Data Mining In Modern Astronomy Sky Surveys:

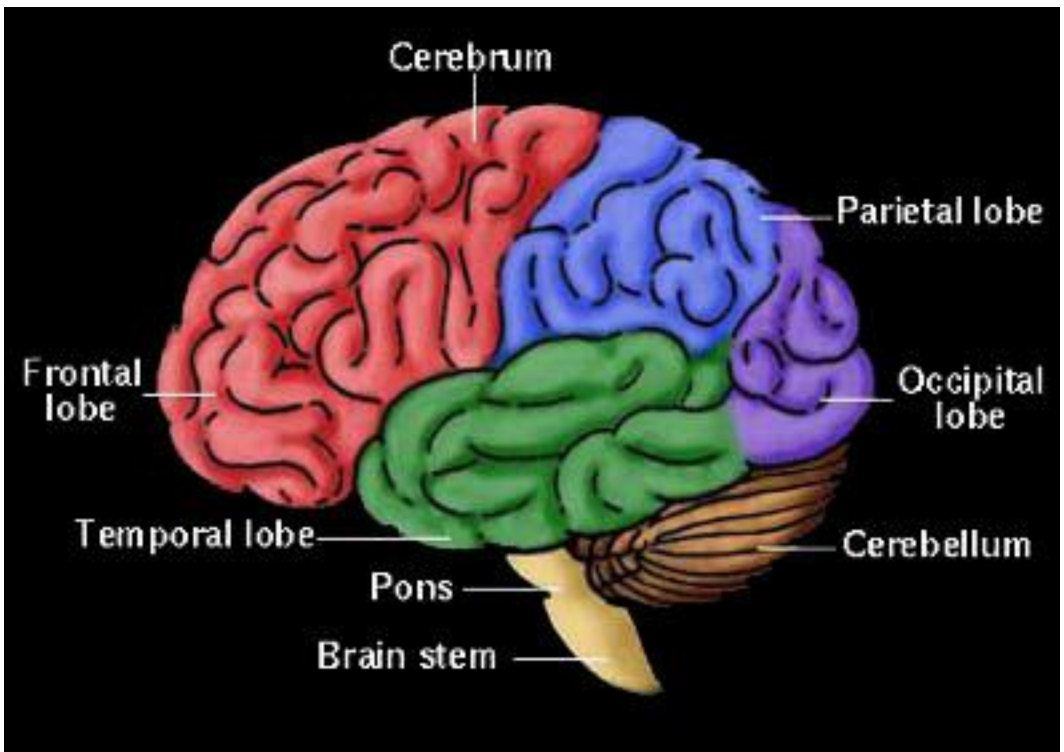
Concepts in Machine Learning, Unsupervised Learning & Astronomy Applications

Ching-Wa Yip

cwyip@pha.jhu.edu; Bloomberg 518

Human are Great Pattern Recognizers

- Sensors: look, smell, touch, hear
- Computation: 100 billions (10^{11}) neurons



Estimated 300 million pattern recognizers (*How to create a mind,* Kurzweil)

Machine Learning

- We want computers to perform tasks.
- It is difficult for computers to “learn” like the human do.
- We use algorithms:
  - Supervised, e.g.:
    - Classification
    - Regression
  - Unsupervised, e.g.:
    - Density Estimation
    - Clustering
    - Dimension Reduction

Machine Learning

- We want computers to perform tasks.
- It is difficult for computers to “learn” like the human do.
- We use algorithms:
  - Supervised, e.g.:
    - Classification
    - Regression
  - Unsupervised, e.g.:
    - Density Estimation
    - Clustering
    - Dimension Reduction

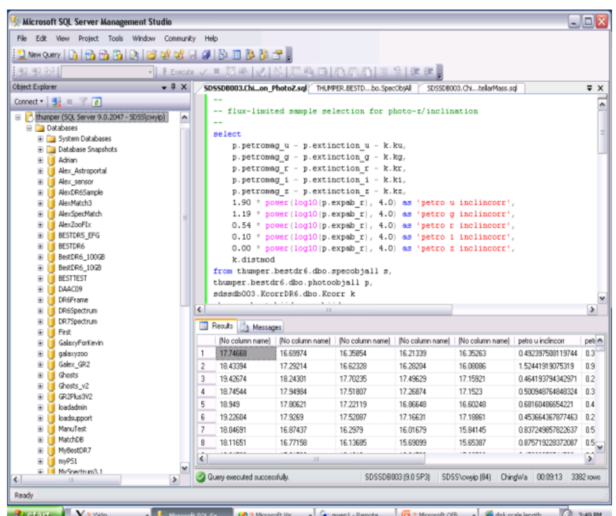
From Data to Information

- We don’t just want data.
- We want information from the data.

Information ← **Database** ← **Sensors**

Machine Learning

- *Not Data Mining by Human*



Expert could be Biased (*Thinking Fast and Slow*, Kahneman)

challenges economic theory, according to which prices should reflect all the available information, including the weather. Ashenfelter's formula is extremely accurate—the correlation between his predictions and actual prices is above .90.

Why are experts inferior to algorithms? One reason, which Meehl suspected, is that experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity. Simple combinations of features are better. Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not. According to Meehl, there are few circumstances under which it is a good idea to substitute judgment for a formula. In a famous thought experiment, he described a formula that predicts whether a particular person will go to the movies tonight and noted that it is proper to disregard the formula if information is received that the individual broke a leg today. The name "broken-leg rule" has stuck. The point, of course, is that broken legs are very rare—as well as decisive.

Another reason for the inferiority of expert judgment is that humans are incorrigibly inconsistent in making summary judgments of complex information. When asked to evaluate the same information twice, they fre-

Expert could be Biased (*Thinking Fast and Slow*, Kahneman)

challenges economic theory, according to which prices should reflect all the available information, including the weather. Ashenfelter's formula is extremely accurate—the correlation between his predictions and actual prices is above .90.

Why are experts inferior to algorithms? One reason, which Meehl suspected, is that experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity. Simple combinations of features are better. Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not. According to Meehl, there are few circumstances under which it is a good idea to substitute judgment for a formula. In a famous thought experiment, he described a formula that predicts whether a particular person will go to the movies tonight and noted that it is proper to disregard the formula if information is received that the individual broke a leg today. The name "broken-leg rule" has stuck. The point, of course, is that broken legs are very rare—as well as decisive.

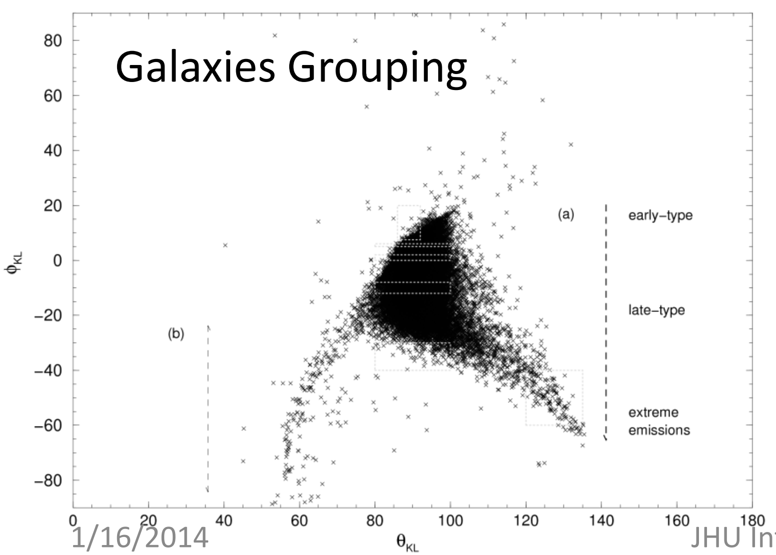
Another reason for the inferiority of expert judgment is that humans are incorrigibly inconsistent in making summary judgments of complex information. When asked to evaluate the same information twice, they fre-

Applications of Unsupervised Learning

Consumer Clustering

Human Network Analysis

Galaxies Grouping



and more…

Basic Concepts in Machine Learning

- Label and Unlabeled data
- Datasets: training set and test set
- Feature space
- Distance between points
- Cost Function (or called Error Function)
- Shape of the data distribution
- Outliers

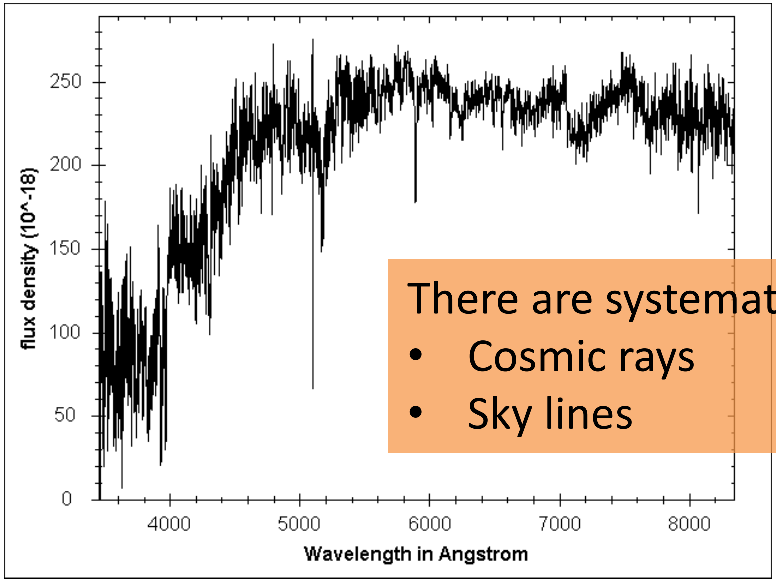
Basic Concepts in Machine Learning

- Label and Unlabeled data
- Datasets: training set and test set
- Feature space
- Distance between points
- Cost Function (or called Error Function)
- Shape of the data distribution
- Outliers

Feature Space

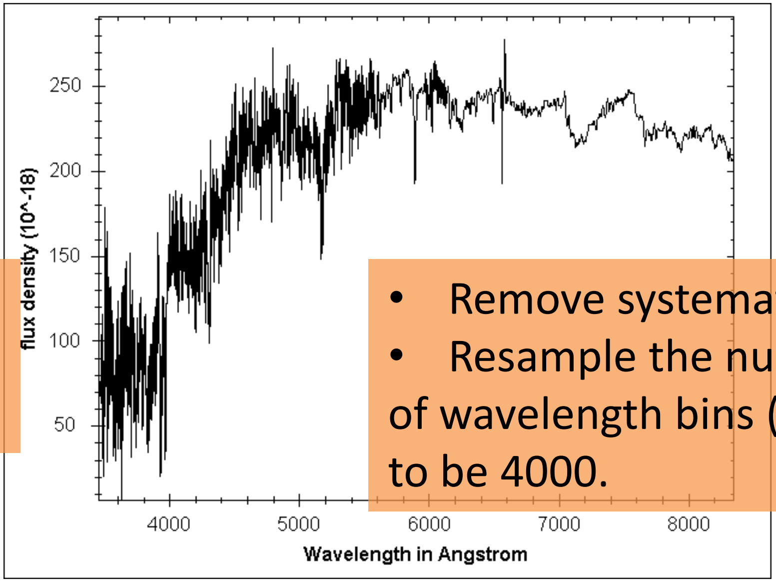
- The raw data may not be immediately suitable for pattern recognition.
- Feature space is spanned by the resultant data after pre-processing the raw data.
- The data usually N points (or vectors), each has M variables (or components).

Galaxy Spectrum: Raw Data



There are systematics:
- Cosmic rays
- Sky lines

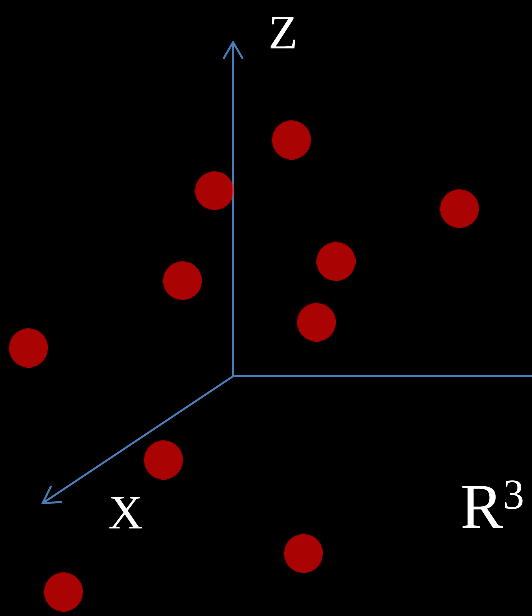
Galaxy Spectrum: After Pre-processing



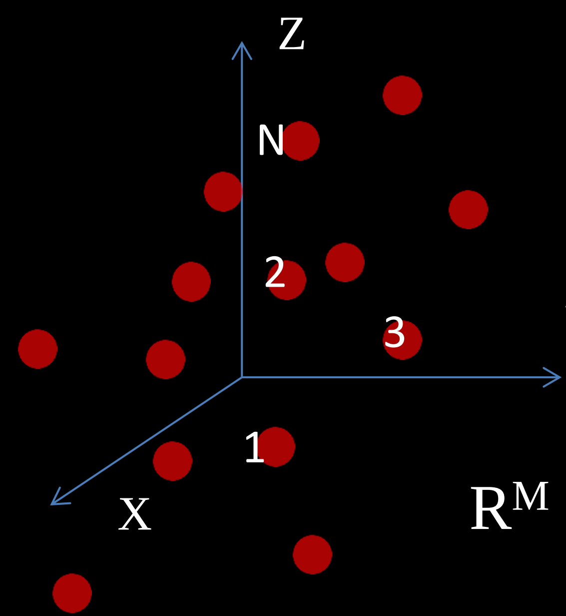
- Remove systematics.
- Resample the number of wavelength bins (M) to be 4000.

Intuition to High-Dimensional Data: N points, M components

N = 10, M = 3



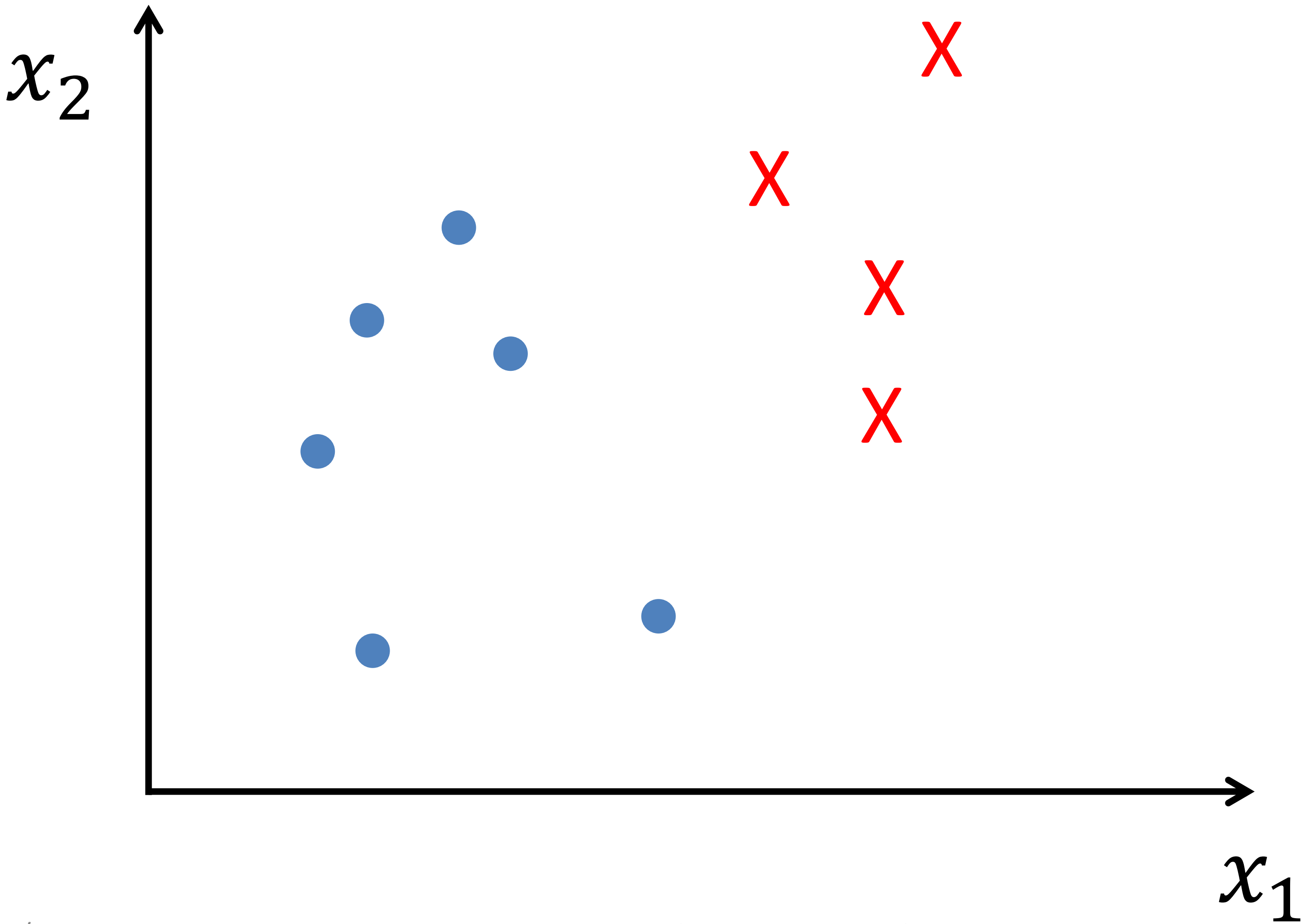
N, M



Unlabeled vs. Labeled Data

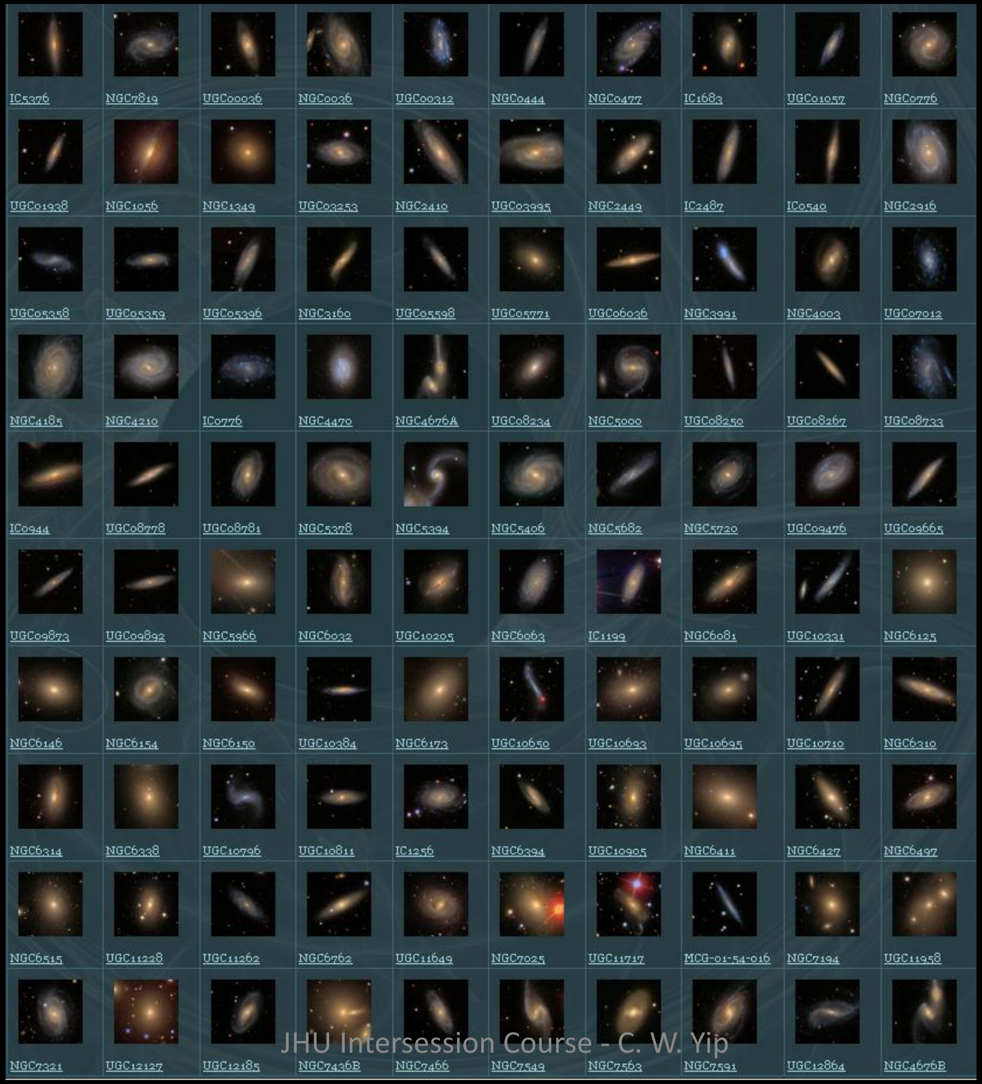
- Labeled data has an extra "label" compared with the unlabeled.
- Data labeling can be expensive (e.g., human expert).

Labeled Data



Unlabeled Data

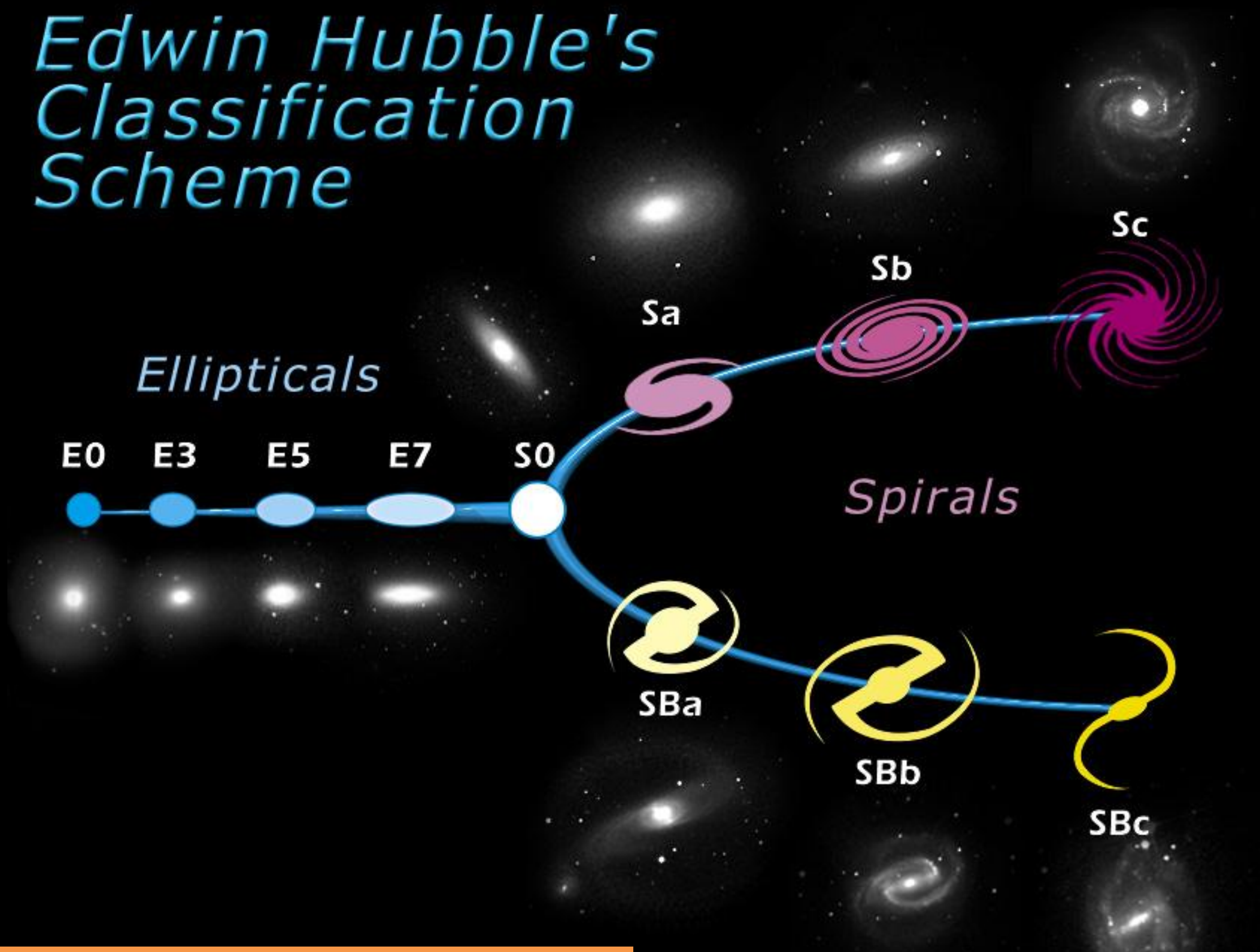
Galaxy Images are Unlabeled Data (well, before Labeling)



(SDSS images of CALIFA DR1; Husemann et al. 2012, Sanchez et al. 2012, Walcher et al. in prep.)

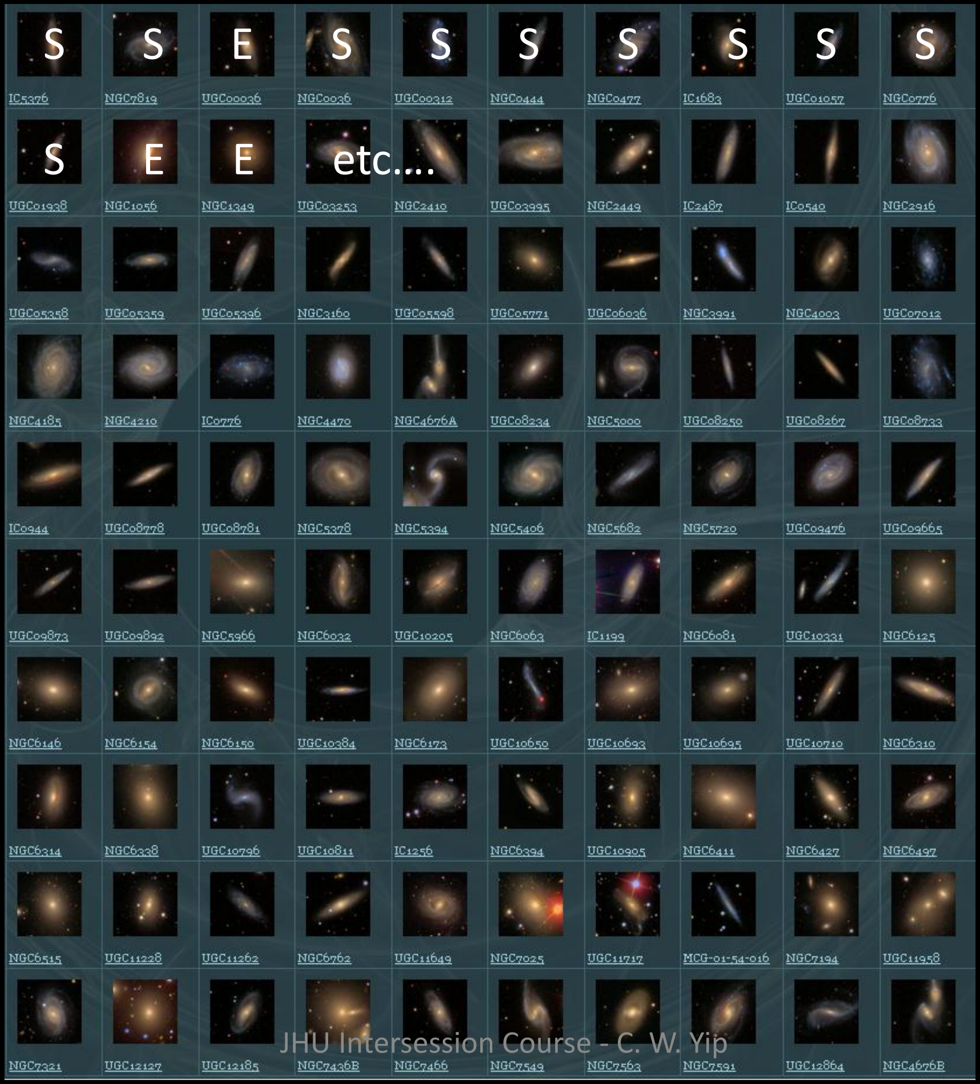
Galaxy Zoo Project

- Web users classify galaxy morphologies.
- Use majority vote to decide on the answer.



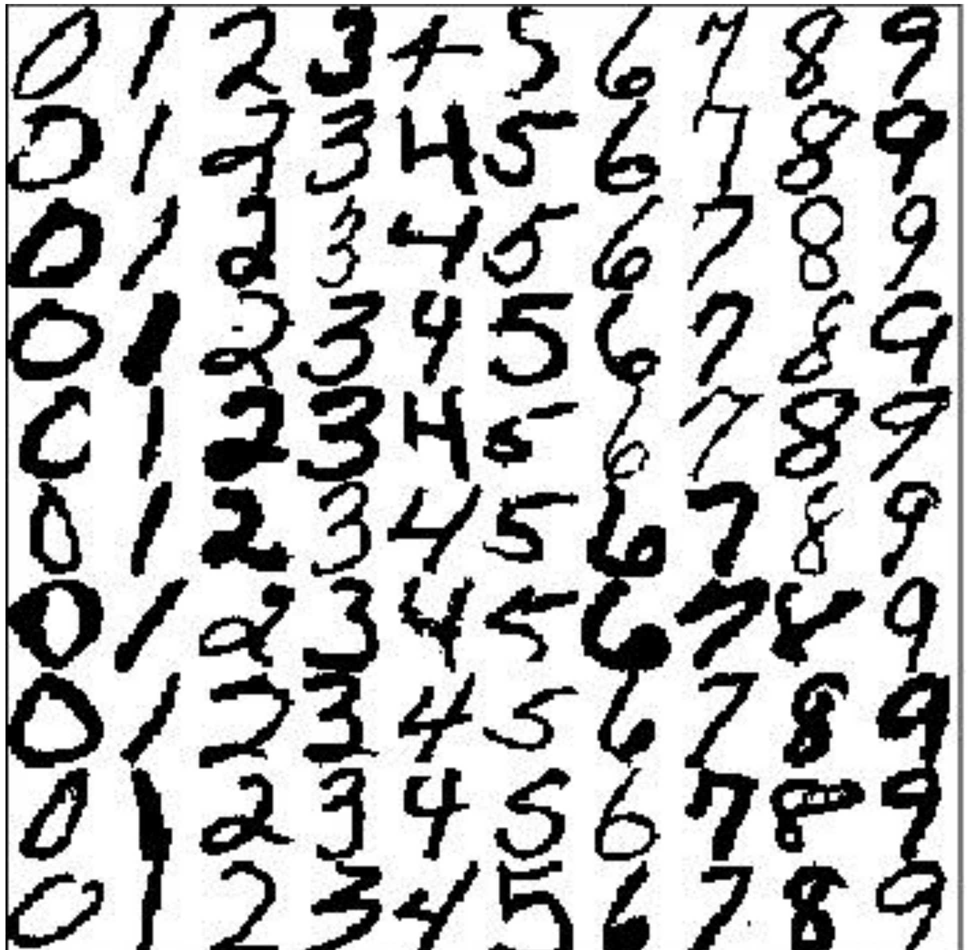
E0, E2, S0, … etc. are the labels.

Galaxy Images are Unlabeled Data (well, before Labeling)



(SDSS images of CALIFA DR1; Husemann et al. 2012, Sanchez et al. 2012, Walcher et al. in prep.)

Digit Recognition



Labels: 0 1 2 3 4 5 6 7 8 9

Distance between 2 Points in Space

- The set of data points spans a space.
- We can measure the distance between 2 points in such space.
  - E.g., Chi-sq measures the sum of distance squared between a point and model.
- If the points are close to each other:
  - They are "neighbors" with similar features.
- The best definition of distance (to yield concepts like "very close" and "far away") is data-dependent.

Cost Function

- In many machine learning algorithms, the idea is to find the model parameters $\boldsymbol{\theta}$ which minimize the cost function $J(\boldsymbol{\theta})$:

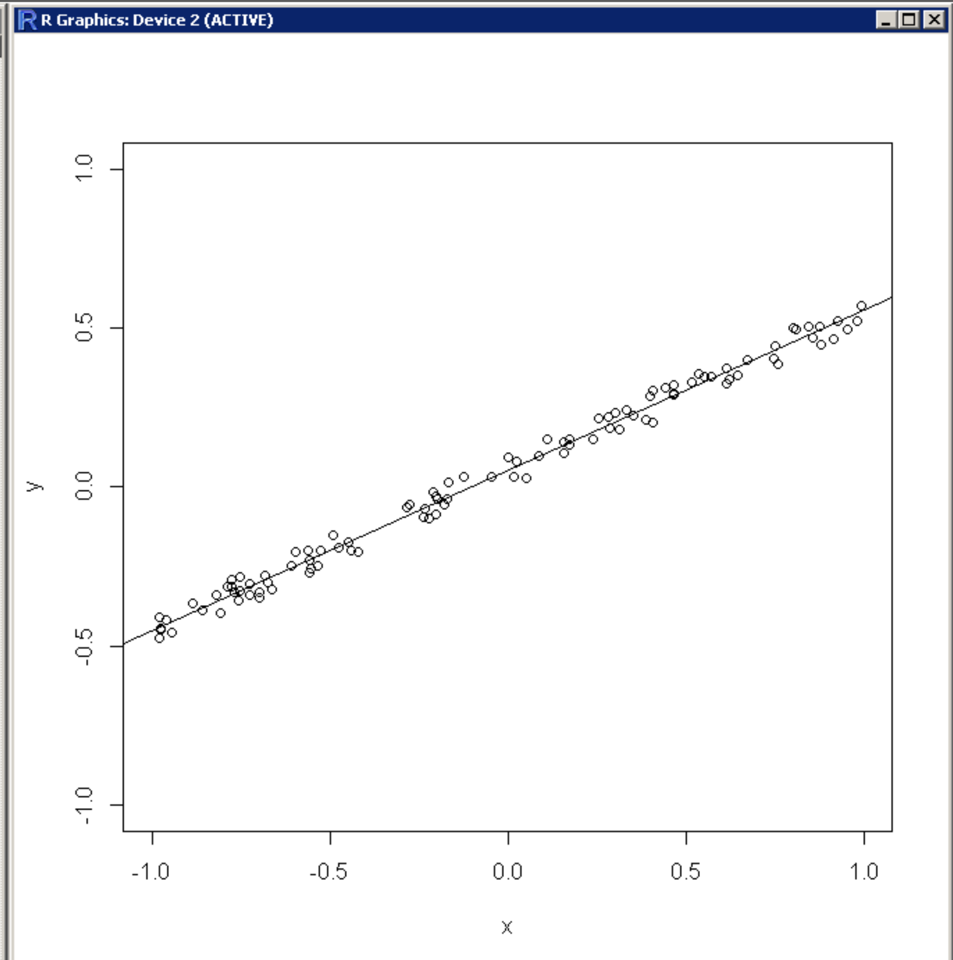
$$J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(model\left(data^{i}\right) - data^{i}\right)^{2}$$

m is the size of the dataset or the training set.

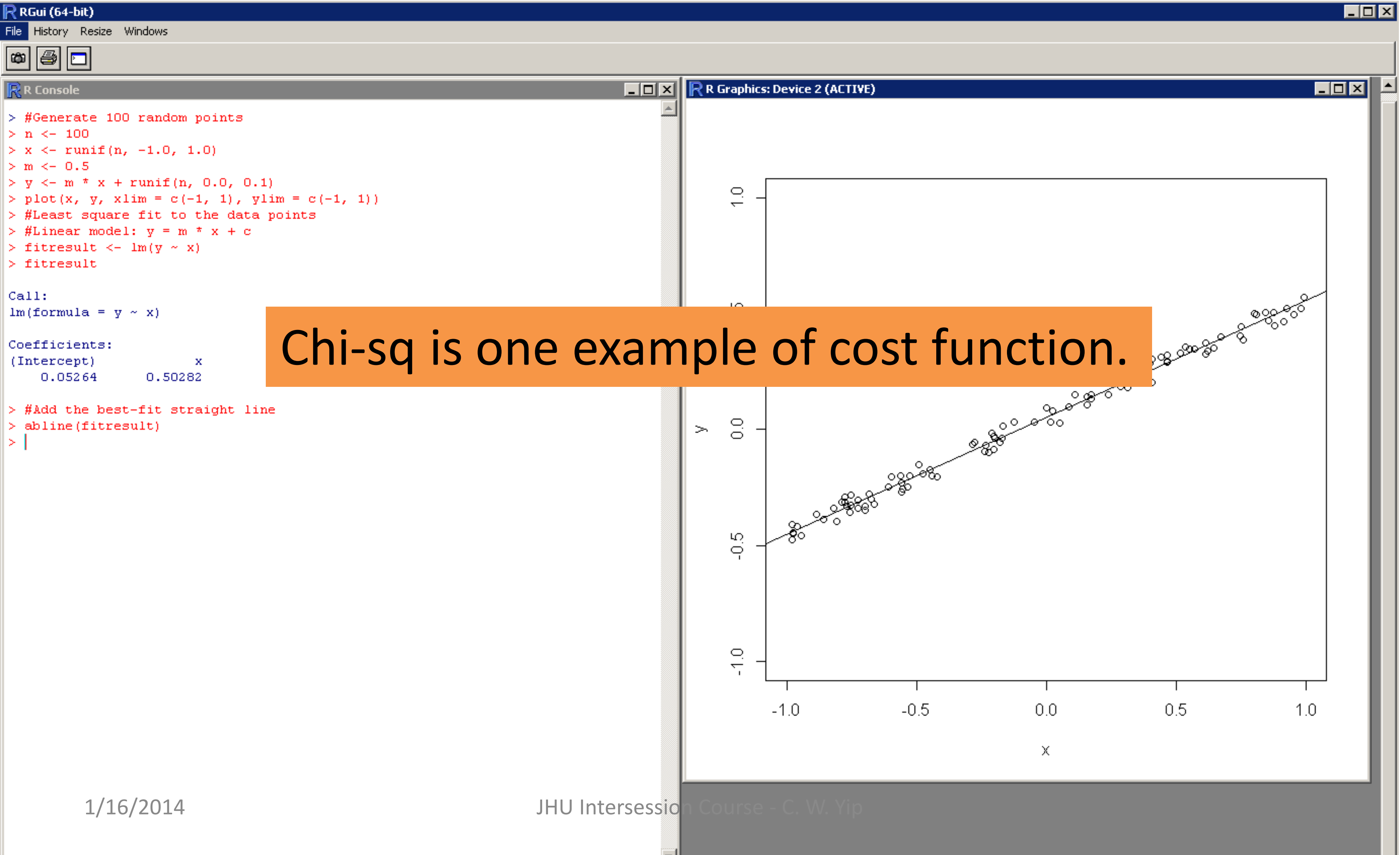
- That is, we want $model$ as close to $data$ as possible.
- Note that $model$ depends on $\boldsymbol{\theta}$.

Recall: LSQ Fitting

```
> #Generate 100 random points
> n <- 100
> x <- runif(n, -1.0, 1.0)
> m <- 0.5
> y <- m * x + runif(n, 0.0, 0.1)
> plot(x, y, xlim = c(-1, 1), ylim = c(-1, 1))
> #Least square fit to the data points
> #Linear model: y = m * x + c
> fitresult <- lm(y ~ x)
> fitresult

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    0.05264      0.50282

> #Add the best-fit straight line
> abline(fitresult)
>
```



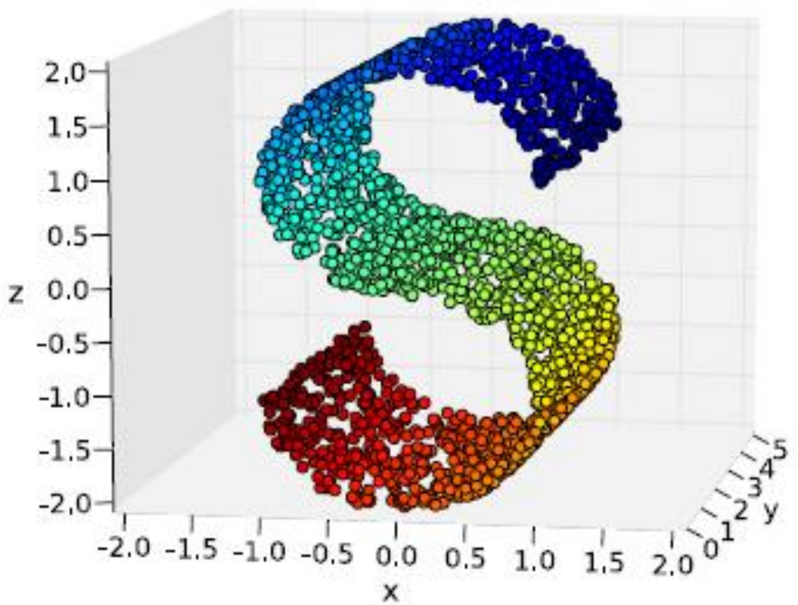
Recall: LSQ Fitting



Chi-sq is one example of cost function.

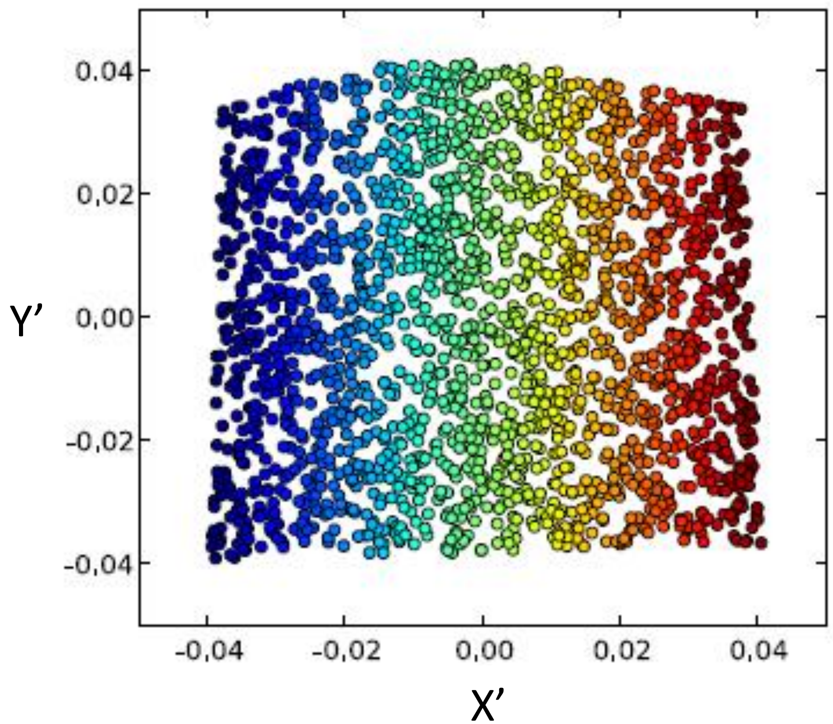
Distance between Points is Non-trivial: "S Curve"

3D Space



$$x = \sin(\theta)$$
$$z = \frac{\theta}{|\theta|}(\cos(\theta) - 1))$$
$$-1.5\pi \leqslant \theta \leqslant 1.5\pi$$
$$0 < y < 5.$$

2D Embedding Space



(Vanderplas & Connolly, 2009)

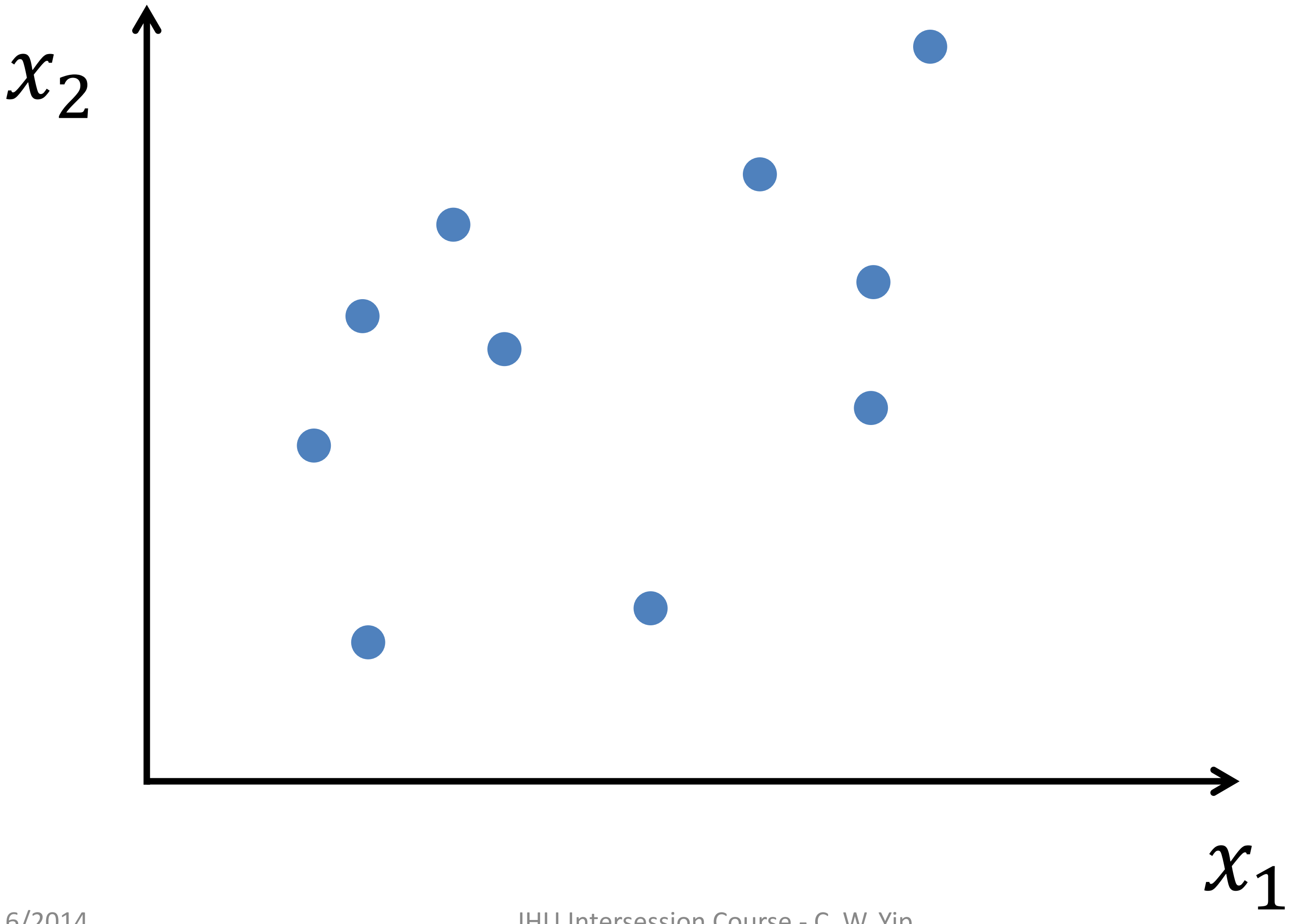
Unsupervised vs. Supervised Learning

- Unsupervised:
  - Given data $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3, \cdots, \boldsymbol{x}^n\}$ find patterns.
  - The description of a pattern may come in the form of a function (say, $g(\boldsymbol{x})$).
- Supervised:
  - Given data $\{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), (\boldsymbol{x}^3, \boldsymbol{y}^3), \cdots, (\boldsymbol{x}^n, \boldsymbol{y}^n)\}$ find a function such that $f(\boldsymbol{x}) = \boldsymbol{y}$.
  - $\boldsymbol{y}$ are the labels.

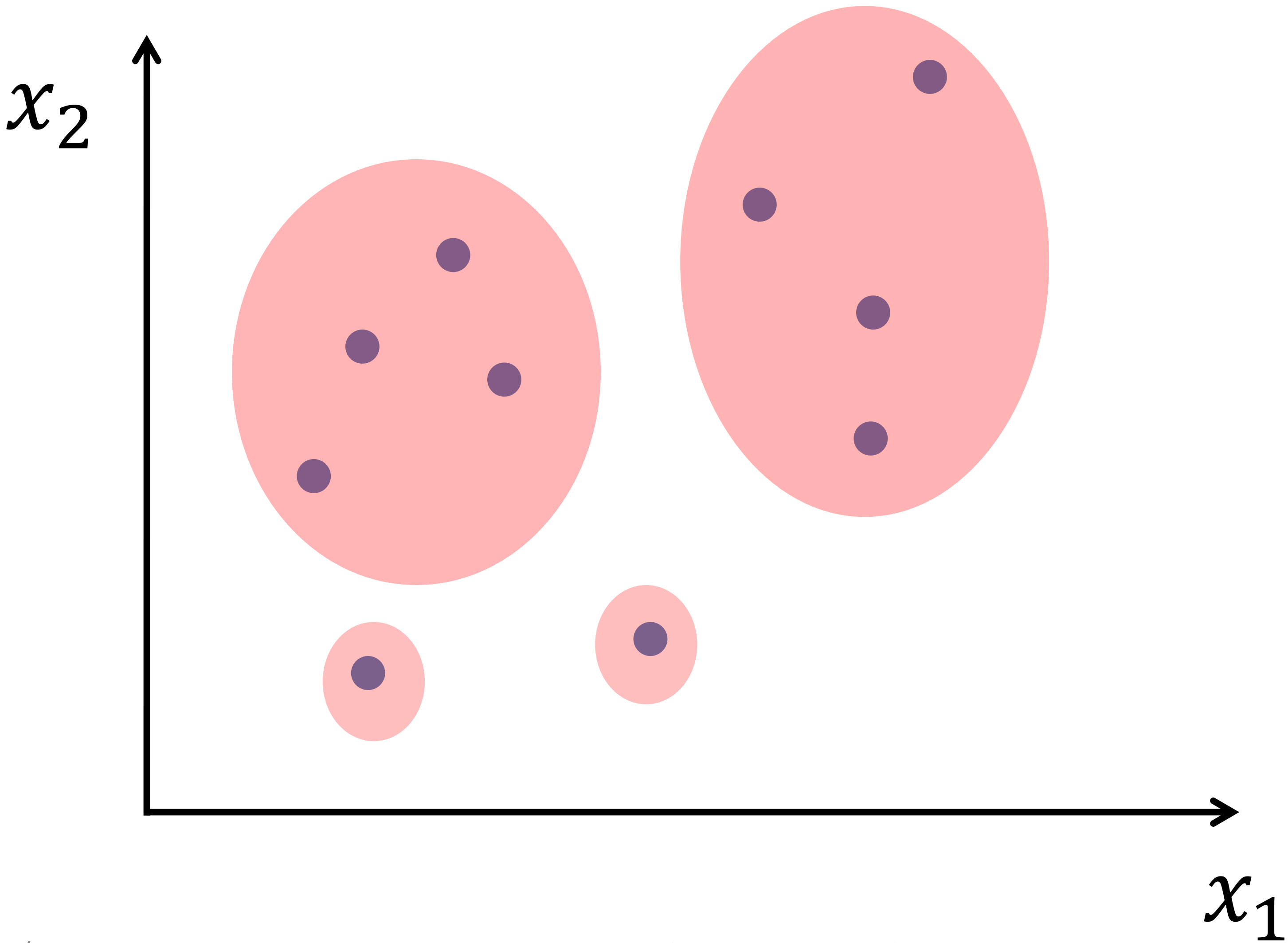
Some Areas in Unsupervised Learning

- Density Estimation
  - Kernel Density Estimation
  - Mixture of Gaussians
- Clustering
  - K Nearest Neighbor
- Dimension Reduction
  - Principal Component Analysis (Linear Technique)
  - Locally Linear Embedding (Non-Linear Technique)

Unsupervised Learning: Find Patterns in the Unlabeled Data



Unsupervised Learning: Find Patterns in the Unlabeled Data



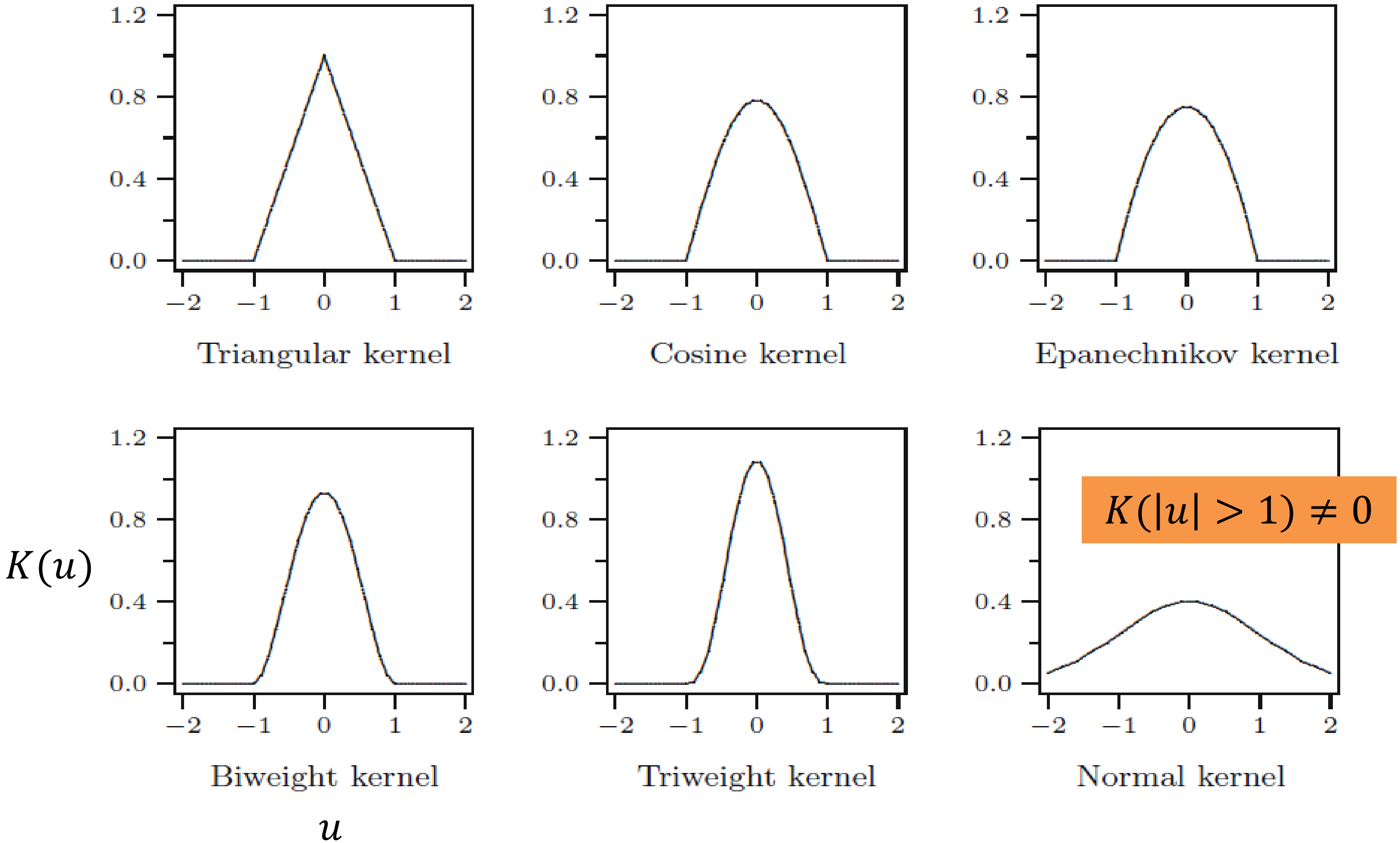
Basis Function

- Basis function allows us to parameterize the data in some handy or/and meaningful ways.
- The decomposition of a data point into basis function is usually quick.
- Examples:
  - Gaussian functions
  - Kernel functions
  - Eigenfunctions (such as "Eigenspectra" in Astronomical Spectroscopy)

Density Estimation: Kernel Density Estimation

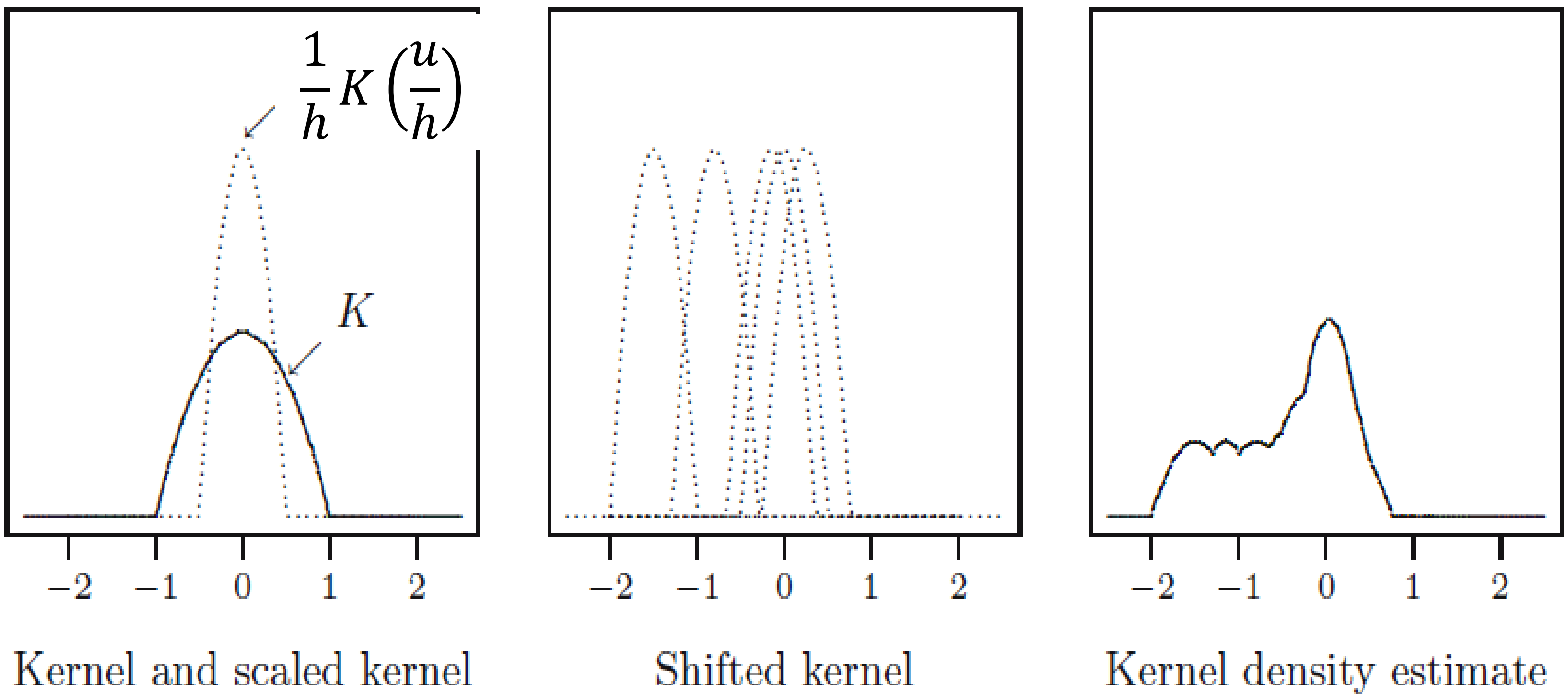
- Approximate a distribution by the sum of kernels (K’s) of various bandwidths (h).
- Properties of kernel:
  - $K\ is\ a\ probabilty\ density\ function, K(u) > 0\ and\ \int_{-\infty}^{\infty} K(u)du = 1.$
  - $K\ is\ symmetric\ around\ zero: K(-u) = K(u)$
  - (Not Always) $K(u) = 0\ for\ |u| > 1$

Example Kernel Functions K(u)



(Dekking et al. 2005)

Constructing the Kernel Density Estimate

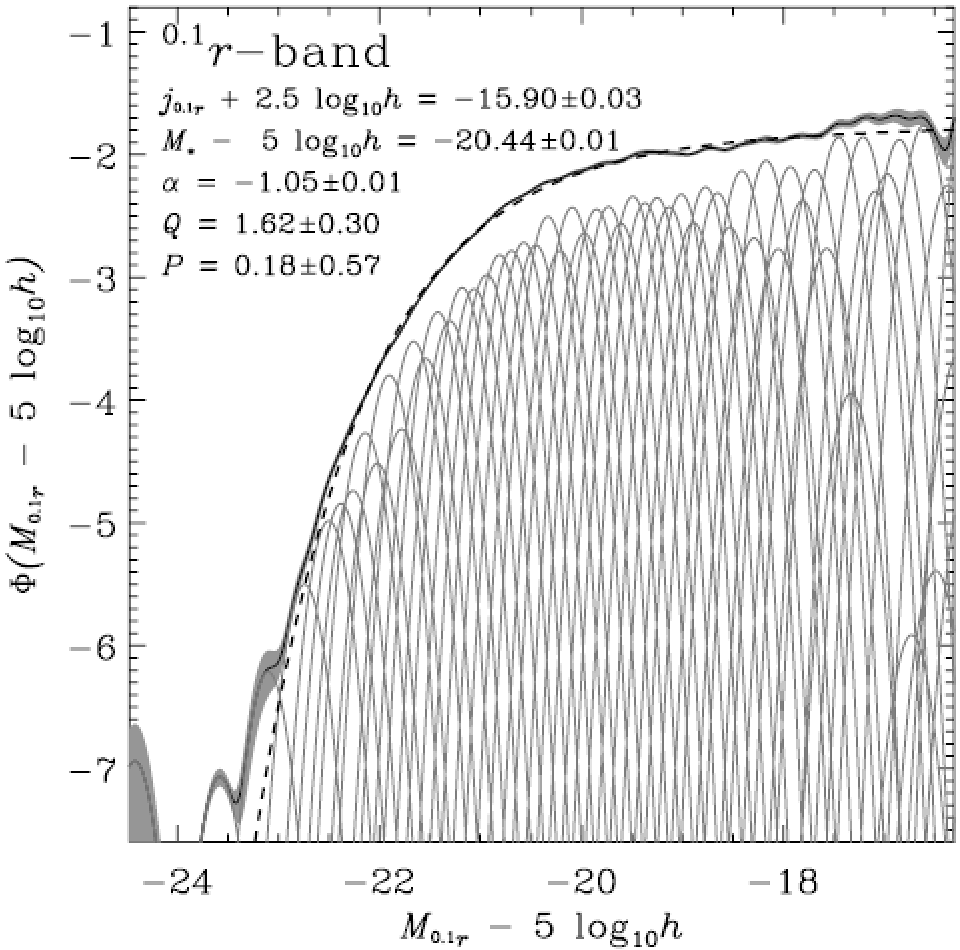


$$f = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{u - x^i}{h}\right)$$

$$\boxed{Note \int_{-\infty}^{\infty} f(u)du = 1}$$

Luminosity Function of Nearby SDSS Galaxies

Spatial Volume Density (Higher→)

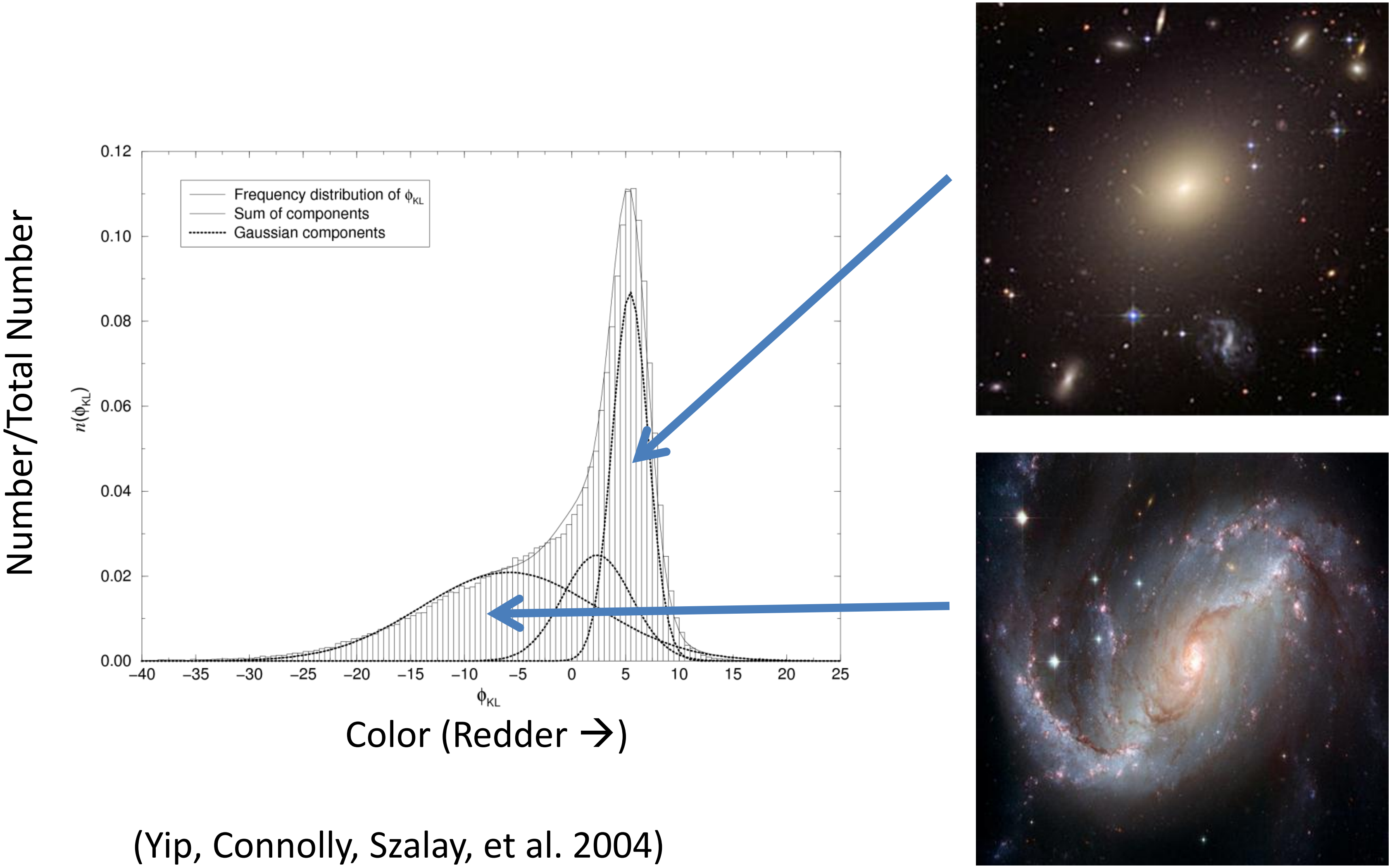


Notice the height of the kernels actually vary, slightly different from the discussed kernel density estimation.

(Blanton et al. 2003)

Magnitude of Galaxy (Dimmer→)

Color Distribution of Nearby Galaxies



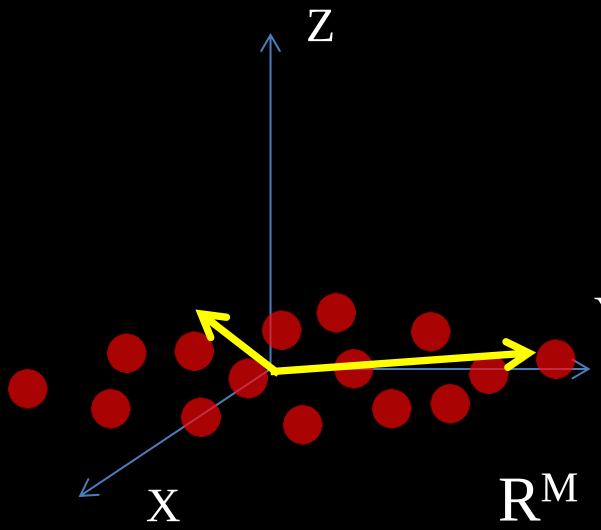
(Yip, Connolly, Szalay, et al. 2004)

Principal Component Analysis

- Perhaps the most used technique to parameterize high-dimensional data.
- Best for linear data distribution.
- Many applications in both academia and industry: dimension reduction, data parameterization, classification problems, image decomposition, audio signal separation, …, etc.

High-Dimensional Data may lie in Lower-Dimensional Space (or Manifold)

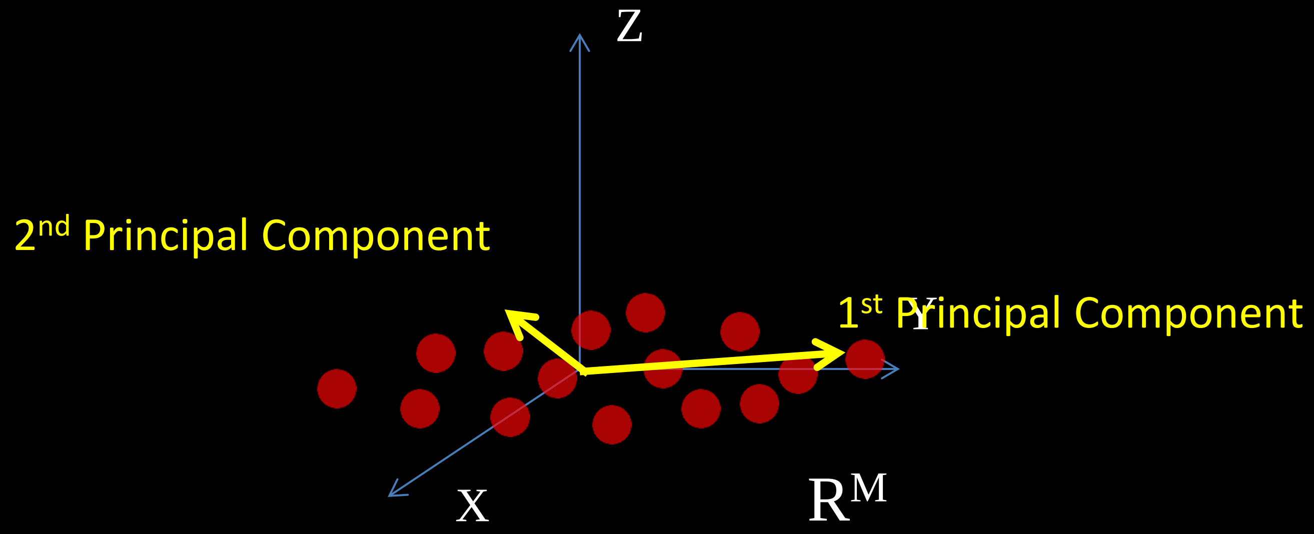
N points, M components



PCA finds the orthogonal directions in the data space which encapsulate maximum sample variances.

High-Dimensional Data may lie in Lower-Dimensional Space (or Manifold)

N points, M components

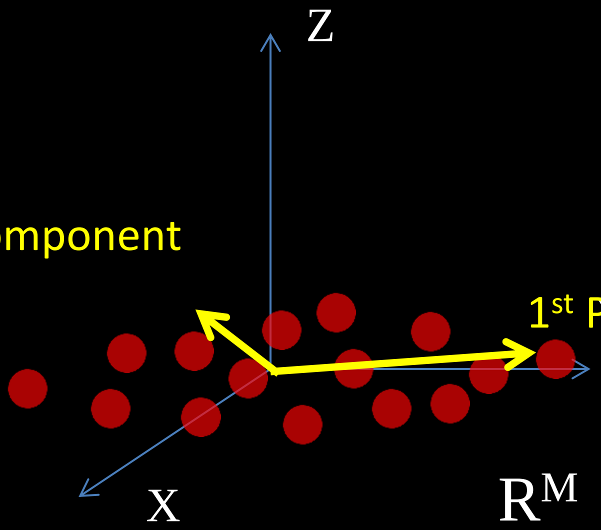


PCA finds the orthogonal directions in the data space which encapsulate maximum sample variances.

A principal component is also called eigenvector, or eigenfunction.

High-Dimensional Data may lie in Lower-Dimensional Space (or Manifold)

N points, M components



We reduce the dimension of the problem from M to 2.

PCA finds the orthogonal directions in the data space which encapsulate maximum sample variances.

A principal component is also called eigenvector, or eigenfunction.

Properties of PCA

- PCA decomposes the data into a set of eigenfunctions.
- The eigenfunctions are orthogonal (perpendicular) to each other:
  - The dot product between two eigenvectors is zero.
- The set of eigenfunctions maximize the sample variance of the data. Therefore, the data can be decomposed into a handful of eigenfunctions (for linear data distribution).
- For non-linear data, PCA may fail (i.e., we need many orders of eigenfunctions).
- In galaxy spectroscopy, the basis functions are called “eigenspectra” (Connolly et al. 1995).

PCA Eigenspectra (e_{i_λ}) Representation of Galaxy Spectra

$$\boxed{f_\lambda = \sum_i a_i e_{i_\lambda}}$$

(i runs from 1 to the number of eigenspectra)

Minimize reconstruction error with respect to a_i's:

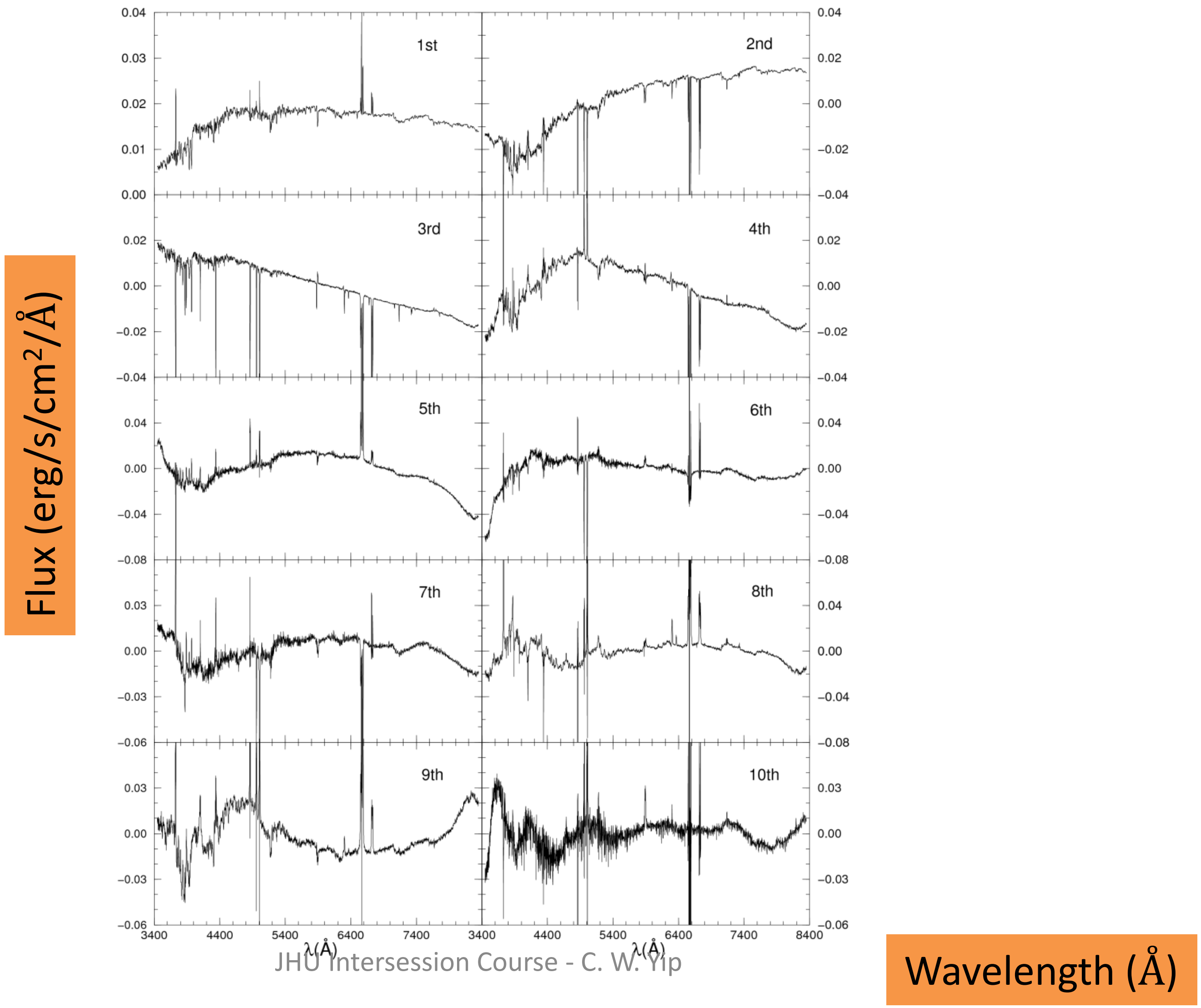
$$x^2 = \sum_\lambda w_\lambda \left(f_\lambda - \sum_i a_i e_{i_\lambda}\right)^2$$

Get the weights for each basis function:

$$a_i = a_i(w_\lambda, e_{i_\lambda}, f_\lambda)$$

(Connolly & Szalay 99)

Eigenspectra of Nearby Galaxies (SDSS: N = 170,000; M = 4000)

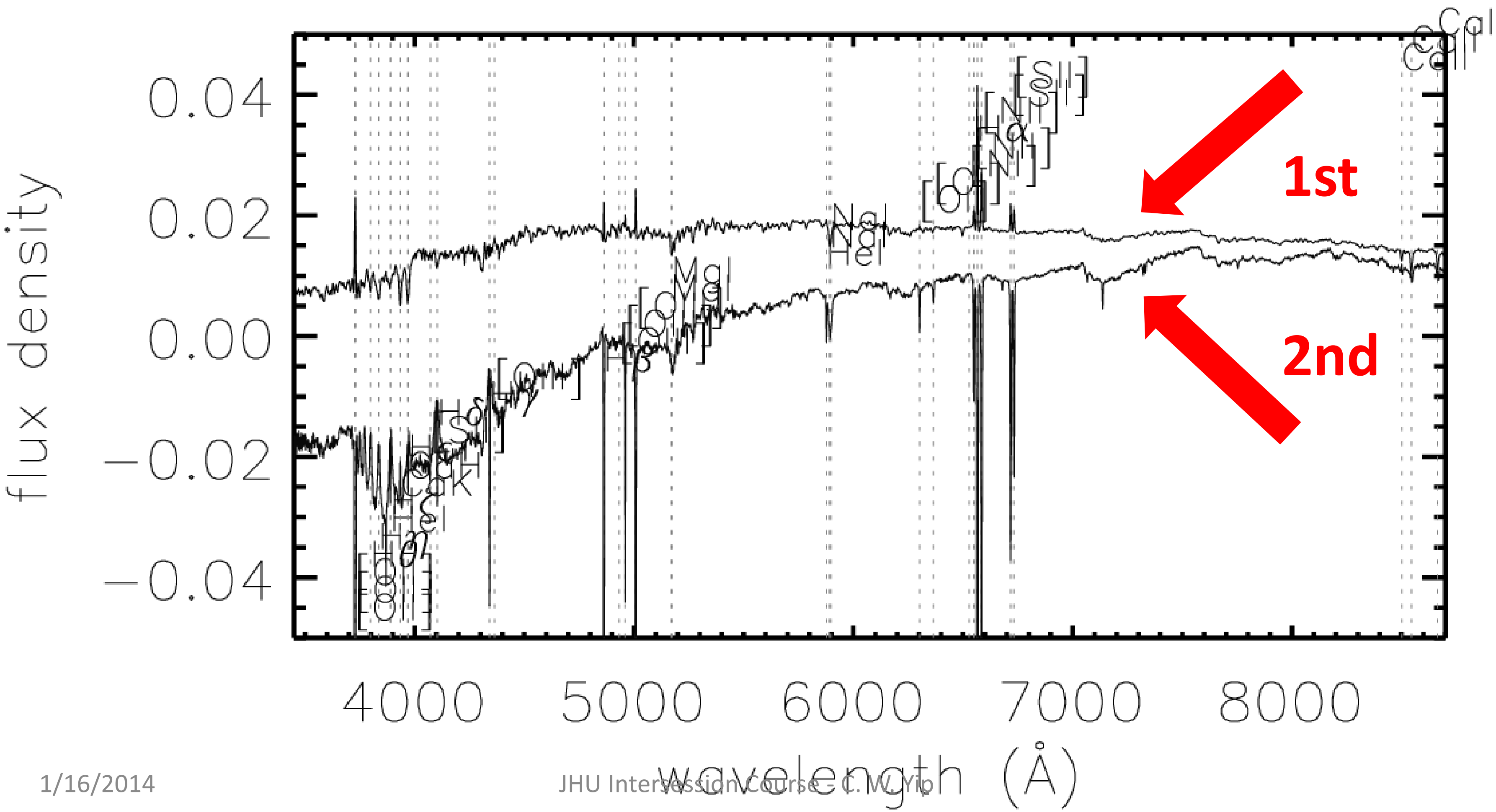


PCA as a way for Signal Separation

- The basis functions has physical meanings. Hence each order (or called “mode”) of function can be considered as a signal.
- In PCA, the basis functions are orthogonal. They point to different direction in the data space and are statistically independent.

Eigenspectra of Half a Million Galaxy Spectra

- 2nd mode: galaxy type (steepness of the spectral slope)



Eigenspectra of Half a Million Galaxy Spectra

- 3rd mode: post-starburst activities (stronger the absorptions, weaker the Hα)

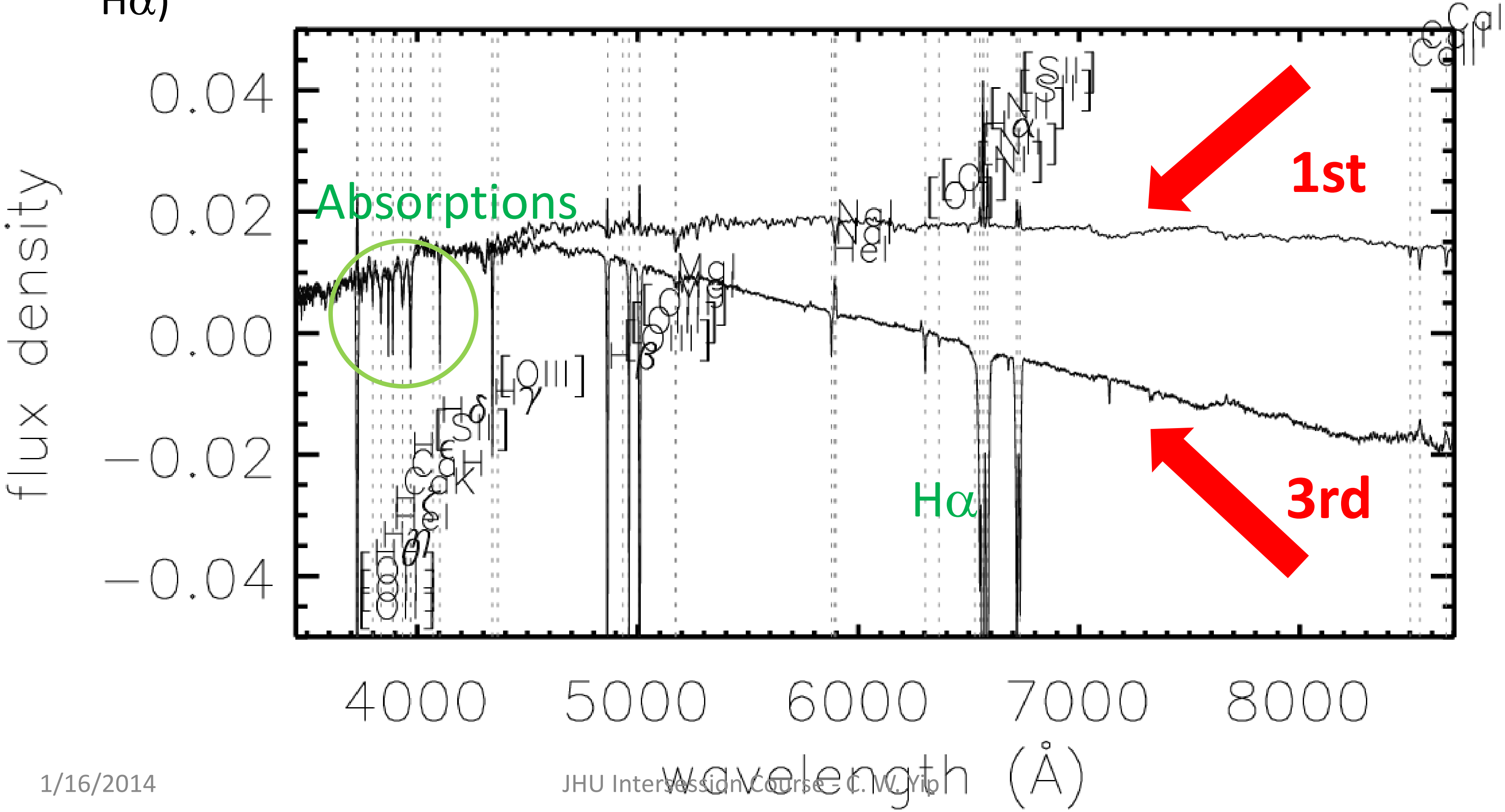


Image Reconstruction

Original Reconstruction using 50 Eigenfaces

(Everson & Sirovich 1995)

Matrix Representation of Data

- Many datasets are made up with N objects and M variables.
- Matrix provides a handy way to represent the data.
- An added advantage is that many algorithms can be expressed conveniently in matrix form.

Example:
Size M x N Data Matrix for Galaxy Spectra

$$A_{ij} = Flux\ at\ the\ ith\ wavelength\ for\ the\ jth\ galaxy$$

$$A = \begin{array}{cc} & \textbf{Galaxy ID} \\ \rotatebox{90}{\textbf{Wavelength}} & \begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2N} \\ \vdots & \ddots & & & \vdots \\ A_{M1} & A_{M2} & A_{M3} & \cdots & A_{MN} \end{pmatrix} \end{array}$$

** PCA in matrix notation: see handout.*

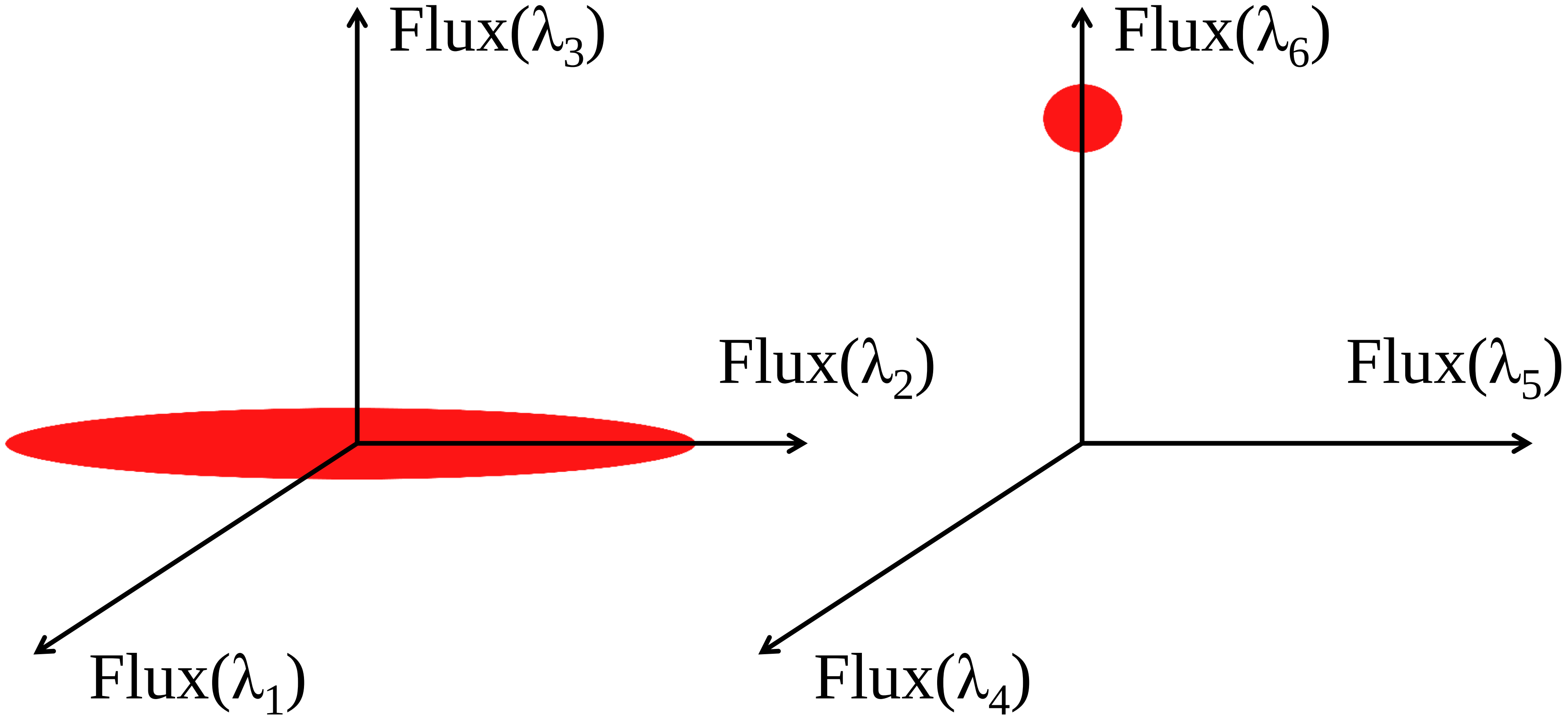
Compressing the Object Space: CUR Matrix Decomposition

- In PCA, the number of components in the data vectors remain intact.
- CUR Matrix Decomposition provides a dramatically new way to compress big data. This approach compresses the variable space:
  - The number of components in the data vectors decreases.

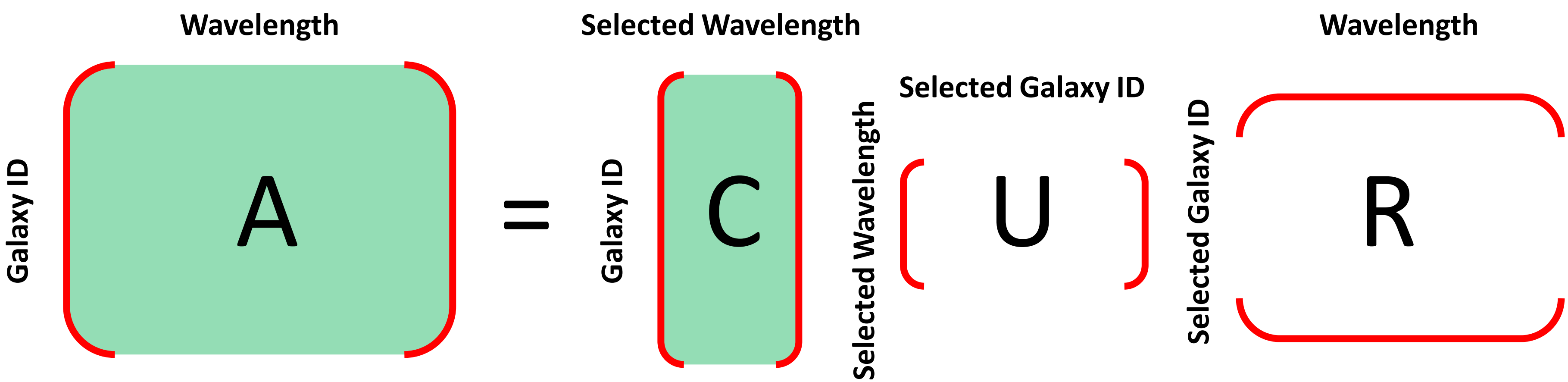
Some Wavelengths are More Informative: Leverage Score per Each Variable

Sample Variance

Sparsity



CUR Matrix Decomposition (Mahoney & Drineas 2009)



CUR approximates data matrix:

$$\min \| A - CUR \|_F$$

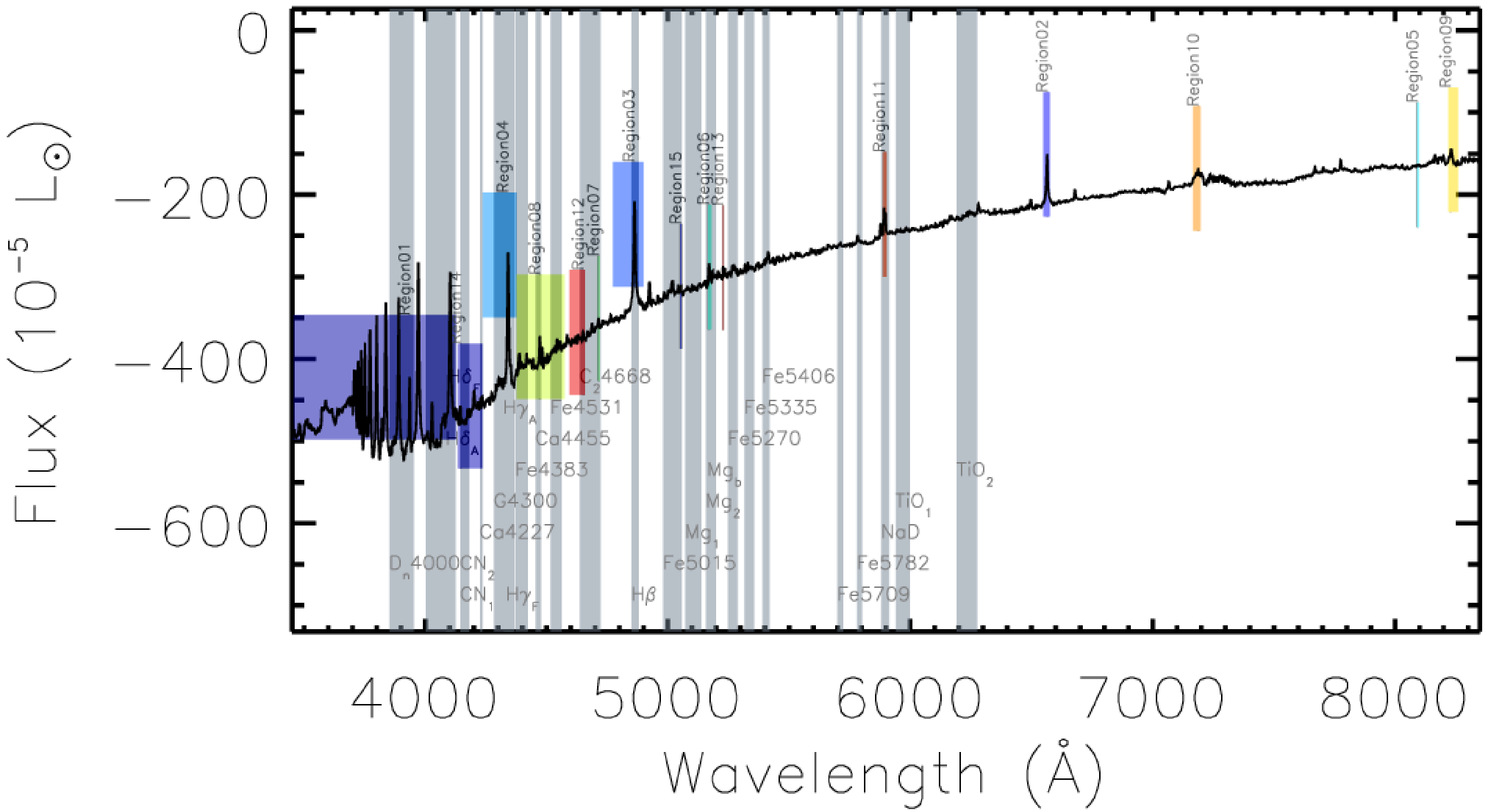
Frobenius Norm

- A matrix norm is a number for representing the amplitude of a matrix.
- The Frobenius norm is defined as follows:

$$\|A\|_F = \sqrt{\sum_{i=1}^{M} \sum_{i=1}^{N} \left|A_{ij}\right|^2}$$

- In CUR Matrix Decomposition, the matrix A is the difference between the data matrix and its approximated matrix. The Frobenius norm therefore measures the “distance” between the two matrices.

Find Important Regions in Multi-Dimensional Data: Galaxy Spectra



Discussion of Jan 7 Homework

- Total number of pixels in the CCD = 1024 x 1024 pixels = 1,048,576 pixels ~ 1 MPixels.
- GAIA needs $10^9/60/60/24/365$ years = 31.7 years =31 years 8 months ~ 32 years.