



האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

Artificial Intelligence in Medicine

Clustering

Nir Friedman and Tommy Kaplan

5/12/22

“In my defense, I was left unsupervised”

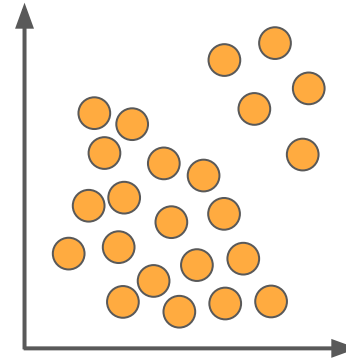
Lee St. John



Course outlook (so far)

The basics of machine learning

- Classifiers (= rules, predictions)
- Parameters learning
- Model selection
- What if data unlabeled?



Learning

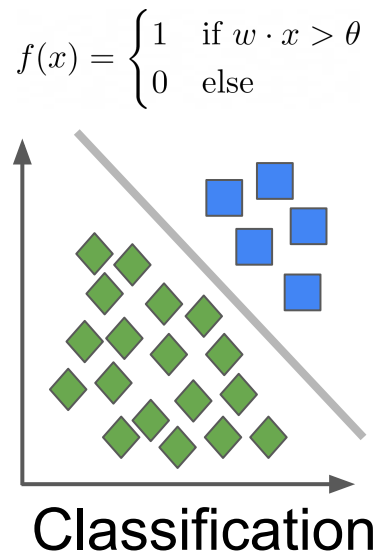
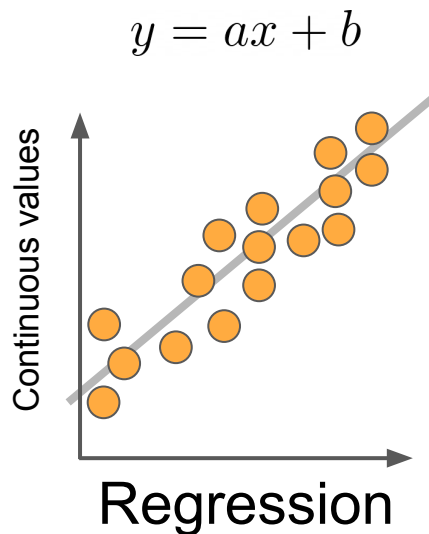
Finding the **optimal** model

model type, parameters,
simple, general, interpretable

Model

Set of rules
for prediction

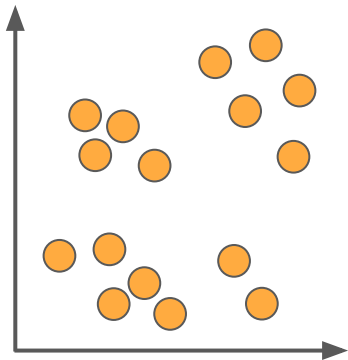
Data



Unsupervised learning

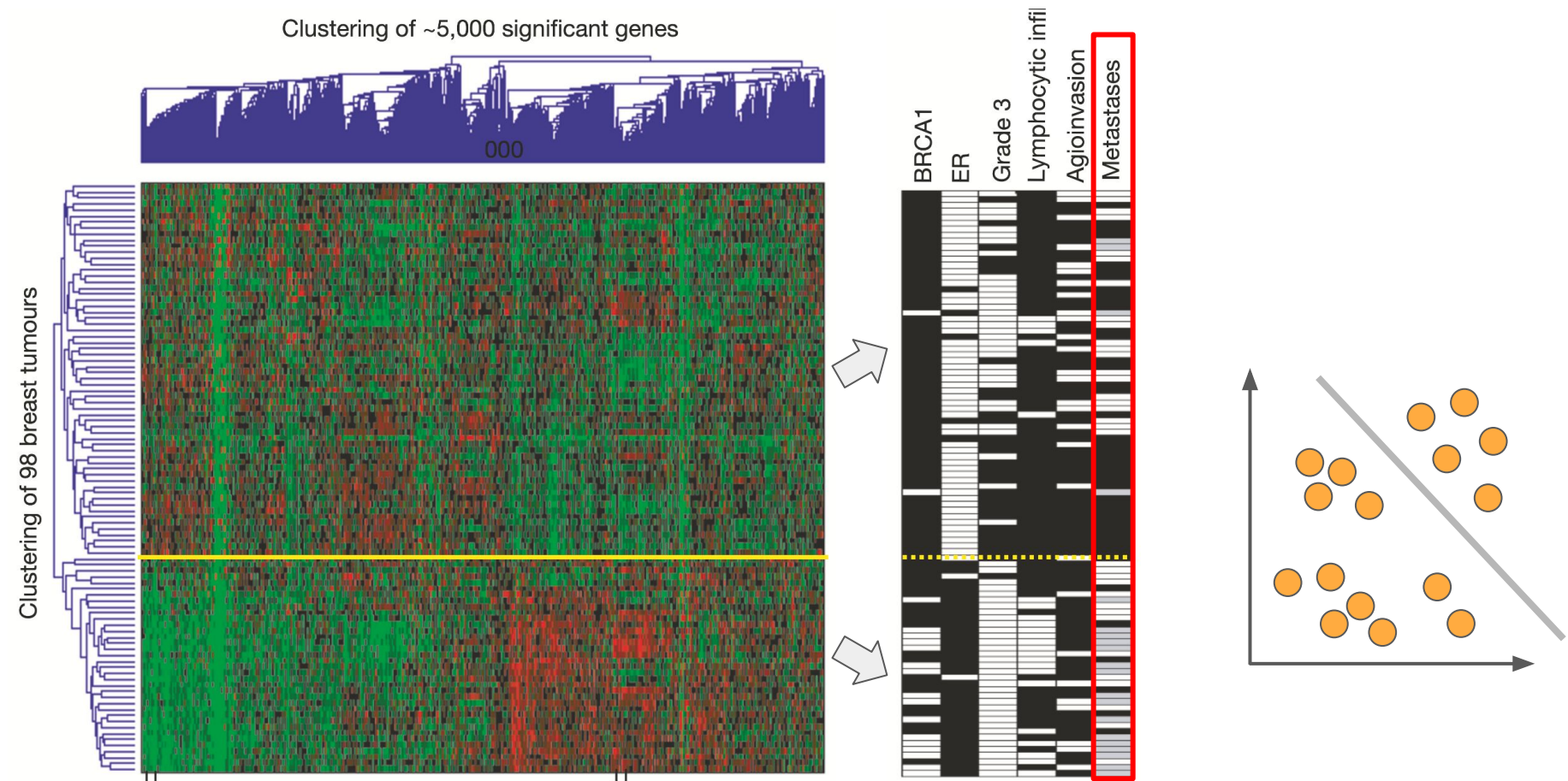
In real life, data is often:

- Unlabeled
- High-dimensional
- Unorganized (missing/errors)
- Unfamiliar
- Unexpected



Previously, we predicted metastases presence

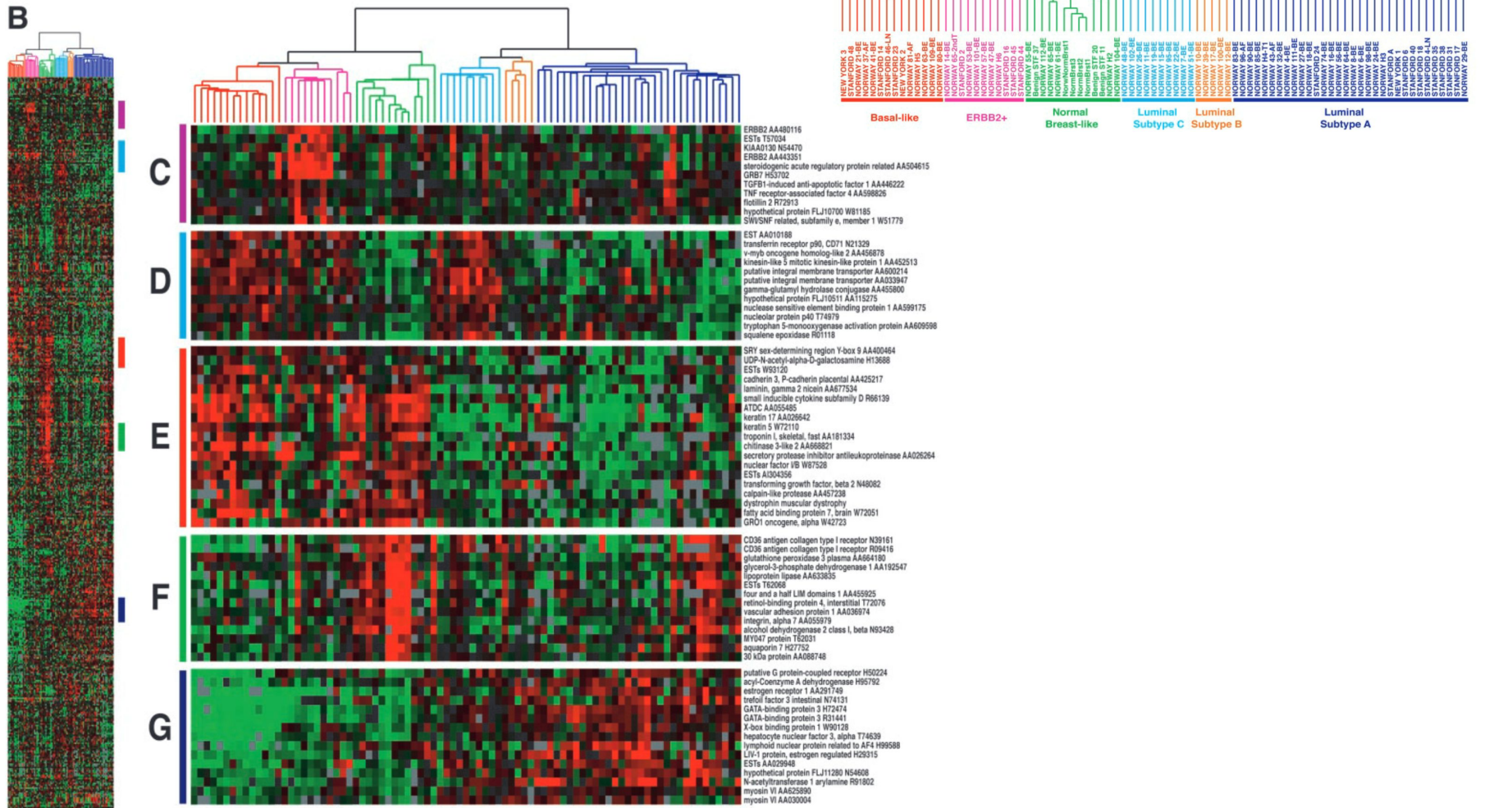
Sub-types? Different treatment? Prognosis?



Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørli^{a,b,c}, Charles M. Perou^{a,d}, Robert Tibshirani^e, Turid Aas^f, Stephanie Geisler^g, Hilde Johnsen^b, Trevor Hastie^e, Michael B. Eisen^h, Matt van de Rijnⁱ, Stefanie S. Jeffrey^j, Thor Thorsen^k, Hanne Quist^l, John C. Matese^c, Patrick O. Brown^m, David Botstein^c, Per Eystein Lønning^g, and Anne-Lise Børresen-Dale^{b,n}

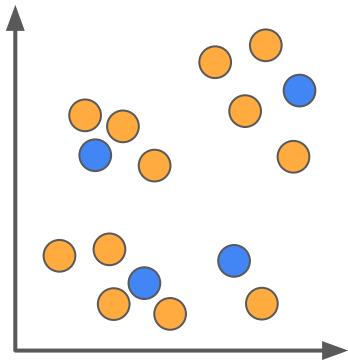
Departments of ^bGenetics and ^cSurgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; ^dDepartment of Genetics and Lineberg Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of ^eHealth Research and Policy and Statistics, ^fGenetics, ^gPathology, ^hSurgery, and ⁱBiochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of ^jMedicine (Section of Oncology), ^kSurgery, and ^lBiochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and ^mLife Sciences Division, Lawrence Orlando Berkeley National Laboratories, and Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720



Unsupervised learning

Clustering allows:

- Grouping
- Qualitative - find archetypes
- Quantitative 1 - how many flavors
- Quantitative 2 - common vs rare/exceptional cases
- Axes by which sample vary

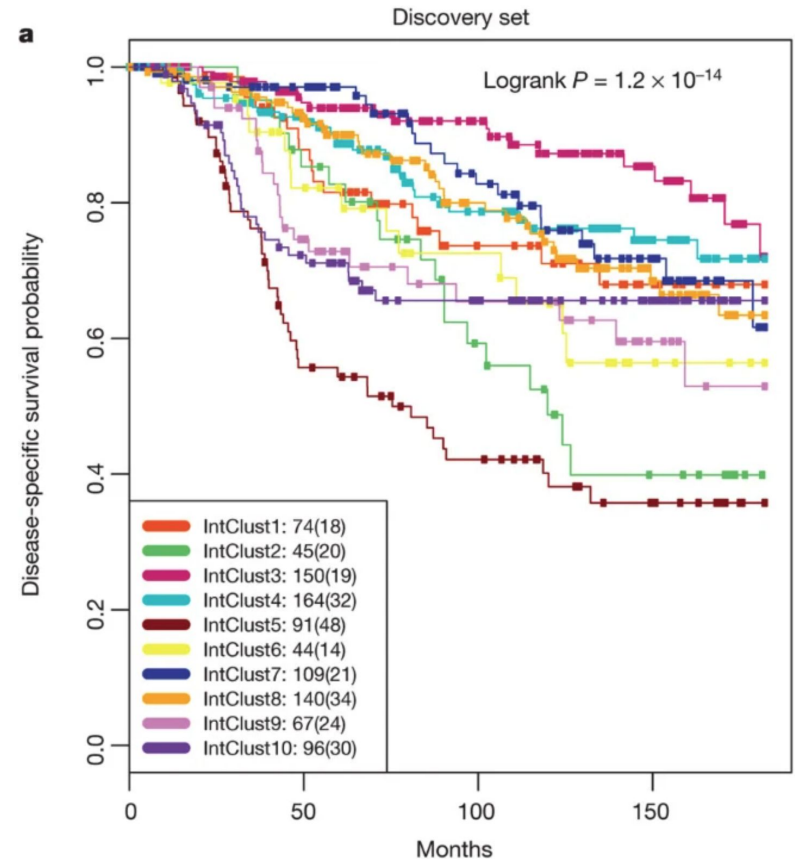


Clustering in medicine / medical research

- Group patients by sub-types
- Single-cell data - better understanding of disease-associated cells
(qualitative, quantitative, cellular/disease dynamics)
- Aging / neurodegenerative examples
- Metagenomics assembly

Christina Curtis^{1,2,*}, Sohrab P. Shah^{3,4,*}, Suet-Feung Chin^{1,2*}, Gulisa Turashvili^{3,4*}, Oscar M. Rueda^{1,2}, Mark J. Dunning², Doug Speed^{2,5,*}, Andy G. Lynch^{1,2}, Shamith Samarajiwala^{1,2}, Yinyin Yuan^{1,2}, Stefan Gräßl^{1,2}, Gavin Ha³, Gholamreza Haffari³, Ali Bashashati³, Roslin Russell², Steven McKinney^{3,4}, METABRIC Group[†], Anita Langerød⁶, Andrew Green⁷, Elena Provenzano⁸, Gordon Wishart⁸, Sarah Pinder⁹, Peter Watson^{3,4,10}, Florian Markowetz^{1,2}, Leigh Murphy¹⁰, Ian Ellis⁷, Arnie Purushotham^{9,11}, Anne-Lise Børresen-Dale^{6,12}, James D. Brenton^{2,13}, Simon Tavaré^{1,2,5,14}, Carlos Caldas^{1,2,8,13} & Samuel Aparicio^{3,4}

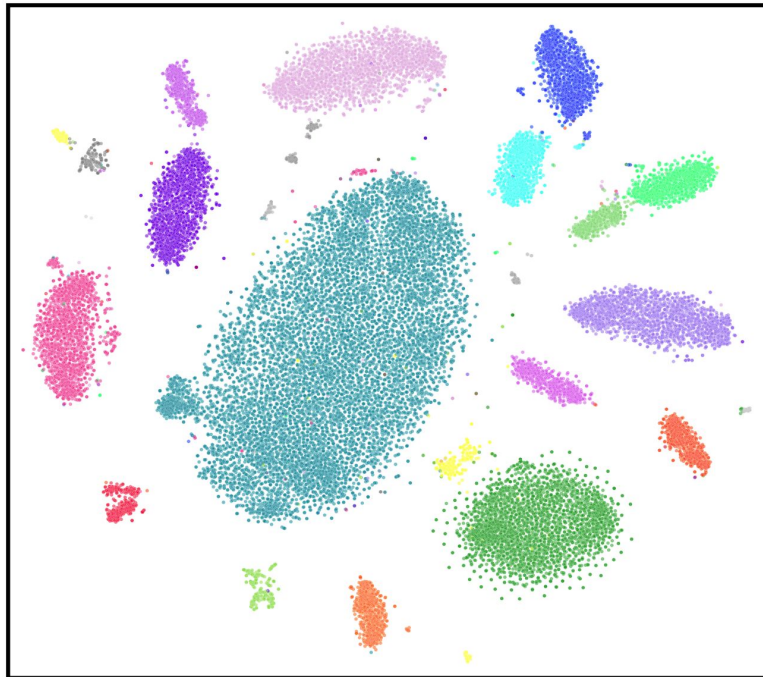
Christina Curtis^{1,2,*}, Sohrab P. Shah^{3,4,*}, Suet-Feung Chin^{1,2*}, Gulisa Turashvili^{3,4*}, Oscar M. Rueda^{1,2}, Mark J. Dunning², Doug Speed^{2,5,*}, Andy G. Lynch^{1,2}, Shamith Samarajiwala^{1,2}, Yinyin Yuan^{1,2}, Stefan Gräßl^{1,2}, Gavin Ha³, Gholamreza Haffari³, Ali Bashashati³, Roslin Russell², Steven McKinney^{3,4}, METABRIC Group[†], Anita Langerød⁶, Andrew Green⁷, Elena Provenzano⁸, Gordon Wishart⁸, Sarah Pinder⁹, Peter Watson^{3,4,10}, Florian Markowetz^{1,2}, Leigh Murphy¹⁰, Ian Ellis⁷, Arnie Purushotham^{9,11}, Anne-Lise Børresen-Dale^{6,12}, James D. Brenton^{2,13}, Simon Tavaré^{1,2,5,14}, Carlos Caldas^{1,2,8,13} & Samuel Aparicio^{3,4}



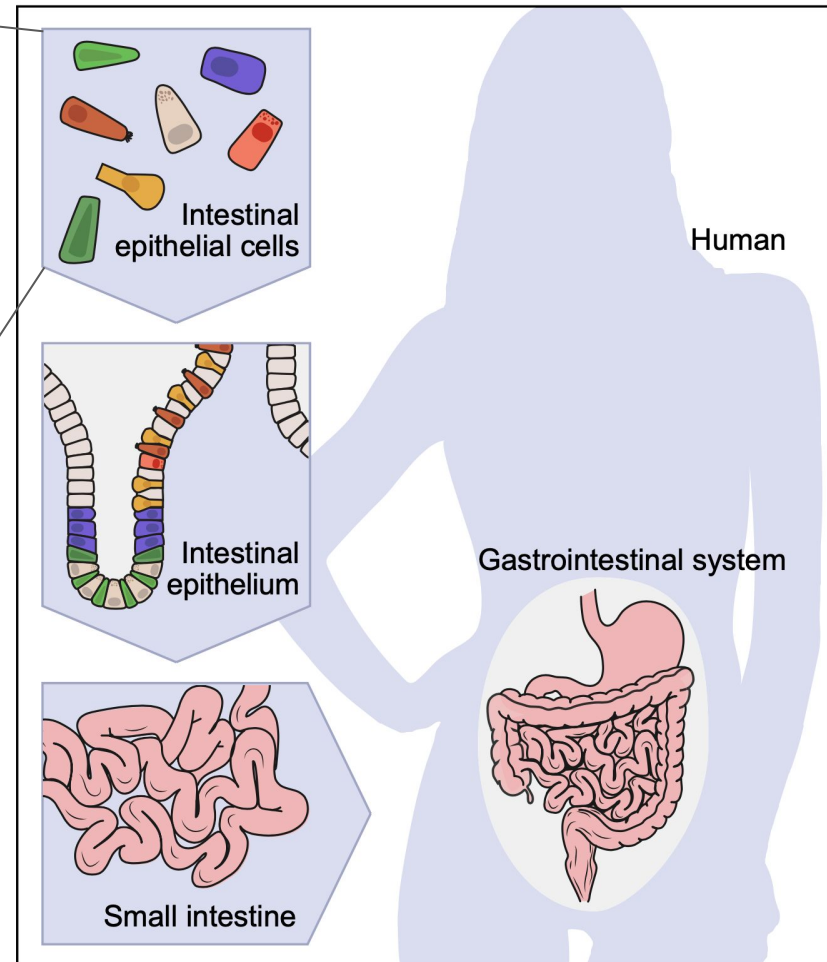
SCIENCE FORUM

The Human Cell Atlas

AVIV REGEV*, SARAH A TEICHMANN*, ERIC S LANDER*



Clustered single-cell gene expression



Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics

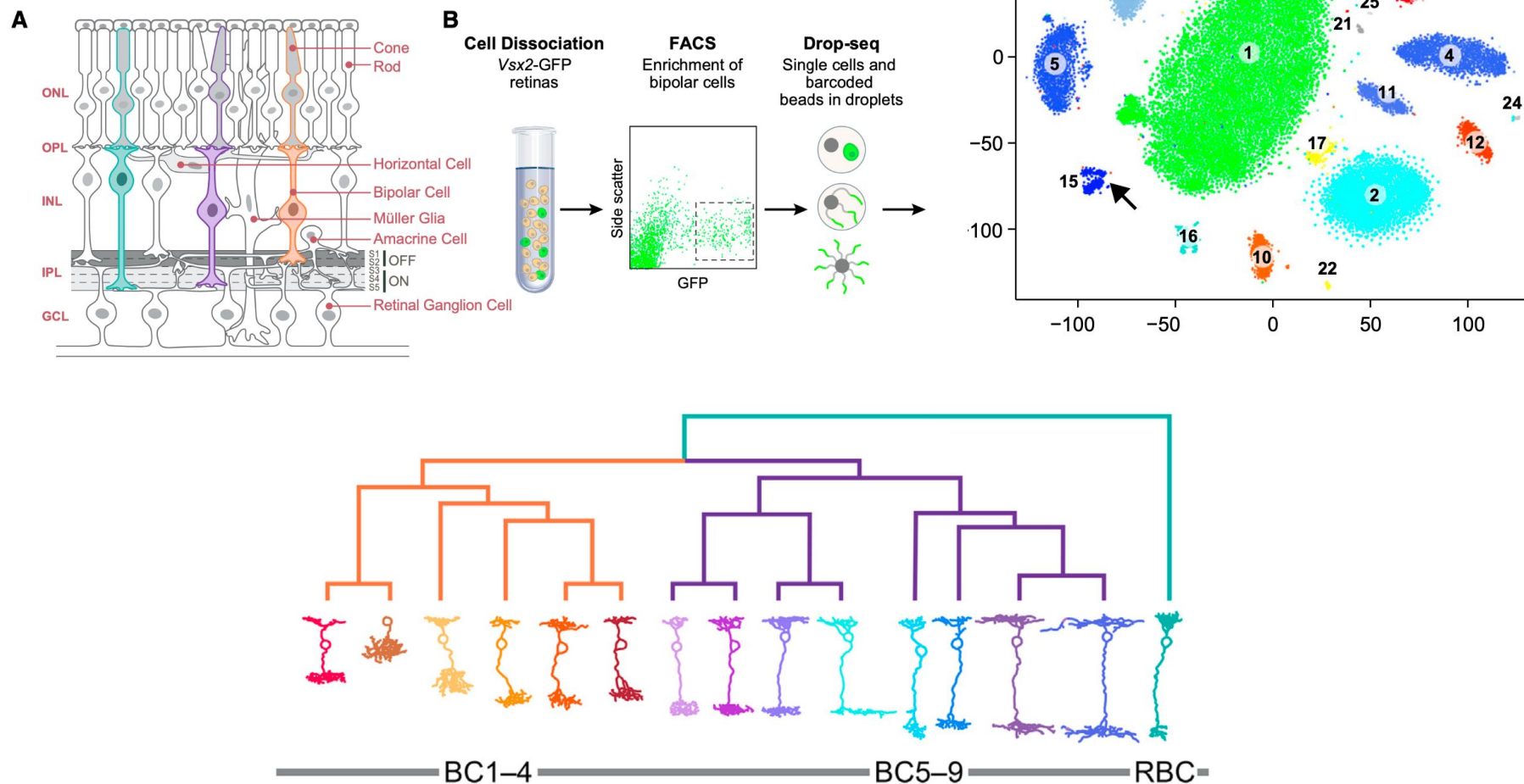
Karthik Shekhar,^{1,9} Sylvain W. Lapan,^{2,7,9} Irene E. Whitney,^{4,9} Nicholas M. Tran,⁴ Evan Z. Macosko,^{2,5,6} Monika Kowalczyk,¹ Xian Adiconis,^{1,5} Joshua Z. Levin,^{1,5} James Nemesh,^{2,5,6} Melissa Goldman,^{2,5} Steven A. McCarroll,^{2,5,6} Constance L. Cepko,^{2,3,7,*} Aviv Regev,^{1,7,8,*} and Joshua R. Sanes^{4,10,*}

¹Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

²Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

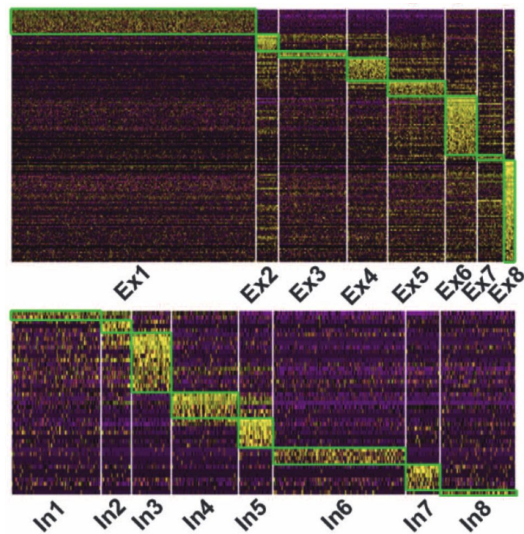
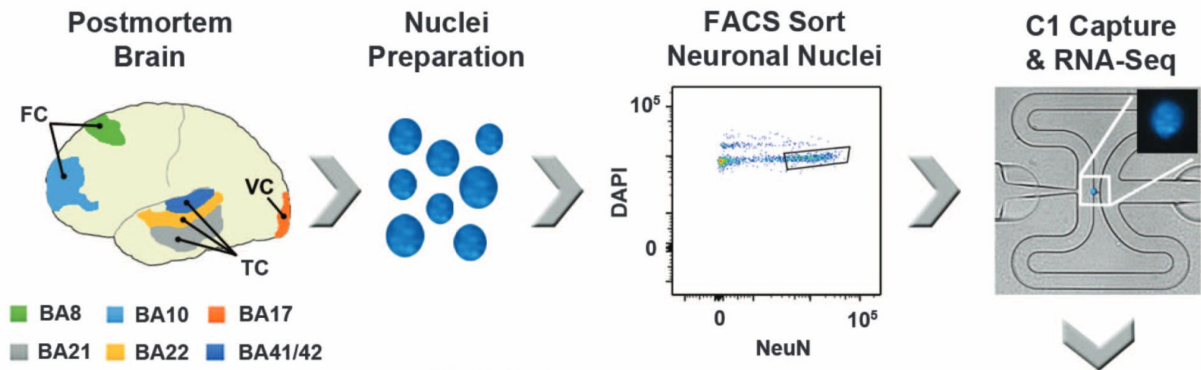
³Department of Ophthalmology, Harvard Medical School, Boston, MA 02115, USA

⁴Center for Brain Science and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02130, USA

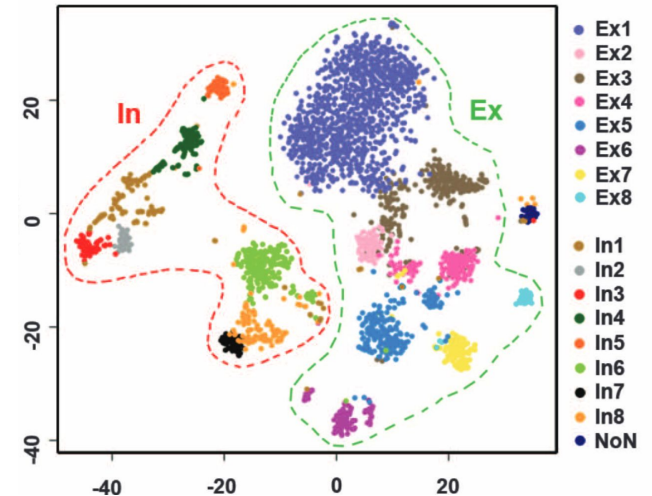


Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain

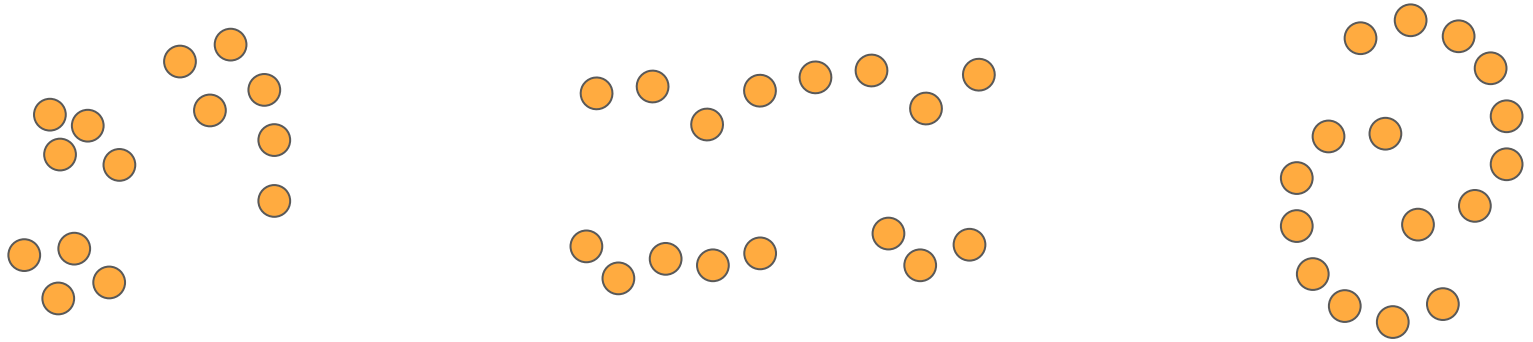
Blue B. Lake,^{1*} Rizi Ai,^{2*} Gwendolyn E. Kaeser,^{3,4*} Neeraj S. Salathia,^{5*} Yun C. Yung,³ Rui Liu,¹ Andre Wildberg,² Derek Gao,¹ Ho-Lim Fung,¹ Song Chen,¹ Raakhee Vijayaraghavan,⁵ Julian Wong,³ Allison Chen,³ Xiaoyan Sheng,³ Fiona Kaper,⁵ Richard Shen,⁵ Mostafa Ronaghi,⁵ Jian-Bing Fan,^{5†} Wei Wang,^{2†} Jerold Chun,^{3†} Kun Zhang^{1†}



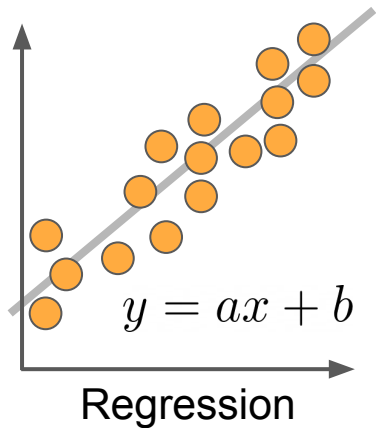
Neuronal Subpopulations



People are good in low-dimensional clustering

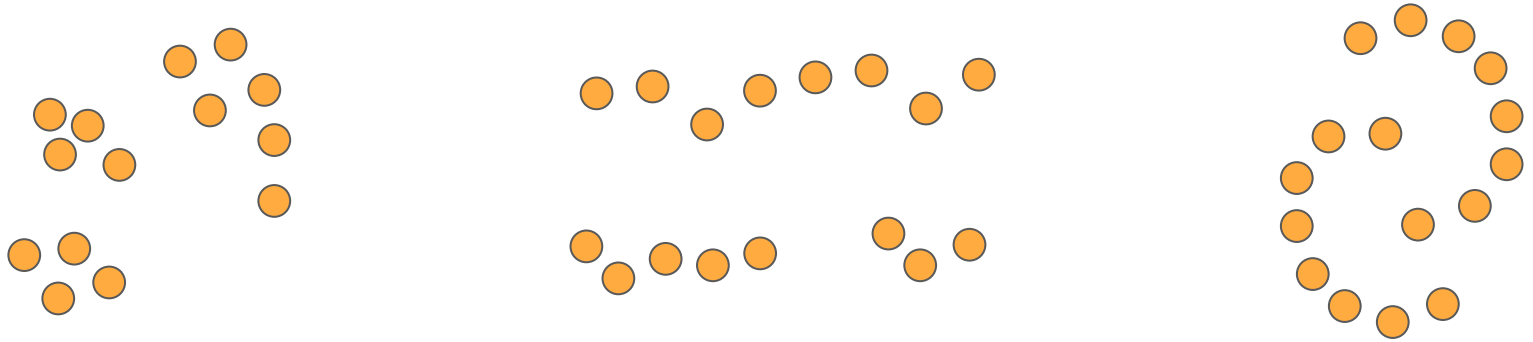


But how would we formally define a good clustering?

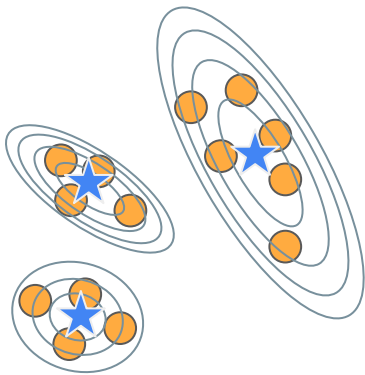


$$\arg \min_{a,b} \sum (y_i - (ax_i + b))^2$$

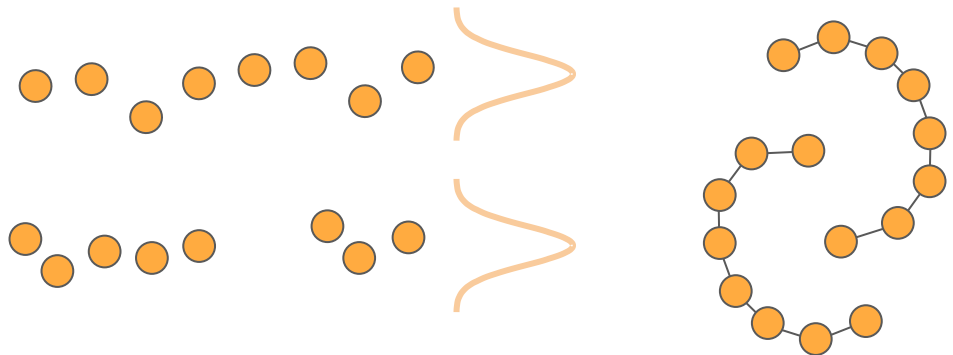
People are good in low-dimensional clustering



But how would we formally define a good clustering?



Top-down approach

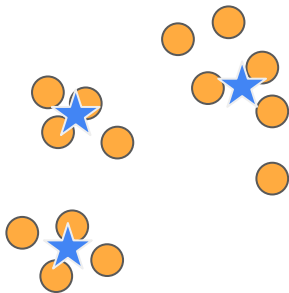


Bottom-up approach

Formal definition - top down

- Density-based clustering
- Minimize distances from “centroid”
- Each point x “belongs” to class C_i whose center is at μ_i

$$\arg \min_{C, \mu} \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2$$



How to find C, μ ?

K-means

Developed by Lloyd, 1957

Init:

Choose at random K data points, as centroids

Loop:

1. Re-assign each point to nearest centroid
2. Move each centroid to mean (or “center of mass”) of assigned data points

Stop:

Upon convergence (no assignment changes)

K-means

- Stochasticity

Init:

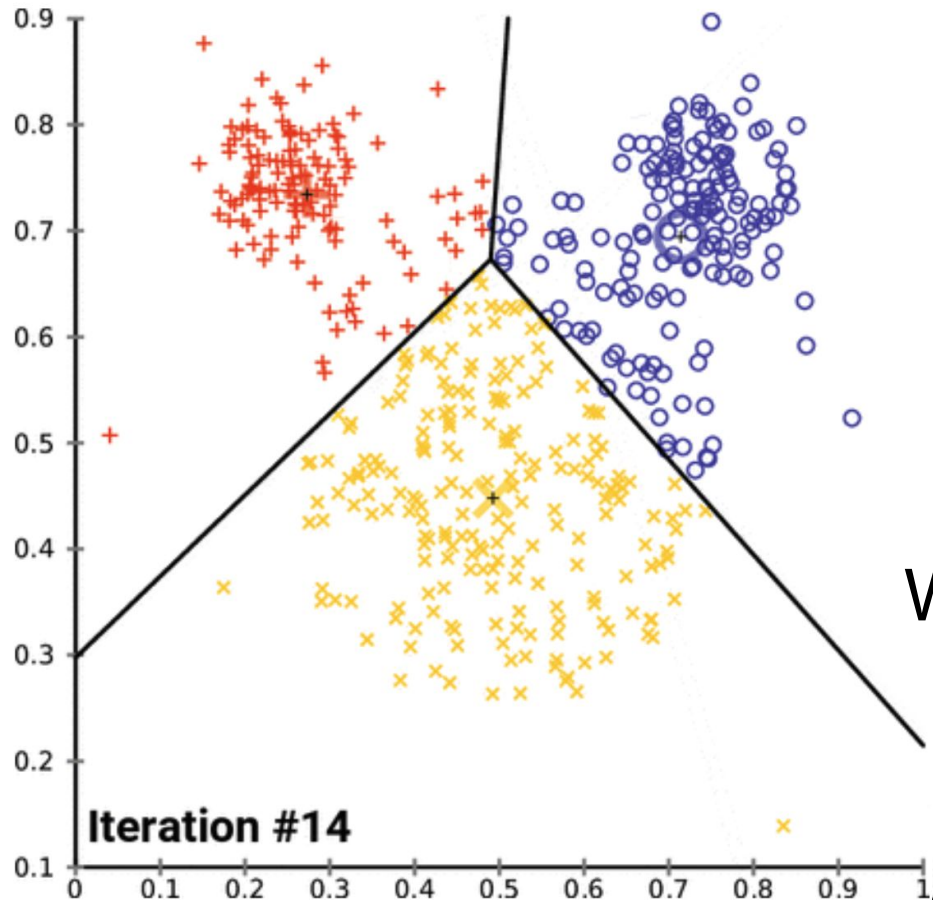
Choose at random K data points, as centroids

Loop:

1. Re-assign each point to nearest centroid
2. Move each centroid to mean (or “center of mass”) of assigned data points

Stop:

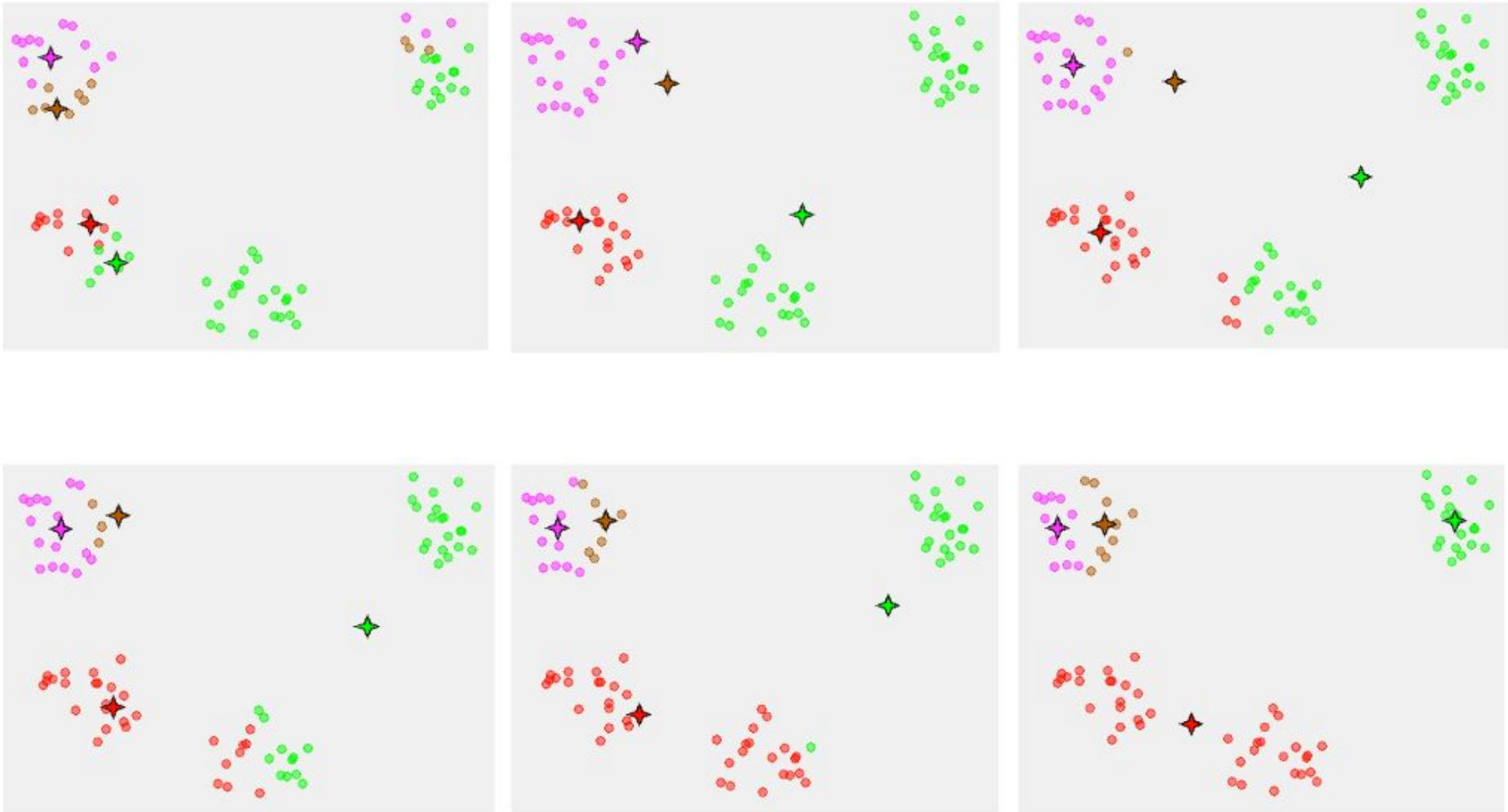
Upon convergence (no assignment changes)



Would you cluster differently?

K-means

- Stochasticity - might not converge to optimal solution



K-means

- Stochasticity
- Random restarts
- Initialization
 - K-means++
 - Cluster subset of data
- Running time
 - Each iteration: $\mathcal{O}(KN)$
 - But how many iterations are typically needed?
- Possible data transformations
 - Feature selection
 - Data transformation

Init:

Choose at random K data points, as centroids

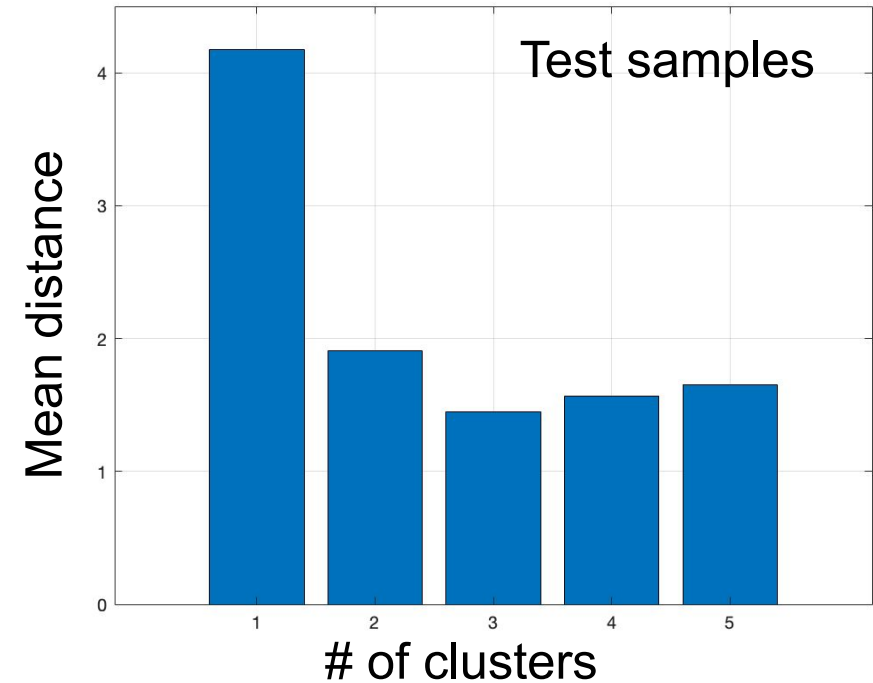
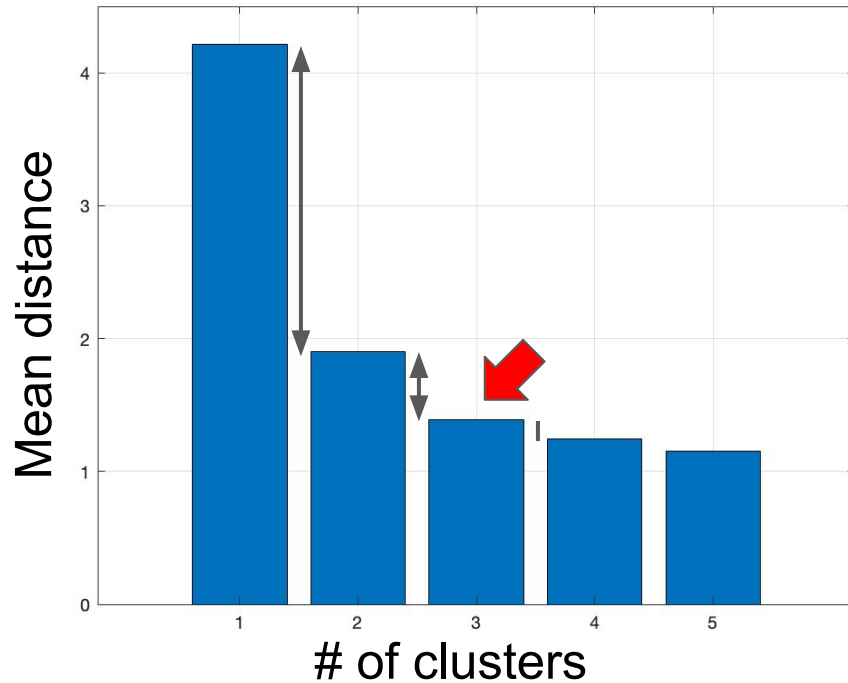
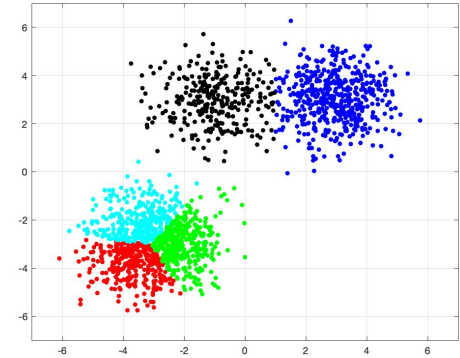
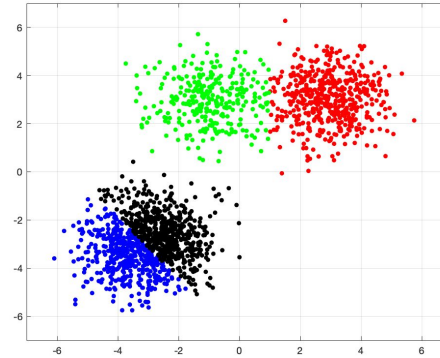
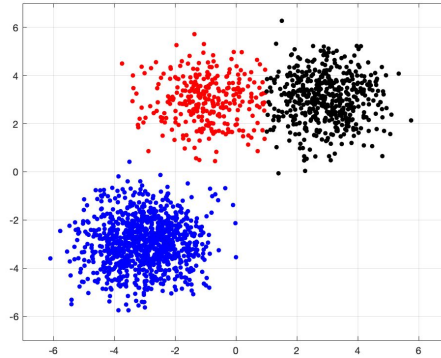
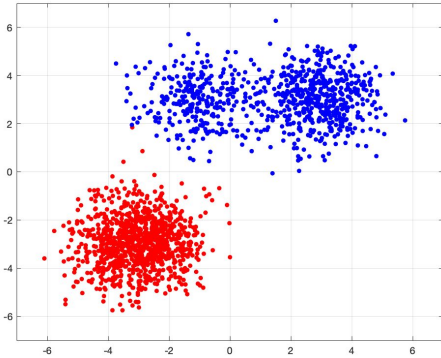
Loop:

1. Re-assign each point to nearest centroid
2. Move each centroid to mean (or “center of mass”) of assigned data points

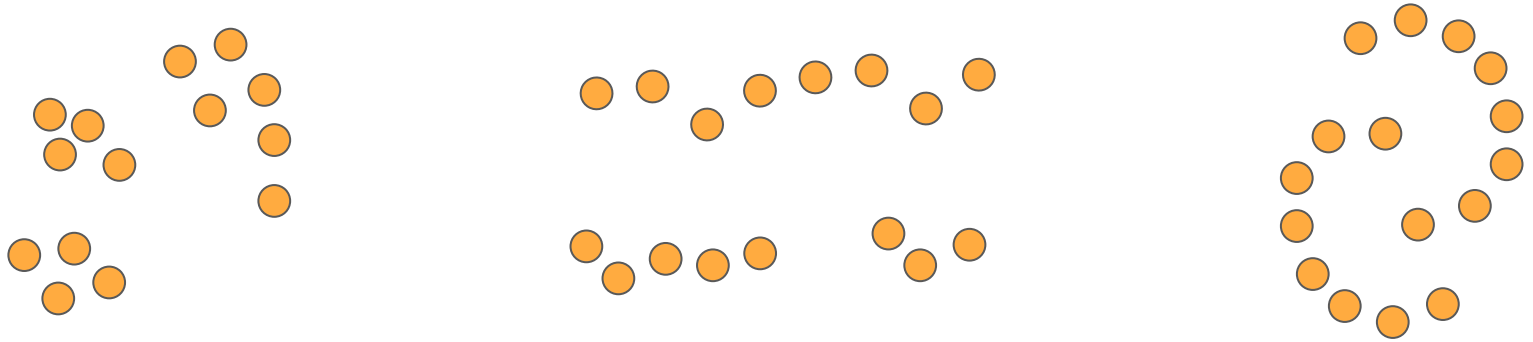
Stop:

Upon convergence (no assignment changes)

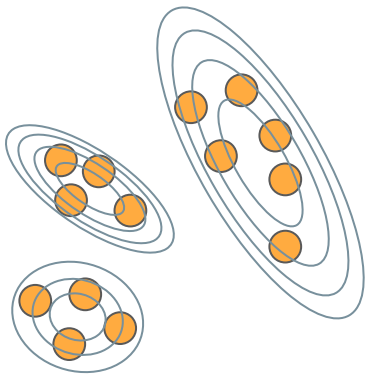
Choosing K using the Elbow method



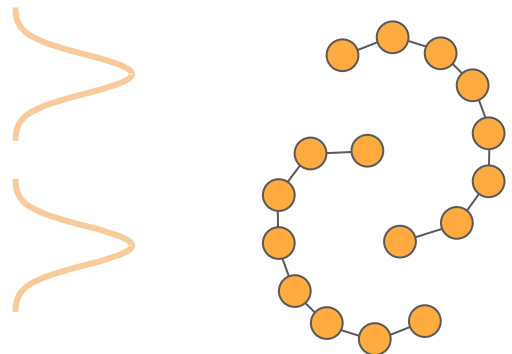
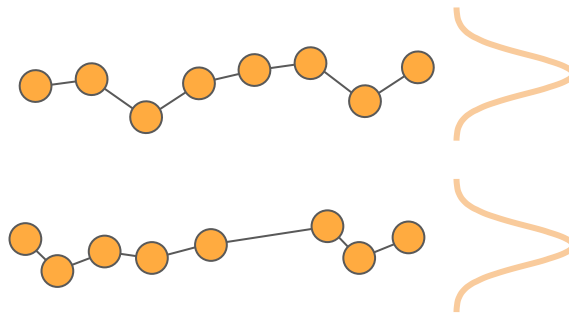
People are good in low-dimensional clustering



But how would we formally define a good clustering?



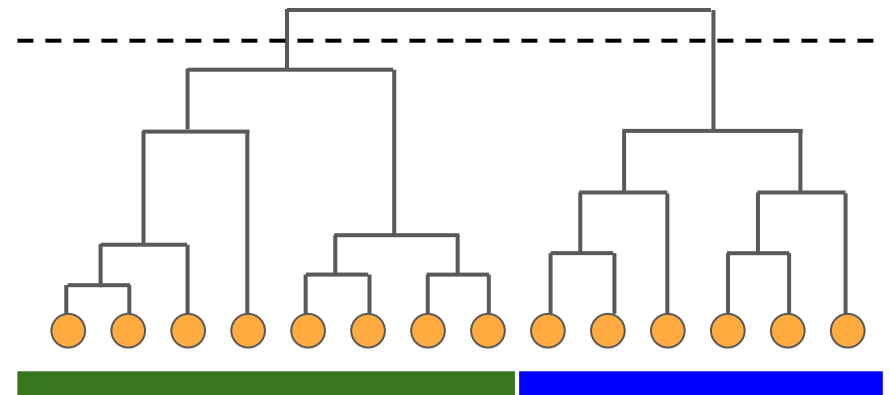
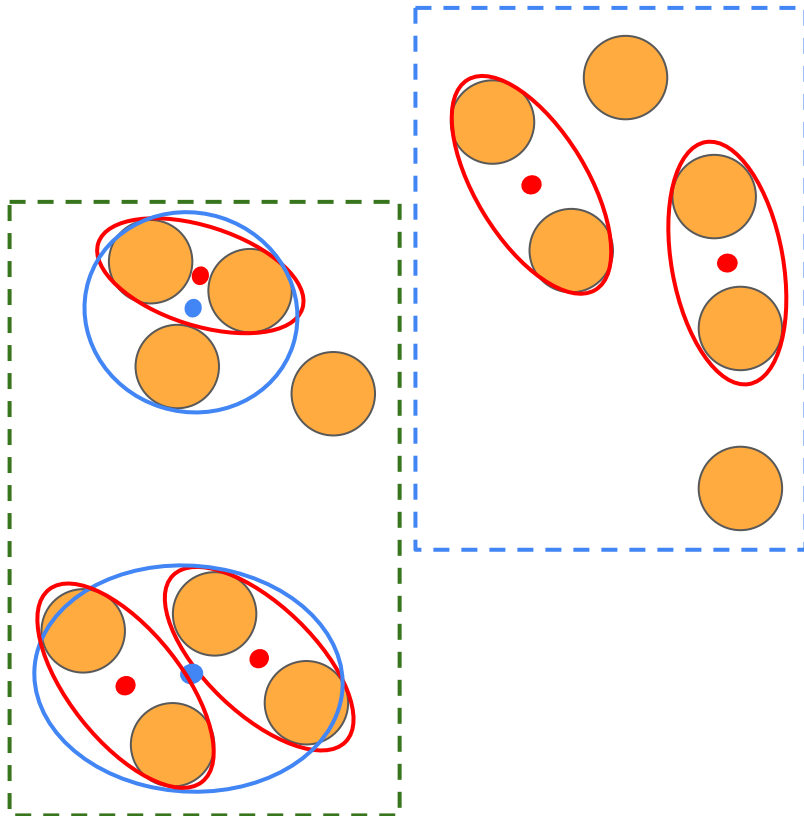
Top-down approach



Bottom-up approach

Hierarchical Agglomerative Clustering

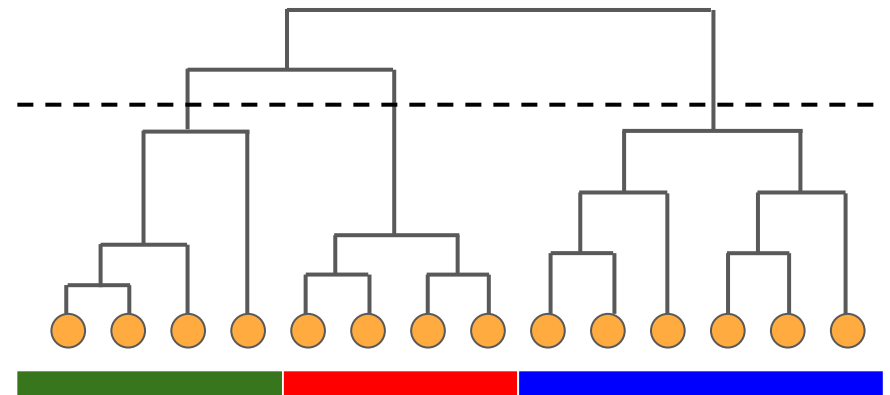
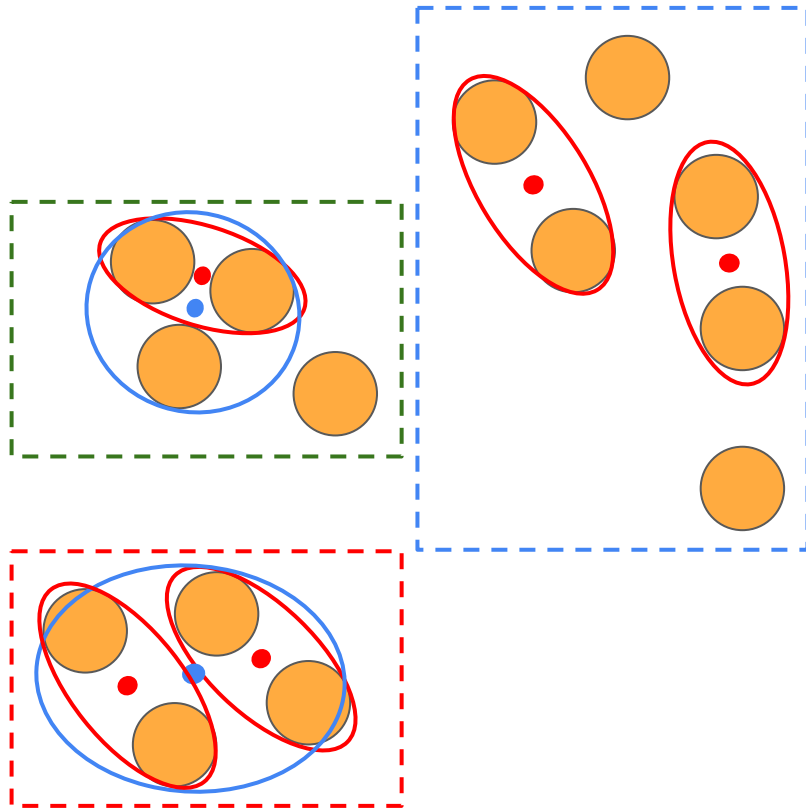
- Iteratively, merge similar points / sub-groups



dendrogram

Hierarchical Agglomerative Clustering

- Iteratively, merge similar points / sub-groups

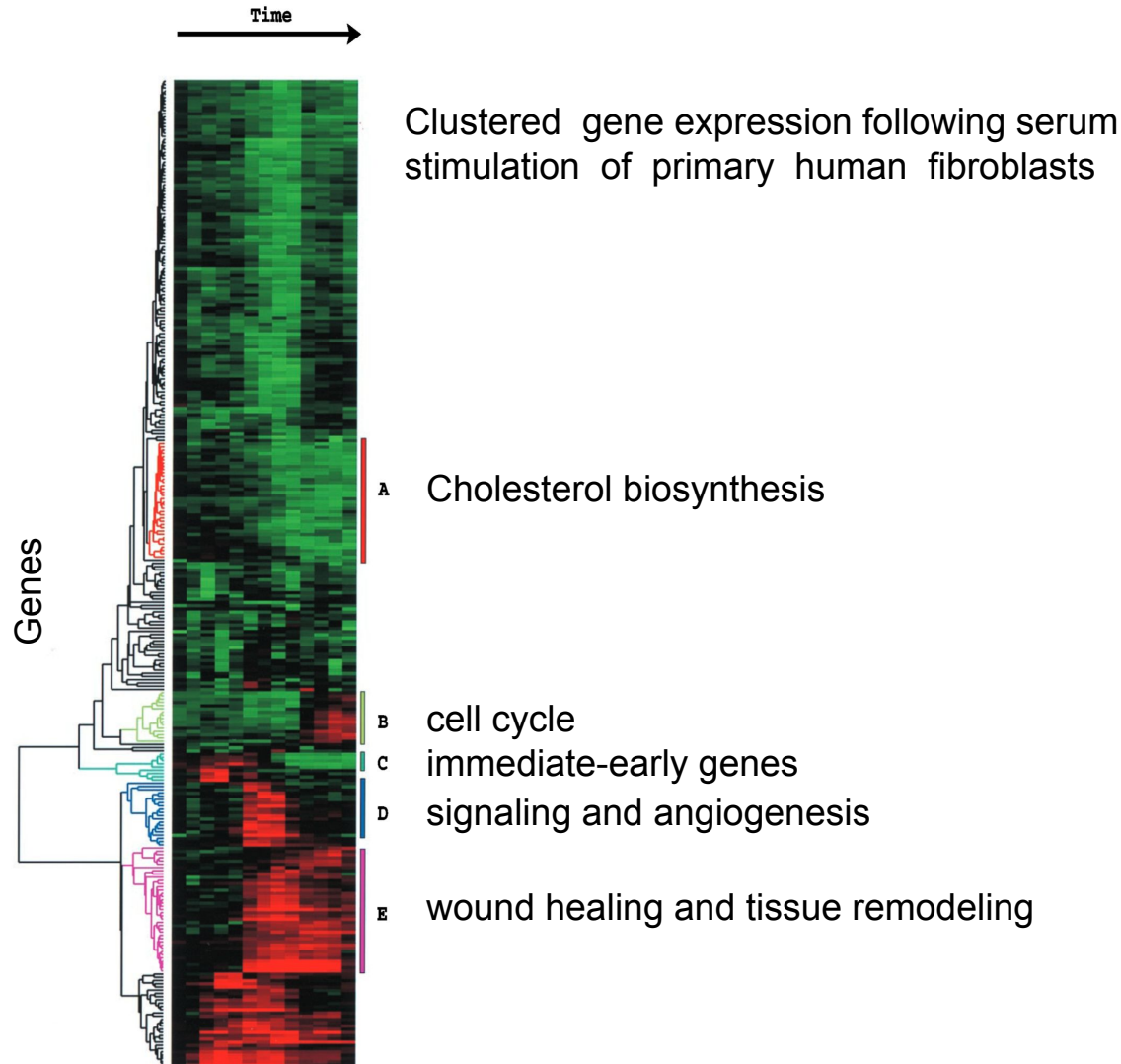


dendrogram

Cluster analysis and display of genome-wide expression patterns

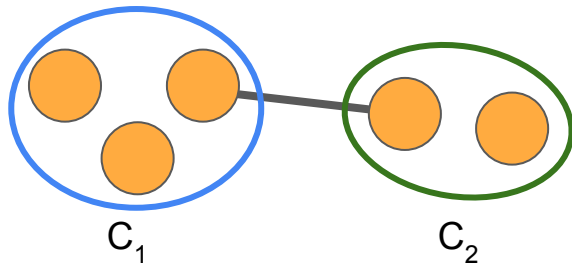
MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

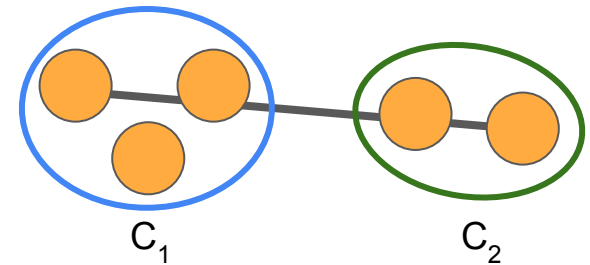


Hierarchical Agglomerative Clustering

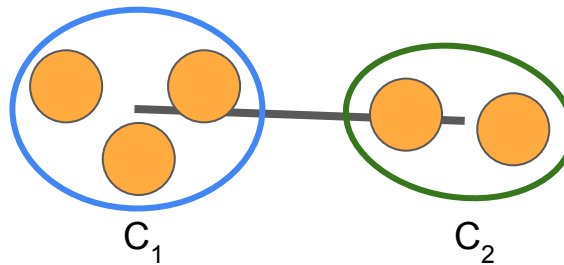
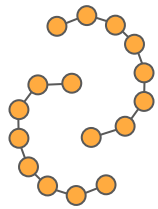
How to define distance between clusters?



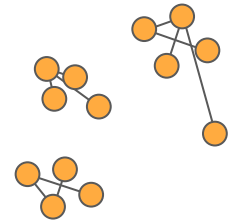
Single linkage
(closest pair)



Complete linkage
(farthest pair)



Average linkage
(mean distance)



Running time: $\mathcal{O}(N^3)$

What have we learned?

- Unsupervised data
- How to approach?
- Cluster to find typical samples (archetypes)
- Top-down (divisive) clustering
 - K-means (+ useful tricks)
- Bottom-up (agglomerative) clustering
 - Hierarchical clustering
 - Interpretation
 - Variations

Syllabus

1. Introduction
2. Classification
3. Learning 1
4. AI in ophthalmology (Prof. Itay Chowers)
5. Learning 2
6. Regression
7. Clustering
8. Visualization (and dimensionality reduction)
9. Deep learning in image analysis (Prof. Leo Joskowicz)
10. Missing data, statistical dependencies
11. Natural language in medicine (Dr. Gabi Stanovsky)
12. Decisions (utility)
13. Longitudinal Data / Project