COURSE INTRODUCTION

J. Elder

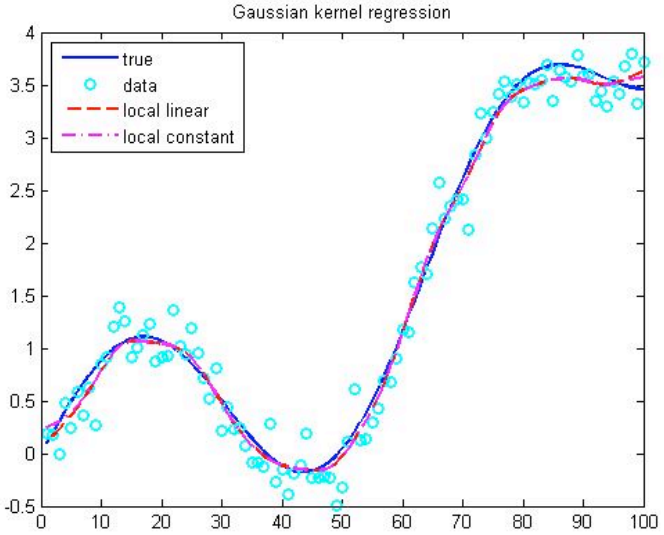
CSE 4404/5327 Introduction to Machine Learning and Pattern Recognition

What is Machine Learning?

- Machine learning is the study of algorithms that learn how to perform a task from prior experience.

What is Pattern Recognition?

- Machine learning can be used to predict a continuous variable, for example, the location of a car now, given its observation from a highway cam 5 minutes ago.
- This kind of prediction is called **regression** or **estimation** in statistics.



What is Pattern Recognition?

- Machine learning can also be used to predict a categorical variable, for example, whether the vehicle is a car or a truck.
- This is pattern recognition (aka **classification**).

?

Supervised vs Unsupervised Learning

- A classification learning problem is said to be **supervised** if labelled training data that associates input observations with class labels are available.

Input	Label
Image 1	Car
Image 2	Truck
Image 3	Truck
Image 4	Car
Image 5	Truck
⋮	⋮

Supervised vs Unsupervised Learning

- A classification learning problem is said to be **unsupervised** if the training data consist of input observations of unknown class.
- In the purely unsupervised case, we may not be able to learn to assign semantic labels, but we may be able to at least identify when inputs are of the same class.
- This is called **clustering**.

Input	Label
Image 1	??
Image 2	??
Image 3	??
Image 4	??
Image 5	??
⋮	⋮

Input	Label
Image 1	Class 1
Image 2	Class 2
Image 3	Class 2
Image 4	Class 1
Image 5	Class 2
⋮	⋮

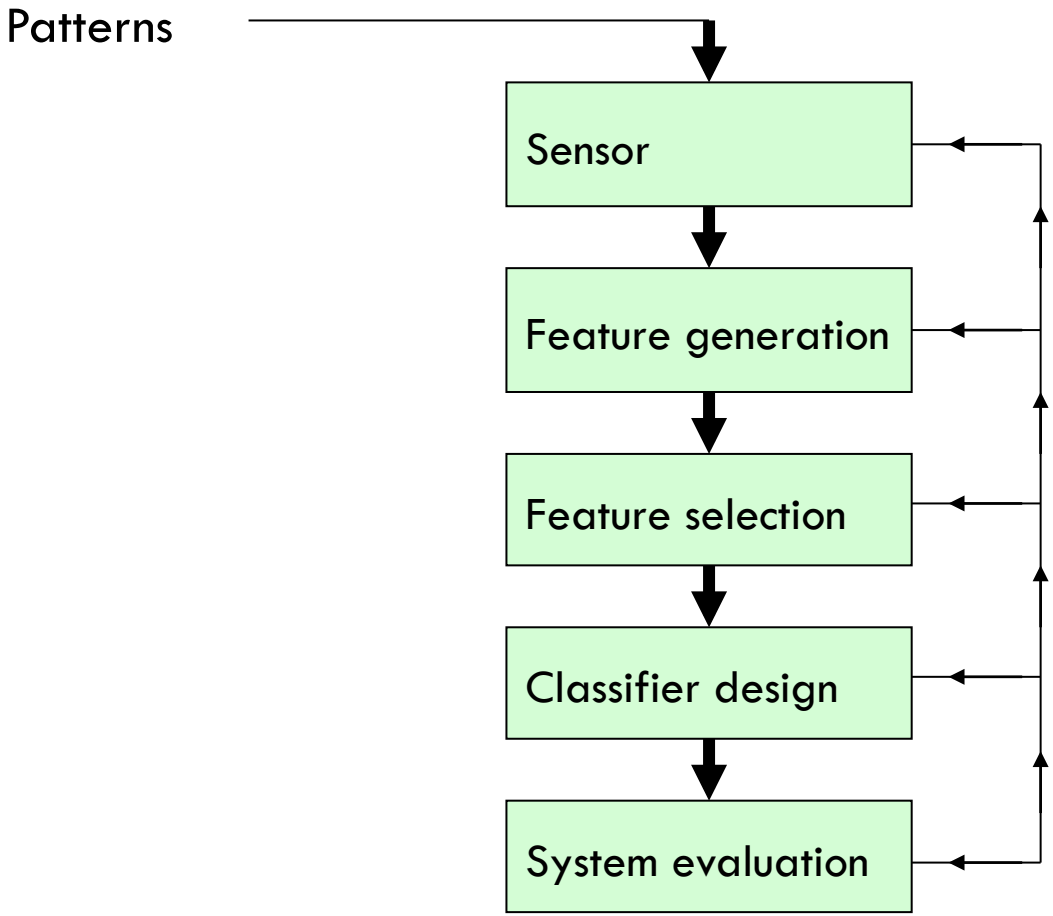
Supervised vs Unsupervised Learning

- **In this course we will focus on supervised learning.**
- CSE 4412 Data Mining covers some topics in unsupervised learning (clustering)

Learning Objectives

- Upon completing this course the student will, through the assignments, test, and final project, have demonstrated an ability to:
  - Use probabilistic modeling and statistical analysis of data to develop powerful pattern recognition algorithms.
  - Identify machine learning models and algorithms appropriate for solving specific problems.
  - Explain the essential ideas behind core machine learning models and algorithms
  - Identify the main limitations and failure modes of core machine learning models and algorithms
  - Program moderately complex machine learning algorithms
  - Manage data and evaluate and compare algorithms in a supervised learning setting
  - Access and correctly employ a variety of machine learning toolboxes currently available.
  - Identify a diversity of pattern recognition applications in which machine learning techniques are currently in use.

Classification System Design

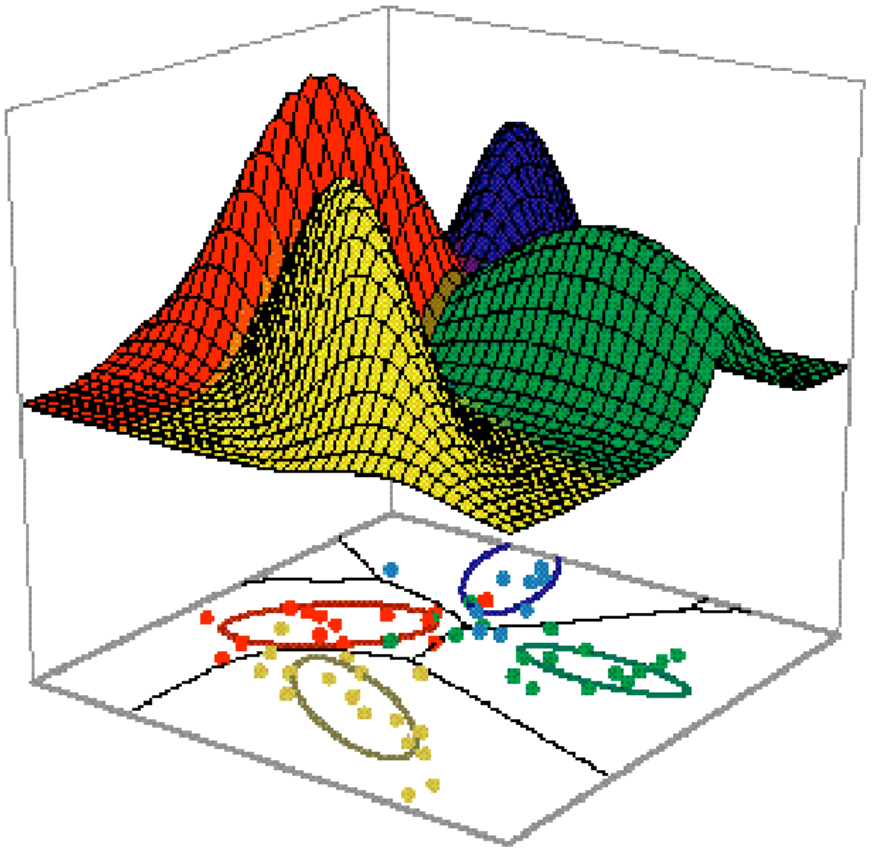


Topics

- Bayesian Decision Theory
- Training & Evaluation
- Linear Regression
- Linear Classifiers
- Feature Selection
- Dimensionality Reduction
- Nonlinear Classifiers
- Markov Models

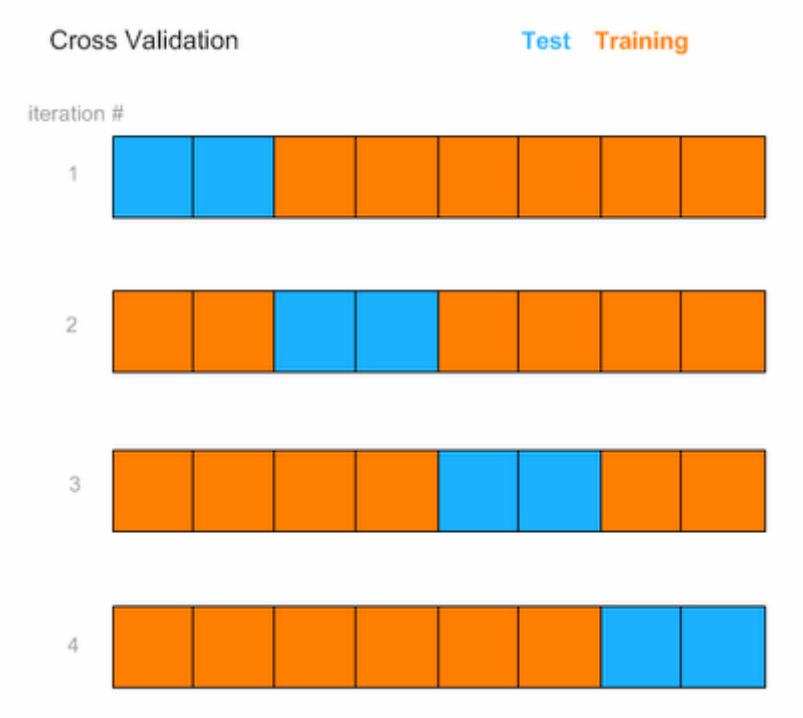
Bayesian Decision Theory

- In the context of pattern recognition, Bayesian Decision Theory specifies how to optimally classify new inputs based upon the prior probability of each class and on observed features.
- Note that the theory is optimal but not always tractable! Hence the need for additional theory and algorithms.



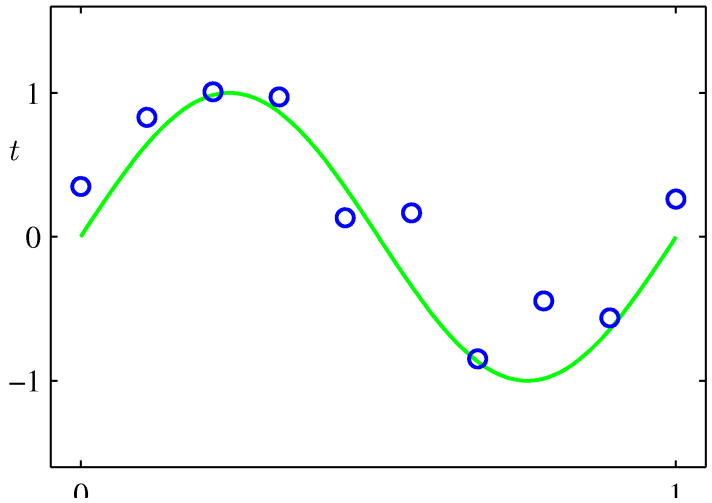
Training and Evaluation

- Test data used for evaluation must be independent of the training data used for learning.
- Cross-validation can be used to prevent over-learning.



Linear Regression

- Regression is a method for determining parameters **w** for a function $y = f(\mathbf{x})$ that predicts the value of a continuous dependent variable y given observation of an input vector **x**.
- Under linear regression, the dependent variable y is a linear function of these parameters **w**.
- Note that the relationship between y and **x** can be highly nonlinear.



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

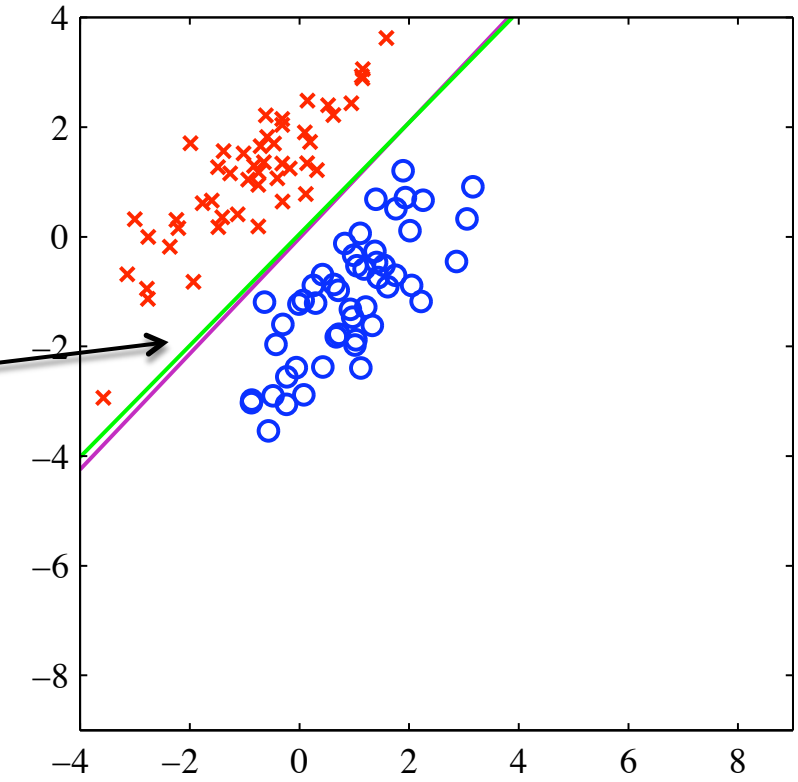
Linear Classifiers

- Linear models for classification separate input vectors into classes using linear *decision boundaries*.
  - Example:

Input vector **x**

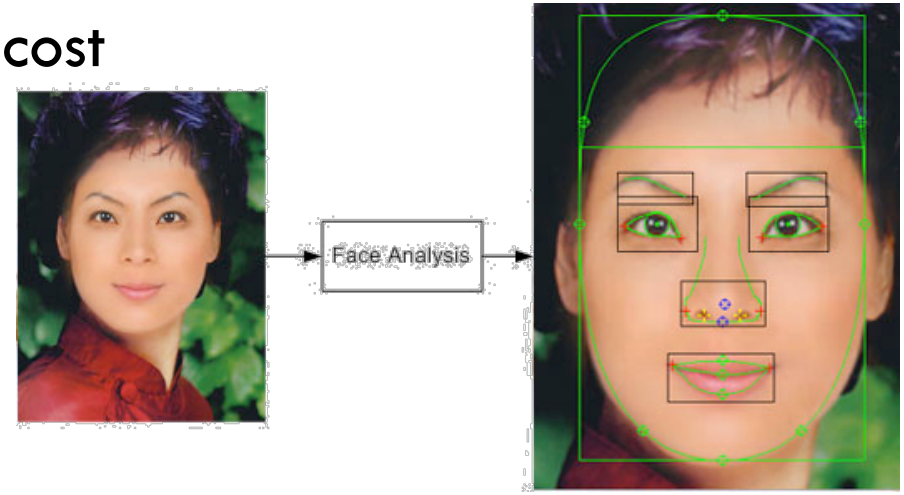
Two discrete classes C_1 and C_2

Decision boundary



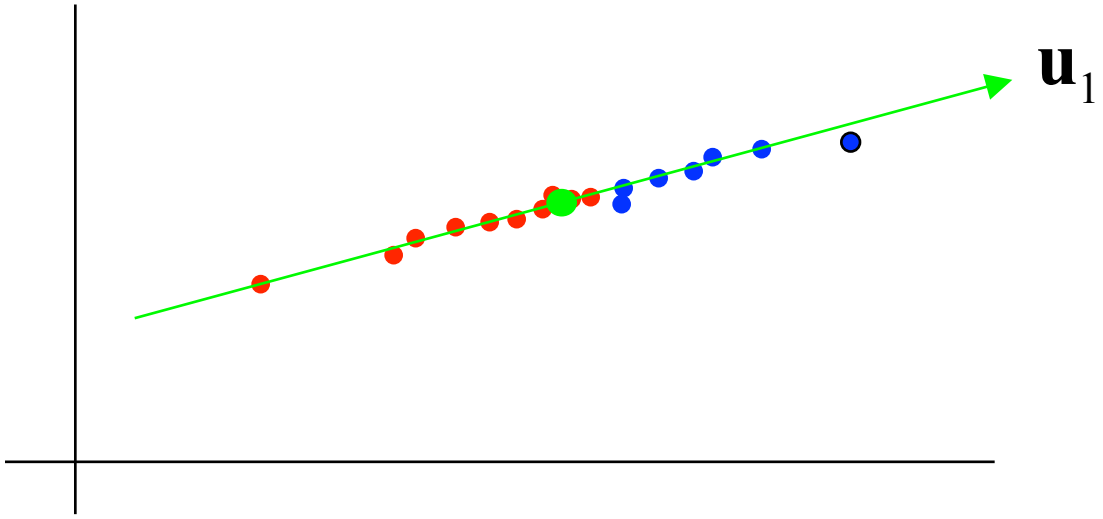
Feature Selection

- The dimensionality of the input vector **x** is often high (e.g., ~10^7 for images)
- Classifiers are typically based on a relatively small number of features extracted from each input. This is important:
  - To prevent overlearning
  - To reduce computational cost



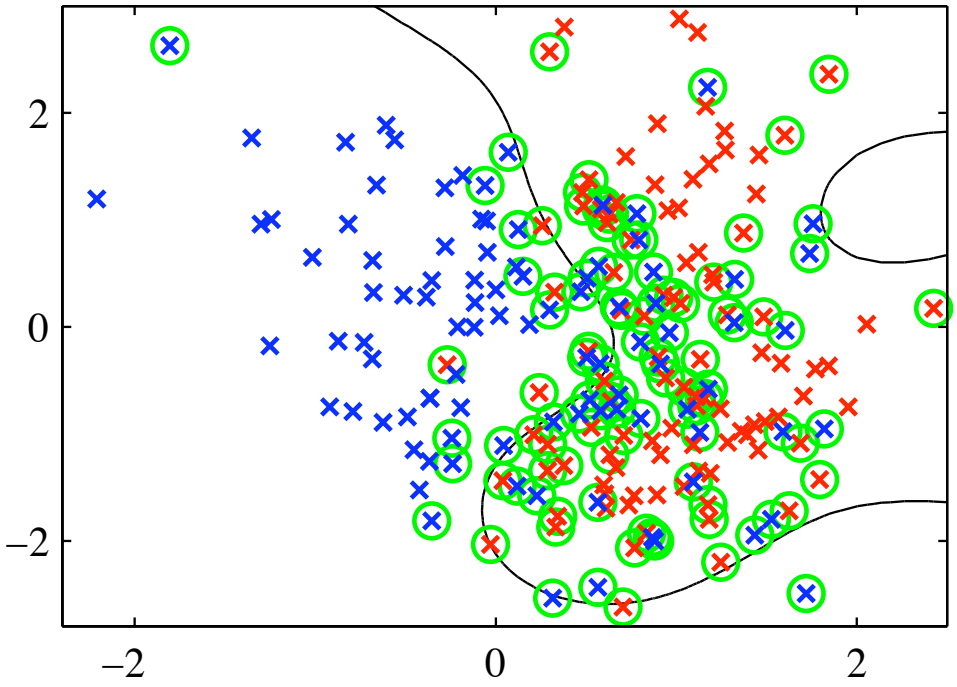
Dimensionality Reduction

- Dimensionality reduction is a key step in extracting features.
- The goal is to identify a low-dimensional subspace of the input space in which the discriminative features lie.



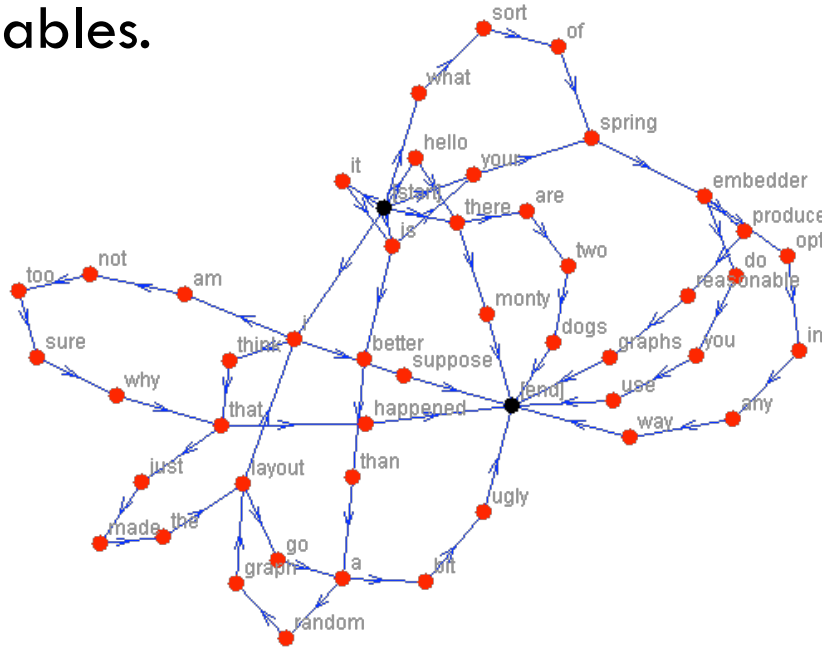
Nonlinear Classifiers

- In practice, classes are rarely linearly separable.
- Fortunately there are many good techniques that extend linear algorithms to these nonlinear problems.



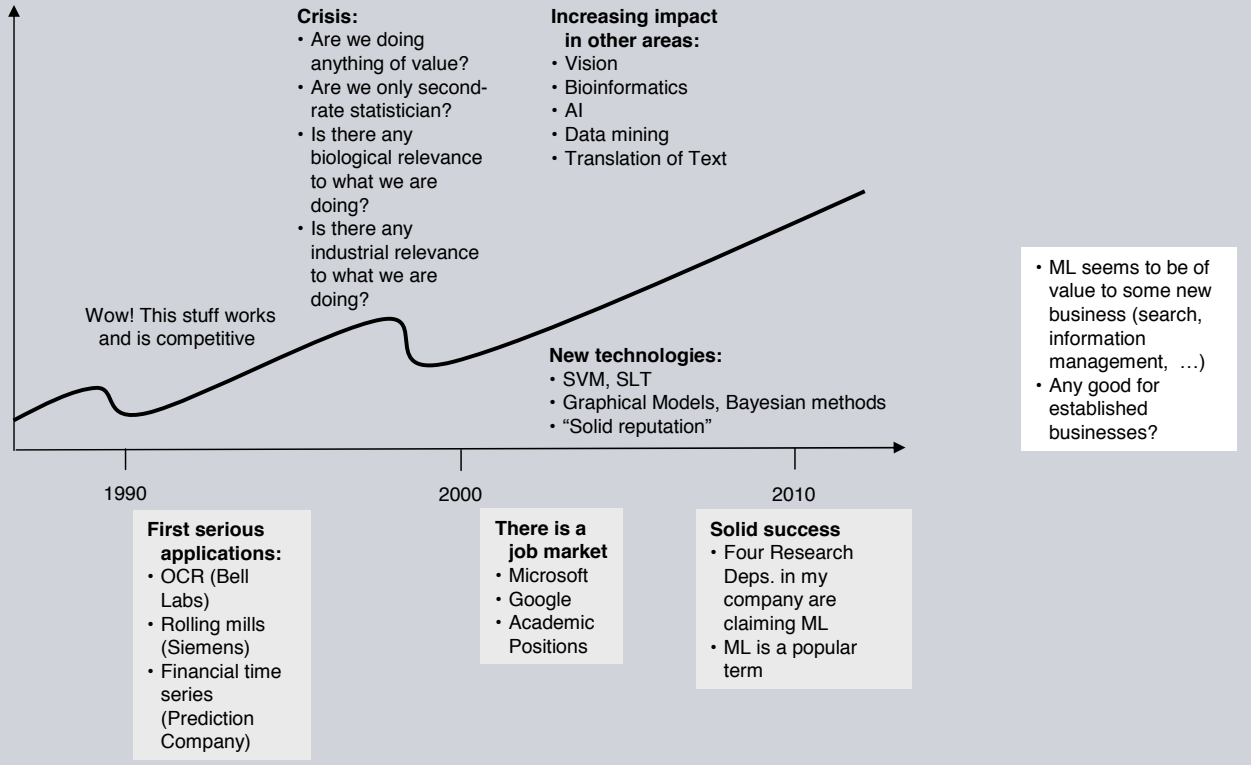
Markov Models

- In many situations, the labels y and hence input vectors **x** are not statistically independent.
- These dependencies must be modeled for accurate classification.
- In a Markov model these dependencies are modeled through local interactions between the variables.



Is Machine Learning Important?

Personal Time-Line of Machine Learning



Page 3

Slide courtesy of Volker Tresp, Siemens Corporate Research

The Department of Machine Learning

Applications of Machine Learning

Neural Nets for Process Control

Machine Learning in Steel Processing

- One of the first shipped machine learning solution in a serious industrial application world wide (early 1990's)
- An online adaptive control solution was realized
- The solution is installed in more than 200 sites world wide!

Page 9

Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Process Control

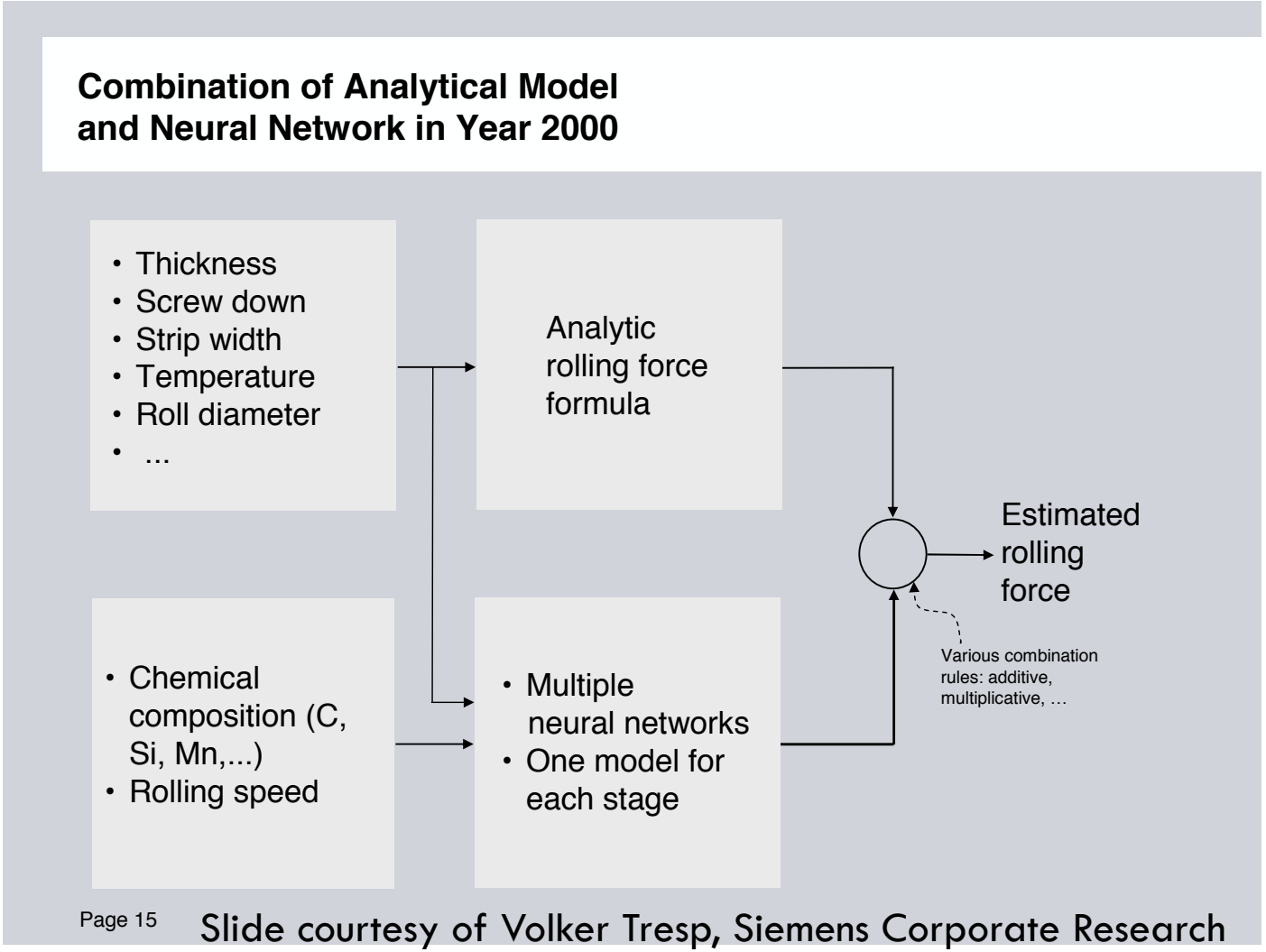
Prediction of Rolling Force (early 1990s)

Problem:
- What is the force required to obtain a certain reduction in thickness based on approximately a dozen features?

- An analytical formula with table based adaptation was being used
- Goal: increase accuracy by neural networks (instead of table)
- Online adaptation was essential ('Tagesform der Anlage')

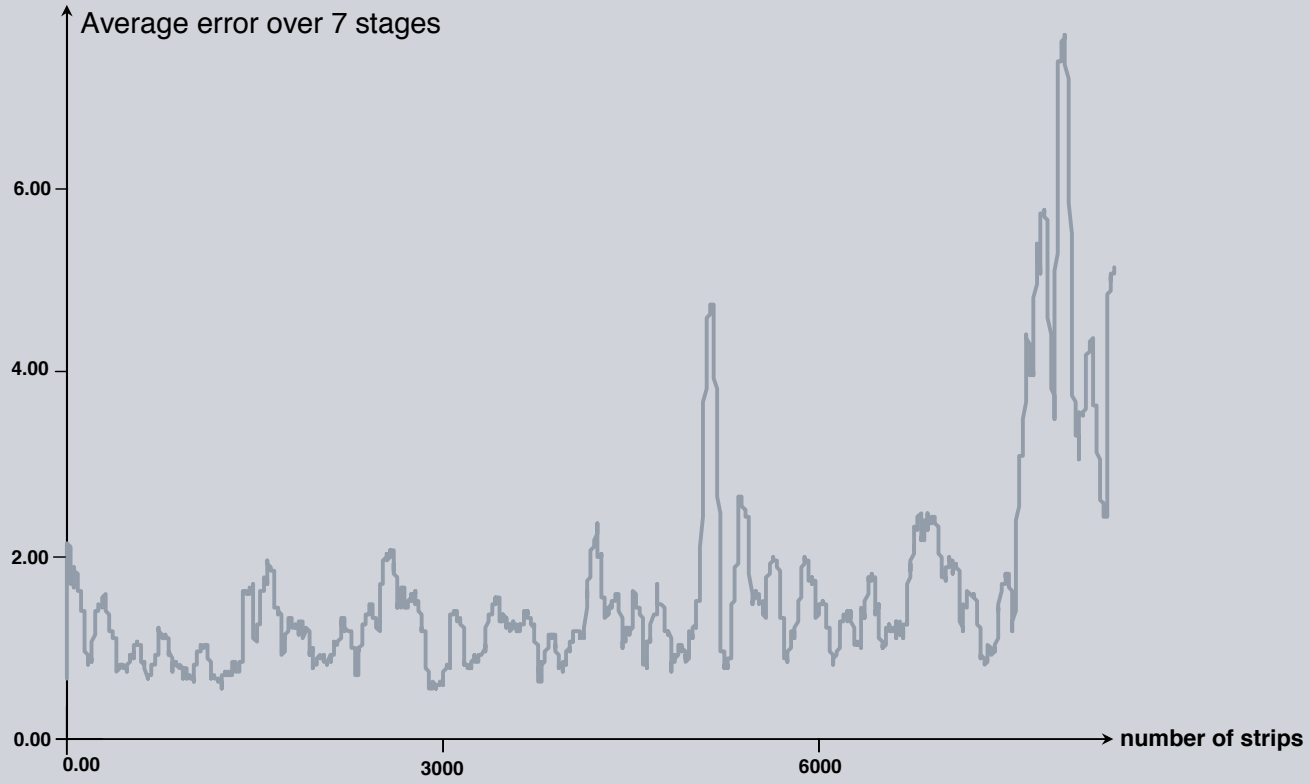
Page 11 Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Process Control



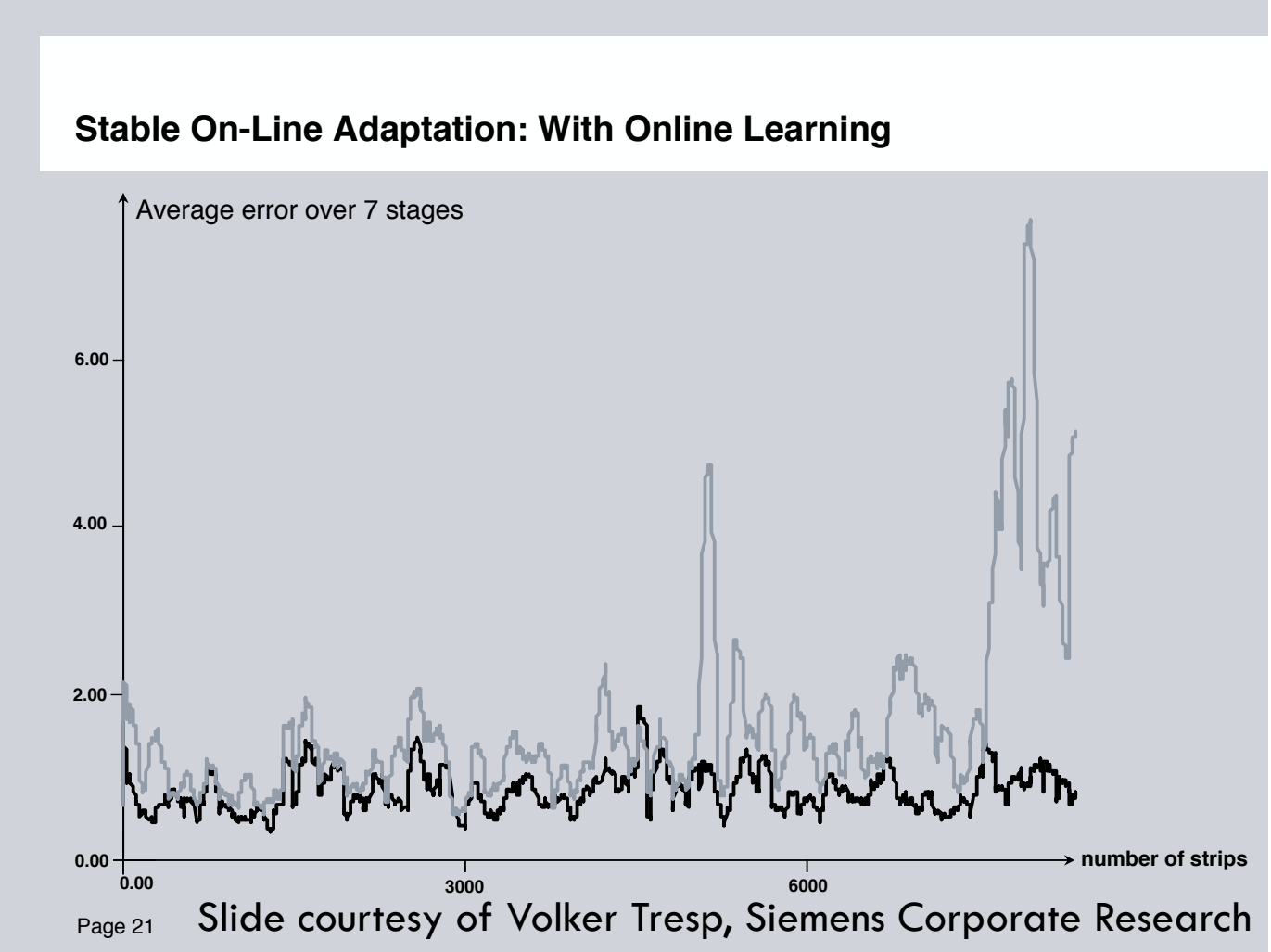
Neural Nets for Process Control

Stable On-Line Adaptation: Without Online Learning



Page 20 Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Process Control



Neural Nets for Process Control

Model Selection and Requires Service

Multi-layer Perceptron (MLP) versus Radial-Basis Functions (RBF)

- On a fixed data set, both approaches gave good performance
- The RBF model was more appropriate to deal with the online learning aspect: only the linear parameters are adapted online
- A modified Widrow-Hoff learning rule was used

How much service is required?

- The system runs autonomously

Page 23

Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

Postal Automation

Page 34 Slide courtesy of Volker Tresp, Siemens Corporate Research

Handwriting Recognition

- Over 85% of handwritten mail in the US is sorted automatically, using handwriting analysis software trained to very high accuracy using machine learning over a very large data set.

Neural Nets for Handwriting Recognition

Overview of Methods

Machine Learning vs Heuristic methods

- Polynomial Classifiers (J. Schürmann)
- Neural networks
- Support Vector Machines

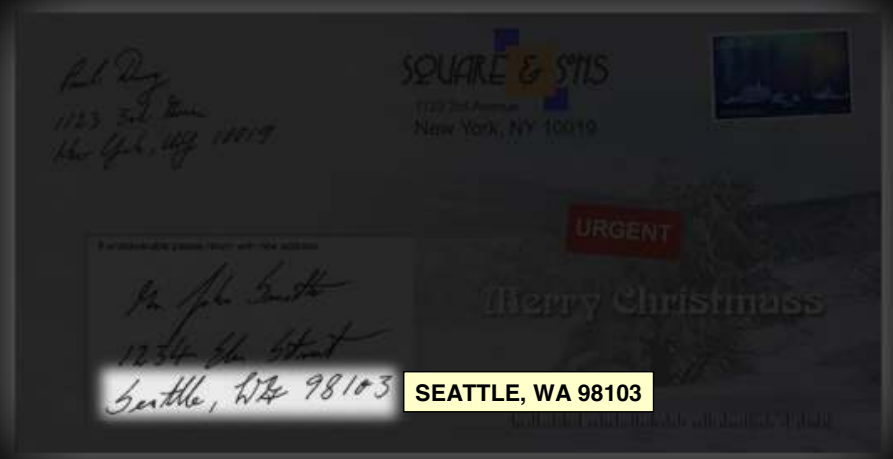
- Hidden Markov Models
- Conditional Random Fields

Page 50

Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

1978: First Postal Code Reader Worldwide



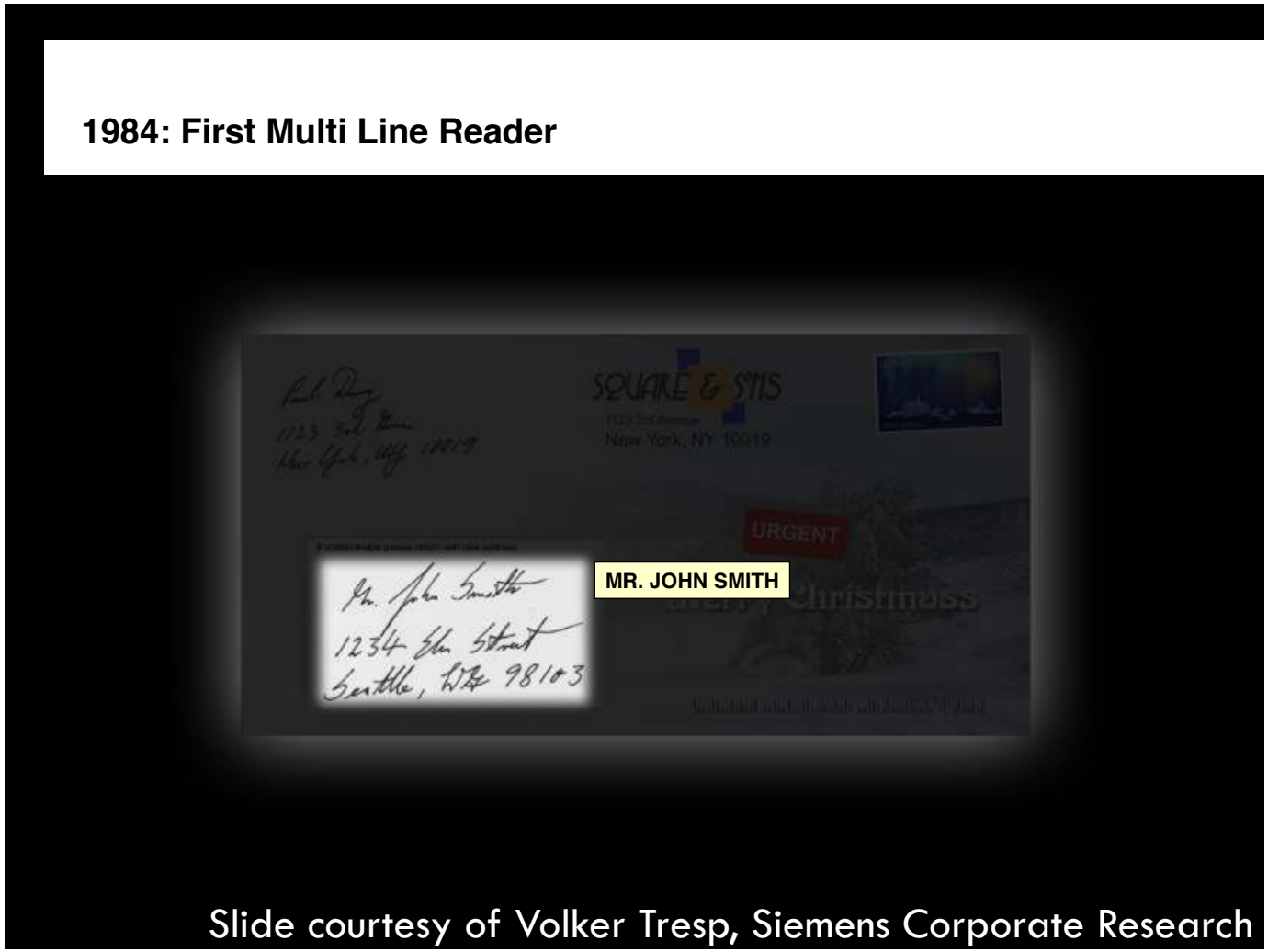
Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

1982: First Address Reader Worldwide

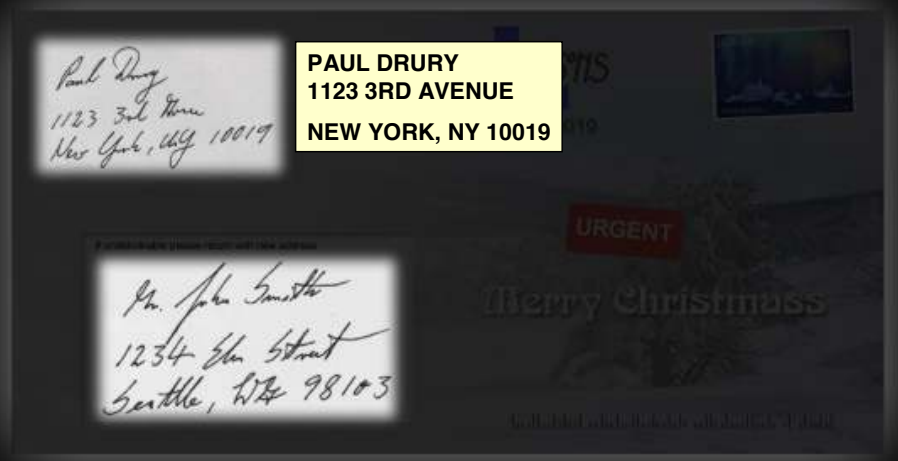
Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition



Neural Nets for Handwriting Recognition

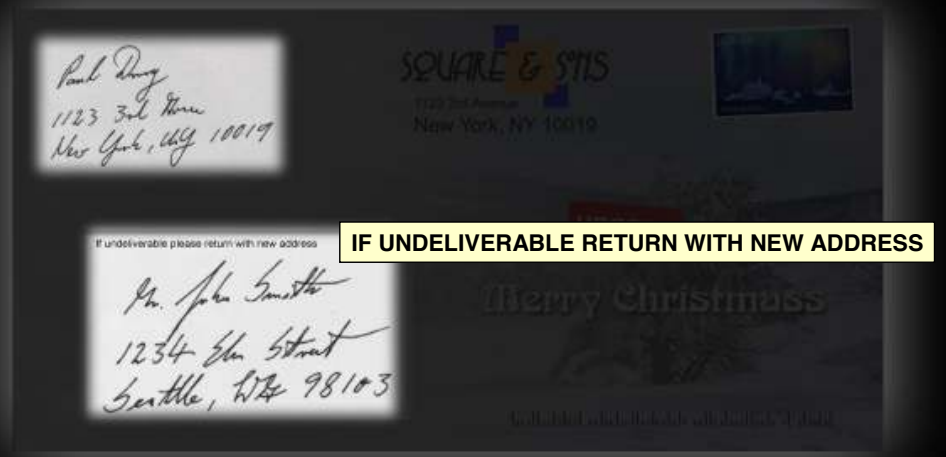
1996: First Sender's Address Reader



Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

1998: First Full Text Reading



Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

2000: First Graphics Recognition

Slide courtesy of Volker Tresp, Siemens Corporate Research

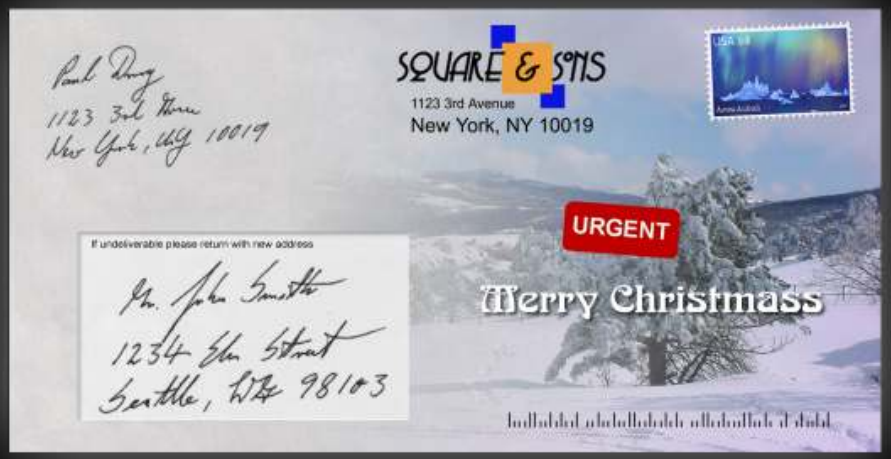
Neural Nets for Handwriting Recognition

2004: First Full Recognition

Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

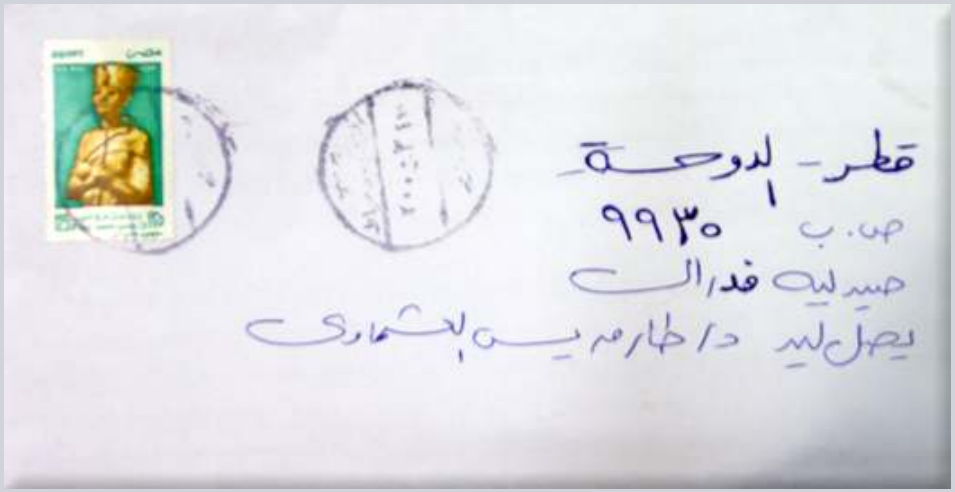
2008: Recognition on Both Sides of Envelope



Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

2000: First Multilingual Readers: e.g. Arabic …

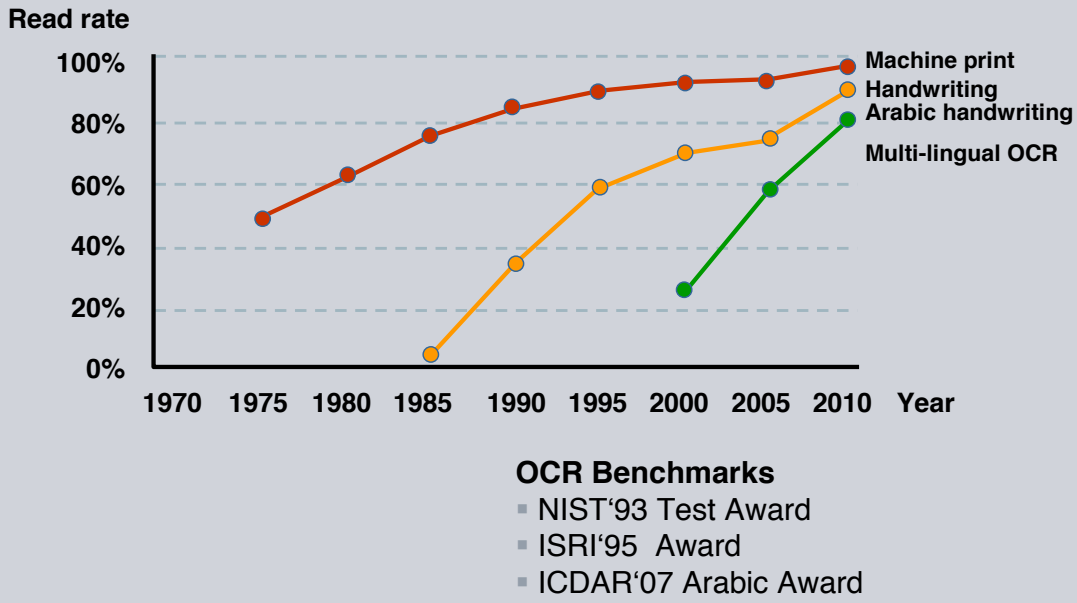


Page 44

Slide courtesy of Volker Tresp, Siemens Corporate Research

Neural Nets for Handwriting Recognition

Steep Increase in Read Rates



OCR Benchmarks

- NIST'93 Test Award
- ISRI'95 Award
- ICDAR'07 Arabic Award

Page 45

Slide courtesy of Volker Tresp, Siemens Corporate Research

Medical Imaging

Detection of Vascular Land Marks

- Automatic extraction of image segments showing vasular regions of interest
- Automatic labelling of vascular segments
- Part of syngo.via since October 2010

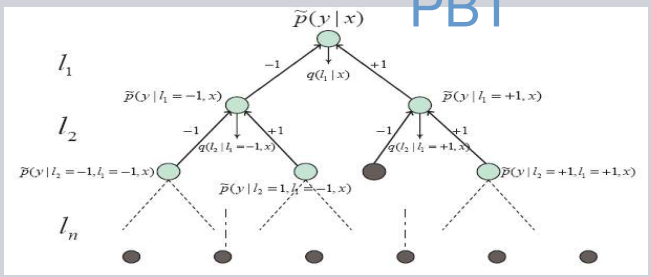
Page 54 Slide courtesy of Volker Tresp, Siemens Corporate Research

Medical Imaging

Applied Methods

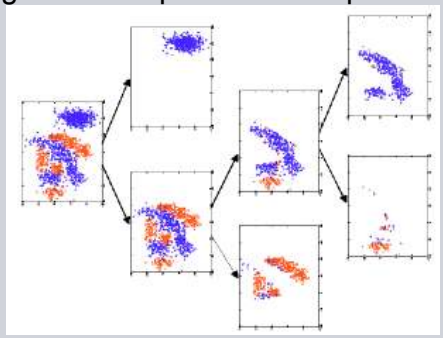
Probabilistic Boosting Tree

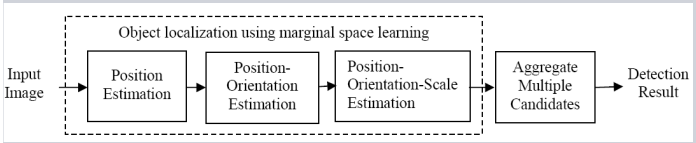
PBT



[Tu, Z.: Probabilistic boosting-tree: Learning discriminative methods for classification, recognition, and clustering. ICCV, 2005.]

Histogram-Example of a 2000-points dataset





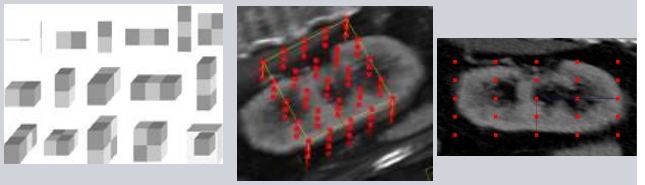
Suche in Unterräumen (Marginalräumen)
→Dimensionsreduktion
→Elimination zahlreicher Hypothesen
→Hohe Effizienz

MSL

[Zheng, Y.; Barbu, A.; Georgescu, B.; Scheuering, M. & Comaniciu, D.: Fast Automatic Heart Chamber Segmentation from 3D CT Data Using Marginal space Learning and Steerable Features, ICCV2007.]

3D Haar-like features

3D steerable features

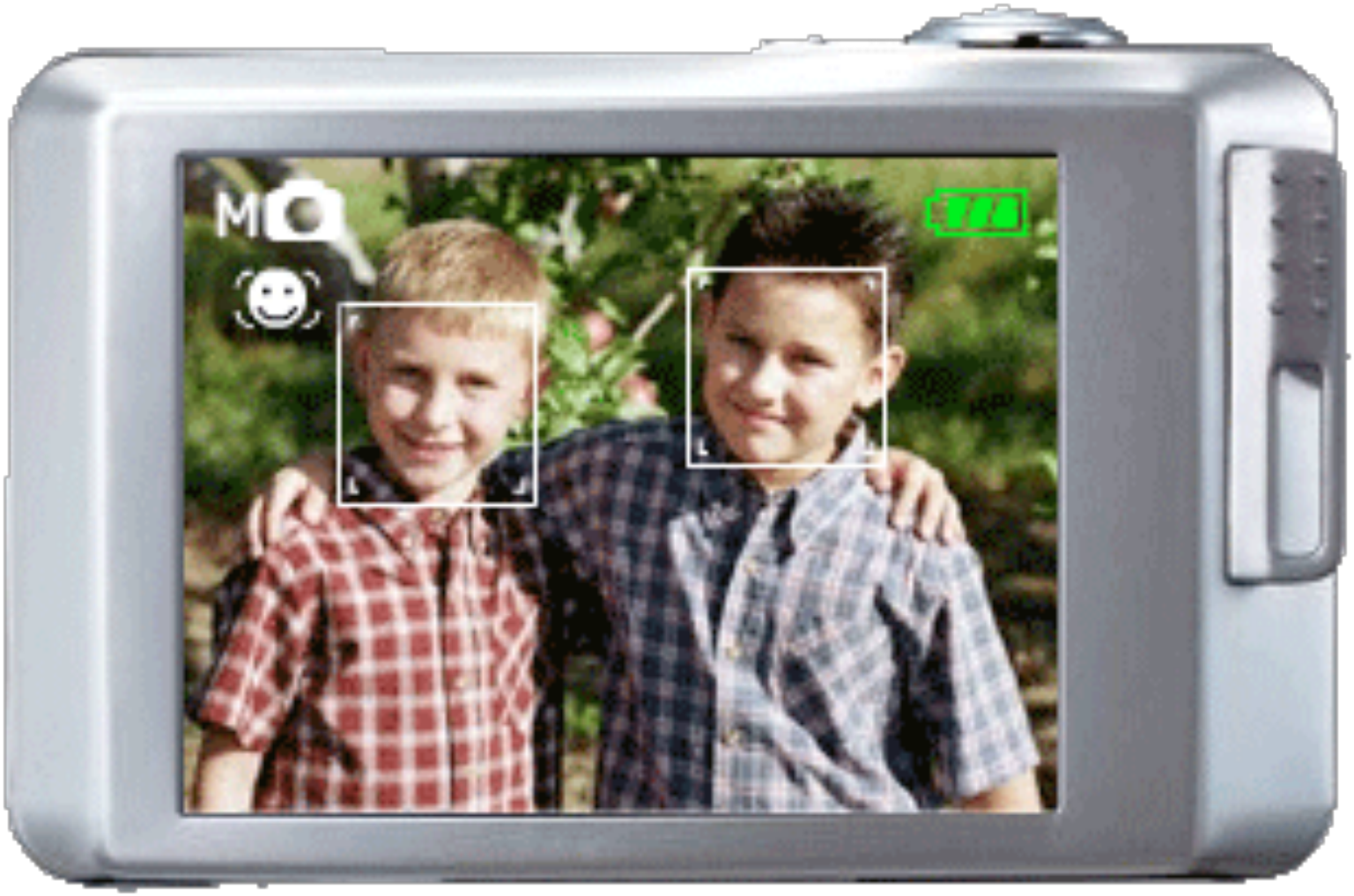


[Tu, Z.; Zhou, X.; Barbu, A.; Bogoni, L. & Comaniciu, D.: Probabilistic 3D Polyp Detection in CT Images: The Role of Sample Alignment, CVPR 2006.]

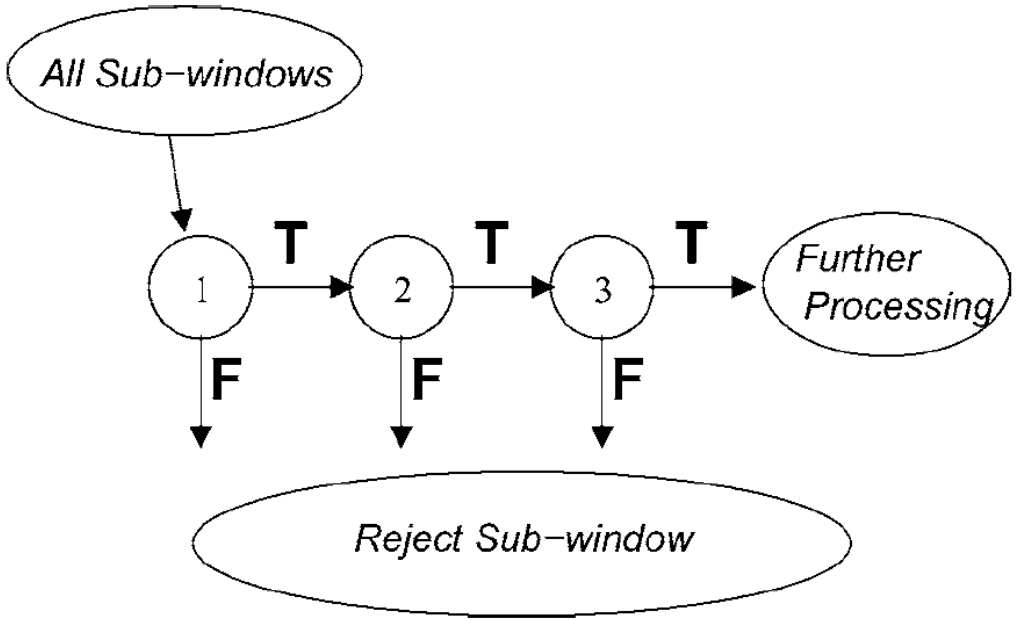
Page 55

Slide courtesy of Volker Tresp, Siemens Corporate Research

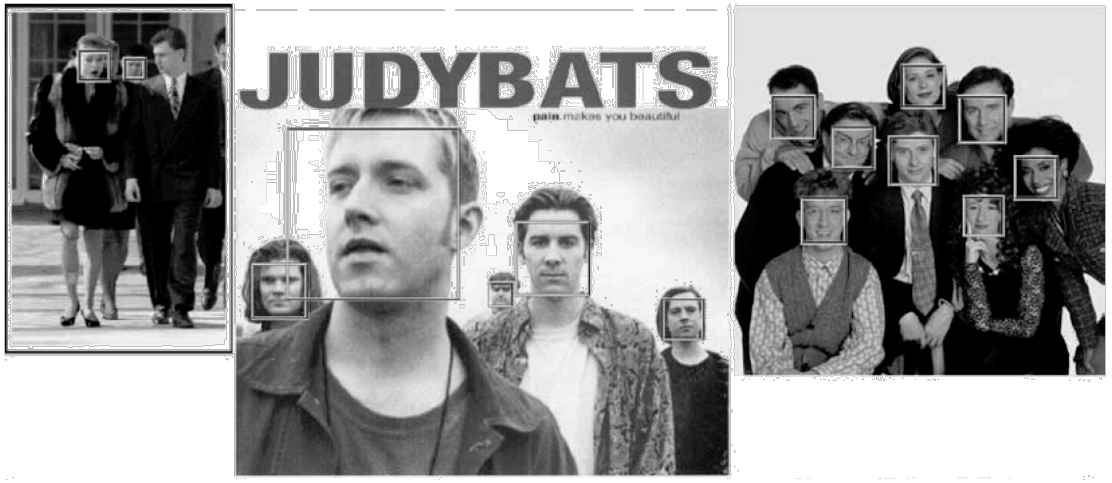
Face Detection in Cameras (2007)



Boosting for Face Detection (2004)



Viola & Jones 2004



Spam Filters

proofpoint
Control tomorrow's email risks today

News & Events

OVERVIEW
PRESS RELEASES
SECURITY, COMPLIANCE & CLOUD
NEWS
IN THE NEWS
EVENT CALENDAR
ACCOLADES
EMAIL SECURITY BLOG

GET STARTED
Build Your Solution

LIVE DEMO
Email Security & Archiving

TRY PROOFPOINT
Free 45-Day Evaluation

Press Release

Proofpoint PR Contact: Orlando De Bruce, 408-338-6870, **pr@proofpoint.com**

0

Proofpoint Introduces Industry's First Dynamic, Machine-learning Based Email Reputation Service

Proofpoint Dynamic Reputation and netMLX Machine Learning Fabric Deliver Highest-accuracy, Highest-performance Reputation Analysis and Connection Management Capabilities

Cupertino, Calif. – April 30, 2007 – Proofpoint, Inc., the leader in large-enterprise messaging security solutions, today introduced the industry's first dynamic, machine-learning based email reputation solution, Proofpoint Dynamic Reputation™. Proofpoint Dynamic Reputation is the only email reputation service that uses a combination of local, predictive behavioral data and globally-observed reputation—analyzed by powerful machine learning algorithms—to block incoming connections from malicious IP addresses. The system provides Proofpoint customers with an accurate, first line of defense against spam, directory harvest attacks, denial of service attacks and other email-borne threats while delivering substantial bandwidth savings.

"Unlike reactive, static reputation services that are forced to make critical trade-offs between connection shed rates, detection accuracy and response time, Proofpoint Dynamic Reputation delivers the highest performance in all three areas at once," said Sandra Vaughan, senior vice president of products and marketing for Proofpoint. "Our solution can block more than 80% of inbound connections with a false positive rate of less than one in one million. Global reputation profiles are updated every minute, ensuring the fastest response to new botnets and malicious IP addresses the moment that they emerge."

Machine Learning Creates Most Accurate, Up-to-date Source of Global IP Reputation

Boosting for Webpage Ranking

From Wikipedia

Commercial web search engines **began using machine learned ranking systems since 2000s**. One of the first search engines to start using it was **AltaVist**a (then Overture, now part of Yahoo), which launched a gradient boosting-trained ranking function in April 2003.

Bing's search is said to be powered by **RankNet** algorithm,which was invented at Microsoft Research in 2005.

…

As of 2008, Google's Peter Norvig **denied that their search engine exclusively relies** on machine-learned ranking.

Page 6

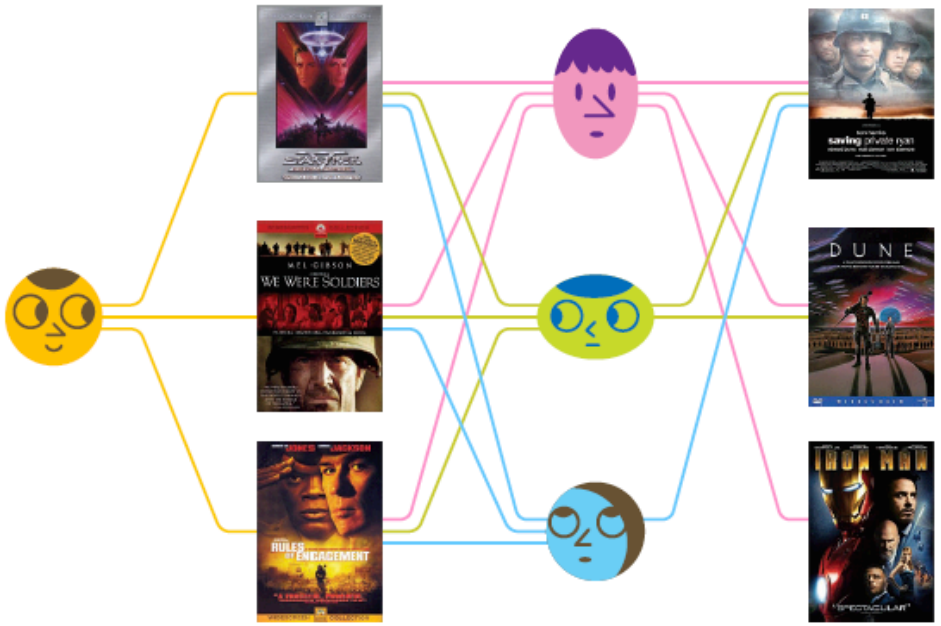
Slide courtesy of Volker Tresp, Siemens Corporate Research

Entertainment: Collaborative Filtering for Movie Recommendations

The Neighborhood Model

THE NEAREST-NEIGHBOR METHOD works on the principle that a person tends to give similar ratings to similar movies. Joe likes the three movies on the left, so to make a prediction for him, find users who also liked those movies and see what other movies they liked. Here the three other viewers all liked *Saving Private Ryan*, so that is the top recommendation. Two of them liked *Dune*, so that's ranked second, and so on.

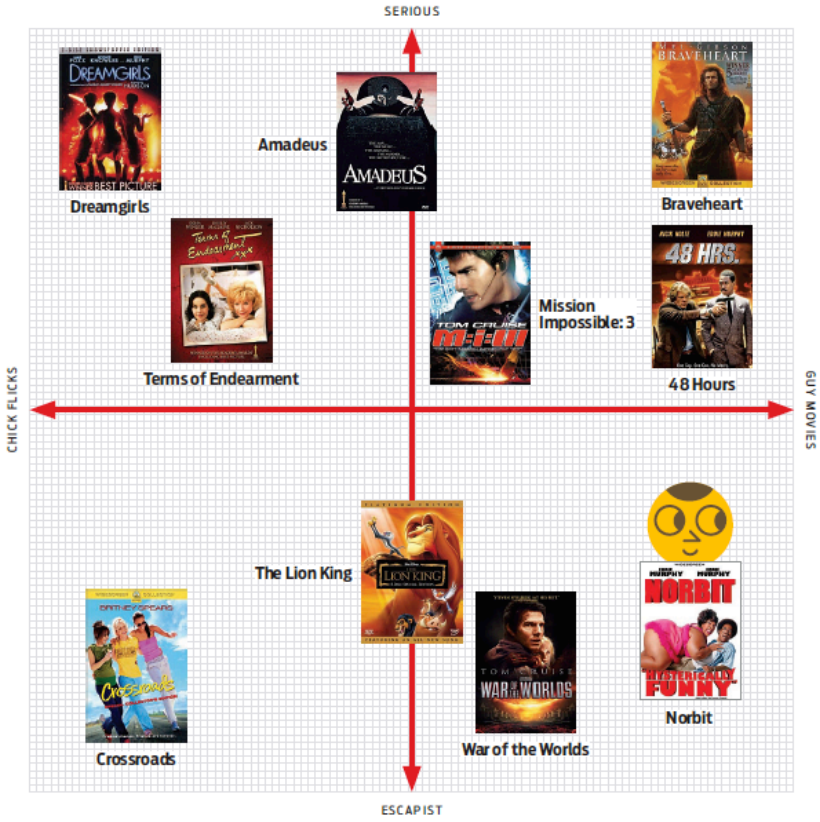
PHOTOS: PARAMOUNT PICTURES



The Latent-Factor Approach

A SECOND, COMPLEMENTARY method scores both a given movie and viewer according to latent factors, themselves inferred from the ratings given to all the movies by all the viewers. The factors define a space that at once measures the characteristics of movies and the viewer's interest in those characteristics. Here we would expect the fellow in the southeast corner of the graph to love *Norbit*, to hate *Dreamgirls*, and, perhaps, to rate *Braveheart* about average.

PHOTOS: AMADEUS, THE SAUL ZAENTZ COMPANY; ALL OTHERS, PARAMOUNT PICTURES



From Bell et al (May 2009), *IEEE Spectrum*

Bespoke News (2008)

- Google News tackles this problem by using a technique called hierarchical agglomerative clustering.
- Because of the enormous number of articles and users on Google News, traditional clustering methods were impractical.
- So Google tested three more advanced algorithms for generating news recommendations: MinHash clustering, Probabilistic Latent Semantic Indexing (PLSI), and covisitation counts.

From Linden (2008) *IEEE Spectrum*

Gaming: Kinect (2010)

- **MACHINE LEARNING BREAKTHROUGH**
  Microsoft's Xbox group in Redmond, Wash., which had been developing Kinect for several years, enlisted [Andrew] Blake's help in 2008. Redmond had come up with a rough prototype for tracking a player in real time, but the system had encountered problems. It relied on a computer graphics program to construct an avatar, which the program would then constantly adjust to match images of the player taken by Kinect's 3-D Web camera as he or she moved.
- But the method had flaws. The system would lose track of a person after a short time, could generally only track someone who was about the same size (relative in terms of scale) and shape as the avatar and couldn't process rapid movements.
- Help came in the form of machine learning. "The Xbox team had used a computer graphics approach, but they didn't know about machine learning," notes Blake. "The researchers at our lab are expert in this area."
- Blake himself had been working for years on real-time motion tracking techniques using machine learning. He and Microsoft researcher Kentaro Toyama published a paper in 2001, "Probabilistic Tracking in a Metric Space," describing a novel approach that assigned a probabilistic likelihood that each movement would lead to another specific type of movement. This data was fed into a computer program that automatically calculated the most likely next move. This was a breakthrough for machine learning, but it wasn't able to fix Kinect.
- Jamie Shotton, a Cambridge researcher, proposed a solution. He suggested **teaching the machine learning system to distinguish one part of a person's body from another**. It only took about three months for Shotton to demonstrate that his method would work. Over the next few months, Blake's team collaborated with the Xbox group's engineers in an extensive machine learning project. The machine learning system was "taught" to recognize people in all shapes and sizes and in many different poses. To do this, the researchers uploaded more than a million images of different people in different positions. To teach the system how to recognize body parts, the team used a computer graphics algorithm to render color-coded images representing the different body parts. They eventually created a machine learning algorithm that could analyze each pixel in an image and determine which limb it was.

Source: Bogdanowicz (Sept 2011), The IEEE Institute

Speech Recognition

- Currently available commercial systems for speech recognition all use machine learning in one fashion or another to train the system to recognize speech. The reason is simple: the speech recognition accuracy is greater if one trains the system, than if one attempts to program it by hand. In fact, many commercial speech recognition systems involve two distinct learning phases: one before the software is shipped (training the general system in a speaker-independent fashion), and a second phase after the user purchases the software (to achieve greater accuracy by training in a speaker-dependent fashion).

Bio-Surveillance

- A variety of government efforts to detect and track disease outbreaks now use machine learning. For example, the RODS project involves real-time collection of admissions reports to emergency rooms across western Pennsylvania, and the use of machine learning software to learn the profile of typical admissions so that it can detect anomalous patterns of symptoms and their geographical distribution.
- Also see the MITACS Centre for Disease Modeling, led by York's Jianhong Wu.

Autonomous Vehicles (2005)

- Stanford's vehicle Stanley won the 2005 DARPA Grand Challenge, relying upon a range of machine learning algorithms.
- Example: Classifying driveable terrain:
  - Adaptive mixture of Gaussians colour model
  - Expectation-Maximization

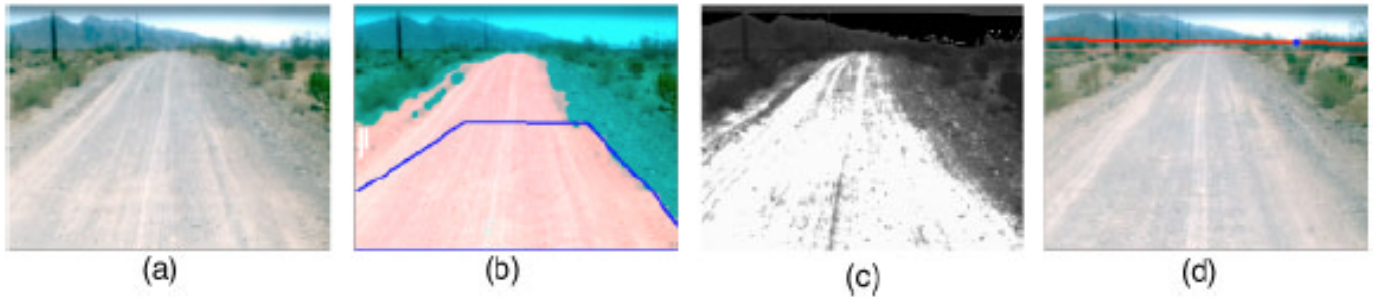


Figure 15. This figure illustrates the processing stages of the computer vision system: (a) a raw image; (b) the processed image with the laser quadrilateral and a pixel classification; (c) the pixel classification before thresholding; and (d) horizon detection for sky removal.

Source: Thrun et al (2005), Journal of Field Robotics

Google's Driverless Car Project (2009)

Empirical Sciences

- Many data-intensive sciences now make use of machine learning methods to aid in the scientific discovery process. Machine learning is being used to
  - learn models of gene expression in the cell from high-throughput data
  - discover unusual astronomical objects from massive data collected by the Sloan sky survey
  - to characterize the complex patterns of brain activation that indicate different cognitive states of people in fMRI scanners.

Is Machine Learning Important?

- Yes!

For more details…

- Supplementary readings are posted at http://moodle.yorku.ca:
  - Mitchell, T.M. (2006). The Discipline of Machine Learning, CMU-ML-06-108
  - Bogdanowicz, A. (2011). The Motion Tech Behind Kinect, *IEEE Institute*, Jan 6, 2011
  - Bell, R.M., Bennett J., Koren, Y. & Volinsky, C. (2009) The Million Dollar Programming Prize, *IEEE Spectrum*, May 2009, 28-33
  - Linden, G. (2008) People who read this article also read..., *IEEE Spectrum*, March 2008, 46-60

More on Moodle

- I will create a Discussion Forum for every lecture.
- Please use this as a handy resource to discuss the material with your fellow students.