Introduction to Modeling

6.872/HST950

Why build Models?

- To predict (identify) something
  - Diagnosis
  - Best therapy
  - Prognosis
  - Cost
- To understand something
  - Structure of model *may* correspond to structure of reality

Where do models come from?

- Pure induction from data
  - Even so, need some "space" of models to explore
  - Maximum A-posteriori Probability (MAP)

$$P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)P(h_i)$$

  - Maximum Likelihood (ML)

$$P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)$$

    - Assumes uniform priors over all hypotheses in the space
- A-priori knowledge, expressed in
  - Structure of the space of models
  - $P(h_i)$
  - Adjustments to observed data

An Example

(Russell & Norvig)

- Surprise Candy Corp. makes two flavors of candy: *cherry* and *lime*
- Both flavors come in the same opaque wrapper
- Candy is sold in large bags, which have one of the following distributions of flavors, but are visually indistinguishable:
  - h_1: 100% cherry
  - h_2: 75% cherry, 25% lime
  - h_3: 50% cherry, 50% lime
  - h_4: 25% cherry, 75% lime
  - h_5: 100% lime
- Relative prevalence of these types of bags is (.1, .2, .4, .2, .1)
- As we eat our way through a bag of candy, predict the flavor of the next piece; actually a probability distribution.

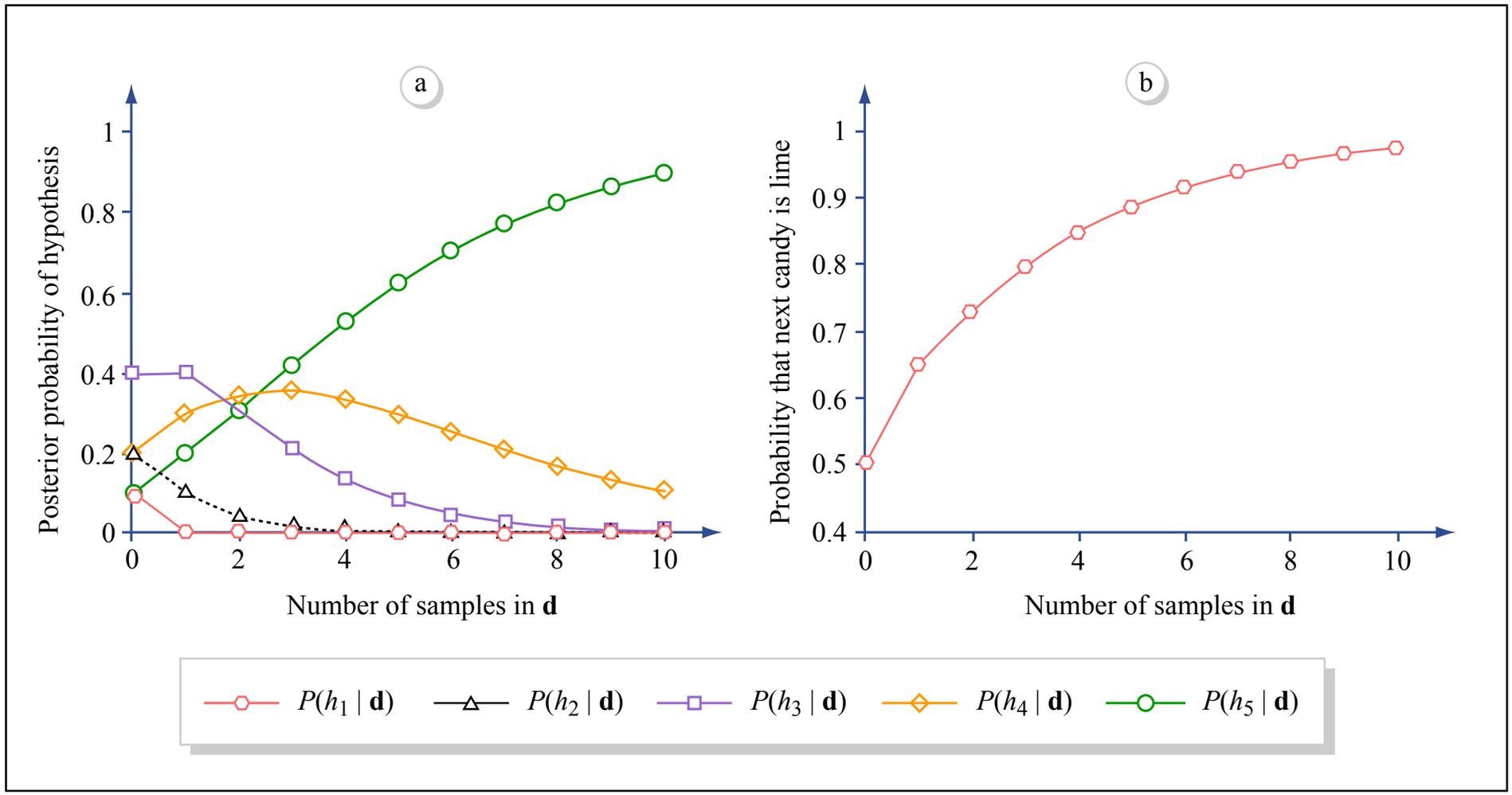
Bayesian Learning

- Calculate the probability of each hypothesis given the data
  $P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)P(h_i)$
- To predict the probability distribution over an unknown quantity, *X*,
  $\boldsymbol{P}(X|\boldsymbol{d}) = \sum_i \boldsymbol{P}(X|\boldsymbol{d}, h_i)\boldsymbol{P}(h_i|\boldsymbol{d}) = \sum_i \boldsymbol{P}(X|h_i)P(h_i|\boldsymbol{d})$
- If the observations ***d*** are independent, then
  $P(\boldsymbol{d}|h_i) = \prod_j P(d_j|h_i)$
- E.g., suppose the first 10 candies we taste are all lime
  $P(\boldsymbol{d}|h_3) = 0.5^{10} \approx 0.001$

h_1: 100% cherry
h_2: 75% cherry, 25% lime
h_3: 50% cherry, 50% lime
h_4: 25% cherry, 75% lime
h_5: 100% lime

Learning Hypotheses and Predicting from Them

- (a) probabilities of h_i after k *lime* candies; (b) prob. of next *lime*



- MAP prediction: predict just from most probable hypothesis
  - After 3 limes, h_5 is most probable, hence we predict *lime*
  - Even though, by (b), it's only 80% probable

Observations

- Bayesian approach asks for prior probabilities on *hypotheses*!
  - Natural way to encode bias against complex hypotheses: make their prior probability very low
- Choosing h_{MAP} to maximize $P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)P(h_i)$
  - is equivalent to minimizing $-\log P(\boldsymbol{d}|h_i) - \log P(h_i)$
    - but as we know that entropy is a measure of information, these two terms are
      - # of bits needed to describe the data given hypothesis
      - # bits needed to specify the hypothesis
    - Thus, MAP learning chooses the hypothesis that maximizes *compression* of the data; *Minimum Description Length* principle
    - *Regularization* is similar to 2nd term—penalty for complexity
- Assuming uniform priors on hypotheses makes MAP yield h_{ML}, the *maximum likelihood hypothesis*, which maximizes $P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)$

Learning More Complex Hypotheses

- Input:
  - Set of *cases*, each of which includes
    - numerous *features*: categorical labels, ordinals, continuous
    - these correspond to the *independent variables*
- Output:
  - For each case, a result, prediction, classification, etc., corresponding to the *dependent variable*
    - In *regression problems*, a continuous output
      - a designated feature the model tries to predict
    - In *classification problems*, a discrete output
      - the category to which the case is assigned
- Task: learn function *f*(input)=output
  - that minimizes some measure of error

Linear Regression

- General form of the function

  $y = f(x_1, x_2, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

- For each case:

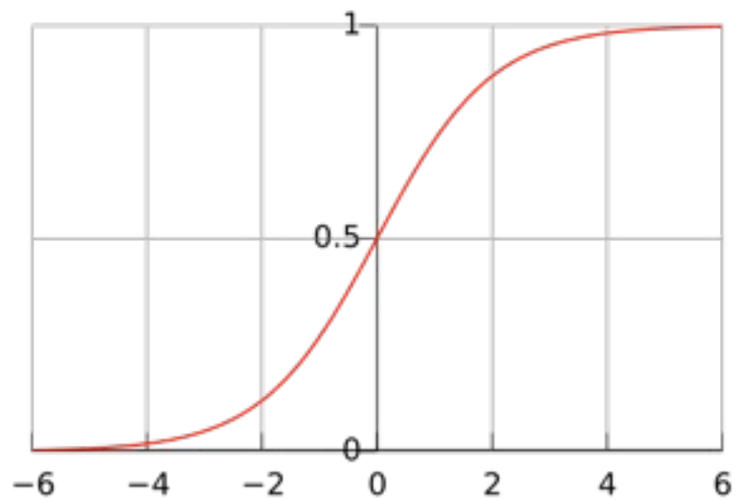
  $\hat{y}_i = f(x_{1,i}, x_{2,i}, \ldots, x_{n,i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i}$

- Find β_j to minimize some function of $(y_i - \hat{y}_i)$ over all y_i
  - e.g., mean squared error: $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

Logistic Regression

- Logistic function: $f(z) = \dfrac{1}{1+e^{-z}}$

$$
\begin{aligned}
y_i &= f(z_i) \\
z_i &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i}
\end{aligned}
$$



- E.g, how risk factors contribute to probability of death
- β_i are the *log odds ratios* $\log O(y_i|x_i)$

More sophisticated models

- Nearest Neighbor Methods
- Classification Trees
- Artificial Neural Nets
- Support Vector Machines
- Bayes Networks (much on this, later)
- Rough Sets, Fuzzy Sets, etc. (see 6.873/HST951 or other ML classes)

How?

- Given: pile of *training data*, all cases labeled with gold standard outcome
- Learn "best" model
- Gather new *test data*, also all labeled with outcomes
- Test performance of model on new test data

- Simple, no?

Simplest Example

- Relationship between a diagnostic conclusion and a diagnostic test

	Test Positive	*Test Negative*	
Disease Present	True Positive	False Negative	TP+FN
Disease Absent	False Positive	True Negative	FP+TN
	TP+FP	FN+TN	

Definitions

	Test Positive	*Test Negative*	
Disease Present	True Positive	False Negative	TP+FN
Disease Absent	False Positive	True Negative	FP+TN
	TP+FP	FN+TN	

Sensitivity (true positive rate): TP/(TP+FN)

False negative rate: 1-Sensitivity = FN/(TP+FN)

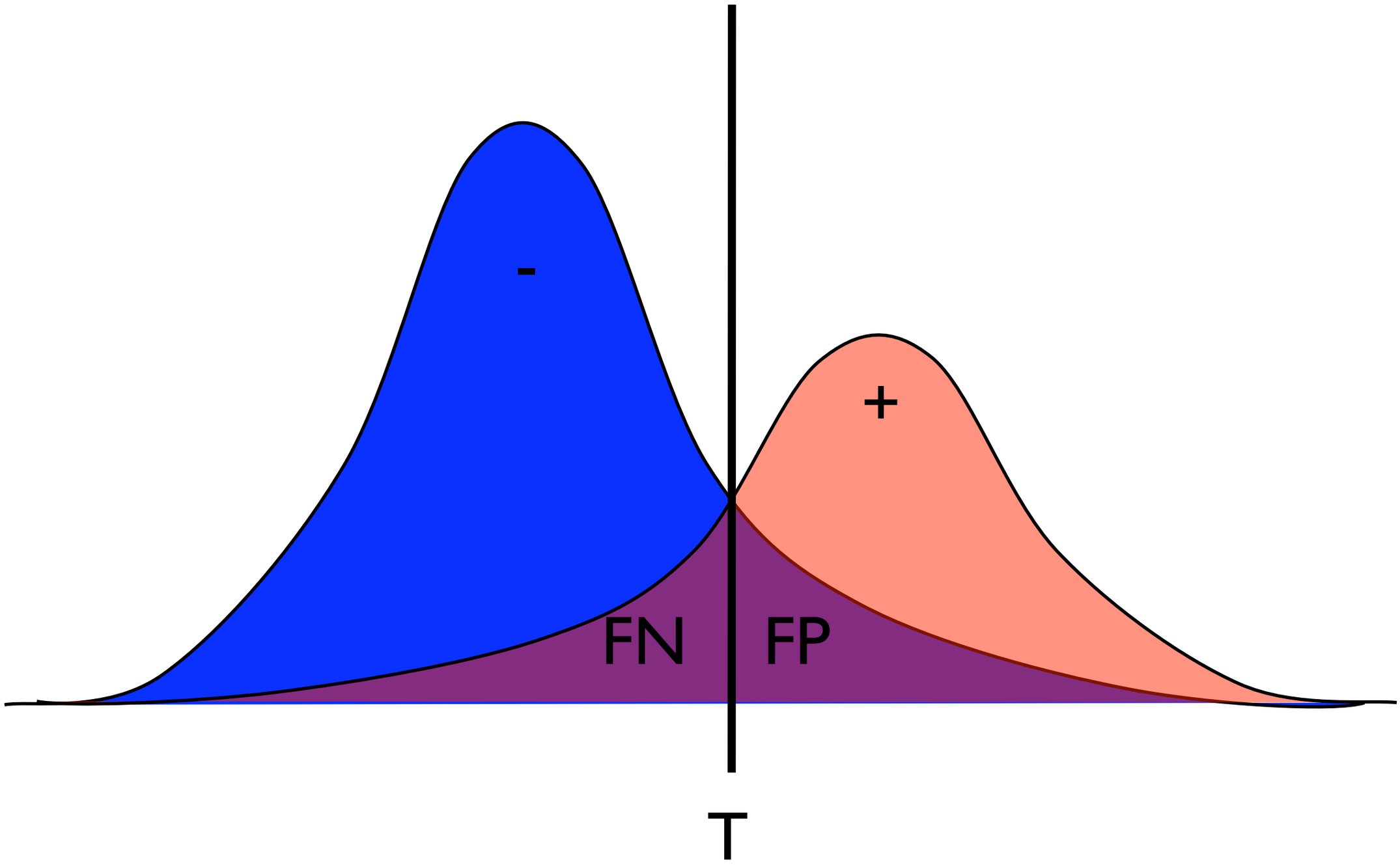
Specificity (true negative rate): TN/(FP+TN)

False positive rate: 1-Specificity = FP/(FP+TN)

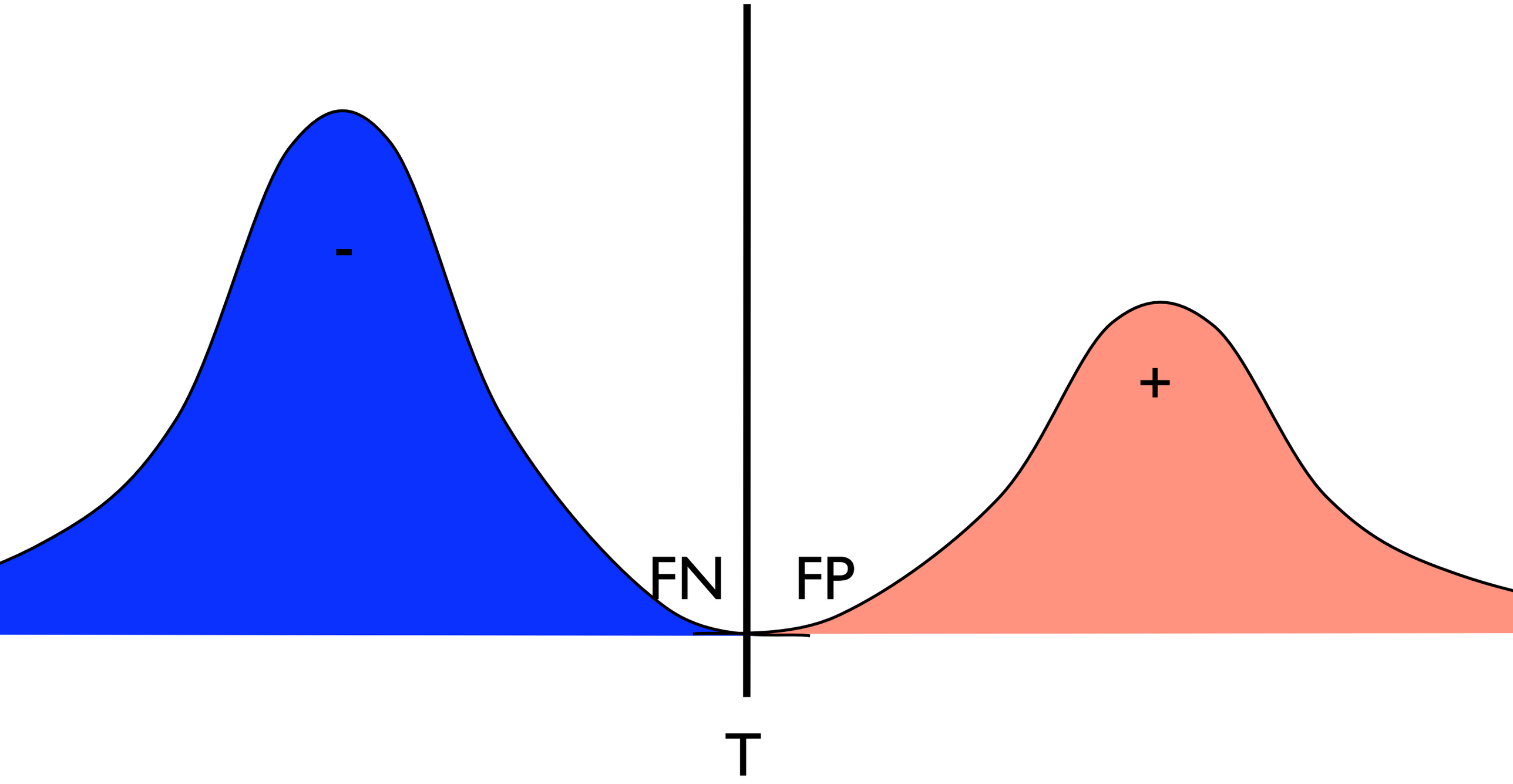
Positive Predictive Value (PPV): TP/(TP+FP)

Negative Predictive Value (NPV): TN/(FN+TN)

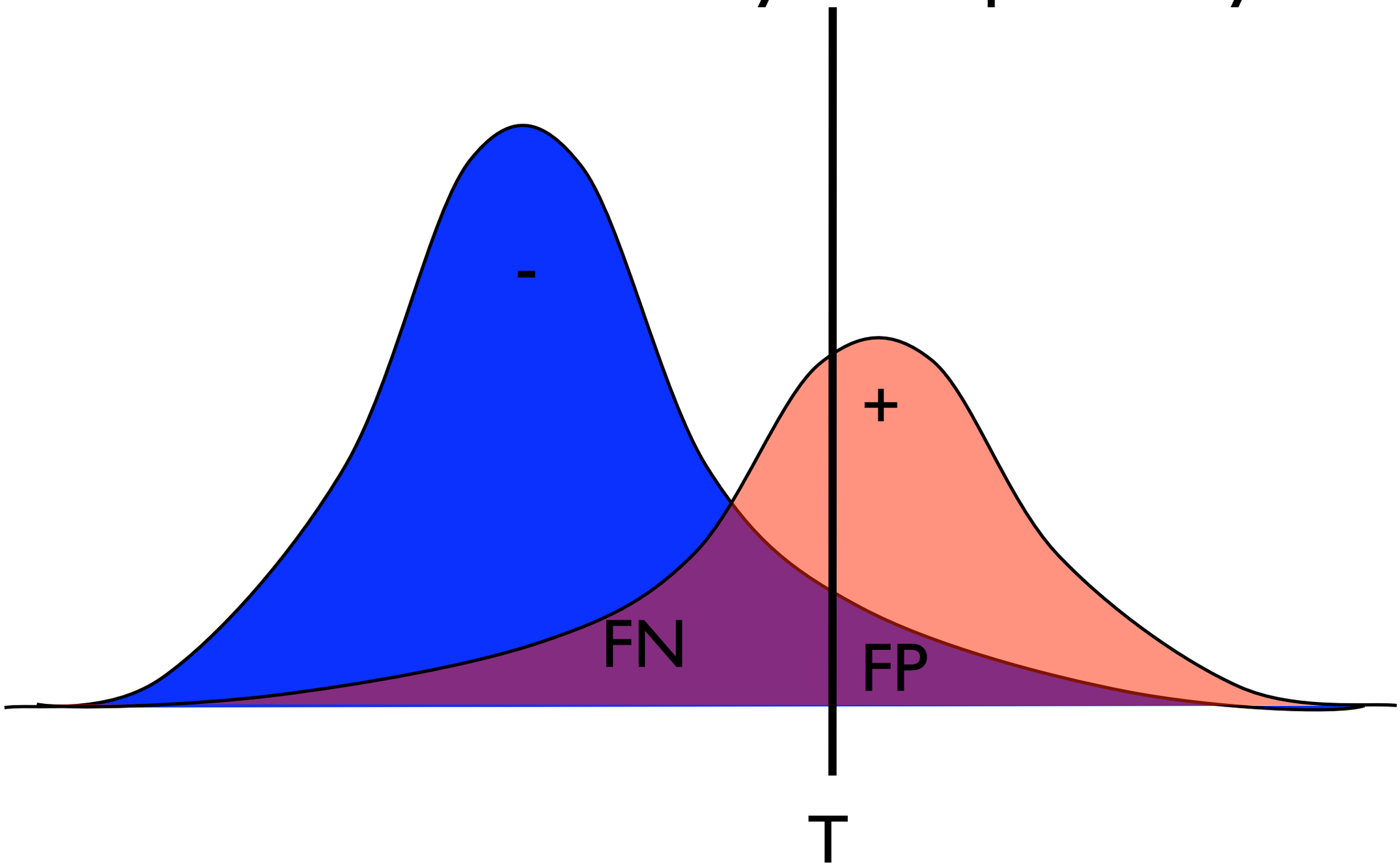
Test Thresholds



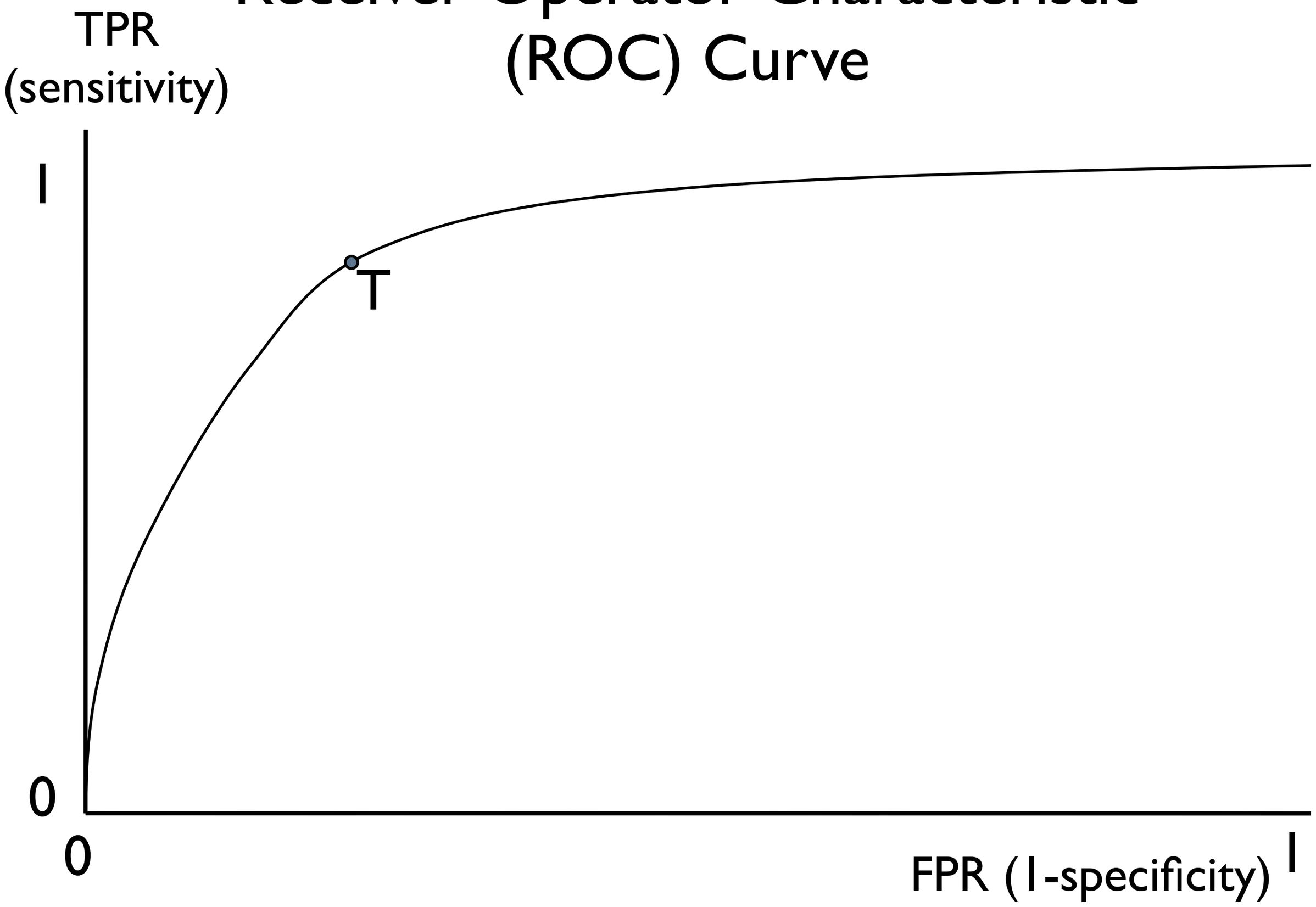
Wonderful Test



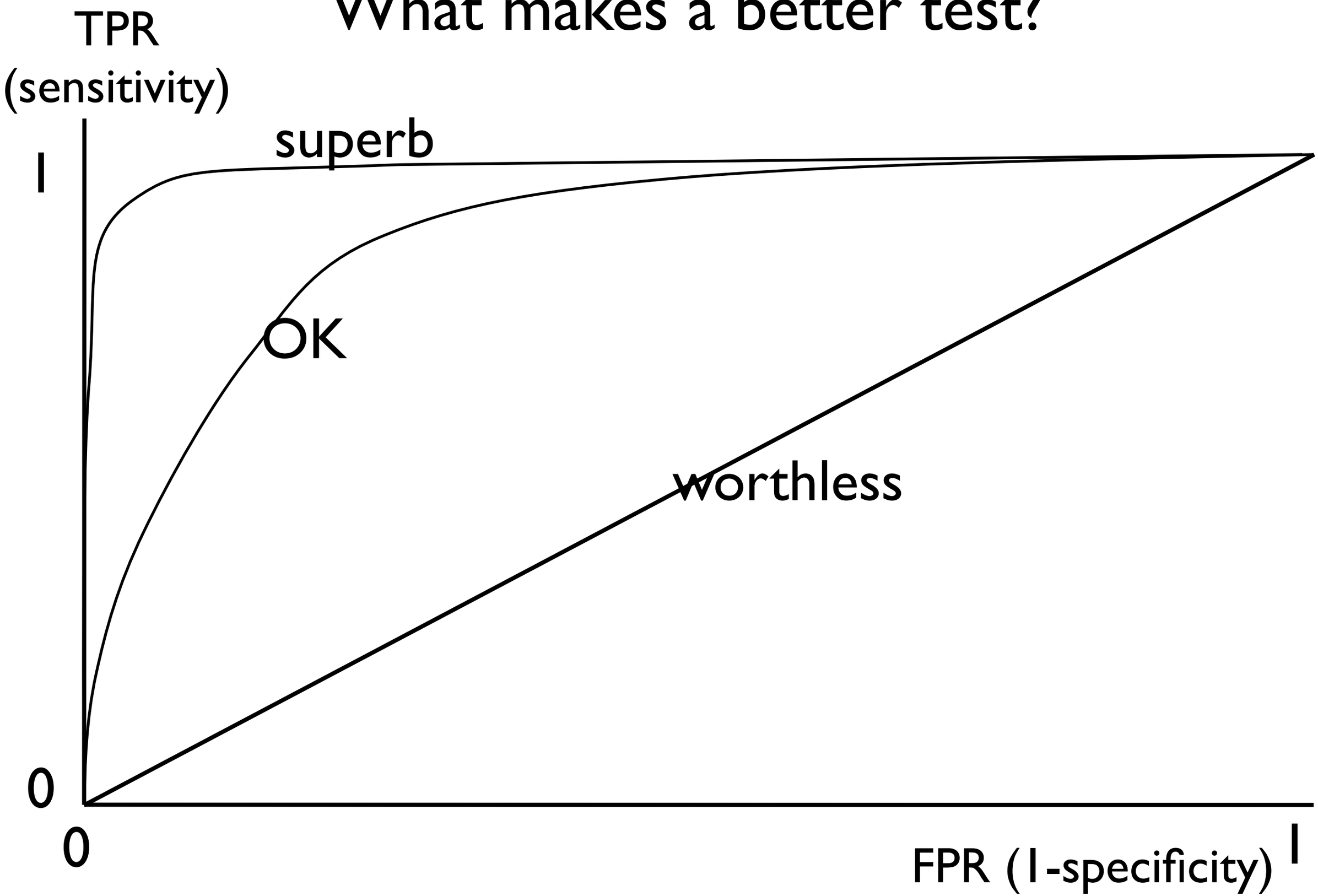
Test Thresholds Change Trade-off between Sensitivity and Specificity



Receiver Operator Characteristic (ROC) Curve



What makes a better test?



Need to explore many models

- Remember:
  - training set => model
  - model + test set => measure of performance
- But
  - How do we choose the best family of models?
  - How do we choose the important features?
  - Models may have structural parameters
    - Number of hidden units in ANN
    - Max number of parents in Bayes Net
- Parameters (like the betas in LR), and *meta-parameters*
- *Not* legitimate to "try all" and report the best !!!!!!!!!!!!!!!!!

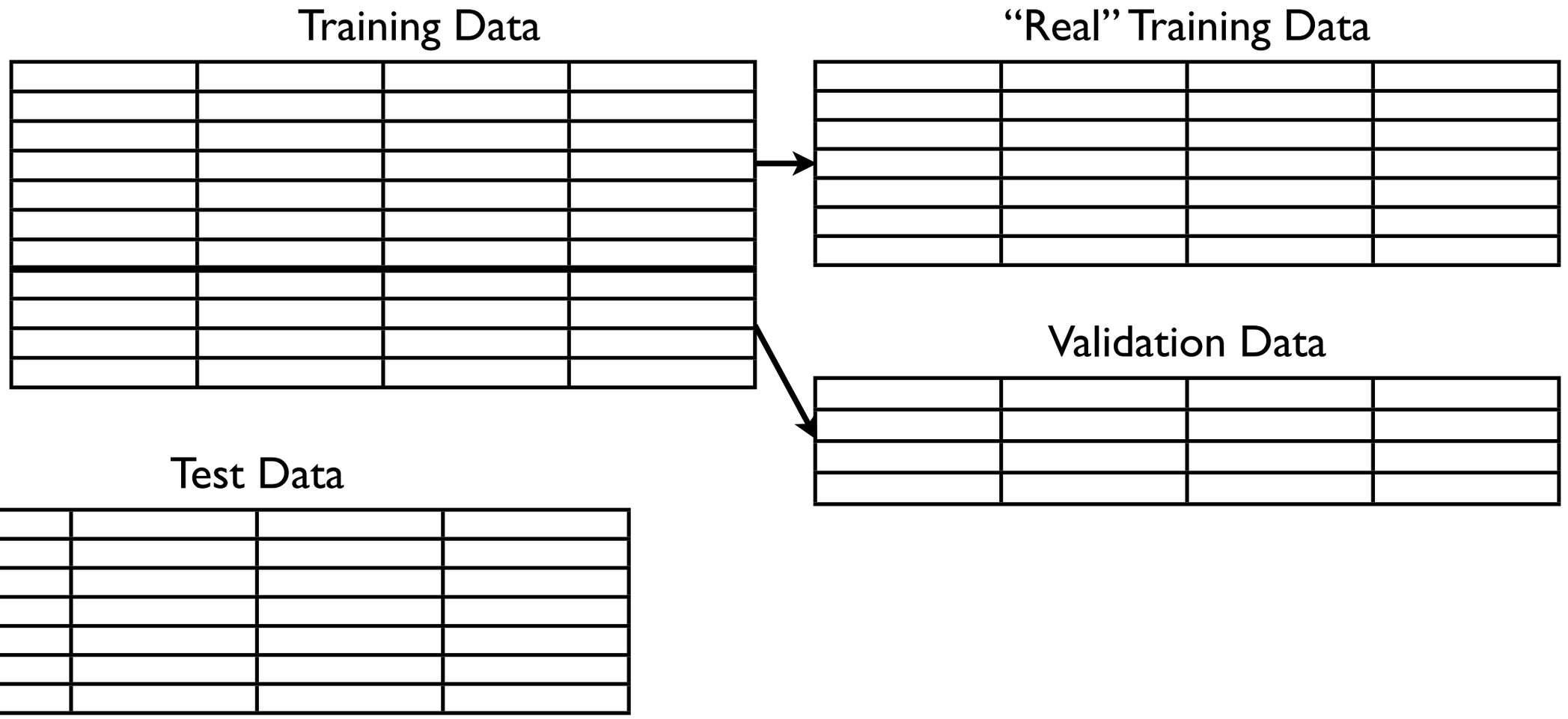
The Lady Tasting Tea

- R. A. Fisher & the Lady
  - B. Muriel Bristol claimed she prefers tea added to milk rather than milk added to tea
  - Fisher was skeptical that she could distinguish
- Possible resolutions
  - Reason about the chemistry of tea and milk
    - Milk first: a little tea interacts with a lot of milk
    - Tea first: vice versa
  - Perform a "clinical trial"
    - Ask her to determine order for a series of test cups
    - Calculate probability that her answers could have occurred by chance guessing; if small, she "wins"
    - ... Fisher's Exact Test
    - Significance testing
      - Reject the *null hypothesis* (that it happened by chance) if its probability is < 0.1, 0.05, 0.01, 0.001, ..., 0.000001, ..., ????

How to deal with multiple testing

- Suppose Ms. Bristol had tried this test 100 times, and passed once. Would you be convinced of her ability to distinguish?
- *Bonferroni correction*: for *n* trials, insist on a p-value that is 1/*n* of what you would demand for a single trial

Cross-validation



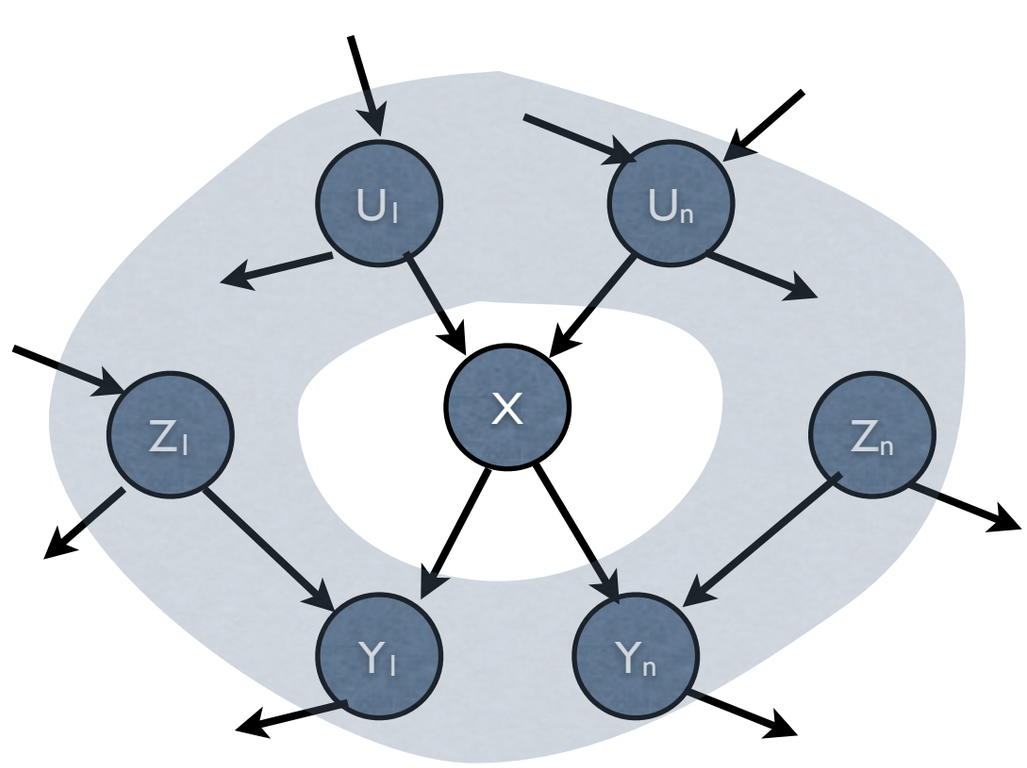
- *Any number of times*
  - Train on some subset of the training data
  - Test on the remainder, called the validation set
- Choose best *meta-parameters*
- Train, with those meta-parameters, on all training data
- Test on Test data, *once!*

Aliferis lessons (part)

- Overfitting
  - bias, variance, noise
  - O = optimal possible model over all possible learners
  - L = best model learnable by this learner
  - A = actual model learned
  - Bias = O - L (limitation of learning method or target model)
  - Variance = L - A (error due to sampling of training cases)
  - Compare against learning from randomly permuted data
- Curse of dimensionality
  - Feature selection
  - Dimensionality reduction

Causality

- Suppes, 1950's
  - Statistical association
  - Temporal succession
  - No confounders (!)
    - hidden variables
- A node, *X*, is *conditionally independent* of all other nodes in the network given its *Markov blanket*: its <u>parents</u>, U_i, <u>children</u>, Y_i, and <u>children's parents</u>, Z_i.



Using MIMIC data to build predictive models

- Mortality
  - Comparison to SAPS II
  - Daily Acuity Scores
  - Real-time Acuity Scores
- Other outcomes
  - Good
    - Weaning from Ventilator
    - Weaning from Intra-Aortic Balloon Pump
    - Weaning from Vasopressors
  - Bad
    - Septic shock
    - Hypotension
    - Acute kidney injury

- Caleb Hug's 2009 PhD thesis: http://dspace.mit.edu/handle/1721.1/46690

Cleaning the data—half the research time

- Missing values
  - Some values are not measured for some clinical situations
  - Failures in data capture process
- Episodically measured variables
- Unclear/undefined clinical states
- Imprecise timing of meds, ...
- Partially measured i/o
- Proxies: e.g., which ICU⇒what disease
- Derived variables: integrals, slopes, ranges, frequencies, etc.
- Transformed variables: square root, log, etc.
- Select subset of data with enough data!

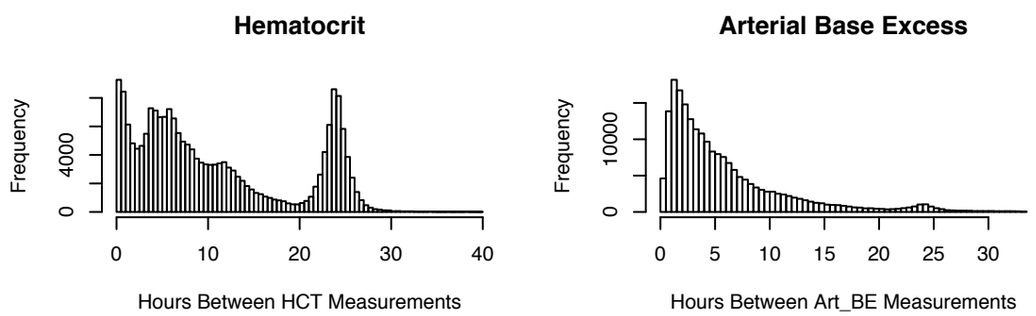
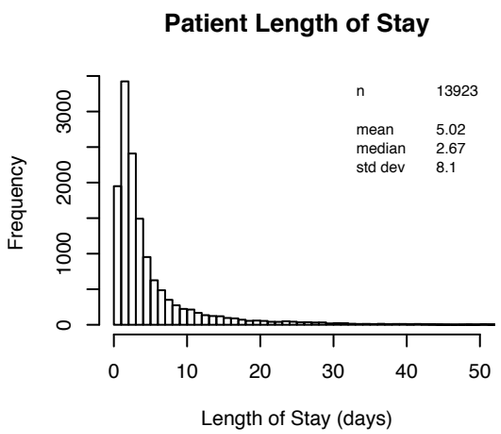
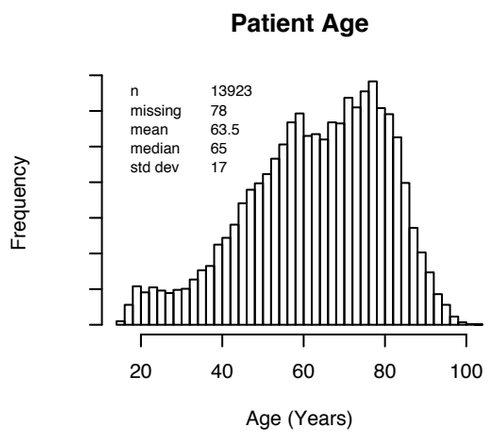
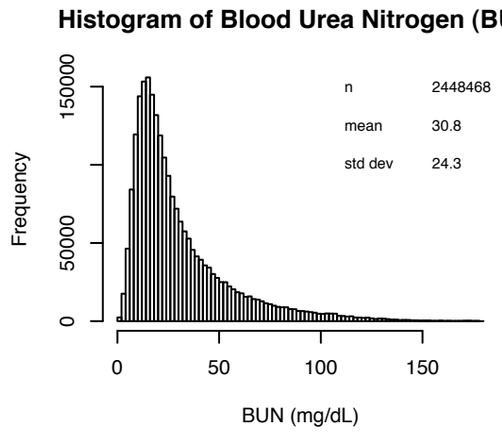


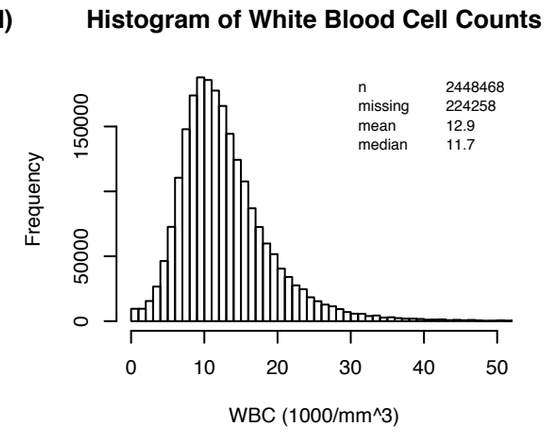
Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

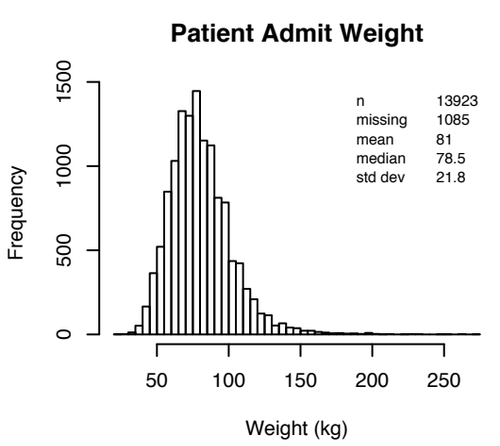
Descriptive look

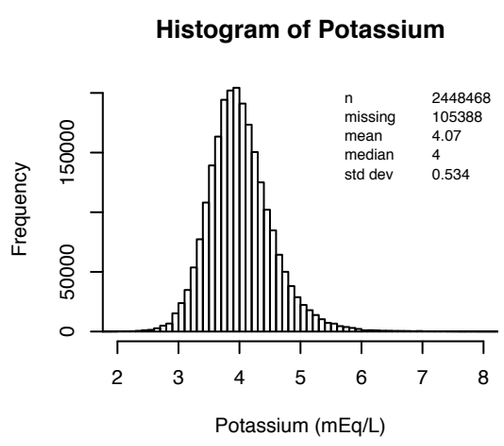


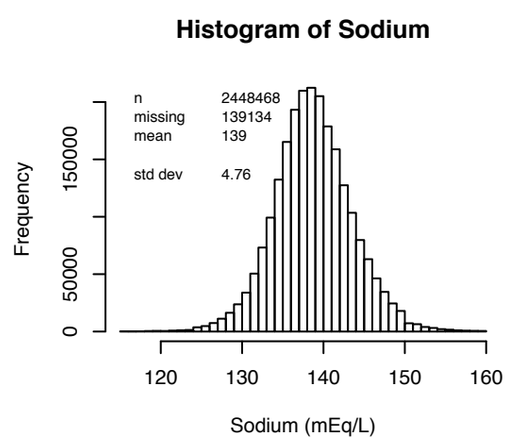


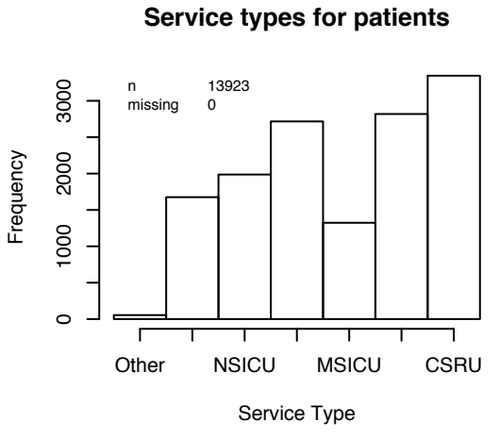


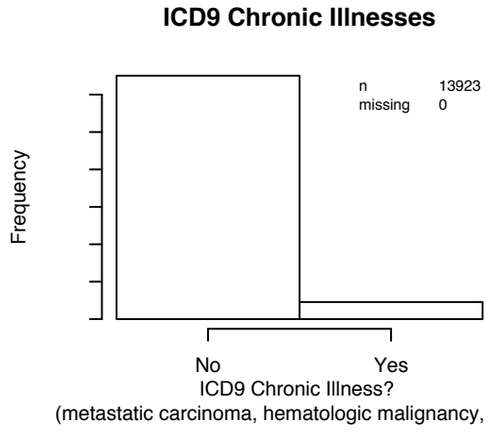


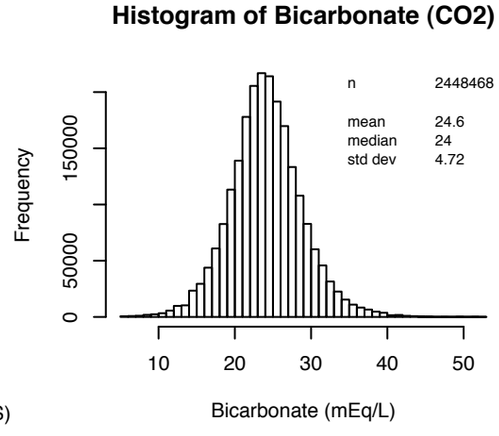












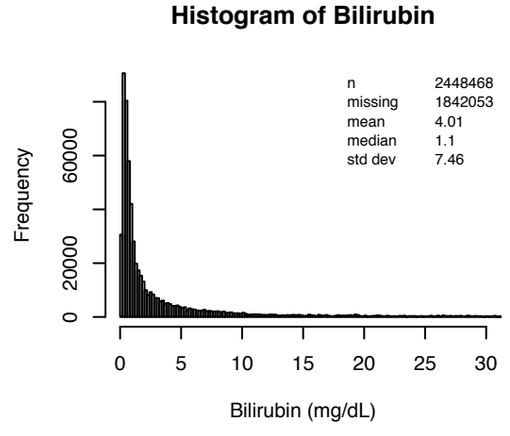
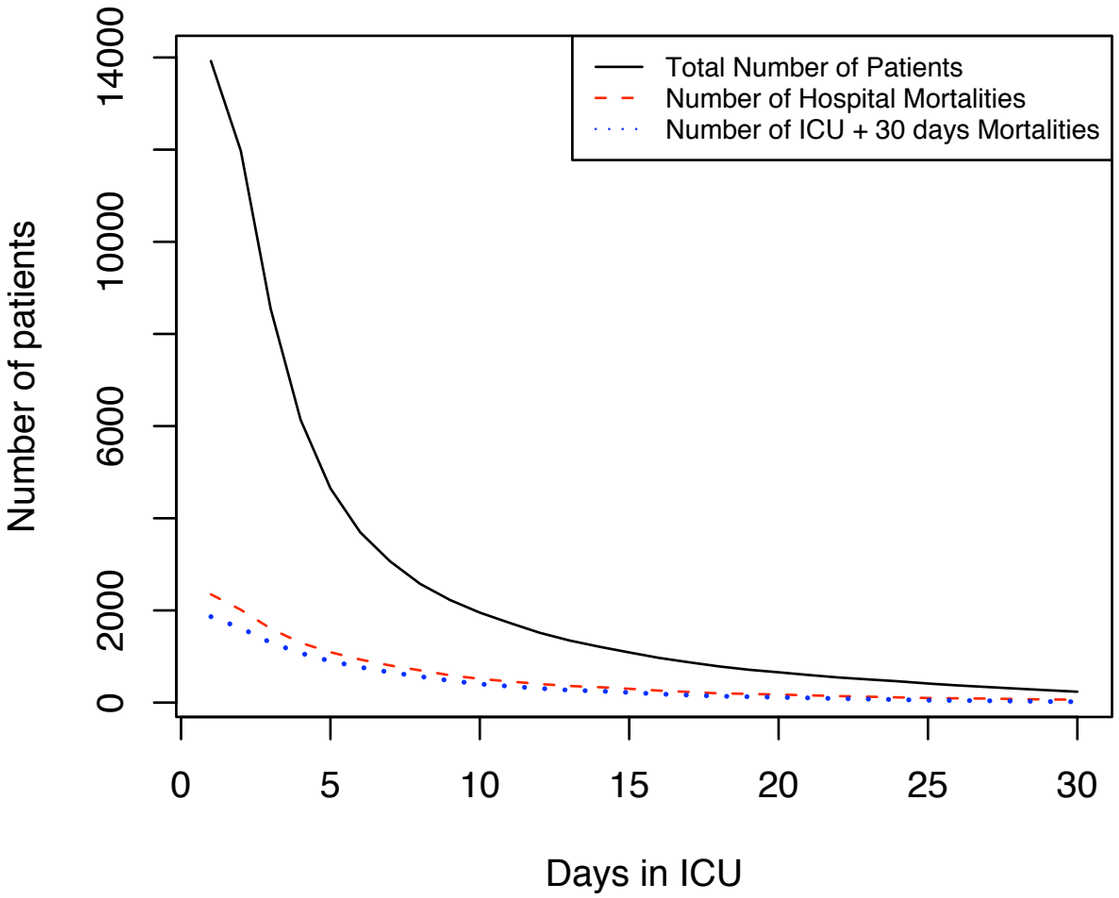


Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

Outcomes



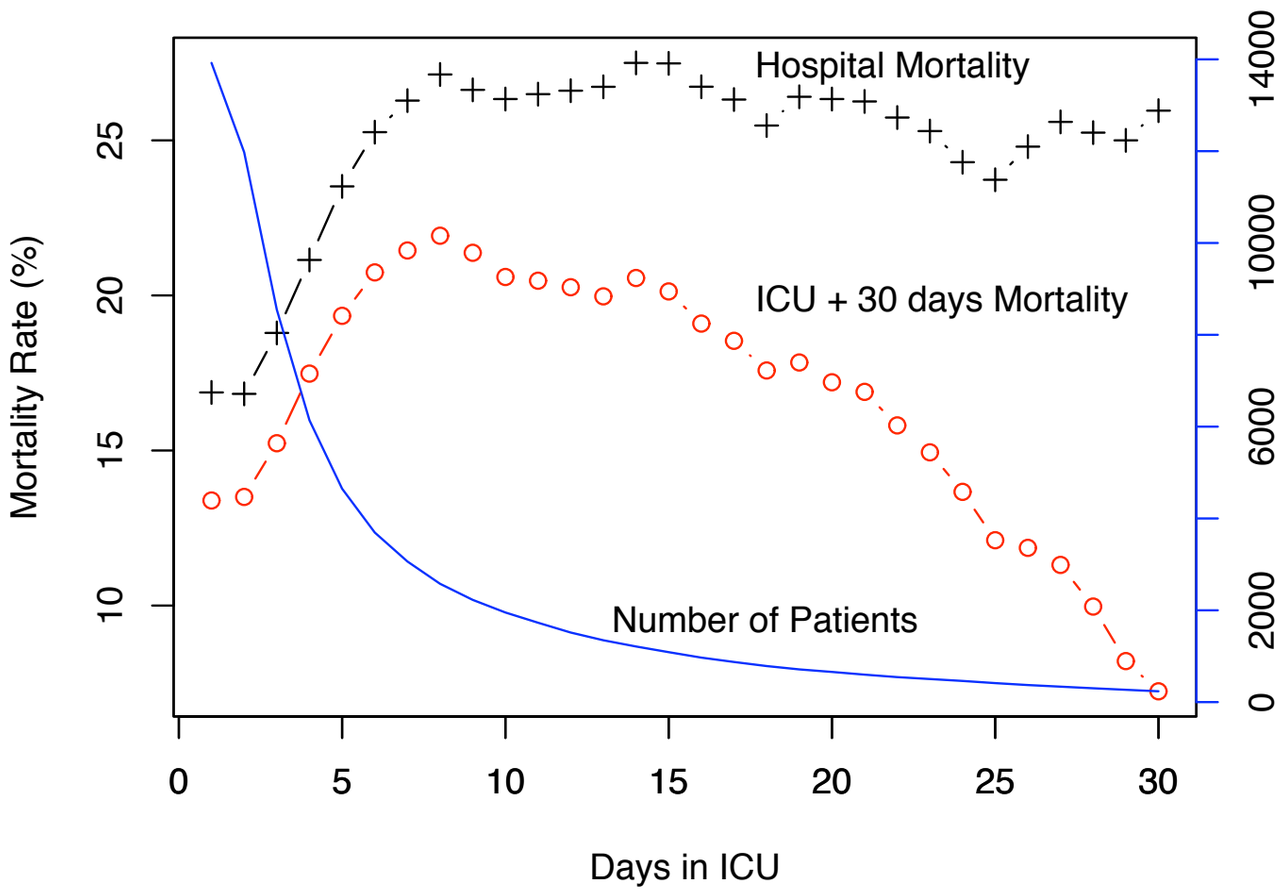


Table 3.15: Preprocessed Data

Number of Patients	10,066
Number of Rows	1,044,982
Number of Features	438

Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

SAPS II

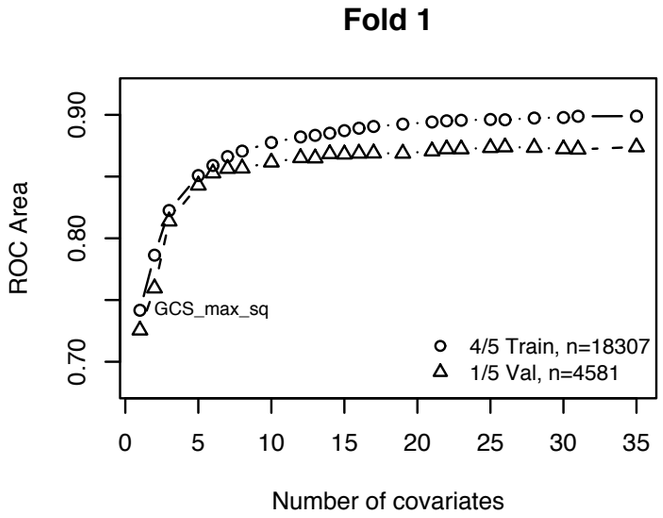
Table 4.1: SAPS II Variables

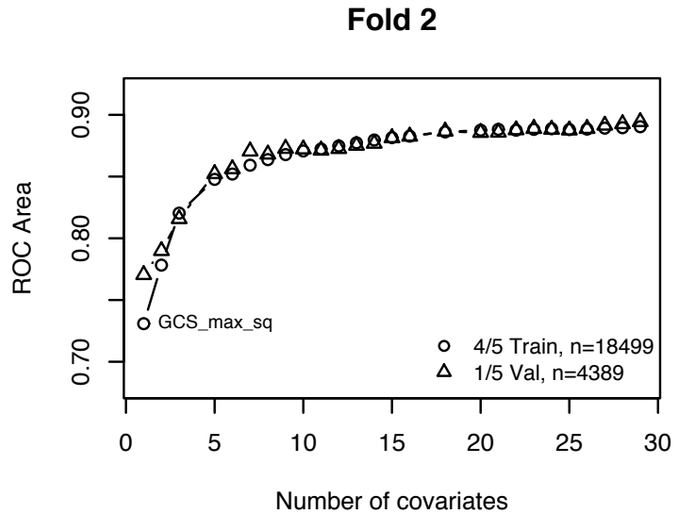
Variable	Max Points
Age	18
Heart rate	11
Systolic BP	13
Body temperature	3
PaO2:FiO2 (if ventilated or continuous positive airway pressure)	11
Urinary output	11
Serum urea nitrogen level	10
WBC count	12
Serum potassium	3
Serum sodium level	5
Serum bicarbonate level	6
Bilirubin level	9
Glasgow Coma Score[a]	26
Chronic diseases	17
Type of admission	8

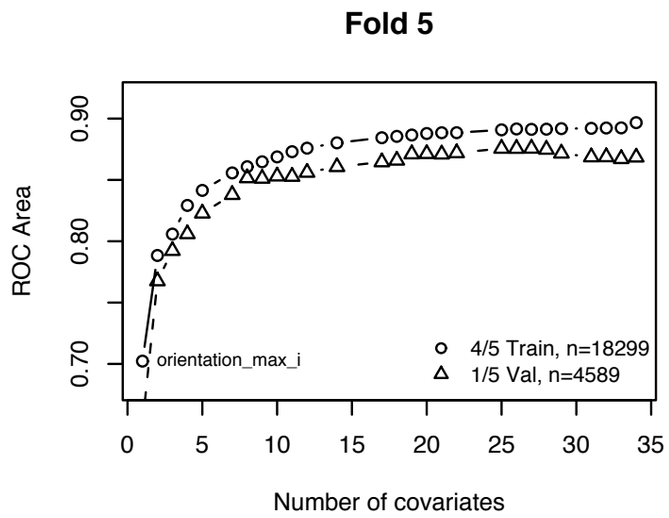
[a]If the patient is sedated, the estimated GCS prior to sedation

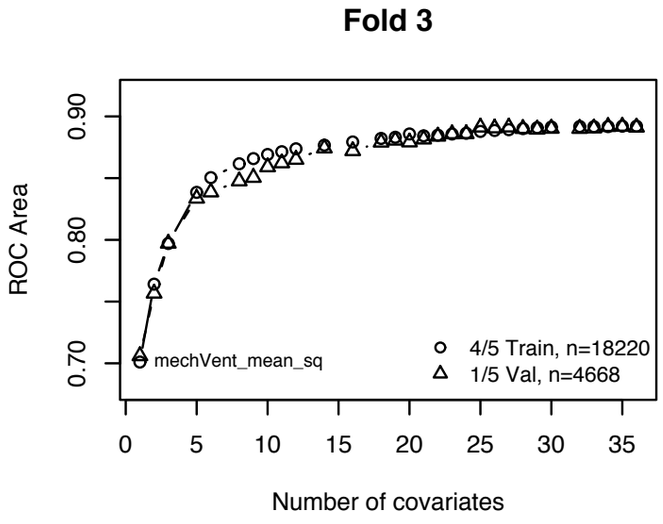
Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

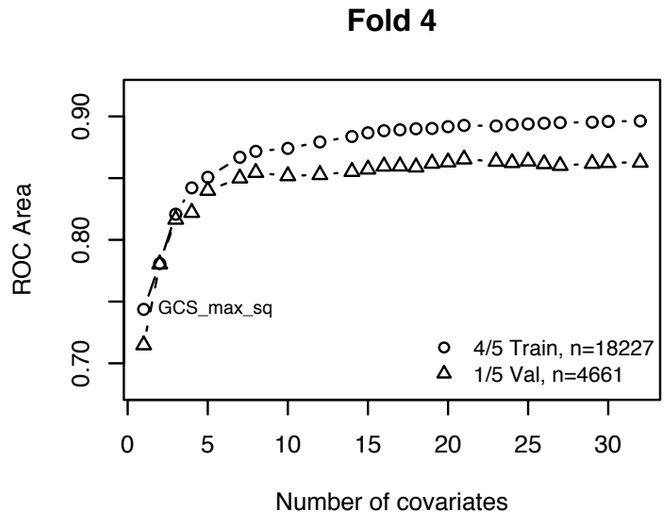
Training models—5-fold cross validation











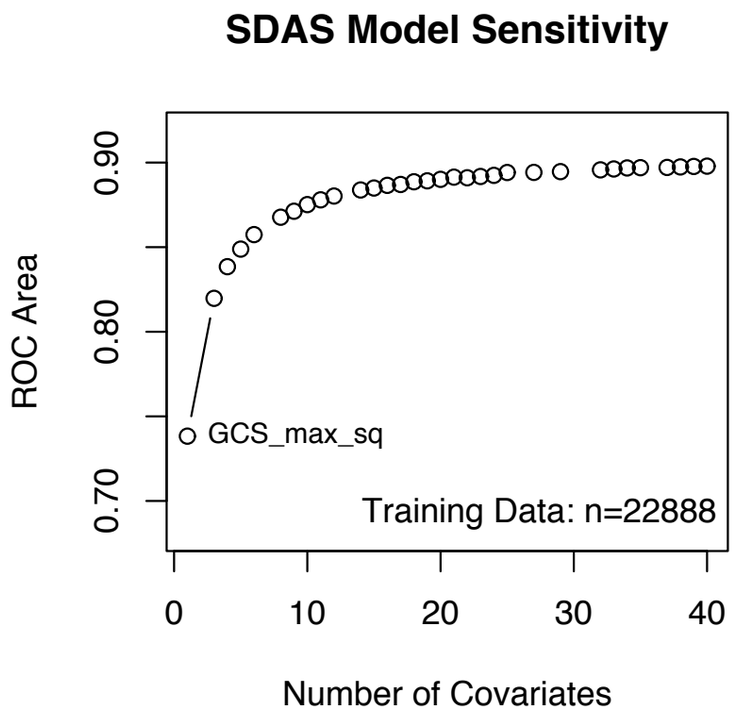


Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

Many univariate analyses

Model 5.1 SDAS Model for Fold 2 with 30 Covariates

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
20172	1e-09	5415.11	30	0	0.893	0.785
Gamma	Tau-a	R2	Brier			
0.787	0.176	0.439	0.076			

	Coef	S.E.	Wald Z	P
INR_mean_i	-1.795e+00	1.423e-01	-12.61	0.0000
GCS_max_sq	-7.485e-03	6.000e-04	-12.47	0.0000
OutputB_60_mean_sqrt	-6.561e-02	6.885e-03	-9.53	0.0000
pacemkr_max	-1.084e+00	1.183e-01	-9.16	0.0000
svCSRU_max	-9.516e-01	1.208e-01	-7.88	0.0000
GCSrdv_mean	-1.138e-01	1.528e-02	-7.45	0.0000
pressD01_mean_am	-2.774e+00	3.893e-01	-7.13	0.0000
Platelets_Slope_1680_min	-5.493e+00	8.615e-01	-6.38	0.0000
pressD01_sd_sq	-5.085e+00	8.678e-01	-5.86	0.0000
sedatives_mean_sq	-4.375e-01	8.455e-02	-5.17	0.0000
Bal24_max	-4.493e-05	1.222e-05	-3.68	0.0002
CV_HRrng_max	-3.267e-03	1.083e-03	-3.02	0.0026
Intercept	4.292e-01	4.085e-01	1.05	0.2934
Milrinone_perKg_min_sq	3.523e+00	1.113e+00	3.17	0.0015
LOSBal_max	2.247e-05	5.703e-06	3.94	0.0001
hrmVA_max	3.410e-01	6.767e-02	5.04	0.0000
MBPm.pr_min_am	1.904e+00	3.711e-01	5.13	0.0000
Mg_min_sq	1.067e-01	1.798e-02	5.93	0.0000
beta.Blocking_agent_mean_lam	2.418e-01	3.955e-02	6.11	0.0000
Na_mean_am	5.214e-02	8.415e-03	6.20	0.0000
mechVent_mean_sq	7.183e-01	1.047e-01	6.86	0.0000
RESP_mean_sq	9.226e-04	1.293e-04	7.13	0.0000
Platelets_mean_i	2.512e+01	3.512e+00	7.15	0.0000
Lasix_max_lam	2.550e-01	3.457e-02	7.38	0.0000
CO2_mean_i	2.038e+01	2.741e+00	7.43	0.0000
jaundiceSkin_mean_la	1.523e-01	2.014e-02	7.56	0.0000
hospTime_min_sqrt	6.860e-03	7.939e-04	8.64	0.0000
pressorSum.std_mean_sqrt	7.758e-01	7.225e-02	10.74	0.0000
SpO2.oor30.t_mean_sqrt	4.929e-01	4.095e-02	12.04	0.0000
BUNtoCr_min_sqrt	2.867e-01	2.323e-02	12.34	0.0000
Age_min_sq	2.258e-04	1.450e-05	15.57	0.0000

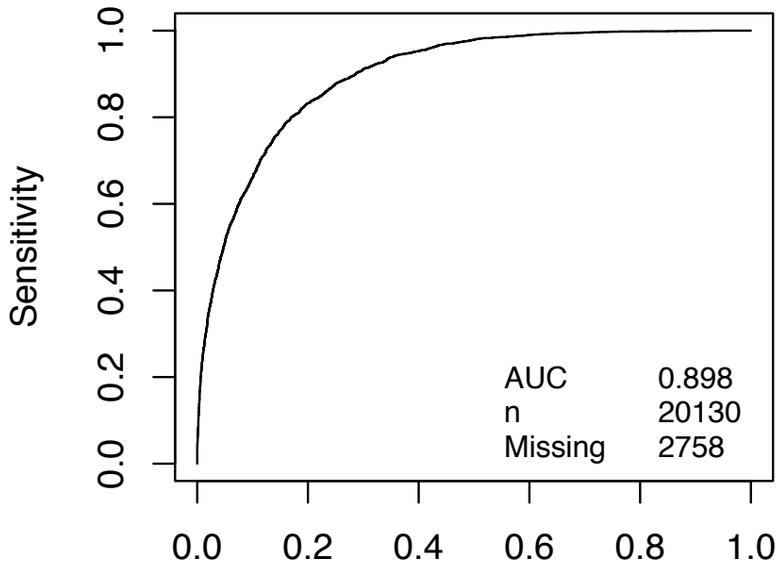
Model 5.2 Final SDAS model

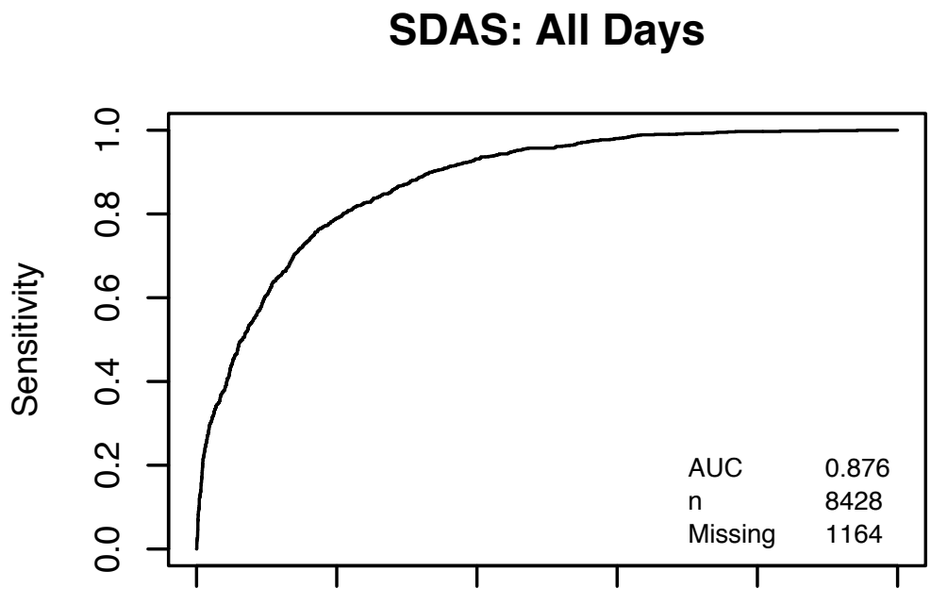
Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
20130	3e-10	5619.28	35	0	0.898	0.797
Gamma	Tau-a	R2	Brier			
0.798	0.177	0.456	0.074			

	Coef	S.E.	Wald Z	P
GCS_max_sq	-0.0064668	5.032e-04	-12.85	0.0000
INR_mean_i	-1.8734049	1.458e-01	-12.85	0.0000
pacemkr_max	-0.9337190	1.179e-01	-7.92	0.0000
svCSRU_max	-0.9137522	1.250e-01	-7.31	0.0000
RikerSAS_mean	-0.3430971	5.151e-02	-6.66	0.0000
Platelets_Slope_1680_min	-5.8856843	8.839e-01	-6.66	0.0000
urineByHr_mean_sqrt	-0.0584113	9.453e-03	-6.18	0.0000
GCSrdv_mean	-0.0902717	1.552e-02	-5.82	0.0000
GCSrng_min_am	-0.0812232	1.459e-02	-5.57	0.0000
pressD01_mean_am	-1.6132643	3.005e-01	-5.37	0.0000
CV_HRrng_max	-0.0061979	1.216e-03	-5.10	0.0000
Insulin_sd_sq	-2.1686950	4.372e-01	-4.96	0.0000
alloutput_max_la	-0.0890330	2.265e-02	-3.93	0.0001
MetCarcinoma_min	0.4468763	1.567e-01	2.85	0.0043
WBC_mean_am	0.0147036	5.149e-03	2.86	0.0043
AIDS_min	0.5954305	1.991e-01	2.99	0.0028
Intercept	1.5314512	4.529e-01	3.38	0.0007
MBPm.pr_min_am	1.4601630	3.518e-01	4.15	0.0000
HemMalig_min	0.6032027	1.212e-01	4.98	0.0000
RESP_mean_sq	0.0006615	1.324e-04	5.00	0.0000
hrmVA_max	0.3520834	6.823e-02	5.16	0.0000
PaO2toFiO2_mean	0.2672376	4.336e-02	6.16	0.0000
Na_mean_am	0.0549066	8.506e-03	6.45	0.0000
Mg_min_sq	0.1173220	1.815e-02	6.46	0.0000
ShockIdx_max	0.5742182	8.853e-02	6.49	0.0000
Platelets_mean_i	24.0719462	3.560e+00	6.76	0.0000
hospTime_min_sqrt	0.0057514	8.158e-04	7.05	0.0000
day_min_sq	0.0170075	2.372e-03	7.17	0.0000
jaundiceSkin_mean_la	0.1469141	2.045e-02	7.18	0.0000
CO2_mean_i	19.3845272	2.682e+00	7.23	0.0000
Lasix_max_lam	0.2523702	3.444e-02	7.33	0.0000
beta.Blocking_agent_mean_lam	0.2918077	3.923e-02	7.44	0.0000
Sympathomimetic_agent_min	0.8576883	9.254e-02	9.27	0.0000
SpO2.oor30.t_mean_sqrt	0.4059329	4.128e-02	9.83	0.0000
BUNtoCr_min_sqrt	0.2829088	2.348e-02	12.05	0.0000
Age_min_sq	0.0002601	1.495e-05	17.40	0.0000

Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

Evaluating the models





Devel

Validation

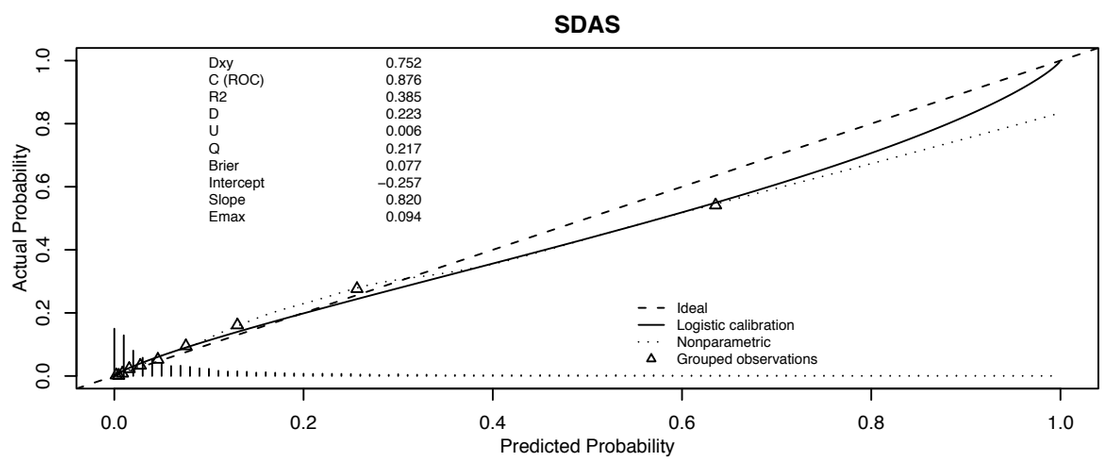


Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

Selected features for each day of ICU stay

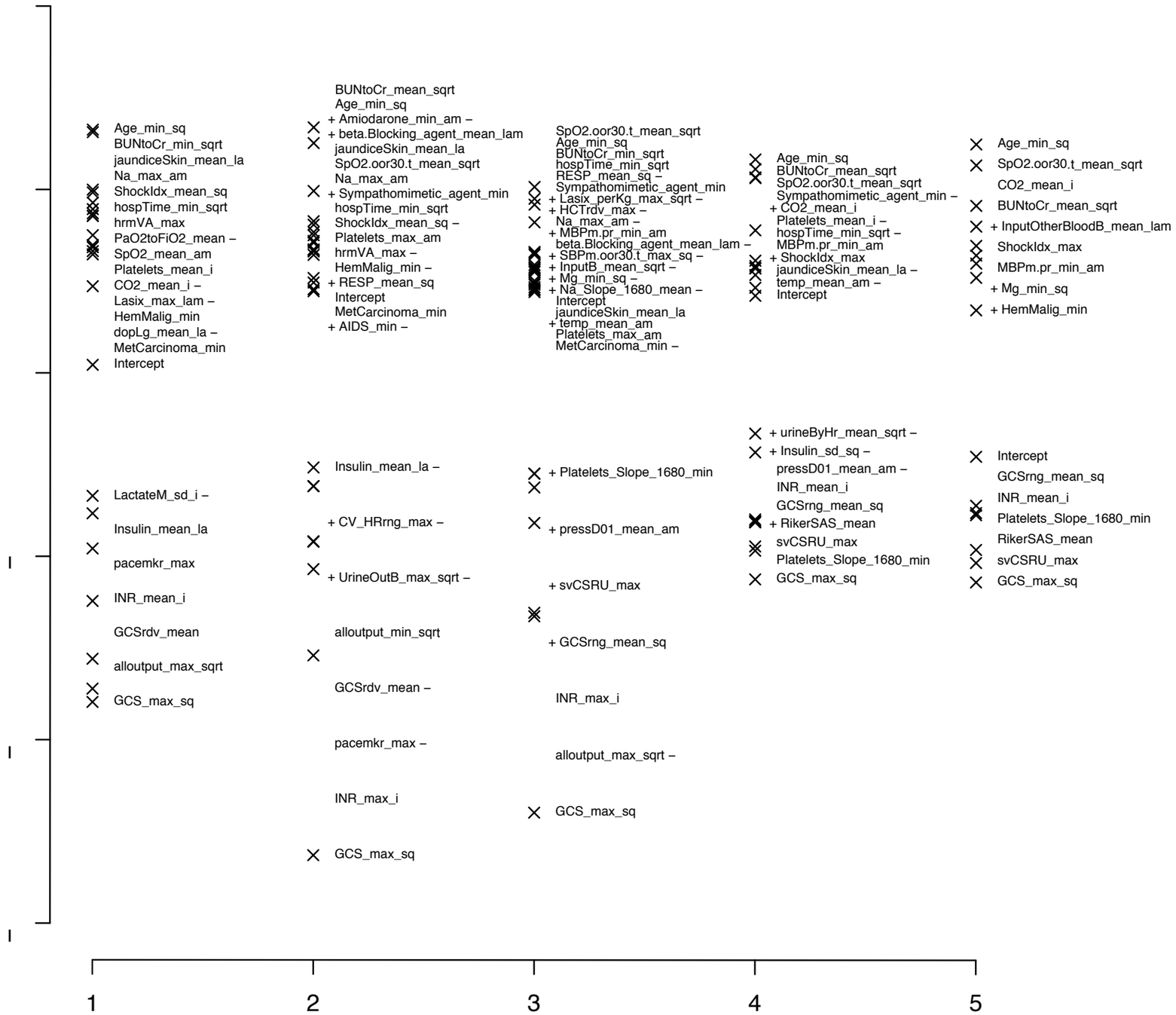


Figure by Hug, Caleb Wayne. "Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models." *Massachusetts Institute of Technology*, 2009.

HST.950J / 6.872 Biomedical Computing
Fall 2010