Exploring Hyperspace

Hobson Lane
hobs@totalgood.com

total
GOOD

What is High-D?

- **10-D**
  - Physics breaks (string theory = 12D)
  - My mind breaks

- **Fails**
  - Bounding boxes are useless (empty or full)
  - Indexing
  - Locality-sentiive hashes

total
GOOD

Examples

Images

Time Series

Natural Language

total
GOOD

Mental Gymnastics

"Imagine a 3D space and then say 12 dimensions to yourself forcefully over and over again."

- Geoffrey Hinton

“Imagine a 3D sphere at the edge of universe, your vectors are out there.”

- Anonymous

total GOOD

Why Squash?

- Feature Extraction
- Abstraction
- Dimension Reduction
- Generalization
- Stereotyping?

total GOOD

"Issues" for AI

Features are far from each other

Features on a thin shell

Manifolds have many saddle points

Costs have local minima (though fewer than saddles)

Classes far away in feature space

total
GOOD

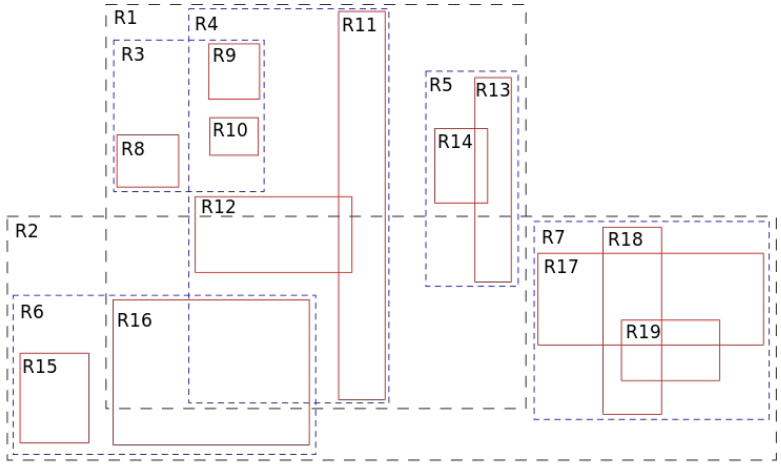
Indexes? Hash tables? Trees?

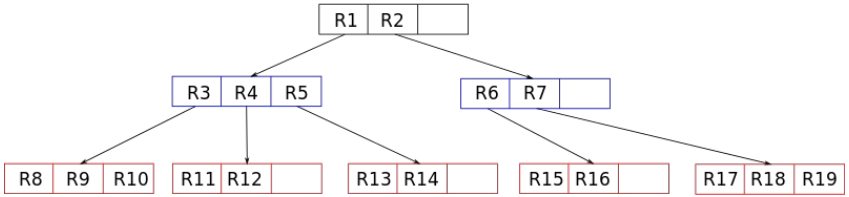
- Great for discrete or sparse vectors
  - ➔ [1 2 3 1 1 2 3 4 1 2 1 3 4]
  - ➔ [42 0 0 3 0 9 0 0 12 0 1 0 0]
  - ➔ [“A” 0 0 0 0 0 0 0 0 “and” 0 0 0 0 0 0 0 0 “cat”]
  - ➔ [0.123 0 0 0 0 0 0.567 0 0 0 0 -.42 0 0 0 2.718 0 0 0]

- Bad for dense, real-valued vectors
  - ➔ [0.123 0.456 0.789 -0.1011 …]

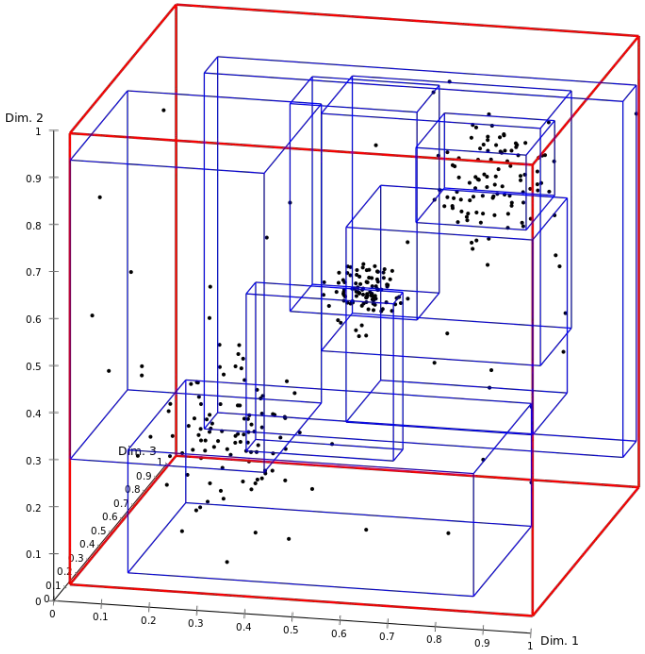
total
GOOD

R-Tree (PostGIS)





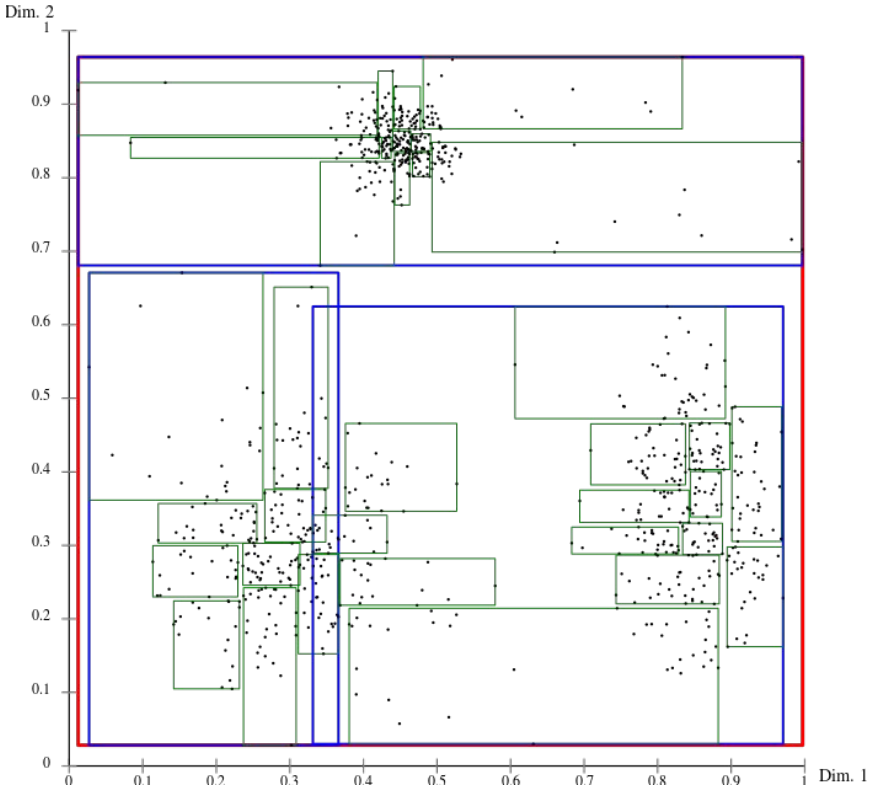
Wikimedia Creative Commons



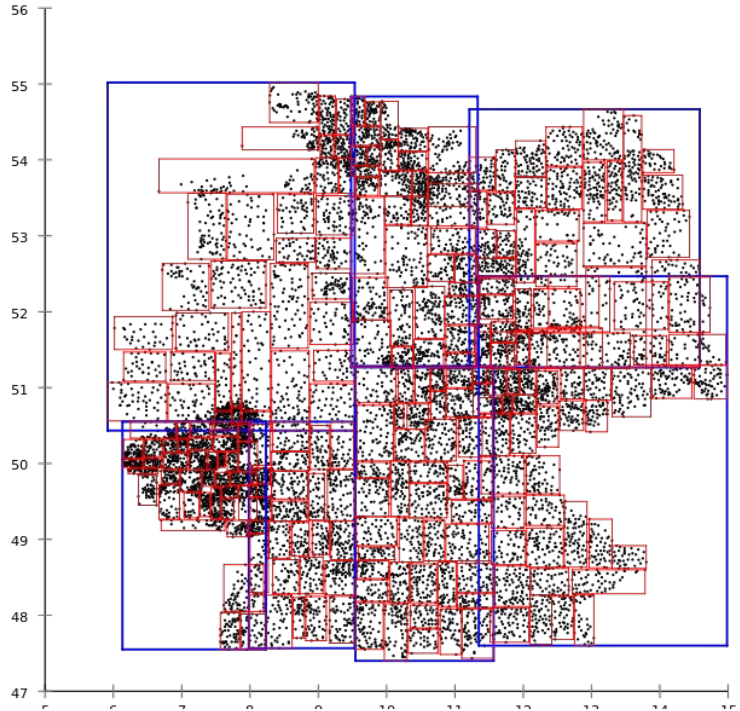
Wikimedia Creative Commons

total GOOD

R*-Tree



Wikimedia Creative Commons



Wikimedia Creative Commons

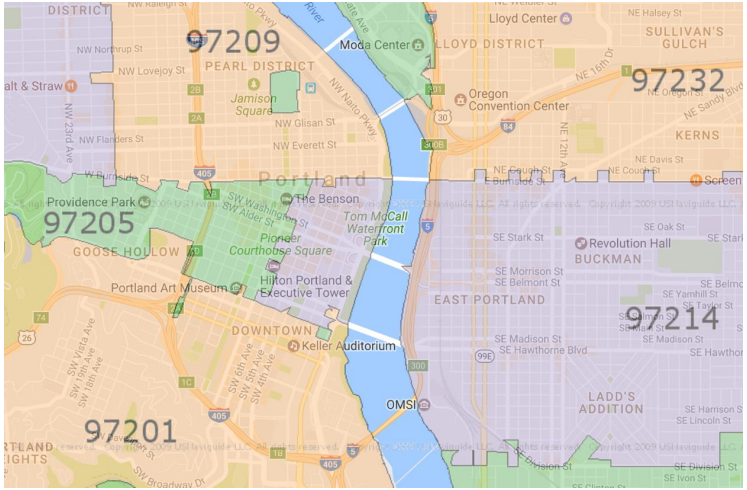
total GOOD

Hashes (`dict`)

Constant time lookup
Pseudorandom (no “distance”)

Locality-Sensitive Hashes?

Zip Codes work in 2-D, right?



total GOOD

Semantic Search Index

D	N	100th Cosine Distance	Top 1 Correct	Top 2 Correct	Top 10 Correct	Top 100 Correct
2	4254	0	TRUE	TRUE	TRUE	TRUE
3	7727	0.0003	TRUE	TRUE	TRUE	TRUE
4	12198	0.0028	TRUE	TRUE	TRUE	TRUE
5	9920	0.0143	TRUE	TRUE	TRUE	TRUE
6	11310	0.0166	TRUE	TRUE	TRUE	TRUE
7	12002	0.0246	TRUE	TRUE	TRUE	FALSE
8	11859	0.0334	TRUE	TRUE	TRUE	FALSE
9	6958	0.0378	TRUE	TRUE	TRUE	FALSE
10	5196	0.0513	TRUE	TRUE	FALSE	FALSE
11	3019	0.0695	TRUE	TRUE	TRUE	FALSE
12	12263	0.0606	TRUE	TRUE	FALSE	FALSE
13	1562	0.0871	TRUE	TRUE	FALSE	FALSE
14	733	0.1379	TRUE	FALSE	FALSE	FALSE
15	6350	0.1375	TRUE	TRUE	FALSE	FALSE
16	10980	0.0942	TRUE	TRUE	FALSE	FALSE

```
pip install lshash3
```

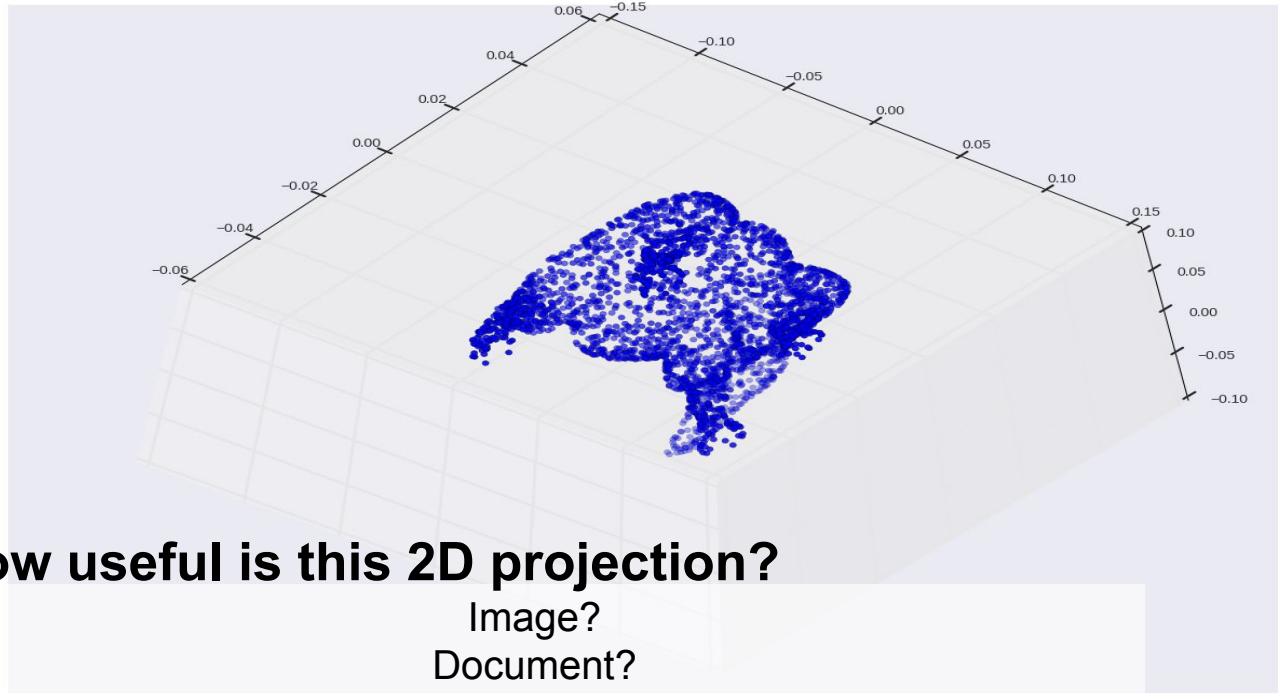
total GOOD

Variance Maximizing D-Reducers

- **PCA**: Principal Component Analysis
- **LSI**: Latent Semantic Indexing
- **LSA**: Latent Semantic Analysis
- **SVD**: Singular Value Decomposition
-
- One way to "squash" our hyperspace would be to **"lemmatize" LSI**, **LSA**, **SVD, and PCA together**, eliminating 12 dimensions

total GOOD

Some 3D Vectors

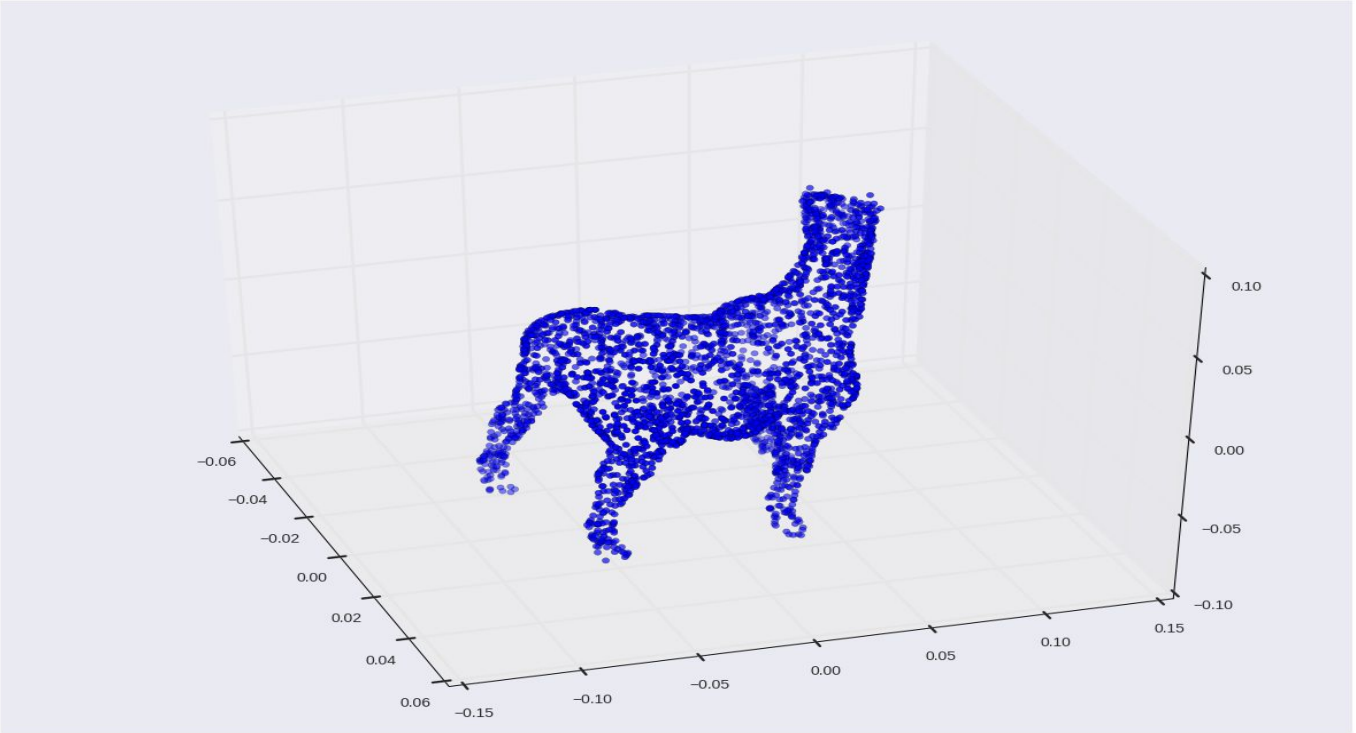


How useful is this 2D projection?

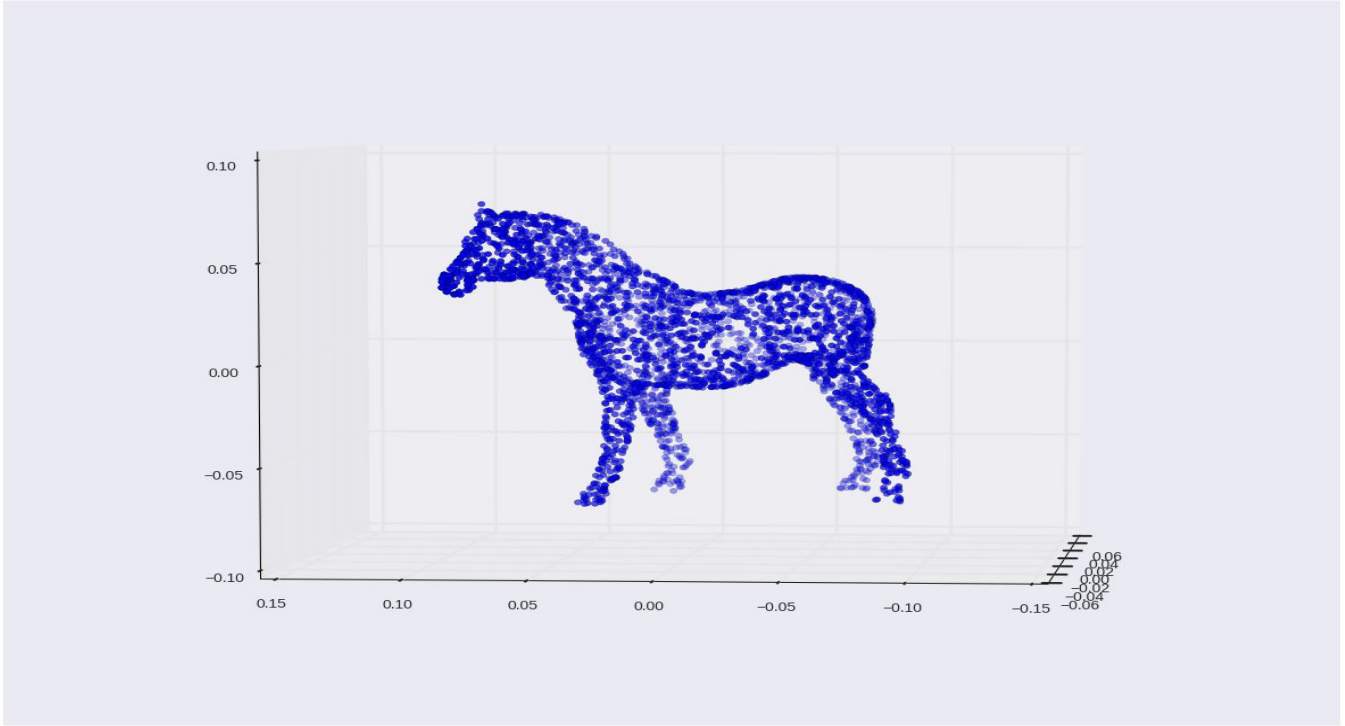
Image?
Document?
Time series?
Person, place, or thing?

total GOOD

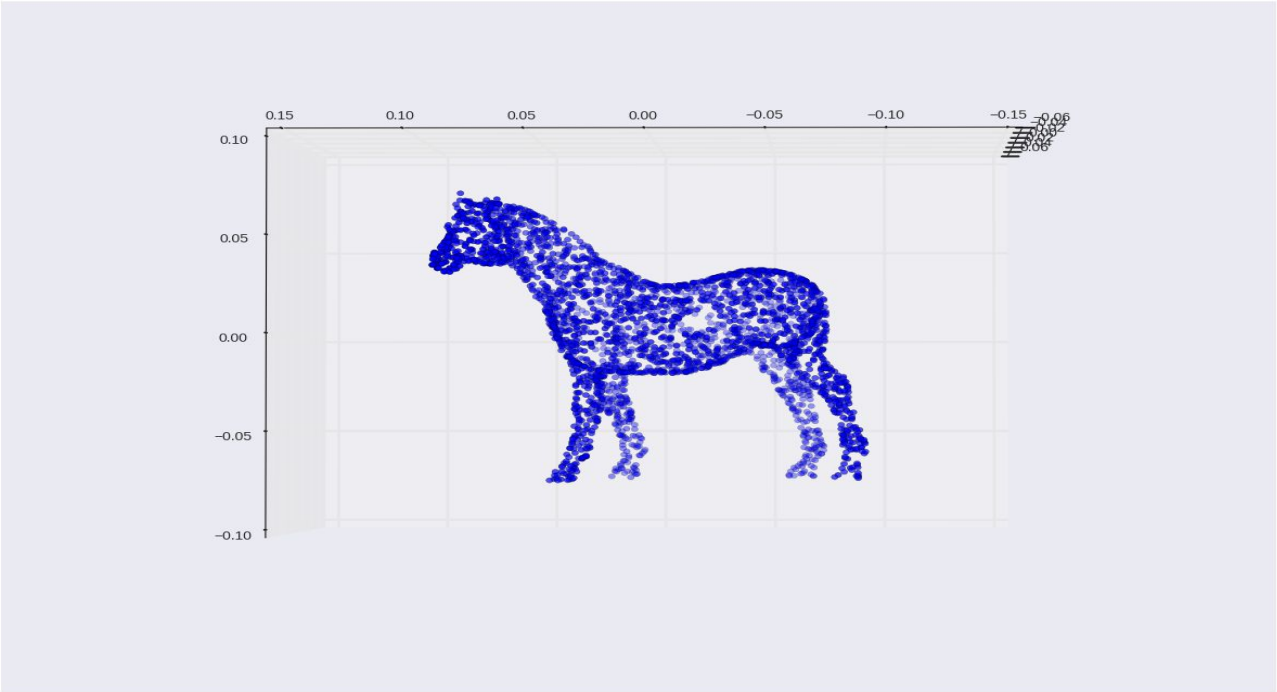
Better projection



Better Projection

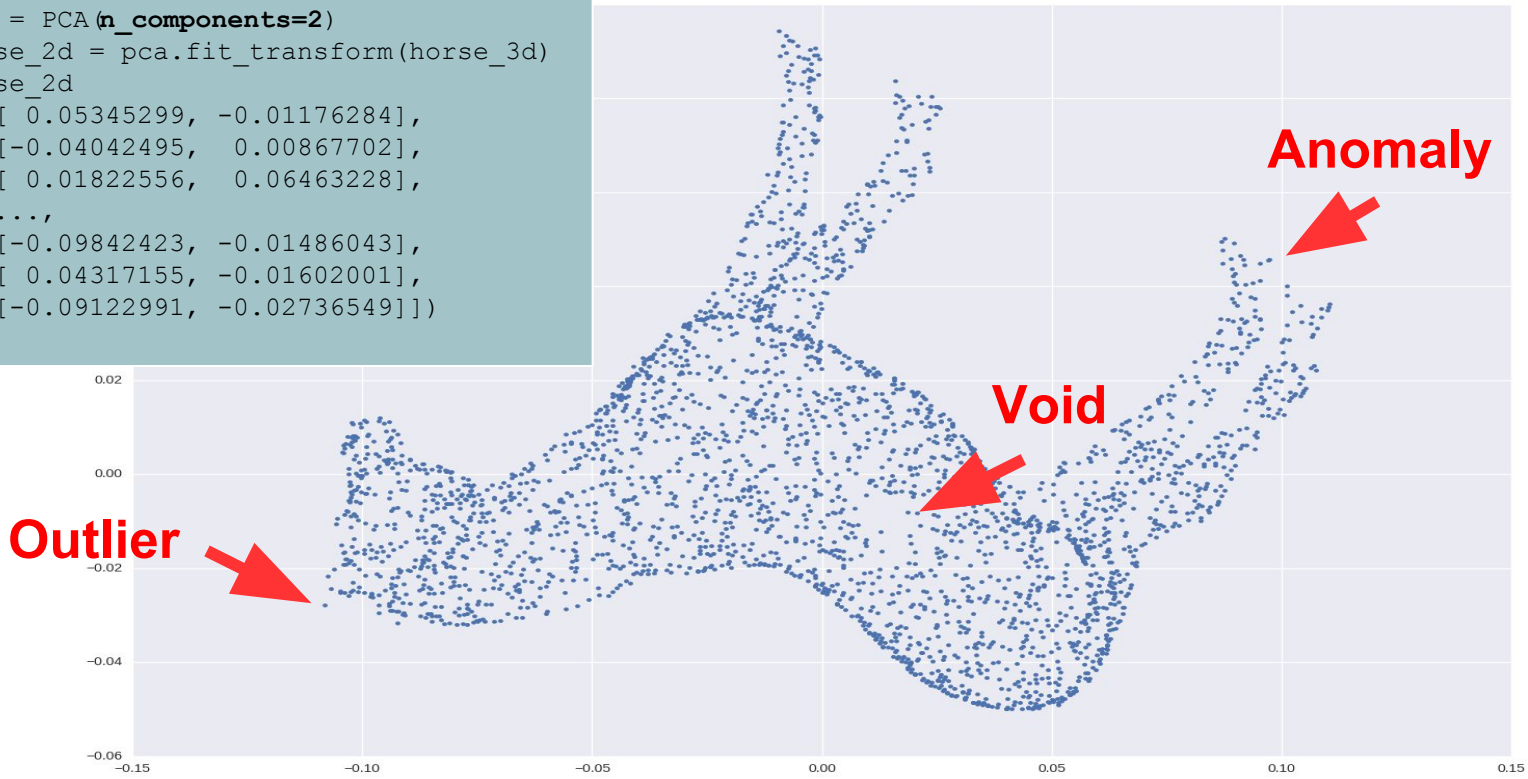


Best Projection by Human

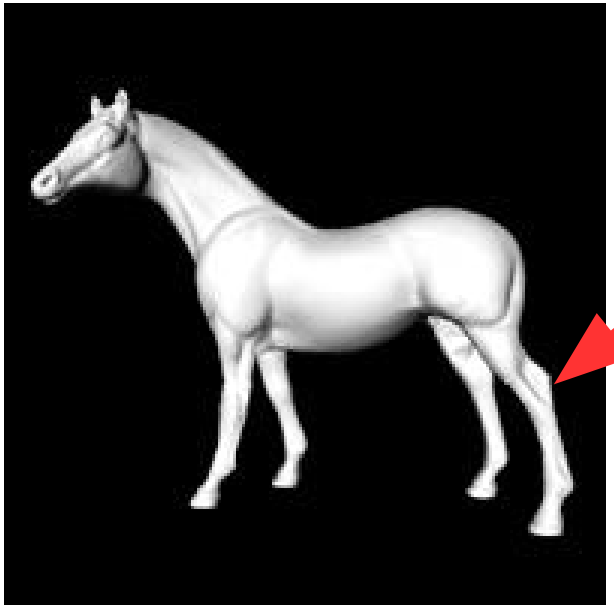


What PCA "sees"

```
>>> from sklearn.decomposition import PCA
>>> pca = PCA(n_components=2)
>>> horse_2d = pca.fit_transform(horse_3d)
>>> horse_2d
array([[ 0.05345299, -0.01176284],
       [-0.04042495,  0.00867702],
       [ 0.01822556,  0.06463228],
       ...,
       [-0.09842423, -0.01486043],
       [ 0.04317155, -0.01602001],
       [-0.09122991, -0.02736549]])
```



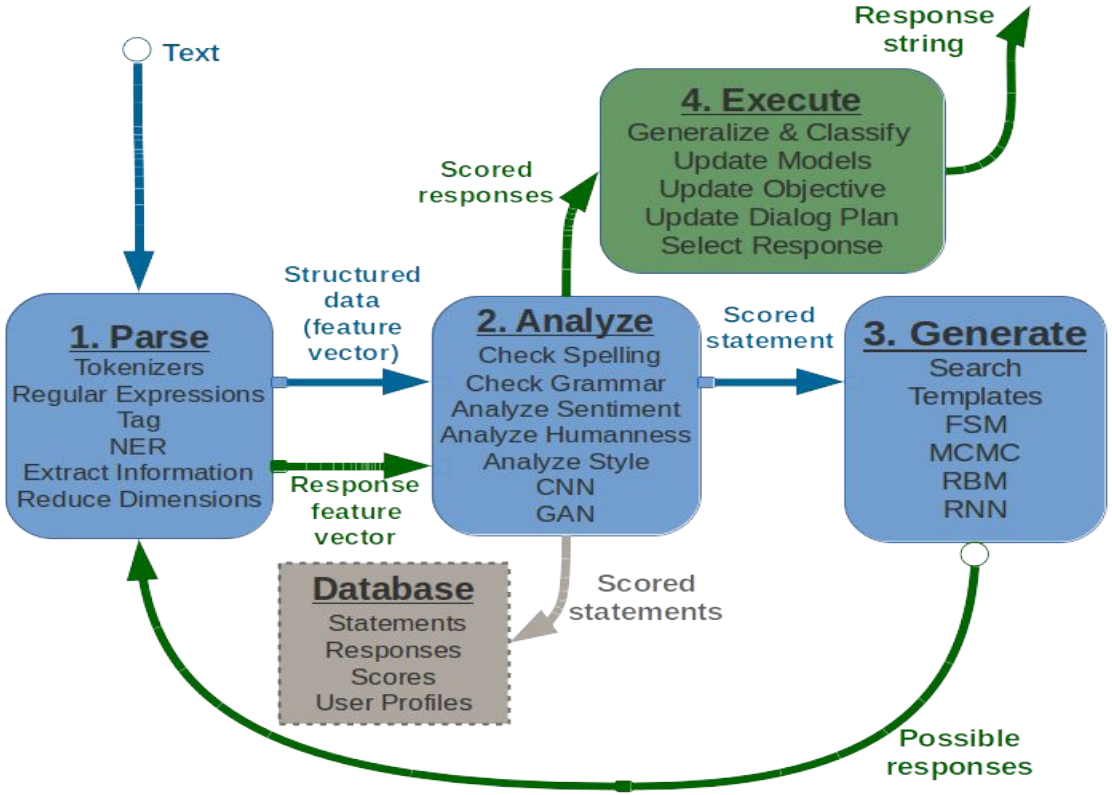
A Horse is a Horse unless…



… It's **Mr. Ed,**
or a plastic toy on a 3D scanner.

total
GOOD

Recurrent Chatbot

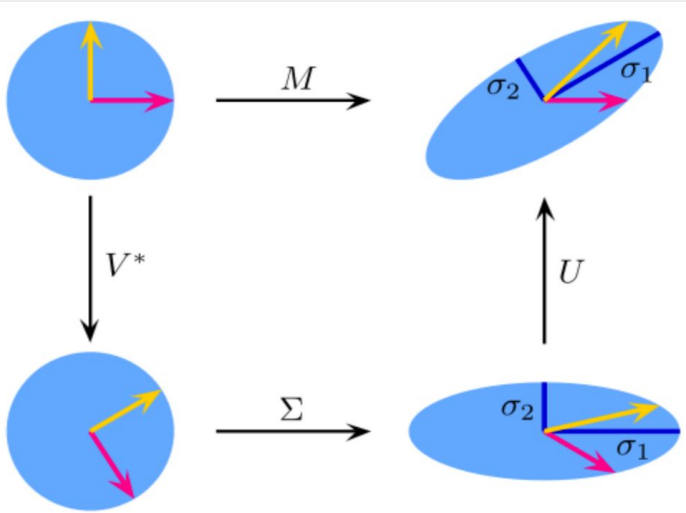


total GOOD

Deep Dive into SVD

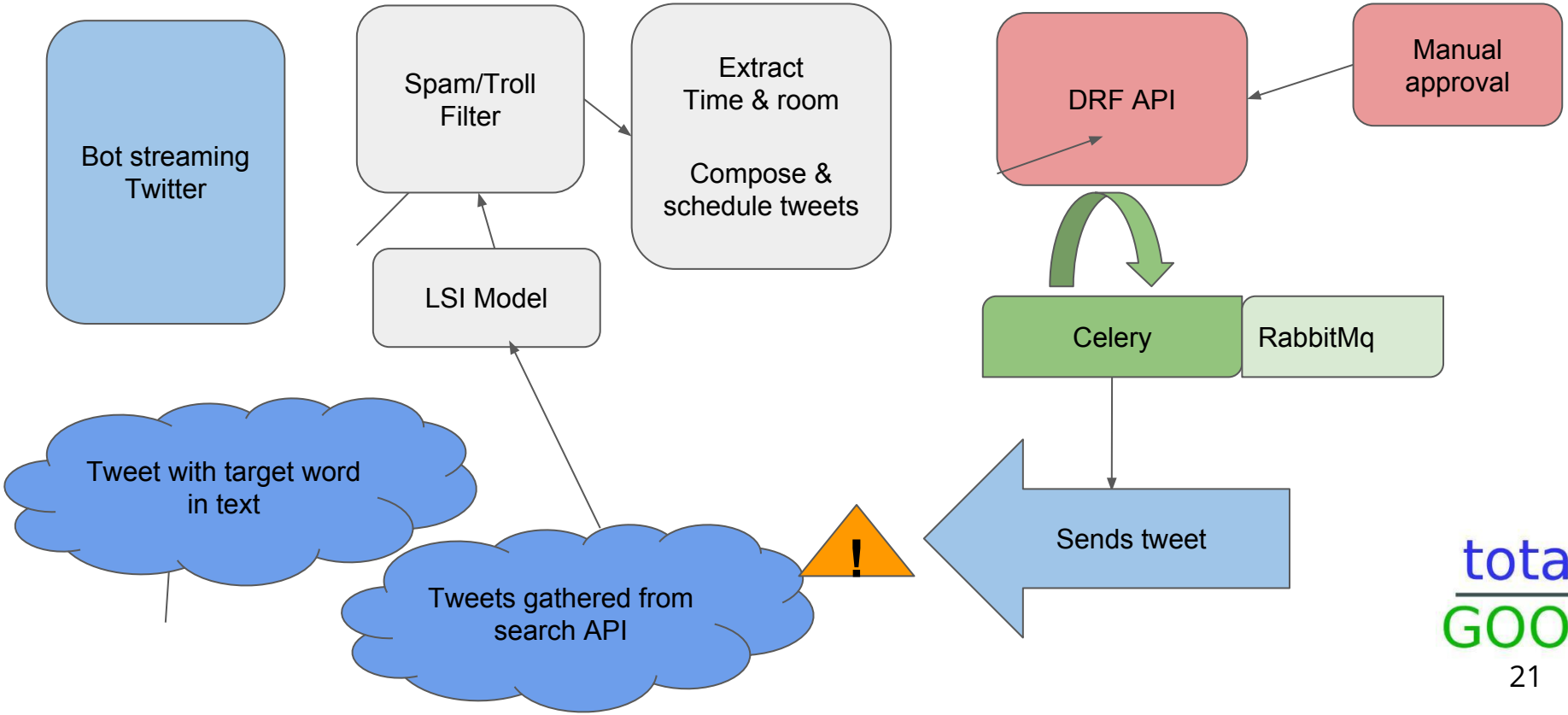
Singular Value Decomposition of a rectangular matrix

$M = U \Sigma V^T$



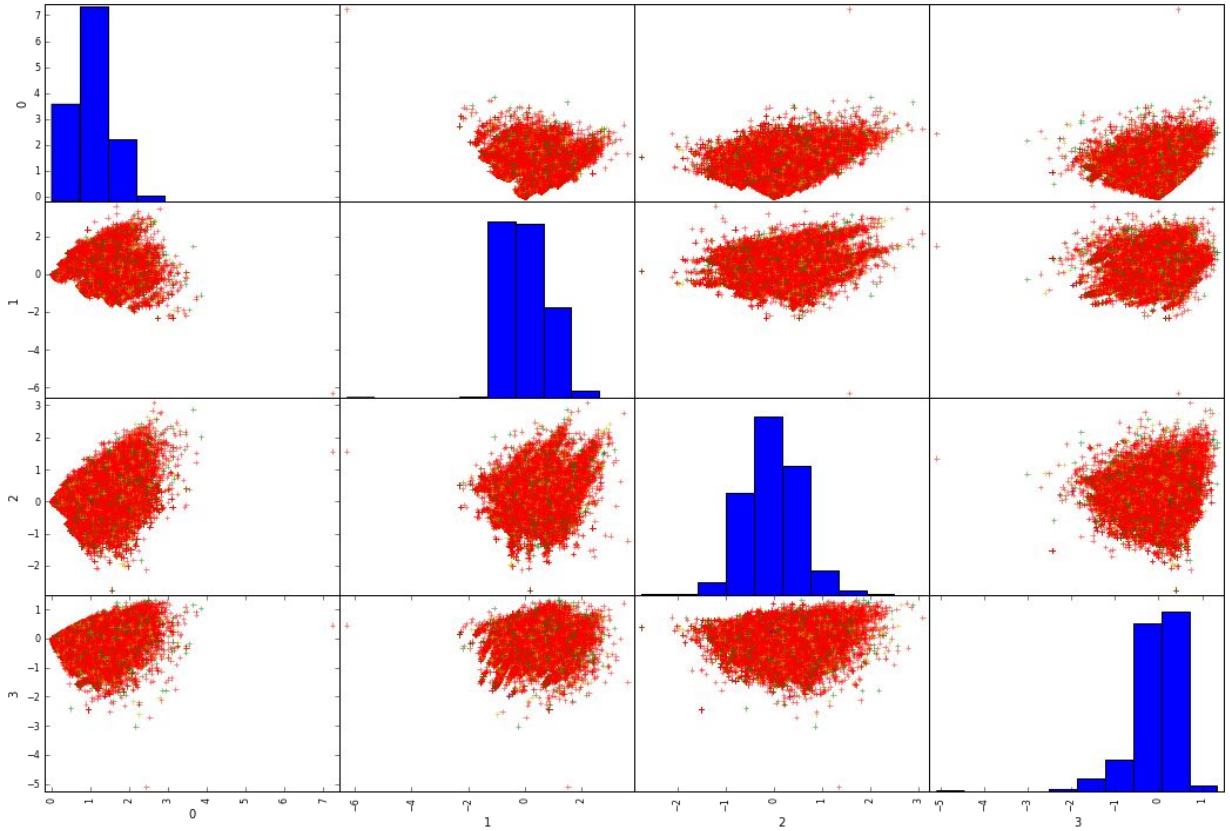
total GOOD

Open Spaces Twitter Bot



total GOOD

gensim.models.LsiModel



total
GOOD

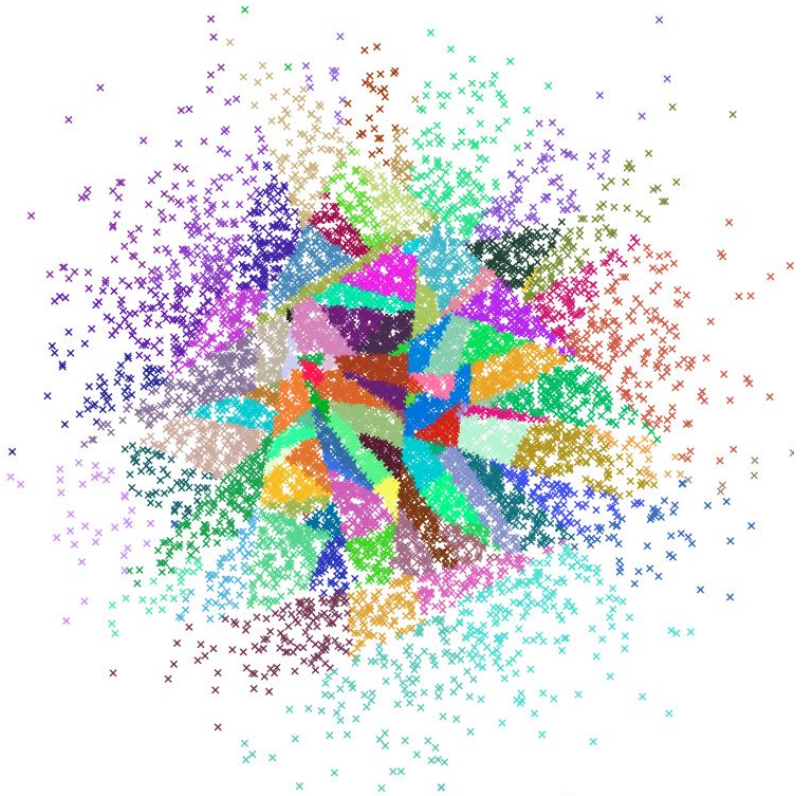
PCA Generalizes Well

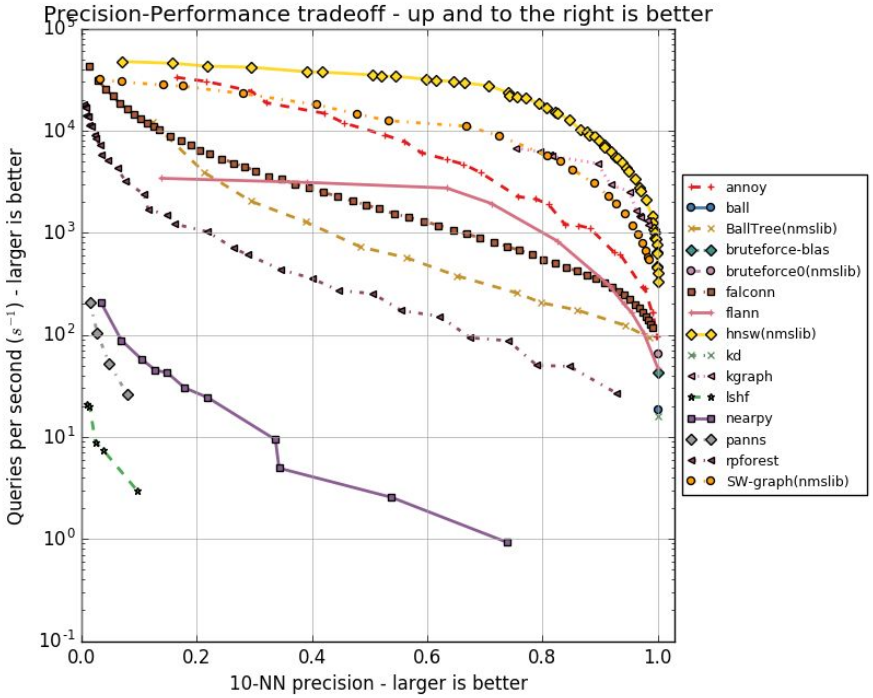
LSA on **5.5 M** tweets
81.2% accuracy for **50 k** bot/human tweet examples
79.6% accuracy for **3.3 k** bot/human examples

Only 1.6% gain for **15x** labeled data!

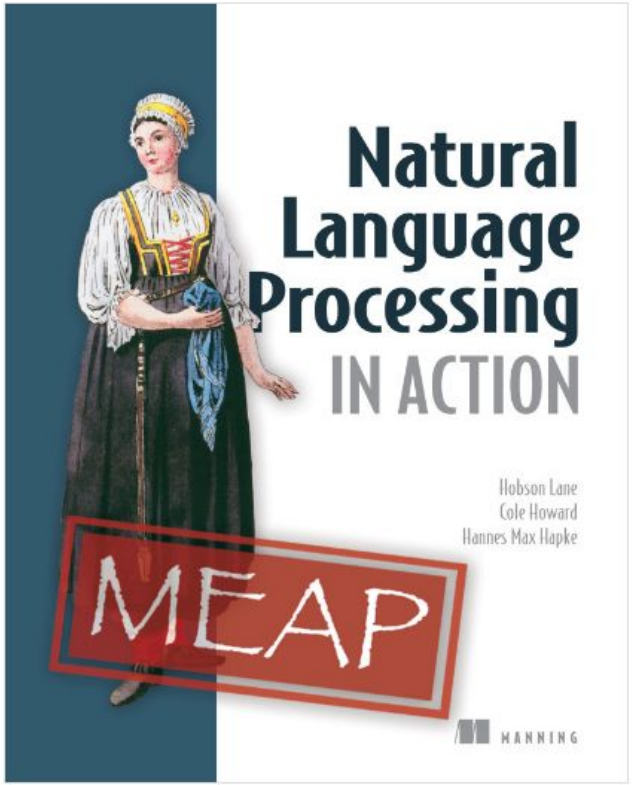
total
GOOD

pip install annoy





NLP in Action



NL Search
NL Modeling
NL Generation

CNNs, RNNs, and LSTMs
Word2Vec, GloVe

total
GOOD