Lecture 1: Introduction

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MATH 373

Plan for this Lecture

- What is Machine Learning?
- Notation and the Learning Problem
- Supervised vs. Unsupervised Learning
- Motivation and Applications

What is Machine Learning?

1. "[A] branch of artificial intelligence [that] concerns the construction and study of systems that can learn from data." - Wikipedia

2. " A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its perfromance at tasks in T, as measured by P, improves with experience E." - T. Mitchell

3. "Machine learning is a powerful artificial intelligence tool that enables us to crunch petabytes of data and make sense of a complicated world... It's solving previously unsolved problems." - Forbes

4. "Statistical learning refers to a vast set of tools for *understanding* data" - ISL book

A Major Component in Modern Data Science

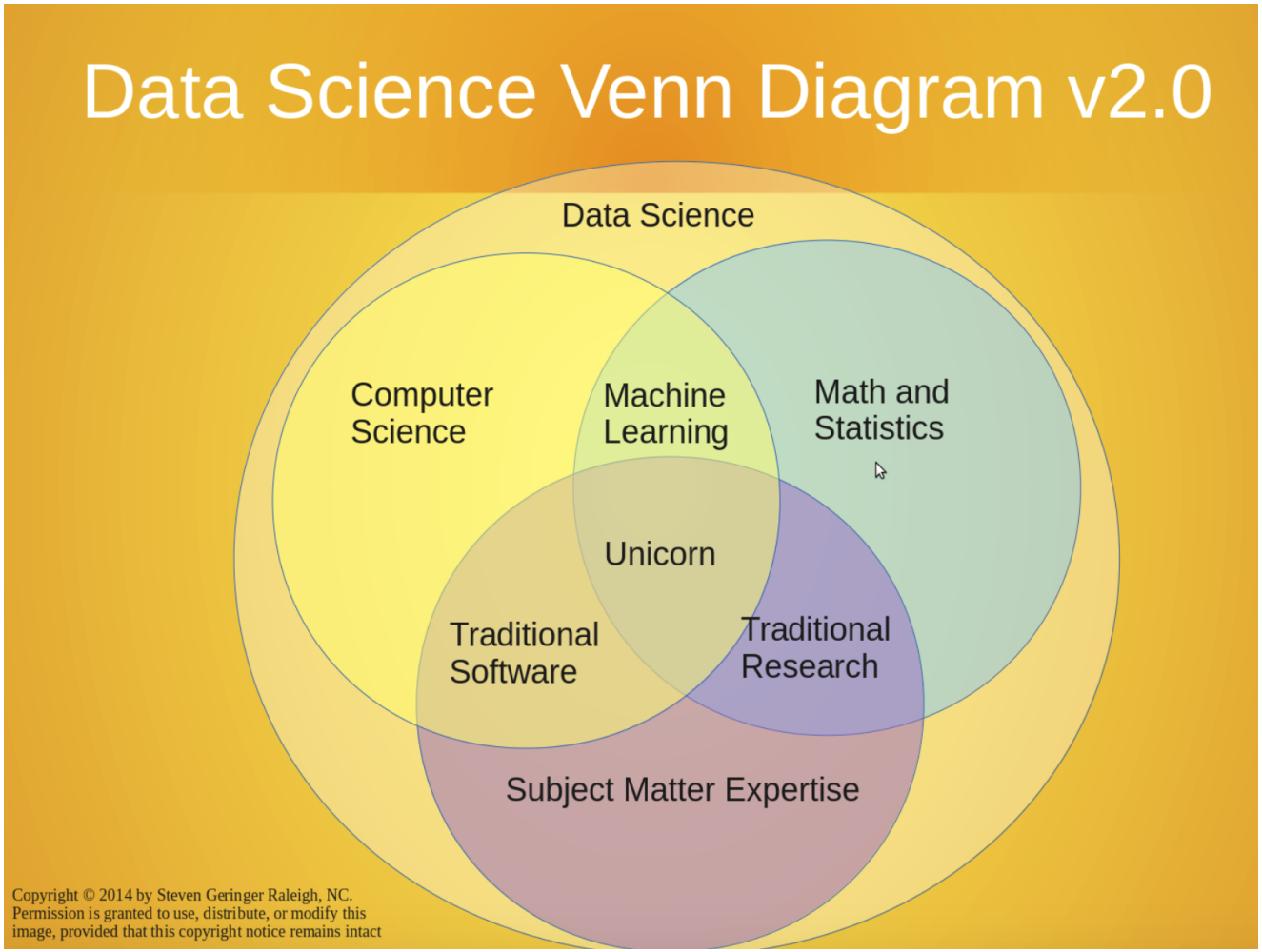


Figure: From *www.datasciencecentral.com*

Sample / Measurement Data

Experiment: Make p measurements on each of n samples.

Result: Data matrix / table X with n rows and p columns

- ith row of X is the vector of measurements on the ith sample
- jth column of X is the vector of values of the jth variable (measurement) across all samples

Different Perspectives on data:

- $n \times p$ matrix X
- n vectors of dimension $p \Leftrightarrow$ samples
- p vectors of dimension $n \Leftrightarrow$ variables

Notation

Data matrix: $$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Rows of X: p variable measurements for each observation.

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$$

Columns of X: n observations of each variable.

$$\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$$

Can write X as: $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = (x_1^T, x_2^T, \ldots, x_n^T)^T$

Data Dimensionality

Old Paradigm: More samples than variables ($n >> p$)

- Number of samples moderate (10s or 100s)
- Number of variables small (1s or 10s)

High Dimensional Paradigm: More variables than samples ($p >> n$)

- Number of samples moderate or large (100s or 1Ks)
- Number of variables *very* large (10Ks or 1Ms)

Big Data Paradigm: Many samples and/or many variables

Source of data: high-throughput measurement technologies for microarray analysis, e-commerce data, click-through rates, etc.

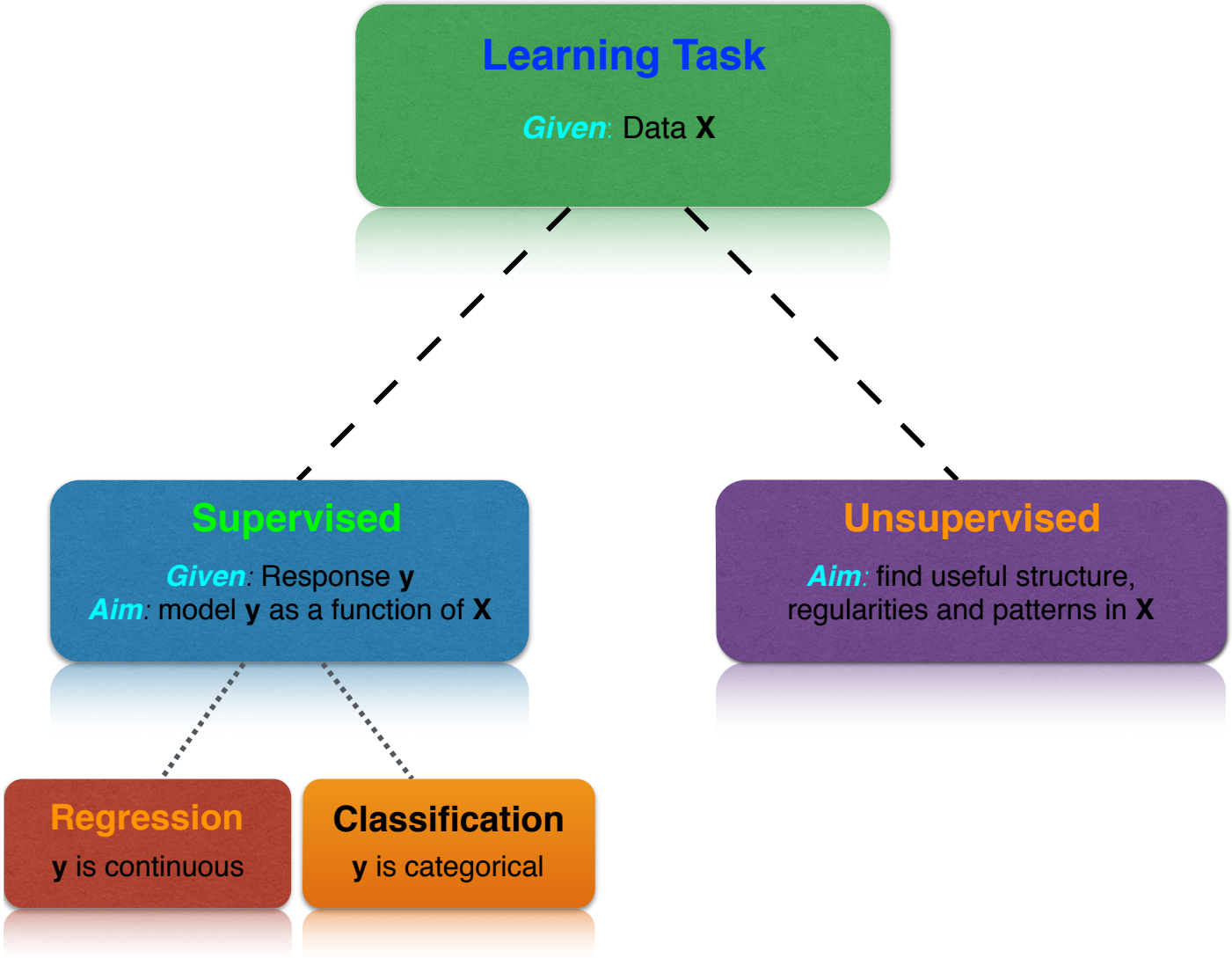
Response / Label

A response is an externally defined quantity of interest whose value is measured for each sample.

Notation: y_i denotes the response for the ith sample (or the ith row of the data matrix X). Vector of responses: $y = (y_1, \ldots, y_n)^T$

Note: when y_i is categorical, e.g. binary, it is referred to as a label.

Machine Learning: High-level Taxonomy



Supervised vs. Unsupervised Learning

Supervised Learning: Data matrix X and response y. In this case, y *supervises* the model and knowledge that is gained.

Goal: *model the response as a function of the sample*. That is, find the function f that represents the relationship between X and y:

$$y = f(X) + \epsilon$$

Here, ϵ is the error associated with the model.

- Classification: y takes discrete values
- Regression: y takes continuous values (e.g. linear regression)

Challenges: model selection and model assessment

Supervised vs. Unsupervised Learning

Unsupervised Learning: Data matrix X, but no response to supervise any model!

Goal: identify structure, regularities, and patterns in X

- Data mining: finding "special" subsets of a large data set
- Clustering: finding patterns in X through partitioning the data

Challenges:

- what structure are we looking for?
- how do we evaluate a method?
- theoretical properties?

The Machine Learning way of Thinking

1. Ask what kind of data? Supervised or unsupervised problem? What question are we trying to answer?
2. Prepare / clean data: imputation, outlier removal, etc.
3. Explore data $\rightarrow$ hypotheses about X and/or model f
4. Apply models and algorithms to answer question
5. Validation of approach

Example: Housing data

Samples: houses *Response*: cost of house

Variables: features of each house

- size (sq. ft.), distance to public transportation
- # bedrooms, # bathrooms
- attached garage? good school district?

Goals:

- *Prediction:* Find function f(variables) to accurately predict the cost of a house that is not in the data set.
- *Variable selection:* Identify a (small) set of important variables that can be used to predict housing cost.

Example: Medical tests

Samples: patients *Response*: disease state (0 or 1)

Variables: results of diagnostic tests

- blood pressure (cystolic, diastolic)
- temperature
- heart rate
- age
- do any relatives have disease?

Goals: predict disease state of new patient (personalized medicine), identify variables needed for accurate prediction.

Relational Data: Networks and Graphs

Data matrix $X = \{x_{ij}\}$ is square, with x_{ij} representing the existence, or degree, of association between sample i and sample j.

Examples:

- x_{ij} = distance or correlation between samples s_i and s_j
- x_{ij} = 1 if s_i likes s_j, and 0 otherwise

Goals:

- Assess causality and conditional independence
- Find subgraphs or communities of samples with a high level of interaction or association

Fisher's Iris Data

Figure: Iris setosa. Courtesy of `www.wikipedia.org`

- Four attributes are measured for $n = 150$ flowers
  - 50 samples each of *Iris Setosa, Iris Virginica, Iris Versicolor*
  - $p = 4$ measurements: length and width of sepals and petals

Fisher's Iris Data: Scatterplots

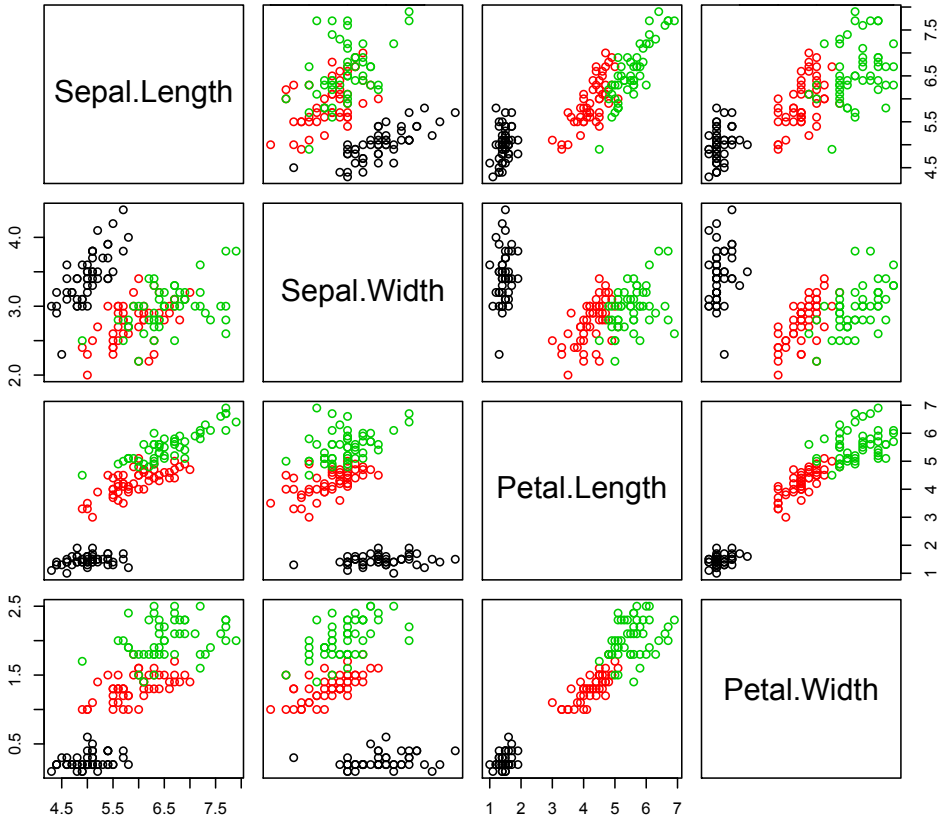


Figure: Pairwise scatterplot of Iris measurements. Colors: *Setosa*, *Virginica*, *Versicolor*.

Fisher's Iris Data: PCA and Clustering

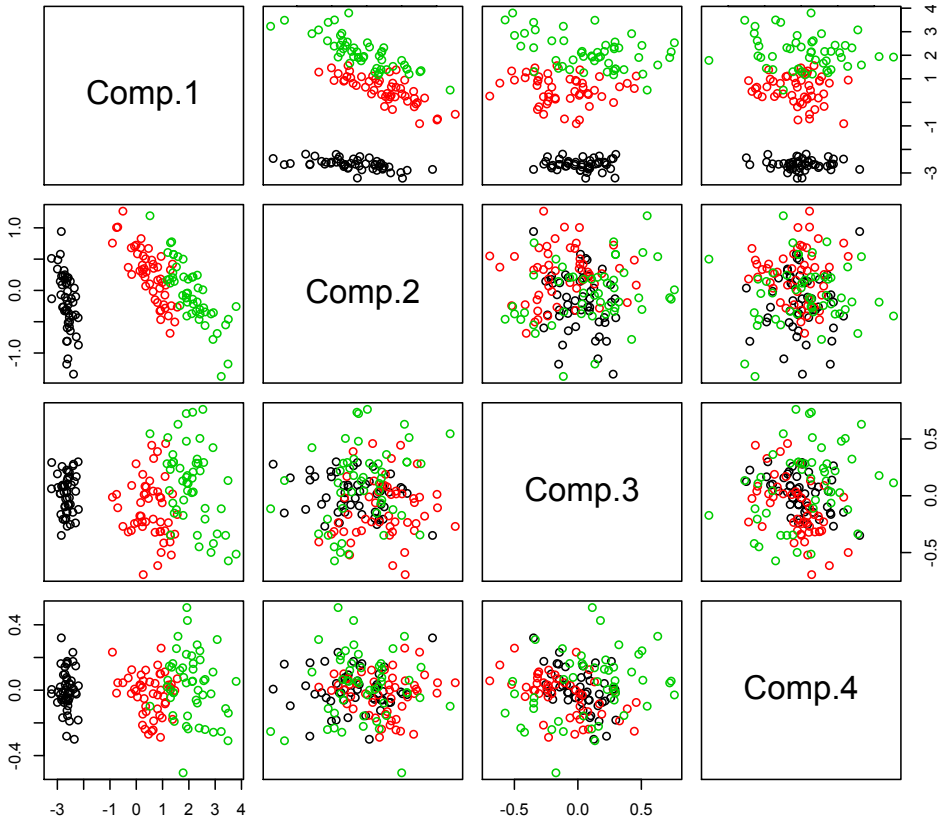
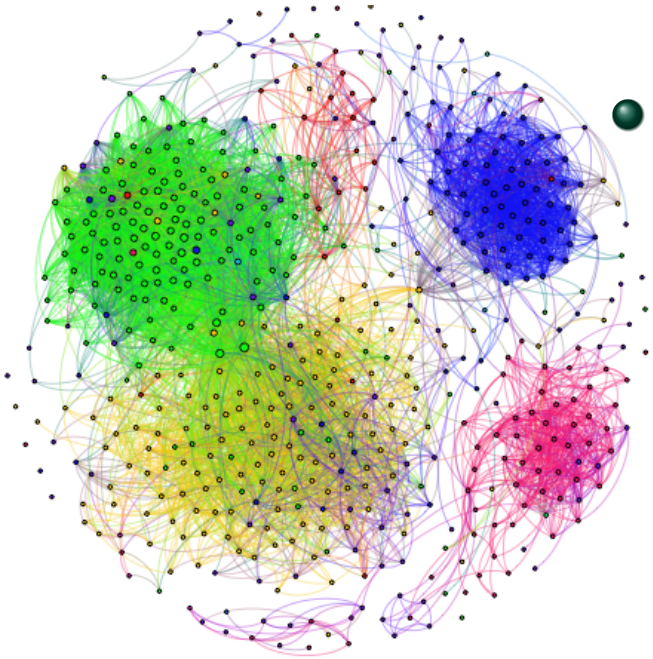


Figure: Pairwise scatterplot of principal components of Iris measurements.

Colors: *Setosa*, *Virginica*, *Versicolor*.

Force Directed Layout of My Facebook Data



- Network of 561 vertices and 8375 edges
  - Each vertex is a friend of mine
  - Each edge represents a friendship between two individuals
  - Each individual is categorized into one of 8 unique locations

Facebook Data: Community Detection

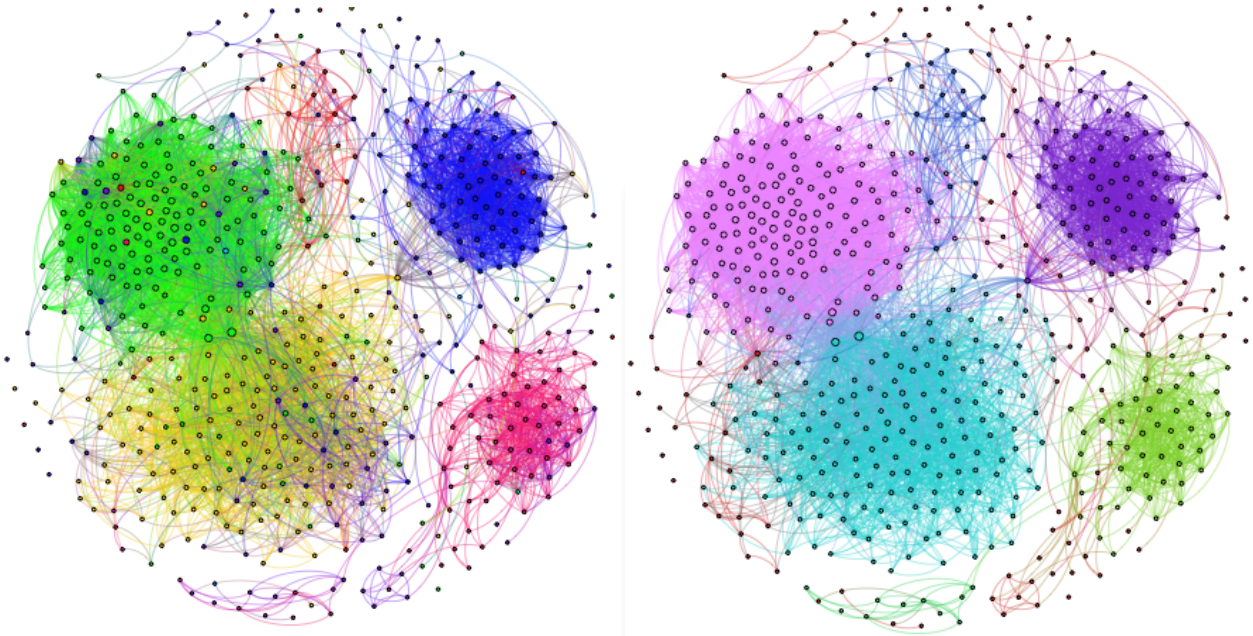


Figure: (Left): Original Network colored by location. (Right): Application of community detection finds clusters that closely match the features of the original network. Vertices are colored by community.

Gene Expression Arrays and Clustering

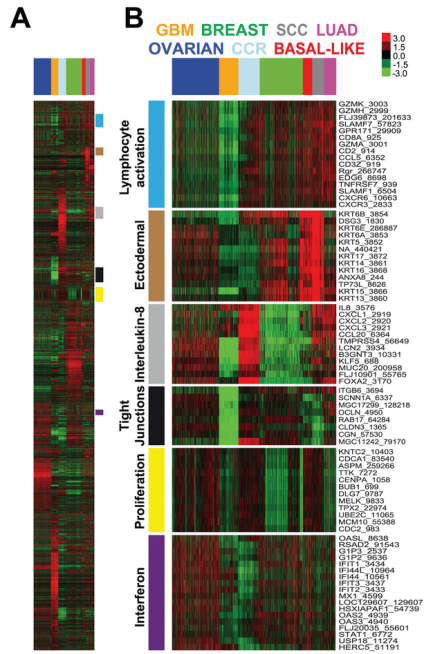


Figure: Gene expression array from *Nature* article

- *Samples*: subjects *Variables*: genes
- **Goal**: cluster subjects according to gene expression to identify relationship among cancer subtypes and genes

Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- Coursera.org
  - Free online courses in data science and machine learning
  - Recent notable course: ”The Data Scientist’s Toolbox”

Focus of this course

Principles of machine learning

- Algorithms: regression, classification, clustering, tree-based methods
- Theory: when to use what algorithm and why. Probability, statistics, and linear algebra
- Software: python and R
- Data-driven