Data Science, Statistics, and Health

with a focus on Statistical Learning and Sparsity

and applications to biomedicine

Rob Tibshirani
Departments of Biomedical Data Science & Statistics
Stanford University

SDSS 2020

Outline

1. Some general thoughts about data science and health
2. Some comments about supervised learning, sparsity, and deep learning.
3. General advice for data scientists
4. **Example**: Predicting platelet usage at Stanford Hospital
5. **Example**: The Delphi project- COVID19 forecasting

There is a lot of excitement about Data Science and health

- Artificial intelligence, predictive analytics, precision medicine are all hot areas with huge potential
- A wealth of data is now available in every area of public health and medicine, from new machines and assays, smart phones, smart watches ....
- Already, there have been good successes in data science in pathology, radiology, and other diagnostic specialties.

For Statisticians: 15 minutes of fame

- 2009: " I keep saying the sexy job in the next ten years will be statisticians." Hal Varian, Chief Economist Google

- 2012 "Data Scientist: The Sexiest Job of the 21st Century" Harvard Business Review

Some obstacles to this success

- **Data siloing** is a problem. In the US health care system researchers tend not to share data openly. Access to large, representative datasets is essential for this work.
- **The bar is higher in health.** There is more more at stake in the health area than in say a recommendation system for movies. Errors are much more costly
- The public may be **less tolerant** of data science/machine errors than human errors (analogous to self-driving cars).

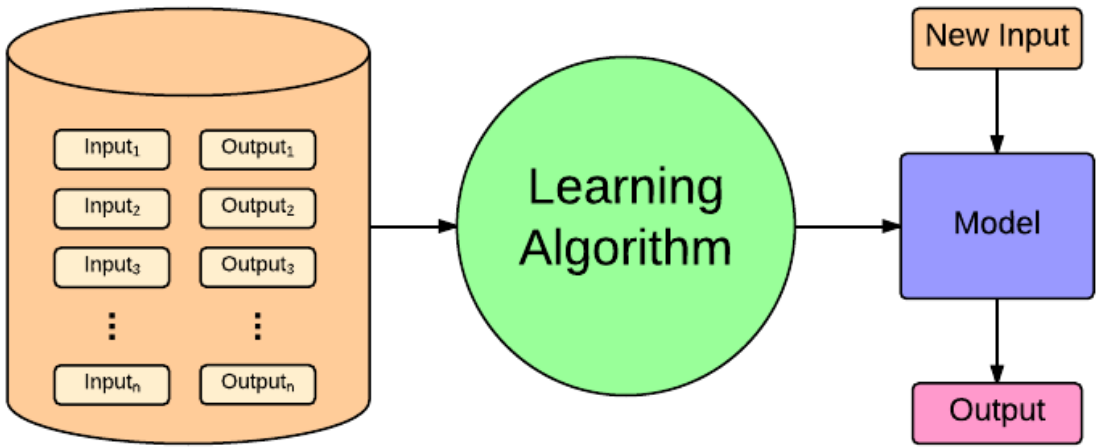
High level comments about Data Science, Statistics and Machine learning

- Data Science and Statistics involve **much more** than simply running a **machine learning algorithm** on data
- For example:
  - What is the question of interest? How can I collect data or run a experiment to address this question?
  - What inferences can be drawn from the data?
  - What action should I take as a result of what I've learned?
  - Do I need to worry about bias, confounding, generalizability, concept drift ...?
- $\rightarrow$**Statistical concepts and training are essential**

Important points for data science applications in health

- Models should be kept as **simple as possible**, as they are easier to understand. *And we can more easily anticipate how/when they might break.* See Brad Efron's recent paper "Prediction, Estimation, and Attribution"
- **Uncertainly quantification** is essential. We must build trust in our models.
- **Algorithmic fairness** is important
- Example of a key unsolved problem: in classification, with high dimensional features, how can we arrange for our classifier to sometimes "**abstain**"? (because it is extrapolating).
- Estimation of **heterogeneous treatment effects (HTE)**, eg for personalized medicine, is a very important area in need of research. A simple unsolved problem: how can we do internal cross-validation of a model for HTE?

The Supervising Learning Paradigm



Training Data Fitting Prediction

Traditional statistics: domain experts work for 10 years to learn good features; they bring the statistician a small clean dataset

Today's approach: we start with a large dataset with many features, and use a machine learning algorithm to find the as good ones. A huge change.

This talk is about supervised learning: building models from data for predicting an outcome using a collection of input features.
Big data vary in *shape*. These call for different approaches.

Wide Data

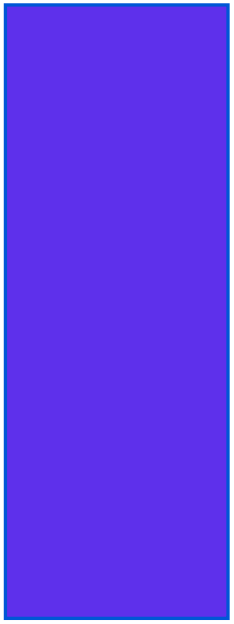
Thousands / Millions of Variables

Hundreds of Samples

Lasso & Elastic Net

We have too many variables; prone to overfitting.
Lasso fits linear models to the data that are *sparse* in the variables.
Does automatic variable selection.

Tall Data



Tens / Hundreds of Variables

Thousands / Tens of Thousands of Samples

Random Forests &
Gradient Boosting

Sometimes simple models (linear) don't suffice.
We have enough samples to fit nonlinear models with many interactions, and not too many variables.
A Random Forest is an automatic and powerful way to do this.

The Lasso

The **Lasso** is an estimator defined by the following optimization problem:

$$\underset{\beta_0,\beta}{\text{minimize}}\ \frac{1}{2}\sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 \qquad \text{subject to} \quad \sum |\beta_j| \le s$$

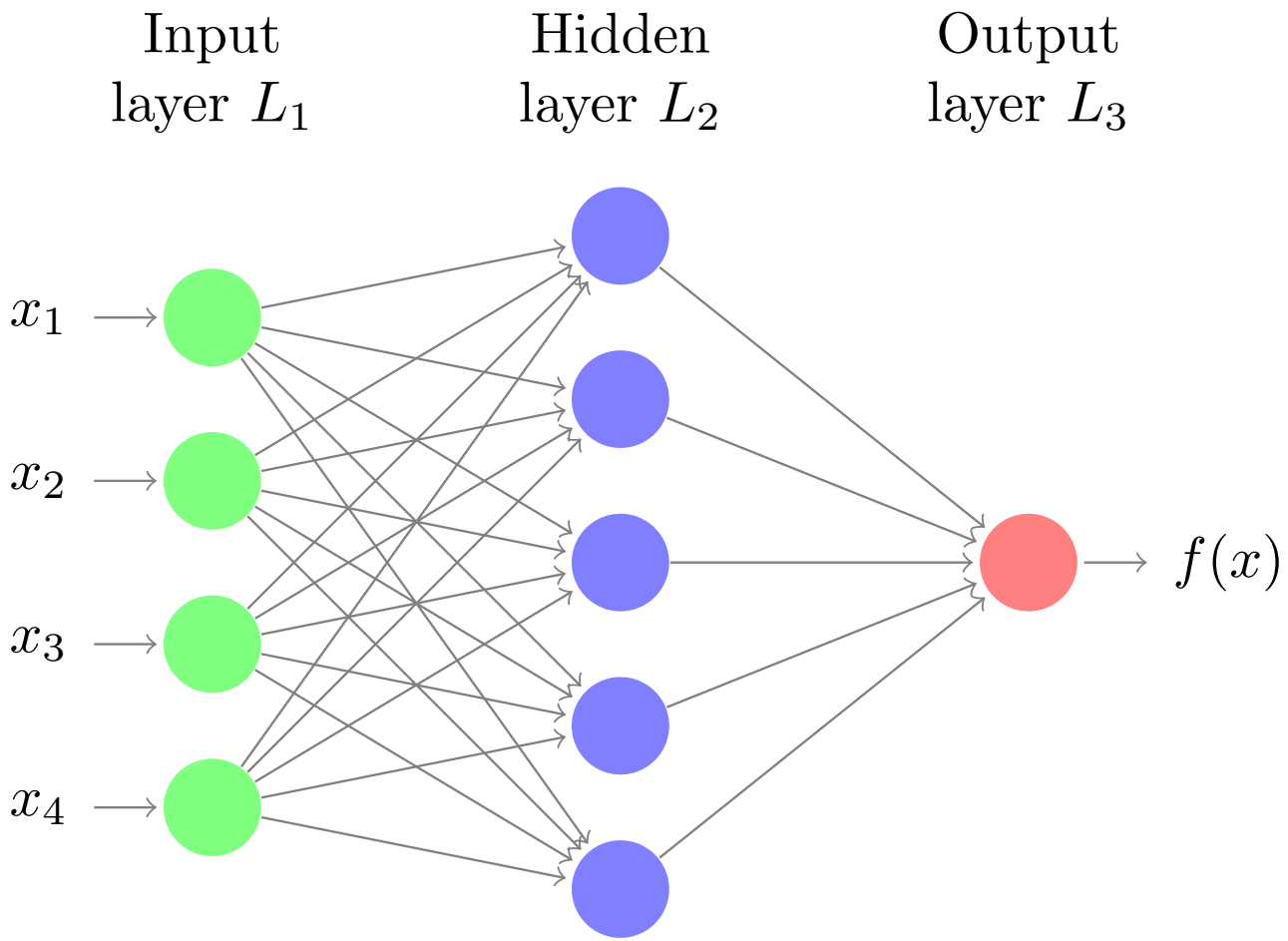
- Penalty $\implies$ sparsity (feature selection)
- Convex problem (good for computation and theory)
- Our lab has written a open-source R language package called **glmnet** for fitting lasso models (Friedman, Hastie, Simon, Tibs). Available on CRAN.
- glmnet v3.0 (just out) now features the *relaxed lasso* and other goodies (like a progress bar!)

  Almost 2 million downloads

The Elephant in the Room: DEEP LEARNING

Will it eat the lasso and other statistical models?

Deep Nets/Deep Learning



Neural network diagram with a single hidden layer. The hidden layer derives transformations of the inputs — nonlinear transformations of linear combinations — which are then used to model the output

What has changed since the invention of Neural nets?

- Much bigger training sets and faster computation (especially GPUs)
- Many clever ideas: convolution filters, stochastic gradient descent, input distortion ...
- Use of **multiple layers** (the "Deep" part): Tommy Poggio says that the universal approximation theorems for single layer NNs in the 1980s **set the field back 30 years**
- **Confession:** I was Geoff Hinton's colleague at Univ. of Toronto (1985-1998) and didn't appreciate the potential of Neural Networks!

What makes Deep Nets so powerful

(and challenging to analyze!)

It's not one "mathematical model" but a **customizable framework**– a set of **engineering tools** that can exploit the special aspects of the problem (weight-sharing, convolution, feedback, recurrence ...)

Will Deep Nets eat the lasso and other statistical models?

Deep Nets are especially powerful when the features have some spatial or temporal organization (signals, images), and SNR is high

But they are not a good approach when

- we have moderate #obs or wide data (#obs $<$ #features),
- SNR is low, or
- interpretability is important
- It's difficult to find examples where Deep Nets beat lasso or GBM in low SNR settings, with "generic " features

General advice for Data Scientists

Details matter!

Q: *Have you tried method X?*
A: I tried that last year and it didn't work very well!

Q: *What do you mean? What exactly did you try? How did you measure performance?*
A: I can't remember.

General tips

- Try a number of methods and use **cross-validation** to tune and compare models (our "lab" is the computer-experiments are quick and free)
- Be **systematic** when you run methods and make comparisons: compare methods on the same training and test sets.
- Keep **carefully documented scripts** and data archives (where practical).
- Your work should be **reproducible** by **you and others**, even a few years from now!

In Praise of Simplicity

'Simplicity is the ultimate sophistication" — Leonardo Da Vinci

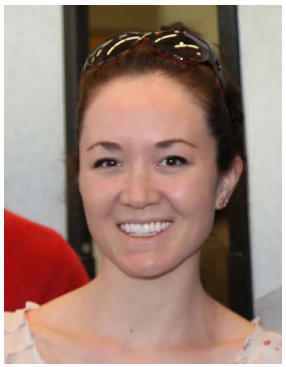
- Many times I have been asked to review a data analysis by a biology postdoc or a company employee. Almost every time, they are unnecessarily complicated. Multiple steps, each one poorly justified.
- Why? I think we all like to justify– internally and externally— our advanced degrees. And then there's the "hit everything with deep learning" problem
- Suggestion: Always try simple methods first. Move on to more complex methods, only if necessary

Outline again

1. Some general thoughts about data science and health
2. Some comments about supervised learning, sparsity, and deep learning
3. General advice for data scientists
4. $\rightarrow$ **Example**: Predicting platelet usage at Stanford Hospital
5. **Example** The Delphi project- COVID19 forecasting

How many units of platelets will the Stanford Hospital need tomorrow?

Allison Zemek

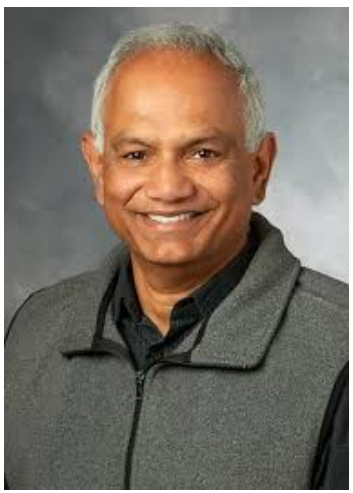


Tho Pham

Saurabh Gombar

Leying Guan

Xiaoying Tian



Balasubramanian Narasimhan

PNAS

Big data modeling to predict platelet usage and minimize wastage in a tertiary care system

Leying Guan[a,1], Xiaoying Tian[a,1], Saurabh Gombar[b], Allison J. Zemek[b], Gomathi Krishnan[c], Robert Scott[d], Balasubramanian Narasimhan[a], Robert J. Tibshirani[a,e,2], and Tho D. Pham[b,d,f,2]

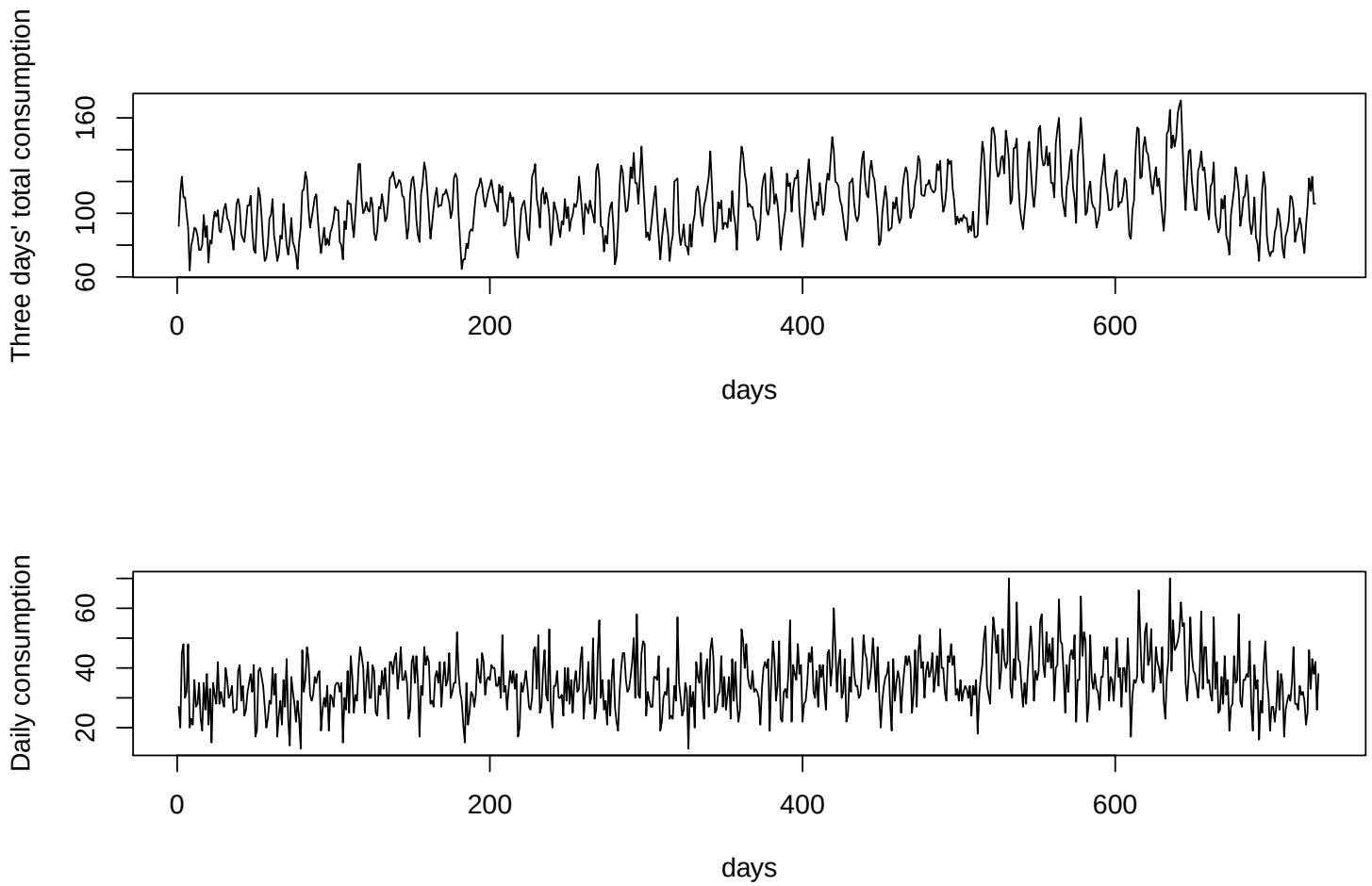
[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Pathology, Stanford University, Stanford, CA 94305; [c]Stanford for Clinical Informatics, Stanford University, Stanford, CA 94305; [d]Stanford Hospital Transfusion Service, Stanford Medicine, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Stanford Blood Center, Stanford Medicine, Stanford, CA

Contributed by Robert J. Tibshirani, August 10, 2017 (sent for review June 25, 2017; reviewed by James Burner, Pearl Toy, and Minh-Ha Tran)

Background

- Each day Stanford hospital orders some number of units (bags) of platelets from Stanford blood center, based on the estimated need (roughly 45 units)
- The daily needs are estimated "manually"
- Platelets have just 5 days of shelf-life; they are safety-tested for 2 days. Hence are **usable for just 3 days.**
- Currently about **1400** units (bags) are wasted each year. That's about **8%** of the total number ordered.
- There's rarely any shortage (shortage is bad but not catastrophic)
- Can we do better?

Data overview



Data description

Daily platelet use from 2/8/2013 - 2/8/2015.

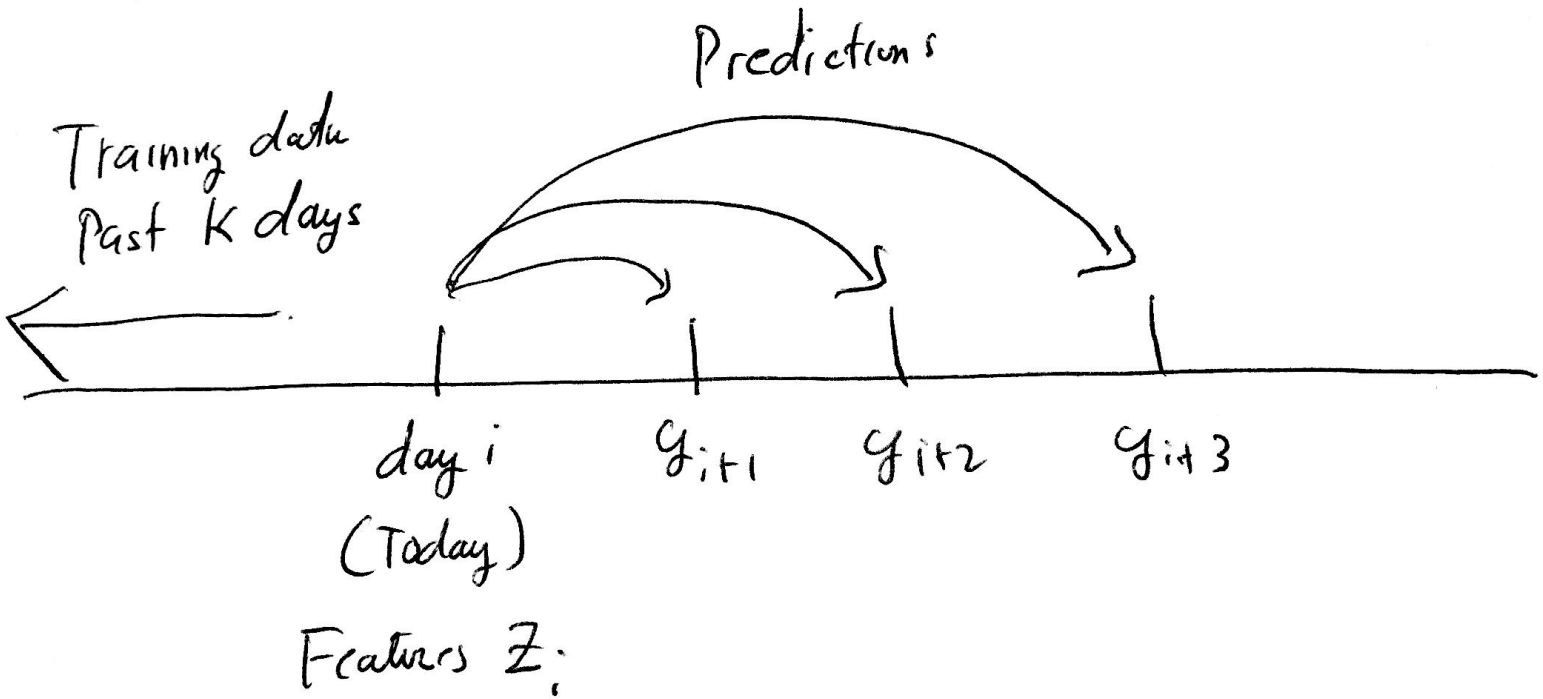
- Response: number of platelet transfusions on a given day.
- Covariates:
  1. **Complete blood count (CBC) data**: Platelet count, White blood cell count, Red blood cell count, Hemoglobin concentration, number of lymphocytes, ...
  2. **Census data**: location of the patient, admission date, discharge date, ...
  3. **Surgery schedule data**: scheduled surgery date, type of surgical services, ...
  4. ...

Notation

y_i : actual PLT usage in day i.
x_i : amount of new PLT that arrives at day i.
$r_i(k)$: remaining PLT which can be used in the following k days, $k = 1, 2$
w_i : PLT wasted in day i.
s_i : PLT shortage in day i.

- **Overall objective**: waste as little as possible, with little or no shortage

Our first approach



Our first approach

- Build a supervised learning model (via lasso) to predict use y_i for next three days (other methods like random forests or gradient boosting didn't give better accuracy).
- Use the estimates $\hat{y}_i$ to estimate how many units x_i to order. Add a buffer to predictions to ensure there is no shortage. Do this is a "rolling manner".
- Worked quite well- reducing waste to 2.8%- - but the loss function here is not ideal

More direct approach

This approach minimizes the waste directly:

$$J(\beta) = \sum_{i=1}^{n} w_i + \lambda ||\beta||_1 \tag{1}$$

where

$$\text{three days}' \text{ total need} \quad t_i = z_i^T \beta, \quad \forall i = 1, 2, .., n \tag{2}$$

$$\text{number to order :} \quad x_{i+3} = t_i - r_i(1) - r_i(2) - x_{i+1} - x_{i+2} \tag{3}$$

$$\text{waste} \quad w_i = [r_{i-1}(1) - y_i]_+ \tag{4}$$

$$\text{actual remaining} \quad r_i(1) = [r_{i-1}(2) + r_{i-1}(1) - y_i - w_i]_+ \tag{5}$$

$$r_i(2) = [x_i - [y_i + w_i - r_{i-1}(2) - r_{i-1}(1)]_+]_+ \tag{6}$$

$$\text{Constraint : fresh bags remaining} \quad r_i(2) \geq c_0 \tag{7}$$

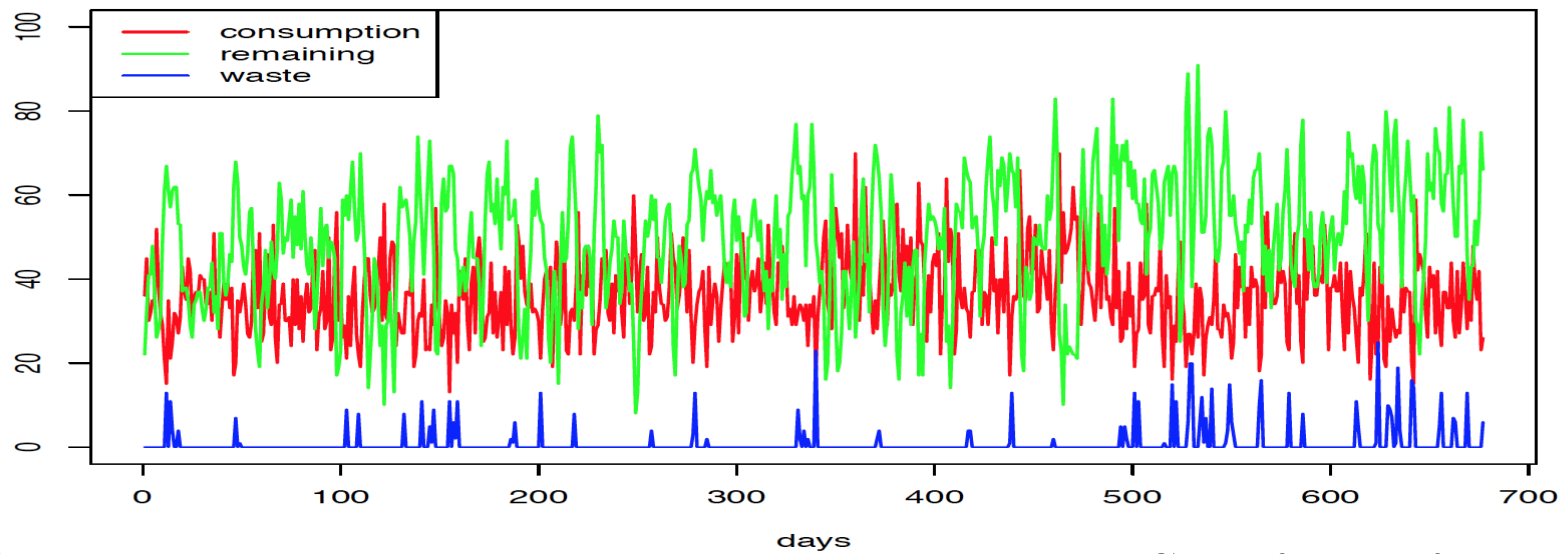
(8)

This can be shown to be a convex problem (LP).

Results

Chose sensible features- previous platelet use, day of week, # patients in key wards.

Over 2 years of backtesting: no shortage, reduces waste from **1400** bags/ year (8%) to just **339** bags/year (1.9%)



Corresponds to a predicted direct savings at Stanford of $350,000/year. If implemented nationally could result in approximately $110 million in savings.

Moving forward

- System has just been deployed at the Stanford Blood center (R Shiny app). **But yet not in production**
- We are distributing the software around the world, for other centers to train and deploy
- see Platelet inventory R package `https://bnaras.github.io/pip/`
- Just this week we learned that the predictions have been way off for the past month! Data problem? Model problem? Not sure yet.
- **A remaining challenge:** How can we detect if/when the system is no longer working well?

Covidcast: A map of Real-time COVID-19 Indicators

- For the past month, I have been working with Roni Rosenfeld (Chair of ML), Ryan Tibshirani (Statistics+ML) and their **Delphi** flu prediction team at Carnegie Mellon University.
- The Delphi group has been doing influenza forecasting for the CDC for the past five years, as part of the CDC national influenza forecasting challenge.
- They have done well- finishing first in the CDC national influenza forecasting challenge 3 of the past 4 years.
- They were awarded a CDC Center of Excellence in September 2019, along with U. Mass.

Then covid arrived

- In March the CDC asked Delphi (and other groups) to make covid-19 predictions, and this led to the current effort involving about 25 software engineers and statisticians
- The team includes, professors, research assistants and current grad students. All are at CMU except Lester MacKay (MS research, formerly Stanford stat), David Farrow (engineer on loan from Google) and myself.
- We launched a national covid-19 symptoms map last week, and work continues on Phase 2: **forecasting hospitalization usage**

Ryan Tibshirani

Lead Researcher, Delphi COVID-19 Response Team

Associate Professor, Department of Statistics and Machine Learning Department

Carnegie Mellon University

Roni Rosenfeld

Lead Researcher, Delphi COVID-19 Response Team

Professor and Head, Machine Learning Department

School of Computer Science

Carnegie Mellon University

Delphi COVID-19 Response Team

Getting good data is the key

- We needed data on covid-19 symptoms (not just confirmed cases) in order to forecast hospitalization needs
- Ryan approached Google, Facebook, Amazon and other companies, asking them to conduct surveys. He has spent about a month in negotiations.
- Main problem: legal concerns about health data privacy. FB solution: rather than having the survey run by FB on their site, they put just a link to an external survey at CMU.
- With this model, Amazon and Google joined. We have good data so far from Google Surveys (600K/day) and FB (100K/day).

The questions

FaceBook survey

1. In the past 24 hours, have you or anyone in your household had (yes/no for each):
   a. Fever (of 100 degrees or higher)
   b. Sore throat
   c. Cough
   d. Shortness of breath
   e. Difficulty breathing
2. How many people in your household (including yourself) are sick (fever, along with at least one other symptom from the above list)?
3. How many people are there in your household in total (including yourself)?
4. What is your current ZIP code?

From this, covid positive is defined as fever and at least one of (cough, shortness of breath, difficulty breathing)
We also measure influenza positive as fever and at least one of (cough, sore throat)

Google asks : *how many people in your community that you know are sick with fever, along with one of sore throat, shortness of breath, cough or difficulty breathing?*

Other data sources

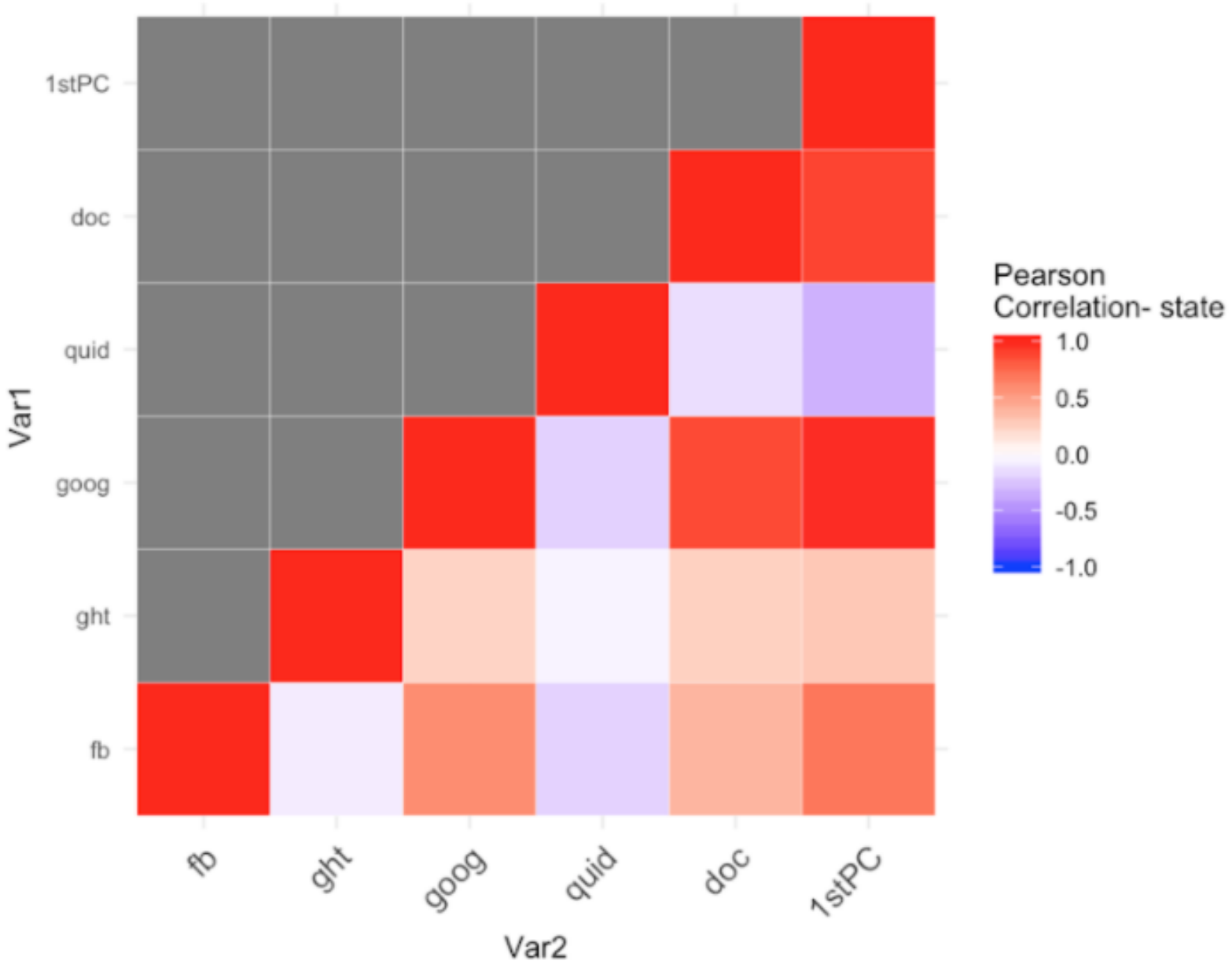
- A major medical/hospital provider is giving us real time data on doctors visits, hospital admissions, and ICU admissions
- Google Health trends is providing estimates of the percentage of Google Searches that relate to covid symptoms
- Quidel (diagnostic healthcare manufacturer) is giving us data on the number of lab tests for influenza

The experience for me

- **Exciting!**: like being in a startup.
- daily zoom meetings, github, slack, a lot of coding
- I had no idea the amount of work involved in such a project; statistical (survey sampling, estimating trends, estimating uncertainty) and lots of software engineering
- The data is complicated and at many resolutions: state, metropolitan area, hospital referral region and county.
- We launched thursday morning. At around 10pm wed, the entire map/site was broken; eventually we figured out it was because someone had decided to change a file name without telling anyone!

SHOW THE MAP `https://covidcast.cmu.edu/`

Correlation and consensus



Things I've learned

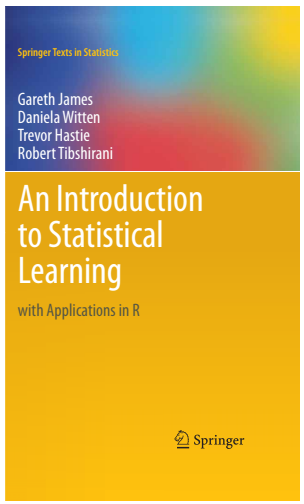
- An greater appreciation for R, R markdown, github and CRAN. Also: we are making substantial use `glmnet`.
- BUT: many CRAN libraries need improvement- numerics, defaults, documentation
- Spend time on this: it's really, really important

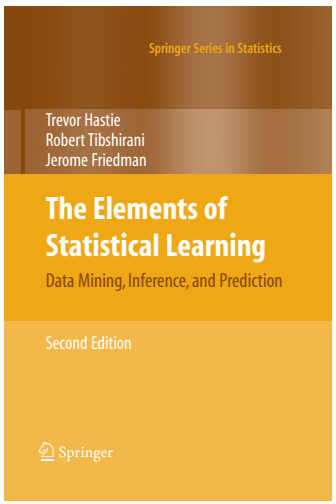
The next stage

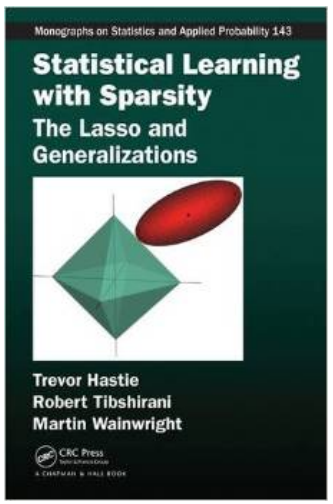
- Use current signals, as well as data on hospital admissions to forecast hospitalization needs (in each Hospital referral region) a few weeks ahead. Our goal is to launch this in 2 weeks.
- Try to assess the effects of interventions like Shelter at Home
- We will have other data sources
- Help inform public health officials, to make better decisions. Ryan + Roni have already been speaking to senators in Penn, and some officials in Washington. Ryan & I spoke to the California Dept of Public Health today

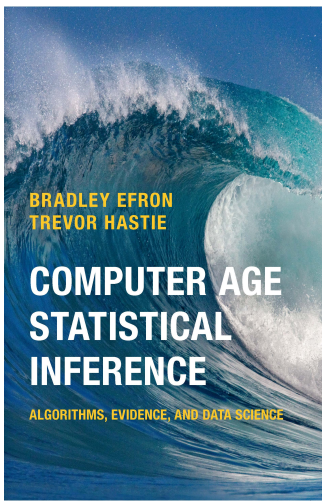
For further reading

Many of the methods used are described in detail in our books on Statistical Learning: (last one by Efron & Hastie)









All available online for free