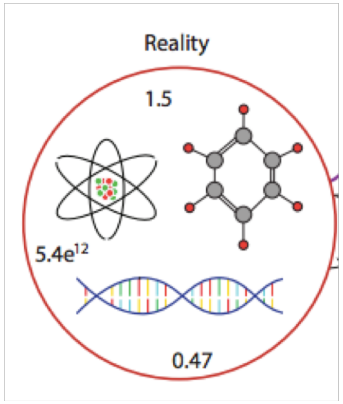
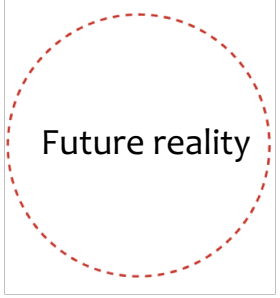
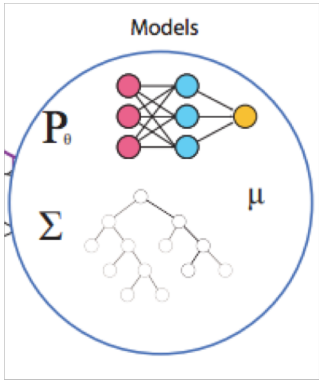






Statistics 154, Spring 2019

Modern Statistical Prediction and Machine Learning

Lecture 6: CV for PCA, NMF, and hierarchical

Instructor: Bin Yu
(binyu@berkeley.edu); office hours: Tu: 9:30-10:30 am; Wed:**1:30-2:30 pm (change)**
office: 409 Evans

GSIs: Yuansi Chen (Mon: 10-12; 4-6); Raaz Dwivedi (Mon: 12-2; 2-4)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu
(Yuansi: Tuesday 1-3; Raaz: Monday 10:30-11:30, Thurs. 9:30-10:30)

Taxonomy

of Machine Learning/Statistics

Labeled Data

Indirect (reward)

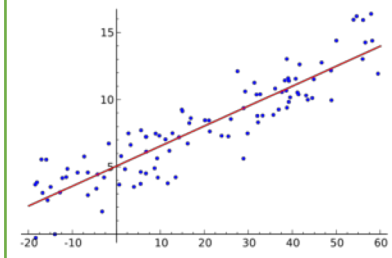
Unlabeled Data

Supervised Learning

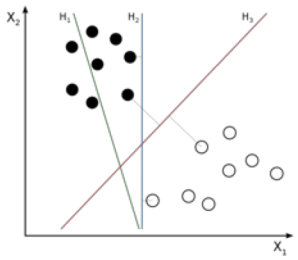
Reinforcement & Bandit Learning

Unsupervised Learning

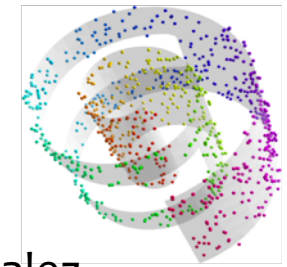
Regression



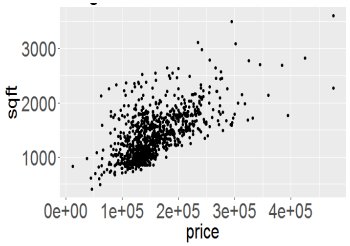
Classification



Dimensionality Reduction



Clustering



Thanks to J. Gonzalez

How do we visualize beyond 4 dimensions? Ames data has 80 features

- We can at most "look" at 3-dim data (or 4-dim with movies of data)
- Visual cortex takes up about 30% of our brain so we want to see projections of data into low dim spaces (2d or 3d, for example)
- How to project?
- Random projection? Or?

Dimension reduction for more interpretable results of high-dim data

Dimensionality reduction is needed for

- Visualization
- Fast computation
- Smaller storage and faster communication
- One form for regularization: Simpler models (to reduce variance of estimation)

PCA: Core Idea

The central idea of PCA is :

- **to reduce the dimensionality** of a data set that has a large number of **interrelated** variables,
- while **retaining as much as possible of the variation** present in the data set.

This is achieved by **transforming** to a new set of variables (PCs)

- which are **uncorrelated (orthogonal)**, and
- which are **ordered** so that **(hopefully)** the first few retain most of the variation present in all of the original variables.

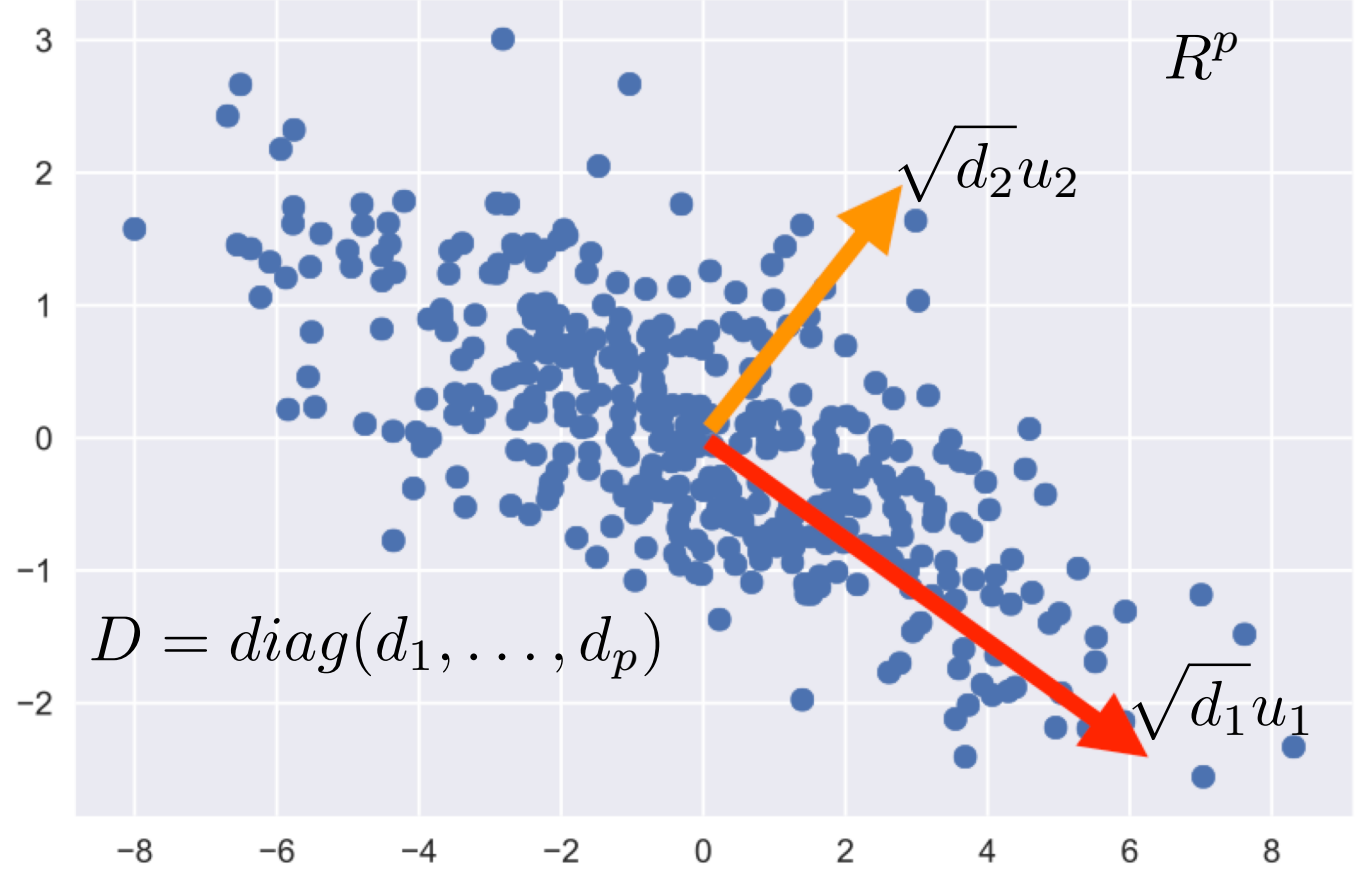
Jolliffe, Principal Component Analysis, 2nd edition

PCA recap, illustrated in p=2 dim

Data matrix $X_{n\times p}$

$G = X'X = UDU^T$ (eigen decomposition)

$U = (u_1, ..., u_p)$ u_j is the jth PC direction, unit vector in R^p



$Z_{n\times p} = XU$

is the data matrix in the new coordinate system spanned by u_1 and u_2

The ith row of Z is called the "principal components" of the ith data point

PCA: maximizing variance

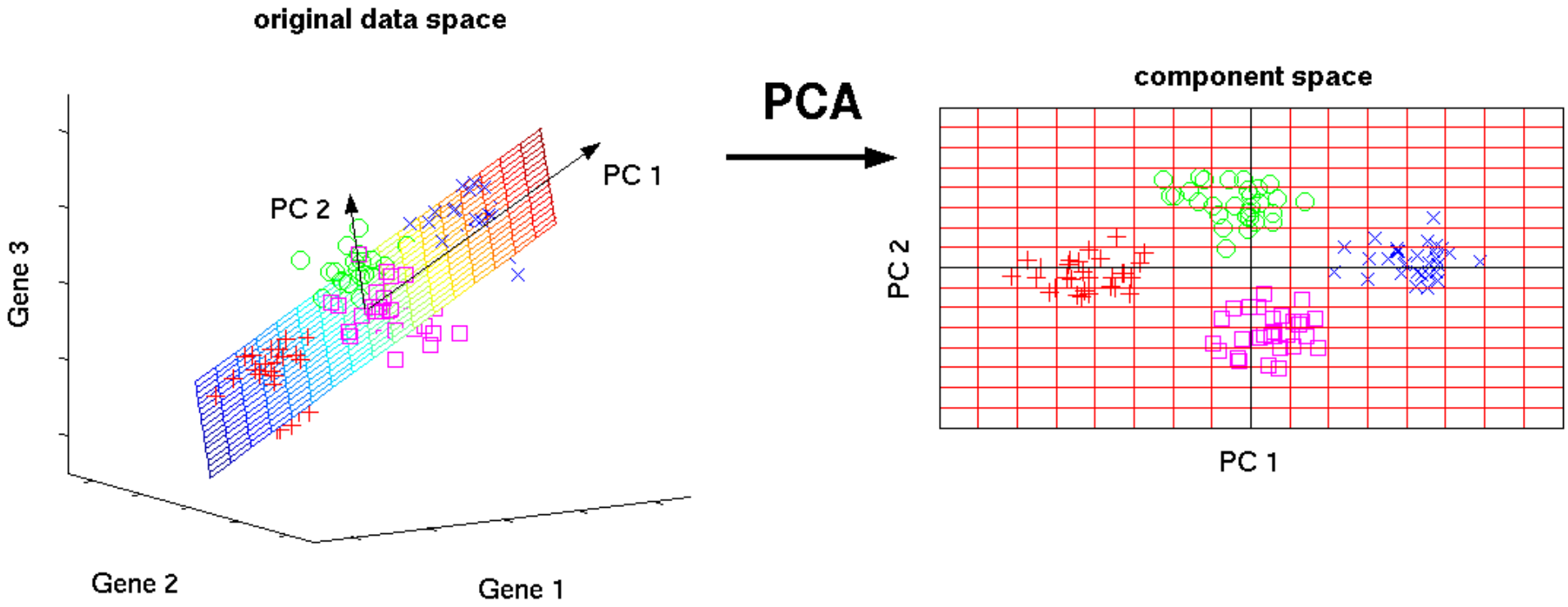
As we have shown, an equivalent way to define principal components is to first find the linear combination (loading) with norm 1 of the columns of X such that its variance is maximized.

Such a loading forms the first principal component vector.

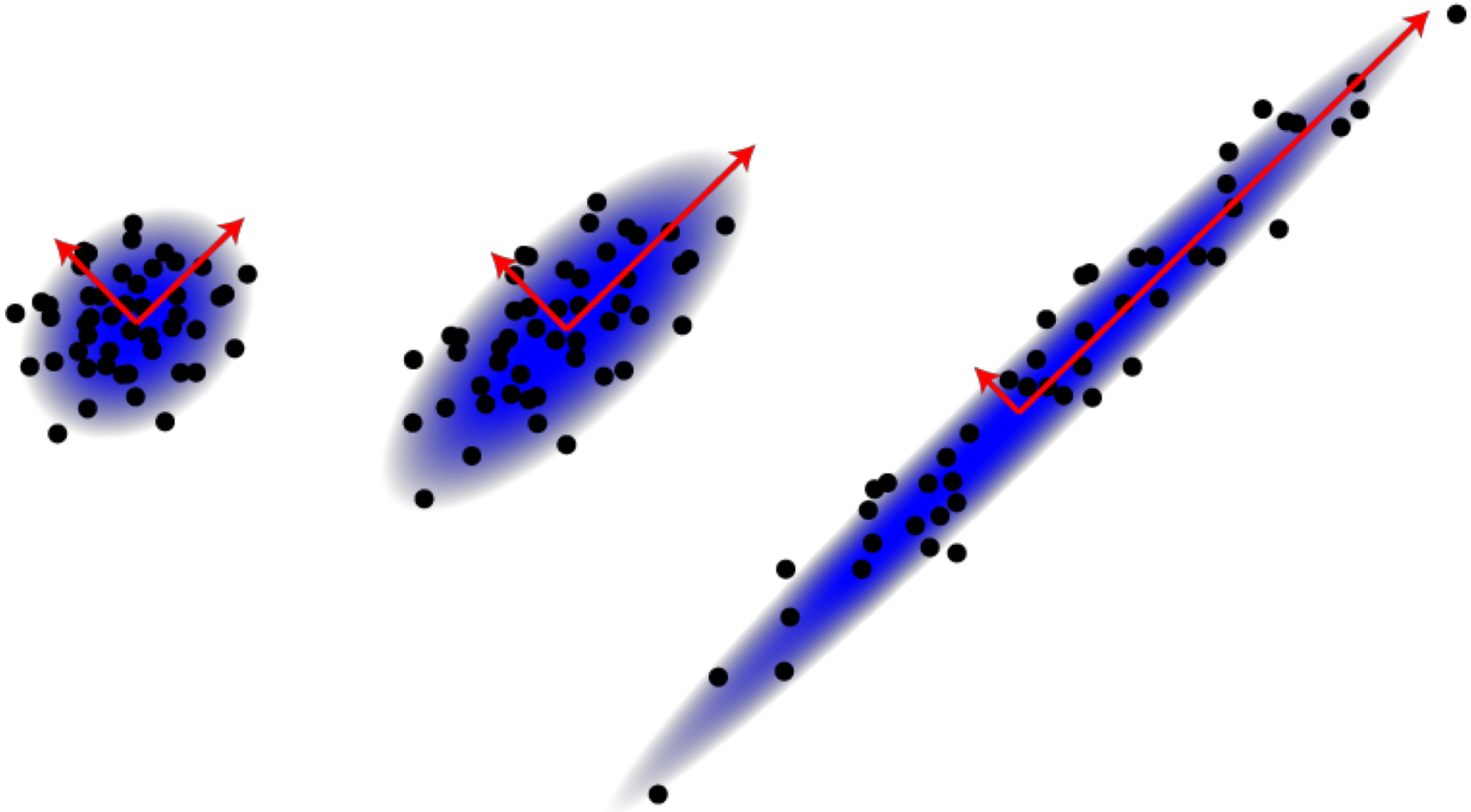
Similarly, we can define the second principal component vector that is the orthogonal to the first and has norm 1 and maximize the variance of the resulted linear combination of the columns of X.

…

PCA in 3-dim to 2-dim: 1-dim lost when is it not a bad idea?



Eigen values matter



PCA: keeping the variance in data

An important property of the eigen values is that they add up to the trace of the sample covariance matrix or the variance of the original column vectors of X (with columns centered).

$$\sum_{j=1}^{p} d_j = tr(G) = \sum_{j=1}^{p} Var(X_j)$$

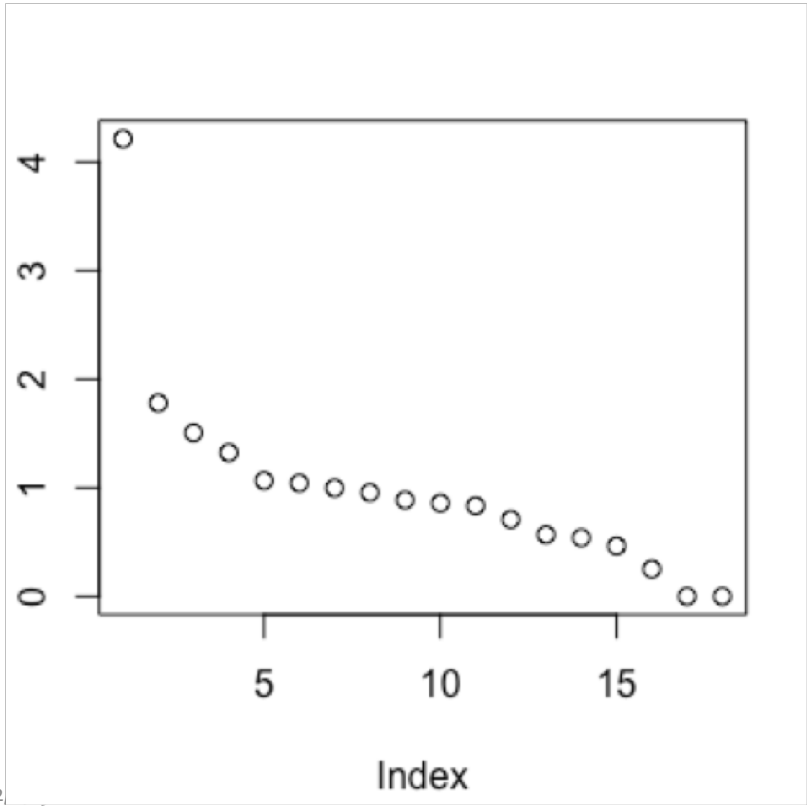
(Proof on blackboard)

Hence, in a sense PCA keeps all the variation in the original data.

Screeplot

Screeplot: plot the eigenvalues in decreasing order

Screeplot is commonly inspected visually to find a gap and decide how many principal components to keep



X : a 2421 by 18 matrix

2421 Ames house sales prices with 18 continuous features

1st PC: 33% of total variation
1st and 2nd: 42%
1st, 2nd, 3rd: 49% (corrected)
1st, 2nd, 3rd, 4th: 54% (corrected)
…

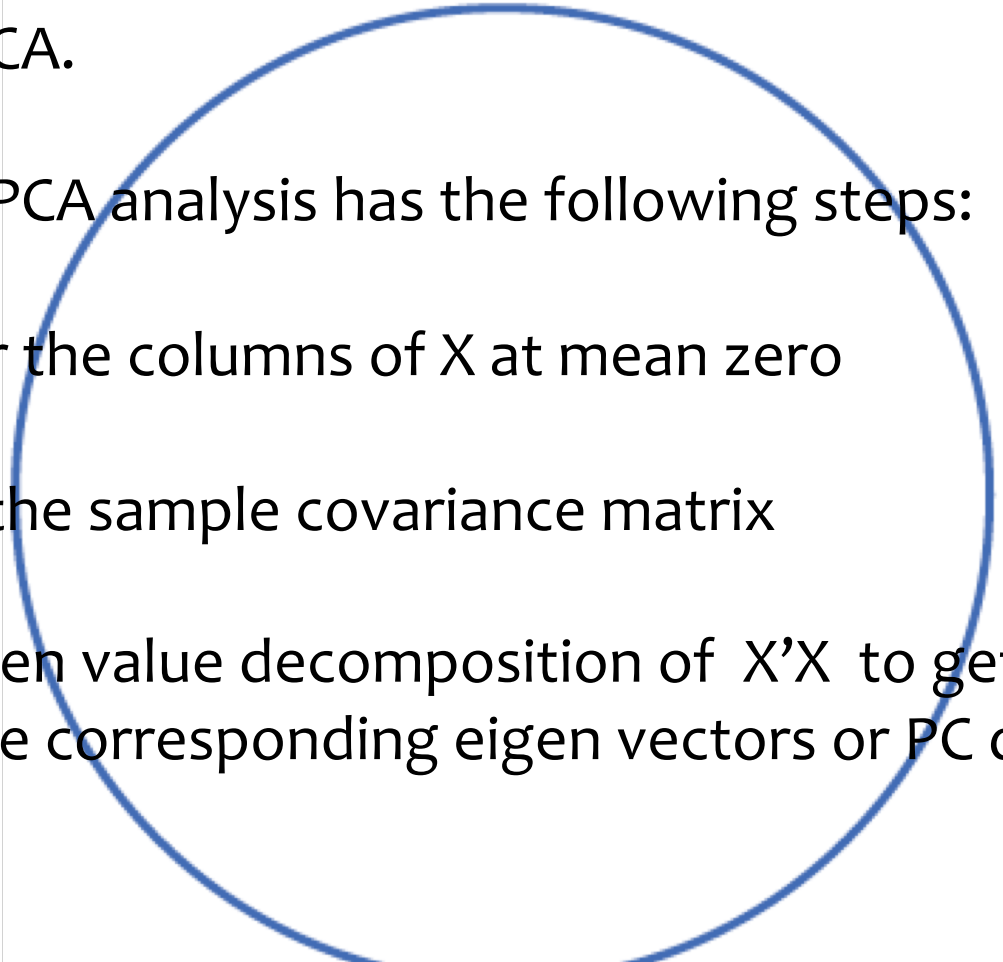
PCA: centering and normalization

(ask first why it is a good idea with your data)

In order to avoid one column of X having an undue influence on the principal components (PCs), it is common to first standardize the column vectors into mean zero and variance 1 before PCA.

Hence a PCA analysis has the following steps:

- Center the columns of X at mean zero
- Form the sample covariance matrix
- Do eigen value decomposition of X'X to get eigenvalues and the corresponding eigen vectors or PC directions

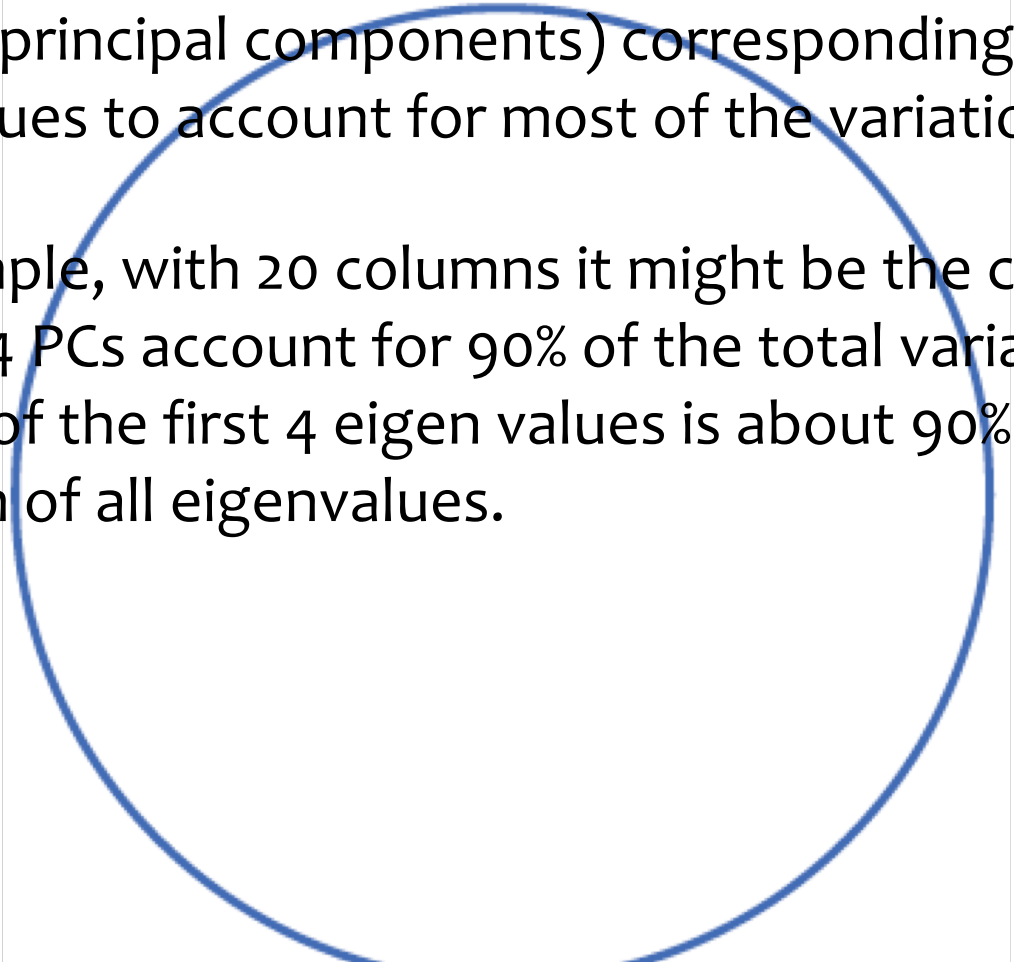


PCA: centering and normalization

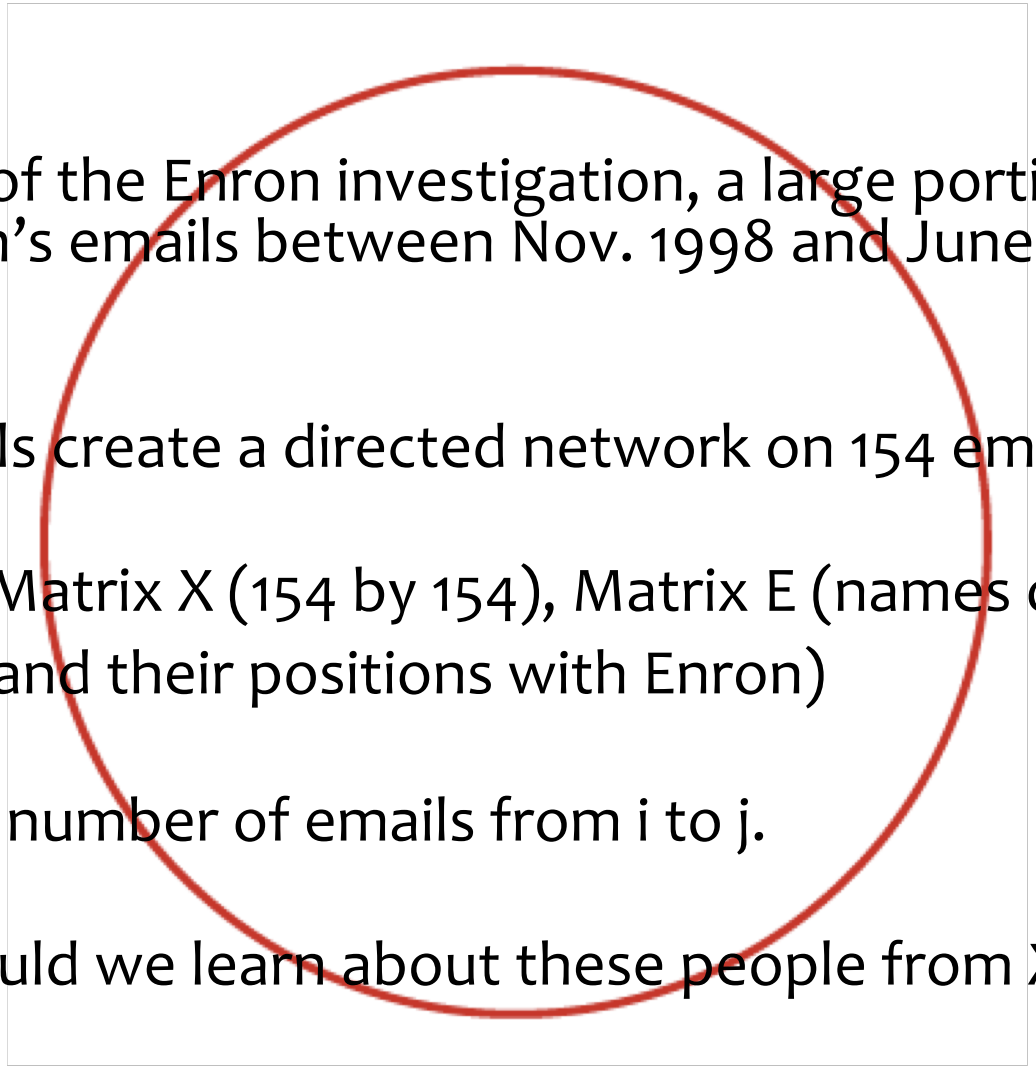
(ask first why it is a good idea with your data)

Keep the K PC directions (or the K leading columns of Z or K principal components) corresponding to K large eigen values to account for most of the variation in the data.

For example, with 20 columns it might be the case that the first 4 PCs account for 90% of the total variance or the sum of the first 4 eigen values is about 90% of the total sum of all eigenvalues.

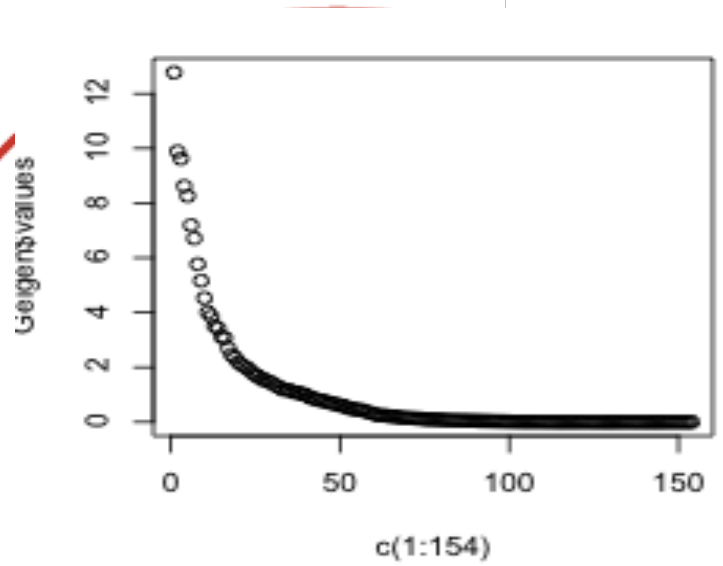


What can we learn via PCA about the Enron data?



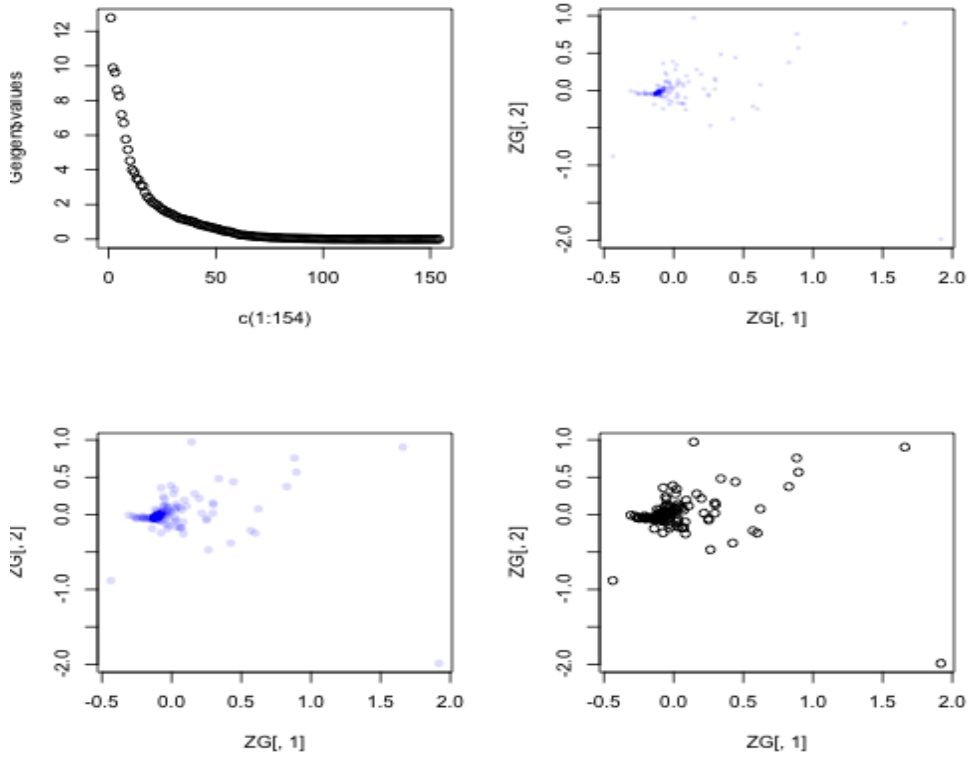
- As a result of the Enron investigation, a large portion of the corporation's emails between Nov. 1998 and June 2002 became public.
- These emails create a directed network on 154 employees.
- Our data: Matrix X (154 by 154), Matrix E (names of these employees and their positions with Enron)
- X_{ij} is the number of emails from i to j.
- Q: what could we learn about these people from X?

PCA on X with centering and with scaling: sum of first two eigenvalues is 14% of total



- Screeplot showing one large eigenvalue and many smaller ones with no big gaps.
- If we take the top 5, we get to 32%. Top 10 gives us 50%.
- Hence we need many dims here to represent the full variability in data. Top 2 is not representing data well at all.

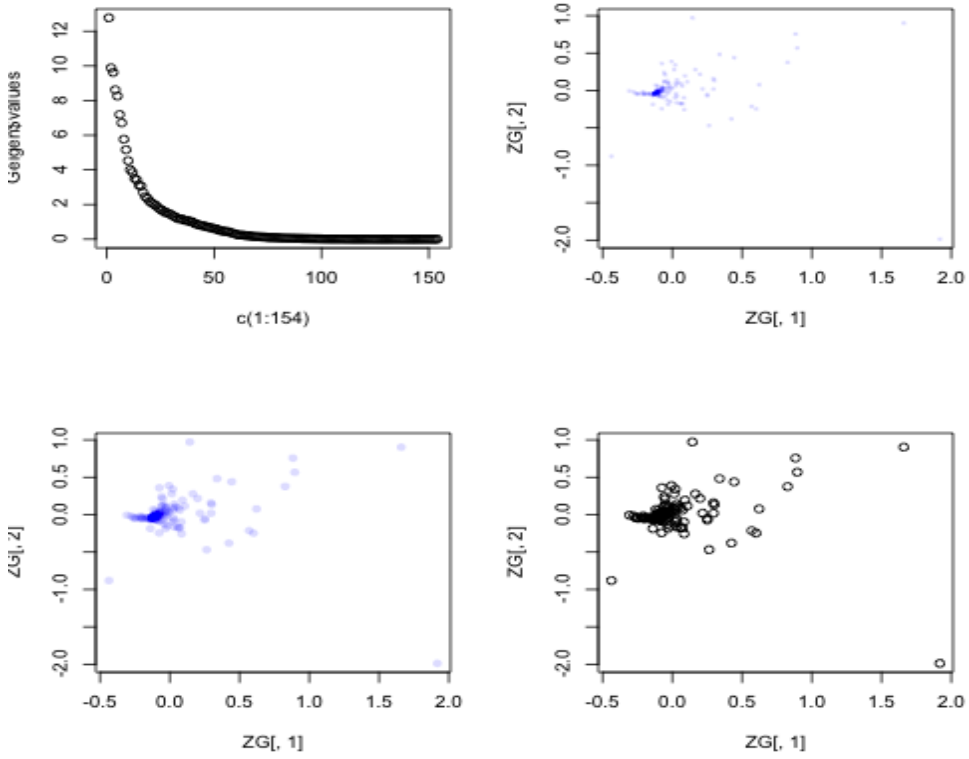
PCA on X with centering and with scaling
Sum of first two eigenvalues is 14% of total



What are the diff. among the three scatter plots?

- The outliers on the first comp. direction are employees 20 37 41 61 63 65 72 88. They do not share the same receivers of emails. Except for 20, the rest are all from trading or other depts.

PCA on X with centering and with scaling
Sum of first two eigenvalues is 14% of total

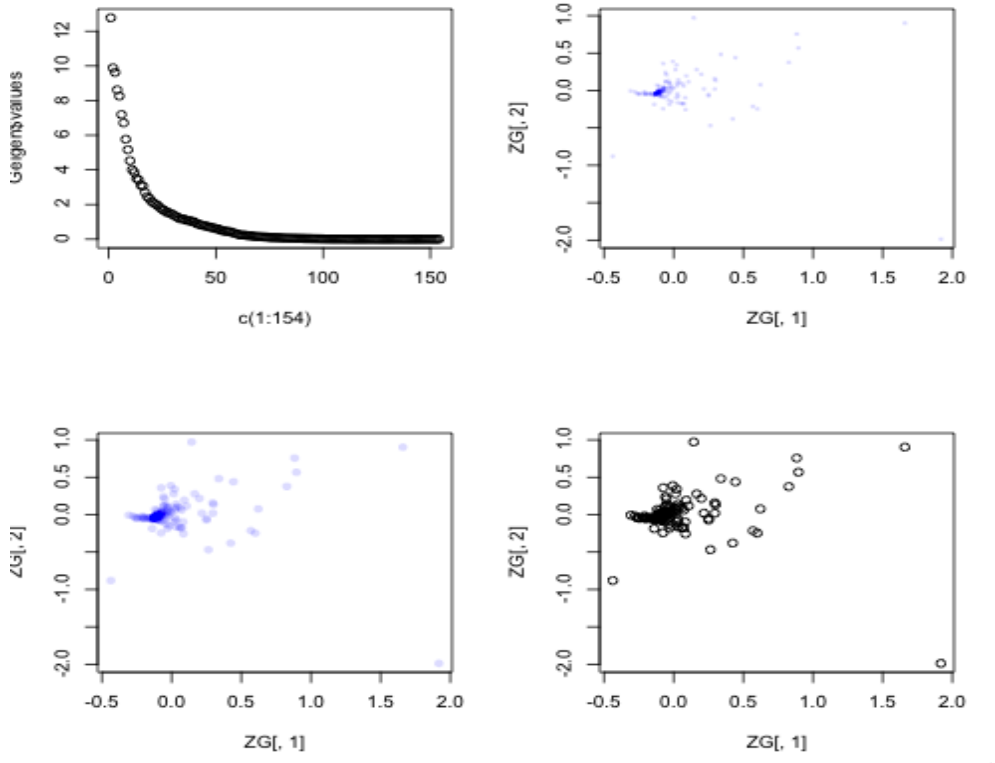


The three scatter plots plot the same data: first two columns of Z.

The two blue plots use transparency in plotting with different point sizes. They are good with giving good global views, but not with individual points such as outliers.

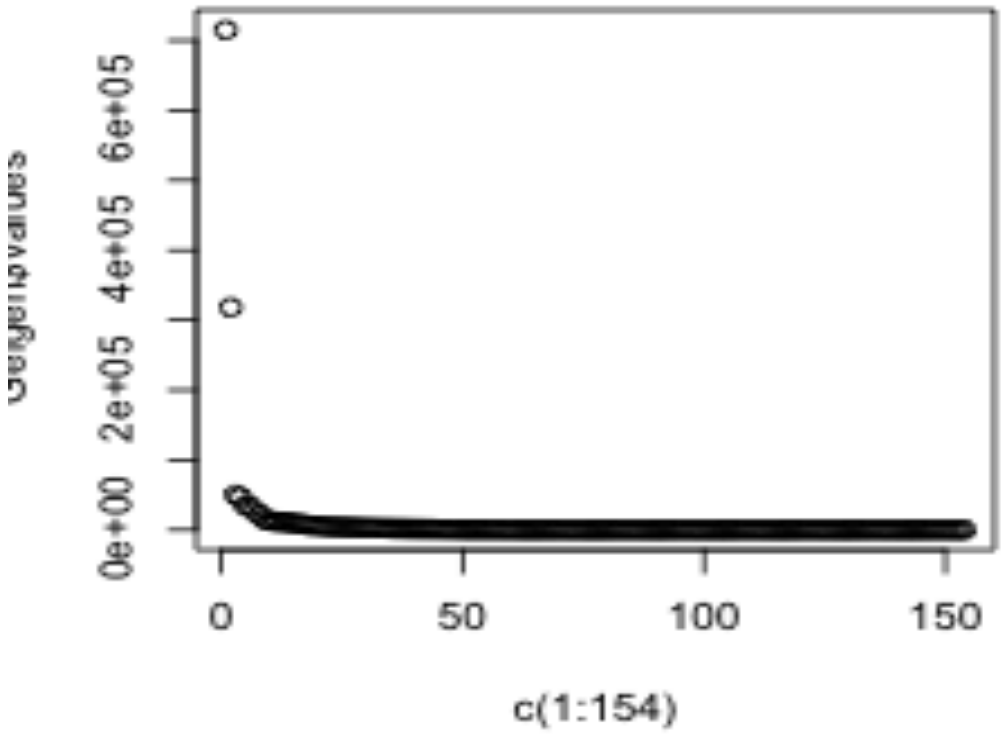
- The outliers on the second comp. direction are 24 37 57 63 65 96. They are all from trading and other depts except for **57** whom we will see later too.

PCA on X with centering and with scaling
Sum of first two eigenvalues is 14% of total



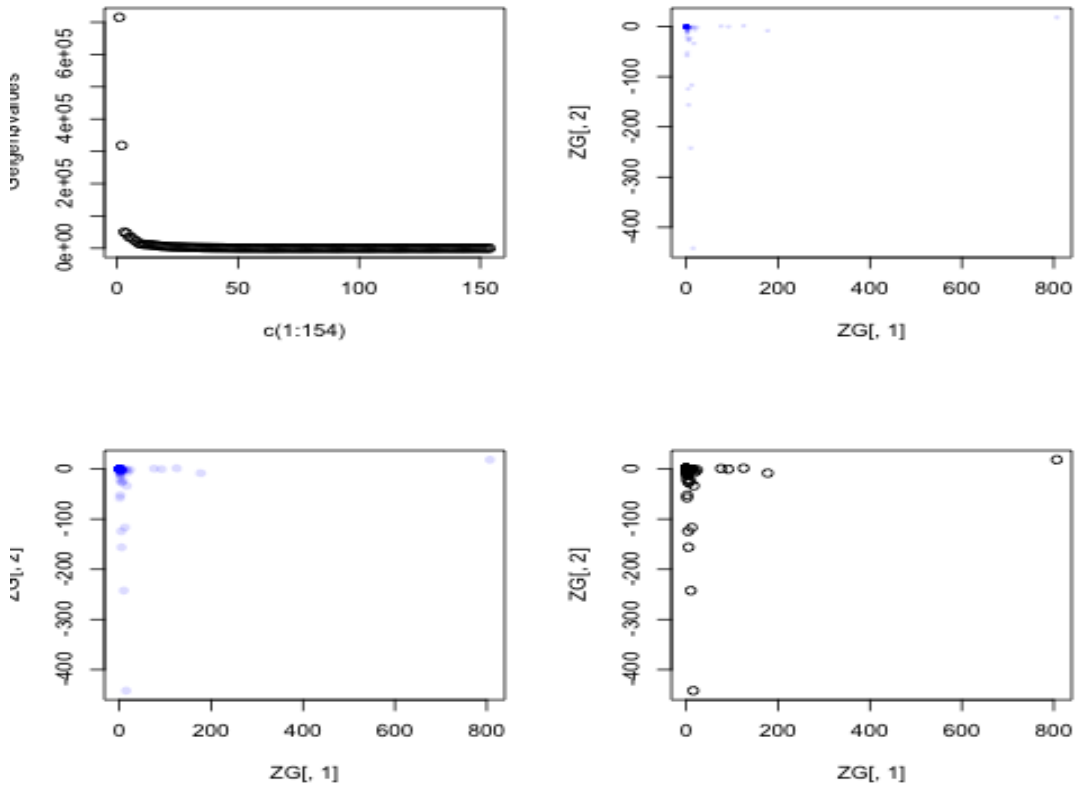
- It is also not clear what L2 scaling does is meaningful. L1 scaling seems to make more sense since we are dealing with count data.

PCA on X without centering and scaling: sum of first two eigenvalues is 71.6% of total



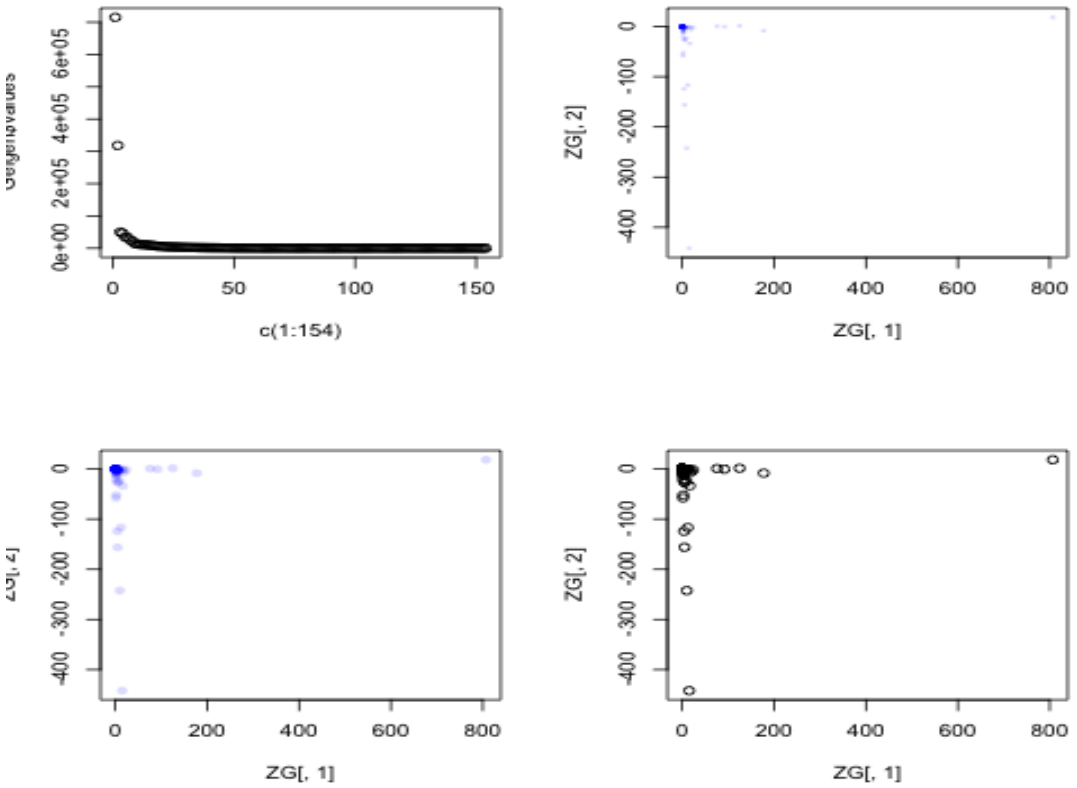
Clearly there are two large eigenvalues

PCA on X without centering and scaling: Sum of first two eigenvalues is 71.6% of total



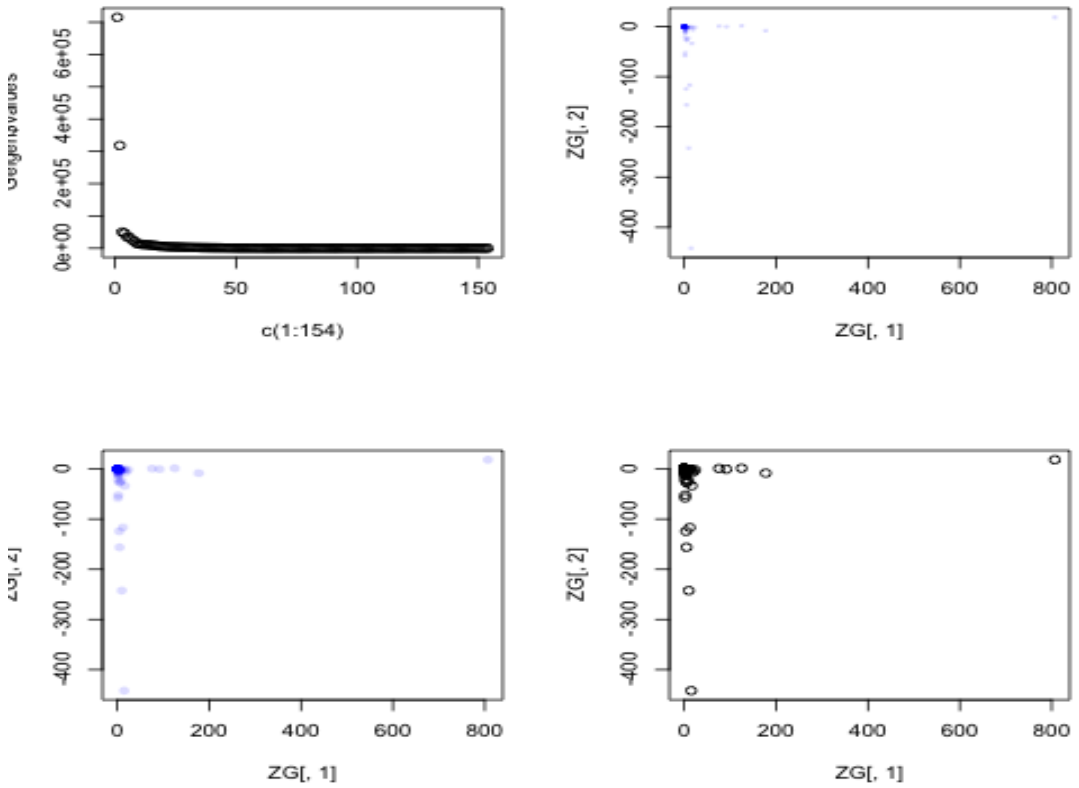
- Notice that the default scatter plot (bottom right) is the best for spotting outliers.
- The first comp. is dictated by the upper right corner outlier or employee 20.
- The second comp. is dictated by the lower left corner outlier or employee 57;

PCA on X without centering and scaling: Sum of first two eigenvalues is 71.6% of total



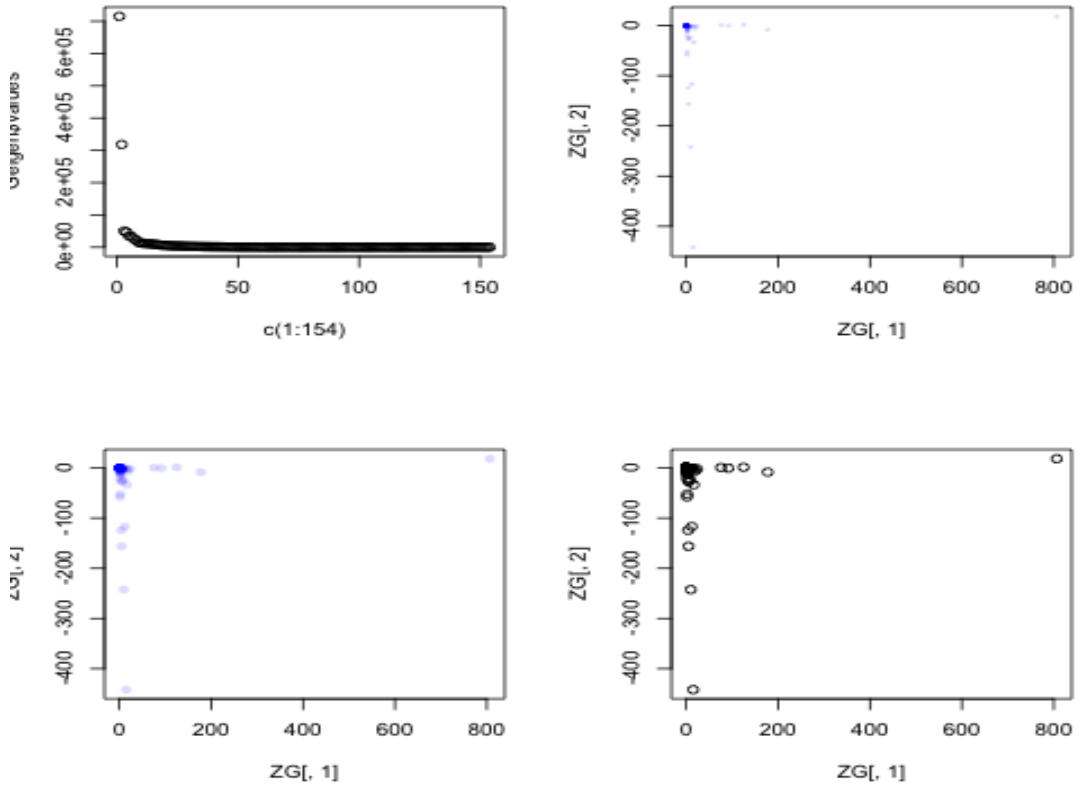
- If we look at the other two outliers on the first comp. direction, they are employee 44 and 129. All three are legal people. They have sent relative large number of emails to the same people. This explains why they show up along the same direction of the first comp.

PCA on X without centering and scaling: Sum of first two eigenvalues is 71.6% of total



- Employees 3 and 59 received many emails from all three and they are a VP and a manager at West Power.

PCA on X without centering and scaling: Sum of first two eigenvalues is 71.6% of total



- Second comp. dir outliers: 57 90 118 128 136. They sent many emails to the same People: 4,136, for example and these two are also legal people.

Employee 20 sent large numbers of emails to 4-5 people

- X[20,]

```
 [1]  0  0 78  0  0  0  0  0  0  0  2  0  0  0  0  0  9  1  0  8  0  0  0
  28  1  0  1  0  0  1  0  1  7  0  1  0  0  2  1  0  5  0  2 85 13  0  2  0  0
   0  1  0  4  0  1  1  2  4 342  0  0  0 23  1  8  3  1  0  0  1  1  0  0  0  0
   0  0  0  0  1  0  0  0  0  1  0  0  0  1 10  0  1  0  0  2  0  0  0  1  0
   0  1  0  0  0  2  0

[109]  0  1 124  0  0  0 19  3  1  1  1 449  0  1  1  1  0 11  1 551  0  1  0
   3  3  2  2  0 10  0  0 36  1  1  0 11  0  2  0  0  1  0  2  0 14
```

- E[20,]
  - 20 Jeff Dasovich Legal Regulatory and Government Affairs Director Male Senior
  - Employee 20 dictated first principal component due to its large L2 norm.

Employee 57 also sent large numbers of emails to 4-5 people, but not as large as 20

- X[57,]

```
  [1] 15  0  0 106  1  0  1  0  0  0  0  0  0  0  0  0  1  1
 [19]  0  0  2  0  0  0  0 126  2  0  0  0  0  0  1  0  0  0
 [37]  0  0  0  0  0  0  1  1  2  0 202  1  0  0 200  0  0  0
 [55]  0  5  1  0  1  0  0  0 28  0  1  0  1  0  0  0  0  1
 [73]  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 68
 [91]  0  0  2  0  0  2  0 15  0  0  0  0  0  0  0  1 75  0
[109]  0  1  1  0  0  0  1  0  0 146  0  2  0  0  0  0  0  0
[127]  0 112  1  0  0  0  0  5  0 240  0  0  0  1  0  0  0  0
[145]  0  0  0  0  0  1  0 16  3  0
```

- E[57,]
  - Tana Jones Legal ENA Legal Specialist Legal Female Junior

- Comparing the two vectors X[20,] and X[57,] shows that they sent frequently emails to different people.

PCA with Enron data

- The above PCA analyses are suggestive to give leads for further follow-up studies. They are not conclusive in any sense.

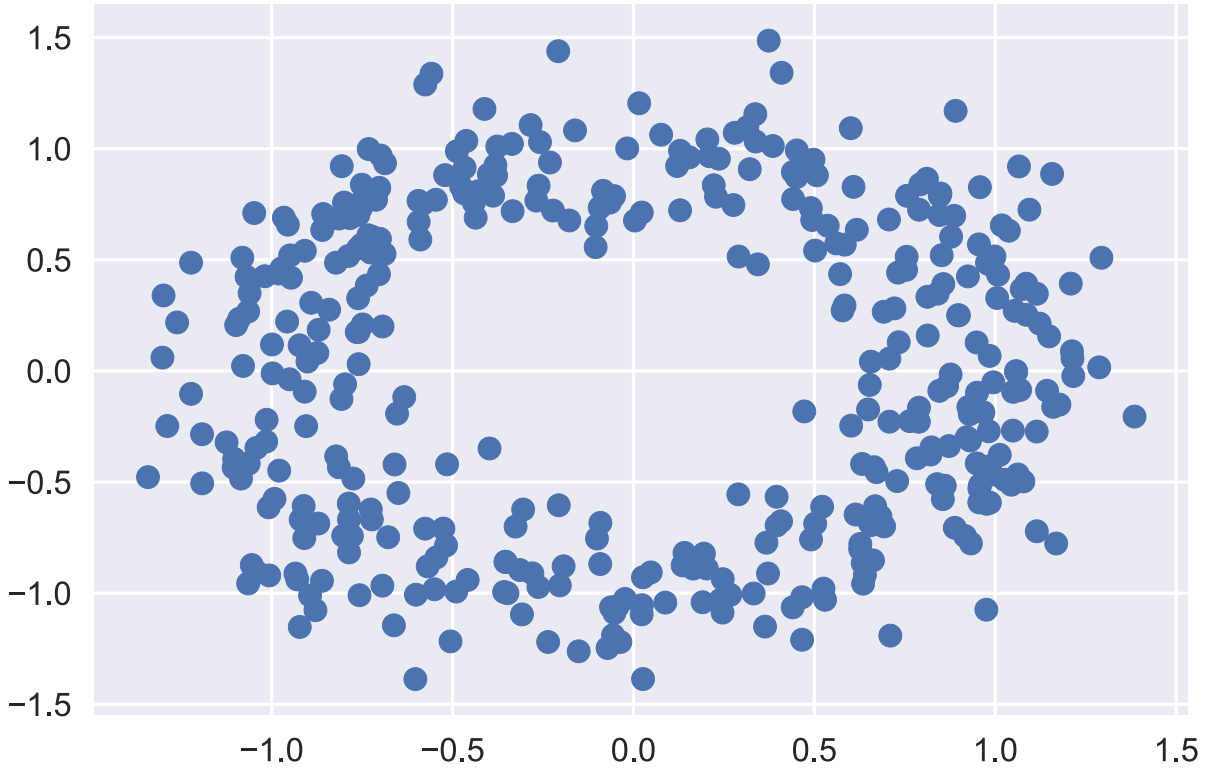
- Employee 20 is
  - Jeff Dasovich, Legal, Regulatory and Government Affairs, Director, Male Senior
  - His name appeared in some news report during Enron investigation.

- So, what do we learn: What is better? PCA with or without centering?
- Depends on the data at hand!

PCA: Assumptions and limitations

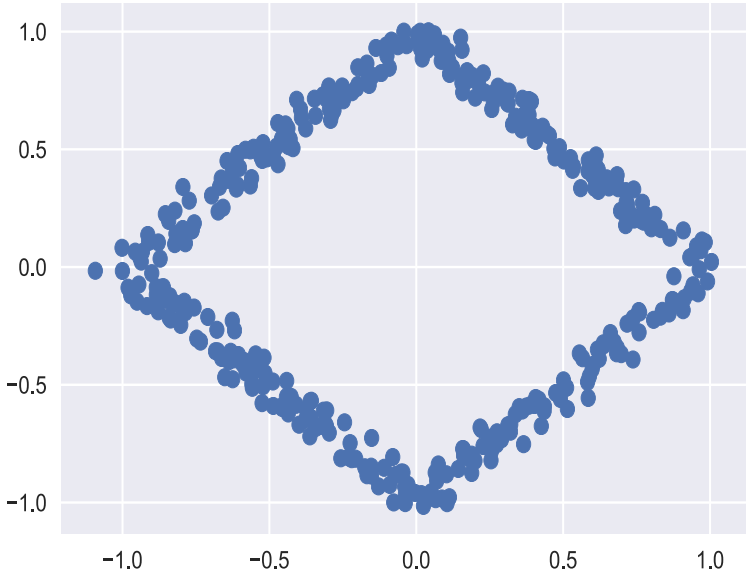
- Gaussianity of data for theoretical guarantees of optimality
  - For non-Gaussian data, often works! Not always.
- Orthogonality of PCs:
  - What if the data has non-orthogonal components
  - ICA
  - What if the data lies on manifold?
  - Kernel PCA
- The data has to be real and continuous valued data
  - Categorical data?
- What if the data is non-negative valued (astronomy)
  - PCs and Mean-removal may not be a good ideaI
  - NMF

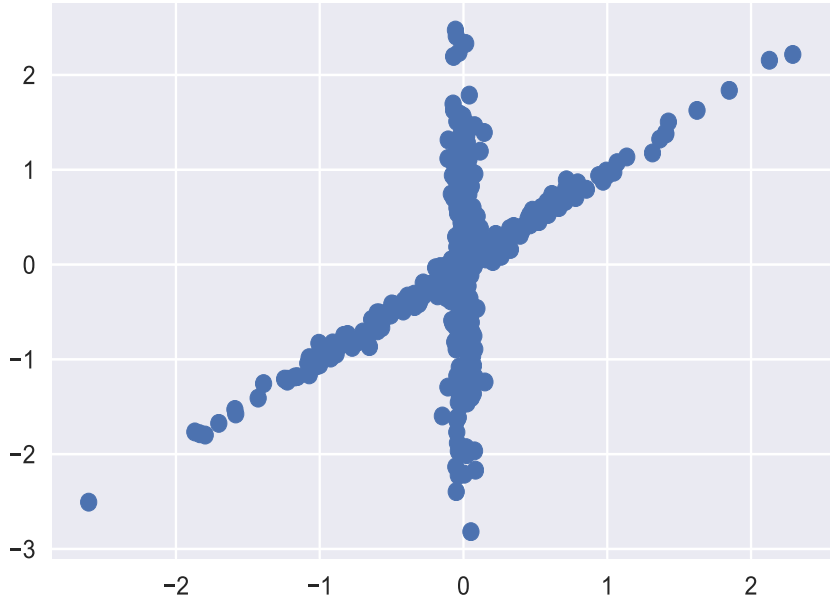
PCA: Limitations



What are the principal components here? Are they meaningful?

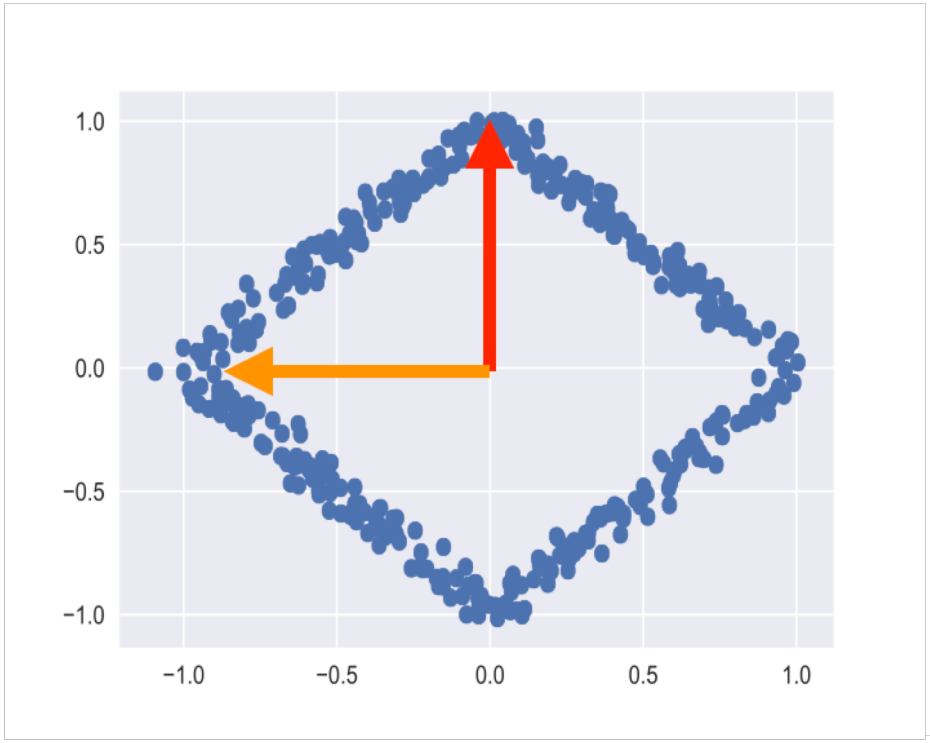
PCA: Limitations

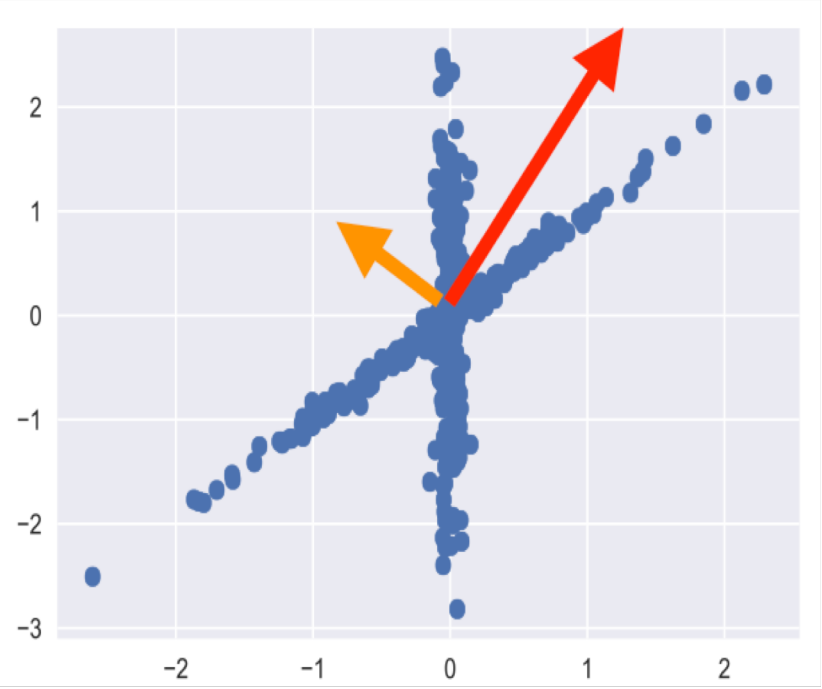




What are the principal components here? Are they meaningful?

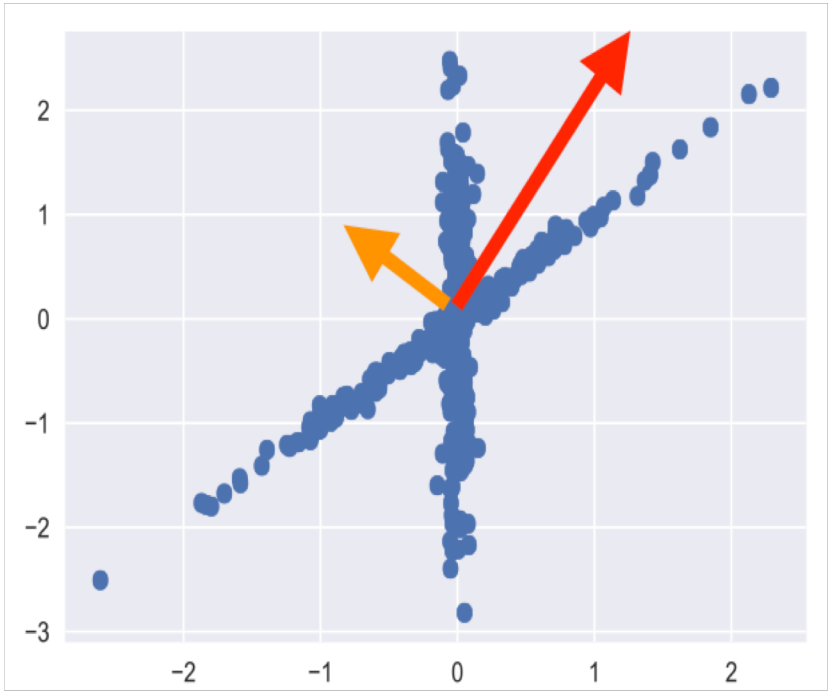
PCA: Limitations



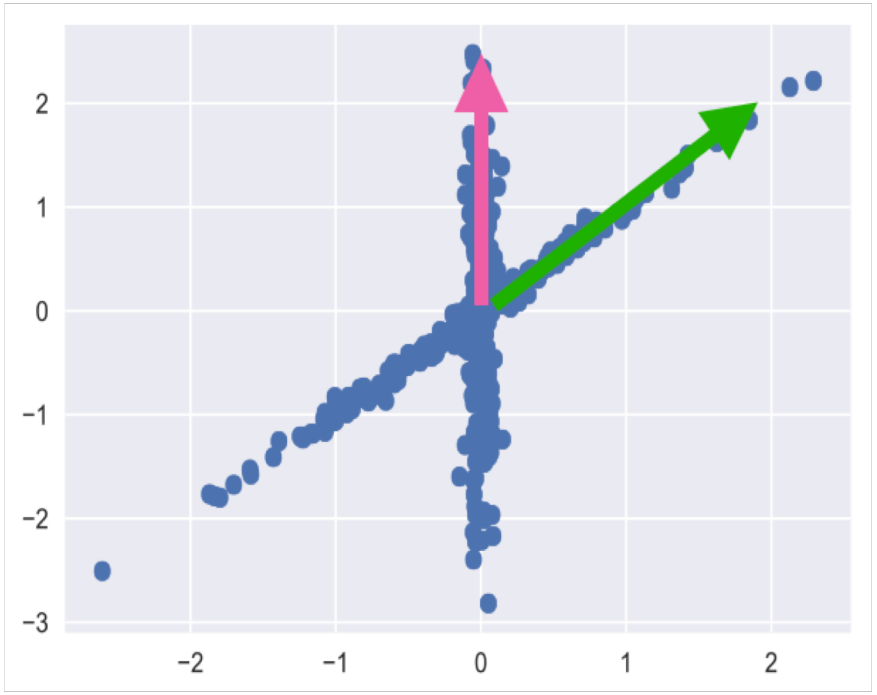


What are the principal components here? Are they meaningful?

Beyond PCA: Kernel PCA, ICA, NMF, …



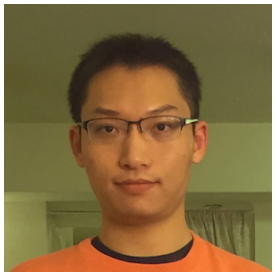
PCA



ICA

A research project where PCA doesn't work

Statistics, UCB



Siqi Wu

A. Joseph

B. Yu

Erwin Frise

A. Hammonds

S. Celniker

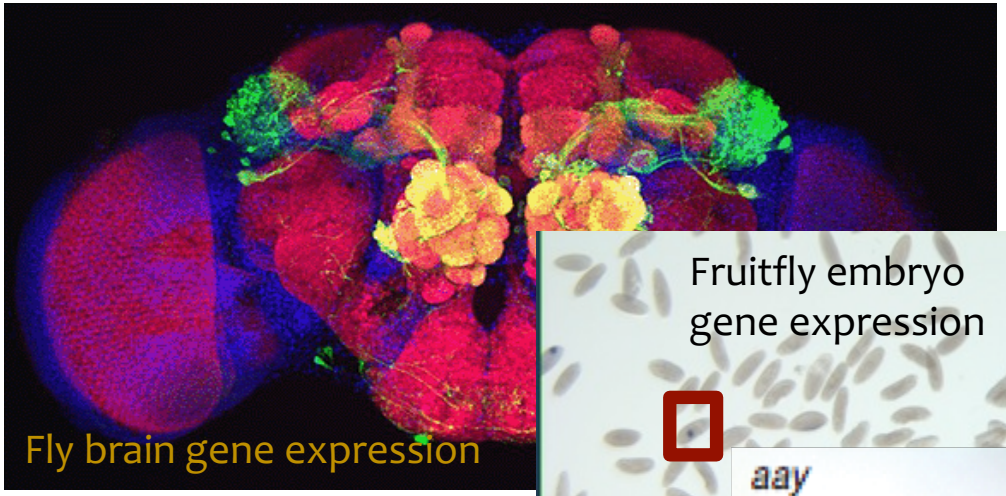
PNAS

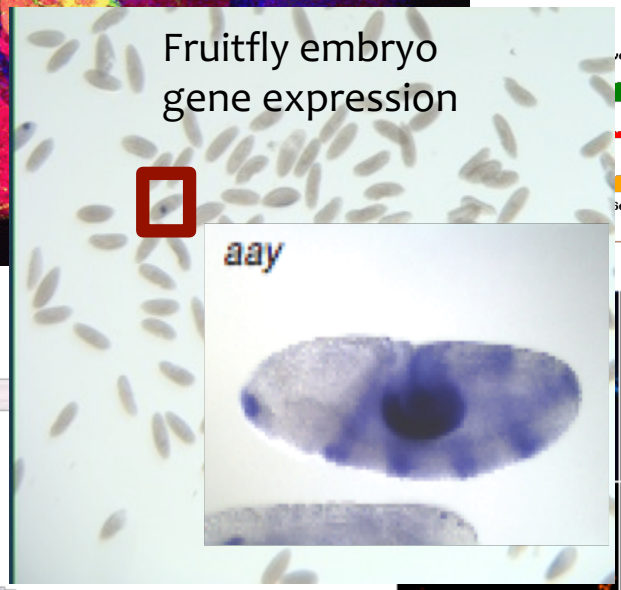
Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks

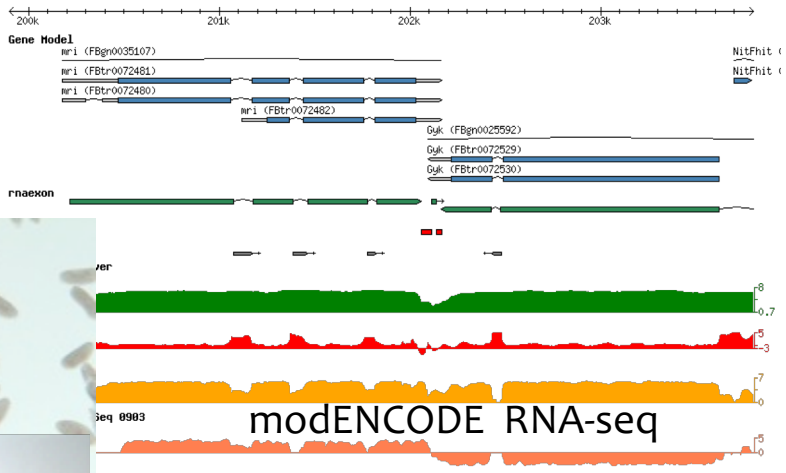
Siqi Wu[a,b], Antony Joseph[a,b,c], Ann S. Hammonds[b], Susan E. Celniker[b], Bin Yu[a,d,1], and Erwin Frise[b,1]

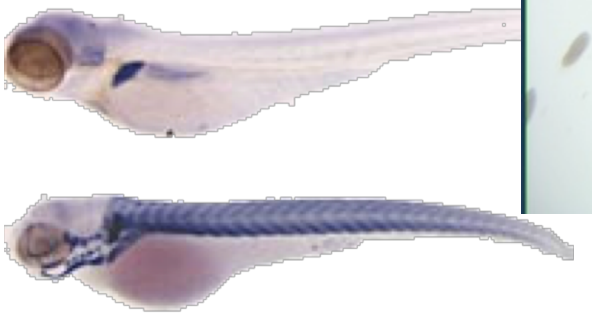
[a]Department of Statistics, University of California, Berkeley, CA 94720; [b]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [c]Walmart Labs, San Bruno, CA 94066; and [d]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

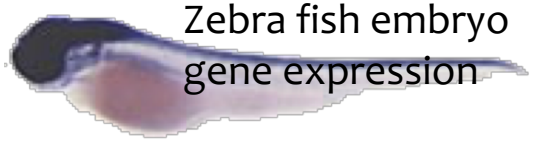
Recent image genomics data aims to answer:
where and **how** do **genes interact?**

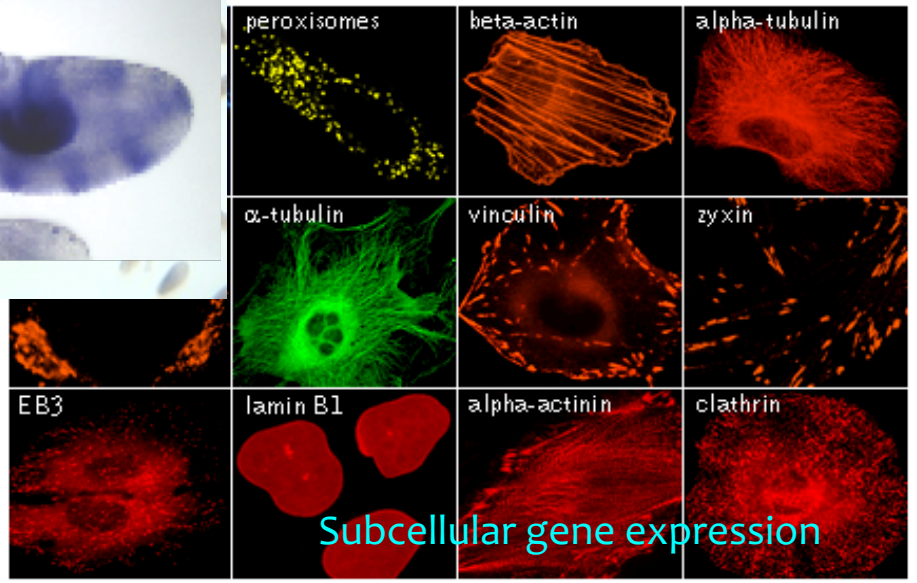












Wu, Joseph, Hammonds, Celniker, Yu*, Frise* (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *PNAS*.

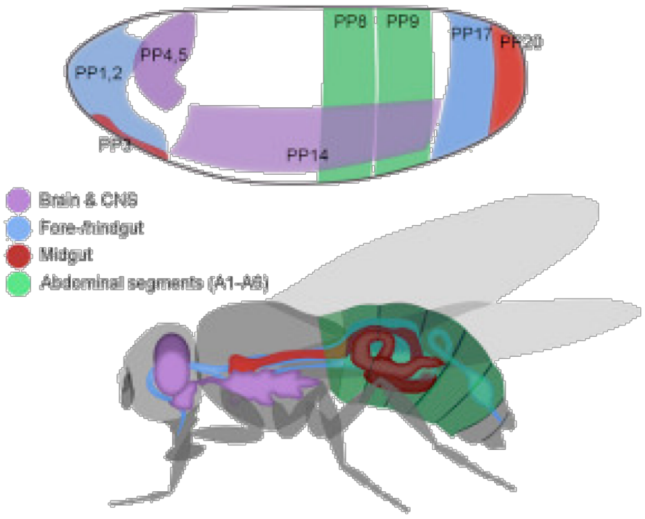
(e) Science News About

Updated by artificial intelligence 49 min ago Learn more

Astronomy Space | Biology Nature | Environment Climate | Health Medicine | Economics Math | Paleontol Archaeol

Mapping a cell's destiny: New tool speeds discovery of spatial patterns in gene networks

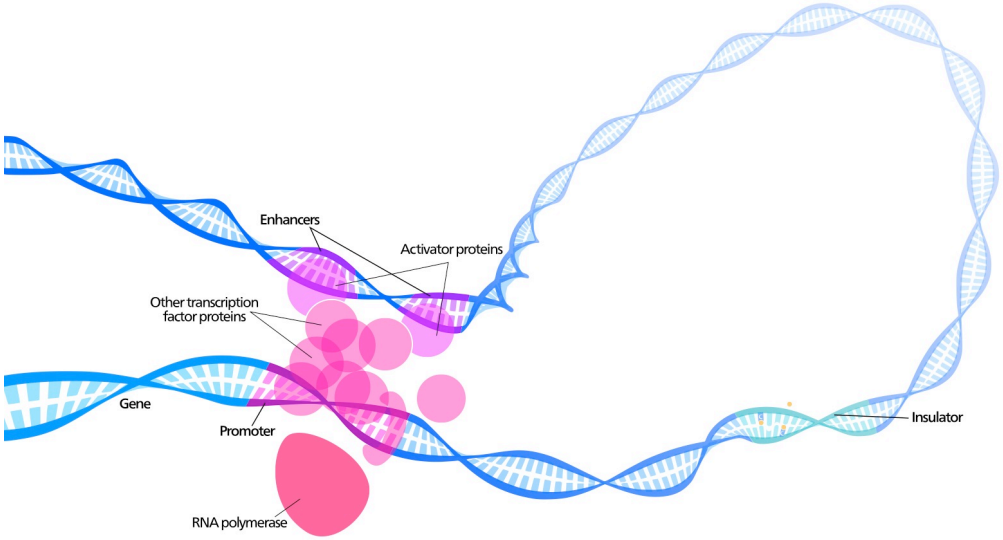
Thursday, May 5, 2016 - 06:30 in Biology & Nature

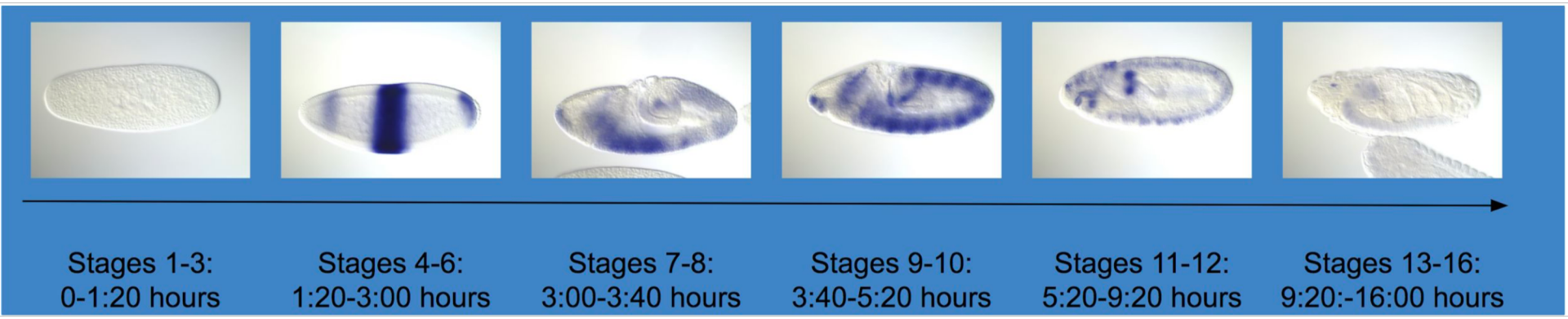


Transcription Factors (TFs)

Transcription Factors (TF): DNA binding molecules. "On-off" switch for triggering gene expression.

TFs drive a cascade of gene expression that partitions developing embryos into pre-organ regions

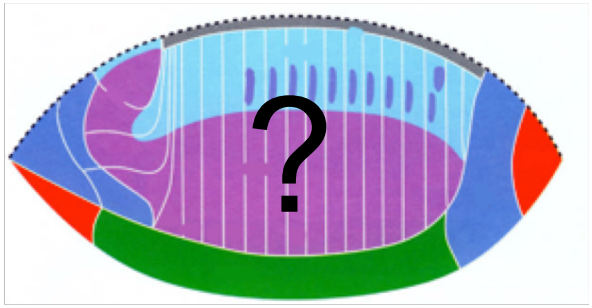




The Berkeley Drosophila Genome Project (BDGP) (cont'd)

7K+ genes examined – about 1 TB data

We seek answers to the following questions:



Drosophila (fruitfly) embryo

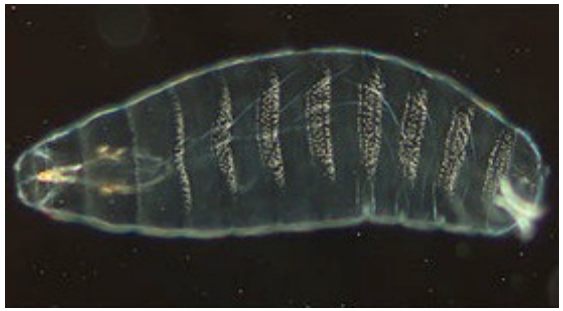
How many functional regions are there?

Gene A Gene B

What are the new gene functions and gene-gene interactions in these regions?

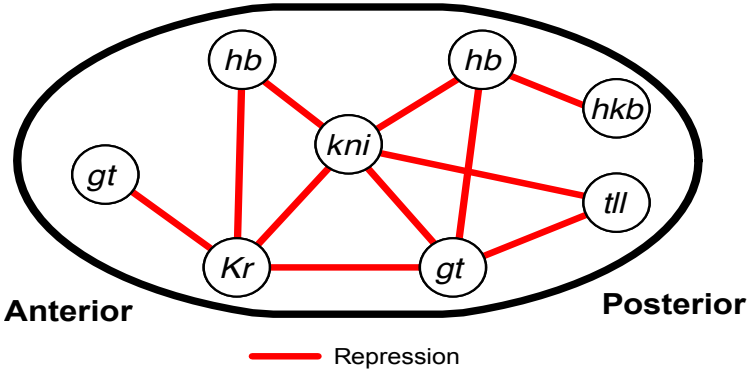
The gap gene network: genes interact locally in space

Segmentation (vertical "coordinate"): work of the gap gene network



Fruitfly embryo: segmentation

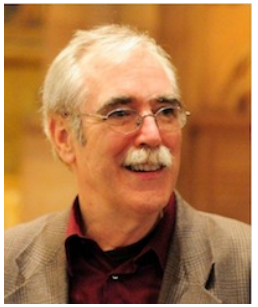
Human embryo: segmentation



1995 Nobel Prize in Physiology or Medicine for work on gap gene genetic control of early embryonic development

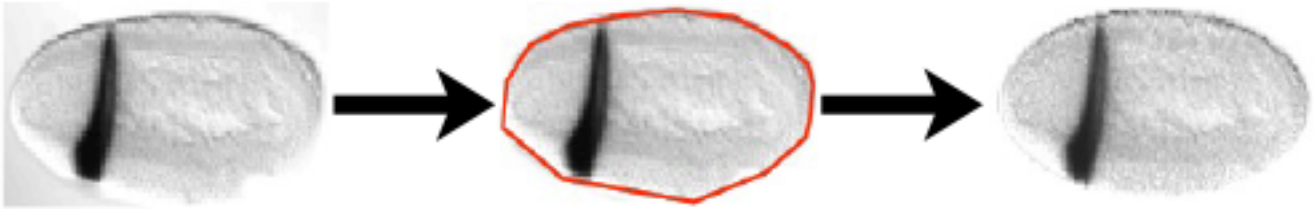
Christiane Nusslein-Volhard

Edward B. Lewis



Eric Wiechaus

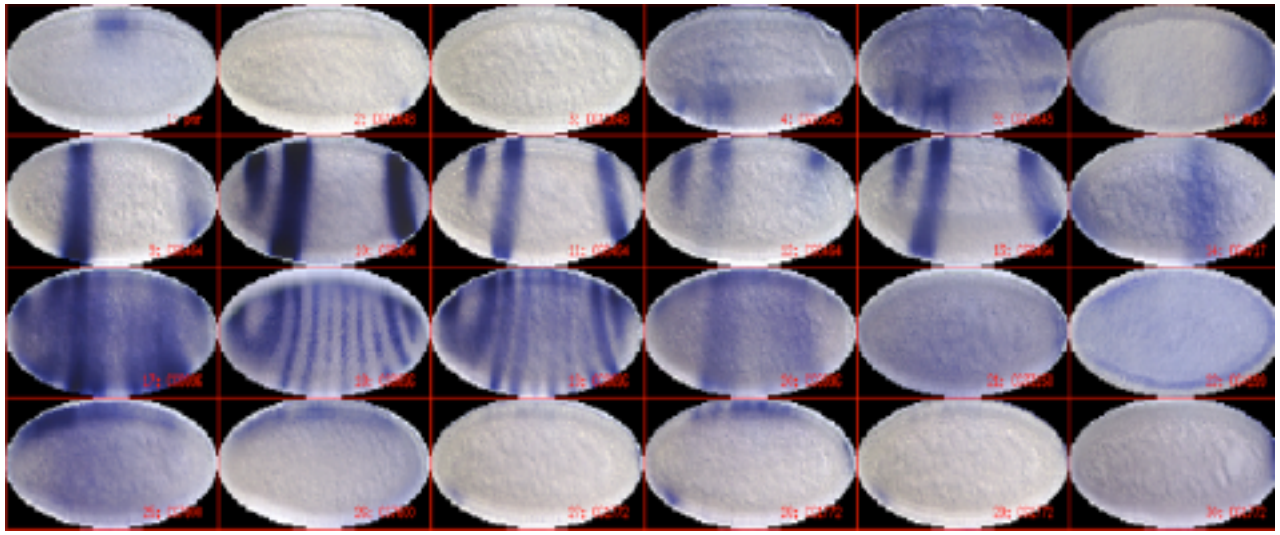
Data preprocessing of stage 4-6 images



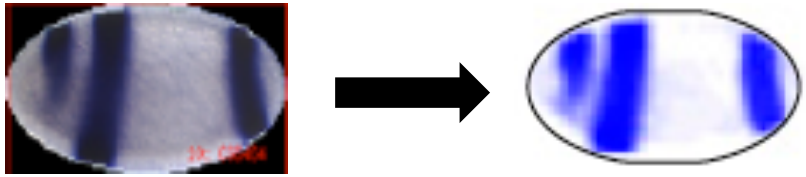
Outline detection

Register onto an ellipse

Registered images



Stain extraction



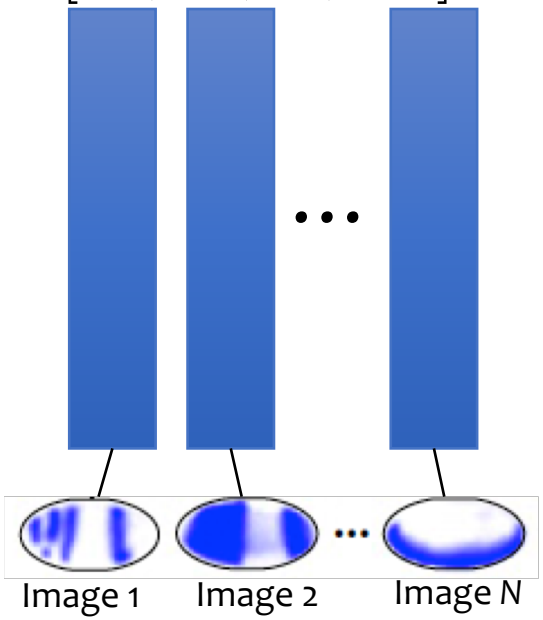
PCA

$$\min_{\mathbf{D},\mathbf{A}:\mathbf{D}^T\mathbf{D}=\mathbf{I_K}} ||\mathbf{X}-\mathbf{DA}||_F^2$$

where $||B||_F^2 = \sum_i \sum_j b_{ij}^2$ for a matrix $B = (b_{ij})$

Data matrix

$\mathbf{X} = [x_1, x_2, ..., x_N]$



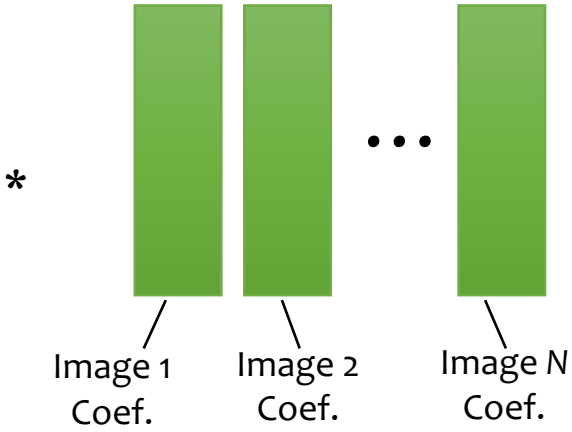
$\approx$

Dictionary

$\mathbf{D} = [d_1, d_2, ..., d_K]$

Coefficient matrix

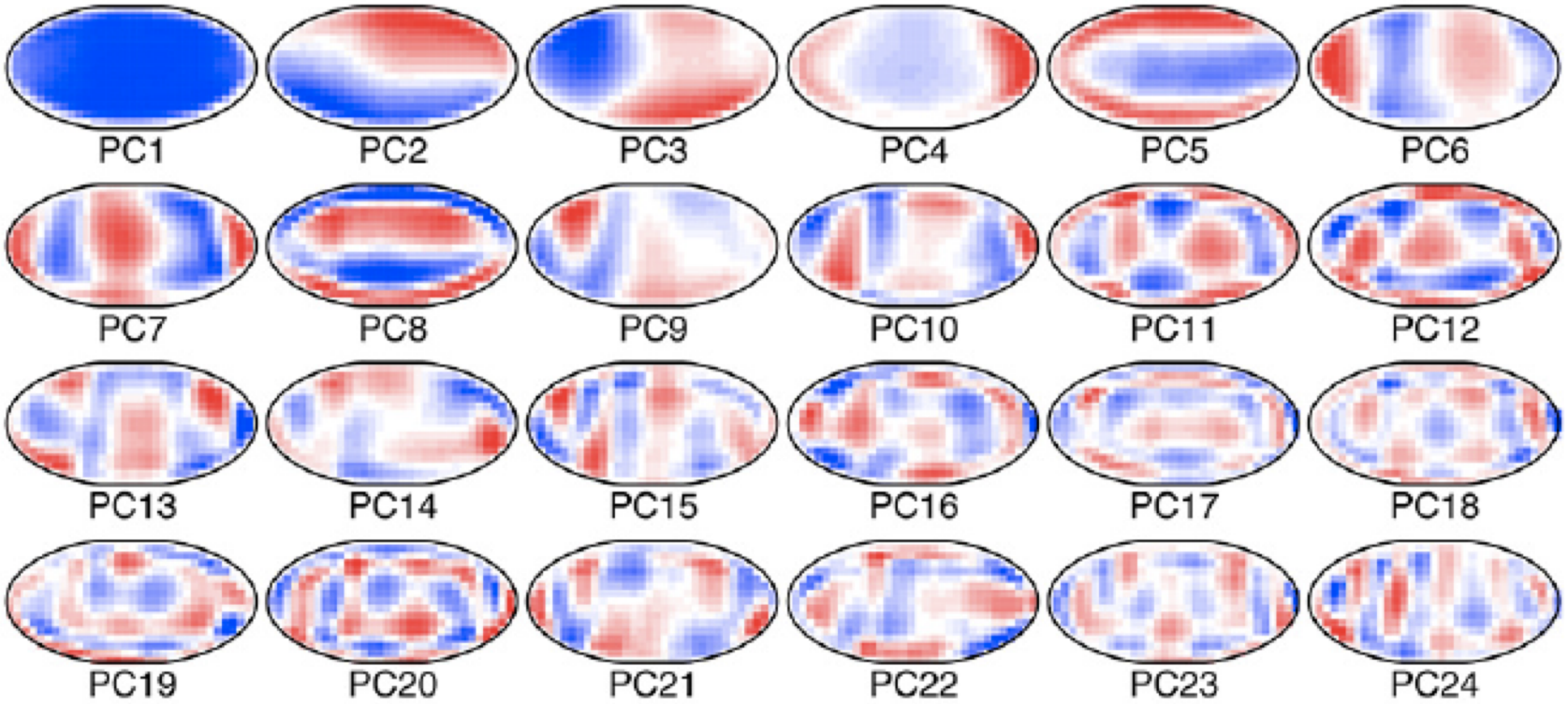
$\mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_N]$



PC = Principal Component

Principal components found:
not biologically meaningful as pre-organ regions

A



Positivity is a natural biological constraint.

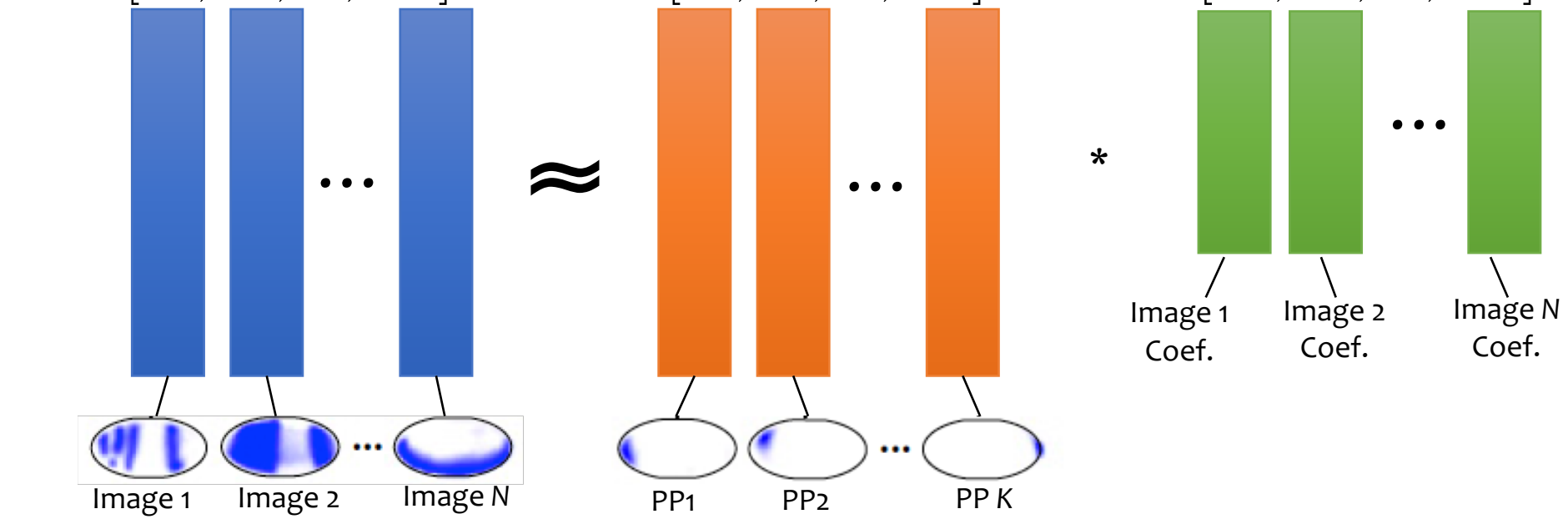
Interpretable data representation via stability driven nonnegative matrix factorization (staNMF)

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \|\mathbf{X}-\mathbf{DA}\|_F^2$$

Data matrix $\mathbf{X} = [x_1, x_2, ..., x_N]$

Dictionary $\mathbf{D} = [d_1, d_2, ..., d_K]$

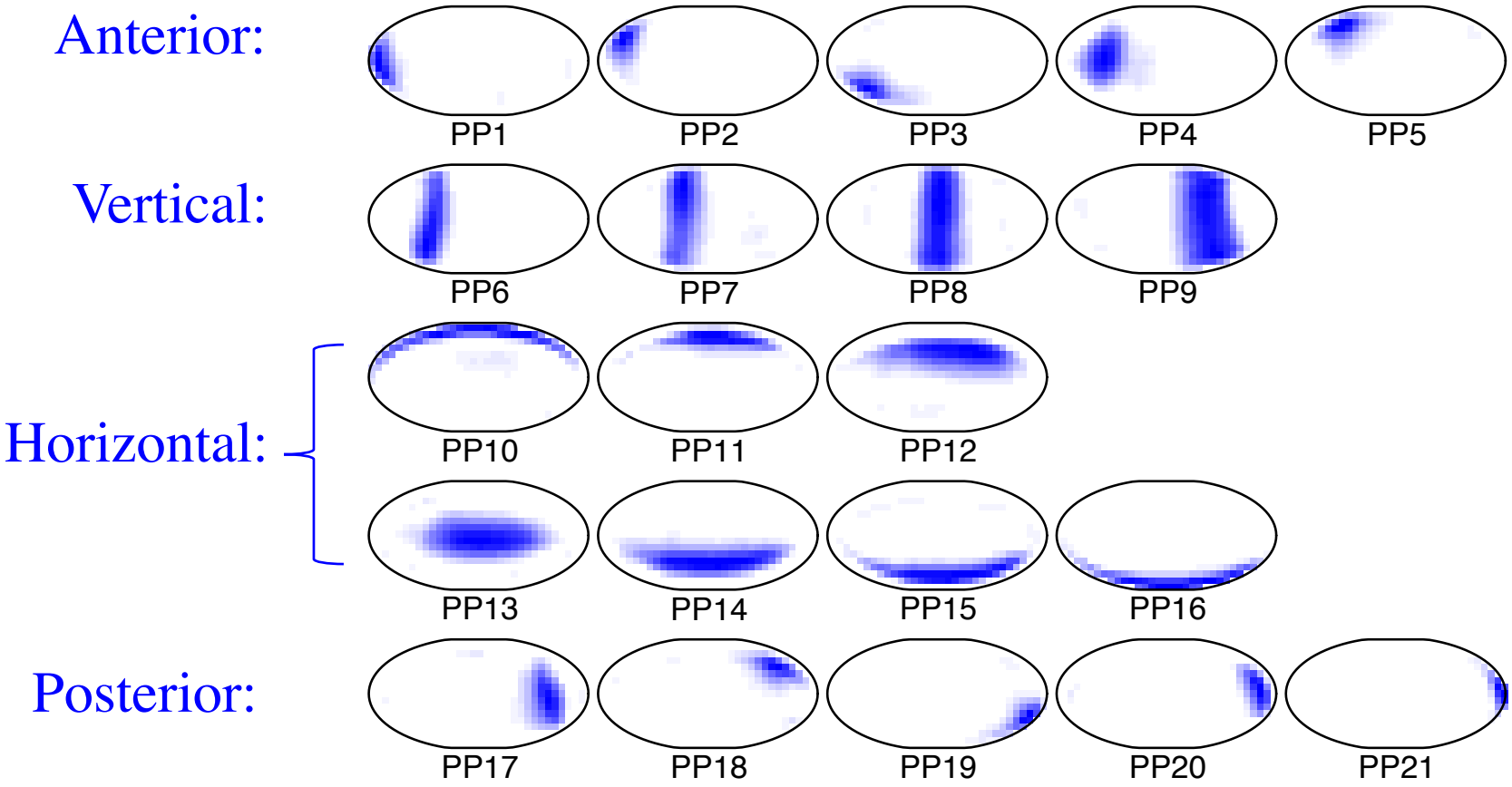
Coefficient matrix $\mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_N]$



PP = Principle Pattern (or atom in general)

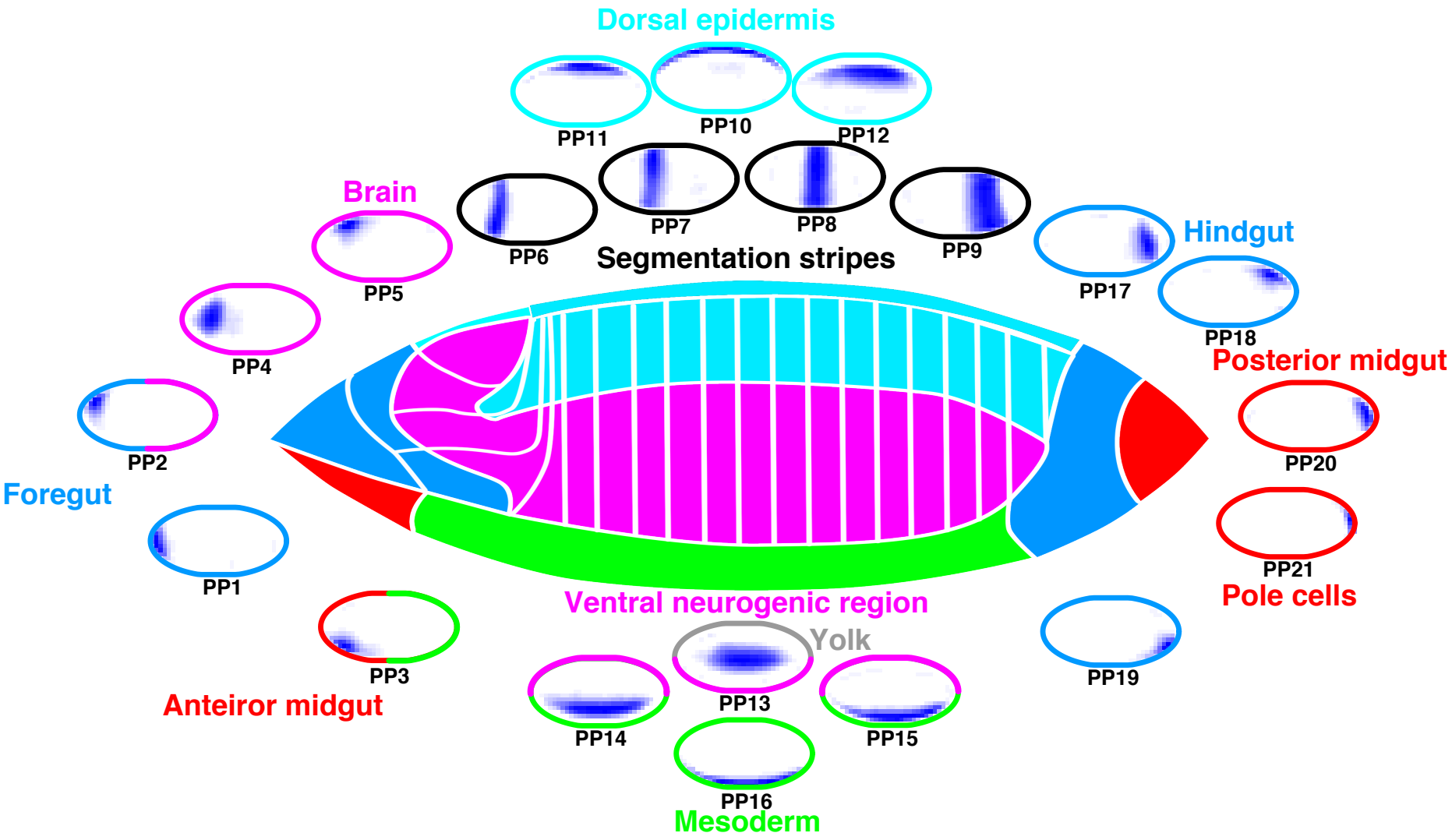
References: … Lawton and Sylvestre (1971), Lee and Seong (1999), (SPAMS) by Marial et.al (2010).

21 Principal Patterns found via staNMF: biologically meaningful as pre-organ regions



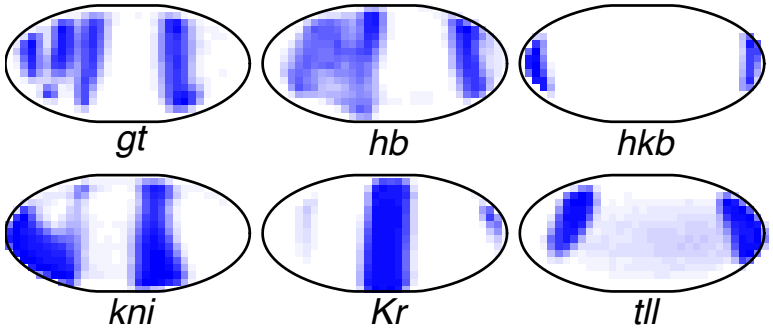
The four types of patterns work together to define a rough "coordinate" system in early Drosophila embryonic development.

Principal patterns (PP) as pre-organ regions



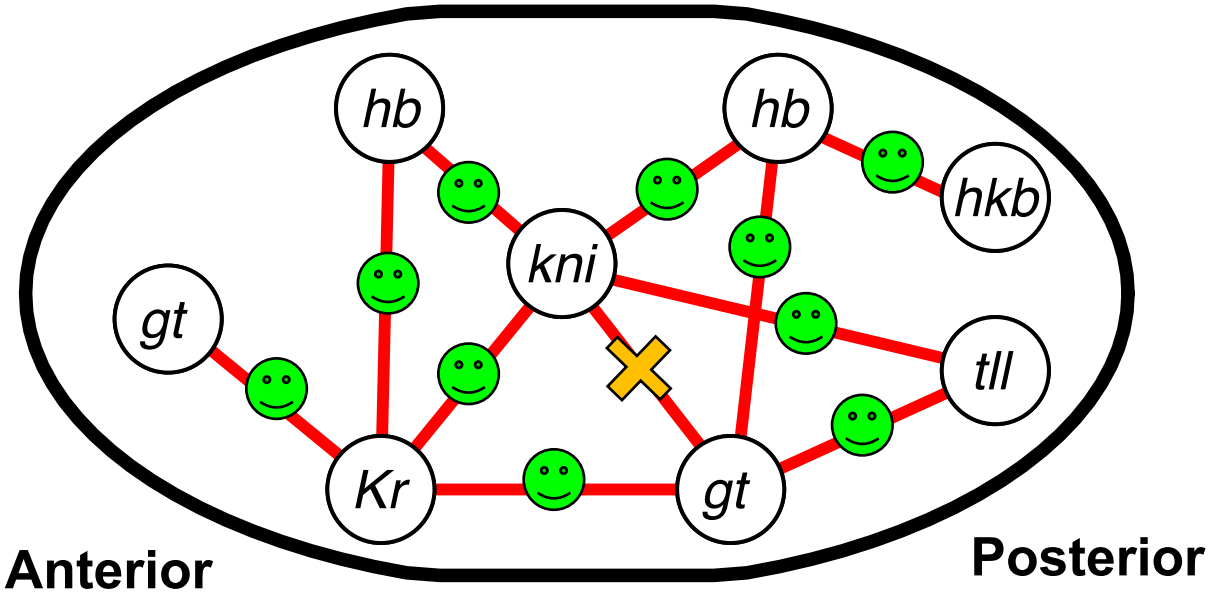
Local correlation networks reproduce 10 out of 11 gap gene links

Gap Genes



: Interactions correctly predicted

: Interactions that we missed



Taxonomy

of Machine Learning/Statistics

Labeled Data

Indirect (reward)

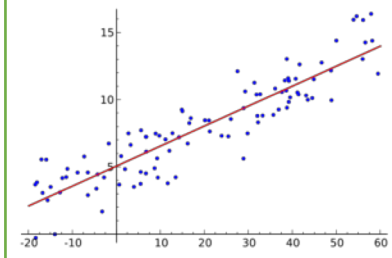
Unlabeled Data

Supervised Learning

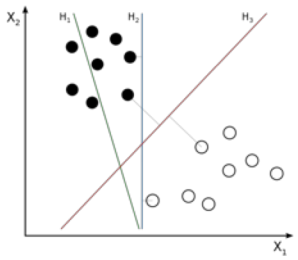
Reinforcement & Bandit Learning

Unsupervised Learning

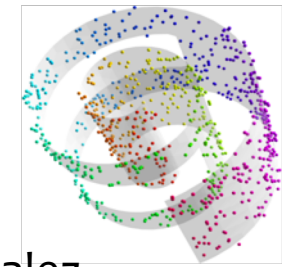
Regression



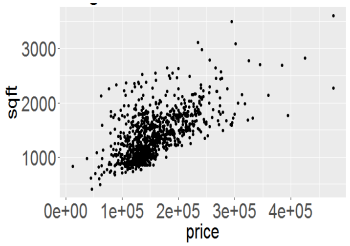
Classification



Dimensionality Reduction



Clustering



Thanks to J. Gonzalez

Organizing data through **clustering**

How do we know our clustering is meaningful and for what?

cluster | ˈklə-stər |

noun

a group of similar things or people positioned or occurring closely together: ***clusters of*** *creamy-white flowers* | ***a cluster of*** *antique shops*.

- Astronomy a group of stars or galaxies forming a relatively close association.
- Linguistics (also **consonant cluster**)a group of consonants pronounced in immediate succession, as *str* in *strong* .
- a natural subgroup of a population, used for statistical sampling or analysis.
- Chemistry a group of atoms of the same element, typically a metal, bonded closely together in a molecule.

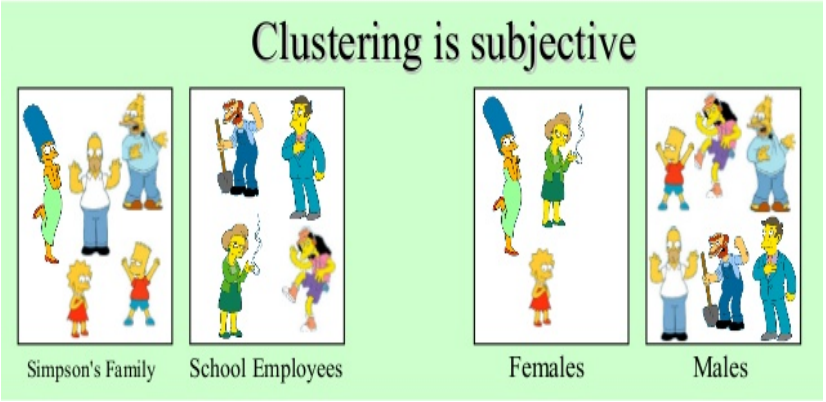
-- Oxford dictionary

Clustering, why bother?

Humans clustered similar things (objects and animals and people) way before statistics and machine learning…

We even gave terms to the clusters: red, blue; big, small; good, bad…
Language is clustering of "reality" into words…

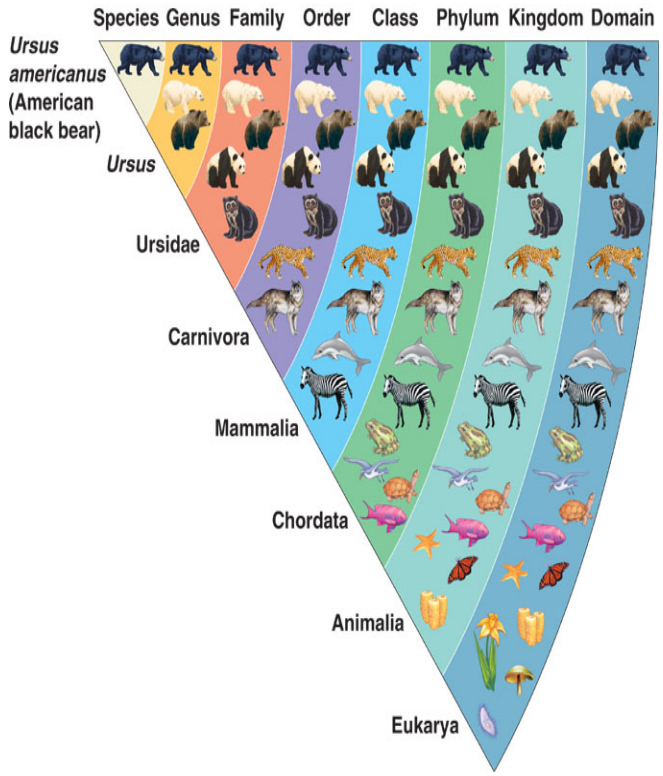
Clustering is old, vague, and subjective
…

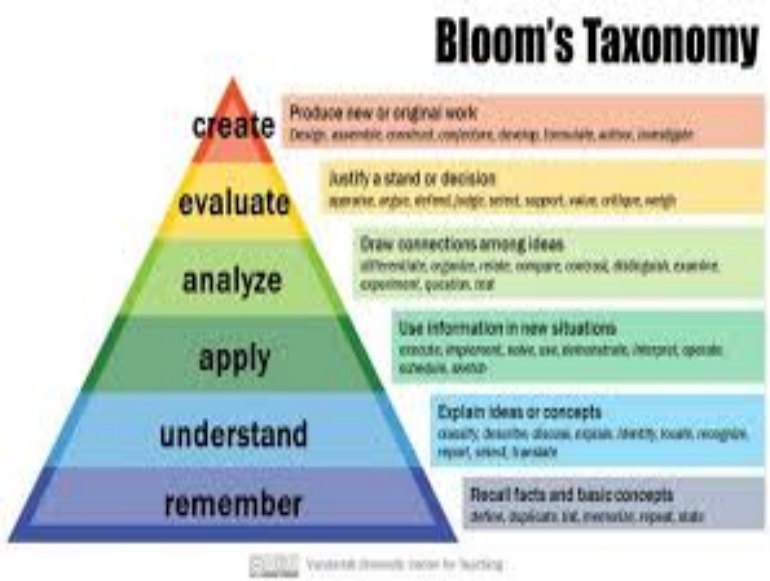


Taxonomy is clustering

Taxonomy (biology), a branch of science that encompasses the description, identification, nomenclature, and classification of organisms

-- Wikipedia





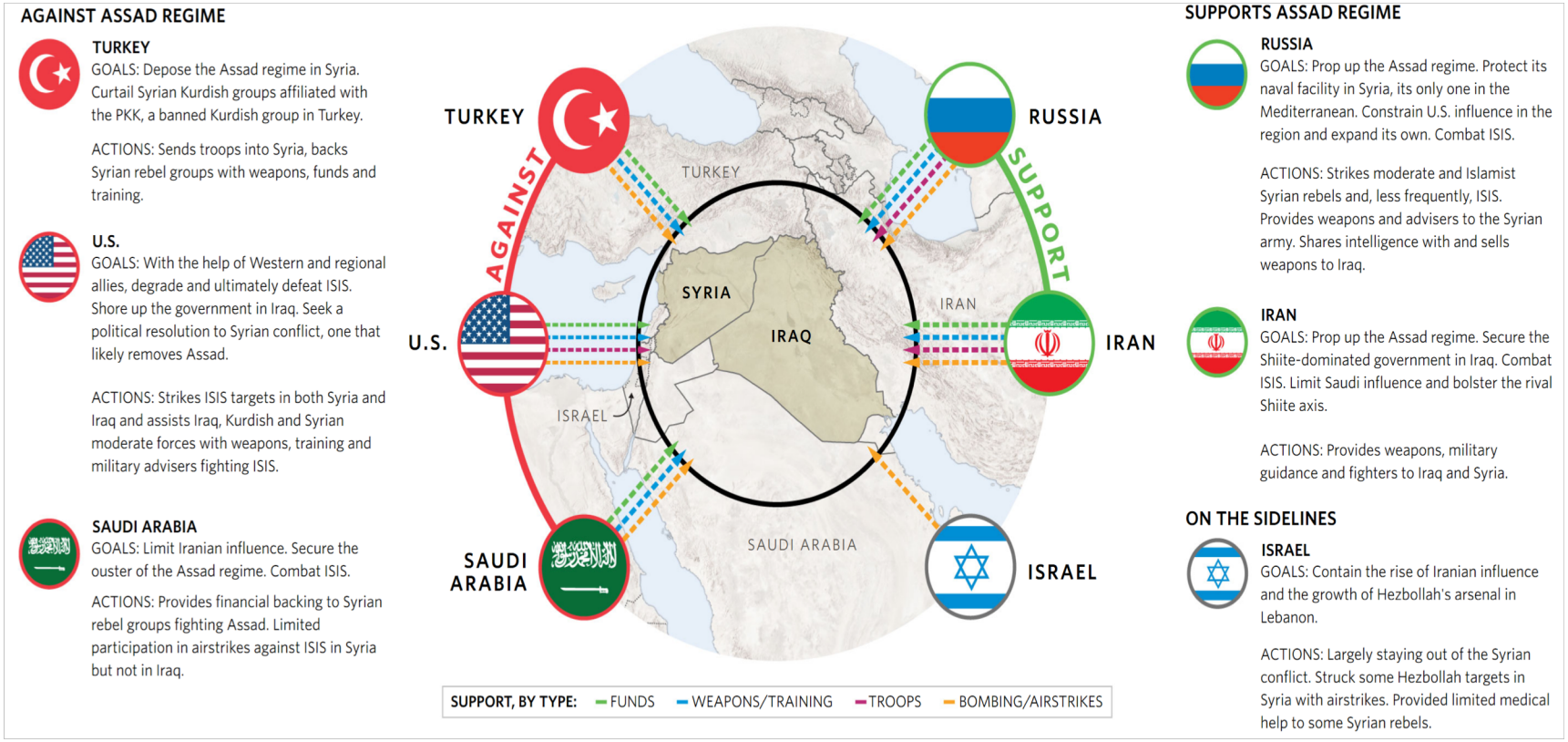
Clustering is a form of information reduction/organization

- To store in human finite memory (or computer's finite memory) and facilitate understanding

- To communicate between people (or processors) for more effective understanding between people and collective decisions

Effective decision-making is impossible based on raw big data

Understanding the Syria conflict



Tangled Alliances Graphic In "Airstrike Raises Tension with Tehran" in WSJ, April 8, 2017

Very helpful **clustered** information in terms of countries and important dimensions to consider. One criticism: country names could be placed better, not associated with the bars.

Red catches attention; size matters too.

Syria conflict

Tangled Alliances Graphic in

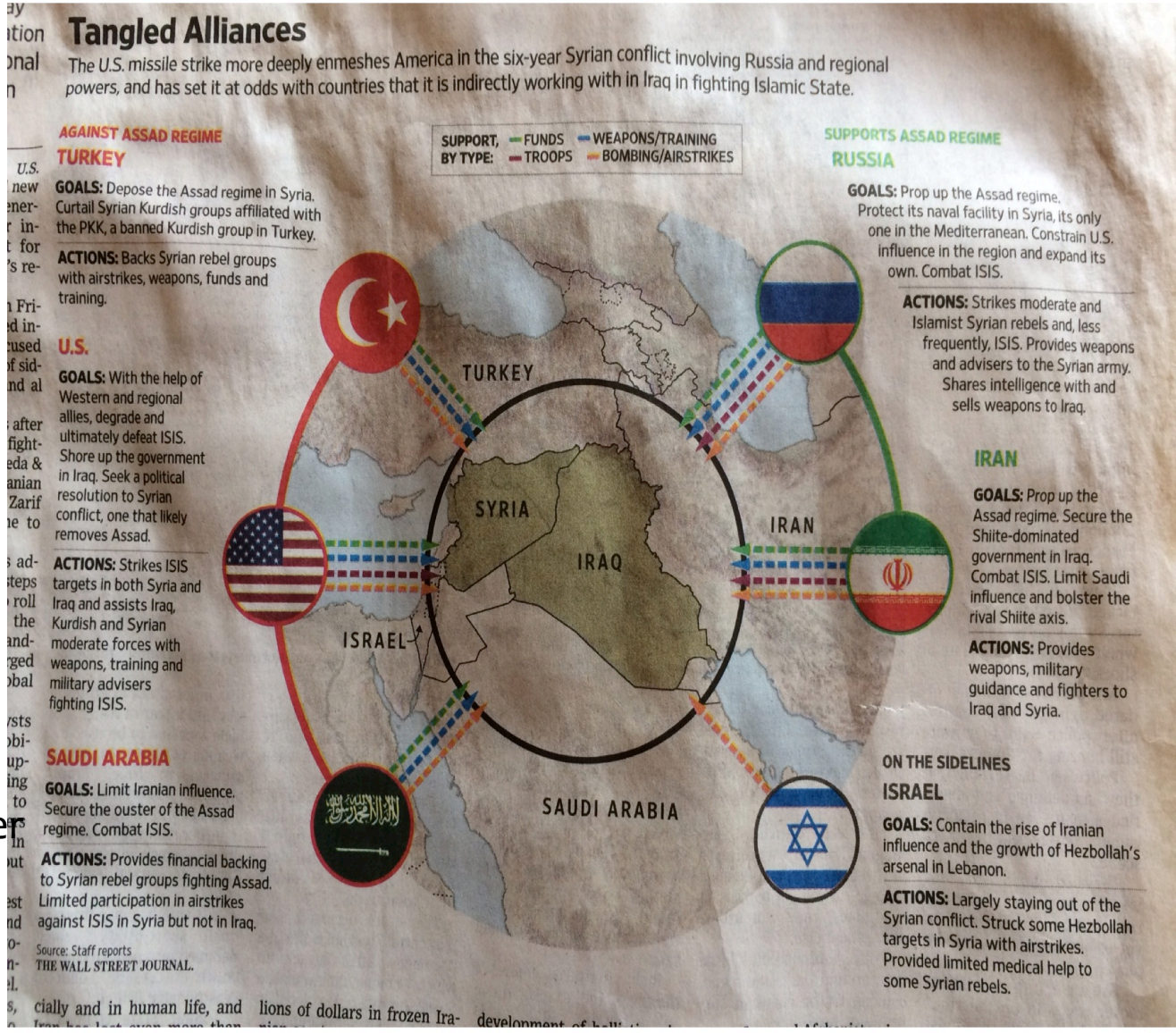
"Airstrike Raises Tensions with Tehran"

WSJ, April 8, 2017

Very helpful **clustered** information in terms of countries and important dimensions to consider.

One criticism: country names could be placed better, not associated with the bars- more pronounced in the paper version

Medium also matters.



Reading assignments

- Sections 10.2 and 10.3 in James et al
  on PCA, Hierarchical clustering and K-means