האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

Artificial Intelligence in Medicine

Clustering

Nir Friedman and Tommy Kaplan 19/2/24

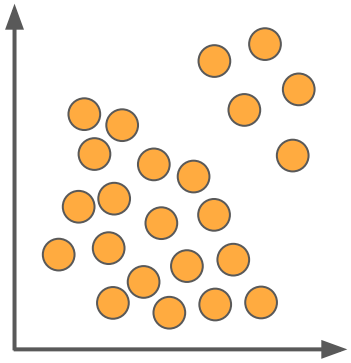
“In my defense, I was left unsupervised”

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

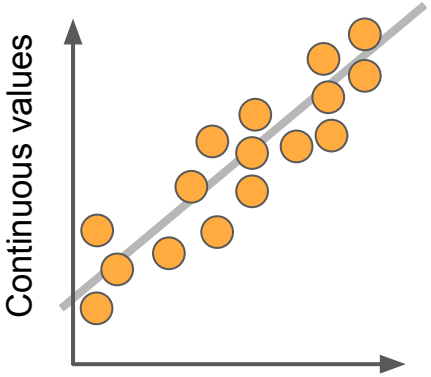
model type, parameters,
simple, general, interpretable

Model

Set of rules
for prediction

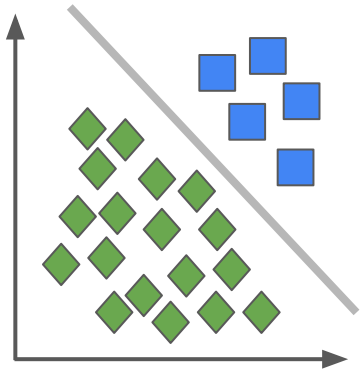
Data

$$y = ax + b$$



Regression

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x > \theta \\ 0 & \text{else} \end{cases}$$

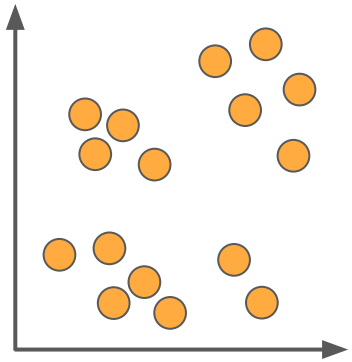


Classification

Unsupervised learning

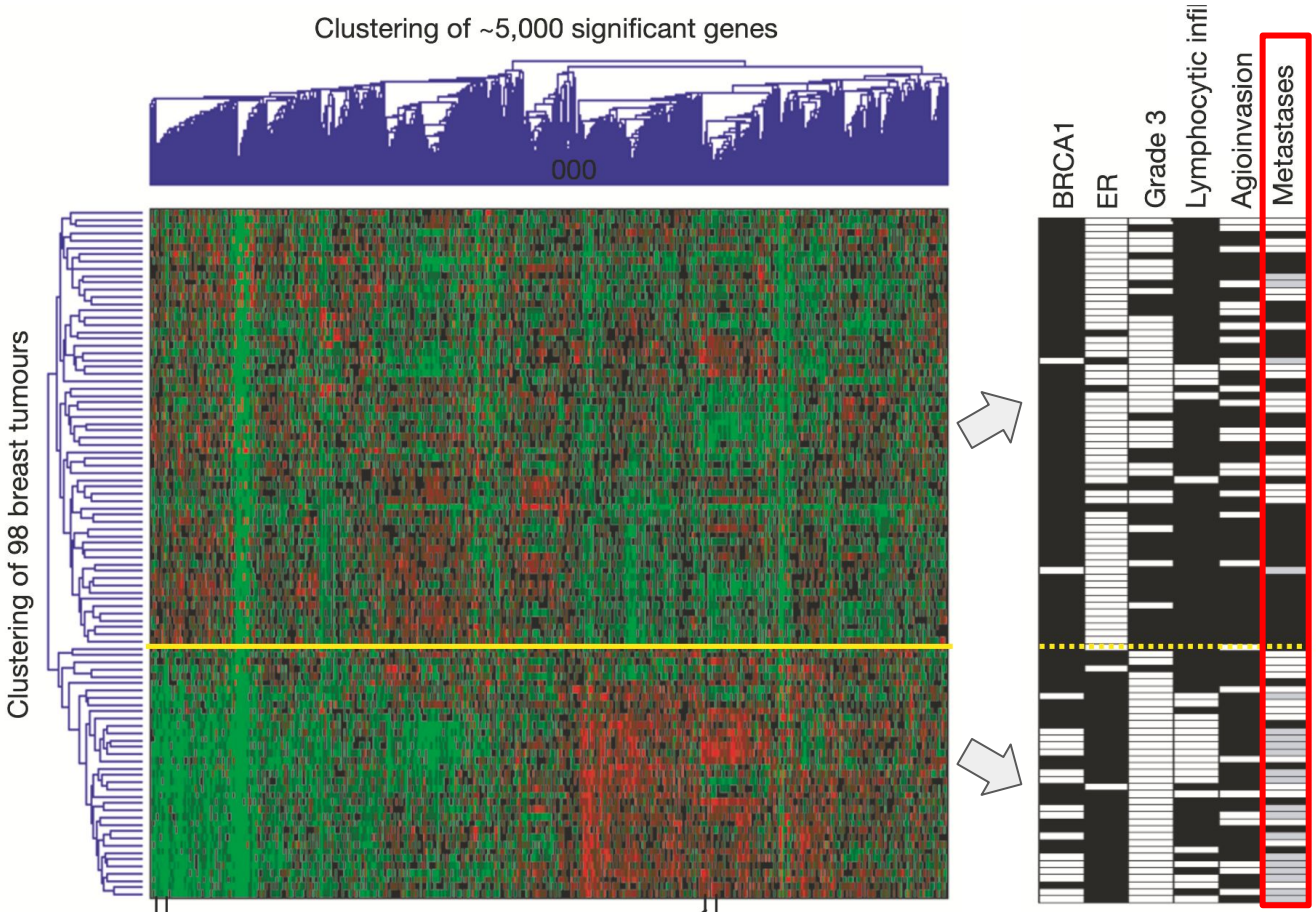
In real life, data is often:

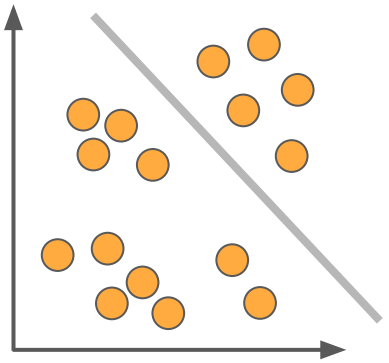
- Unlabeled
- High-dimensional
- Unorganized (missing/errors)
- Unfamiliar
- Unexpected



Previously, we predicted metastases presence

Sub-types? Different treatment? Prognosis?

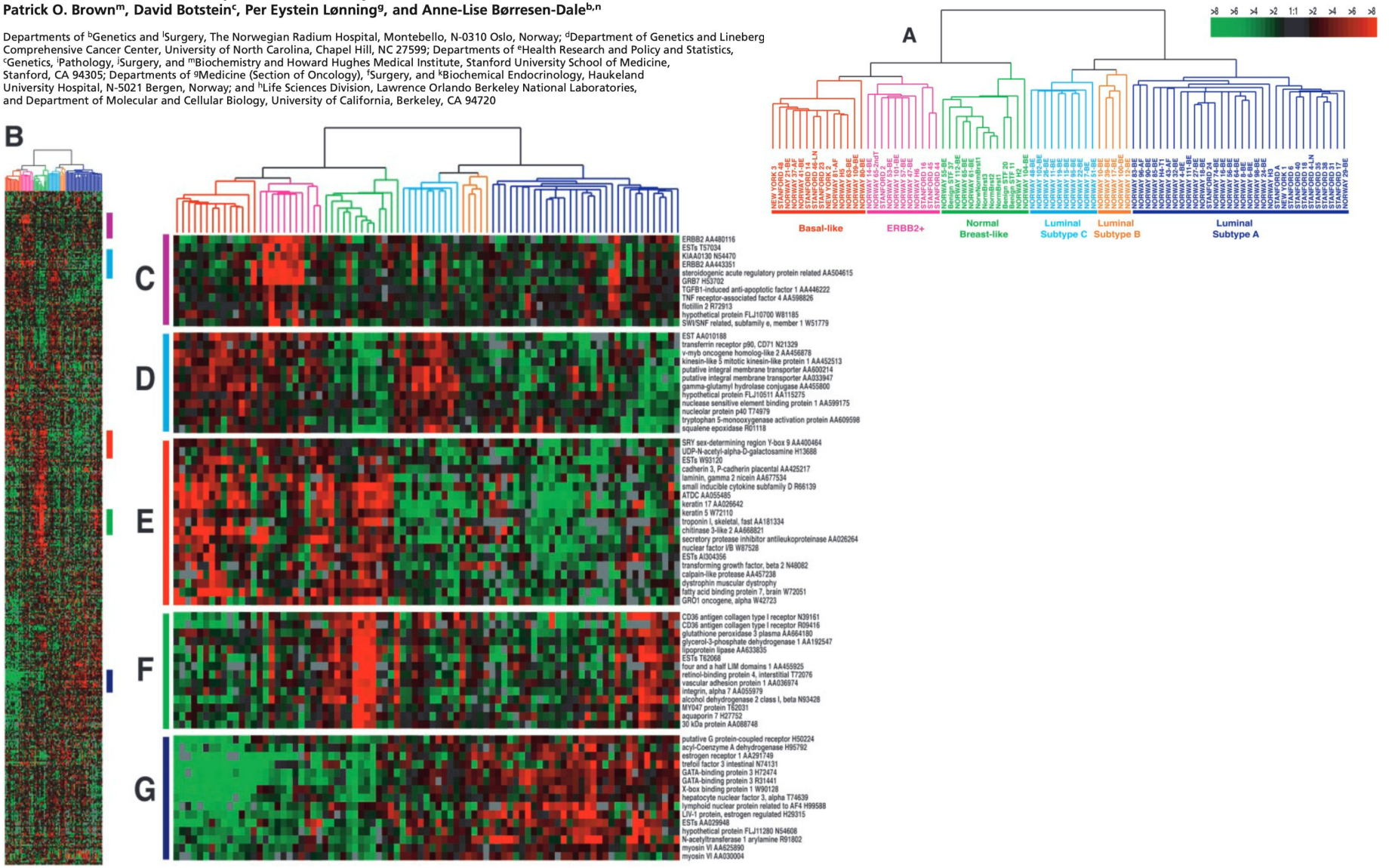




Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[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 Rijn[i], 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]

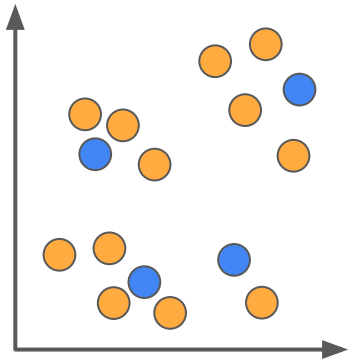
Departments of [b]Genetics and [l]Surgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; [d]Department of Genetics and Lineberg Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of [e]Health Research and Policy and Statistics, [c]Genetics, [i]Pathology, [j]Surgery, and [m]Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [g]Medicine (Section of Oncology), [f]Surgery, and [k]Biochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and [h]Life 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

ARTICLE

doi:10.1038/nature10983

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

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. Dunning[2], Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Stefan Gräf[1,2], Gavin Ha[3], Gholamreza Haffari[3], Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anita Langerød[6], Andrew Green[7], Elena Provenzano[8], Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Leigh Murphy[10], Ian Ellis[7], 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]

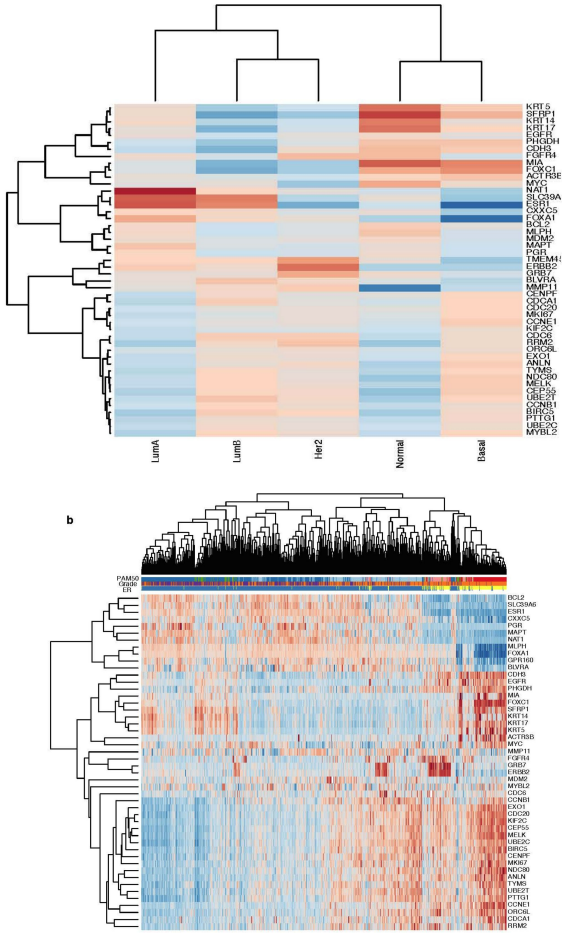
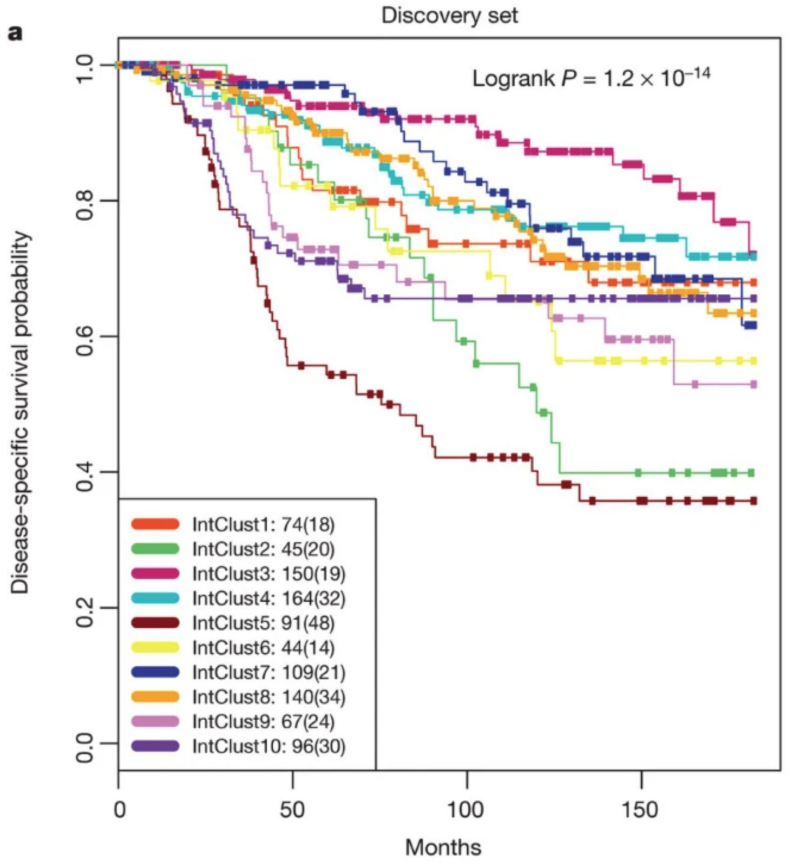


Figure 5: The integrative subgroups have distinct clinical outcomes.



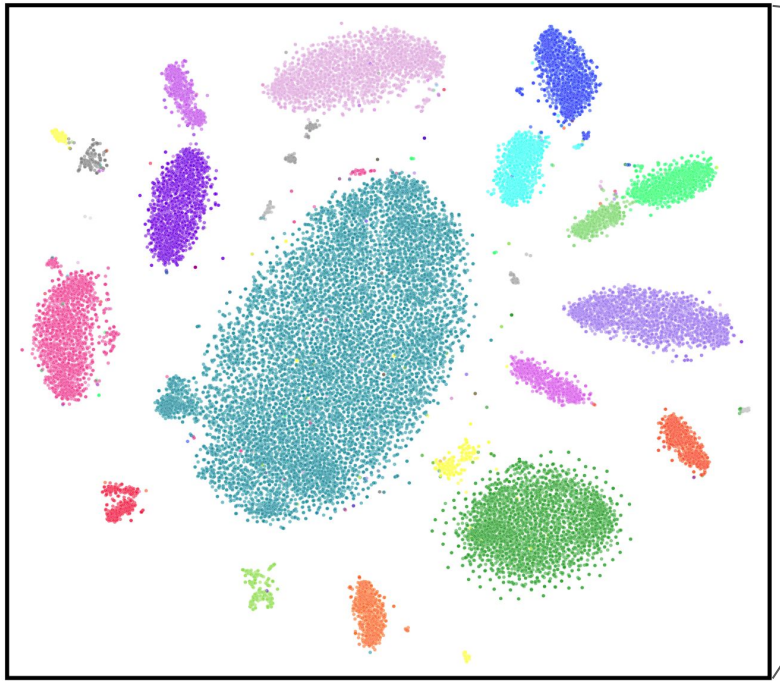
eLIFE
elifesciences.org

FEATURE ARTICLE

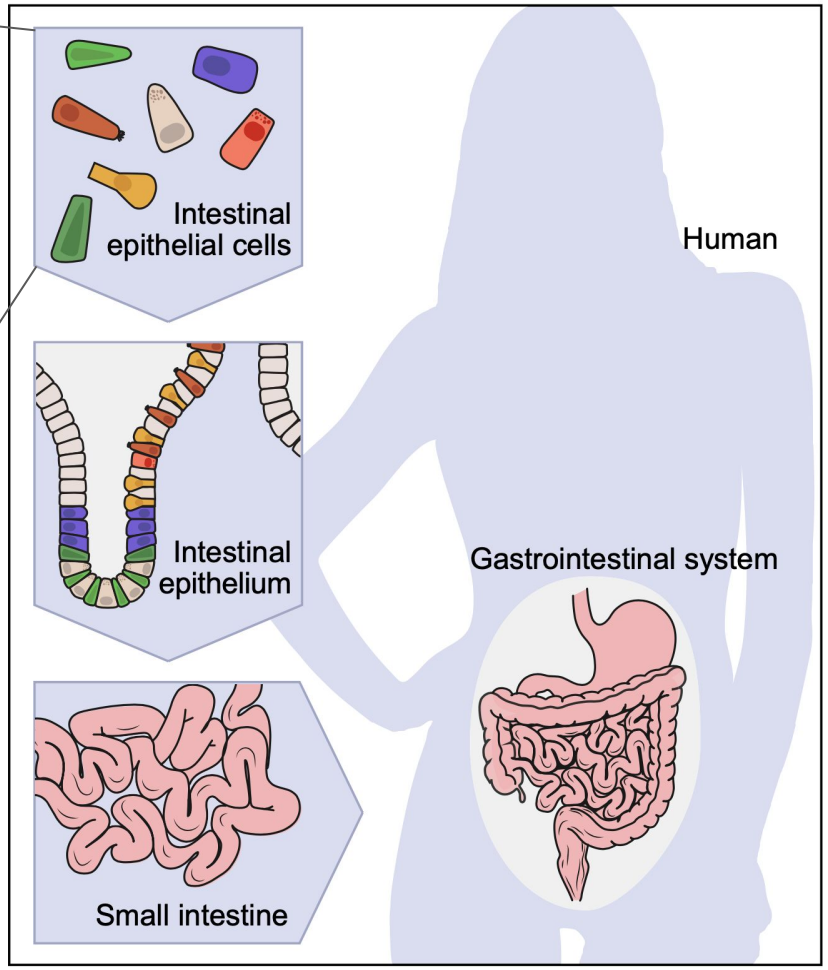
SCIENCE FORUM

The Human Cell Atlas

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



Clustered single-cell gene expression



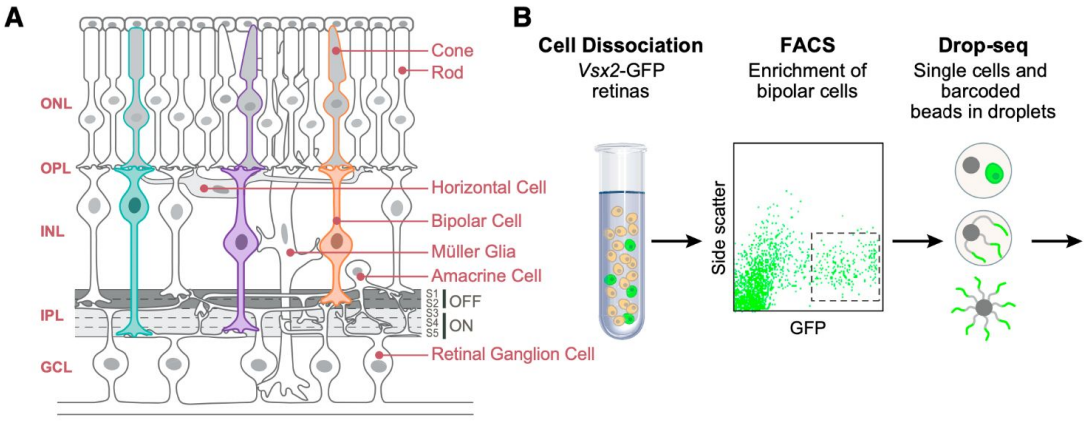
Cell

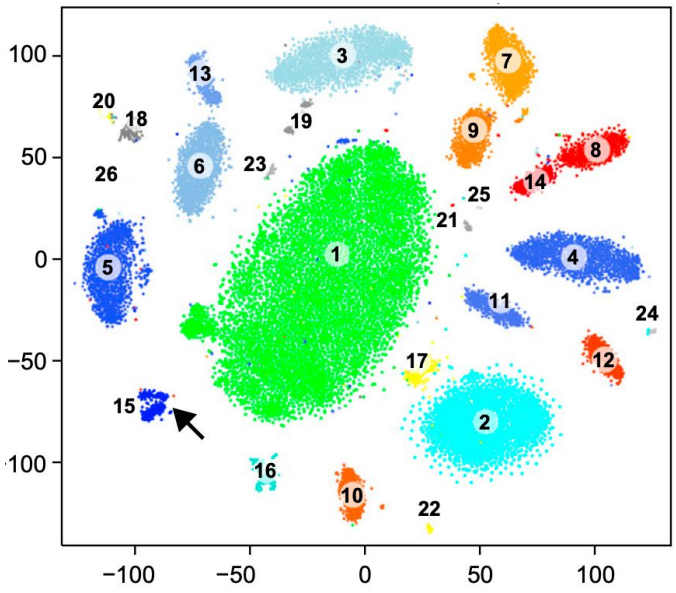
Resource

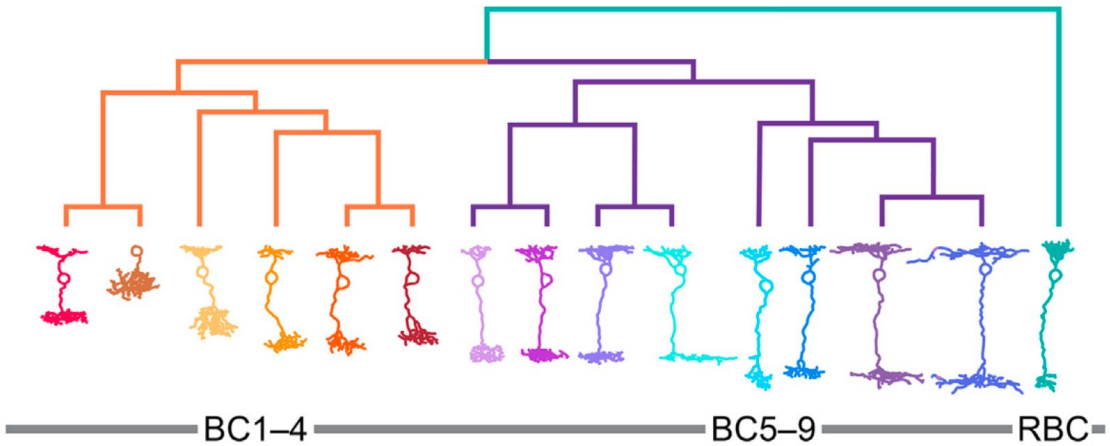
Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics

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,*]

[1]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Ophthalmology, Harvard Medical School, Boston, MA 02115, USA
[4]Center for Brain Science and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02130, USA





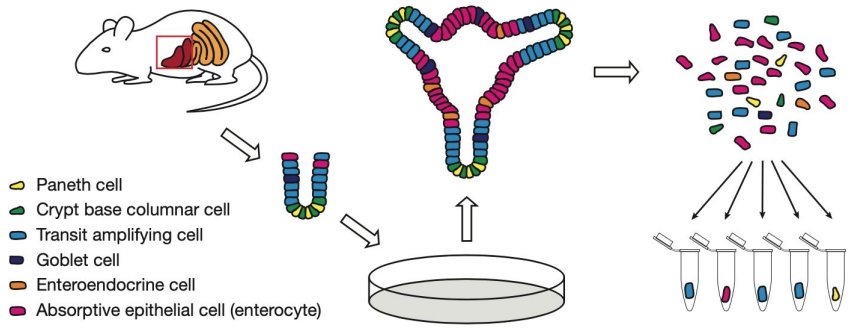


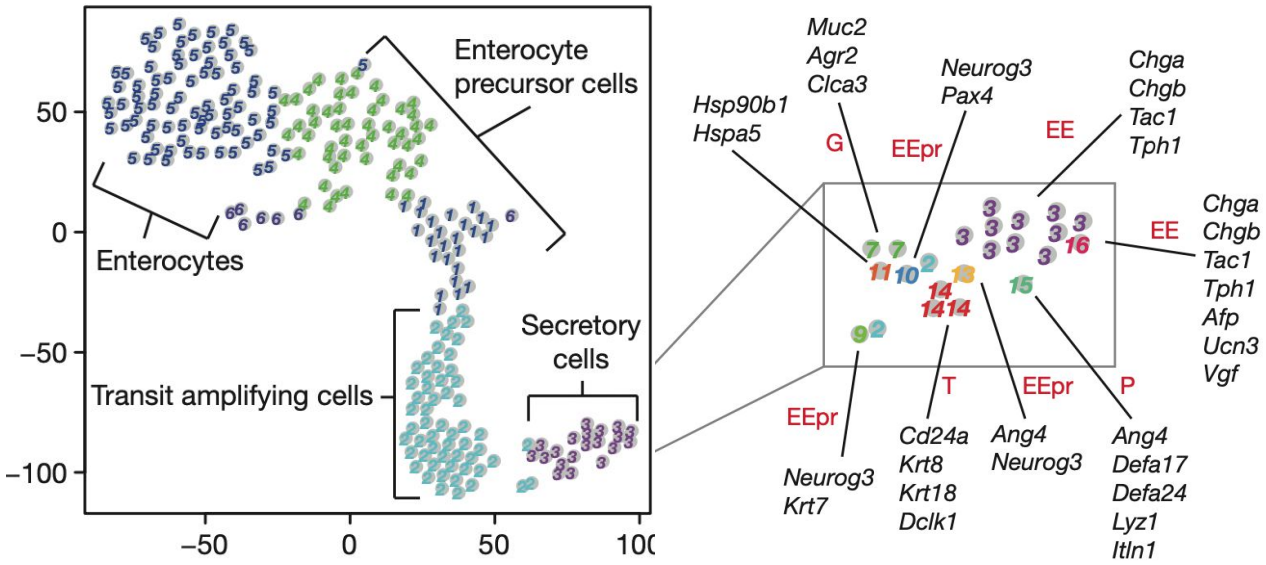
LETTER

doi:10.1038/nature14966

Single-cell messenger RNA sequencing reveals rare intestinal cell types

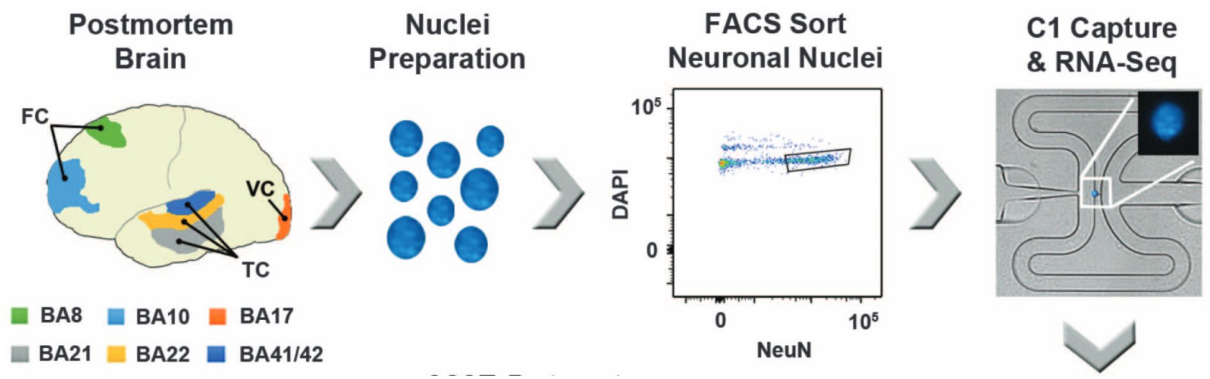
Dominic Grün[1,2]*, Anna Lyubimova[1,2]*, Lennart Kester[1,2], Kay Wiebrands[1,2], Onur Basak[1,2], Nobuo Sasaki[1,2], Hans Clevers[1,2] & Alexander van Oudenaarden[1,2]

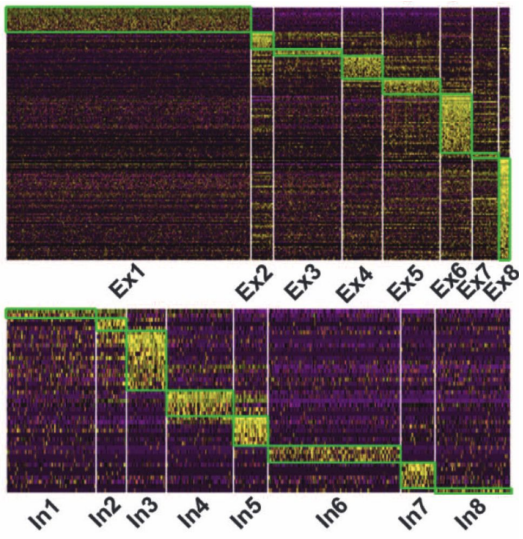


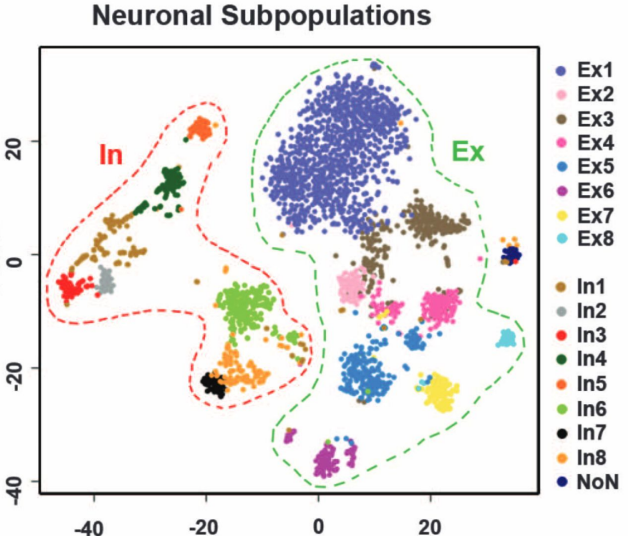


Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain

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]†





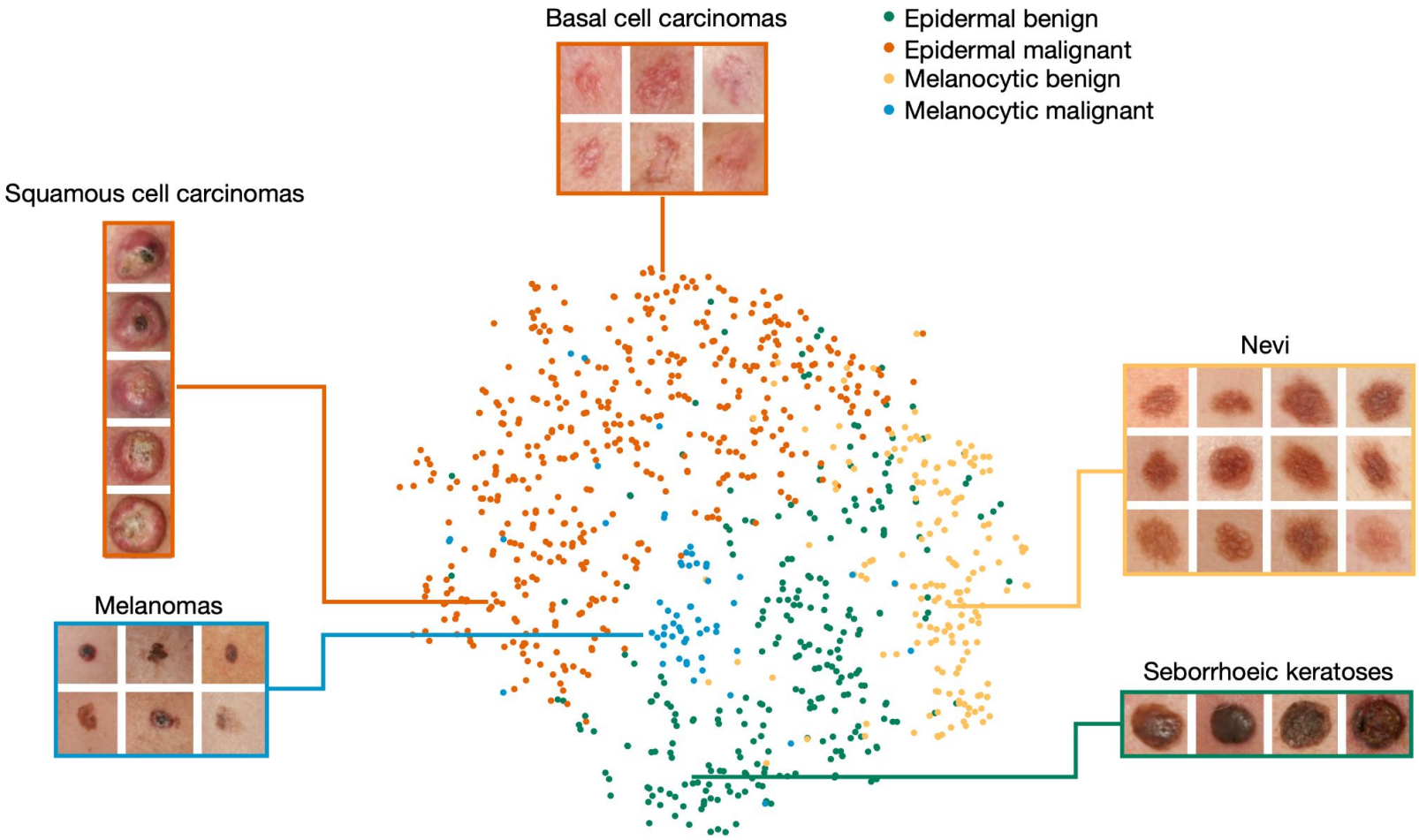


LETTER

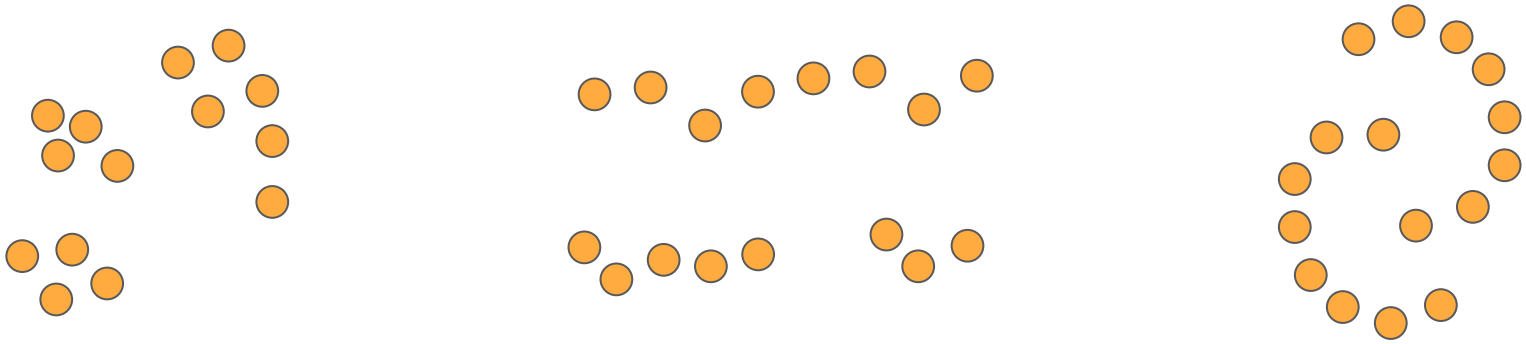
doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

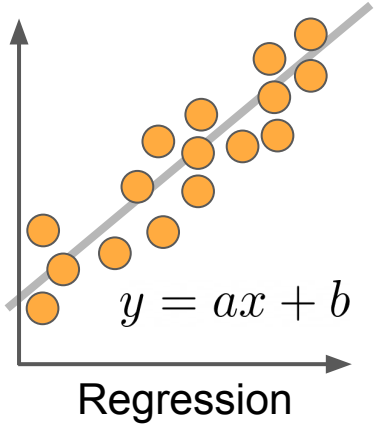
Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]



People are good in low-dimensional clustering

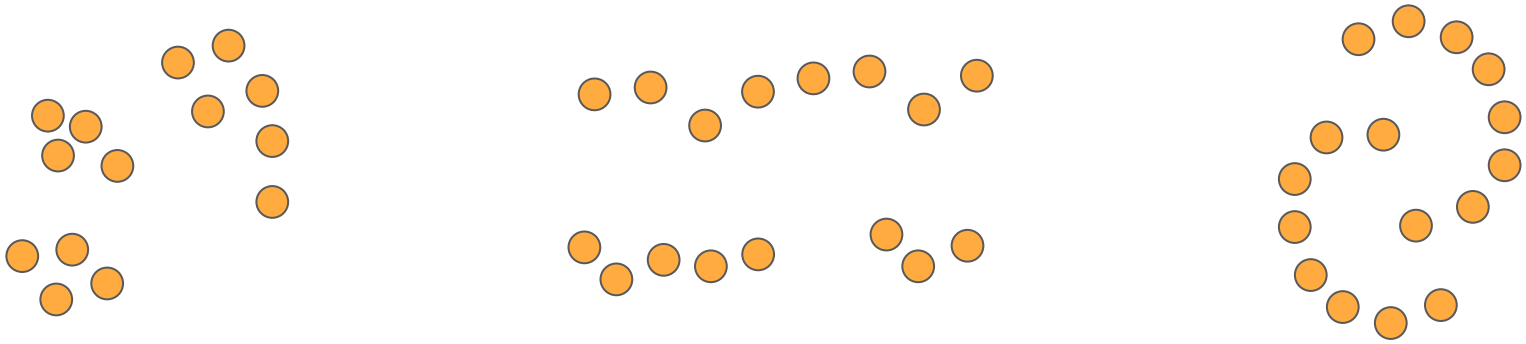


But how would we formally define a good clustering?

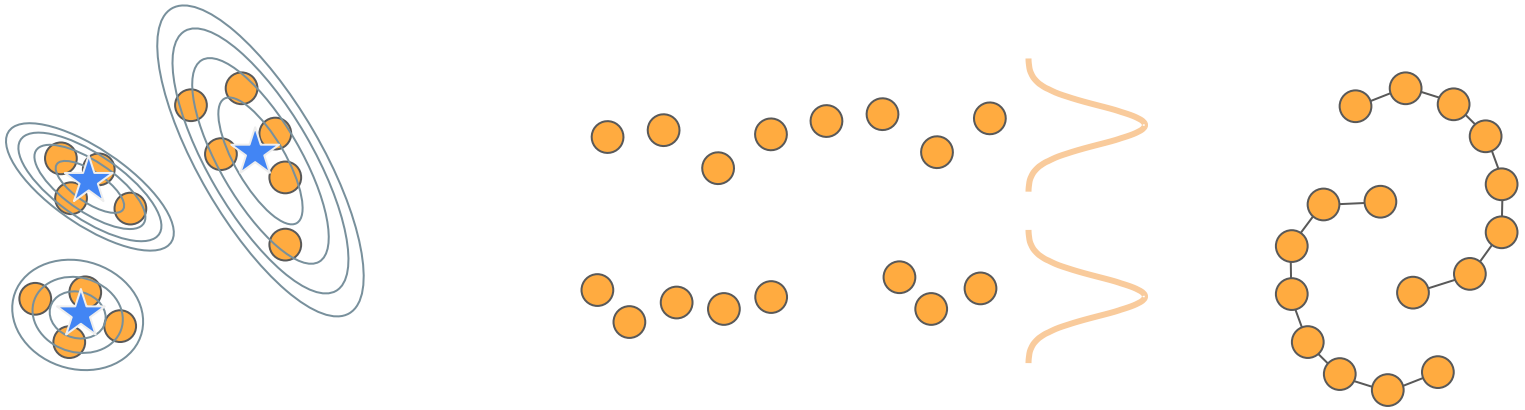


$$\arg\min_{a,b} \sum \left(y_i - (ax_i + b)\right)^2$$

People are good in low-dimensional clustering



But how would we formally define a good clustering?



Top-down approach

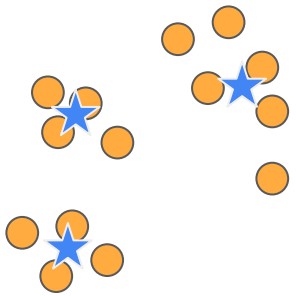
Bottom-up approach

Formal definition - top down

- Density-based clustering
- Minimize distances from “centroid”

- Each point x “belongs” to class $\mathbf{C_i}$ whose center is at $\boldsymbol{\mu}_i$

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



How to find C,μ?

K-means

Developed by Lloyd, 1957

Init:

Choose at random K data points, as centroids

Loop:

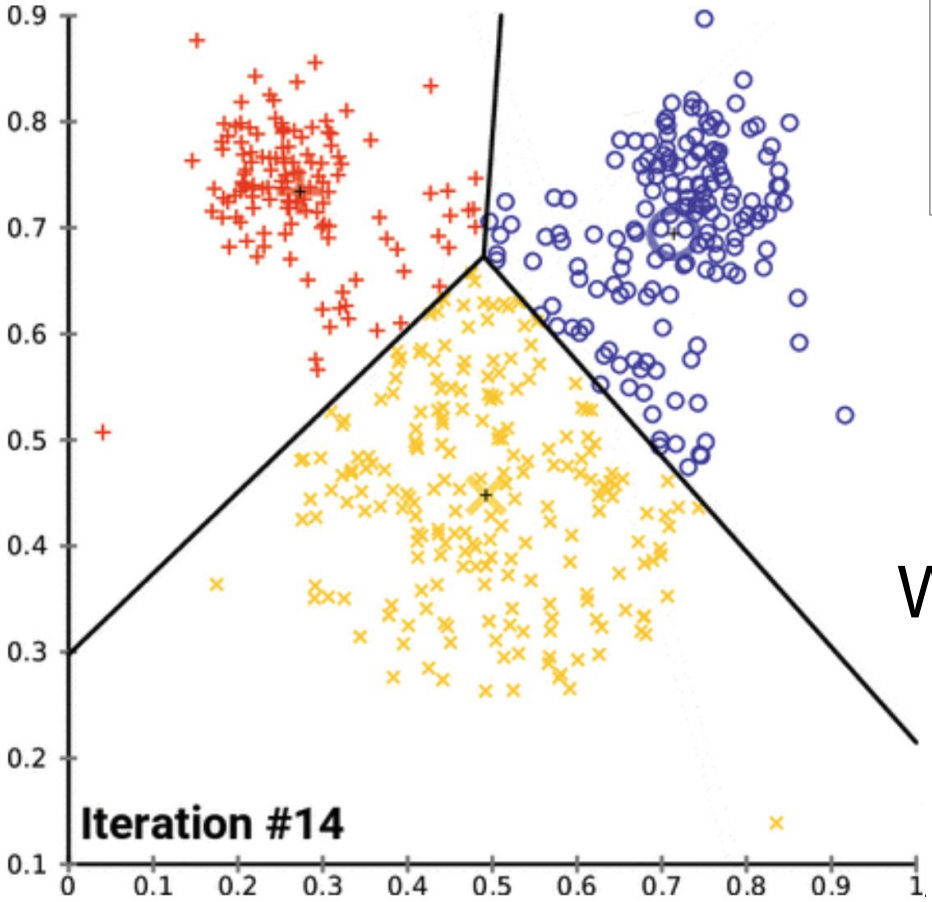
1. Re-assign each point to nearest centroid
2. Move each centroid to mean (or “center of mass”) of assigned data points

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

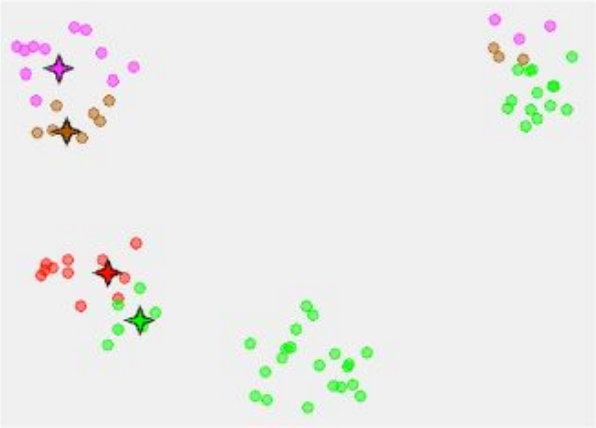
Stop:

Upon convergence (no assignment changes)

Would you cluster differently?

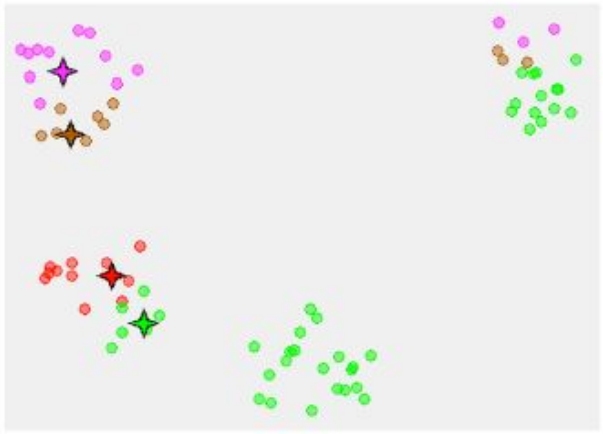
K-means

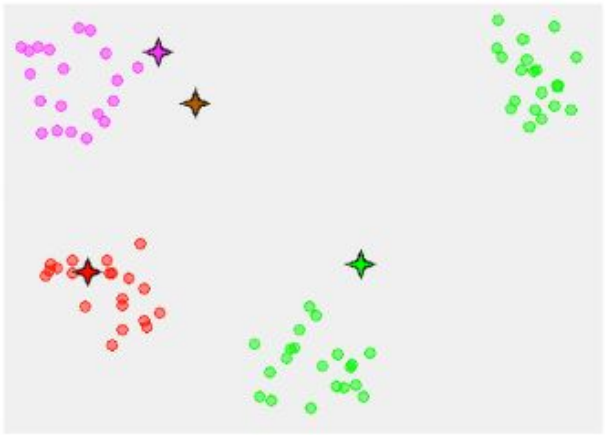
- Stochasticity - might not converge to optimal solution



K-means

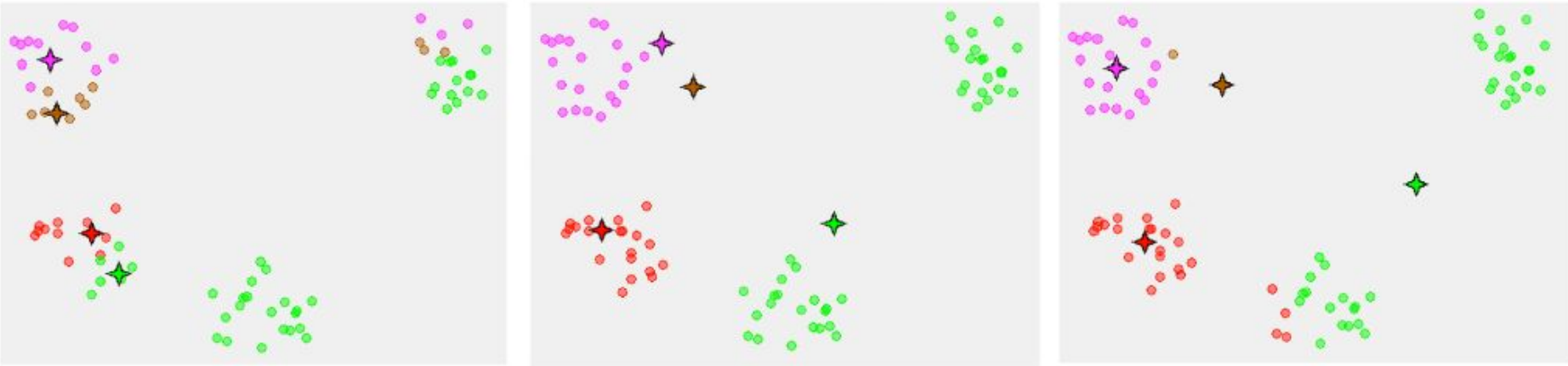
- Stochasticity - might not converge to optimal solution





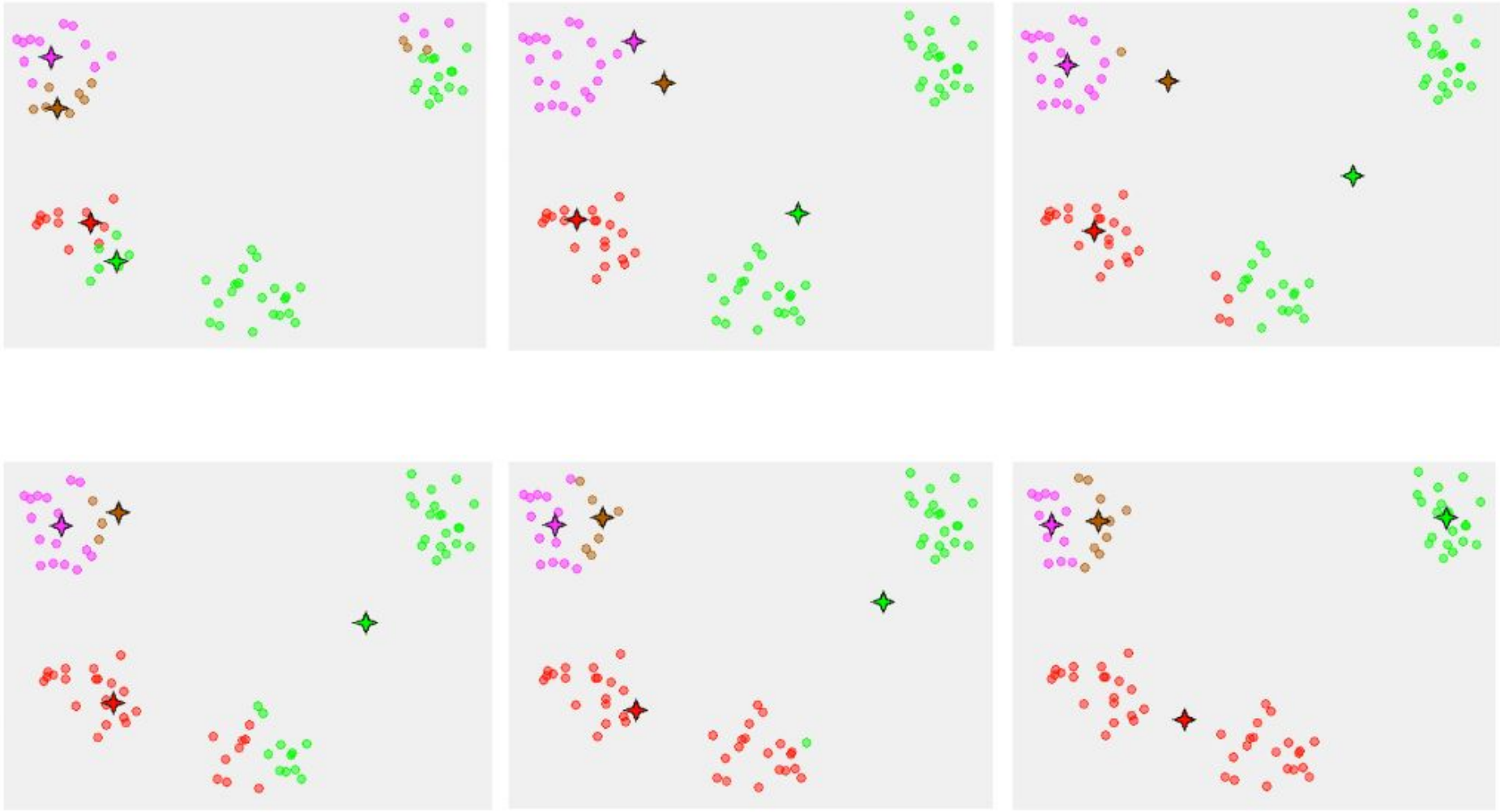
K-means

- Stochasticity - might not converge to optimal solution



K-means

- Stochasticity - might not converge to optimal solution



K-means

- Stochasticity
  - Convergence to local optimum?

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)

- YES!

Step 1 assigns each point to closest centroid (better or equal)

Step 2 minimizes sum-of-squared-distances for all clusters

K-means

- Stochasticity
- Random restarts

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
- Random restarts
- Initialization
  - K-means++
  - Cluster subset of data

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

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
- 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

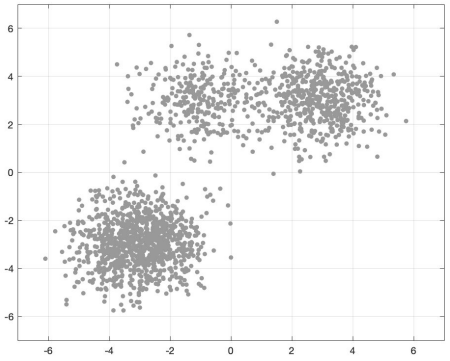
Loop:

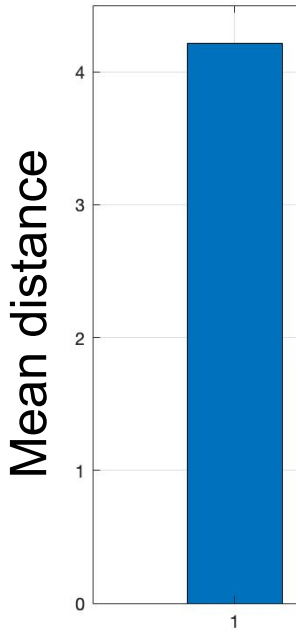
1. Re-assign each point to nearest centroid
2. Move each centroid to mean (or “center of mass”) of assigned data points

Stop:

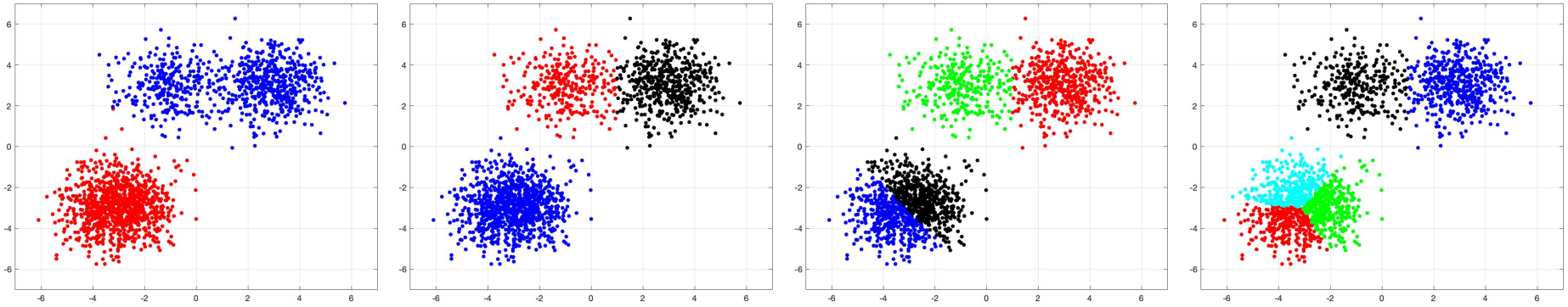
Upon convergence (no assignment changes)

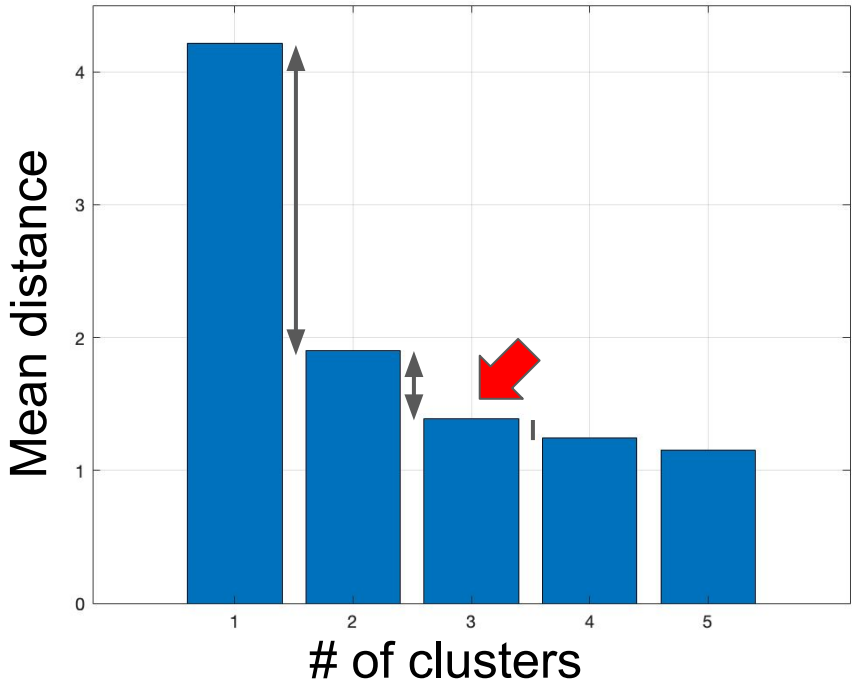
Choosing K using the Elbow method

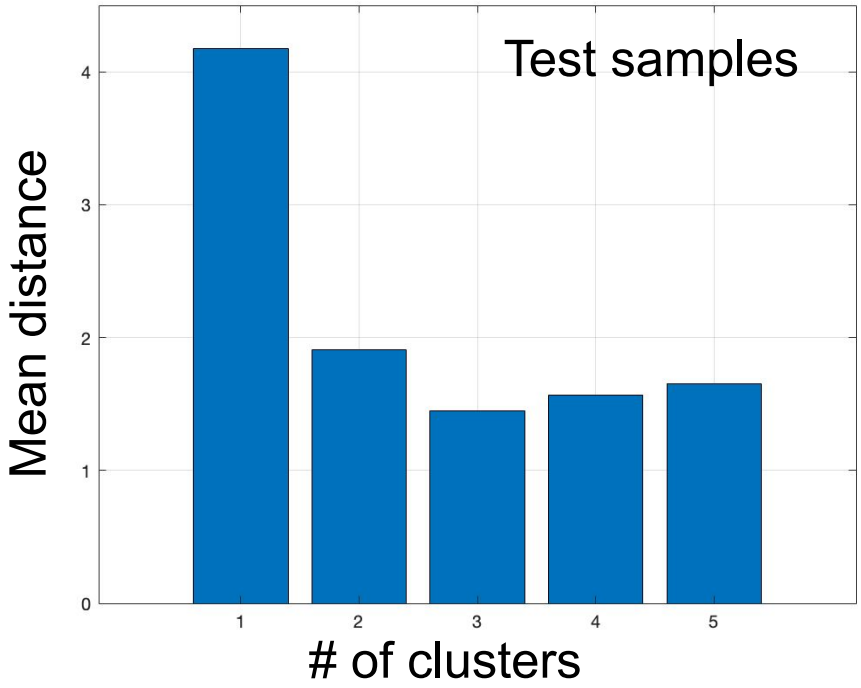




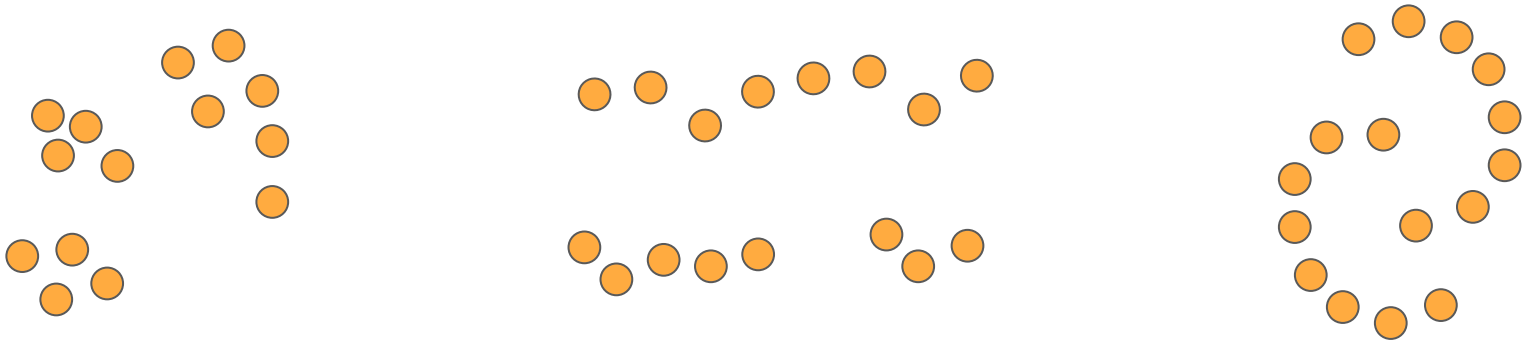
Choosing K using the Elbow method



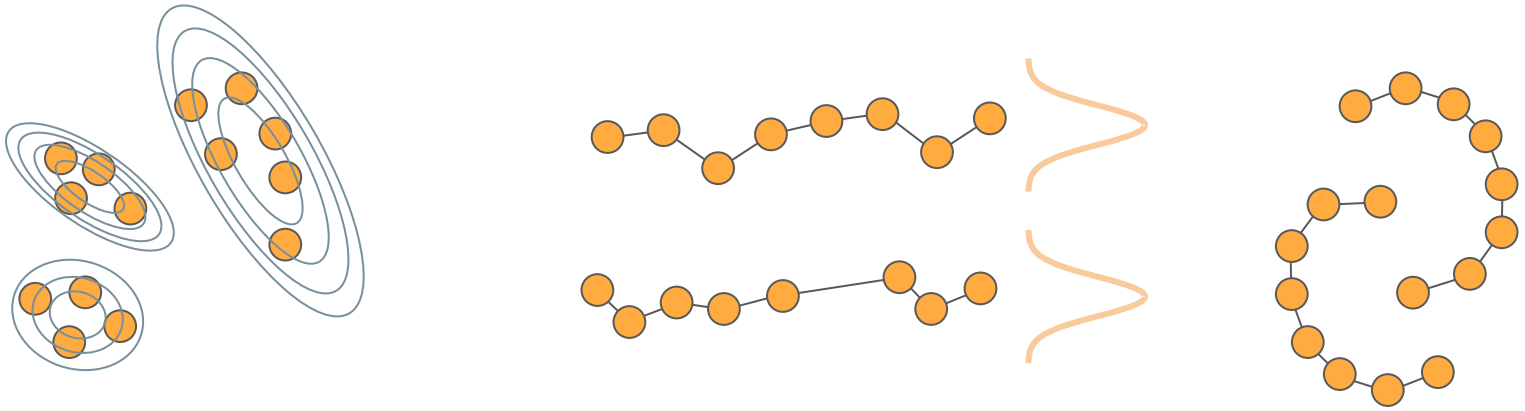




People are good in low-dimensional clustering



But how would we formally define a good clustering?

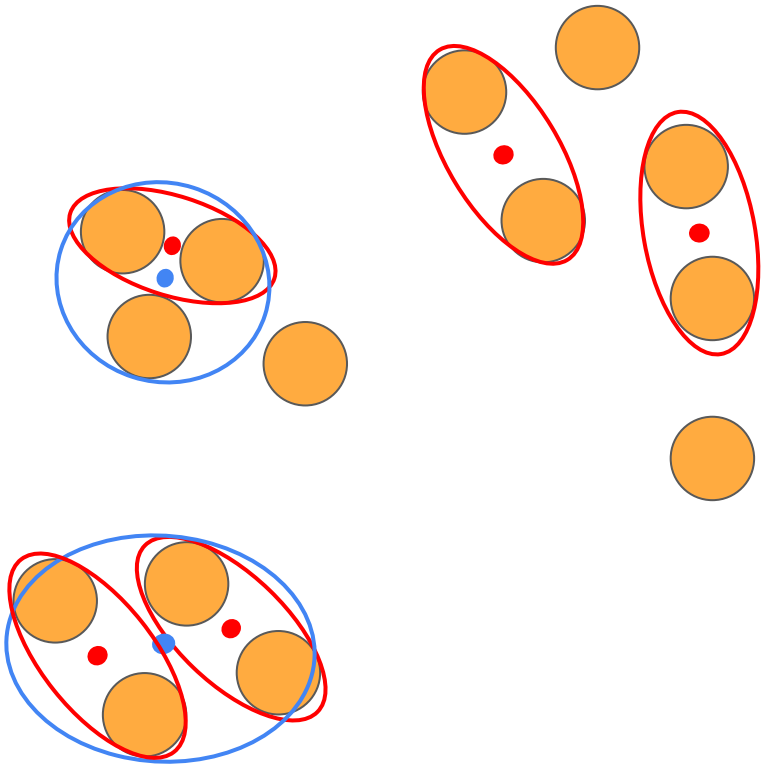


Top-down approach

Bottom-up approach

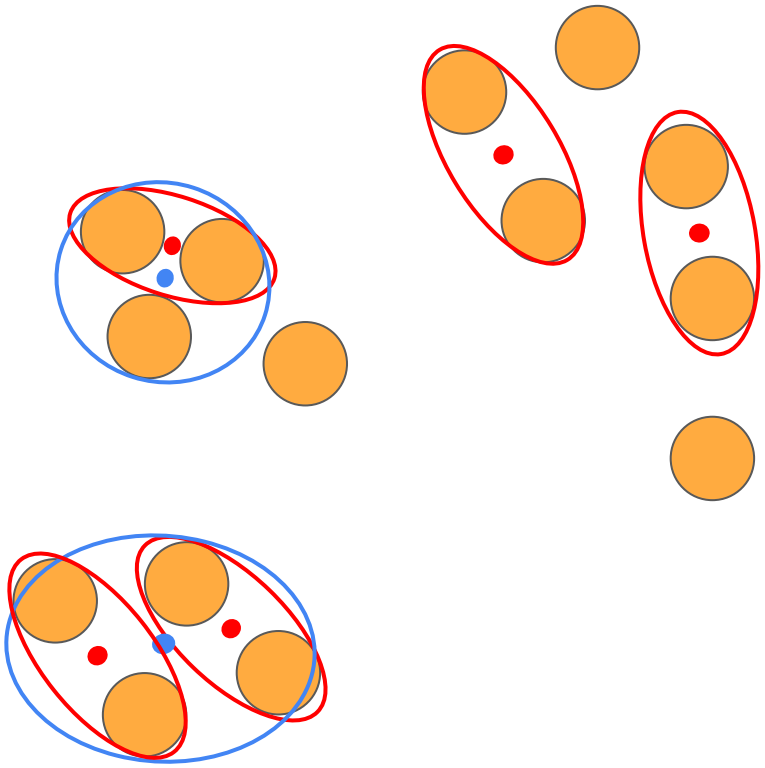
Hierarchical Agglomerative Clustering

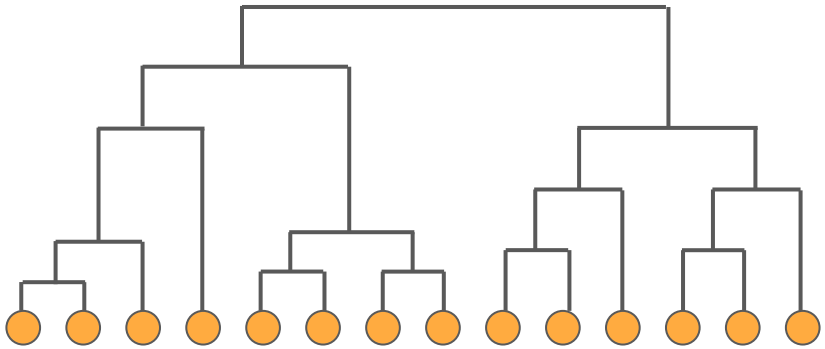
- Iteratively, merge similar points / sub-groups



Hierarchical Agglomerative Clustering

- Iteratively, merge similar points / sub-groups

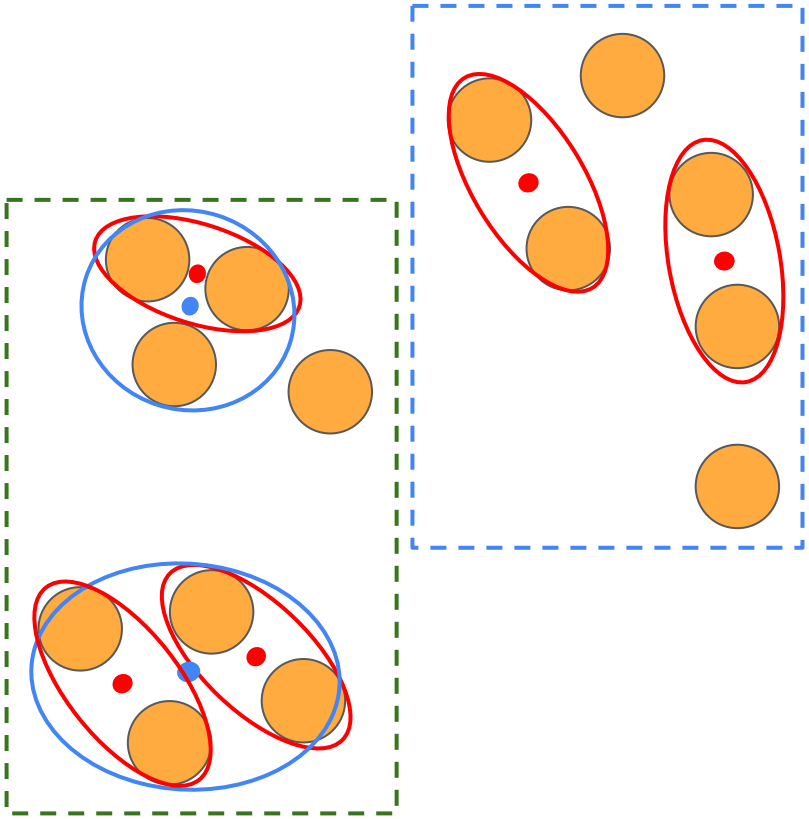


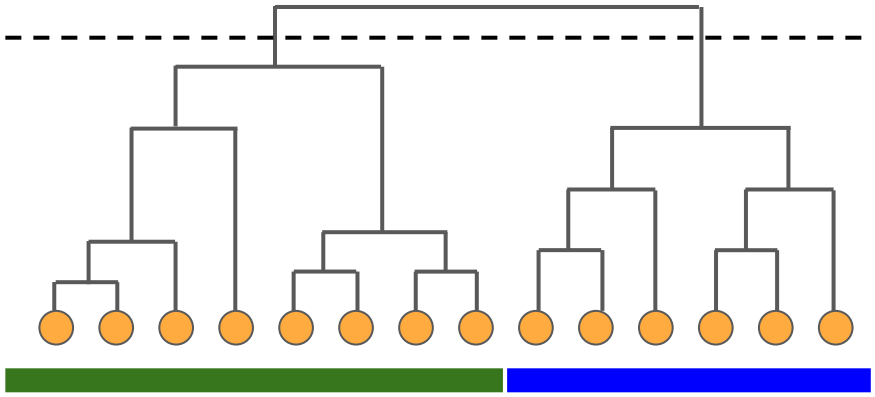


dendrogram

Hierarchical Agglomerative Clustering

- Iteratively, merge similar points / sub-groups

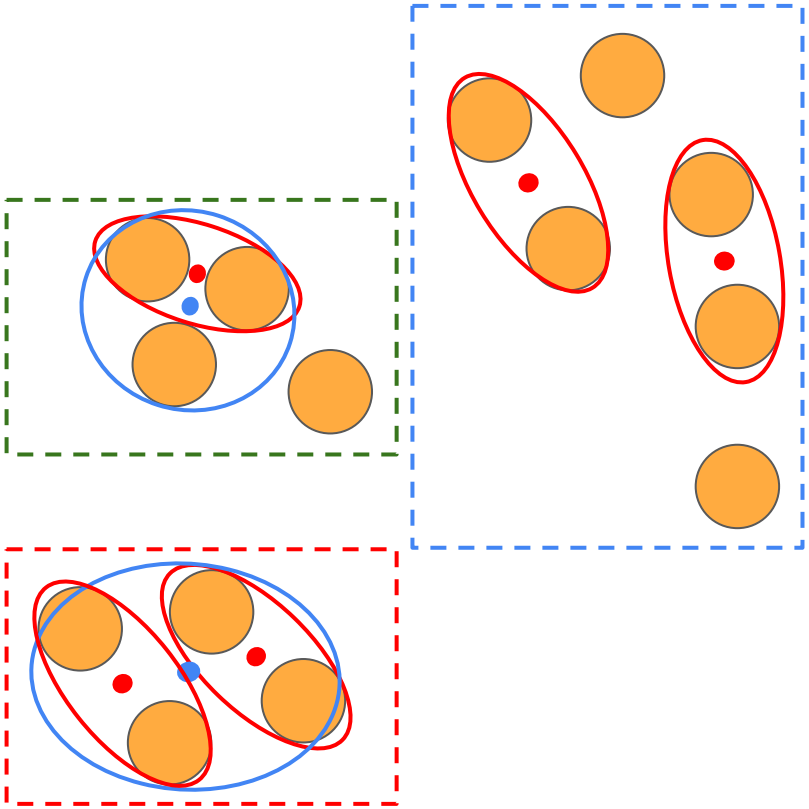


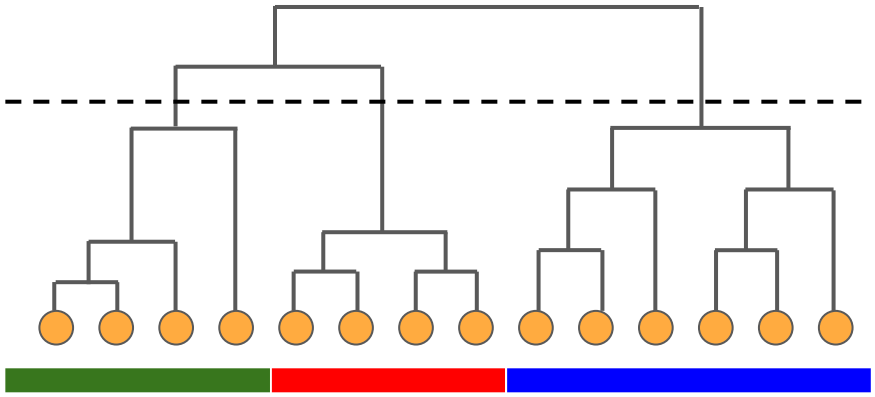


dendrogram

Hierarchical Agglomerative Clustering

- Iteratively, merge similar points / sub-groups





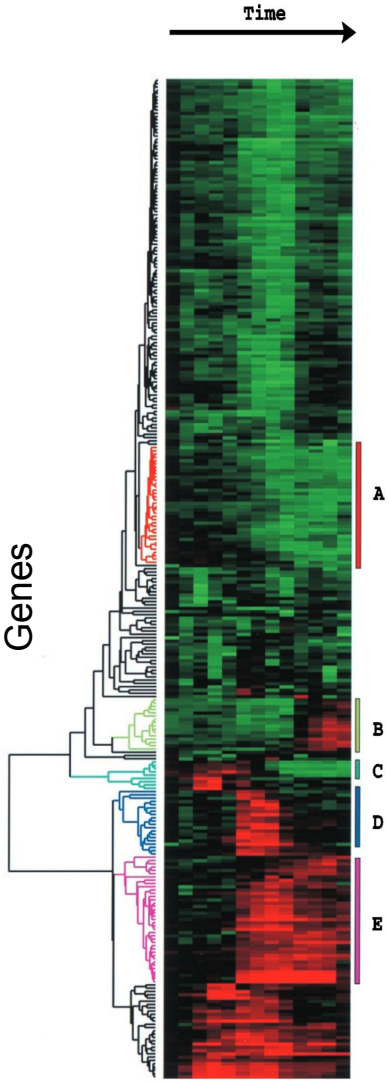
dendrogram

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

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



Clustered gene expression following serum stimulation of primary human fibroblasts

A Cholesterol biosynthesis

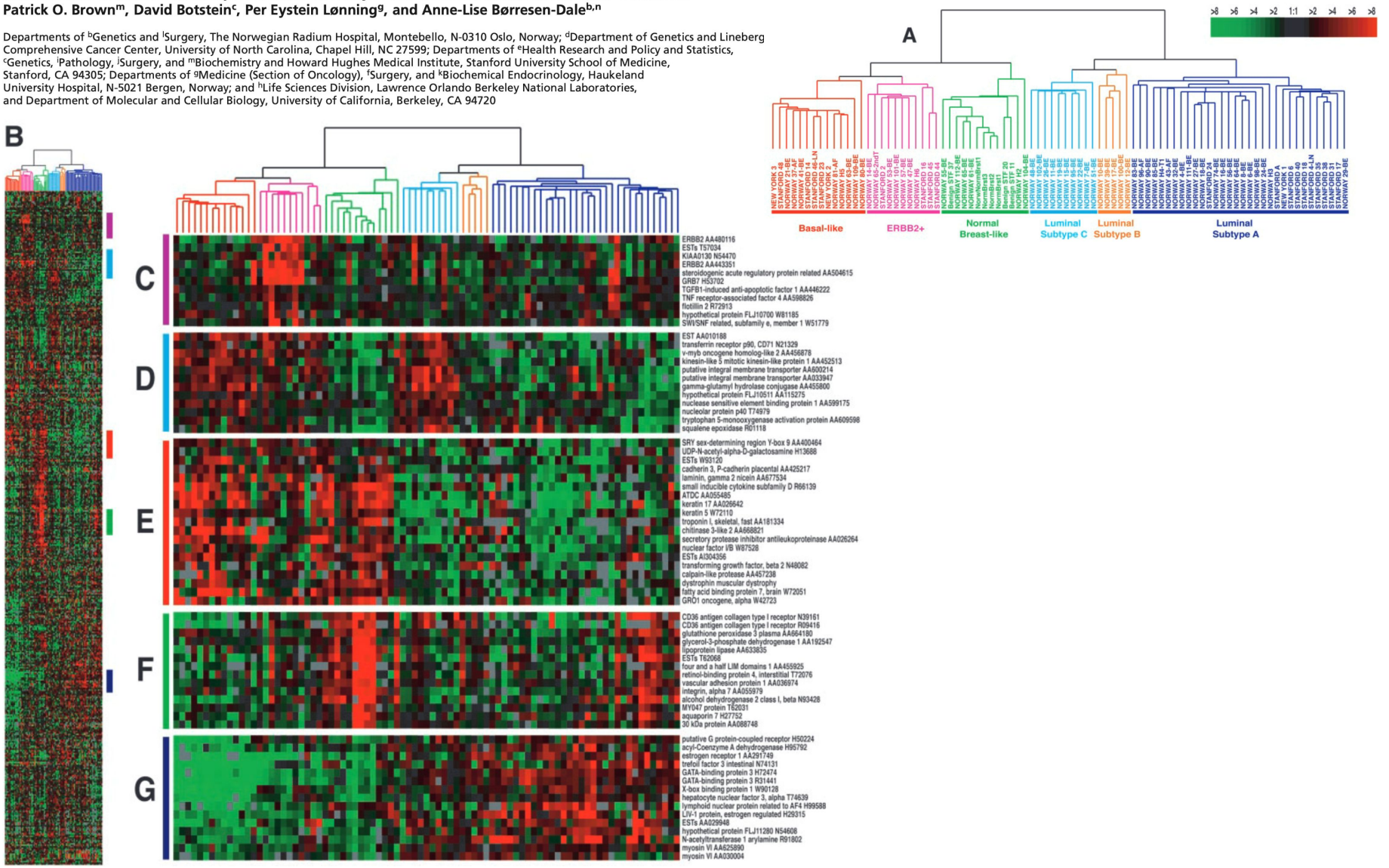
B cell cycle
C immediate-early genes
D signaling and angiogenesis

E wound healing and tissue remodeling

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

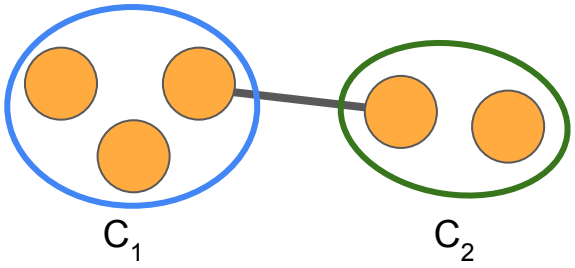
Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per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

Departments of [b]Genetics and [l]Surgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; [d]Department of Genetics and Lineberg Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of [e]Health Research and Policy and Statistics, [c]Genetics, [i]Pathology, [j]Surgery, and [m]Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [g]Medicine (Section of Oncology), [f]Surgery, and [k]Biochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and [h]Life Sciences Division, Lawrence Orlando Berkeley National Laboratories, and Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720



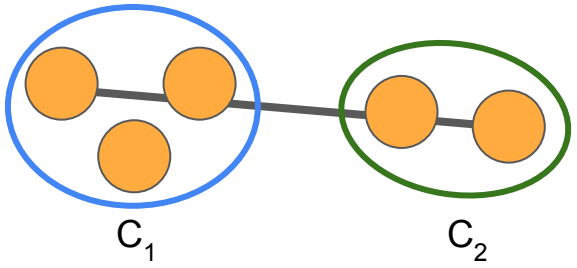
Hierarchical Agglomerative Clustering

How to define distance between clusters?



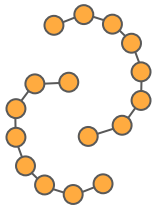
Single linkage

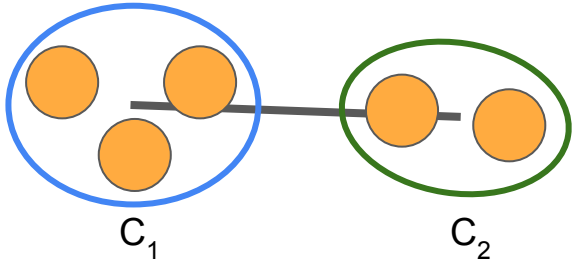
(closest pair)



Complete linkage

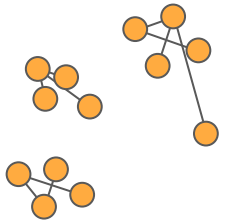
(farthest pair)





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

Often there is more than one “good” solution

“Clustering: science or art?”

Syllabus

1	1/1	AI in ophthalmology (Prof. Itay Chowers)
2	8/1	Classification
3	15/1	Learning 1
4	22/1	Learning 2
5	7/2	Regression (Wed.)
6	12/2	Deep learning in image analysis (Prof. Leo Joskowicz)
7	19/2	Clustering
8	26/2	Dimensionality reduction and visualization
9	28/2	Deep learning, Missing data (Wed.)
10	4/3	Natural language in medicine (Dr. Gabi Stanovsky)
11	11/3	?