

Big Data and Machine Learning

David Strömberg

Agenda

- Overview of recent methods in Machine Learning and their use in economics.
- No question on exam or problem set.
- More detailed coverage in Applied Economics I in fall term.

Where to learn ML

- An Introduction to Statistical Learning with Applications in R
 - Gareth, Witten, Hastie, Tibshirani. Less technically advanced.
Each chapter ends with an R lab, in which examples are developed.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction
 - More technical and covers a broader range of topics.
- Above books available free from authors' websites.
- Online course:
<https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about>

Outline

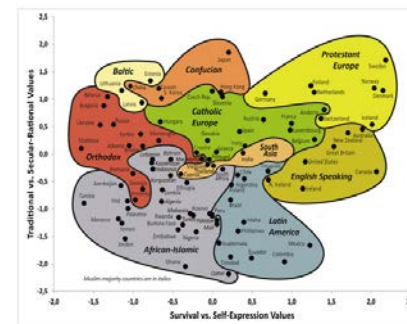
- ML
 - Unsupervised – organization
 - Supervised - prediction
- New Data
 - Satellite
 - Text
- ML-predictions in policy.

Unsupervised learning

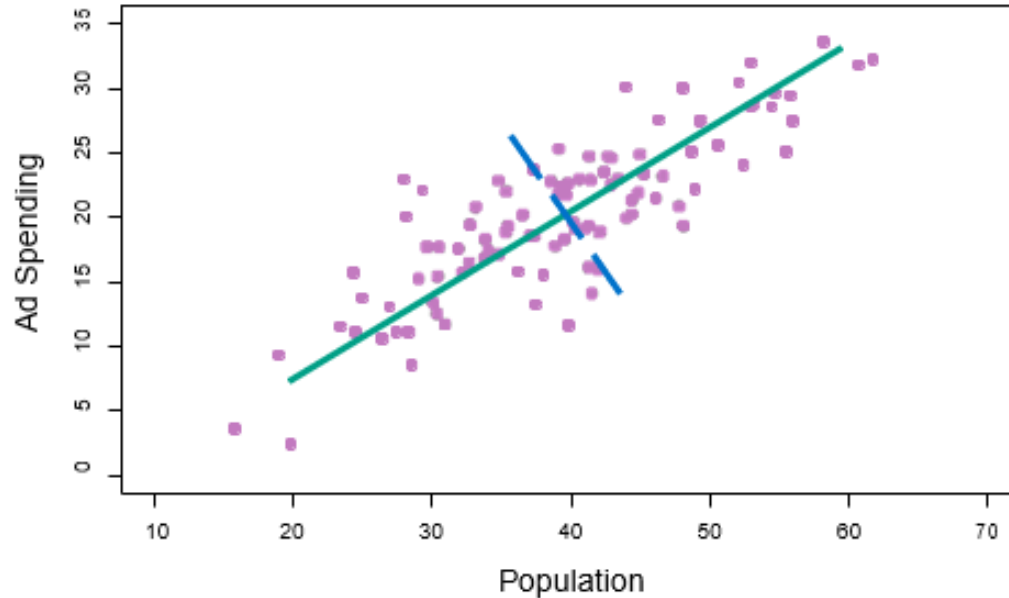
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy— find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
 - Difficult to know how well you are doing.
- Can be useful as a pre-processing step for supervised learning.
- Discussed here: PCA and clustering

Principal Components Analysis (PCA)

- PCA produces a low-dimensional representation of a dataset.
 - It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Produces derived variables for use in supervised learning problems
 - Example: face recognition.
- Serves as a tool for data visualization.
 - Example: World Values Survey



PCA: example

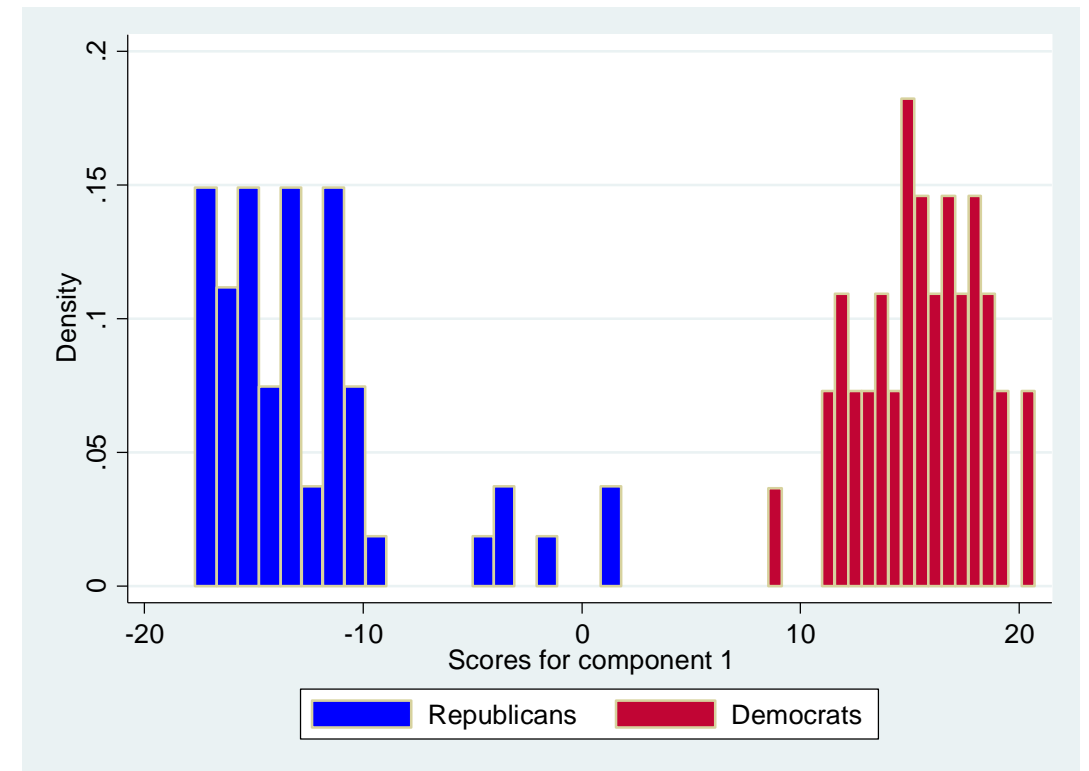


The population size (**pop**) and ad spending (**ad**) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

More dimensions: Voting in the US Senate

- 105th Senate: PCA on 612 roll call votes by each of 100 Senators.

lstate	party	name	V1	V519
ALABAMA	200	SESSIONS	1	1
ALABAMA	200	SHELBY	1	1
ALASKA	200	MURKOWSKI	1	1
ALASKA	200	STEVENS	1	1
ARIZONA	200	KYL	1	1
ARIZONA	200	MCCAIN	1	1
ARKANSAS	100	BUMPERS	1	6
ARKANSAS	200	HUTCHINSON,	1	1
CALIFOR	100	BOXER	1	6
CALIFOR	100	FEINSTEIN	1	6

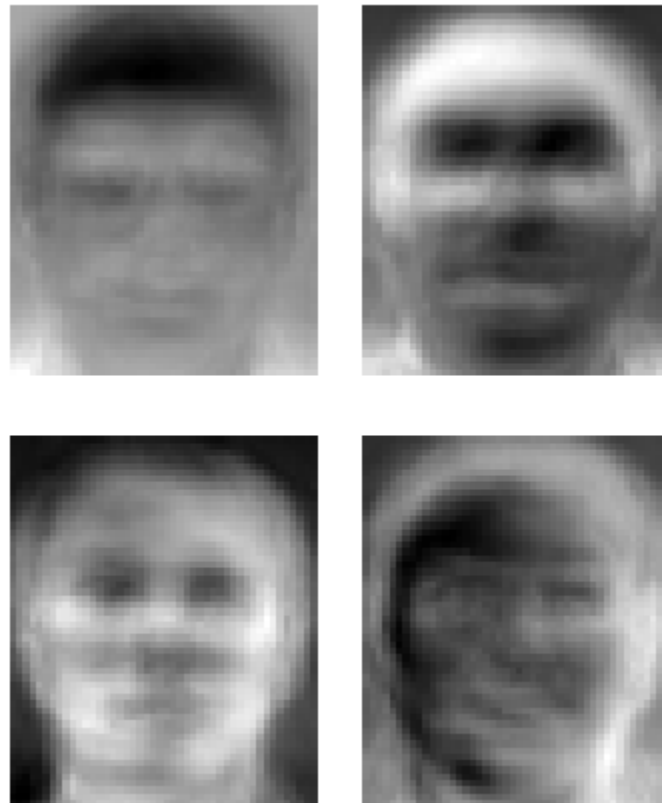


Even more dimensions: Face recognition

- Large set of digitized images of human faces is taken under the same lighting conditions.
- The images are normalized to line up the eyes and mouths.
- The eigenvectors of the covariance matrix of the statistical distribution of face image vectors are then extracted.
- These eigenvectors are called eigenfaces.

Eigenfaces

- The principal eigenface looks like a bland androgynous average human face



PCA 3: Face recognition



`coefficients =`

```
{-6.85693, 23.7498, -11.4515, -3.43352, 5.24749, -7.1615,  
8.09015, -9.7205, -0.660834, -2.4148, -10.3942, 3.33424,  
2.94988, -2.75981, 3.02687, -2.4499, -2.09885, -5.98832,  
-4.22564, -0.65014, 2.20144, -5.43782, -9.61821, -3.25227,  
7.49413, -0.145002, 7.61483, -0.696994, -3.7731, 3.23569,  
-1.78853, 0.0400116, -3.86804, -2.02456, 2.20949, -1.86902,  
1.23445, 0.140996, 0.698304, -0.420466, 2.30691, 3.70434,  
1.02417, 0.382809, 0.413049, -0.994902, 0.754145, 0.363418,  
-0.383865, 1.46379, 1.96381, -2.90388, -2.33381, -0.438939,  
-0.30523, -0.105925, 0.665962, -0.729409, -1.28977, 0.150497,  
0.645343, 0.30724, -1.04942, 1.0462, -0.60808, 0.333288,  
1.09659, -1.38876, 0.33875, 0.278604, 1.0632, -0.0446148,  
0.24526, -0.283482, -0.236843, 0.312122};
```

Clustering

- *Clustering* refers to a very broad set of techniques for finding *subgroups*, or *clusters*, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other.
- PCA vs Clustering
 - PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
 - Clustering looks for homogeneous subgroups among the observations.

Clustering algorithms are simple

- Randomly assign a number, from 1 to K , to each of the observations.
- Calculate cluster centroid.
- Allocate each obs to closest centroid.
- Iterate.

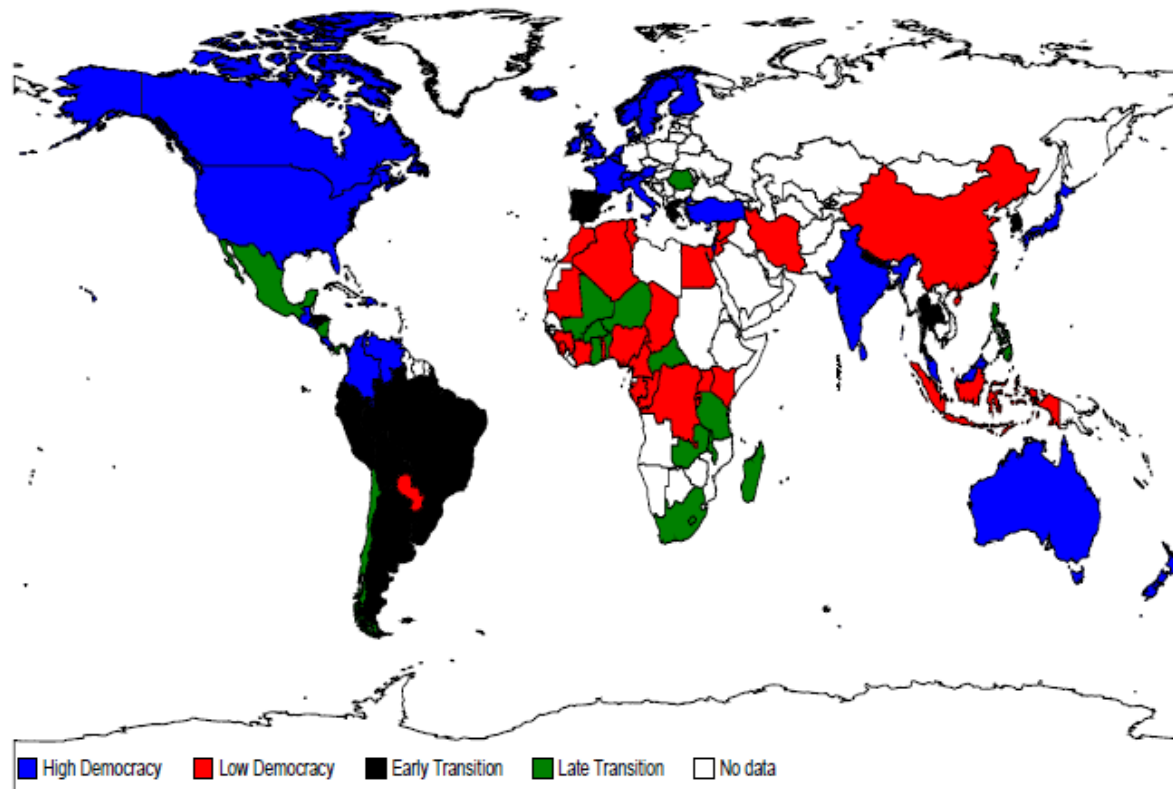


Example: K-means clustering: specification design for causal estimates

- Many studies in economics include group-fixed effects:
how should these groups be defined?
- Bonhomme and Manresa (EMA, 2015)
 - Select group structure to minimize squared residual with respect to all possible groupings of the cross-sectional units.

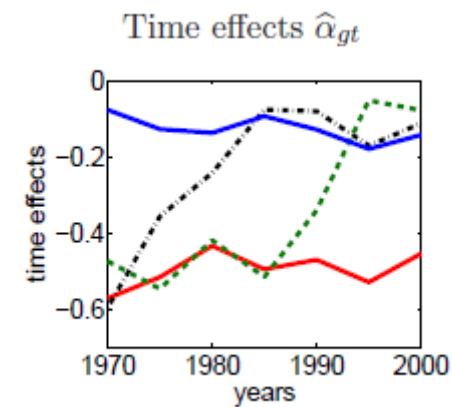
Example: Bonhomme and Manresa (2015)

Figure 2: Patterns of heterogeneity, $G = 4$



- Example: "Income and Democracy", Acemoglu et al. (AER, 2008)

$$democracy_{it} = \theta_1 democracy_{it-1} + \theta_2 \log GDP_{pcit-1} + \alpha_{gt} + \nu_{it}.$$

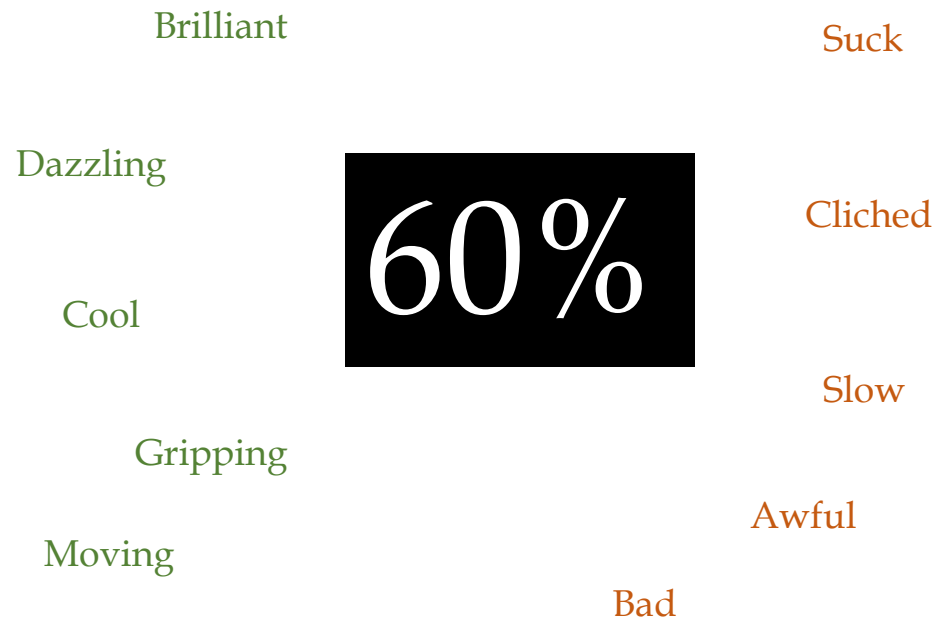


Supervised learning

- Economics:
 - $\hat{\beta}$: estimate one/few coefficients of interest (causal effect)
 - Use one main specification (linear), show robustness to alternative specification and placebo tests.
 - Model is evaluated on in-sample-properties (e.g. R²).
- Supervised learning: \hat{y}
 - \hat{y} : predict outcome.
 - Use data-driven model selection.
 - Model is evaluated out-of-sample (e.g. cross validation).
 - When model produce $\hat{\beta}$ estimates (e.g. Lasso), these are typically not consistent.

Supervised learning: how does it work?

- Movie reviews (Pan, Lee, Vaithyanathan)
 - Intelligent methods, trying to copy humans, perform poorly.

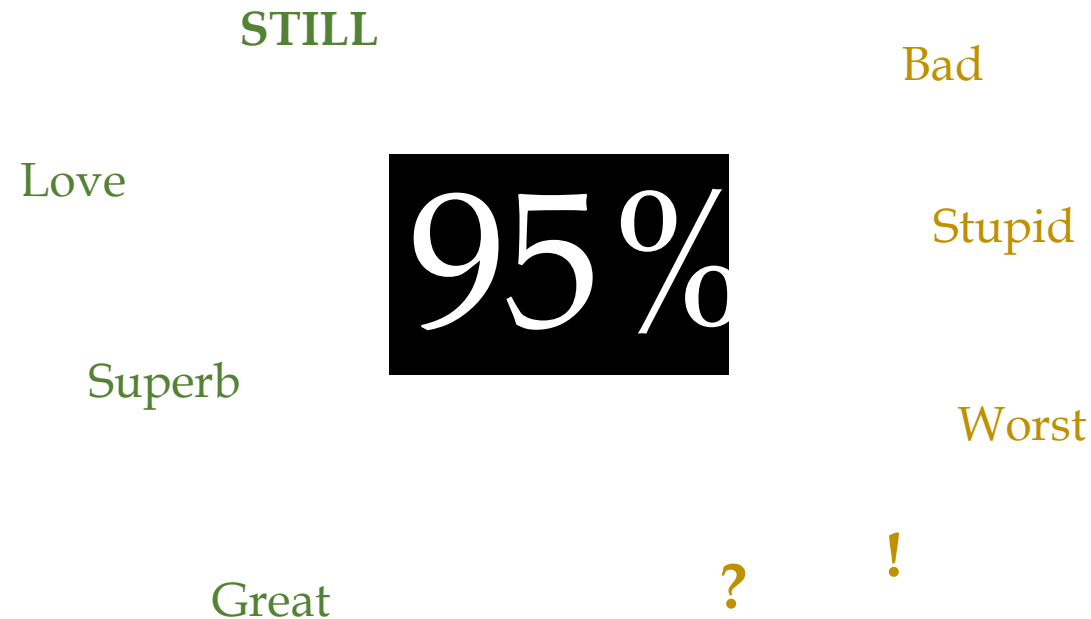


What is so hard?

- Decide what words makes for a positive review
 - What combination of words do you look for?
 - “Some people say this talk was great”
- This problem was endemic to every problem
 - Driving a car: What is a tree?
 - Language: Which noun does this pronoun modify?
- This approach stalled
 - “Trivial” problems proved impossible
 - Forget about the more complicated problems like language

ML Approach

- Collect example dataset:
 - 2000 movie reviews
 - 1000 good and 1000 bad reviews
- Now just ask what combination of words predicts being a good review.



What is the magic trick?

- Stop trying to “figure it out”
- Stop looking for an insight
- Just look at the data
- Treat the known like we treat the unknown

Widespread use

- Post office uses machines to read addresses
- Voice recognition (Siri)
- Spam filters
- Movie and other recommendations

- String together many smaller tasks
 - Driverless cars

- *Empirical* intelligence

The problem

- Find function $f(x) = \hat{y}$ to minimize $L(y, \hat{y})$ in new data.
 - Loss function: $L(y, \hat{y})$, typically MSE of prediction error.
- Two components
 - Choose function f class: regression tree, linear, etc.
 - Select regularizer: bias-variance trade-off.

Examples

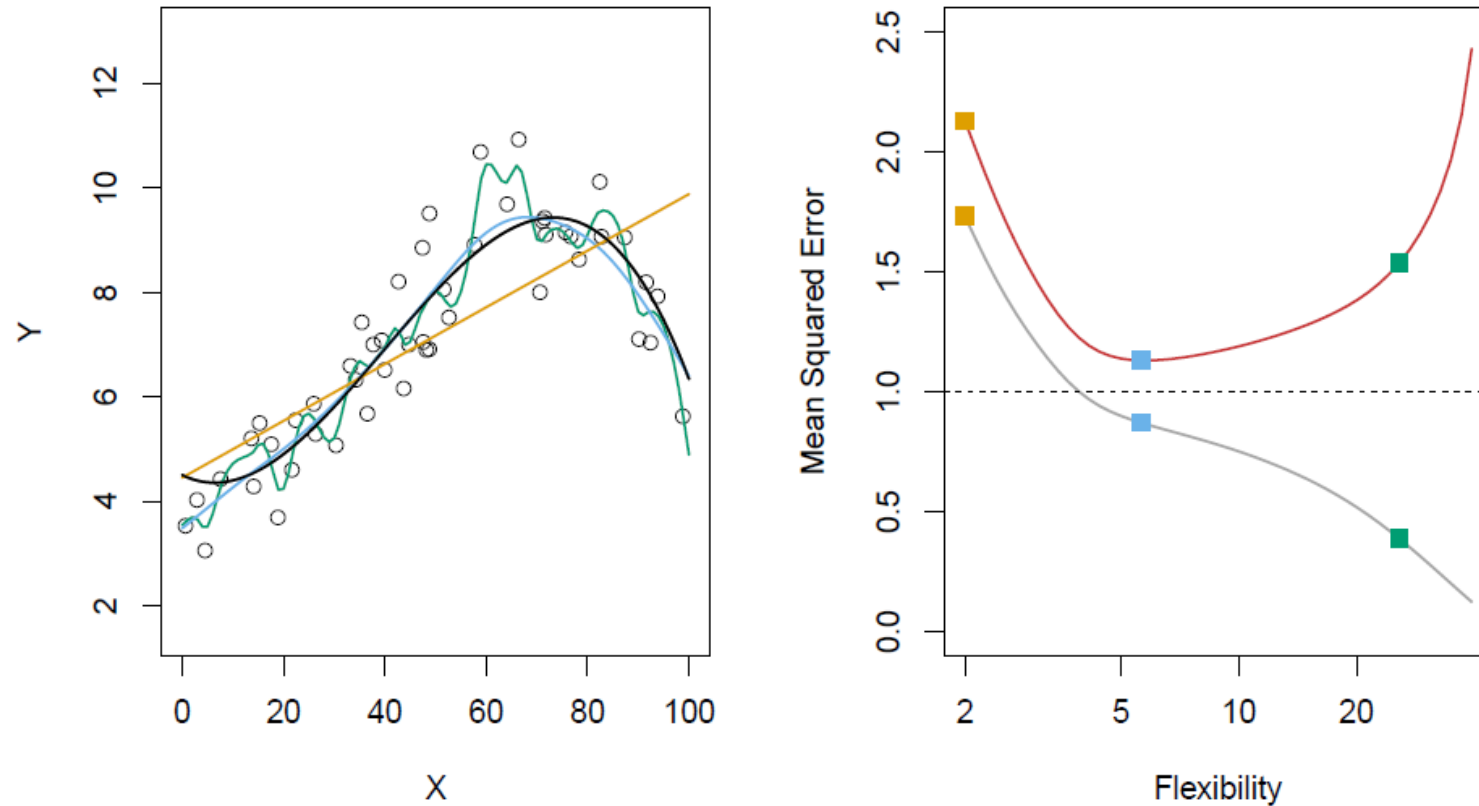
Table 2
Some Machine Learning Algorithms

<i>Function class \mathcal{F} (and its parametrization)</i>	<i>Regularizer $R(f)$</i>
Global/parametric predictors	
Linear $\beta'x$ (and generalizations)	Subset selection $\ \beta\ _0 = \sum_{j=1}^k \mathbf{1}_{\beta_j \neq 0}$ LASSO $\ \beta\ _1 = \sum_{j=1}^k \beta_j $ Ridge $\ \beta\ _2^2 = \sum_{j=1}^k \beta_j^2$ Elastic net $\alpha \ \beta\ _1 + (1 - \alpha) \ \beta\ _2^2$
Local/nonparametric predictors	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth
Mixed predictors	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
Combined predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)

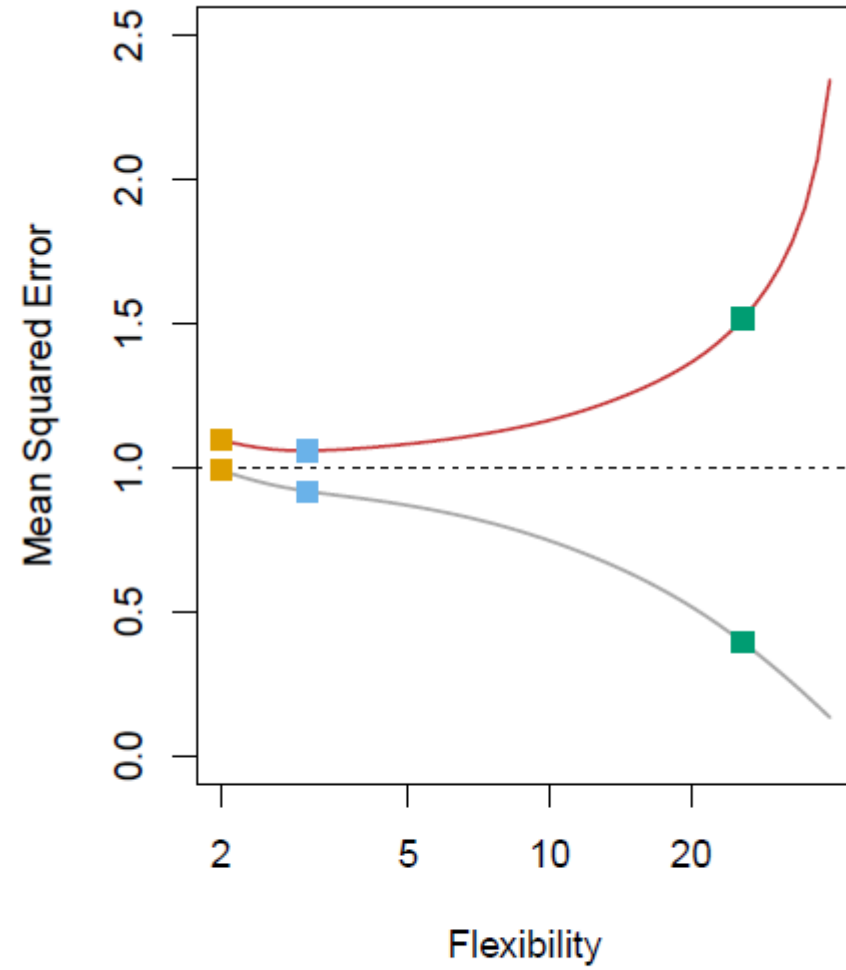
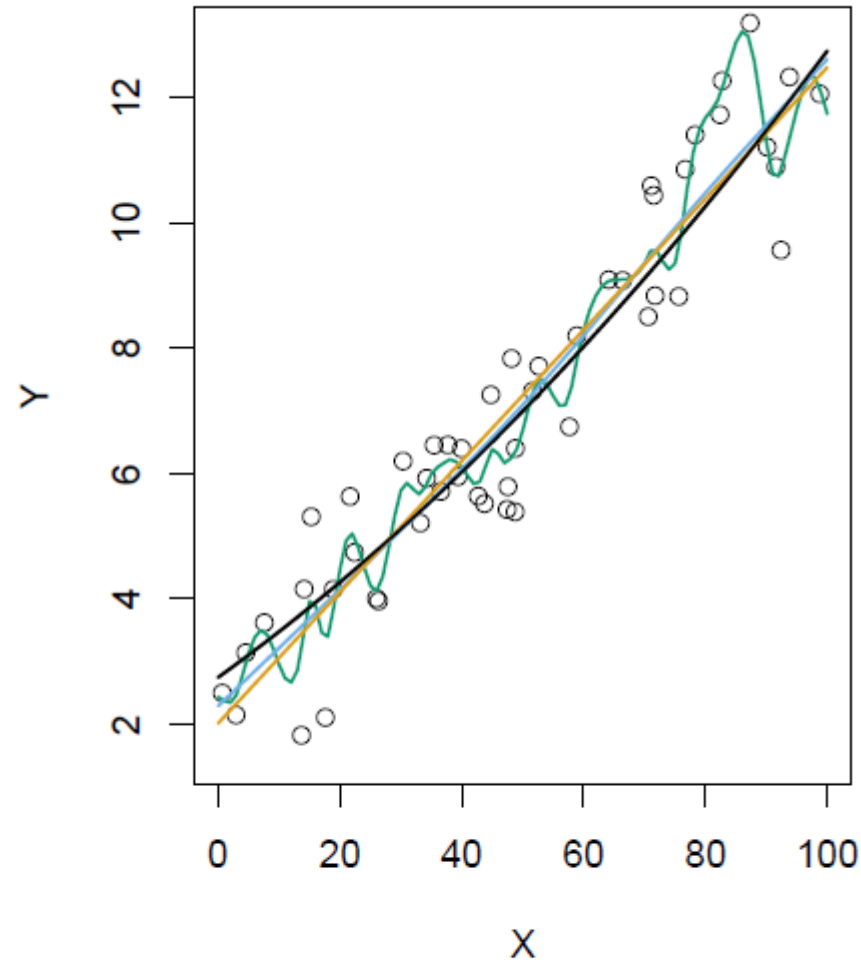
Overfit problem

- OLS looks good with the sample you have
 - It's the best you can do *in this sample*
- Problem is OLS by construction overfits
 - We overfit in estimation
 - But in low-dimensional this is not a major problem
- In wide data, the problem is huge
 - We'll fit very well (perfectly if $k > n$) **in sample**
 - But arbitrarily badly **out of sample**

Regularization: choice of function flexibility



Black curve is truth. Red curve on right is MSE_{Te} , grey curve is MSE_{Tr} . Orange, blue and green curves/squares correspond to fits of different flexibility.



Here the truth is smoother, so the smoother fit and linear model do really well.

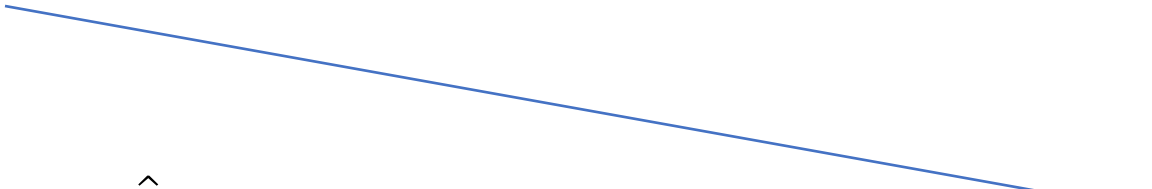
Restricting flexibility

Unconstrained

$$\hat{f}_A = \arg \min_{f \in \mathcal{F}_A} \mathbb{E}_H L(f(x), y)$$

Constrained:

$$\begin{aligned} \arg \min_{f \in \mathcal{F}} \mathbb{E}_H L(f(x), y) \\ \text{s.t. } R(f) \leq c \end{aligned}$$


$$\hat{f}_{A_\lambda} = \arg \min_{f \in \mathcal{F}_A} \mathbb{E}_H L(f(x), y) + \lambda R(f)$$

Tuning parameter

Tuning parameter sets price on flexibility


$$\hat{f}_{A_\lambda} = \arg \min_{f \in \mathcal{F}_\mathcal{A}} \mathbb{E}_H L(f(x), y) + \lambda R(f)$$

Regularizers for linear functions

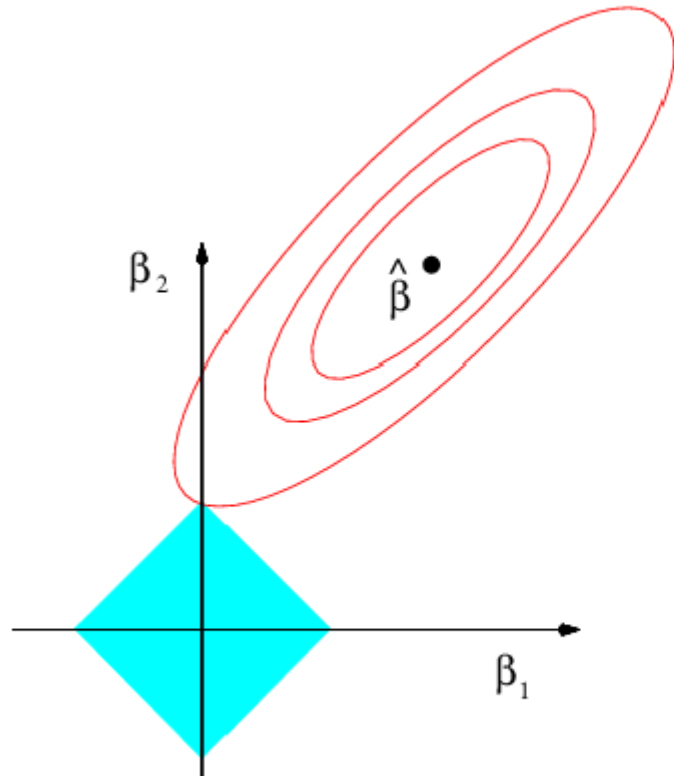
Lasso

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

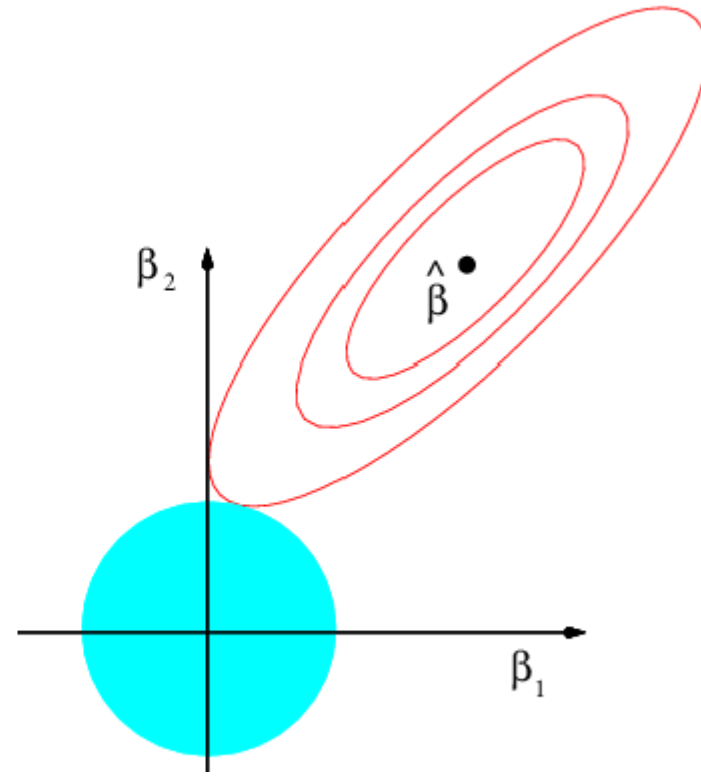
Ridge

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Lasso selects subset of regressors with positive weights



$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

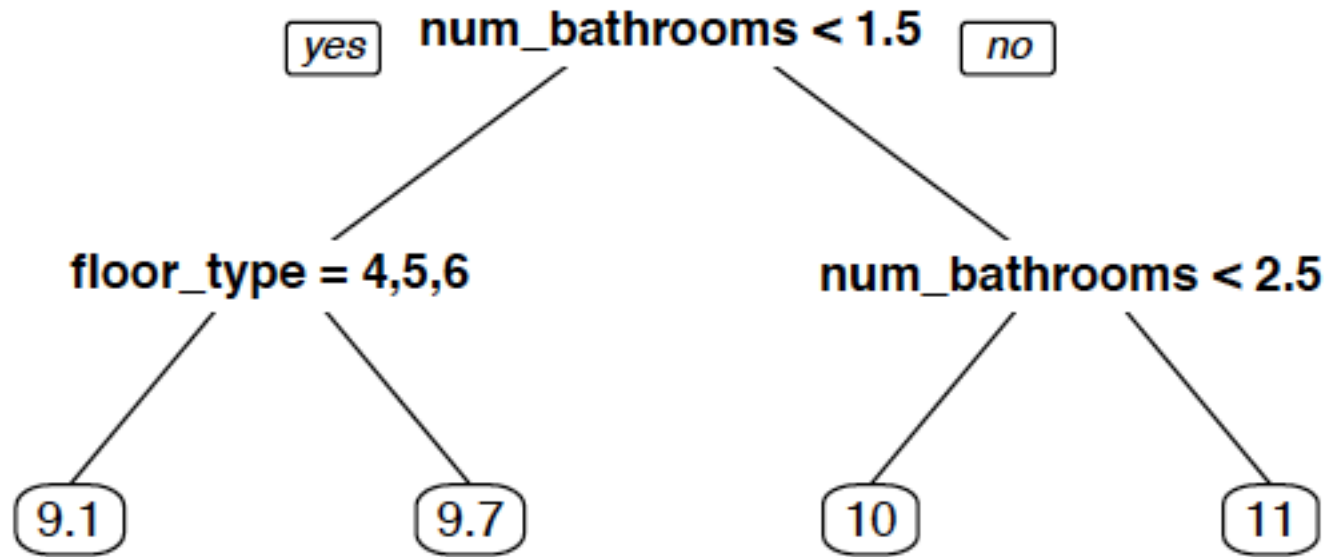


$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

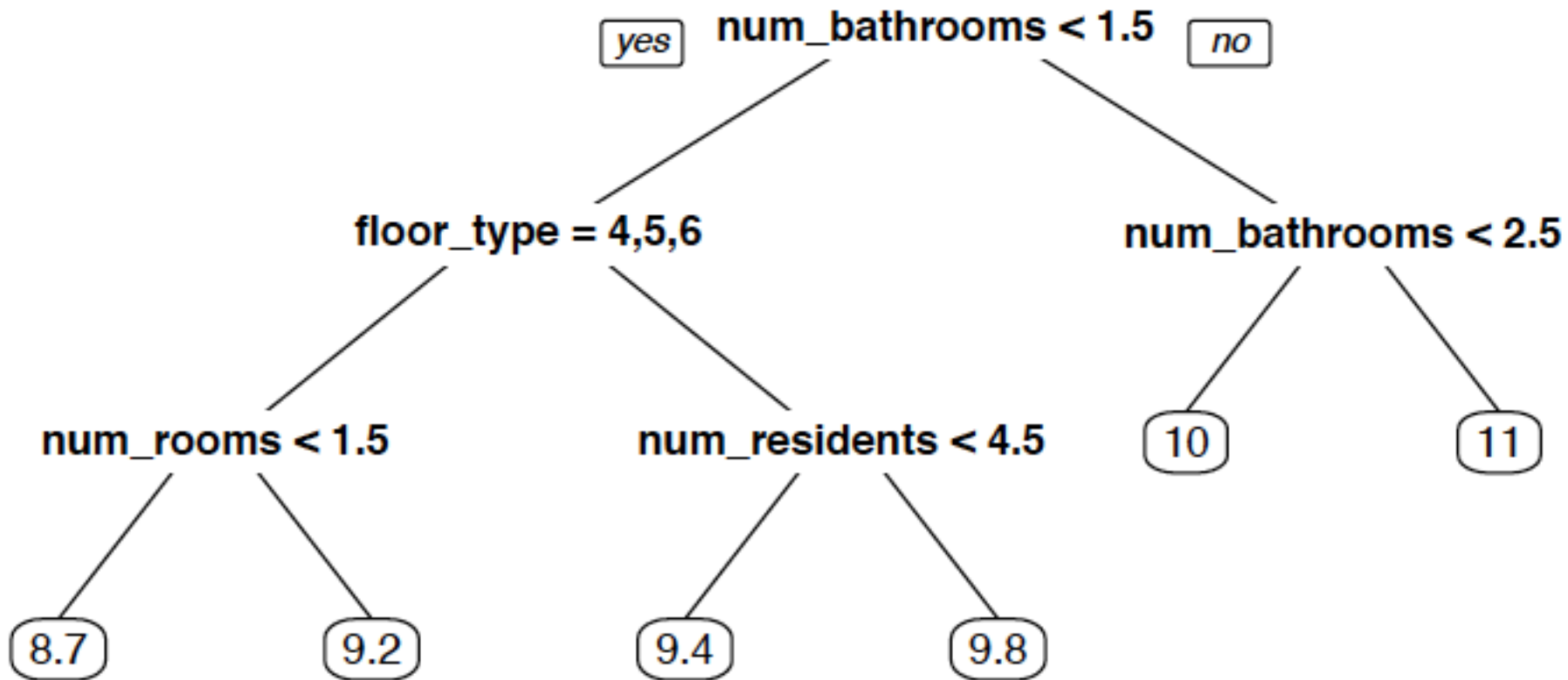
Decision tree example: housing value



Decision tree example: housing value

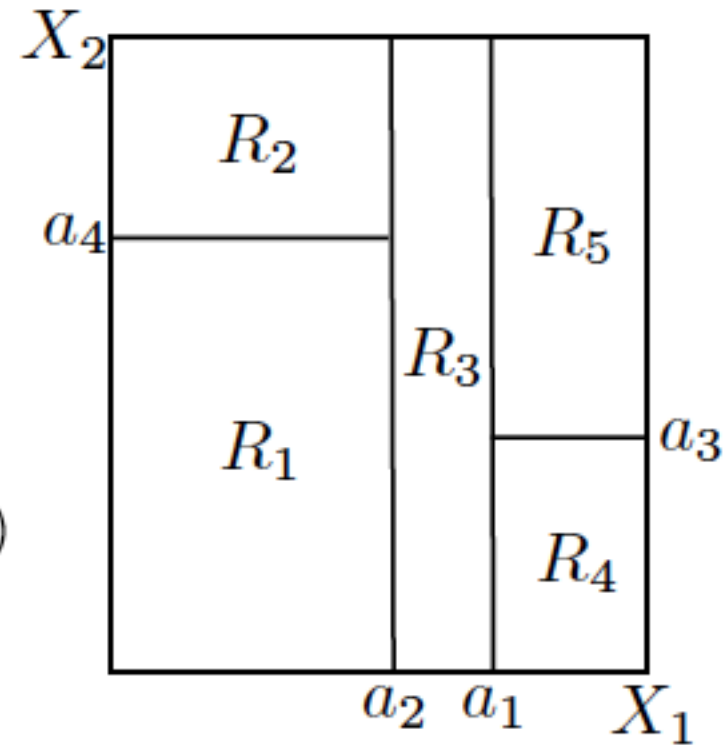
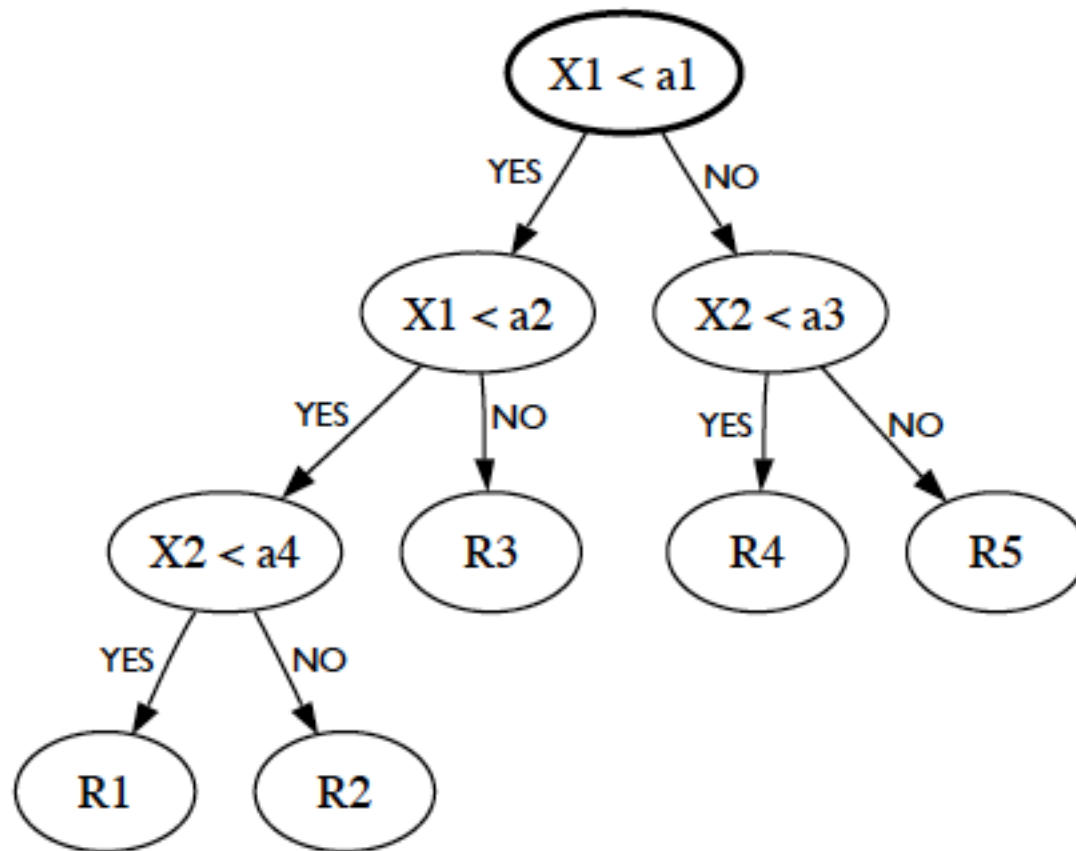


Decision tree example: housing value



Another way to understand trees

- Partitioning predictors -- allowing for non-linearities



Fitting

- Suppose we fit the best tree we could do to some dataset
- What would we get?
- Regularization parameters
 - Max number of nodes/leaves/depth.
 - Min information gain at split

Model assessment and selection

Model selection:

- Estimate performance of models to choose best one.

Model assessment:

- Given the final chosen model, estimate prediction error on new data.

Best approach in data rich environment:
randomly split data three ways.



Less data: approx. validation step by analytic formula (e.g. AIC)
or re-use sample (cross-validation and bootstrap).

K -fold cross validation

- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k , fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out k th part.
- This is done in turn for each part $k = 1, 2, \dots, K$, and then the results are combined.

Validation	Train	Train	Train	Train
------------	-------	-------	-------	-------

Creating Out-of-Sample In Sample

- Major point:
 - Not many assumptions
 - Don't need to know true model.
 - Don't need to know much about algorithm
- We use the data itself to choose complexity

What separates prediction from estimation

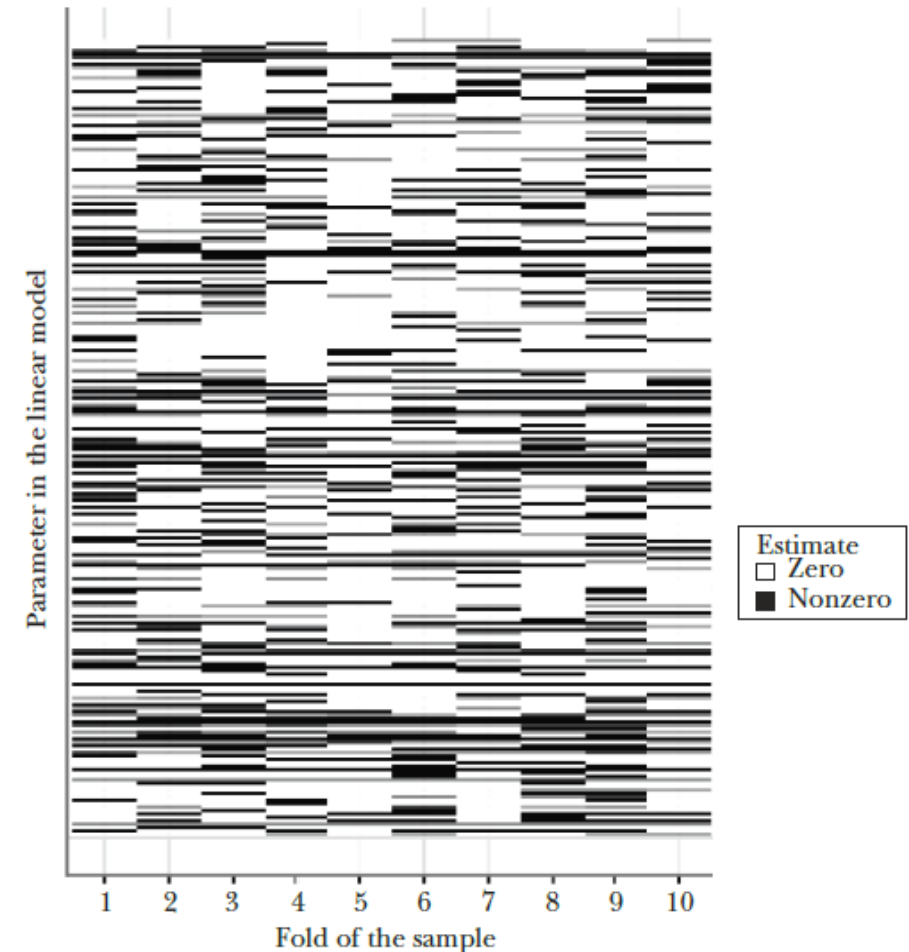
- Validity of predictions are measurable:
prediction quality is observable
- Validity of coefficient estimators require structural knowledge.
 - Whether the world is truly drawn from a linear model.
Not observable.

What is ML not good for?

- Learning about the underlying model.
 - LASSO produces a sparse prediction function: many coefficients are zero.
 - This selection is sensitive to multicollinearity. The prediction is stable but not the choice of included regressors.

Figure 2

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



ML in causal inference

Active research area

- Selecting controls or control function

$$y = \alpha D + f(x, \beta) + \varepsilon$$

- Post double selection: Select control variables in OLS (Belloni et al, 2014)
- Selection among many instruments in first stage in IV (Belloni et al, 2013, 2015)
- Non-linear (e.g. random forest) control functions (Chernozhukov et al., 2017)

- Heterogenous effects

$$y = \alpha_g D + f(x, \beta) + \varepsilon$$

- Athey, Imbens (2016)
- Chernozhukov et al. (2018)

Post double selection (Belloni et al., 2014)

Determine which controls not to exclude from regression

$$y = \alpha D + \mathbf{x}'\beta + e$$

1. Estimate Lasso to determine which variables can be dropped from the standpoint of predicting y and key x -variable.
 - Standard cross-validation not optimal for setting the tuning parameter when prediction is not end goal.
2. Estimate standard OLS included selected controls.
 - Lasso coefficients from step 1 are biased.
3. Reported standard errors are consistent.
 - Tuning parameter set to avoid over fitting.

Post double selection (Belloni et al., 2014)

- Strong key assumption:
underlying model is **sparse** (few relevant variables).
 - **If** there is this structure then Lasso will find it **as n goes to infinity**.
 - Consistent estimation of β in (not just α) also requires that no irrelevant variable is too correlated with the relevant variables.
- Use to avoid omitting key controls (cherry-picking).
Not to find exclusive set of controls.

Instrument selection

IV-regression x \hat{x}

$$y = x'\beta + \varepsilon$$

$$x = z'\gamma + \omega$$

- First-stage is a prediction problem: predict \hat{x} and plug into main eq.
- Finite sample bias due to over-fitting: \hat{x} biased towards x .
 - Small sample size, large number of weak instruments (Bound et al 1995, Bekker 1994, Stager and Stock 1997).
- Avoid overfitting
 - Split sample IV (Angrist and Krueger 1995)
 - Use Lasso and regularization to select instruments (e.g. Belloni et al., 2013, 2015)

Applications of Machine Learning

- New Data
 - Satellite images
 - Use ML to extract meaningful information: crop yield, economic activity from luminosity, man-made structures, pollution, house values, etc.
 - Text
 - Stock market mood from bulletin boards, firm evaluations from financial information, job postings, policy documents, etc.
 - Cell-phone use, digital economy: Google searches, online consumption data, etc.
- Prediction as input to policy



Poverty mapping in Uganda (Xie et al 2016)

- Predict night-time lights from daytime imagery.
- Identifies different terrains and man-made structures, including roads, buildings, and farmlands, without any supervision beyond nighttime lights.
- Poverty mapping approaching the predictive performance of survey data collected in the field.

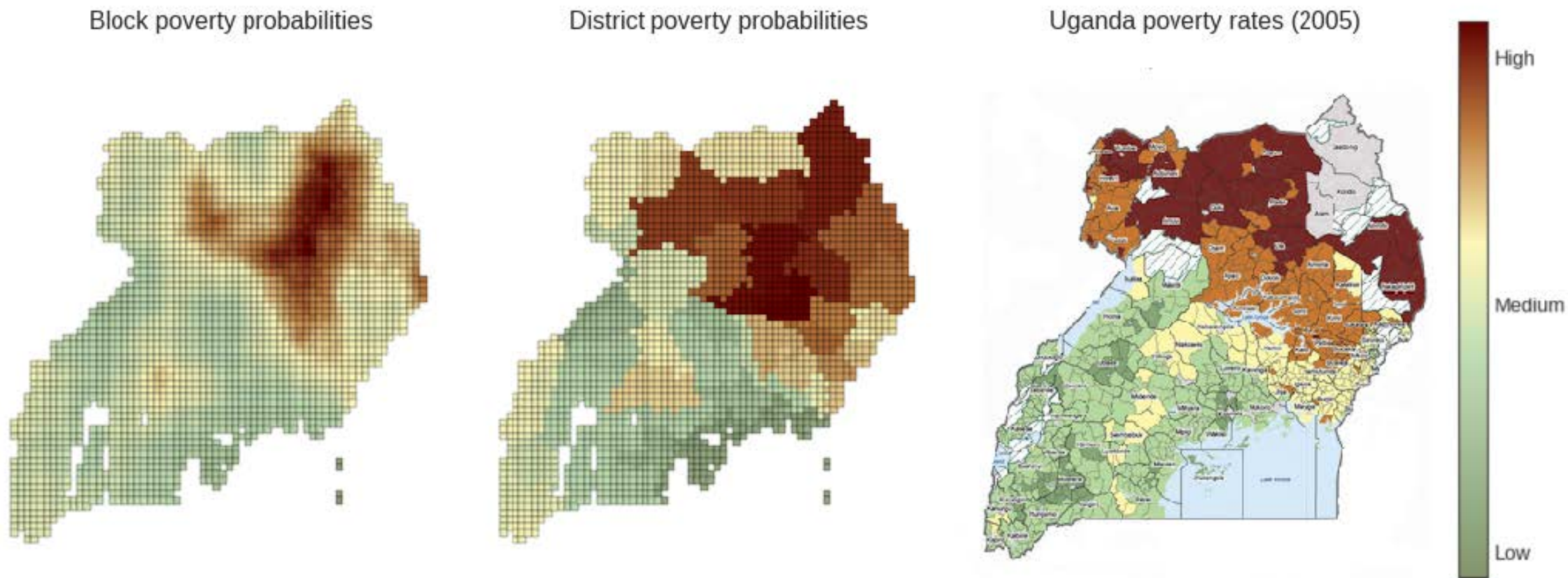


Figure 3: **Left:** Predicted poverty probabilities at a fine-grained 10km \times 10km block level. **Middle:** Predicted poverty probabilities aggregated at the district-level. **Right:** 2005 survey results for comparison (World Resources Institute 2009).

Predicting poverty and wealth from mobile phone data (Blumenstock et al 2015)

- Anonymized records of billions of mobile phone interactions.
- Follow-up phone surveys geographically stratified random sample of 856 individual subscribers.
- Predict individuals' socioeconomic status.
- Predicted attributes of millions of individuals
 - reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households.

Table 1. Summary statistics for primary data sets. Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic and Health Survey (DHS) data were collected by the Rwandan National Institute of Statistics. N/A, not applicable.

Summary statistic	Phone survey	Call detail records	DHS (2007)	DHS (2010)
Number of unique individuals	856	1.5 million	7377	12,792
Data collection period	July 2009	May 2008–May 2009	Dec. 2007–Apr. 2008	Sept. 2010–Mar. 2011
Number of questions in survey	75	N/A	1615	3396
Primary geographic units	30 districts	30 districts	30 districts	30 districts
Secondary geographic units	300 cell towers	300 cell towers	247 clusters	492 clusters

Predicting survey responses

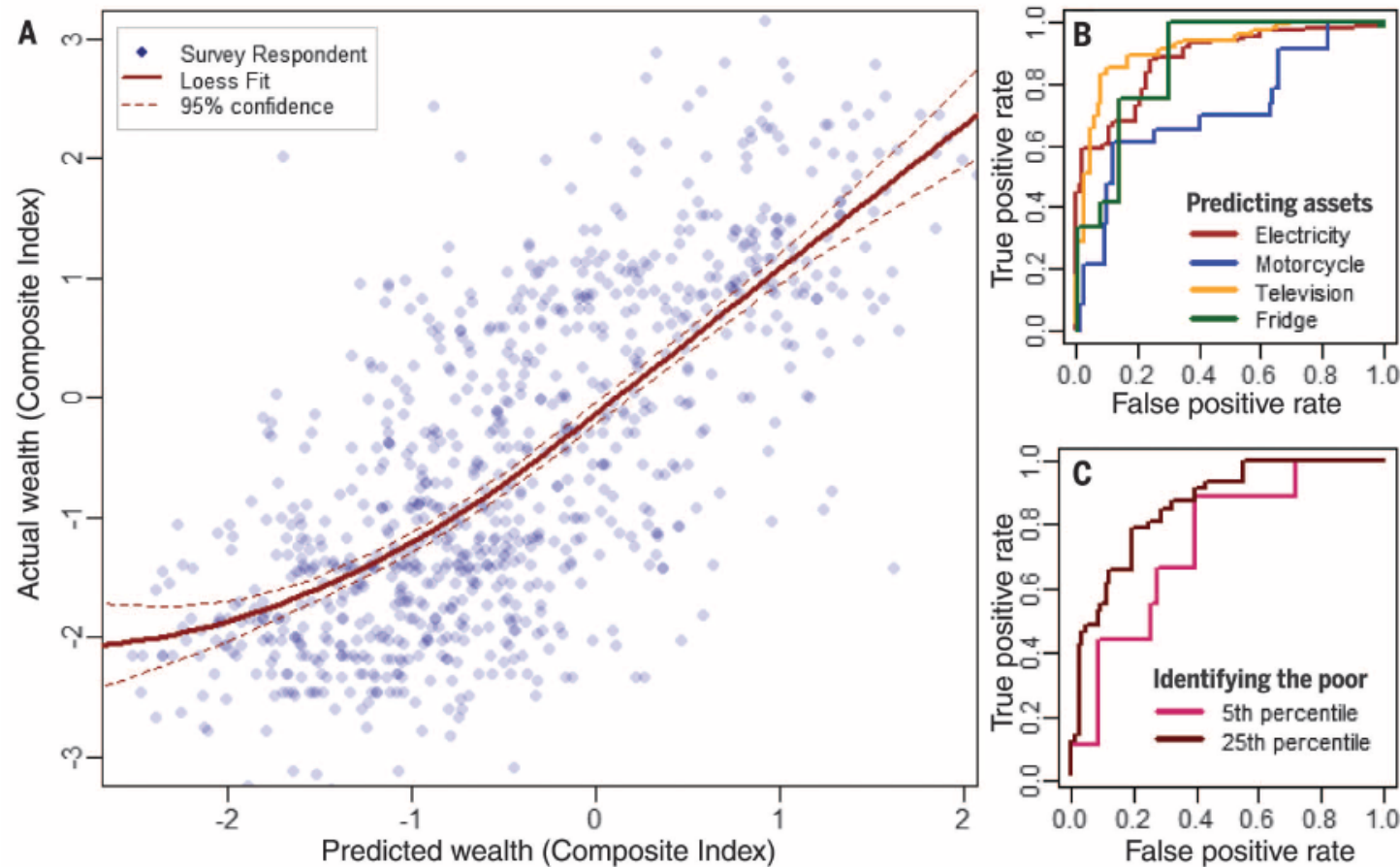
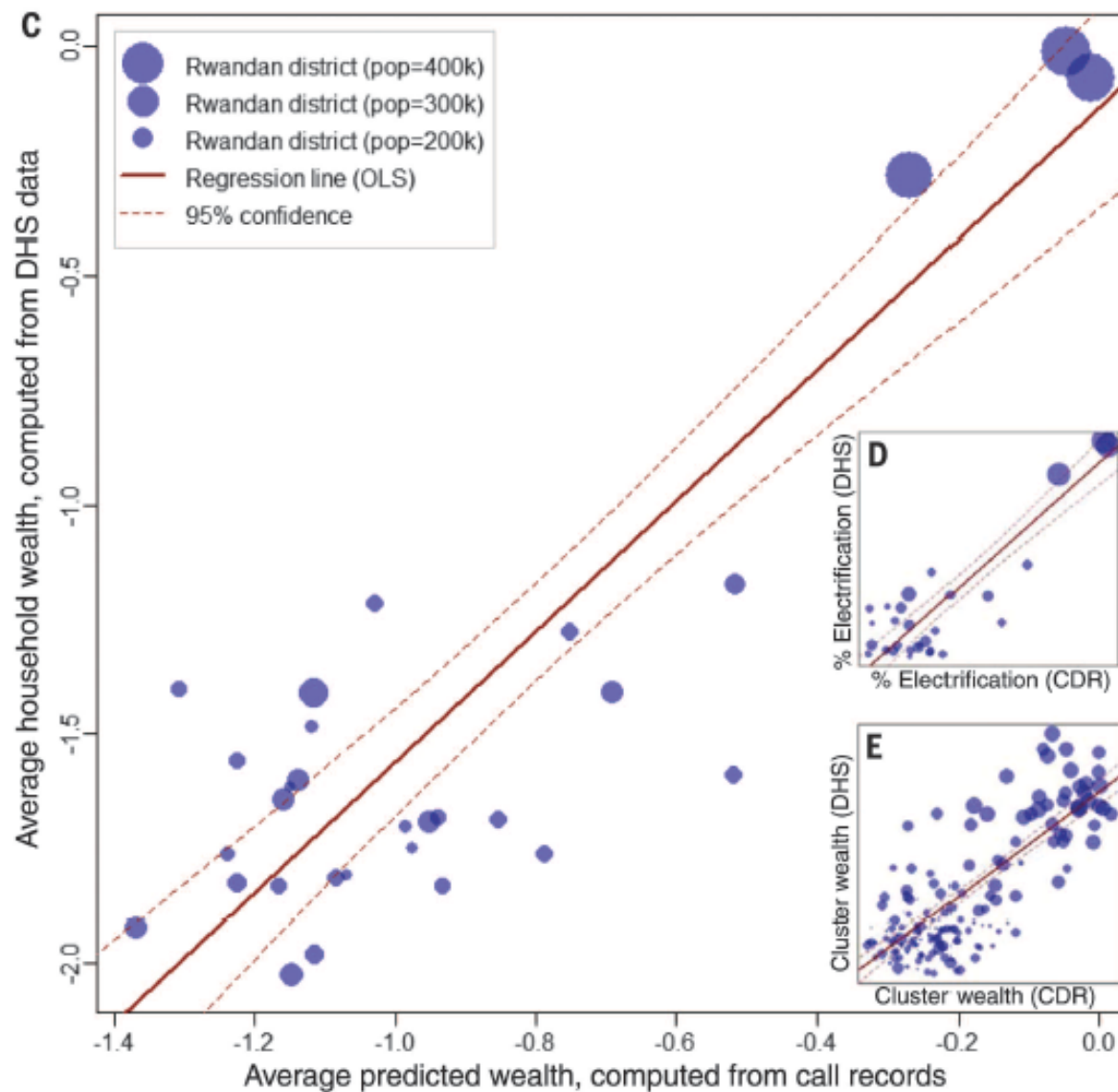


Fig. 1. Predicting survey responses with phone data. (A) Relation between actual wealth (as reported in a phone survey) and predicted wealth (as inferred from mobile phone data) for each of the 856 survey respondents. (B) Receiver operating characteristic (ROC) curve showing the model's ability to predict whether the respondent owns several different assets. AUC values for electricity, motorcycle, television, and fridge, respectively, are as follows: 0.85, 0.67, 0.84, and 0.88. (C) ROC curve illustrates the model's ability to correctly identify the poorest individuals. The poor are defined as those in the 5th percentile (AUC = 0.72) and the 25th percentile (AUC = 0.81) of the composite wealth index distribution.

Predicting district wealth



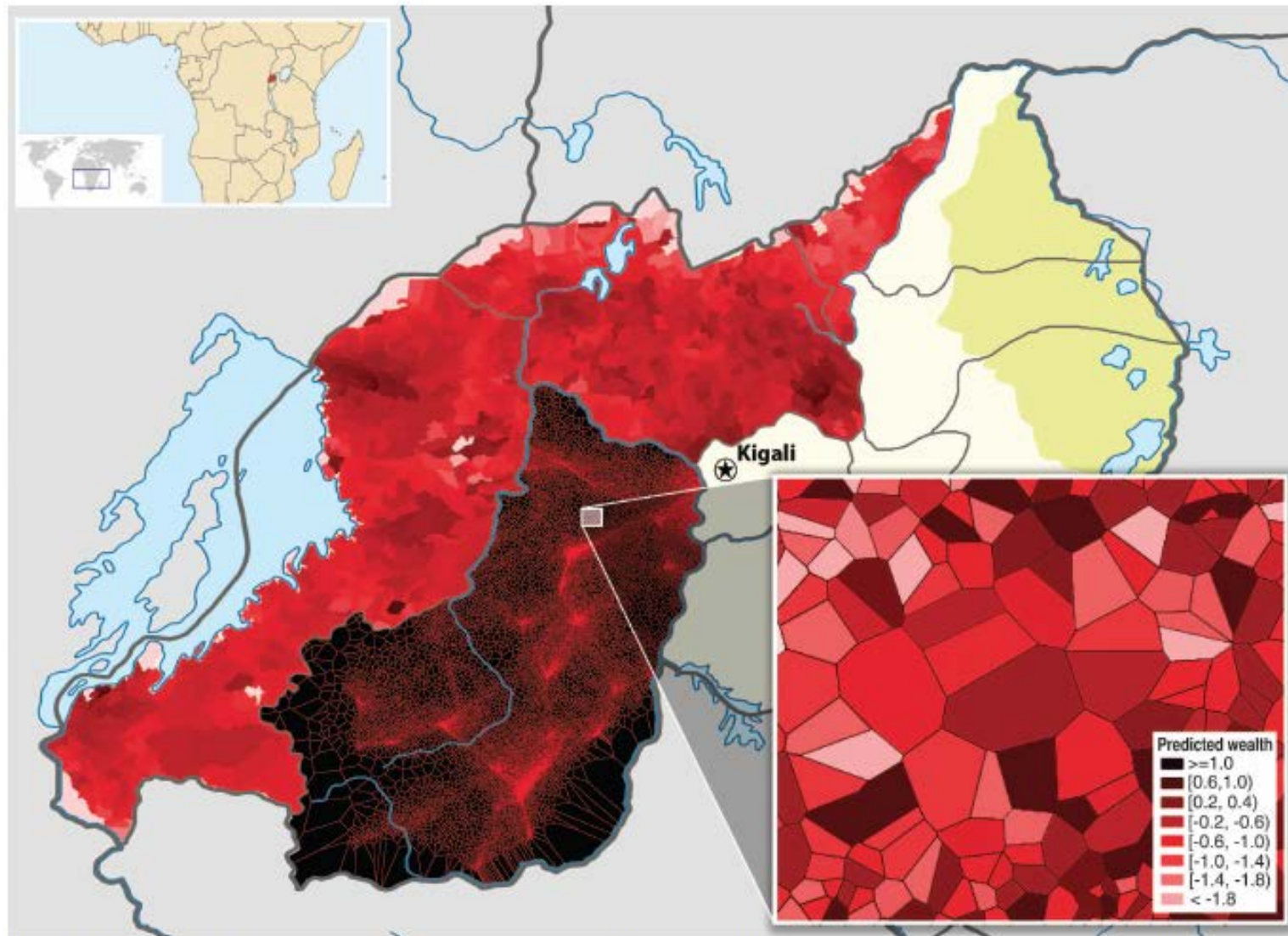
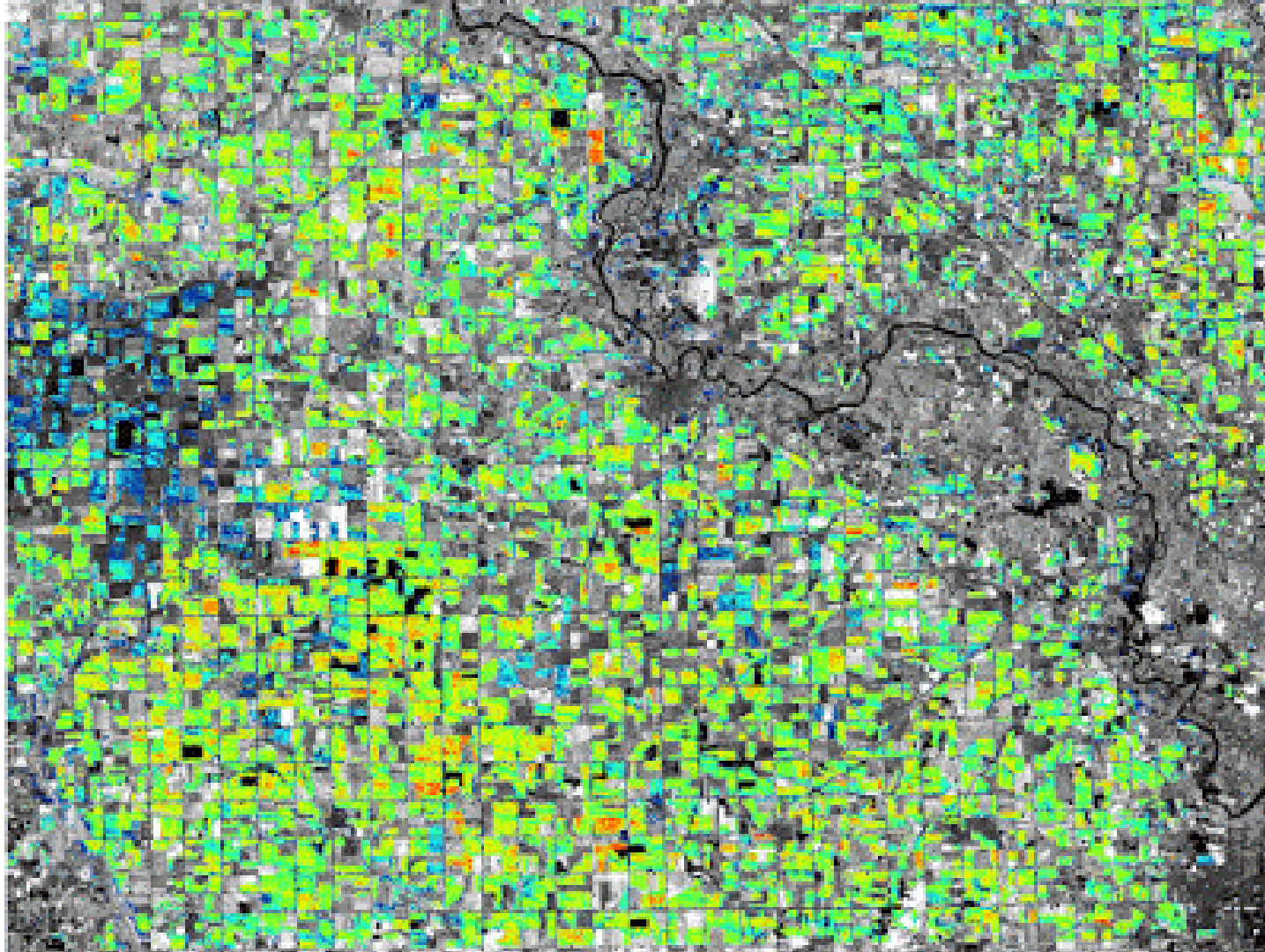


Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km² region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

Crop yield



Text as data

- Gentzkow and Shapiro (EMA, 2010)
 - Words as predictors of ideology on members of congress and newspapers.
- Media sentiment and the stock market
 - Giving Content to Investor Sentiment: The Role of Media in the Stock Market (Tetlock, JoF, 2007)
 - When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks (Loughran and McDonald, JoF, 2011)
- Transparency and Deliberation within the FOMC: a Computational Linguistics Approach (Hansen and Prat, 2014)
- My work with Bei Qin and Yanhui Wu
 - Measure government control of newspapers
 - Use Chinese social media data
 - to identify corruption
 - to measure amount of propaganda.

Text as data

1. Document representation

- Bag of words: vector of word frequencies or tf-idf.

2. Classifier construction.

References

1. TM
high-level

Introduction to the **tm** Package Text Mining in R

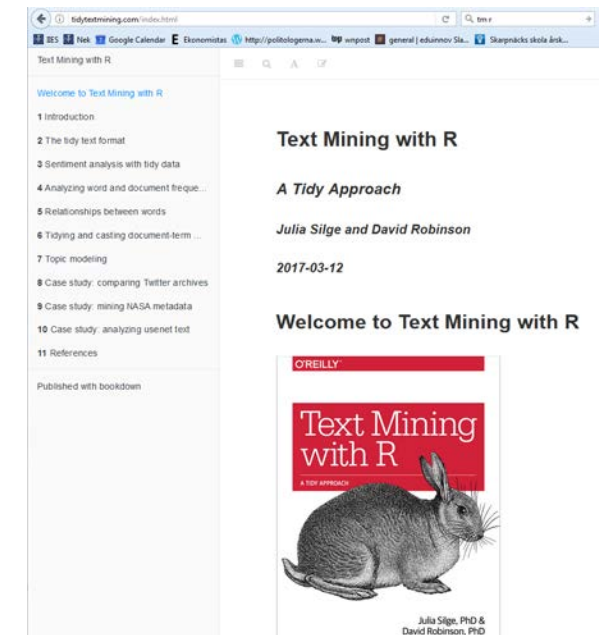
Ingo Feinerer

March 2, 2017

Introduction

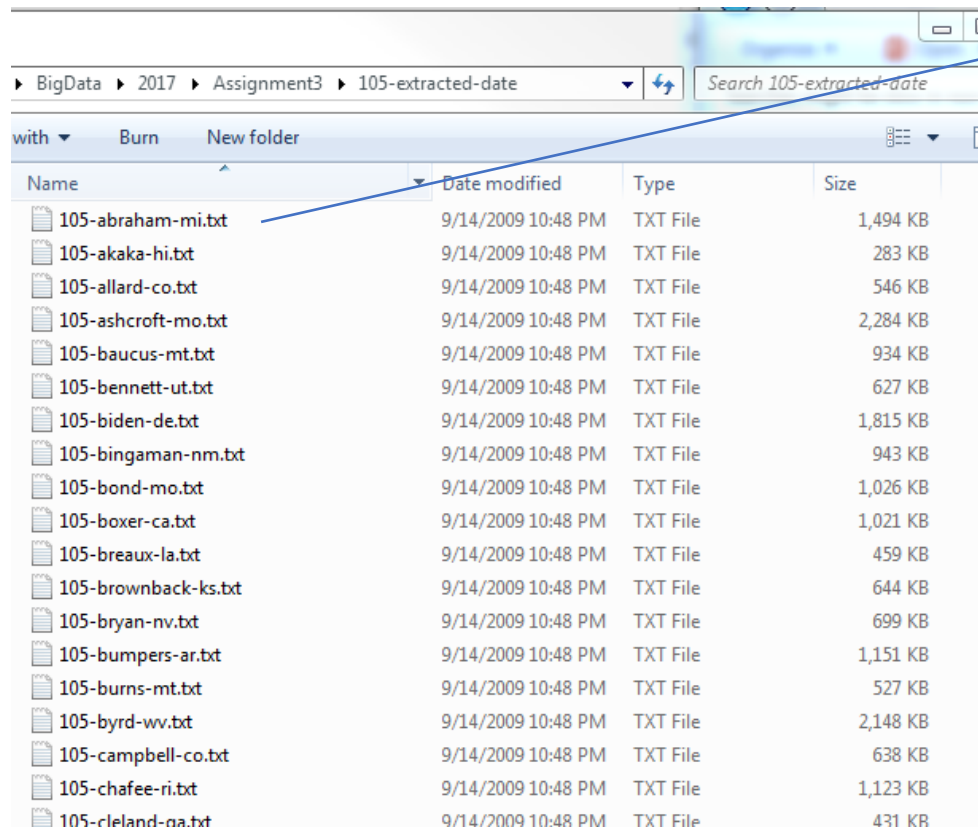
This vignette gives a short introduction to text mining in R utilizing the text mining framework provided by the **tm** package. We present methods for data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices. Our focus is on the main aspects of getting started with text mining in R—an in-depth description of the text mining infrastructure offered by **tm** was published in the *Journal of Statistical Software* (Feinerer et al., 2008). An introductory article on text mining in R was published in *IT News* (Feinerer, 2008).

2. Tidytext
low-level, standard commands.
<https://www.tidytextmining.com>
3. Introduction to Statistical Learning
Ch 4 (Classification), Ch 9 (SVM).
4. Other: stringr, stringi, wordcloud
dplyr, tidyr
slam
SparseM
e1071

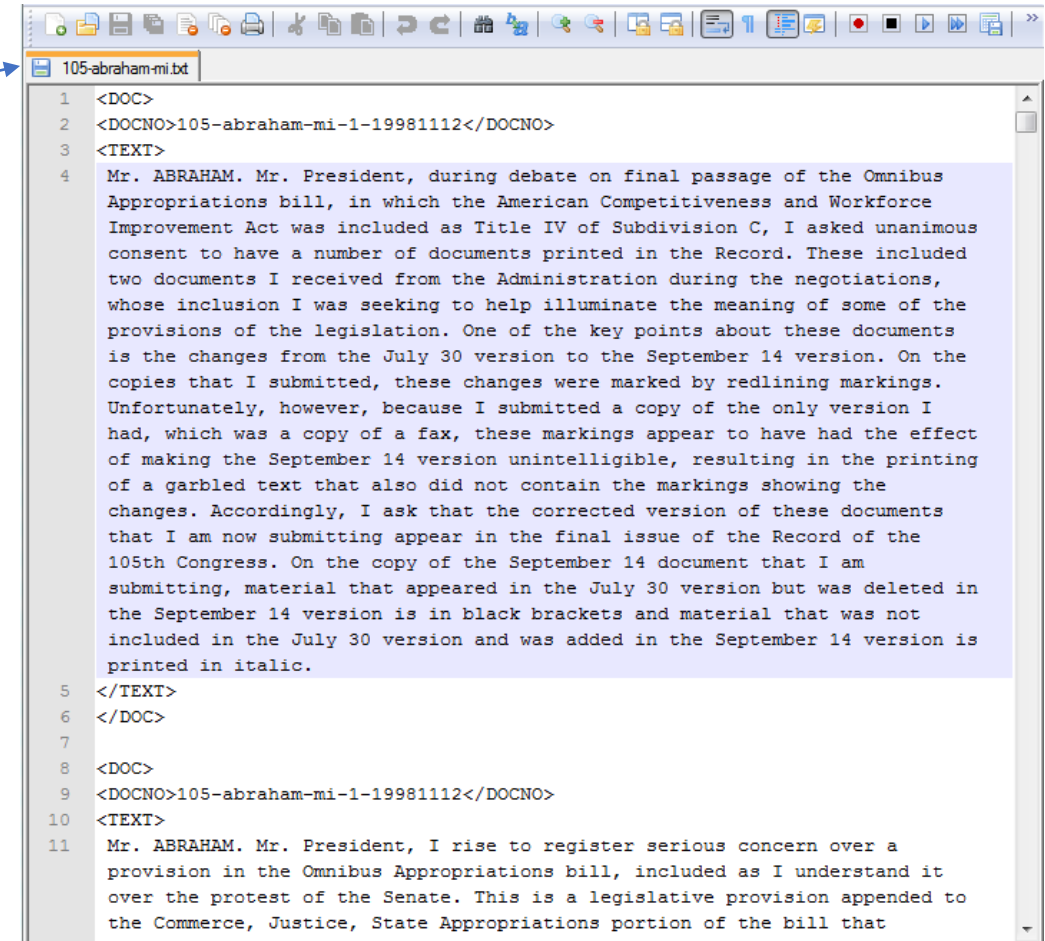


1. Read Corpus Collection of texts into R.

Raw data with documents of senator speeches.



BigData > 2017 > Assignment3 > 105-extracted-date				
Search 105-extracted-date				
with Burn New folder				
Name	Date modified	Type	Size	
105-abraham-mi.txt	9/14/2009 10:48 PM	TXT File	1,494 KB	
105-akaka-hi.txt	9/14/2009 10:48 PM	TXT File	283 KB	
105-allard-co.txt	9/14/2009 10:48 PM	TXT File	546 KB	
105-ashcroft-mo.txt	9/14/2009 10:48 PM	TXT File	2,284 KB	
105-baucus-mt.txt	9/14/2009 10:48 PM	TXT File	934 KB	
105-bennett-ut.txt	9/14/2009 10:48 PM	TXT File	627 KB	
105-biden-de.txt	9/14/2009 10:48 PM	TXT File	1,815 KB	
105-bingaman-nm.txt	9/14/2009 10:48 PM	TXT File	943 KB	
105-bond-mo.txt	9/14/2009 10:48 PM	TXT File	1,026 KB	
105-boxer-ca.txt	9/14/2009 10:48 PM	TXT File	1,021 KB	
105-breaux-la.txt	9/14/2009 10:48 PM	TXT File	459 KB	
105-brownback-ks.txt	9/14/2009 10:48 PM	TXT File	644 KB	
105-bryan-nv.txt	9/14/2009 10:48 PM	TXT File	699 KB	
105-bumpers-ar.txt	9/14/2009 10:48 PM	TXT File	1,151 KB	
105-burns-mt.txt	9/14/2009 10:48 PM	TXT File	527 KB	
105-byrd-wv.txt	9/14/2009 10:48 PM	TXT File	2,148 KB	
105-campbell-co.txt	9/14/2009 10:48 PM	TXT File	638 KB	
105-chafee-ri.txt	9/14/2009 10:48 PM	TXT File	1,123 KB	
105-cleland-na.txt	9/14/2009 10:48 PM	TXT File	431 KB	



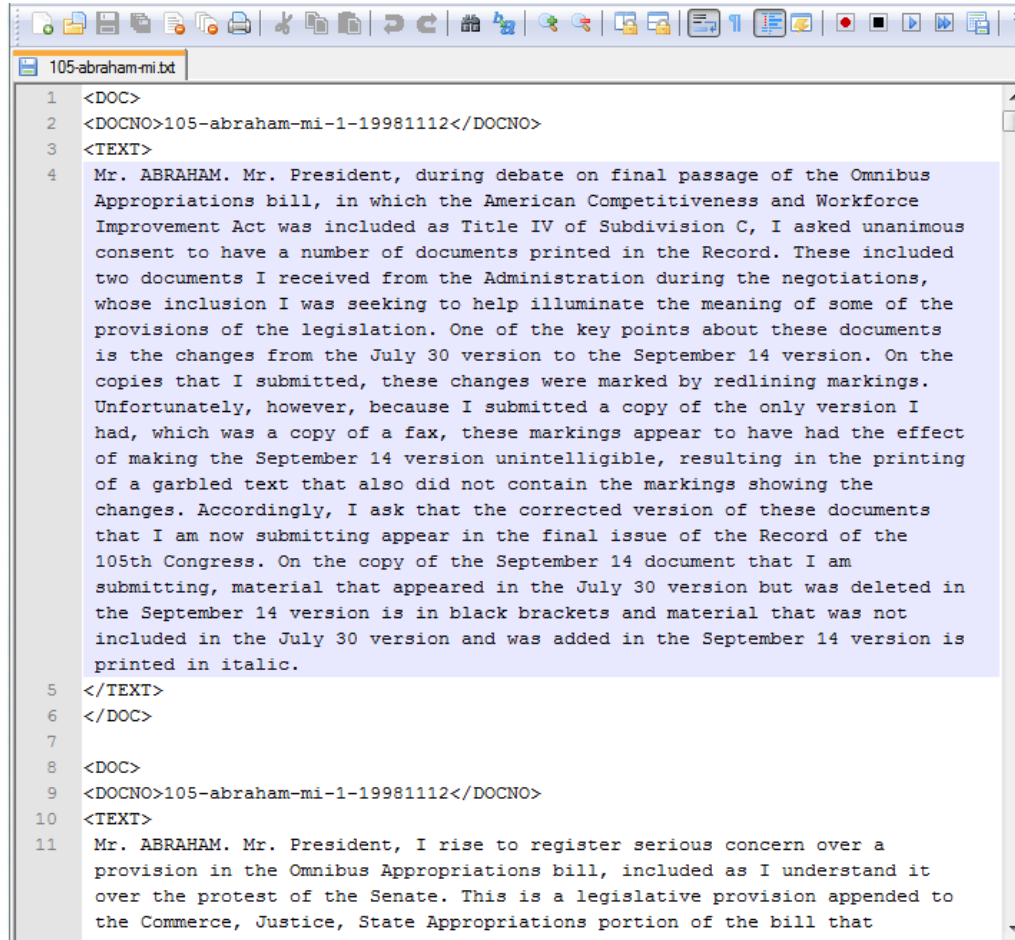
```
1 <DOC>
2 <DOCNO>105-abraham-mi-1-19981112</DOCNO>
3 <TEXT>
4 Mr. ABRAHAM. Mr. President, during debate on final passage of the Omnibus
Appropriations bill, in which the American Competitiveness and Workforce
Improvement Act was included as Title IV of Subdivision C, I asked unanimous
consent to have a number of documents printed in the Record. These included
two documents I received from the Administration during the negotiations,
whose inclusion I was seeking to help illuminate the meaning of some of the
provisions of the legislation. One of the key points about these documents
is the changes from the July 30 version to the September 14 version. On the
copies that I submitted, these changes were marked by redlining markings.
Unfortunately, however, because I submitted a copy of the only version I
had, which was a copy of a fax, these markings appear to have had the effect
of making the September 14 version unintelligible, resulting in the printing
of a garbled text that also did not contain the markings showing the
changes. Accordingly, I ask that the corrected version of these documents
that I am now submitting appear in the final issue of the Record of the
105th Congress. On the copy of the September 14 document that I am
submitting, material that appeared in the July 30 version but was deleted in
the September 14 version is in black brackets and material that was not
included in the July 30 version and was added in the September 14 version is
printed in italic.
5 </TEXT>
6 </DOC>
7
8 <DOC>
9 <DOCNO>105-abraham-mi-1-19981112</DOCNO>
10 <TEXT>
11 Mr. ABRAHAM. Mr. President, I rise to register serious concern over a
provision in the Omnibus Appropriations bill, included as I understand it
over the protest of the Senate. This is a legislative provision appended to
the Commerce, Justice, State Appropriations portion of the bill that
```

2. Tokenization

Raw text.



Text in vector form: one word one row.

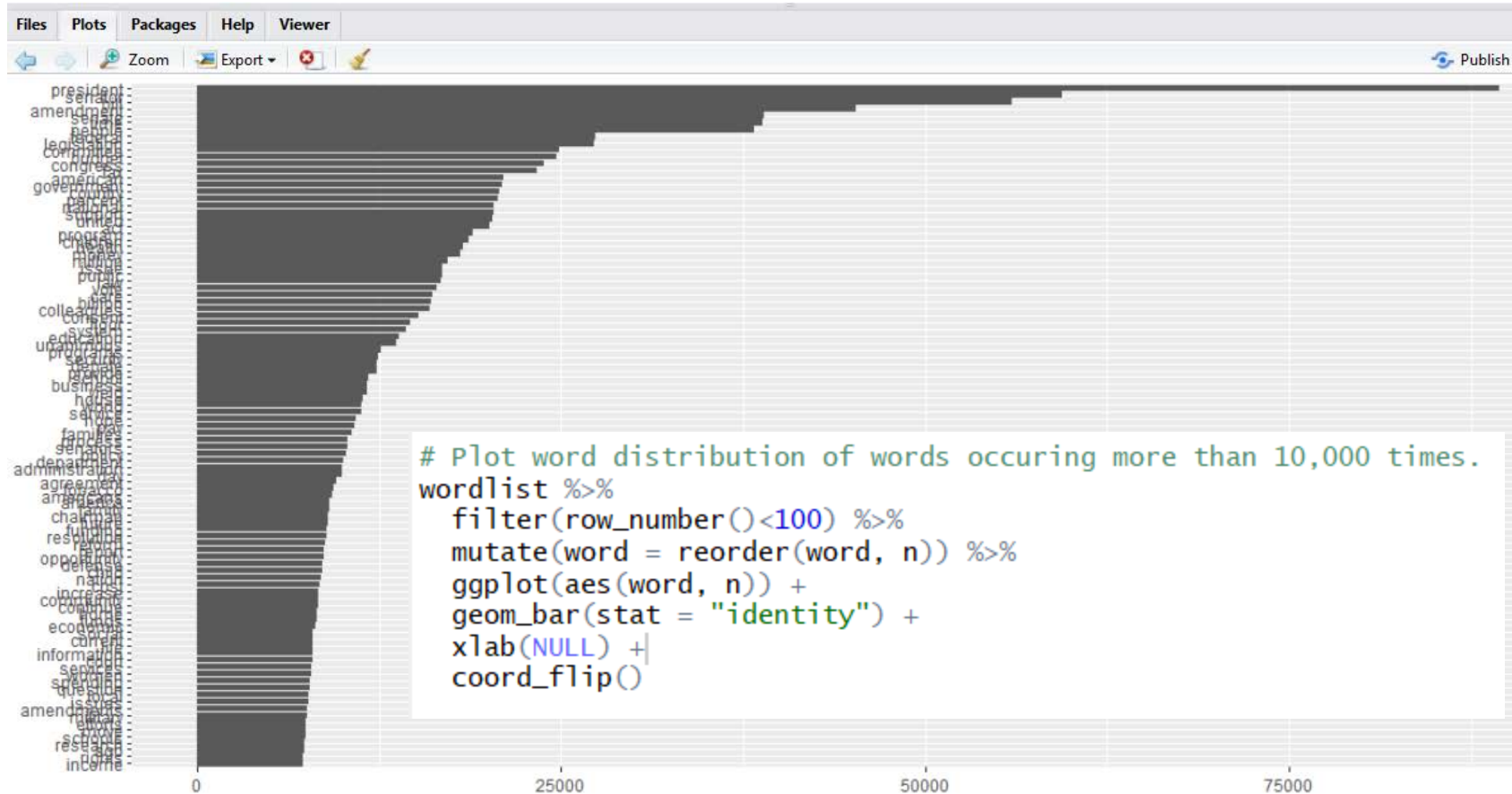


```
1 <DOC>
2 <DOCNO>105-abraham-mi-1-19981112</DOCNO>
3 <TEXT>
4 Mr. ABRAHAM. Mr. President, during debate on final passage of the Omnibus
  Appropriations bill, in which the American Competitiveness and Workforce
  Improvement Act was included as Title IV of Subdivision C, I asked unanimous
  consent to have a number of documents printed in the Record. These included
  two documents I received from the Administration during the negotiations,
  whose inclusion I was seeking to help illuminate the meaning of some of the
  provisions of the legislation. One of the key points about these documents
  is the changes from the July 30 version to the September 14 version. On the
  copies that I submitted, these changes were marked by redlining markings.
  Unfortunately, however, because I submitted a copy of the only version I
  had, which was a copy of a fax, these markings appear to have had the effect
  of making the September 14 version unintelligible, resulting in the printing
  of a garbled text that also did not contain the markings showing the
  changes. Accordingly, I ask that the corrected version of these documents
  that I am now submitting appear in the final issue of the Record of the
  105th Congress. On the copy of the September 14 document that I am
  submitting, material that appeared in the July 30 version but was deleted in
  the September 14 version is in black brackets and material that was not
  included in the July 30 version and was added in the September 14 version is
  printed in italic.
5 </TEXT>
6 </DOC>
7
8 <DOC>
9 <DOCNO>105-abraham-mi-1-19981112</DOCNO>
10 <TEXT>
11 Mr. ABRAHAM. Mr. President, I rise to register serious concern over a
  provision in the Omnibus Appropriations bill, included as I understand it
  over the protest of the Senate. This is a legislative provision appended to
  the Commerce, Justice, State Appropriations portion of the bill that
```

	id	word	row
1	abraham-mi	doc	1
2	abraham-mi	docno	2
3	abraham-mi	105	3
4	abraham-mi	abraham	4
5	abraham-mi	mi	5
6	abraham-mi	1	6
7	abraham-mi	19981112	7
8	abraham-mi	docno	8
9	abraham-mi	text	9
10	abraham-mi	mr	10
11	abraham-mi	abraham	11
12	abraham-mi	mr	12
13	abraham-mi	president	13
14	abraham-mi	during	14
15	abraham-mi	debate	15

Zipf's law: word frequency approx $1/n$.

Words like president is not very informative since every document contains it.



Tf-idf

```
#Compute word frequency, by senator
wordlist_s <- senators_td2 %>%
  inner_join(sen105_party) %>%
  count(id, party, word, sort=TRUE) %>%
  ungroup()

#Compute tf-idf, each senator is a "document"
wordlist_s <- wordlist_s %>%
  bind_tf_idf(word, id, n)
```

	id	party	word	n	share	tf	idf	tf_idf
1	lott-ms	200	president	3030	0.020277460	0.020277460	0	0
2	lott-ms	200	senate	2780	0.018604402	0.018604402	0	0
3	lott-ms	200	senator	2560	0.017132111	0.017132111	0	0
4	wellstone-mn	100	people	2355	0.016220796	0.016220796	0	0

“president” used by all senators: idf=0.

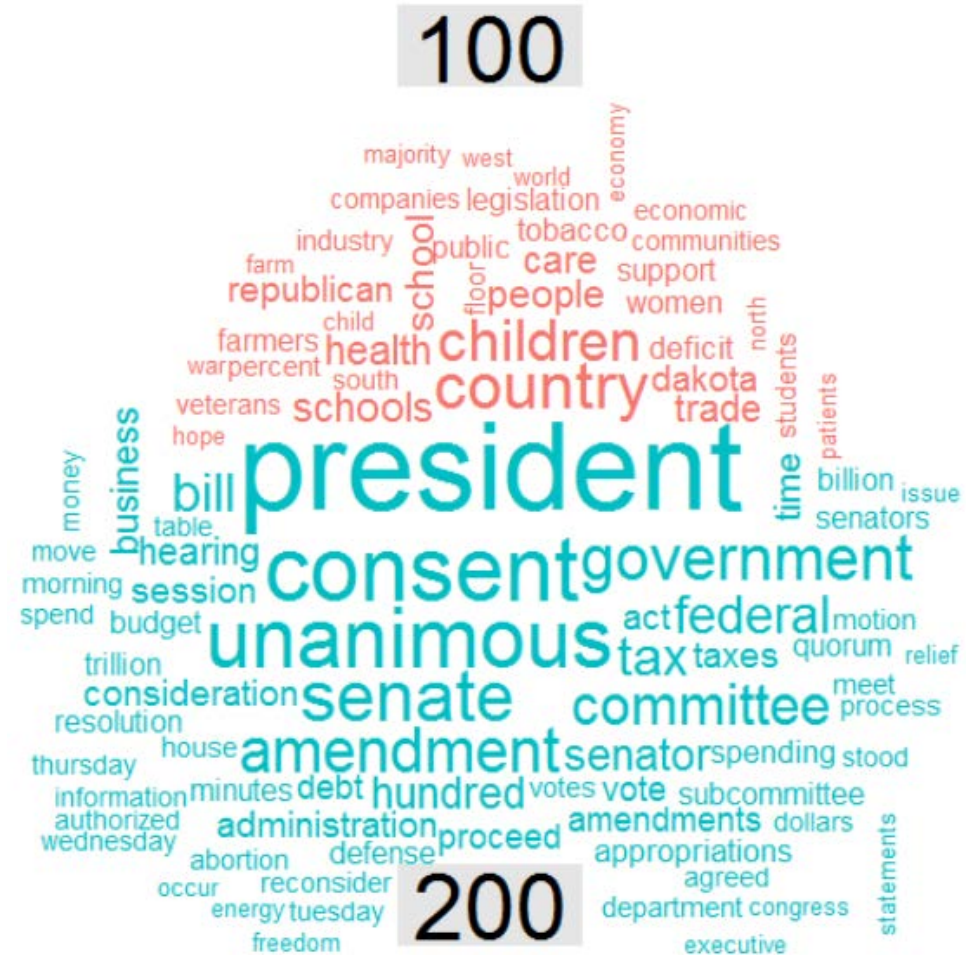
	id	party	word	n	share	tf	idf	tf_idf
824885	akaka-hi	100	hawaii's	45	0.0021577559	0.0021577559	3.21887582	0.006945548
824884	dewine-oh	200	haitian	196	0.0024831501	0.0024831501	2.12026354	0.005264933
824883	conrad-nd	100	forks	166	0.0025595165	0.0025595165	1.83258146	0.004690522
824882	wellstone-mn	100	blanca	177	0.0012191426	0.0012191426	3.50655790	0.004274994
824881	levin-mi	100	atr	105	0.0010871704	0.0010871704	3.91202301	0.004253035
824880	akaka-hi	100	monk	31	0.0014864541	0.0014864541	2.81341072	0.004182006

3. Analysis

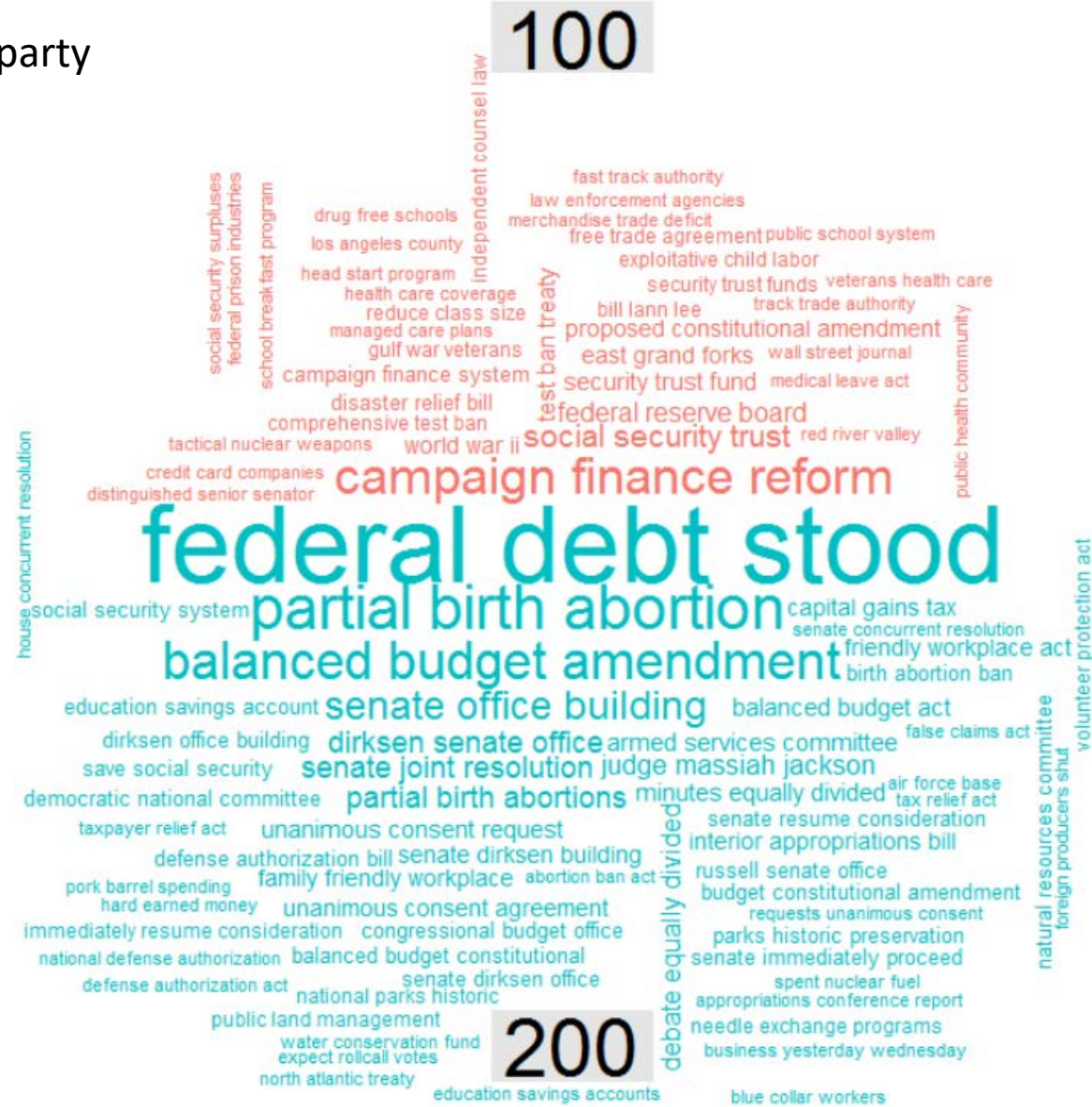
- Descriptive: frequency by party.
- Sentiment analysis.
- Classification (SVM).

Descriptive: Words by party

	party	word	n	share	rank
	<int>	<chr>	<int>	<dbl>	<int>
1	200	president	51613	0.014386891	1
2	100	president	37879	0.011503566	1
3	200	senator	32291	0.009000971	2
4	200	bill	30986	0.008637208	3
5	100	senator	27100	0.008230065	2
6	200	amendment	25756	0.007179369	4
7	100	bill	24981	0.007586541	3
8	200	senate	22907	0.006385223	5
9	200	time	21165	0.005899648	6
10	100	amendment	19452	0.005907425	4
#	... with 98,724 more rows				



Trigrams by party



Sentiment analysis:

Wordcount using the sentiments lexicons in tidytext.

```
> # Sentiments, word count
> library(tidytext)
> sentiments
# A tibble: 23,165 × 4
  word sentiment lexicon score
  <chr>      <chr>   <chr> <int>
1   abacus    trust    nrc    NA
2  abandon    fear    nrc    NA
3  abandon  negative  nrc    NA
4  abandon  sadness  nrc    NA
5 abandoned  anger    nrc    NA
6 abandoned  fear    nrc    NA
7 abandoned  negative  nrc    NA
8 abandoned  sadness  nrc    NA
9 abandonment anger    nrc    NA
10 abandonment fear    nrc    NA
# ... with 23,155 more rows
```

```
> table(lexicon)
lexicon
AFINN  Bing  nrc
2476  6788 13901
> table(sentiment[lexicon=="nrc"])
      anger anticipation    disgust      fear
      1247           839       1058      1476
      joy      negative    positive    sadness
      689          3324       2312       1191
      surprise      trust
      534          1231
> table(sentiment[lexicon=="bing"])
negative positive
  4782      2006
> table(score[lexicon=="AFINN"])
-5  -4  -3  -2  -1   0   1   2   3   4   5
16  43 264 965 309   1 208 448 172  45   5
```

Sentiment analysis: implement by merge

```
> get_sentiments("nrc")[1:10,]  
# A tibble: 10 × 2  
  word sentiment  
  <chr>      <chr>  
1   abacus    trust  
2  abandon    fear  
3  abandon negative  
4  abandon sadness  
5 abandoned  anger  
6 abandoned  fear  
7 abandoned negative  
8 abandoned sadness  
9 abandonment anger  
10 abandonment fear
```

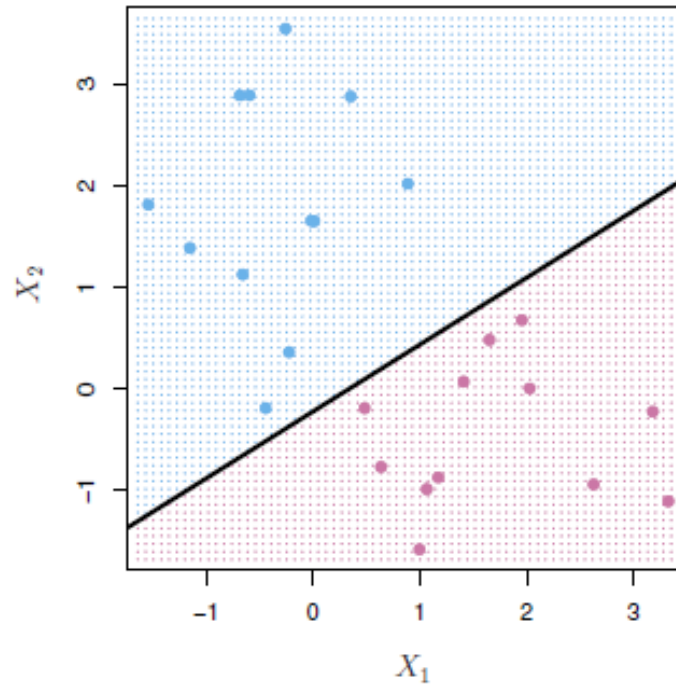
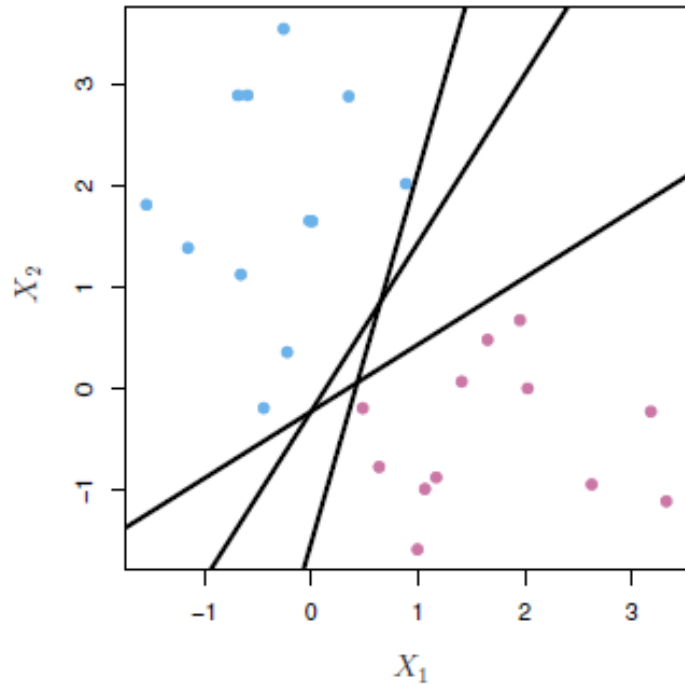
```
> wordlist_s %>%  
+   inner_join(get_sentiments("nrc")) %>%  
+   group_by(party) %>%  
+   mutate(total=sum(n)) %>%  
+   group_by(party, sentiment) %>%  
+   summarise(n2=sum(n/total)) %>%  
+   spread(party, n2)  
Joining, by = "word"  
# A tibble: 10 × 3  
  sentiment      `100`      `200`  
*      <chr>      <dbl>      <dbl>  
1      anger 0.05536695 0.05335390  
2 anticipation 0.10183271 0.10087415  
3      disgust 0.03113449 0.02888124  
4      fear 0.06904045 0.06815292  
5      joy 0.06772064 0.06639535  
6     negative 0.11874833 0.11591421  
7     positive 0.26543140 0.27106587  
8      sadness 0.05392743 0.05254787  
9     surprise 0.03358197 0.03209497  
10     trust 0.20321562 0.21071953
```

Predict party based on words

- Support Vector Machine (SVM)
 - y-variable is party.
 - x-variables is document (senator) – trigram matrix
 - Parameters
 - kernel="linear"
 - cost argument: ten-fold cross-validation

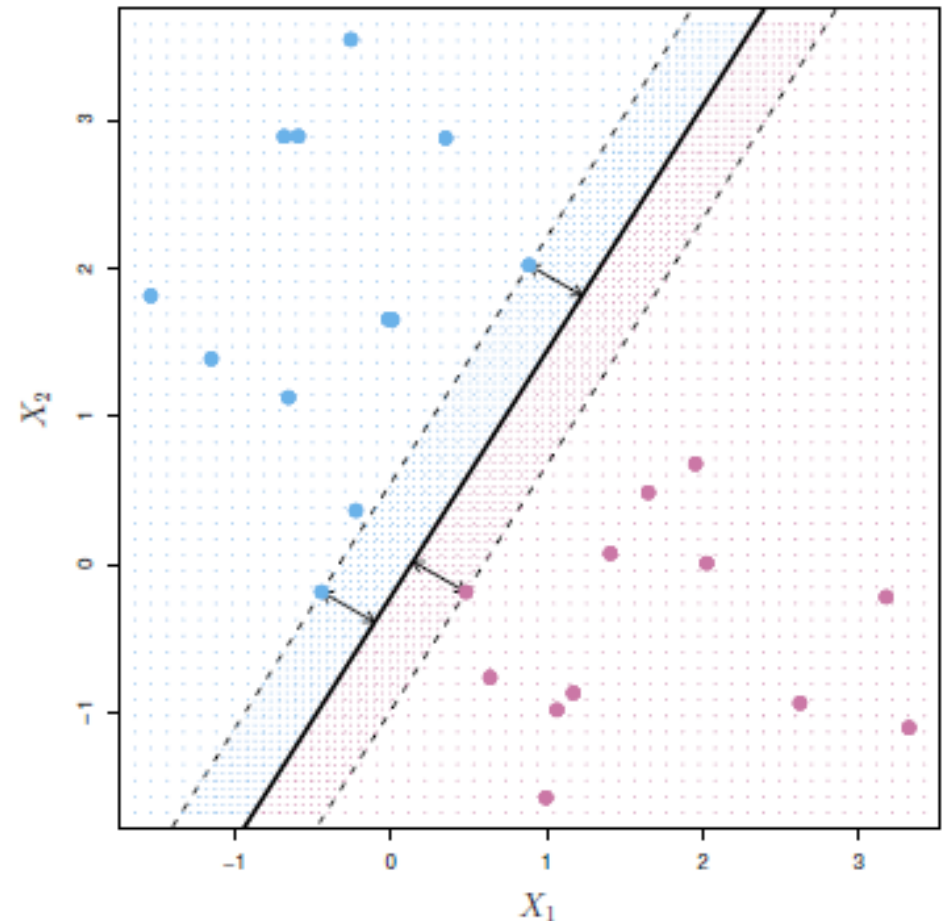
SVM based on separating hyperplanes

- Classification problem: find hyperplane that separates Democrats from Republicans in word space.



Maximal Margin Classifier

- Among all separating hyperplanes, find the one that makes the biggest gap or margin between the two classes.
- SVM does this while allowing a few senators to be misclassified
- SVM works well when classes are (nearly or completely) separated.



Retrieve beta-coefficients on trigrams

	beta
campaign finance reform	-0.009416556
world war ii	-0.007504979
test ban treaty	-0.006341214
el camino real	-0.005978151
social security trust	-0.005900607

senate dirksen office	0.006196373
debate equally divided	0.006375586
social security system	0.007054818
capital gains tax	0.008592982
senate office building	0.008645145
partial birth abortion	0.010126529

Senators with most ideological language

party	id	200/100
100	dorgan-nd	-1.7261673
100	feingold-wi	-1.5034754
100	ford-ky	-1.0002059
100	bryan-nv	-1.0001969

200	ashcroft-mo	1.6163157
200	santorum-pa	1.6557853
200	hatch-ut	2.0381536
200	lott-ms	2.0437390

Predictions in policy

Can prediction be directly useful in policy?

- Policy decisions seem inherently causal
 - “Should we do policy X”?
 - “What will X do?”
 - “What happens with and without X?”
- Predictions may still provide policy relevant information.

Policy-relevant prediction problems

- Who will commit crime if released on bail?
- Where and when will the next famine occur?
- Who will default on loan?
- Who is evading taxes?
- Who will be long-term unemployed?
- Who can work / should receive benefits?

Issues

- Accountability.
- Transparency.
- Statistical discrimination.

Nigeria's food crisis: by the time famine is declared, it's too late

events Australia edition ▾

theguardian



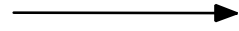
A Policy Problem in the US (Kleinberg et al 2018)

- Each year police make over 12 million arrests
- Where do people wait for trial?
- Release vs. detain high stakes
 - Pre-trial detention spells avg. 2-3 months (can be up to 9-12 months)
 - Nearly 750,000 people in jails in US
 - Consequential for jobs, families as well as crime

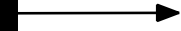
Judge's Problem

- Judge must decide whether to release or not (bail)
- Defendant when out on bail can behave badly:
 - Fail to appear at case
 - Commit a crime
- The judge is making a *prediction*

New Input
Defendant
history



Judge



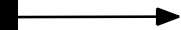
Prediction
Crime?

Data on past defendants
IN: Defendant history
OUT: Crime?

New Input
Defendant
history



Algorithm



Prediction
Crime?

Predict using variant of decision tree

Variables Used	Variables Not Used
Number of Arrest Charges	Outcome Variables
Age	Days Detained
Most Serious Prior Arrest	Released
Number of Prior Arrests	Bail Amount
Number of Prior Felony Arrests	FTA
Number Prior Misd. Arrests	Re-arrest Category
Any Prior FTA	New Violent Charge
Most Serious Prior Conviction	
Number of Prior Convictions	
Number Prior Felony Convictions	Prohibited
Number of Prior Misd. Convictions	Gender
Number of Adult Felony Convictions	Race
Number prior Prison incarcerations	Hispanic Origin
Number prior Jail incarcerations	
Charge Category	
Released on Prior Case at Arrest	
On Probation at time of Arrest	
On Parole at time of Arrest	
In Custody at time of Arrest	
In Diversion Program at time of Arrest	
Fugitive at time of arrest	
Other CJ status at time of Arrest	
County	
State	
Year	

What outcomes are observed?

Both Jail

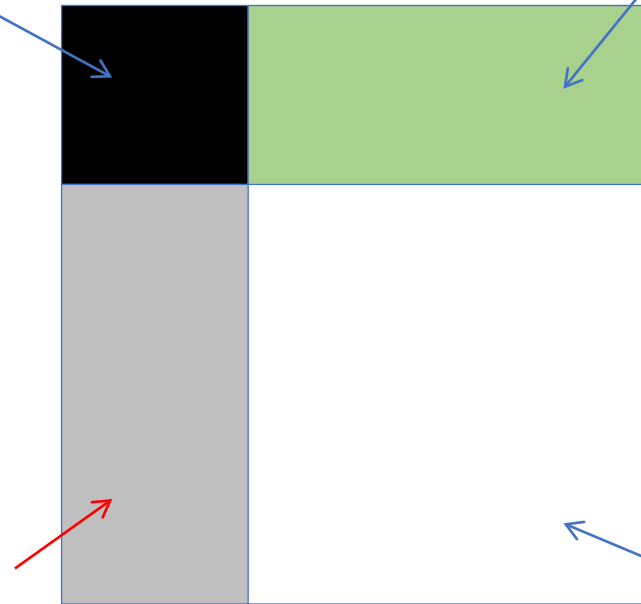
Judge R
Machine Jails
(ground truth)

Machine Release

Machine R
Judge Jails
(NO ground truth)

Both Release

Judge Release



But the crime rate could be very high in this case if the judge observes e.g. gang tatoos.

Observed differences

- Judges release high risk defendants.
- Adjusted rule
 - Don't release these.

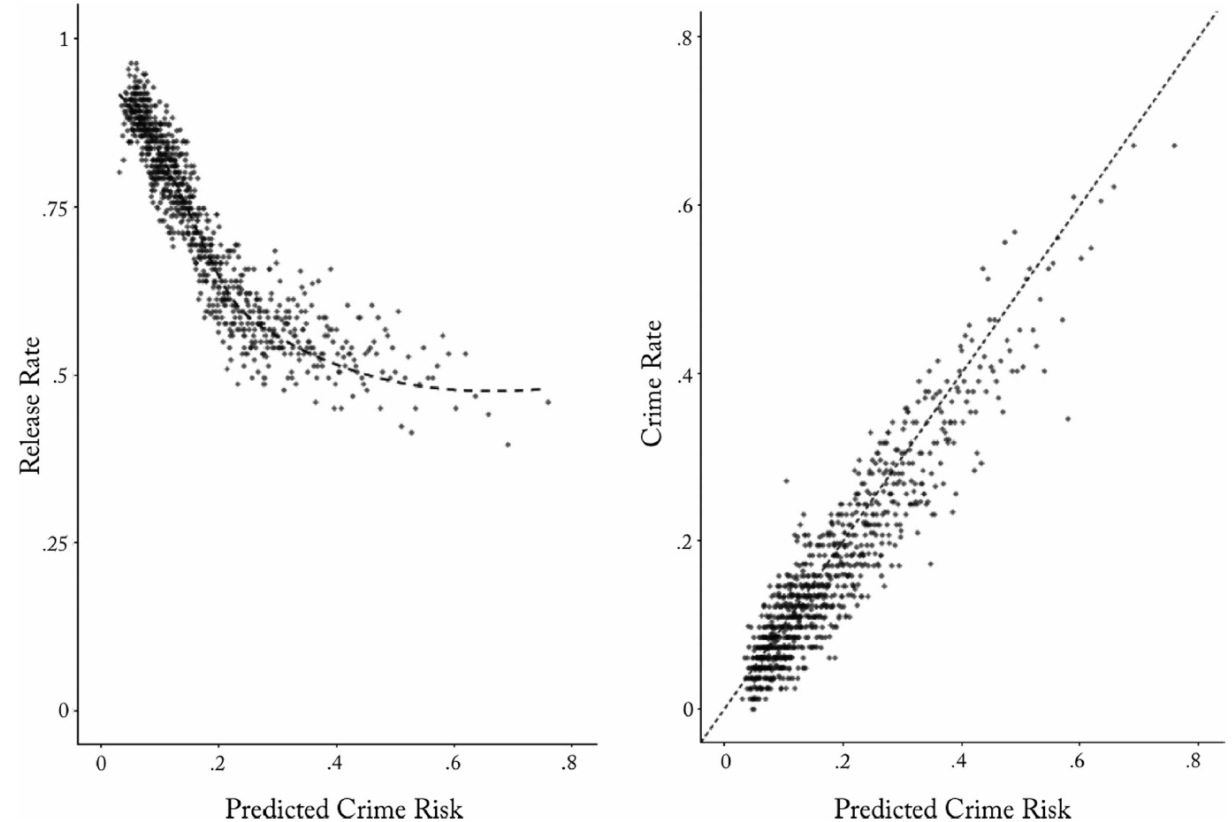


FIGURE II

How Machine Predictions of Crime Risk Relate to Judge Release Decisions and Actual Crime Rates

What Makes Bail “Work”

- Clear goal
- Core problem is predictive
 - Not asking “what works?”
 - Asking “who” and “when”?

Algorithms are not a panacea

- Human decision aid, not substitute.
- Are biases baked in?
- Prediction accuracy and black box versus parsimony and interpretability.
 - Linear models are easy to interpret; thin-plate splines and random forests are not.
 - Simpler model involving fewer variables may be preferred over a black-box predictor involving them all.

Causal trees and forests: heterogenous causal effects

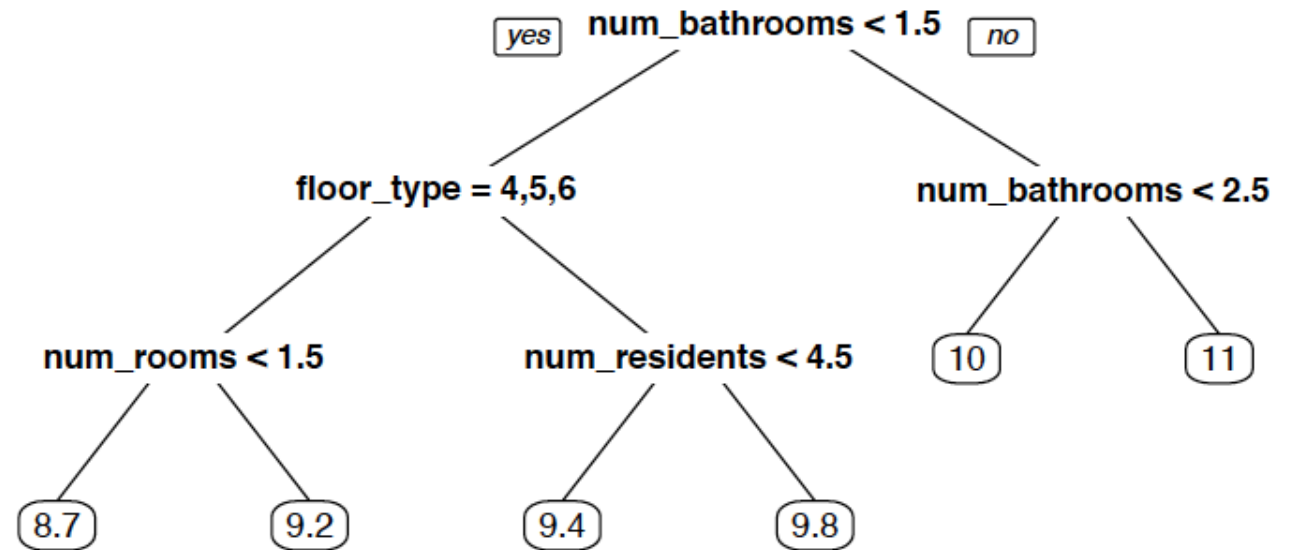
- Concerns about ex-post “data-mining”
 - In medicine, scholars required to pre-specify analysis plan
 - In economic field experiments, calls for similar protocols
- How predict all relevant heterogeneity in an environment with many covariates?
- Goal 1:
 - Allow researcher to specify set of potential covariates
 - Data-driven search for heterogeneity in causal effects with valid standard errors
- Goal 2:
 - Estimate of treatment effect heterogeneity needed for optimal decision-making
 - Heterogenous effects for policy targeting is a prediction problem (not causation).

References

- Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113.27 (2016): 7353-7360.
- Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113.523 (2018): 1228-1242.
- Athey, Susan, and Stefan Wager. "Policy learning with observational data." *Econometrica* 89.1 (2021): 133-161.
- https://gsbdbi.github.io/ml_tutorial/hte_tutorial/hte_tutorial.html

Causal trees

- Recall prediction tree:
 - Prediction: mean y in leaf
 - Split criterion: min MSE



Causal trees

- Prediction: treatment effect in leaf $\tau(l) = \mu(1, l) - \mu(0, l)$
- Split criterion. Let an empirical estimate of $\tau(l)$ be $\bar{\tau}_l$. The standard tree split criterion minimizes the MSE

$$\frac{1}{N} \sum_{i=1}^N (\bar{\tau}_l - \tau_i)^2$$

But this is infeasible since τ_i is not observed!

Solutions: Compute expected MSE (EMSE), similar to CP, AIC, etc.
(Athey and Imbens, 2016)

Causal forests

1. Bootstrap averaging (bagging) of many trees
 - On average, each bagged tree uses $2/3$ of observations.
 - Out-of-bag (OOB) estimates of τ_i for $1/3$ of obs not in bootstrap sample
 - Averages of these used as out-of-sample predictors.
2. Random subset of m predictors is chosen in each split, typically $m=\text{sqrt}(p)$

Inference

Independent samples for

- Training
 - ▶ S^{tr} , tree construction based on EMSE,
 - ▶ S^{est} , estimation of within leaf means,
 - ▶ Bootstrap samples of (S^{tr}, S^{est}) .
- S^{te} , inference (testing)
 - ▶ Holding tree fixed,
standard methods and asymptotic theory within a leaf.
 - ▶ No assumptions about sparsity of true data-generating process.

Example: sickness insurance misuse (Yakymovych, 2022).

Randomized experiment for sickness insurance recipients July-Dec 1988

- Sample: 70,000 in Jämtland 240,000 in Gothenburg.
 - ▶ After restrictions: 123,429 spells by 77,672 workers
- Treatment: days of absence spell before doctor's certificate required
 - ▶ 7 for odd birth dates (control)
 - ▶ 14 for even dates (treated)

Outcomes

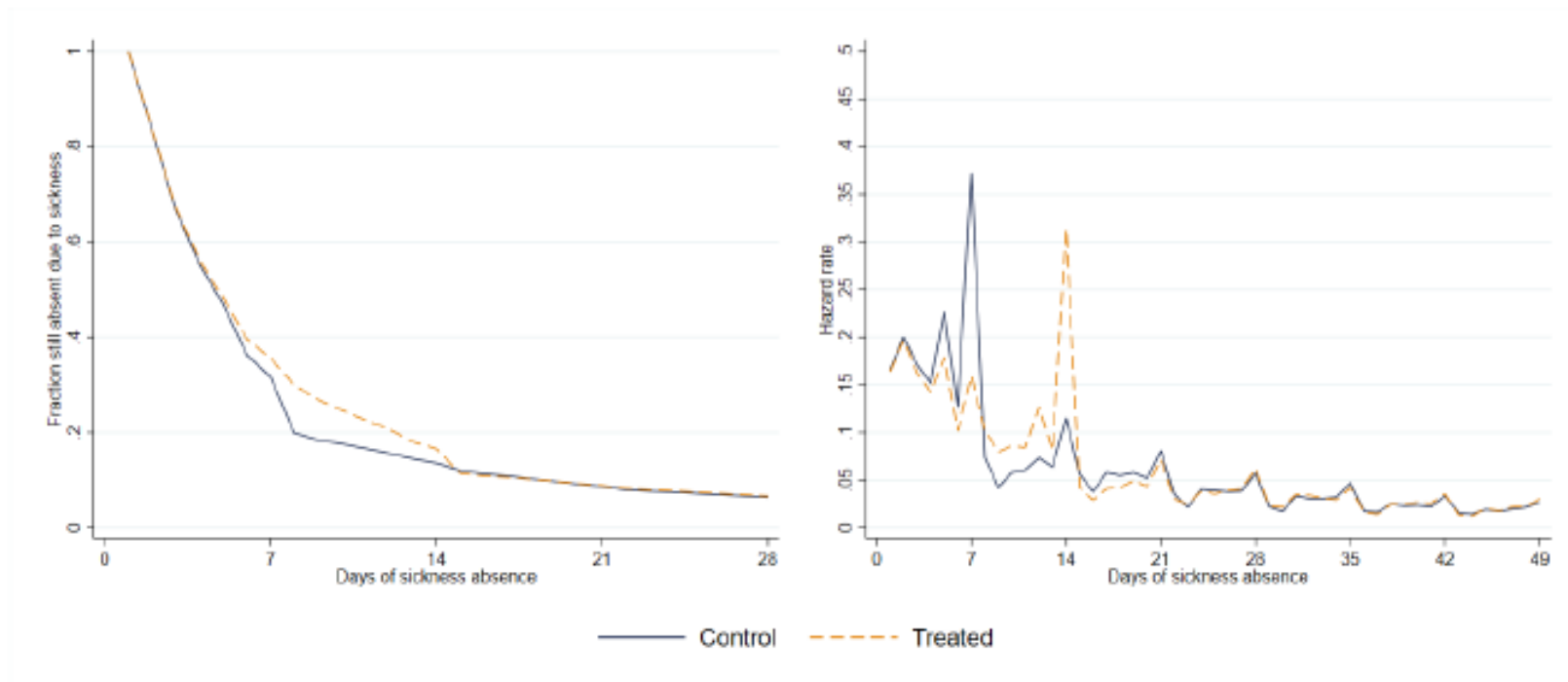
- Duration of sickness spell in days
- Sickness spell 8-14 days

Worker Characteristics

- Health, demographic, family, education, neighborhood, career, workplace, etc

Main effect

FIGURE 3. SURVIVAL AND HAZARD RATES FOR SICKNESS ABSENCE SPELLS TAKEN BY TREATED AND CONTROLS WORKERS IN GOTHENBURG AND JÄMTLAND DURING THE EXPERIMENT

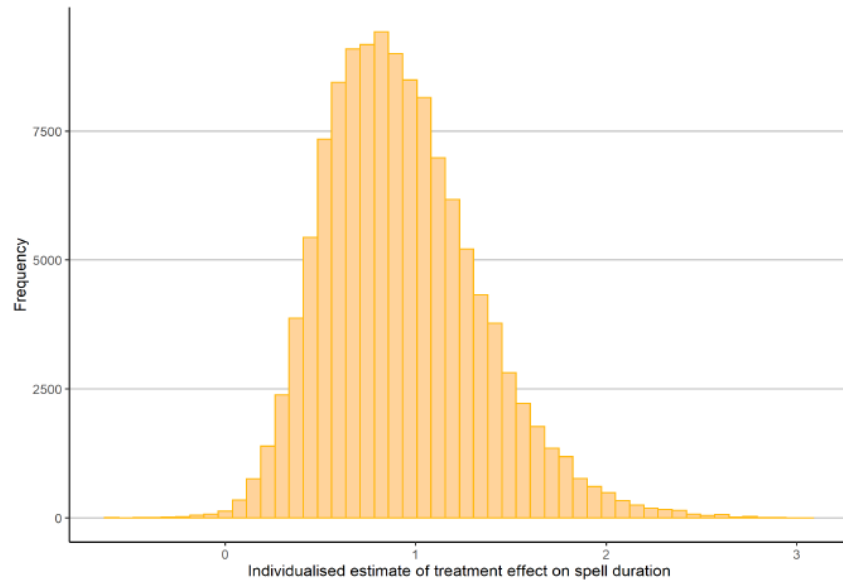


Note: Spells which began between July 1st and December 31st 1988. The hazard rate represents the probability that a worker who has been absent for a given number of days returns to work on the next day.

How strong is the heterogeneity

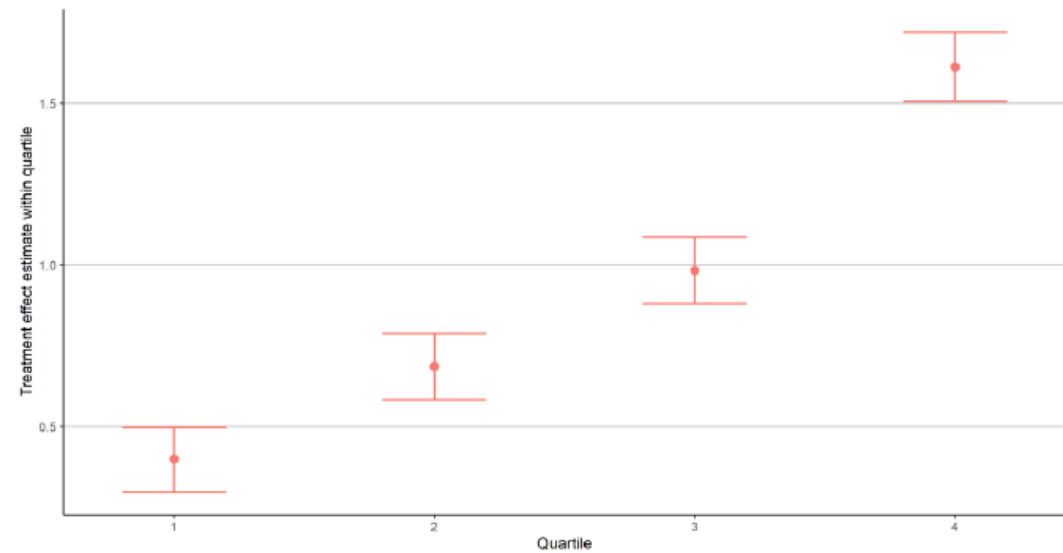
Out-of-bag CATE

FIGURE 4. DISTRIBUTION OF PREDICTED TREATMENT EFFECTS ON SICKNESS ABSENCE SPELL DURATION.



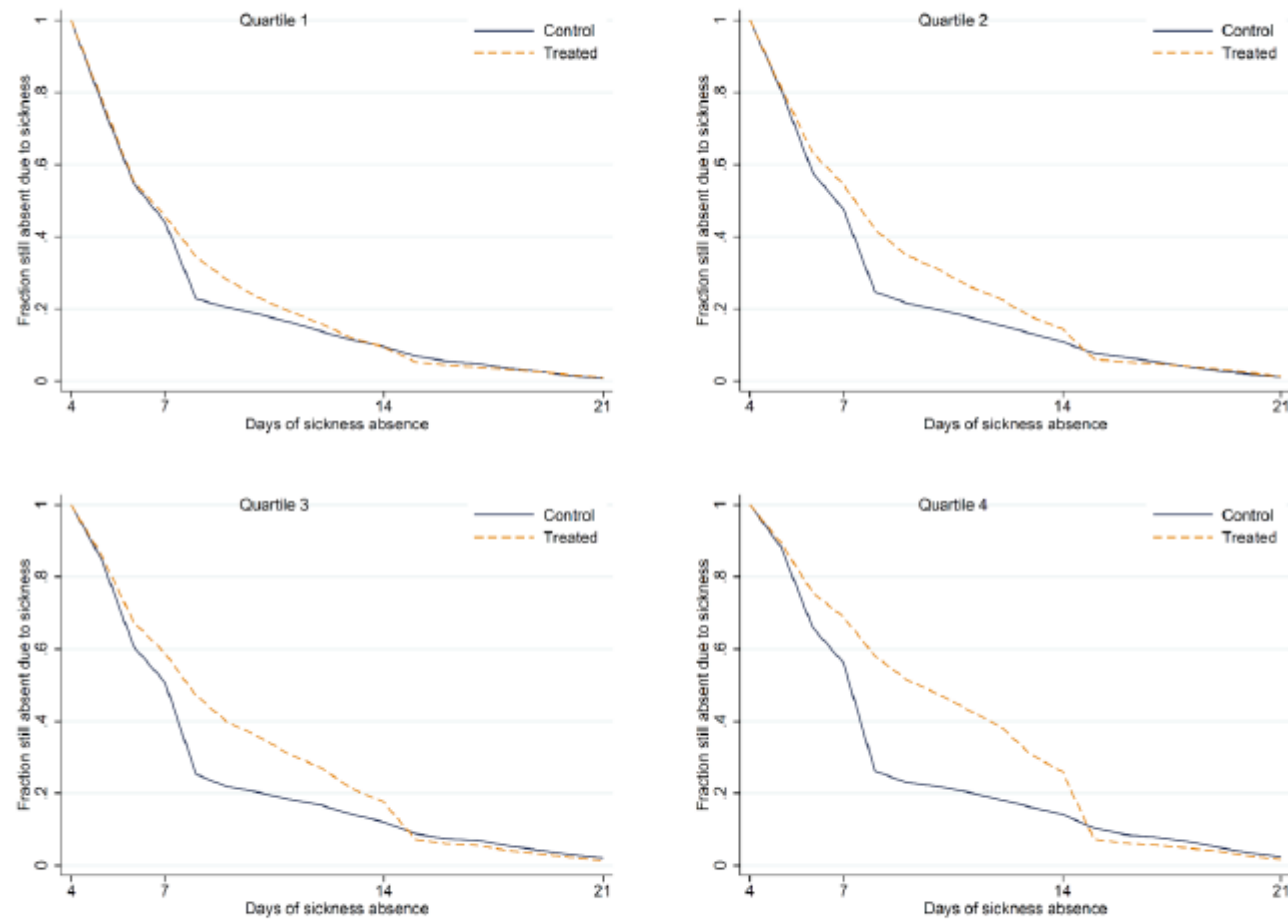
Note: Frequency represents the number of sickness spells with estimated causal forest treatment effects that fall within each bin.

FIGURE 5. ESTIMATED TREATMENT EFFECTS FOR WORKERS WITHIN EACH QUARTILE OF PREDICTED CAUSAL FOREST $\hat{\tau}_x$ ESTIMATES



Note: Quartiles ranked according to causal forest estimated treatment effects, with Q1 containing those estimated to be least affected and Q4 those estimated to be most affected. Treatment effects within each of the quartiles estimated as $\hat{\tau} = \bar{y}_i|(W_w = 1) - \bar{y}_i|(W_w = 0)$. Confidence intervals at the 95 percent level shown.

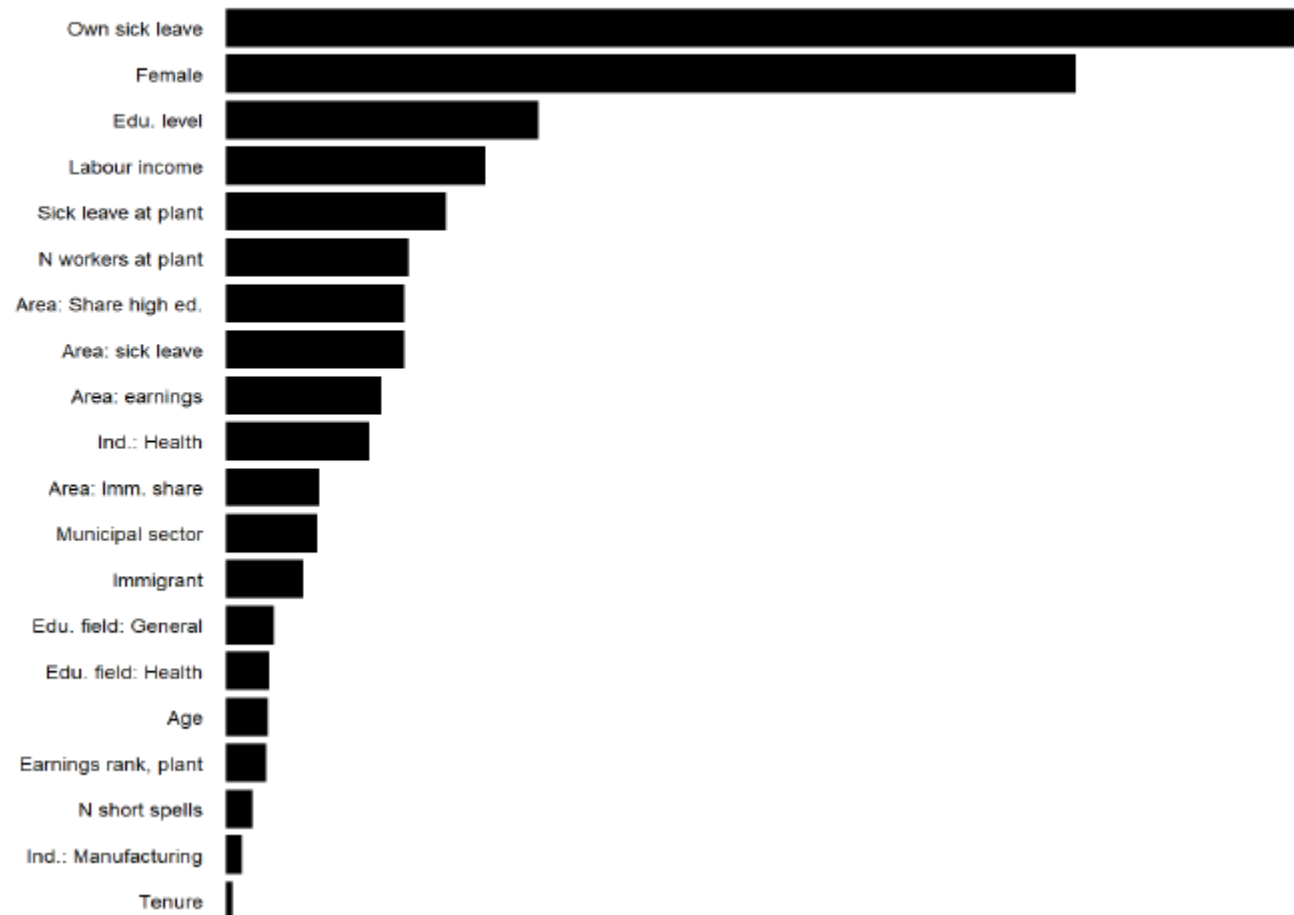
FIGURE 6. SURVIVAL GRAPHS FOR ABSENCE SPELLS AMONG WORKERS IN THE HELD-OUT TEST SET, RANKED BY QUARTILES OF PREDICTED TREATMENT EFFECTS



Note: Survival rates for absence spells of the 20 percent of workers randomised into the held-out test set. Workers divided into quartiles based on out-of-bag causal forest predictions. Quartiles ranked according to size of predicted effect, with Q1 containing those estimated to be least affected and Q4 those estimated to be most affected.

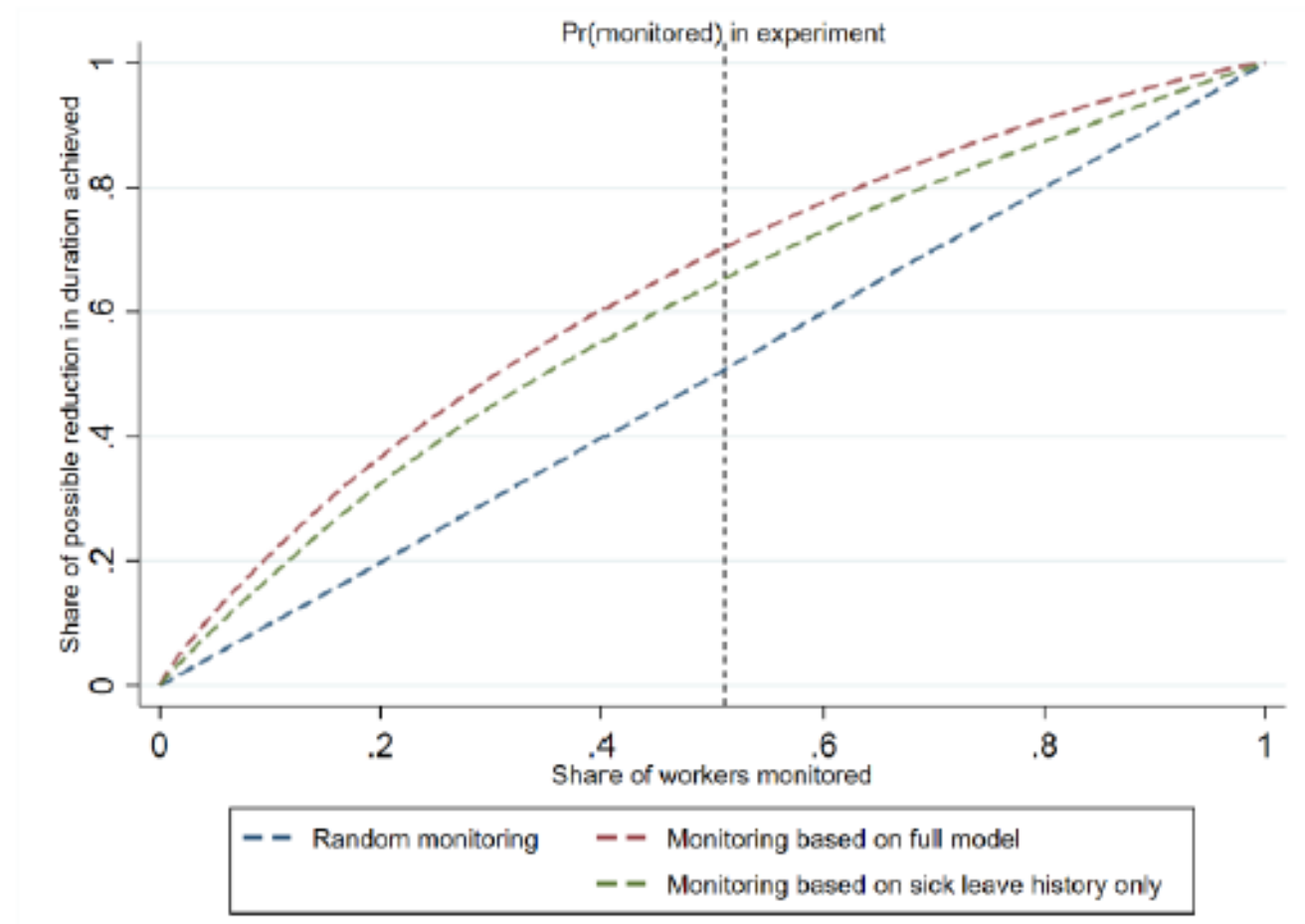
What characteristics predict heterogeneity

FIGURE 8. IMPORTANCE OF WORKER CHARACTERISTICS FOR HETEROGENEITY, BASED ON THE NUMBER OF TIMES THE CAUSAL FOREST'S TREES SPLIT ON THE CHARACTERISTIC



Gains from targeting

FIGURE 11. EFFECT IN TERMS OF REDUCIBLE SICKNESS ABSENCE DURATION FOR GIVEN SHARE OF WORKERS MONITORED ACCORDING TO DIFFERENT MONITORING POLICIES



Flexibility vs interpretability

