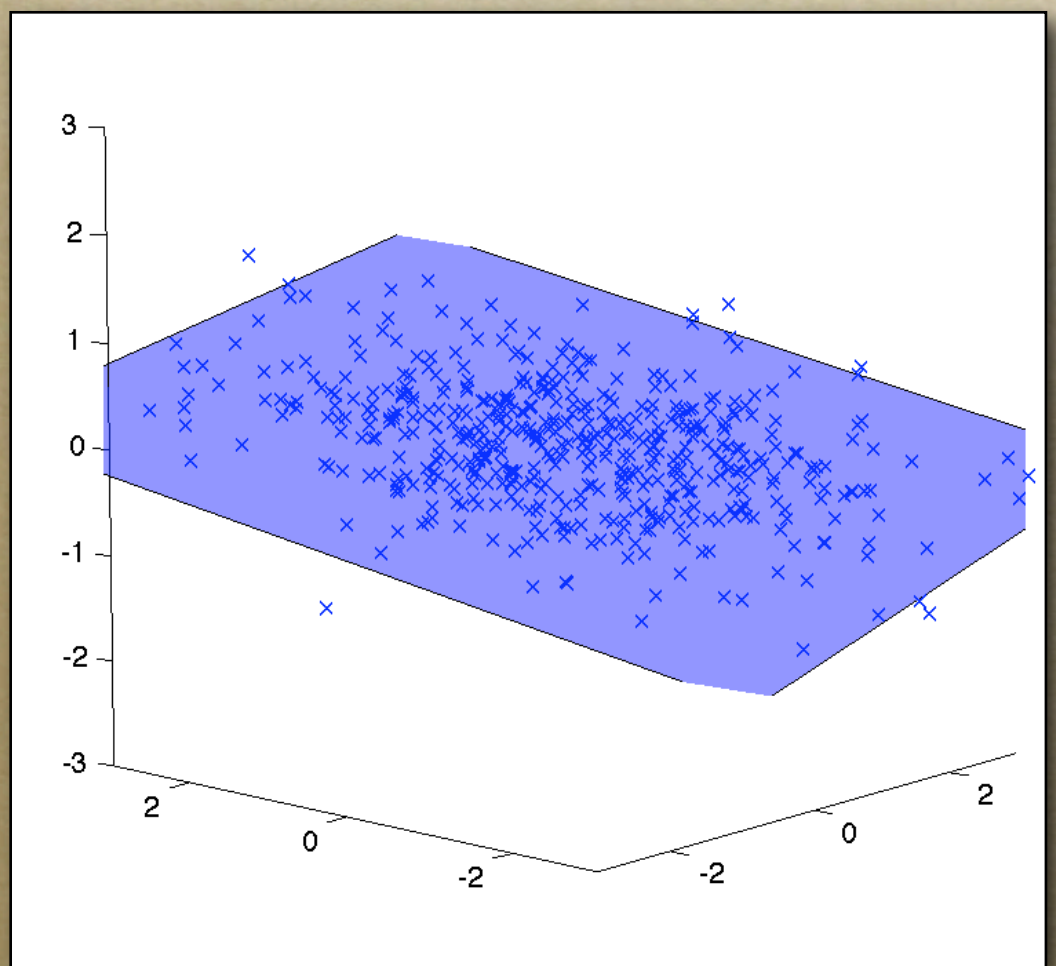
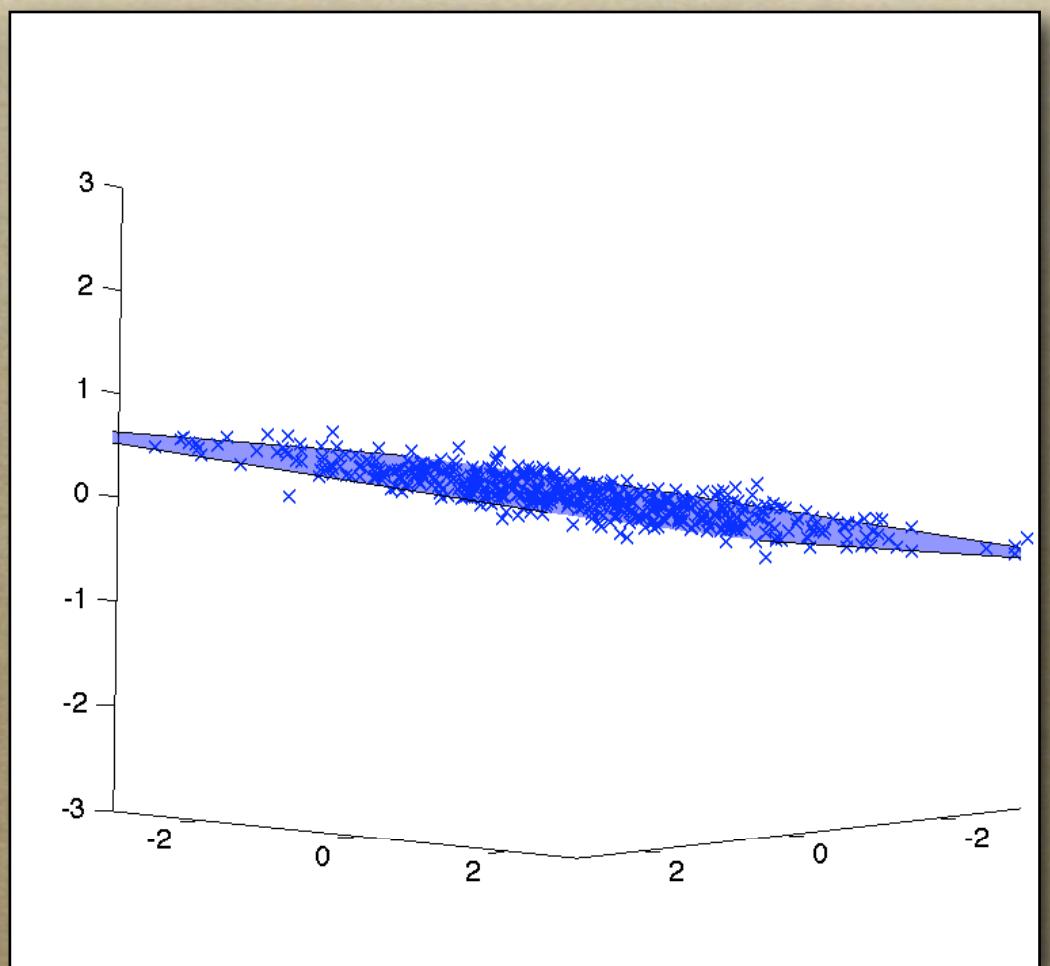
It's a Small(-Rank) World, After All

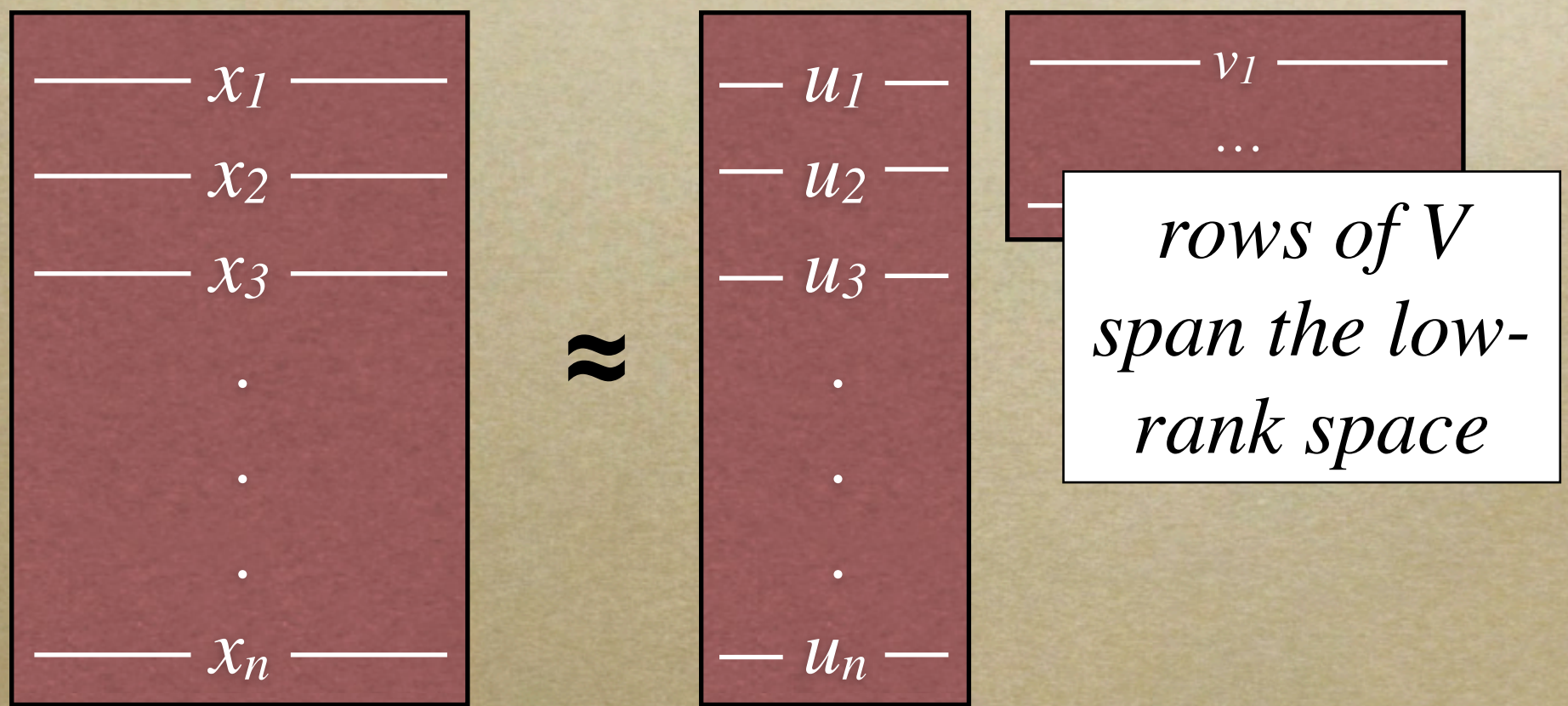
Geoff Gordon
SELECT Lab
Machine Learning Department
Joint work with Ajit Singh, Francisco Pereira, Nick Roy, Byron Boots, Sajid Siddiqi

Principal Components Analysis





PCA: the equation



Examples: topic analysis

CHAMPION PRODUCTS <CH> APPROVES STOCK SPLIT

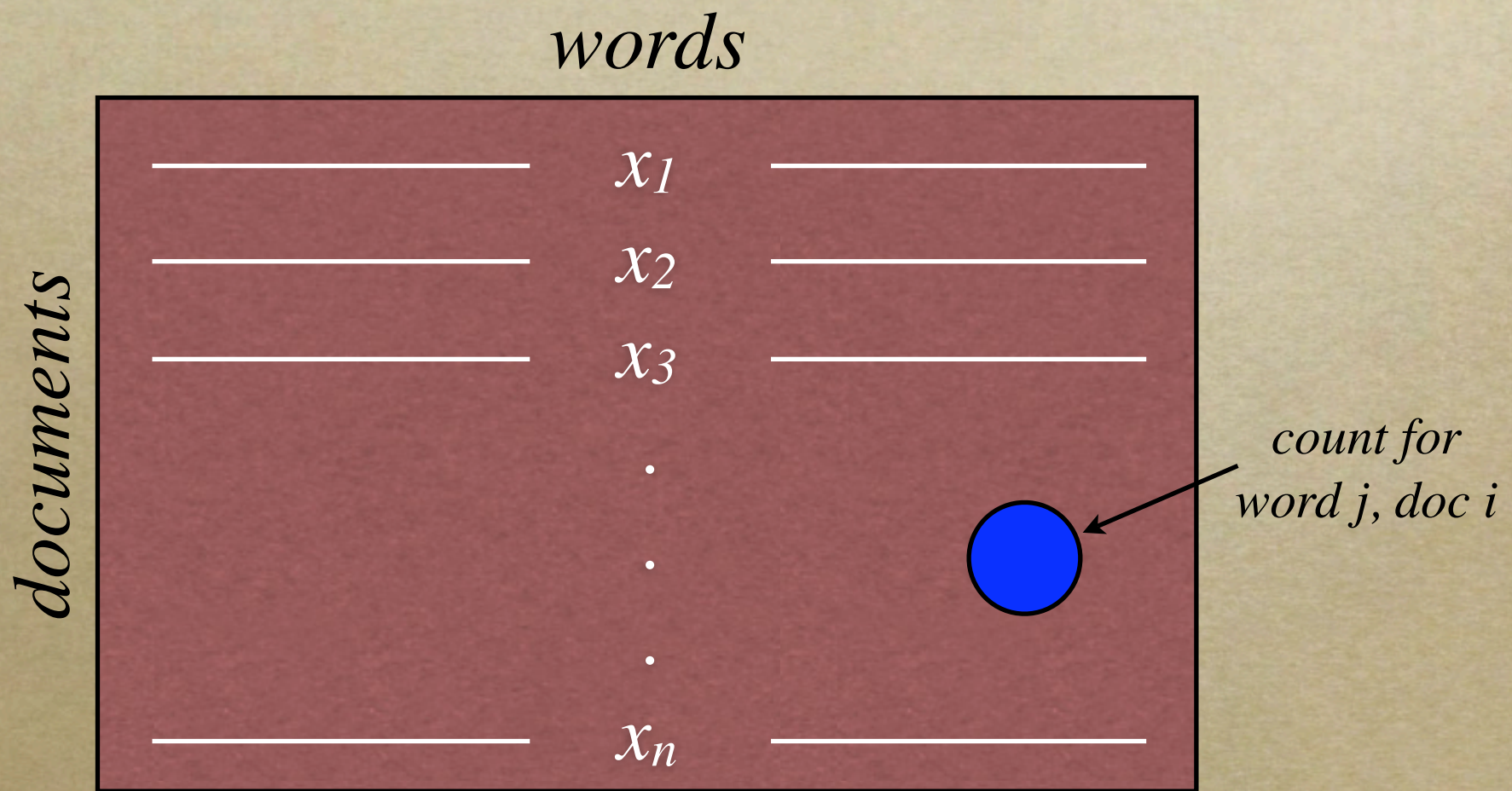
ROCHESTER, N.Y., Feb 26 - Champion Products Inc said its board of directors approved a two-for-one stock split of its common shares for shareholders of record as of April 1, 1987.

The company also said its board voted to recommend to shareholders at the annual meeting April 23 an increase in the authorized capital stock from five mln to 25 mln shares.

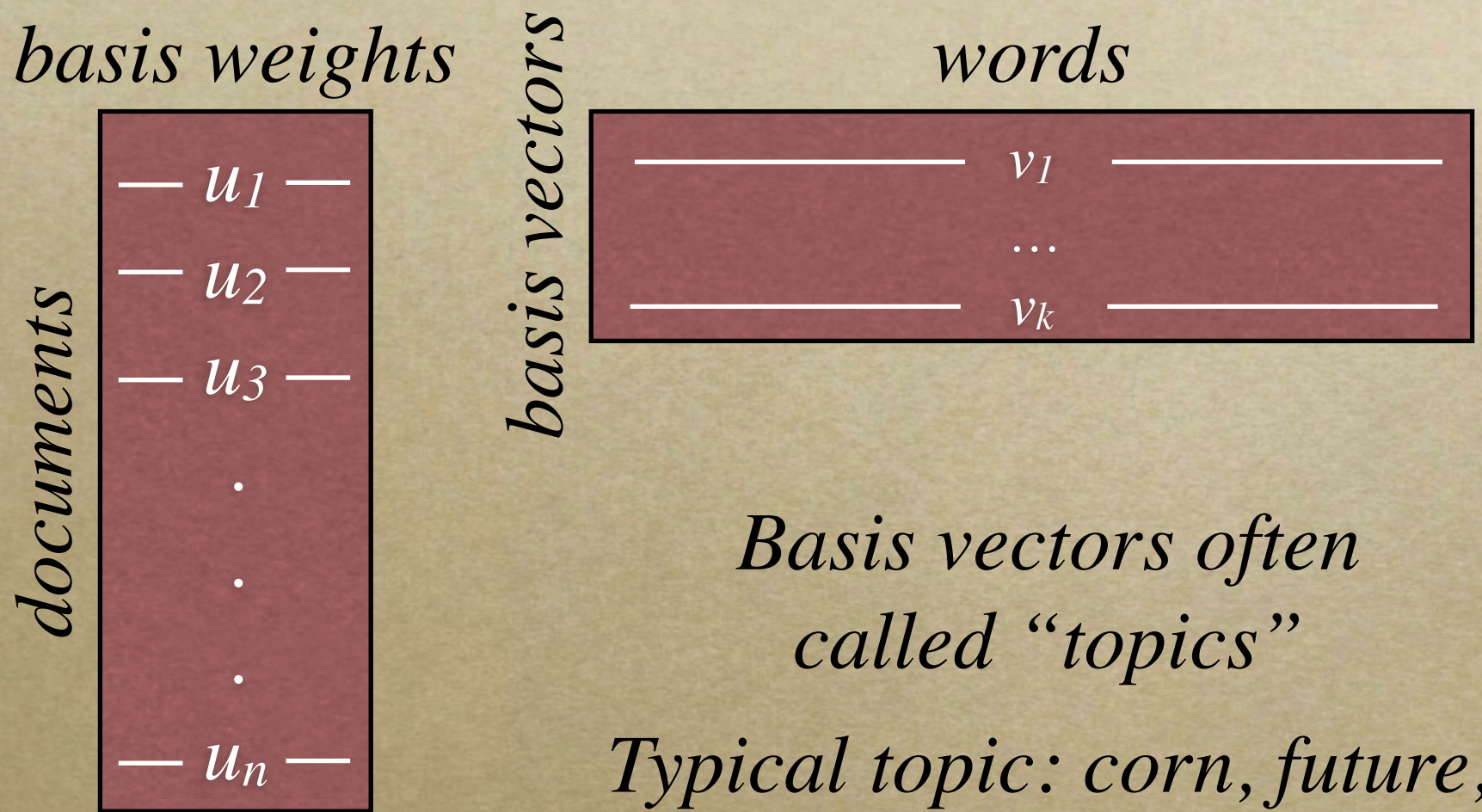
Reuter

article: Reuters data set

Data matrix: “bag of words”



Result of factoring



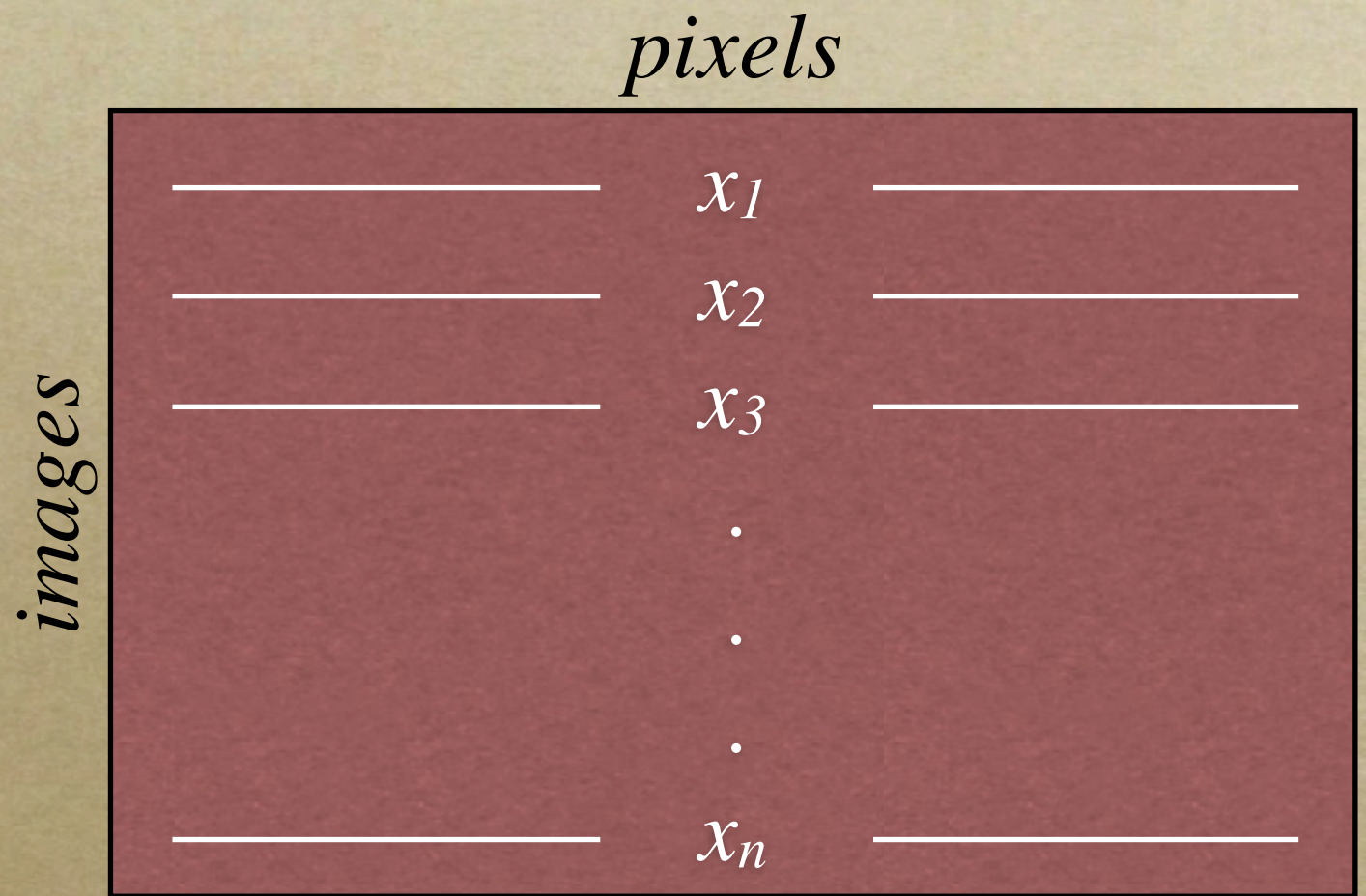
Basis vectors often called "topics"

Typical topic: corn, future, price, weather, market, …

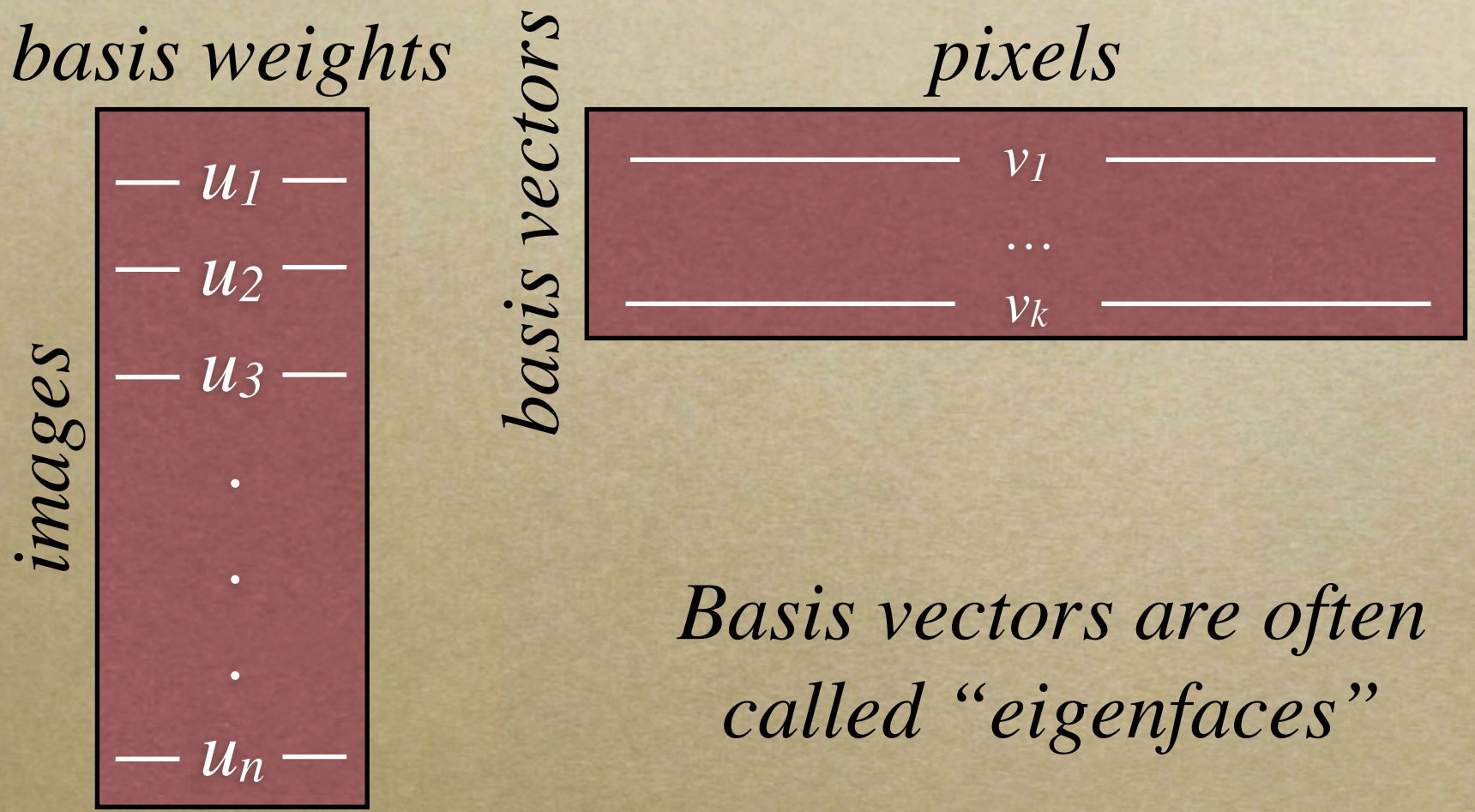
Example: images

face images from Groundhog Day, extracted by Cambridge face DB project

Data matrix



Result of factoring



Basis vectors are often called "eigenfaces"

Eigenfaces

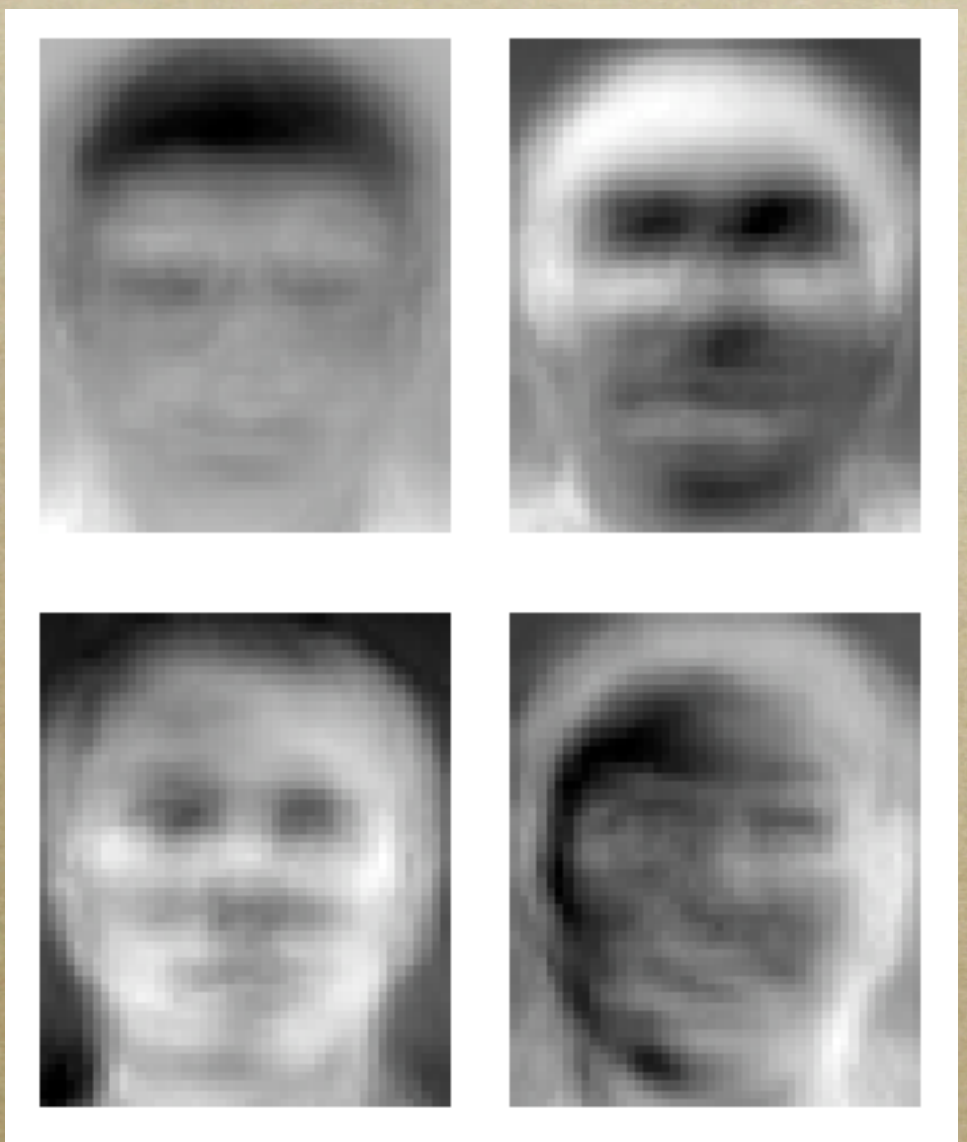
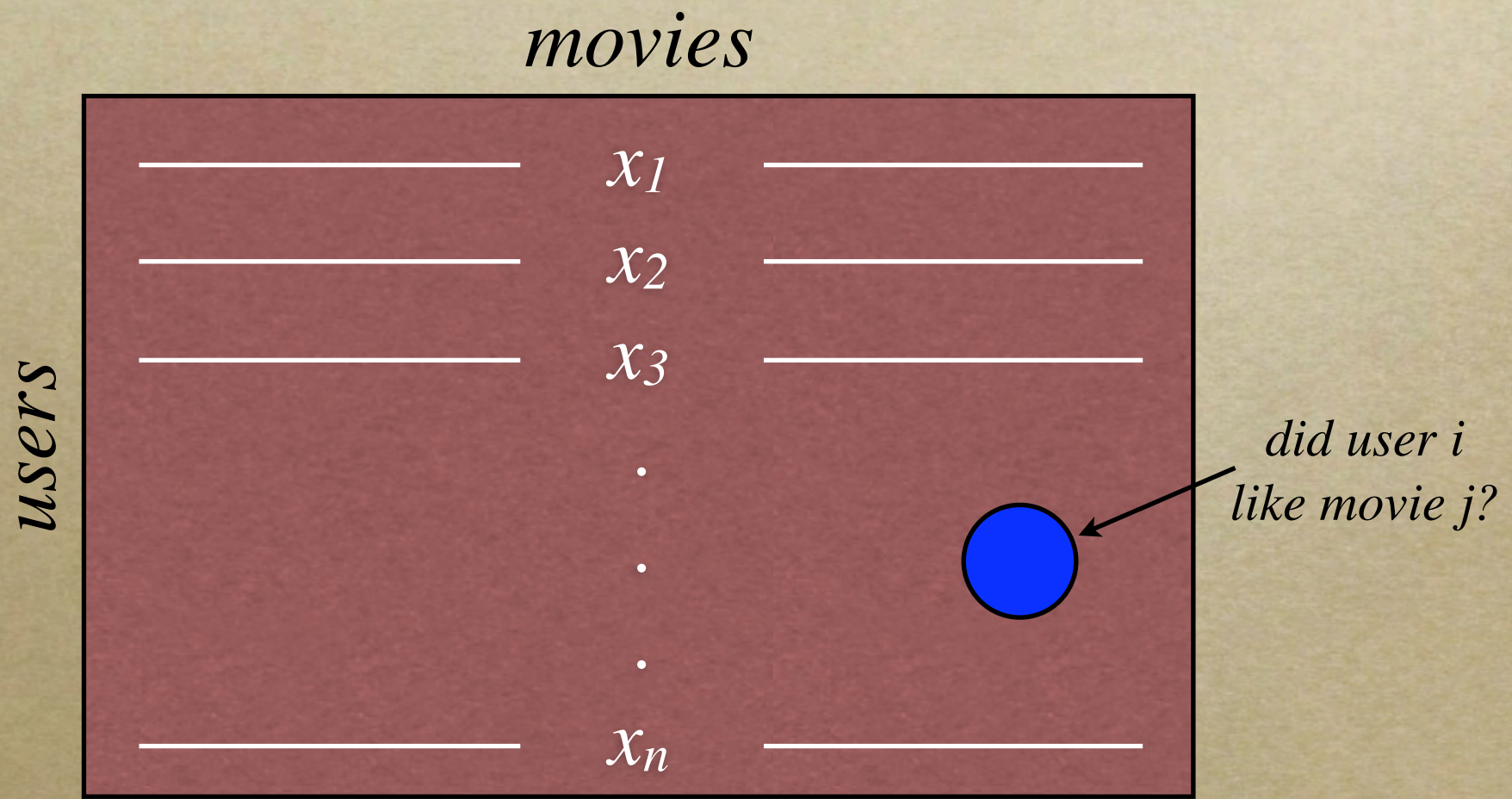


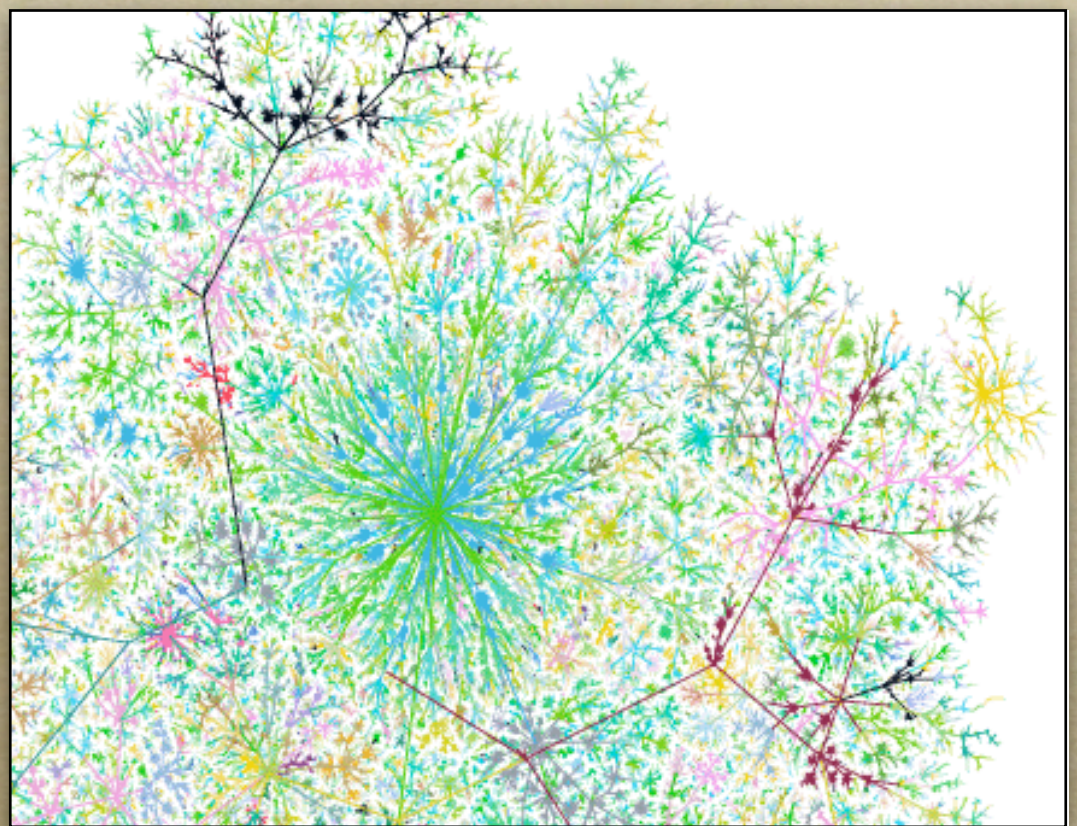
image credit: AT&T Labs Cambridge

Example: collaborative filtering

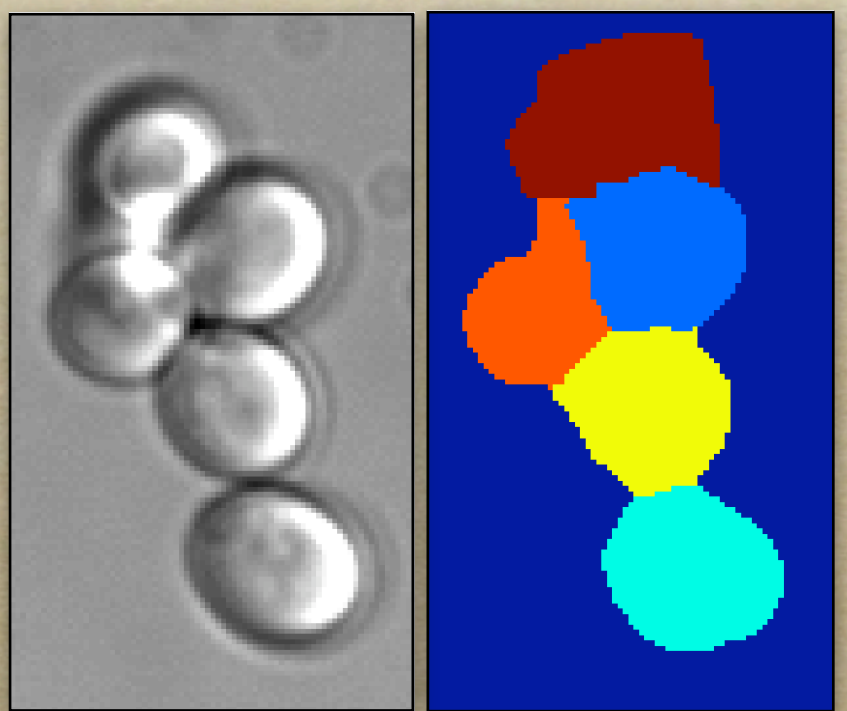


And so forth…

web graph



cells *segmented*



- *bibliometrics, spectral clustering, speech recognition, PageRank, …*

image credit: http://www.cheswick.com/ches/

If you have a hammer

image credit: Wikipedia

This talk: some more nails

image credit: Wikipedia

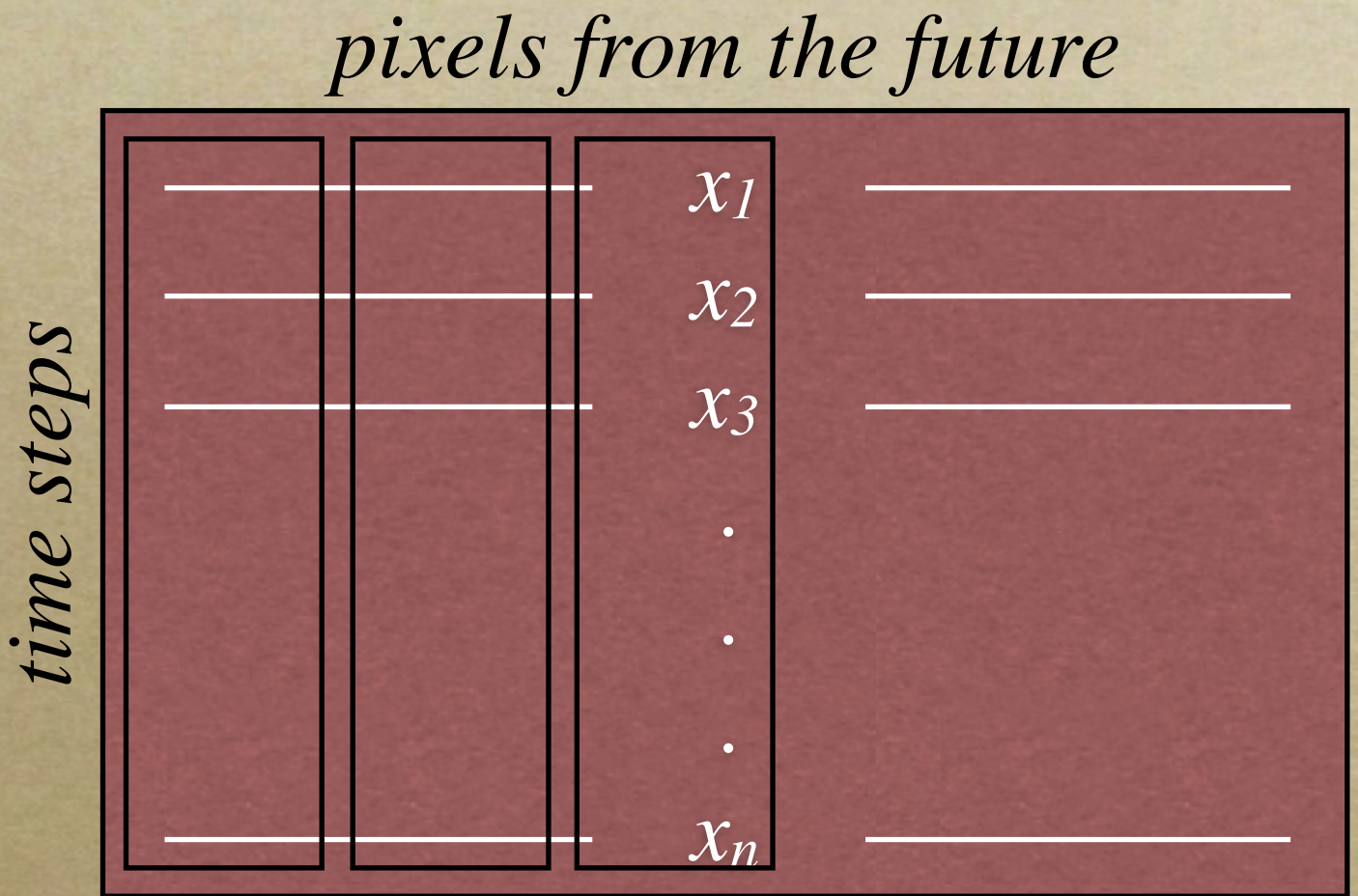
…and some bigger hammers

image credit: Wikipedia

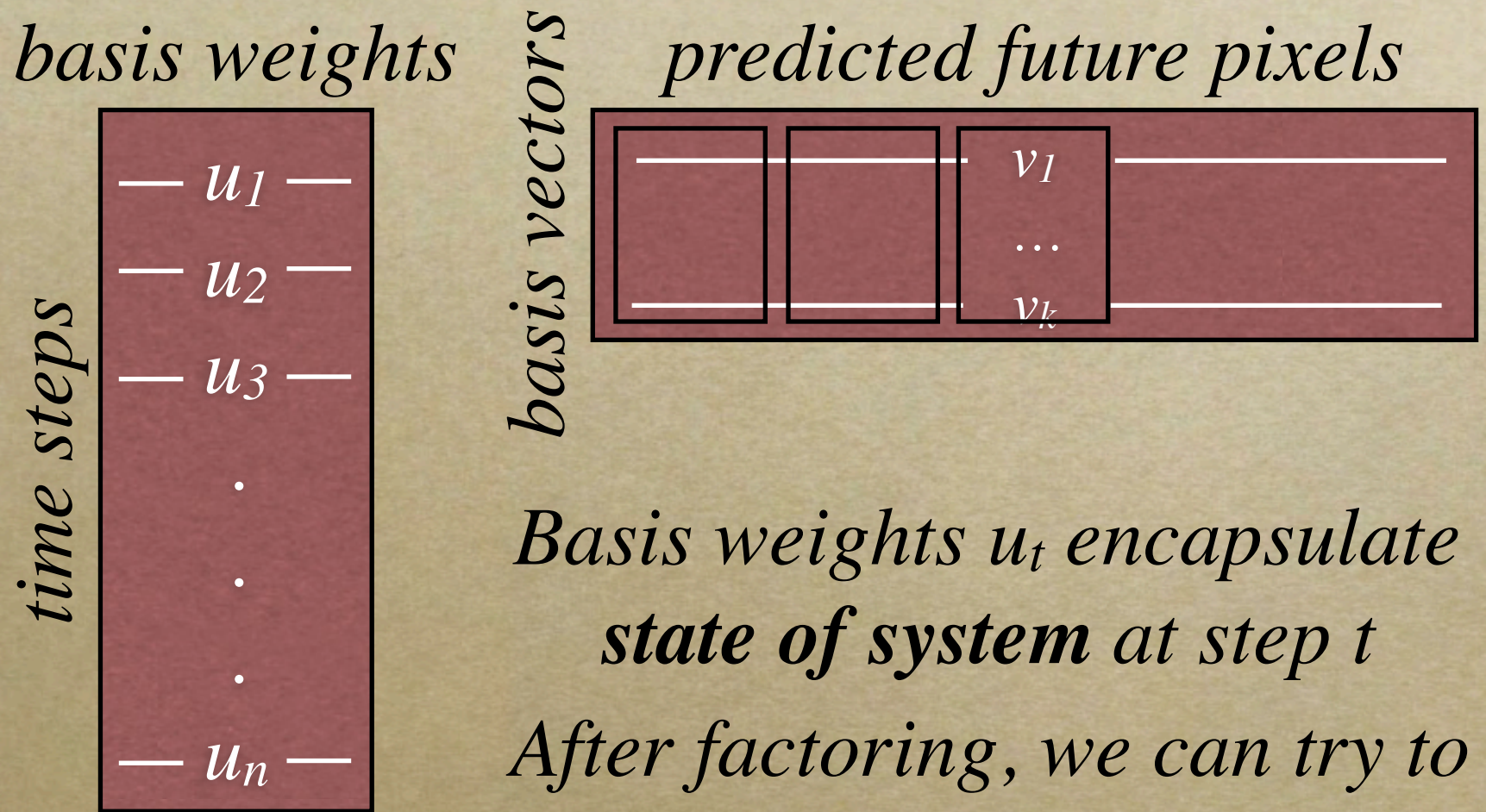
Interesting nail #1: movies

Can we learn to predict future frames?

Data matrix: pixel sequences



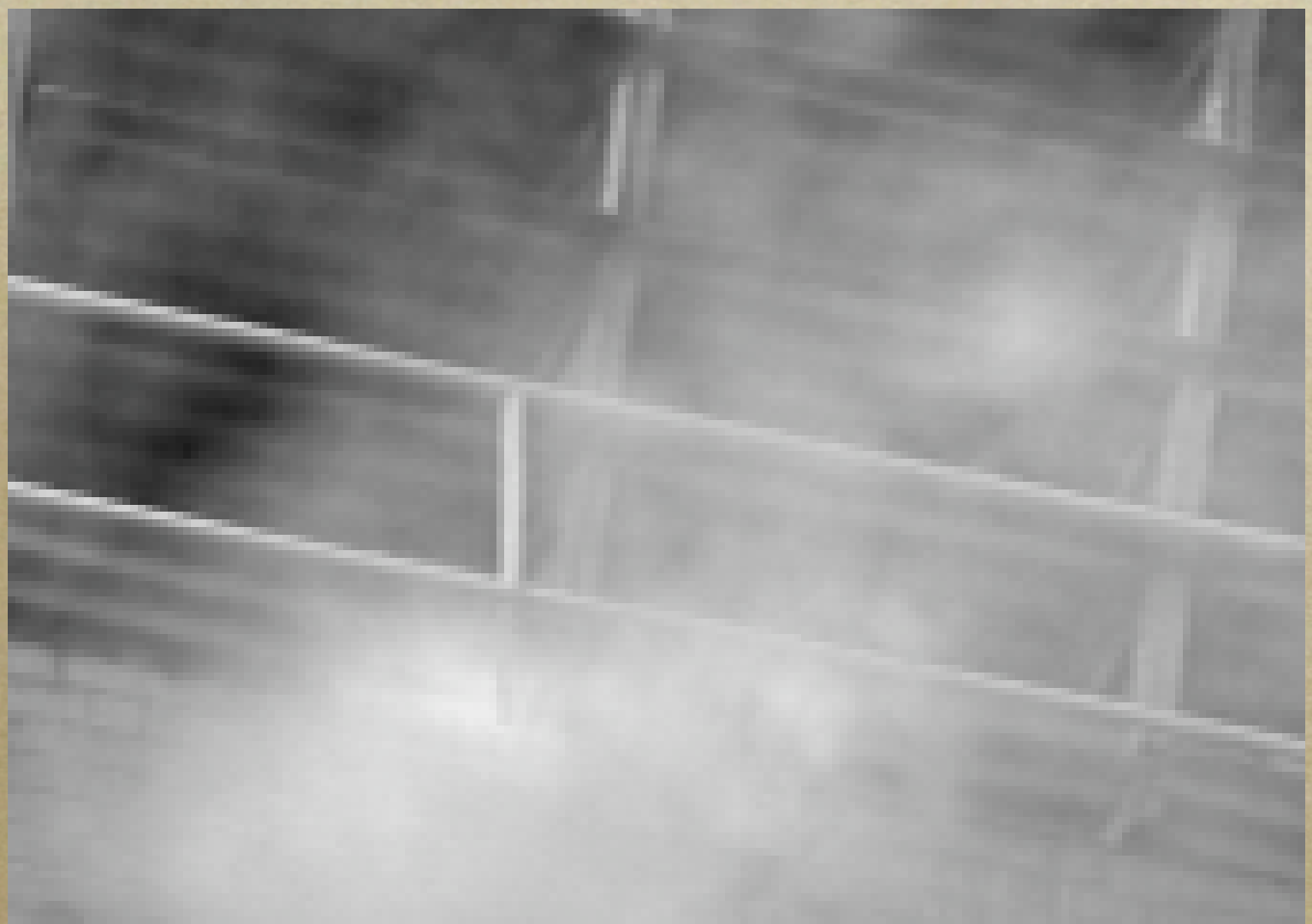
Result of factoring



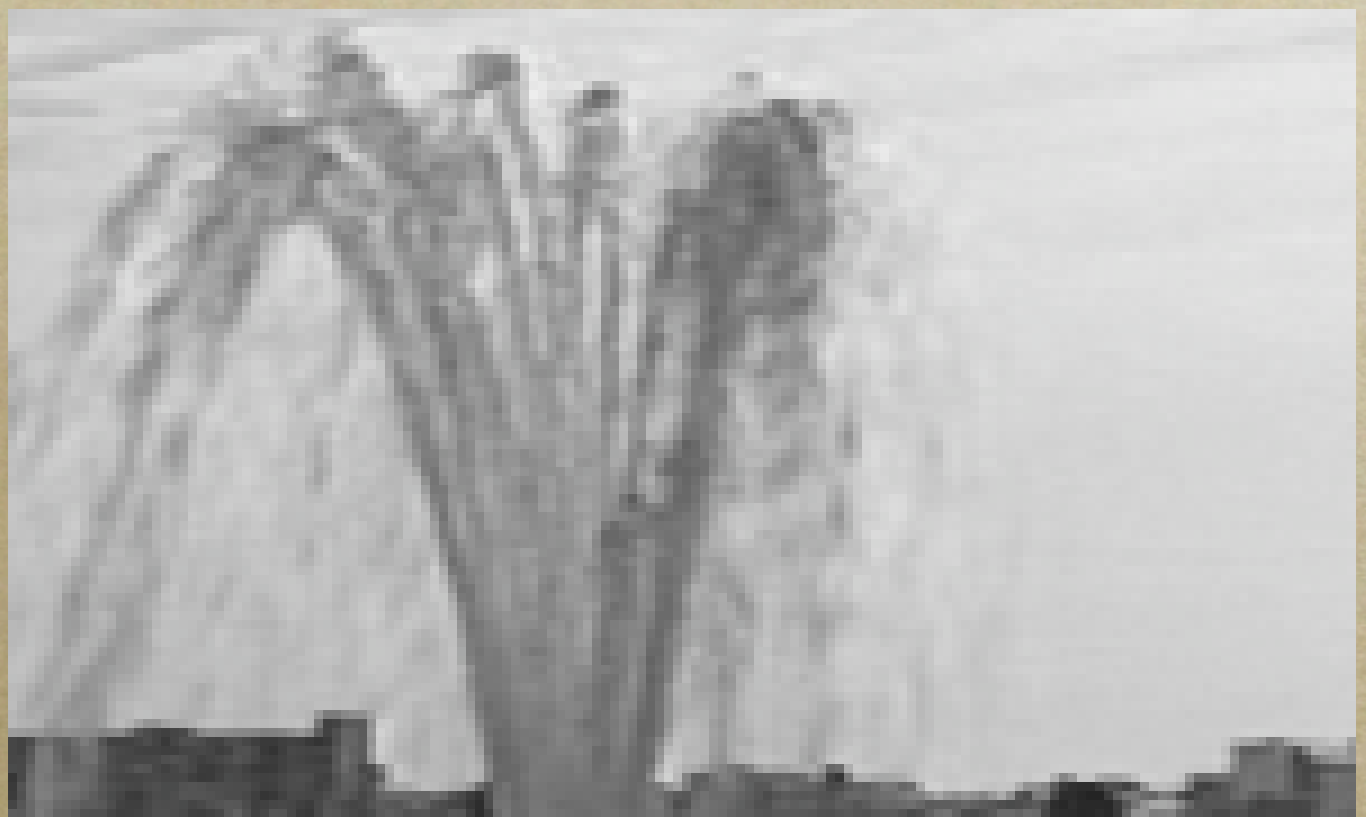
Basis weights u_t encapsulate ***state of system*** *at step t*

After factoring, we can try to learn to predict u_{t+1} from u_t

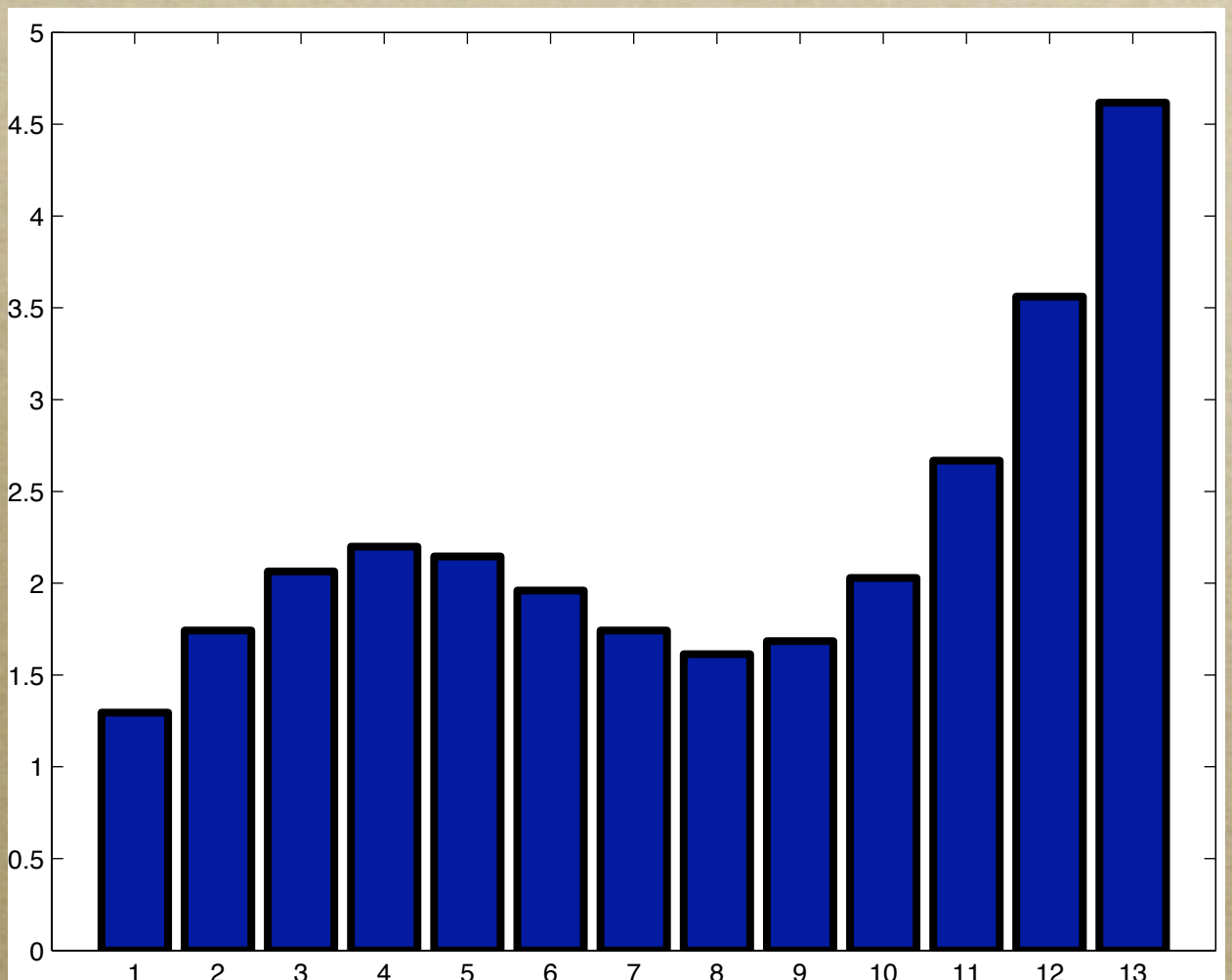
Movie example



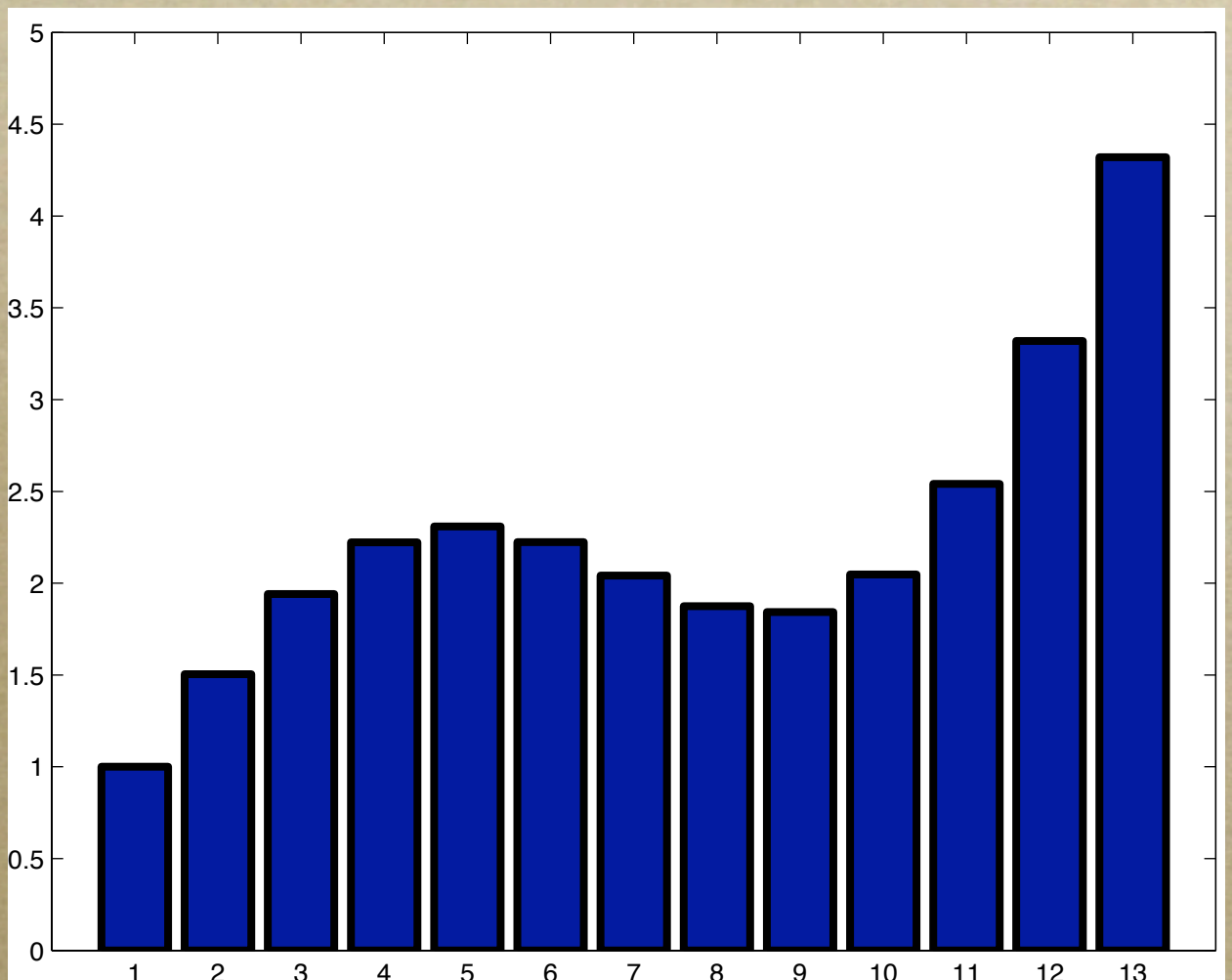
Movie example



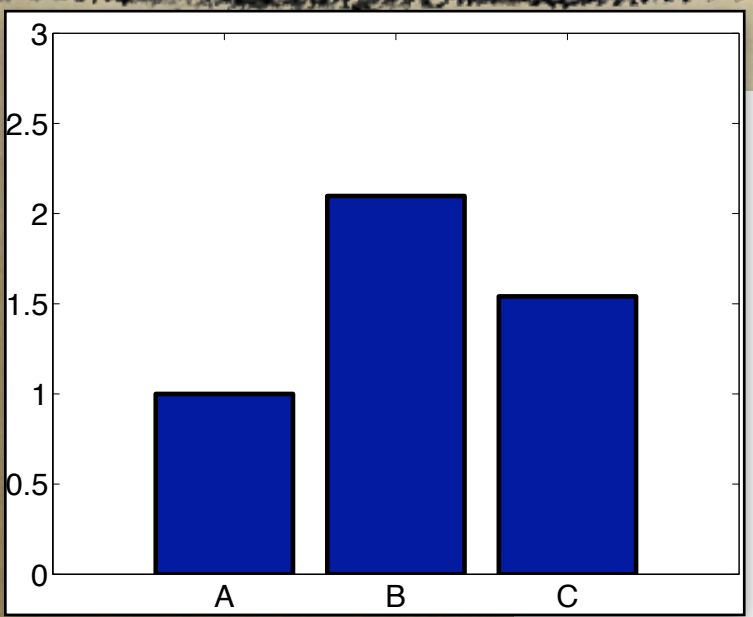
Interesting nail #2: histograms

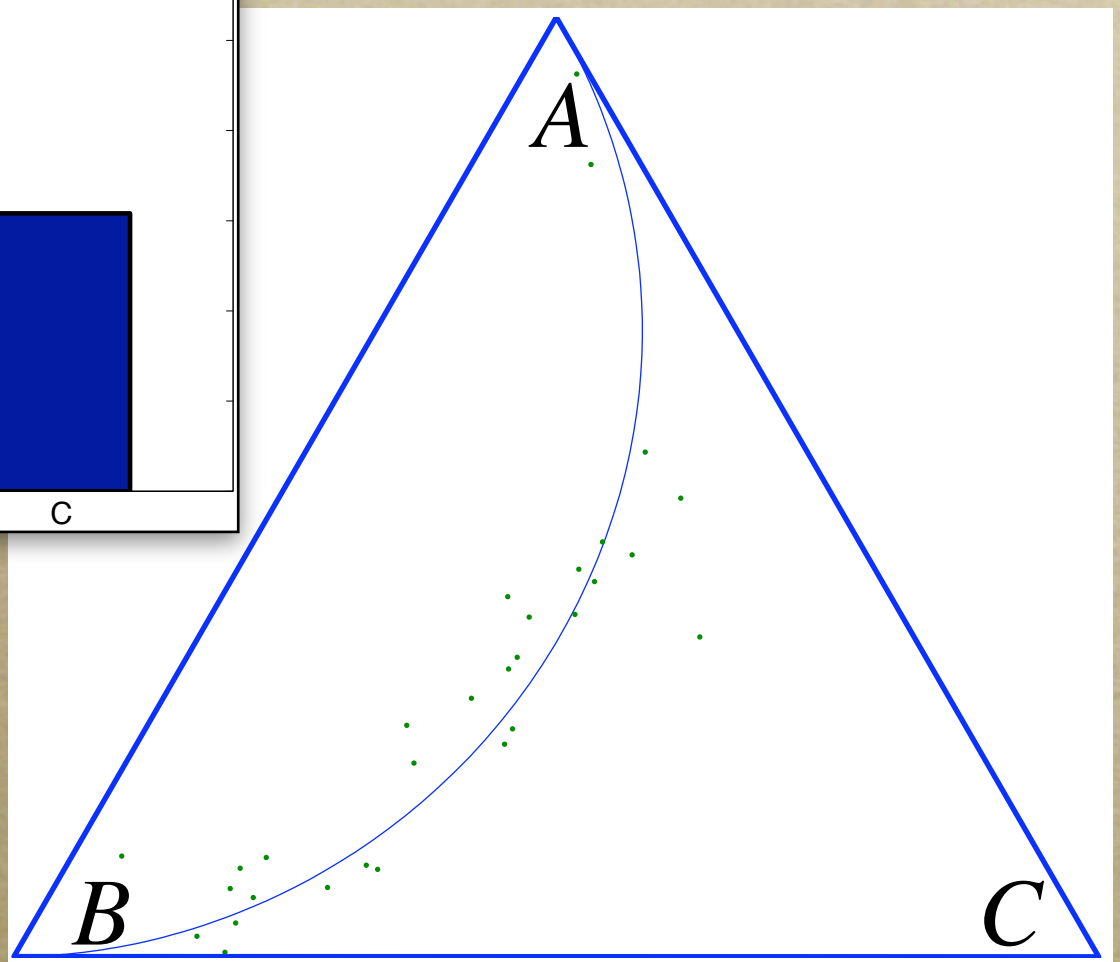


Interesting nail #2: histograms



Compressing a set of histograms





Points: observed histograms in $\mathbb{R}^3$

Line: a 1-parameter family of multinomials

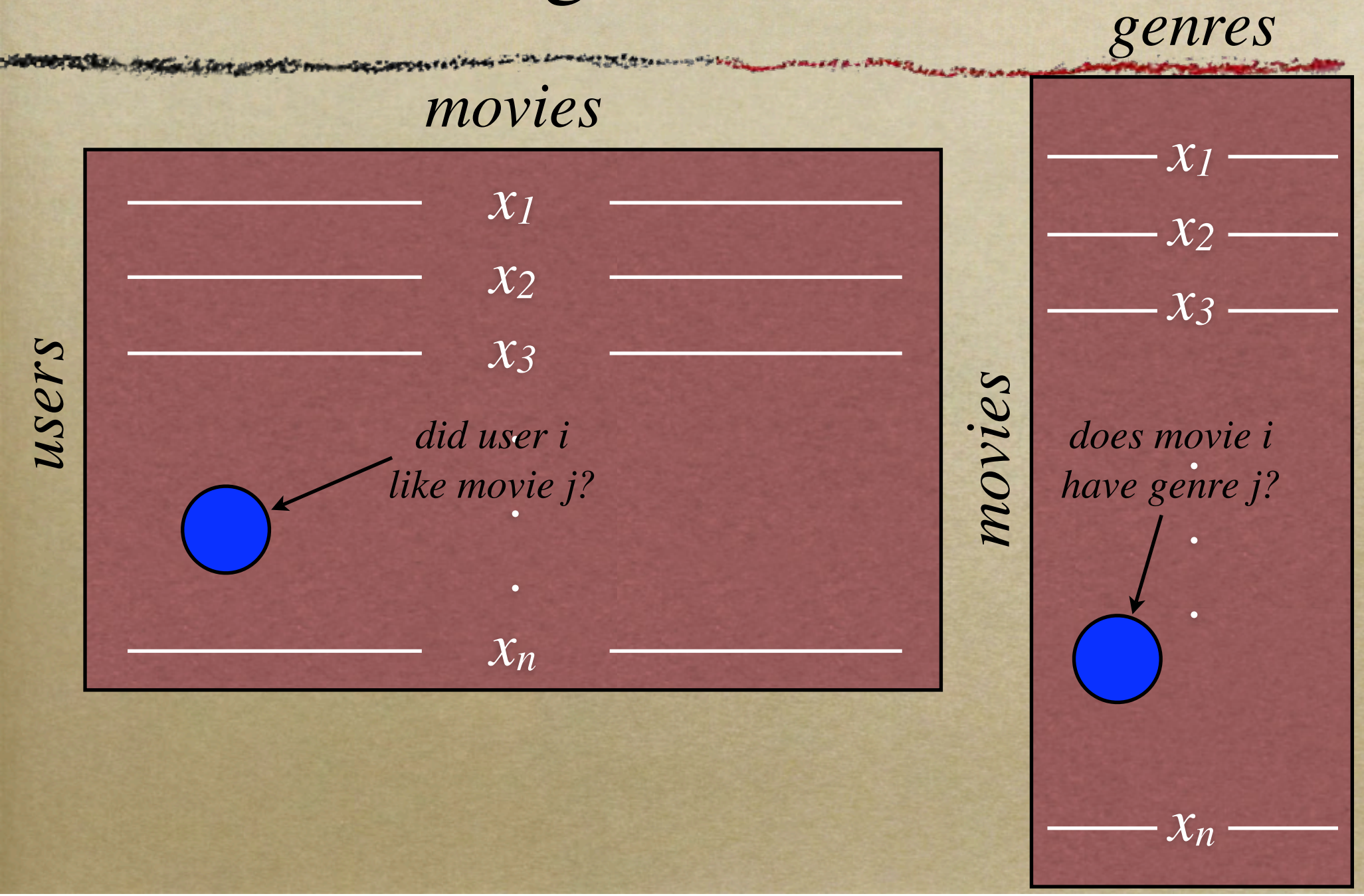
Why compress histograms?

Robot beliefs

Why compress histograms?

- *Single histogram = test performance distribution (of multiple students on one test, or one student on multiple tests)*

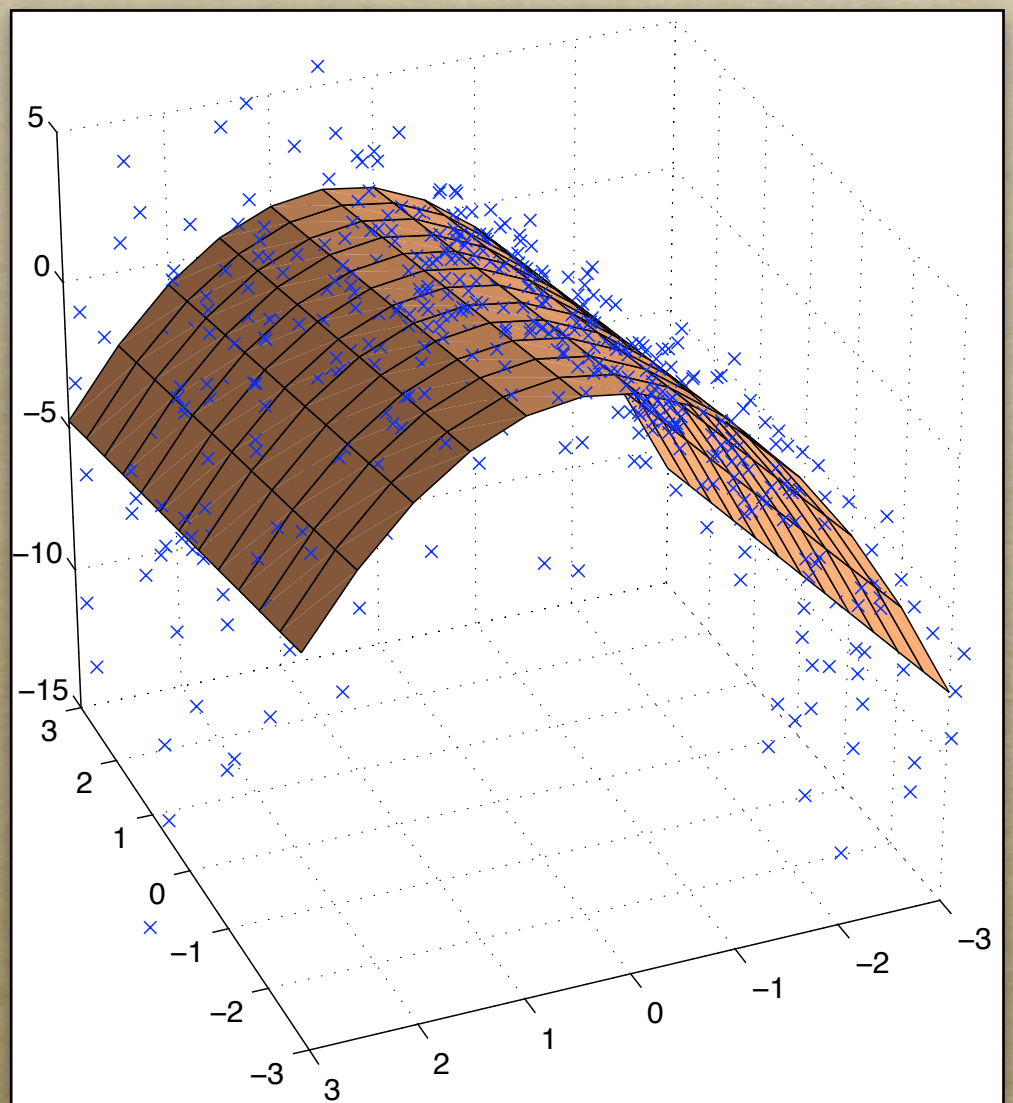
Interesting nail #3: relations



Bigger hammers

- *Nonlinearity*
- *Relational factorization*

Bigger hammer: nonlinearity



And its brother: non-normality

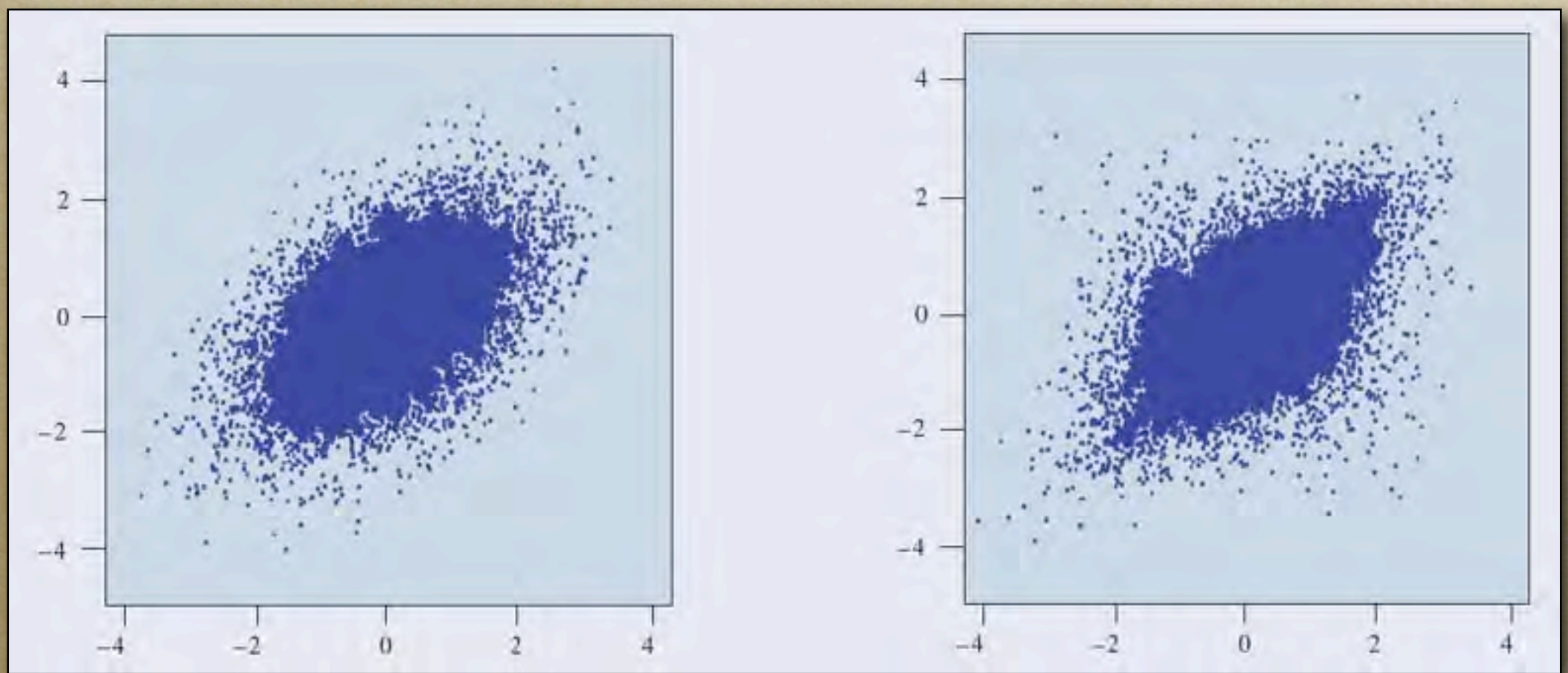


image credit: risklab.ch

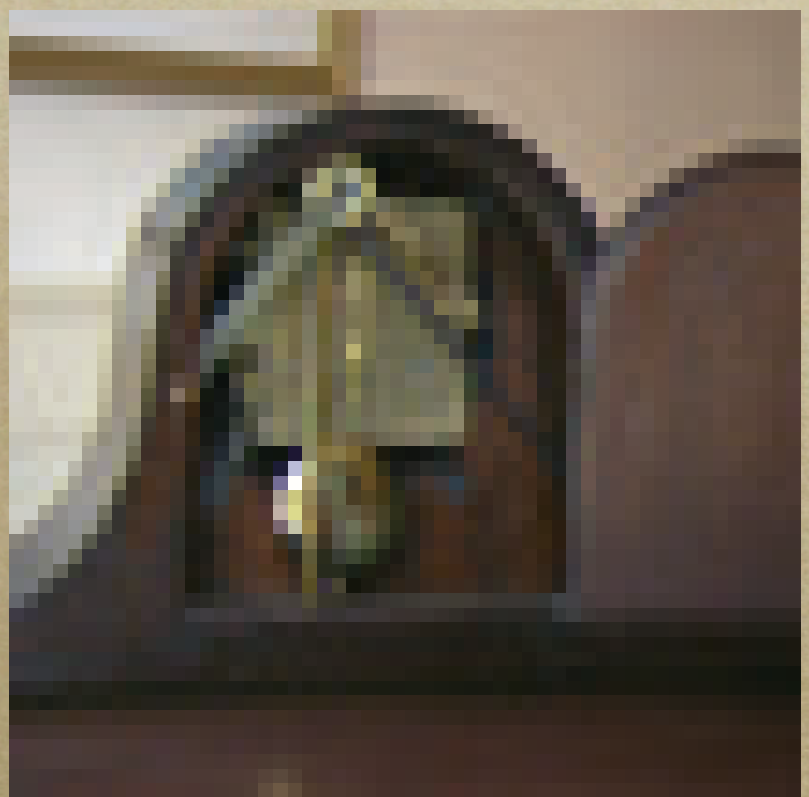
Where do we need nonlinearity?

- *Bounds:*
  - *Images: can't have pixel >1 or <0*
  - *Word counts: can't be negative*

Where do we need nonlinearity?

- *Consistency of interpolation*
  - *Suppose x_1, x_2 are samples*
  - *Linearity means $x_1 + \lambda(x_2 - x_1)$ should make sense* ***for any λ***

Consistency of interpolation



training
data

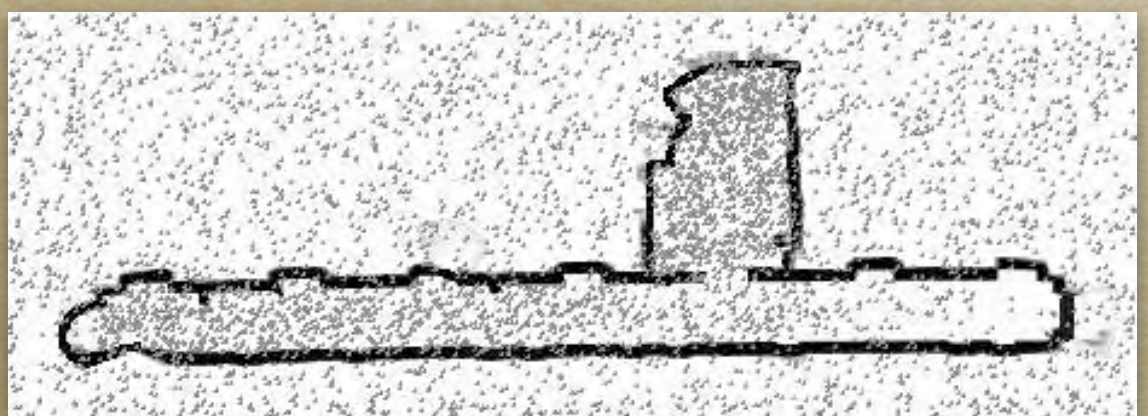
predicted
image

Where do we need non-normality?

- *Uneven variance*
  - *Error of ±3 in count of 5 vs. 50 vs. 500*
- *Asymmetric errors*
  - *E.g., counts often have long tail*

Uneven variance, asymmetric errors

Original



Learned (PCA)

Generalizing PCA

- *In PCA, had* $X_{ij} \approx U_i \cdot V_j$
- *What if*
  - $X_{ij} \approx exp(U_i \cdot V_j)$
  - $X_{ij} \approx logit(U_i \cdot V_j)$
  - …

Generalizing PCA

- *In PCA, had* $X_{ij} \sim Normal(\mu)$, $\mu = U_i \cdot V_j$
- *What if*
  - $X_{ij} \sim Poisson(\mu)$
  - $X_{ij} \sim Binomial(p)$
  - *…?*

Exponential family review

- *Exponential family of distributions:*

$$P(X \mid \theta) = P_0(X)\, \exp(X\theta - g(\theta))$$

- *$g(\theta)$ is always strictly convex, differentiable on interior of domain*
  - *means g' is strictly monotone (in 1D)*

Exponential family review

- *Exponential family PDF:*

  $$P(X \mid \theta) = P_0(X)\, \exp(X\theta - g(\theta))$$

- $g'(\theta) = E(X \mid \theta)$
  - g' & $(g')^{-1}$ = *"link function"*
  - $E(X \mid \theta)$ = *"expectation parameter"*

Examples

- *Normal(mean)*
- *Poisson(log rate)*
- *Binomial(log odds)*

Solving both problems at once

- *Let $P(X \mid \theta)$ be an exponential family with natural parameter θ*
- *Predict $X_{ij} \sim P(X \mid \theta_{ij})$, where $\theta_{ij} = U_i \cdot V_j$*
  - *e.g., in Poisson, $E(X_{ij}) = exp(\theta_{ij})$*
  - *e.g., in Binomial, $E(X_{ij}) = logit(\theta_{ij})$*

More precisely,

$$\max_{U,V} \sum \log P(X_{ij} \mid \theta_{ij}) + \log P(U) + \log P(V)$$
$$\text{s.t.} \quad \theta_{ij} = U_i \cdot V_j$$

- *"Generalized linear" or "exponential family" PCA*
  - *all P(…) terms are exponential families*
  - *analogy to GLMs*

[Collins et al, 2001]
[Gordon, 2002]
[Roy & Gordon, 2005]

Theorem

In GL PCA, finding U which maximizes likelihood (holding V fixed) is a convex optimization problem

And, finding best V (holding U fixed) is a convex problem

Further, Hessian is block diagonal

- *So, an efficient and effective optimization algorithm: alternately improve U and V*

Proof

- *Exponential family PDF:*

$$P(X \mid \theta) = P_0(X)\, exp(X\theta - g(\theta))$$

- $\sum log\, P(X_{ij} \mid \theta_{ij}) =$

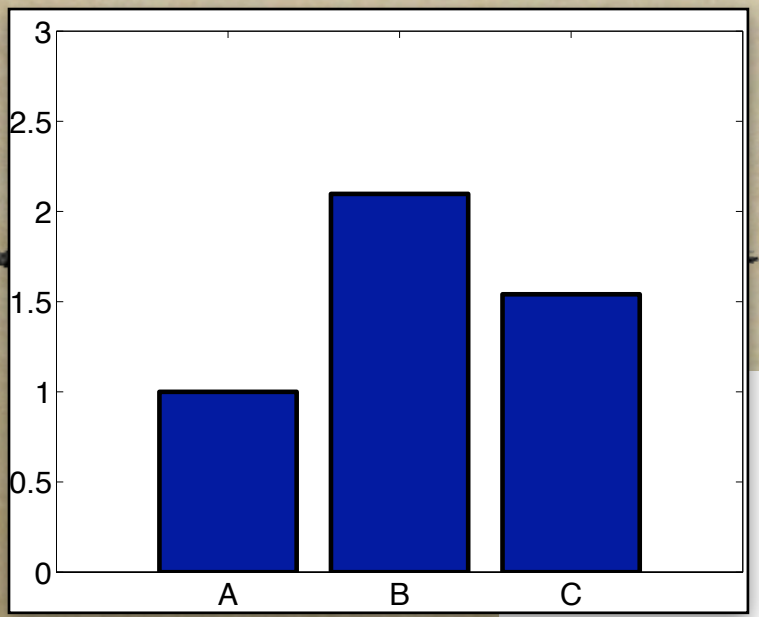
$\sum [log\, P_0(X_{ij}) - X_{ij}\theta_{ij} - g(\theta_{ij})] =$

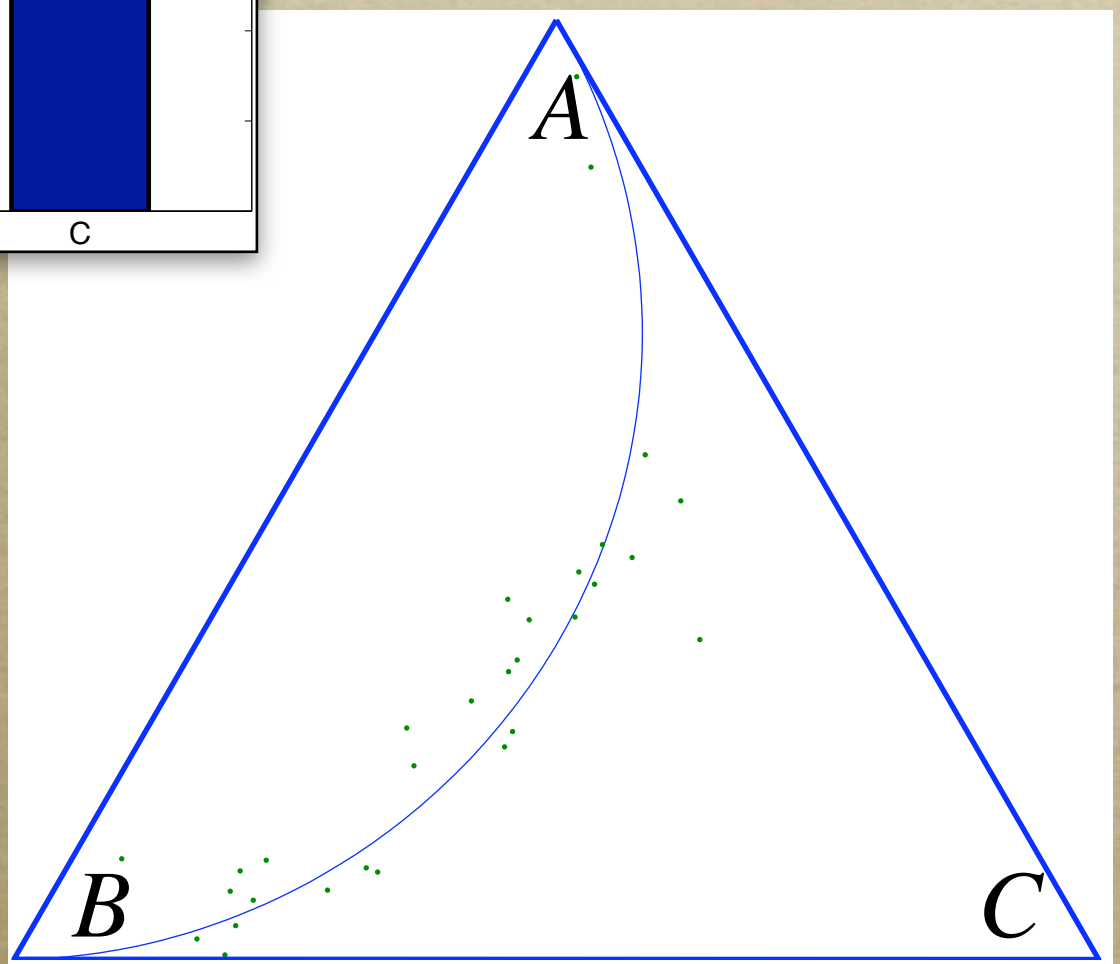
$\sum [log\, P_0(X_{ij}) - X_{ij}(U_i^T V_j) - g(U_i^T V_j)]$

Special cases

- *PCA, probabilistic PCA*
- *k-means clustering*
- *Independent components analysis (ICA)*
- *Poisson PCA*
- *Max-margin matrix factorization (MMMF)*
- *Almost: pLSI, pHITS, NMF*

Example

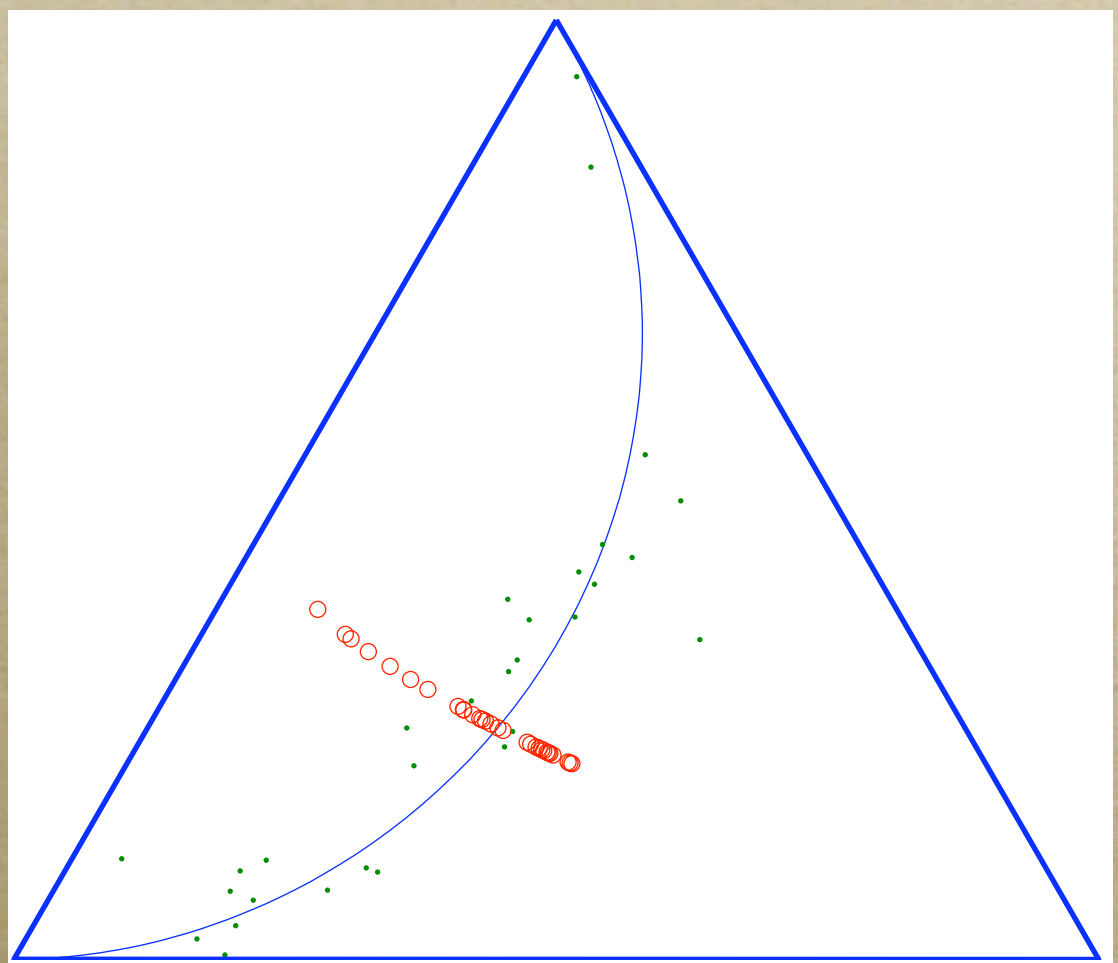




Points: observed frequencies in $\mathbb{R}^3$

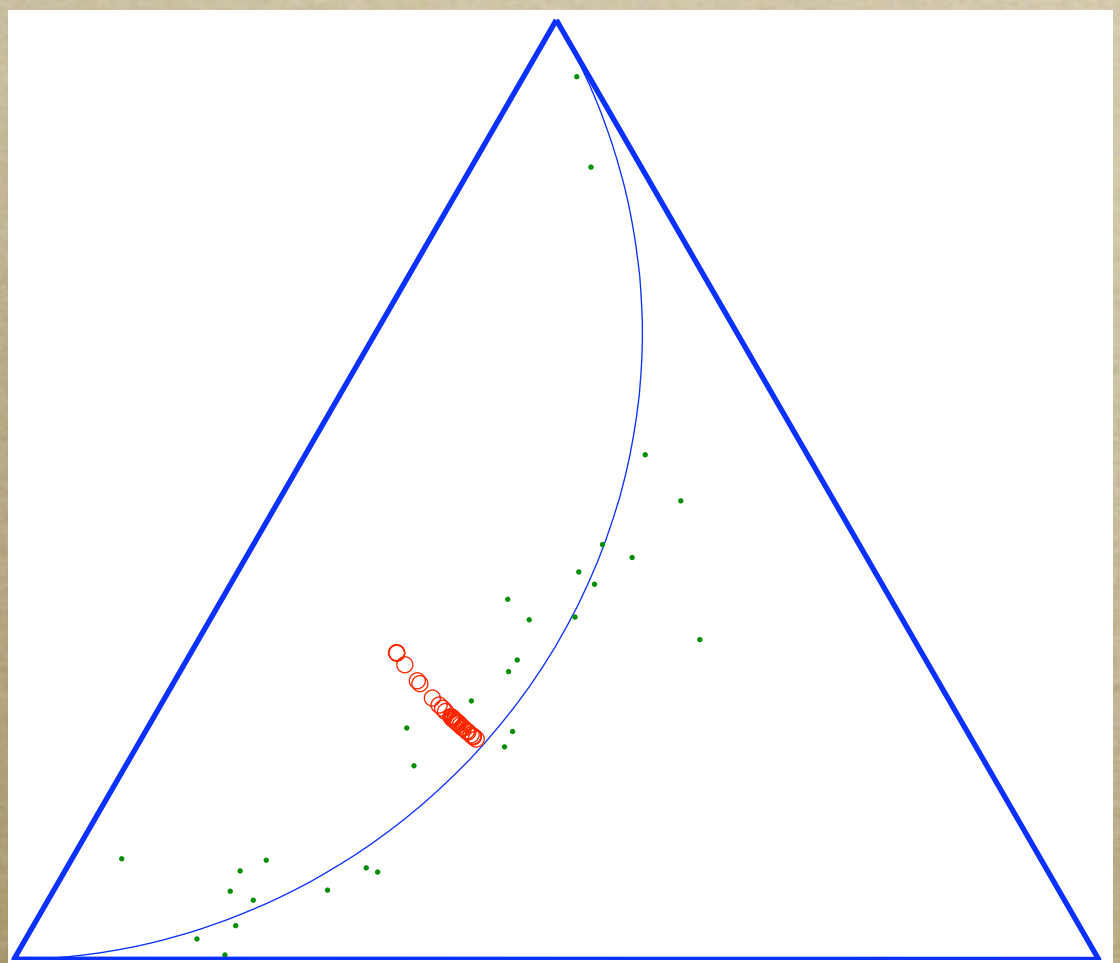
Hidden manifold: a 1-parameter family of multinomials

Example



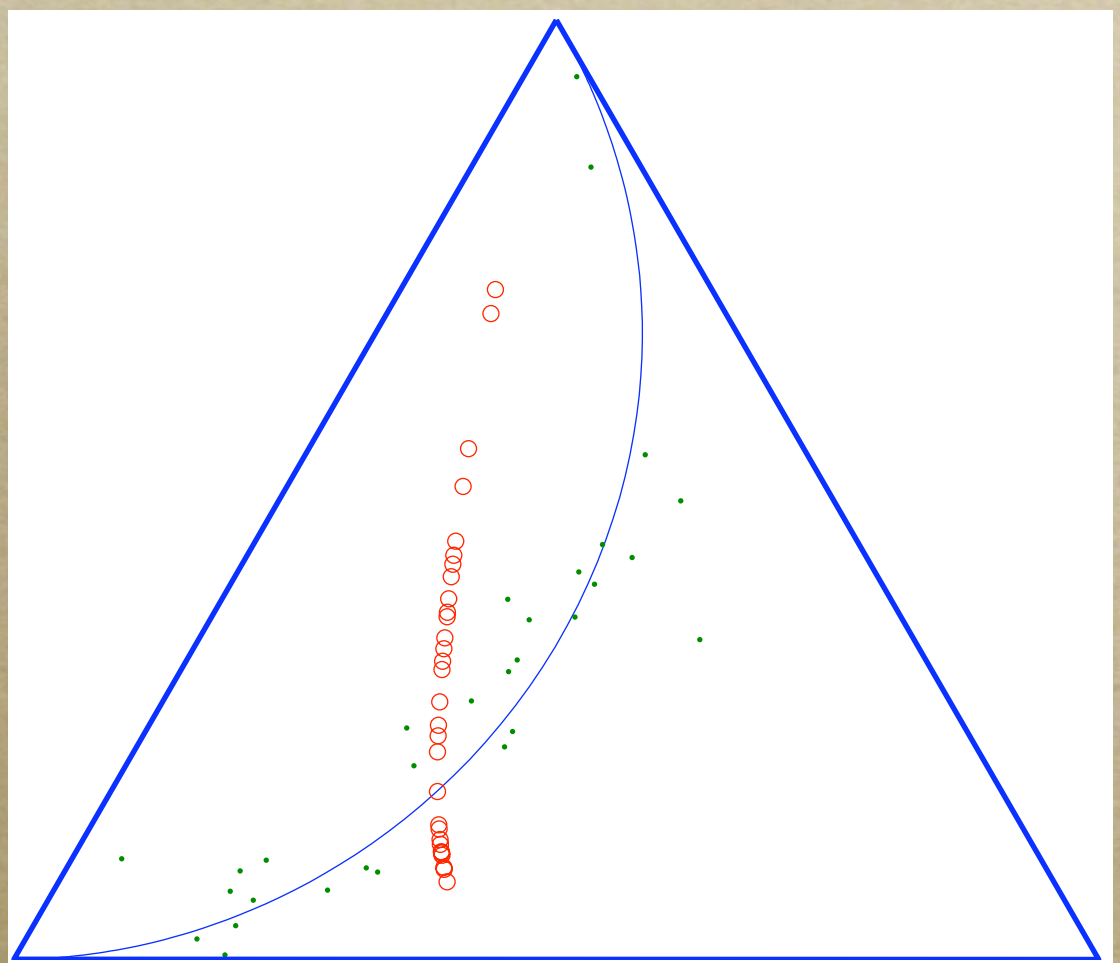
Iteration 1

Example



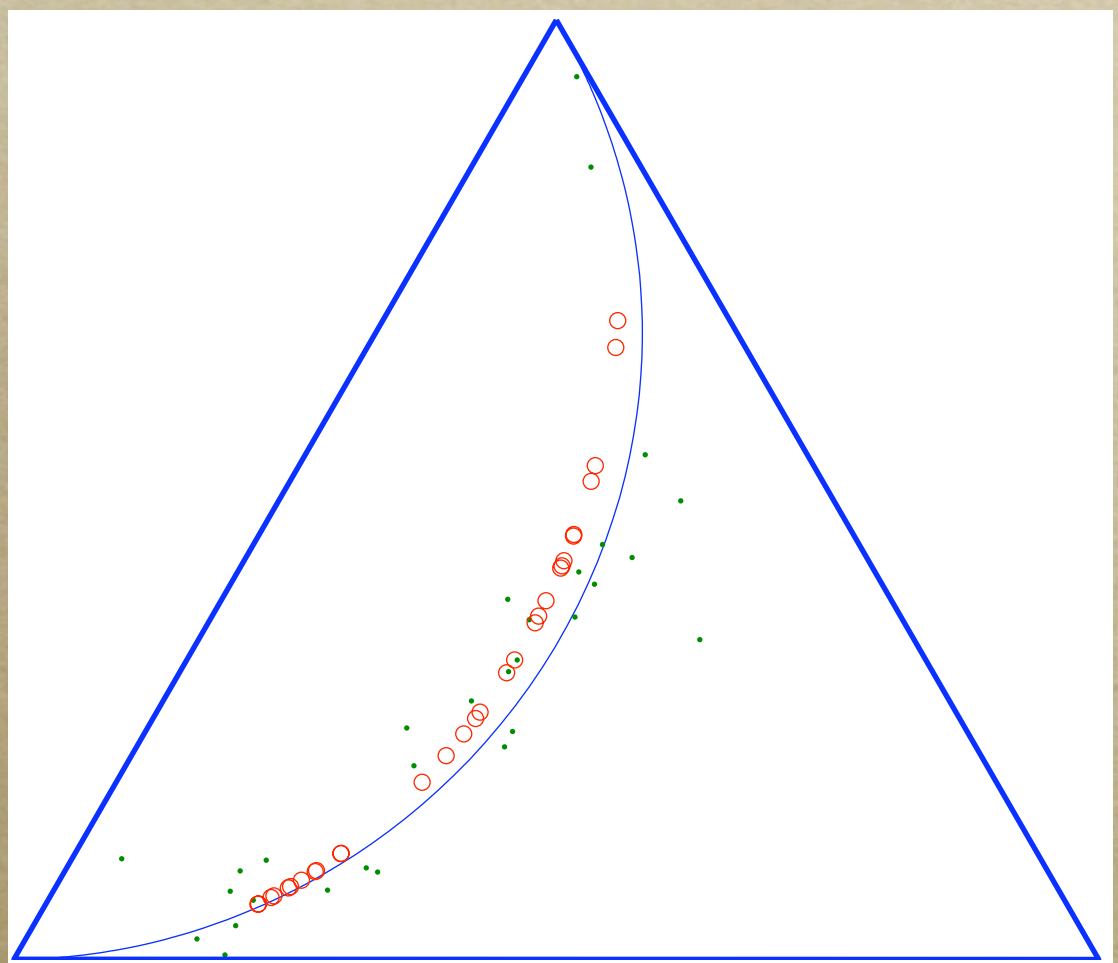
Iteration 2

Example



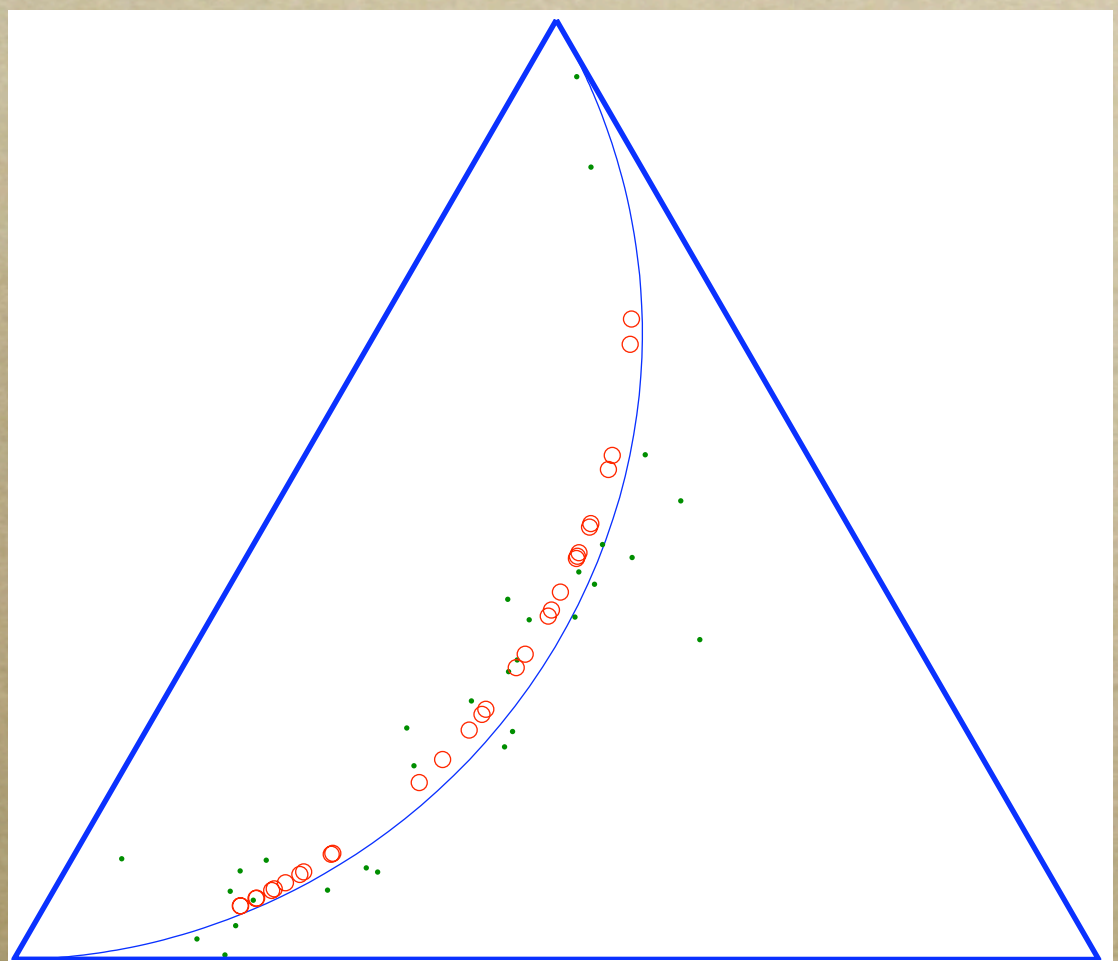
Iteration 3

Example



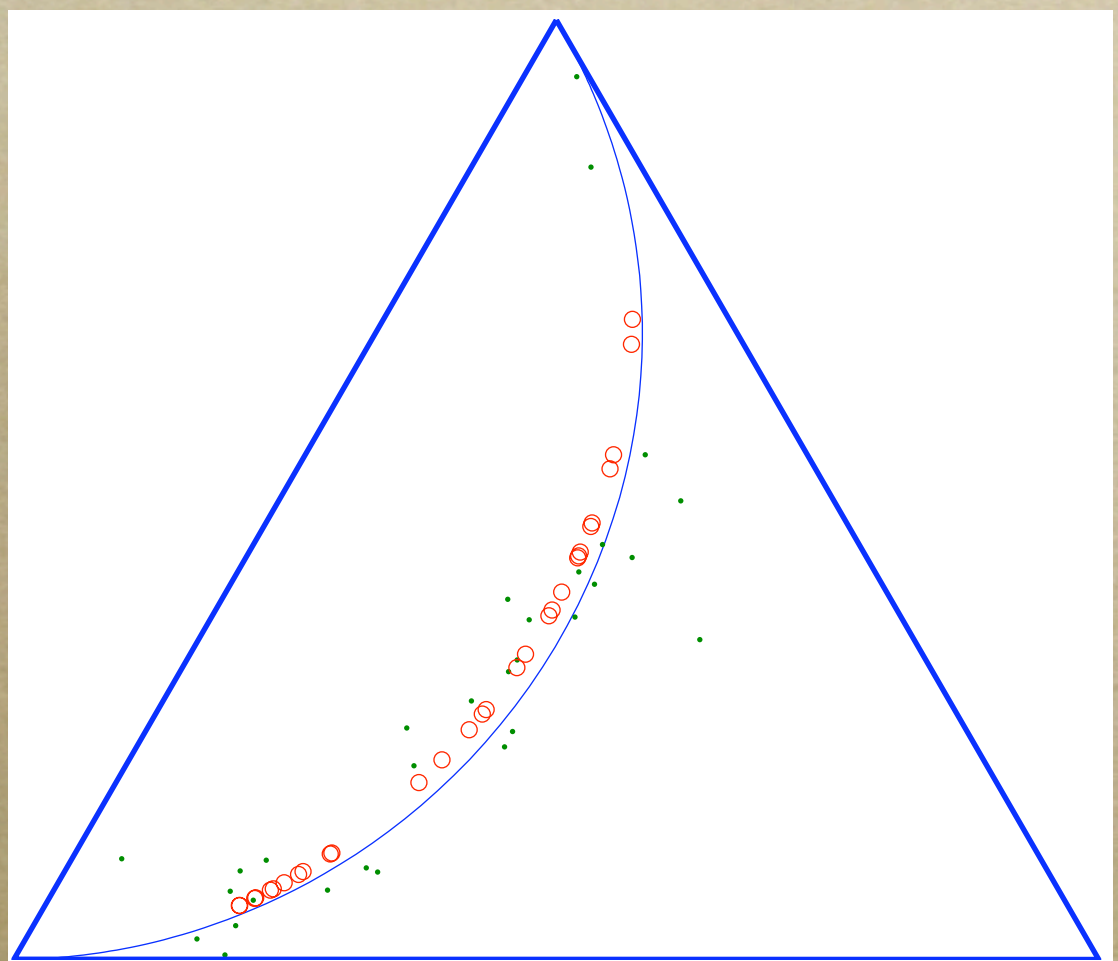
Iteration 4

Example



Iteration 5

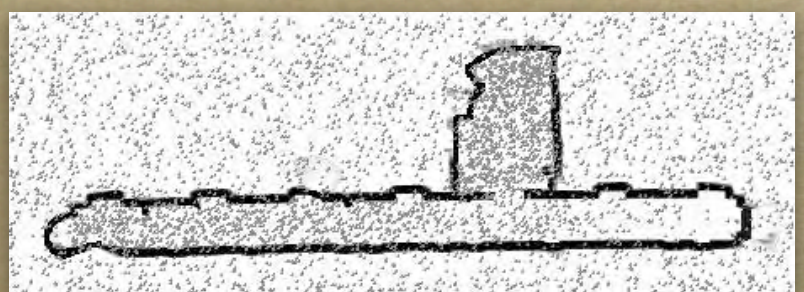
Example



Iteration 9

Robot beliefs example

Original



Learned (PCA)

Learned (Poisson PCA)