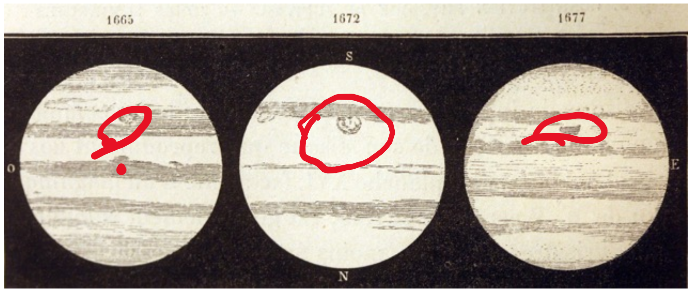
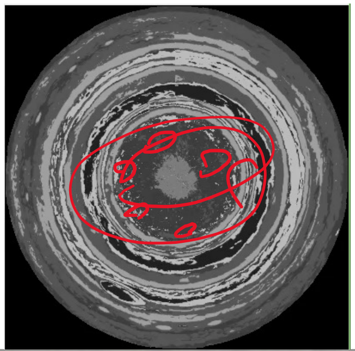
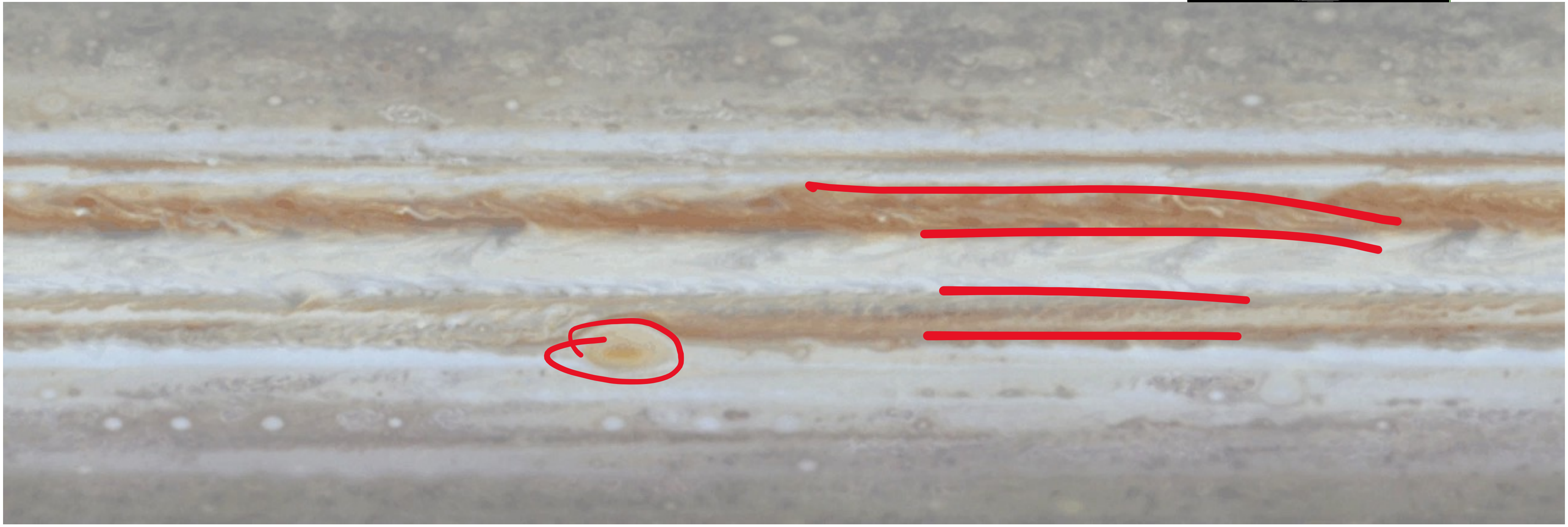
EE520 Data Driven Analysis of Complex Systems

Erik Bollt







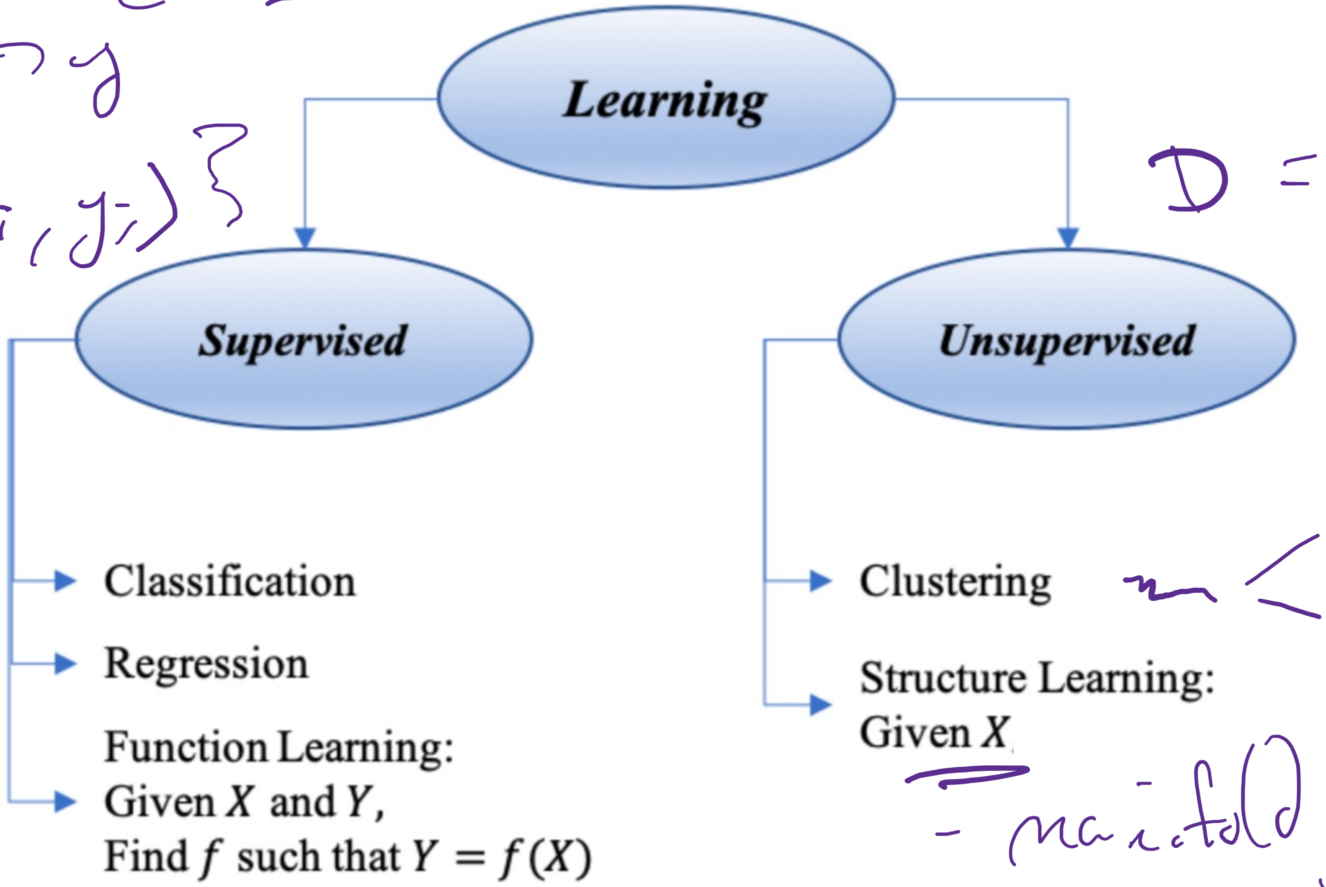
Ch 5 - Clustering and Classification

attributes & labels .

$f : x \to y$

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

$D = \{x_i\}$



Learning

Supervised

- Classification
- Regression
- Function Learning: Given X and Y, Find f such that $Y = f(X)$

Unsupervised

- Clustering
- Structure Learning: Given X

Kmeans

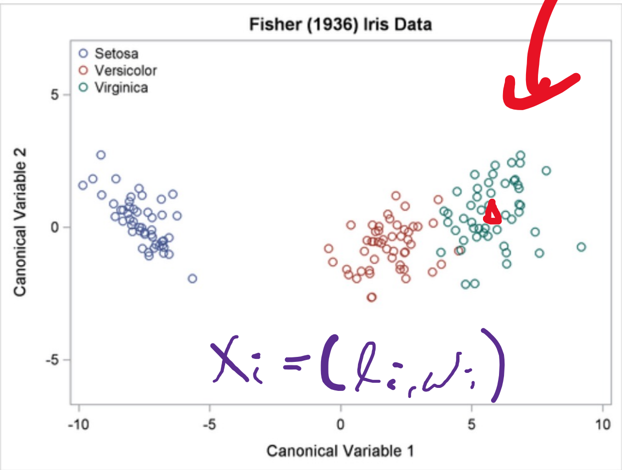
spectral

- manifold
- tree-hierarchy.

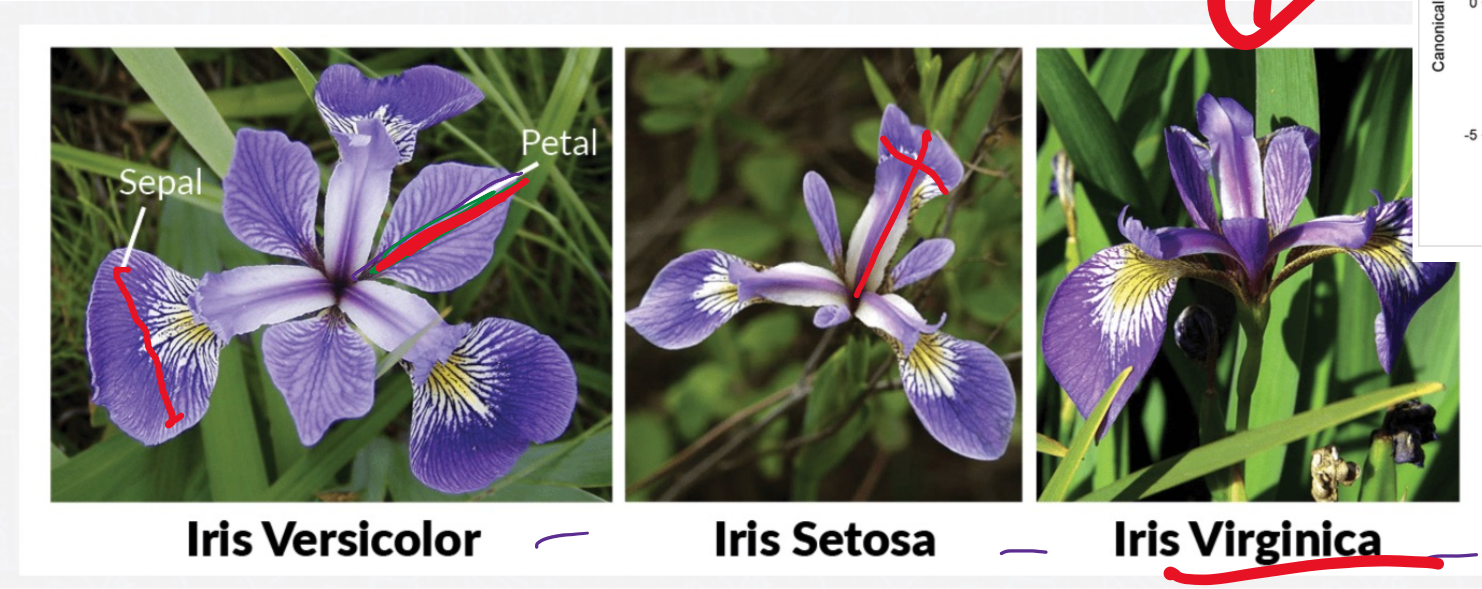
Discriminating Fisher's iris data by using the petal areas

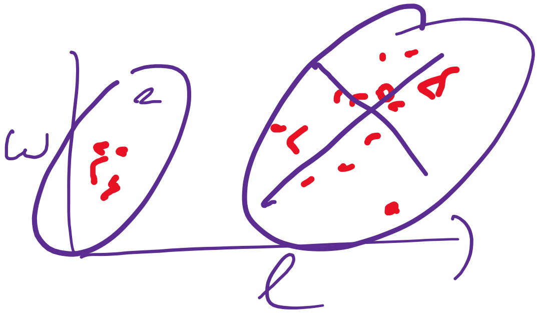
The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification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

Sp

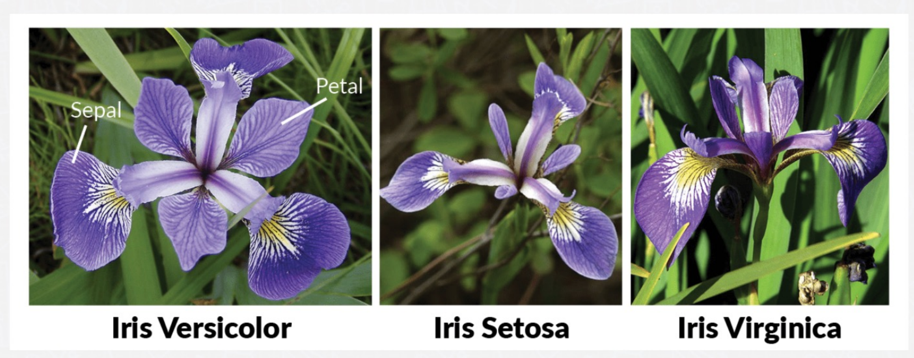


$X_i = (l_i, w_i)$





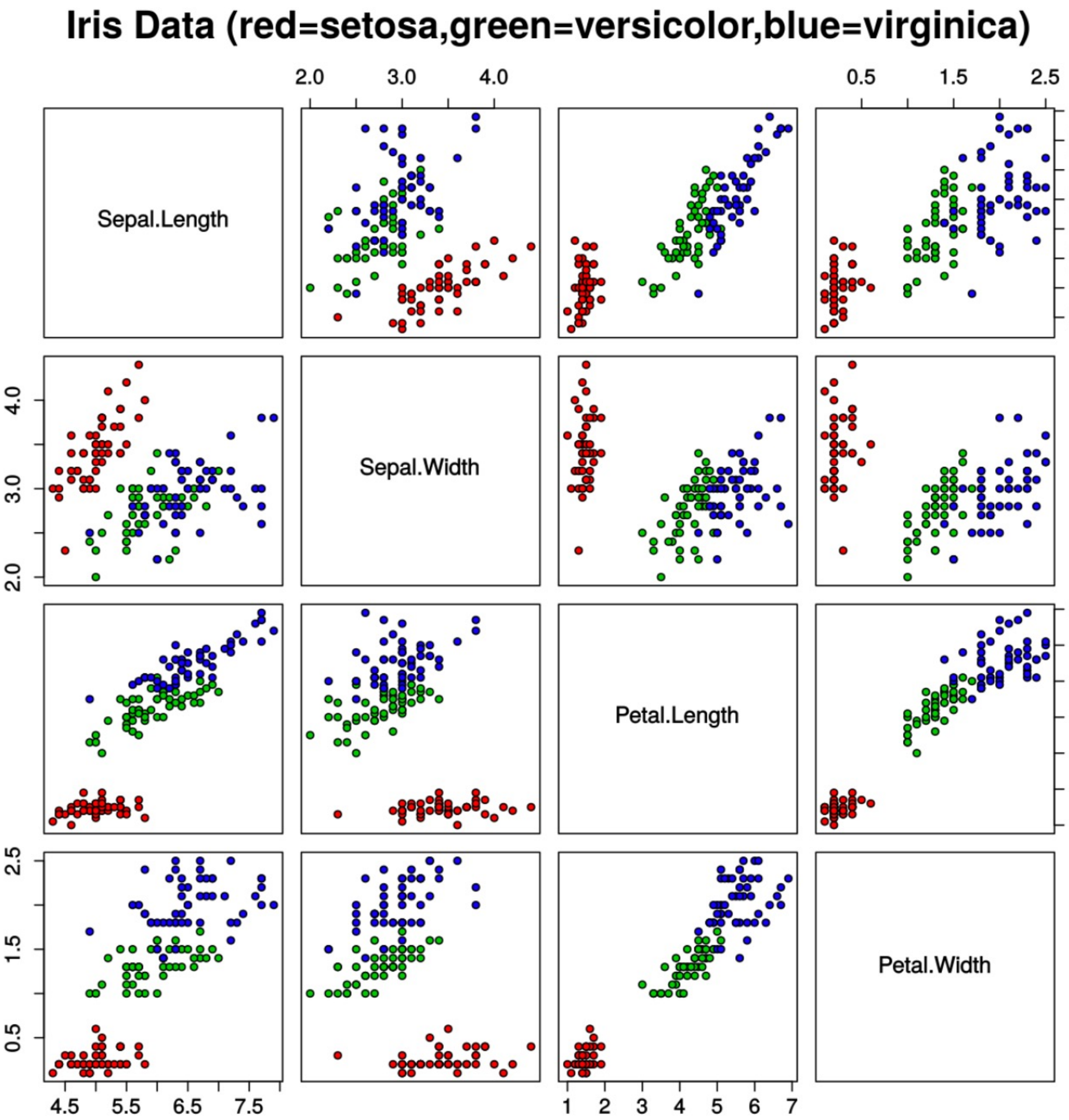
Ronald Fischer 1936



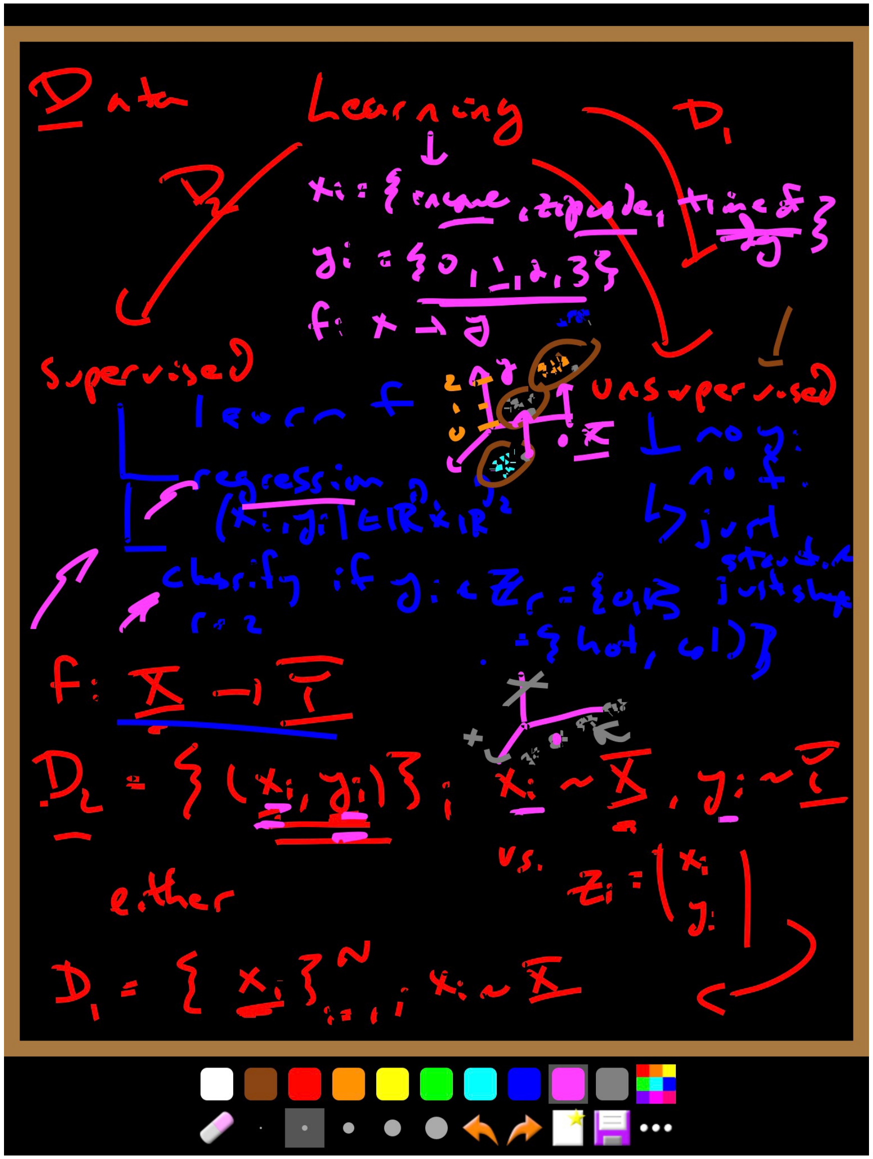
Iris Versicolor

Iris Setosa

Iris Virginica



Iris Data (red=setosa,green=versicolor,blue=virginica)



Data Learning

D_2 D_1

$x_i = \{\text{name}, \text{zipcode}, \text{time of day}\}$

$y_i = \{0, 1, 2, 3\}$

$f: x \to y$

supervised

learn f

regression $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

classify if $y_i \in \mathbb{Z}_r = \{0,1\}$ $r=2$

$= \{\text{hot}, \text{cold}\}$

unsupervised

no y_i

no f

just structure

just shape

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

$D_2 = \{(x_i, y_i)\}_i$; $x_i \sim \underline{X}$, $y_i \sim \underline{Y}$

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

either

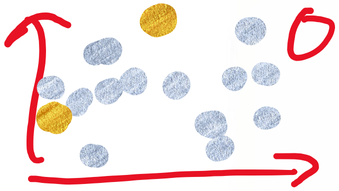
$D_1 = \{x_i\}_{i=1}^{N}$, $x_i \sim \underline{X}$

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

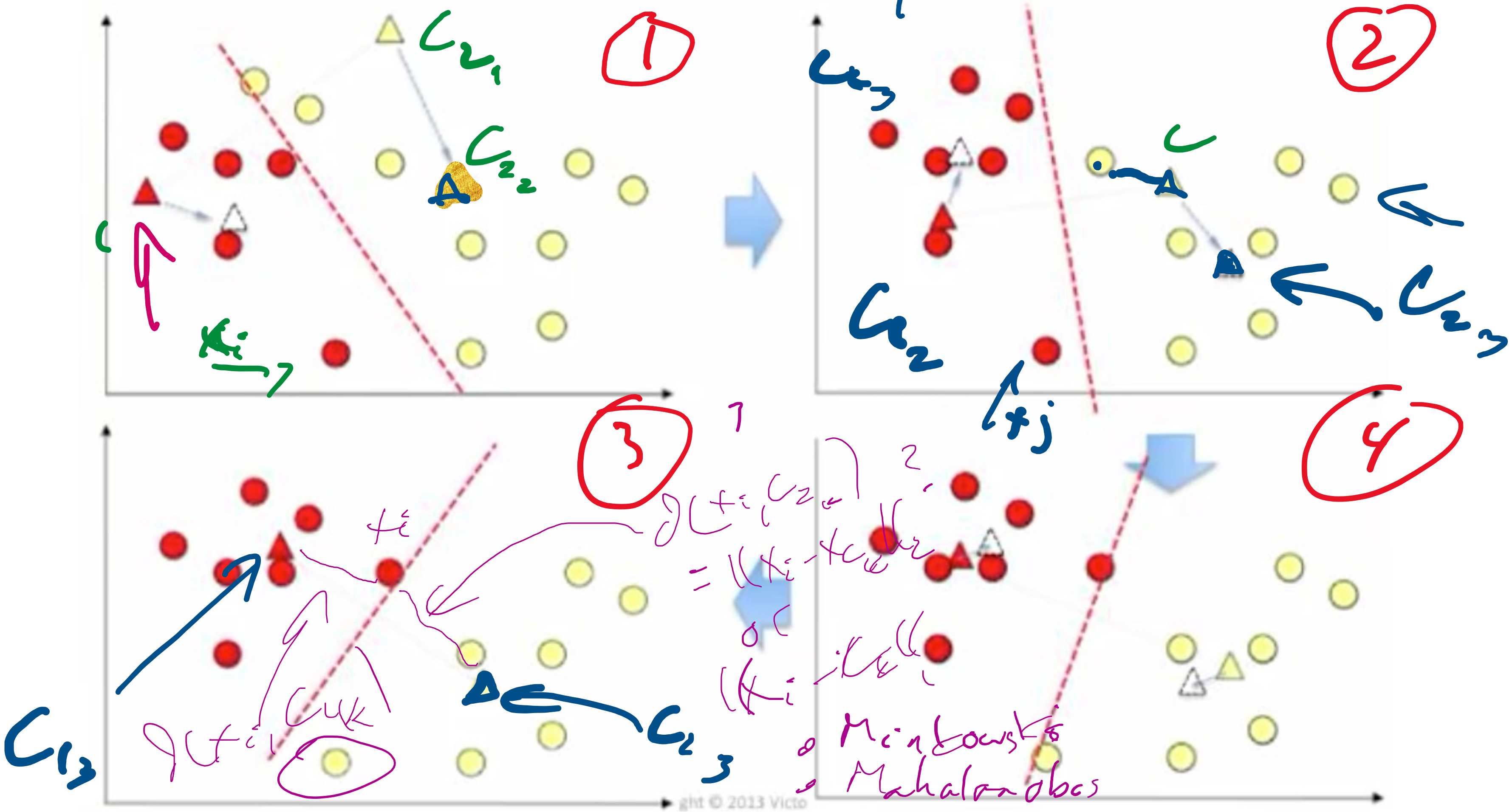
-Supervised discrete output (labels)
-supervised continuous output

-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

$x_i = (x_{i_1}, x_{i_2}, \dots)$

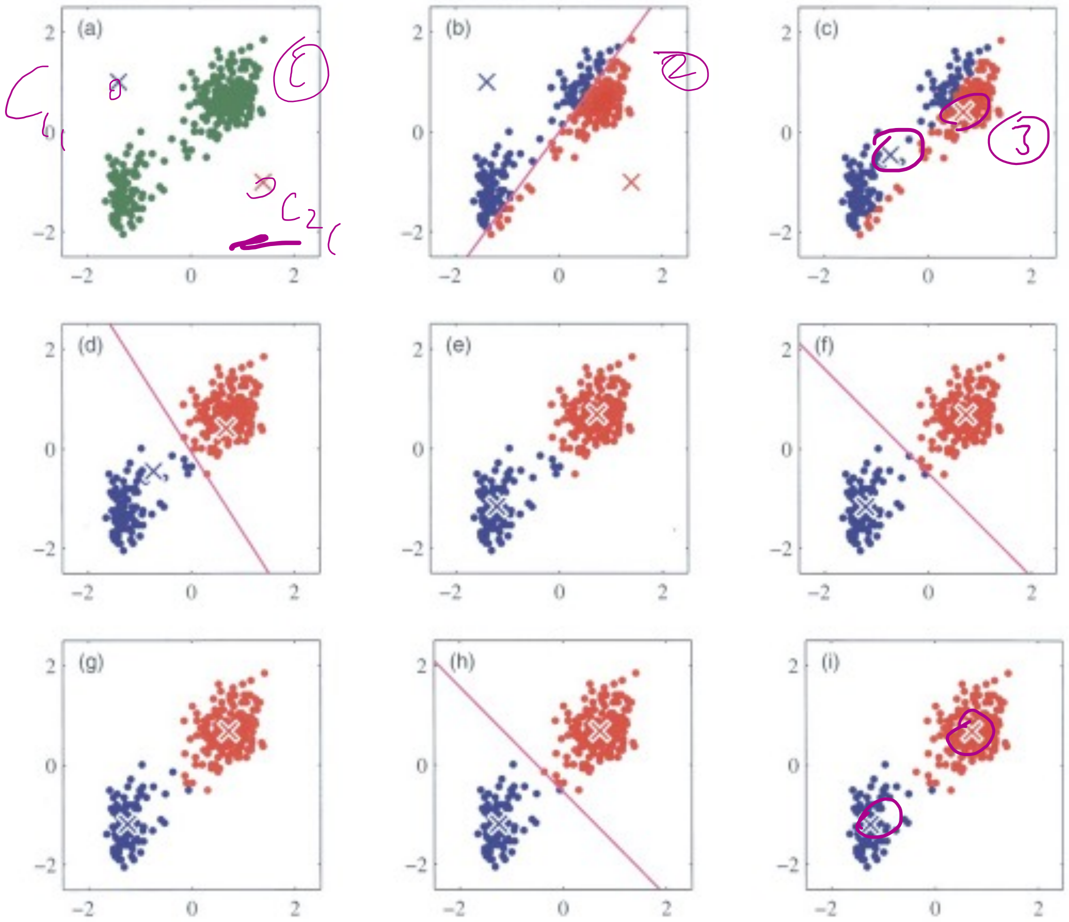


Illustration of K-means algorithm from [Bishop 2006]

Kmeans Convergence

$C_i \equiv \mu_i$

$C_2 \equiv \mu_2$
μ_1
C_2
C_1

Objective

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

C

1. Fix μ, optimize C:

Step 1 of kmeans

$$\min_{C}\sum_{i=1}^{k}\sum_{x\in C_i}|x-\mu_i|^2 = \min_{c}\sum_{i}^{n}|x_i-\mu_{x_i}|^2$$

2. Fix C, optimize μ:

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

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

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

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

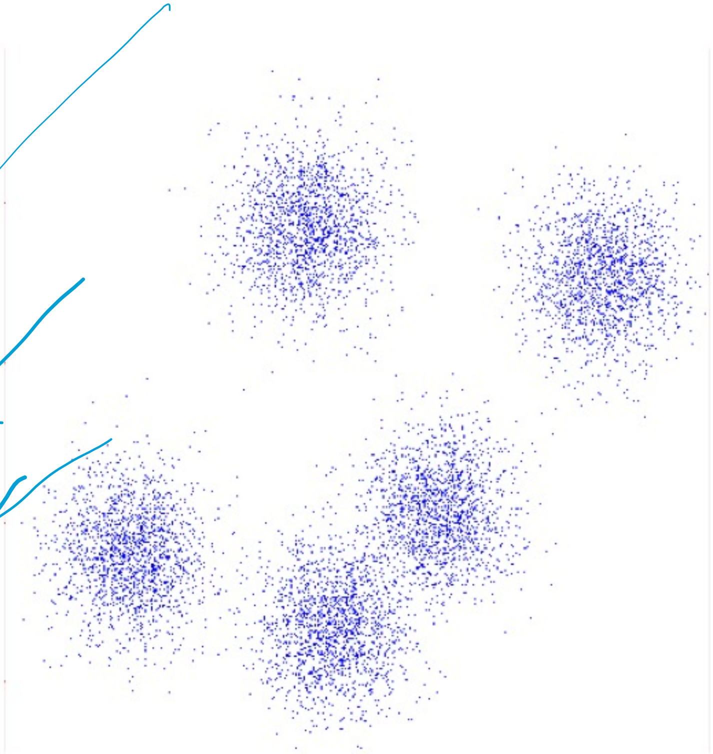
Step 2 of kmeans

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

[Slide from Alan Fern]

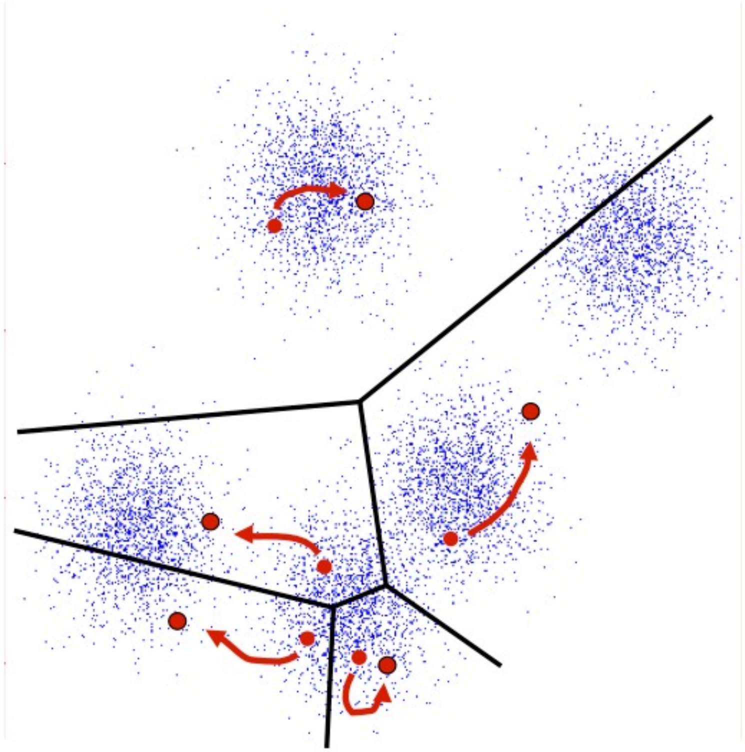
K-Means

- 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

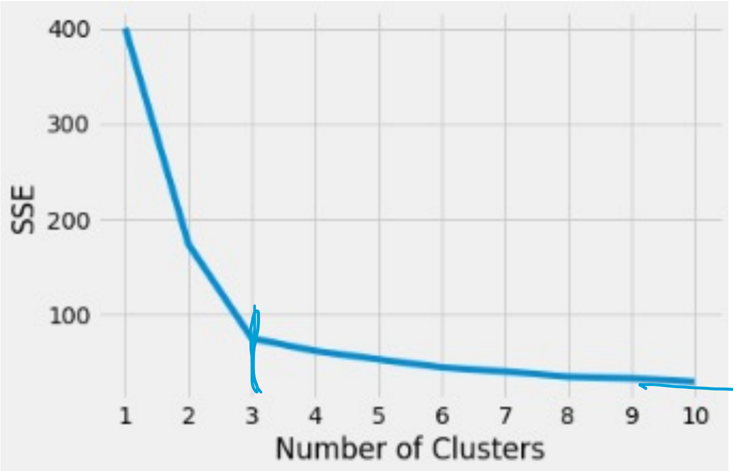


K-Means

- 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



K



elbow

$X = (x_1, x_2)$

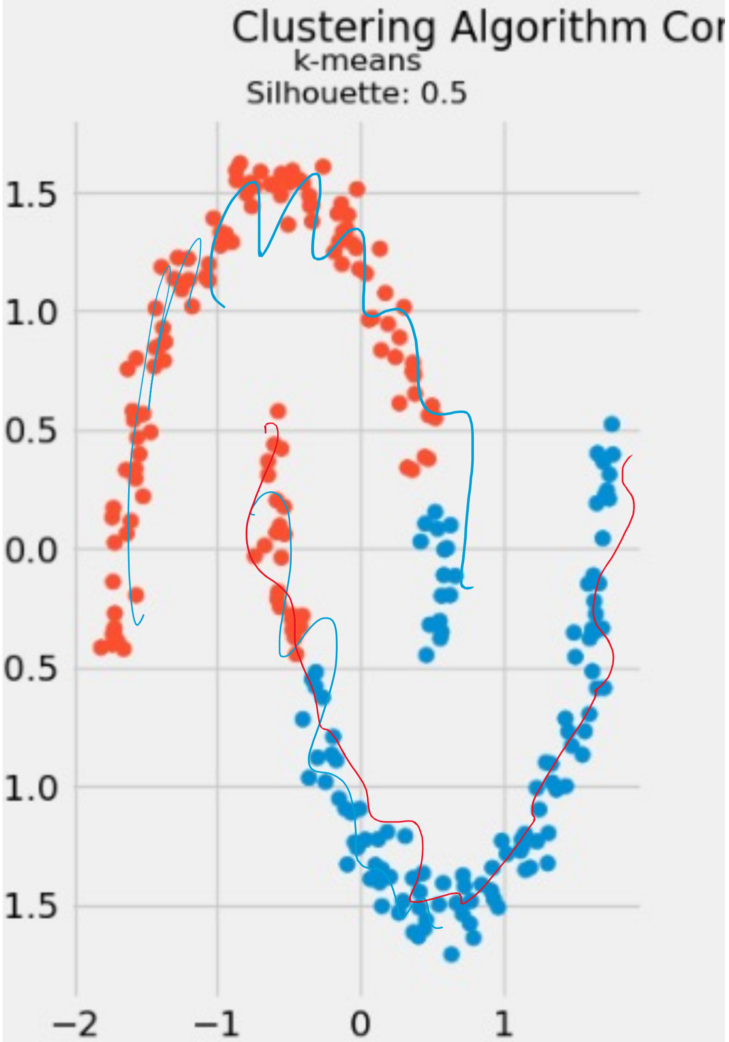
$C = (C_1, C_2)$

$$d(X, C)^2$$

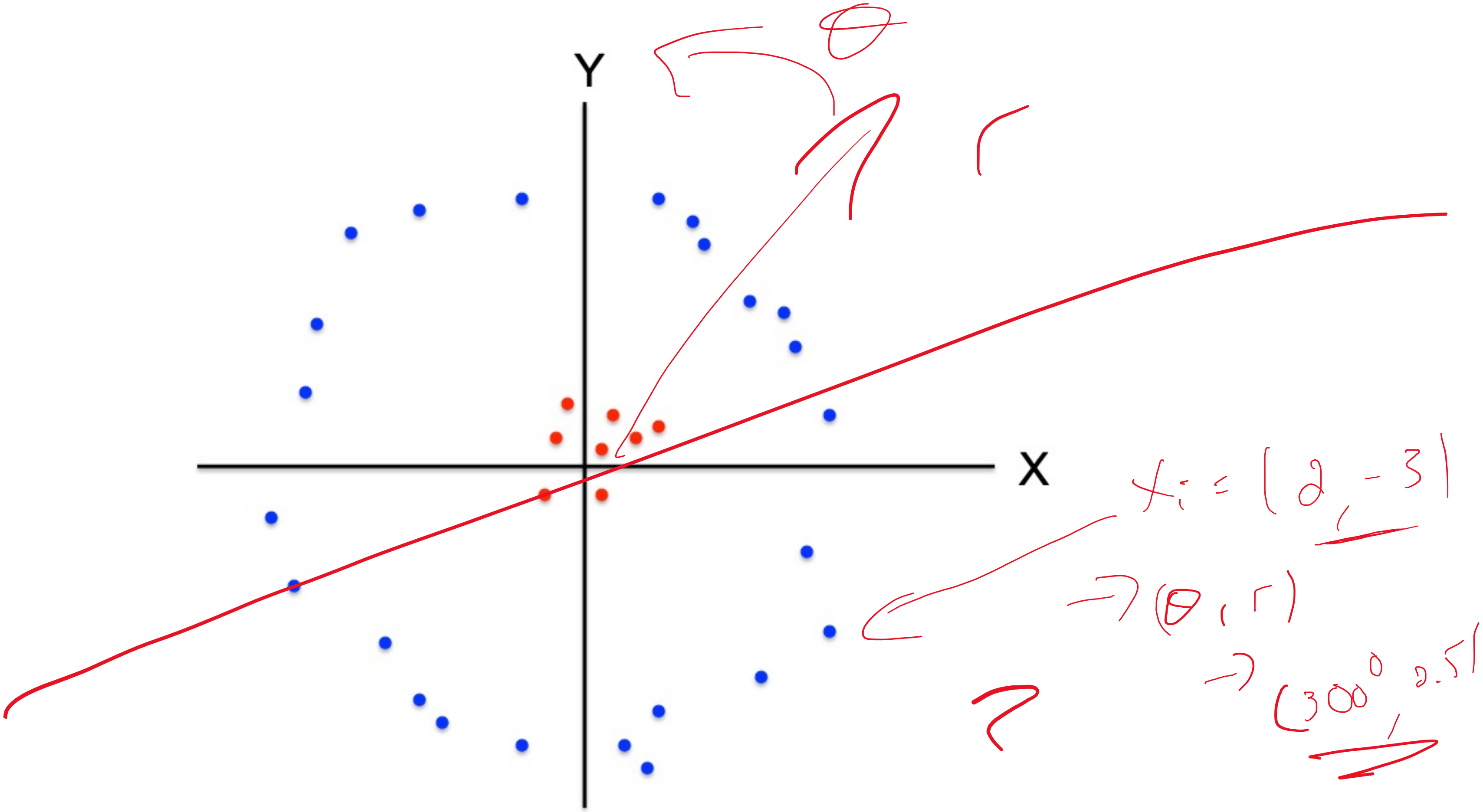
$$= \cancel{\sum_i (X_i - C_i)^2}$$

$$= \cancel{\sum (x_i - c_i)^4}$$

$$\sum e^{t(x_i - c_i)} \quad ?$$



K-means not able to properly cluster

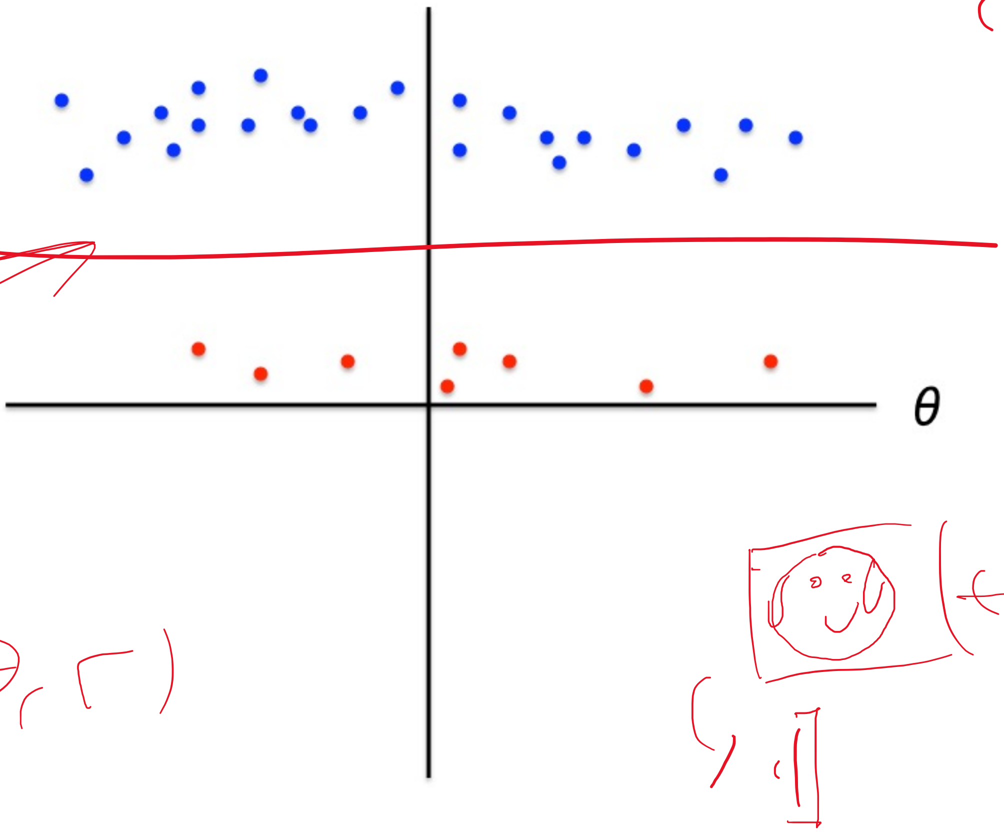


$x_i = (2, -3)$

$\to (\theta, r)$

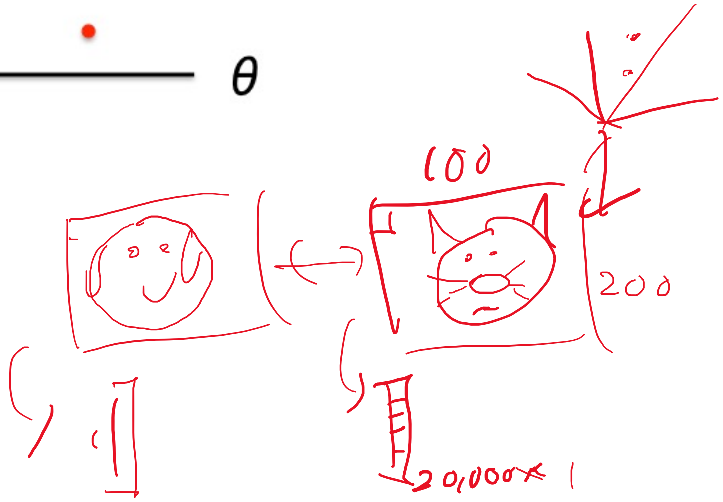
$\to (300^0, 0.5)$

Changing the features (distance function) can help

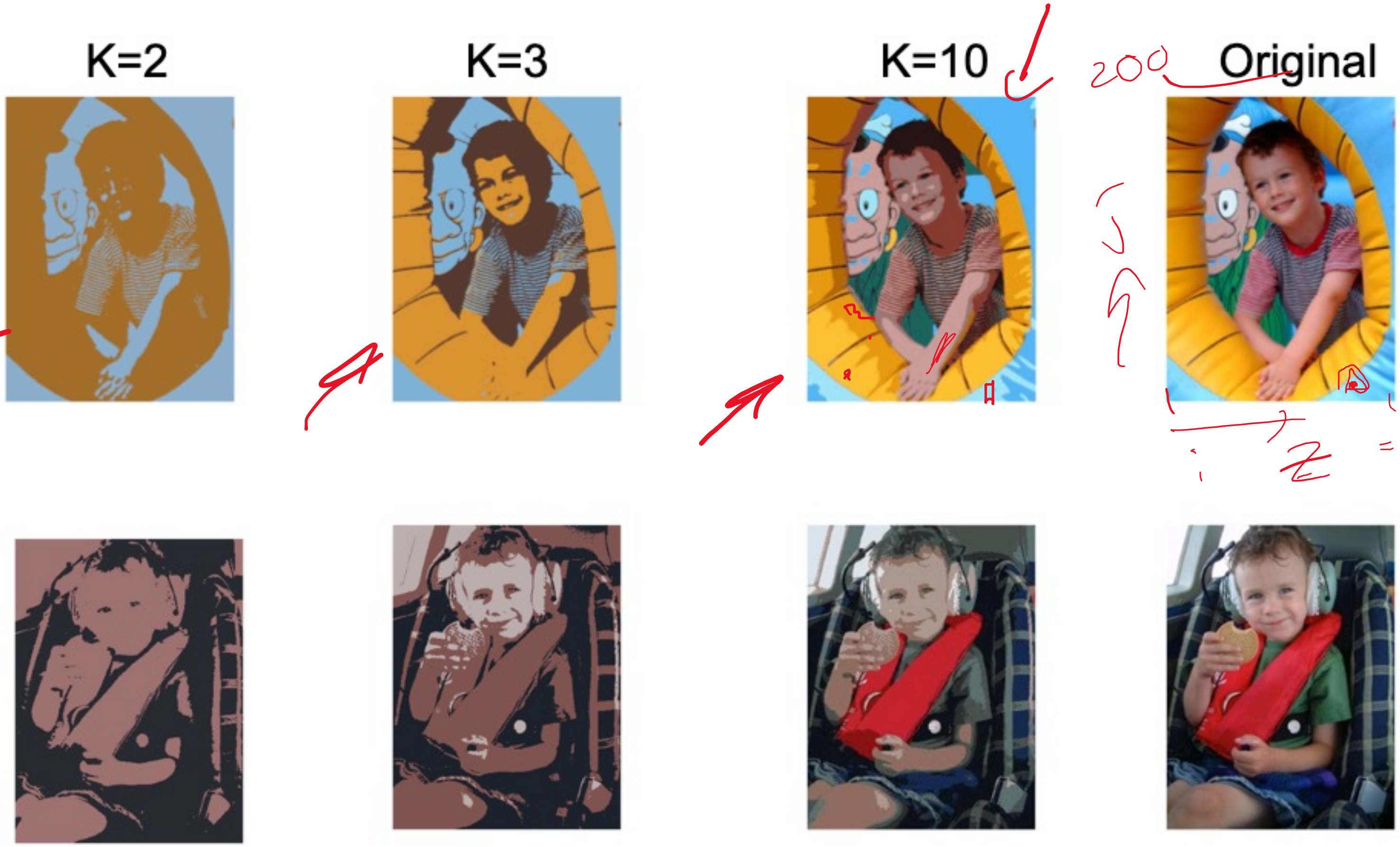


radial as features

$(x, y) \to (\theta, r)$



Example: K-Means for Segmentation



K=2

K=3

K=10

Original

200

j

i

z

rgb

0-255

0-255

0-255

20,000

$\{z_k\}_{k=1}^{20,000}$

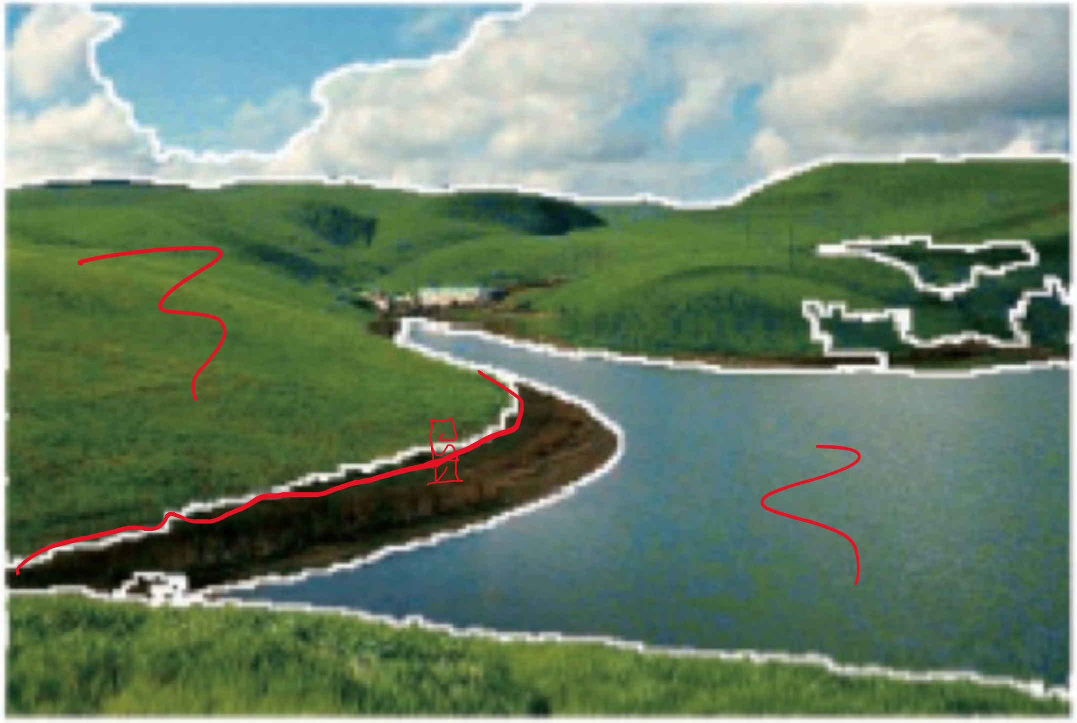
100

$z = (i, j, r, g, b)$

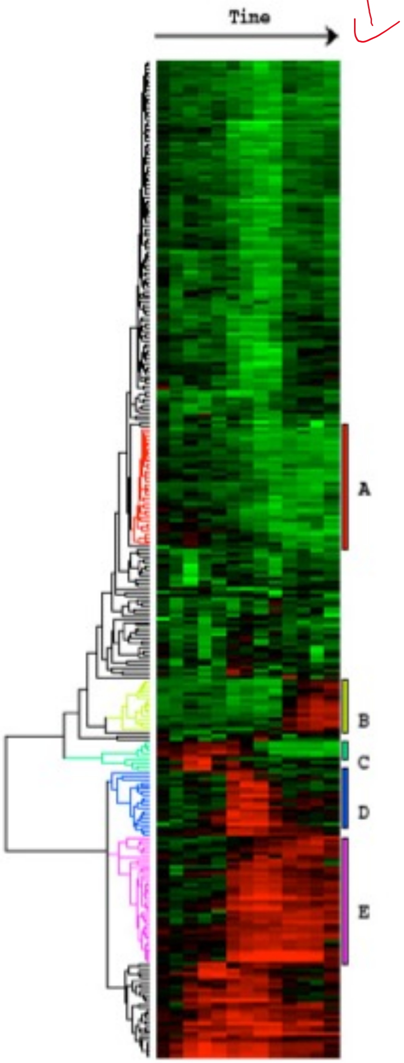
75,8

Image segmentation

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

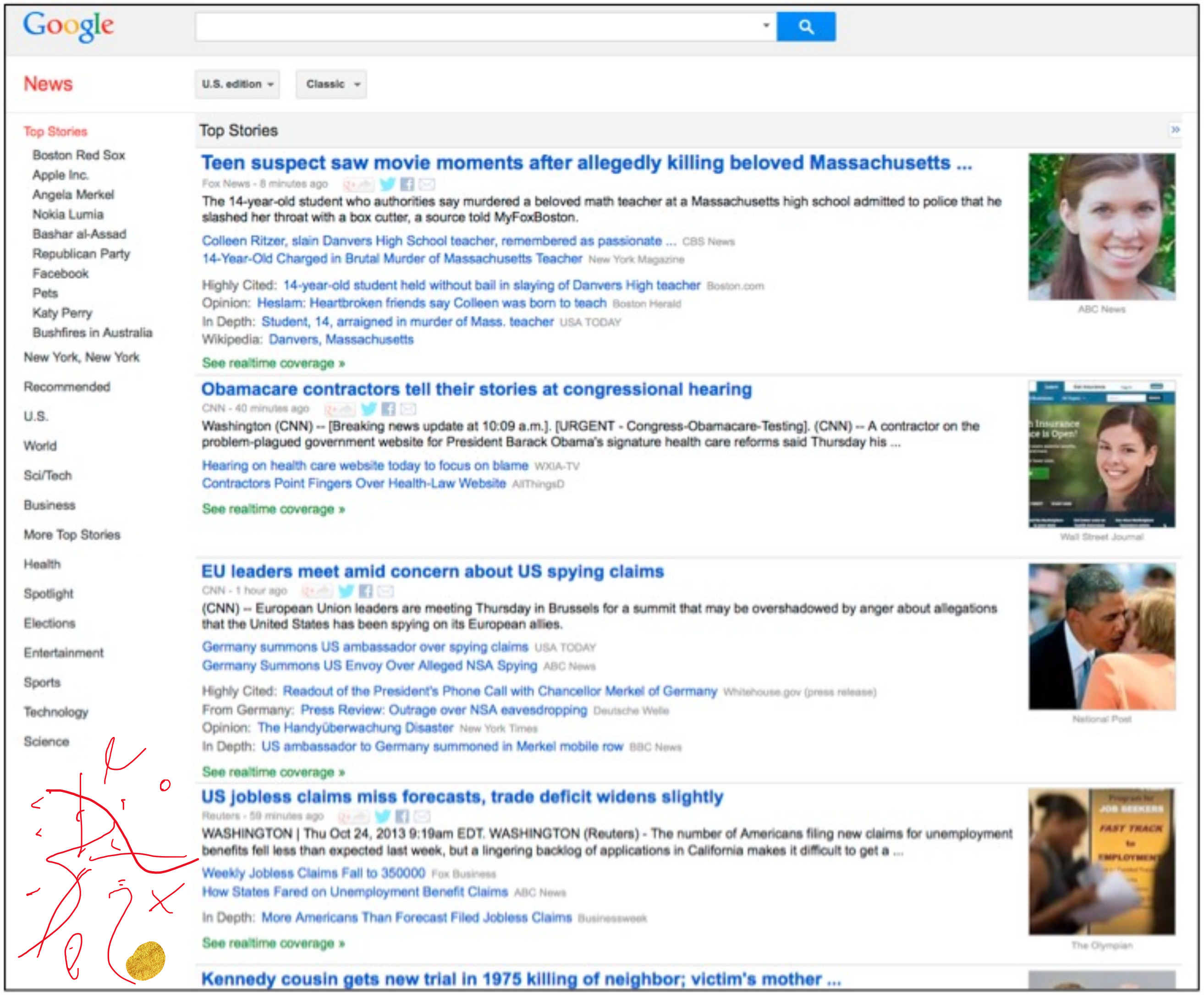


Clustering gene expression data



Eisen et al, PNAS 1998

Cluster news articles



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 »

ABC News

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 WXIA-TV
Contractors Point Fingers Over Health-Law Website AllThingsD

See realtime coverage »

Wall Street Journal

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 »

National Post

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 350000 Fox Business
How States Fared on Unemployment Benefit Claims ABC News

In Depth: More Americans Than Forecast Filed Jobless Claims Businessweek

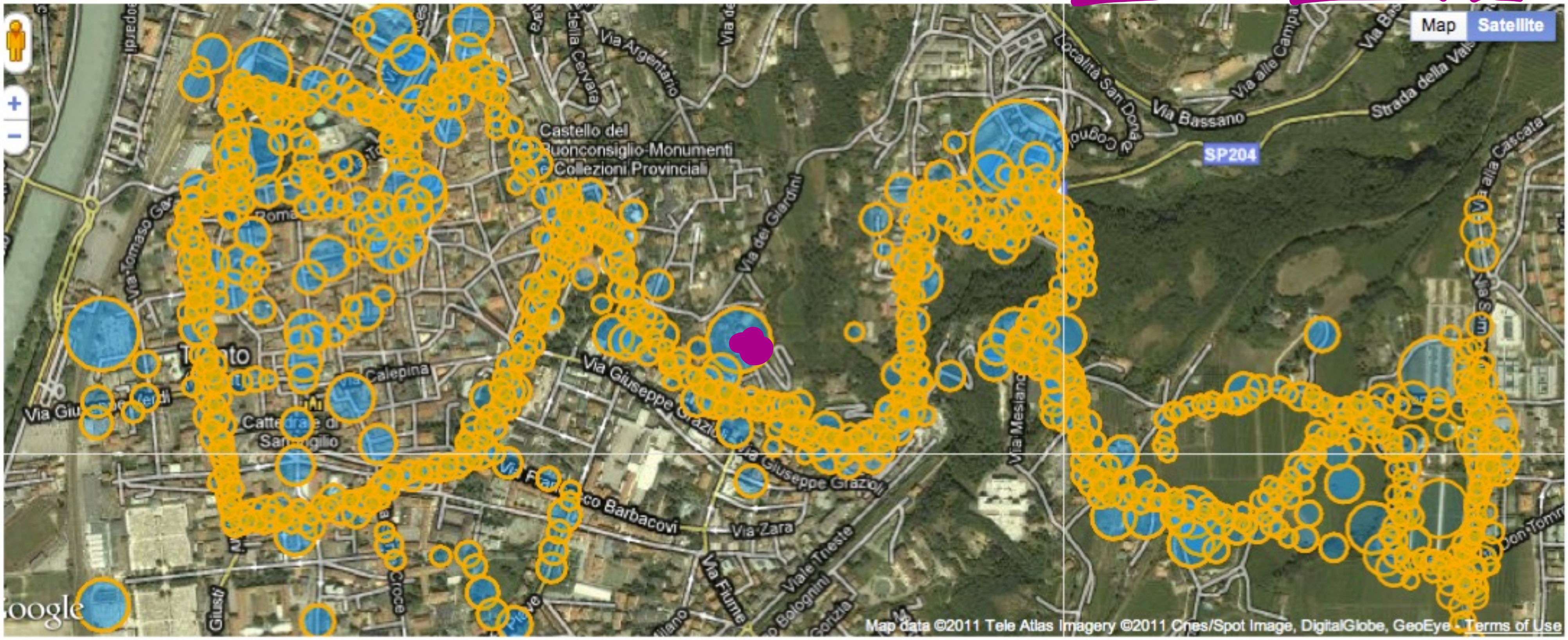
See realtime coverage »

The Olympian

Kennedy cousin gets new trial in 1975 killing of neighbor; victim's mother ...

Cluster people by space and time

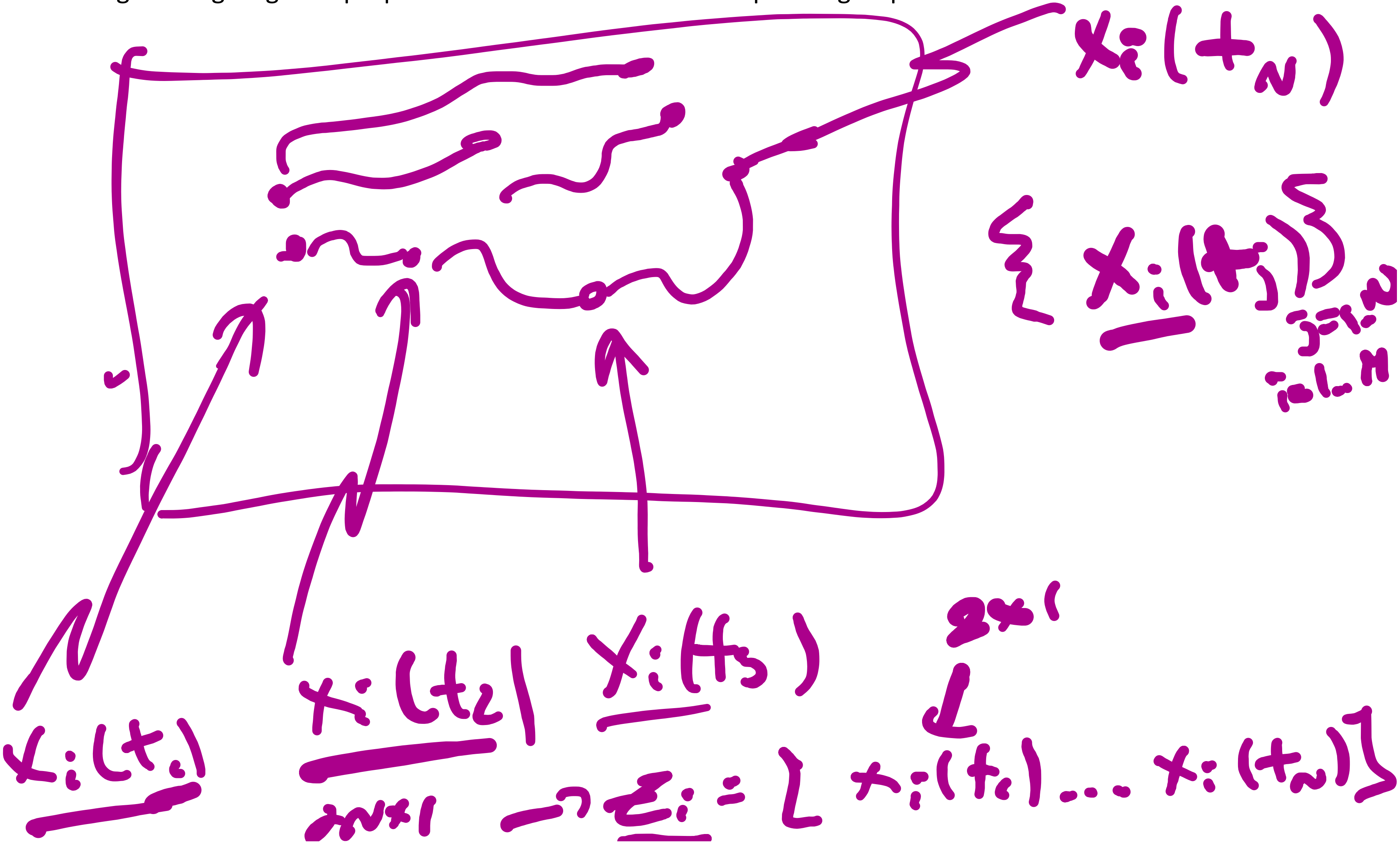
$z^i = (x_1^i, y_1^i, t_1^i), (x_2^i, y_2^i, t_2^i), (x_3^i, y_3^i, t_3^i)$


Map
Satellite
Castello del Buonconsiglio-Monumenti e Collezioni Provinciali
Via Bassano
SP204
Via Giuseppe Grazioli
Via Calepina
Cattedrale di San Vigilio
Via Francesco Barbacovi
Via Zara
Via dei Giardini
Google
Map data ©2011 Tele Atlas Imagery ©2011 Cnes/Spot Image, DigitalGlobe, GeoEye - Terms of Use

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

[Image from Pilho Kim]

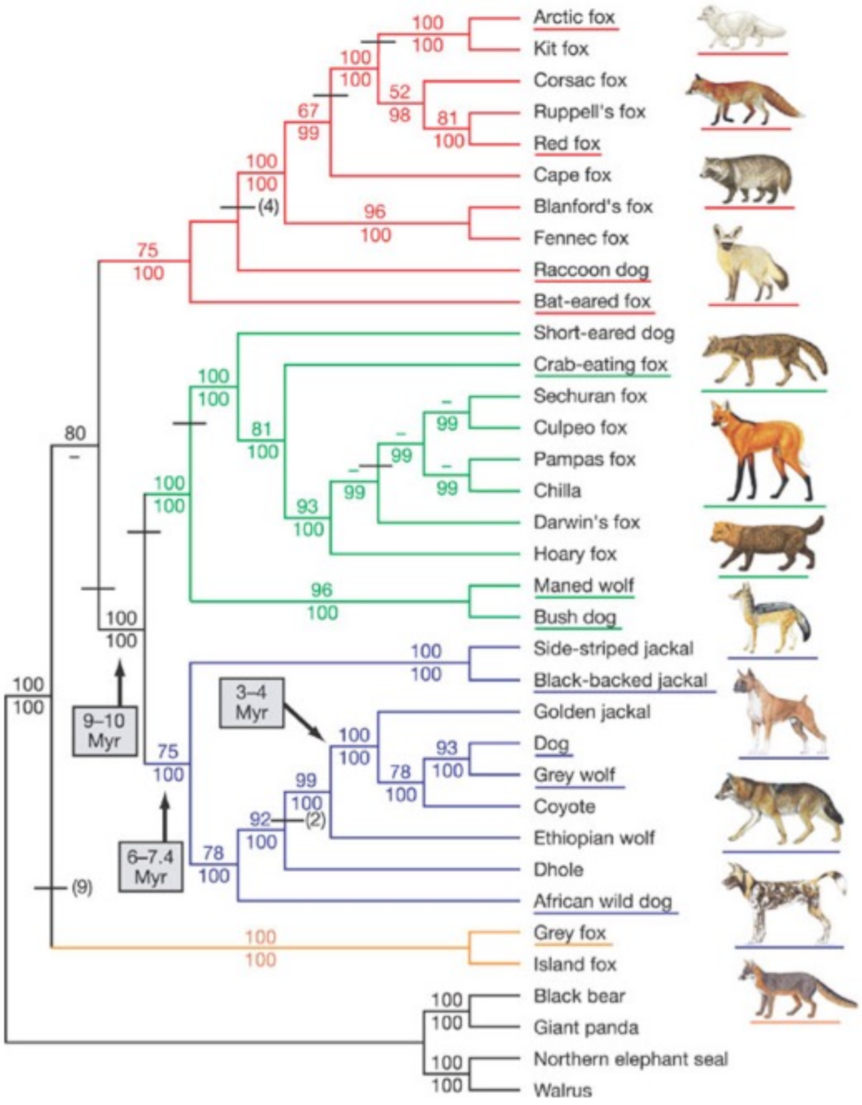
Clustering on tracking moving targets – prepare a vector – feature - corresponding to position over time.



$\{x_i(t_j)\}$ $j=1..N$ $i=1..M$

$\rightarrow z_i = [x_i(t_1) \ldots x_i(t_N)]$

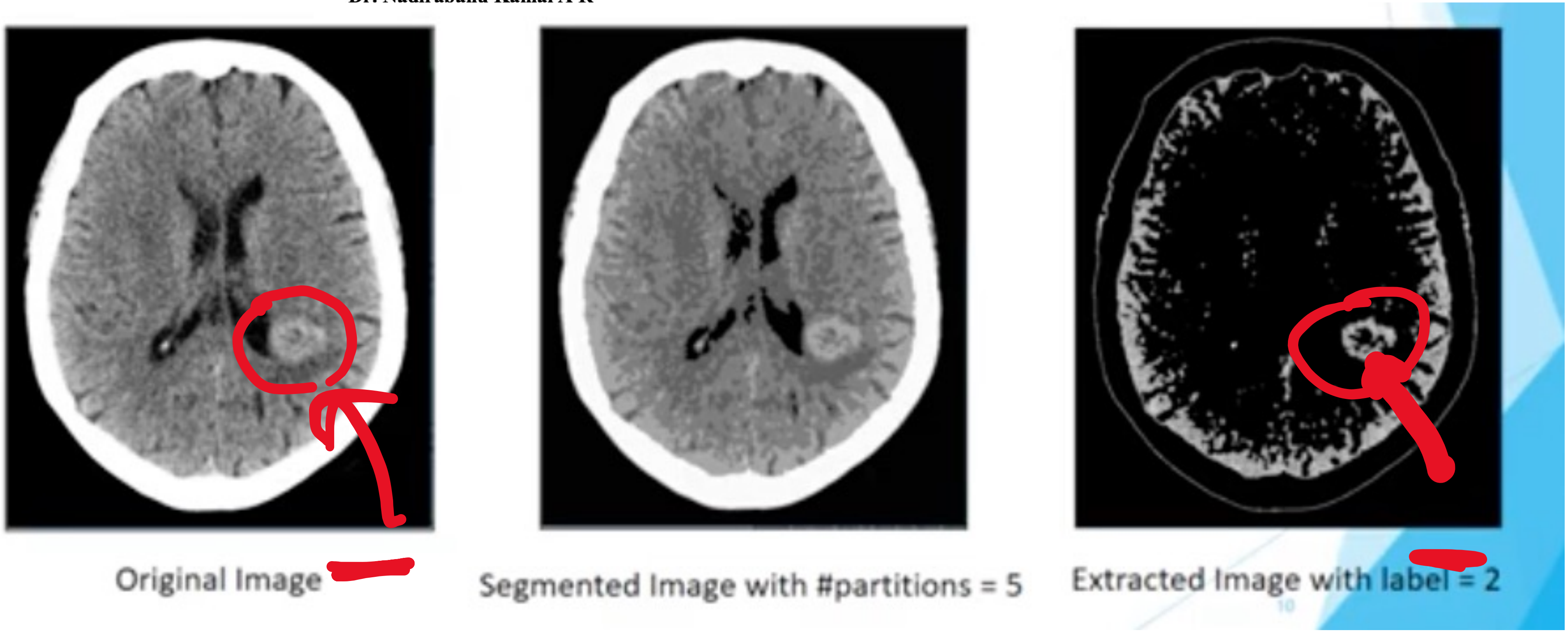
Clustering species ("phylogeny")



[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

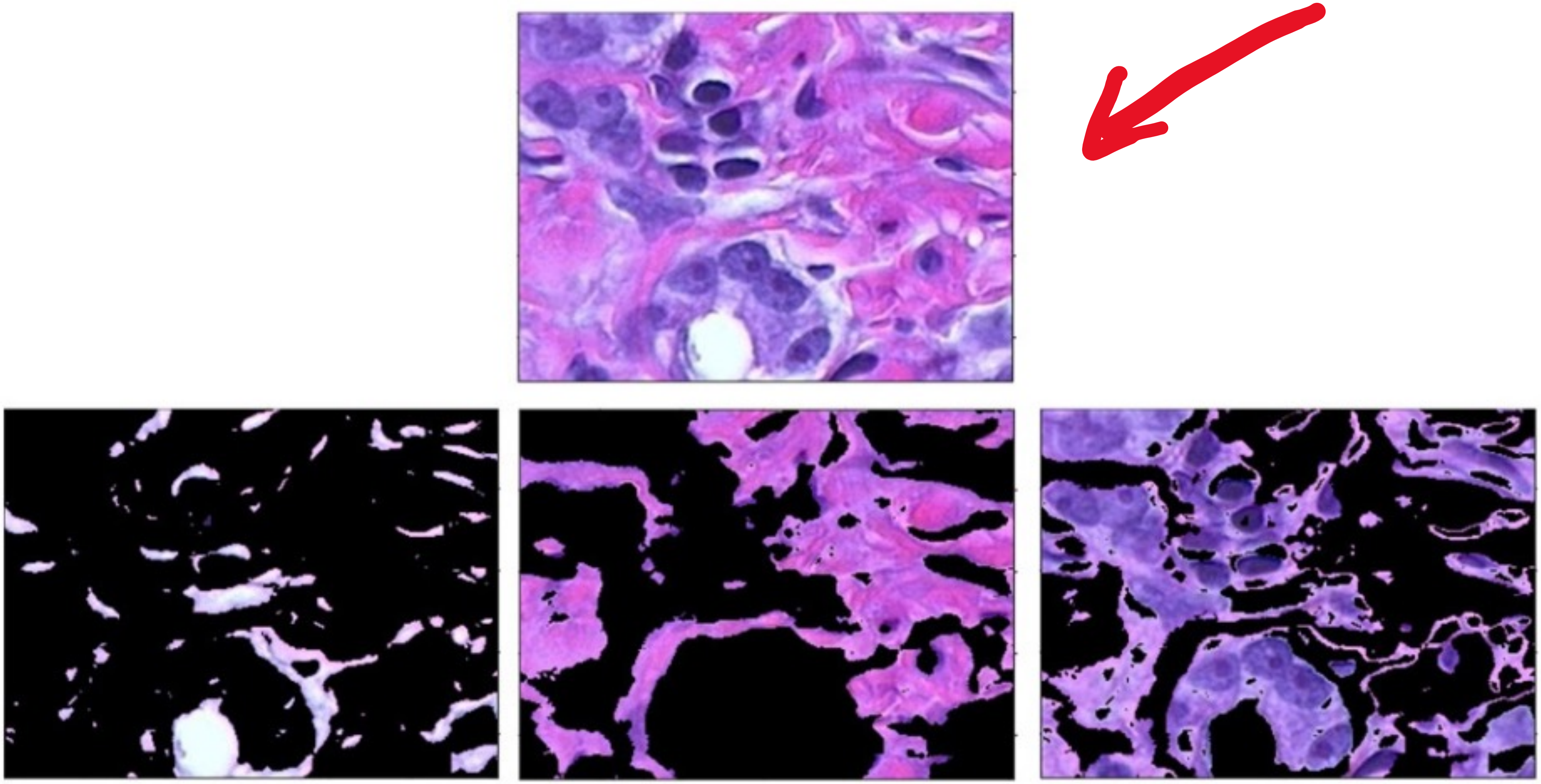


Original Image

Segmented Image with #partitions = 5

Extracted Image with label = 2

Medical imaging – cancer tissue



IJEEE, Vol. 1, Issue 3 (June, 2014) e-ISSN: 1694-2310 | p-ISSN: 1694-2426

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

[1]Neha Sharma, [2]Nishant Kinra
[1]Indira Gandhi Delhi Technical University for Women, New Delhi, India
[2]Deenbandhu Chhotu Ram University of Science and Technology, Haryana, India
[1]neha.sksharma@yahoo.co.in, [2]nishant.kinra@gmail.com

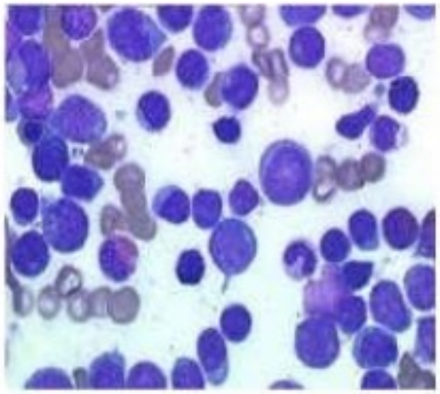


Fig (1): Example of Leukocytosis

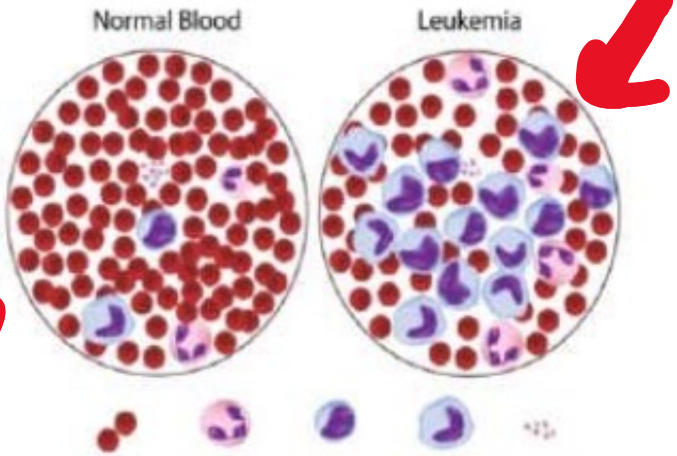


Fig (2): Comparison between Normal Blood and Leukemia

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)

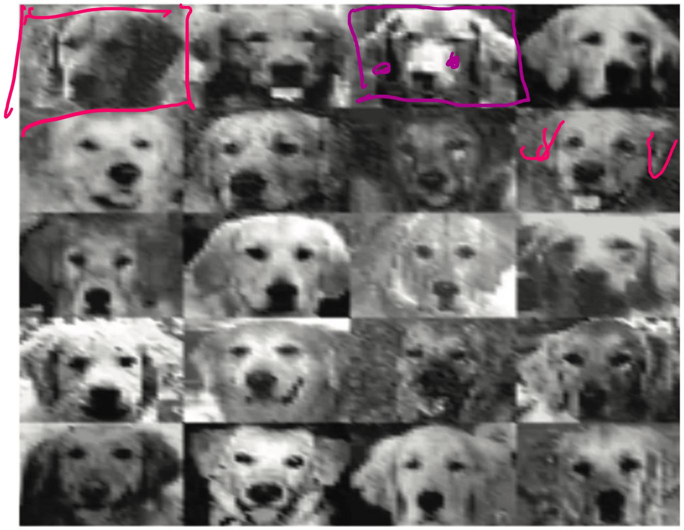
Then Chapter 6 a grandly popular method – artificial neural nets.

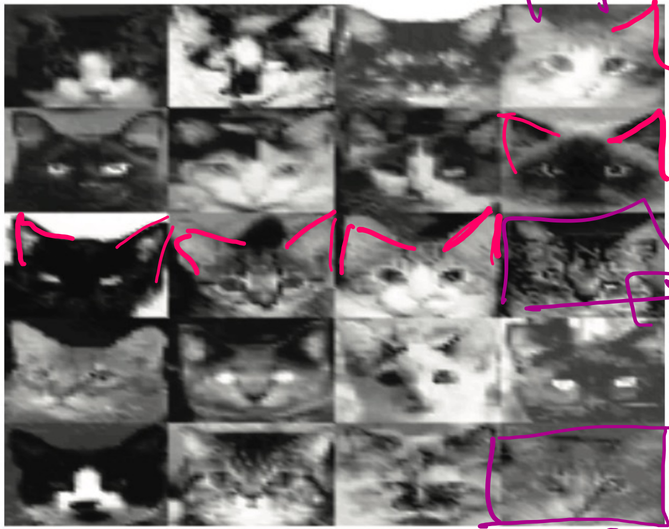
Supervised Learning – Cat or Dog?

y_i = Dog

x_i = ith picture q p

$[x_i]_{qp\times 1}$ =





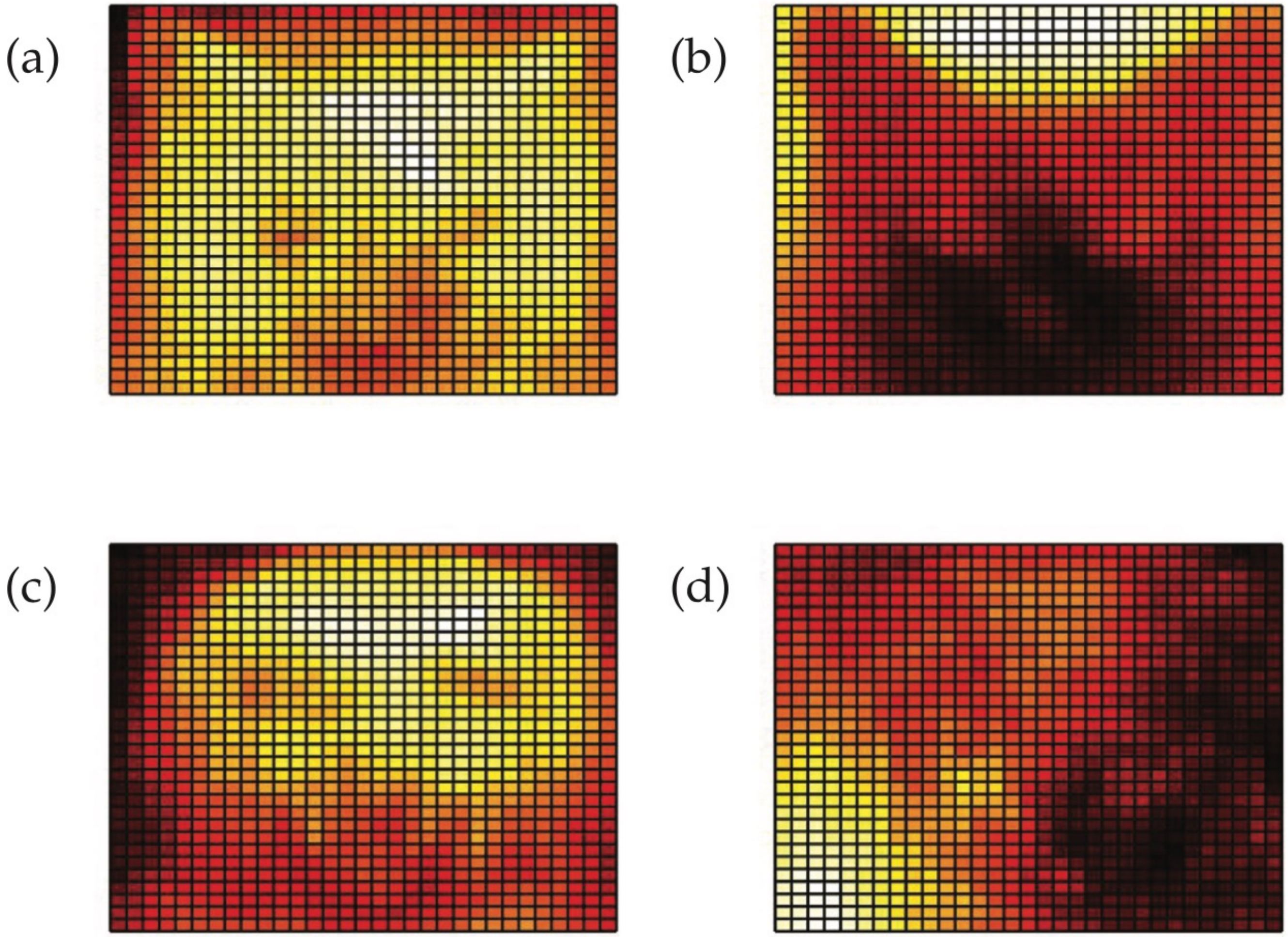
$D = \{(x_i, y_i)\}_{i=1}^N$

x_i = ith picture

y_i = ith label

$f: X \to Y$; $f: \mathbb{R}^{20,000} \to \mathbb{Z}_2$

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.

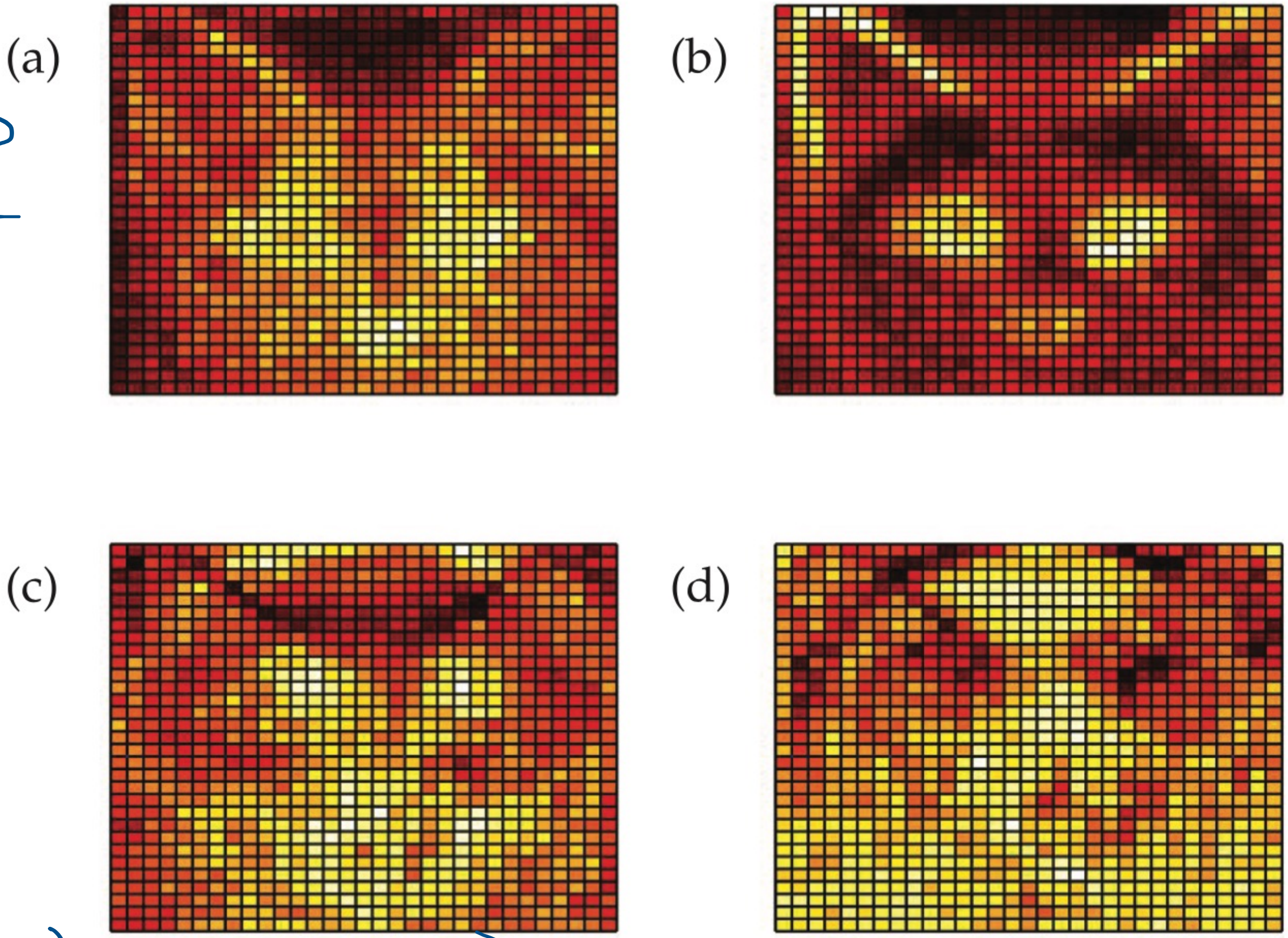
unsupervised.

$D = \{(x_i)\}$

structure

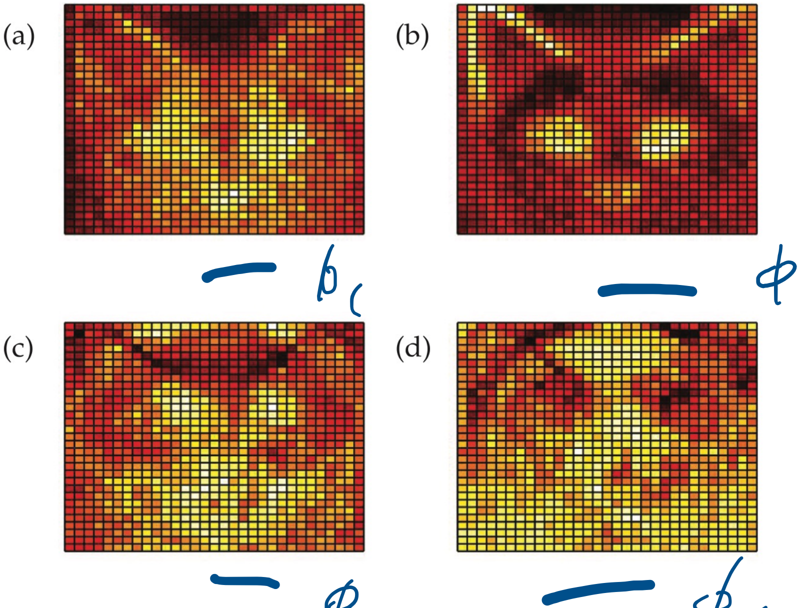
cluster

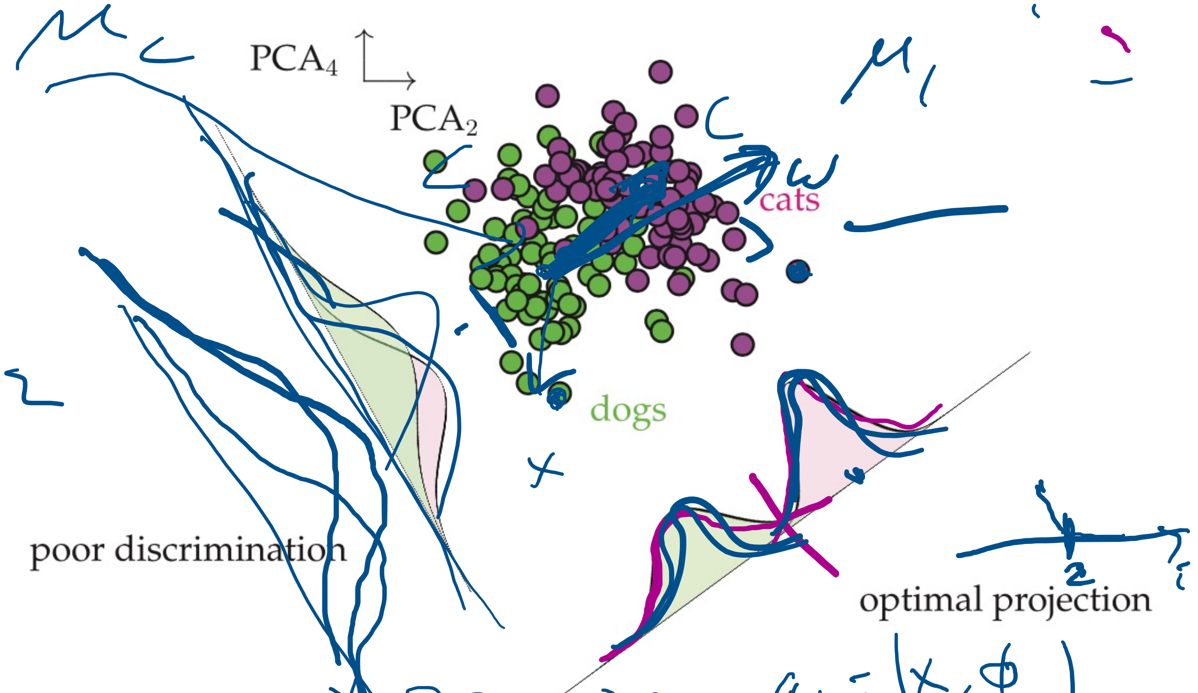
ROM



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

Fischer’s Linear Discriminant Analysis LDA





$x \to a \to y$ $\quad a_1 = \langle x, \phi_1 \rangle$

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

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

$$= \int_\Omega x(s)\phi_1(s)\,ds$$

$$\mathbf{S}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T = \underbrace{(\mu_2 - \mu_1)}_{2\times 1}\underbrace{(\mu_2 - \mu_1)^T}_{1\times 2}$$

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

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

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

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

Haar Wavelet

Define

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

Mother function.

Mathworld ~ mathematica

and

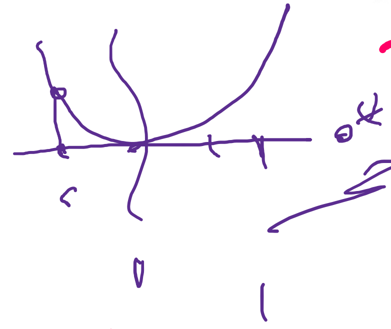
$$\psi_{jk}(x) \equiv \psi(2^j x - k)$$

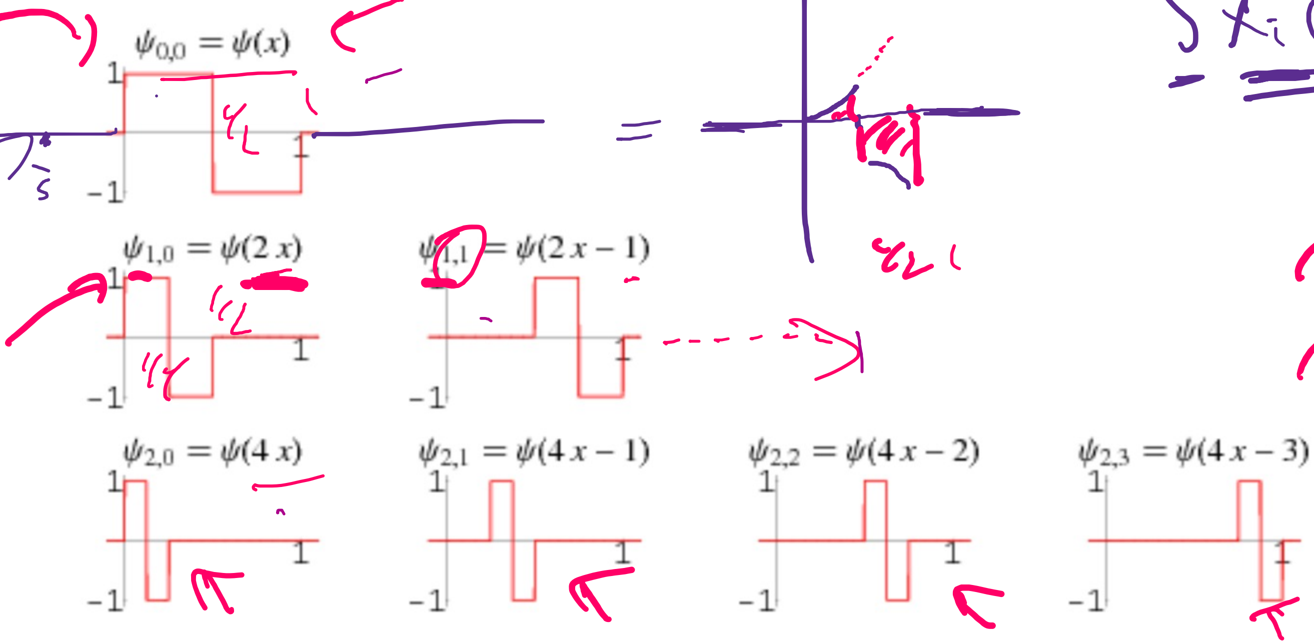
dyadic

b_k (x_i, ψ_k)

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

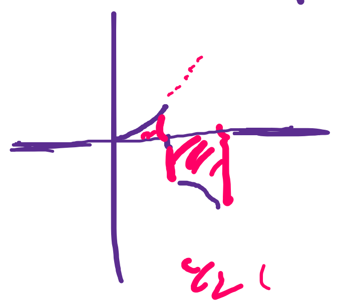
$x_i(s) = s^2$





=

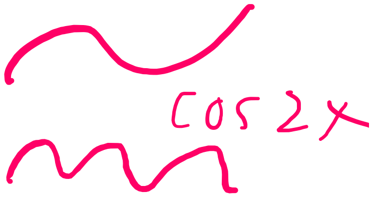
$x_i(s)\psi(s)$



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

cos x

cos 2x



$\phi_1(s) = \psi_{00}$

$\phi_2(s) = \psi_{10}$

$\phi_3(s) = \psi_{11}$

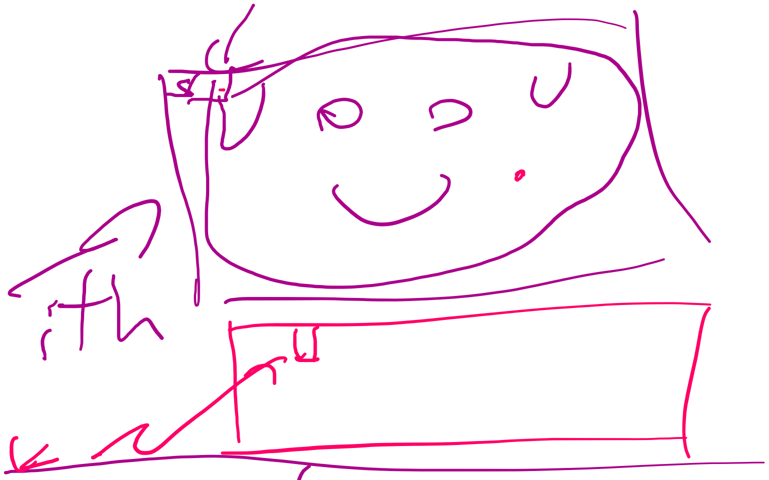
Better features

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

ϕ_k

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

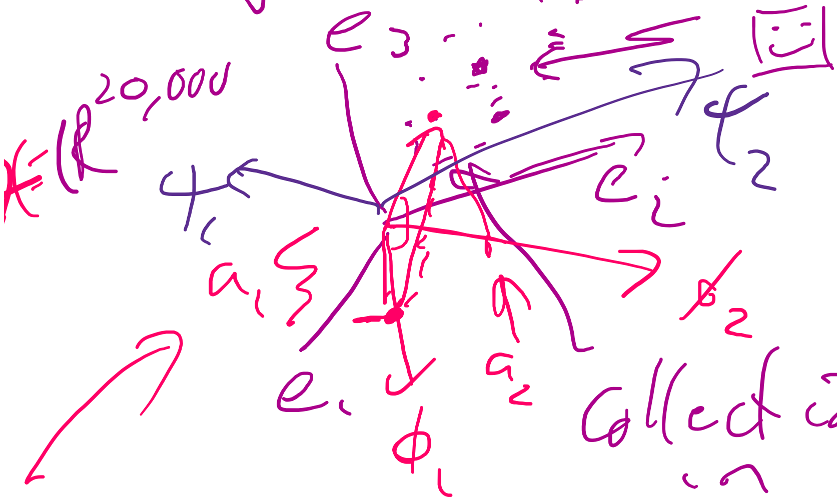
$$a_k = (X_k, \phi_k) = \int X_k(s)\phi_k(s)\,ds = \sum X_k(s)\phi_k(s)$$



= $\underline{20,000}$ the values in the pixels each pixel is considered a feature.

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

Project onto basis function.

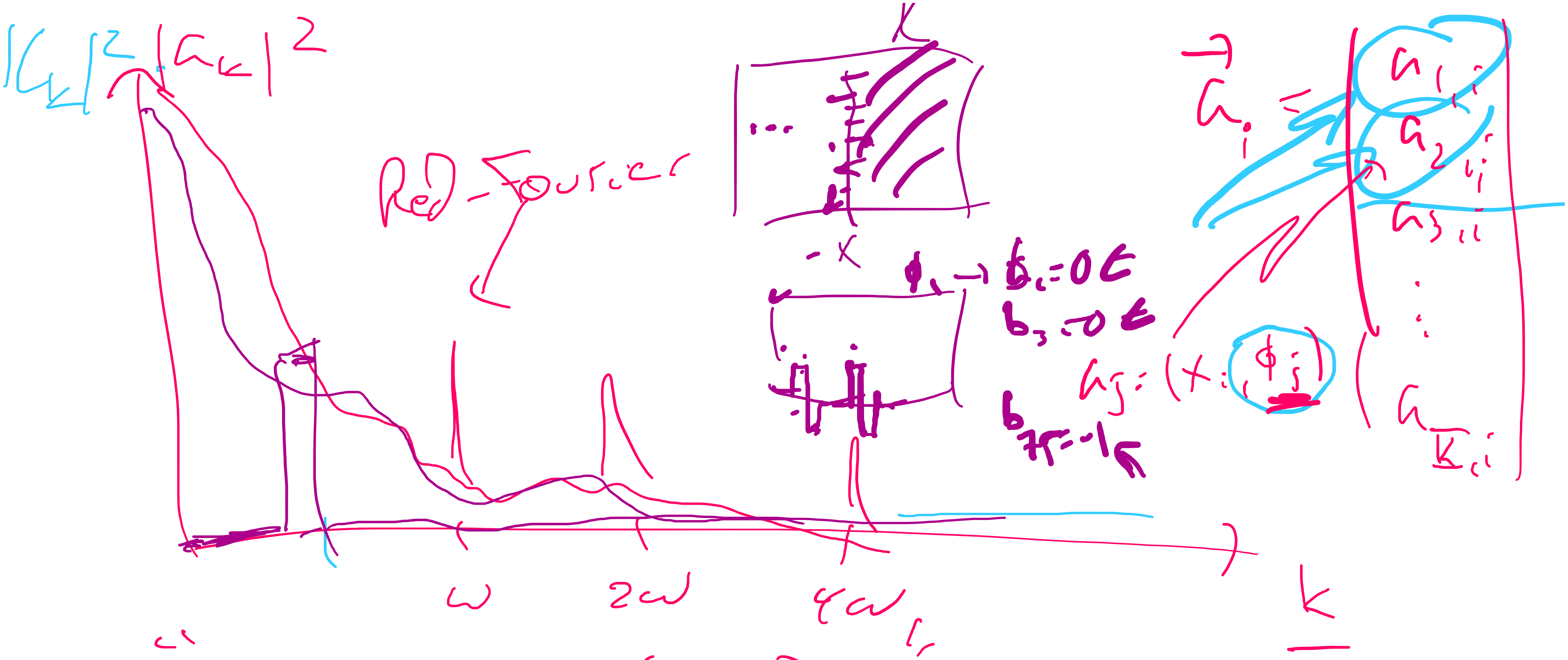


$$p_k = (x_i, \chi_k) = \int x_i(s)\,\chi_k(s)\,ds$$

↑ features

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

Collection of pics as $\underline{\text{data cloud}}$. in $\mathcal{H} = \mathbb{R}^{20,000}$



"Power spectrum".

$$\int |x(s)|^2 ds \overset{\leq}{\simeq} \sum |a_k|^2 \longleftarrow$$ Parseval

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

$$w^* = \arg\max_w J(w)$$

I

$$\min -\frac{1}{2} w^T S_B w$$

$$\text{s.t. } w^T S_W w = 1$$

$$\text{I} \Rightarrow \text{II}$$

$$\mathcal{L} = -\frac{1}{2} w^T S_B w + \frac{\lambda}{2}(w^T S_W w - 1)$$

$$(\mu_2-\mu_1)\cdot v$$

$$= (\mu_2-\mu_1)^T v$$

$$= v \cdot (\mu_2-\mu_1)$$

$$= v^T(\mu_2-\mu_1)$$

$$((\mu_2-\mu_1)\cdot w)^2 = w^T \underline{(\mu_2-\mu_1)\cdot(\mu_2-\mu_1)^T} w$$

let $\mathcal{L} = -\frac{1}{2} w^T S_B w + \frac{1}{2}\lambda(w^T S_W w - 1)$

$\frac{\partial \mathcal{L}}{\partial w_1} = 0$; $\frac{\partial \mathcal{L}}{\partial w_2} = 0, \cdots$; $\nabla_w \mathcal{L} = 0$

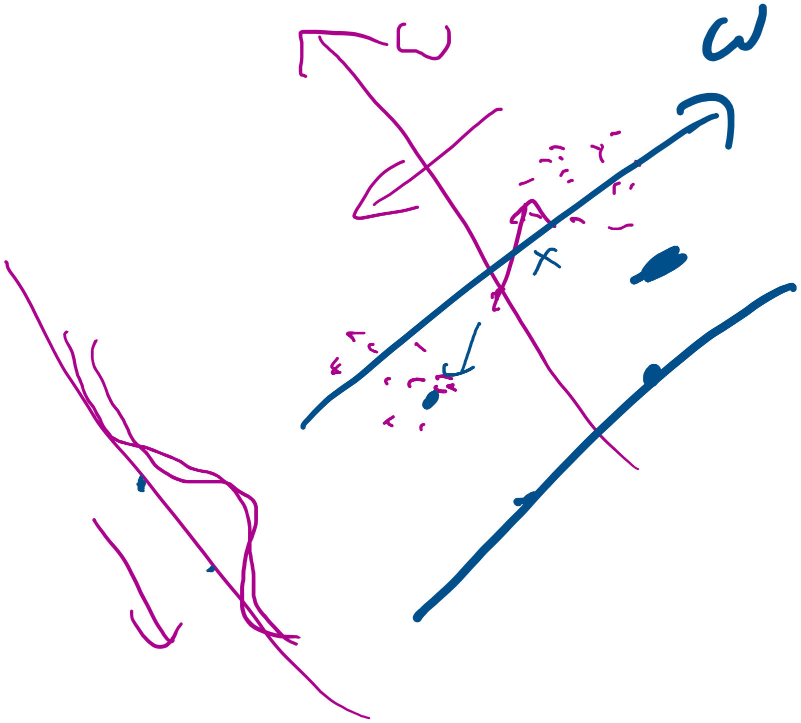
$\boxed{\text{KKT}}$

$\nabla_w \mathcal{L} = -S_B w + \lambda S_W w = 0$

$\boxed{S_B w = \lambda S_W w}$

$\nabla(\frac{1}{2} w^T w) = w$

$\boxed{Ax = \lambda Bx \text{ generalized eigenvector/value statement!}}$



$f : \underline{\overline{X}} \to \underline{\overline{Y}}$

$f : \underline{\overline{X}} \to ?$

(x, ω)

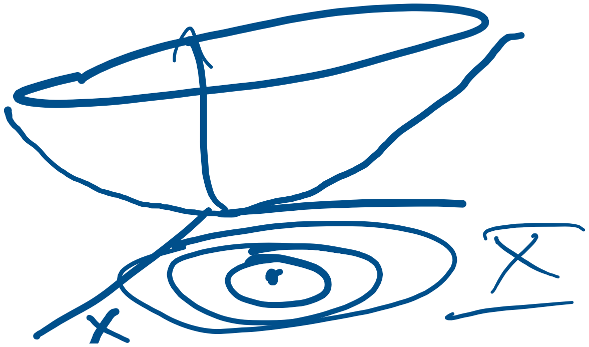
KKT :

$$x^* = \underset{x}{\arg\min}\ f(x)$$

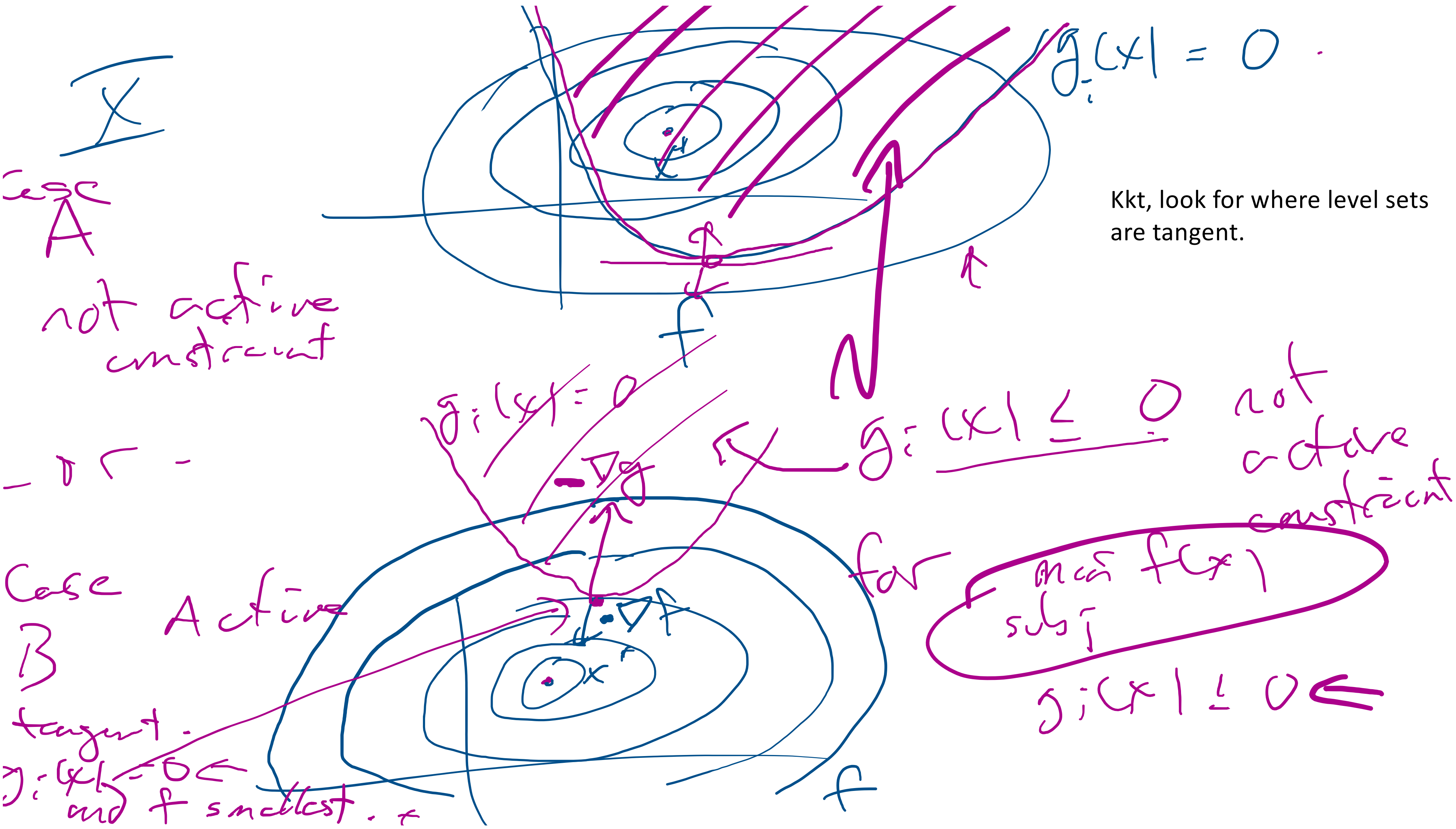
subj. $h_i(x) = 0, \forall i = 1, \dots, m$; $\vec{h}(\vec{x}) = \vec{0}$

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

$$x^* = \underset{\substack{x \\ \lambda, \mu}}{\arg\min}\ \mathcal{L}(x, \lambda, \mu) = \underset{x}{\arg\min}\ f(x) + \sum_{i}^{m} \lambda_i h_i(x) + \sum_{i}^{n} \mu_j g_j(x)$$



$f : \mathbb{X} \to \mathbb{R}$



Kkt, look for where level sets are tangent.

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

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

$$\nabla f + \lambda \nabla g = 0$$ .... Lagrange mult.

Equality constr.

→ $\nabla_x f + \sum \nabla_x \lambda_i h_i(x) + \sum \mu_i \nabla_x g_i(x) = 0$

- stationarity. ←
- ineq. constr true.

Support Vector Machines (SVM) – 5.7 Linear

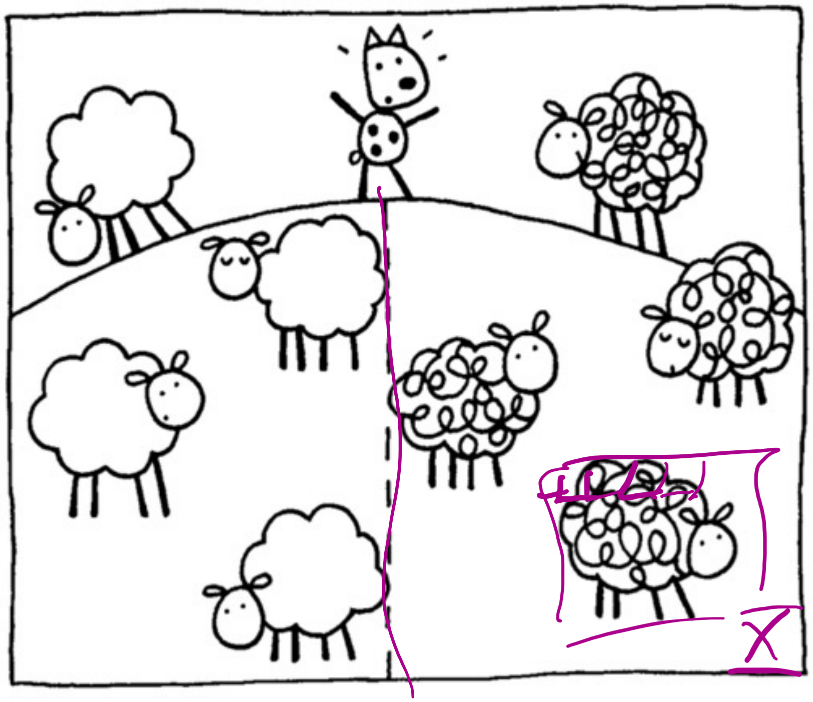
Then

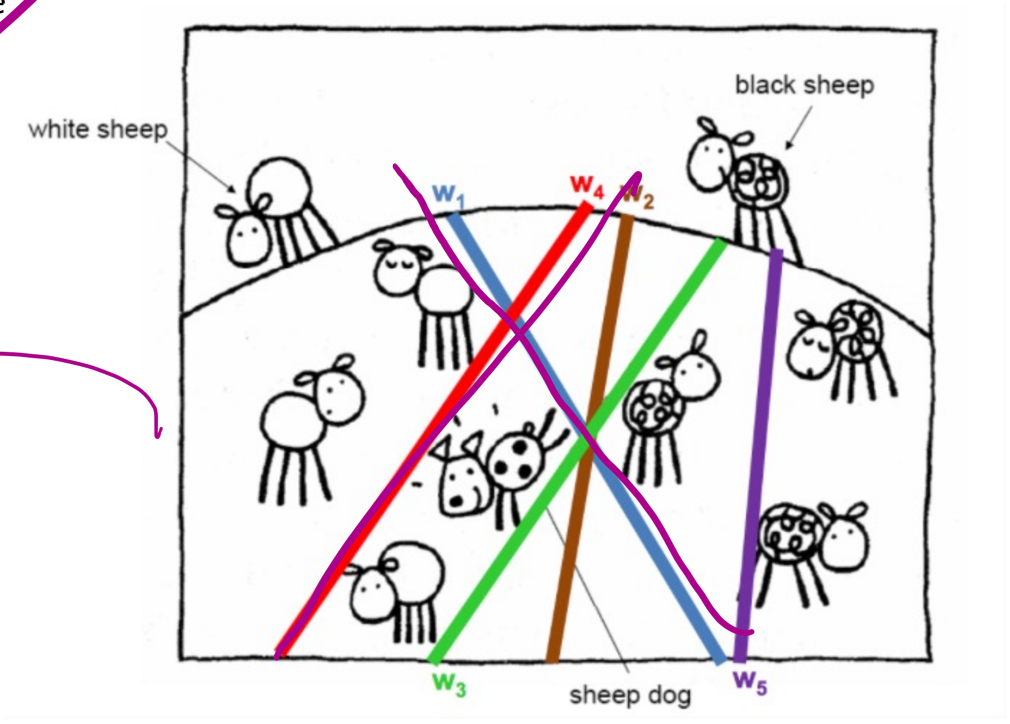
Nonlinear (kernelized) SVM (KSVM)

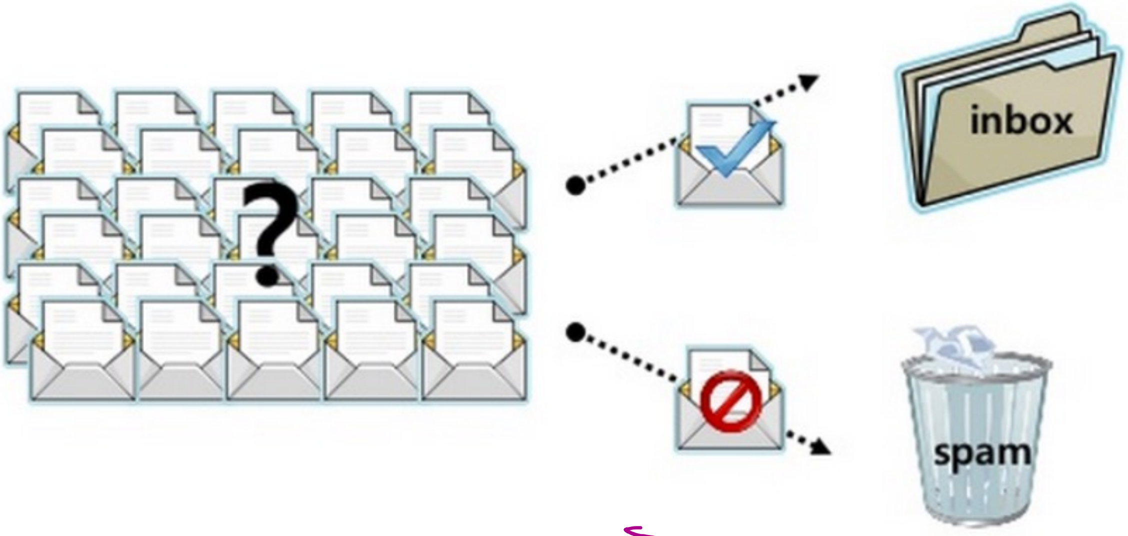
Smola - Schökopff.

Wide Margin Decision Hyperplane for
Supervised - Learning Classification

First a linear binary classification – decision boundary/hyperplane







$y_i = \mathbb{Z}_\xi$

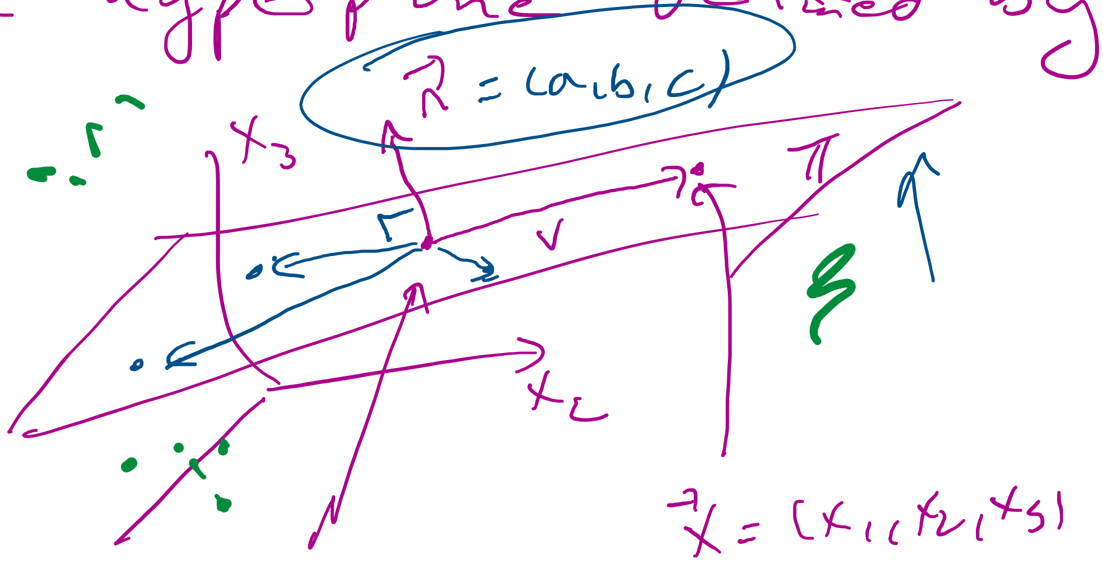
$\xi = 2$

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

- Instance space: $x \in X$ ($|X| = n$ data points)
  - Binary or real-valued feature vector x of word occurrences
  - d features (words + other things, d~100,000+)
- Class: $y \in Y$
  - y: Spam (+1), Ham (-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 eq'n is a co-dim-1 restriction of space

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

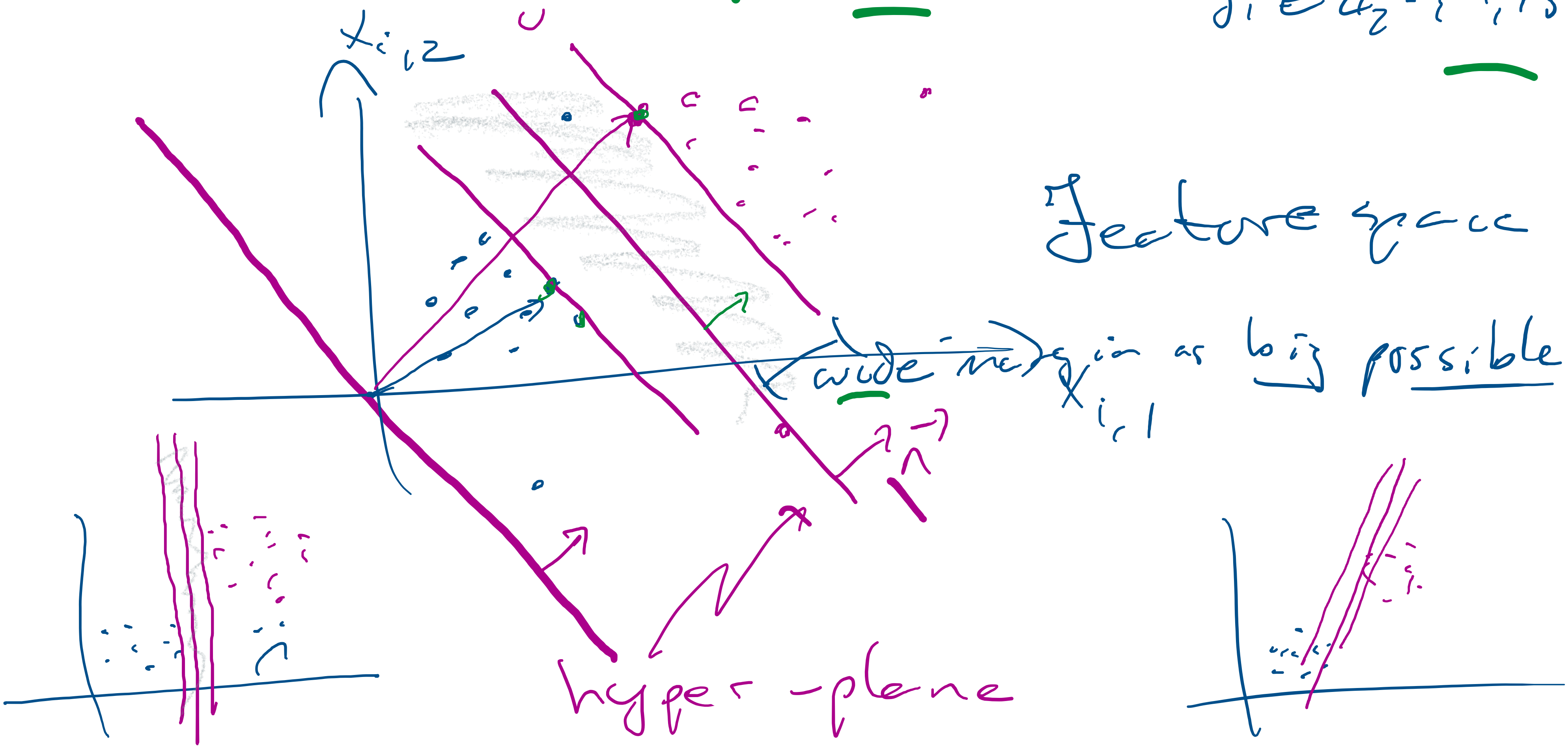
$v = \vec{x} - \vec{x}_0 = \langle (x_1 - x_{1,0}), (x_2 - x_{2,0}), (x_3 - x_{3,0}) \rangle$

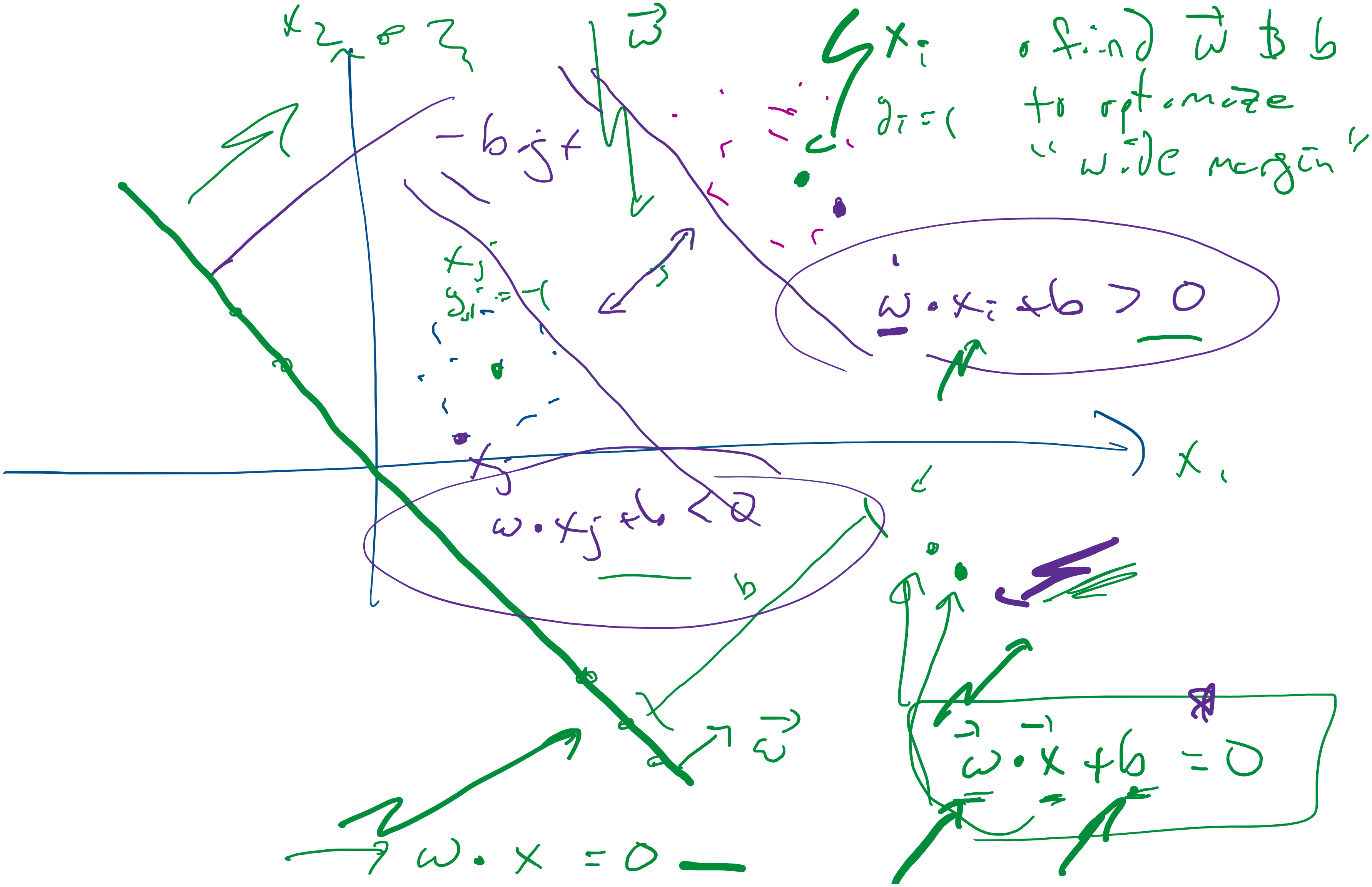
$\Pi := \{ x = (x_1, x_2, x_3) : \vec{v} = \underline{\vec{x} - \vec{x}_0} \perp \vec{n} \}$

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

$n \cdot v = \langle a, b, c \rangle \cdot \langle x_1 - x_{1,0}, x_2 - x_{2,0}, x_3 - x_{3,0} \rangle = a(x_1 - x_{1,0}) + b(x_2 - x_{20}) + c(x_3 - x_{30}) = 0$

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



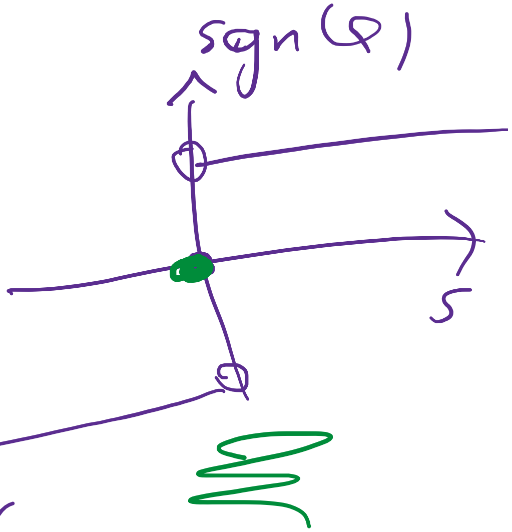


find $\vec{w}$ & b
to optimize
"wide margin"

$(\bar{x}_i, y_i)$, $y_i = -1, \boxed{1}$

($y_i = 0$ or 1
apple or orange
dog or cat.

$y_j(w \cdot x_j + b) = \text{sgn}(w \cdot x_j + b)$ good label.



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

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

a loss function ·

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

$\bar{y}_i$

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

Total loss $\quad \displaystyle\sum_{i=1}^{N} \ell(y_i, \bar{y}_i)$

Cost : ... ~~loss~~

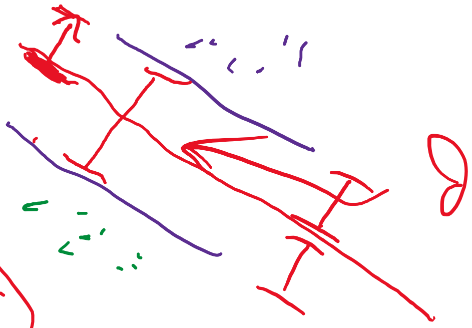
$w \cdot x_j$

$\underline{\text{argmin}} \; \boxed{\frac{1}{2} \|w\|_2^2}$ $\underline{\text{subj}}$ $\underline{y_i (\underline{w^T x_i - b}) - 1 = 0}$

small $\|w\|$

subj every matches truth

dist between



dist big

$\frac{2}{\|w\|_2^2}$ big $\approx$ $\boxed{\frac{\|w\|_2^2}{2}}$ small constrained opt.

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

s_i

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

$\{(x_i, y_i)\}$

$\frac{1}{2} w \cdot w$

$y_1(w^T x_1 - b) - 1 = 0$

$y_2(w^T x_2 - b) - 1 = 0$

$\vdots$

$y_n(w^T x_n - b) - 1 = 0$

m

$$= \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} s_i y_i (w^T \vec{x}_i - b) + \sum_{i=1}^{n} s_i$$

$\nabla_\theta \mathcal{L} = \vec{0}$

$\nabla_w \mathcal{L}(x, s, \theta) = w - \sum s_i y_i \vec{x}_i = \vec{0} \leftarrow$ d-eqns for d featu

$\nabla_b \mathcal{L} = \frac{\partial}{\partial b} \mathcal{L} = \sum s_i y_i = 0$

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

$x_i \in \mathbb{R}^d$

$s \in \mathbb{R}^n$ $\in \mathbb{R}$

n-values.

"Trick" - KKT - Primal Dual form.

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

s_i $\times$

Dual Form

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

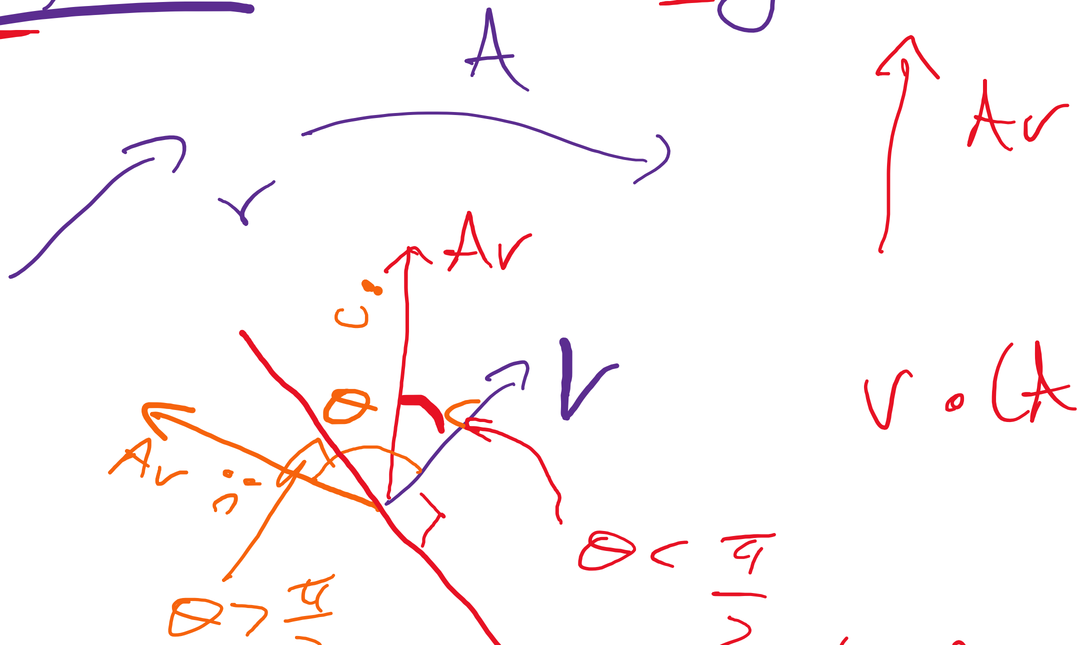
$\phi(x_i) \cdot \phi(x_j)$

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

$\sum_{i=1}^{n} s_i y_i = 0$

Pos. Semi-Defn Definitions -

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



$$v\cdot(Av) = \overset{+,}{\|v\|}\,\overset{+,}{\|Av\|}\cos\theta$$

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

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

vector space

- a Hilbert space is a set $\mathcal{H}$ of vectors such that there is a complete inner product.

limits work properly.

dot

Spectral Decomp. thm:

Suppose $A_{n \times n}$ is pos. defn. symm. matrix with eigenvectors & eigenvalues λ_i, v_i, $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$

$$\underline{A} = \sum_{i=1}^{n} \underline{\lambda_i} \boxed{v_i v_i^T}$$

$v_i \otimes v_i = \boxed{P_i}$ rank-1 projector.

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

$\leftarrow \underline{X} = \mathbb{R}^2$

∘ $x_i \cdot x_j$ need the dot product between $\underline{x}_i$ & $\underline{x}_j$ to do SVM.

∘ $\underline{K(x_i, x_j)} = \phi(x_i) \cdot \phi(x_j)$

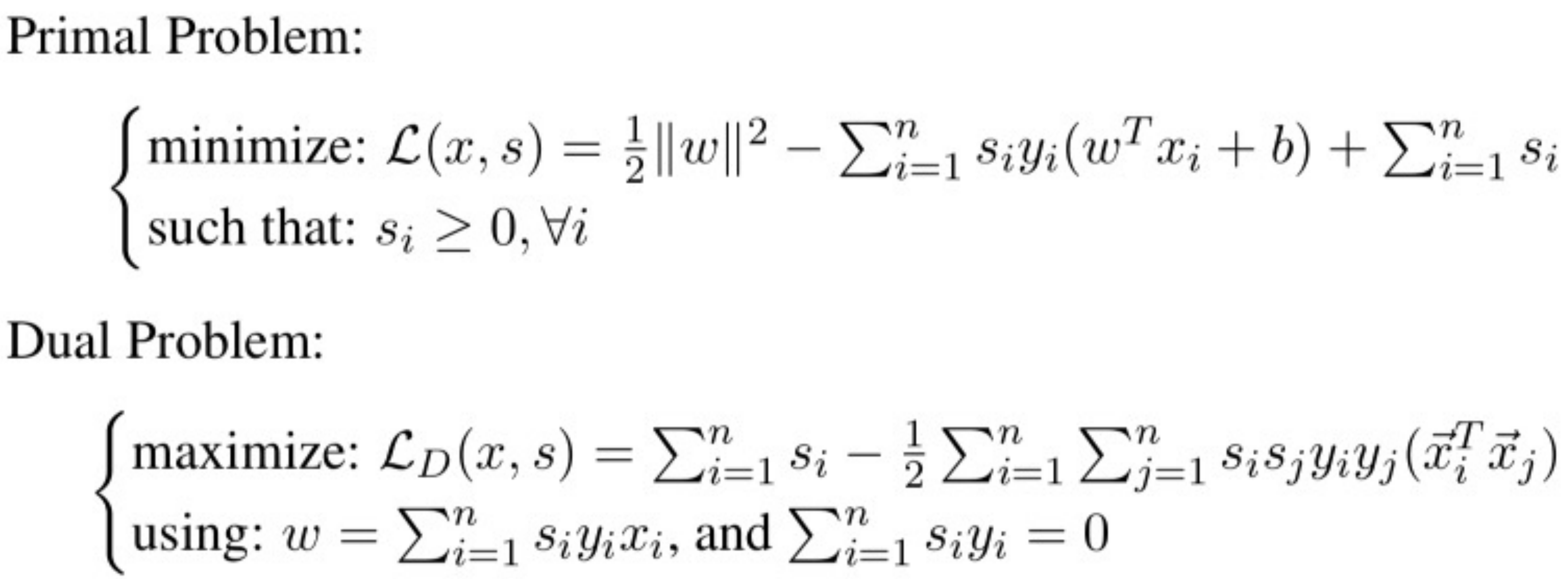
$\uparrow$ corresponds. $\uparrow$

$\phi : \underline{X} \to \mathcal{H} \leftarrow \mathbb{R}^6$

Gram matrix. $\leftarrow$ symmetric

$$\underline{K} = \begin{pmatrix} \underline{K(x_1, x_2)} & K(x_2, x_3) & \cdots & \underline{K(x_1, x_n)} \\ \vdots & & & \\ \underline{K(x_n, x_1)} & \cdots & & K(x_n, x_n) \end{pmatrix}$$

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 kerekl function

$\textcircled{0}$ $\mathbb{X} = \mathbb{R}^2$

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: $K(x, \tilde{x}) = (x^T \tilde{x} - y)^d$. $+1$
- Gaussian RBF: $K(x, \tilde{x}) = e^{-\frac{Vertx - \tilde{x}\|^2}{2\sigma^2}}$ $e^{-\frac{\|x-\tilde{x}\|}{2\sigma^2}}$

Kernel is

- $\textcircled{0}$ —
- $\textcircled{1}$ symmetric
- $\textcircled{2}$ pos. semi-defn
- $\textcircled{3}$ cts. w.r.t. both inputs.

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

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

$$\begin{aligned} K(x, \tilde{x}) &= (x \cdot \tilde{x} + 1)^d \\ &= (x^T \tilde{x} + 1)^d \\ &= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + 1)^2 \\ &= x_1^2 \tilde{x}_1^2 + 2x_1 \tilde{x}_1 + 2x_2 \tilde{x}_2 + x_1 \tilde{x}_1 x_2 \tilde{x}_2 + 1 + x_2^2 \tilde{x}_2^2 \end{aligned}$$

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

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

d ↑ D ↑↑ maybe

D = ∞ .

$$
\begin{aligned}
K(x,\tilde{x}) &= (x\cdot\tilde{x}+1)^d \\
&= (x^T\tilde{x}+1)^d \\
&= (x_1\tilde{x}_1+x_2\tilde{x}_2+1)^2 \\
&= x_1^2\tilde{x}_1^2+2x_1\tilde{x}_1+2x_2\tilde{x}_2+x_1\tilde{x}_1x_2\tilde{x}_2+1 \quad + x_2^2\tilde{x}_2^2
\end{aligned}
$$

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

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

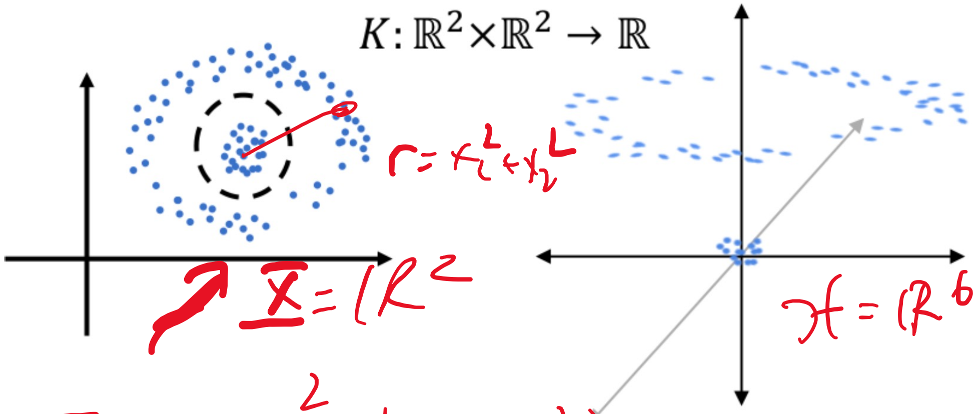
n → n-1

$\phi\colon \mathbb{R}^2 \to \mathbb{R}^6$

$K\colon \mathbb{R}^2\times\mathbb{R}^2 \to \mathbb{R}$

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

$\mathbb{R}^2$ $\mathbb{R}$ C D

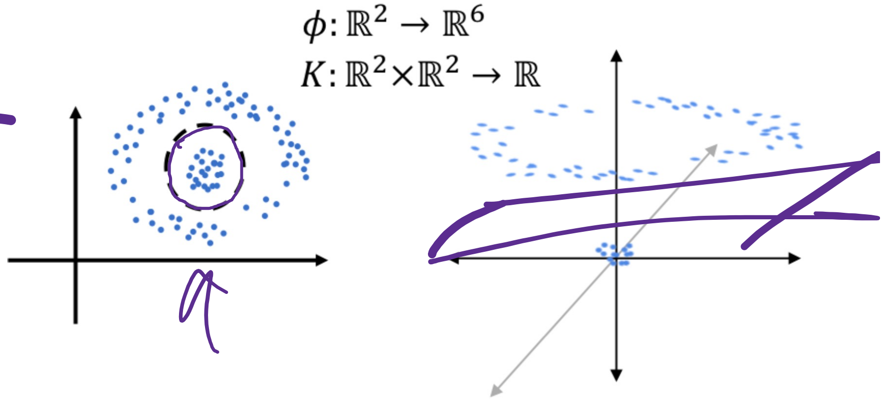


$$\phi : X \longrightarrow H$$

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

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

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



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

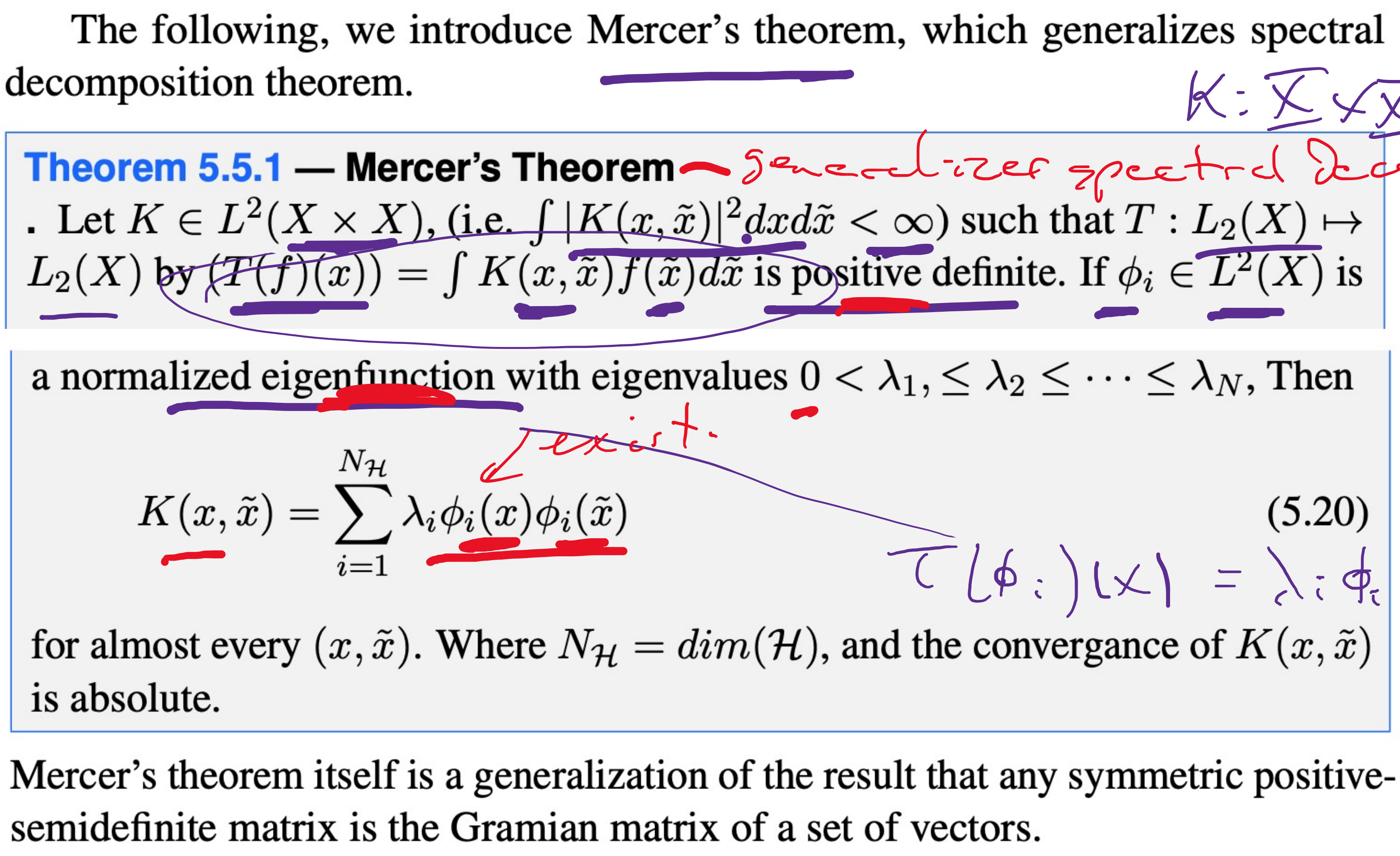
$K\colon \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$

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.

$K : X \times X \to \mathbb{R}$

Theorem 5.5.1 — Mercer's Theorem — generalizes spectral decomp thm.

. 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 $0 < \lambda_1, \leq \lambda_2 \leq \cdots \leq \lambda_N$, Then

usual

$Av = U$

e.g. eigen

$Aw = \lambda w$

exist.

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

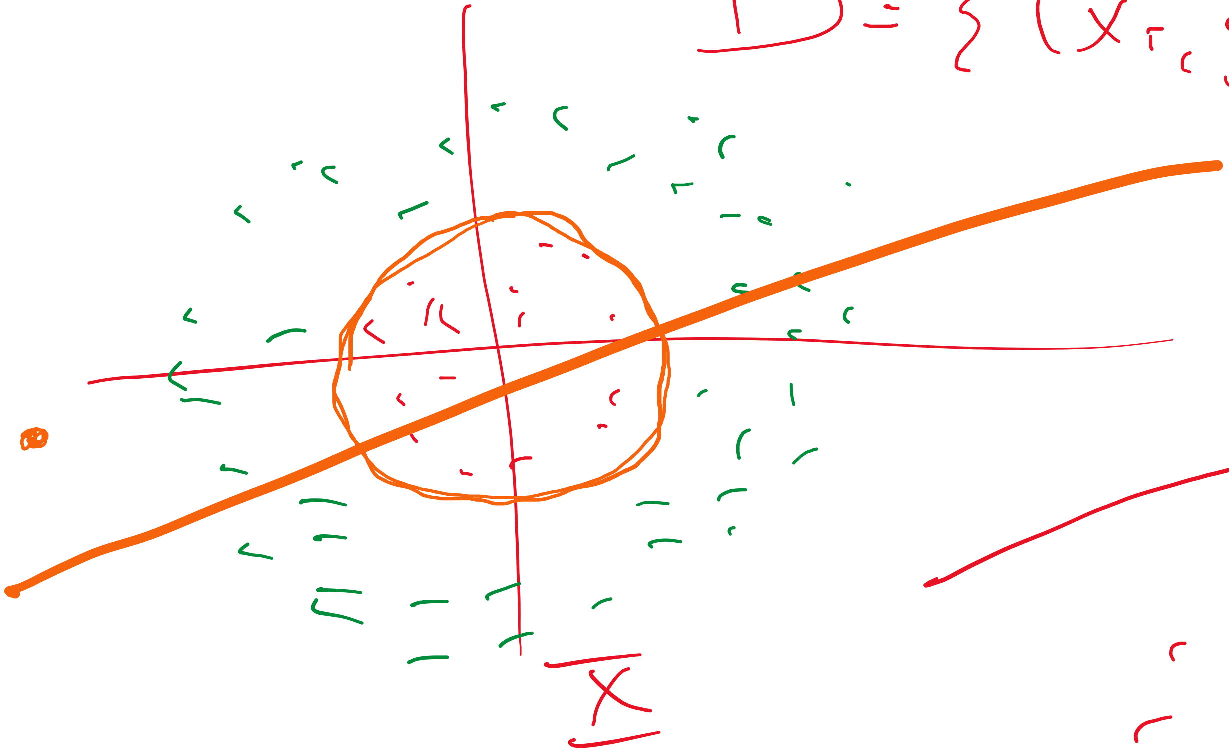
$T(\phi_i)(x) = \lambda_i \phi_i(x) \leftarrow$ eigen fn.

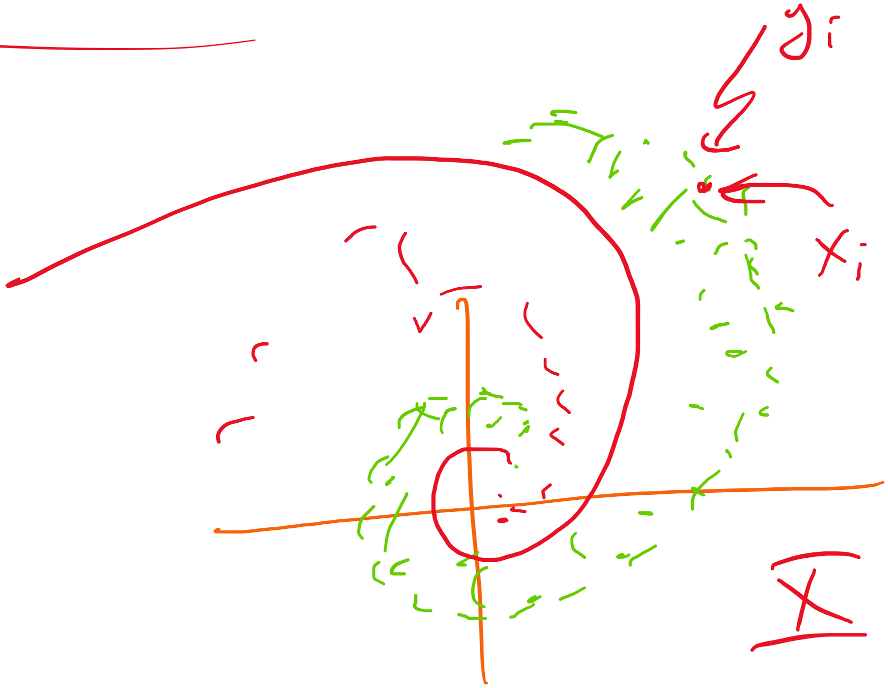
for almost every $(x,\tilde{x})$. Where $N_{\mathcal{H}} = dim(\mathcal{H})$, and the convergance of $K(x,\tilde{x})$ is absolute.

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 KSVM.

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





And now for something completely different