000 / 54 000 = 50%

1 – ***FDR*** = precision

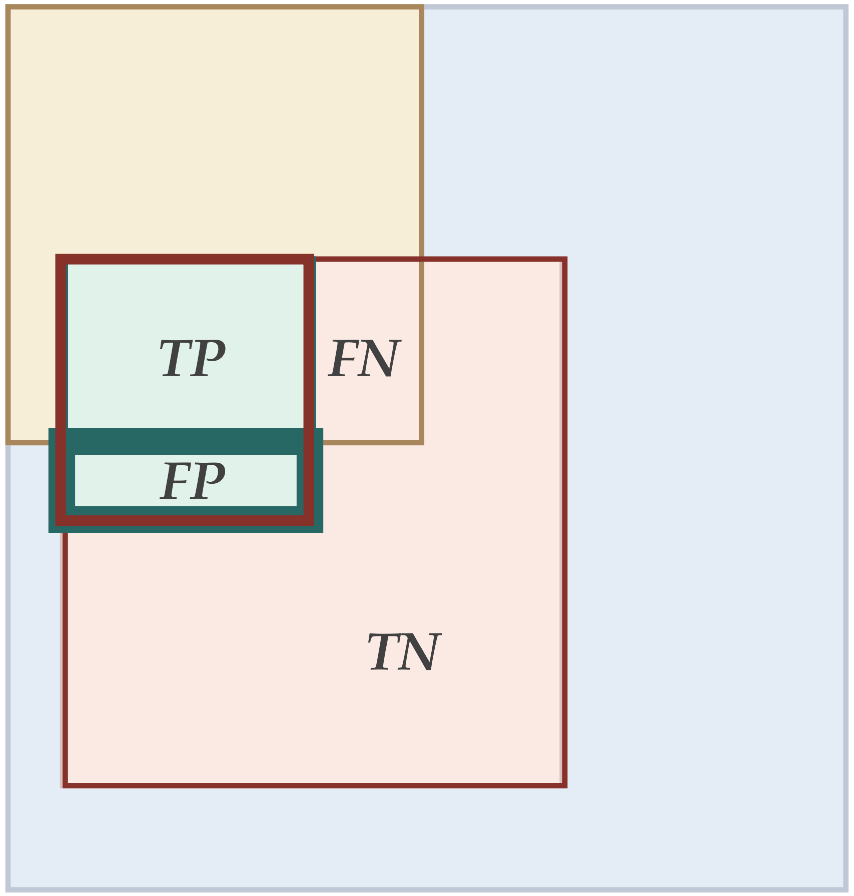
All possible protein pairs in yeast S. cerevisiae


99.7% pairs do not interact physically
0.3% pairs share a physical interaction

FPR vs FDR



$$FPR = 1 - TNR = \frac{\boxed{FP}}{\boxed{TN + FP}}$$

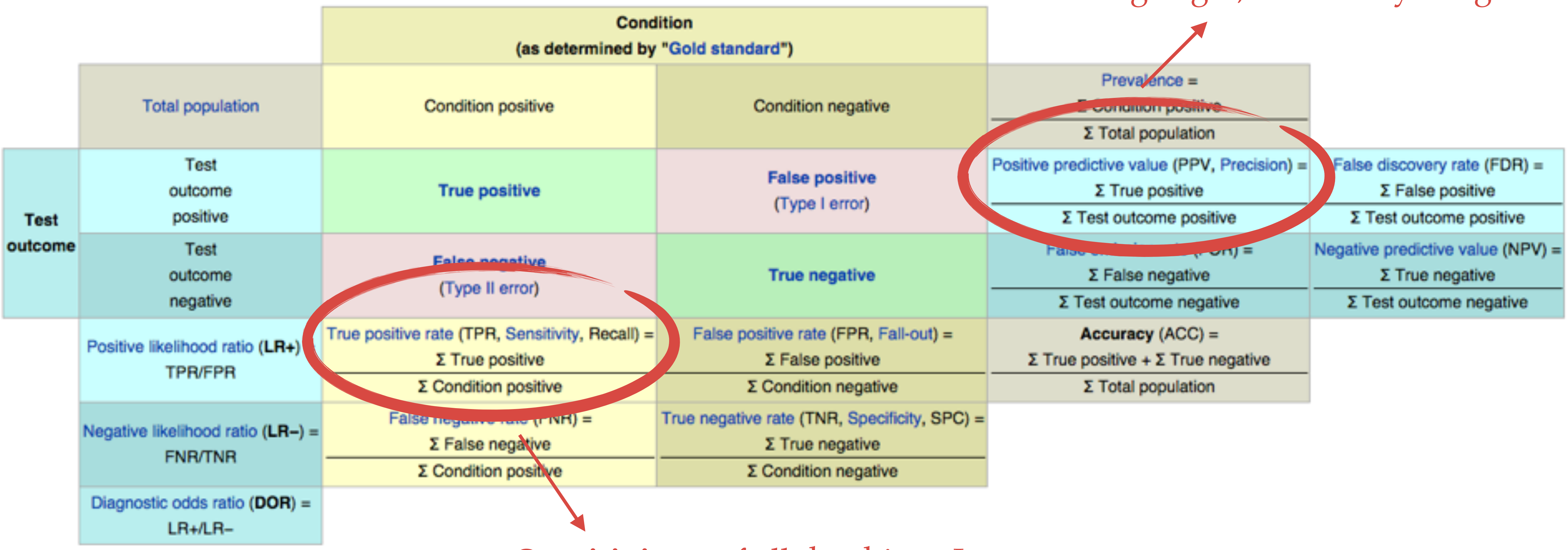


$$FDR = \frac{\boxed{FP}}{\boxed{FP + TP}}$$

Cheat sheet (nobody ever remembers all of this)

http://en.wikipedia.org/wiki/Accuracy_and_precision

Precision = 1 – *FDR* = of all the things I got, how many are good?



		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive Σ Total population	
Test outcome	Test outcome positive	True positive	False positive (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive Σ Test outcome positive	False discovery rate (FDR) = Σ False positive Σ Test outcome positive
	Test outcome negative	False negative (Type II error)	True negative	False omission rate (FOR) = Σ False negative Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative Σ Test outcome negative
	Positive likelihood ratio (LR+) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative Σ Total population	
	Negative likelihood ratio (LR−) = FNR/TNR	False negative rate (FNR) = Σ False negative Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative Σ Condition negative		
	Diagnostic odds ratio (DOR) = LR+/LR−				

Sensitivity = of all the things I was supposed to get, how many did I get?

Print it, tape it on a wall & look at it every time you read a paper.

Evaluating binary vs quantitative data

Binary data

Test	Ref		
1	1	P	**TP**
0	1	N	**FN**
0	1	N	**FN**
0	0	N	**TN**
1	0	P	**FP**
0	1	N	**FN**
0	1	N	**FN**
0	0	N	**TN**
1	0	P	**FP**
1	1	P	**TP**
0	1	N	**FN**
0	1	N	**FN**
0	0	N	**TN**

$$\text{Recall} = TPR = \frac{\#\boldsymbol{TP}}{\#\boldsymbol{TP} + \#\boldsymbol{FN}} = \frac{2}{8} = 0.25$$

$$\text{Precision} = 1 - FDR = \frac{\#\boldsymbol{TP}}{\#\boldsymbol{TP} + \#\boldsymbol{FP}} = \frac{2}{4} = 0.5$$

Evaluating binary vs quantitative data

Quantitative data

Test	Ref	
0.21	0	**?**
0.65	0	**?**
0.37	0	**?**
0.42	1	**?**
0.54	1	**?**
0.11	0	**?**
0.69	1	**?**
0.75	1	**?**
0.83	1	**?**
0.31	1	**?**
0.22	1	**?**
0.46	1	**?**
0.17	0	**?**

Sort data from highest to lowest confidence

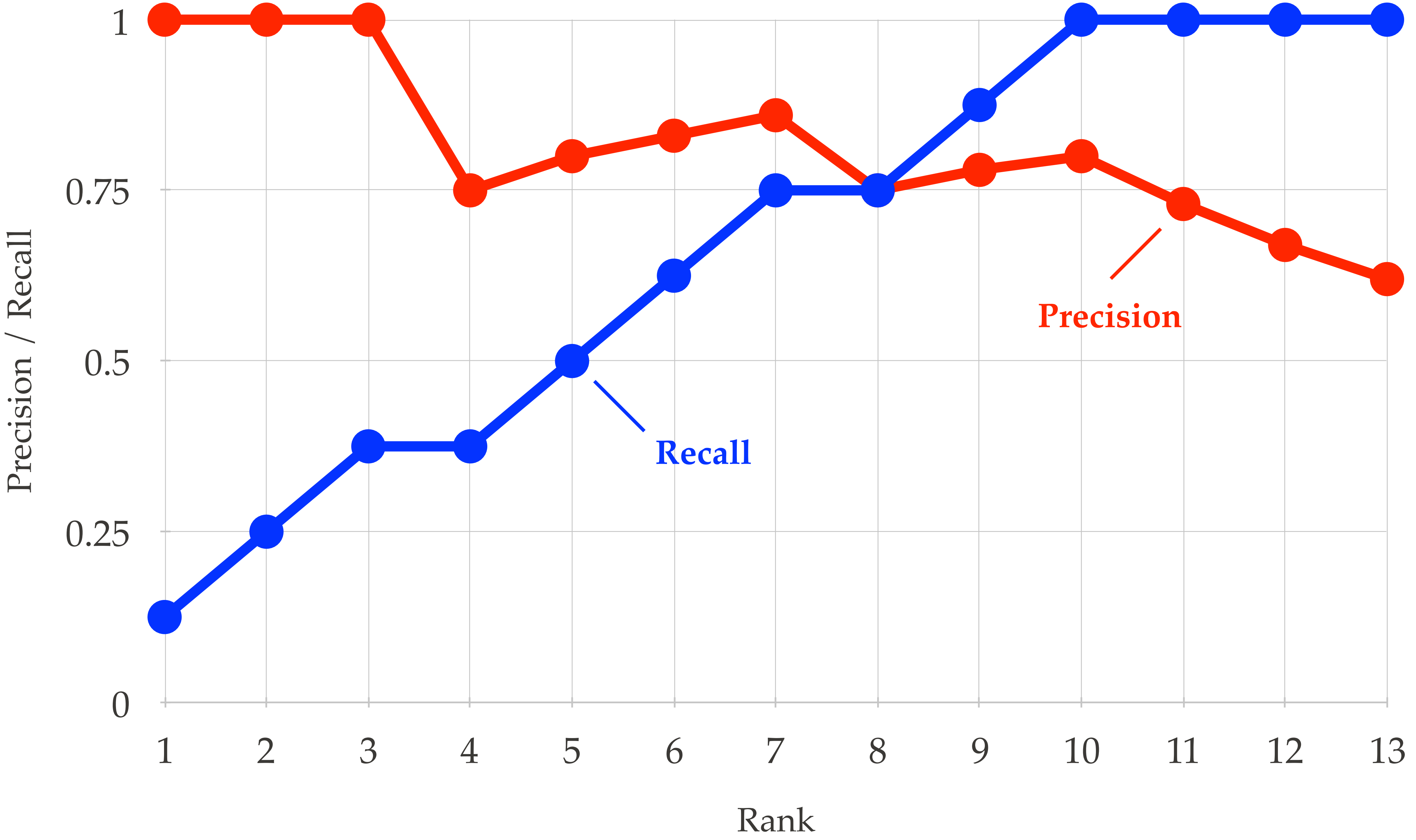
Ask the question:

If I draw the cutoff here, how good would my dataset be?

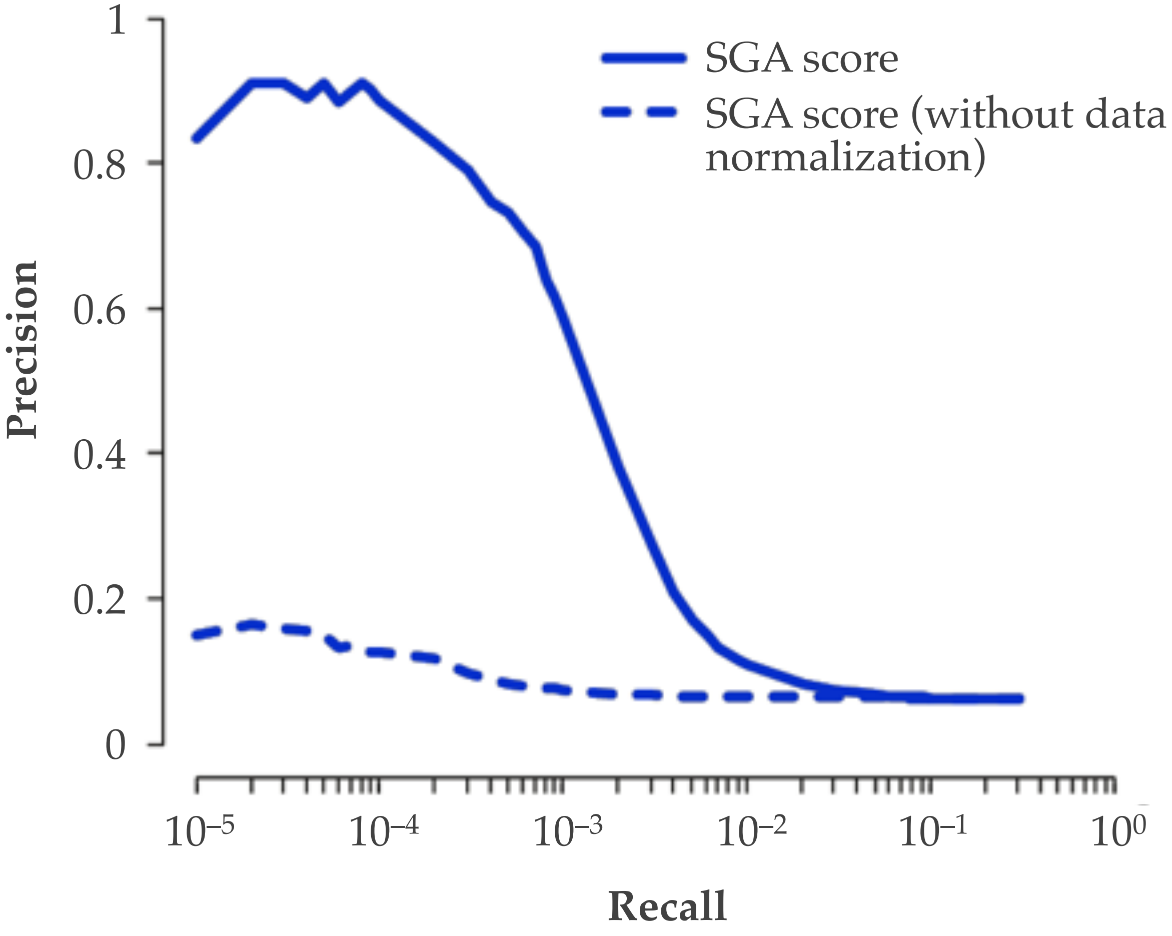
Quantitative data sorted

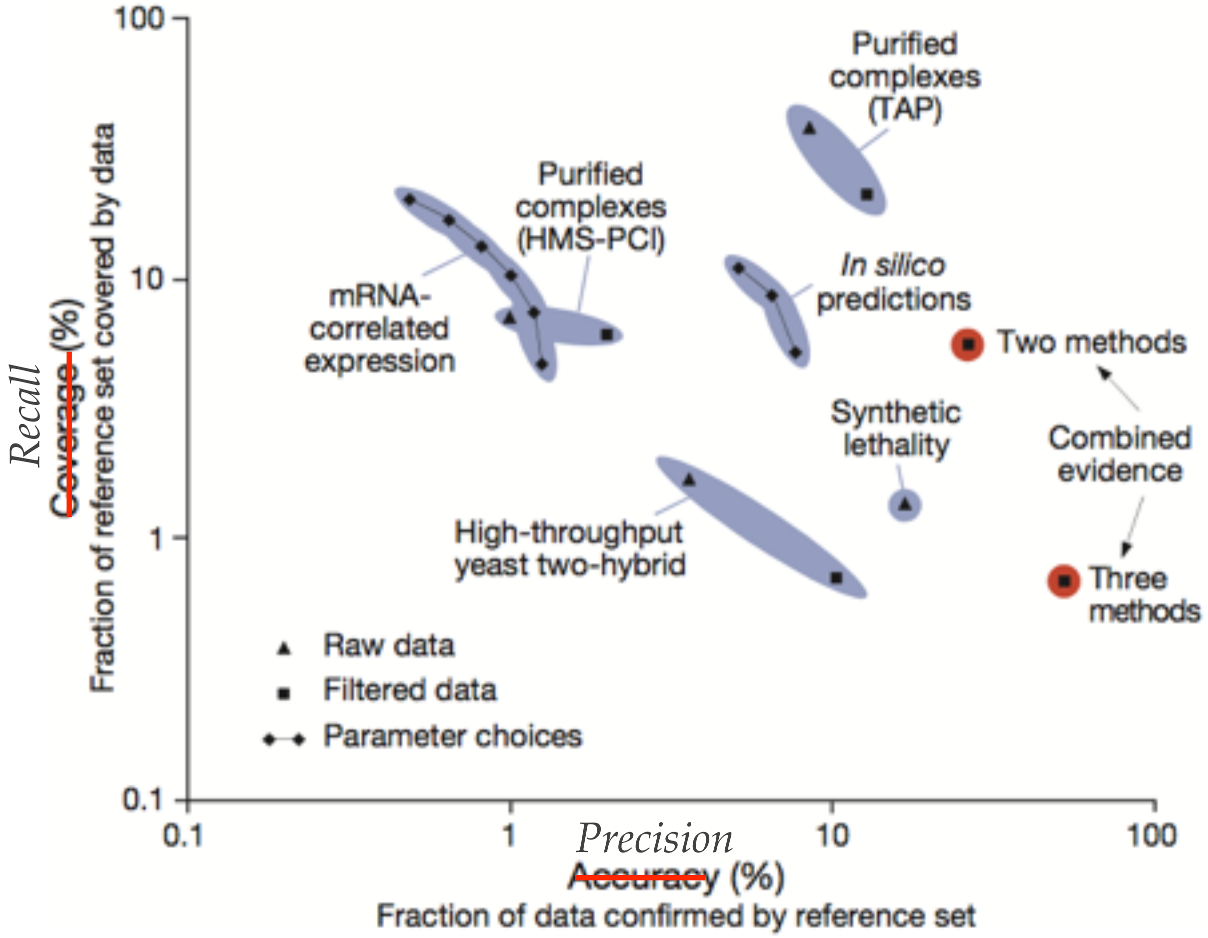
Test	Ref	P	TP	Recall	Precision
0.83	1	**1**	**1**	**1/8**	**1/1**
0.75	1	**2**	**2**	**2/8**	**2/2**
0.69	1	**3**	**3**	**3/8**	**3/3**
0.65	0	**4**	**3**	**3/8**	**3/4**
0.54	1	**5**	**4**	**4/8**	**4/5**
0.46	1	**6**	**5**	**5/8**	**5/6**
0.42	1	**7**	**6**	**6/8**	**6/7**
0.37	0	**8**	**6**	**6/8**	**6/8**
0.31	1	**9**	**7**	**7/8**	**7/9**
0.22	1	**10**	**8**	**8/8**	**8/10**
0.21	0	**11**	**8**	**8/8**	**8/11**
0.17	0	**12**	**8**	**8/8**	**8/12**
0.11	0	**13**	**8**	**8/8**	**8/13**

Precision/recall analysis for defining data thresholds



Precision/recall analysis for comparing datasets

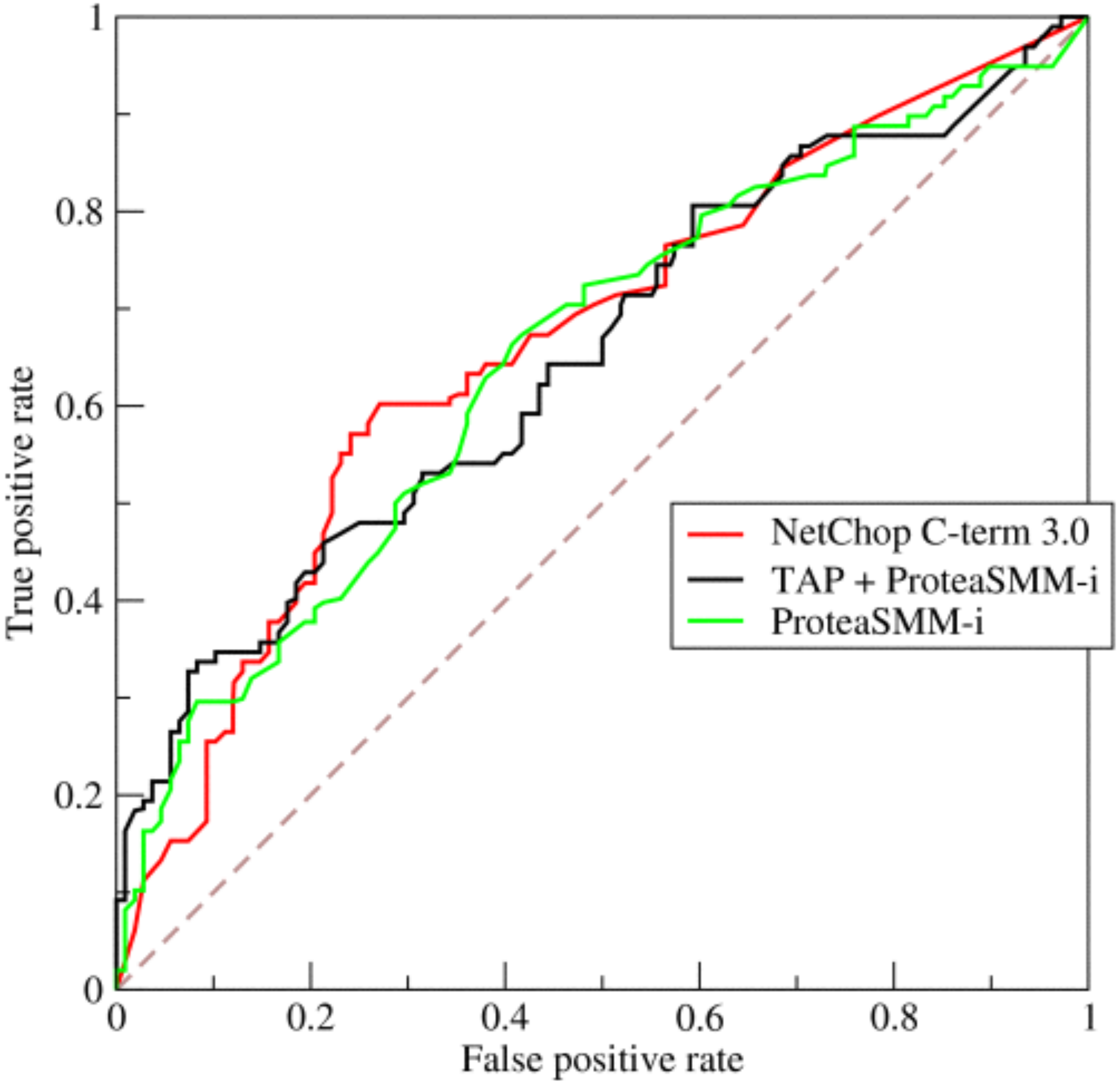




Baryshnikova~Myers, Nat Methods, 2010

von Mering~Bork, Nature, 2002

Receiver operating curve (ROC)



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

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

$$FPR = 1 - TNR = \frac{FP}{TN + FP}$$

AUC = Area Under the ROC Curve

Often used to associate a method with a single number, instead of a plot.

A few final thoughts about precision & recall

- Precision & recall only tell you how well your data aligns with a reference standard.
- The estimate of your data quality will therefore depend on the standard you choose.
- For example, if you reference standard is incomplete, any novel finding will be labelled as False Positive. If a new dataset uncovers a lot of novel biology, it might perform poorly in the precision/recall analysis.
- When using precision/recall to compare datasets, make sure you are comparing them on a common ground (same tested universe, same standard).
- Any estimate is more reliable if supported by multiple standards.
- Any estimate is more reliable if compared to an alternative hypothesis (p-values).