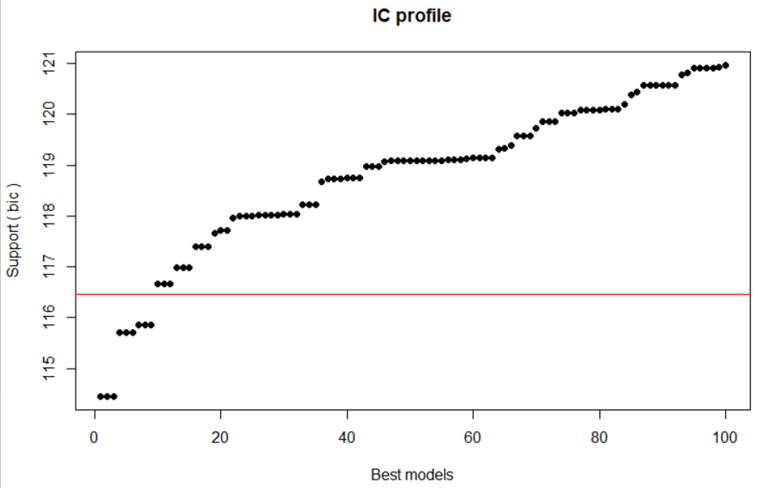
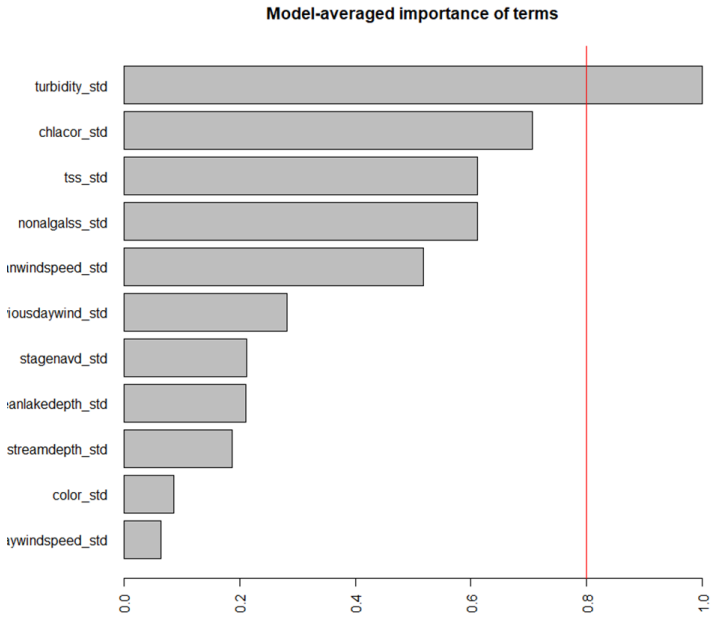
Advanced R: Statistical Machine Learning

Dan Schmutz, MS
Chief Environmental Scientist

Zoom Workshop for SJRWMD
September 24, 2020





Topics

- What is Statistical Machine Learning?
- Fundamental Challenges for Appropriate Application of Statistical Machine Learning
- Supervised Learning Algorithms and Approaches
  - Linear Regression
  - Regularized Regression
  - Logistic Regression
  - k-nearest neighbors
  - Decision Trees
  - Random Forest
- Overview of Unsupervised Learning Algorithms
- Learning Competition

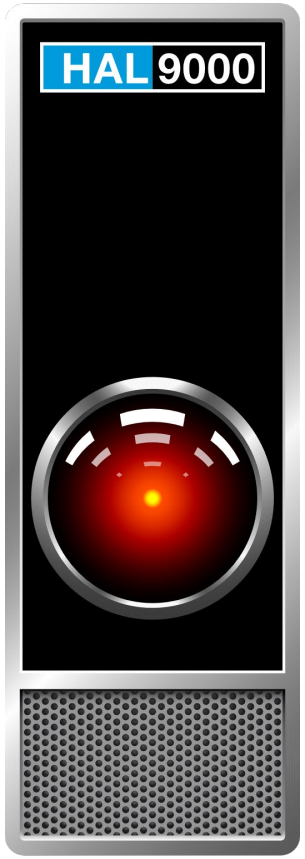
GPI

What is Statistical Machine Learning?

GPI

What is Artificial Intelligence?

What 2001 was supposed to look like



By Grafiker61 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=46424786

What 2004 actually looked like

By DARPA - This file was derived from: DARPA Strategic Plan (2007).pdf, Public Domain, https://commons.wikimedia.org/w/index.php?curid=20798332

GPI

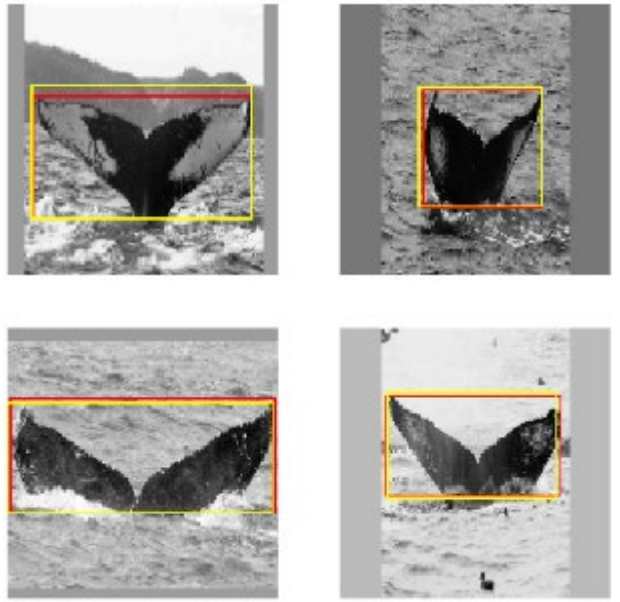
What is Artificial Intelligence (AI)?

- The study of "intelligent agents": devices that perceive their environment and take actions to maximize their chances of successfully achieving their goals.
- Once challenging problems for machines are now commonplace: e.g., Optical Character Recognition (OCR)
- Current examples of AI problem areas
  - understanding human speech
  - dominating strategic games (e.g., chess and Go)
  - autonomously operating cars
- Many diverse subfields, including:
  - Machine Learning
  - Machine Vision and Hearing
  - Robotics
  - Information Retrieval
  - Expert Systems

GPI

What is Machine Learning?

- A subfield of artificial intelligence that develops algorithms that can learn from and make predictions on data*
- Focus is more on prediction than explanation
- Applied widely in business and scientific research
  - Price estimation (Zillow)
  - Product recommendations (e.g., Netflix, Amazon)
  - Google image segmentation and identification
  - Consumer credit scoring
  - Handwriting and object recognition
  - Drug development
  - Bioinformatics



*Alpaydin, E. 2009. Introduction to Machine Learning, Second Edition. MIT Press.

GPI

What is an algorithm?

- a set of rules to be followed in problem-solving operations (e.g., long division).

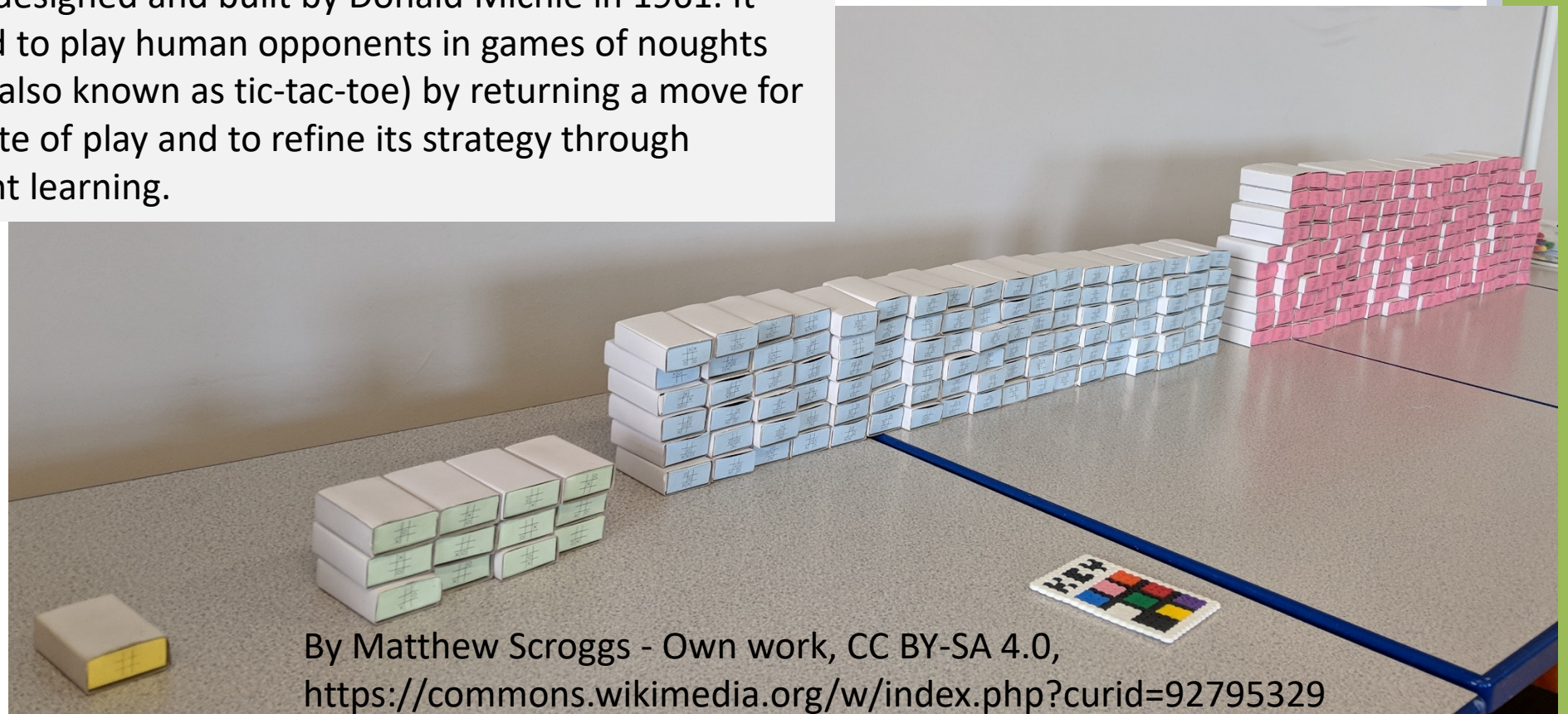
```
  31.75
4)127.00
  12          (12 ÷ 4 = 3)
   07         (0 remainder, bring down next figure)
    4         (7 ÷ 4 = 1 r 3)
    3.0       (bring down 0 and the decimal point)
    2.8       (7 × 4 = 28, 30 ÷ 4 = 7 r 2)
      20      (an additional zero is brought down)
      20      (5 × 4 = 20)
       0
```

https://en.wikipedia.org/wiki/Long_division

GPI

MENACE

The Matchbox Educable Noughts and Crosses Engine (sometimes called the Machine Educable Noughts and Crosses Engine) or MENACE was an analogue computer made up of 304 matchboxes designed and built by Donald Michie in 1961. It was designed to play human opponents in games of noughts and crosses (also known as tic-tac-toe) by returning a move for any given state of play and to refine its strategy through reinforcement learning.



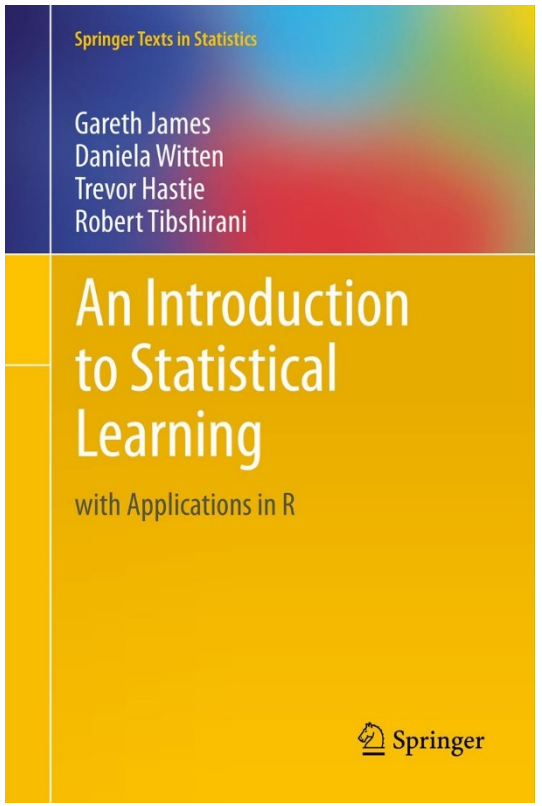
By Matthew Scroggs - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=92795329

GPI

What is “Statistical” Machine Learning?

- Arguably the most current buzzword, acknowledging the mashup of machine learning and statistics
- Statistical Learning describes the paradigm of “learning from data” of algorithms and techniques that learn from observed data by constructing stochastic models that can be used for making predictions and decisions.

Wow! Another Amazing Free Book



http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

GPI

To explain or predict? Must we choose?

- Statistical model building can be useful for
  - Explanation – testing casual hypotheses or at least identifying important factors worthy of further experimental manipulation
  - Prediction – simply making algorithms that perform well on out-of-sample (i.e., test data)

- In my opinion, statistical machine learning supports both perspectives

Galit Shmueli. 2010. To Explain or to Predict? Statistical Science 2010, Vol. 25, No. 3, 289–310 DOI: 10.1214/10-STS330© Institute of Mathematical Statistics. Online resource, accessed 9/20/2020: https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf

GPI

Modeling Y as a function of X

Y

- “target variable”
- “dependent variable”
- “response”
- “outcome measurement”
- “output”

X

- “predictor variable(s)”
- “independent variable(s)”
- “attribute(s)”
- “feature(s)”
- “predictor(s)”
- “auxiliary variables(s)”

X can represent one, but usually many variables in machine learning algorithms.

GPI

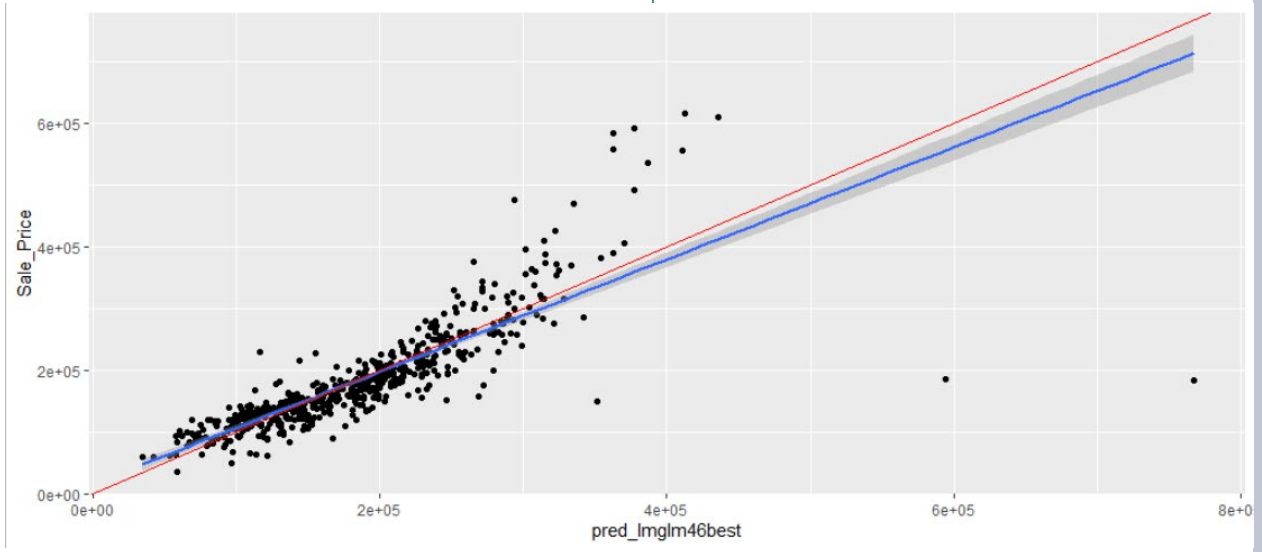
Types of Statistical Machine Learning

- Supervised – labeled cases are given to the algorithm
- Unsupervised – algorithm finds structure in the data

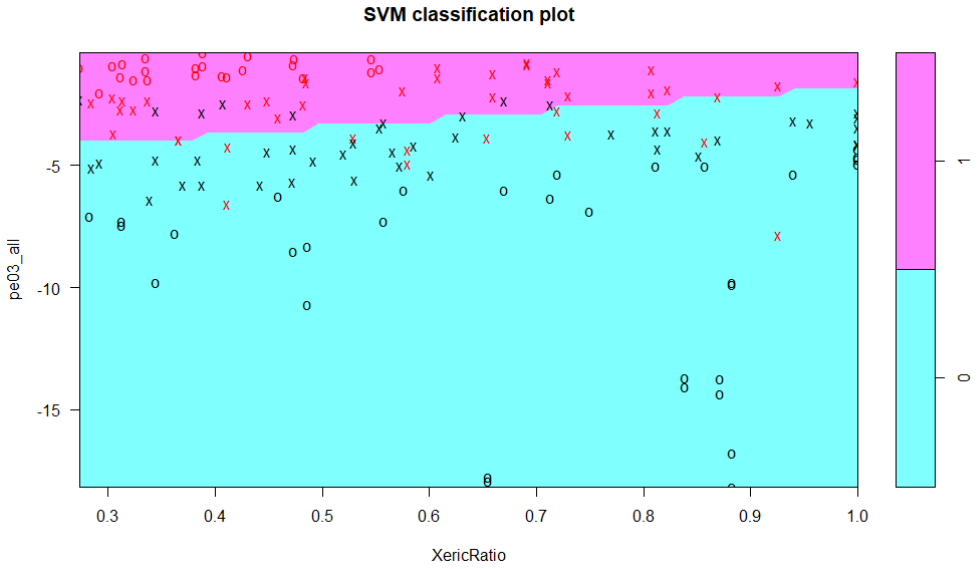
GPI

Supervised Learning

- Regression -continuous output



- Classification – categorical output



Unsupervised Learning

- Clustering – segment observations (i.e., rows of dataframe) into similar groups based on the observed variables, e.g.:
  - K-means clustering
  - Hierarchical clustering

- Dimension reduction – reducing the number of variables in a data set (i.e., columns of dataframe), e.g.,:
  - PCA (Principal Components Analysis)