

AIN311

Fundamentals of Machine Learning

Lecture 3: Kernel Regression, Distance Metrics, Curse of Dimensionality



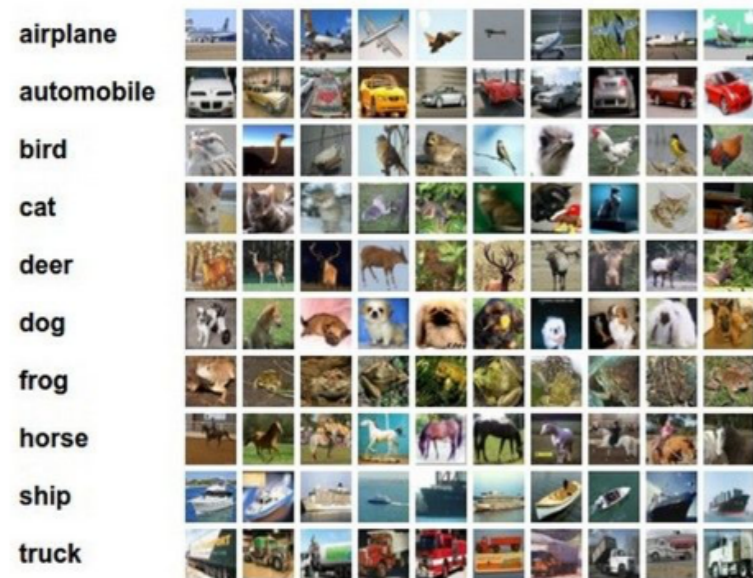
Recall from last time... Nearest Neighbors

Example dataset: **CIFAR-10**

10 labels

50,000 training images

10,000 test images.

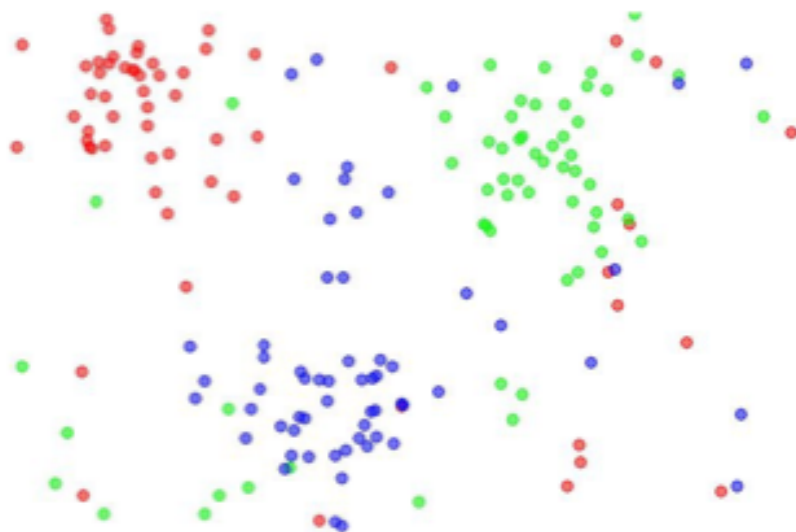


For every test image (first column),
examples of nearest neighbors in rows

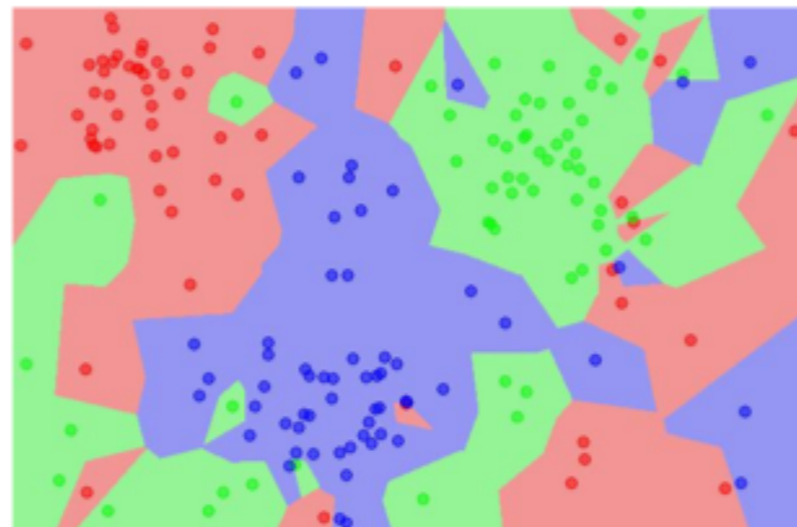


- Very simple method
- Retain all training data
 - It can be slow in testing
 - Finding NN in high dimensions is slow
- Metrics are very important
- Good baseline

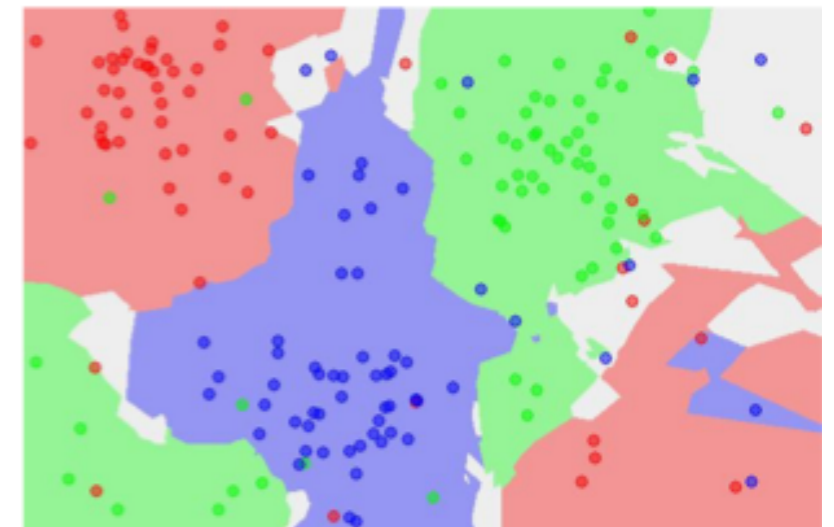
the data



NN classifier



5-NN classifier



Classification

- Input: X
 - Real valued, vectors over real.
 - Discrete values (0,1,2,...)
 - Other structures (e.g., strings, graphs, etc.)
- Output: Y
 - Discrete (0,1,2,...)

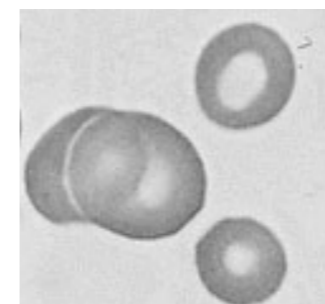
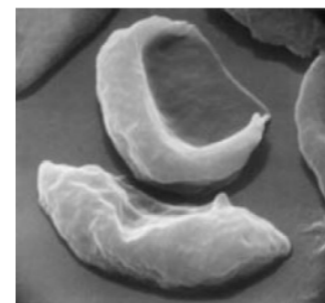


X = Document



Sports
Science
News

Y = Topic



X = Cell Image



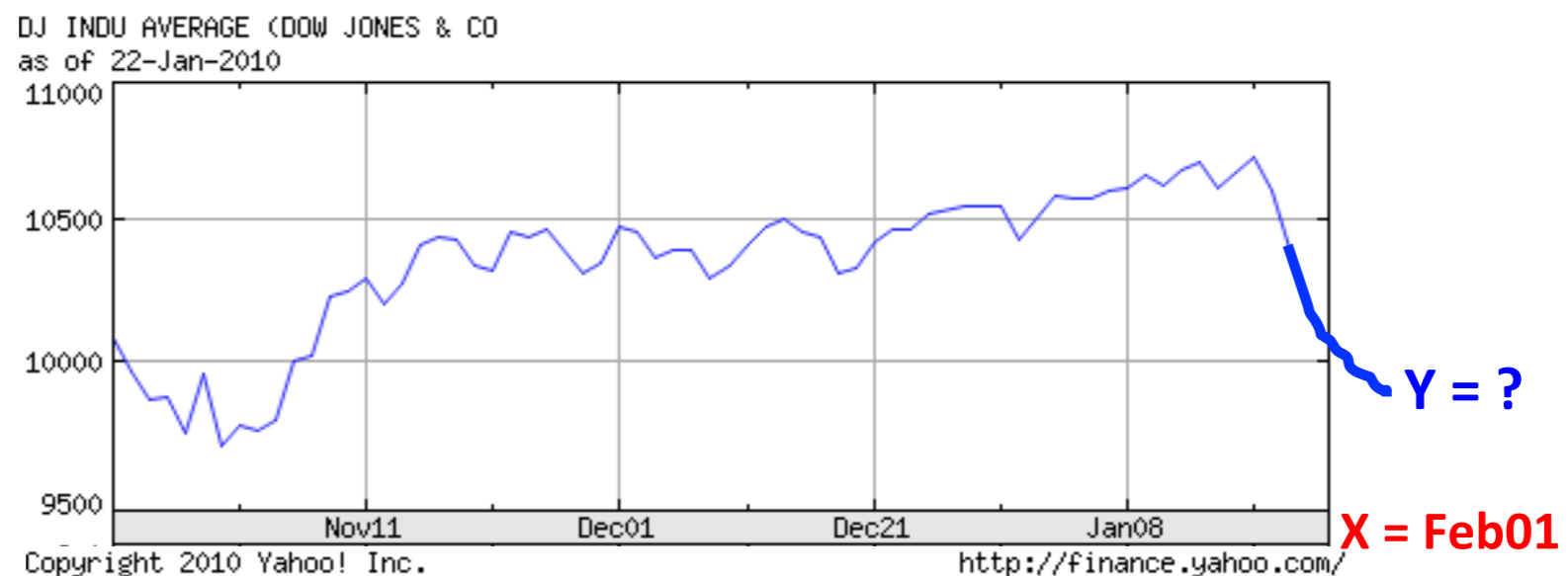
Anemic cell
Healthy cell

Y = Diagnosis

Regression

- Input: X
 - Real valued, vectors over real.
 - Discrete values (0,1,2,...)
 - Other structures (e.g., strings, graphs, etc.)
- Output: Y
 - Real valued, vectors over real.

Stock Market
Prediction



What should I watch tonight?

IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist

The Martian (2015) 9

PG-13 | 144 min | Adventure, Comedy, Drama | 2 October 2015 (USA)

8.1 Your rating: ★★★★★★☆☆☆☆ -/10
Ratings: **8.1/10** from 271,829 users Metascore: 80/100
Reviews: 750 user | 499 critic | 46 from Metacritic.com

During a manned mission to Mars, Astronaut Mark Watney is presumed dead after a fierce storm and left behind by his crew. But Watney has survived and finds himself stranded and alone on the hostile planet. With only meager supplies, he must draw upon his ingenuity, wit and spirit to subsist and find a way to signal to Earth that he is alive.

Director: Ridley Scott
Writers: Drew Goddard (screenplay), Andy Weir (book)
Stars: Matt Damon, Jessica Chastain, Kristen Wiig | See full cast and crew »

+ Watchlist ▼ Watch Trailer Share...

BRING HIM HOME
MATT DAMON
THE MARTIAN
OCTOBER

See More on IMDb Pro »

What should I watch tonight?



IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist



Point Break (2015) 15

PG-13 | 114 min | Action, Crime, Sport | 25 December 2015 (USA)

Your rating: ★★★★★★ -/10

5.4 Ratings: 5.4/10 from 7,322 users Metascore: 34/100
Reviews: 60 user | 84 critic | 19 from Metacritic.com

A young FBI agent infiltrates an extraordinary team of extreme sports athletes he suspects of masterminding a string of unprecedented, sophisticated corporate heists. "Point Break" is inspired by the classic 1991 hit.

Director: [Ericson Core](#)

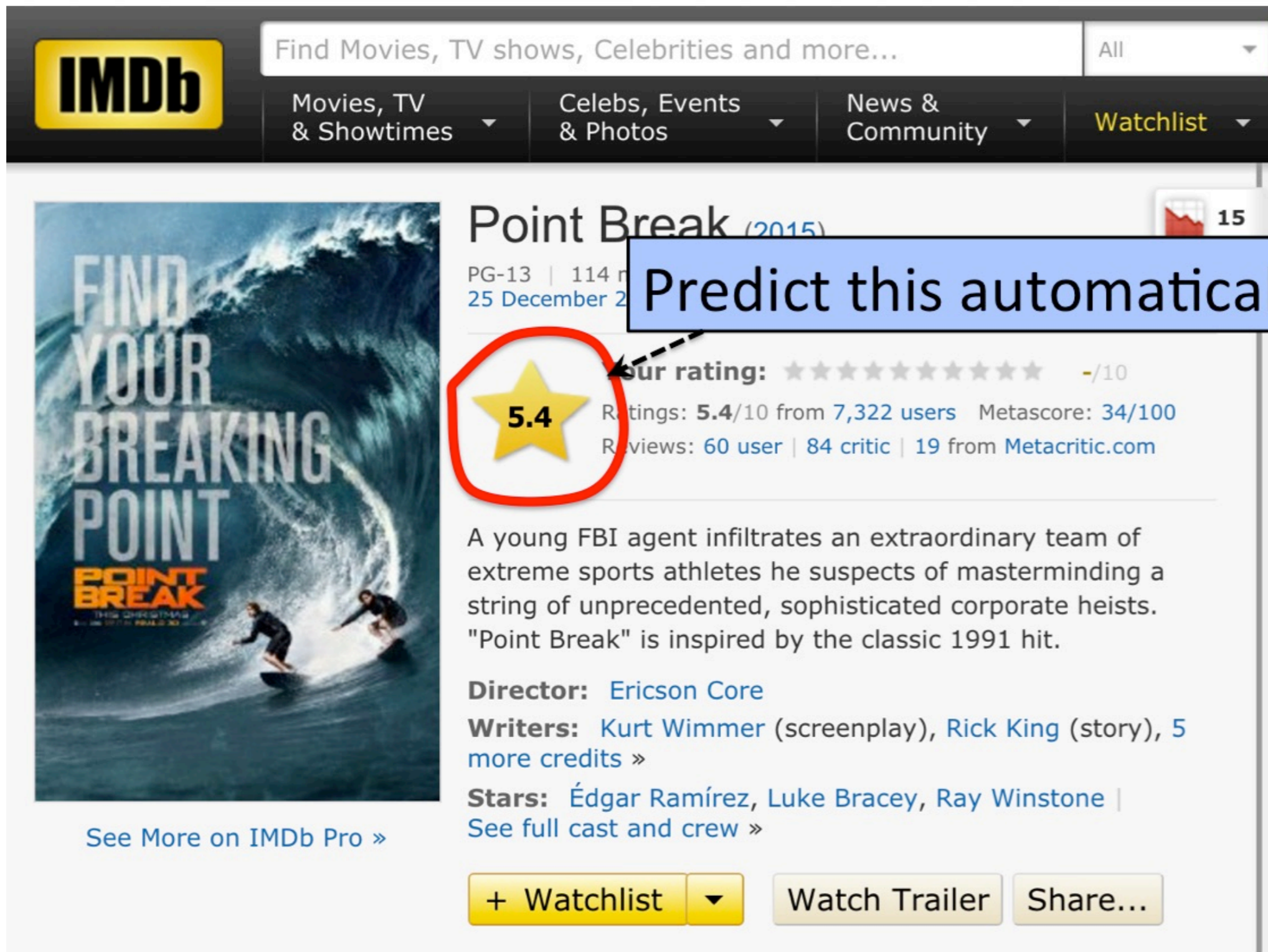
Writers: [Kurt Wimmer](#) (screenplay), [Rick King](#) (story), [5 more credits](#) »

Stars: [Édgar Ramírez](#), [Luke Bracey](#), [Ray Winstone](#) | [See full cast and crew](#) »

[+ Watchlist](#) [Watch Trailer](#) [Share...](#)

[See More on IMDb Pro](#) »

What should I watch tonight?



The image shows a screenshot of the IMDb website for the movie "Point Break (2015)". The IMDb logo is in the top left. A search bar contains the text "Find Movies, TV shows, Celebrities and more...". Navigation tabs include "Movies, TV & Showtimes", "Celebs, Events & Photos", "News & Community", and "Watchlist". The movie title "Point Break (2015)" is displayed with a "15" rating icon. A blue box with the text "Predict this automatically!" has a dashed arrow pointing to a yellow star with the number "5.4" inside, which is circled in red. Below the star, the text reads "Your rating: ★★★★★★ -/10", "Ratings: 5.4/10 from 7,322 users", "Metascore: 34/100", and "Reviews: 60 user | 84 critic | 19 from Metacritic.com". A synopsis follows: "A young FBI agent infiltrates an extraordinary team of extreme sports athletes he suspects of masterminding a string of unprecedented, sophisticated corporate heists. 'Point Break' is inspired by the classic 1991 hit." Credits for Director (Ericson Core), Writers (Kurt Wimmer, Rick King), and Stars (Édgar Ramírez, Luke Bracey, Ray Winstone) are listed. At the bottom are buttons for "+ Watchlist", "Watch Trailer", and "Share...".

IMDb

Find Movies, TV shows, Celebrities and more...

Movies, TV & Showtimes

Celebs, Events & Photos

News & Community

Watchlist

Point Break (2015)

PG-13 | 114 min | 25 December 2015

5.4

Your rating: ★★★★★★ -/10

Ratings: 5.4/10 from 7,322 users Metascore: 34/100

Reviews: 60 user | 84 critic | 19 from Metacritic.com

A young FBI agent infiltrates an extraordinary team of extreme sports athletes he suspects of masterminding a string of unprecedented, sophisticated corporate heists. "Point Break" is inspired by the classic 1991 hit.

Director: Ericson Core

Writers: Kurt Wimmer (screenplay), Rick King (story), 5 more credits »

Stars: Édgar Ramírez, Luke Bracey, Ray Winstone | See full cast and crew »

+ Watchlist

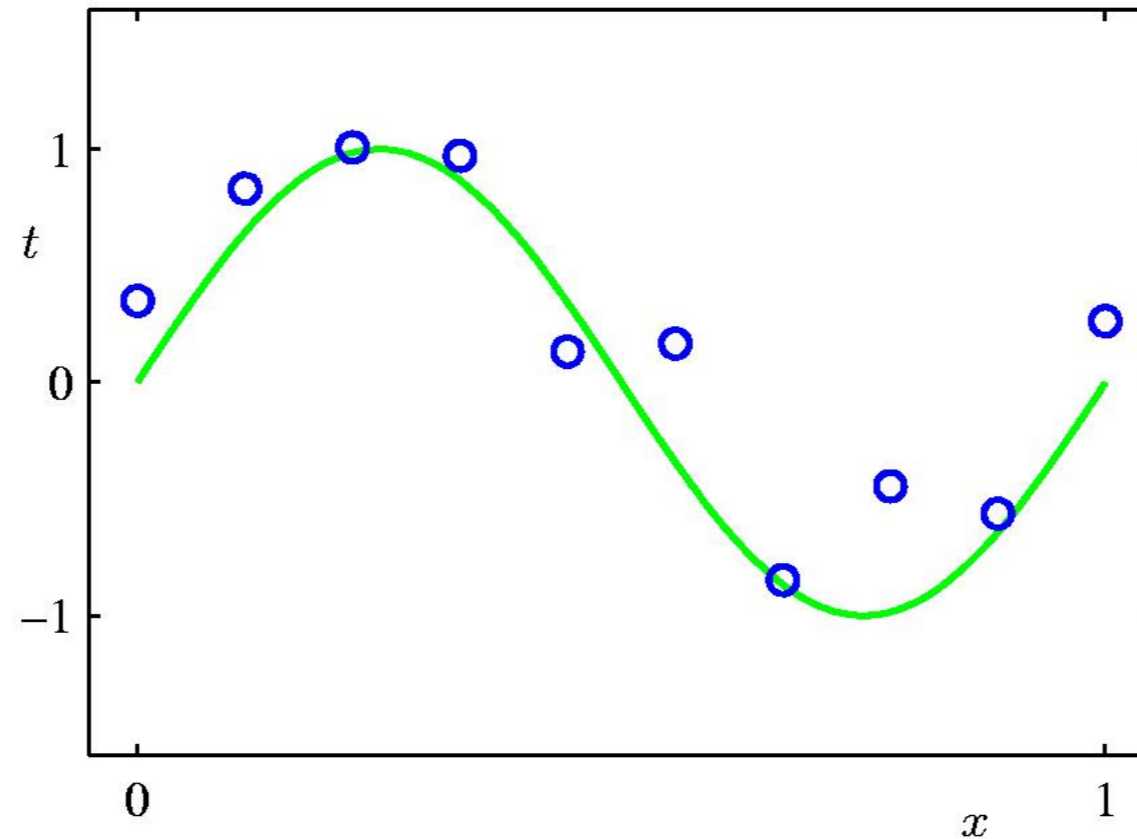
Watch Trailer

Share...

Today

- Kernel regression
 - nonparametric
- Distance metrics
- Linear regression (*more on our next lecture*)
 - parametric
 - simple model

Simple 1-D Regression



- Circles are data points (i.e., training examples) that are given to us
- The data points are uniform in x , but may be displaced in y

$$t(x) = f(x) + \varepsilon$$

with ε some noise

- In **green** is the “true” curve that we don’t know

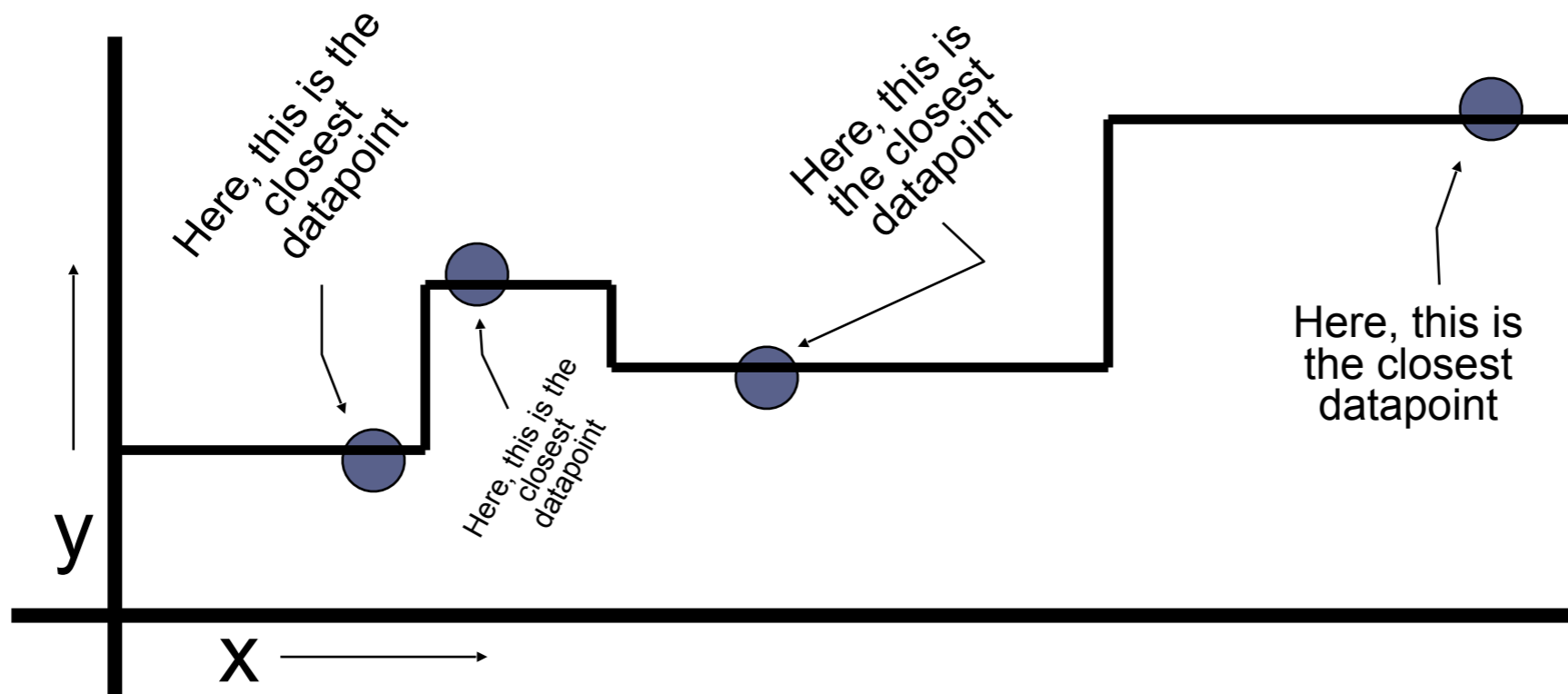
Kernel Regression

K-NN for Regression

- Given: Training data $\{(x_1, y_1), \dots, (x_n, y_n)\}$
 - Attribute vectors: $x_i \in X$
 - Target attribute $y_i \in \mathcal{R}$
- Parameter:
 - Similarity function: $K : X \times X \rightarrow \mathcal{R}$
 - Number of nearest neighbors to consider: k
- Prediction rule
 - New example x'
 - K-nearest neighbors: k train examples with largest $K(x_i, x')$

$$h(\vec{x}') = \frac{1}{k} \sum_{i \in k\text{nn}(\vec{x}')} y_i$$

1-NN for Regression



1-NN for Regression

- Often bumpy (overfits)

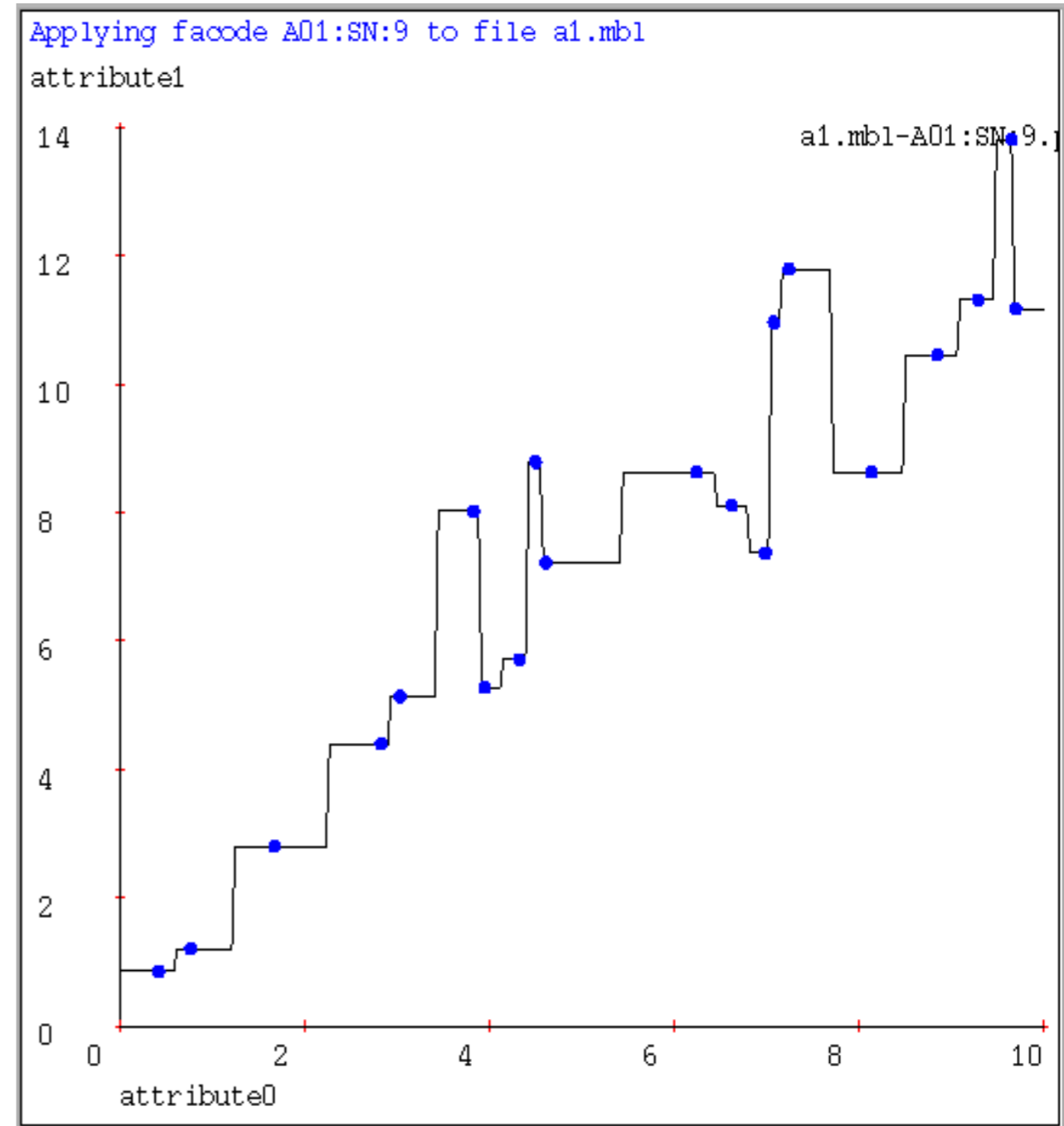
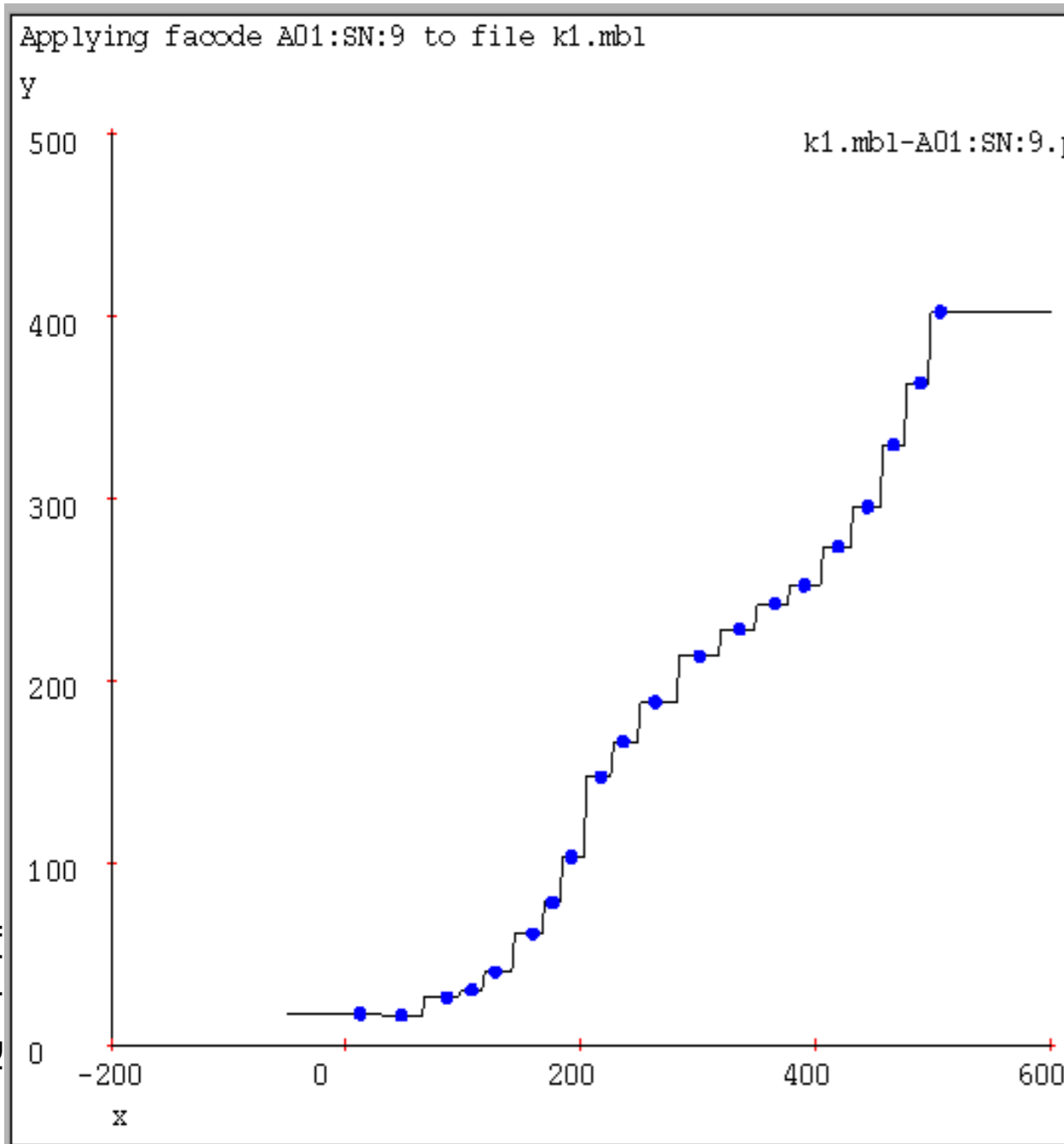


Figure Credit: Andrew Moore

9-NN for Regression

- Often bumpy (overfits)

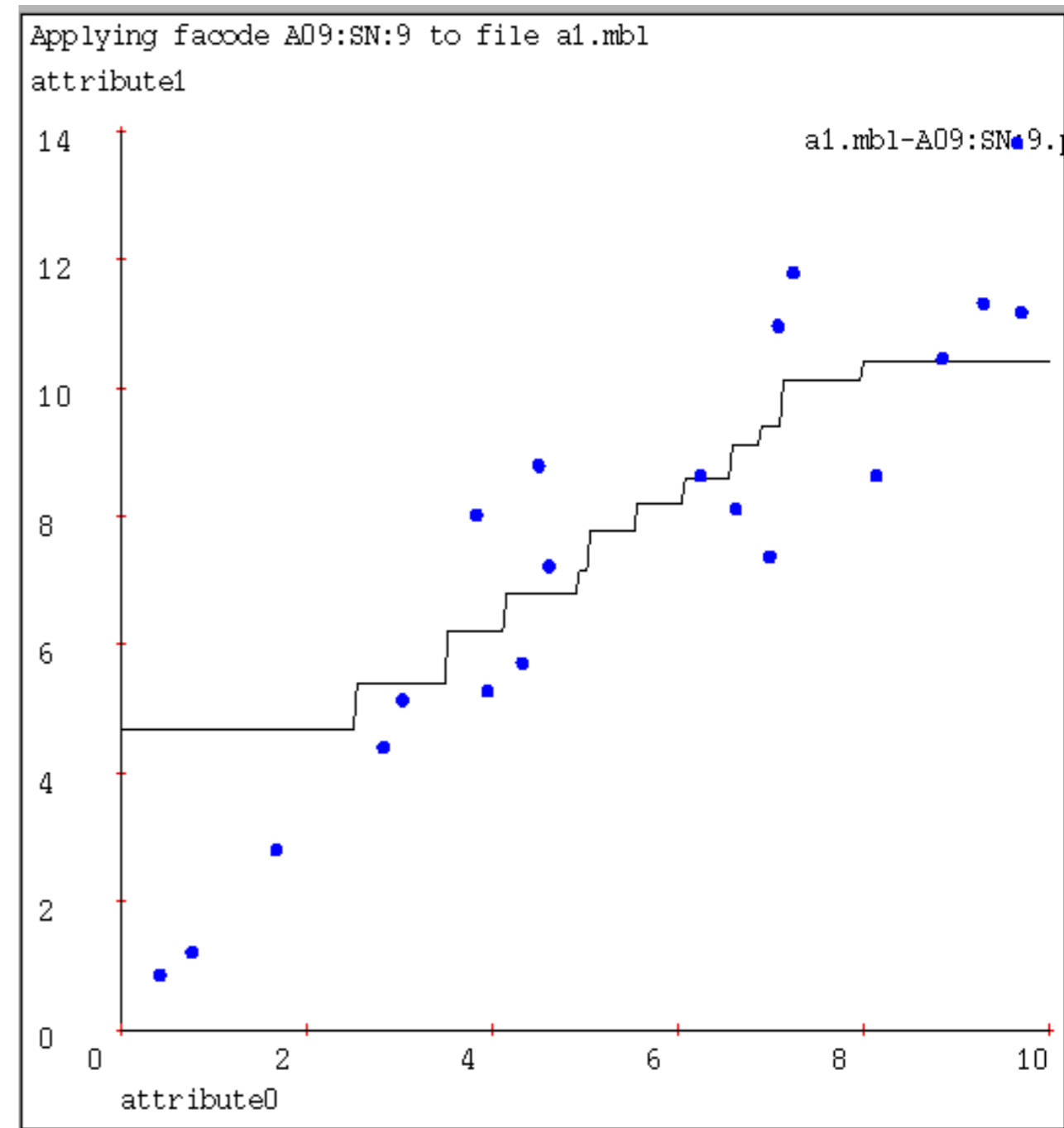
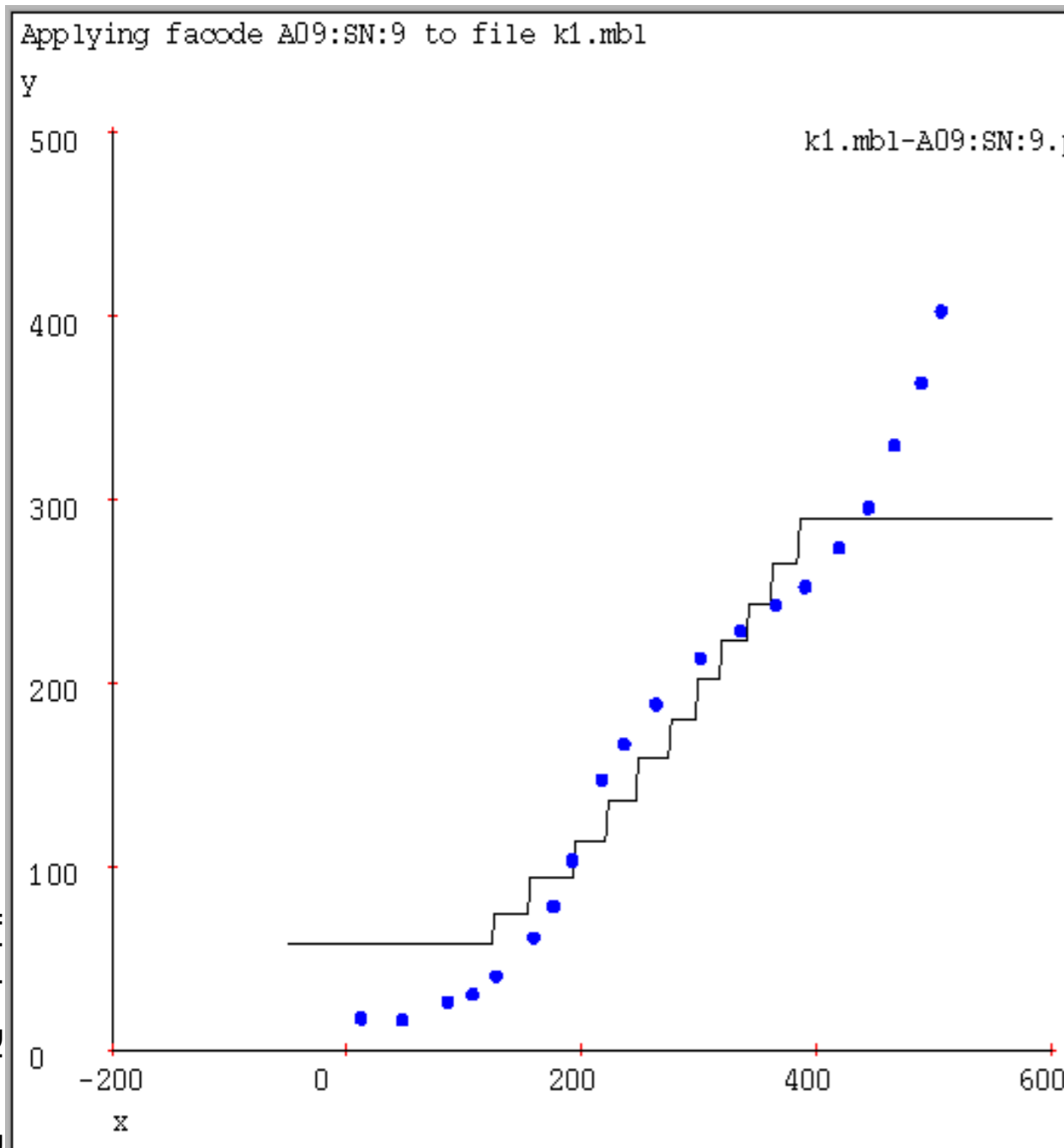
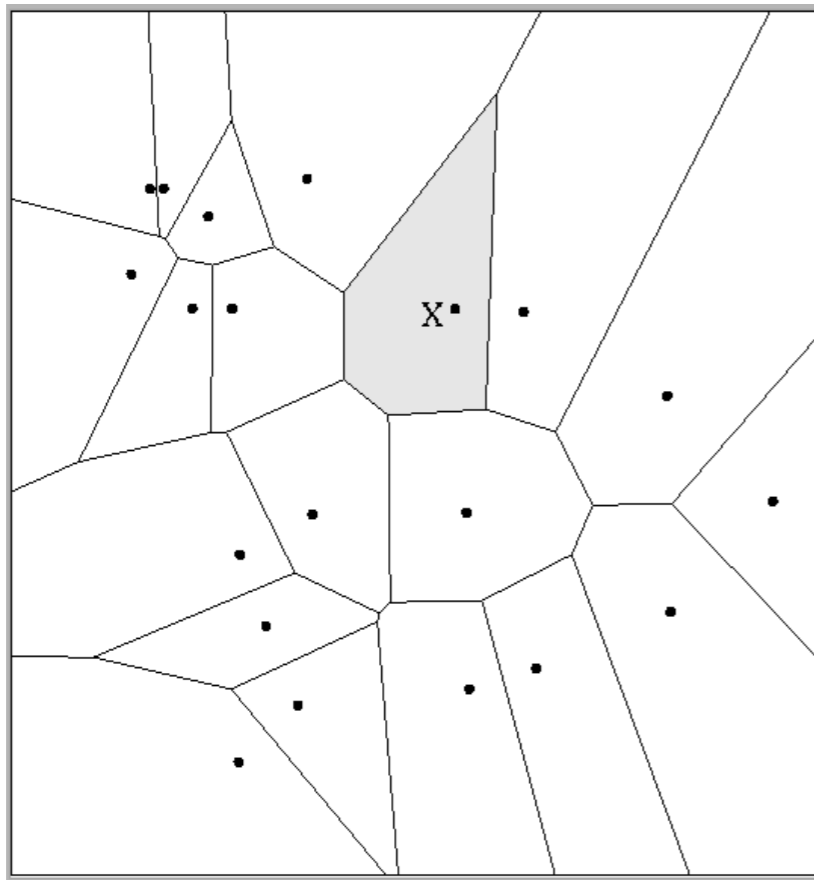


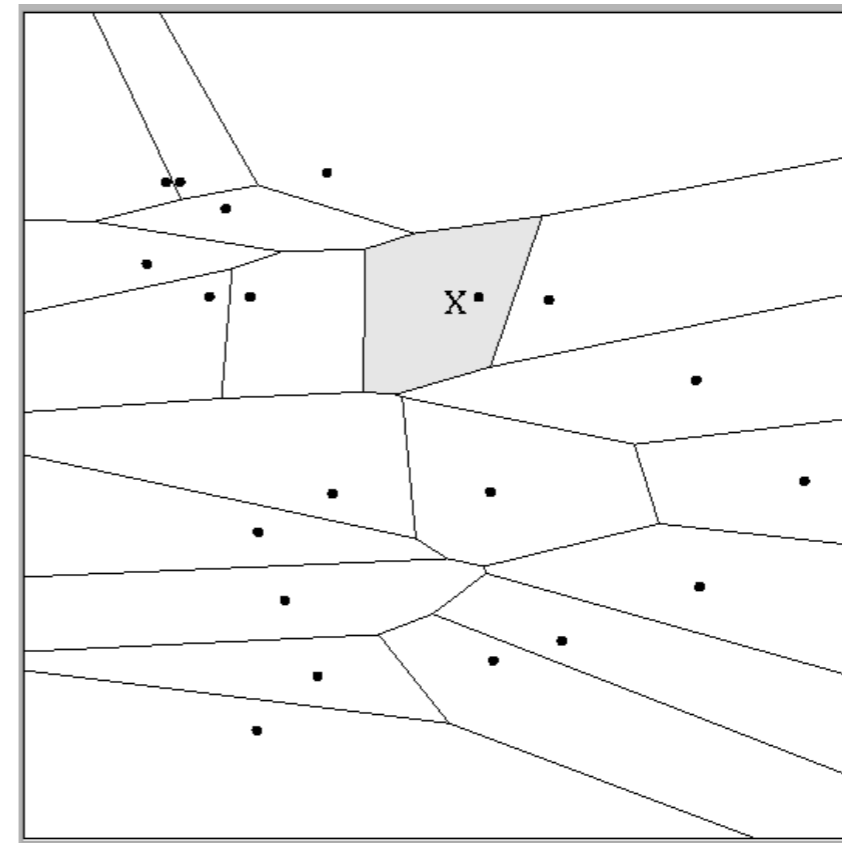
Figure Credit: Andrew Moore

Multivariate distance metrics

- Suppose the input vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are two dimensional:
 $\mathbf{x}_1 = (x_{11}, x_{12}), \mathbf{x}_2 = (x_{21}, x_{22}), \dots, \mathbf{x}_N = (x_{N1}, x_{N2})$.
- One can draw the nearest-neighbor regions in input space.



$$\text{Dist}(\mathbf{x}_i, \mathbf{x}_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$$



$$\text{Dist}(\mathbf{x}_i, \mathbf{x}_j) = (x_{i1} - x_{j1})^2 + (3x_{i2} - 3x_{j2})^2$$

The relative scalings in the distance metric affect region shapes

Example: Choosing a restaurant

- In everyday life we need to make decisions by taking into account lots of factors
- The question is what weight we put on each of these factors (how important are they with respect to the others).

Reviews (out of 5 stars)	\$	Distance	Cuisine (out of 10)
4	30	21	7
2	15	12	8
5	27	53	9
3	20	5	6



Euclidean distance metric

$$D(x, x') = \sqrt{\sum_i \sigma_i^2 (x_i - x'_i)^2}$$

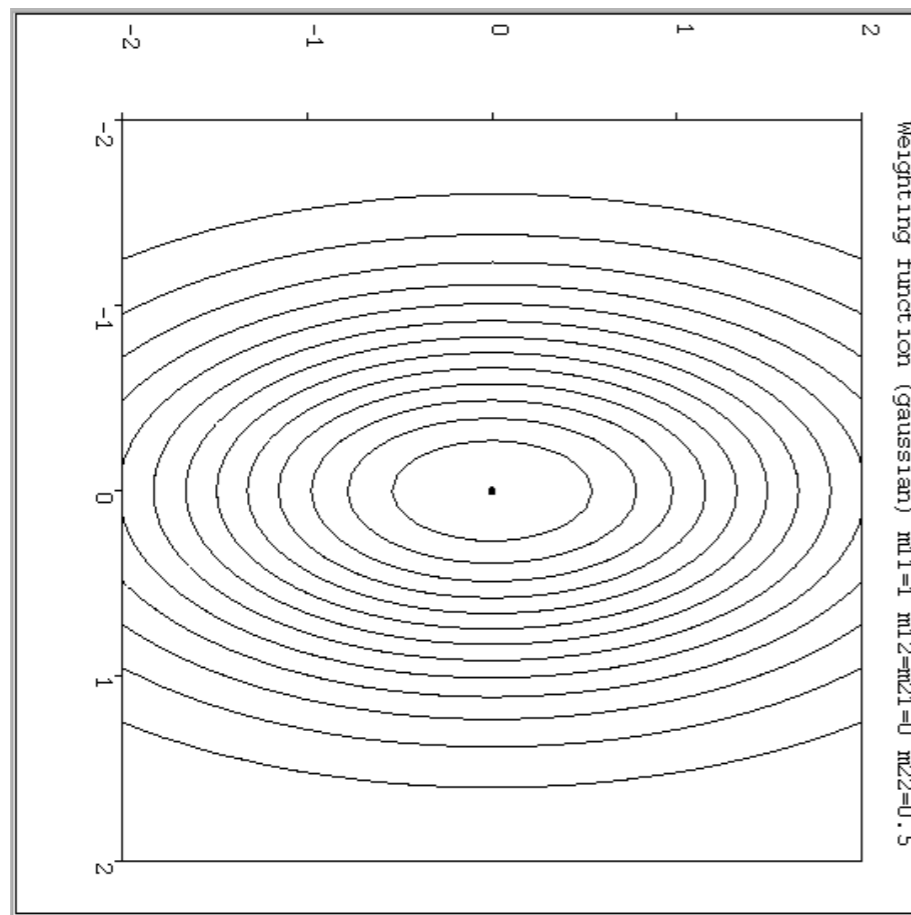
Or equivalently,

$$D(x, x') = \sqrt{(x_i - x'_i)^T A (x_i - x'_i)}$$

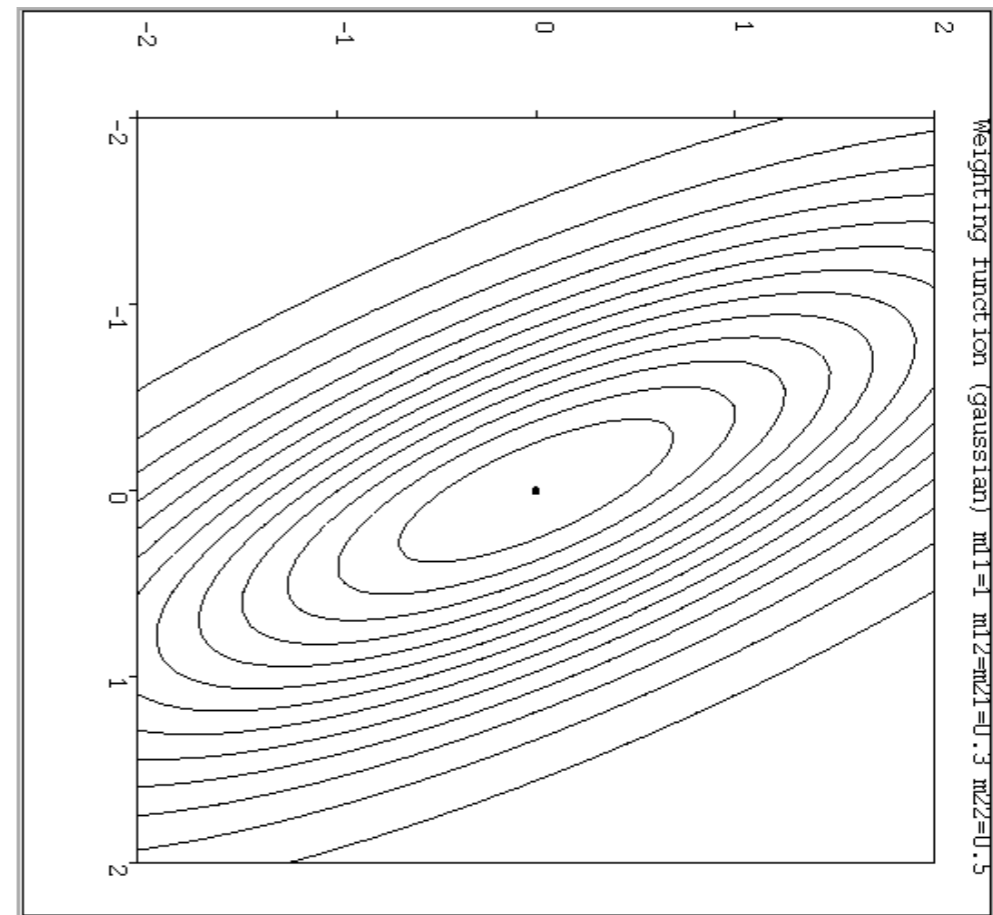
where

$$A = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

Notable distance metrics (and their level sets)

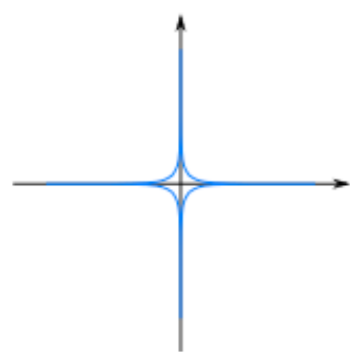


Scaled Euclidian (L_2)

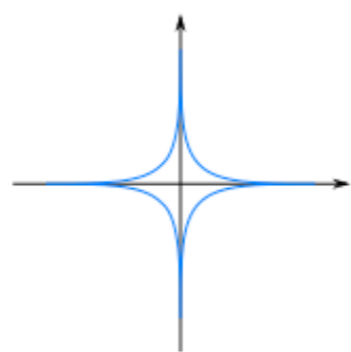


**Mahalanobis
(non-diagonal A)**

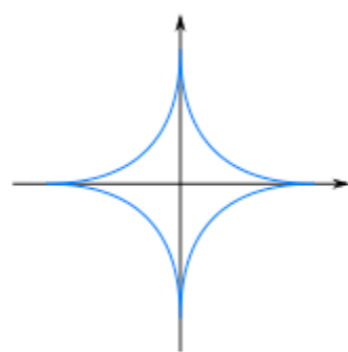
Minkowski distance



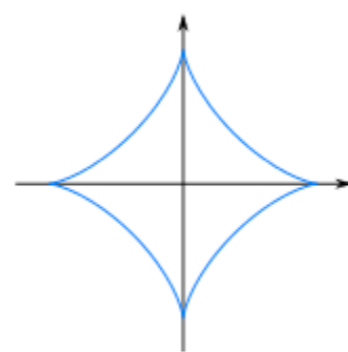
$$p = 2^{-2} \\ = 0.25$$



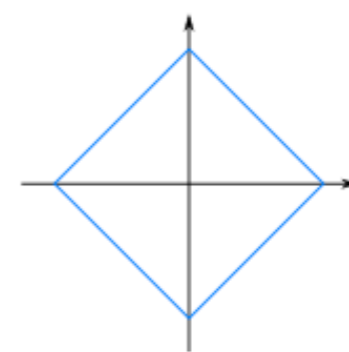
$$p = 2^{-1.5} \\ = 0.354$$



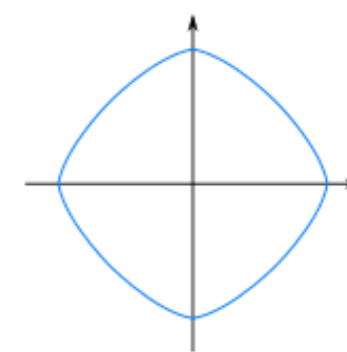
$$p = 2^{-1} \\ = 0.5$$



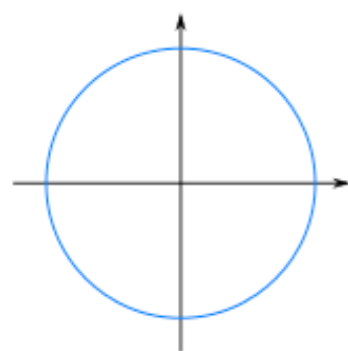
$$p = 2^{-0.5} \\ = 0.707$$



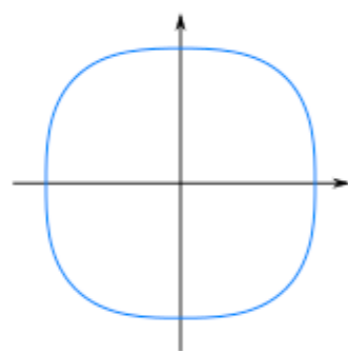
$$p = 2^0 \\ = 1$$



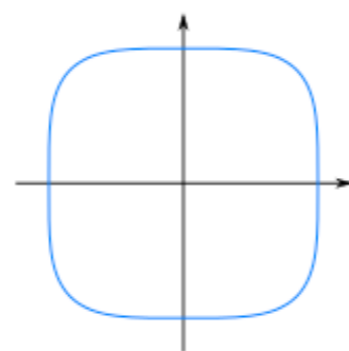
$$p = 2^{0.5} \\ = 1.414$$



$$p = 2^1 \\ = 2$$

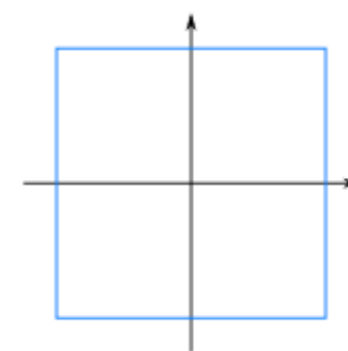


$$p = 2^{1.5} \\ = 2.828$$



$$p = 2^2 \\ = 4$$

...



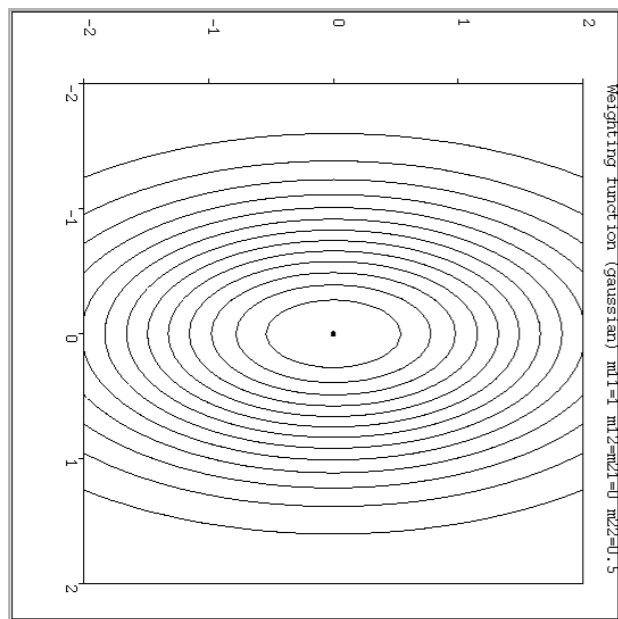
$$p = 2^\infty \\ = \infty$$

$$D = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

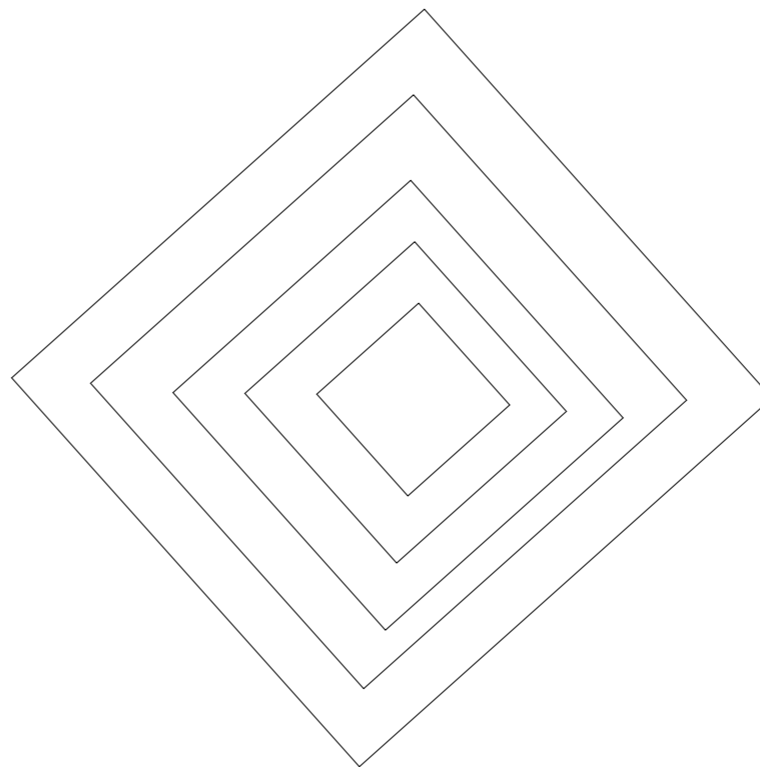
Image Credit: By Waldir (Based on File:MinkowskiCircles.svg)

[CC BY-SA 3.0 (<http://creativecommons.org/licenses/by-sa/3.0/>)], via Wikimedia Commons

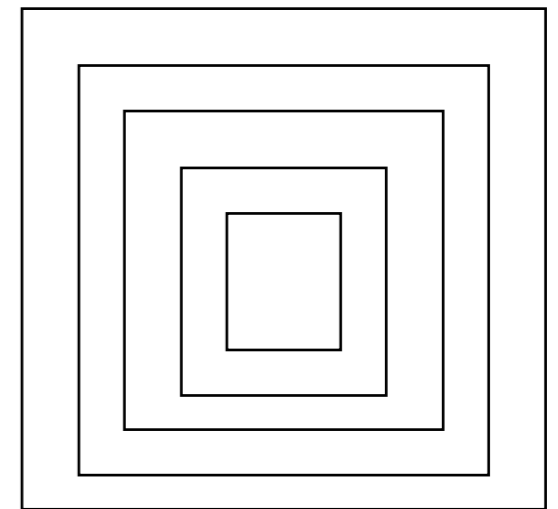
Notable distance metrics (and their level sets)



Scaled Euclidian (L₂)



L₁ norm (absolute)



L_{inf} (*max*) norm

Weighted K-NN for Regression

- Given: Training data $\{(x_1, y_1), \dots, (x_n, y_n)\}$
 - Attribute vectors: $x_i \in X$
 - Target attribute $y_i \in \mathcal{R}$
- Parameter:
 - Similarity function: $K : X \times X \rightarrow \mathcal{R}$
 - Number of nearest neighbors to consider: k
- Prediction rule
 - New example x'
 - K-nearest neighbors: k train examples with largest $K(x_i, x')$

$$h(\vec{x}') = \frac{\sum_{i \in knn(\vec{x}')} y_i K(\vec{x}_i, \vec{x}')}{\sum_{i \in knn(\vec{x}')} K(\vec{x}_i, \vec{x}')}$$

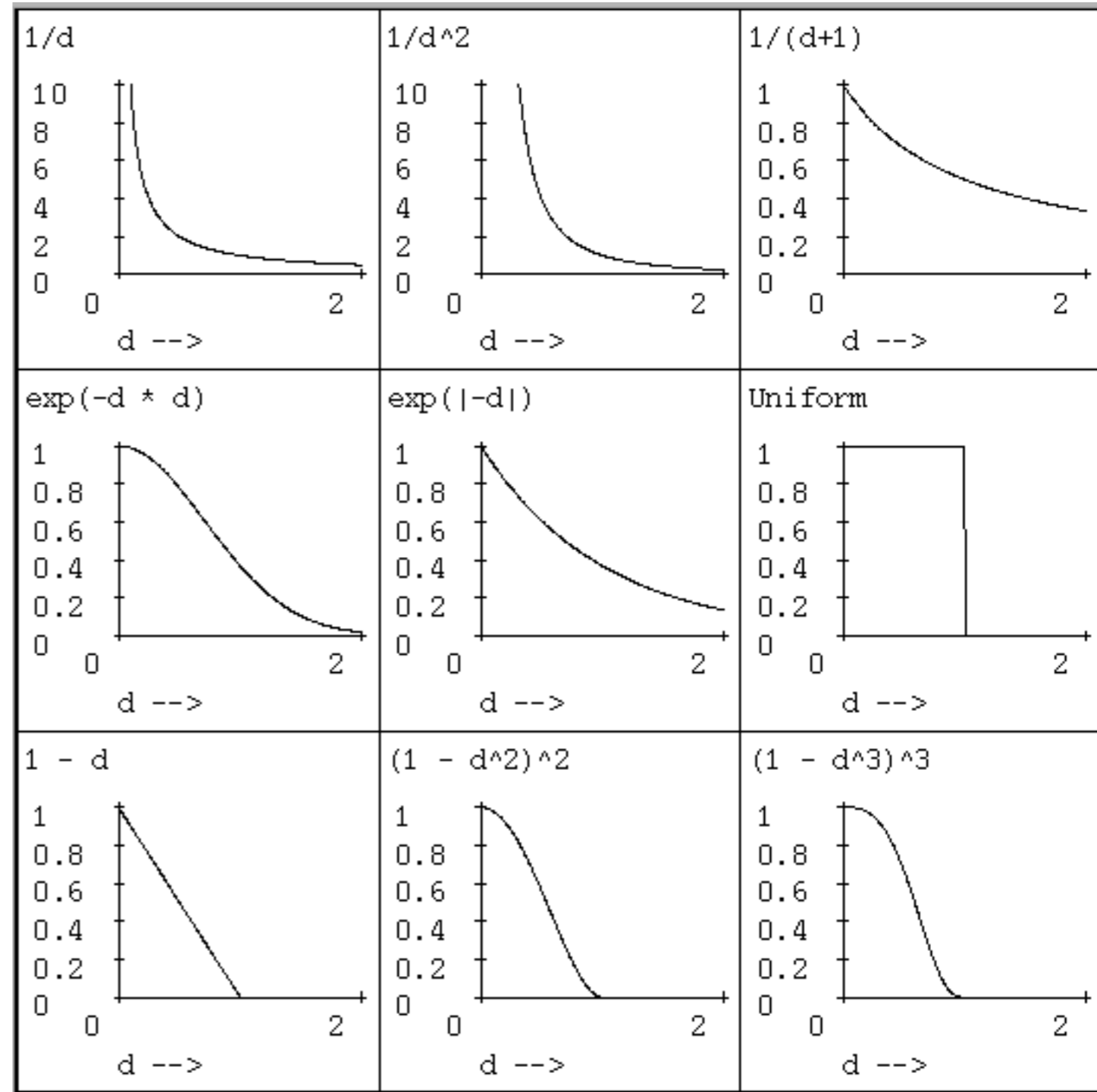
Kernel Regression/Classification

Four things make a memory based learner:

- **A distance metric**
 - Euclidean (and others)
- **How many nearby neighbors to look at?**
 - All of them
- **A weighting function (optional)**
 - $w_i = \exp(-d(x_i, query)^2 / \sigma^2)$
 - Nearby points to the query are weighted strongly, far points weakly. The σ parameter is the Kernel Width. Very important.
- **How to fit with the local points?**
 - Predict the weighted average of the outputs
predict = $\sum w_i y_i / \sum w_i$

Weighting/Kernel functions

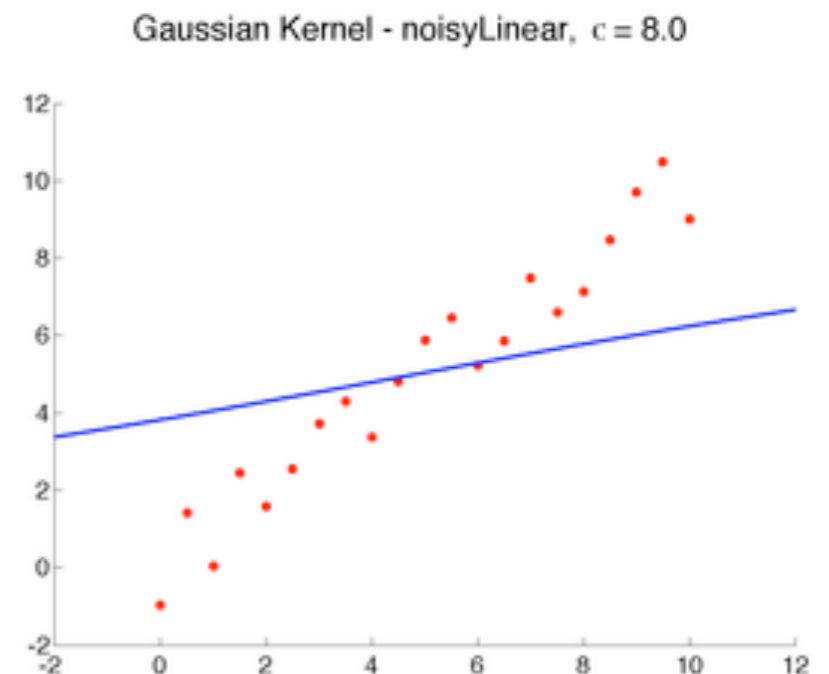
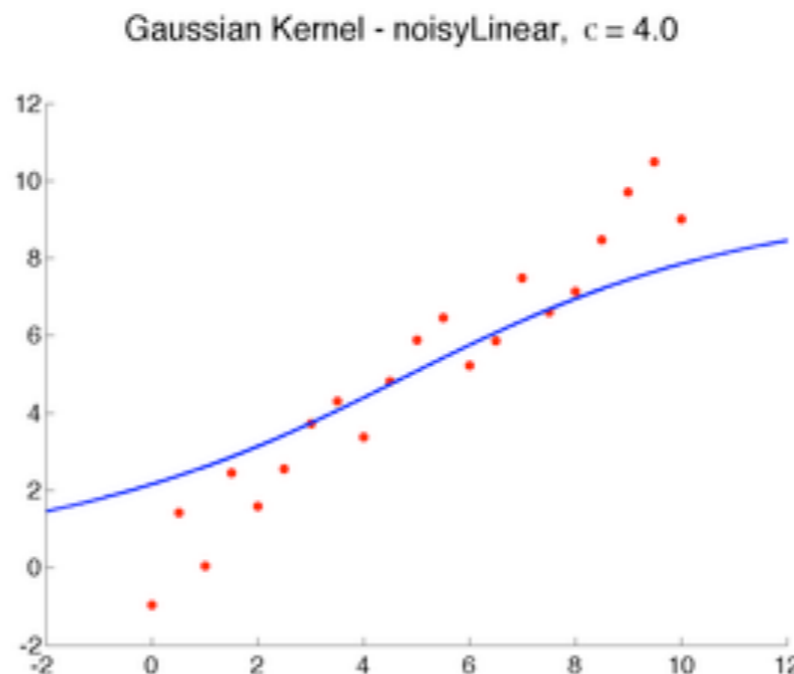
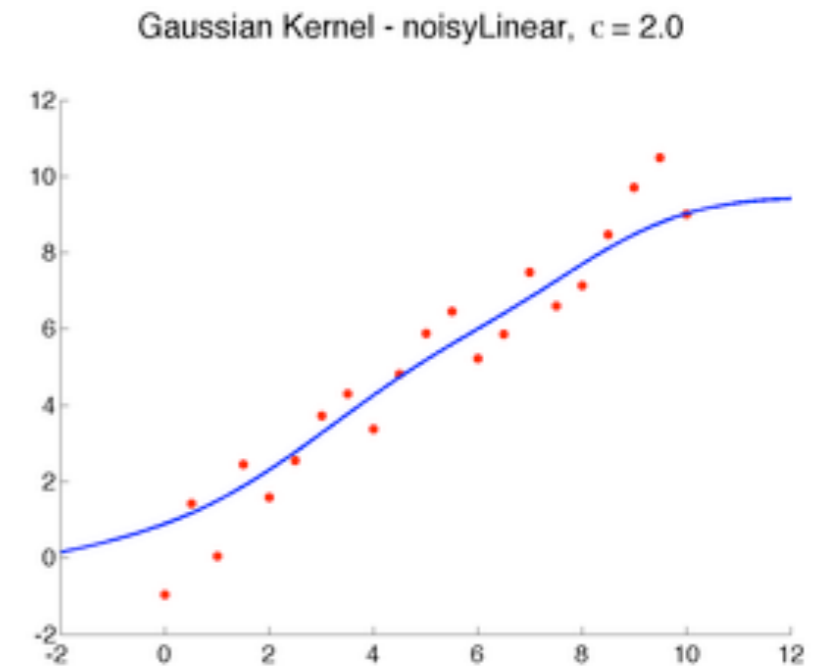
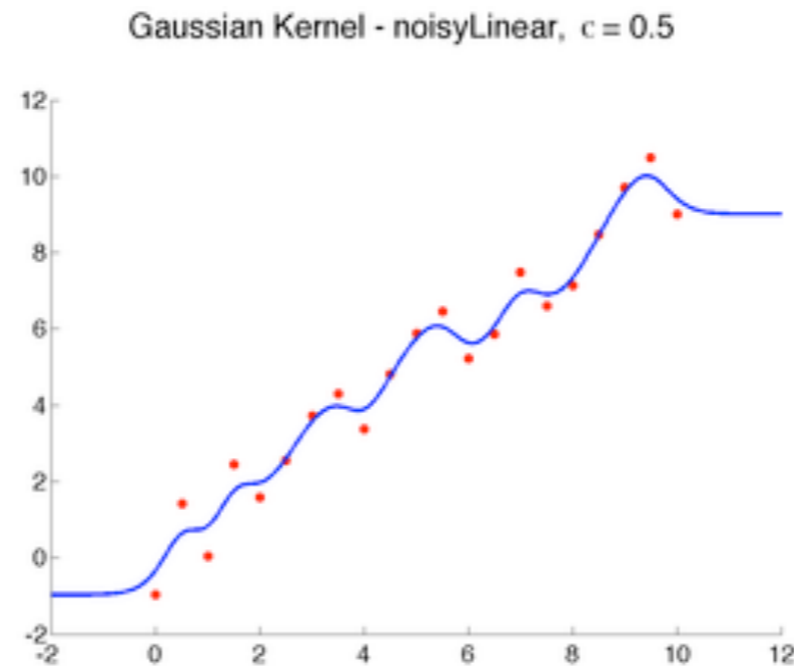
$$w_i = \exp(-d(x_i, \text{query})^2 / \sigma^2)$$



(Our examples use Gaussian)

Effect of Kernel Width

- What happens as $\sigma \rightarrow \text{inf}$?
- What happens as $\sigma \rightarrow 0$?



Problems with Instance-Based Learning

- Expensive
 - No Learning: most real work done during testing
 - For every test sample, must search through all dataset
 - very slow!
 - Must use tricks like approximate nearest neighbour search

Problems with Instance-Based Learning

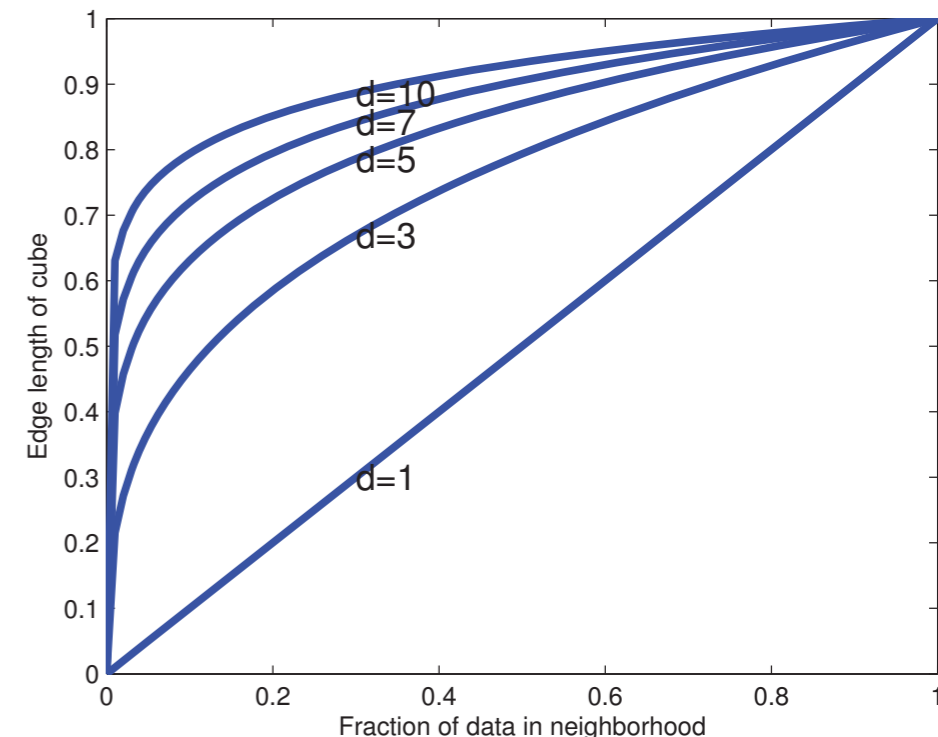
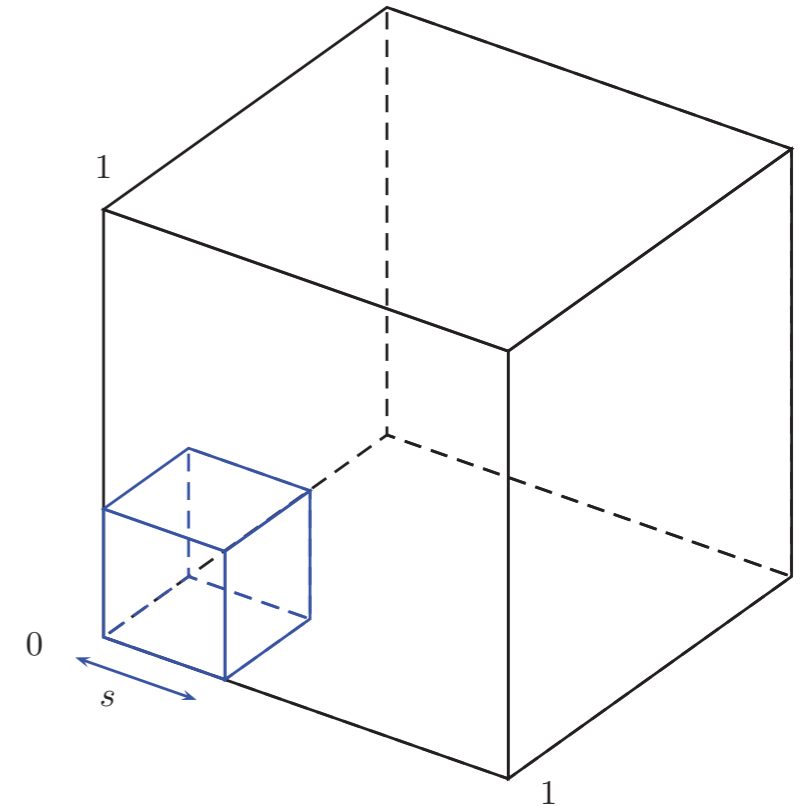
- Expensive
 - No Learning: most real work done during testing
 - For every test sample, must search through all dataset
 - very slow!
 - Must use tricks like approximate nearest neighbour search
- Doesn't work well when large number of irrelevant features
 - Distances overwhelmed by noisy features

Problems with Instance-Based Learning

- Expensive
 - No Learning: most real work done during testing
 - For every test sample, must search through all dataset
 - very slow!
 - Must use tricks like approximate nearest neighbour search
- Doesn't work well when large number of irrelevant features
 - Distances overwhelmed by noisy features
- Curse of Dimensionality
 - Distances become meaningless in high dimensions

Curse of Dimensionality

- Consider applying a KNN classifier/regressor to data where the inputs are uniformly distributed in the D -dimensional unit cube.
- Suppose we estimate the density of class labels around a test point x by “growing” a hyper-cube around x until it contains a desired fraction f of the data points.
- The expected edge length of this cube will be $e_D(f) = f^{1/D}$.
- If $D = 10$, and we want to base our estimate on 10% of the data, we have $e_{10}(0.1) = 0.8$, so we need to extend the cube 80% along each dimension around x .
- Even if we only use 1% of the data, we find $e_{10}(0.01) = 0.63$. — **no longer very local**

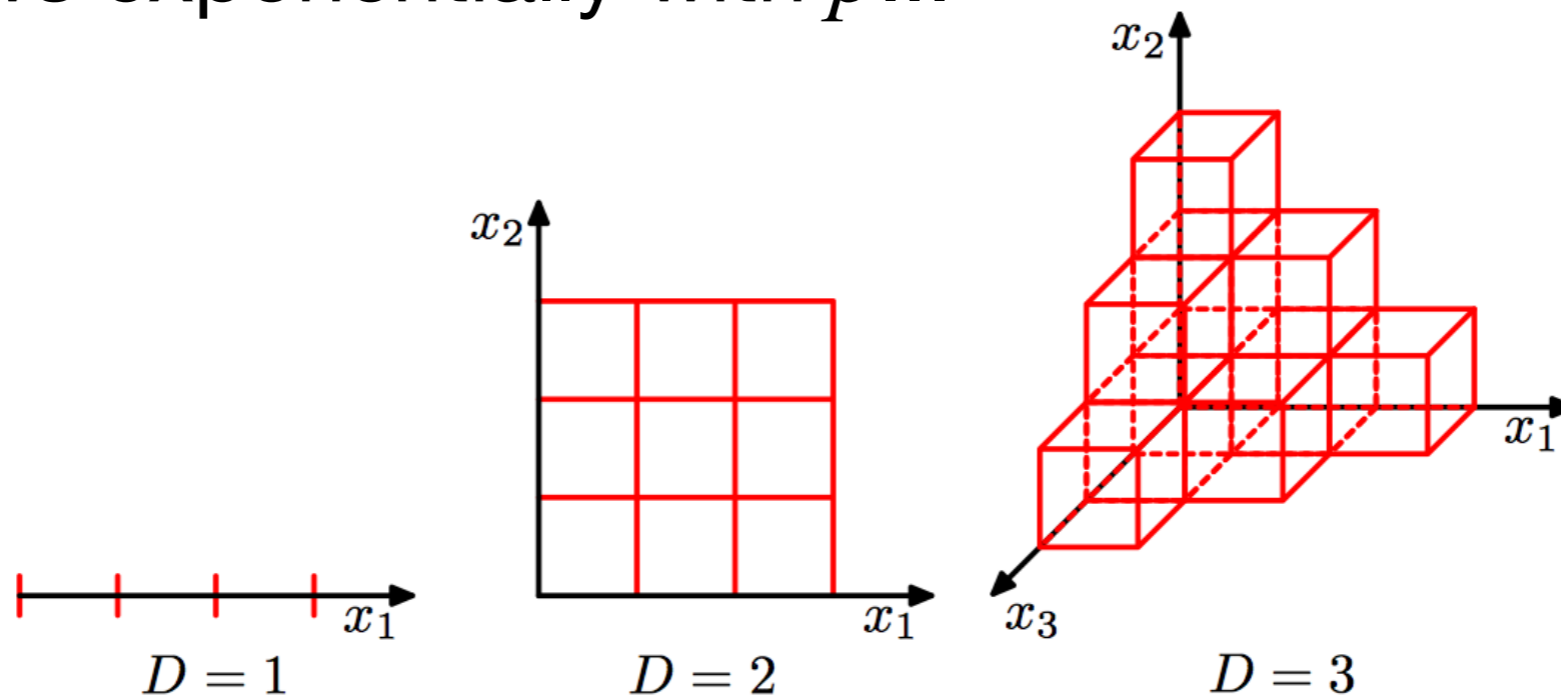


High dimensional spaces are empty

- Assume your data lives in $[0, 1]^p$. The volume of an hypercube with an edge length of $r = 0.1$ is 0.1^p
→ when p grows, it quickly becomes so small that the probability to capture points from your database becomes very very small...

Points in high dimensional spaces are isolated

- To overcome this limitation, you need a number of sample which grows exponentially with p ...



High dimensional spaces are empty

- X, Y two independent variables, with uniform distribution on $[0, 1]^p$. The mean square distance $\|X - Y\|^2$ satisfies

$$\mathbb{E}[\|X - Y\|^2] = p/6 \quad \text{and} \quad \text{Std}[\|X - Y\|^2] \simeq 0.2\sqrt{p}$$

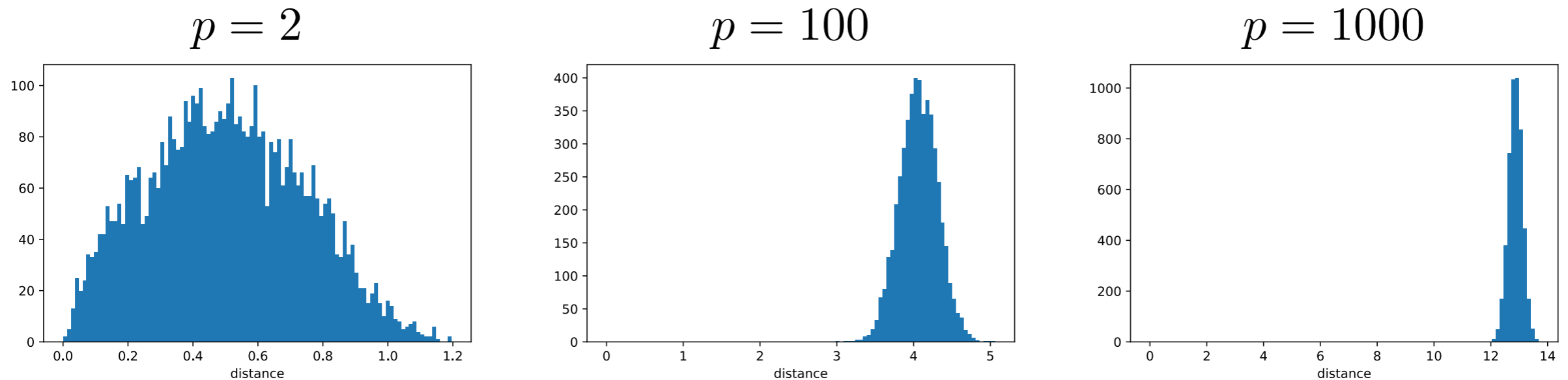


Figure: Histograms of pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$

The notion of nearest neighbors vanishes.

Parametric vs Non-parametric Models

- Does the capacity (size of hypothesis class) grow with size of training data?
 - Yes = Non-parametric Models
 - No = Parametric Models

Poll Time!

- In k-NN, the curse of dimensionality will likely result in the overfitting problem.

Which of the following option(s) would you consider to handle this problem?

(a) Dimensionality reduction

(b) Feature selection

- a
- b
- a and b
- None of these

Ways to avoid the curse of dimensionality

- **Dimension reduction:**
 - the problem comes from that p is too large,
 - therefore, reduce the data dimension to $d \ll p$,
 - such that the curse of dimensionality vanishes!
- **Regularization:**
 - The problem comes from that parameter estimates are unstable,
 - therefore, regularize these estimates,
 - such that the parameter are correctly estimated!
- **Parsimonious models:**
 - the problem comes from that the number of parameters to estimate is too large,
 - therefore, make restrictive assumptions on the model,
 - such that the number of parameters to estimate becomes more “decent”!

Next Lecture:
Linear Regression,
Least Squares Optimization,
Model complexity, Regularization