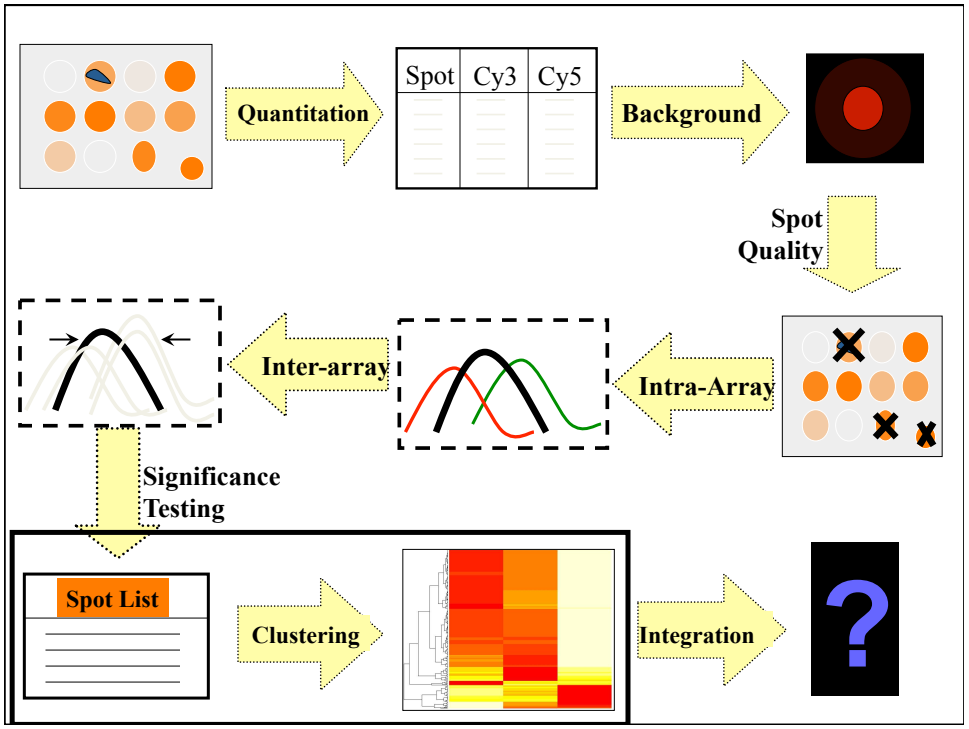
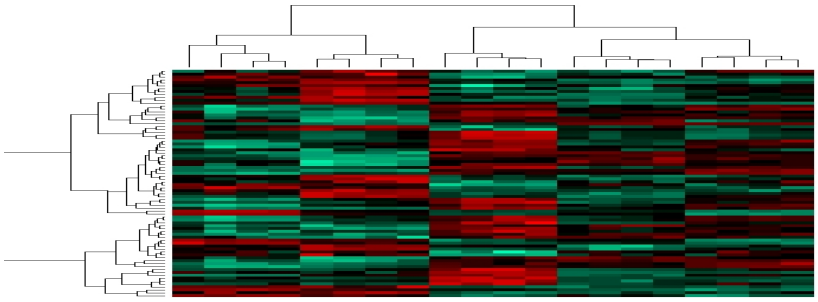


Clustering

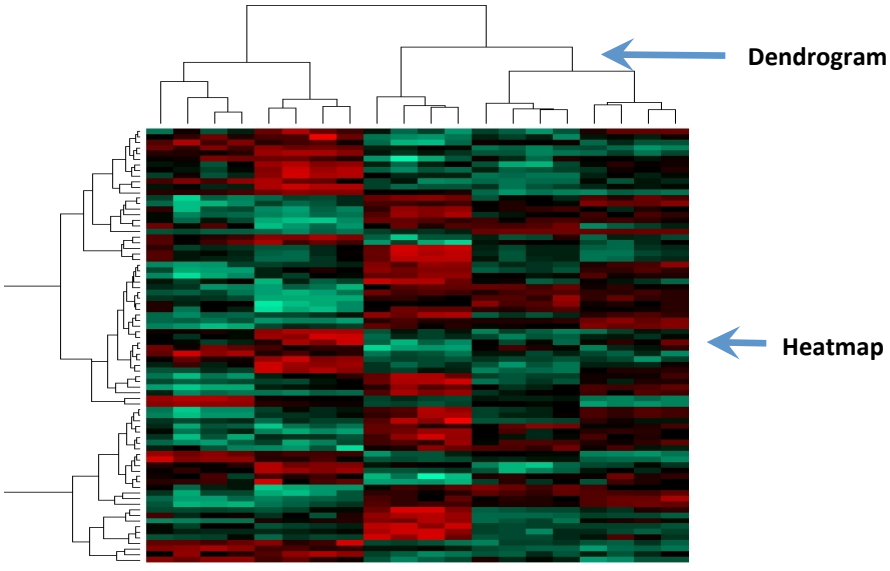
Why?	Finding patterns in the data
How?	Unsupervised machine-learning
Difficulty?	+
Research?	+++++



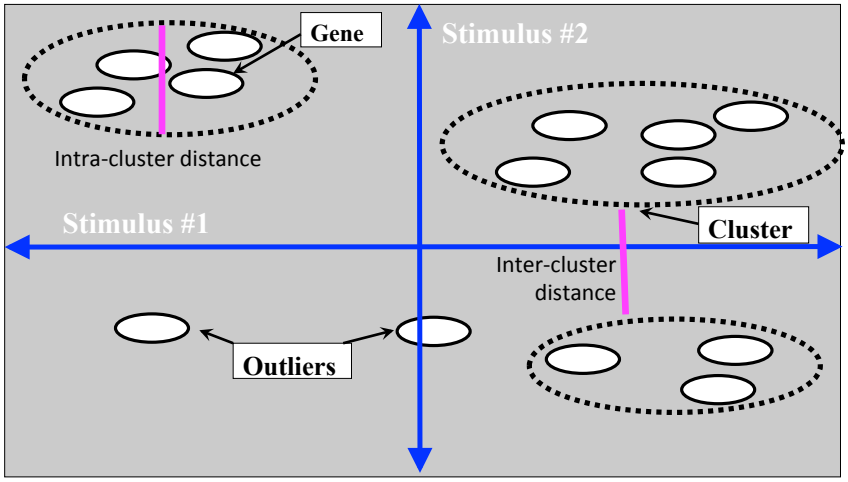
What is Clustering?

- Clustering: finding patterns in data
- Each "pattern" is a cluster
- A (small) branch of "machine learning"
- A (very) overused part of bioinformatics

Anatomy of a Clustergram



How is Clustering Done (Simple)?



Why is Clustering Used?

I. Data visualization

II. To predict class assignment

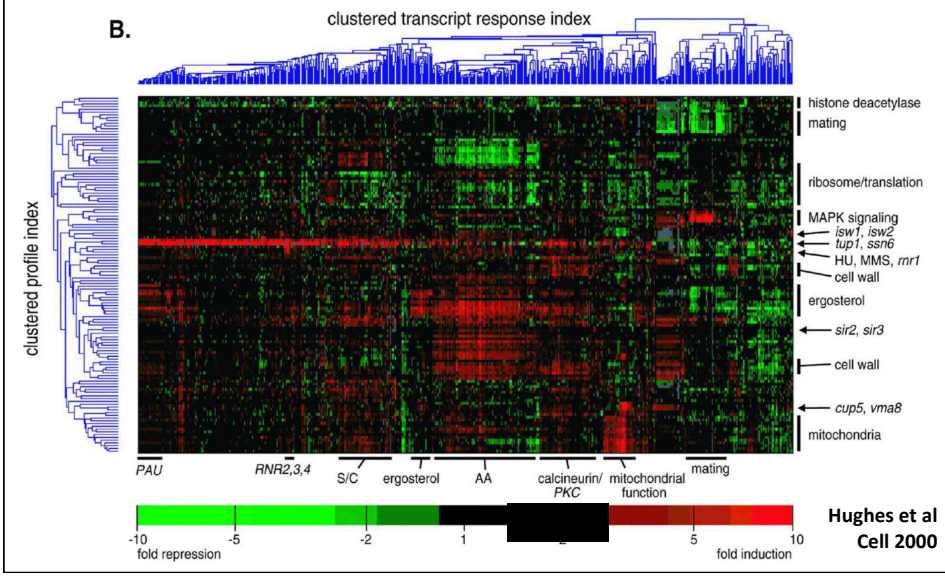
III. To identify co-regulation

IV. Quality Control

Example:
Predicting Gene Function

- Most genes have NO functional annotation
  - 1,500 / 7,000 yeast genes
  - 12,000 / 20,000 human genes
- Can we automatically estimate their function based on their patterns of expression?

Solution:
Clustering of Expression Profiles



Abuses of Clustering?

- Clustering pre-selected data
  - Clustering after significance analysis is only for visualization
- Detecting differential expression
  - Clustering cannot replace significance-testing
- No assessment of chance
  - How likely is a given pattern to be observed by chance alone? Statistics exist to test this!