BD2K Module, Cancer Genomics

Exploratory Data Analysis

Andrew Nobel

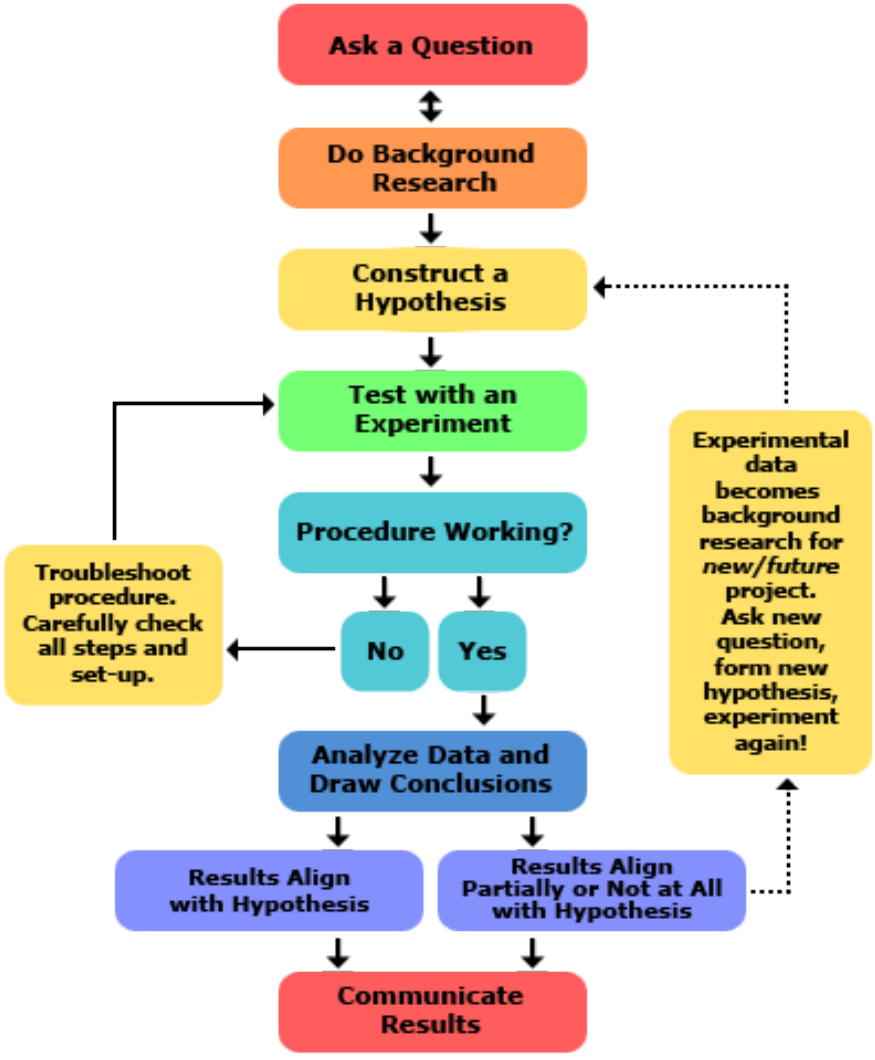
February 26, 2018

Overview

- The Scientific Method: Then and Now
- Reproducible Research
- Exploratory Data Analysis
- Clustering
- Biclustering
- Community Detection
- Correlation Mining

The Scientific Method

The Scientific Method: Flow Chart (from science buddies.org)



Paradigm Shift

Traditional Scientific Method: Hypothesis Driven

- Formulate a hypothesis
- Collect data to confirm/refute hypothesis

Modern Scientific Method: Data Driven

- Acquire data from high-throughput measurement technologies
- Mine the data for possible hypotheses
- Use the data again to test selected hypotheses

Needles and Haystacks

General Principle: If you have enough data, and you ask enough questions, you are bound to find something interesting, **just by chance.**

Bob: I found a needle in a haystack!

Amy: That seems very surprising. How many haystacks did you look in?

Bob: A thousand.

Amy: Oh, maybe that's not so surprising.

Two Facets of Reproducible Research

I. **Reproducibility of scientific analysis:** Can we replicate the analysis?

- Public access to raw data and preprocessing steps
- Public access to general and special purpose software
- Careful step-by-step documentation of data analysis

II. **Reproducibility of scientific conclusions:** Are the conclusions true?

- Are data, methods, and assumptions of initial study sound?
- Are results of initial study robust?
- Do similar experiments with different data yields the same conclusion?

Reproducibility Crisis

2015: Re-examination of 100 psychology studies

- About 33 studies were reproducible

2012: Re-examination of 53 landmark studies in oncology and hematology.

- Only 6 studies were reproducible

2009: Re-examination of 18 gene expression studies

- Only 2 studies were reproducible

Lack of Reproducibility: Some Causes

Experimental Process

- Cognitive bias: Favor supporting data over contradictory data
- Fabrication of data and/or mis-use of data analysis (infrequent)
- Change the hypothesis after seeing the data
- Try out lots of hypotheses until you find one supported by data

Publication Process

- Submission bias (of researcher): Only submit positive results
- Publication bias (of journal): Only publish positive results

Survey of 2000 US Psychologists (2012)

- 50% selectively reported only studies that were successful
- 58% looked at initial results, and then decided if they should collect more data
- 43% threw out “bad” data
- 35% reported unexpected findings as predicted from the outset

Exploratory Data Analysis

Exploratory Data Analysis

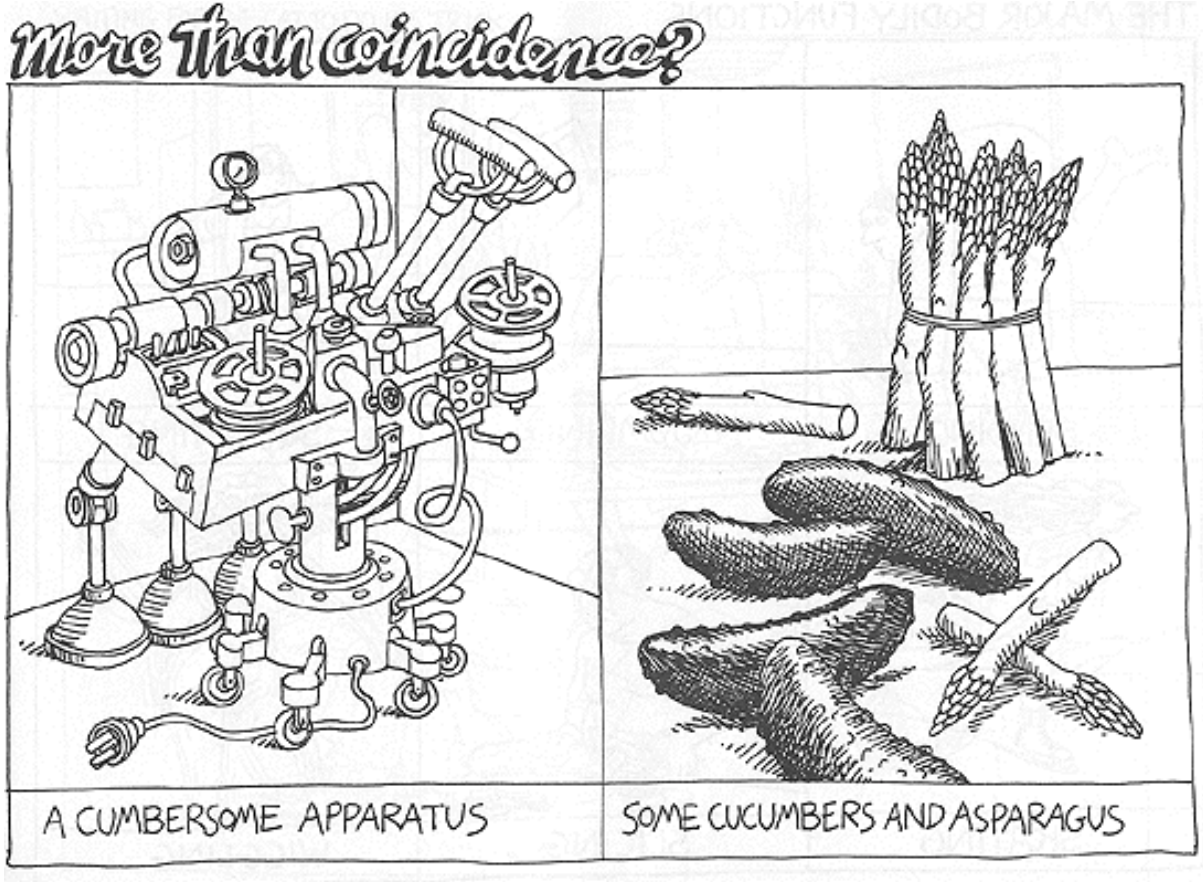
First look at a data set, typically in the form of a matrix of numbers.

- Visualization
- Identifying patterns or regularities of interest

Preliminaries:

- Identifying and addressing outliers and extreme values
- Imputing missing values
- Normalization: removing systematic differences between samples
- Transforming data values using logarithm or other functions
- Checking distributional/model assumptions

Finding Patterns



Drawing by B. Kliban

Univariate Sample $x = x_1, \ldots, x_n$

Statistics

- Sample mean $m(x) = \overline{x} = n^{-1} \sum_{i=1}^{n} x_i$
- Sample variance $s^2(x) = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ and SD $s(x)$
- Standardized sample $\tilde{x}$ with $\tilde{x}_i = (x_i - \overline{x})/s(x)$
- Quantiles, percentiles, and order statistics

Visualization

- Histogram/density plots
- Bar and whisker plots, QQ plots

Bivariate Sample $(x, y) = (x_1, y_1), \ldots, (x_n, y_n)$

Statistics

- Sample covariance of x and y

$$s(x, y) \;=\; n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \;=\; n^{-1} \sum_{i=1}^{n} x_i \, y_i - \overline{x} \, \overline{y}$$

- Sample correlation of x and y

$$r(x, y) \;=\; \frac{s(x, y)}{s(x) \, s(y)} \;\in\; [-1, 1]$$

Visualization

- Scatter-plot $\{(x_i, y_i) : 1 \leq i \leq n\} \subseteq \mathbb{R}^2$

Aside: Regression Line and R-squared

Def'n: Sample regression line of y on x is the line $\ell^*(x)$ minimizing

$$\text{MSE}(\ell) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \ell(x_i))^2$$

over all linear functions $\ell(x) = ax + b$.

Fact: Sample regression line ℓ^* of y on x is given by

$$\ell^*(x) = m(y) + \frac{s(x,y)}{s^2(x)}[x - m(x)]$$

and satisfies $\text{MSE}(\ell^*) = s^2(y)[1 - r^2(x,y)]$.

Note: $s^2(y) =$ MSE of straight line $l(x) = m(y)$.

High-throughput Genomic Data

Represented as a $p \times n$ data matrix $\mathbf{X} = \{x_{i,j}\}$ with n columns and p rows

- n columns corresponding to n samples
- p rows corresponding to p genomic variables
- $x_{i,j} =$ value of variable i in sample j

Common Examples

- gene expression data
- copy number data
- methylation data
- genotype data

Exploratory Analysis of Genomic Data

Step 1a: Univariate analysis of columns and rows of data matrix $\mathbf{X}$

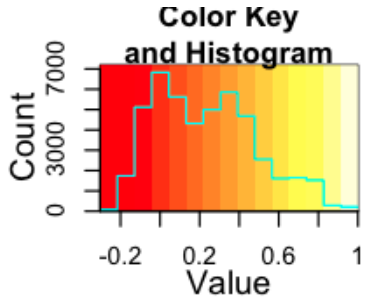
- sample/variable means and standard deviations
- histograms of these

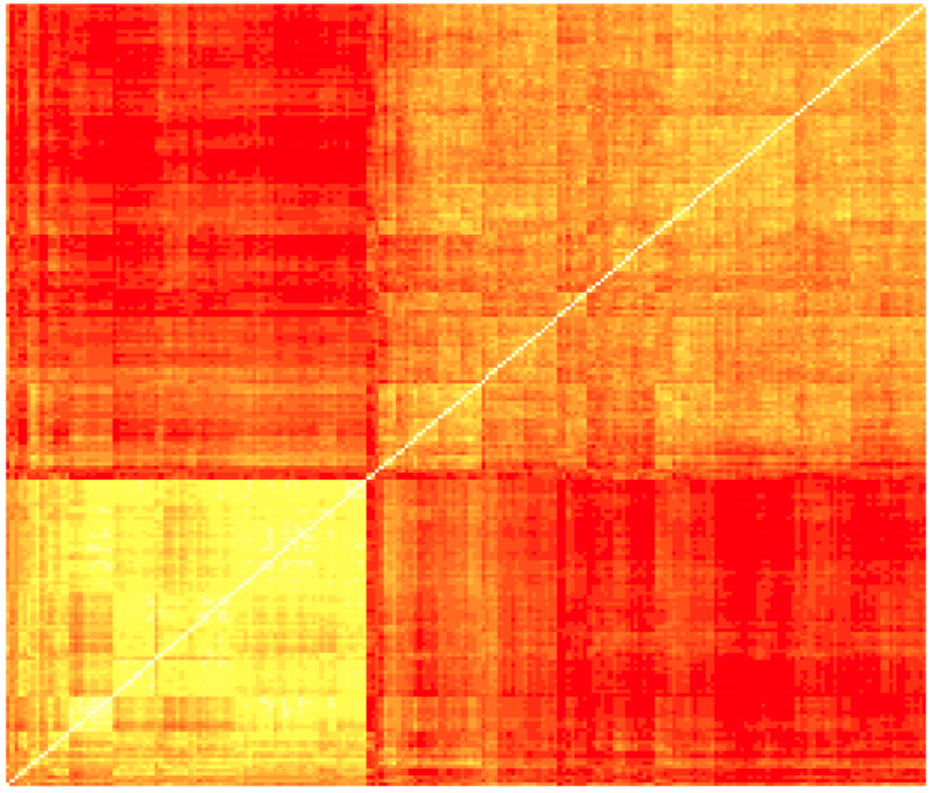
Step 1b: Bivariate analysis of columns and rows of data matrix $\mathbf{X}$

- heatmap of $n \times n$ matrix of correlations between samples
- heatmap of $p \times p$ matrix of correlations between variables
- scatter plots

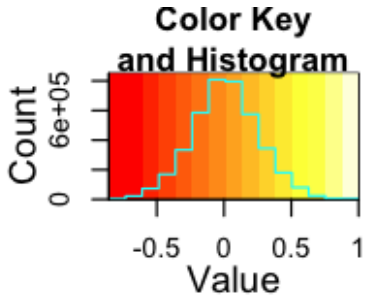
Next steps: Principal component analysis (PCA), clustering, biclustering

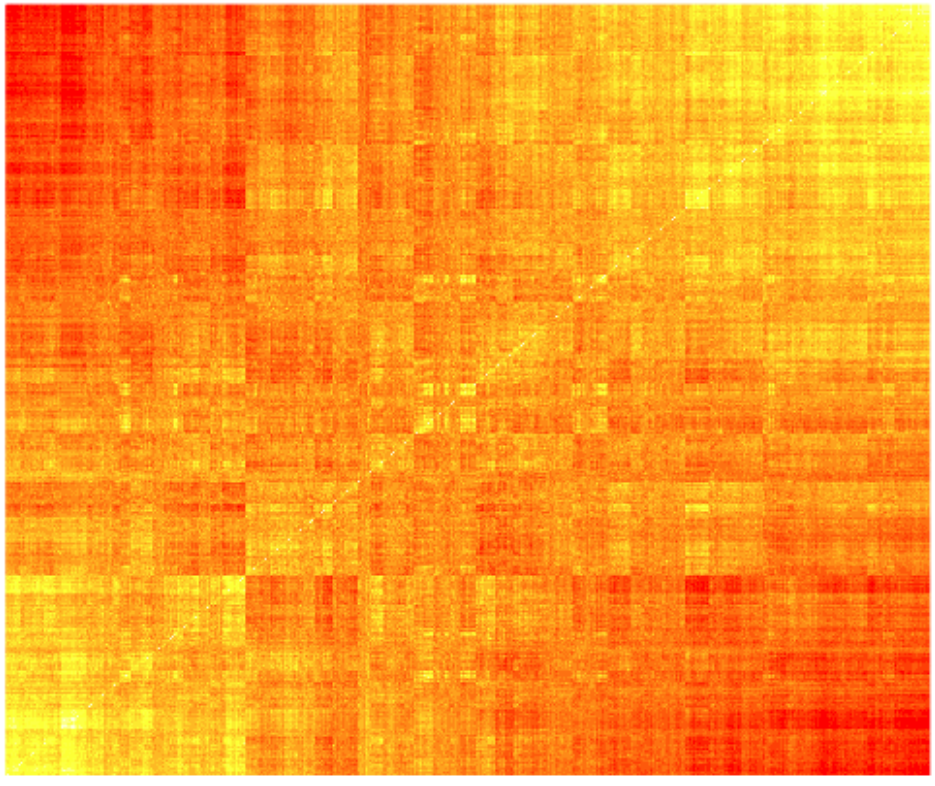
Heatmap: Correlation Matrix of Samples ($n \times n$)



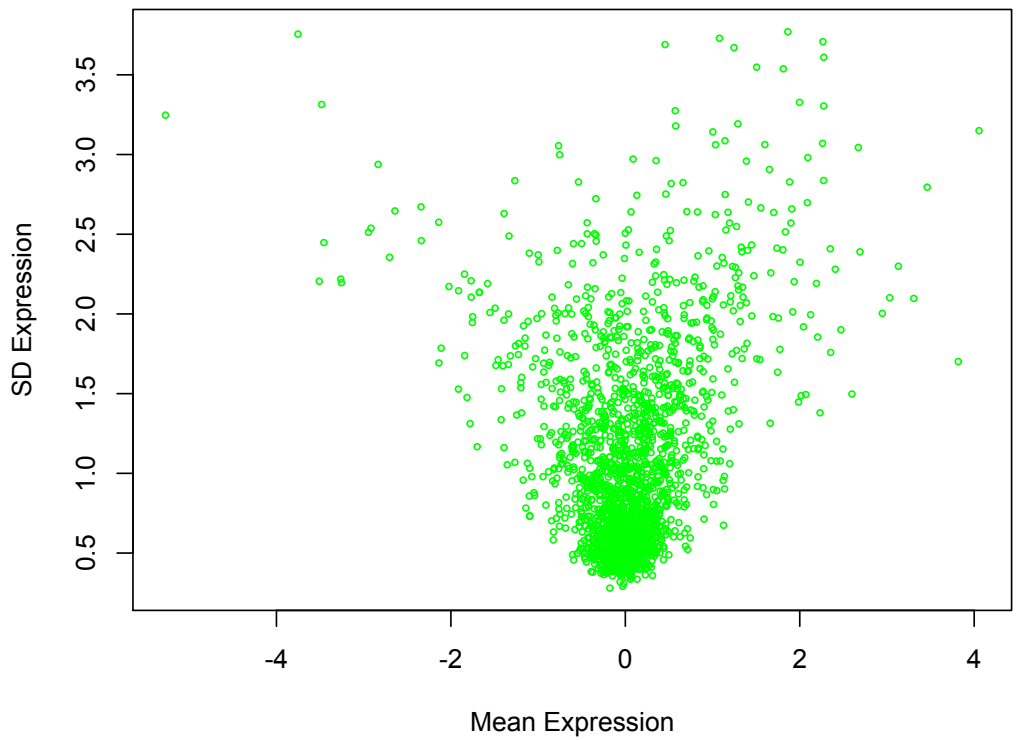


Heatmap: Correlation Matrix of Genes ($p \times p$)

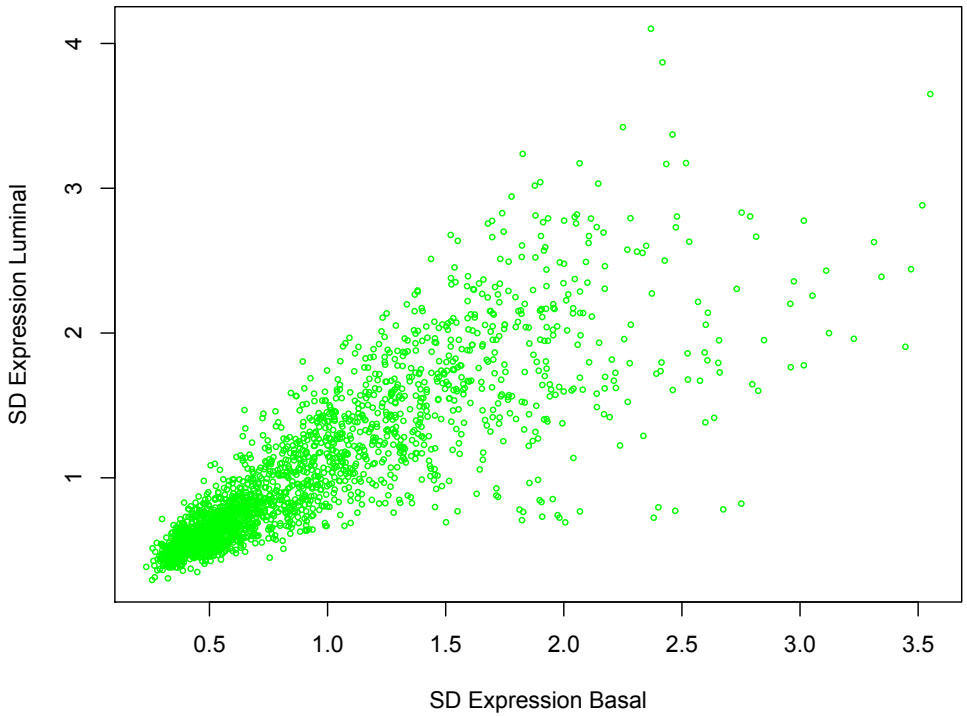




Scatterplot of Mean and SD of Expression



Scatterplot of SD(expression) for Two Subtypes



- Correlation: $r = 0.8384$

Principal Component Analysis

Principal Component Analysis (PCA)

Given: High dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ with $\sum_i \mathbf{x}_i = \mathbf{0}$

Goal: Find a subspace V of $\mathbb{R}^p$ meeting two criteria

- *Dimension reduction:* the dimension of V is small (much less than p, n)
- *Approximation:* sample $\mathbf{x}_j$ is close to its projection onto V

Goal: Subspace V is a good low dimensional approximation of the data.

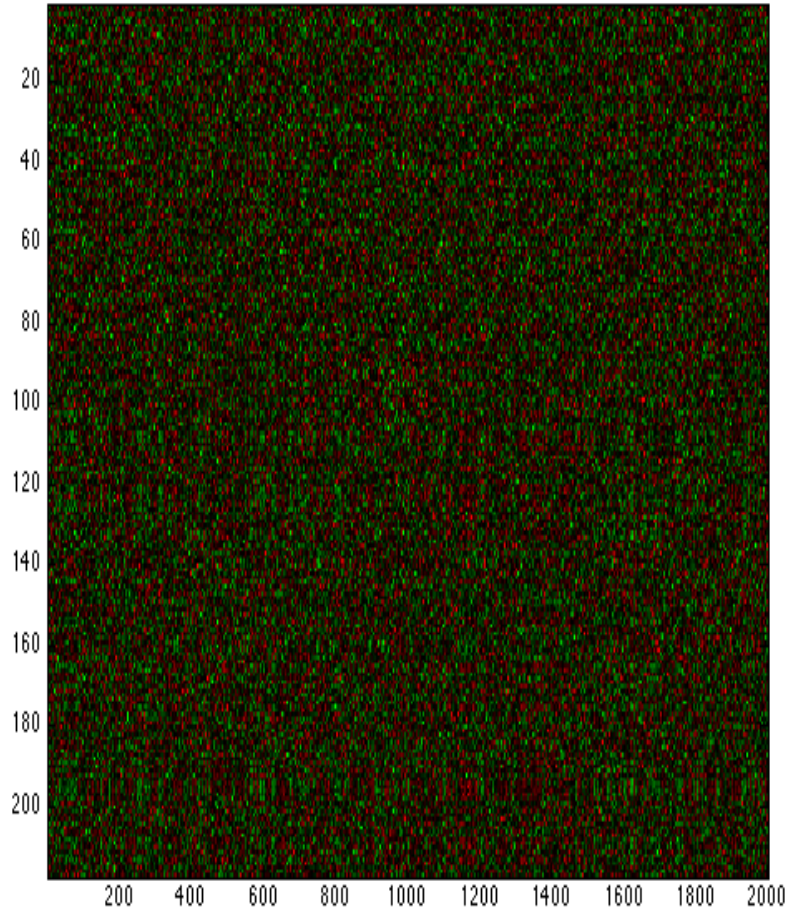
PCA, cont.

Simplest case: Approximating subspace V is one-dimensional, that is, a line in $\mathbb{R}^p$ determined by a unit vector $\mathbf{v}$.

Turns out

- Finding a good direction is equivalent to maximizing the variance of the projections of the samples $\mathbf{x}_1, \ldots \mathbf{x}_n$ onto $\mathbf{v}$.
- The best direction $\mathbf{v}_1$ corresponds to leading eigenvector of the $p \times p$ sample covariance matrix $\mathbf{S} = n^{-1}\mathbf{X}\,\mathbf{X}^T$, with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$.
- Other directions $\mathbf{v}_2, \mathbf{v}_3, \ldots$ can be obtained from other eigenvectors of $\mathbf{S}$.

Example TCGA Gene Expression Data



Heat map of gene expression data from
The Cancer Genome Atlas (TCGA)

- Samples $n = 117$, two groups
  - 95 Luminal A breast tumors
  - 122 Basal breast tumors
- Variables: $p = 2000$ randomly selected genes

PCA on TCGA Expression Data

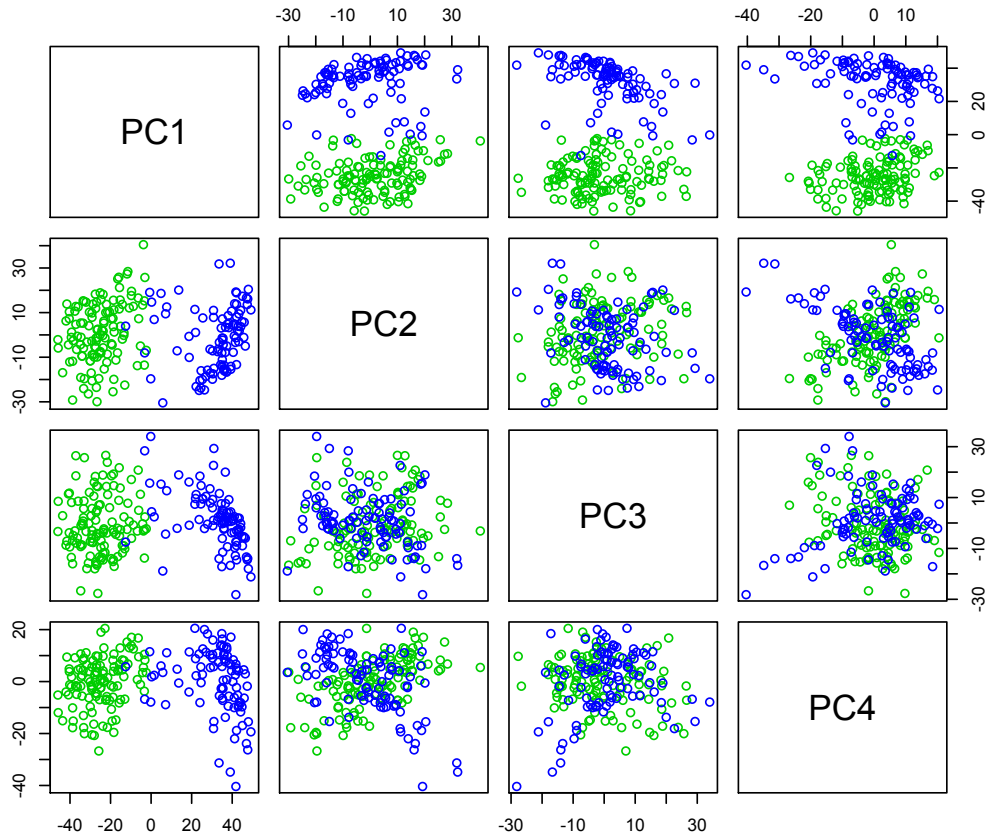
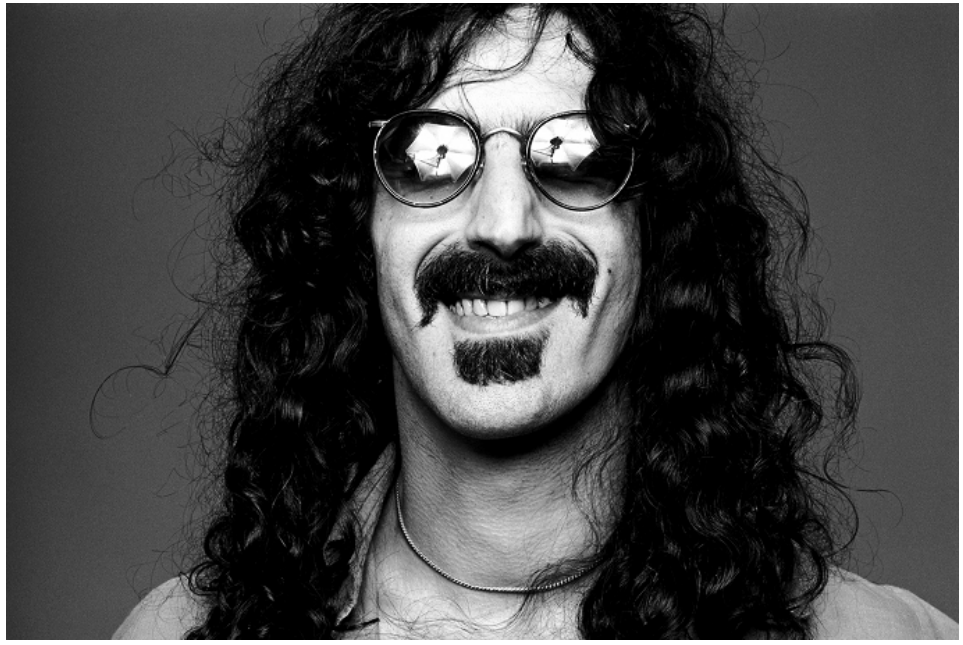


Figure: Projections of Sample data onto the first four principal components of the TCGA dataset. Colors represent subtype of cancer: Luminal A and Basal

Image Data



- **Data**: $\mathbf{X} = 458 \times 685$ matrix of pixel intensities
- **Idea**: Project columns of the image onto $\mathbf{d}$ leading eigenvectors of their sample covariance matrix. Consider quality of reconstruction.

Proportion of Variation Explained

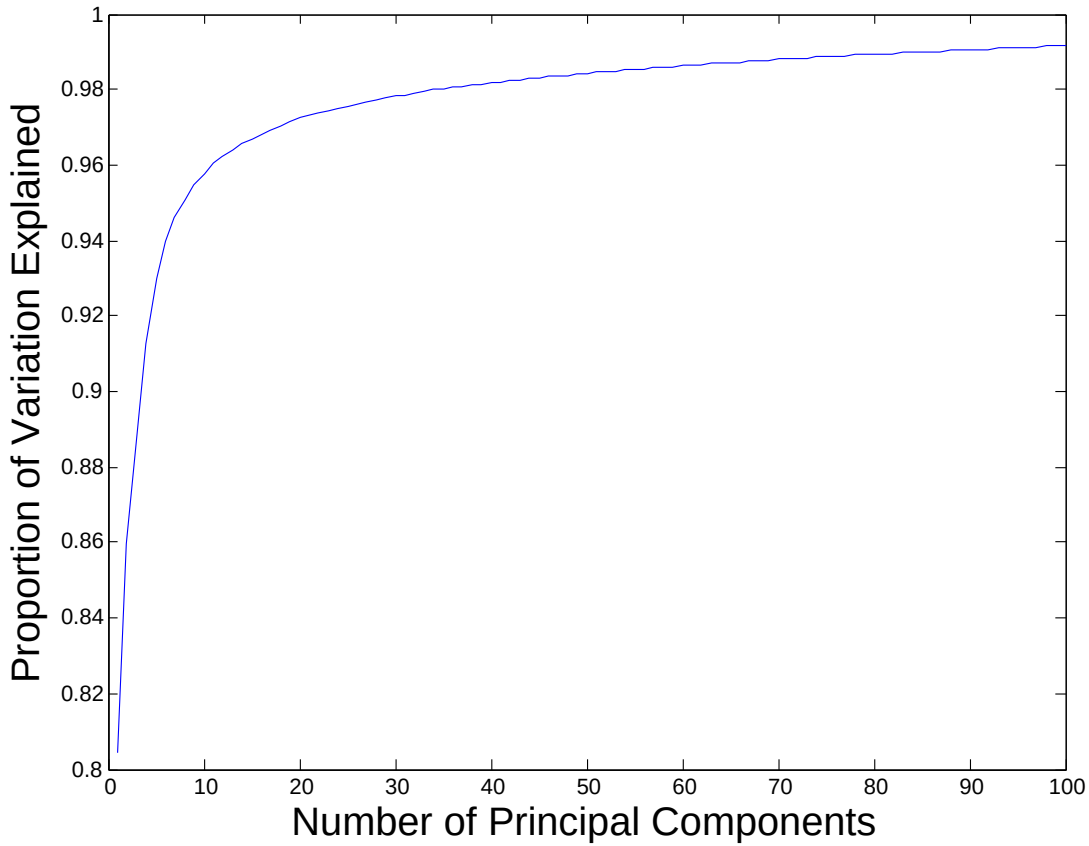


Image Reconstruction

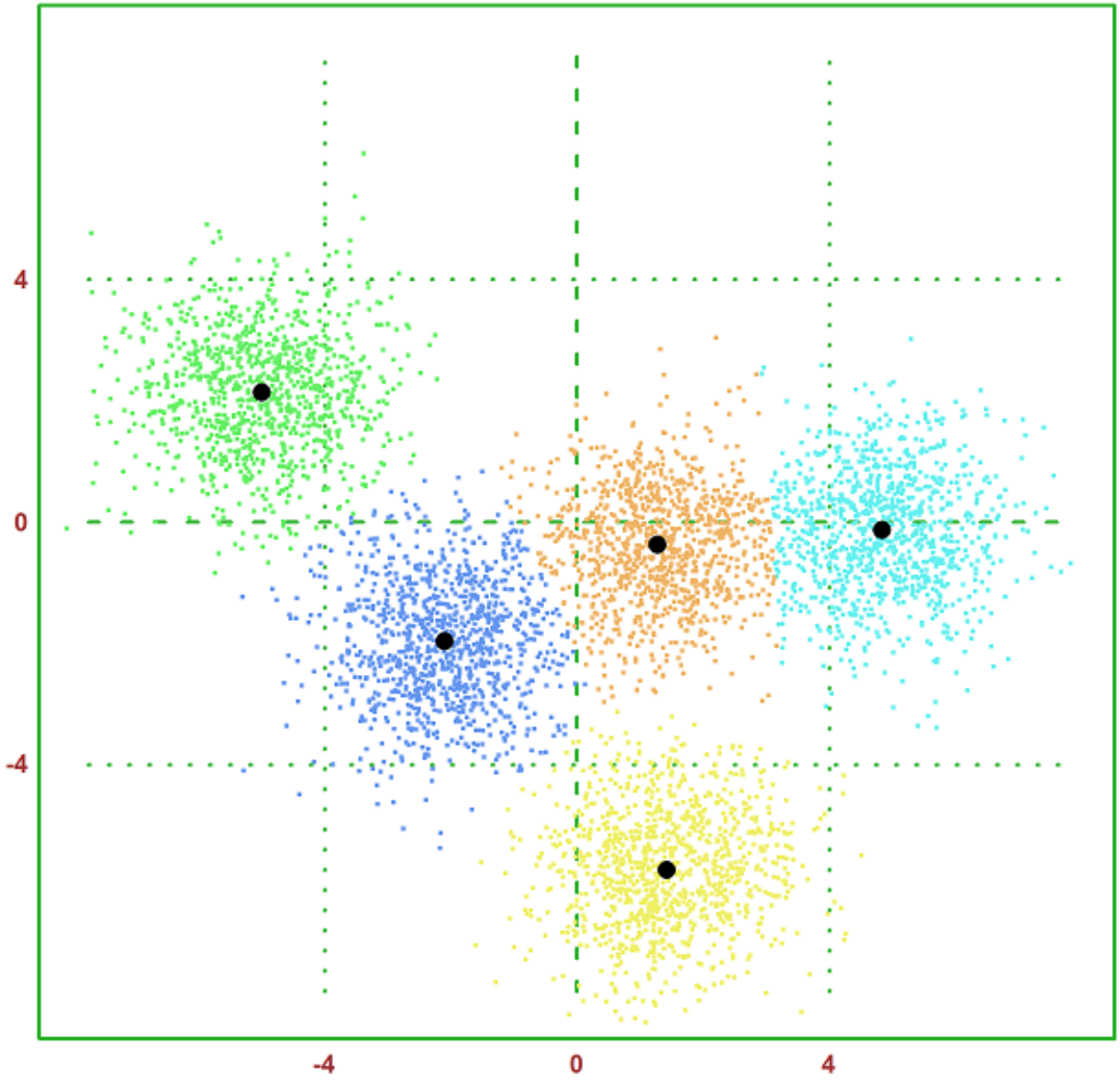
d = 10, PVE = 95.79

d = 20, PVE = 97.24

d = 40, PVE = 98.18

Clustering

Example (http://rosettacode.org)



General Setting

Given: Vectors $x_1, \ldots, x_n \in \mathbb{R}^d$

Goal: Identify group structure. Divide vectors into a small number of disjoint groups, called *clusters*, such that

- distances between vectors in the same cluster are small
- distances between vectors in different clusters are large

Areas of application

- Genomics and Biology
- Computer Science
- Psychology and Social Sciences

Some Clustering Approaches

Hierarchical: Candidate divisions of data described by a binary tree

- *Agglomerative (bottom-up)
- Divisive (top-down)

Iterative: Search for local minimum of simple cost function

- *k-means and variants
- Partitioning around medioids

Model-based: Fit feature vectors by a mixture of Gaussians

Spectral: Cluster top eigenvectors of Laplacian of dissimilarity matrix

The k-Means Algorithm

Given: Observations $x_1, \ldots, x_n \in \mathbb{R}^d$ *and* desired number of clusters k

Initialize: Cluster centers $\mathcal{C}_0 = c_0(1), \ldots, c_0(k) \in \mathbb{R}^d$

Iterate: For $m = 1, 2, \ldots$ do:

- Let π_m be the nearest neighbor partition of the centers $\mathcal{C}_{m-1}$.
- Let $\mathcal{C}_m$ be the centroids (averages) of the vectors in each cell of π_m

Stop: When $\text{Cost}(\mathcal{C}_m) = \sum_{i=1}^{n} \min_{1 \le j \le k} ||x_i - c_m(j)||^2$ stabilizes

Agglomerative Clustering

Stage 0: Assign each object x_i to its own cluster

Stage k:

- Find the two *closest* clusters at stage $k-1$
- Combine them into a single cluster

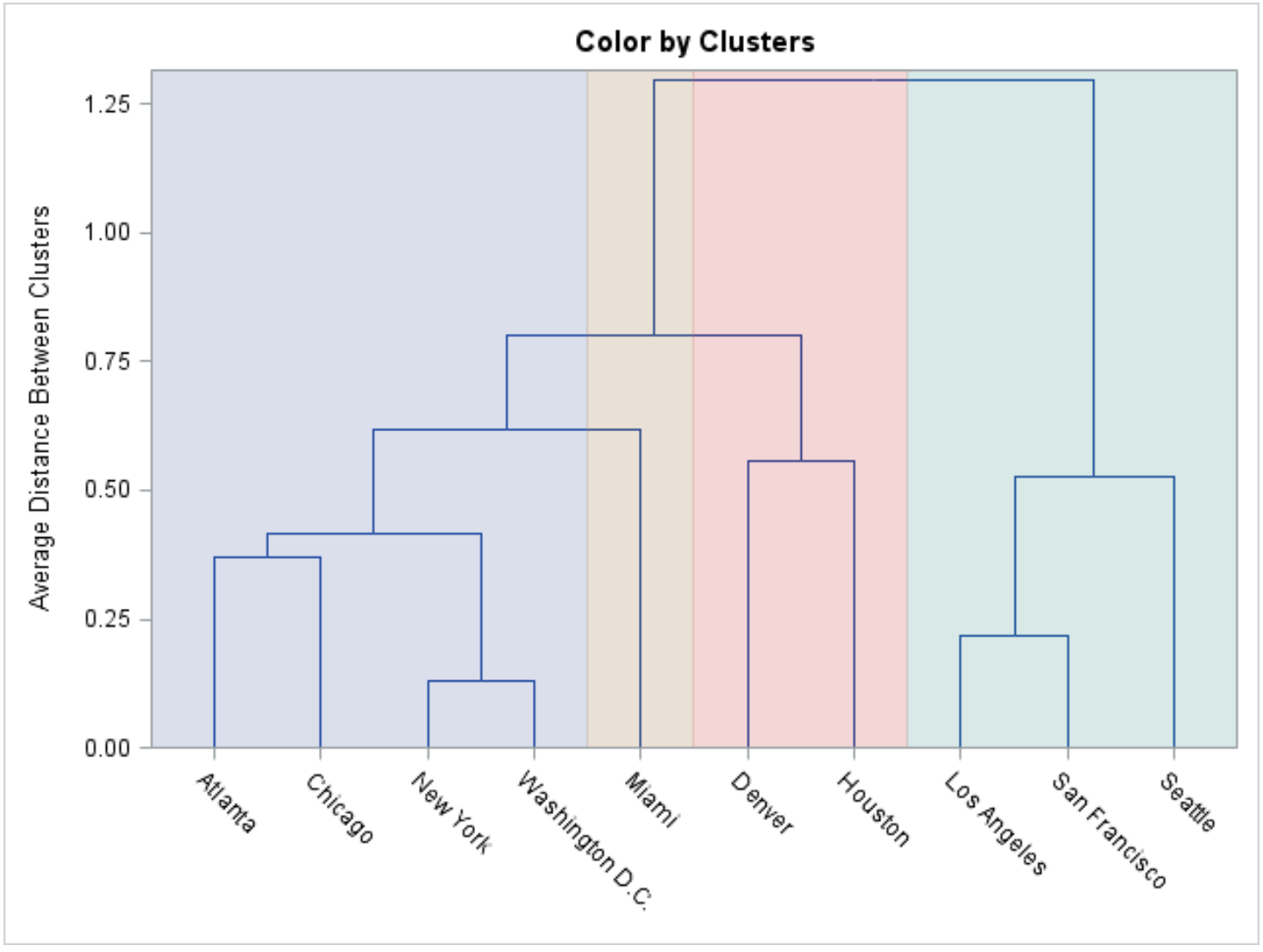
Stop: When all objects x_i belong to a single cluster

Output: Dendrogram $=$ binary tree where every node corresponds to a cluster, height of a node is distance between its children.

Note: Distance $d(C, C')$ between clusters C, C' measured in different ways

$$\min_{x_i \in C,\ x_j \in C'} d(x_i, x_j) \quad \text{or} \quad \frac{1}{|C|\,|C'|} \sum_{x_i \in C,\, x_j \in C'} d(x_i, x_j)$$

Cities by Distance (blogs.sas.com)

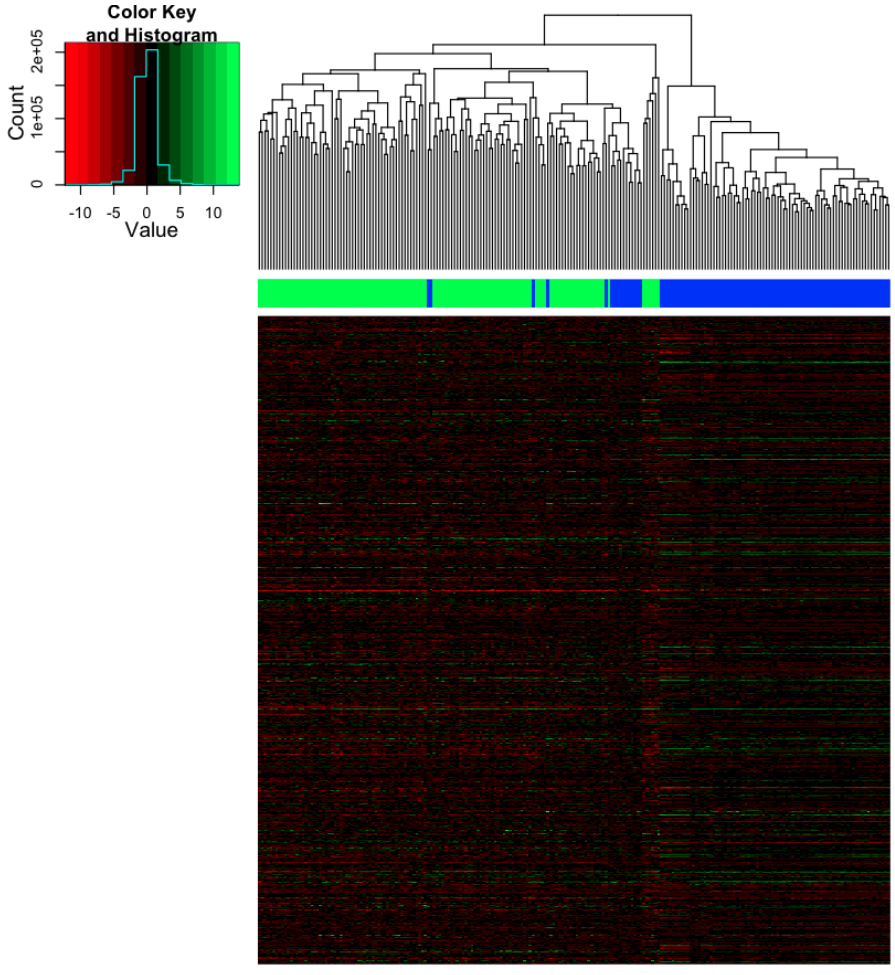


TCGA Data

Gene expression data from The Cancer Genome Atlas (TCGA)

- **Samples**
  - 95 Luminal A breast tumors
  - 122 Basal breast tumors
- **Variables**: 2000 randomly selected genes

TCGA Data



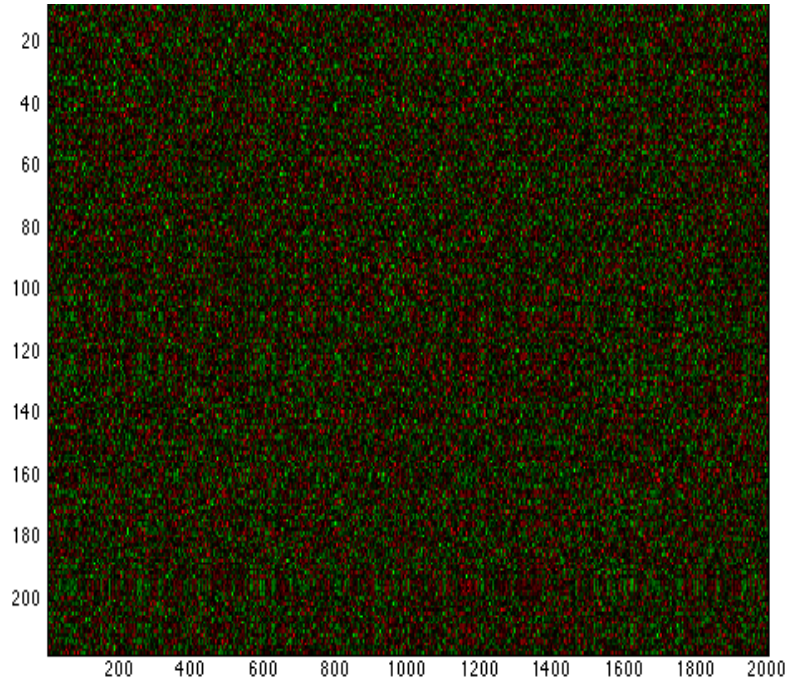
- Clustered samples (breast tumor subtype)
- Colors: Luminal A and Basal

Important Questions

- What is the right number of clusters?
- What is right measure of distance?
- Which clustering method to use?
- How robust is an observed clustering to small perturbations of the data?
- What significance can be assigned to the clusters?

Co-Clustering and Biclustering

TCGA Gene Expression Data

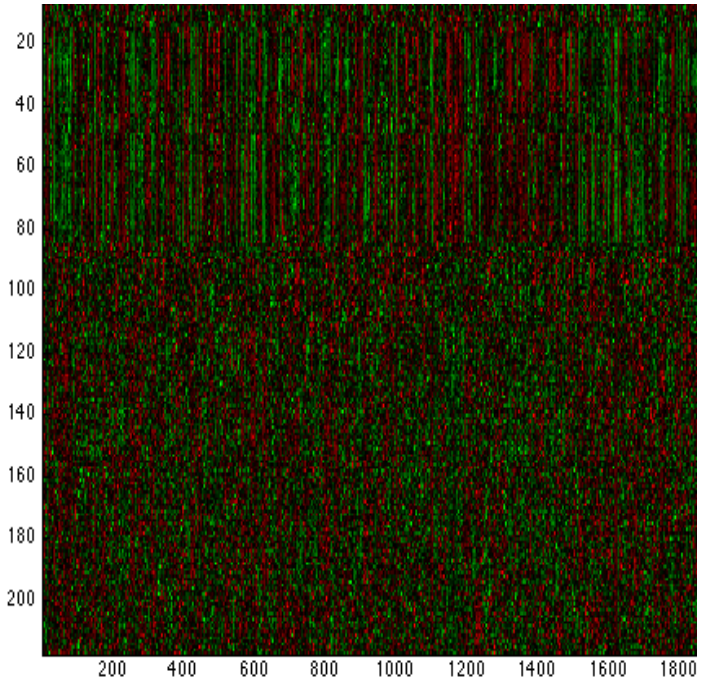


Heat map of gene expression data from The Cancer Genome Atlas (TCGA)

- Samples
  - 95 Luminal A breast tumors
  - 122 Basal breast tumors
- Variables: 2000 randomly selected genes

Row and Column Clustering

Row Clustering



Column Clustering

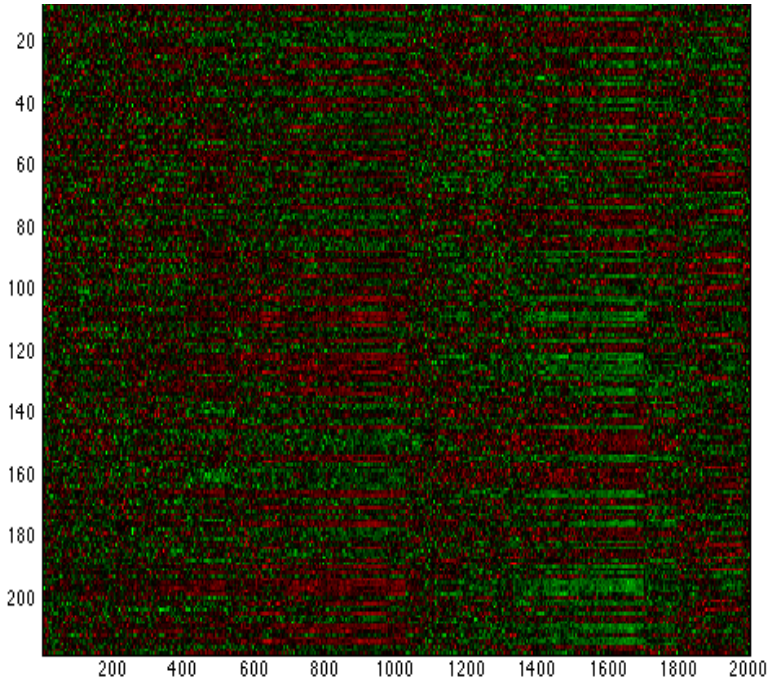
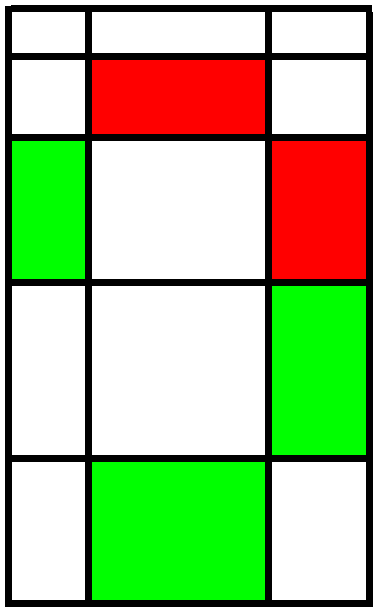


Figure: (Left) Rows reordered according to hierarchical clustering. (Right) Columns reordered according to hierarchical clustering.

Co-Clustering



Independently cluster rows and columns of the data matrix.

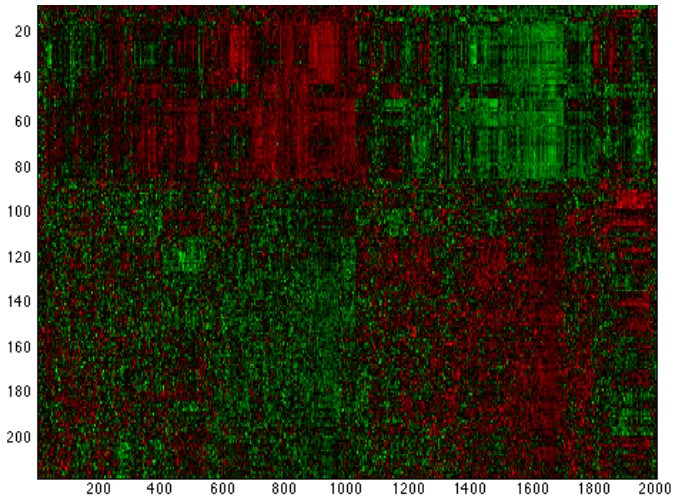
Result is a checkerboard partition

Note: Red, green blocks correspond to *large average submatrices* representing sample-variable interactions. Potential

- disease subtypes
- regulatory pathways

Co-Clustering and Biclustering

Co-Clustering



Biclustering

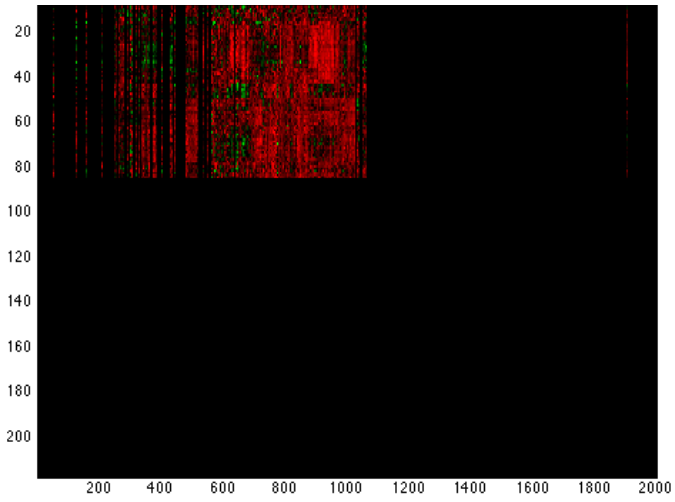


Figure: (Left): Co-Clustering: Rows and Columns of data matrix are separately reordered by clustering. (Right) The first bicluster extracted from this data.

Biclustering

Basic Idea: Search directly for a set of rows A and a set of columns B such that the entries of the submatrix

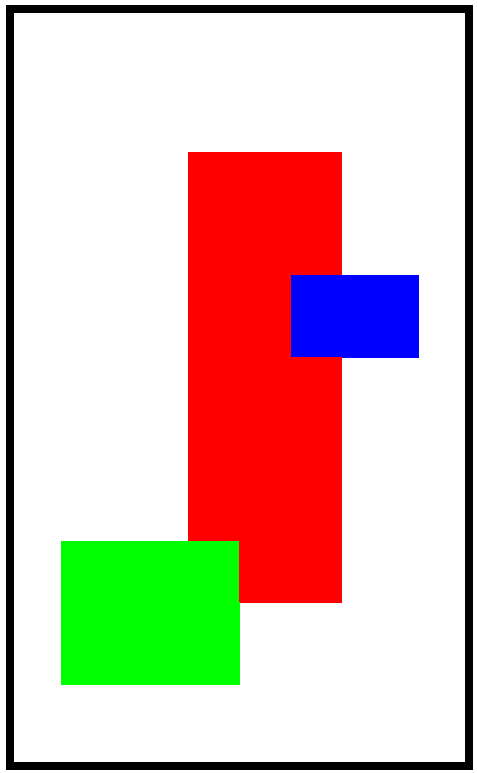
$$C = \{x_{i,j} : i \in A, j \in B\}$$

have large average. Rows and columns of C need *not* be contiguous.

Advantages over (co)clustering

- Direct search for sample-variable interactions
- Clusters may overlap and need not cover the entire data matrix: better reflects underlying biology.
- Local: Inclusion of samples/variables in a block depends only on their expression values inside the block.

Biclustering



Three overlapping Biclusters.

LAS Search Procedure (Shabalin et al. 2010)

Input: An $n \times n$ matrix $\mathbf{X}$ and integer $1 \leq k \leq n$.

Loop: Select k columns J at random. Iterate until convergence.

Let $I := k$ rows with largest sum over columns in J.

Let $J := k$ columns with largest sums over rows in I.

Output: *Locally optimum* submatrix associated with I, J.

In Practice

- Repeat 1000 times, adaptively choosing submatrix dimensions
- Output submatrix with largest average
- Residualize and repeat

Community Detection in Networks

Undirected Networks

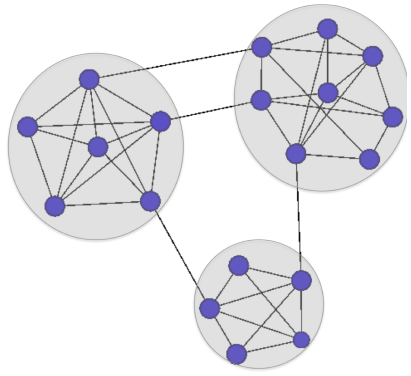
Simple Graph $G = (V, E)$ where

- Node set $V = [n] = \{1, \ldots, n\}$
- Edge set E with $\{u, v\} \in E$ if u is linked to v
- No self-loops or multi-edges

Degree Sequence $\mathbf{d} = \{d(1), \ldots, d(n)\}$ with

$$d(u) = \sum_{v \in V} \mathbb{I}(\{u, v\} \in E) \; = \; \text{number of edges incident on } u$$

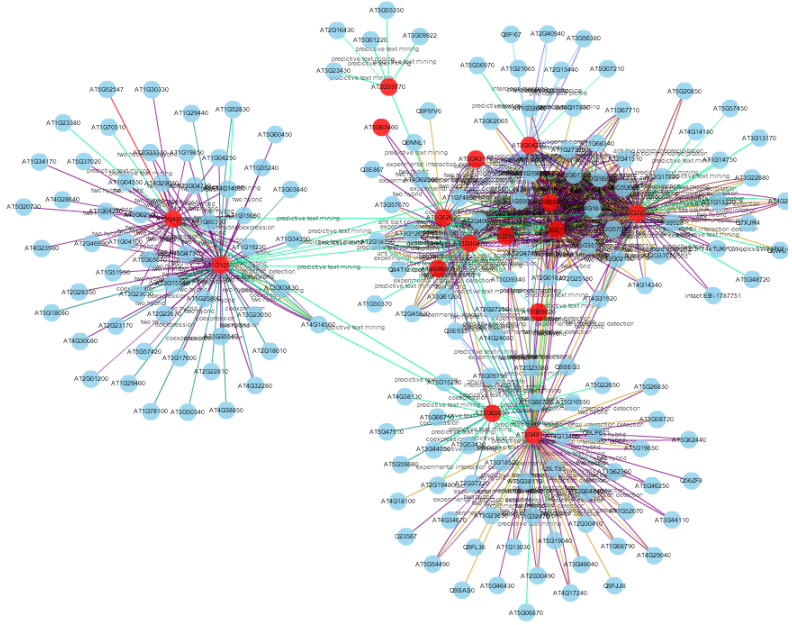
Community Detection (Informal)



Given $G = (V, E)$ identify sets $C_1, \ldots, C_k \subseteq V$ such that

- Edge density within sets C_i is large
- Edge density between sets C_i is small
- Sets C_i called *communities*

Community Detection: Applications



Exploratory Analysis of

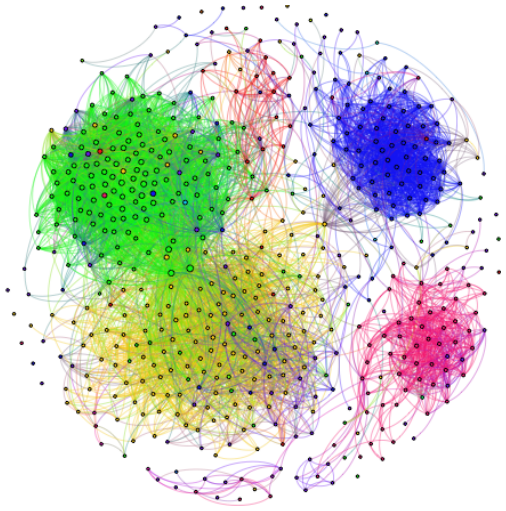
- Social networks
- Genetic networks
- Communication networks

Community Detection vs. Clustering

Community detection and clustering share common goal of grouping objects, but differ in fundamental ways:

Clustering	Community Detection
Feature Vectors	Nodes
Similarity (continuous)	Connectivity (binary)
Metric structure	Relational structure

Application: Facebook Network



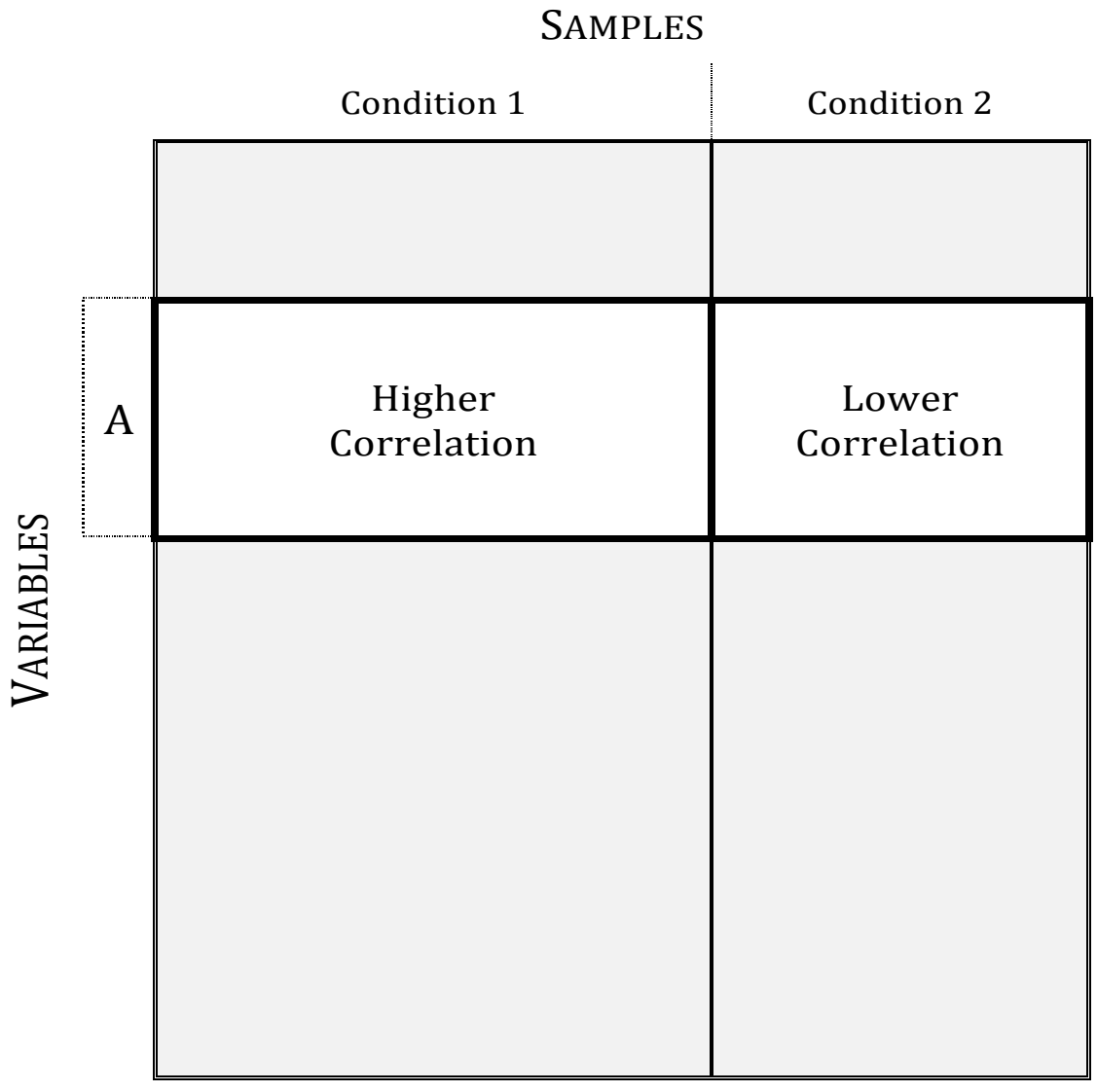
- Nodes $=$ friends of JW on FB (561)
- Edges between FB friends (8375)
- Friends divided into 8 different groups

Results of community detection (ESSC)

- 7 communities detected
- Match score $= .87$ out of 1

Mining Differential Correlation

Mining Differential Correlation



Mining Differential Correlation

Overall Goal: Adaptively identify differentially correlated variable sets A.

- Candidate variable set(s) A *not* specified in advance
- Special case of differential analysis for weighted networks

Non-Assumptions: Correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ potentially complex

- May *not* be diagonal, banded, or sparse.

Note: *Differential correlation distinct from differential expression, clustering*

Application areas: Genomics, Connectomics, Economics

Example: TCGA

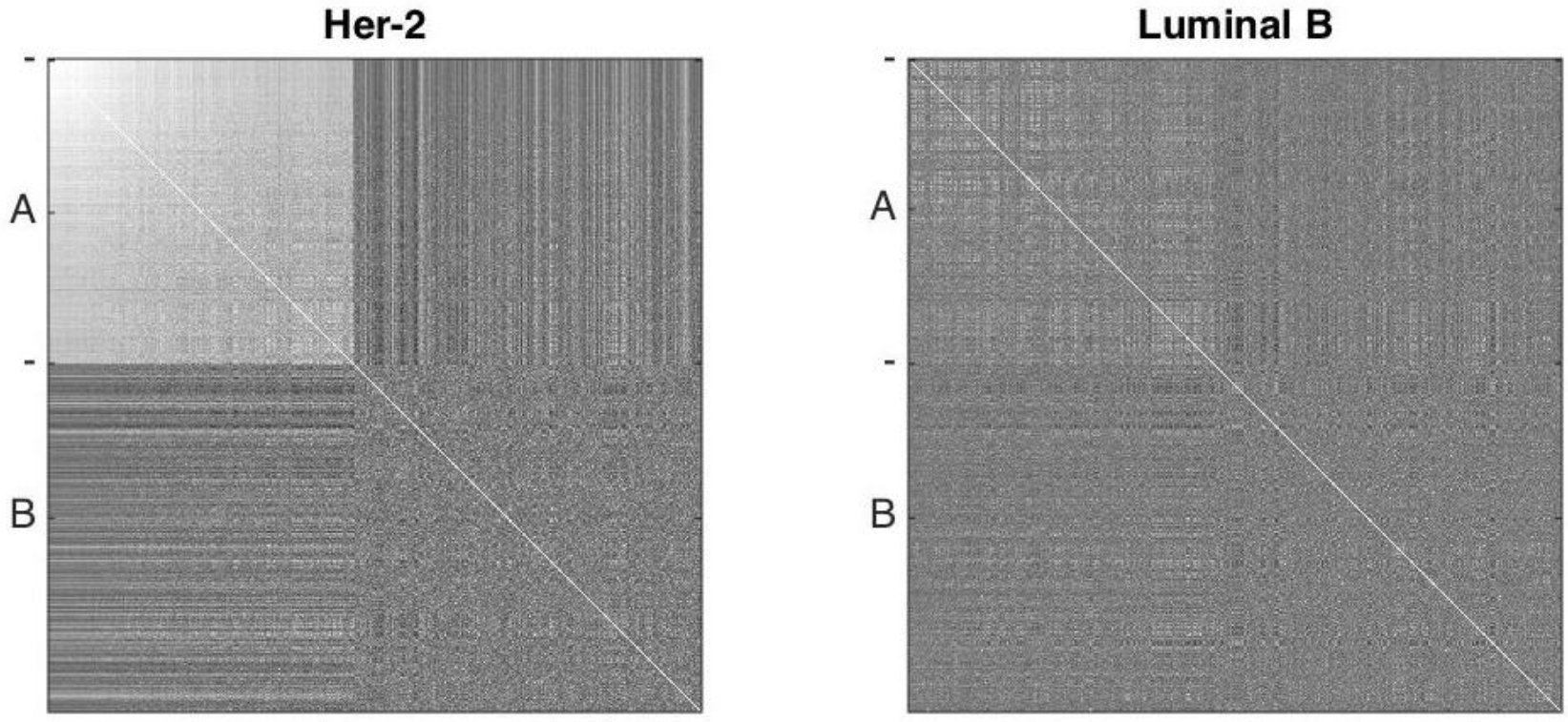


Figure: Sample correlation matrices from Her-2 and Luminal B cancer subtypes. Differentially correlated set of 165 genes (A) and 200 randomly chosen genes (B).

Application: Brain Connectome

FMRI data from Human Connectome Project (www.humanconnectome.org)

Single subject: 97K brain locations (37K voxels + 60K greyordinates)

- Condition 1: 316 language tasks
- Condition 2: 284 motor tasks

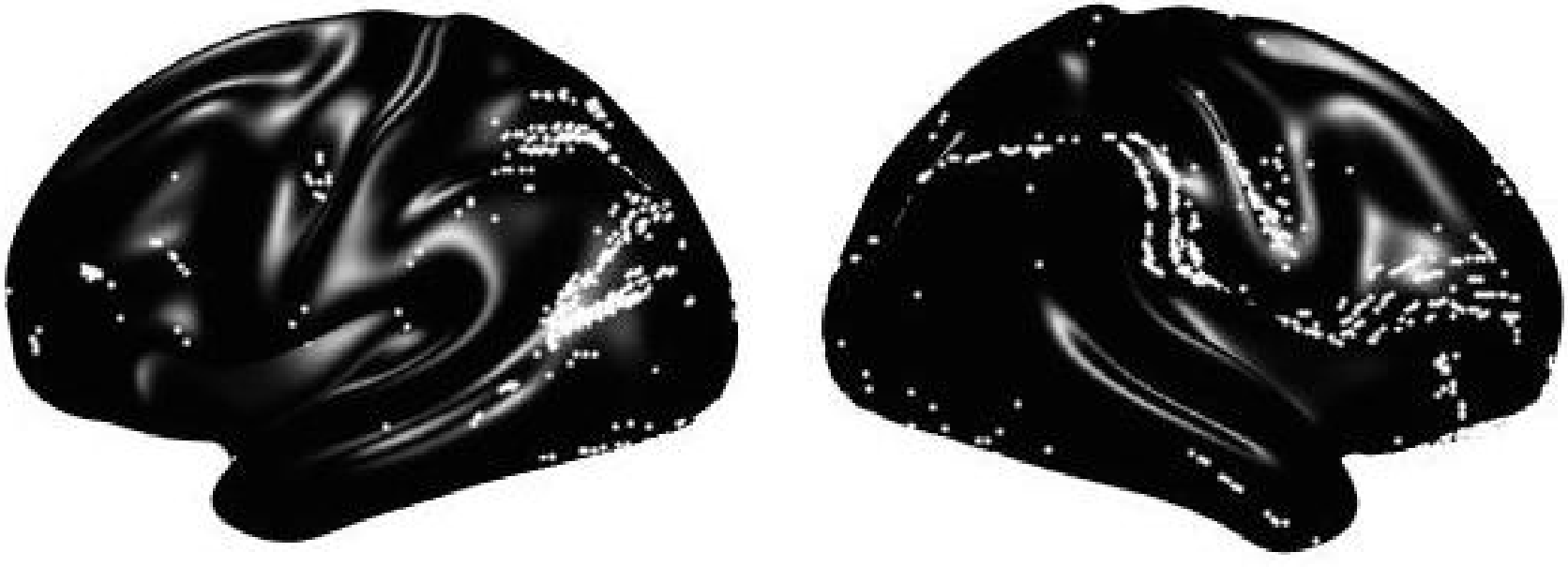
DCM output: 5 sets of brain locations

Time per DC set: 1-3 minutes (in Matlab)

Brain Connectome: Differential Correlation

First DC set: 1200 locations with $\overline{r}(C_1) = .24$ and $\overline{r}(C_2) = .05$

Visualization: DC locations on L/R hemisphere show clear spatial structure



Brain Connectome: Differential Expression

Visualization: Top 1200 locations as ranked by standard t-test

Conclusion

Recap

- The Scientific Method: Then and Now
- Reproducible Research
- Exploratory Data Analysis
- Principal Component Analysis
- Clustering and Biclustering
- Community Detection
- Correlationg Mining