

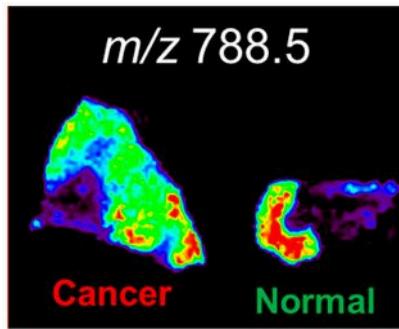
Machine Learning & Data Mining

CS/CNS/EE 155

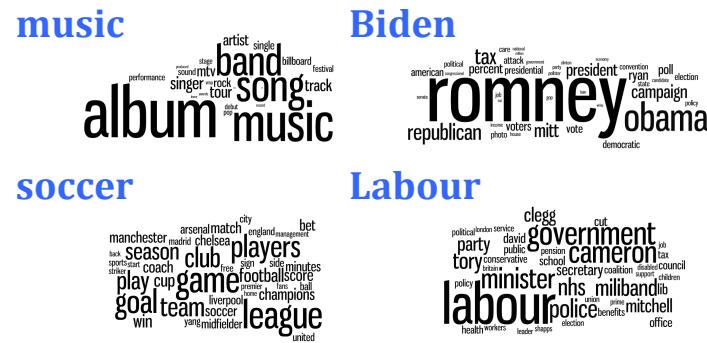
Lecture 4:
Recent Applications of Lasso

Today: Two Recent Applications

Cancer Detection



Personalization via twitter



- Applications of Lasso (and related methods)
 - Think about the data & modeling goals
 - Some new learning problems

Slide material borrowed from Rob Tibshirani and Khalid El-Arini

Image Sources: <http://www.pnas.org/content/111/7/2436>

<https://dl.dropboxusercontent.com/u/16830382/papers/badgepaper-kdd2013.pdf>

Aside: Convexity

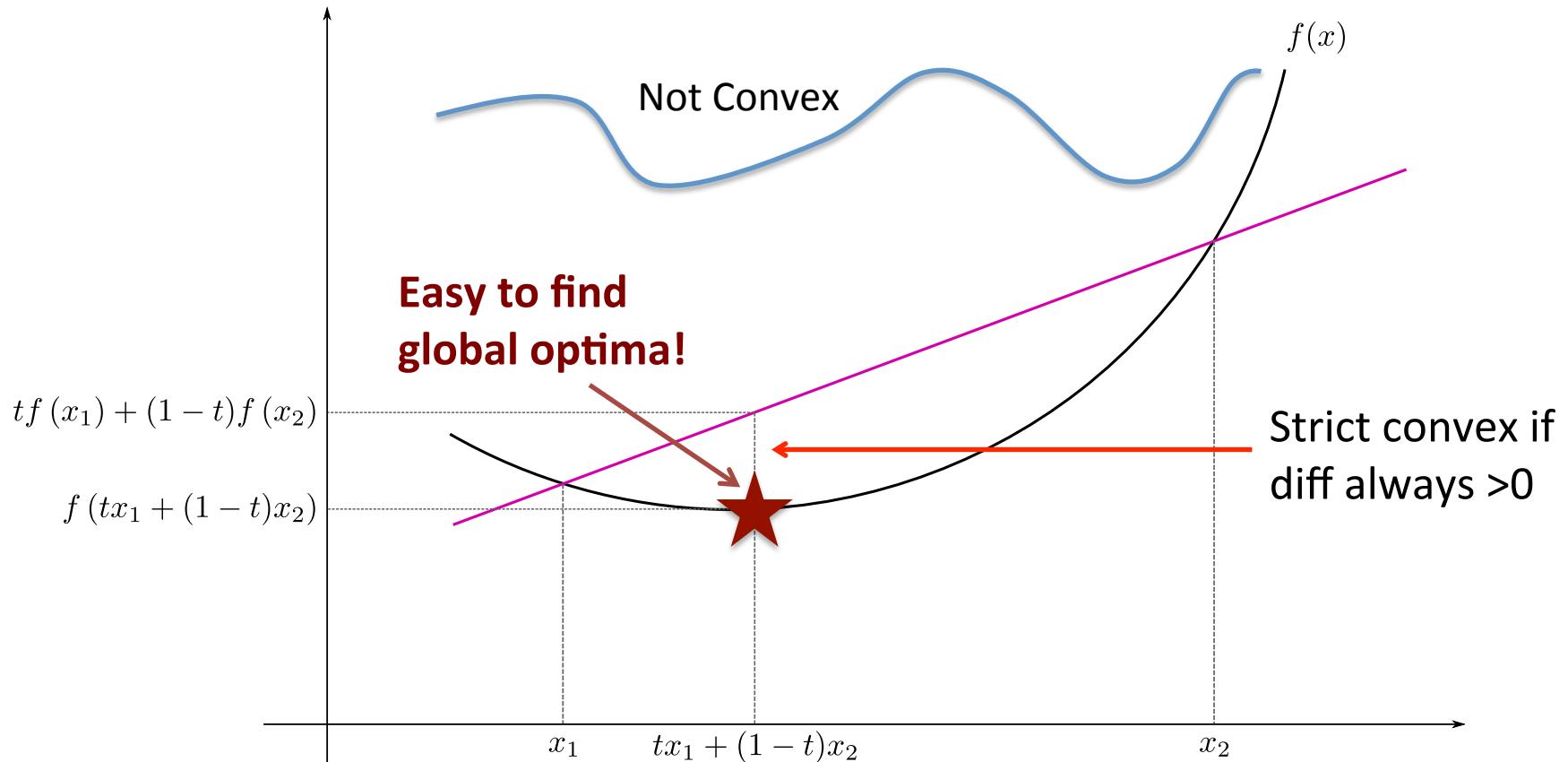


Image Source: http://en.wikipedia.org/wiki/Convex_function

Aside: Convexity

- All local optima are global optima:

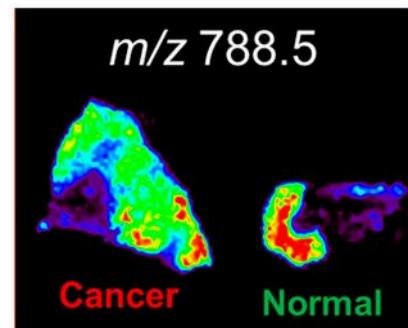


- Strictly convex: unique global optimum:



- Almost all objectives discussed are (strictly) convex:
 - SVMs, LR, Ridge, Lasso... (except ANNs)

Cancer Detection



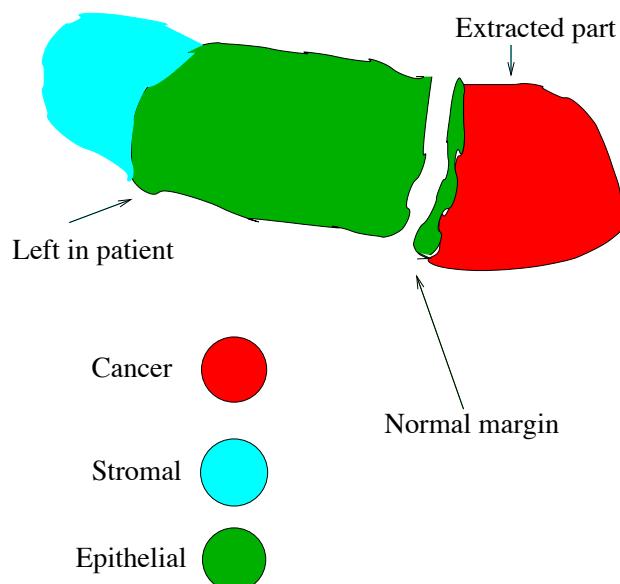
“Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging”

Proceedings of the National Academy of Sciences (2014)

Livia S. Eberlin, Robert Tibshirani, Jialing Zhang, Teri Longacre, Gerald Berry, David B. Bingham, Jeffrey Norton, Richard N. Zare, and George A. Poulsides

<http://www.pnas.org/content/111/7/2436>

<http://statweb.stanford.edu/~tibs/ftp/canc.pdf>

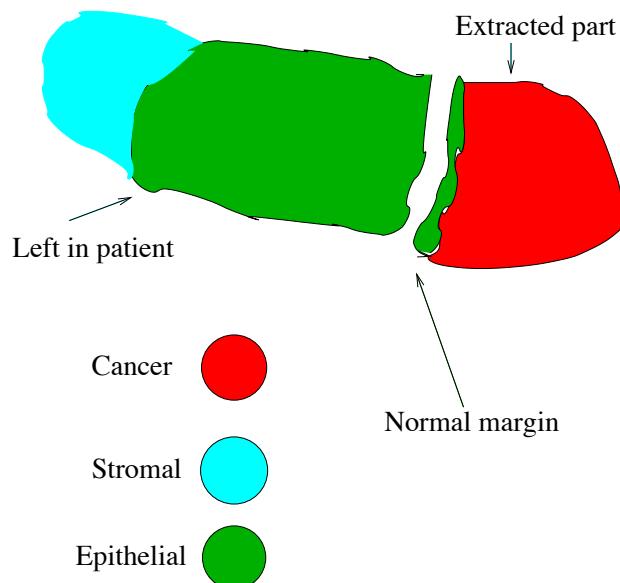


Gastric (Stomach) Cancer

1. Surgeon removes tissue
2. Pathologist examines tissue
 - Under microscope
3. If no margin, GOTO Step 1.

Drawbacks

- **Expensive:** requires a pathologist
- **Slow:** examination can take up to an hour
- **Unreliable:** 20%-30% can't predict on the spot



Gastric (Stomach) Cancer

1. Surgeon removes tissue
2. Pathologist examines tissue
 - Under microscope
3. If no margin, GOTO Step 1.

Machine Learning to the Rescue!

(actually just statistics)

- Lasso originated from statistics community.
 - **But we machine learners love it!**

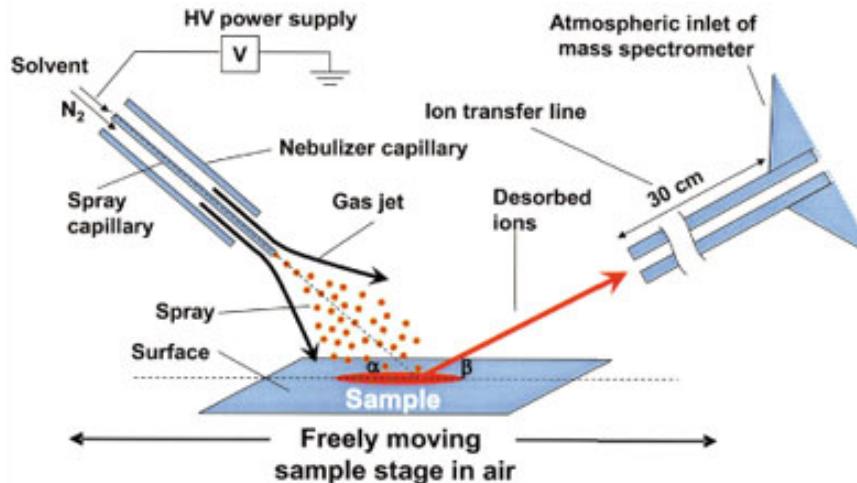
Basic Lasso:

$$\operatorname{argmin}_{w,b} \lambda |w| + \sum_{i=1}^N L(y_i, w^T x_i - b)^2$$

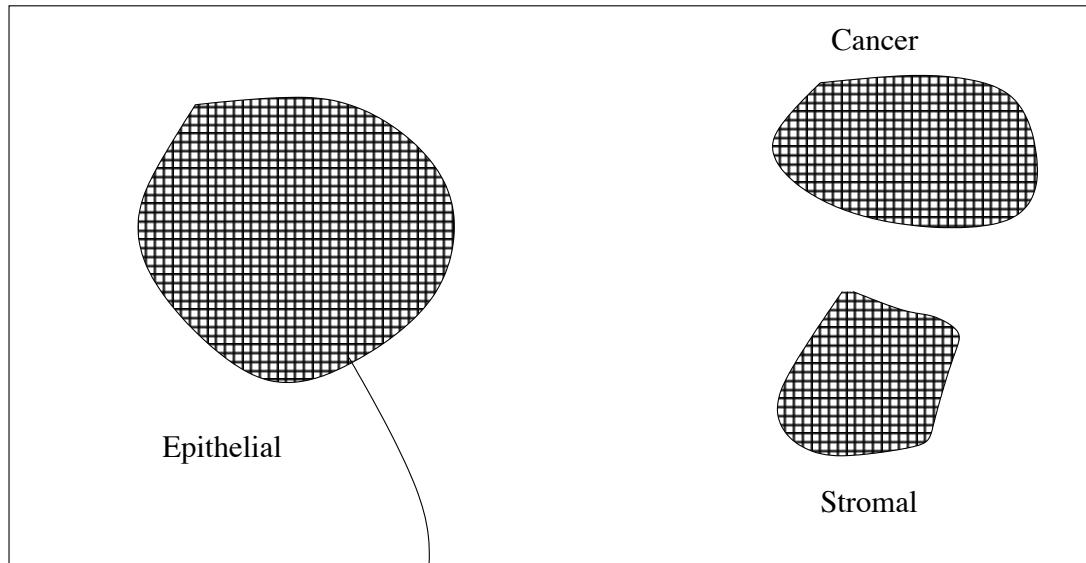
- Train a model to predict cancerous regions!
 - $Y = \{C, E, S\}$ (How to predict 3 possible labels?)
 - What is X ?
 - What is loss function?

Mass Spectrometry Imaging

- DESI-MSI (Desorption Electrospray Ionization)

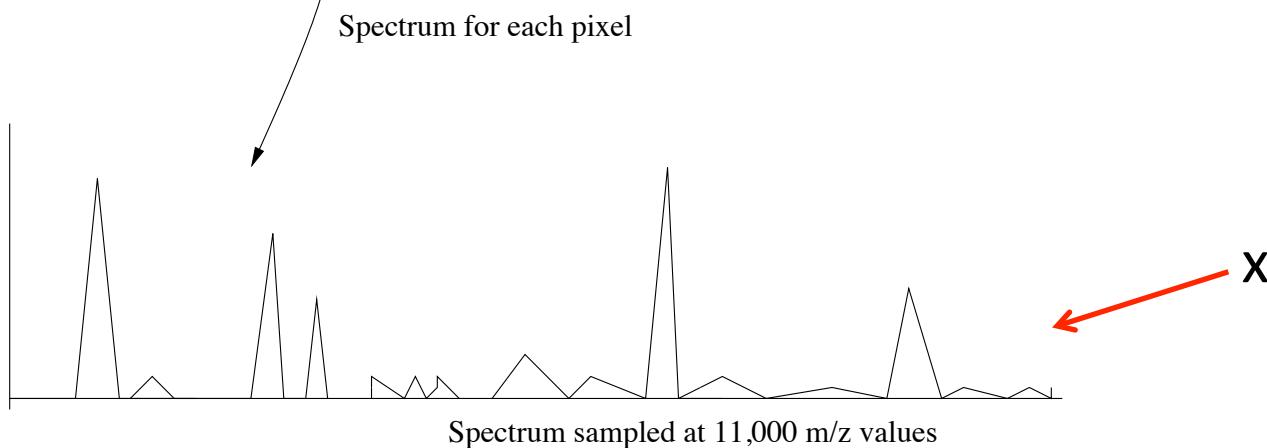


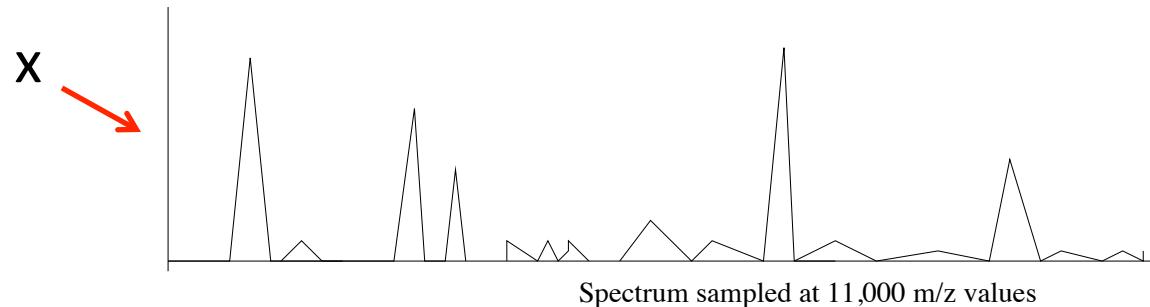
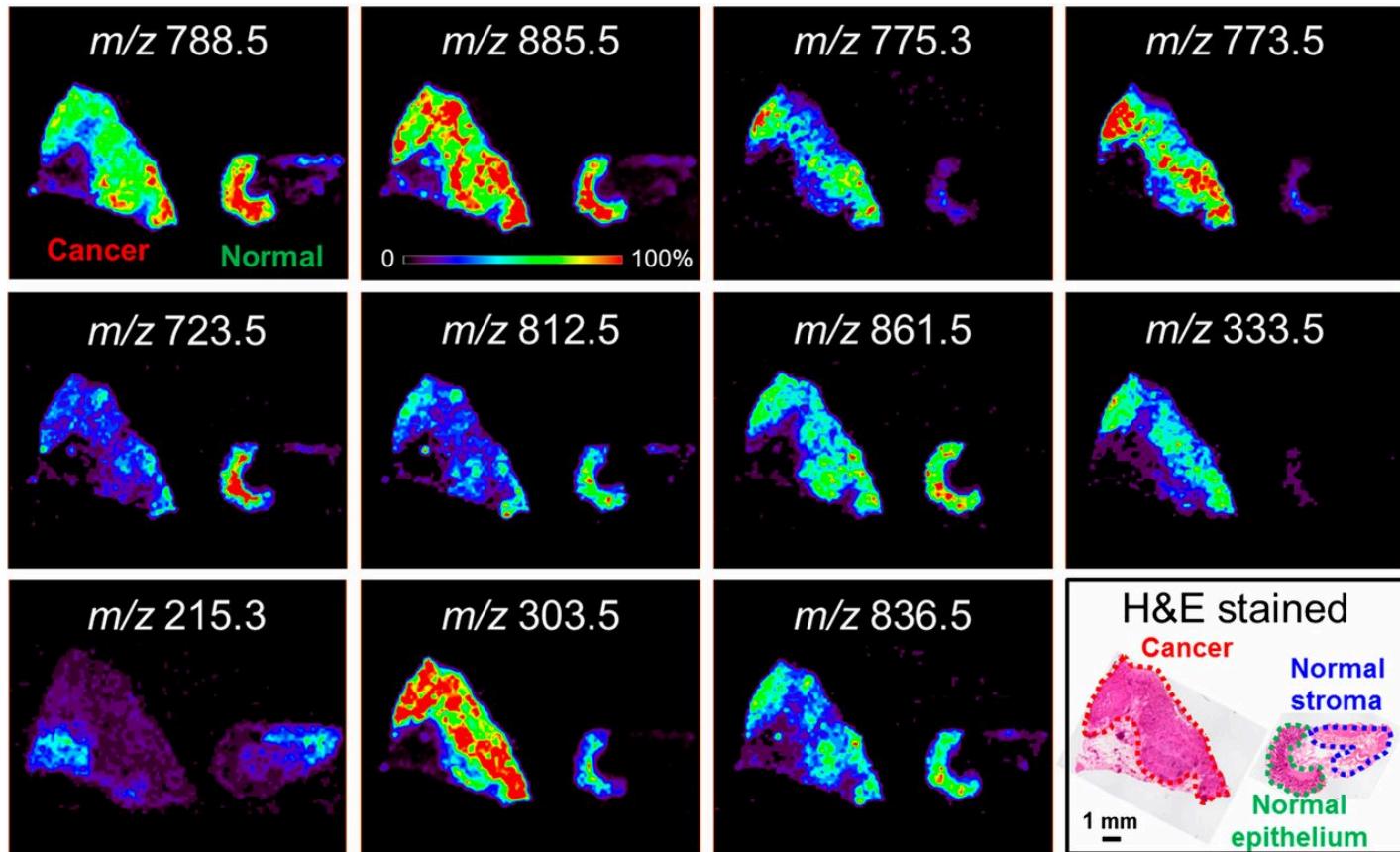
- Effectively runs in real-time (used to generate x)
http://en.wikipedia.org/wiki/Desorption_electrospray_ionization



Each pixel is data point

x via spectroscopy
y via cell-type label





Each pixel has
11K features.
Visualizing a
few features.

Multiclass Prediction

- Multiclass y:

$$S = \{(x_i, y_i)\}_{i=1}^N \quad \begin{array}{l} x \in R^D \\ y \in \{1, 2, \dots, K\} \end{array}$$

- Most common model:

Replicate Weights:

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

Score All Classes:

$$f(x | w, b) = \begin{bmatrix} w_1^T x - b_1 \\ w_2^T x - b_2 \\ \vdots \\ w_K^T x - b_K \end{bmatrix}$$

Predict via Largest Score:

$$\operatorname{argmax}_k \begin{bmatrix} w_1^T x - b_1 \\ w_2^T x - b_2 \\ \vdots \\ w_K^T x - b_K \end{bmatrix}$$

- Loss function?

Multiclass Logistic Regression

Binary LR: $P(y|x, w, b) = \frac{e^{y(w^T x - b)}}{e^{y(w^T x - b)} + e^{-y(w^T x - b)}}$ $y \in \{-1, +1\}$

“Log Linear” Property: $P(y|x, w, b) \propto e^{y(w^T x - b)}$ $(w_1, b_1) = (-w_{-1}, -b_{-1})$

Extension to Multiclass: $P(y = k|x, w, b) \propto e^{w_k^T x - b_k}$ Keep a (w_k, b_k) for each class

Multiclass LR: $P(y = k|x, w, b) = \frac{e^{w_k^T x - b_k}}{\sum_m e^{w_m^T x - b_m}}$

Referred to as Multinomial Log-Likelihood by Tibshirani

<http://statweb.stanford.edu/~tibs/ftp/canc.pdf>

Multiclass Log Loss

$$\operatorname{argmin}_{w,b} \sum_i -\ln P(y_i | x_i, w, b) \quad \begin{array}{l} x \in R^D \\ y \in \{1, 2, \dots, K\} \end{array}$$

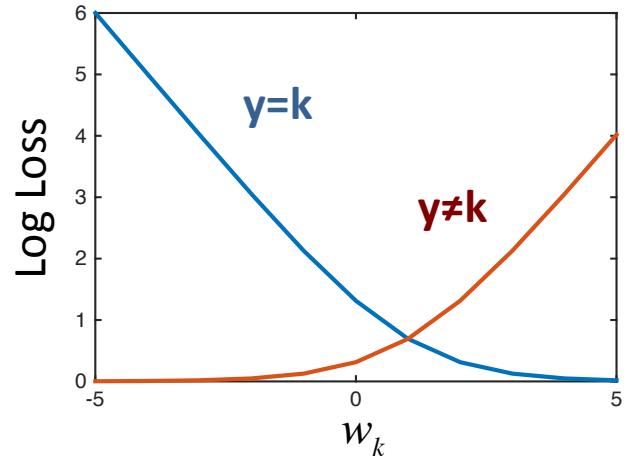
$$P(y | x, w, b) = \frac{e^{w_y^T x - b_y}}{\sum_m e^{w_m^T x - b_m}} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

$$-\ln P(y | x, w, b) = -w_y^T x + b_y + \ln \left(\sum_m e^{w_m^T x - b_m} \right)$$

$$\partial_{w_k} -\ln P(y | x, w, b) = \begin{cases} (-1 + P(y | x, w, b))x & \text{if } y = k \\ P(y | x, w, b)x & \text{if } y \neq k \end{cases}$$

Multiclass Log Loss

- Suppose $x=1$ & ignore b
 - Model score is just w_k
 - Vary one weight, others = 1



$$-\ln P(y|x, w, b) = -w_y^T x + b_y + \ln \left(\sum_m e^{w_m^T x - b_m} \right)$$

$$\partial_{w_k} -\ln P(y|x, w, b) = \begin{cases} (-1 + P(y|x, w, b))x & \text{if } y = k \\ P(y|x, w, b)x & \text{if } y \neq k \end{cases}$$

Lasso Multiclass Logistic Regression

$$\operatorname{argmin}_{w,b} \lambda |w| + \sum_i -\ln P(y_i | x_i, w, b) \quad \begin{array}{l} x \in R^D \\ y \in \{1, 2, \dots, K\} \end{array}$$

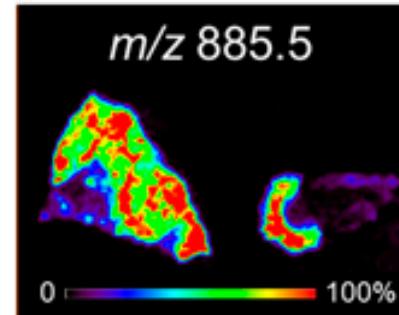
$$|w| = \sum_k |w_k| = \sum_k \sum_d |w_{kd}|$$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

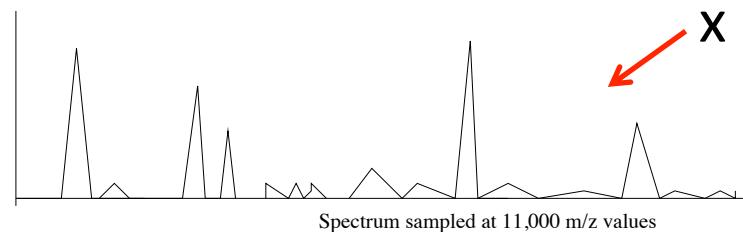
- Probabilistic model
- Sparse weights

Back to the Problem

- Image Tissue Samples
- Each pixel is an x
 - 11K features via Mass Spec
 - Computable in real time
 - 1 prediction per pixel
- y via lab results
 - ~2 weeks turn-around



Visualization of all pixels for one feature

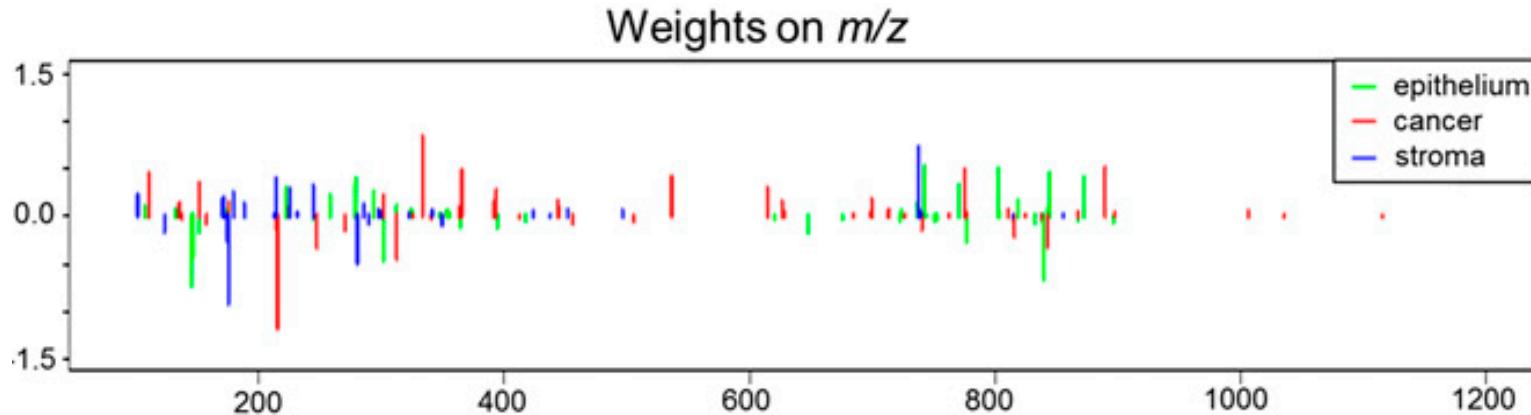


Learn a Predictive Model

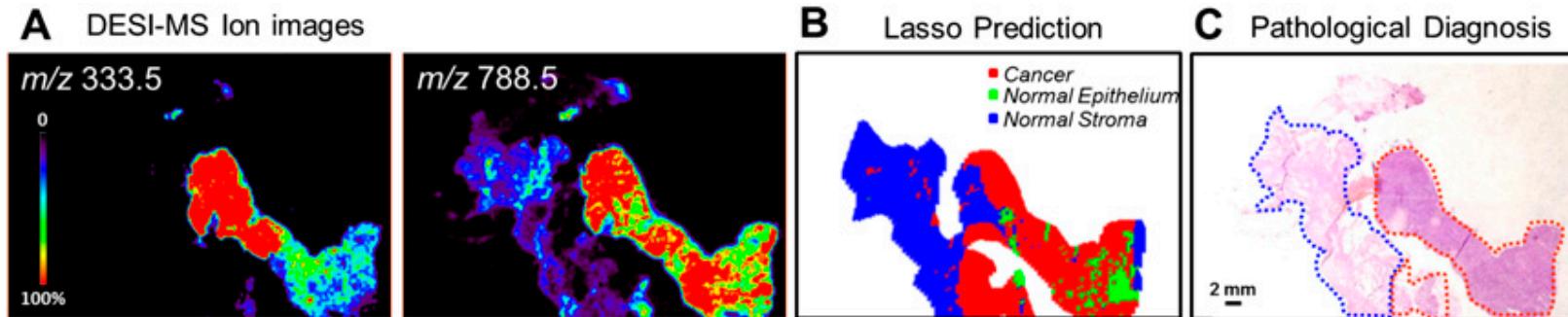
- Training set: 28 tissue samples from 14 patients
 - Cross validation to select λ
- Test set: 21 tissue samples from 9 patients
- Test Performance:

Pathology	Predicted				Agreement, %	Overall agreement, %
	Cancer	Epithelium	Stroma	Don't know		
Cancer	5,809	114	2	230	97.0	97.2
Epithelium	134	3,566	118	122	96.8	
Stroma	25	82	2,630	143	96.1	
Cancer		Normal			Agreement, %	Overall agreement, %
Cancer	5,809	116		230	97.0	98.4
Normal	159	6,396		265	99.7	

≥ 0.2 margin
in probability



- **Lasso yields sparse weights! (Manual Inspection Feasible!)**
- Many correlated features
 - Lasso tends to focus on one



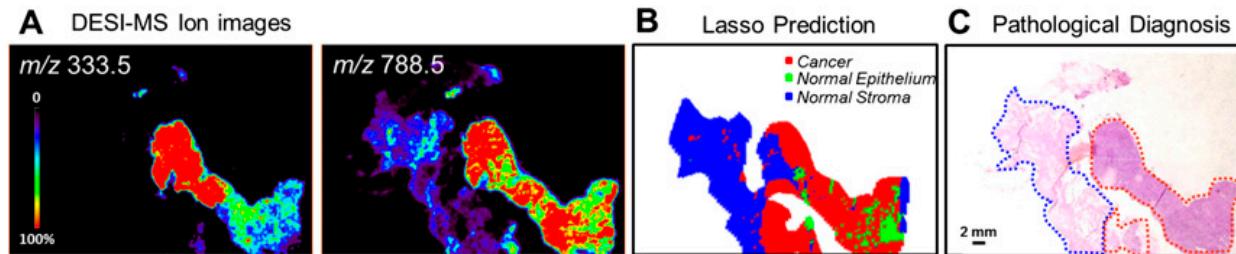
Extension: Local Linearity

$$P(y|x, w, b) = \frac{e^{w_y^T x - b_y}}{\sum_m e^{w_m^T x - b_m}}$$

- Assumes probability shifts along straight line
 - Often not true
- **Approach:** cluster based on x
 - Train customized model for each cluster

Patient	1	2	3	4	5	6	Overall
Standard training	0.29%	4.56%	6.78%	0.00%	13.76%	2.77%	3.58%
Customized training	0.71%	1.89%	0.82%	0.40%	9.43%	0.92%	1.89%

Recap: Cancer Detection



- Seems Awesome! What's the catch?
 - Small sample size
 - Tested on 9 patients
 - Machine Learning only part of the solution
 - Need infrastructure investment, etc.
 - Analyze the scientific legitimacy
 - Social/Political/Legal
 - If there is mis-prediction, who is at fault?

Personalization via twitter

music



band
song
track
album
music

A word cloud centered around the word "music". Other visible words include band, song, track, album, and music.

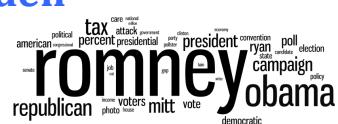
soccer



manchester
arsenal
match
city
bet
season
club
players
coach
striker
play
cup
game
goal
team
win
soccer
league
united

A word cloud centered around the word "soccer". Other visible words include manchester, arsenal, bet, season, club, players, coach, striker, play, cup, game, goal, team, win, soccer, league, and united.

Biden



american
political
percent
presidential
photo
voters
mitt
vote
republican
democratic

A word cloud centered around the word "Biden". Other visible words include american, political, percent, presidential, photo, voters, mitt, vote, republican, and democratic.

Labour



clegg
political
border
services
party
tory
minister
labour
police
health workers
leader
slaps
election
benefits
miliband
mitchell
office

A word cloud centered around the word "Labour". Other visible words include clegg, political, border, services, party, tory, minister, labour, police, health workers, leader, slaps, election, benefits, miliband, mitchell, and office.

“Representing Documents Through Their Readers”

Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (2013)

Khalid El-Arini, Min Xu, Emily Fox, Carlos Guestrin

<https://dl.dropboxusercontent.com/u/16830382/papers/badgepaper-kdd2013.pdf>

The Washington Post

THE HUFFINGTON POST
THE INTERNET NEWSPAPER: NEWS BLOGS VIDEO COMMUNITY

The New York Times

the guardian



ft.com/frontpage UK All times are London time
FINANCIAL TIMES

THE DAILY BEAST

FOX NEWS



Slate

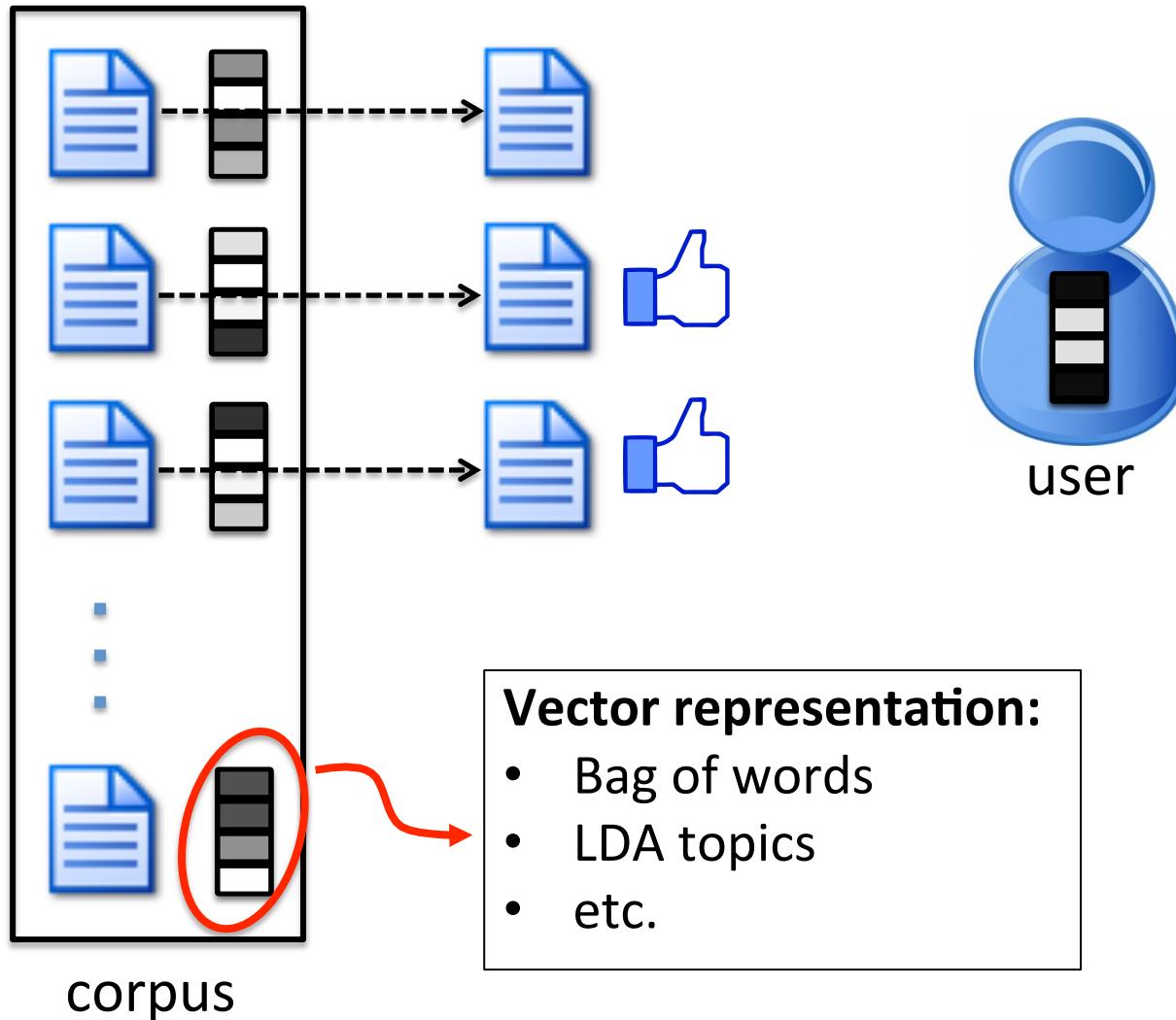
TC

overloaded by news

≥ 1 million news articles & blog posts generated every hour*

* [www.spinn3r.com]

News Recommendation Engine



Challenge

Most common representations don't naturally line up with user interests



Fine-grained representations (**bag of words**) **too specific**

Haqqani network is considered most ruthless branch of Afghan insurgency
Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion



High-level topics (e.g., from a topic model)

- too fuzzy and/or vague
- can be inconsistent over time

Goal

Improve recommendation
performance through a
more natural document
representation

An Opportunity: News is Now Social

- In 2012, Guardian announced more readers visit site via Facebook than via Google search

Other Agencies Clamor for Data N.S.A. Compiles

By ERIC LICHTBLAU and MICHAEL S. SCHWARTZ

Published: August 3, 2013

238 Comments

WASHINGTON — The [National Security Agency's](#) dominant role as the nation's spy warehouse has spurred frequent tensions and turf fights with other federal intelligence agencies that want to use its surveillance tools for their own investigations, officials say.

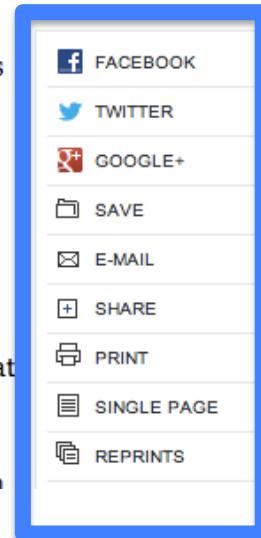
Connect With Us on Twitter



Follow
@NYTNational for breaking news and headlines.

Twitter List: Reporters and Editors

Agencies working to curb drug trafficking, cyberattacks, money laundering, counterfeiting and even copyright infringement complain that their attempts to exploit the security agency's vast resources have often been turned down because their own



Log in to see what your friends are sharing on [Log In With Facebook](#)
[nytimes.com](#). [Privacy Policy](#) | [What's This?](#)

What's Popular Now [f](#)

Cory Booker for Senator  Michael Ansara, Actor Who Played Cochise and Kang, Dies at 91 



Substandard Nerd

@substandardnerd

Gig Going, Festival Attending, Music Loving, Linux Fettling, Perl Hacking, Cycling, Vegan

The Gdansk of Oxfordshire .



<https://www.youtube.com/user/apusskidu/featured>



Substandard Nerd @substandardnerd

13 Jan

Stevie Nicks: the return of Fleetwood Mac

guardian.co.uk/music/2013/jan...

View summary

Approach

Learn a document representation based on how readers publicly describe themselves

Substandard Nerd

@substandardnerd

Gig Going, Festival Attending, Music Loving, Linux Fettling, Perl Hacking, Cycling, Vegan

The Gdansk of Oxfordshire .

<https://www.youtube.com/user/apusskidu/featured>



Substandard Nerd @substandardnerd

13 Jan

Stevie Nicks: the return of Fleetwood Mac

[guardian.co.uk](https://www.theguardian.com/music/2013/jan/13/stevie-nicks-fleetwood-mac-interview)

[View summary](#)

Culture > Music > Stevie Nicks

Stevie Nicks: the return of Fleetwood Mac

Stevie Nicks's tumultuous life as a rock queen led her to addiction, heartbreak and "insanity". As Fleetwood Mac reunite, she tells Caspar Llewellyn Smith why she's going back for more

Using **many** tweets, can we learn
that someone who identifies with

via profile badges → **music**

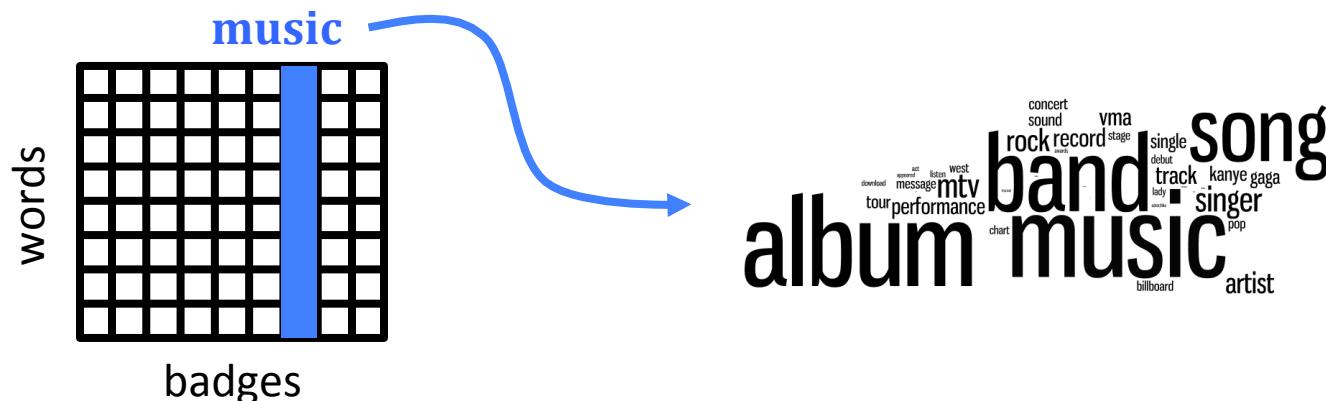
reads articles with these words:



Given: training set of tweeted news articles from a specific period of time

3 million articles

1. Learn a badge dictionary from training set



2. Use badge dictionary to **encode** new articles

Haqqani network is considered most ruthless branch of Afghan insurgency

Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion



afghanistan islam security
east conflict disabled
guardian divorced adult pakistan
adult

Advantages

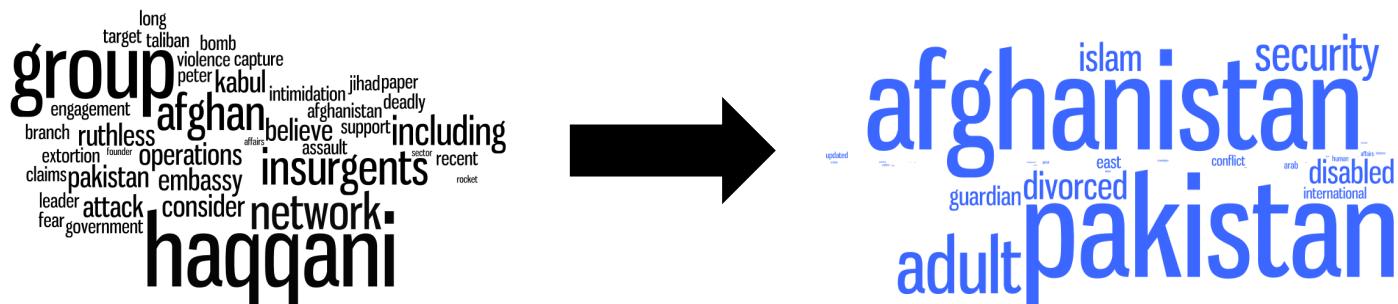
- Interpretable
 - Clear labels
 - Correspond to user interests
- Higher-level than words

Advantages

- Interpretable
 - Clear labels
 - Correspond to user interests

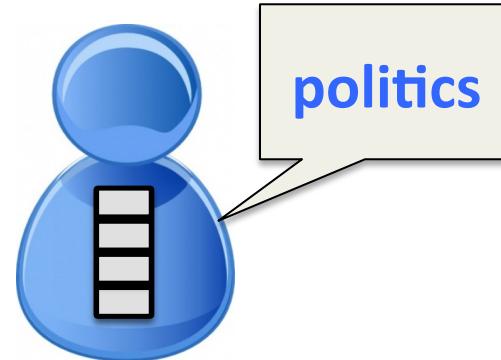
Haqqani network is considered most ruthless branch of Afghan insurgency

Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion



Advantages

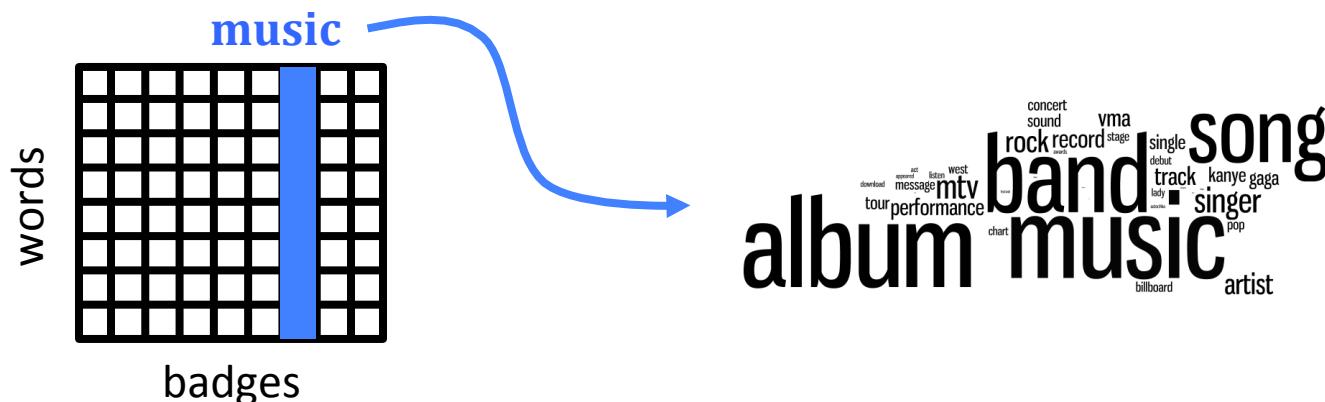
- Interpretable
 - Clear labels
 - Correspond to user interests
- Higher-level than words
- Semantically consistent over time



Given: training set of tweeted news articles from a specific period of time

3 million articles

1. Learn a **badge dictionary** from training set



2. Use badge dictionary to encode new articles

Haqqani network is considered most ruthless branch of Afghan insurgency

Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion



afghanistan islam security
east conflict arab disabled
guardian divorced international
adult pakistan

Dictionary Learning

- Training data :

Identifies badges
in Twitter profile
of tweeter

$$S = \left\{ (z_i, y_i) \right\}_{i=1}^N$$

Bag-of-words
representation of
document

Culture > Music > Stevie Nicks

Stevie Nicks: the return of Fleetwood Mac

Stevie Nicks's tumultuous life as a rock queen led her to addiction, heartbreak and "insanity". As Fleetwood Mac reunite, she tells Caspar Llewellyn Smith why she's going back for more

Substandard Nerd

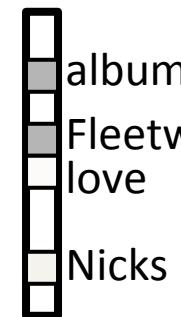
@substandardnerd

Gig Going, Festival Attending, Music Loving, Linux Fettling, Perl Hacking, Cycling, Vegan

The Gdansk of Oxfordshire

<https://www.youtube.com/user/apusskido/featured>

y



z



Normalized!

Dictionary Learning

$$S = \{(z_i, y_i)\}_{i=1}^N$$

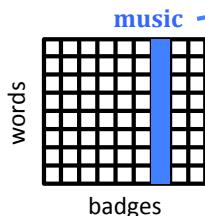
Identifies badges in Twitter profile of tweeter

Bag-of-words representation of document

- Training Objective:

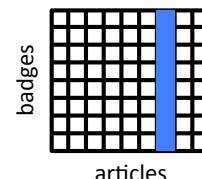
$$\operatorname{argmin}_{B,W} \lambda_B |B| + \lambda_W |W| + \sum_{i=1}^N \|y_i - BW_i\|^2$$

“Dictionary”



Haqqani network is considered most ruthless branch of Afghan insurgency
Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion

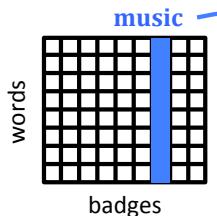
“Encoding”



afghanistan islam security
adult pakistan
guardian divorced sex
disabled person

$$\underset{B, W}{\operatorname{argmin}} \lambda_B |B| + \lambda_W |W| + \sum_{i=1}^N \|y_i - BW_i\|^2$$

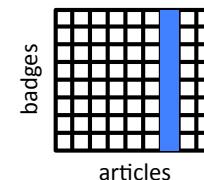
“Dictionary”



music
album band music song
concert tour performance
messengers mta
single track kanye gaga
singer artist

Haqqani network is considered most ruthless branch of Afghan insurgency
Group that started as part of anti-Soviet jihad has moved into mafia-like violence, intimidation and extortion

“Encoding”



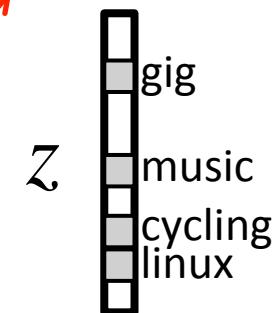
afghanistan islam security
islam guardian divorced est conflict
adult pakistan disabled international

- Not convex! (because of BW term)
- Convex if only optimize B or W (but not both)
- Alternating Optimization (between B and W)
- How to initialize?

Use: $S = \{(z_i, y_i)\}_{i=1}^N$

Initialize:

$$W_i = \frac{z_i}{|z_i|}$$



$$\underset{B,W}{\operatorname{argmin}} \lambda_B |B| + \lambda_W |W| + \sum_{i=1}^N \|y_i - BW_i\|^2$$

- Suppose Badge s often co-occurs with Badge t
 - B_s & B_t are correlated
- From perspective of W , B 's are features.
 - Lasso tends to focus on one correlated feature

Substandard Nerd

@substandardnerd

Gig Going, Festival Attending, Music Loving, Linux Fettling, Perl Hacking, Cycling, Vegan

The Gdansk of Oxfordshire .

<https://www.youtube.com/user/apusskidu/featured>

Many articles might be about Gig, Festival & Music simultaneously.

$$\operatorname{argmin}_{B,W} \lambda_B |B| + \lambda_W |W| + \sum_{i=1}^N \|y_i - BW_i\|^2$$

- Suppose Badge s often co-occurs with Badge t
 - B_s & B_t are correlated
- From perspective of W, B's are features.
 - Lasso tends to focus on one correlated feature
- Graph Guided Fused Lasso:

$$\operatorname{argmin}_{B,W} \lambda_B |B| + \lambda_W |W| + \lambda_G \sum_{i=1}^N \sum_{(s,t) \in E(G)} \omega_{st} |W_{is} - W_{it}| + \sum_{i=1}^N \|y_i - BW_i\|^2$$

↗
Graph G of related Badges
Co-occurrence Rate
On Twitter Profiles

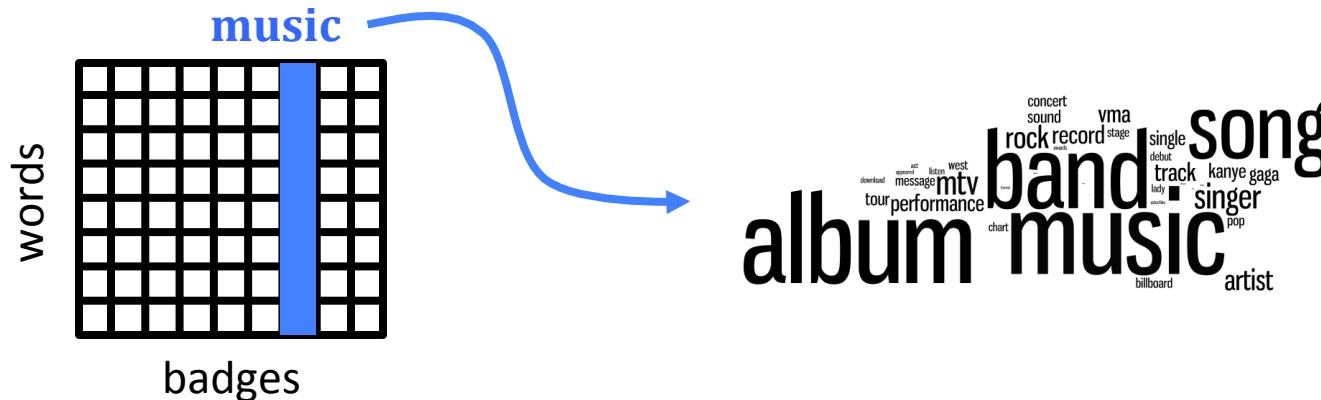
Encoding New Articles

- Badge Dictionary B is already learned
- Given a new document j with word vector y_j
 - Learn Badge Encoding W_j :

$$\operatorname{argmin}_{W_j} \lambda_W |W_j| + \lambda_G \sum_{(s,t) \in G} |W_{js} - W_{jt}| + \|y_j - BW_j\|^2$$

Recap: Badge Dictionary Learning

1. Learn a **badge dictionary** from training set



2. Use badge dictionary to **encode new articles**

Haqqani network is considered most
ruthless branch of Afghan insurgency
Group that started as part of anti-Soviet jihad has moved into
mafia-like violence, intimidation and extortion



afghanistan
islam security
pakistan
adult
divorced
guardian
east
conflict
arab
international
disabled
adult
islam
security
pakistan
adult
divorced
guardian
east
conflict
arab
international
disabled

Examining B

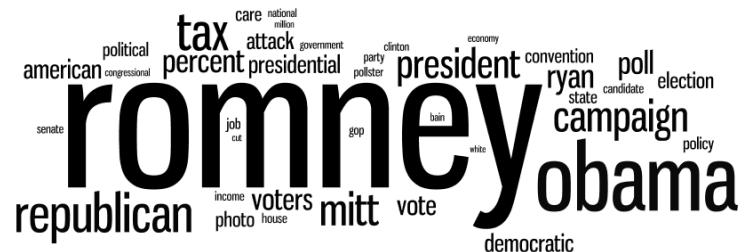
music



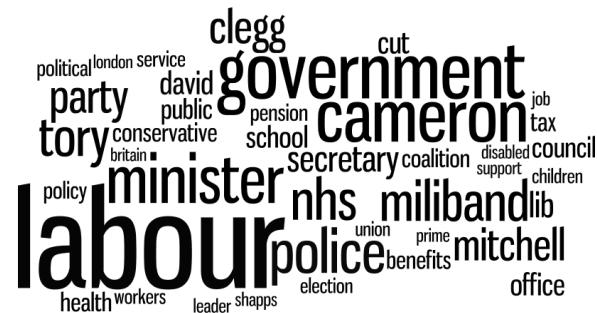
soccer



Biden

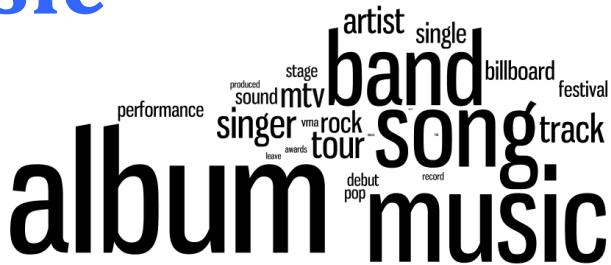


Labour



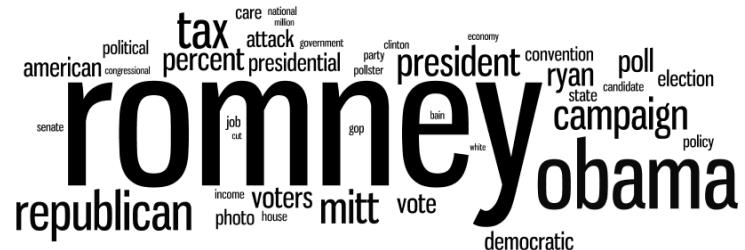
Badges Over Time

music



September 2012

Biden



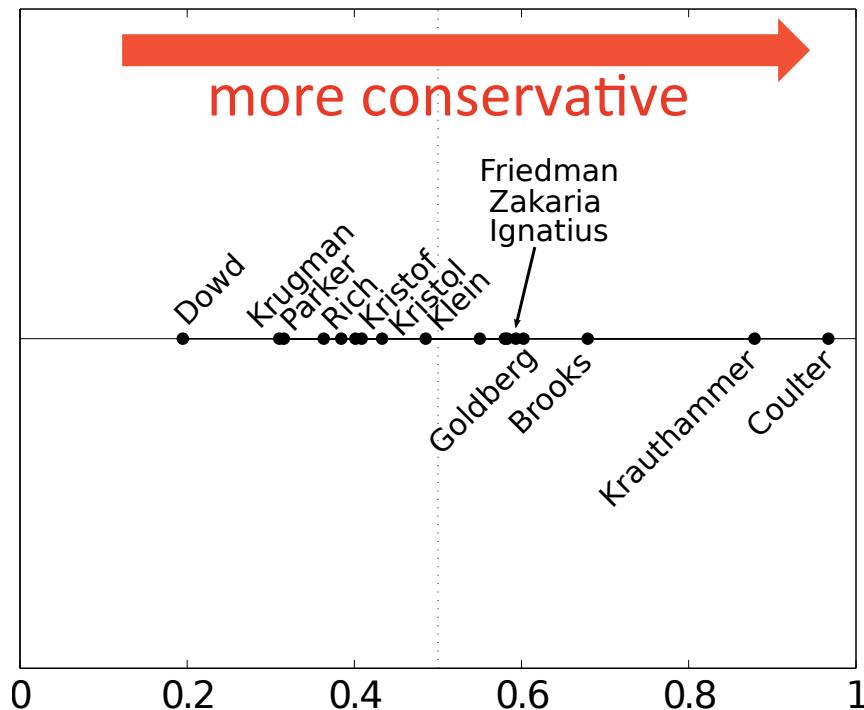
September 2010

biden administration party cowboys削弱 forgotpasswordform house fundraising thanksgiving whining hostess republican senator president screeching vice clegg iraq morgan iran election hamas twinkies white gop register joe voters registerform thanks outfront obama frontpage
democrats post-season union est piers

A Spectrum of Pundits

“top conservatives on Twitter”

- Limit badges to **progressive** and **TCOT**
- Predict political alignments of likely readers?

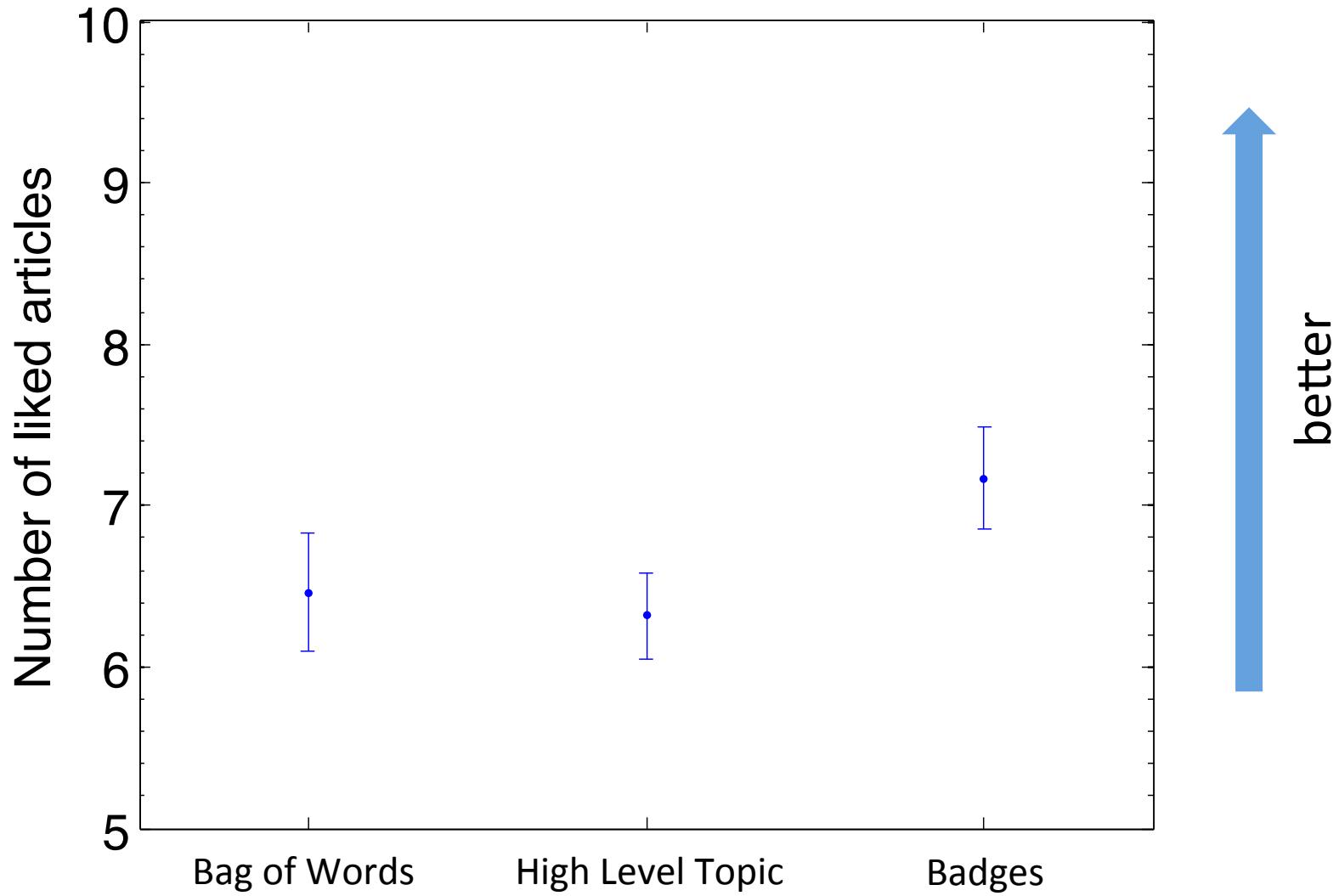


- Took all articles by columnist
- Looked at encoding score
 - progressive vs TCOT
- Average

User Study

- Which representation best captures user preferences over time?
- Study on Amazon Mechanical Turk with 112 users
 1. Show users random 20 articles from Guardian, from time period 1, and obtain ratings
 2. Pick random representation
 - bag of words, high level topic, Badges
 3. Represent user preferences as mean of liked articles
 4. GOTO next time period
 - Recommend according to preferences
 - GOTO STEP 2

User Study



Recap: Personalization via twitter

- Sparse Dictionary Learning
 - Learn a new representation of articles
 - Encode articles using dictionary
 - Better than Bag of Words
 - Better than High Level Topics
- Based on social data
 - Badges on twitter profile & tweeting
 - Semantics not directly evident from text alone

Next Week

- Sequence Prediction
- Hidden Markov Models
- Conditional Random Fields
- Homework 1 due Tues 1/20 @5pm
 - via Moodle