# An Application of the Lasso in Biomedical data sciences

*H2O World 2017*

Rob Tibshirani
Departments of Biomedical Data Science & Statistics
Stanford University

# The Lasso for supervised learning

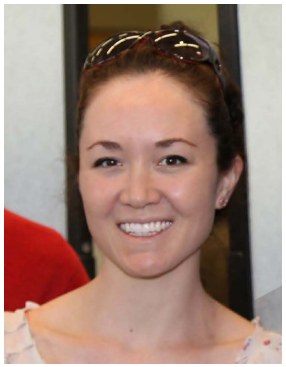
Given features $x_{ij}$ and an outcome measurement $y_i$,
the **Lasso** is an estimator defined by the following optimization problem:

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

- 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. Available on CRAN. Also available in **h2o.glm** in H2O.

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


Allison Zemek



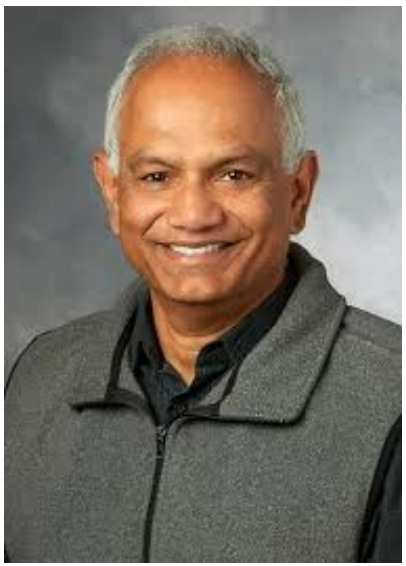
Tho Pham

Saurabh Gombar

Leying Guan

Xiaoying Tian

# Balasubramanian Narashiman is the key person in the deployment phase



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]

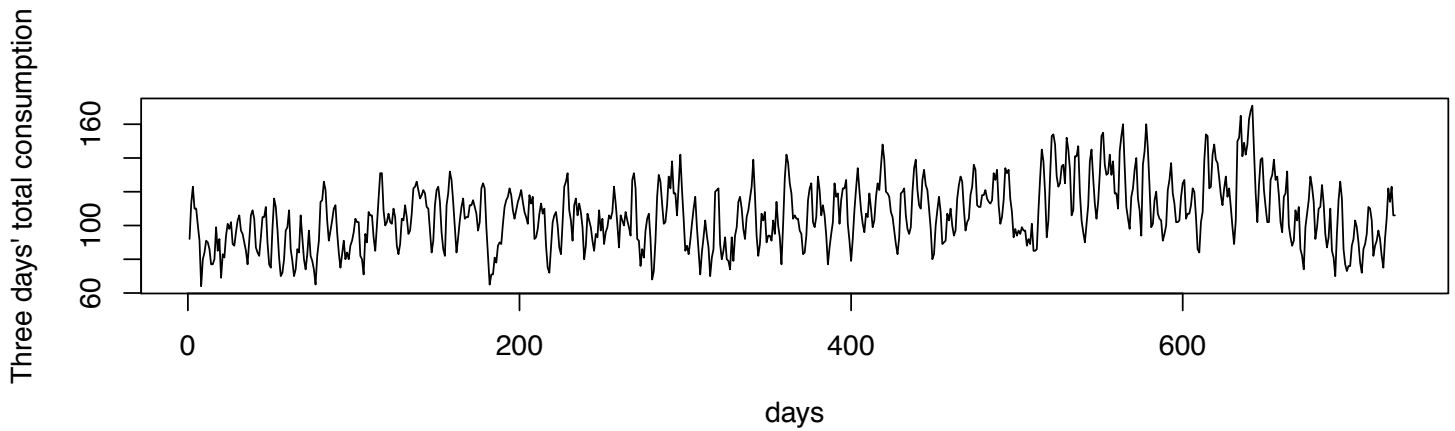
[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Pathology, Stanford University, Stanford, CA 94305; [c]Stanford for Clinical Informatics, Stanford University, Stanford, CA 94305; [d]Stanford Hospital Transfusion Service, Stanford Medicine, Stanford, CA 94305; [e]Department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

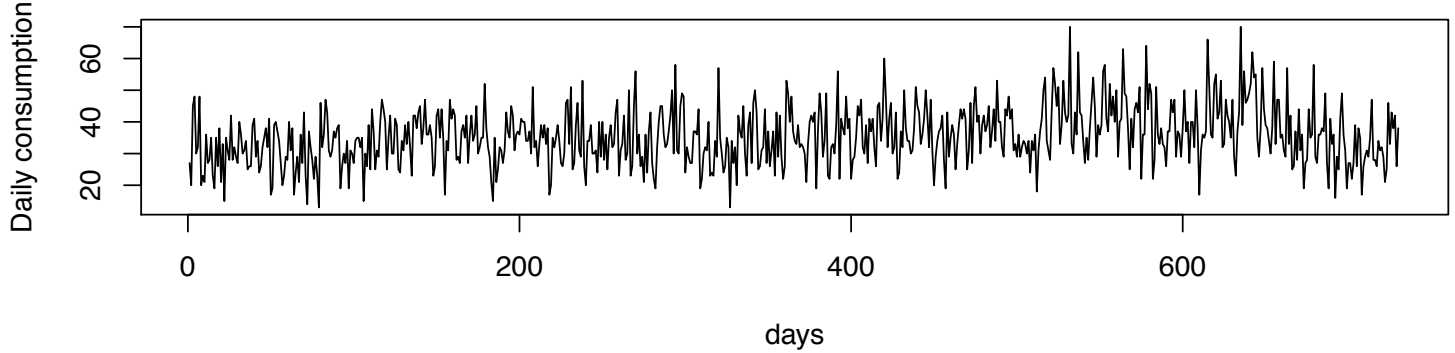
Contributed by Robert J. Tibshirani, August 10, 2017 (sent for review June 25, 2017; reviewed by James Burner, Pearl Toy, and Minh-Ha Tran)

# The current system

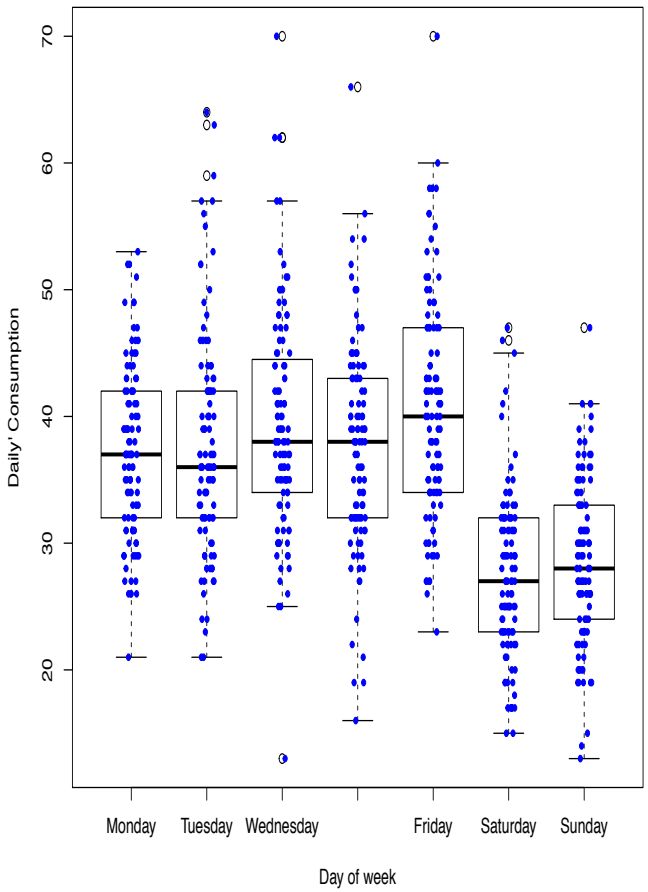
- Each day, the Stanford blood center collects some number of units (bags) of platelets, based on the estimated needs at Stanford Hospital. The daily needs are estimated "manually".
- Platelets have a **5 day shelf life**, and are safety-tested for 2 days. So they are **usable for just 3 days**, and are discarded after that time.
- 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 use available information about the hospital to do better?*

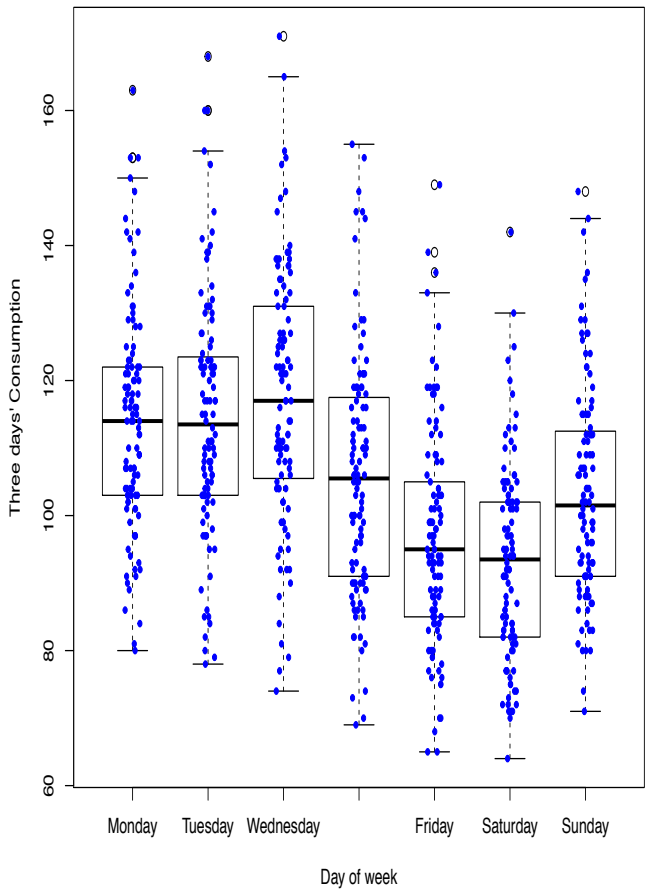
# Data overview



# Data overview- continued



# 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. ...

# Data description

We first tried to work on the individual patient level, but there were many complications:

1. **Complete blood count (CBC) data**:
   - 30% of patients have no CBC measurement at all
   - After being measured, a patient can (1) have a transfusion right away; (2) leave the hospital; (3) come back later in the future but we do not know when.
2. **Census data**:
   - Often there no matching medical record number.
3. **Surgery schedule data**:
   - Often does not match previous data file at the personalized level.
   - Large percentage of missingness.

Conclusion: **Use aggregated features.**

# Feature Construction

- CBC measurement: for each day $i$ and feature $j$, count the number of patients below the first quartile of the population. Use the average of the past week(11 features).
- CENSUS record: for each day $i$, count the total number of patients at a location $j$ in the hospital (26 features).

# Feature Construction— continued

- PLT transfusion record: for each day $i$, let $y_i$ be the total number of PLT used at day $i$. Use the average of past week $\bar{y}_i$ at day $i$ when making prediction(1 feature).
- SURGERY record: for each day $i$, and count the number of scheduled surgeries at day $i + k$ when making prediction for day future $k$ days, $k = 1, 2, 3$ (17 features).
- Day of the week information: Monday,...,Sunday

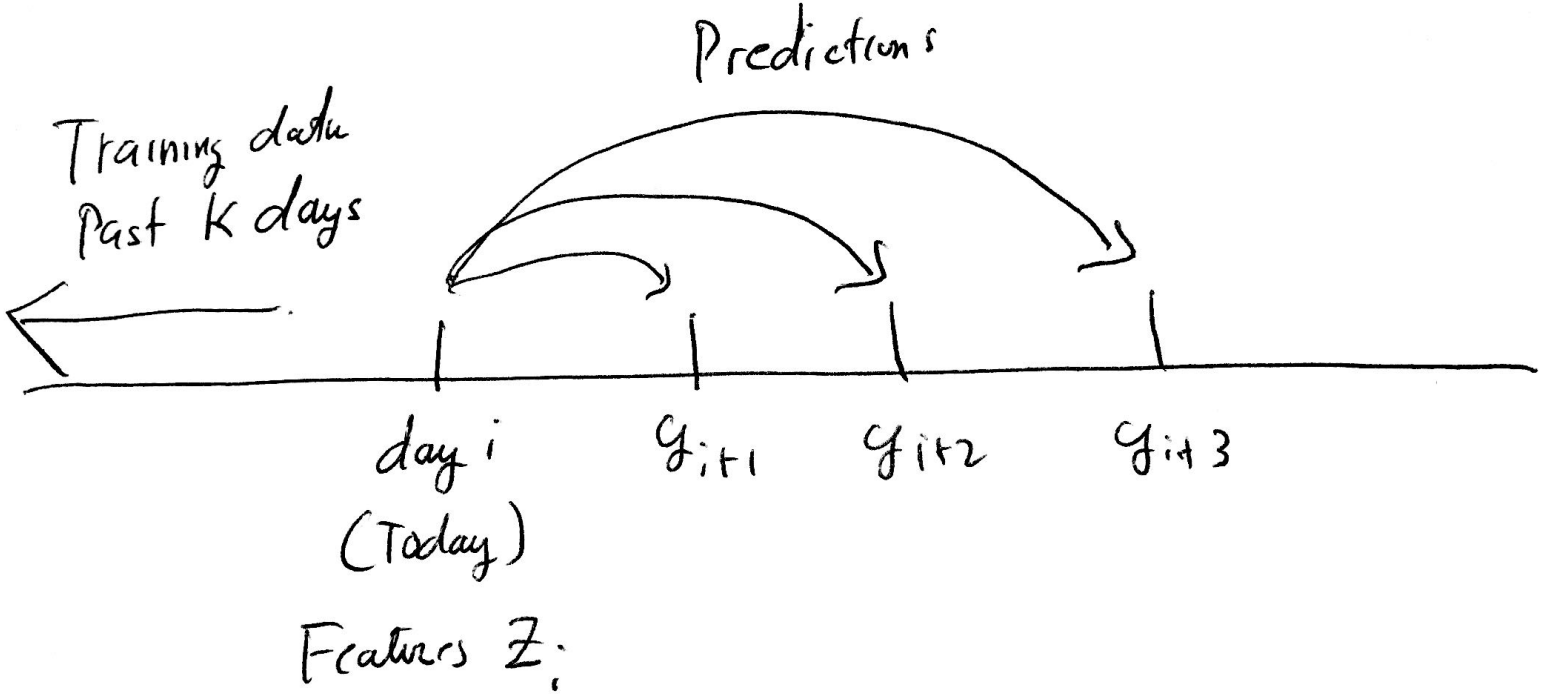
$\Rightarrow$ 61 features in total.

# 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: supervised learning



# 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).
- Starting at day 200, train model. Then use it moving forward, retraining model every month
- We tried training on the prior $k$ days of data. $k$ = all data, or 400, or 150 days.
- 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.
  If $t_i = \hat{y}_i + \hat{y}_{i+1} + \hat{y}_{i+2}$, then amount to order is

$$x_{i+3} = t_i - r_i(1) - r_i(2) - x_{i+1} - x_{i+2}$$

- Works quite well- but (1) choice of buffer is trial and error, and (2) doesn't solve the problem directly (why not?)

# A More direct approach

This approach minimizes the waste directly:

sparsity penalty

linear predictor

total waste

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

where (2)

three days' total need $t_i = z_i^T \beta, \quad \forall i = 1, 2, .., n$ (3)

number to order : $x_{i+3} = t_i - r_i(1) - r_i(2) - x_{i+1} - x_{i+2}$ (4)

waste $w_i = [r_{i-1}(1) - y_i]_+$ (5)

actual remaining $r_i(1) = [r_{i-1}(2) + r_{i-1}(1) - y_i - w_i]_+$ (6)

$r_i(2) = [x_i - [y_i + w_i - r_{i-1}(2) - r_{i-1}(1)]_+]_+$ (7)

Constraint : fresh bags remaining $r_i(2) \geq c_0$ (8)

(9)

No shortage allowed

This can be shown to be a convex problem (LP). We solve it using standard software in R.

# Choice of $\lambda$

We choose $\lambda$ via 8-fold block-wise cross-validation: constraints and targets only involve the remaining 7 folds.
We used the objective function:

$$\sum w_i + 50\{i : r_i < 10\}$$

# Important features selected

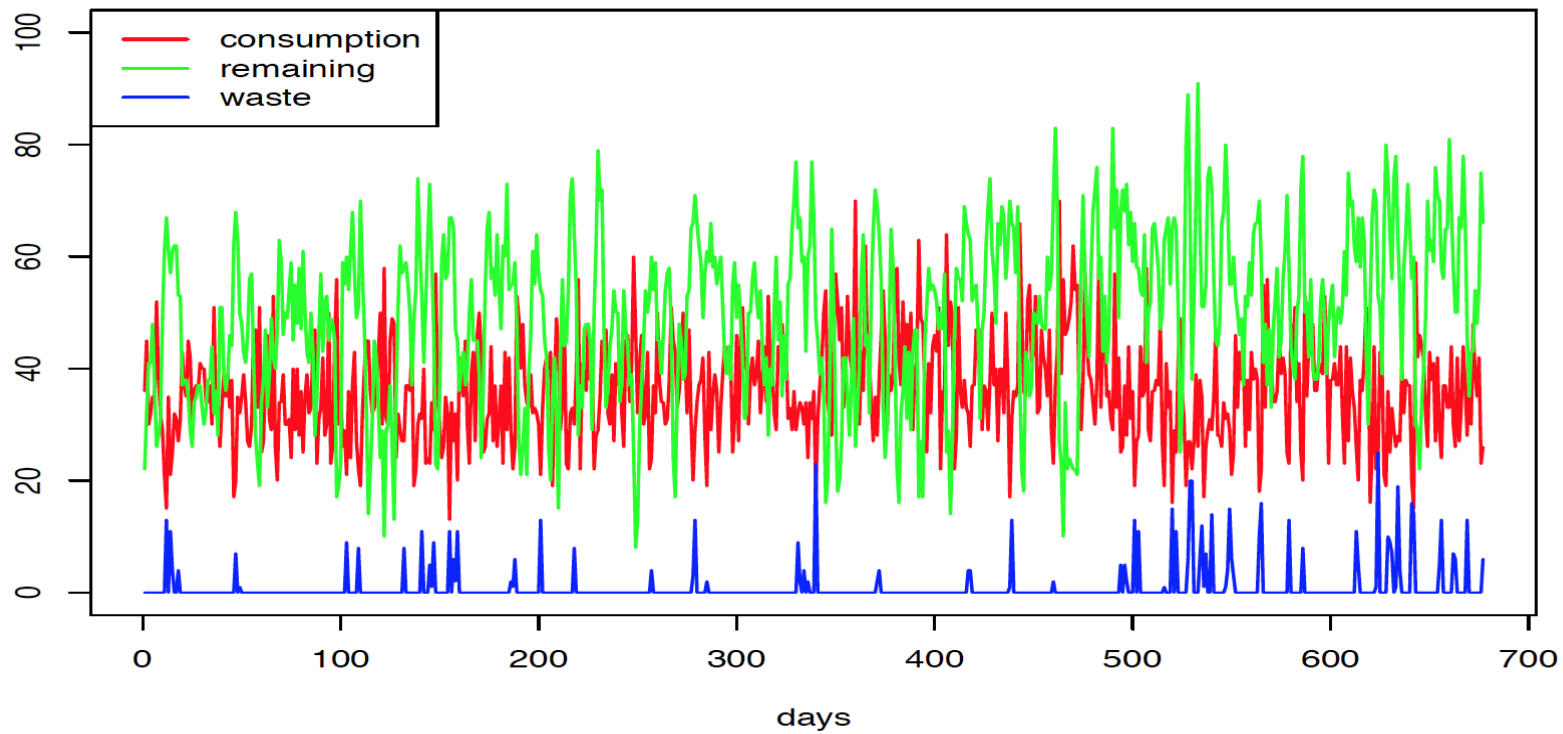
Table : Important features selected

| | |
|---|---|
| PLT transfusion record $\bar{y}_i$ : 5.16 | Day of week:Fri -3, Sun +2 |
| HCT: 1.82 | RDW:+1 |
| MCHC: $-3$ | RBC:$-3$ |
| PLT: $-3.5$ | |
| CENSUS B2: 1 | CENSUS C2: +1.5 |
| CENSUS E3: +2 | CENSUS H1: +6 |
| CENSUS H2: +2 | CENSUS FGR: $-1.5$ |

Others:CENSUS E2.ICU, CATH PACU.....

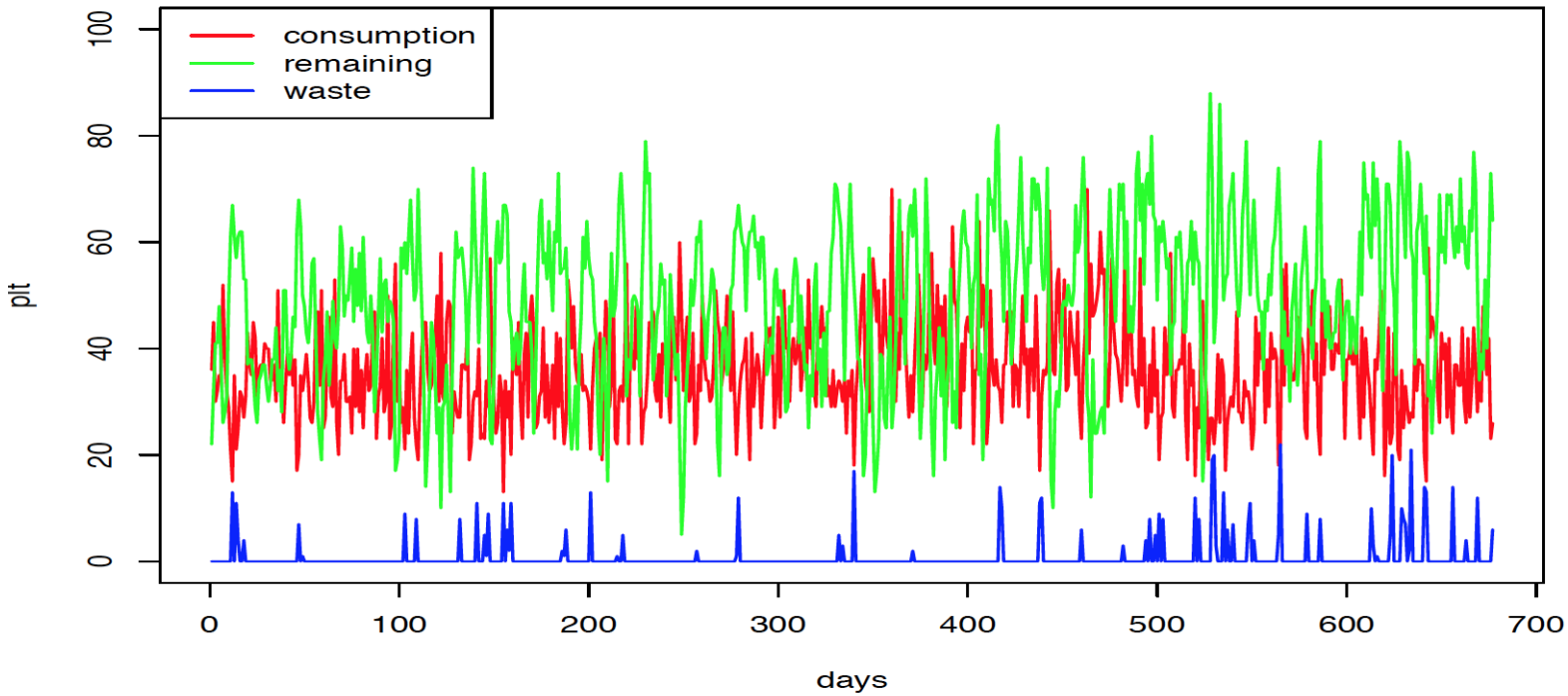
# Results: All data

Using all data points from the past as training data: no shortage, waste 389 bags(2.00%) between 2/08/2013 – 2/08/2015



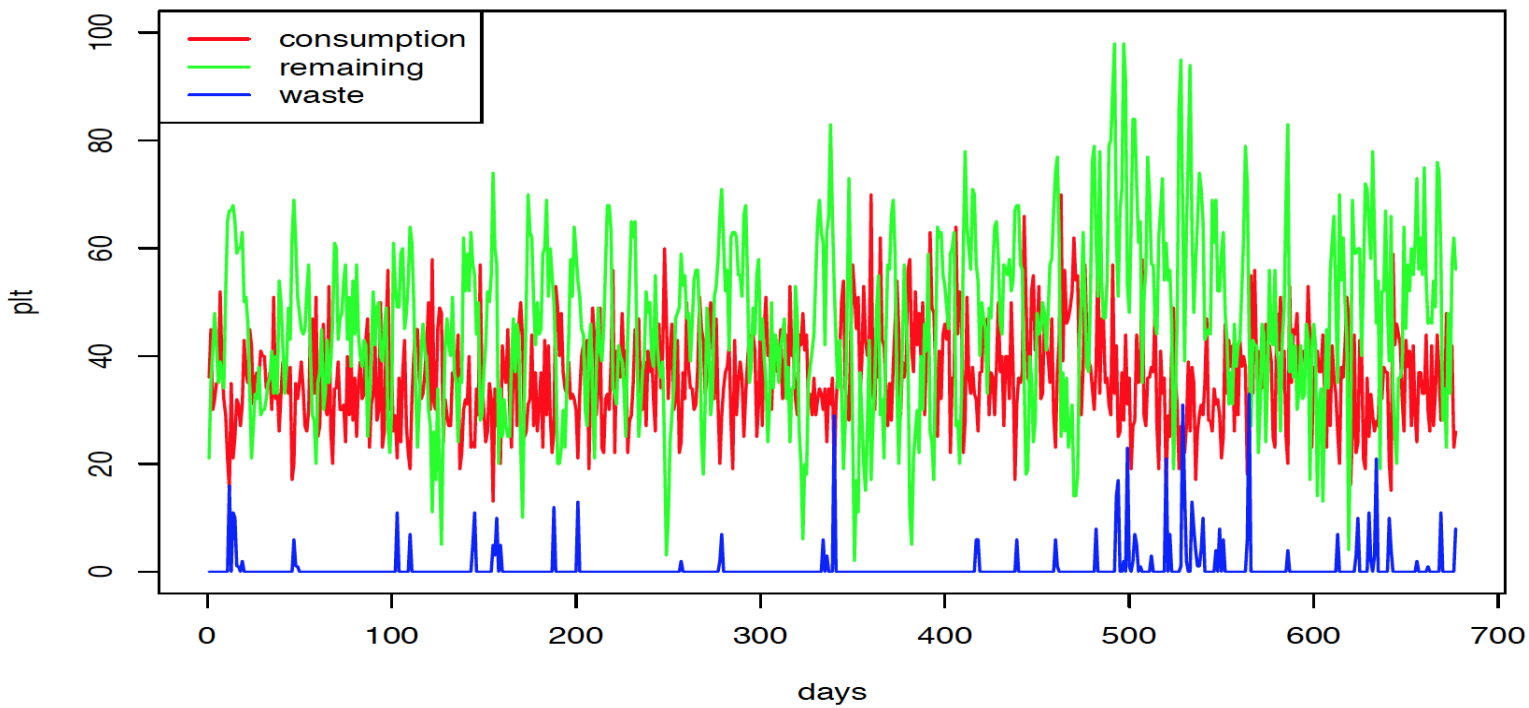
# Time window 400

Using 400 data points from the past as training data: no shortage, waste 359 bags(1.85%) between 2/08/2013 – 2/08/2015

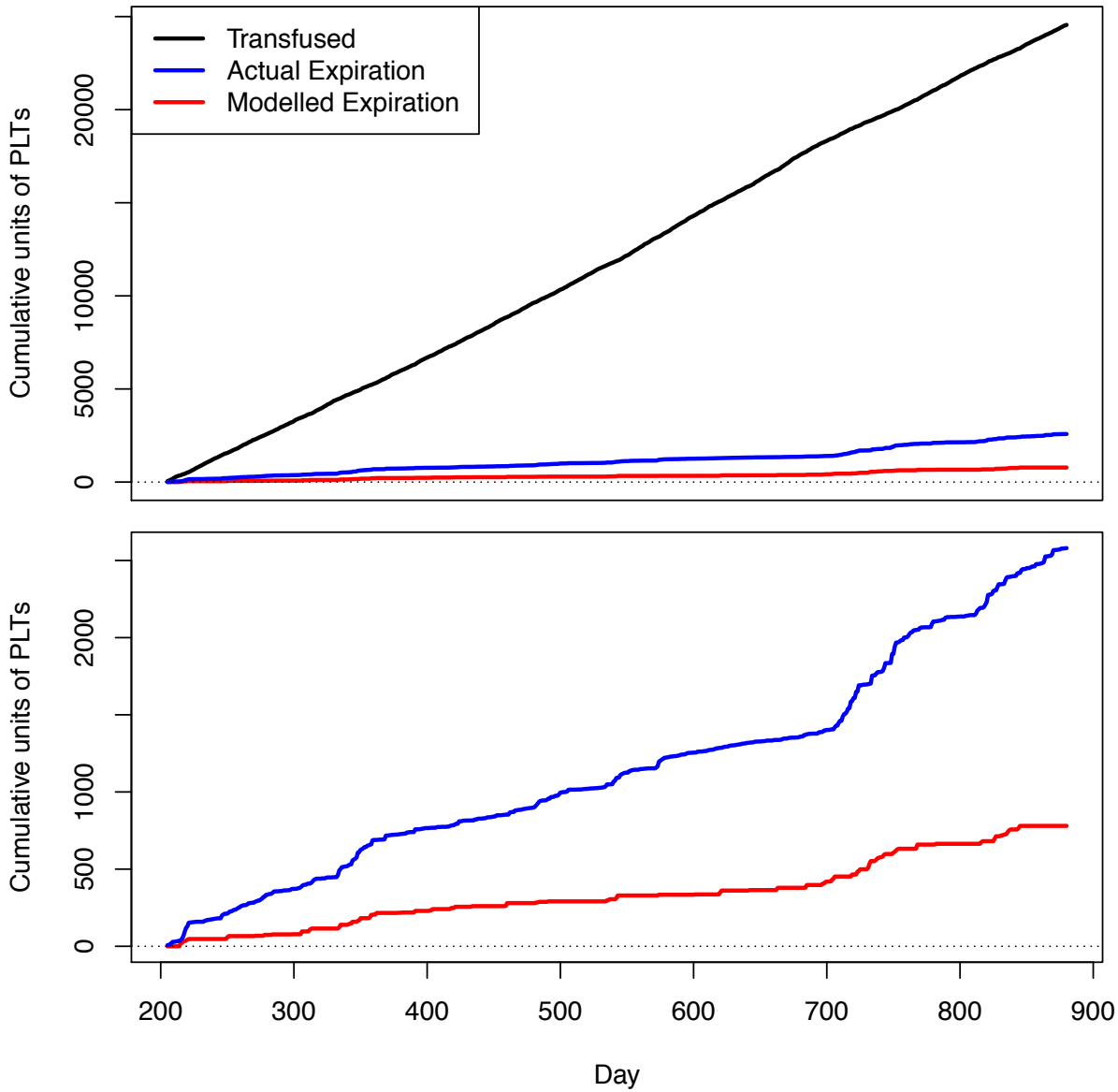


# Time window 150

Using 150 data points from the past as training data: no shortage, waste 383 bags(1.97%) between 2/08/2013 – 2/08/2015



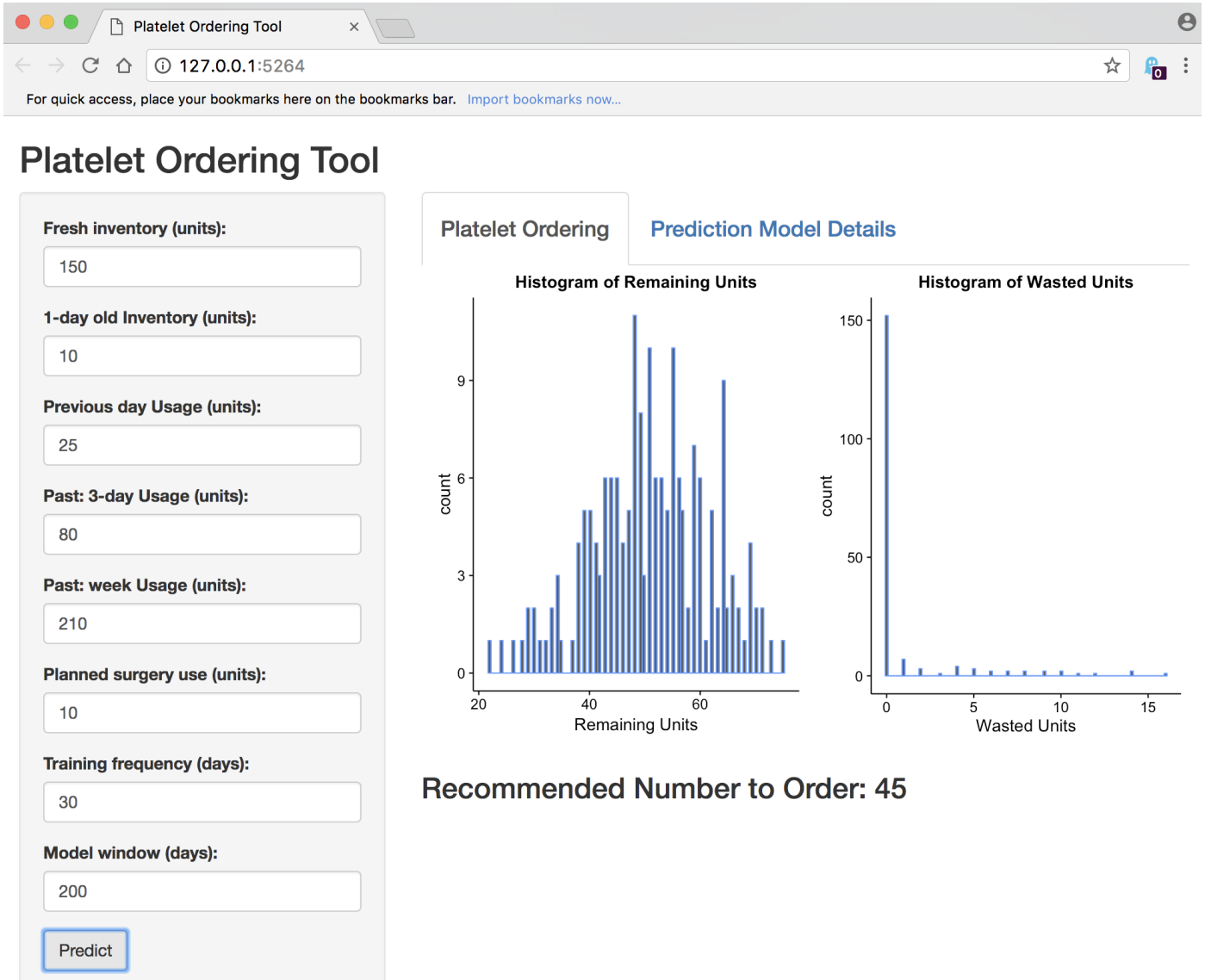
# Results


Transfused
Actual Expiration
Modelled Expiration
Cumulative units of PLTs
0
5000
10000
20000
Cumulative units of PLTs
0
500
1000
2000
200
300
400
500
600
700
800
900
Day

# Summary

- Reducing wastage from 8% to 2% corresponds to a predicted direct savings at Stanford of $350,000/year.
- If implemented nationally could result in approximately $110 million in savings.
- Just published in PNAS
- Future work: deploy!

# Mockup of R Shiny App (Thanks to Naras)



A plug for:

**CVXR**– disciplined convex programming in R

Stephen Boyd, Anqii Fu, Balasubramanian Narashiman

available on CRAN

# For further reading

The methods in this lecture are described in detail in our books on Statistical Learning: