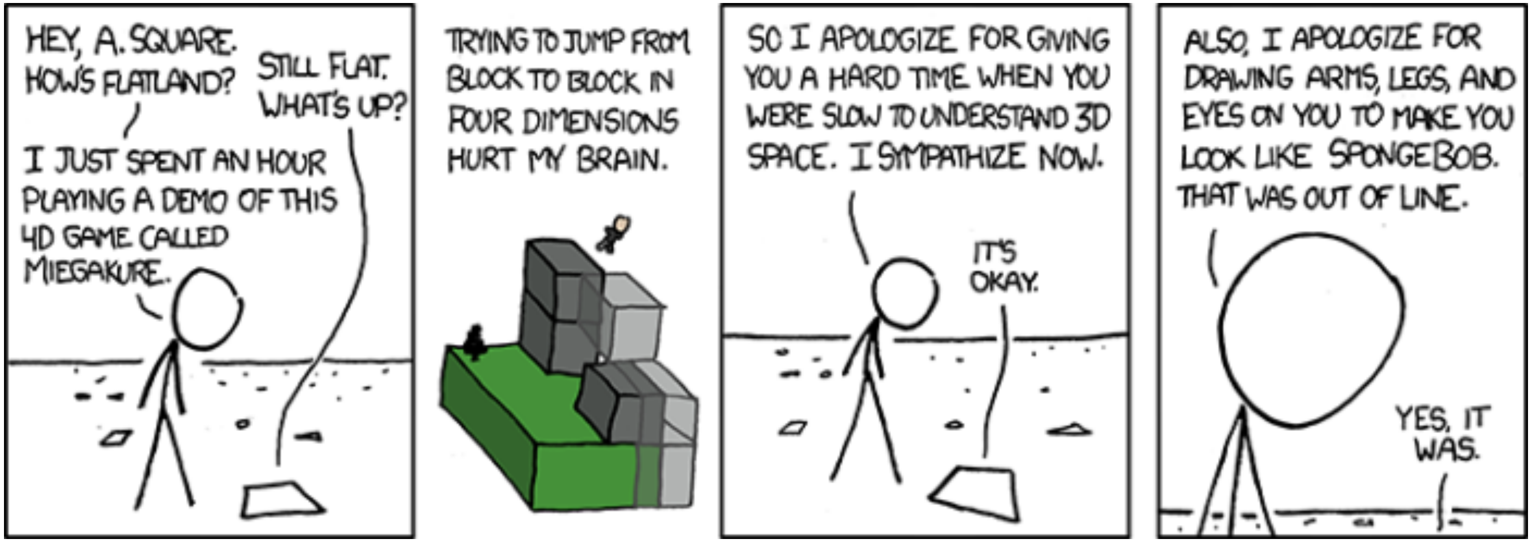
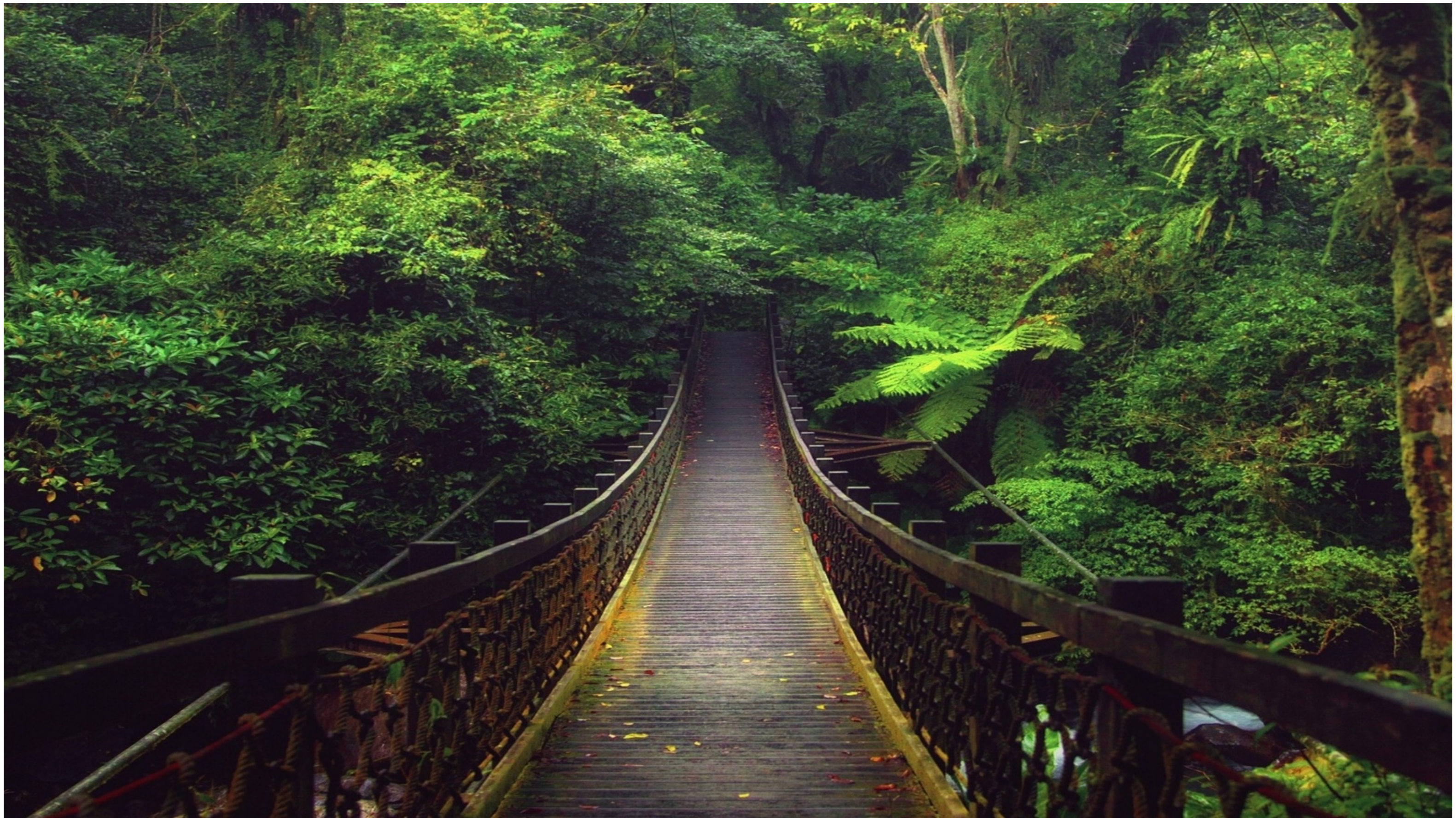
# Module 3

# Understanding data in reduced dimensions


HEY, A. SQUARE. HOW'S FLATLAND?
STILL FLAT. WHAT'S UP?
I JUST SPENT AN HOUR PLAYING A DEMO OF THIS 4D GAME CALLED MIEGAKURE.
TRYING TO JUMP FROM BLOCK TO BLOCK IN FOUR DIMENSIONS HURT MY BRAIN.
SO I APOLOGIZE FOR GIVING YOU A HARD TIME WHEN YOU WERE SLOW TO UNDERSTAND 3D SPACE. I SYMPATHIZE NOW.
IT'S OKAY.
ALSO, I APOLOGIZE FOR DRAWING ARMS, LEGS, AND EYES ON YOU TO MAKE YOU LOOK LIKE SPONGEBOB. THAT WAS OUT OF LINE.
YES, IT WAS.

xkcd



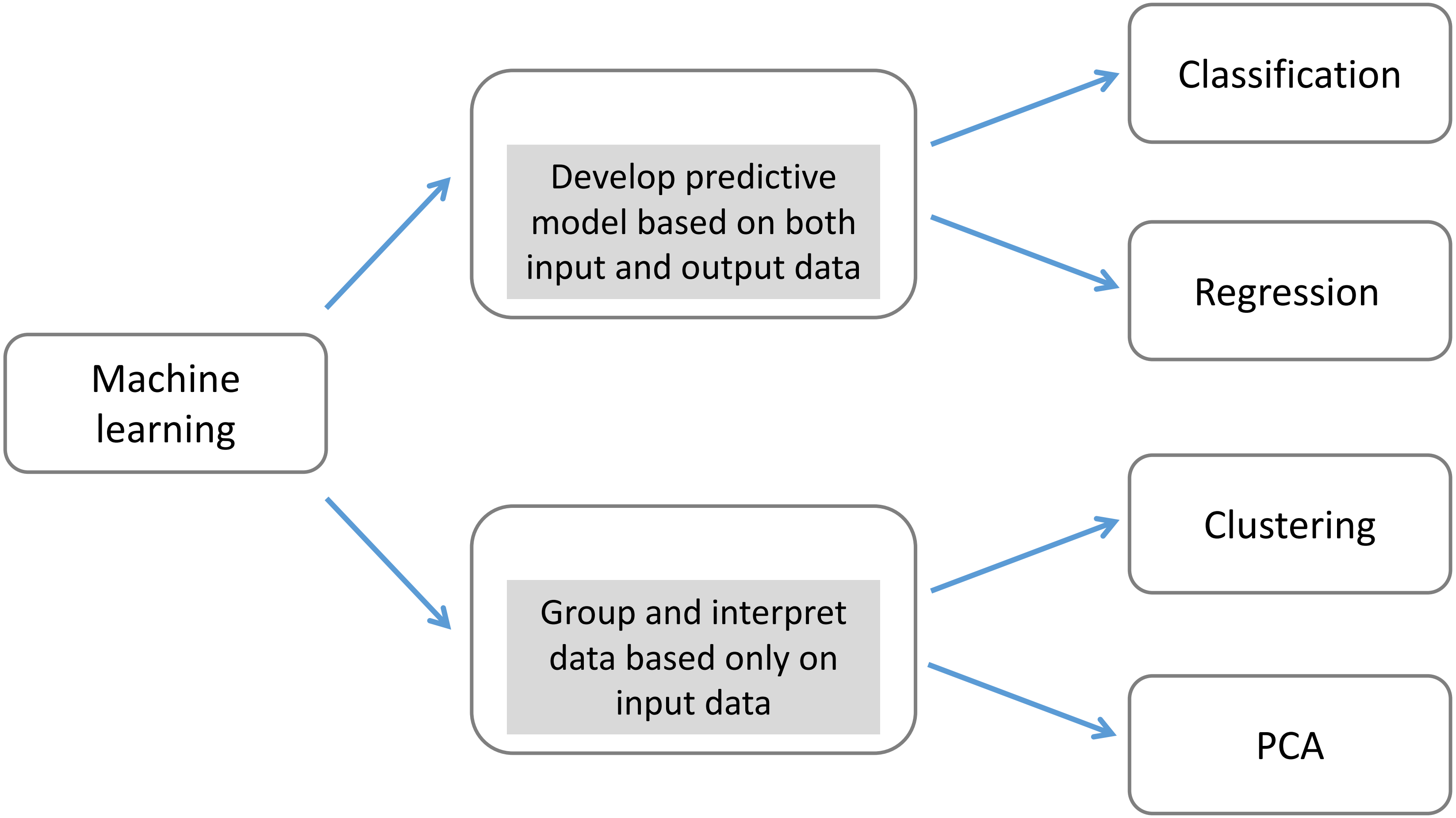
What data reduction is occurring here?

# Visualizing high-dimensional data

## How would you visualize…

- 1 variable?
- 2 variables?
- 3 variables?
- 4 variables?
- Systems biology problem → ~1000s of variables?

# Supervised versus unsupervised learning methods



Machine learning

Develop predictive model based on both input and output data

Group and interpret data based only on input data

Classification

Regression

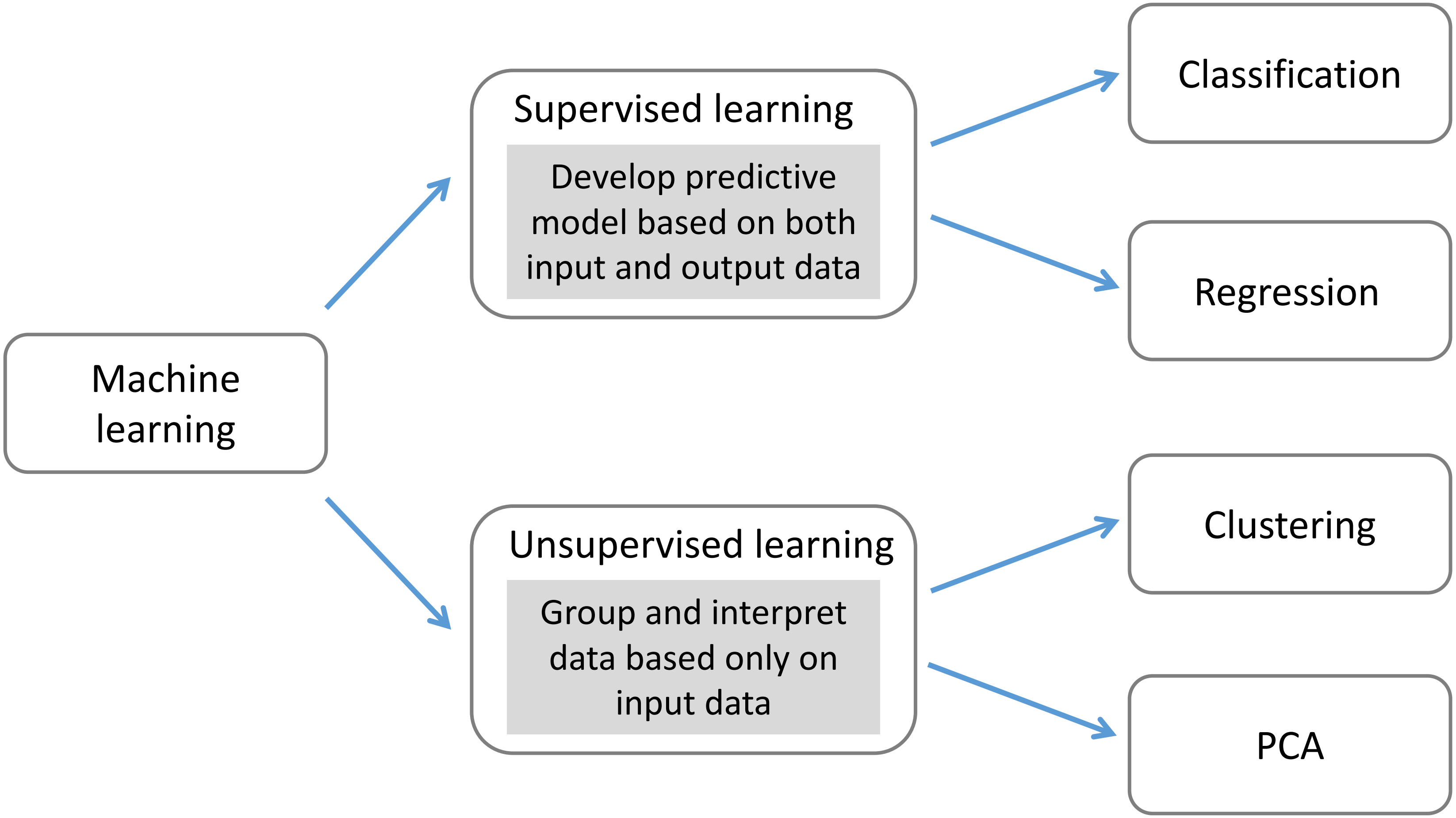
Clustering

PCA

# Supervised versus unsupervised learning methods

Supervised: Infer from training examples

Unsupervised: Exploratory data analysis



Machine learning

Supervised learning

Develop predictive model based on both input and output data

Unsupervised learning

Group and interpret data based only on input data

Classification

Regression

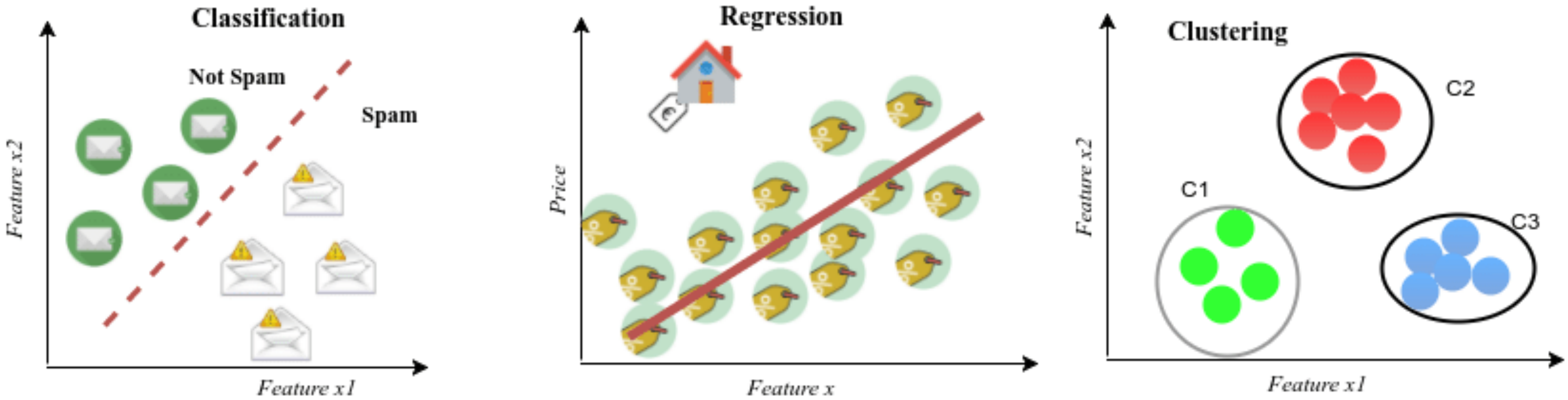
Clustering

PCA

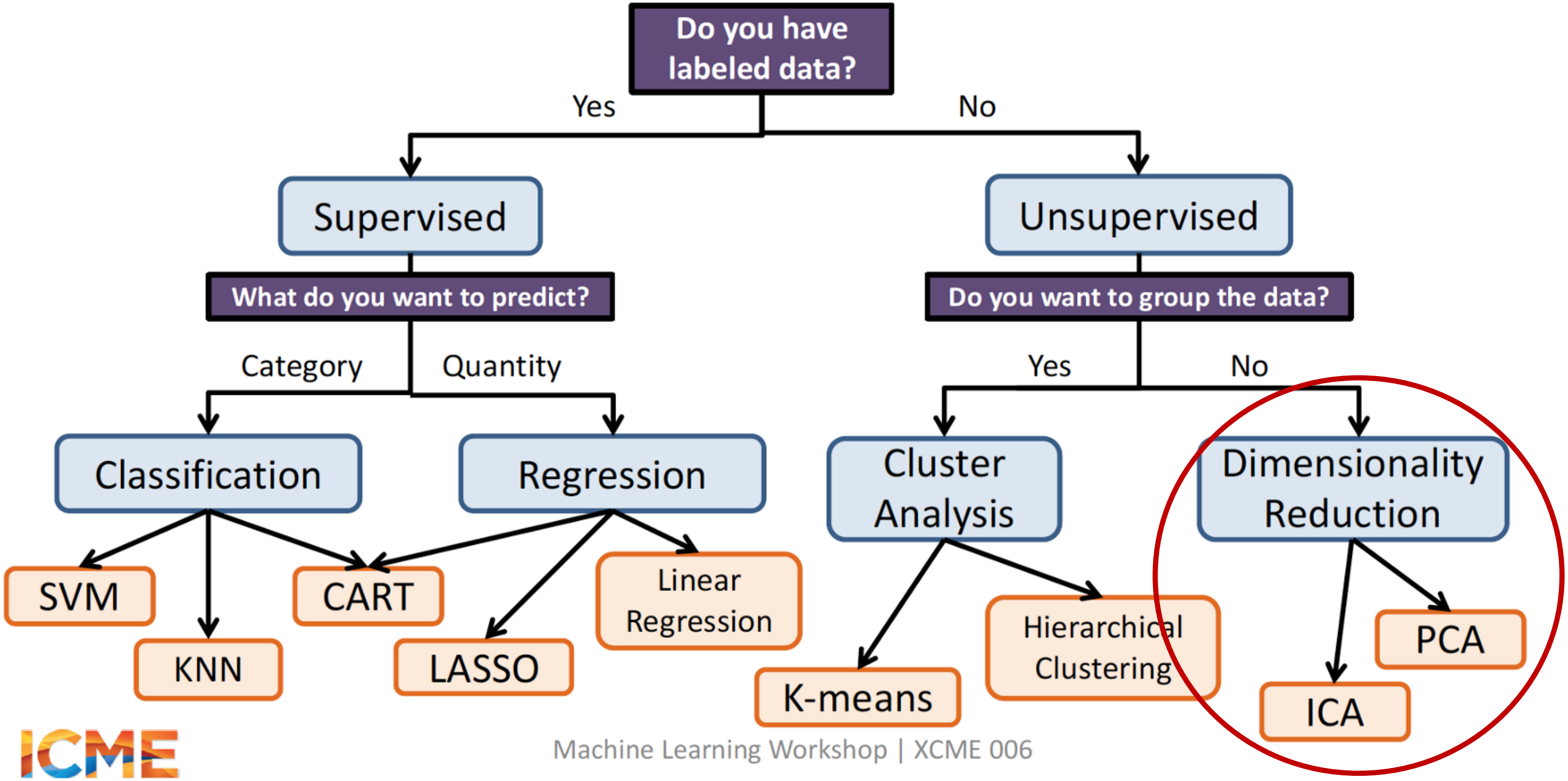
→ Don’t have a priori knowledge of expected relationships.

→ Extract knowledge from our dataset rather than impose a structure on it.

# Supervised versus unsupervised learning methods



https://www.researchgate.net/publication/319093376



ICME

Machine Learning Workshop | XCME 006

https://medium.com/@sandramoerch

# Curse of dimensionality and least squares regression

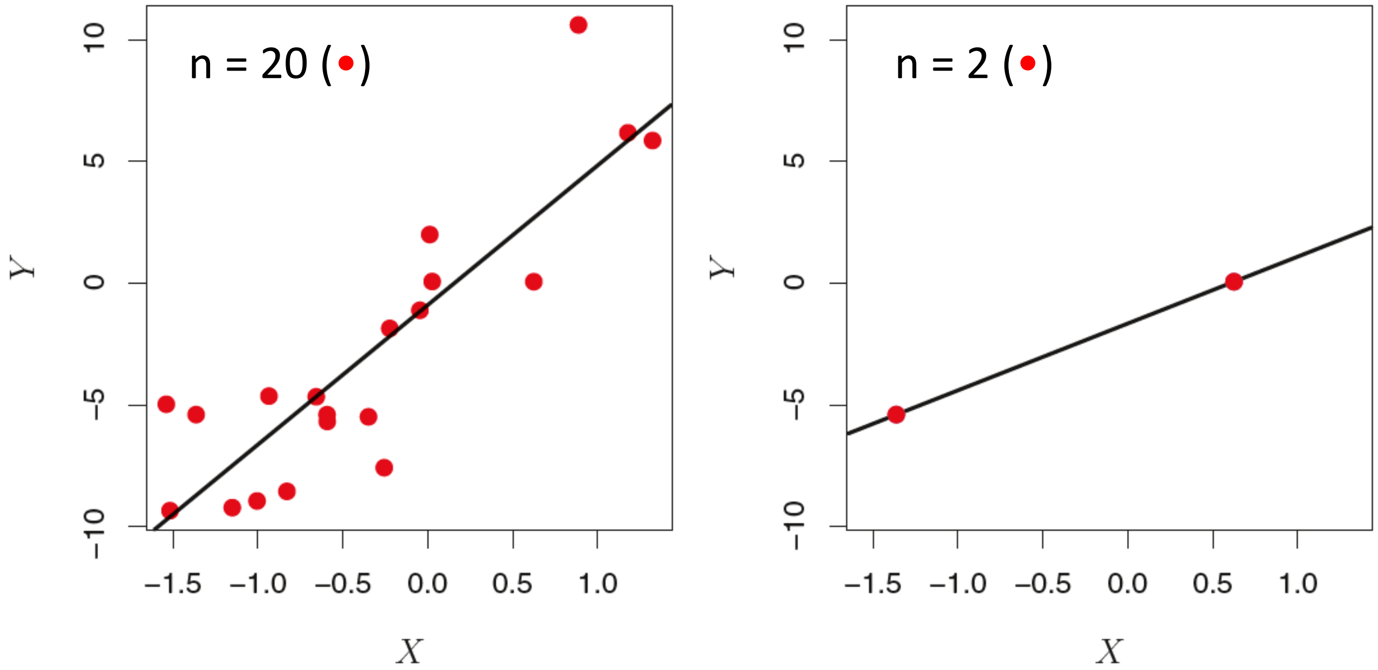
n = number of observations (samples, cell lines)

p = number of variables or features (expressed genes)

## n observations on two features (p)

Low dimension, n >> p

High dimension, n ~ p



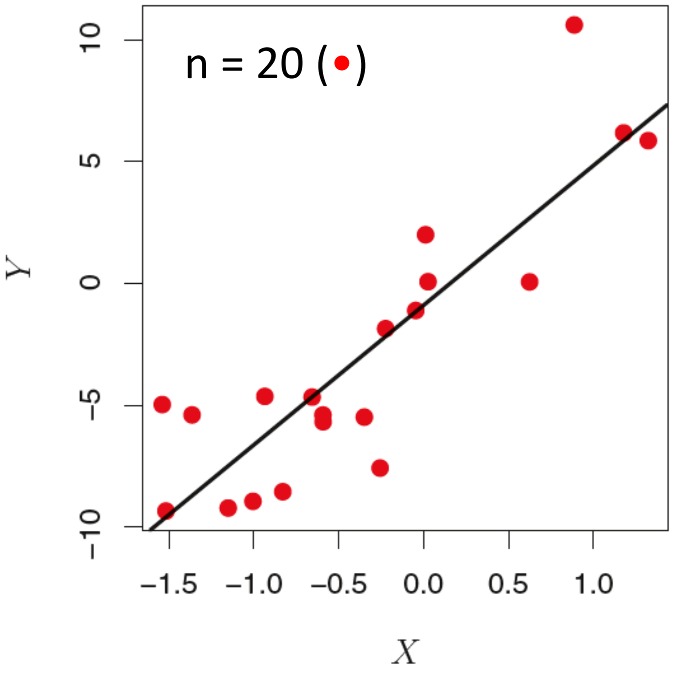
For least squares regression with n = p =2, the fit is **always** perfect: **Overfitting**.

G. James, D. Witten, T. Hastie, R, Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer 2013. ISBN 978-1-4614-7137-0

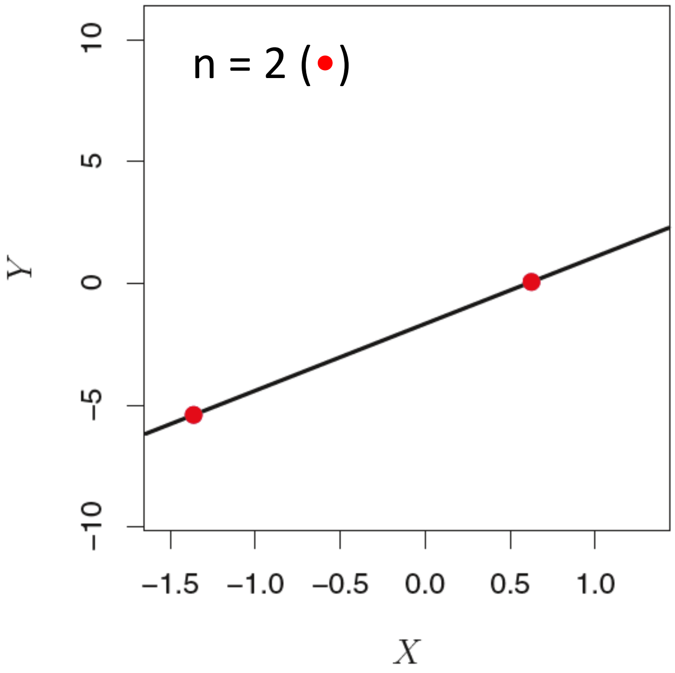
# Which model will do a better job of predicting independent data?

## n observations on two features (p)

Low dimension, n >> p

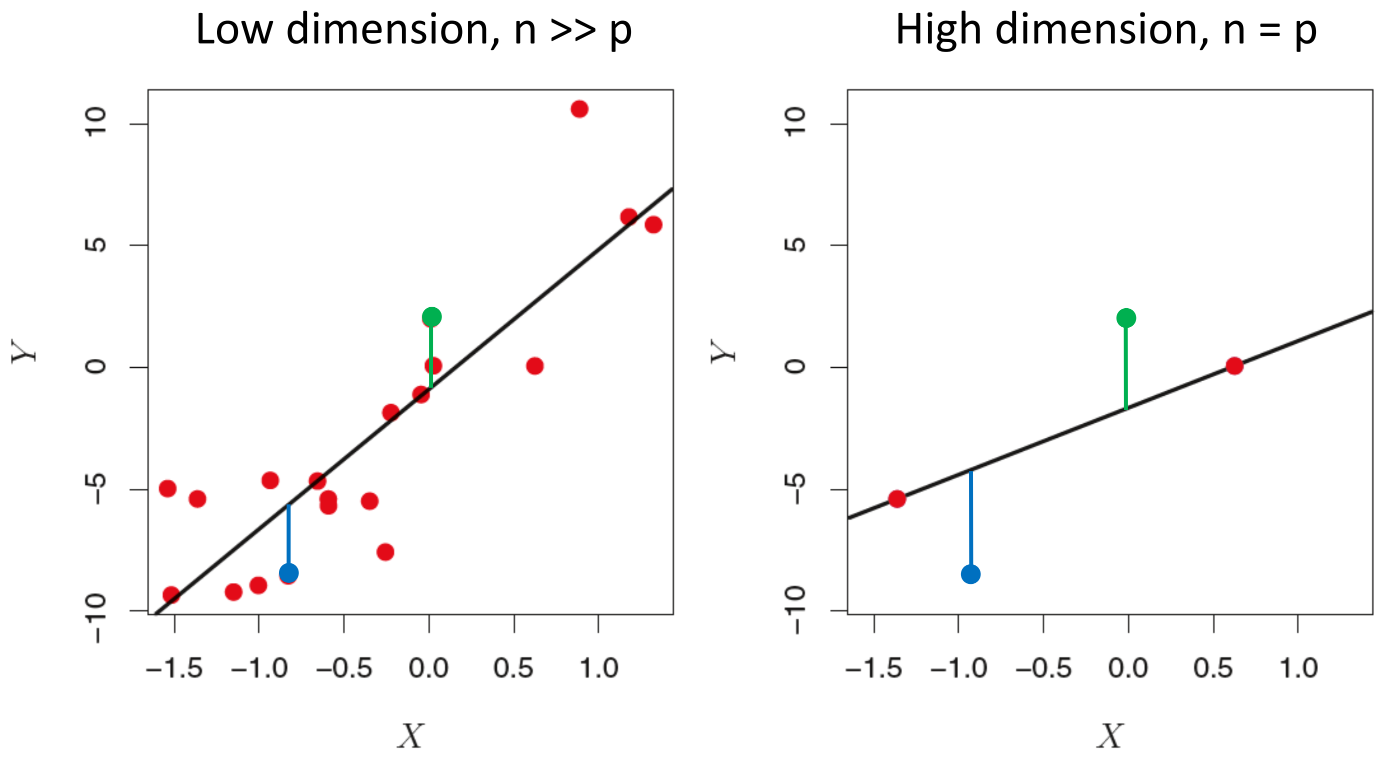


High dimension, n ~ p



More samples than features? OR More features than samples?

# Curse of dimensionality and least squares regression



Low dimension, n >> p

High dimension, n = p

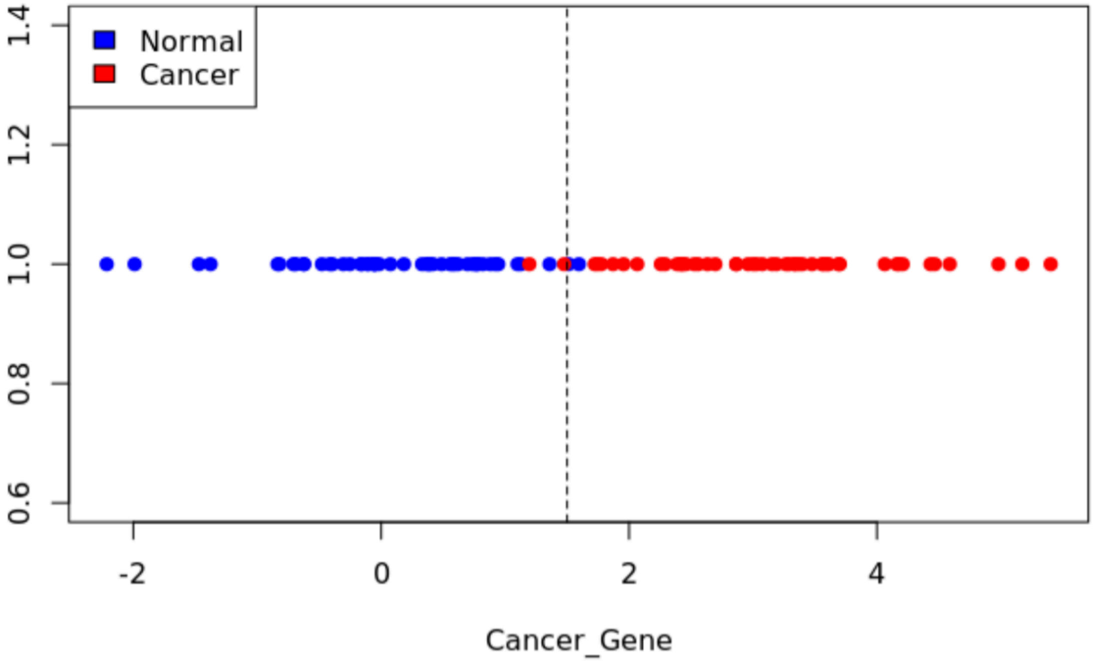
The model obtained when there are more samples than features is more predictive.

Overfitting leads to poor performance on an independent data set.

## Curse of dimensionality: spurious findings

Pretend there's a gene whose expression level correlates with disease state: higher expression indicates a greater likelihood of cancer. We can simulate this condition by sampling from two distributions, one for non-cancer (normal) and one for cancer.

```
normal <- rnorm(50, mean = 0, sd = 1)
cancer <- rnorm(50, mean = 3, sd = 1)
```

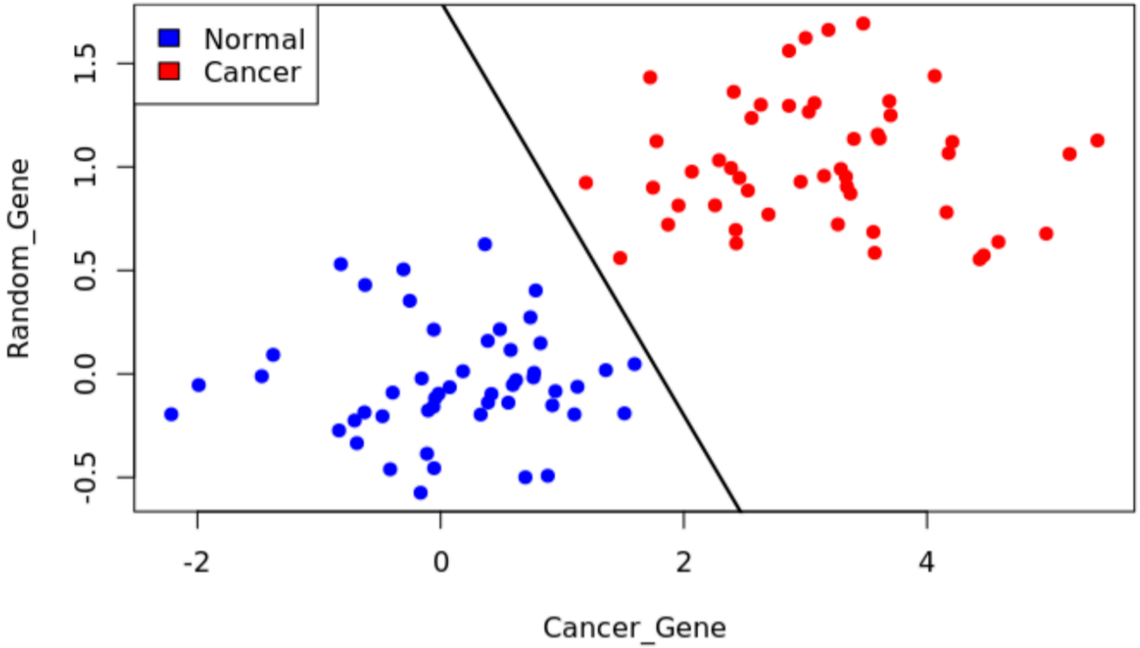


Because there is only 1 gene, this is a 1-dimensional problem. The optimal dividing line between states is not perfect, but separates the vast majority of samples.

## Curse of dimensionality: spurious findings

We can simulate a second, random gene and add it to the plot. We are now operating in 2 dimensions. We can easily find a dividing line that separates all points into cancer / non-cancer, even though only the first gene is a true biomarker.

```
random_gene <- c(rnorm(50, 0, .3), rnorm(50, 1, .3))
```



This is an example of how adding more features (dimensions) can lead to overfitting. The more features one adds, the easier it is to develop a model that over fits the data.

# Low dimensional data example

Number of observations (n) >> number of features (p)

*Develop a model to predict a patient’s blood pressure based on age, gender, and body mass index data from 5,000 patients.*

n = 5,000 (patients)

p = 3 (age, gender, body mass index)

Adding a few additional features, e.g. cholesterol and exercise levels, may improve predictive power of model.

# High dimensional data example

Number of observations (n) << number of features (p)

*Develop a model to predict a patient's blood pressure based on age, gender, body mass index data from 5,000 patients and 500K SNPs.*

n = 5,000 (patients)

p ~ 500K

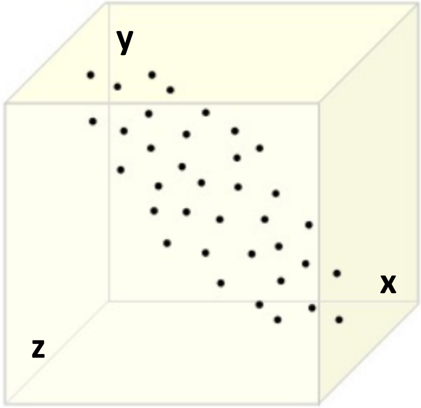
*Predict patient survival from mRNA gene expression, DNA methylation, microRNA and copy number alterations for breast invasive carcinoma samples collected by TCGA.*

n = 400

p ~ 40K

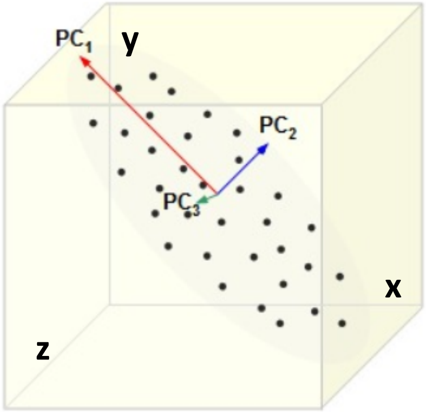
**Dimensionality Reduction** is the process of compressing data into something that captures the essence of the original data

Principal component analysis (PCA) reduces dimensionality, retaining most of the variation



We have a three-dimensional dataset: x-, y-, z-components of spatial coordinates of atoms in a protein structure.

We note that the spread of coordinate values is greater in some dimensions than others.



Principal components are directions along which variation in the data is maximal.

PC1 carries most of the variation.

PC2 carries most of the variation in a direction perpendicular to PC1.

PC3 is perpendicular to PC1 and PC3.

Which PC is the “most important”?

# Principle component analysis for gene expression data.

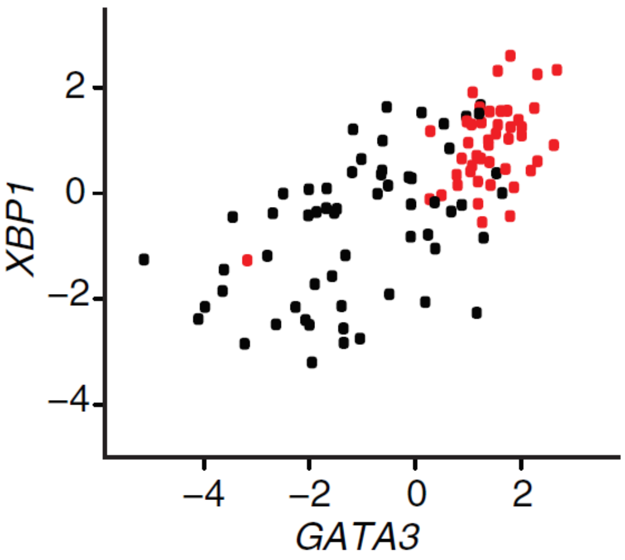
Consider a **simple** problem:

Visualizing the expression levels of **two genes**, XBP1 and GATA3, in 105 breast tumor samples.

Only 2 genes out of 28,000! *Note: We can also do this analysis with transcript counts.*

We assume the data are normalized and have zero mean

Estrogen receptor classification of breast cancer samples, **ER+** and **ER-**



**XBP1**: X-box-binding protein

**GATA3**: Trans-acting T-cell-specific transcription factor

Each dot represents a sample with the coordinates log2(expression level) of XBP1 versus GATA3.

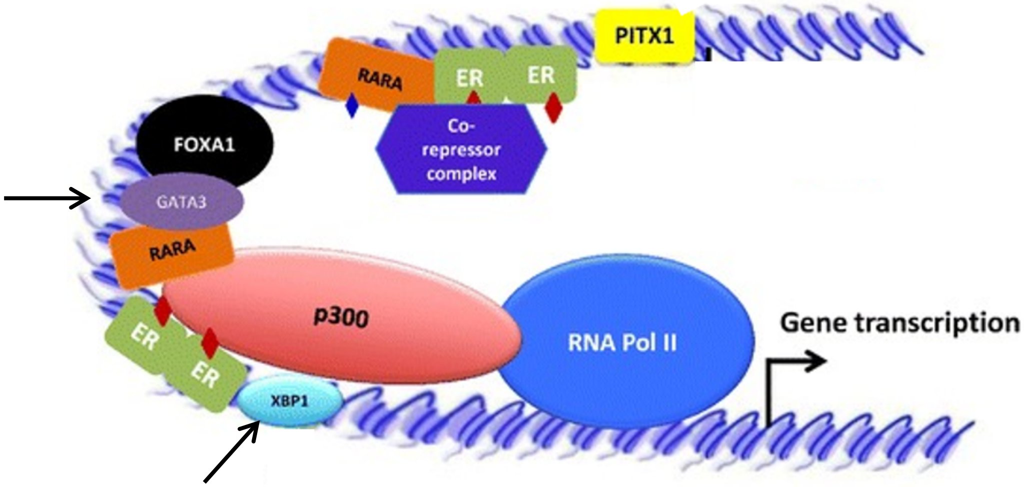
The samples are colored **ER+** and **ER-**

**PMID: 18327243**

## Biological aside 1: Estrogen receptor (ER) status, ER+ and ER-

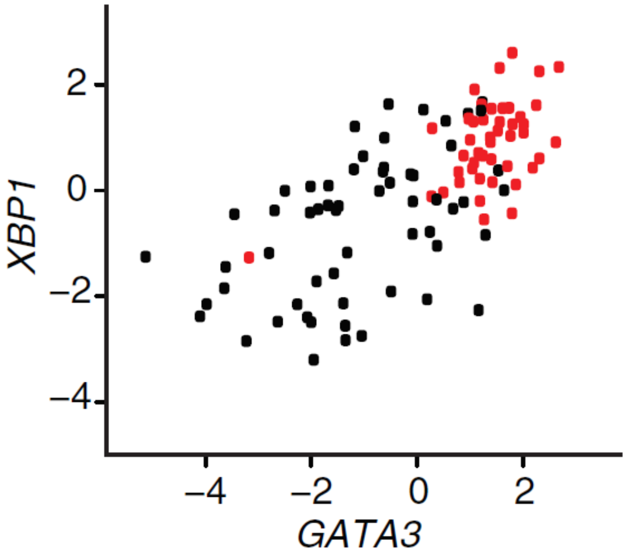
- 70% of all newly diagnosed breast cancers express ERα (ER+).
- Antiestrogens: tamoxifen (TAM) and fulvestrant are widely used.
- Resistance, either de novo or acquired, limits their curative potential.
- More die from ERα+ breast cancer than from any other breast cancer subtype.

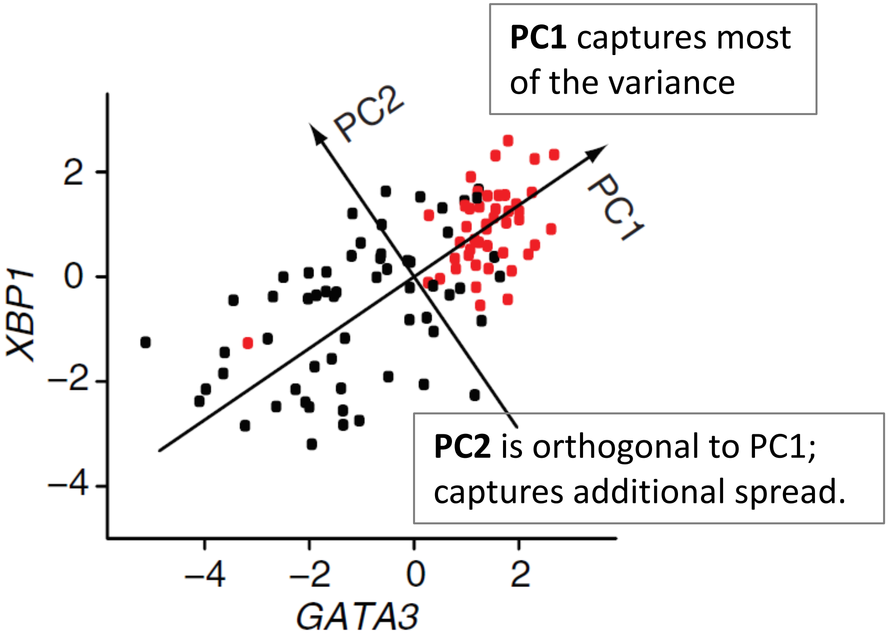
## Biological aside 2: GATA3 and XBP1 and Estrogen-mediated transcription in breast carcinoma cell lines



- FOXA1, GATA3, and ER complex regulates estrogen transcription.
- XBP1 promotes ER transcriptional activity.

# Principle component analysis for gene expression data.





**PC1** captures most of the variance

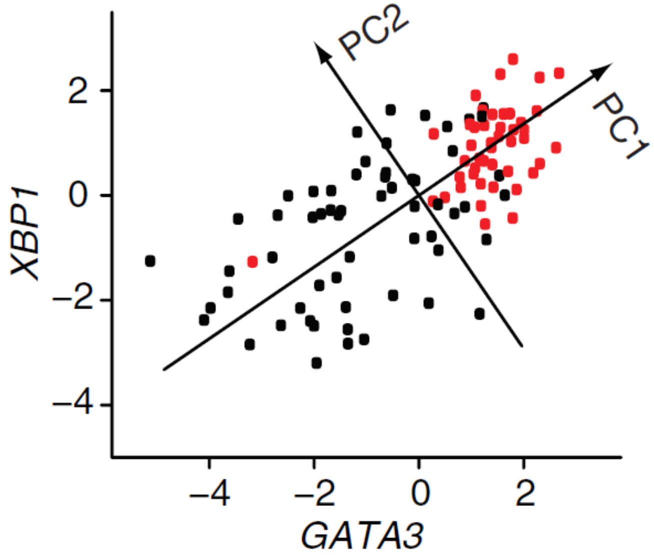
**PC2** is orthogonal to PC1; captures additional spread.

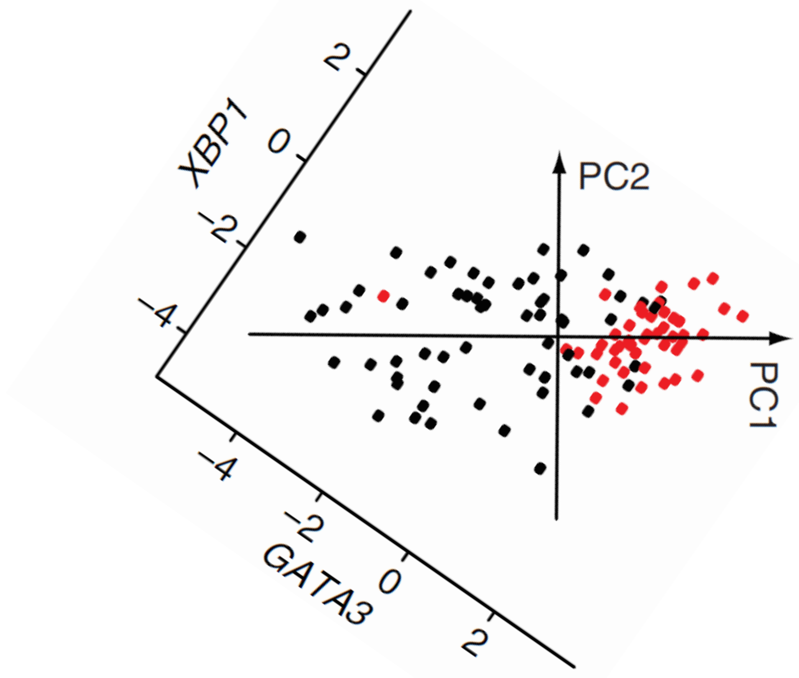
**PC1 = 0.89 × GATA3 + 0.46 × XBP1**

0.89 and 0.46 are ***loading values*** that tell us how important each feature (gene) is in PC1.

**PMID: 18327243**

# Principle component analysis: Loading values





PC1 = **0.89** × GATA3 + **0.46** × XBP1

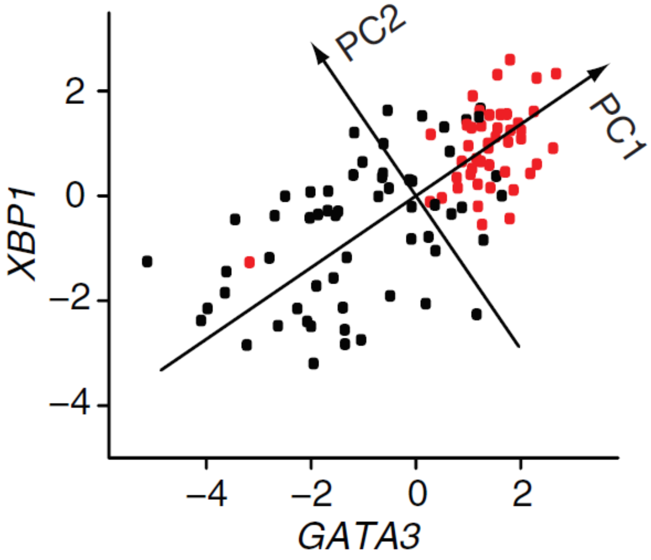
PC2 = **-0.61** x GATA3 + **0.80** x XBP1

**Loading values** that tell us how important each feature (gene) is in each principal component.

Note that the samples have zero average expression, and 0.89^2 + 0.46^2 = 1.

PMID: 18327243

# **Principle component analysis** for gene expression data.



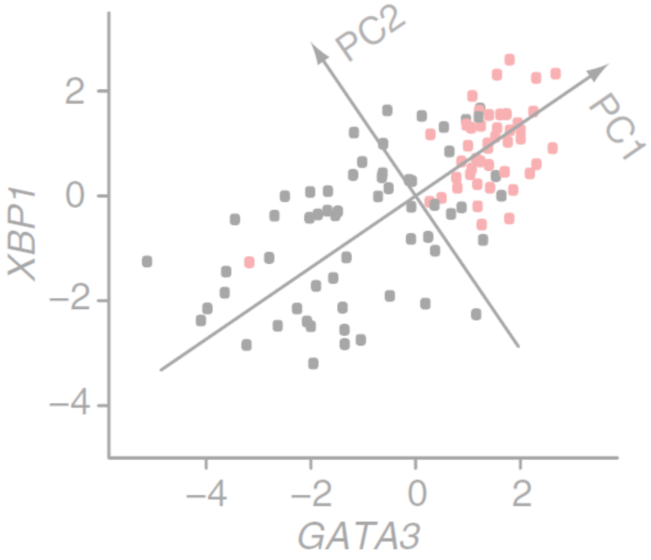
Each component can be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the samples when projected onto the component.

So we have two PCs for 2D data.

Where is the dimensionality reduction?!

**PMID: 18327243**

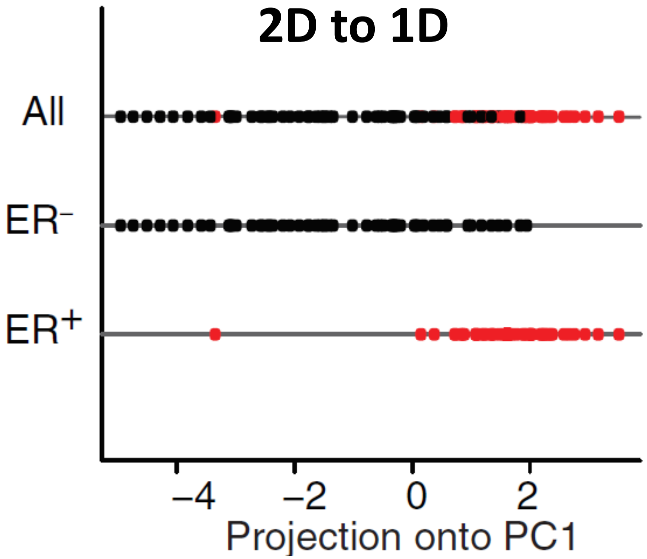
# **Principle component analysis** for gene expression data.



Each component can be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the samples when projected onto the component.

So we have two PCs for 2D data.

Where is the dimensionality reduction?!



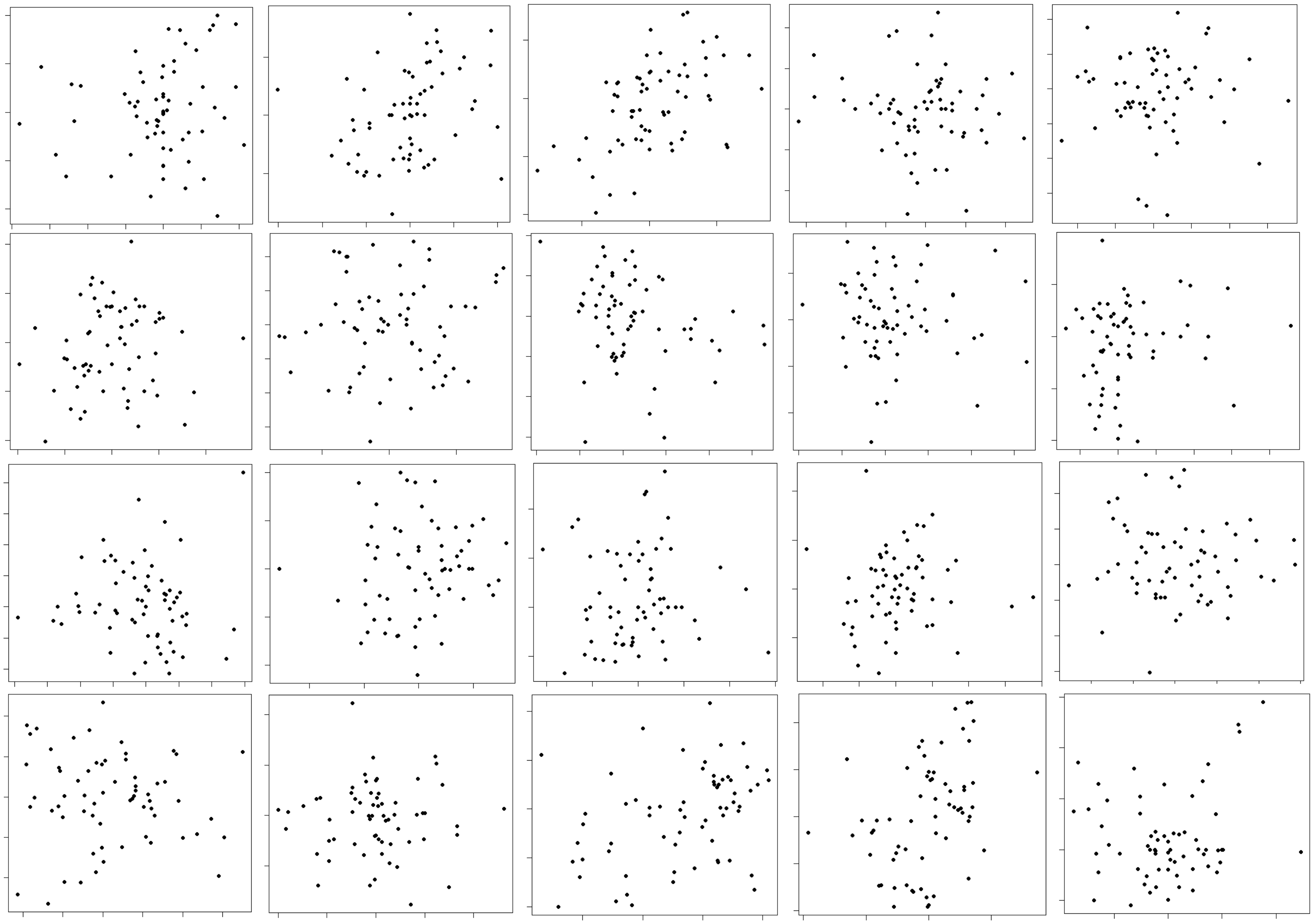
*The projection of the data to 1D retains the separation with respect to estrogen receptor status, **ER+** and **ER-**.*

**PC1 =** 0.89 × **GATA3** + 0.46 × **XBP1**

PC1 can be considered a "***metagene***"

How do we visualize the expression of 8,500 genes (features, p) across 105 breast cancer samples (n).

***Do we have to consider all possible p(p-1)/2 ~ 36M 2D scatterplots?***



How do we visualize the expression of 8,500 genes (features, p) across 105 breast cancer samples (n).

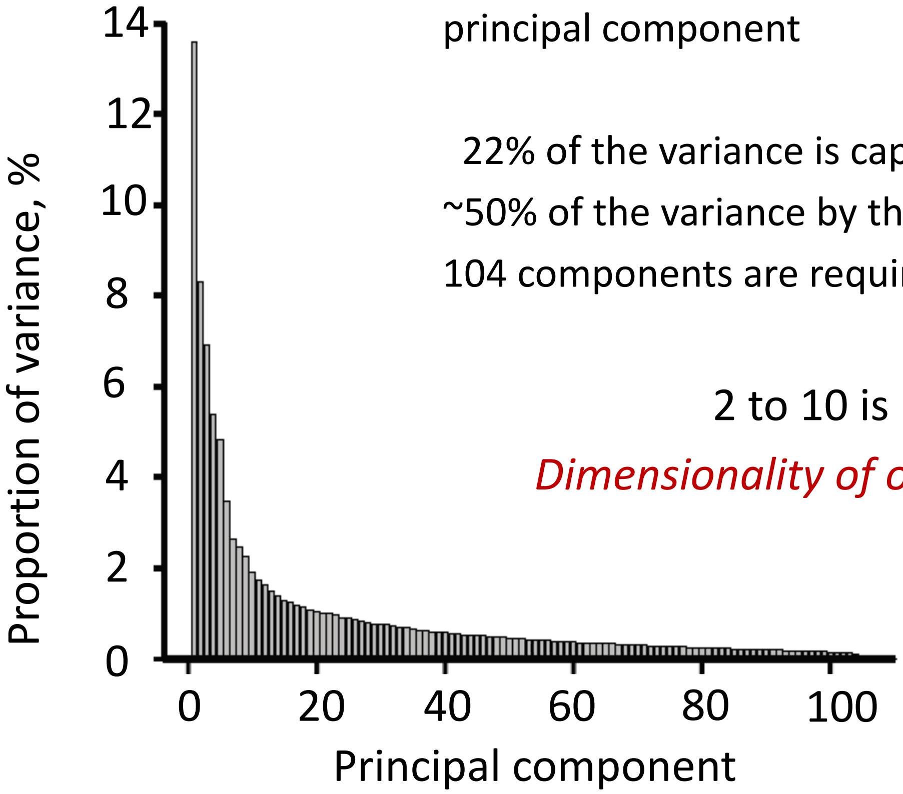
***Do we have to consider all possible p(p-1)/2 ~ 36M 2D scatterplots?***

***No! Let R worry about the computation.***

# Principle component analysis: n =105 breast tumor samples, p = 8,500 genes

We get *min*(n-1, p) principal components, so 104 PCs.

**But a small subset of these contains most of the information.**



Proportion of the variance present in all genes contained within each principal component

22% of the variance is captured by PC1 and PC2

~50% of the variance by the first 10 components

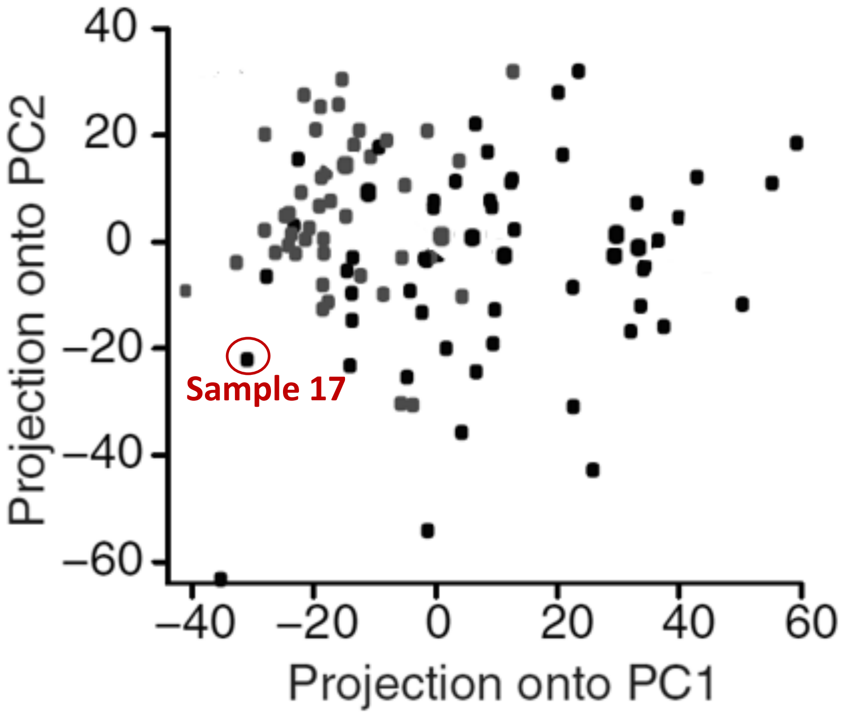
104 components are required to retain the original variance.

2 to 10 is much less than 8,500

*Dimensionality of our problem is greatly reduced.*

**PMID: 18327243**

105 samples are plotted in 2D using their projections onto PC1 and PC2



What does a point on this plot mean?

# 105 samples with 2 genes are plotted in 2D as projections onto PC1 and PC2

Sample 17:

PC1 loading vector

$$[\phi(1,1), \phi(2,1)]^T = \begin{pmatrix} 0.89 \\ 0.46 \end{pmatrix}$$

PC1 = **0.89** · GATA3 + **0.46** · XBP1

$$pc(17,1) = \phi(1,1) \cdot \chi(17,1) + \phi(2,1) \cdot \chi(17,2)$$

$\chi(17,1), \chi(17,2)$ are the expression values for sample 17

PC2 = **-0.61** · GATA3 + **0.80** · XBP1

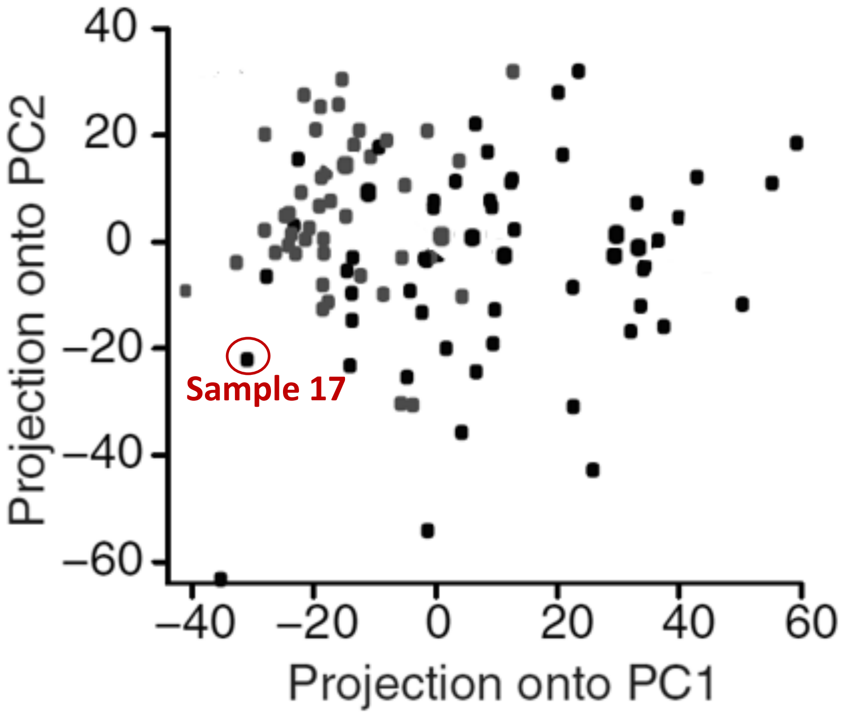
PC2 loading vector

$$[\phi(1,2), \phi(2,2)]^T = \begin{pmatrix} -0.61 \\ 0.80 \end{pmatrix}$$

$$pc(17,2) = \phi(1,2) \cdot \chi(17,1) + \phi(2,2) \cdot \chi(17,2)$$

pc(17,1) and pc(17,2) are PC scores for sample 17

# 105 samples with 8,500 genes are plotted in 2D as projections onto PC1 and PC2



What does a point on this plot mean?

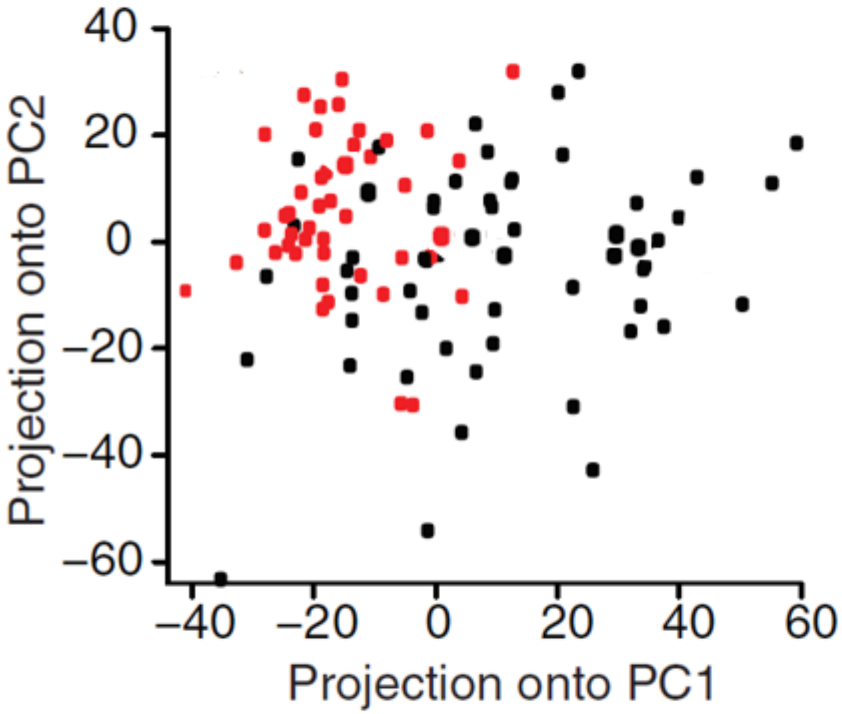
{χ} original data (expression levels of genes

{ϕ} loading values from PCA

$$pc(17,1) = \phi(1,1)\cdot\chi(17,1) + \phi(2,1)\cdot\chi(17,2) + \phi(3,1)\cdot\chi(17,3) + \ldots\ldots + \phi(p,1)\cdot\chi(17,p)$$

$$pc(17,2) = \phi(1,2)\cdot\chi(17,1) + \phi(2,2)\cdot\chi(17,2) + \phi(3,2)\cdot\chi(17,3) + \ldots\ldots + \phi(p,2)\cdot\chi(17,p)$$

# 105 samples are plotted in 2D using their projections onto PC1 and PC2

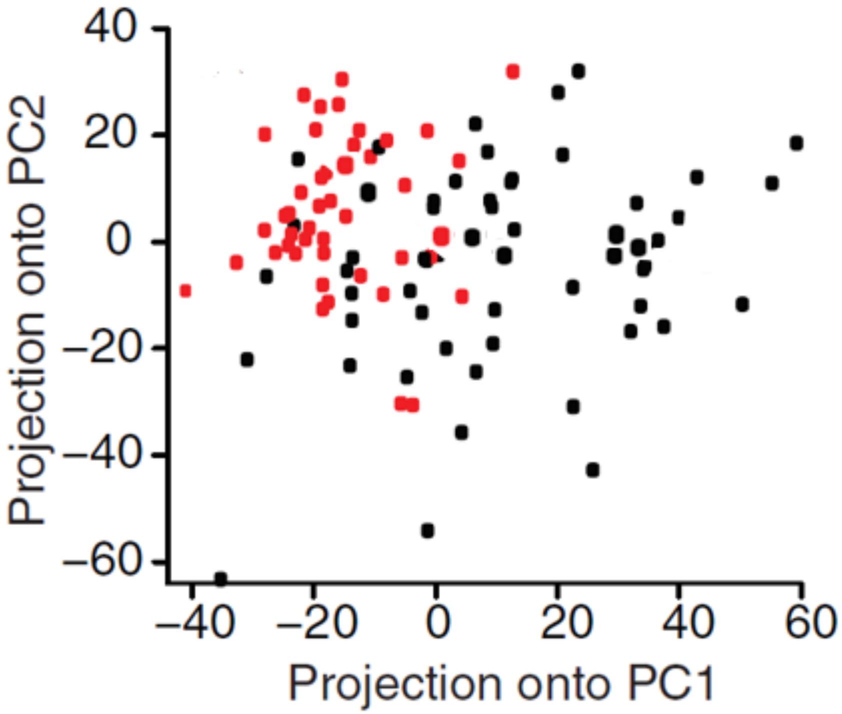


**ER+** and **ER-** samples are projected onto PC1 and PC2, which are linear combinations of the expression levels of genes.

**Do PC1 and PC2 contain relevant information?**

**PMID: 18327243**

105 samples are plotted in 2D using their projections onto PC1 and PC2



**ER+** and **ER-** samples are projected onto PC1 and PC2, which are linear combinations of the expression levels of genes.

**Do PC1 and PC2 contain relevant information?**

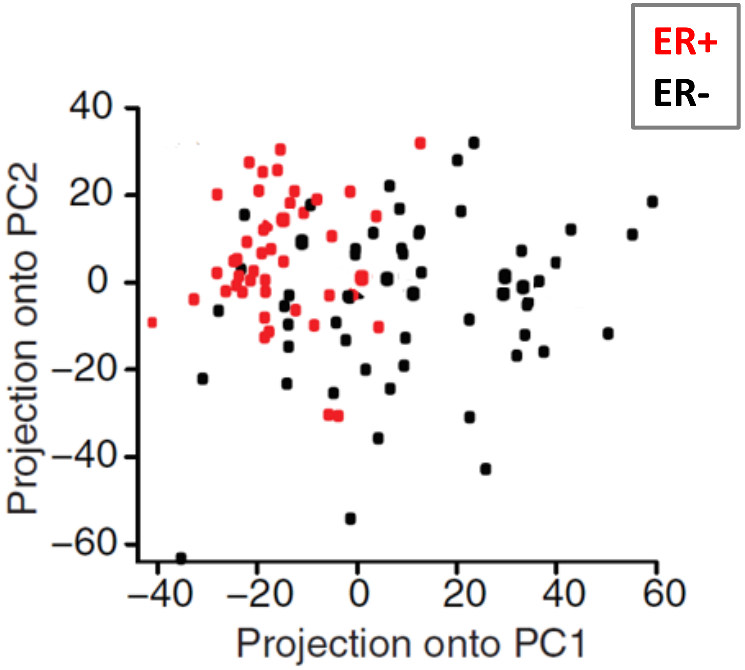
**ER+** and **ER-** are still separated though not strictly distinct clusters

**PMID: 18327243**

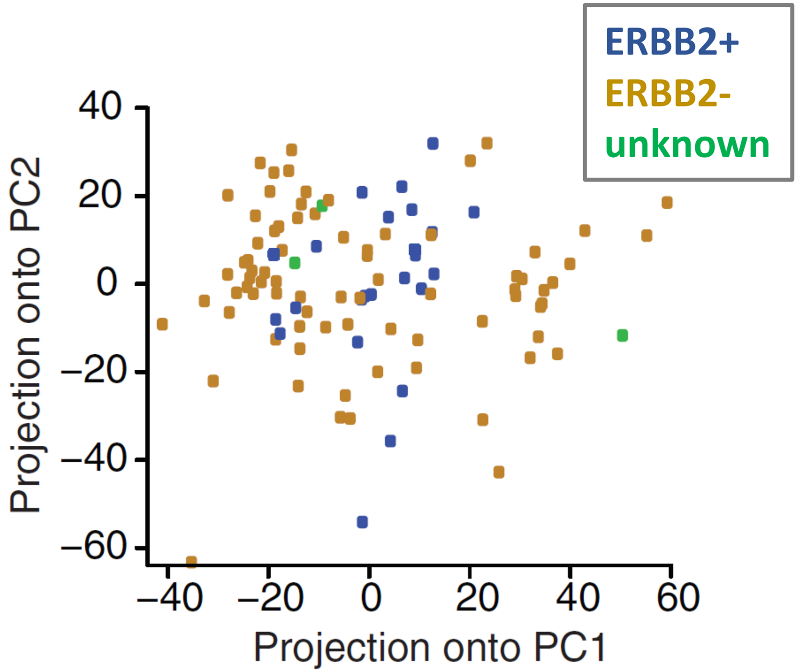
What are some other clinical or phenotypic characteristics of breast cancer?

1. Estrogen receptor status, **ER+** and **ER-**

# PCA and phenotypic relevance



PCA does not generate
two separate clusters



ERBB2 copy number info is lost
in reduced dimension.

***PCA emphasizes sources of greatest variability,***
***but doesn't guarantee interpretability.***

**PMID: 18327243**

# PCA: Limitations

- **Linearity** : PCA assumes that the principle components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.

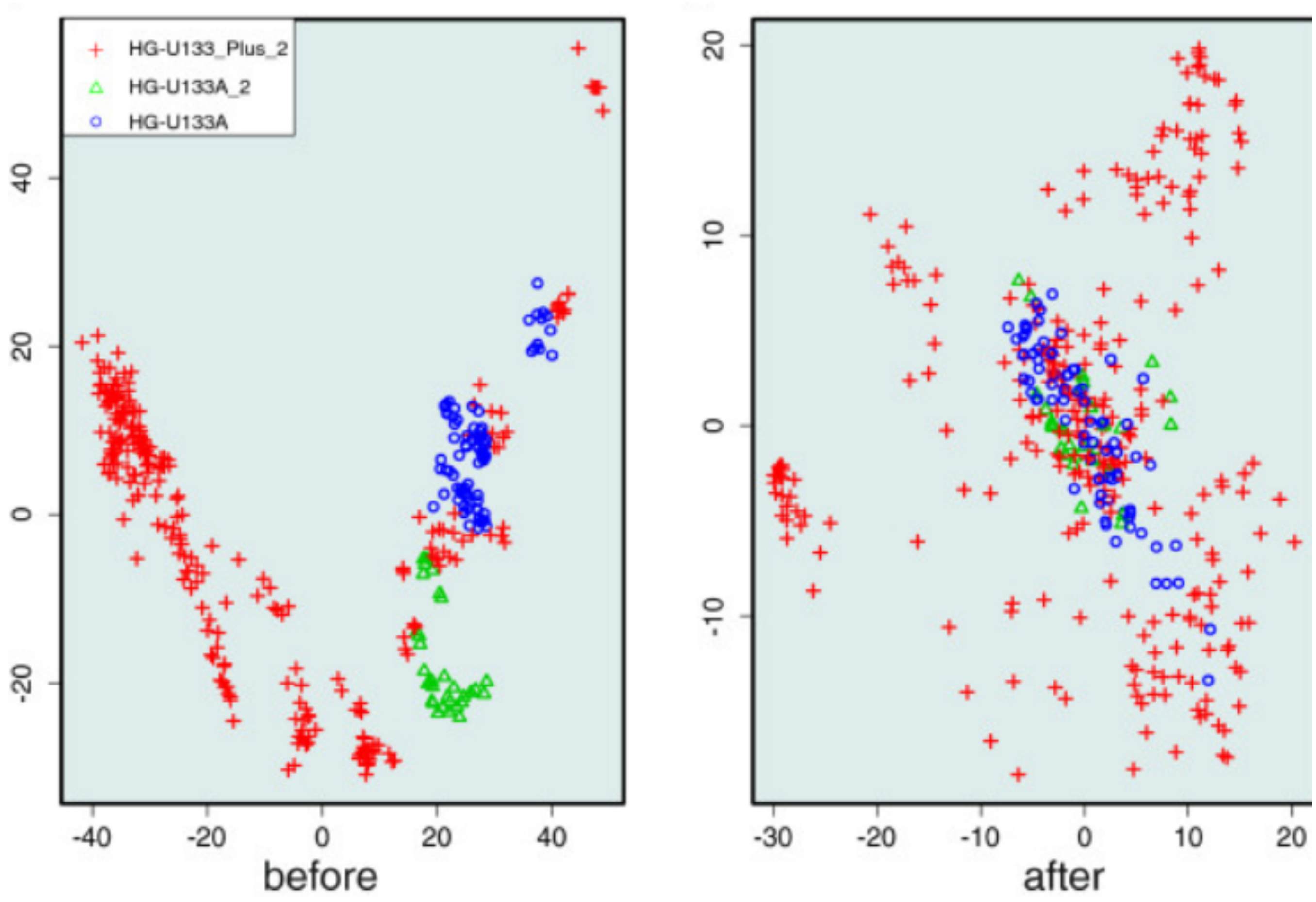
- **Large variance implies more structure**: PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principle components, while low variance axes are treated as noise. *Centering and scaling.*

- **Orthogonality**: PCA assumes that the principle components are orthogonal.

## Principal Component Analysis and batch effect correction algorithms

● Integration and analysis of new high-throughput gene-expression and proteomic-profiling data.

● Technical heterogeneity or batch effects (different experiment times, handlers, reagent lots, etc.) are a major hurdle.



PMID: 24528953

# PCA, TCGA, and histology

**Two major histological classes of solid tumors**:

1) Adenocarcinomas: from cells that form glands
2) Squamous cell carcinomas: from cells that line organs

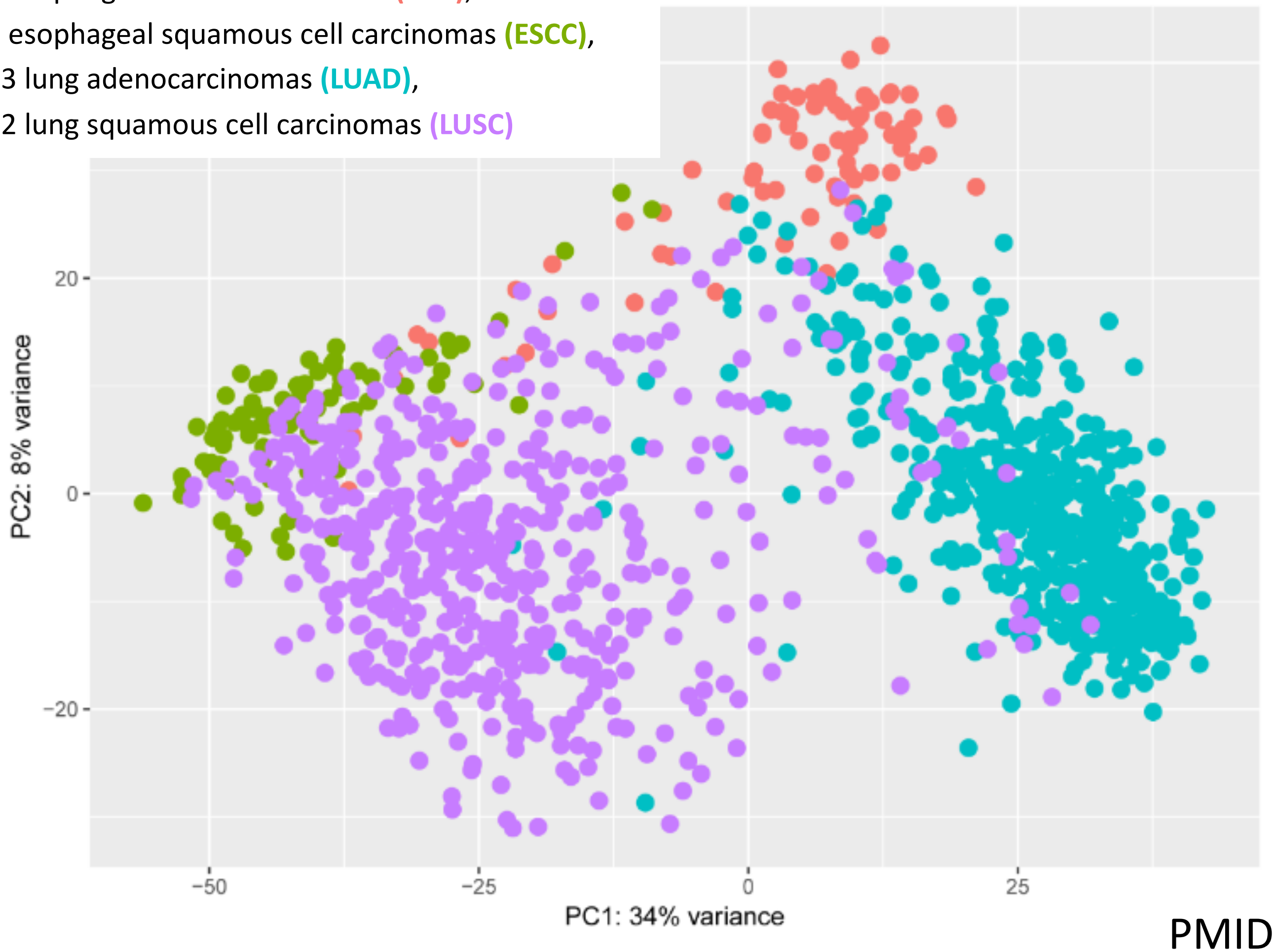
From TCGA:

80 esophageal adenocarcinomas **(EAC)**,

81 esophageal squamous cell carcinomas **(ESCC)**,

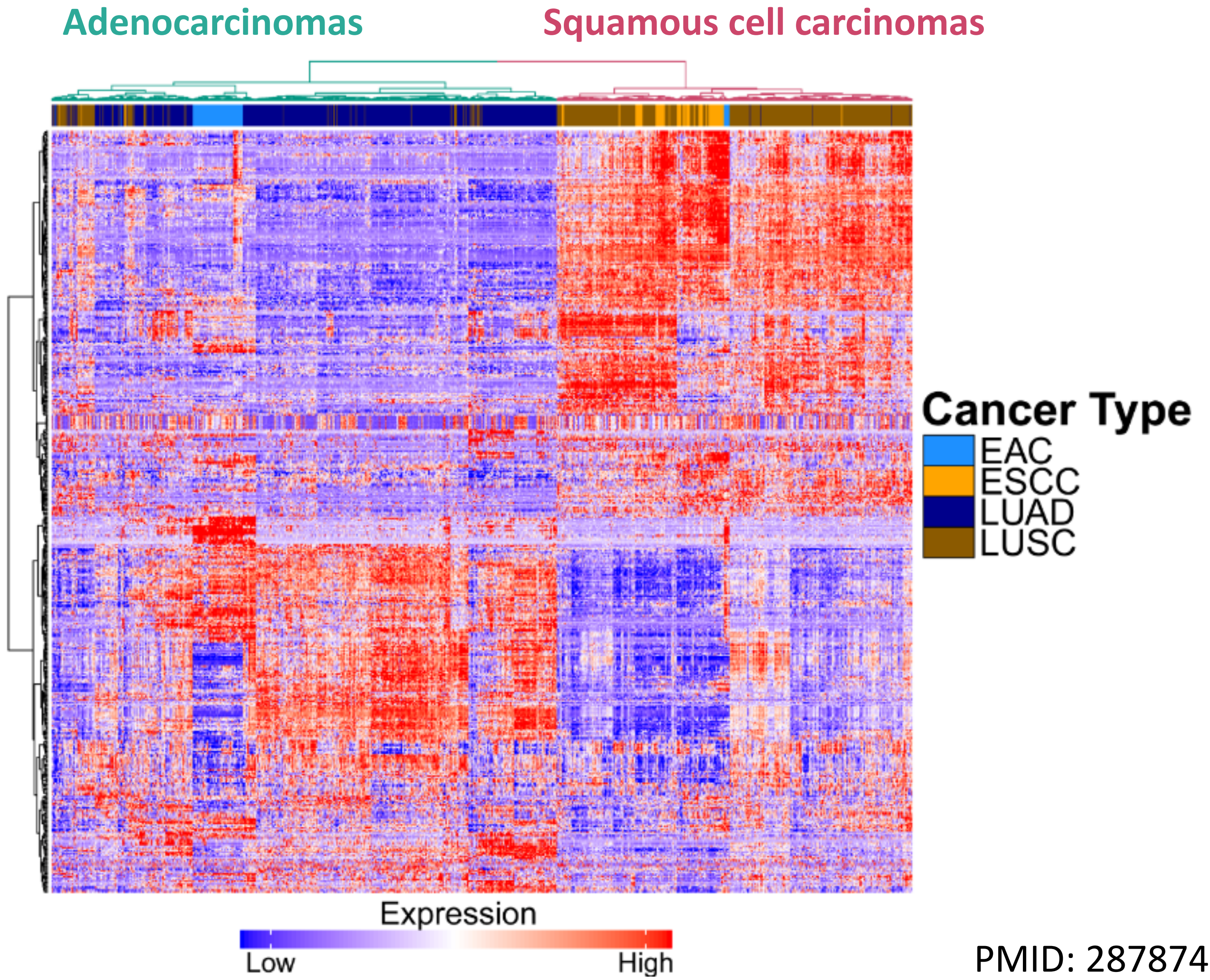
533 lung adenocarcinomas **(LUAD)**,

502 lung squamous cell carcinomas **(LUSC)**



PMID: 28787442

Heatmap of differentially expressed genes between esophageal cancers and lung cancers, with hierarchical clustering.



PMID: 28787442

Diana Murray

dm527@cumc.columbia.edu

Department of Systems Biology
Columbia University Medical Center