Short Course
Robust Optimization and Machine Learning

Lecture 7: Sparse Machine Learning for Text Analytics

Laurent El Ghaoui

EECS and IEOR Departments
UC Berkeley

Spring seminar TRANSP-OR, Zinal, Jan. 16-19, 2012

Outline

Outline

Information Overload

Avalanche of “information” in text format, *e.g.*

- News articles, press releases, RSS feeds, TV captioning data.
- 10-K filings, marketing brochures, financial analyst reports, and other company-related documents.
- Consumer reviews, blogs, emails, and other social media content.
- Scientific papers, patents, law documents, bills, medical reports, literature.

Information Overload

Avalanche of “information” in text format, *e.g.*

- News articles, press releases, RSS feeds, TV captioning data.
- 10-K filings, marketing brochures, financial analyst reports, and other company-related documents.
- Consumer reviews, blogs, emails, and other social media content.
- Scientific papers, patents, law documents, bills, medical reports, literature.

The top *20* most important news sources have generated $\sim$ *40,000* news articles yesterday.

What might be useful?

- *Summarize* large text databases.
- Detect and visualize *trends* in term usage.
- *Compare* how topics of interest are treated across different sources.
- *Group* similar text documents.
- Provide *interpretable* visualizations.

What might be useful?

- *Summarize* large text databases.
- Detect and visualize *trends* in term usage.
- *Compare* how topics of interest are treated across different sources.
- *Group* similar text documents.
- Provide *interpretable* visualizations.

Approach: *sparse machine learning* tools to help in these tasks.

Outline

Sparse Machine Learning

Let X be the term-by-document matrix, y a label vector.

- Cardinality-penalized *least-squares* :

$$\min_{w,b} \|X^T w + b\mathbf{1} - y\|_2^2 + \lambda \mathbf{Card}(w)$$

- Cardinality-penalized *low-rank approximations* (or, sparse PCA):

$$\min_{w,v} \|X - wv^T\|_F + \lambda \mathbf{Card}(w) + \mu \mathbf{Card}(v).$$

Despite the hardness of these problems, we can solve them heuristically, at very high scale.

Why are these problems scalable?

Both of these problems allow to eliminate features (or documents) prior to solving the problem at very cheap cost, in an *embarrassingly parallel* way. This is known as a SAFE elimination procedure.

For example, for the sparse low-rank approximation problem it can be proven that no feature appears if the corresponding variance is less than the penalty parameter λ.

SAFE for sparse PCA

The sparse PCA problem can be expressed as ($n =$ number of features)

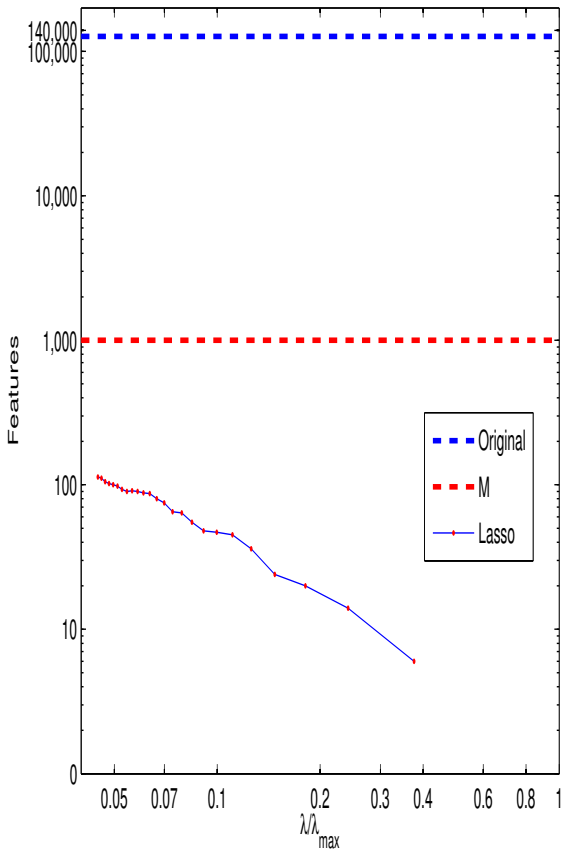
$$\max_{z\,:\,\|z\|_2=1} \sum_{i=1}^{n} \max\left((x_i^T z)^2 - \lambda, 0\right).$$

- x_i is the data for the i-th feature (i-th column of matrix X).
- For no cardinality penalty ($\lambda = 0$), reduces to an eigenvalue problem.
- When $\lambda > 0$, i-th feature can be removed safely when its variance $= \|x_i\|_2 < \lambda$.

In our text applications, cardinality penalty λ is *very high* , allowing to greatly reduce the size of the problem.

SAFE for the LASSO

LASSO is a convex approximation to cardinality-penalized least-squares. There is a SAFE procedure for it.



PubMed data has 3M documents, 150K words in dictionary. SAFE for LASSO brings downs that number to 1000. This allows to load the data and solve the LASSO problem.

Outline

What is topic imaging?

Topic image: A small set of terms that are semantically related to a given topic ("the query").

As a predictive problem: predict appearance of query term in a document given the term use in that document.

What is topic imaging?

Topic image: A small set of terms that are semantically related to a given topic ("the query").

As a predictive problem: predict appearance of query term in a document given the term use in that document.

- Predictive model must be *interpretable:* number of predictors (other terms) must be few (sparse classification).

What is topic imaging?

Topic image: A small set of terms that are semantically related to a given topic ("the query").

As a predictive problem: predict appearance of query term in a document given the term use in that document.

- Predictive model must be *interpretable:* number of predictors (other terms) must be few (sparse classification).
- Model must be obtained *fast* .

Occurence vs. Classification

Two NYT op-ed columnists

Occurence analysis : top 10 words

Nicholas Kristof	Roger Cohen
mr	obama
people	iran
obama	said
said	american
president	president
world	iranian
new	israel
american	states
years	new
united	united

Occurence vs. Classification

Two NYT op-ed columnists

Occurence analysis : top 10 words

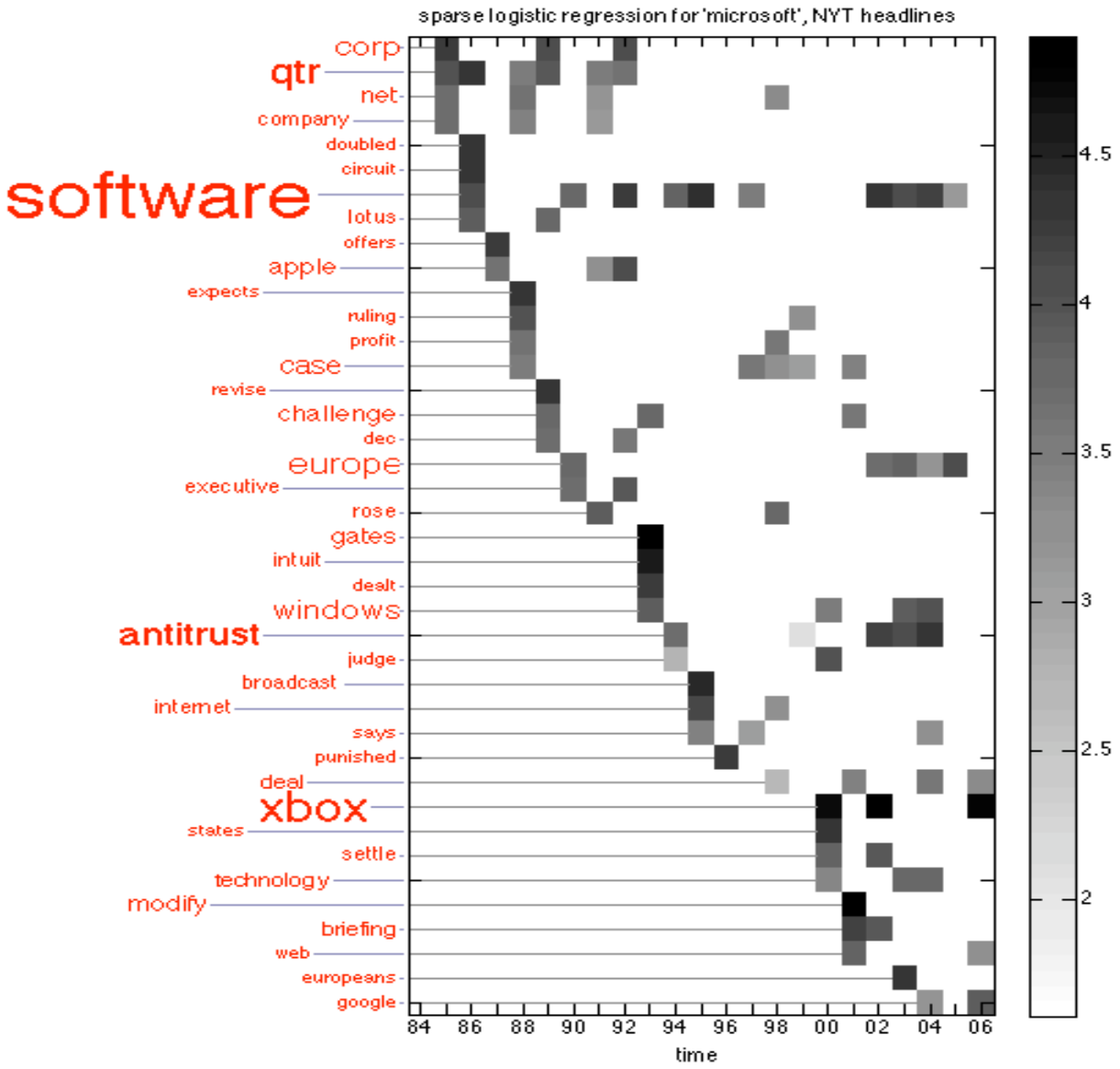
Nicholas Kristof	Roger Cohen
mr	obama
people	iran
obama	said
said	american
president	president
world	iranian
new	israel
american	states
years	new
united	united

Sparse classification

Nicholas Kristof	Roger Cohen
videos	olmert
darfur	persian
antibiotics	chemical
facebook	mohammad
sudanese	ali
janjaweed	dialogue
youtube	cease
sudan	iranian
sweatshops	tehran
invite	holocaust

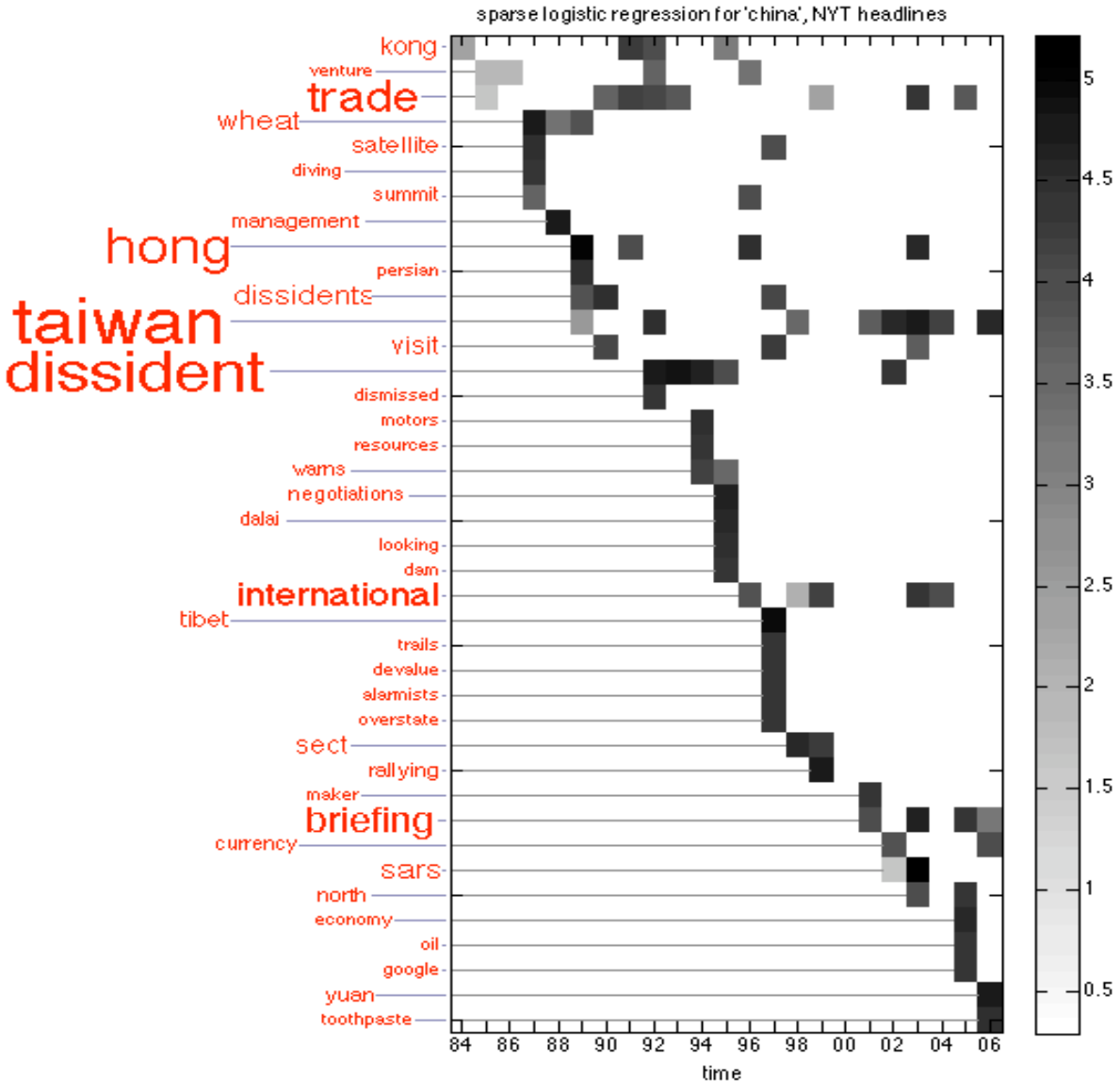
Image across time: “Microsoft”

Data: The New York Times headlines, 1985-2007



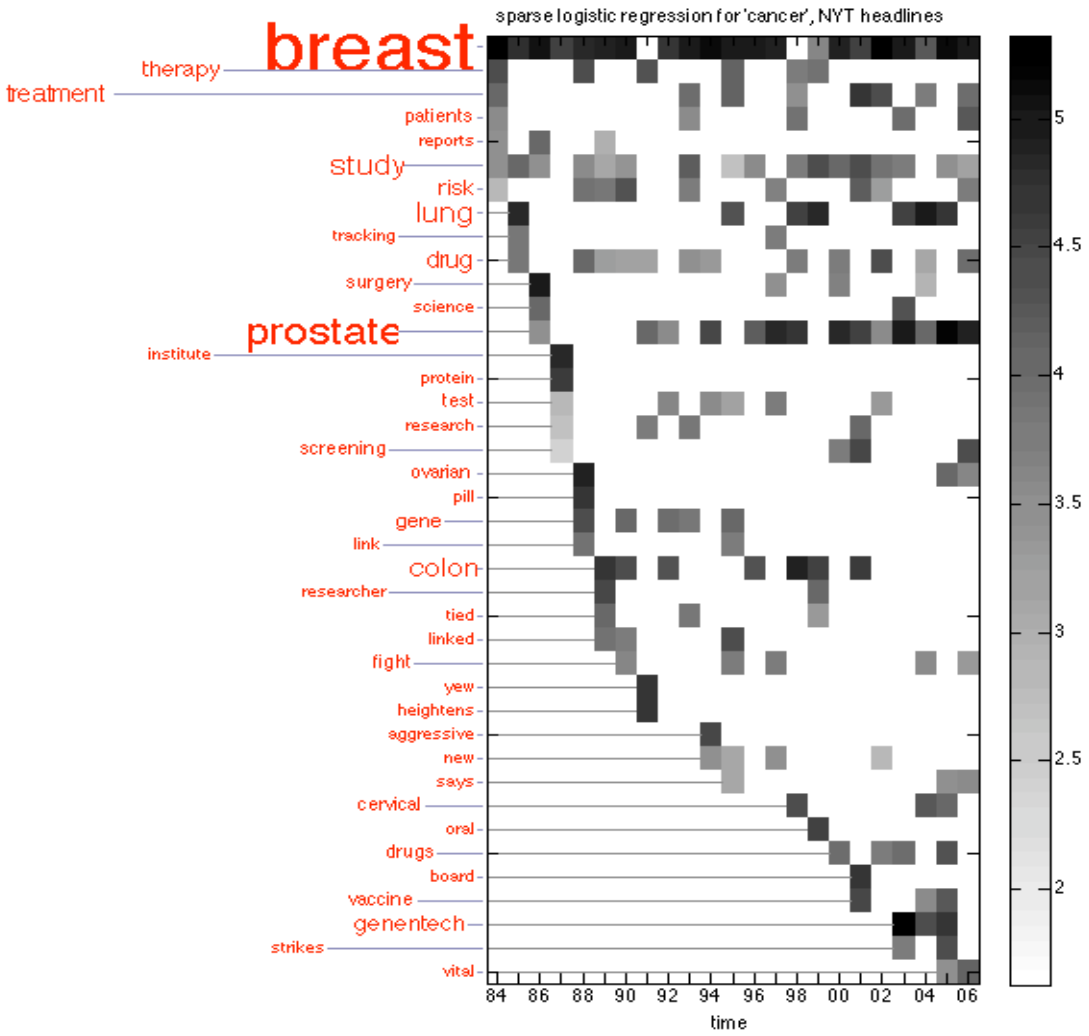
“China”

Data: The New York Times headlines, 1985-2007



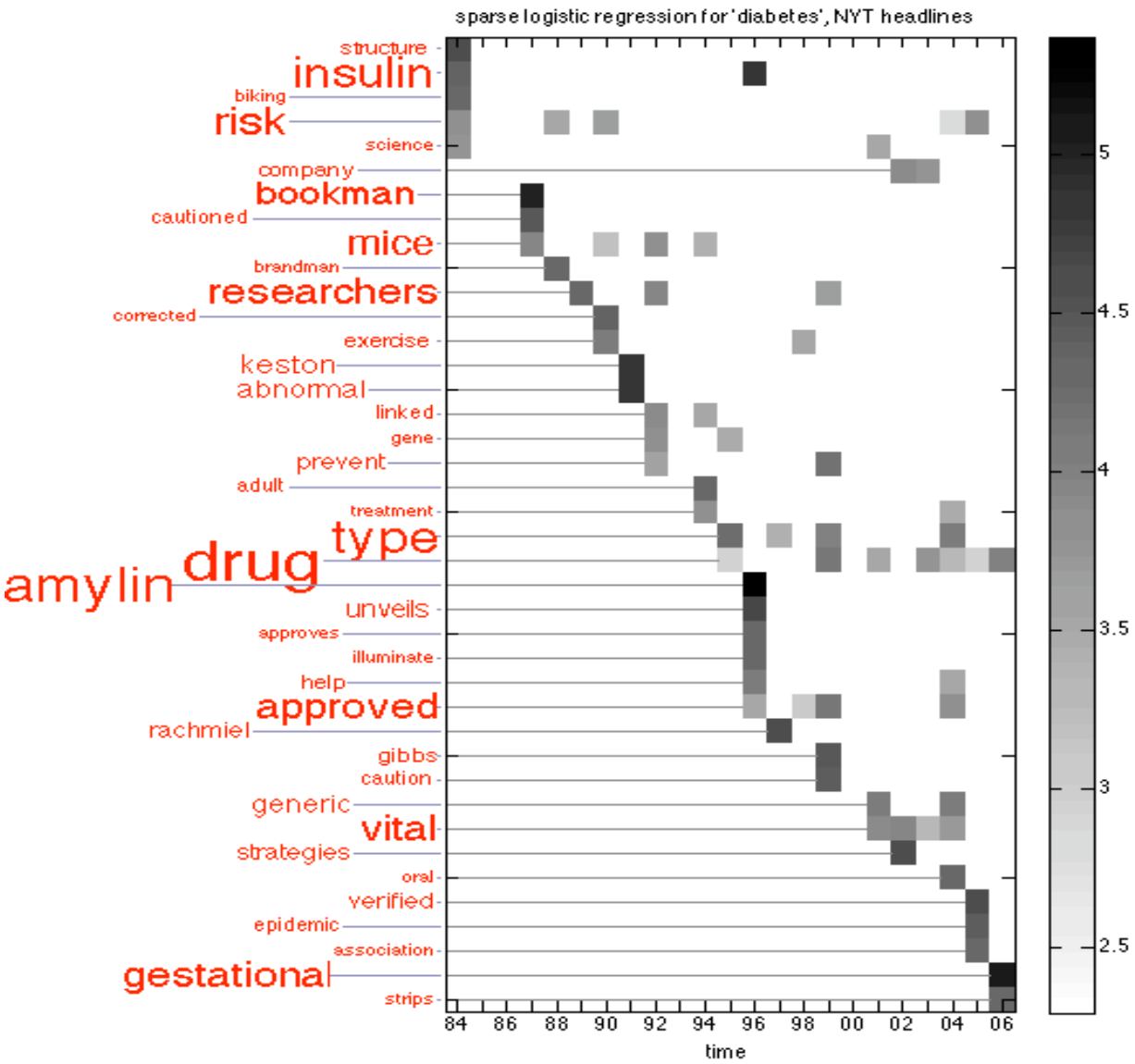
“Cancer”

Data: The New York Times headlines, 1985-2007



“Diabetes”

Data: The New York Times headlines, 1985-2007



Topic imaging in foreign languages

- Translate query term.
- Run topic imaging task on foreign press data in original language.
- Translate the *few* terms in the resulting list.

Avoids huge translation task!

Topic imaging in foreign languages

- Translate query term.
- Run topic imaging task on foreign press data in original language.
- Translate the *few* terms in the resulting list.

Avoids huge translation task!

Query: can you guess?

Source: People's Daily, Feb-Apr 2011.

利比亚 欧佩克 opec
利比亚 武力 force
利比亚 局势 situation
利比亚 行动 action
利比亚 平民 civilians
利比亚 撤出 withdrawal
利比亚 空袭 airstrike
利比亚 北非 french-speaking
利比亚 瓦莱塔 valletta
利比亚 撤离 evacuate
利比亚 军机 planes
利比亚 人道主义 humanitarianism
利比亚 卡扎菲 qadhafi

Outline

Example

Discovery of emerging issues in flight security

After each commercial flight in the US, pilots generate “ASRS reports” to document flight-related issues.

Key problem: detect emerging issues that are not being classified into existing categories, *e.g.*:

- “Wake vortex” problem of the Boeing 757.
- Increased number of runway incursions at LAX.

Example

Discovery of emerging issues in flight security

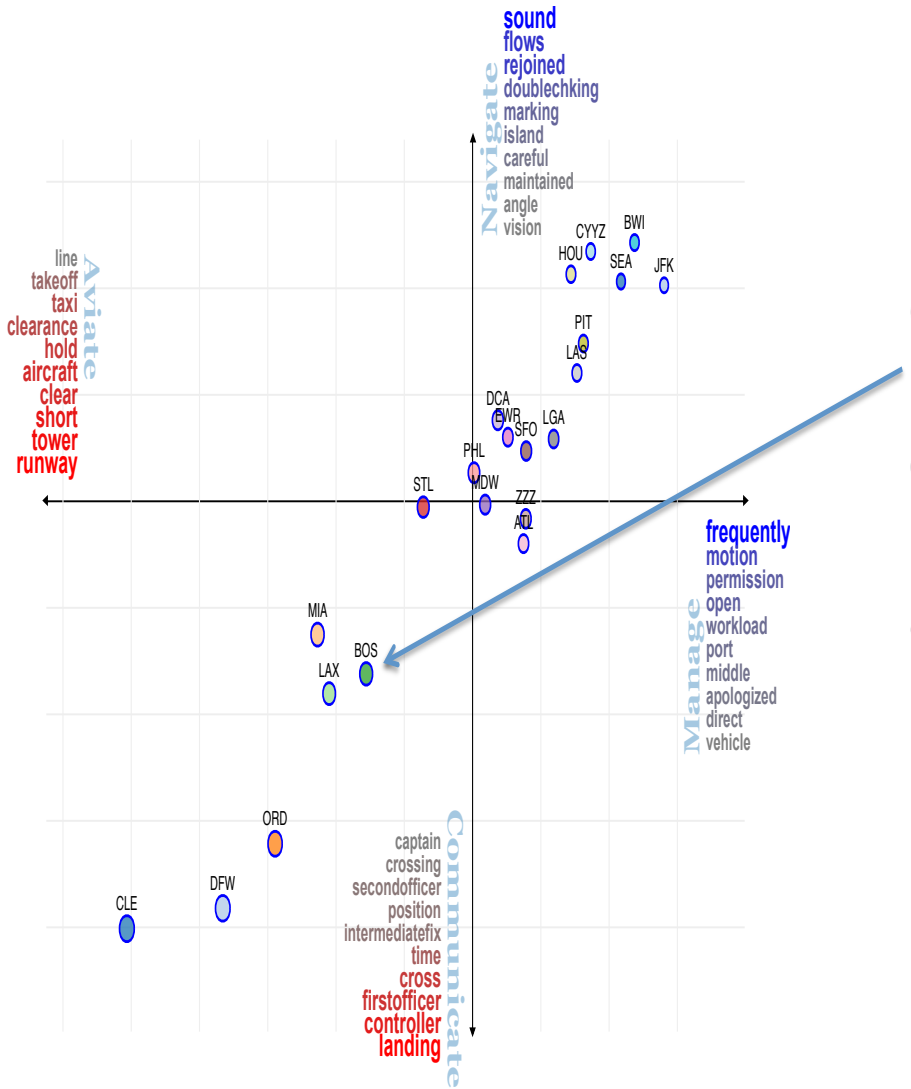
After each commercial flight in the US, pilots generate “ASRS reports” to document flight-related issues.

Key problem: detect emerging issues that are not being classified into existing categories, *e.g.*:

- “Wake vortex” problem of the Boeing 757.
- Increased number of runway incursions at LAX.

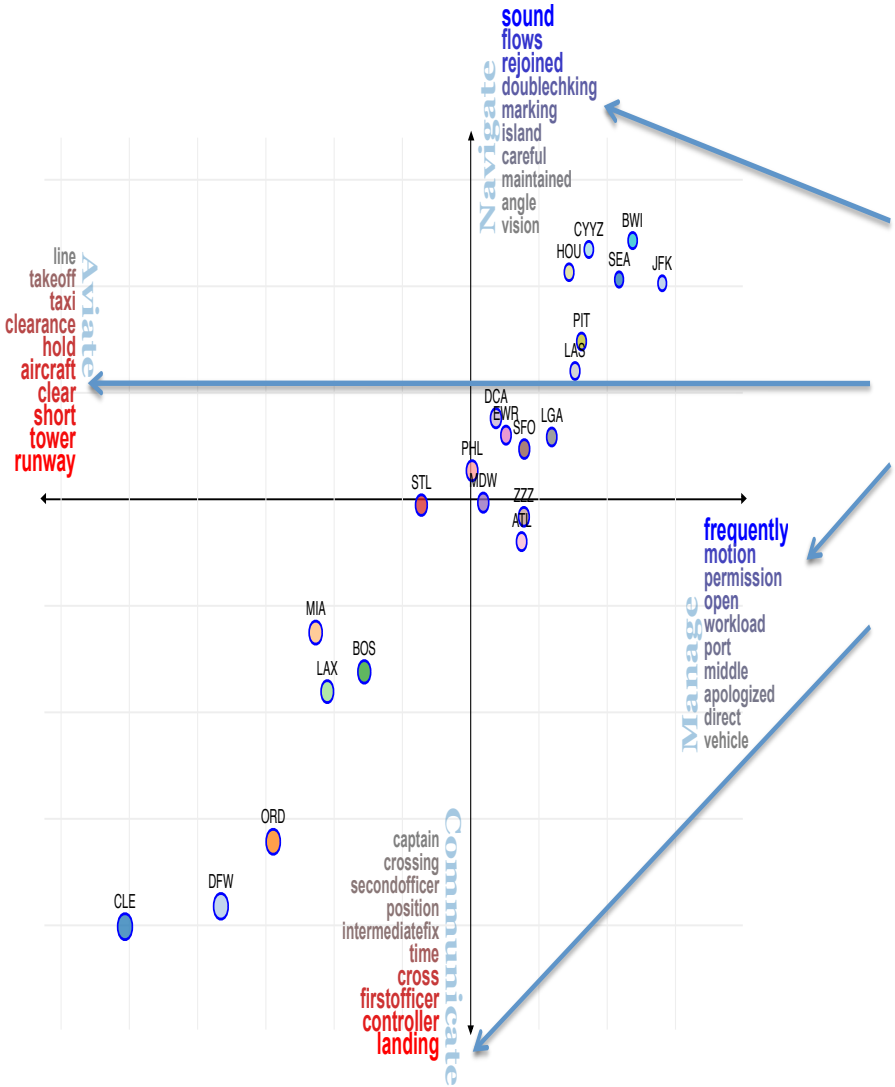
Don’t search for a needle — picture the haystack!

Sparse PCA Imaging



Each of these dots represents reports that originate from a particular airport.

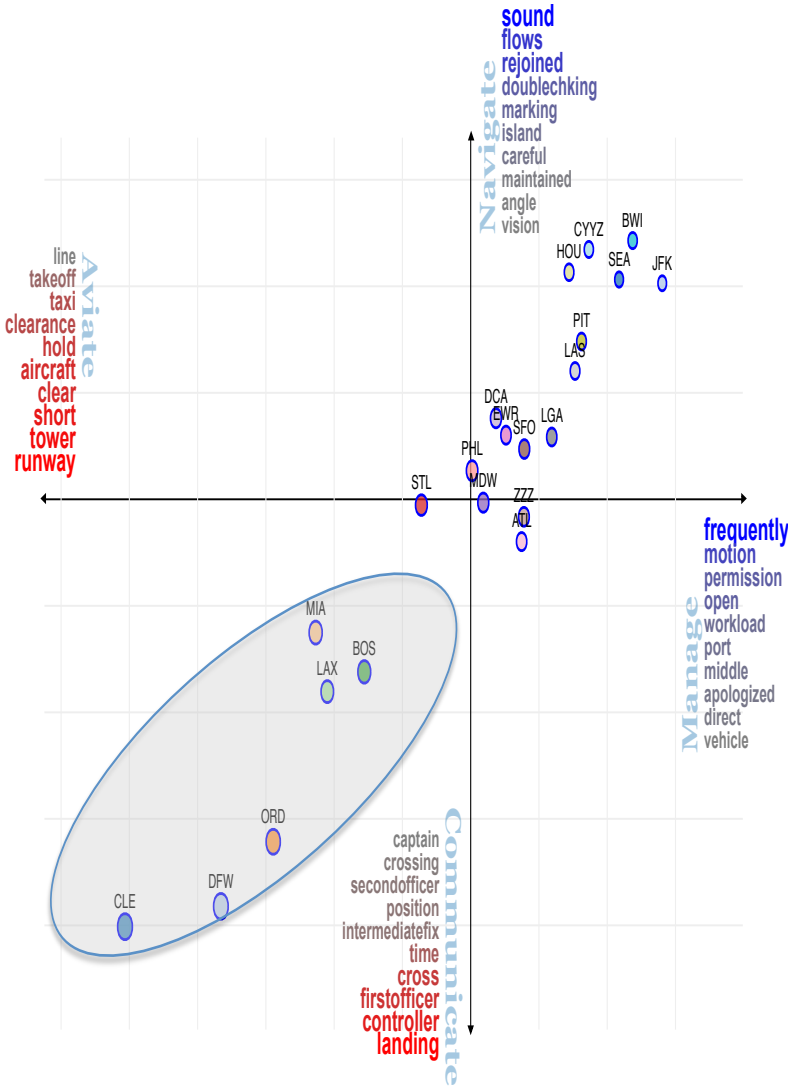
Sparse PCA Imaging



Each of the four directions corresponds to one of the four basic pilot tasks (eg, ``Aviate'').

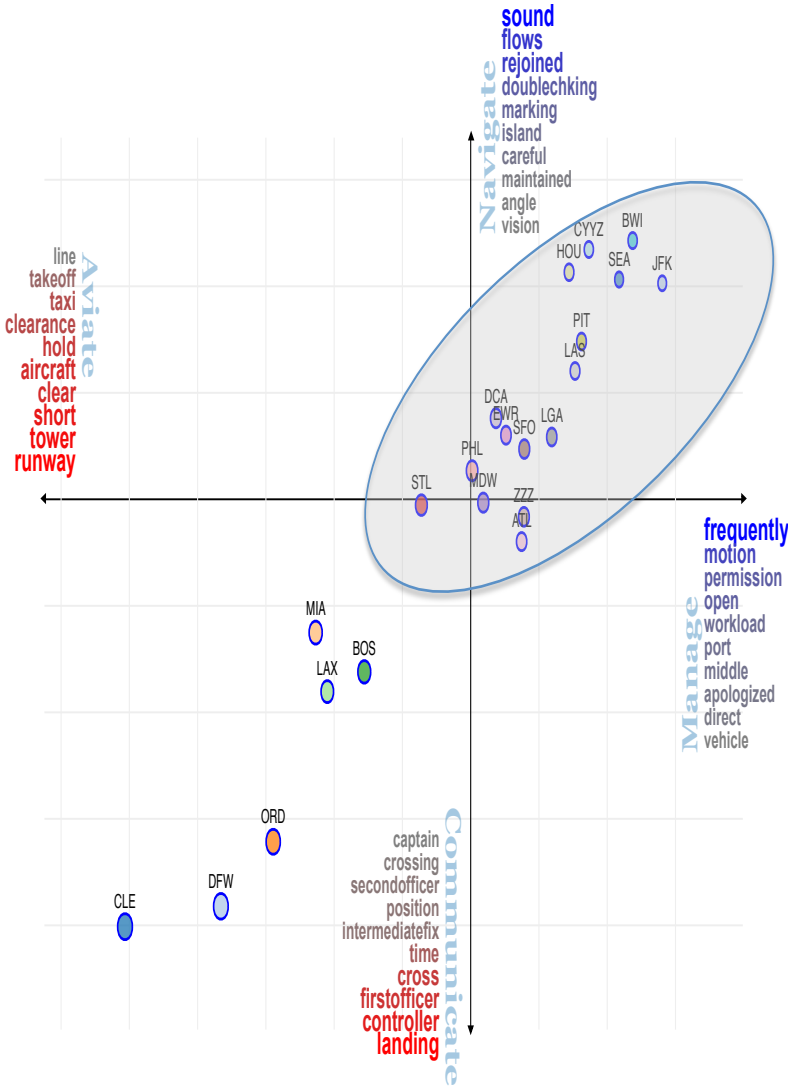
The terms were *automatically* found.

Sparse PCA Imaging



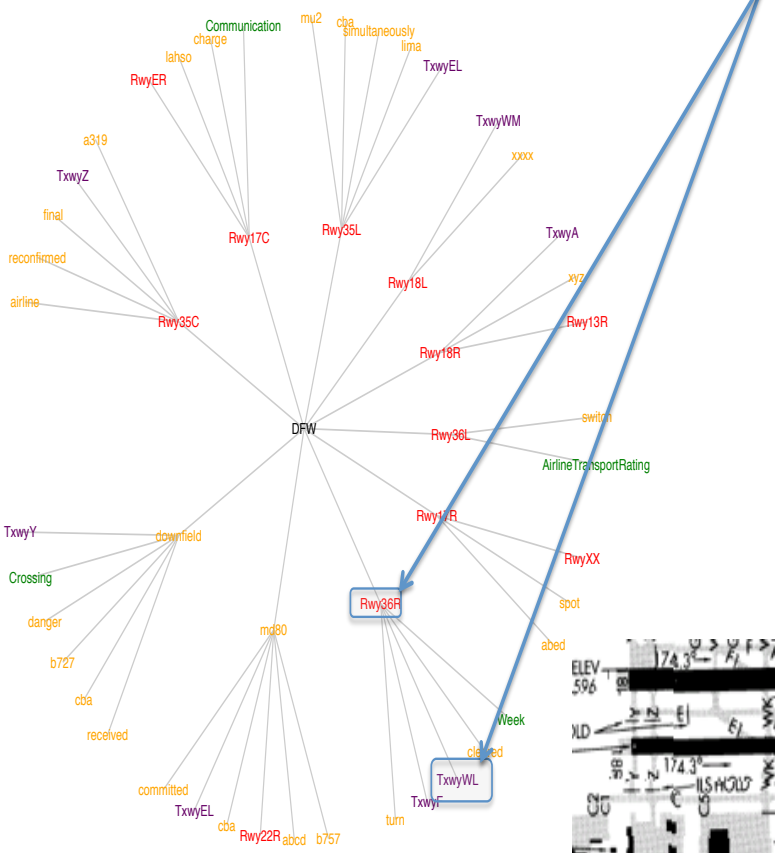
Large-volume airports are mostly exposed to aviation (eg, take-off) and communication issues.

Sparse PCA Imaging



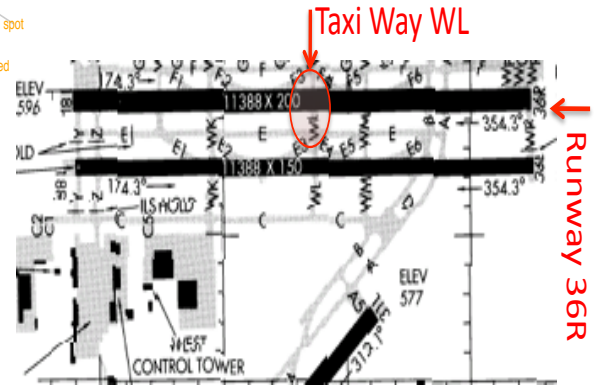
Smaller-volume airports are mostly exposed to management (worload) and navigation (eg, markings on runway) issues.

LASSO for DFW



- Identifies critical runways, in particular runway/taxiway intersections that cause trouble.

- Also identifies communication issues.



Outline

Xinyu Dai, Jinzhu Jia, Laurent El Ghaoui, and Bin Yu.

SBA-term: Sparse bilingual association for terms.
In *Fifth IEEE International Conference on Semantic Computing*, Palo Alto, CA, USA, 2011.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani.

Safe feature elimination for the LASSO.
Submitted to *Journal of Machine Learning Research*, April 2011.
Early draft: EECS Technical Report no. 126, September 2010.

B. Gawalt, J. Jia, L. Miratrix, L. El Ghaoui, B. Yu, and S. Clavier.

Discovering word associations in news media via feature selection and sparse classification.
In *Proc. 11th ACM SIGMM International Conference on Multimedia Information Retrieval*, 2010.

Luke Miratrix, Jinzhu Jia, Brian Gawalt, Bin Yu, and Laurent El Ghaoui.

Summarizing large-scale, multiple-document news data: sparse methods and human validation.
submitted to JASA.

Haipeng Shen and Jianhua Z. Huang.

Sparse principal component analysis via regularized low rank matrix approximation.
J. Multivar. Anal., 99:1015–1034, July 2008.

Y. Zhang, A. d'Aspremont, and L. El Ghaoui.

Sparse PCA: Convex relaxations, algorithms and applications.
In M. Anjos and J.B. Lasserre, editors, *Handbook on Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications*. Springer, 2011.
To appear.