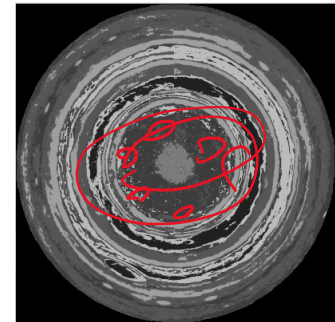
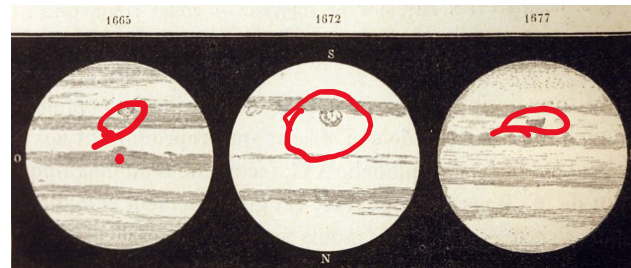


EE520 Data Driven Analysis of Complex Systems

Erik Bollt

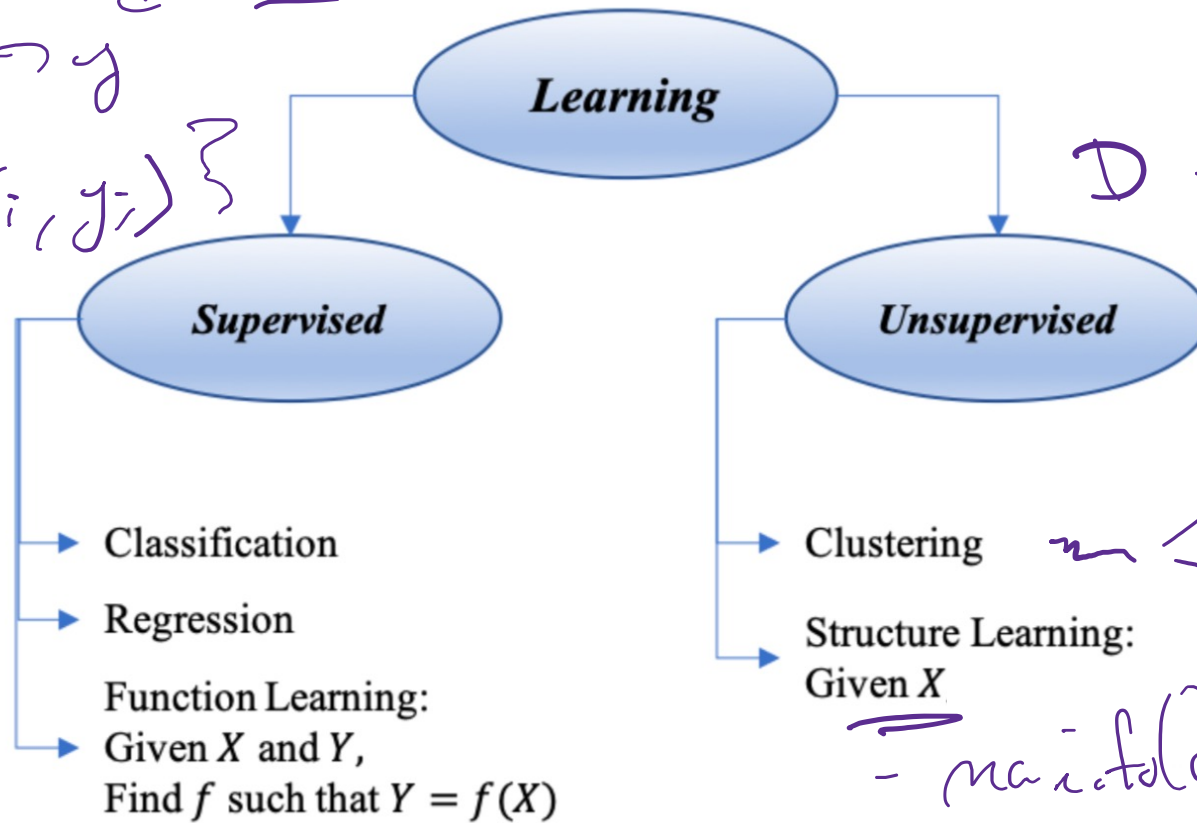
~

~



Ch 5 - Clustering and Classification

attributes
 $f: x \mapsto y$
 $D = \{(x_i, y_i)\}$



$D = \{x_i\}$

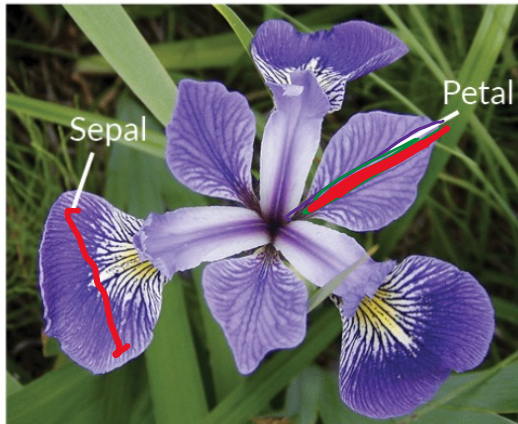
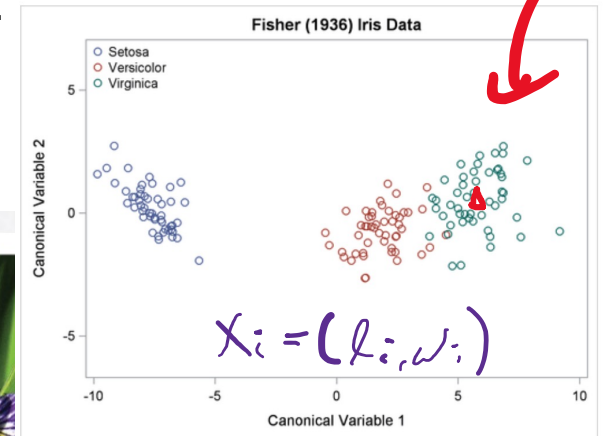
Clustering

Structure Learning:
Given X

\sim $\begin{cases} \text{k means} \\ \text{spectral} \end{cases}$
 \sim manifold
 \sim tree-hierarchy.

Discriminating Fisher's iris data by using the petal areas

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.



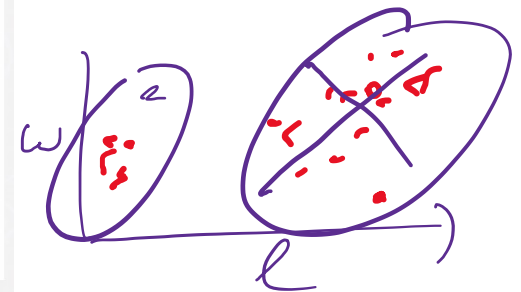
Iris Versicolor



Iris Setosa



Iris Virginica



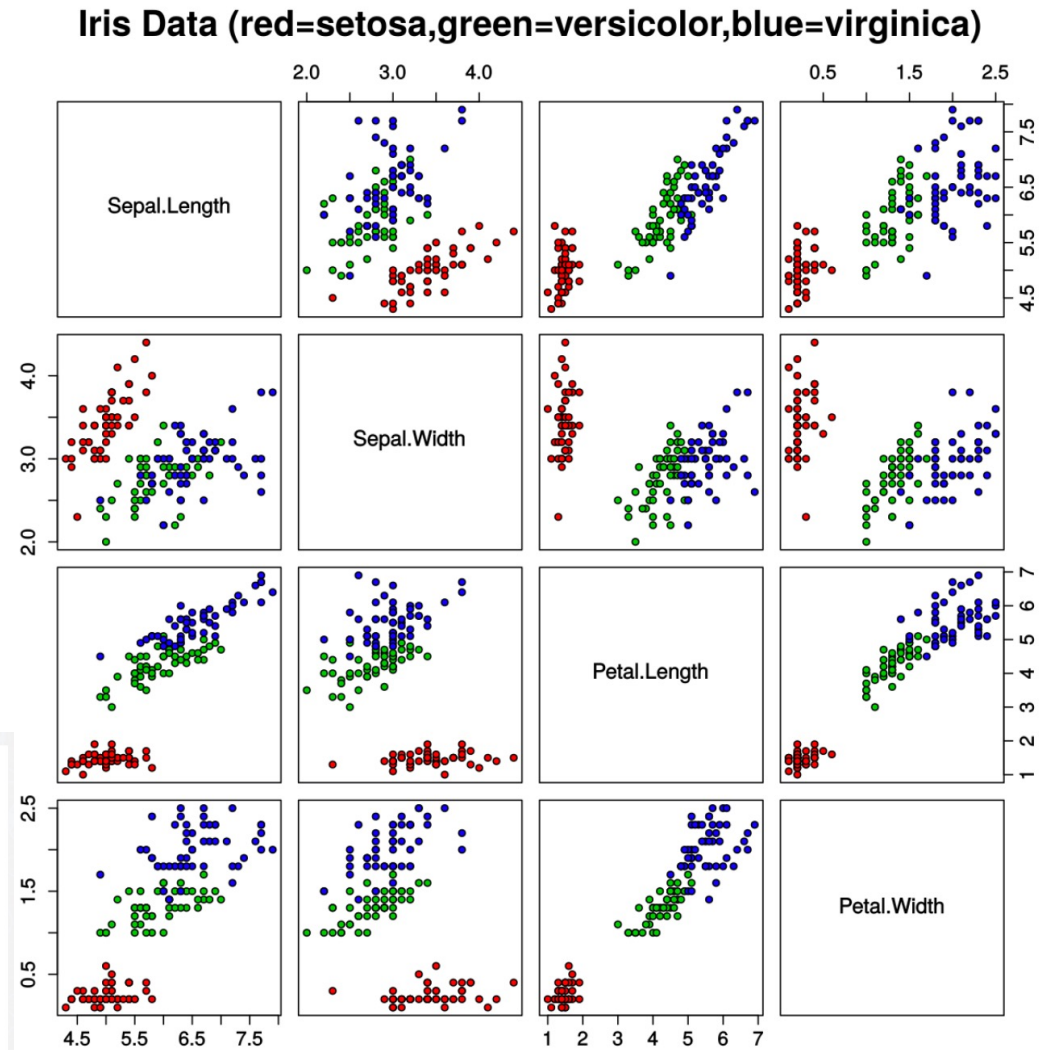
Ronald Fischer 1936



Iris Versicolor

Iris Setosa

Iris Virginica



Data

Learning

D_1

D_2

$x_i = \{\text{name, zip code, time of day}\}$
 $y_i = \{0, 1, 2, 3\}$
 $f: x \rightarrow y$

Supervised

learn f
regression
 $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}^2$

Unsupervised

no y_i
no f

clarify if $y_i \in \mathbb{Z}_r = \{0, 1, 2\}$ just shape
no 2
= {hot, cold}

$f: \underline{X} \rightarrow \underline{Y}$

$\underline{D}_2 = \{(\underline{x}_i, \underline{y}_i)\}$; $\underline{x}_i \sim \underline{X}$, $\underline{y}_i \sim \underline{Y}$

either

vs. $z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$

$D_1 = \{\underline{x}_i\}_{i=1}^N$, $x_i \sim \underline{X}$

If we are going to do some machine learning – we had better get serious about what does learning mean?

-Supervised discrete output (labels) Classification

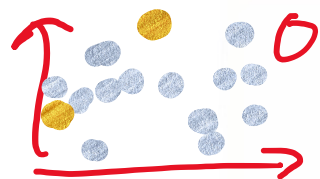
-Supervised continuous output Regression

-Unsupervised discrete output (labels – cluster analysis)

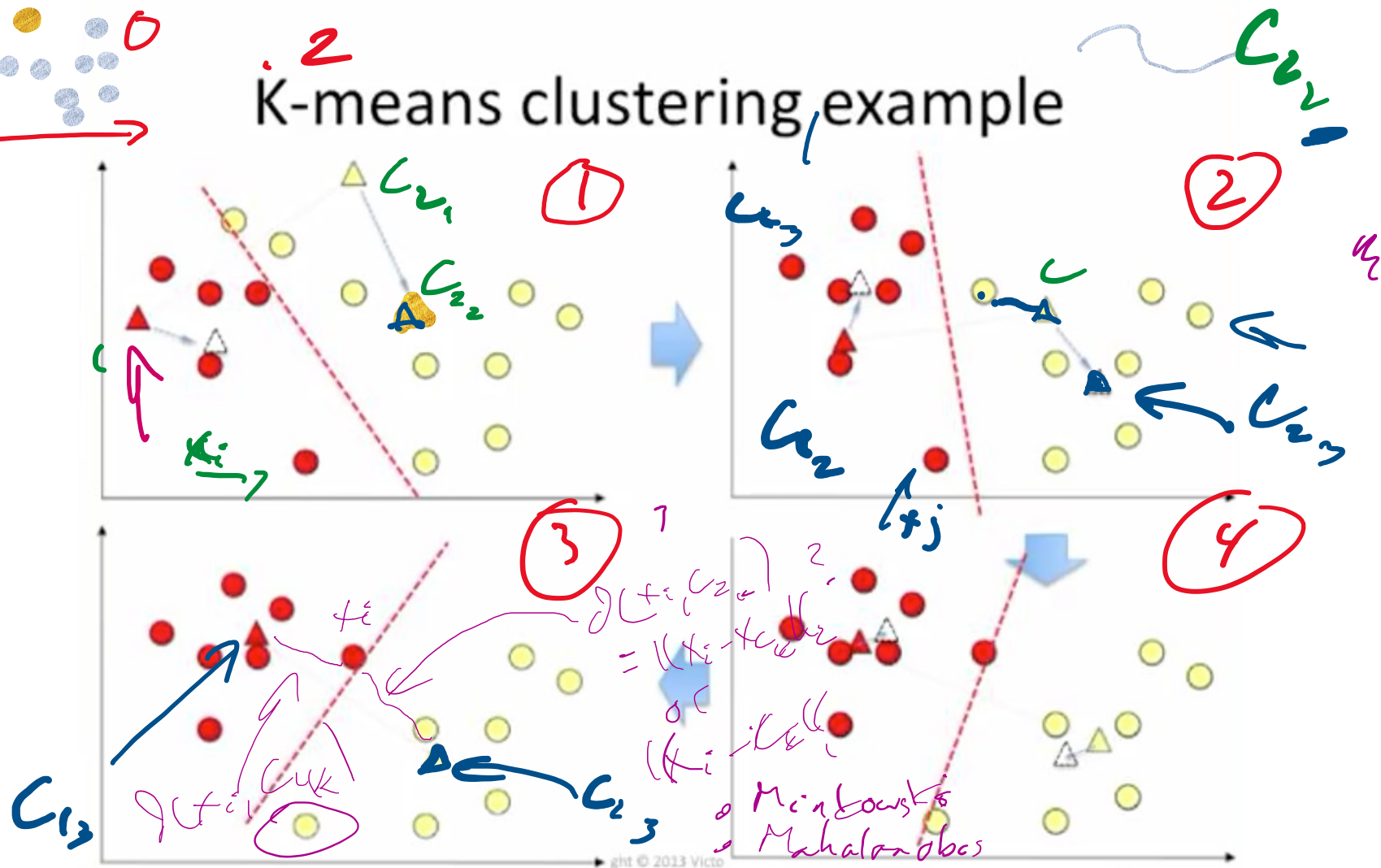
-Unsupervised continuous output (ROM, manifold learning, density estimation)

Learning functions?

Learning structure?



K-means clustering example



https://files.realpython.com/media/centroids_iterations.247379590275.gif

$$x_i = (x_{i1}, x_{i2}, \dots)$$

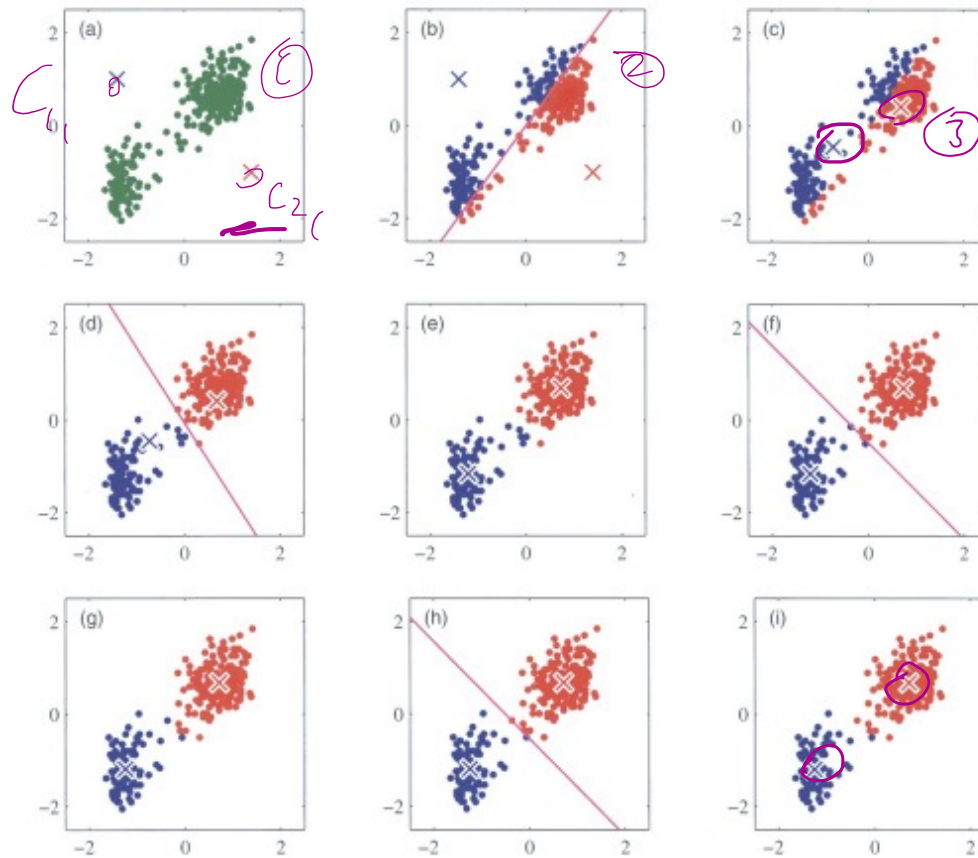


Illustration of K-means algorithm from [Bishop 2006]

Kmeans Convergence

Objective

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



1. Fix μ , optimize C :

$$\min_C \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2 = \min_c \sum_i^n |x_i - \mu_{x_i}|^2$$

Step 1 of kmeans

2. Fix C , optimize μ :

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

- Take partial derivative of μ_i and set to zero, we have

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Step 2 of kmeans

$$\|x - \mu_i\|_2^2 = \partial C(x, \mu_i)$$

Kmeans takes an alternating optimization approach, each step is guaranteed to decrease the objective – thus guaranteed to converge

K-Means

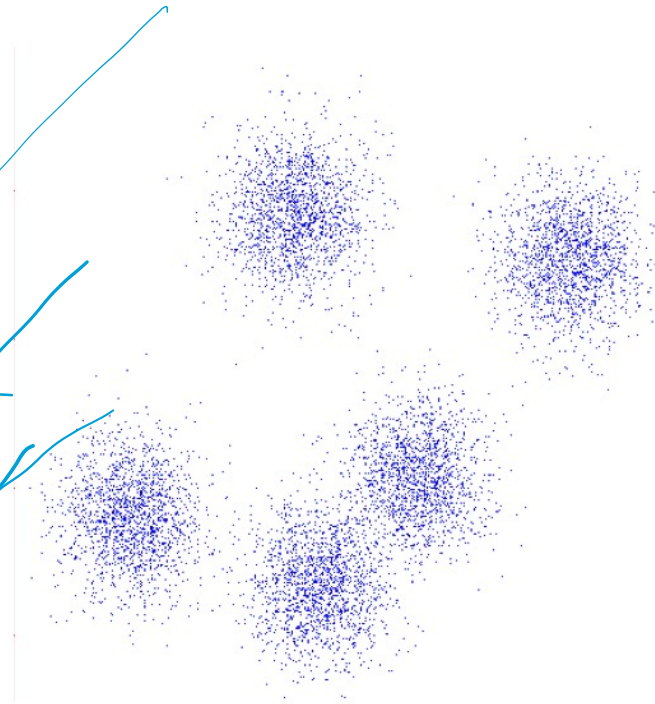
- An iterative clustering algorithm

- **Initialize:** Pick K random points as cluster centers

- **Alternate:**

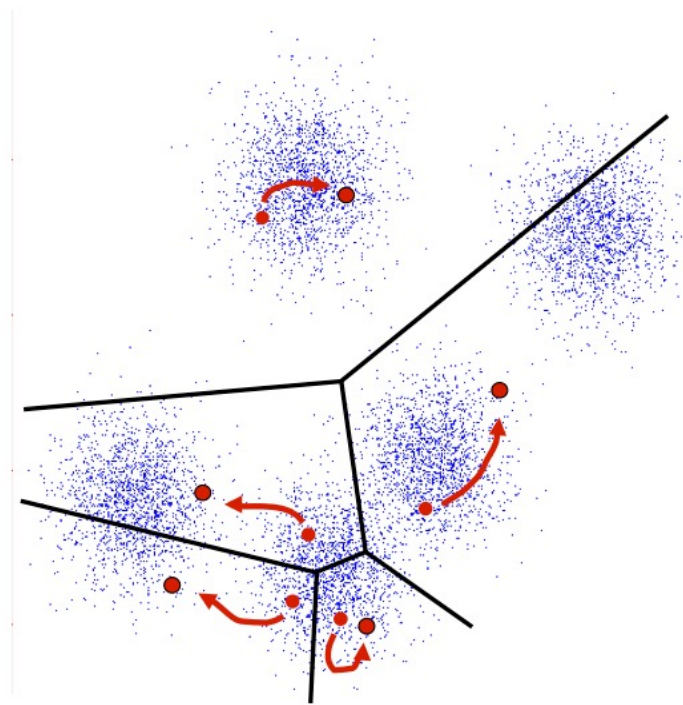
1. Assign data points to closest cluster center
2. Change the cluster center to the average of its assigned points

- **Stop when no points' assignments change**

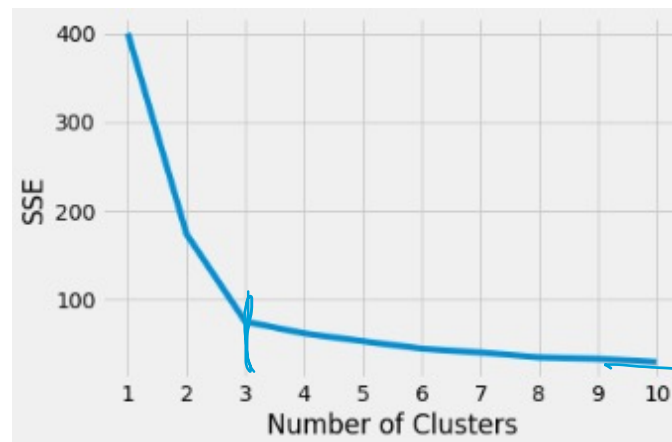


K-Means

- An iterative clustering algorithm
 - **Initialize:** Pick K random points as cluster centers
 - **Alternate:**
 1. Assign data points to closest cluster center
 2. Change the cluster center to the average of its assigned points
 - **Stop** when no points' assignments change



8



elbow

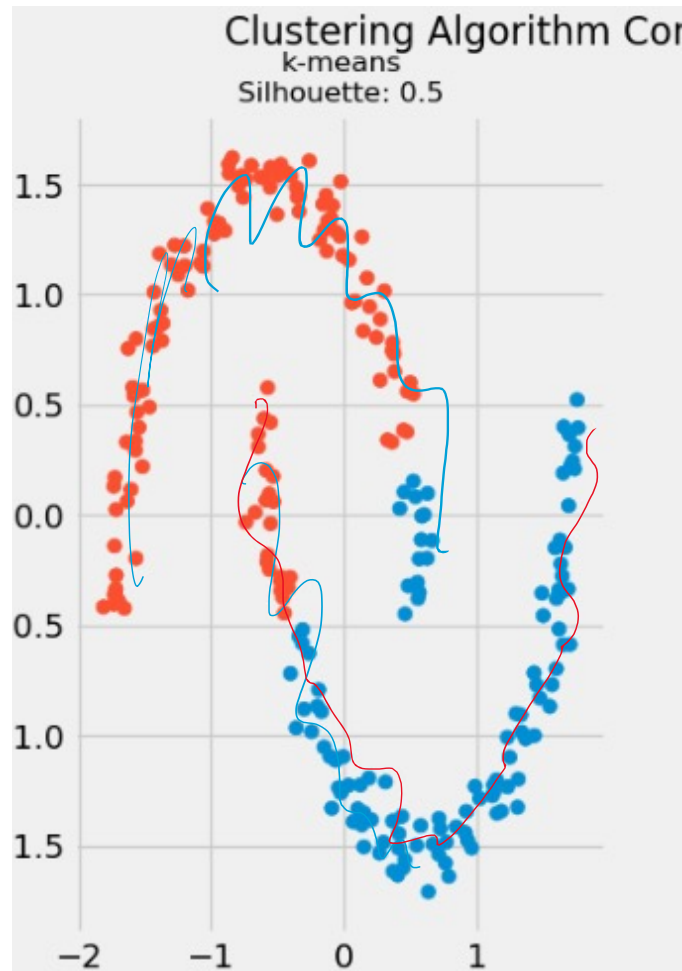
$$X = (x_1, x_2)$$

$$C = (c_1, c_2)$$

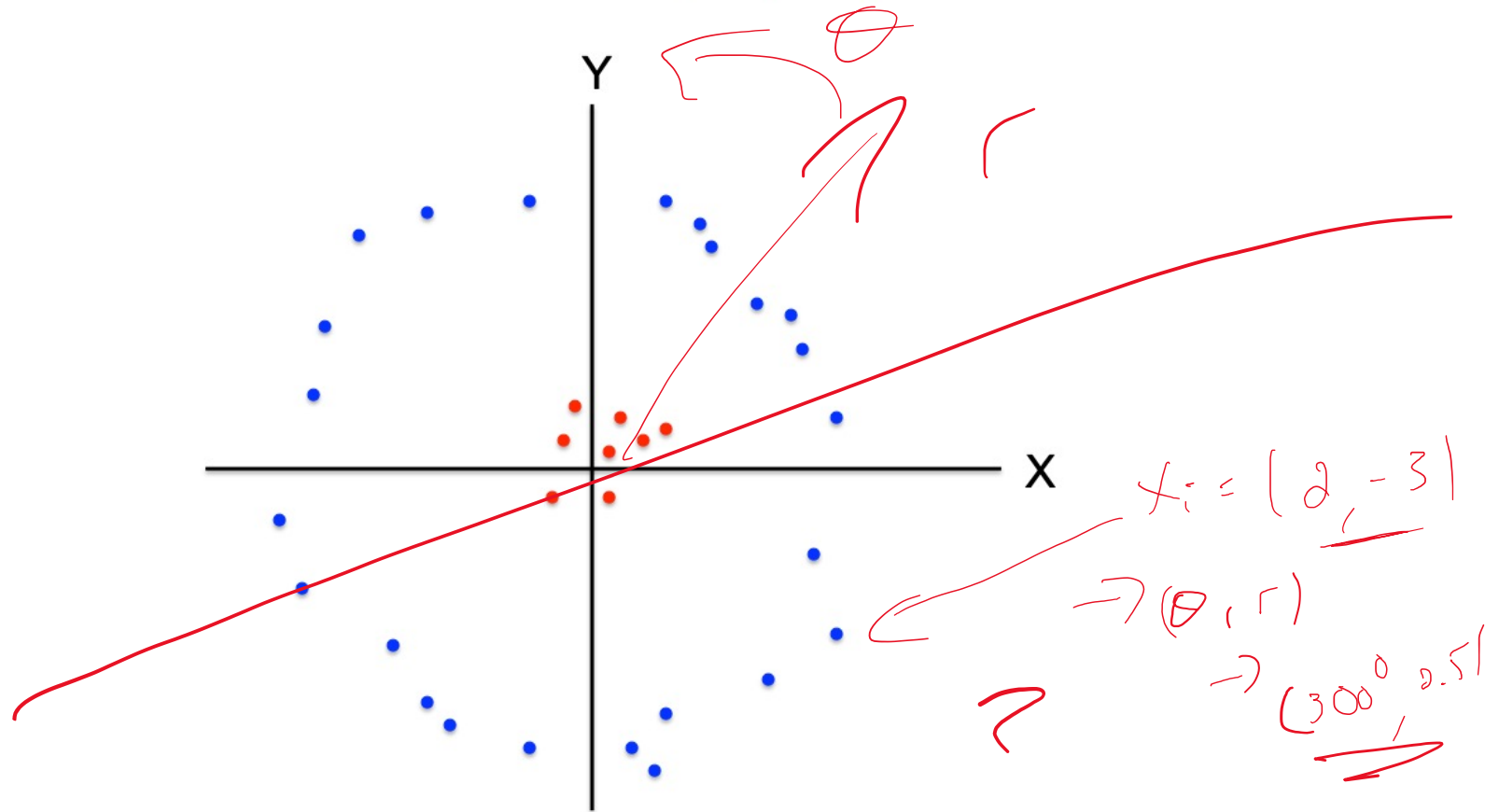
$$Q(X, C)$$

$$= \sum_i (x_i - c_i)^2$$

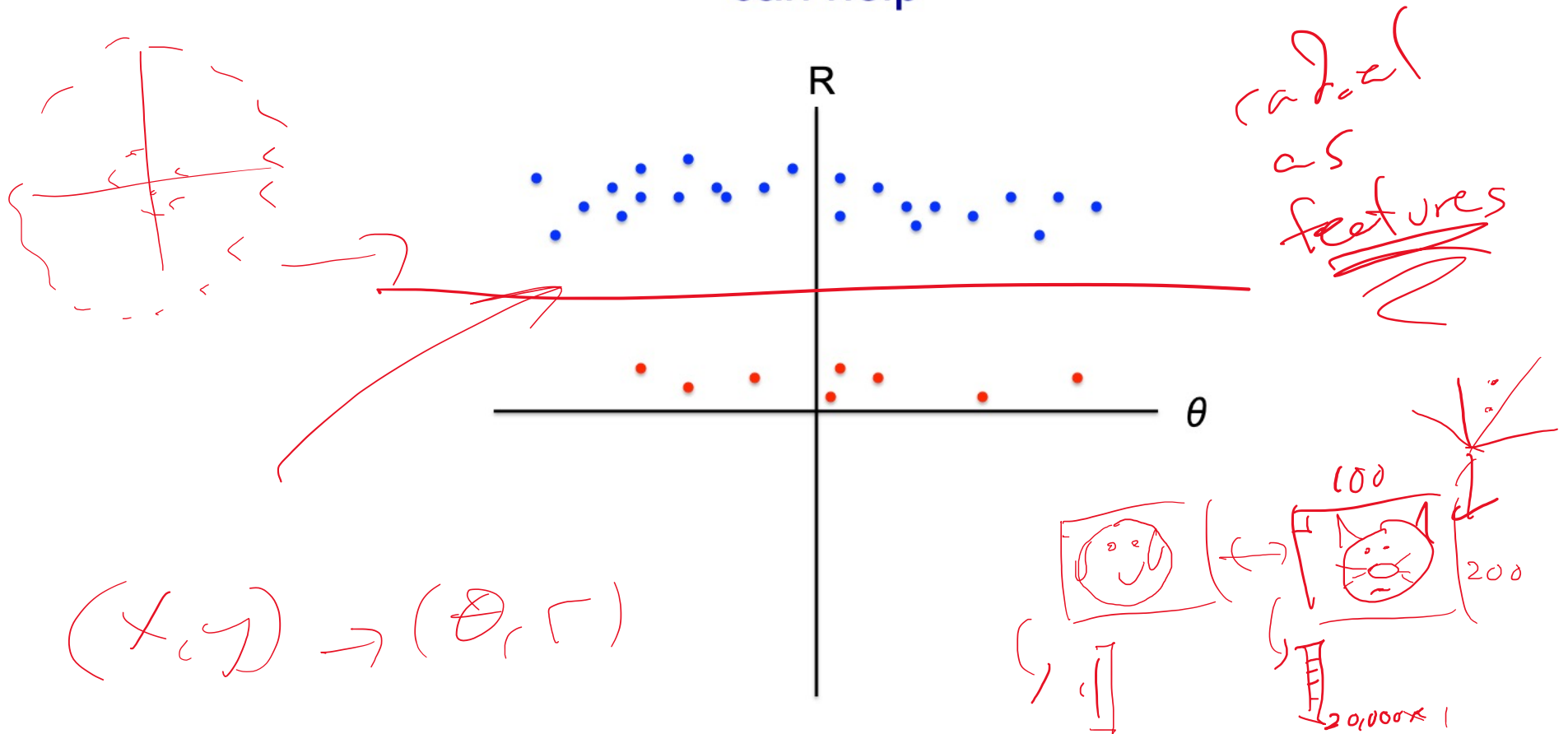
$$= \sum_i (x_i - c_i)^2$$



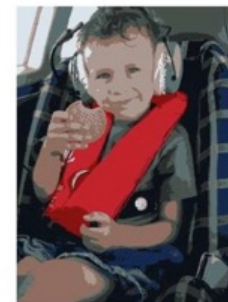
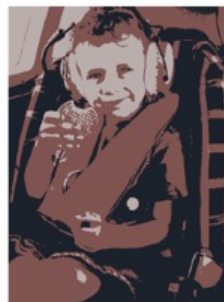
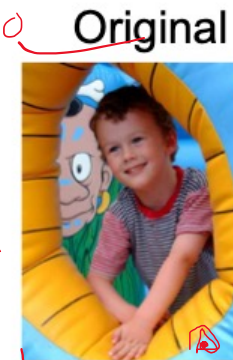
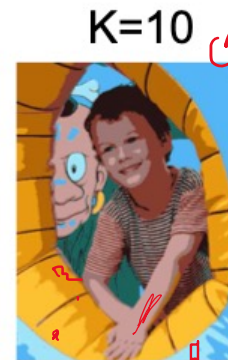
K-means not able to properly cluster



Changing the features (distance function) ~~can~~
can help



Example: K-Means for Segmentation



200 Original

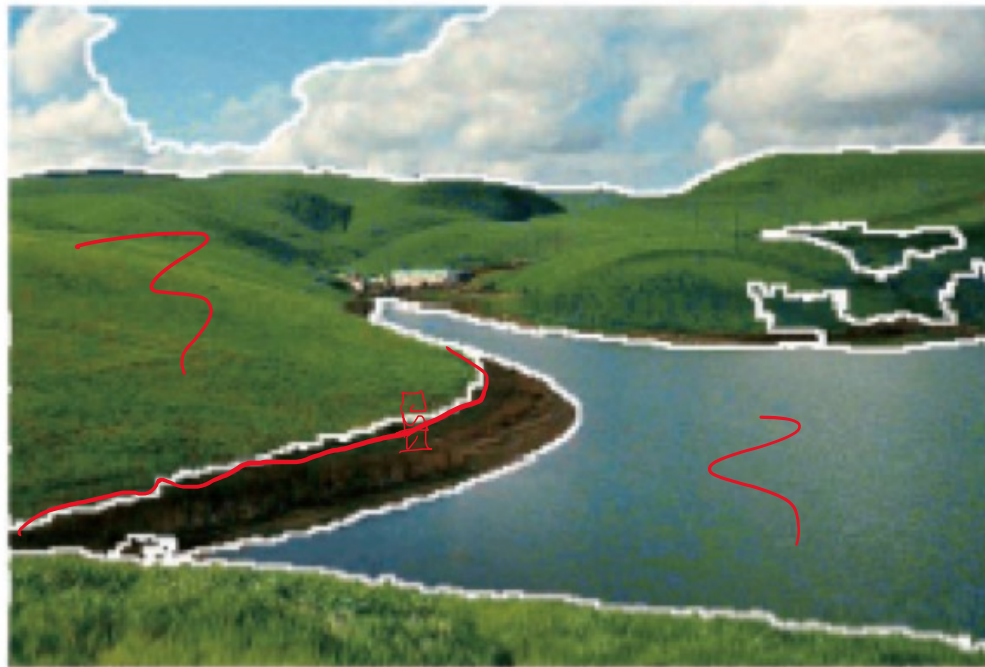
$\sum_{i=1}^n \sum_{j=1}^K z_{ij} = 20,000$

$z = \{z_{ij}\}_{i,j}$

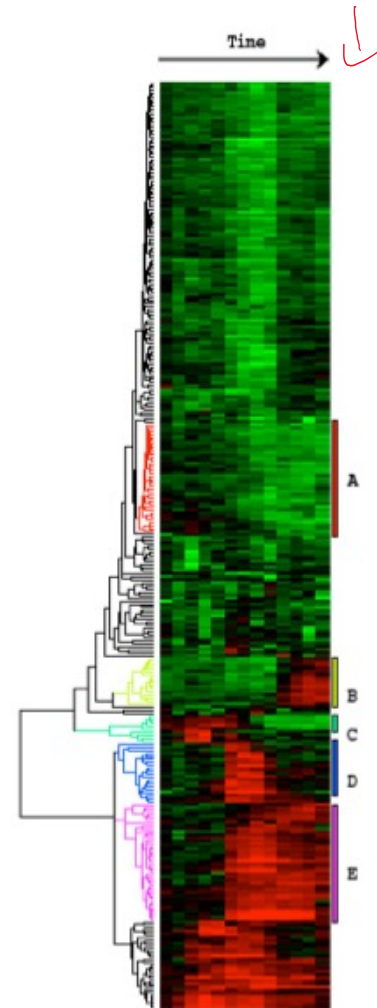
$75, 8$

Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions



Clustering gene expression data



Eisen et al, PNAS 1998

Cluster news articles



Google News interface showing top stories and recommended articles. The interface includes a search bar, navigation tabs (U.S. edition, Classic), and a sidebar with categories like Top Stories, Recommended, U.S., World, Sci/Tech, Business, More Top Stories, Health, Spotlight, Elections, Entertainment, Sports, Technology, and Science.

Top Stories

- Teen suspect saw movie moments after allegedly killing beloved Massachusetts ...**
Fox News - 8 minutes ago
The 14-year-old student who authorities say murdered a beloved math teacher at a Massachusetts high school admitted to police that he slashed her throat with a box cutter, a source told MyFoxBoston.
Colleen Ritzer, slain Danvers High School teacher, remembered as passionate ... CBS News
14-Year-Old Charged in Brutal Murder of Massachusetts Teacher New York Magazine
Highly Cited: 14-year-old student held without bail in slaying of Danvers High teacher Boston.com
Opinion: Heslam: Heartbroken friends say Colleen was born to teach Boston Herald
In Depth: Student, 14, arraigned in murder of Mass. teacher USA TODAY
Wikipedia: Danvers, Massachusetts
[See realtime coverage »](#)
- Obamacare contractors tell their stories at congressional hearing**
CNN - 40 minutes ago
Washington (CNN) -- [Breaking news update at 10:09 a.m.], [URGENT - Congress-Obamacare-Testing]. (CNN) -- A contractor on the problem-plagued government website for President Barack Obama's signature health care reforms said Thursday his ...
Hearing on health care website today to focus on blame WXXA-TV
Contractors Point Fingers Over Health-Law Website AIThingsD
[See realtime coverage »](#)
- EU leaders meet amid concern about US spying claims**
CNN - 1 hour ago
(CNN) -- European Union leaders are meeting Thursday in Brussels for a summit that may be overshadowed by anger about allegations that the United States has been spying on its European allies.
Germany summons US ambassador over spying claims USA TODAY
Germany Summons US Envoy Over Alleged NSA Spying ABC News
Highly Cited: Readout of the President's Phone Call with Chancellor Merkel of Germany Whitehouse.gov (press release)
From Germany: Press Review: Outrage over NSA eavesdropping Deutsche Welle
Opinion: The Handyüberwachung Disaster New York Times
In Depth: US ambassador to Germany summoned in Merkel mobile row BBC News
[See realtime coverage »](#)
- US jobless claims miss forecasts, trade deficit widens slightly**
Reuters - 59 minutes ago
WASHINGTON | Thu Oct 24, 2013 9:19am EDT. WASHINGTON (Reuters) - The number of Americans filing new claims for unemployment benefits fell less than expected last week, but a lingering backlog of applications in California makes it difficult to get a ...
Weekly Jobless Claims Fall to 350,000 Fox Business
How States Fared on Unemployment Benefit Claims ABC News
In Depth: More Americans Than Forecast Filed Jobless Claims Businessweek
[See realtime coverage »](#)
- Kennedy cousin gets new trial in 1975 killing of neighbor; victim's mother ...**

Recommended

- U.S.
- World
- Sci/Tech
- Business
- More Top Stories
- Health
- Spotlight
- Elections
- Entertainment
- Sports
- Technology
- Science

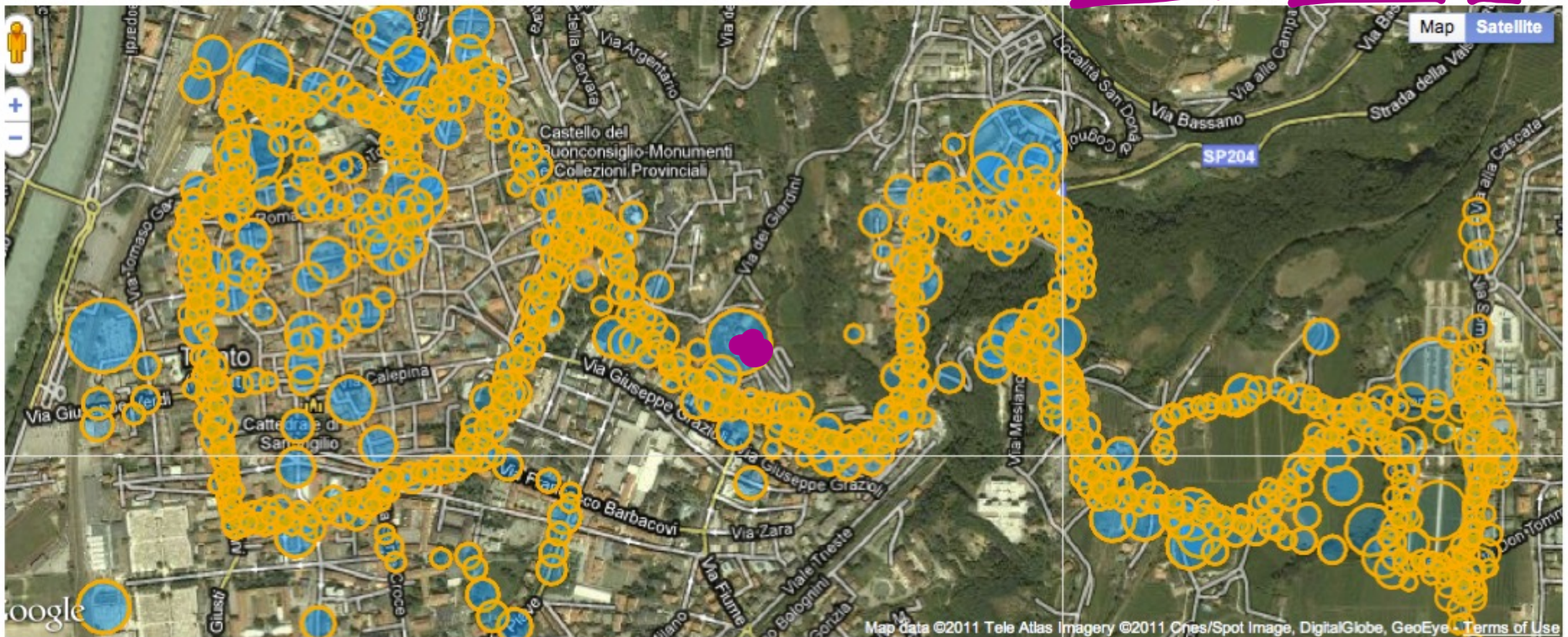
Images:

- ABC News: A woman smiling.
- Wall Street Journal: A woman smiling.
- National Post: Barack Obama and Michelle Obama.
- The Olympian: A person holding a sign that says "FAST TRACK to EMPLOYMENT".



Cluster people by space and time

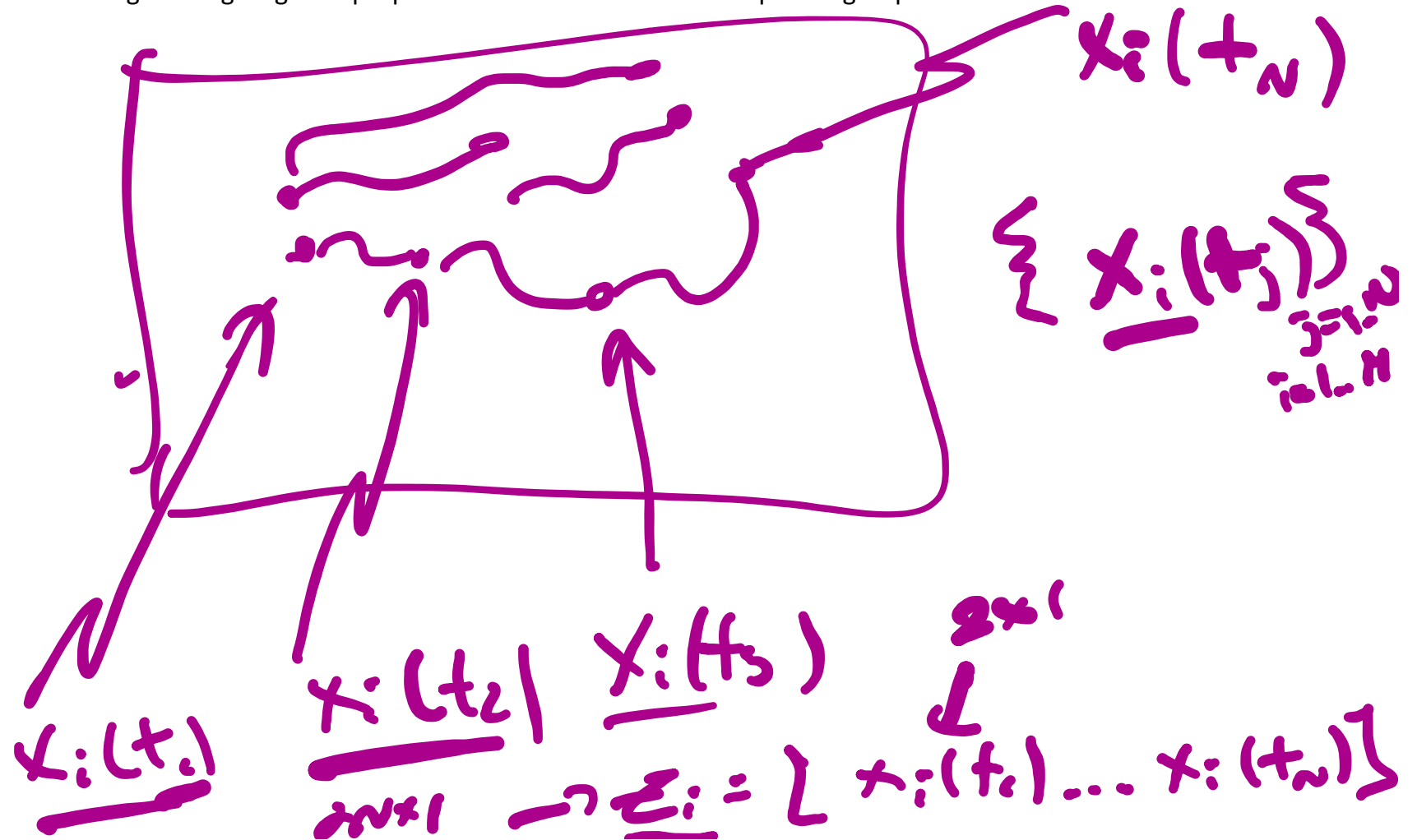
$$Z^i = (x_i^i, y_i^i, t_i^i), (x_2^i, y_2^i, t_2^i), (x_3^i, y_3^i, t_3^i)$$



[Image from Pilho Kim]

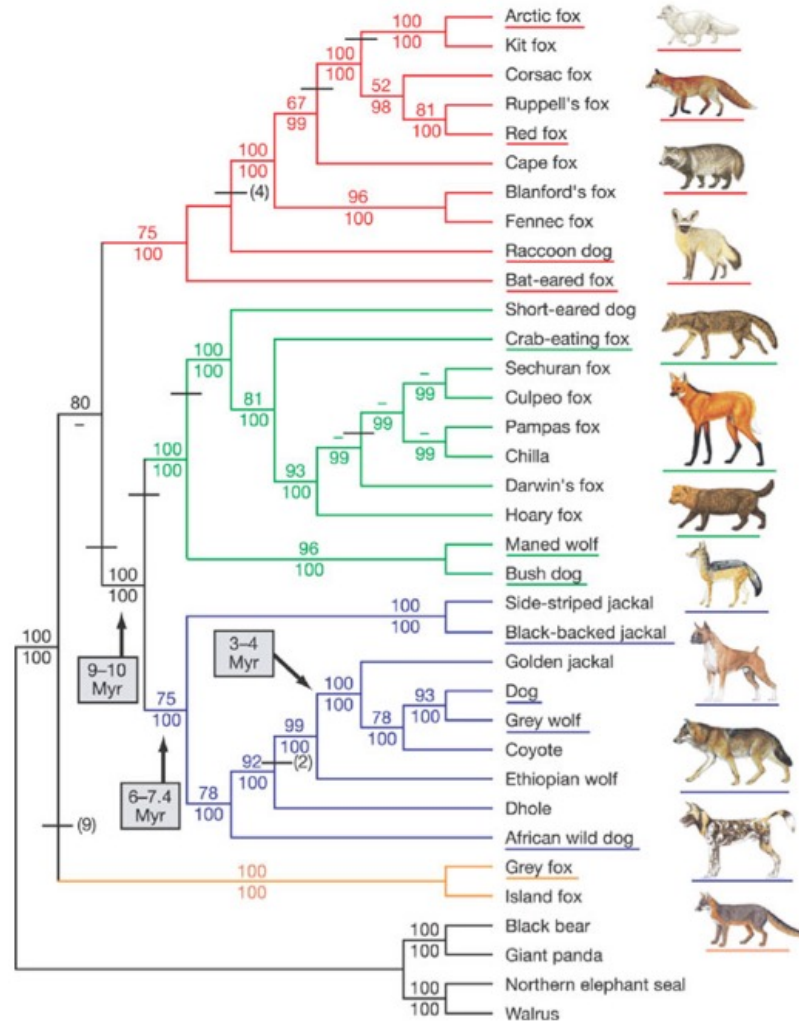
$$\{Z_i\}_{i=1}^N$$

Clustering on tracking moving targets – prepare a vector – feature - corresponding to position over time.



Clustering species ("phylogeny")

[Lindblad-Toh et al., Nature 2005]



Brain Tumor Detection and Identification Using K-Means Clustering Technique

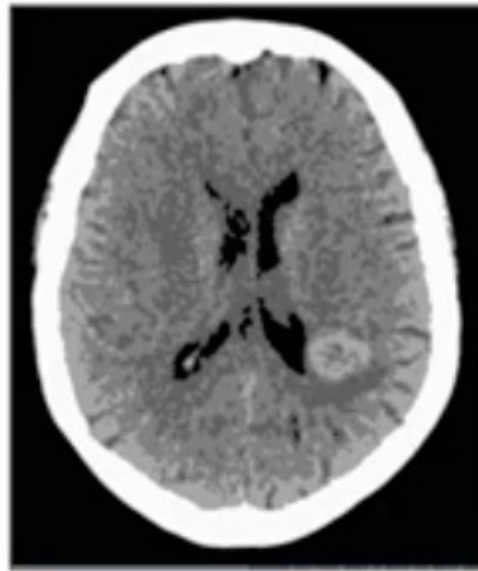
Malathi R

Department of Computer Science, SAAS College, Ramanathapuram, Email: malapraba@gmail.com

Dr. Nadirabanu Kamal A R



Original Image

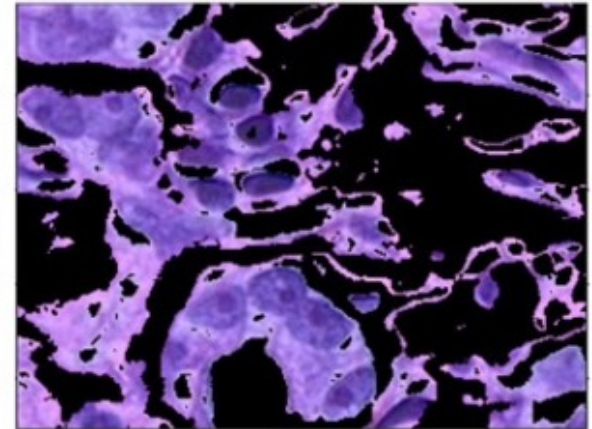
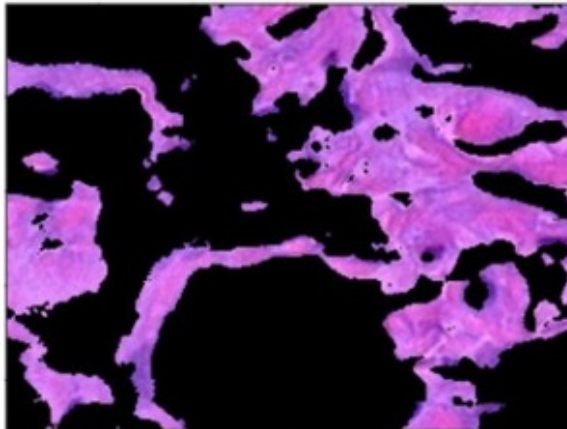
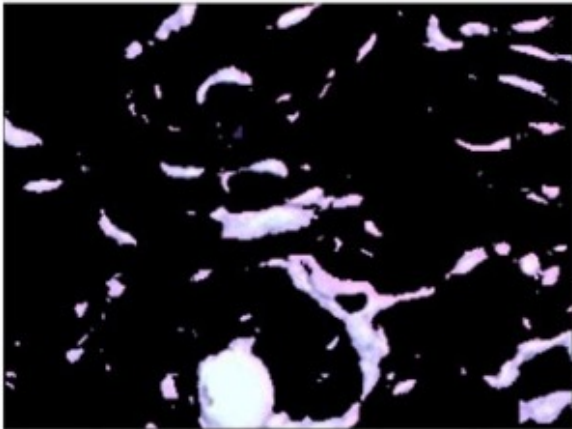
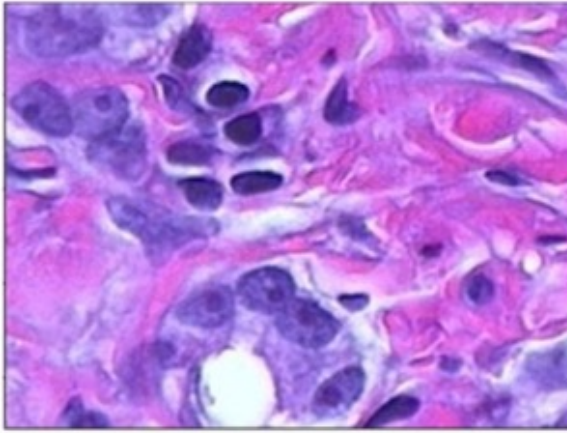


Segmented Image with #partitions = 5



Extracted Image with label = 2

Medical imaging – cancer tissue



DETECTING AND COUNTING THE NO. OF WHITE BLOOD CELLS IN BLOOD SAMPLE IMAGES BY COLOR BASED K-MEANS CLUSTERING

¹Neha Sharma, ²Nishant Kinra

¹Indira Gandhi Delhi Technical University for Women, New Delhi, India

²Deenbandhu Chhotu Ram University of Science and Technology, Haryana, India

¹neha.sksharma@yahoo.co.in, ²nishant.kinra@gmail.com

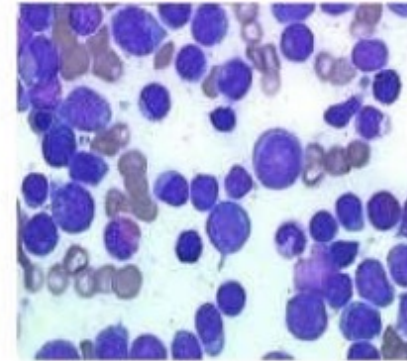


Fig (1): Example of Leukocytosis

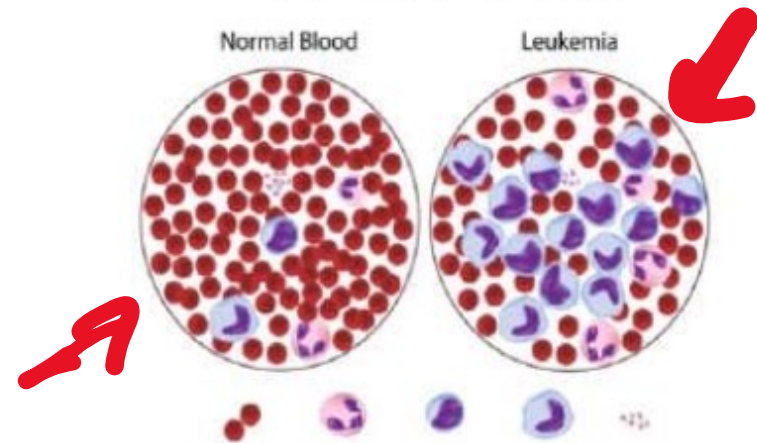


Fig (2): Comparison between Normal Blood and Leukemia

So that's it for the time being with unsupervised learning, a few clustering methods – mostly kmeans,

But in your book are other interesting clustering methods in 5.4, 5.5, including tree methods and also Gaussian Mixture models that are very popular.

Later we will also develop a spectral clustering method that is very general and powerful.

Also in unsupervised learning later we will do “manifold learning”.

But for now.....

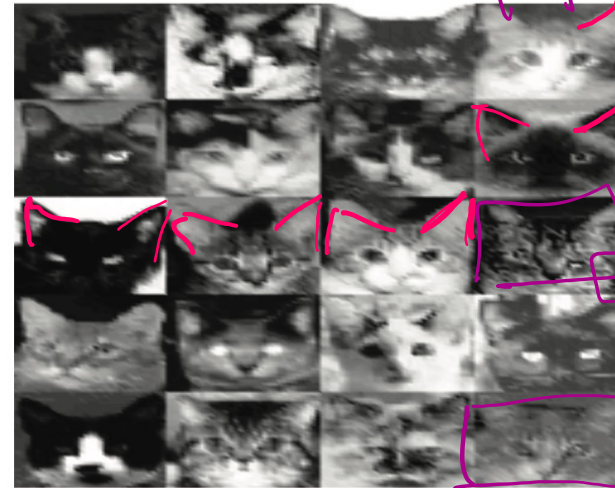
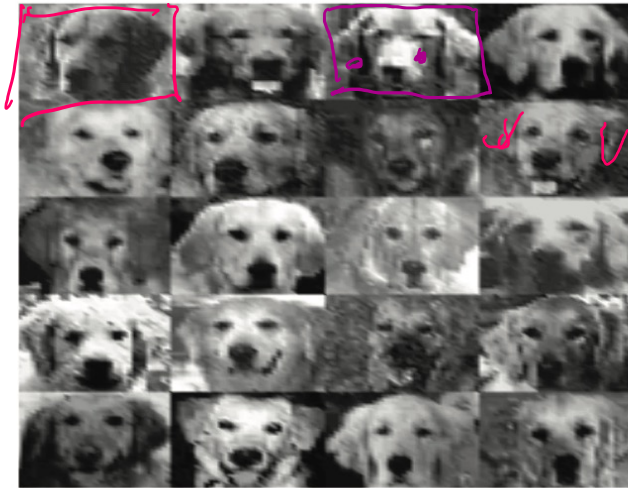
We transition to supervised learning

1st in 5.6 we will cover Fischer's Linear Discriminant (LDA)

2nd in 5.7 on to support vector machines (SVM)

Then Chapter 6 a grandly popular method – artificial neural nets.

Supervised Learning – Cat or Dog?



$$D = \{ (\underline{x_i}, \underline{y_i}) \}_{i=1}^N$$

$\rightarrow f: \underline{x} \mapsto \underline{y}$

$x_i = i^{th} \text{ picture}$
 $y_i = i^{th} \text{ label}$
 $f: \mathbb{R}^{20,100} \rightarrow \mathbb{Z}_2$

$y_i = \text{dog}$

$x_i = i^{th} \text{ picture}$

$\rightarrow [x_i, y_i]$

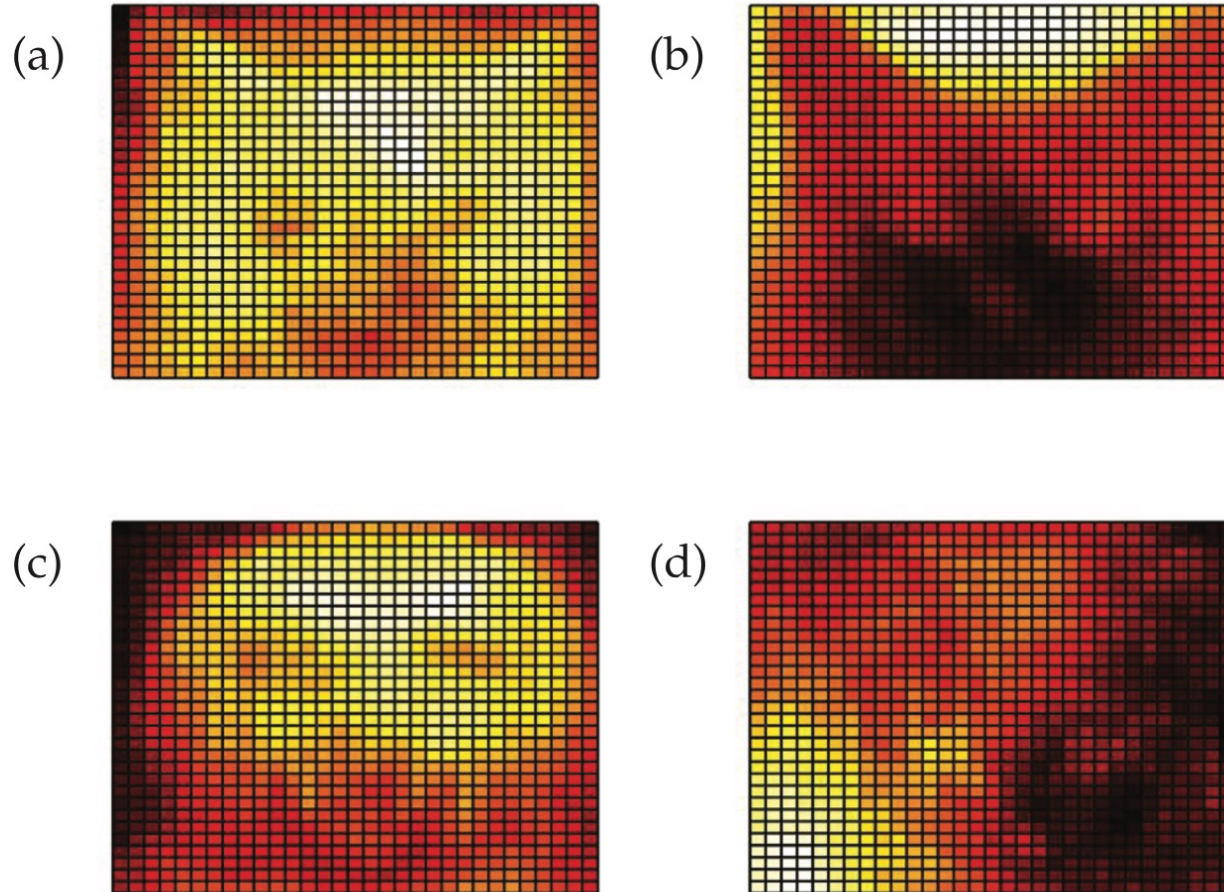
$y_i = \text{cat}$

160

100

200

To distinguish cat's from dogs – first it will be more efficient if **we choose good (efficient) features.**



Wavelets then PCA (SVD) will work well.

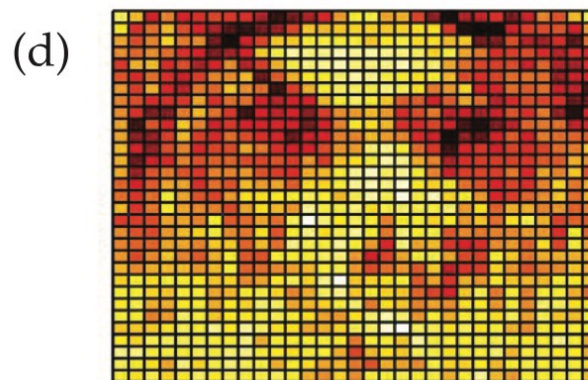
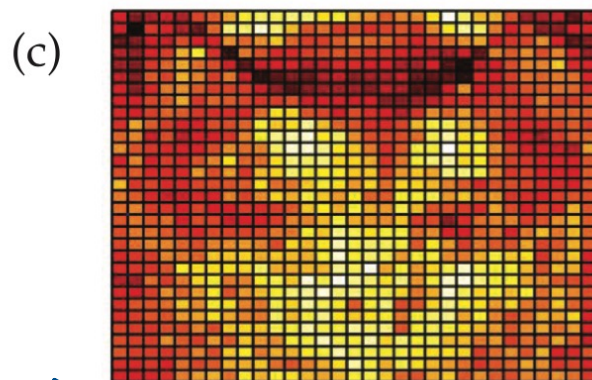
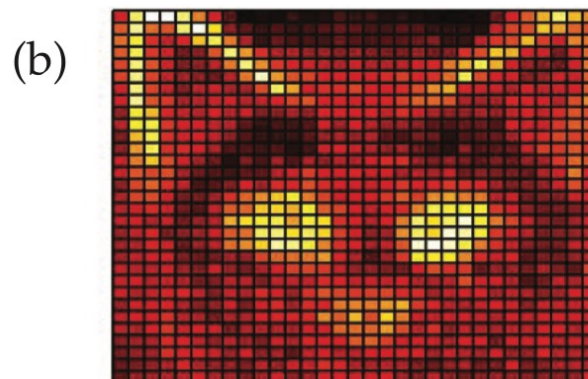
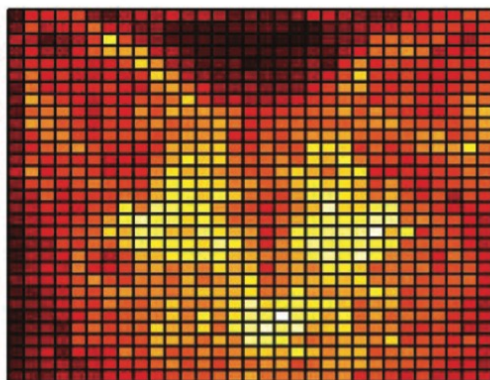
unsupervised -

$$D = \{(x_i)\}$$

structure

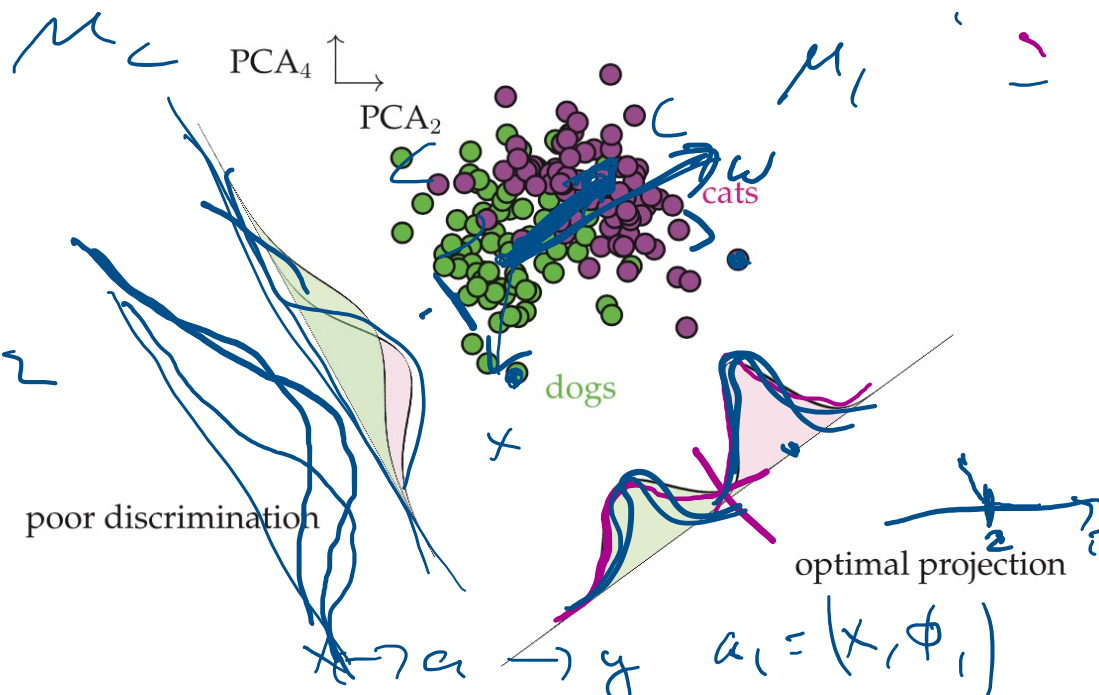
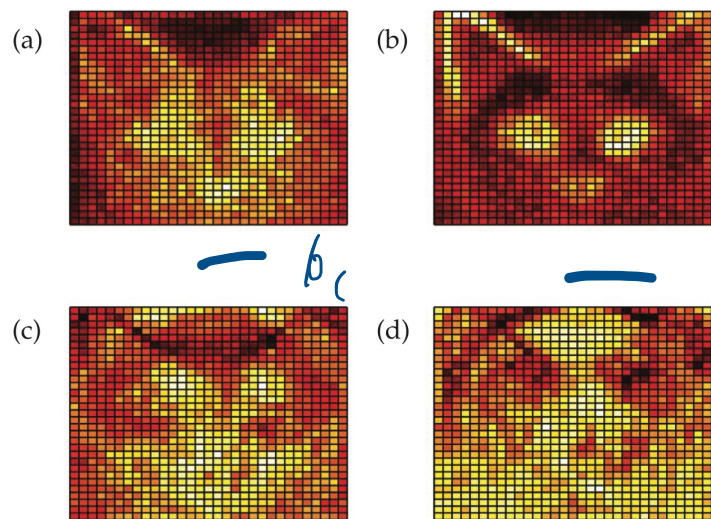
cluster

ROM



$$D = \{(x_i, y_i)\} \text{ supervised - classify, } y = -1 \text{ or } 1.$$

Fischer's Linear Discriminant Analysis LDA



$$\mathbf{w} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

$$\mathbf{S}_W = \sum_{j=1}^2 \sum_{\mathbf{x}} (\mathbf{x} - \mu_j)(\mathbf{x} - \mu_j)^T$$

where the scatter matrices for between-class \mathbf{S}_B and within-class \mathbf{S}_W data are given by

$$\mathbf{S}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

$$= \int_{\Omega} x(s) \phi_1(s) ds$$

$$= \sum x(s_i) \phi_1(s_i)$$

Rayleigh quotient whose solution can be found via the generalized eigenvalue problem

LDA – is a supervised learning problem of type – classification. Assumption – the data are Gaussian within class.

Suppose two classes of observations have **means** $\vec{\mu}_0, \vec{\mu}_1$ and covariances Σ_0, Σ_1 . Then the linear combination of features $\vec{w}^T \vec{x}$ will have **means** $\vec{w}^T \vec{\mu}_i$ and **variances** $\vec{w}^T \Sigma_i \vec{w}$ for $i = 0, 1$. Fisher defined the separation between these two **distributions** to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2} = \frac{(\vec{w} \cdot \vec{\mu}_1 - \vec{w} \cdot \vec{\mu}_0)^2}{\vec{w}^T \Sigma_1 \vec{w} + \vec{w}^T \Sigma_0 \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T (\Sigma_0 + \Sigma_1) \vec{w}}$$

the scatter matrices for between-class S_B and within-class S_W data

$$\mathbf{S}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

$$\mathbf{S}_W = \sum_{j=1}^2 \sum_{\mathbf{x}} (\mathbf{x} - \mu_j)(\mathbf{x} - \mu_j)^T.$$

$$\mathbf{w} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \quad ; \quad \hat{\omega} = \underset{\omega}{\operatorname{argmax}} J(\omega)$$

$$\begin{aligned} & \min -\frac{1}{2} \omega^T S_B \omega \\ & \text{s.t. } \omega^T S_W \omega = 1 \end{aligned}$$

$I \Rightarrow \Pi$



$$\begin{aligned} & (\mu_2 - \mu_1) \cdot v \\ & = (\mu_1 - \mu_1)^T v \\ & = v \cdot (\mu_2 - \mu_1) \\ & = v^T (\mu_2 - \mu_1) \end{aligned}$$

$$((\mu_2 - \mu_1) \cdot v)^2 = v^T \underbrace{(\mu_2 - \mu_1) \cdot (\mu_2 - \mu_1)}_{S_B} v$$

$$J = \frac{1}{2} \omega^T S_B \omega + \frac{\lambda}{2} (\omega^T S_W \omega - 1)$$

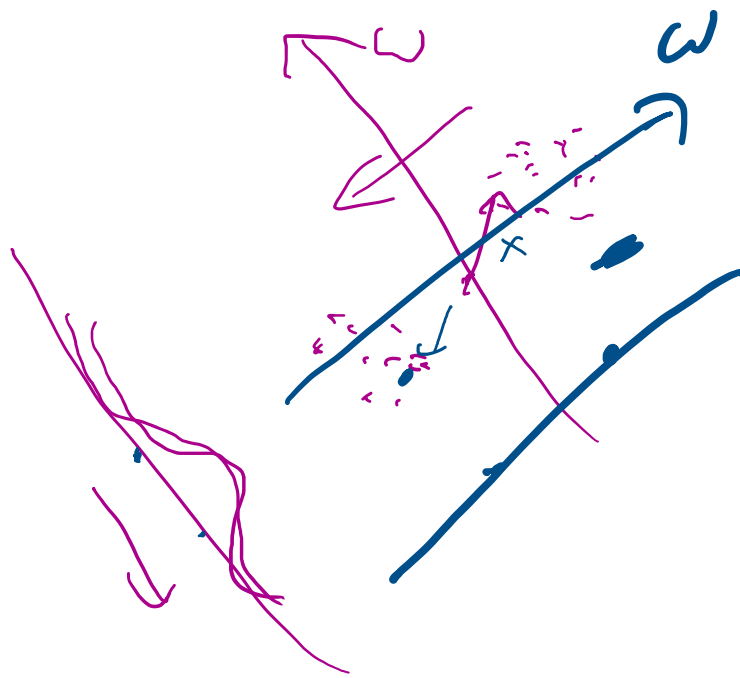
$$\text{let } f = \frac{1}{2} \omega^T S_B \omega + \frac{1}{2} \lambda (\omega^T S \omega \omega - 1) \quad \leftarrow$$

$$\frac{\partial f}{\partial \omega} = 0; \quad \frac{\partial f}{\partial \omega_2} = 0, \dots; \quad \nabla_{\omega} f = 0$$

$$\nabla_{\omega} f = -S_B \omega + \lambda S \omega \omega = 0 \quad \left[\begin{smallmatrix} K & K & T \end{smallmatrix} \right]$$

$$S_B \omega = \lambda S \omega \omega \quad \left[\nabla \left(\frac{1}{2} \omega^T \omega \right) = \omega \right]$$

$Ax = \lambda Bx$ generalized eigenvalue problem statement!



$$f: \underline{X} \rightarrow \underline{Y}$$

$$\uparrow \qquad \qquad \qquad \uparrow$$

$$f: \underline{X} \rightarrow ?$$

$$\uparrow$$

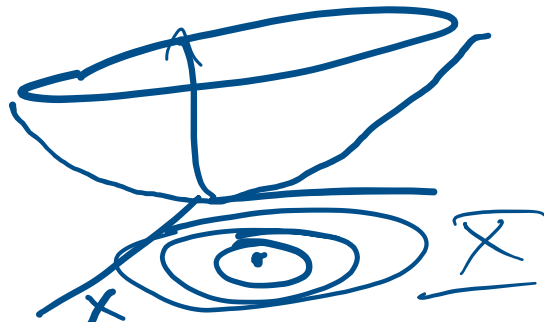
$$(x, \omega)$$

KKT: $x^* = \arg \min_x f(x)$

subj. $h_i(x) = 0, \forall i = 1, \dots, m; \quad \vec{\lambda}^T \vec{h}(x) = 0$

$g_i(x) \leq 0; \forall i = 1, \dots, n$ ~~$\vec{g}(x)^T \vec{\mu} \leq 0$~~

$x^* = \arg \min_{x, \lambda, \mu} \mathcal{L}(x, \lambda, \mu) = \arg \min_x f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^n \mu_j g_j(x)$



$f: \mathcal{X} \rightarrow \mathbb{R}$

X

Case A

not active constraint

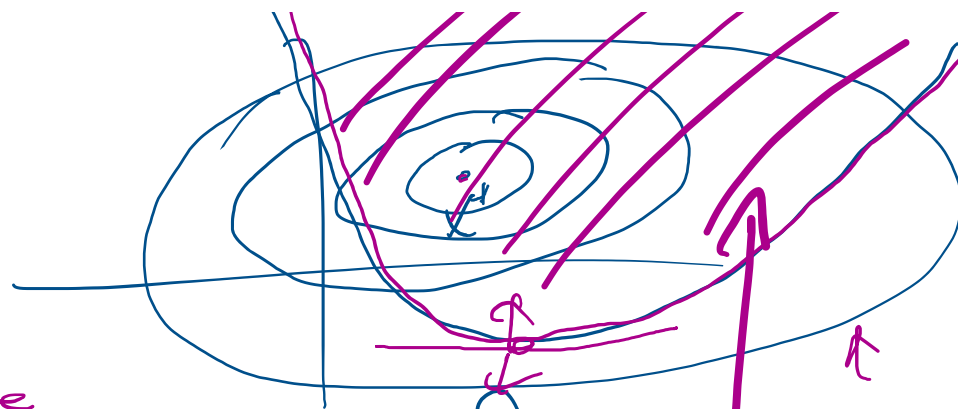
- - -

Case B

Active

tangent.

$g_i(x) = 0$ and f smallest.



$$g_i(x) = 0$$

Kkt, look for where level sets are tangent.



$$g_i(x) \leq 0$$

not active constraint

far

min $f(x)$ subj

$$g_i(x) \leq 0$$

KKT $\nabla g \parallel \nabla f$ at constrained opt.

$$\lambda \nabla g = \nabla f$$

$$\nabla f + \lambda \nabla g = 0 \quad \dots \text{Lagrange multiplier.}$$

Equality constr.

$$\Rightarrow \nabla_1 f + \sum \nabla_2 \lambda_i h_i(x) + \sum \mu_i \nabla_2 g_i(x) = 0$$

• stationarity. \leftarrow

• ineq. constr true.

Haar Wavelet

Define

$$\psi(x) \equiv \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Mother function.

Math word
~ Math notation

and

$$\psi_{jk}(x) \equiv \psi(2^j x - k)$$

gradic

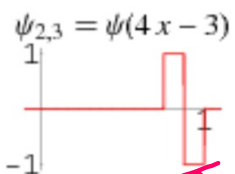
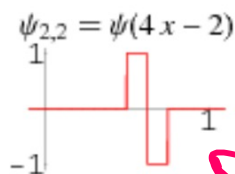
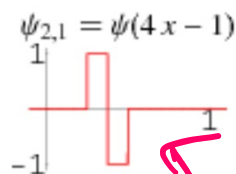
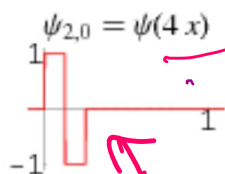
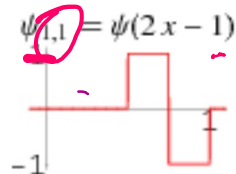
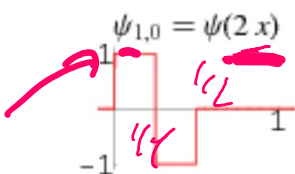
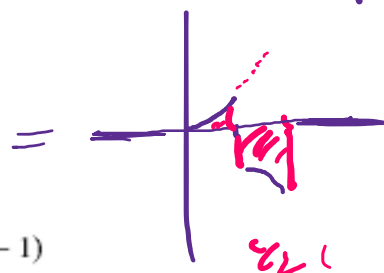
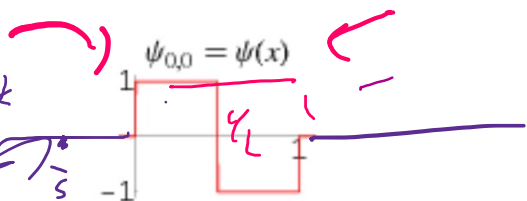
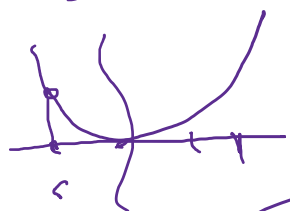
$$\psi_{jk}(x)$$

$$(x_i, \psi_k)$$

for j a nonnegative integer and $0 \leq k \leq 2^j - 1$.

$$\int x_i(s) \psi_k(s) ds$$

$$x(s) = s^2$$



$$\begin{aligned} \phi_1(s) &= \psi_{0,0} \\ \phi_2(s) &= \psi_{1,0} \\ \phi_3(s) &= \psi_{1,1} \end{aligned}$$


cos x

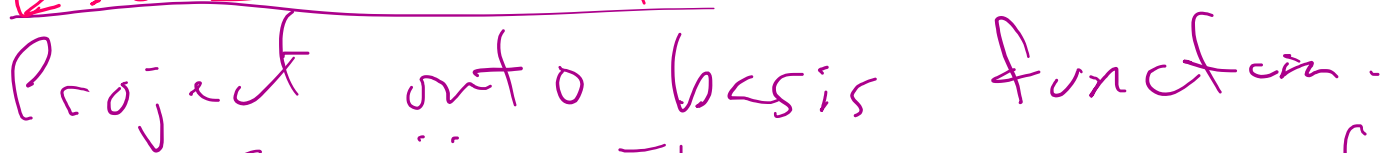
cos 2x

Better features

$$\{\phi_k\}_{k=1}^K = \begin{cases} \{\gamma_k\} & \text{indicator functions} \\ \underline{e^{-iks}} & \text{fourier basis} \leftarrow \\ \underline{\psi_k(s)} & \text{haar wavelet basis} \end{cases}$$

$$e^{-iks} = \cos ks + i \sin ks$$

$$\textcircled{a_k} = (\textcircled{X_k} \cdot \textcircled{\phi_k}) = \int X_k(s) \phi_k(s) ds = \sum X_k(s) \phi_k(s)$$




= 20,000 The values in the pixels
each pixel is considered a feature. (characterist

Characteristic
fn: $\chi_k(s) = \begin{cases} 1 & s \leq k \\ 0 & s > k \end{cases}$

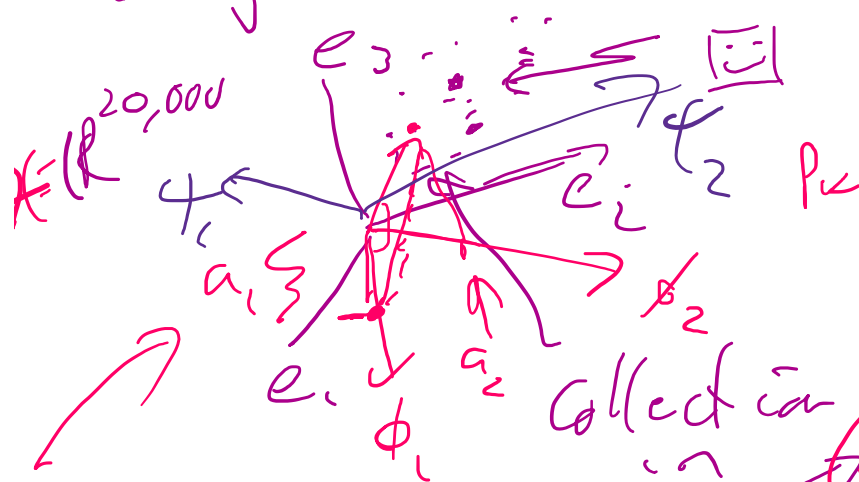
$$\underline{\underline{f}}(x_i, \underline{\underline{f}}_k) = \int \underline{\underline{x}}_i(s) \underline{\underline{f}}_k(s) ds$$

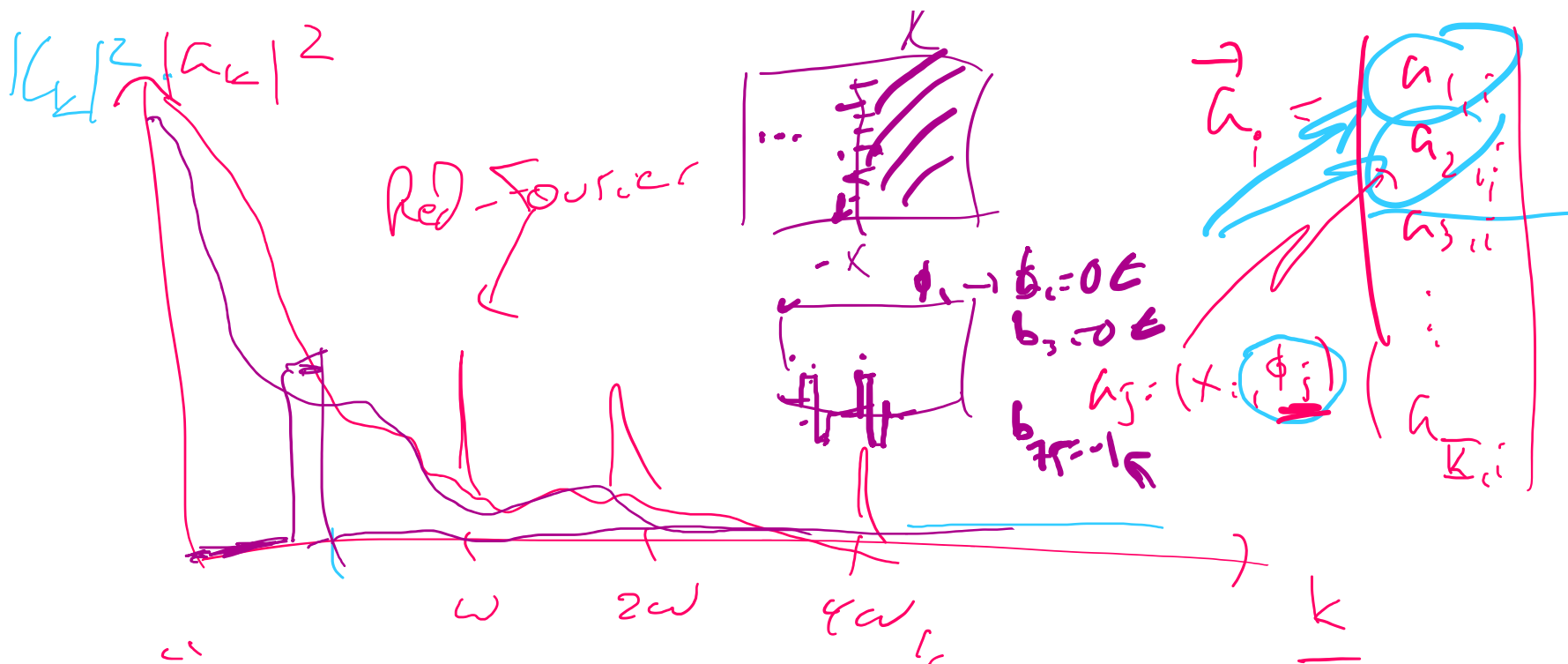
$$\vec{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$$

↑ features

Date done

$$R_H = (R_{20000})$$





Power spectrum.

$$\int |x(s)|^2 ds \stackrel{(*)}{=} \sum |a_k|^2 \leftarrow \text{Parseval}$$

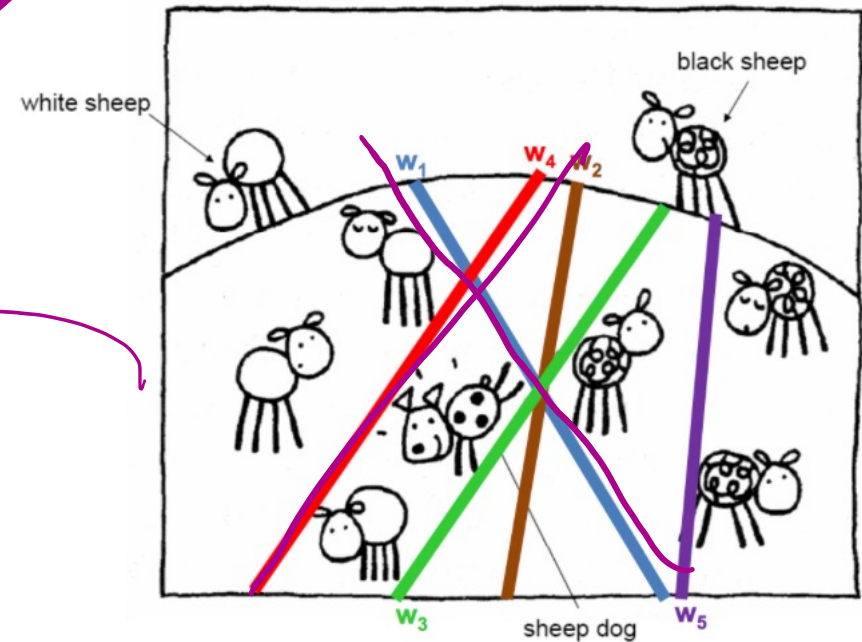
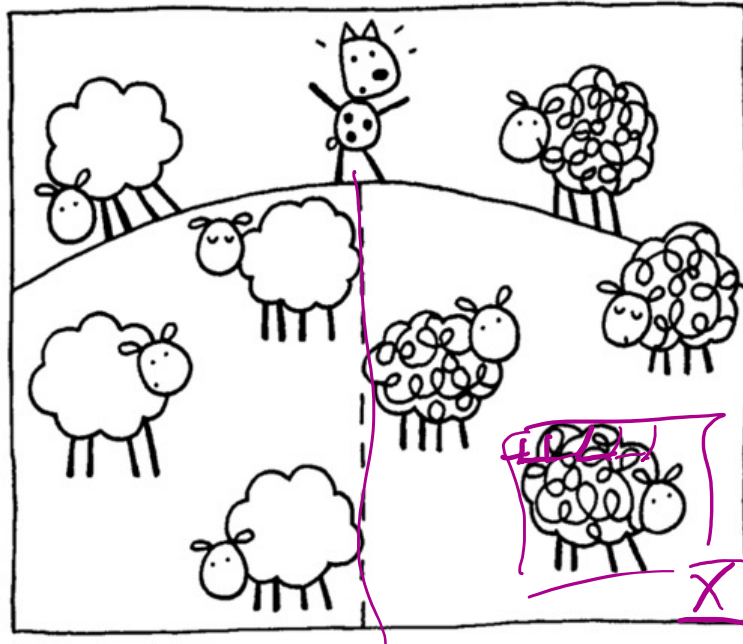
Support Vector Machines (SVM) — 5.7 *linear*

Then

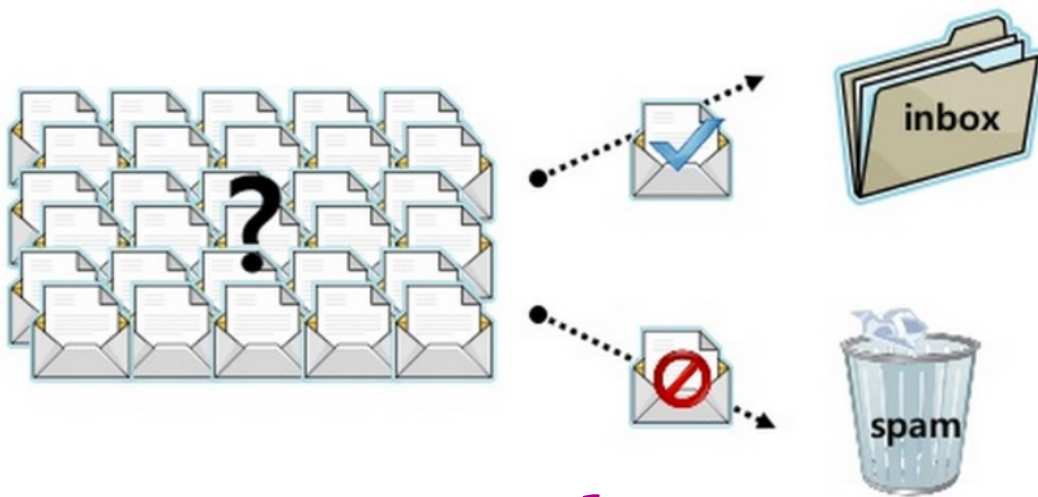
Nonlinear (kernelized) SVM (KSVM)

Wide Margin Decision Hyperplane for
Supervised - Learning Classification

First a linear binary classification – decision boundary/hyperplane



Sauza - Schökopff.



- Instance space: $\mathbf{x} \in X$ ($|X| = n$ data points)
 - Binary or real-valued feature vector \mathbf{x} of word occurrences
 - d features (words + other things, $d \sim 100,000+$)
- Class: $y \in Y$
 - y : Spam (+1), Ham (-1)

$\{(x_i, y_i)\}_{i=1}^n$

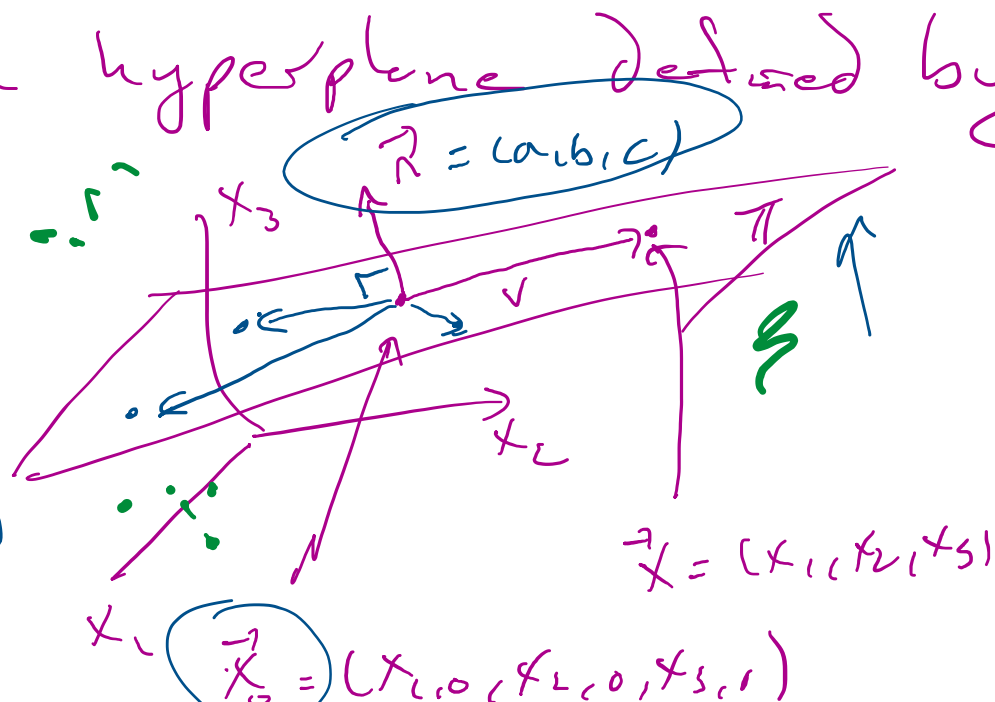
$y := \begin{cases} +1 \\ -1 \end{cases}$
 $\mathcal{Y} = \{+1, -1\}$

Viagra	Learning	The	Dating	Nigeria	Is_spam
1	0	1	0	0	1
0	1	1	0	0	-1
0	0	0	0	1	1



Review - a hyperplane defined by a vector.

• 1 eqn is a co-dim-1 restriction of space



$$\vec{x} = (x_1, x_2, x_3)$$

$$\vec{x}_0 = (x_{1,0}, x_{2,0}, x_{3,0})$$

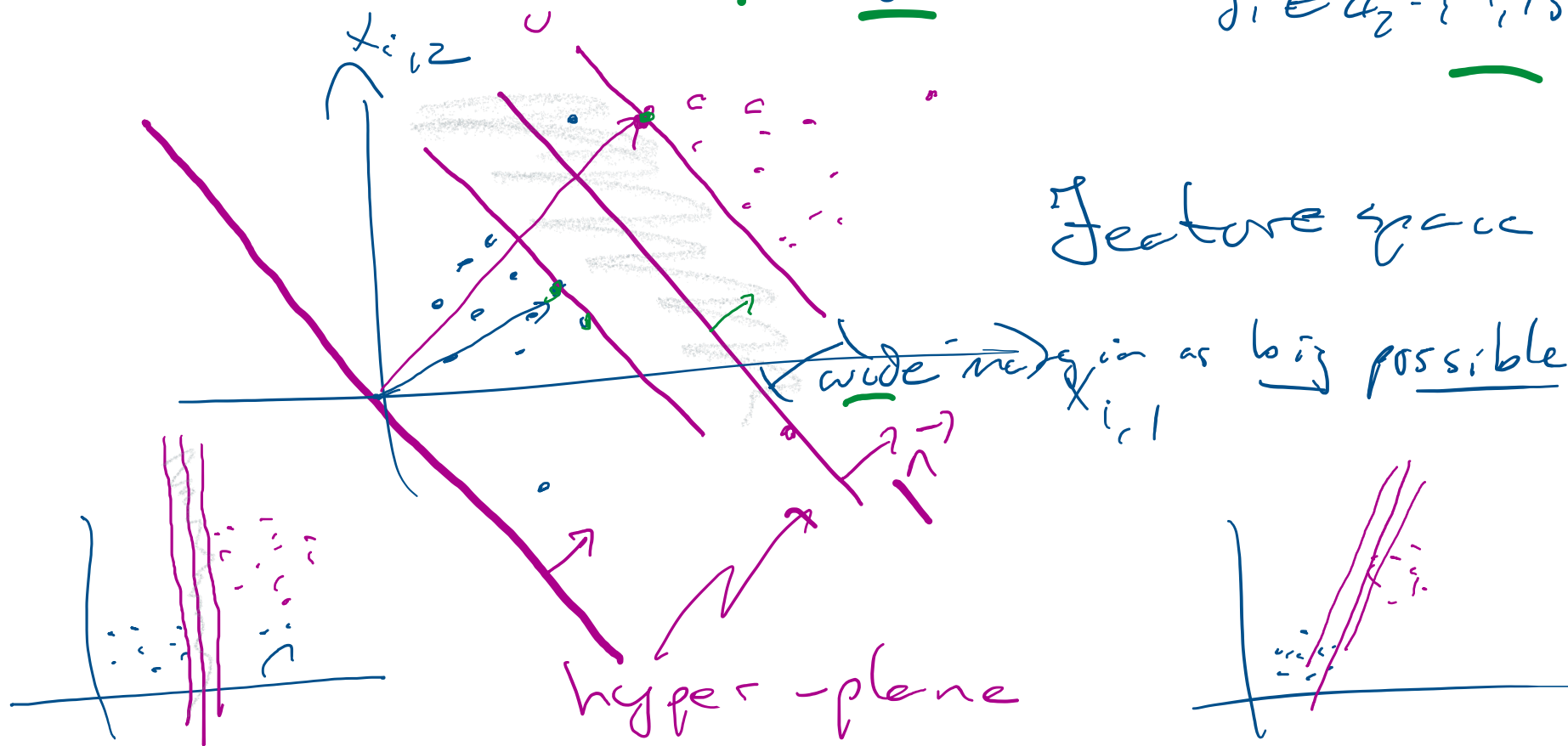
$$\vec{v} = \vec{x} - \vec{x}_0 = (x_1 - x_{1,0}, x_2 - x_{2,0}, x_3 - x_{3,0})$$

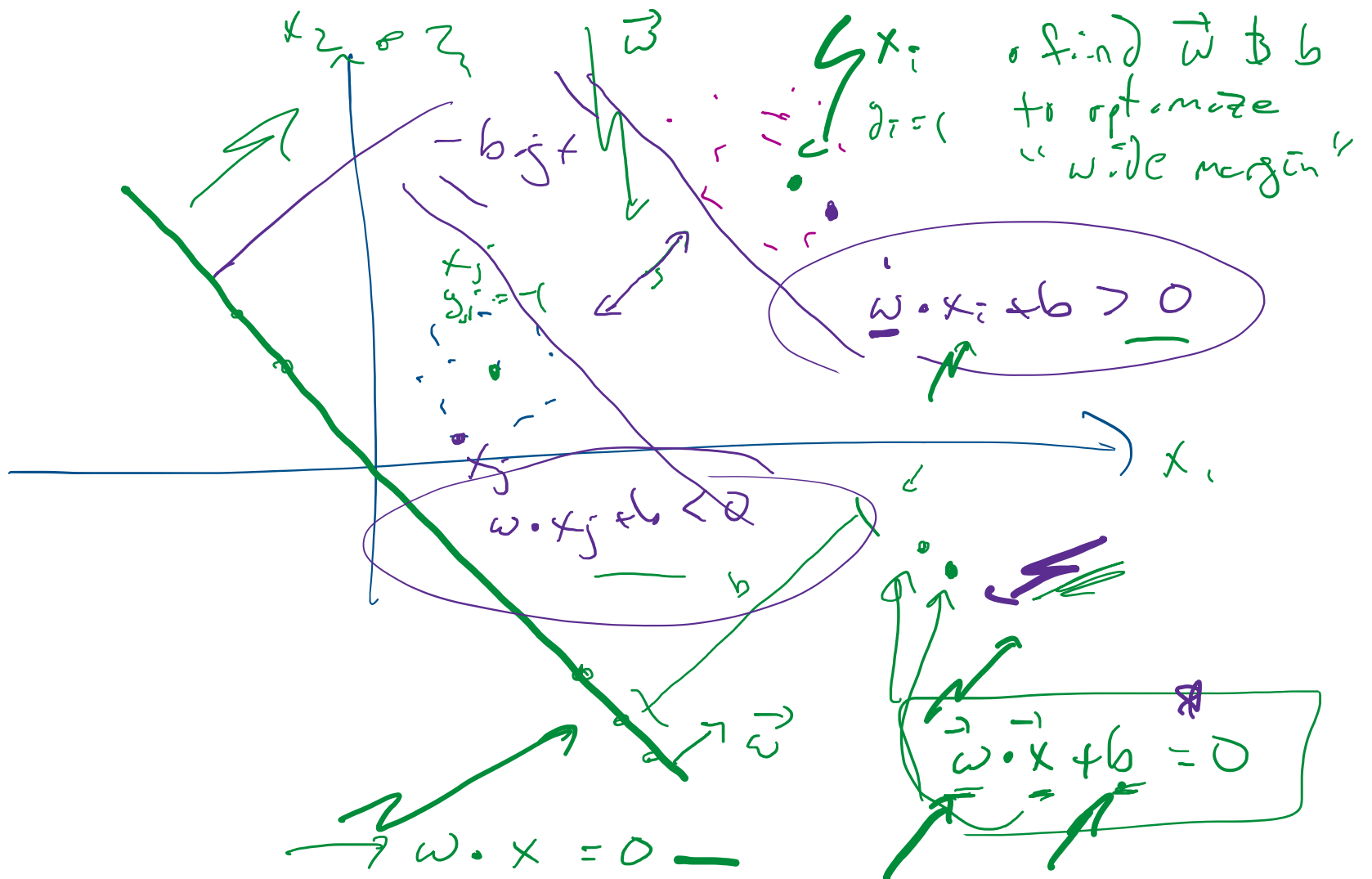
$$\pi := \{ \vec{x} = (x_1, x_2, x_3) : \vec{v} = \vec{x} - \vec{x}_0 \perp \vec{n} \}$$

$$\vec{v} \cdot \vec{n} = 0 \Leftrightarrow \vec{v} \perp \vec{n}$$

$$\vec{n} \cdot \vec{v} = \langle a, b, c \rangle \cdot \langle x_1 - x_{1,0}, x_2 - x_{2,0}, x_3 - x_{3,0} \rangle = a(x_1 - x_{1,0}) + b(x_2 - x_{2,0}) + c(x_3 - x_{3,0}) = 0$$

SVM: $D = \{ \underline{x_i}, \underline{y_i} \}_{i=1}^n$ $\underline{x_i} \in \mathbb{R}^d$ —
 $\underline{y_i} \in \mathbb{Z}_2 = \{-1, 1\}$ —





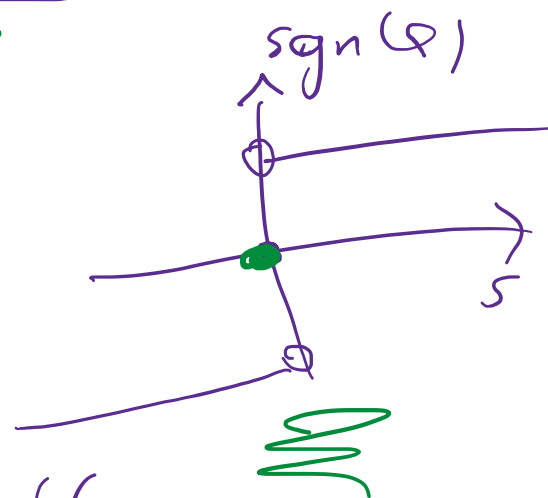
(\bar{x}_i, y_i)

$y_i = -1, 1$

$y_i = 0 \text{ or } 1$
apple or orange
dog or cat.

$y_i (\omega \cdot x_i + b) = \text{sgn}(\omega \cdot x_i + b)$ good label.

$\text{sgn}(s) = \begin{cases} 1 & \text{if } s > 0 \\ 0 & \text{if } s = 0 \\ -1 & \text{if } s < 0 \end{cases}$



$\text{sgn}(\omega \cdot x_i + b) = 1$ - labelled well -
 $\text{sgn}(\omega \cdot x_i + b) = -1$ - mislabelled. -



a loss function -

$$l(y_i, \bar{y}_i) = l(y_i, \text{sgn}(w \cdot \vec{x}_i + b)) = \begin{cases} 0 & \text{if correct label} \\ & y_i = \text{sgn}(w \cdot \vec{x}_i + b) \\ 1 & \text{incorrect label} \end{cases}$$

\bar{y}_i label you infer from \vec{x}_i alone
if you have a good hyperplane \vec{w} & b

Total loss

$$\sum_{i=1}^N l(y_i, \bar{y}_i)$$

Cost : ... ~~loss~~ \nearrow

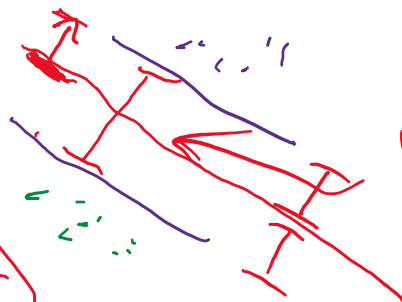
arg min $\frac{1}{2} \|\omega\|_2^2$

small $\|\omega\|$

subj $y_i (\omega^T x_i - b) - 1 = 0$

subj every matches truth

dist between



dist big

$\frac{2}{\|\omega\|_2^2}$
big

\Leftrightarrow

$\frac{\|\omega\|_2^2}{2}$
small

constrained opt. $(\vec{w}, \underline{b}) \leftarrow \vec{w} \in \mathbb{R}^d, d=2$
 $\leftarrow d+1=3$

$$f(\underline{x}, \underline{s}, \underline{\theta}) = \frac{1}{2} \|\underline{w}\|_2^2 - \sum_{i=1}^n s_i (y_i (\underline{w}^T \underline{x}_i - \underline{b}) - 1)$$

$\{(\underline{x}_i, y_i)\}$ (points to \underline{x}_i and y_i)
 $\frac{1}{2} \underline{w} \cdot \underline{w}$ (points to $\|\underline{w}\|_2^2$)
 $y_i (\underline{w}^T \underline{x}_i - \underline{b}) - 1 = 0$
 $y_1 (\underline{w}^T \underline{x}_1 - \underline{b}) - 1 = 0$
 $y_2 (\underline{w}^T \underline{x}_2 - \underline{b}) - 1 = 0$
 \vdots
 $y_n (\underline{w}^T \underline{x}_n - \underline{b}) - 1 = 0$

$$\nabla_{\underline{\theta}} f = \vec{0} = \frac{1}{2} \|\underline{w}\|_2^2 - \sum_{i=1}^n s_i y_i (\underline{w}^T \underline{x}_i - \underline{b}) \leftarrow \sum_{i=1}^n s_i$$

$$\nabla_{\underline{w}} f(\underline{x}, \underline{s}, \underline{\theta}) = \underline{w} - \sum_{i=1}^n s_i y_i \underline{x}_i = \vec{0} \leftarrow d = \text{dims of } \underline{w} \text{ \& feature}$$

$$\nabla_{\underline{b}} f = -\frac{\partial}{\partial \underline{b}} f = \sum_{i=1}^n s_i y_i = 0$$

$$w = \sum_{i=1}^n s_i y_i x_i \quad ; \quad \sum_{i=1}^n s_i y_i = \vec{s} \cdot \vec{y} = 0$$

\uparrow $x_i \in \mathbb{R}^d$
 \uparrow $\vec{s} \in \mathbb{R}^n$ n -values
 \uparrow $y_i \in \mathbb{R}$

Trick - KKT - Primal Dual form.

Primal Form. $\} \min_{\theta, s} \mathcal{L}(x, s, \theta) = \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^n s_i y_i (w^T x_i + b)$

Dual Form

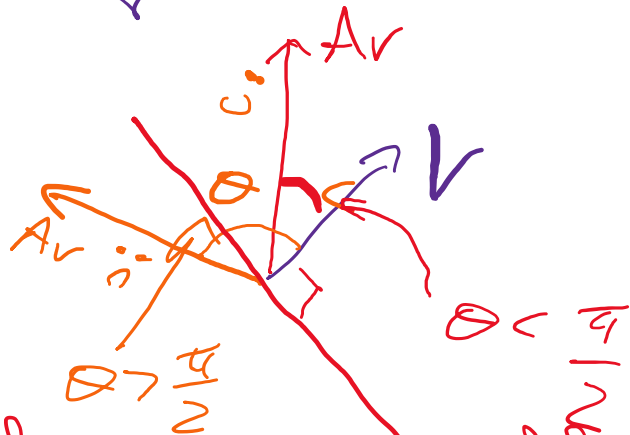
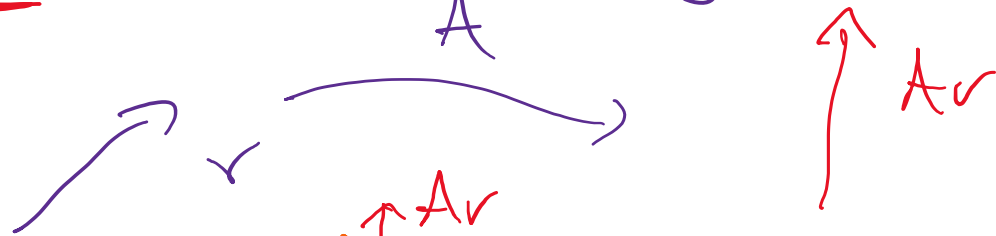
$$\max \mathcal{L}_D(x, s, \theta) = \sum s_i - \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n s_i s_j y_i y_j (x_i \cdot x_j)$$

\swarrow s.t. $w = \sum_{i=1}^n s_i y_i x_i$ and $\sum_{i=1}^n s_i y_i = 0$

$\phi(x_i) \cdot \phi(x_j)$

Pos. Semi-Defn Definition -

• A matrix $A_{n \times n}$ is ^{semi} positive definite if
 $\underline{v \cdot (A \cdot v) \geq 0}$ for any $\underline{v \neq 0}$ in domain of A .



$$v \cdot (Av) = \|v\| \|Av\| \cos \theta$$

• A kernel fn is pos. semi definite if 0, 1, 3, and \overline{K} pos. semi. def matrix for any input set

• \mathcal{H} = Hilbert space – a complete inner product space

• a Hilbert space is a set \mathcal{H} of vectors such that there is a

vector space



complete

inner product.



limits work properly.

Spectral Decomp. thm:

Suppose $A_{n \times n}$ is pos. defn. - symm. matrix with eigenvectors & eigenvalues

$$\lambda_i, v_i, \quad 0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

$$A = \sum_{i=1}^n \lambda_i \underbrace{v_i v_i^T}_{v_i \otimes v_i = P_i}$$

$$\text{not } v_i^T v_i = v_i \cdot v_i$$

= rank-1 projector

• $X_i = X_j \Rightarrow \angle \underline{X} = \mathbb{R}^2$
 need the dot product between
 $\underline{X_i}$ & $\underline{X_j}$ to do SVM.

$$\bullet \underline{K(X_i, X_j)} = \phi(X_i) \cdot \phi(X_j)$$

↑ corresponds.

Gram
matrix - symmetric

$$\phi: \underline{X} \rightarrow \mathcal{H} \subset \mathbb{R}^6$$

$$\underline{K} = \begin{pmatrix} \underline{K(X_1, X_2)} & K(X_1, X_3) & \dots & \underline{K(X_1, X_n)} \\ \vdots & & & \\ \underline{K(X_n, X_1)} & \dots & & K(X_n, X_n) \end{pmatrix}$$

KKT

Primal Problem:

$$\begin{cases} \text{minimize: } \mathcal{L}(x, s) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n s_i y_i (w^T x_i + b) + \sum_{i=1}^n s_i \\ \text{such that: } s_i \geq 0, \forall i \end{cases}$$

Dual Problem:

$$\begin{cases} \text{maximize: } \mathcal{L}_D(x, s) = \sum_{i=1}^n s_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n s_i s_j y_i y_j (\vec{x}_i^T \vec{x}_j) \\ \text{using: } w = \sum_{i=1}^n s_i y_i x_i, \text{ and } \sum_{i=1}^n s_i y_i = 0 \end{cases}$$

The amazing Kernel trick – nonlinear SVM through a kernel and all dot products in the high dimensional space
 Done through a kernel function

⑥ $\mathbb{X} = \mathbb{R}^2$

Now, we define a kernel $K : X \times X \mapsto \mathbb{R}$, which can take different forms such as:

• Linear kernel: $K(x, \tilde{x}) = x^T \tilde{x}$.

• Polynomial kernel: $K(x, \tilde{x}) = (x^T \tilde{x} + 1)^d$.

• Gaussian RBF: $K(x, \tilde{x}) = e^{-\frac{\|x - \tilde{x}\|^2}{2\sigma^2}}$

Consider the polynomial kernel, for $d = 2$, $X = \mathbb{R}^2$, then we have:

$$K(x, \tilde{x}) = (x \cdot \tilde{x} + 1)^d$$

$$= (x^T \tilde{x} + 1)^d$$

$$= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + 1)^2$$

$$= x_1^2 \tilde{x}_1^2 + 2x_1 \tilde{x}_1 + 2x_2 \tilde{x}_2 + x_1 \tilde{x}_1 x_2 \tilde{x}_2 + 1 + x_2^2 \tilde{x}_2^2$$

which interestingly can be re-written in terms of dot product:

- Kernel is
- ① symmetric
 - ② pos. semi-defn
 - ③ cts. w.r.t. both inputs.

$$K(x, \tilde{x}) = K(\tilde{x}, x)$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix}$$

$$K(x, \tilde{x}) = (x \cdot \tilde{x} + 1)^d$$

$$= (x^T \tilde{x} + 1)^d$$

$$= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + 1)^2$$

$$\Rightarrow = \underline{x_1^2 \tilde{x}_1^2 + 2x_1 \tilde{x}_1 + 2x_2 \tilde{x}_2 + x_1 \tilde{x}_1 x_2 \tilde{x}_2 + 1} + \underbrace{x_2^2 \tilde{x}_2^2}$$

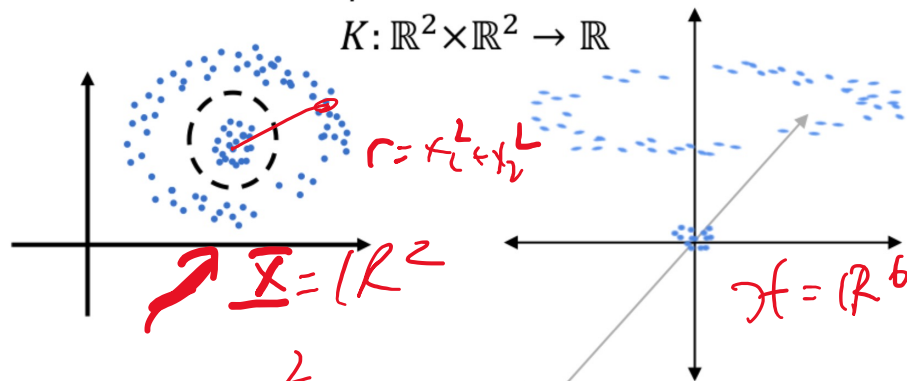
which interestingly can be re-written in terms of dot product:

$$K(x, \tilde{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_1x_2, x_2^2) \cdot (1, \sqrt{2}\tilde{x}_1, \sqrt{2}\tilde{x}_2, \tilde{x}_1^2, \tilde{x}_1\tilde{x}_2, \tilde{x}_2^2)$$

$$\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^6$$

$$K: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$$

$$K(x, \tilde{x}) = (\phi(x) \cdot \phi(\tilde{x}))$$



$$\phi: \underline{X} \rightarrow \mathcal{H}$$

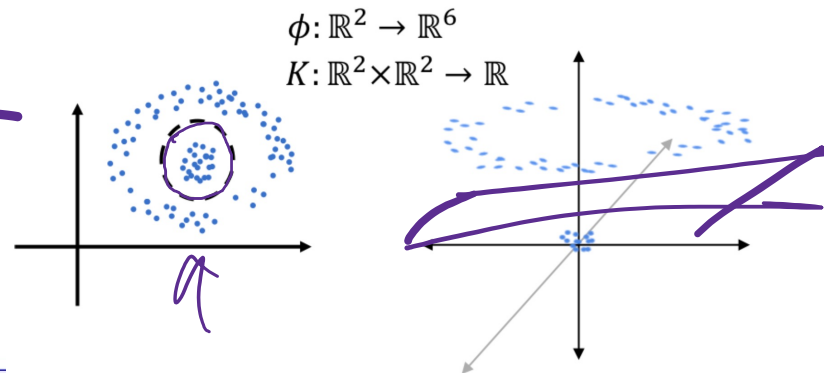
$$x \mapsto \phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_1x_2, x_2^2)$$

$d \uparrow$ DPT maybe
 $D = \infty$.

$\therefore \rightarrow \Delta \uparrow$
 n
 $n-1$

$$\phi(x_1, x_2) = (\phi_1(x_1, x_2), \phi_2(x_1, x_2), \dots, \phi_6(x_1, x_2))$$

where $\phi : X \mapsto \mathcal{H}$.



Note that $X = \mathbb{R}^2$ is the domain, and \mathcal{H} is the Hilbert space, which is (in machine learning literature) the feature space, and a set of features $\phi_i, \forall i$, is called dictionary.

Mantra

A major theme in machine learning is that sometimes things actually get easier in higher dimensions !!!.

- A linear plane in high dimensional feature space \mathcal{H} , may be a nonlinear curves in the domain space.
- \mathcal{H} is a plane, with calculus with dot products is legit.

The following, we introduce Mercer's theorem, which generalizes spectral decomposition theorem.

Theorem 5.5.1 — Mercer's Theorem

Let $K \in L^2(X \times X)$, (i.e. $\int |K(x, \tilde{x})|^2 dx d\tilde{x} < \infty$) such that $T : L_2(X) \mapsto L_2(X)$ by $(T(f))(x) = \int K(x, \tilde{x}) f(\tilde{x}) d\tilde{x}$ is positive definite. If $\phi_i \in L^2(X)$ is

a normalized eigenfunction with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$, Then

$$K(x, \tilde{x}) = \sum_{i=1}^{N_{\mathcal{H}}} \lambda_i \phi_i(x) \phi_i(\tilde{x}) \quad (5.20)$$

for almost every (x, \tilde{x}) . Where $N_{\mathcal{H}} = \dim(\mathcal{H})$, and the convergence of $K(x, \tilde{x})$ is absolute.

Mercer's theorem itself is a generalization of the result that any symmetric positive-semidefinite matrix is the Gramian matrix of a set of vectors.

ϕ_i 's exist & I can use them in KSM.

$$K: \overline{X \times X} \rightarrow \mathbb{R}$$

generalizes spectral decomposition theorem.

usual
 $Ar = U$
 $Aw = \lambda w$

eigen fn.

$$T(\phi_i)(x) = \lambda_i \phi_i(x)$$

$$D = \{ (x_i, y_i) \}$$

