Introduction to Machine Learning

Katherine Heller

Deep Learning Summer School 2018

Outline

- Kinds of machine learning
- Linear regression
- Regularization
- Bayesian methods
- Logistic Regression
- Why we do this

Kinds of Machine Learning

- Supervised Learning
- Reinforcement Learning
- Unsupervised Learning

Supervised Learning

- Takes some set of inputs $x_1, x_2, \ldots, x_d$ and maps them to an output, or label y
- Our goal in supervised learning is to find a good function f which maps $(x_1, x_2, \ldots, x_d) \rightarrow y$
- What is this function? How do we find it?
- Whatever this function is, we'd like it to map our inputs to the correct label all of the time.
- Need some kind of metric to tell us how close we are to this ideal function
  - Loss – number of times we get it wrong, how close we are to y
- Minimize the loss using the data set we have

Examples of Supervised Learning

- Object Detection

- Predicting Surgical Complications

2-week old early wound infection

What kind of surgery?
Patient age
Other illnesses of patient
DNR status
What surgeon?
Nutrition status

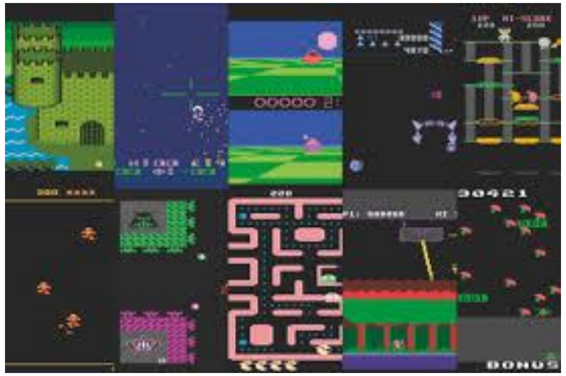
Bleeding complication
Pulmonary complication
Neurological complication
Infection

Reinforcement Learning

- Learning what actions to take based on a reward signal
- Generally involves an assumption of being in a state s_1 and evaluating the chances of changing to another state s_2 if a particular action a is taken.
- In the end (or along the way) a reward signal is received.
- We want to take actions that lead to getting the most reward possible.
- Therefore learn a set of actions, or a policy, which leads to maximizing the expected reward.

Examples of Reinforcement Learning

- Games like Go or Atari



- Medical interventions

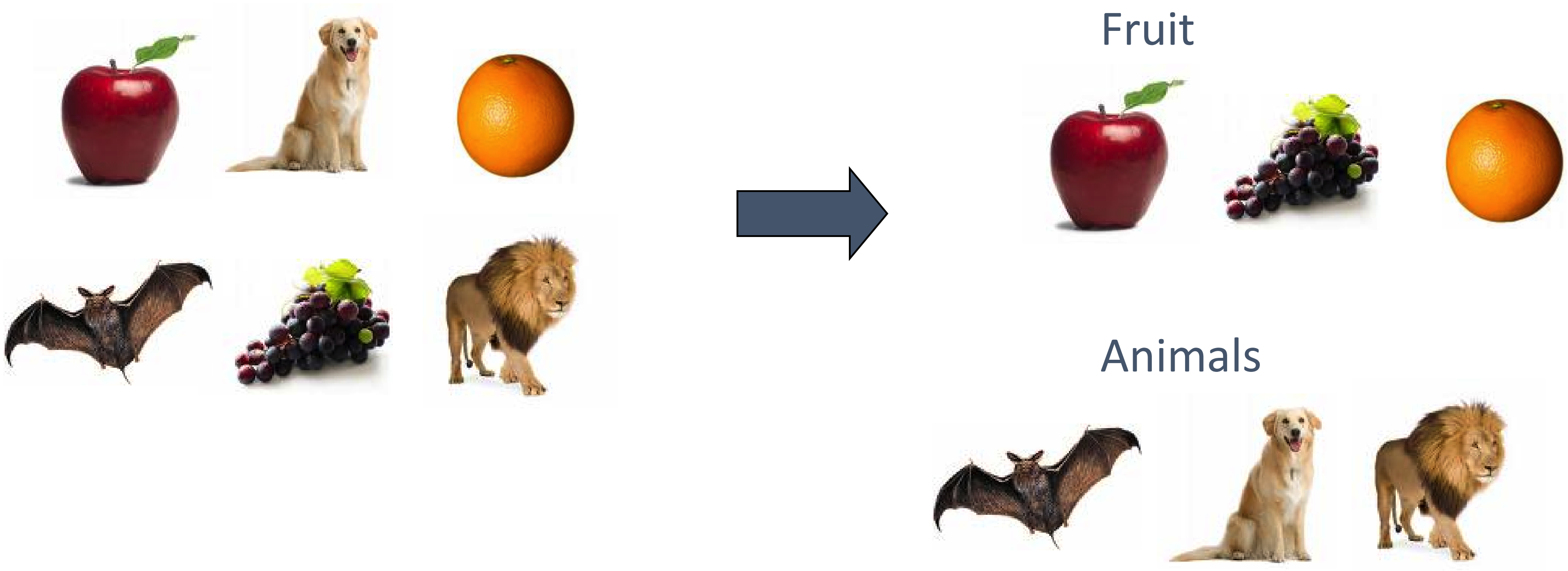
State: Health of the patient (ie vital signs)
Action: Medical intervention (ie administration of IV fluids)
Reward: Did the patient live? +10

Unsupervised Learning

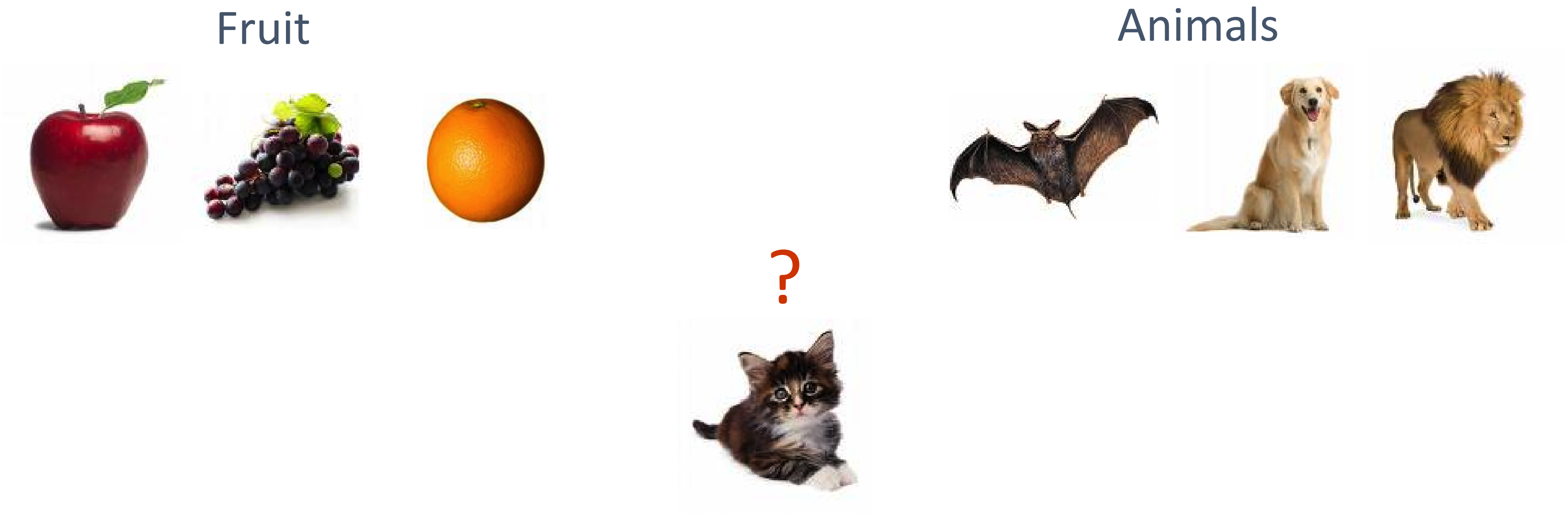
- No labels
- Find interesting and informative patterns in the data
- Examples: clustering or dimensionality reduction

Clustering

- Clustering is the act of grouping objects (or data points) together based on common features.
  - Clustering is natural - people are constantly grouping things together



Clustering



Fruit

Animals

?

- There are no inherently “correct” clusterings, only clusterings that are right for a particular problem.

Hangs on Trees

Doesn’t

Examples of Unsupervised Learning

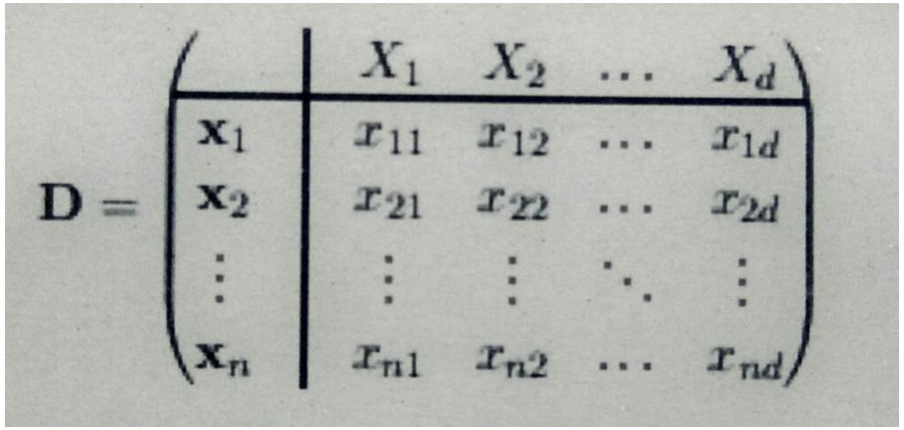
- Learning hidden factors that make patient populations different in different hospitals
- Discovering subpopulations of patients who have a neurological disease from mobile app data.

Why are you tired?

Disease
Medication
Depression
Lack of Sleep

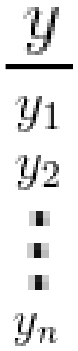
What Does a Data Look Like?

- Learning is typically done on data given in the form of a matrix:



$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- An additional column in this data matrix may be the label:
  - If these labels are discrete – classification
  - If these labels are continuous – regression
  - More complex – structured prediction



$$\begin{array}{c} y \\ \hline y_1 \\ y_2 \\ \vdots \\ y_n \end{array}$$

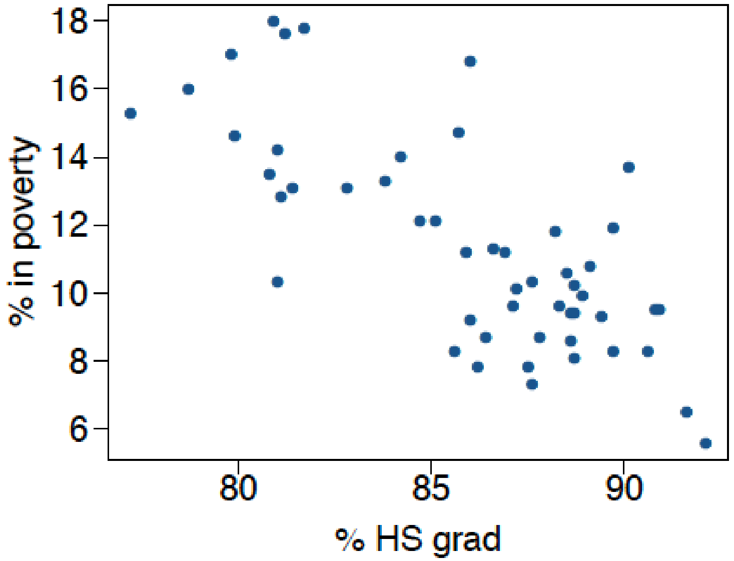
Back to Supervised Learning

- Takes some set of inputs $x_1, x_2, \ldots, x_d$ and maps them to an output, or label y
- Our goal in supervised learning is to find a good function f which maps $(x_1, x_2, \ldots, x_d) \to y$
- What is this function? How do we find it?
- Whatever this function is, we'd like it to map our inputs to the correct label all of the time.
- Need some kind of metric to tell us how close we are to this ideal function
  - Loss – number of times we get it wrong, how close we are to y
- Minimize the loss using the data set we have

What function do we use?

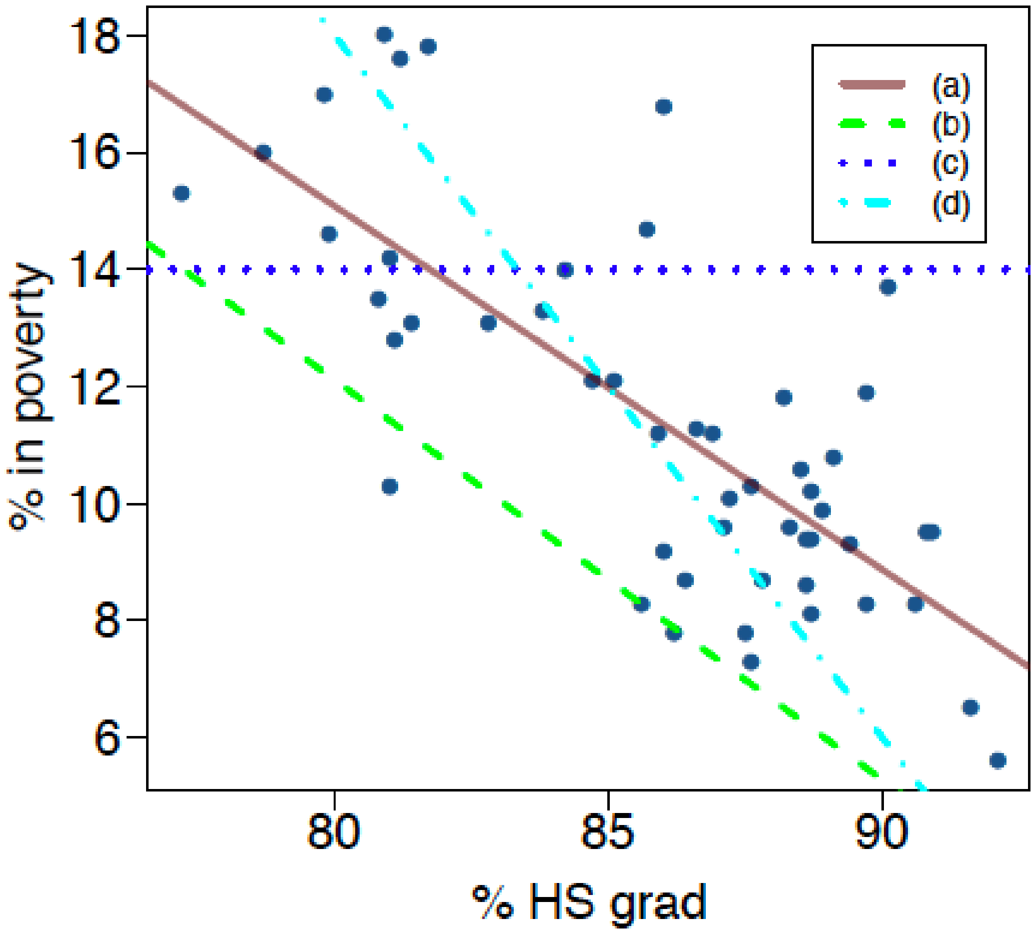
- This function comes from a hypothesis class. The hypothesis class includes certain functions, or functions of a particular type.
  - Linear, polynomial, etc.

- This scatterplot shows the relationship between US HS graduation rate and % of people living in poverty.



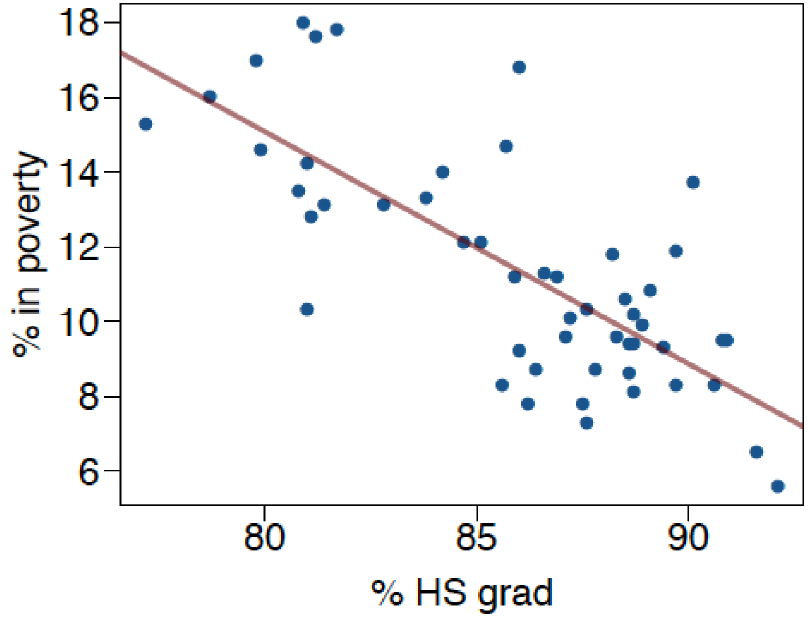
- What kind of hypothesis class might be appropriate?

Which Line?



Fitting a Line

- The line shown can be described by an an equation of the form $y = w_0 + w_1 x$



- In general $y = w_0 + w_1 x_1 + \ldots = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^\top \mathbf{x}$

  where $\mathbf{w}$ and $\mathbf{x}$ are $d+1$ length vectors.

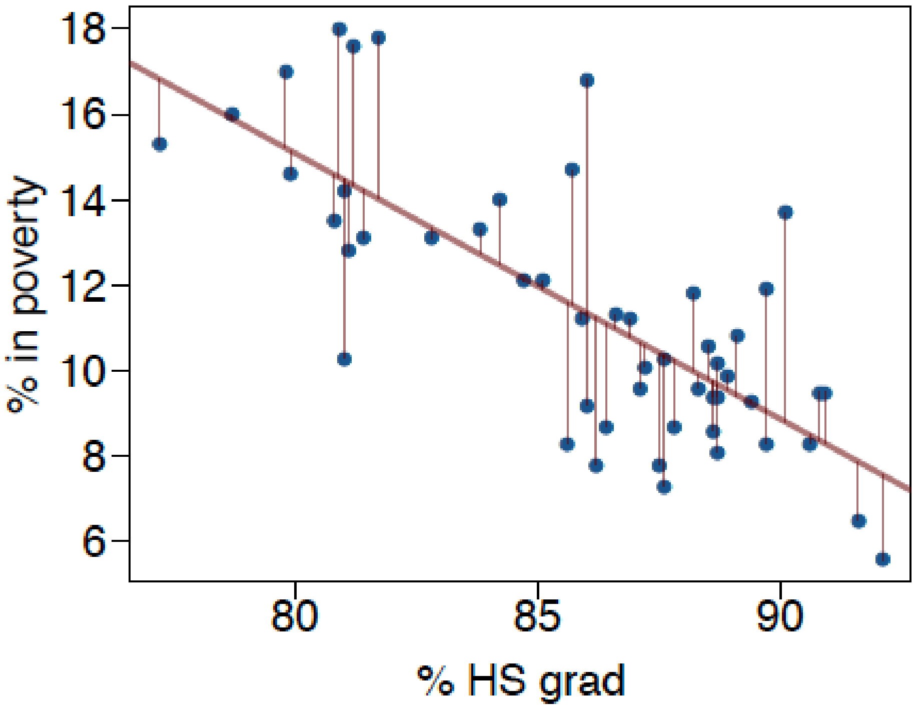
- Here $\mathbf{w}$ is what we want to learn, our parameters, or weights.

- How do we measure the quality of this line’s fit? How do we choose the best $\mathbf{w}$?

Residuals

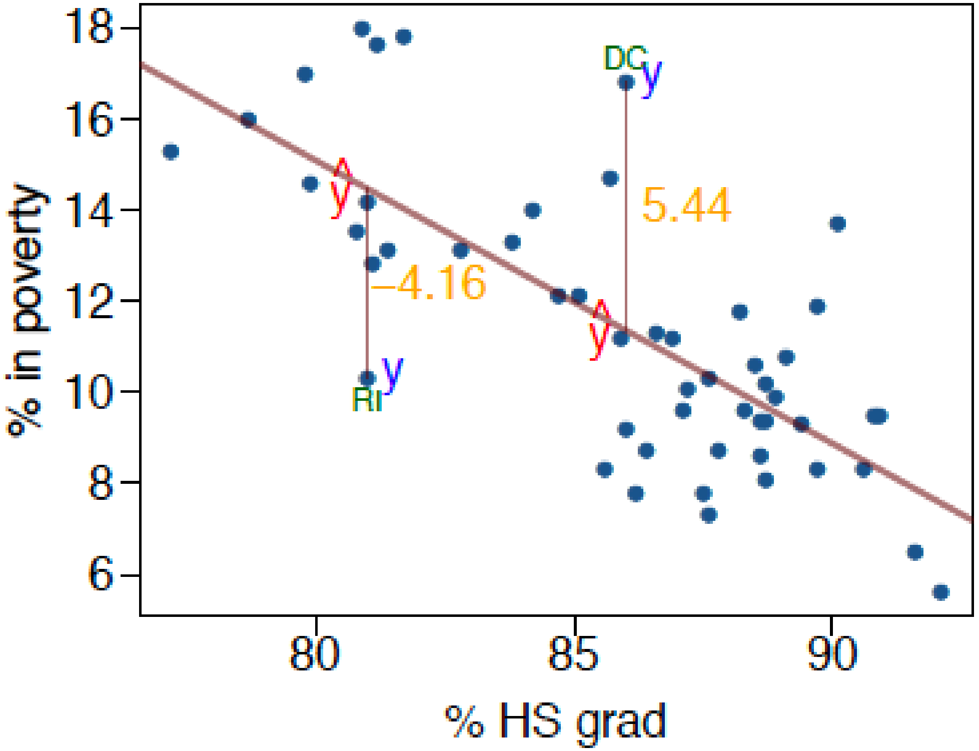
- We can think of each observed value, y_i, as being the result of our model $y = w_0 + w_1 x$ plus an unexplained error, e_i. This error is called a residual.

$$y_i = w_0 + w_1 x_i + e_i$$



Residuals

- The residual can be thought of as the observed value minus the value estimated by our linear function.



Loss – How good or bad is our line?

- We want a line that has small residuals. What should we do?
  - Minimize the sum of squared residuals – least squares

$$e_1^2 + e_2^2 + \ldots + e_n^2$$

- Why least squares?
  - Most commonly used
  - Square is a nicer function than absolute value
  - In many applications, twice the residual is more than twice as bad

Least Squares

- How do we use minimizing the squared residuals to find $\mathbf{w}$?

$$\arg\min_{\mathbf{w}} \sum_i e_i^2 = \arg\min_{\mathbf{w}} \sum_i (y_i - (w_0 + w_1 x_1))^2$$

- As usual, this is computed by taking the gradient with respect to $\mathbf{w}$ and setting equal to zero.

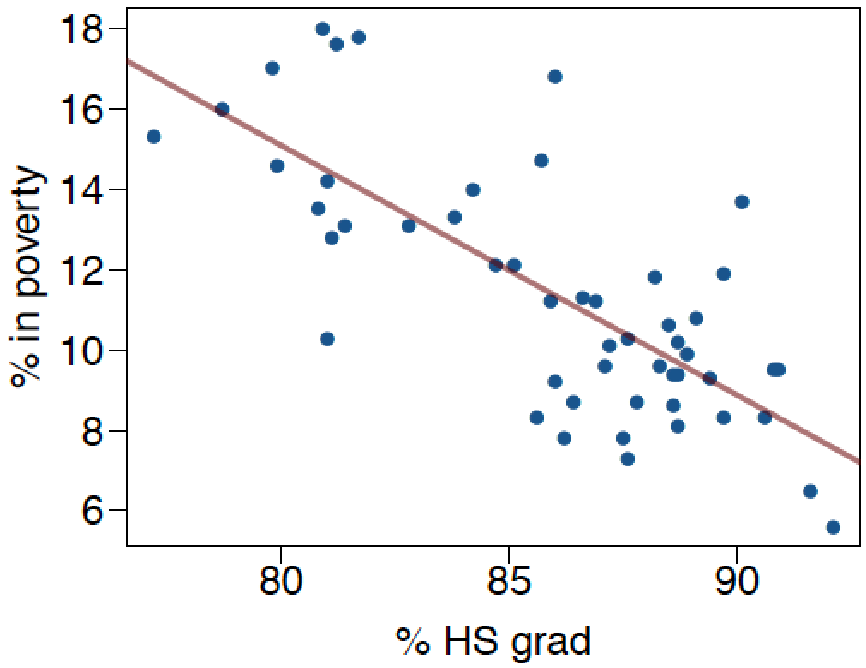
- The solution can be computed in closed form:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

So what are the weights?

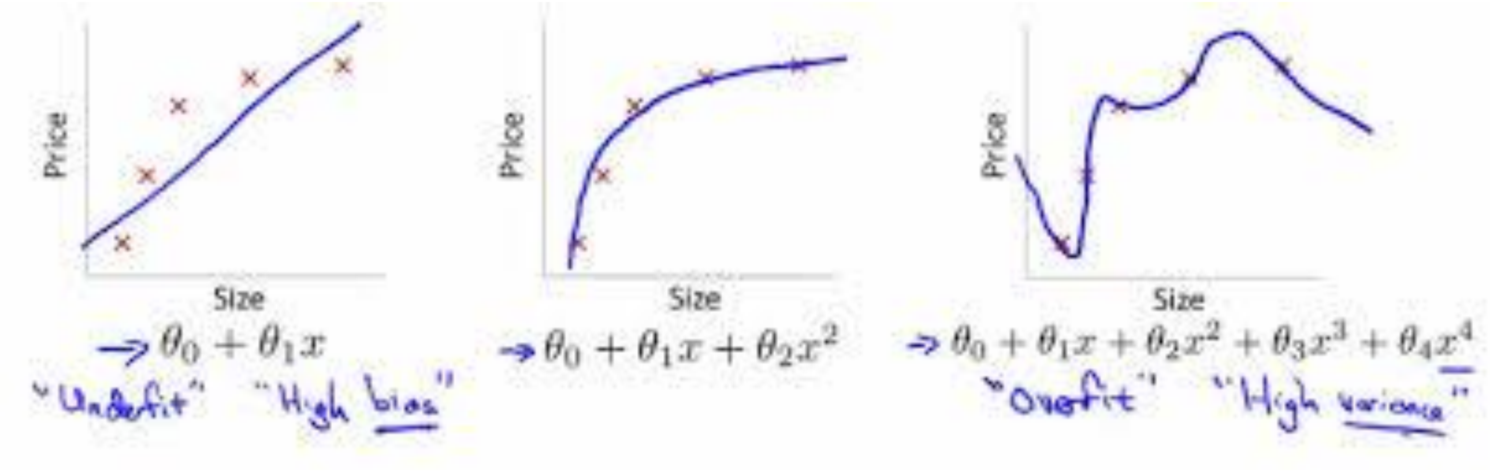
- For our example:

% in poverty = 64.68 – 0.62 (% HS grad)



How do we know a line is right?

- We can underfit or overfit to our data



- If we underfit we usually have high training error
- If we overfit we have very low training error, but high test error

Maximum Likelihood

- As we acquire more data, we can safely consider more complex hypotheses.
- What happens when we consider functions that have many parameters compared to our data set size?
- The approach that we have considered to finding parameters so far is a maximum likelihood approach. The probability of our data given the model is maximized with respect to the parameters.

$$\arg\max_{\theta} p(D|\theta, m)$$

- This can be done the same way as for least squares in linear regression – taking the gradient with respect to θ and setting equal to zero.
- In fact, our least squares solution can be seen as maximizing the likelihood of a Gaussian with mean $\mathbf{w}^\top \mathbf{x}$

Regularization

- Using the maximum likelihood approach we run the risk over overfitting if we have too many parameters.
- We can change the function we are optimizing to penalize complexity.

$$f(\theta) = L_\theta(\mathbf{x}, y) + \lambda R(\theta)$$

λ is the regularization coefficient and controls the tradeoff between goodness of fit and function simplicity.

- There are many different potential regularizers, including L1 and L2 complexity penalties.

L1 and L2 regularization

- Penalize complexity in different ways
- L1 penalizes the number of parameters used, while L2 penalizes the number of parameters squared.
- In our linear regression setting this would be:

$$\min_{\mathbf{w}} \frac{1}{n}||\mathbf{y} - \mathbf{w}^\top \mathbf{x}||^2 + \lambda ||w||_2^2$$

L2

$$\min_{\mathbf{w}} \frac{1}{n}||\mathbf{y} - \mathbf{w}^\top \mathbf{x}||^2 + \lambda ||w||_1$$

L1

- These regularizers correspond to Gaussian and Laplacian priors on the weights. L1 regularization also leads to shrinkage of the weight values towards zero. Still optimizing.

Bayesian Methods

- Bayesian methods provide a coherent framework for reasoning about our beliefs in the face of uncertainty.

Bayes Rule

$$p(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(\theta)$ - our prior beliefs about the state of the world, θ

$P(D|\theta)$ - the probability of observations,D, given a particular state, θ

$P(\theta|D)$ - our updated beliefs about the state of the world,θ, given the observations, D

- Avoid overfitting problem

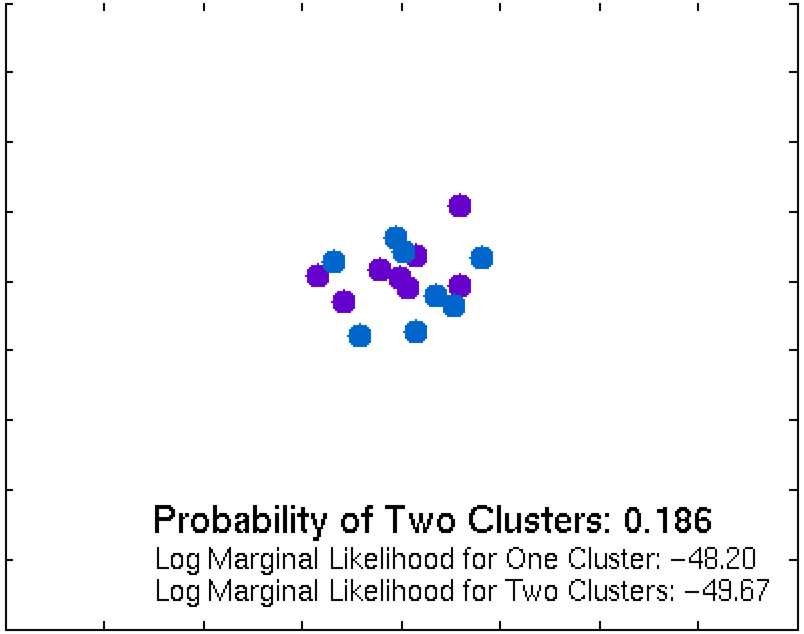
Marginal Likelihoods

- We use marginal likelihoods to evaluate cluster membership

- The marginal likelihood is defined as:

$$P(\mathcal{D}|m) \quad = \quad \int P(\mathcal{D}|\theta, m) P(\theta|m)\, d\theta$$

and can be interpreted as the probability that all data points in $\mathcal{D}$ were generated from the same model with unknown parameters θ

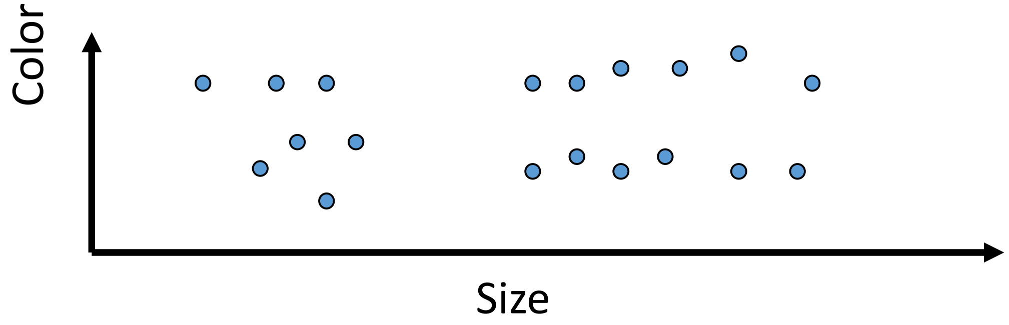


- Used to compare cluster models

m_1 - 1 cluster model $\qquad$ m_2 - 2 cluster model

$$p(m_2|\mathcal{D}) = \frac{p(\mathcal{D}|m_2)}{p(\mathcal{D}|m_1) + p(\mathcal{D}|m_2)}$$

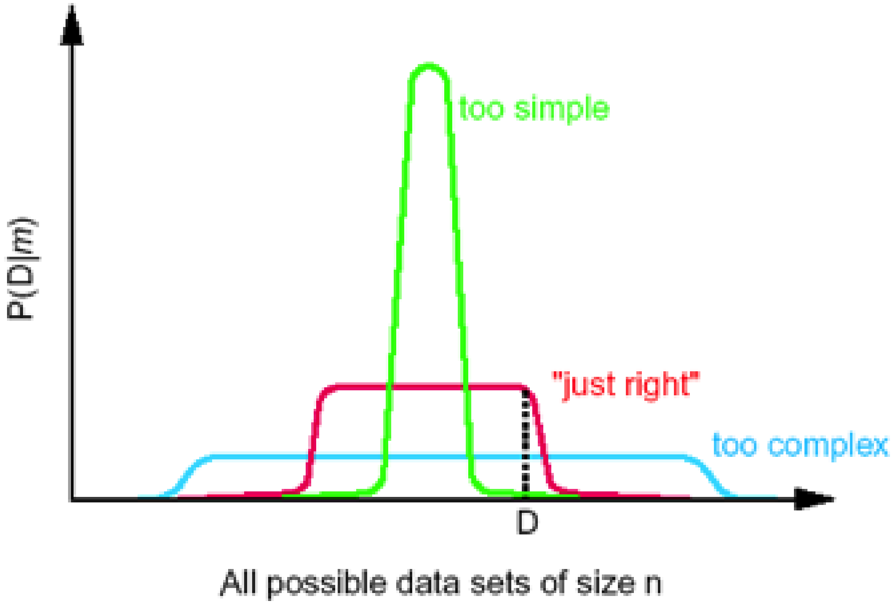
Bayesian Occam’s Razor



- Model with one cluster model, two cluster model, and three cluster model.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Nonparametric Bayesian Models

- How do we know which clustering models to compare?
  - Large numbers of model comparisons are costly.

- Nonparametric Bayesian methods provide flexible priors for clustering models.
  - Allow us to infer the “right” number of clusters for our data.

- Parametric models assume that some finite set of parameters, or clusters, capture everything there is to know about the data.
  - The complexity of the model is bounded.

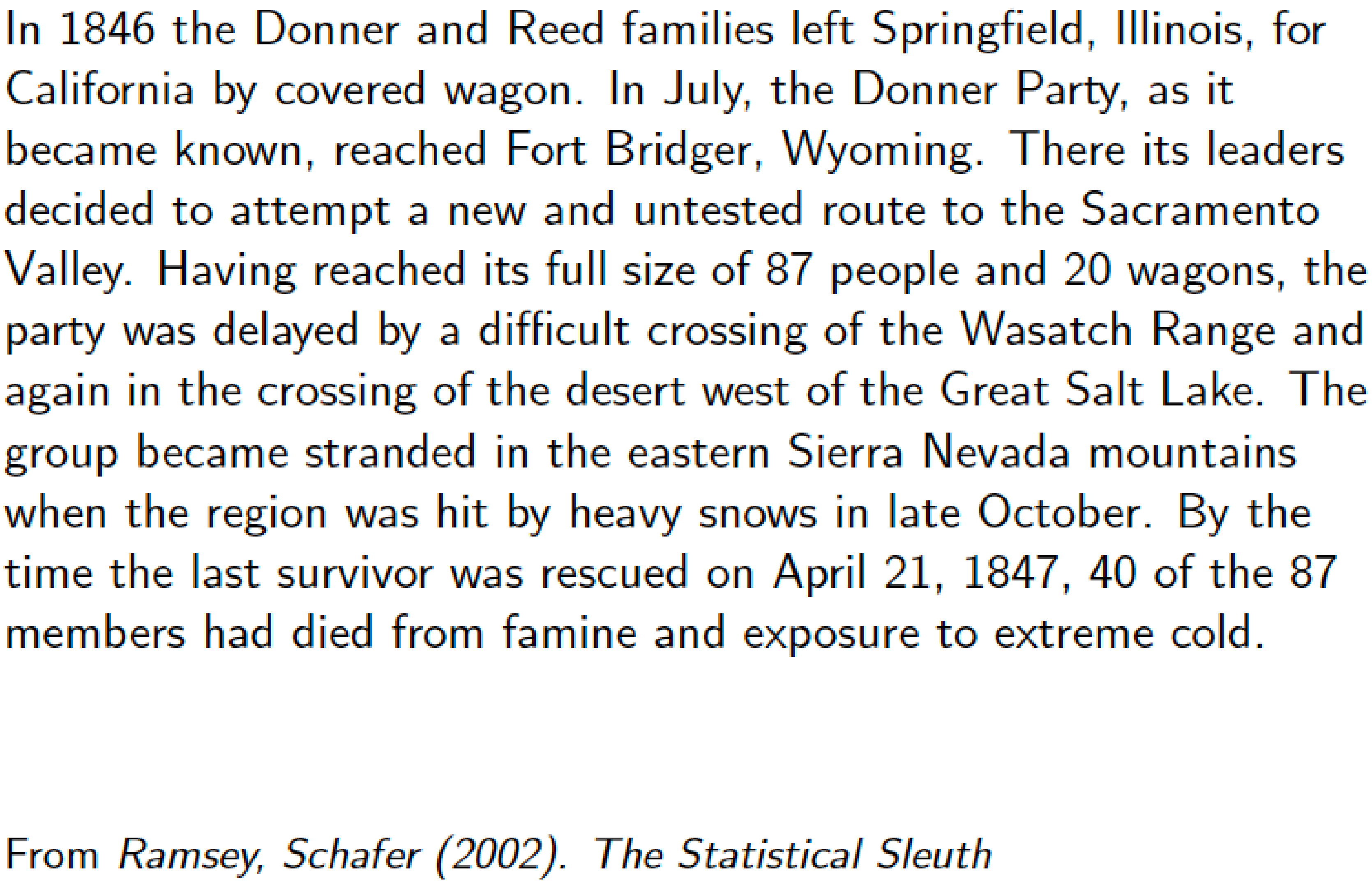
- Nonparametric models assume that an infinite set of parameters is needed.
  - The amount of information captured grows as the data grows.

More Sophisticated Regression Models

- In the same way that we looked at linear regression, we can look at other regression models that allow us to learn nonlinear functions.
  - Generalized Linear Models, Logistic Regression

- Can solve for the parameters here by taking gradients and setting equal to zero as in previous models, but there is no closed form solution and gradient descent must be used. More on optimization techniques to come.

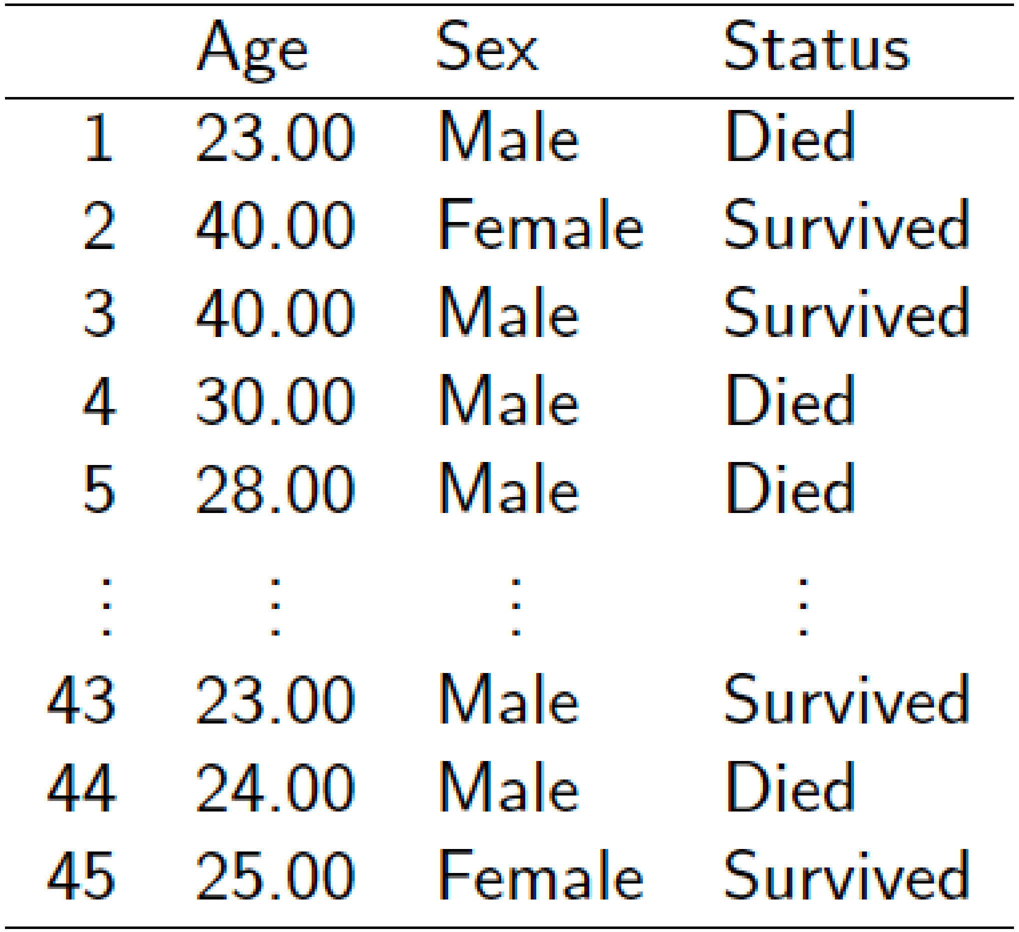
Example: Donner Party



In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, Schafer (2002). The Statistical Sleuth*

Donner Party Data



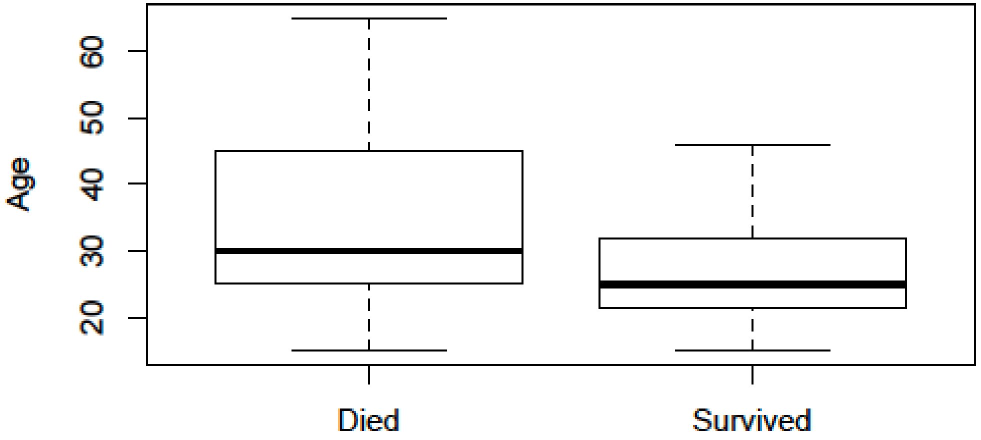
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Donner Party Data

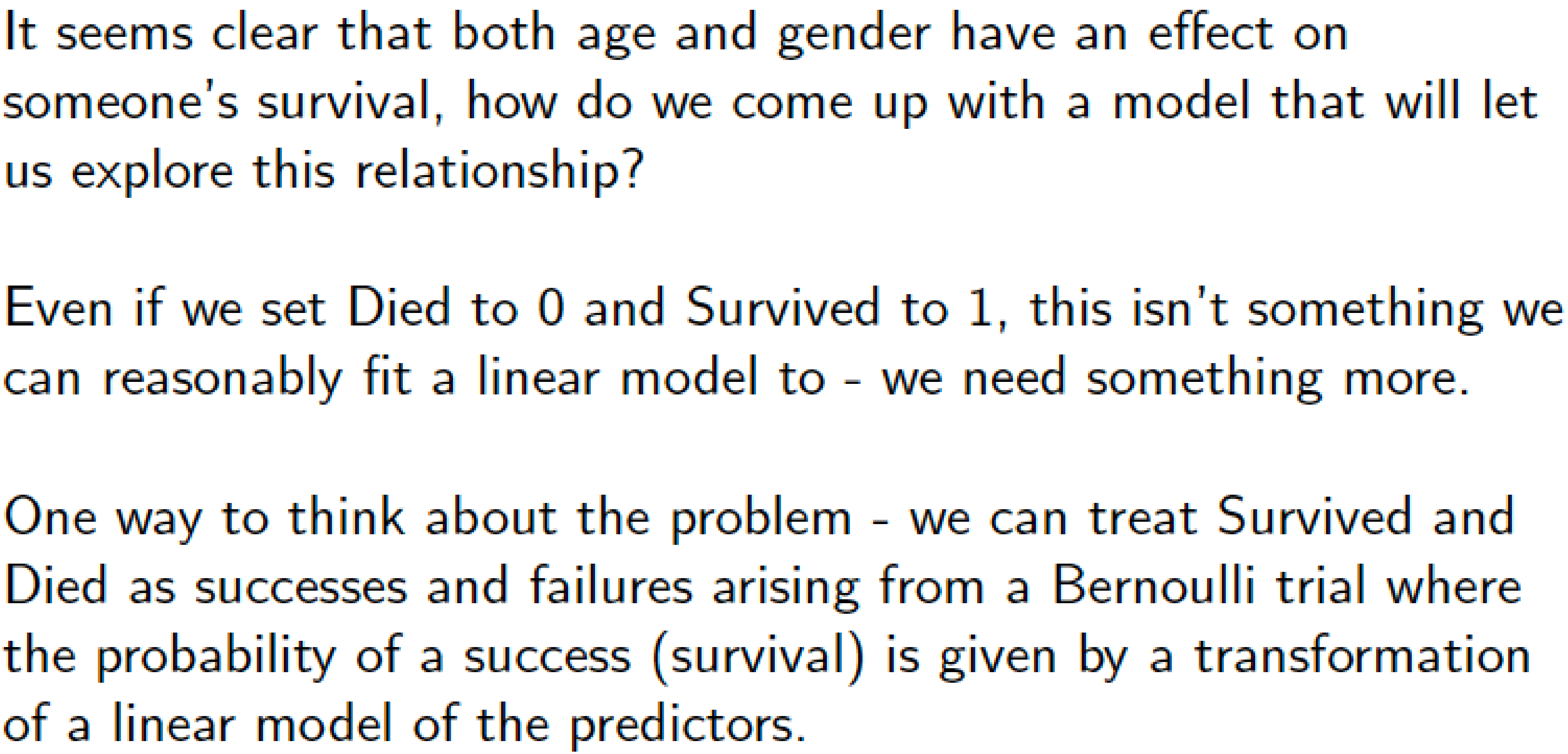
Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



Moving on from Linear Regression



It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can reasonably fit a linear model to - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a Bernoulli trial where the probability of a success (survival) is given by a transformation of a linear model of the predictors.

Generalized Linear Models

- GLMs are a very general way of addressing this type of problem. Logistic regression is just one type of GLM.

- All GLMs have three things:
  1. A probability distribution describing the outcome variable
  2. A linear model: $y = w_0 + w_1 x$
  3. A link function relating the linear model to the outcome distribution:

$$p = g^{-1}(y)$$

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects y to p. There are a variety of options but the most commonly used is the logit function.

Logit function:

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

The Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function:

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation is also useful for interpreting the model, since the logit can be interpreted as the log odds of a success

The Logistic Regression Model

- The three GLM criteria give us:

$$\eta \sim \mathrm{bernoulli}(p))$$

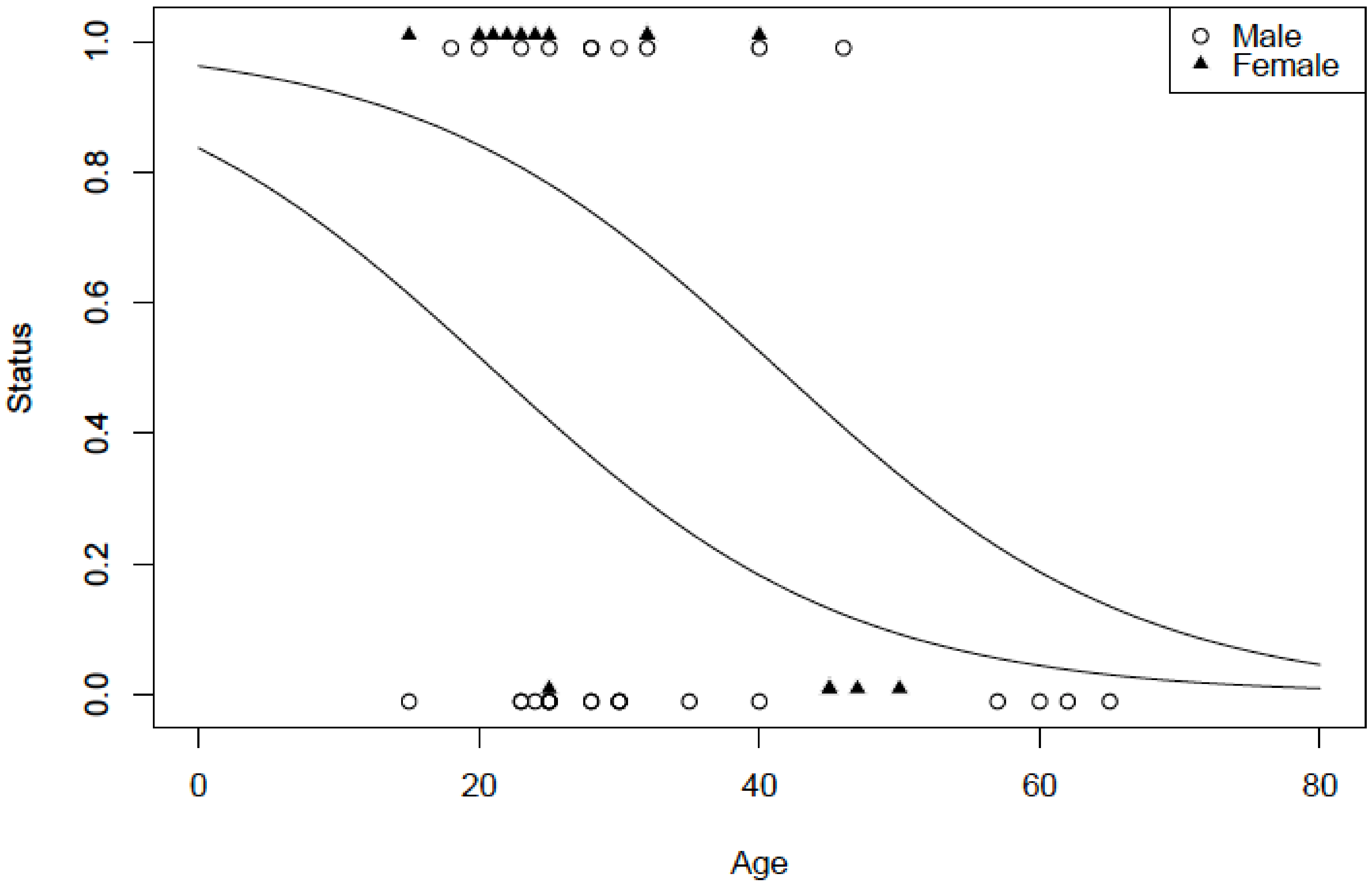
$$y = w_0 + w_1 x$$

$$\mathrm{logit}(p) = y$$

- From which we get:

$$p = \frac{\exp(w_0 + w_1 x)}{1 + \exp(w_0 + w_1 x)}$$

Logistic Regression Plot



Summary

- We looked at different kinds of machine learning
- Linear models and how to fit them
- When models are too simple or complex for our data
- Overfitting
- Regularization
- Bayesian Techniques
- Generalized Linear Models and Logistic Regression

Moving Forward

- These are really just the basics. More detail will be covered going forward in this summer school.

- An important thing to keep in mind is that the primary purpose of this is to develop useful models of real data of interest.

- Lots of cool things being done in the machine learning community, and lots of use of these and more sophisticated methods to have real world impact.

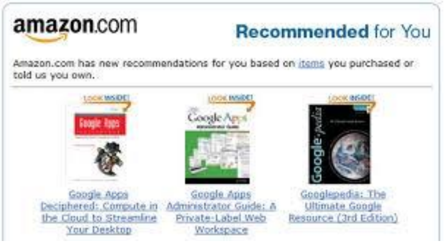
Moving Forward

- These include:

  Self driving cars

  Drones and Robots

  Product recommendation



amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop

Google Apps Administrator Guide: A Private-Label Web Workspace

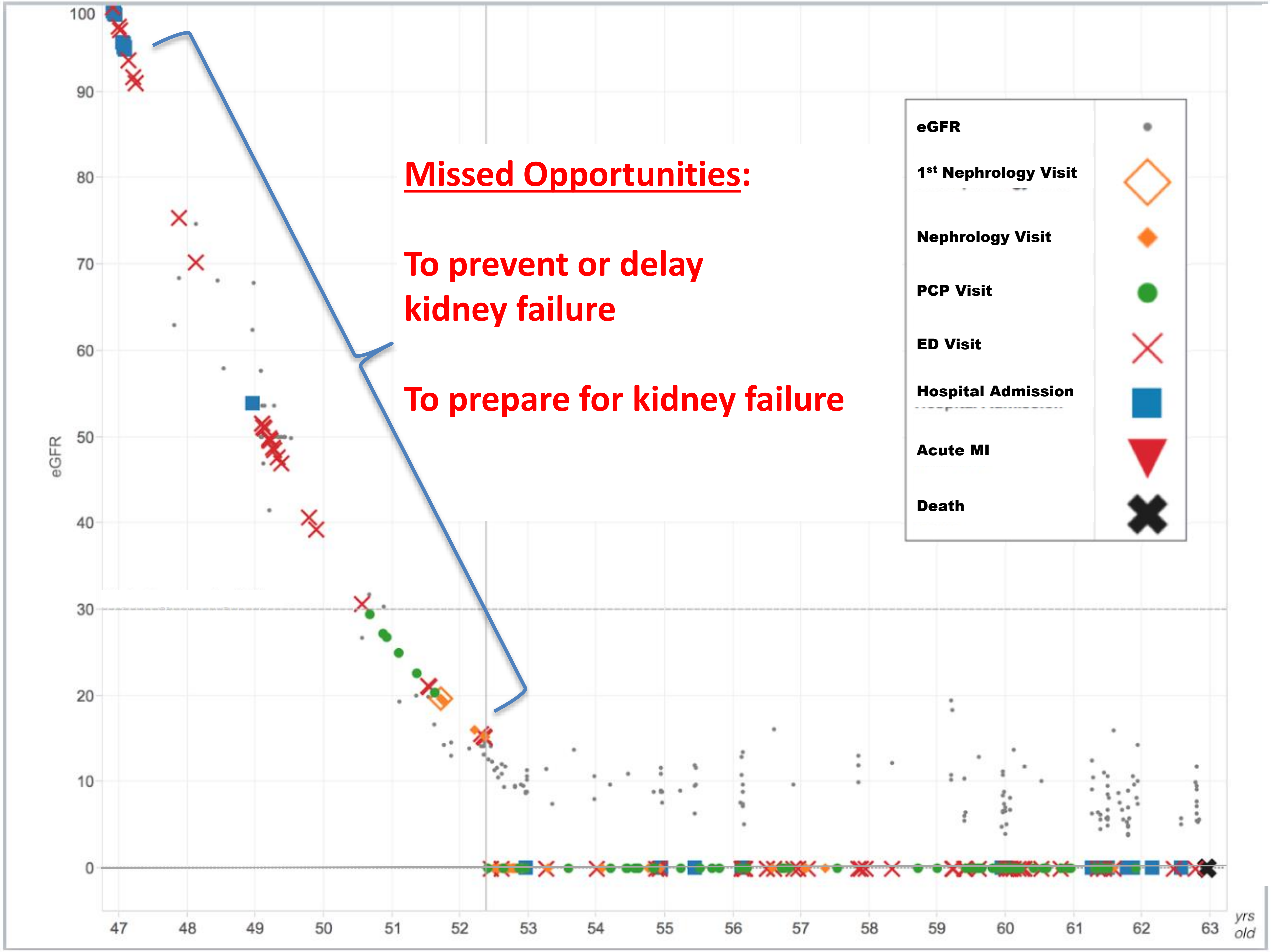
Googlepedia: The Ultimate Google Resource (3rd Edition)

  Criminal justice recommendations

  Health care predictions

But Still..

- There’s a lot that we don’t know. And that we need your help with.
- Where do the gains we are seeing in deep neural networks come from?
- How is regularization working? Early stopping? Dropout?
- How can we speed things up and still have principled approaches to analyzing data?

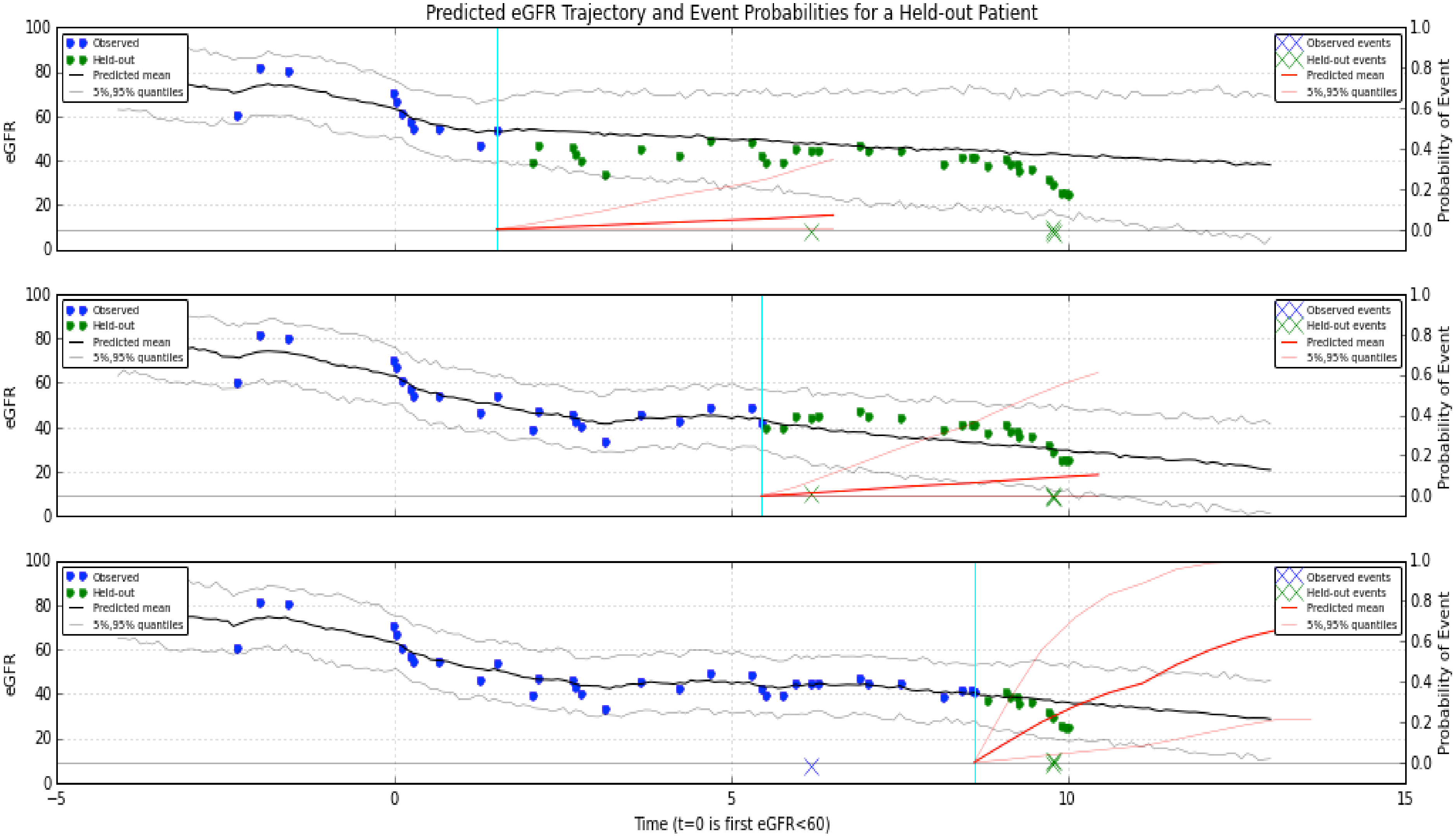


Missed Opportunities:

To prevent or delay kidney failure

To prepare for kidney failure

Joint Model Results



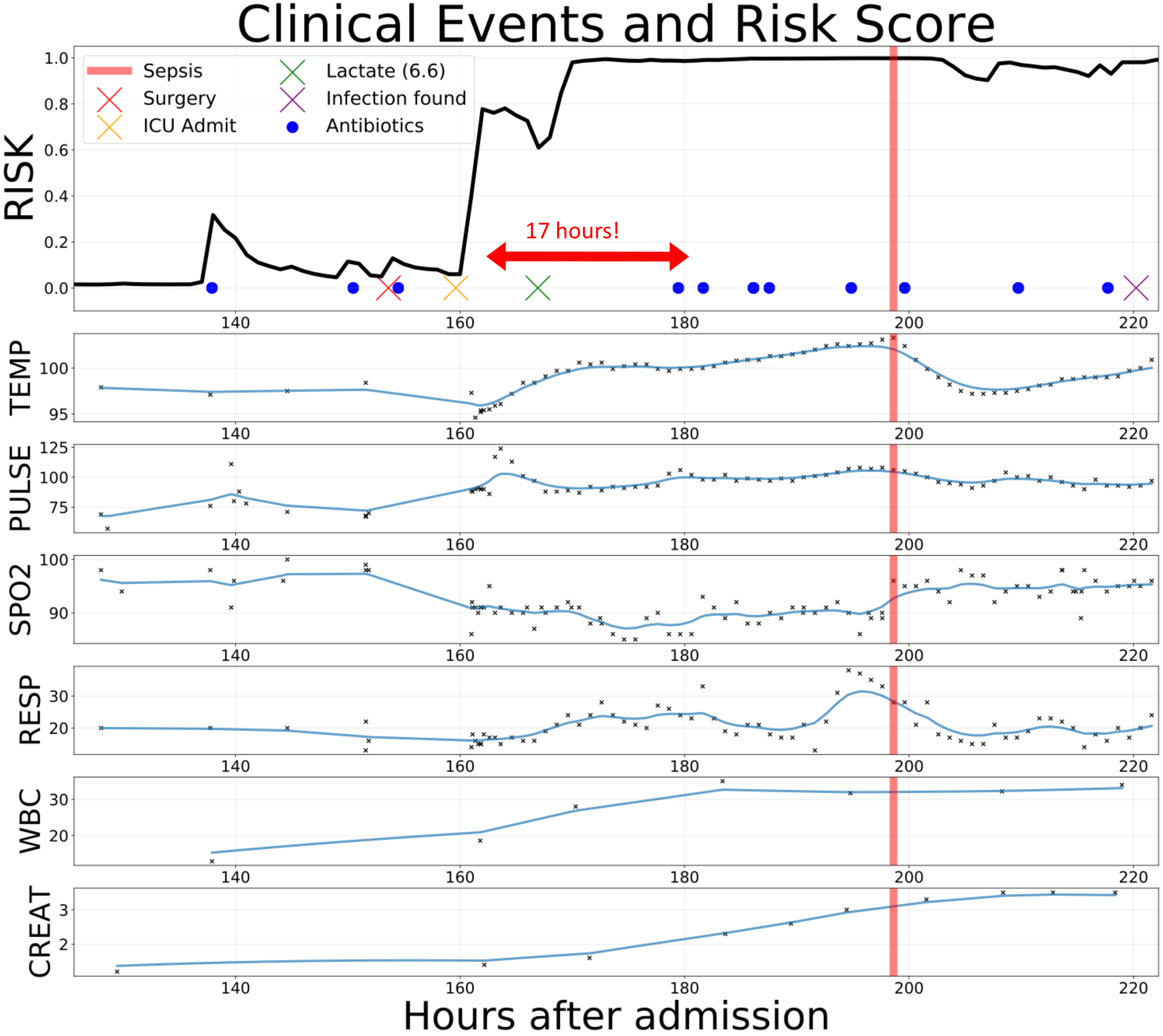
Sepsis

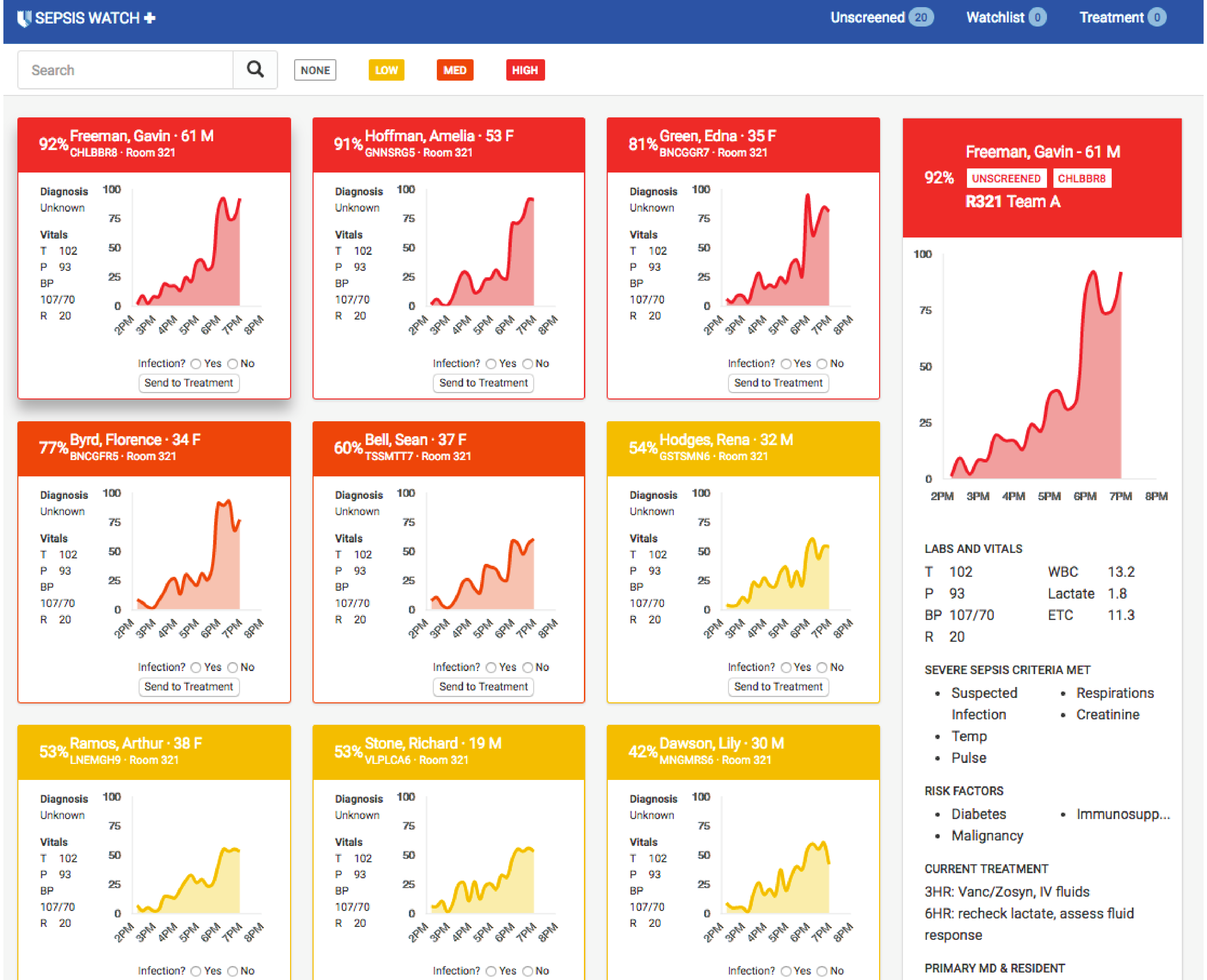
- Severe infection, often acquired in the hospital. High mortality rate (30-50%).

- Developed an algorithm to predict if a patient is becoming septic.
  - Multitask Gaussian process with a Recurrent Neural Net classifier, learned end to end.

- A web app shows patient state and trajectories to physicians. In the process of deploying at Duke University Hospital.

- Exploring ways to use reinforcement learning to make automated treatment recommendations.


Clinical Events and Risk Score
RISK
1.0
0.8
0.6
0.4
0.2
0.0
Sepsis
Surgery
ICU Admit
Lactate (6.6)
Infection found
Antibiotics
17 hours!
140
160
180
200
220
TEMP
100
95
PULSE
125
100
75
SPO2
100
90
RESP
30
20
WBC
30
20
CREAT
3
2
Hours after admission



SEPSIS WATCH ✚

Unscreened 20 Watchlist 0 Treatment 0

Search

NONE LOW MED HIGH

92% Freeman, Gavin · 61 M
CHLBBR8 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

91% Hoffman, Amelia · 53 F
GNNSRG5 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

81% Green, Edna · 35 F
BNCGGR7 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

77% Byrd, Florence · 34 F
BNCGFR5 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

60% Bell, Sean · 37 F
TSSMTT7 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

54% Hodges, Rena · 32 M
GSTSMN6 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Send to Treatment

53% Ramos, Arthur · 38 F
LNEMGH9 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

53% Stone, Richard · 19 M
VLPLCA6 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

42% Dawson, Lily · 30 M
MNGMRS6 · Room 321

Diagnosis
Unknown

Vitals
T 102
P 93
BP
107/70
R 20

Infection? ○ Yes ○ No

Freeman, Gavin - 61 M
92% UNSCREENED CHLBBR8
R321 Team A

LABS AND VITALS

T	102	WBC	13.2
P	93	Lactate	1.8
BP	107/70	ETC	11.3
R	20		

SEVERE SEPSIS CRITERIA MET

- Suspected Infection
- Temp
- Pulse
- Respirations
- Creatinine

RISK FACTORS

- Diabetes
- Malignancy
- Immunosupp...

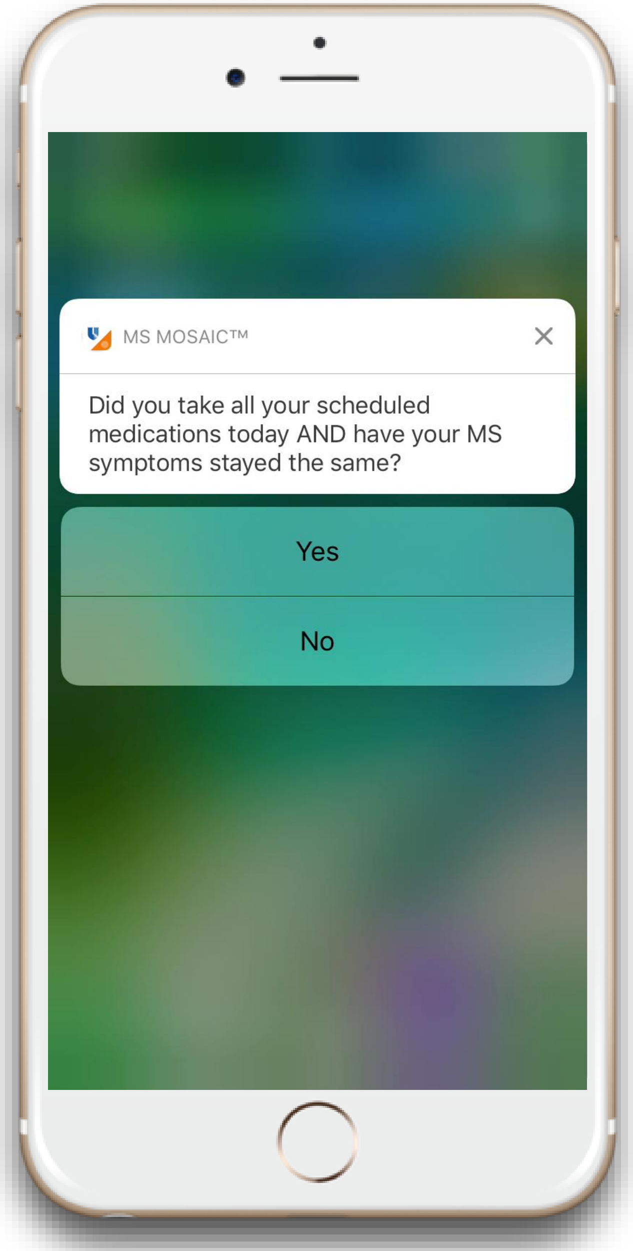
CURRENT TREATMENT

3HR: Vanc/Zosyn, IV fluids
6HR: recheck lactate, assess fluid response

PRIMARY MD & RESIDENT

MS Mosaic App

- For neurology patients with multiple sclerosis.
- Consents subjects
- Collects survey, activity, and phone task data
- Try to reduce work for users
- Dashboard to visualize own data
- Reports for providers
- Learn more section
- Want to combine with other data



Did you take all your scheduled medications today AND have your MS symptoms stayed the same?