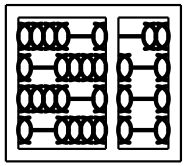


Universidade Estadual de Campinas - UNICAMP
Instituto de Computação - IC

MO444/MC886
Pattern Recognition and Machine Learning

Introduction, problems, data, tools

Prof. Anderson Rocha
Largely based on several materials and slides from other researchers

Campinas, August 12, 2015

Class Presentation

1. 4 credits (60 hrs/class);
2. 1 written exam
3. 4 individual practical assignments
4. 1 larger project

Notes

1. The slides herein are largely based on materials collected from other researchers. This class specifically uses slides prepared by **Prof. Alexander Ihler**, UC/Irvine.

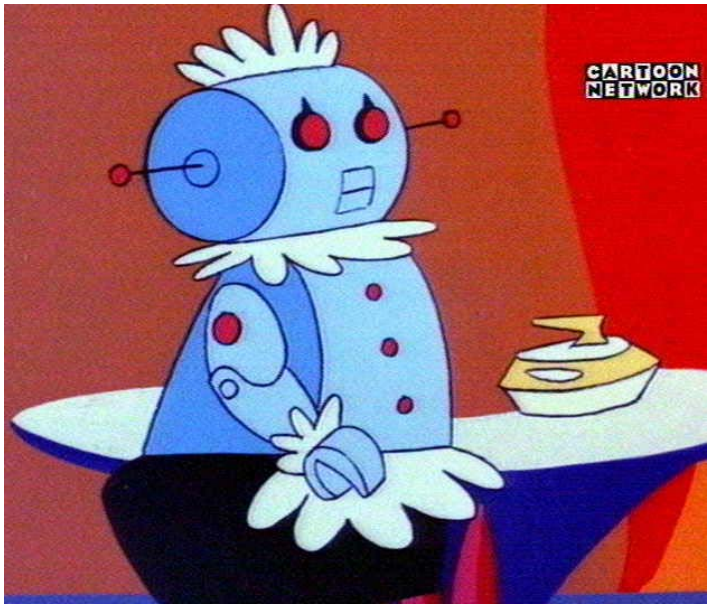
What is machine learning?

- The ability of a machine to improve its performance based on previous results
- Initially, a subspecialty of artificial intelligence
- What is “learning from experience”?
  - Observe the world (data)
  - Change our behavior accordingly
- Typical examples
  - Predicting outcomes
  - Explaining observations
  - Finding “interesting” or unusual data

Examples of machine learning

- Commercial
  - Spam filtering
  - Fraud detection (credit cards, &c)
  - Stock market prediction & trading
  - Advertisements and “suggestions”
- Security
  - Social network analysis
  - Signature & biometric recognition
  - Surveillance
- Information management & retrieval
  - Intelligent search
  - Machine translation
  - Voice to text
- Games
  - Checkers, chess, go …
  - Robo-soccer

What is AI?


CARTOON NETWORK

?

=


Roomba
iRobot

?

=


Now Playing on TiVo
The Apprentice Thu 3/25
Alias Sun 3/28
Survivor Mon 3/29
Stanford vs. Arizona Tue 3/30
The Real World Tue 3/30
American Idol (2) Wed 3/31
Six Feet Under Sun 4/4
Emeril Live (4) Wed 4/7
NBC
CBS
FOX
FOX

History of AI

Some successes:

Chess (Deep Blue v. Kasparov)

RoboCup

Darpa GC (Stanley)

What is ML?

- Less than the whole of AI?
  - Just one part of intelligence…
- More than just AI?
  - Applicable to many “practical” problems
  - Making sense of data automatically
  - Found in
    - Data mining & information retrieval
    - Computational biology
    - Signal processing
    - Image processing & computer vision
    - Data compression and coding

Why is this so important?

- Data available at unprecedented scales
  - Petabyte scale computing…
- Impossible for humans to deal with this information overflow
- True for a wide variety of areas
  - Web pages
  - Images
- Imagine the resources required to
  - look at every image in Flickr and categorize it
  - check every inch of Google earth for changes
  - look through all webpages for the interesting ones

Types of learning

- Supervised learning
  - Specific target signal to predict
  - Training data have known target values
- Unsupervised learning
  - No given target value; looking for structure
  - Ex: clustering, dimensionality reduction
- Semi-supervised learning
  - Some labeled data, some unlabeled
  - Ex: images on the web
  - Try to use unlabeled data to help
- Reinforcement learning
  - Reward signal, possibly delayed
  - Ex: learning to drive, play a game, etc.

Classification

- Discriminating between two (or more) types of data
- Example: Spam filtering

Bad

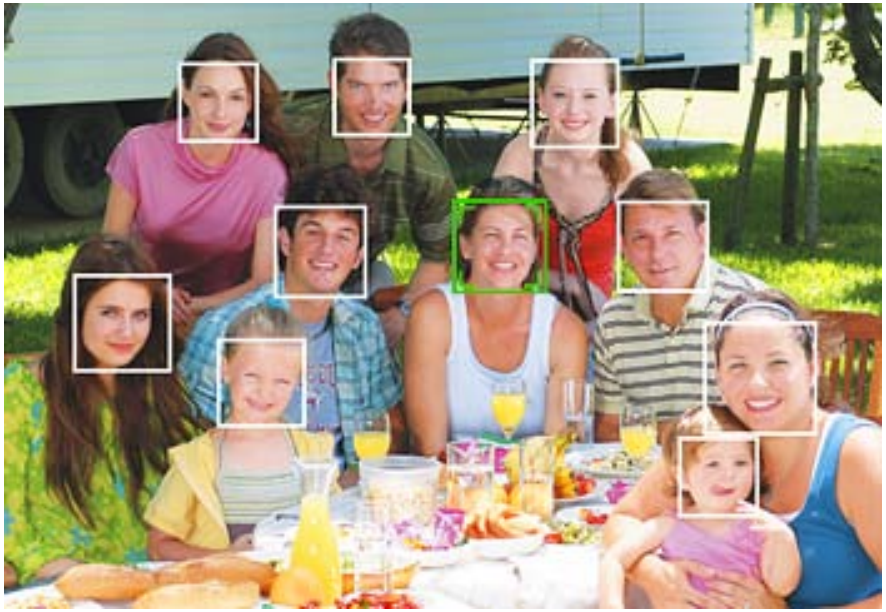
Cures fast and effective! - Canadian *** Pharmacy #1 Internet Inline Drugstore Viagra Our price $1.15 Cialis Our price $1.99 …

Good

Interested in your research on graphical models - Dear Prof. Ihler, I have read some of your papers on probabilistic graphical models. Because I …

Classification

- Example: face detection



Regression

- Based on past history, predict future outcomes

Wall Street

Netflix

Data Mining & Understanding

- Massive volumes of data available
  - Webpages, Google books, …
  - Too large to hand-curate or organize

- How does Google decide the “most relevant” documents?
- How can we look for text documents “about” law, medicine, etc?
- What makes a document “similar”?

- Gets even harder for images, video, …

Slides by Prof. Alexander Ihler, UC/Irvine

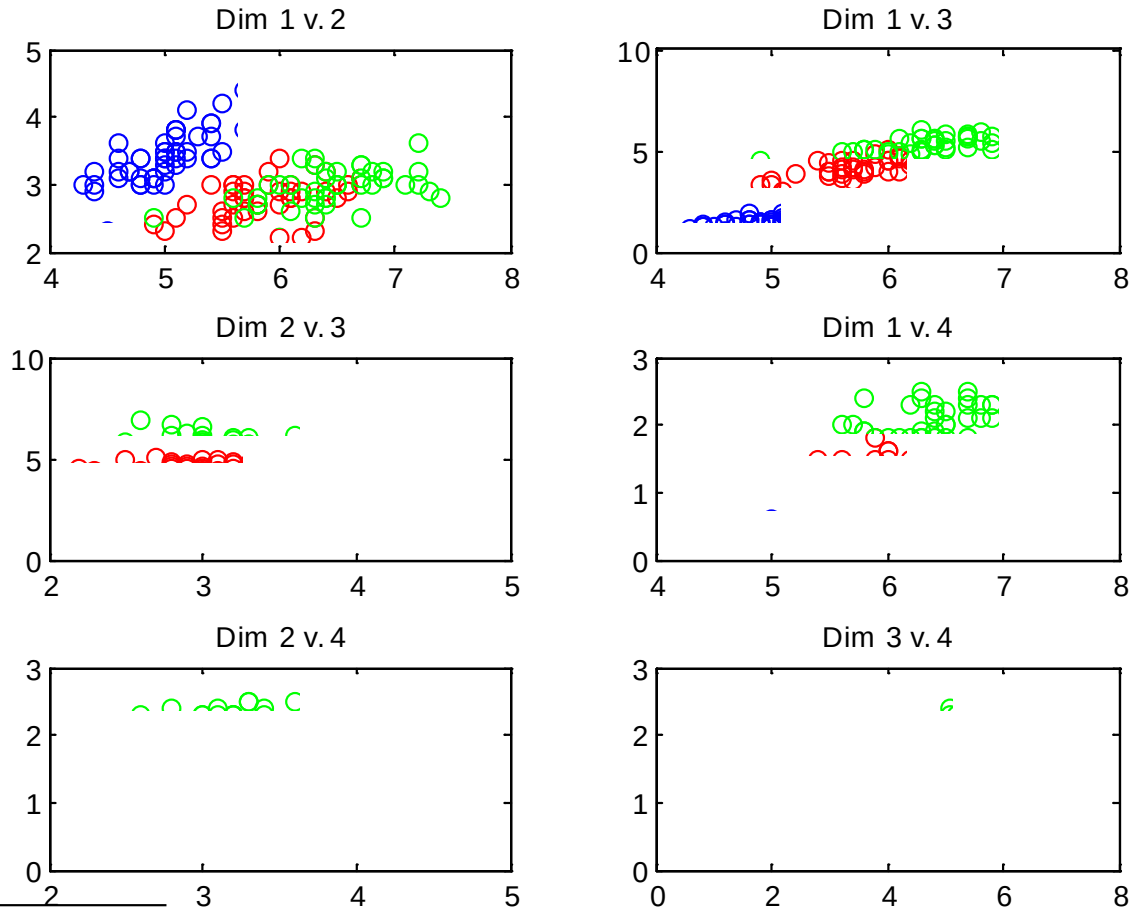
Clustering

- UCI Iris data set

Plot each pair of dimensions

Colors = classes

Data form coherent groups



Collaborative filtering (Amazon)

the ONION

HOME | VIDEO | RADIO | SPORTS | POLITICS | WORLD | ECONOMY | SCI/TECH | ENTE

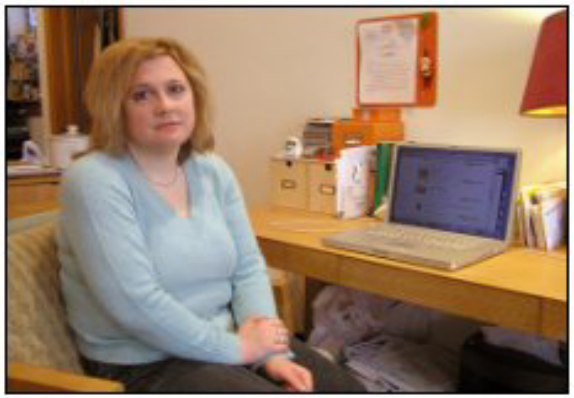
SCIENCE & TECHNOLOGY

Amazon.com Recommendations Understand Area Woman Better Than Husband

JANUARY 9, 2007 | ISSUE 43•02

SANDUSKY, OH—Area resident Pamela Meyers was delighted to receive yet another thoughtful CD recommendation from Amazon.com Friday, confirming that the online retail giant has a more thorough, individualized, and nuanced understanding of Meyers' taste than the man who occasionally claims to love her, husband Dean Meyers.

ENLARGE IMAGE



Meyers said she was pleasantly surprised to receive three e-mails from Amazon today alone.

"To come home from a long day at work and see the message about the new Norah Jones album waiting for me, it just made my week," said Meyers, 36, who claimed she was touched that the company paid such attention to her. "It feels nice to be noticed once in a while, you know?"

Amazon, which has been tracking Meyers' purchases since she first used the site to order *Football For Dummies* in preparation for attending the 2004 Citrus Bowl as part of her husband's 10th wedding anniversary [illegible] has shown impressive accuracy at recommending books, movies, music, and even clothing that perfectly match Meyers' tastes. While the powerful algorithms that power Amazon's

ARTICLE TOOLS

DIGG

FACEBOOK

STUMBLEUPON

TWITTER

REDDIT

EMAIL

PRINT

RELATED ARTICLES

New Linens-N-Shit Opens

LDA and Text Data

Court Allows Scientists to Work at NASA Until Trial Over Background Checks

By JOHN SCHWARTZ
New York Times: January 12, 2008

A group of scientists working at NASA's Jet Propulsion Laboratory won a round in federal court on Friday in their challenge to a Bush administration requirement that they submit to extensive background checks or face losing their jobs.

The United States Court of Appeals for the Ninth Circuit, in California, issued an opinion allowing the scientists to continue working until the question of their privacy challenge can be addressed at a full trial.

They had sued the administration over a new domestic security requirement that all contract workers at the laboratory, which is run jointly by NASA and the California Institute of Technology, undergo background checks and identification requirements. The 26 scientists and engineers filing the suit, whose jobs the government classifies as "low risk," argued that the background checks, which could include information on finances, psychiatric care and sexual practices, constituted an unacceptable invasion of their privacy.

The government, which is requiring the upgraded security review at every federal agency, argued that the contract employees be held to the same standard.

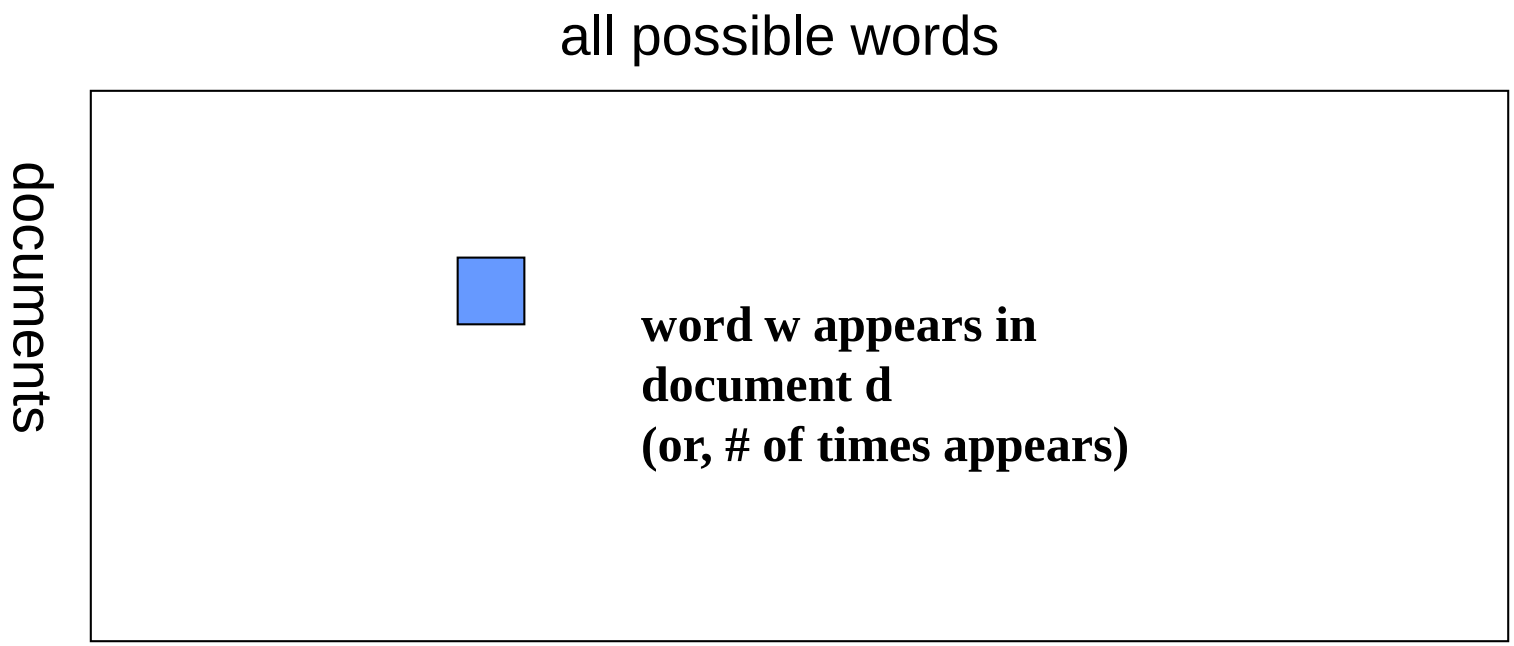
A lower court had denied the scientists' request for an injunction to block the background checks, but in the opinion released Friday, the court of appeals reversed that decision and sent the case back to the lower court...

words related to Legal/ Law

words related to Security/ Privacy

Text data as sparse matrices

- Can represent documents similarly
  - Sparse collection of document word counts



Tools for Machine Learning

- Optimization
  - Use flexible, parameterized models to describe data
  - Use optimization algorithms to fit the models to data
- Probability and Statistics
  - Allows computing with / about uncertainty
  - Combine multiple sources of (uncertain) information
  - Search for “simple” explanations
- Linear algebra
  - Data often represented as matrices;
- Information theory, graph theory, physics, …

Machine learning as statistics

- Key to learning is data
- Goal: find and exploit patterns in data

Ingredients

- Data
  - What kinds of data do we have?
- Prior assumptions
  - What do we know about the problem off the bat?
- Representation
  - How should we represent the data?
- Model / hypothesis space
  - What types of explanations should we consider?
- Feedback / learning signal
  - What signals do we have?
- Learning algorithm
  - How do we update the model given feedback?
- Evaluation
  - How well do we do?

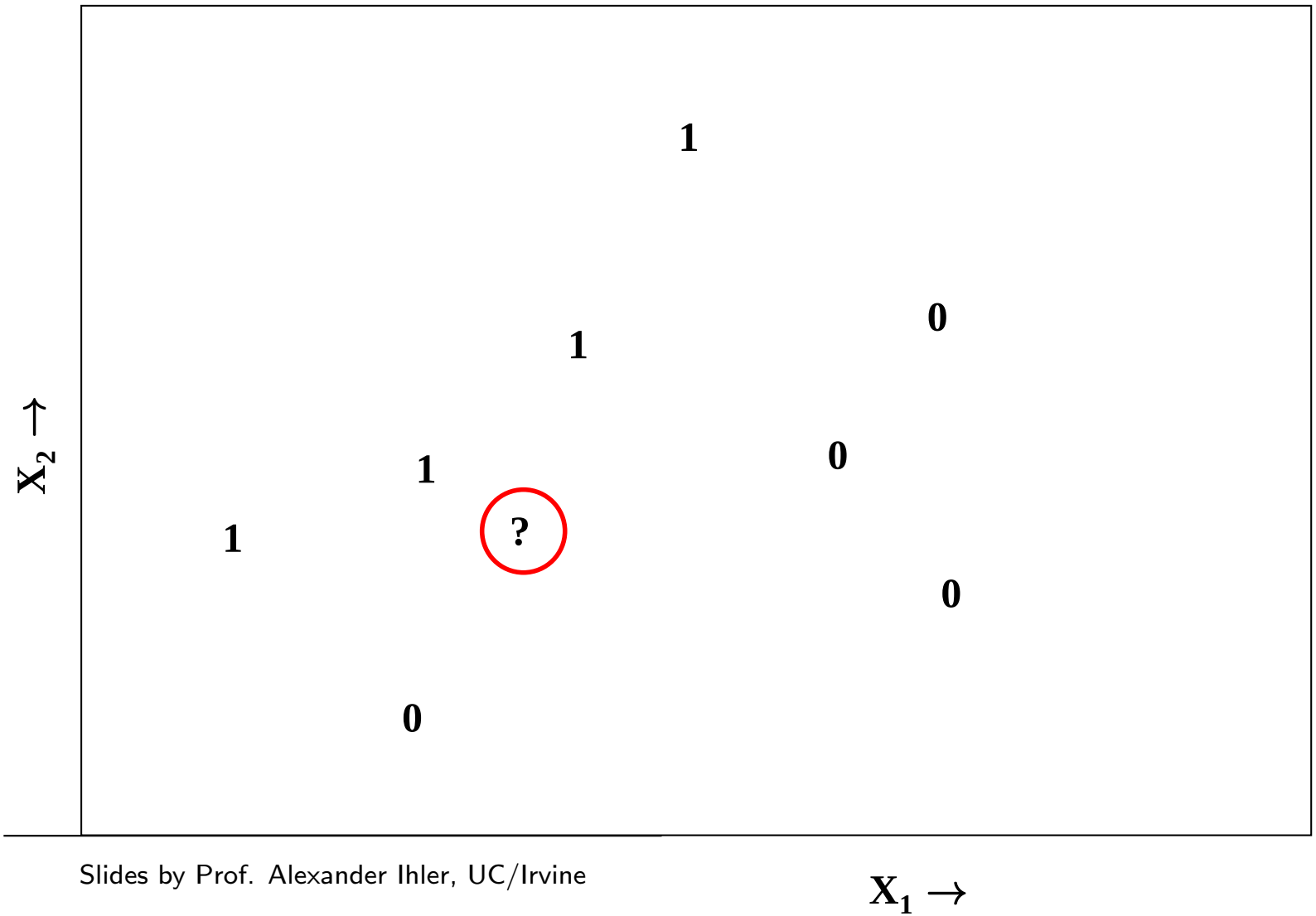
(from Max Welling's slides)

Ingredients

- Data
  - What kinds of data do we have?
- Prior assumptions
  - What do we know about the problem off the bat?
- Representation
  - How should we represent the data?
- Model / hypothesis space
  - What types of explanations should we consider?
- Feedback / learning signal
  - What signals do we have?
- Learning algorithm
  - How do we update the model given feedback?
- Evaluation
  - How well do we do?

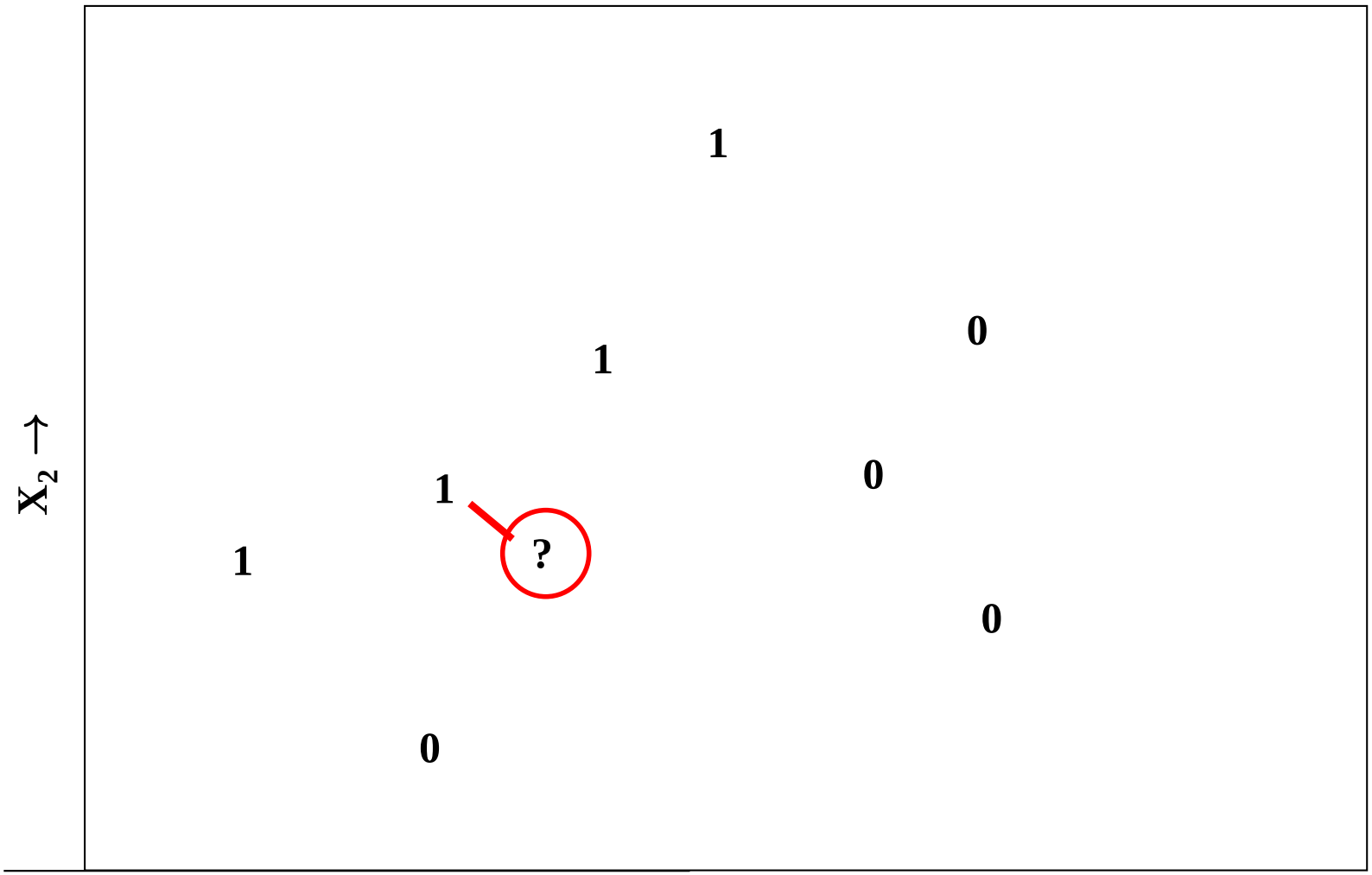
(from Max Welling's slides)

Nearest neighbor classifier



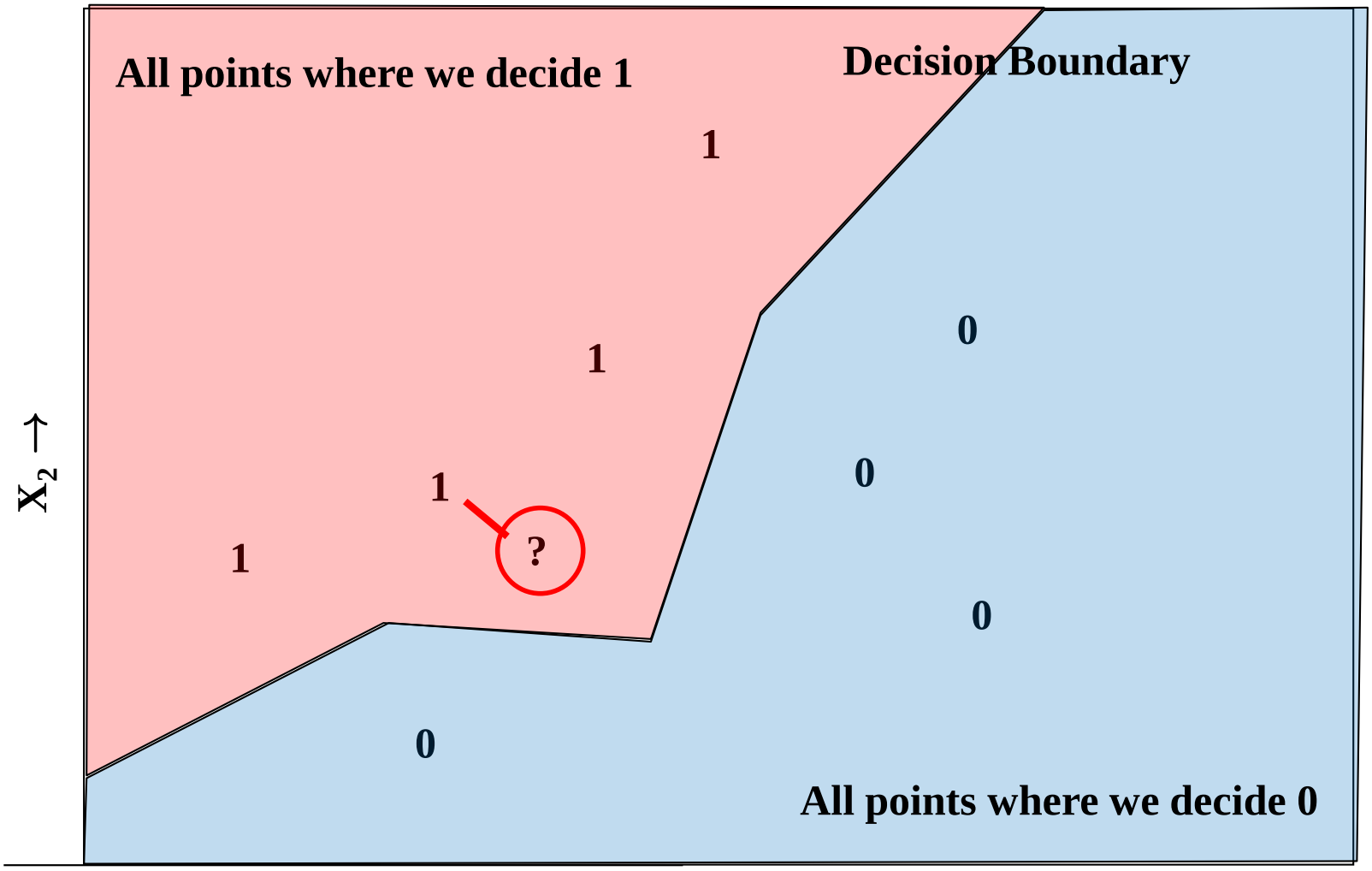
Slides by Prof. Alexander Ihler, UC/Irvine

Nearest neighbor classifier



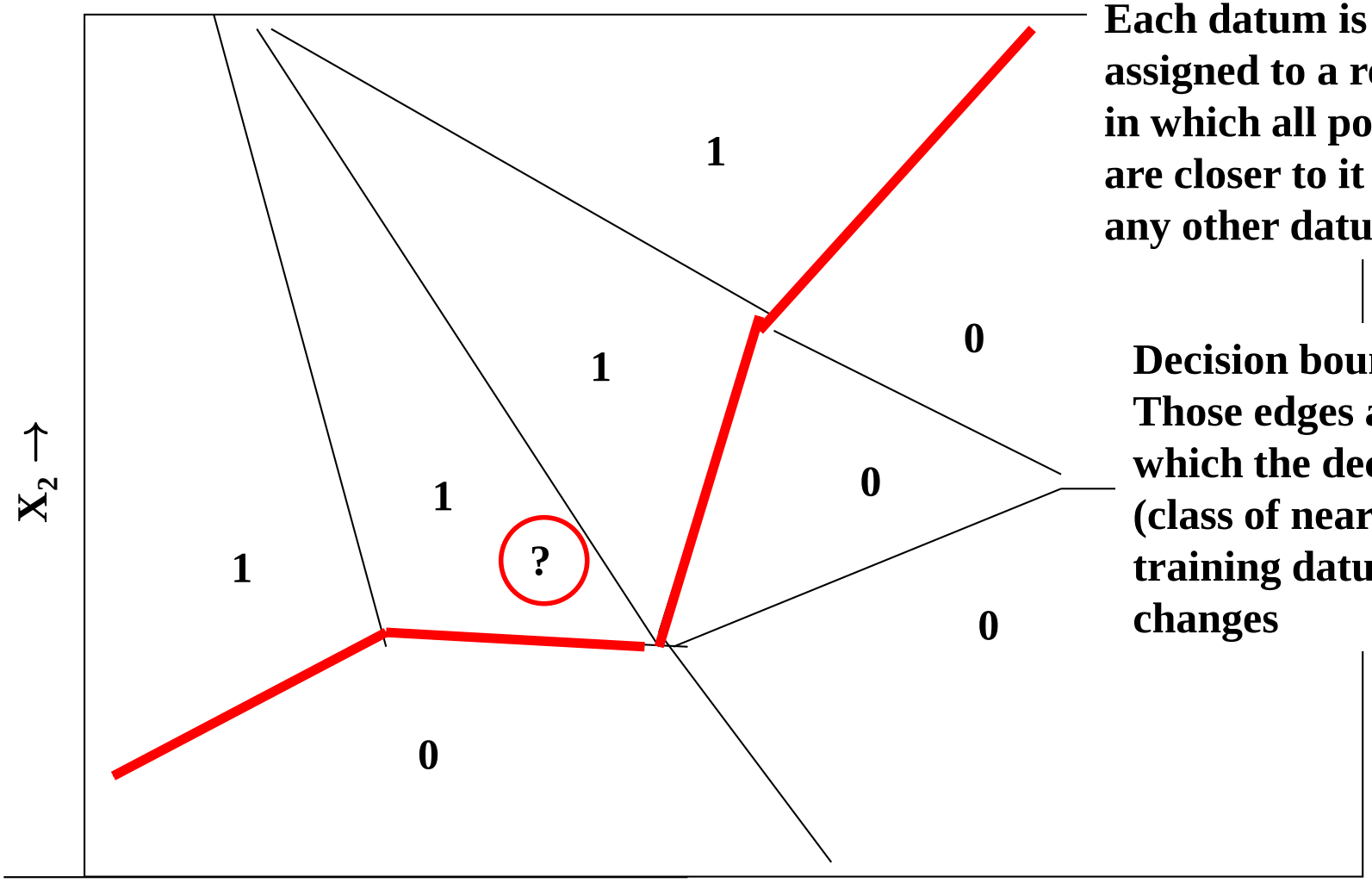
Slides by Prof. Alexander Ihler, UC/Irvine

Nearest neighbor classifier



Slides by Prof. Alexander Ihler, UC/Irvine

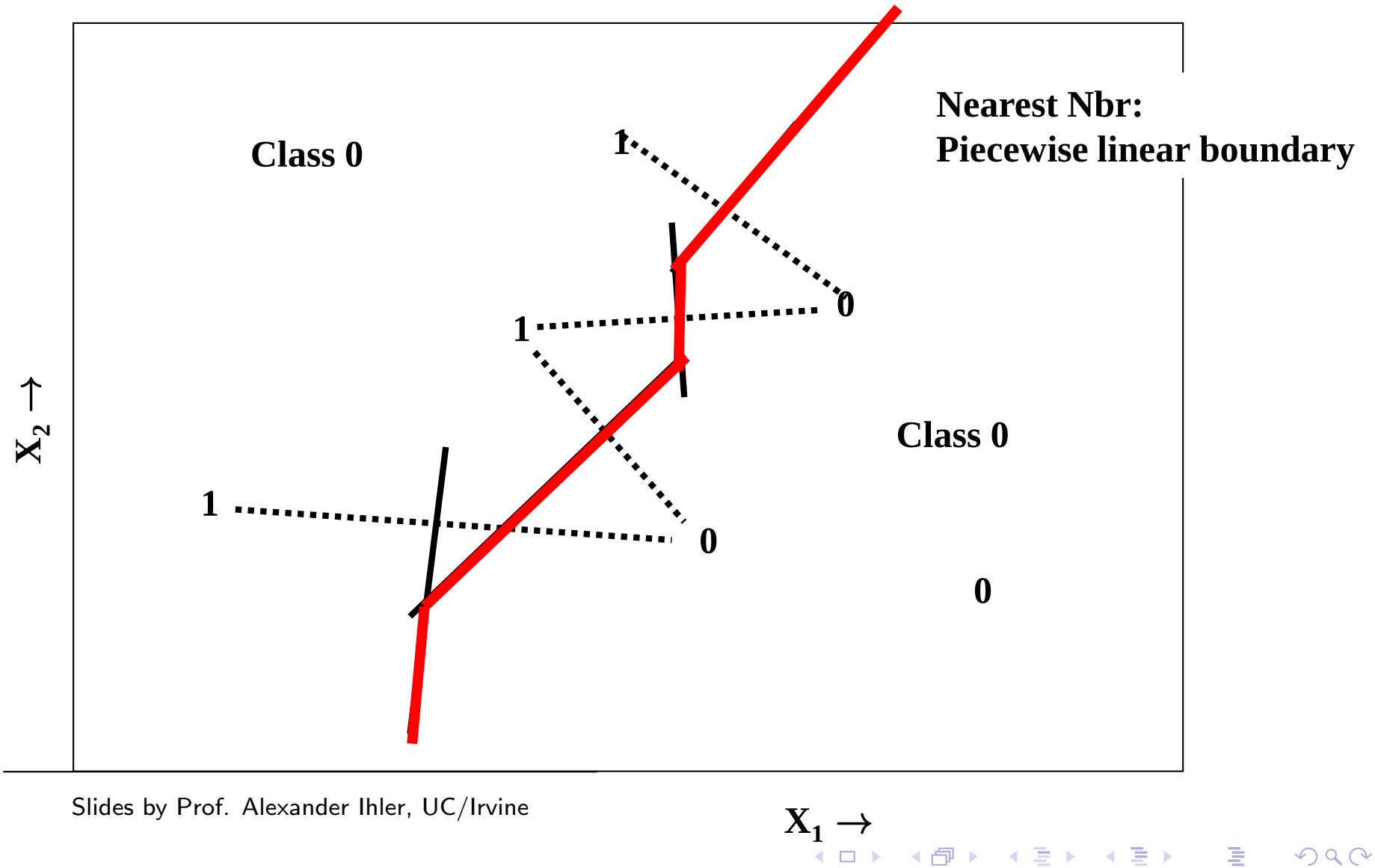
Nearest neighbor classifier



Each datum is assigned to a region, in which all points are closer to it than any other datum

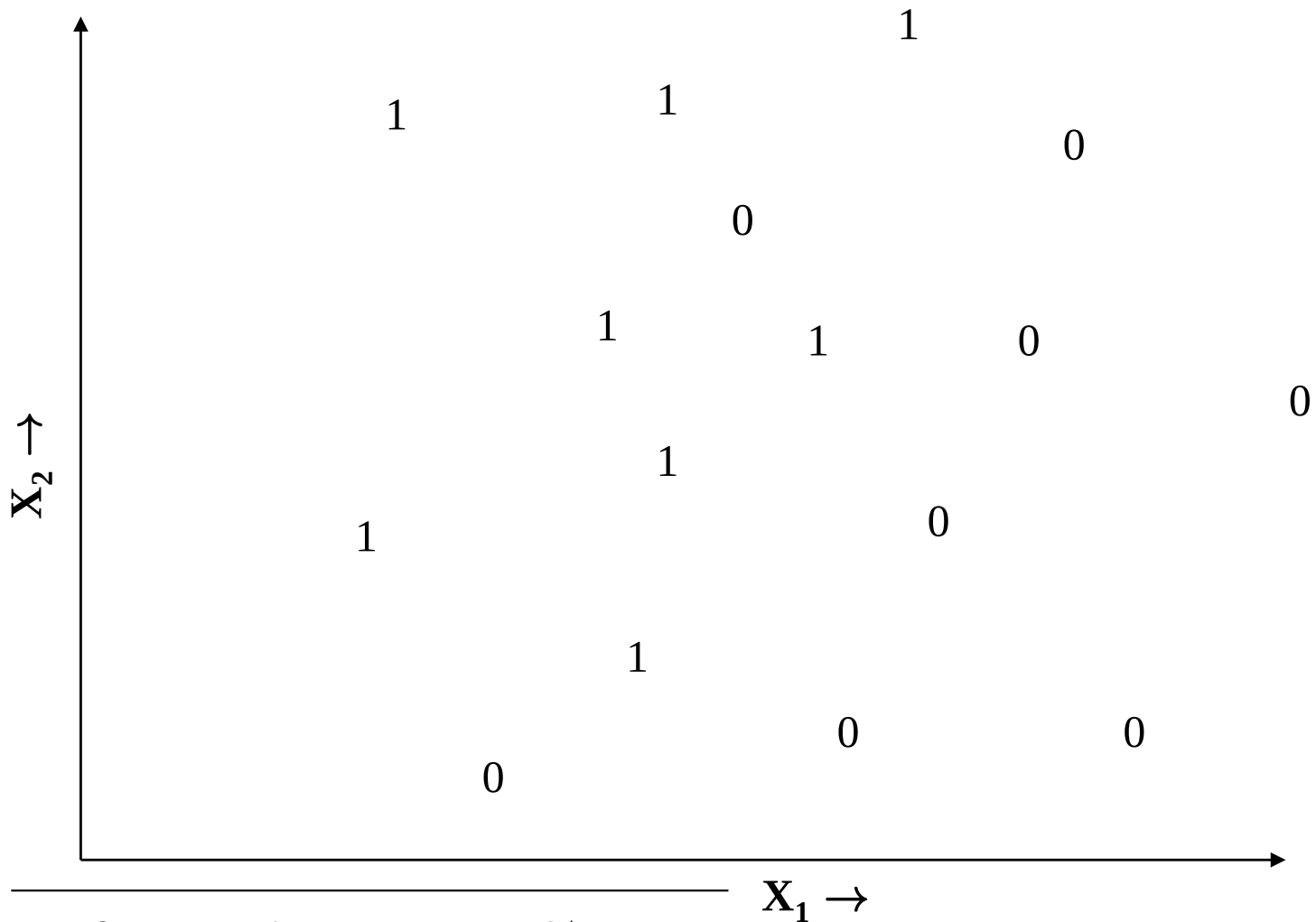
Decision boundary: Those edges across which the decision (class of nearest training datum) changes

Nearest neighbor classifier

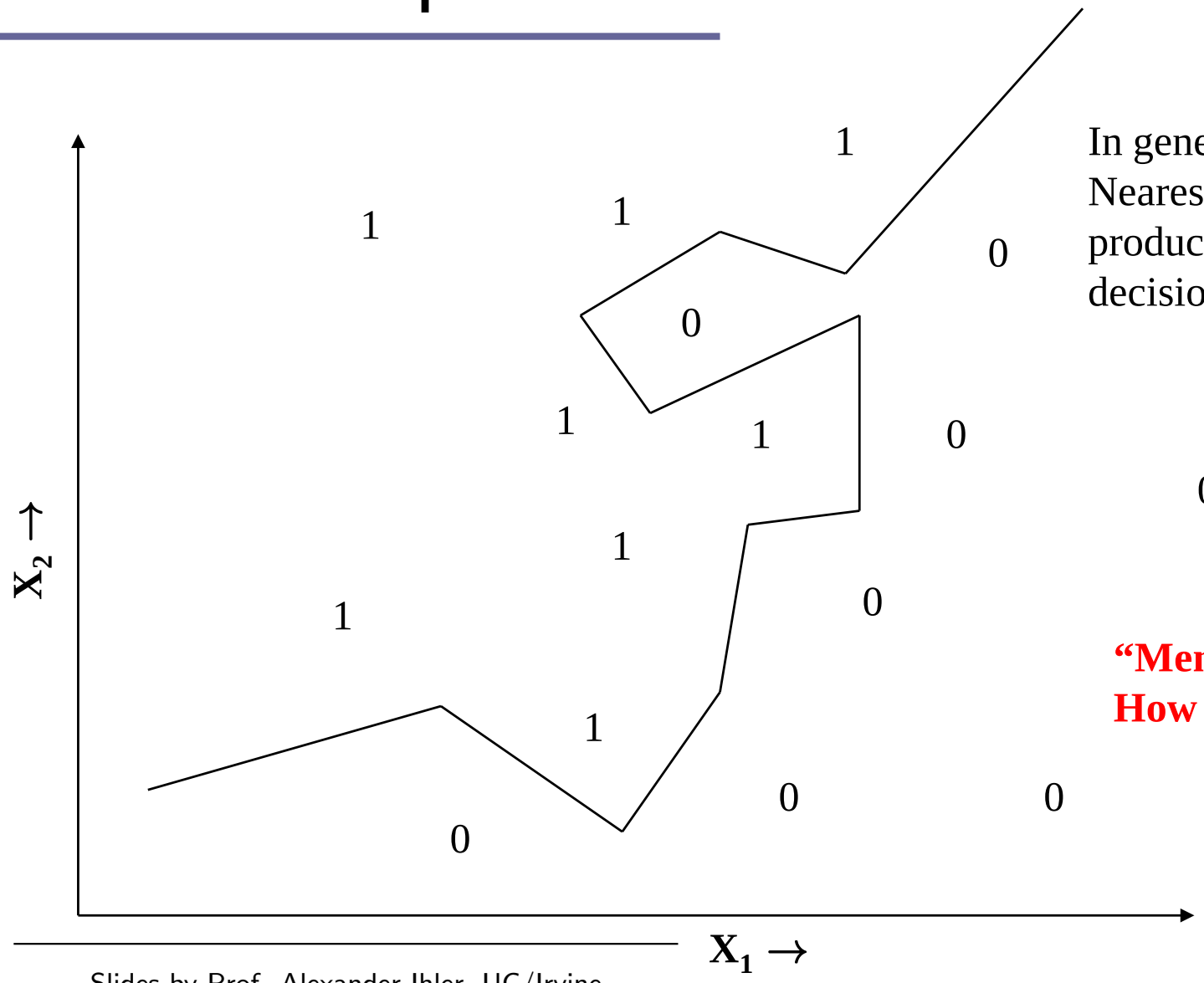


Slides by Prof. Alexander Ihler, UC/Irvine

More Data Points



More Complex Decision Boundary

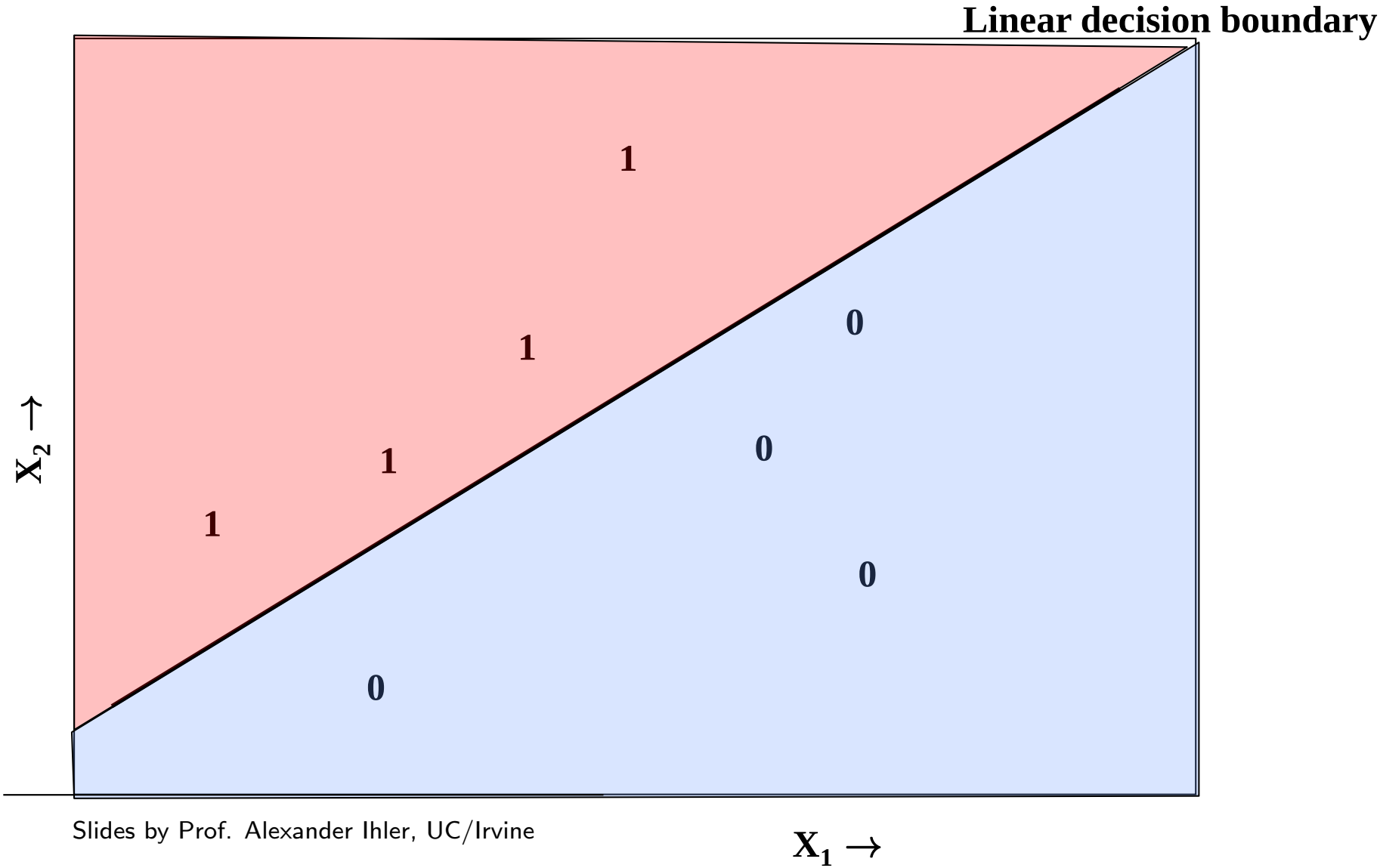


In general:
Nearest-neighbor classifier produces piecewise linear decision boundaries

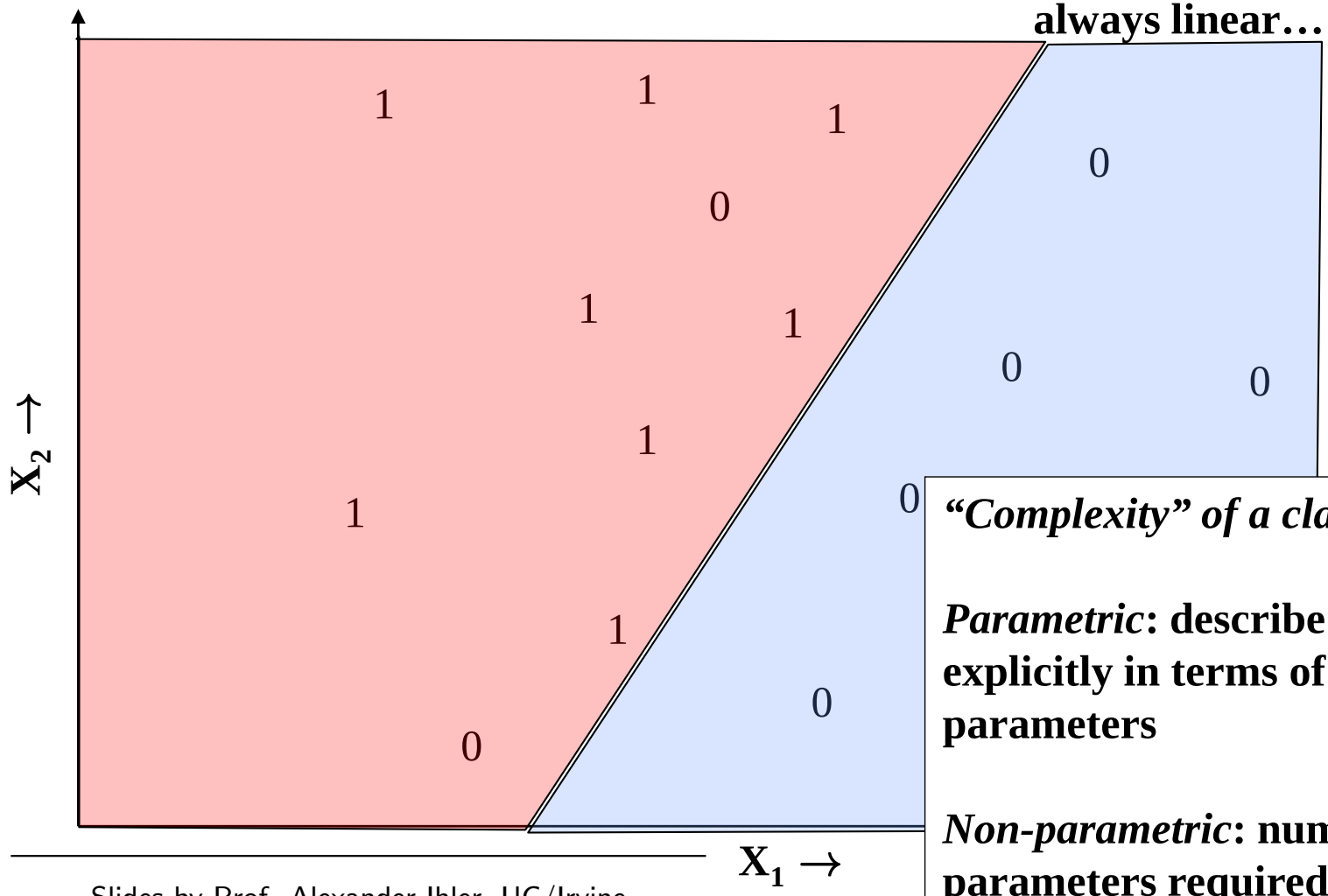
"Memorization"?
How is this learning?

Slides by Prof. Alexander Ihler, UC/Irvine

Contrast: linear classifier



More Data Points?



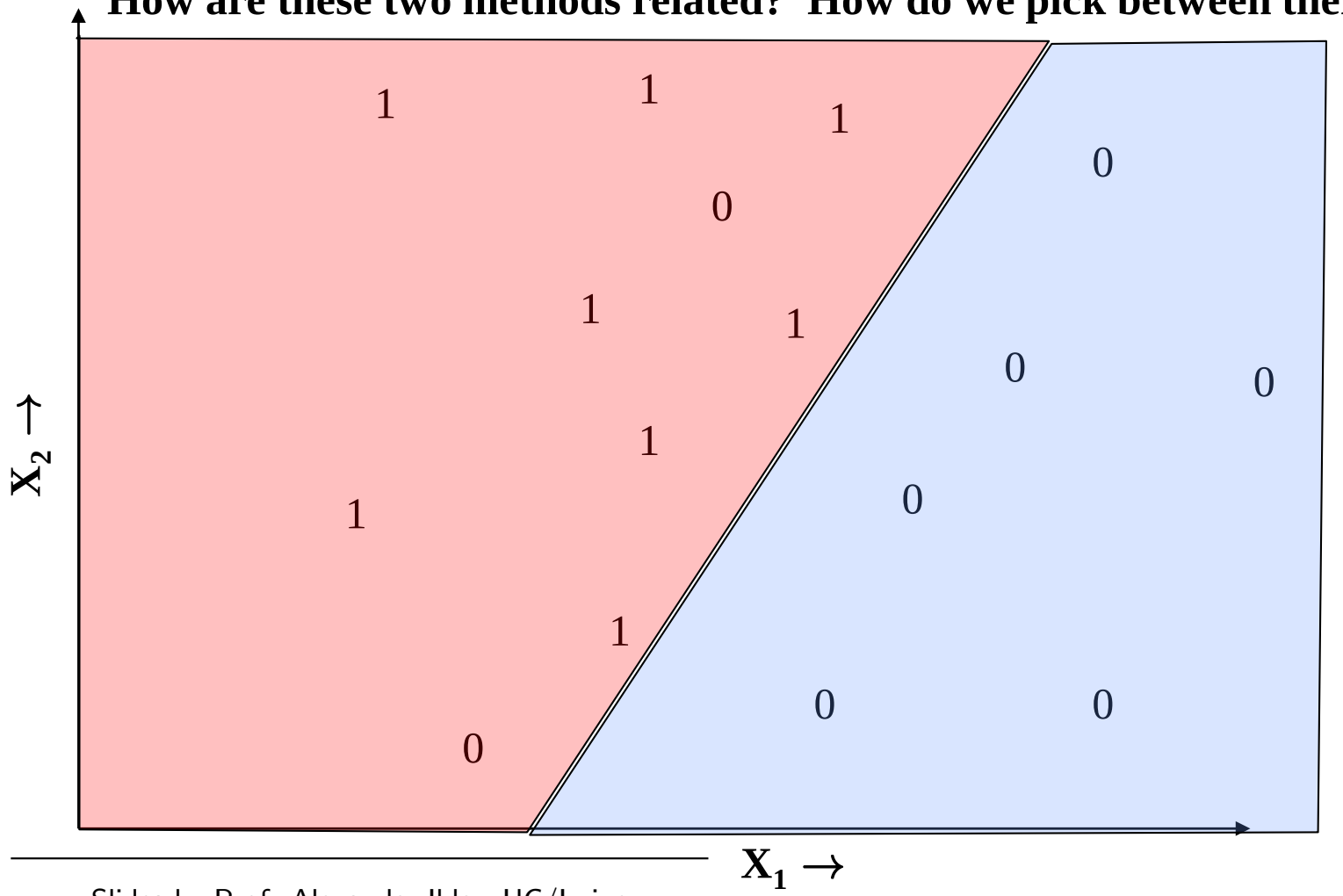
"Complexity" of a classifier

***Parametric*: describe form explicitly in terms of some parameters**

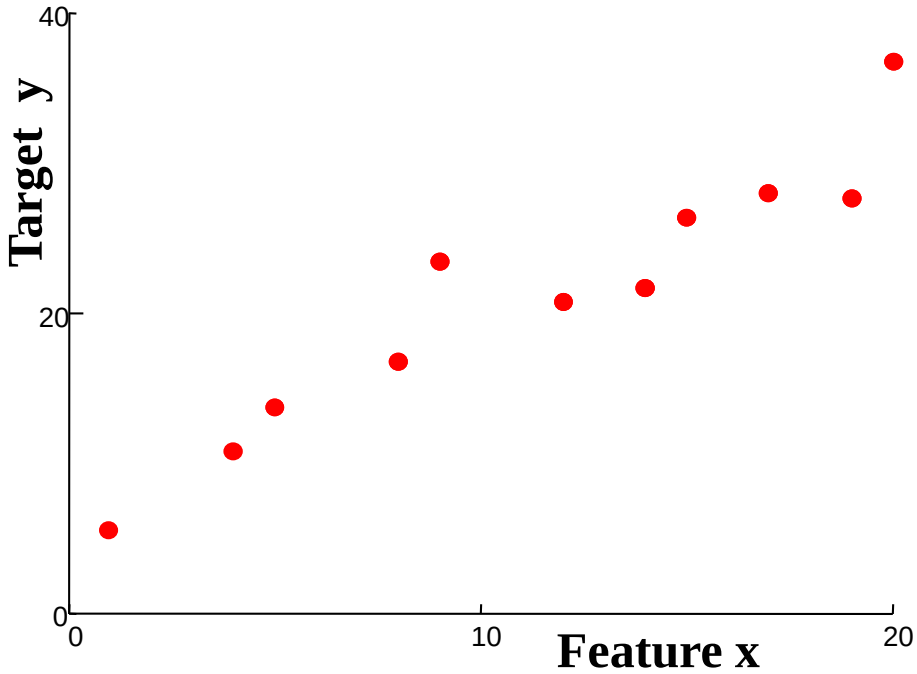
***Non-parametric*: number of parameters required increases with the amount of data**

Questions to consider

How would we select a good linear classifier? (How to measure "error"?)
How are these two methods related? How do we pick between them?

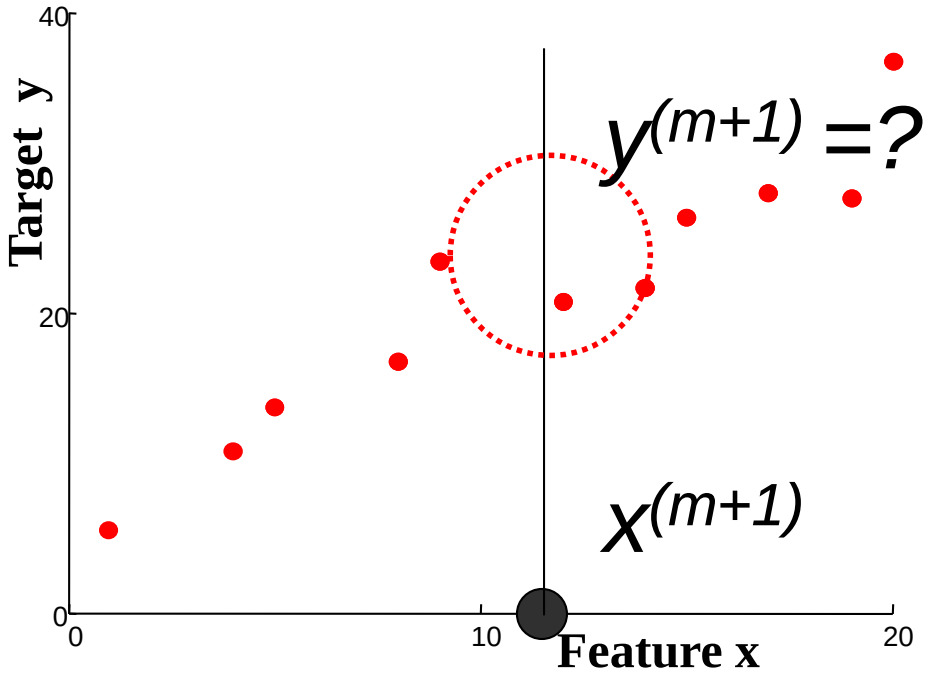


Regression; Scatter plots



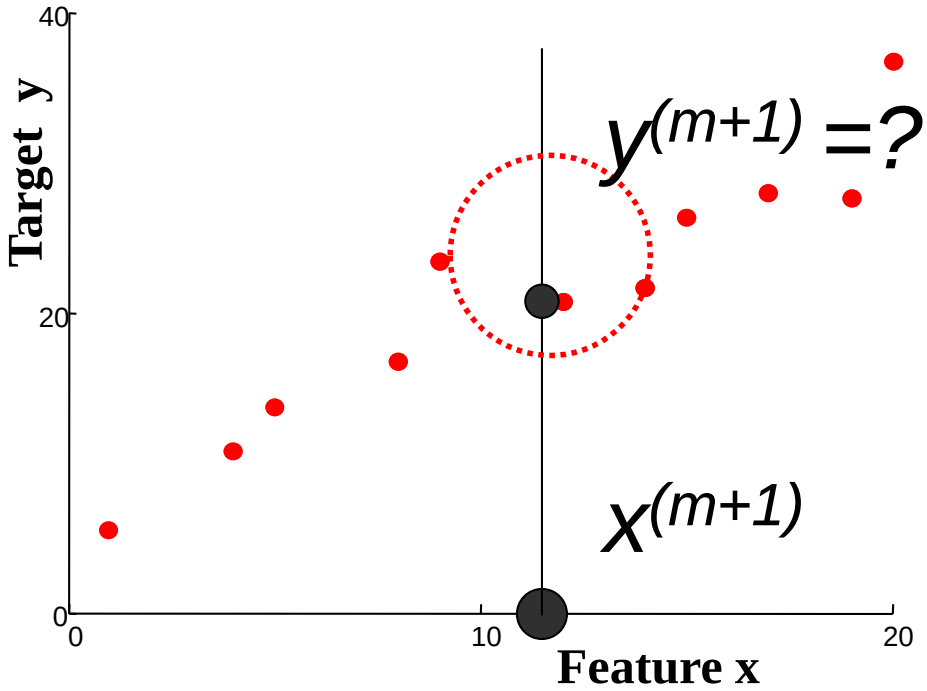
- Suggests a relationship between x and y
- *Prediction*: new x, what is y?

Predicting new examples



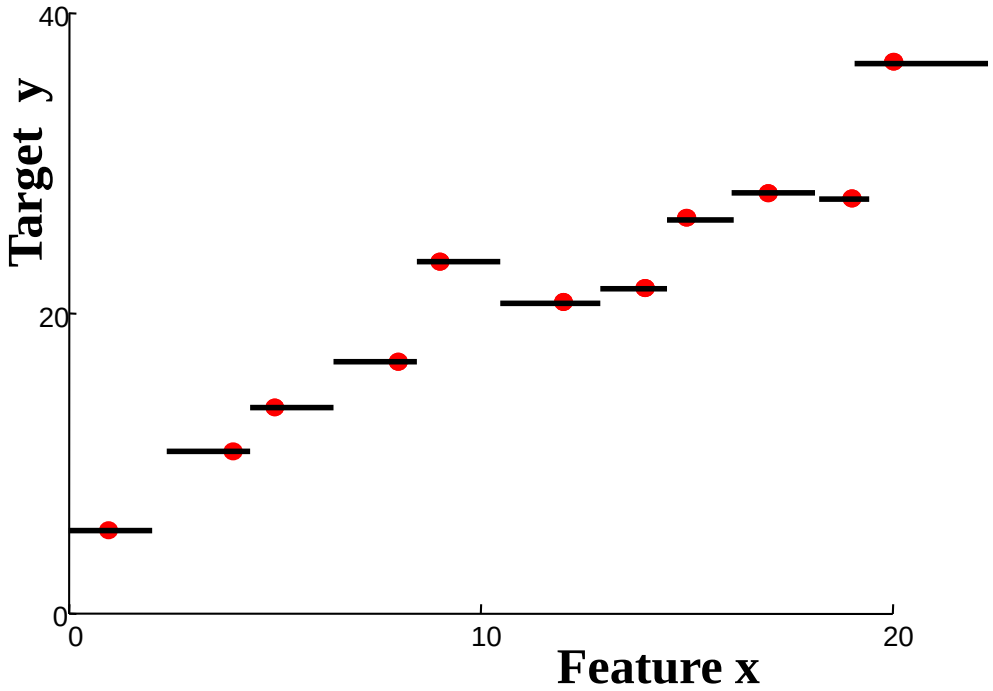
- Regression: given the observed data, estimate $y^{(m+1)}$ given new $x^{(m+1)}$

Nearest neighbor regression



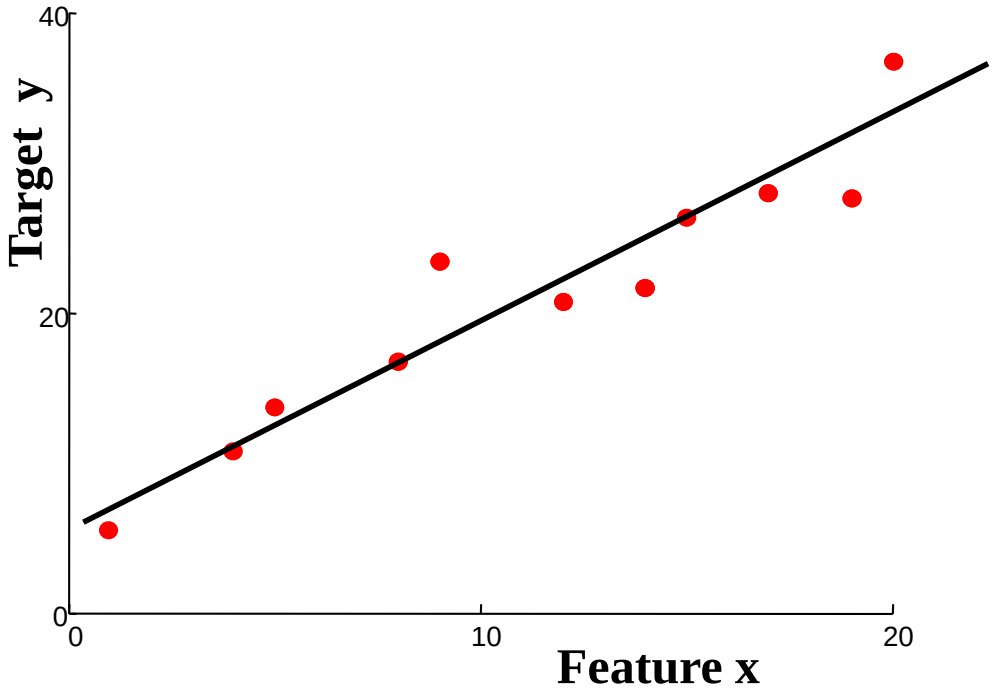
- Find training datum $x^{(i)}$ closest to $x^{(m+1)}$
  Predict $y^{(i)}$

Nearest neighbor regression



- Defines a function f(x) implicitly
- “Form” is piecewise constant

Linear regression



- Define form of function f(x) explicitly
- Find a good f(x) within that family

K-Nearest Neighbor (kNN) Classifier

- Find the k-nearest neighbors to $\underline{x}$ in the data
  - i.e., rank the feature vectors according to Euclidean distance
  - select the k vectors which are have smallest distance to $\underline{x}$

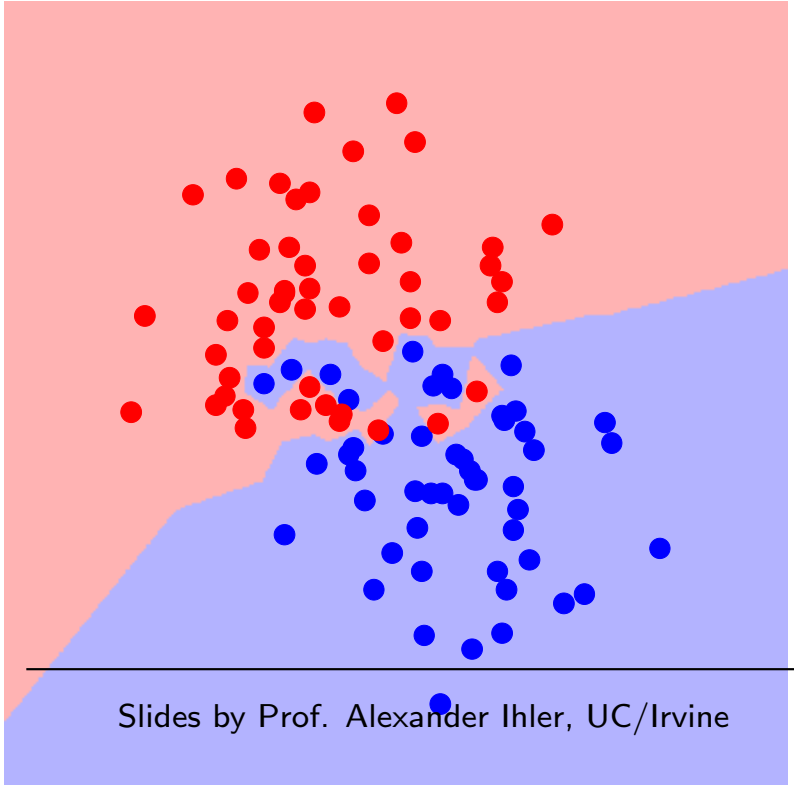
- Classification
  - ranking yields k feature vectors and a set of k class labels
  - pick the class label which is most common in this set ("vote")
  - classify $\underline{x}$ as belonging to this class

- Notes:
  - Nearest k feature vectors from training "vote" on a class label for $\underline{x}$
  - the single-nearest neighbor classifier is the special case of k=1
  - for two-class problems, if we choose k to be odd (i.e., k=1, 3, 5,…) then there will never be any "ties"
  - "training" is trivial for the kNN classifier, i.e., we just use training data as a "lookup table" and search to classify a new datum

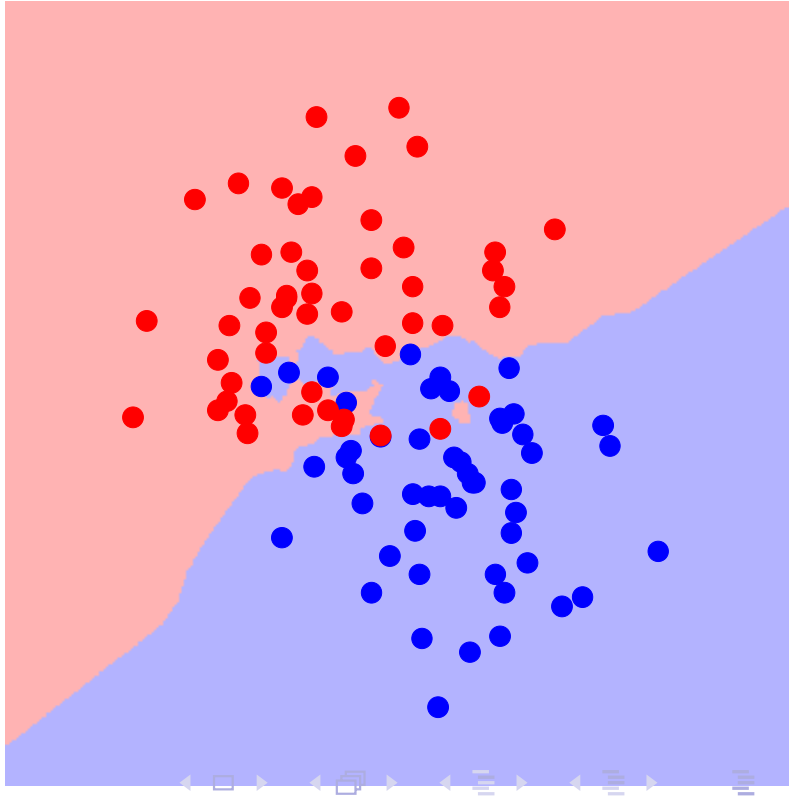
kNN Decision Boundary

- Piecewise linear decision boundary
- Increasing k “simplifies” decision boundary
  - Majority voting means less emphasis on individual points

K = 1



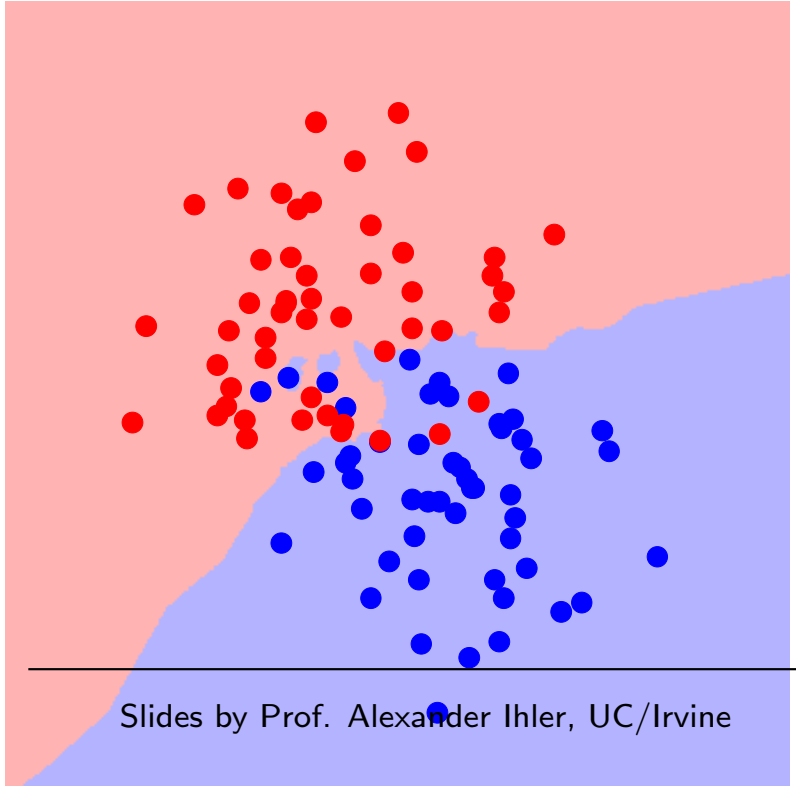
K = 3



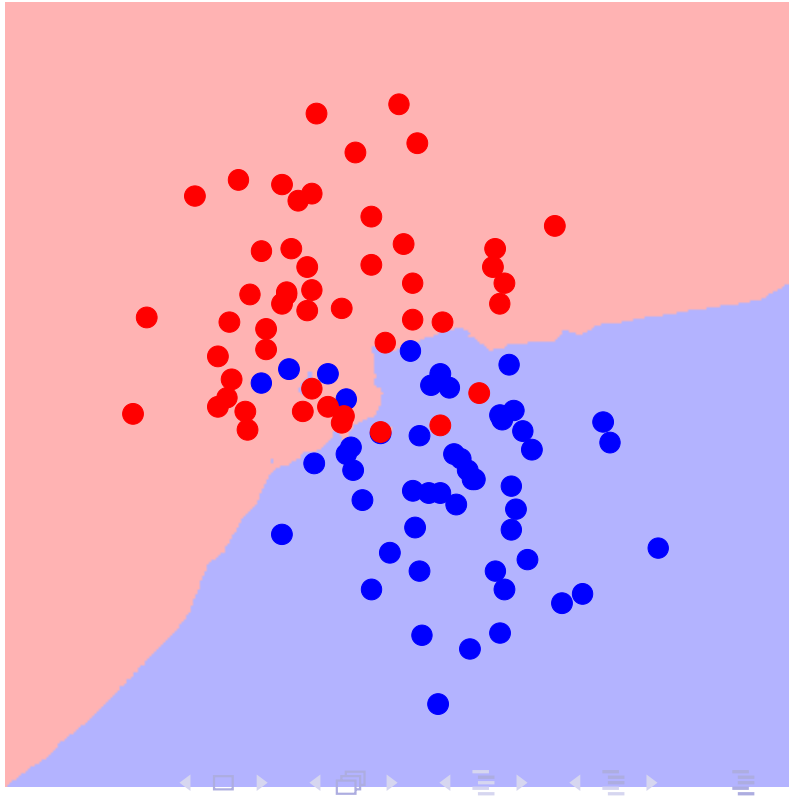
kNN Decision Boundary

- Recall: piecewise linear decision boundary
- Increasing k “simplifies” decision boundary
  - Majority voting means less emphasis on individual points

K = 5



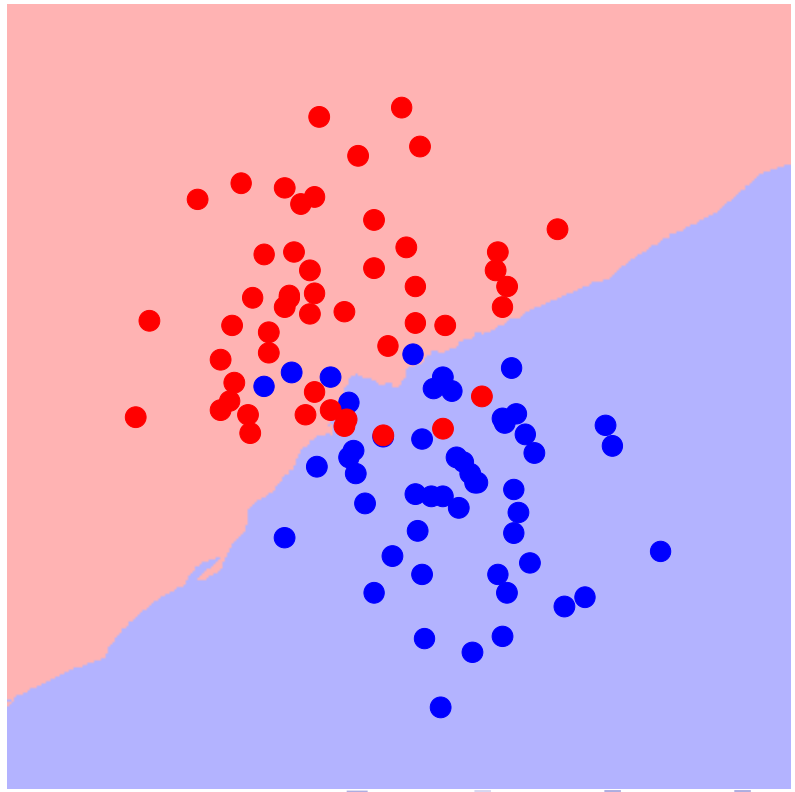
K = 7



kNN Decision Boundary

- Recall: piecewise linear decision boundary
- Increasing k “simplifies” decision boundary
  - Majority voting means less emphasis on individual points

K = 25



Wrapping Up

1. Wrapping Up

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here to register for FREE access »**

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (87)

RELATED

Executive Summary

ALSO AVAILABLE

- Buy PDF

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Editors: Michael Lesk, lesk@acm.org | Jeffrey MacKie-Mason, jmm@umich.edu

Big Data, Big Brother, Big Money

Michael Lesk | Rutgers University

Historically, most of our fears about surveillance and privacy have involved the fear of a controlling government. This is the warning from dystopian fiction such as George Orwell's *1984* or Franz Kafka's *The Trial*, and it is what has happened in Communist states such as East Germany and is now happening in the US. We think of it as purely governmental activity, designed to control what people say and think. What's new today is that surveillance has been outsourced, because most data is in private hands.

Our Former Fears

George Orwell's and Jeremy Bentham's fictional visions of the watched society are well known:[1,2]

> With the development of television, and the technical advance which made it possible to receive and transmit simultaneously on the same instrument, private life came to an end. Every citizen, or

> The more constantly the persons to be inspected are under the eyes of the persons who should inspect them, the more perfectly will the purpose X of the establishment have been attained. Ideal perfection, if that were the object, would require that each person should actually be in that predicament, during every instant of time. This being impossible, the next thing to be wished for is, that, at every instant, seeing reason to believe as much, and not being able to satisfy himself to the contrary, he should conceive himself to be so.

Bentham's model of continuous observation was implemented in the design of the former British Museum Reading Room (1857–1997). The room was circular, with the reading desks laid out along radii. In the center was a raised platform, where one man could stand and see all the readers. When the Library of Congress opened its

Data Science: The Numbers of Our Lives

By CLAIRE CAIN MILLER
Published: April 11, 2013

HARVARD BUSINESS REVIEW calls data science "the sexiest job in the 21st century," and by most accounts this hot new field promises to revolutionize industries from business to government, health care to academia.

Enlarge This Image

40
TRILLION GIGABYTES
Size of digital universe by 2020, up from 130 billion in 2005.

Source: IDC/EMC

The field has been spawned by the enormous amounts of data that modern technologies create — be it the online behavior of Facebook users, tissue samples of cancer patients, purchasing habits of grocery shoppers or crime statistics of cities. Data scientists are the magicians of the Big Data era. They crunch the data, use mathematical models to analyze it and create narratives or visualizations to explain it, then suggest how to use the information to make decisions.

Wrapping Up

- We are living an exciting time to be involved with research in machine learning, information forensics and security
- With more and more people thinking of Machine Learning as the 'Holly Grail' for their problems, our research focuses on studying some key points one has to think of before simply using ML-based solutions

Putting in the Context of Brazil

AVANÇO VIGOROSO

Estudo mostra o crescimento das empresas paulistas de *software*, tecnologia da informação e comunicações e a sua avidez por mão de obra qualificada | **Fabrício Marques**

BAURU
Interiorização e ampliação da cobertura dos serviços fizeram com que a cidade do centro-oeste paulista ampliasse significativamente o número de empregos em telecomunicações entre 2008 e 2012

BARUERI E SANTANA DE PARNAÍBA
Próximas a São Paulo, as cidades especializam-se em empresas de *software*, tratamento de dados, provedores e hospedagem na internet. Boa infraestrutura, logística e incentivos fiscais a empresas explicam o crescimento

Existe um segmento da economia brasileira que cresce a taxas "chinesas" (10,8% em 2012), concentra-se cada vez mais no estado de São Paulo (onde ficam 48,5% das empresas do ramo em operação do país, ante 44,3% em 2008) e se abastece de mão de obra altamente qualificada (47,4% de graduados e pós-graduados, ante 18,8% da média do mercado de trabalho paulista). Trata-se do setor de *software*, tecnologia da informação e de comunicações, esquadrinhado por um estudo lançado em maio pela Fundação Sistema Estadual de Análise de Dados (Seade).

Coordenado pelas pesquisadoras Alda Regina Ferreira de Araújo e Cássia Chrispiniano Adduci, o trabalho faz um mapeamento inédito da distribuição das empresas desse segmento pelos municípios do estado de São Paulo e mostra uma evolução notável entre 2008 e 2012, com a criação de novos polos e a especialização de outros – ainda que a capital paulista siga como centro hegemônico (*ver quadro*). "Nosso interesse em compreender melhor esse segmento, que é intensivo em pesquisa e desenvolvimento, se deve a seu dinamismo e caráter inovador e a sua posição estratégica na promoção do desenvolvimento econômico no estado", diz Alda Ferreira.

A pesquisa também mostra os esforços recentes para formar profissionais capazes de atender às necessidades desse setor – só as instituições públicas de São Paulo aumentaram 93% as vagas em diversos cursos vinculados à computação e às telecomunicações no período analisado pela pesquisa, diante de 32% das instituições particulares – e as dificuldades enfrentadas nesse percurso, como a evasão de alunos. "A ideia é indicar possibilidades para estudos que avancem na discussão sobre a formação de profissionais para o setor e possam contribuir na elaboração de políticas públicas que enfrentem esse desafio", explica Cássia Adduci.

O mercado brasileiro do setor de tecnologia da informação e comunicação é o quarto maior do mundo, atrás de Estados Unidos, China e Japão. Movimentou mais de US$ 230 bilhões em 2012. "O Brasil não chega a ser um *player* mundial no seg-

Wrapping Up

1. We are living an exciting time to be involved with research in Machine Learning and Data Analytics.
2. With more and more people thinking of Machine Learning as the Holy Grail for their problems, our research should focus on studying some key points one has to think of before simply using ML-based solutions.

Questions?

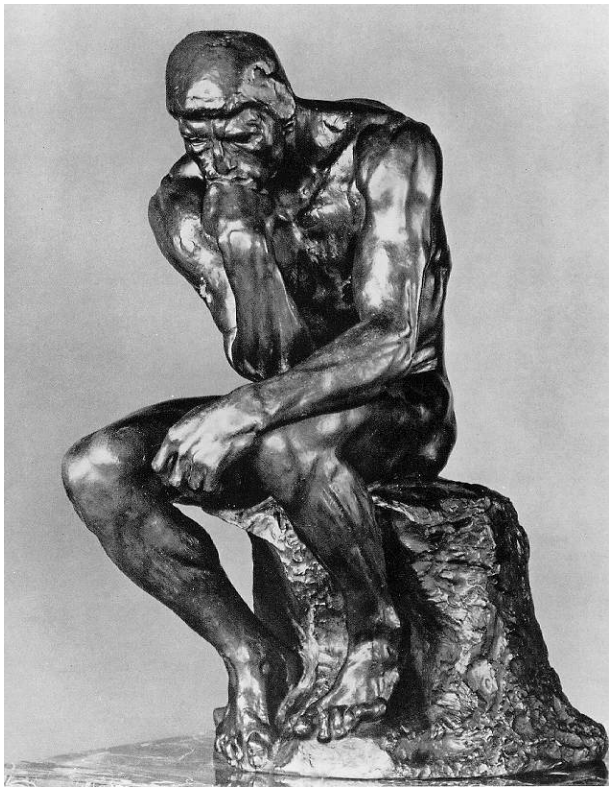


Figure : *The Thinker* - Auguste Rodin.