Introduction to Machine Learning
[Fall 2022]

Introduction to Gradients

October 13, 2022

Lerrel Pinto

Topics for today

- Recap of the class so far
- Bird’s eye view of optimization
- Fundamentals for gradients

Recap of the class so far

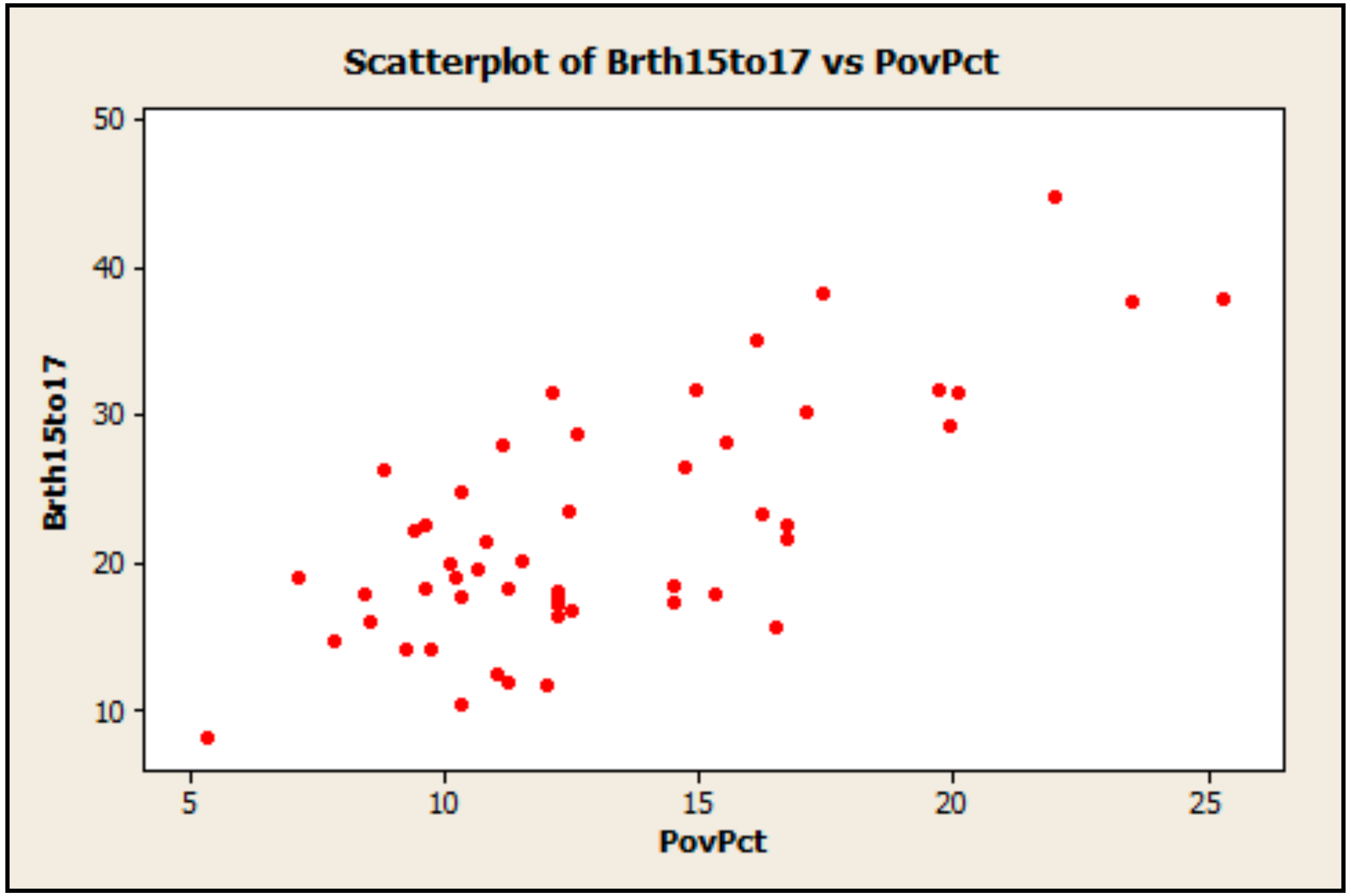
Recap: Linear Regression

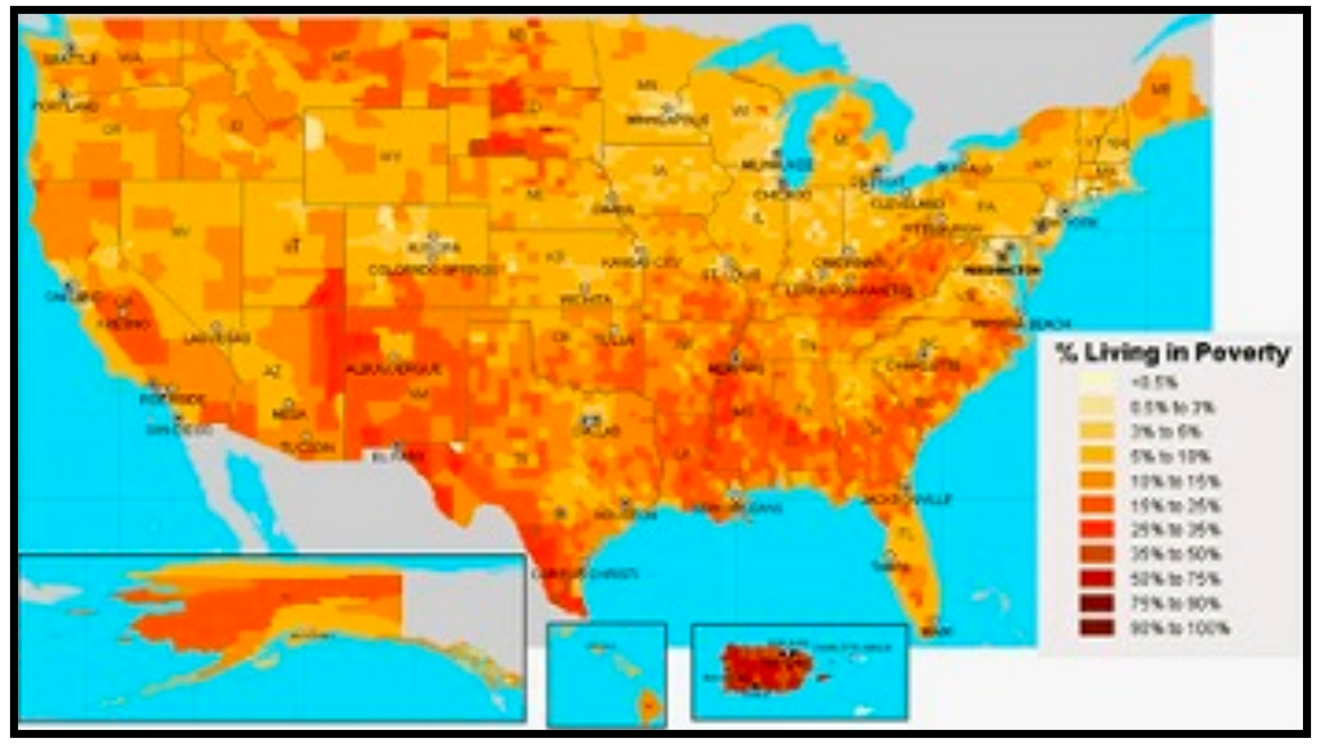
- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^{n}$, where $(\vec{x} \in \mathbb{R}^{d}, y \in \mathbb{R}^{1})$ corresponds to a data point.
  - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.

Recap: Linear Regression

- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^{n}$, where $(\vec{x} \in \mathbb{R}^{d}, y \in \mathbb{R}^{1})$ corresponds to a data point.
  - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.
- Goal: to find $\vec{w} \in \mathbb{R}^{d}$ such that $< \vec{w}, \vec{x} > = y$
  - Minimize $||X^{\top}\vec{w} - Y||^{2}$

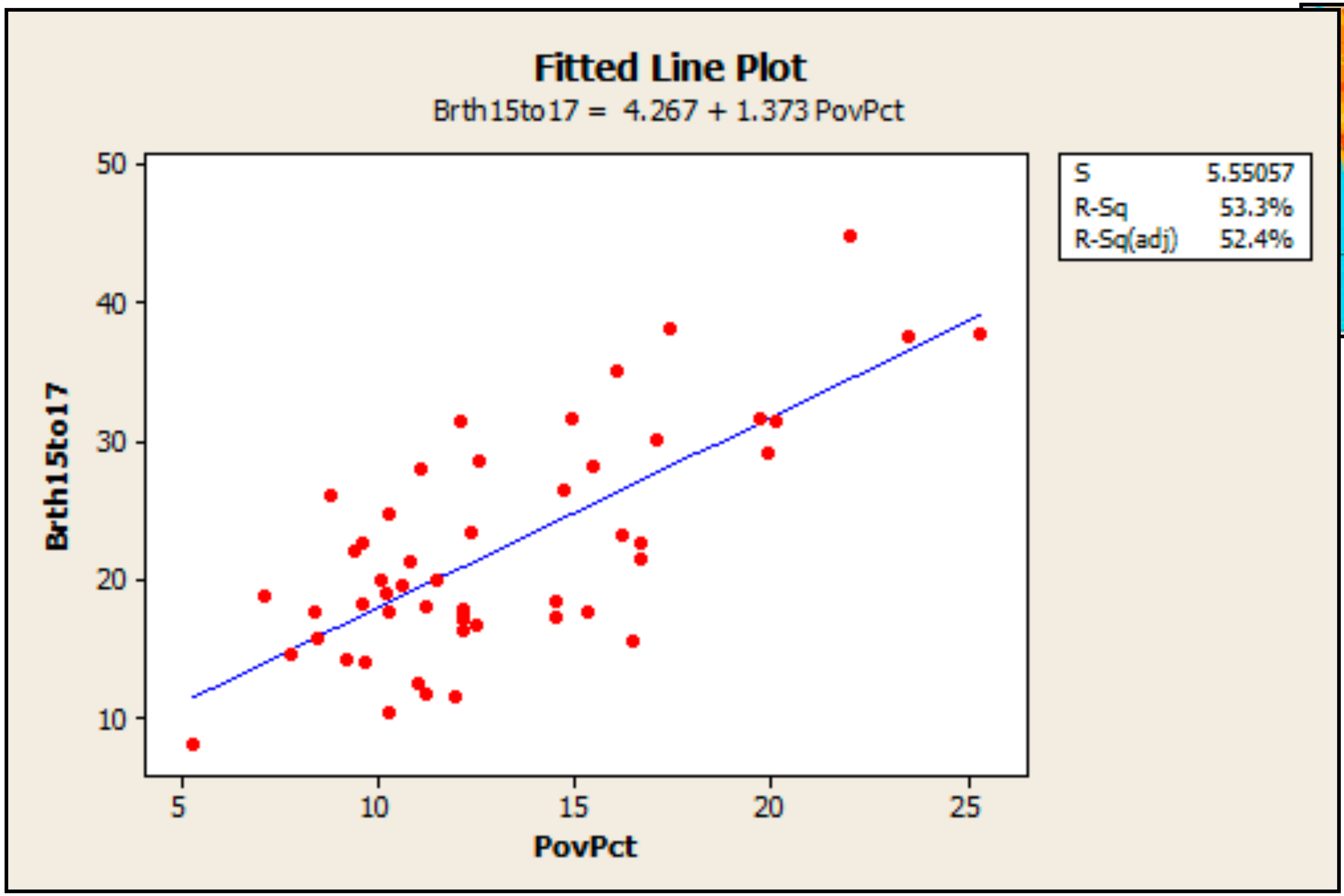
Recap: Linear Regression

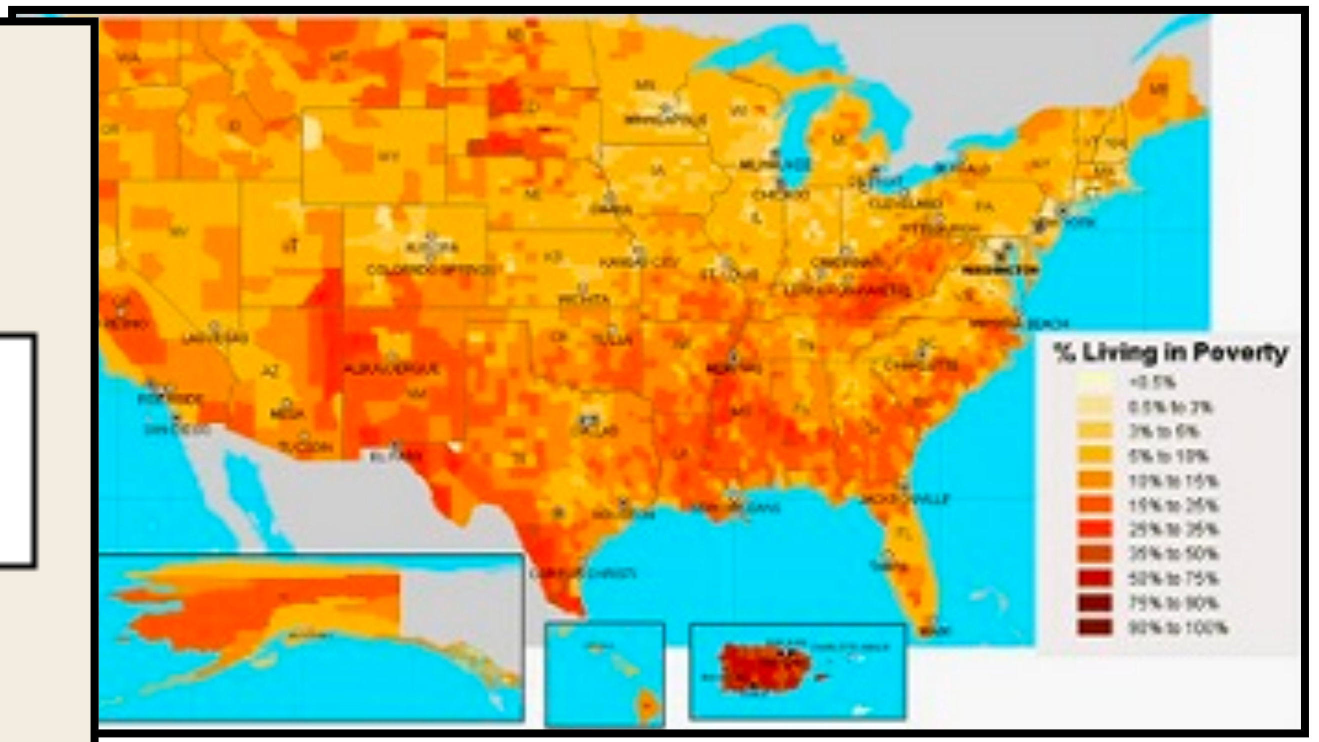




https://online.stat.psu.edu/stat462/node/101/

Recap: Linear Regression





https://online.stat.psu.edu/stat462/node/101/

Recap: Linear Regression

- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^n$, where $(\vec{x} \in \mathbb{R}^d, y \in \mathbb{R}^1)$ corresponds to a data point.
  - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.
- Goal: to find $\vec{w} \in \mathbb{R}^d$ such that $< \vec{w}, \vec{x} > = y$
  - Minimize $||X^\top \vec{w} - Y||^2$
- Solution: $\vec{w} = (XX^\top)^{-1}XY$
  - Easy way to remember $\vec{w} = (X^\top)^+ Y$

Recap: Linear Regression

Diabetes Metab Syndr. 2020 September-October; 14(5): 1467–1474.
Published online 2020 Aug 1. doi: 10.1016/j.dsx.2020.07.045
PMCID: PMC7395225
PMID: 32771920

Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model

Smita Rath,[a,*] Alakananda Tripathy,[a] and Alok Ranjan Tripathy[b]

‣ Author information ‣ Article notes ‣ Copyright and License information Disclaimer

Using News Articles to Predict Stock Price Movements

Győző Gidófalvi
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92037
gyozo@cs.ucsd.edu
2001, June 15, 2001

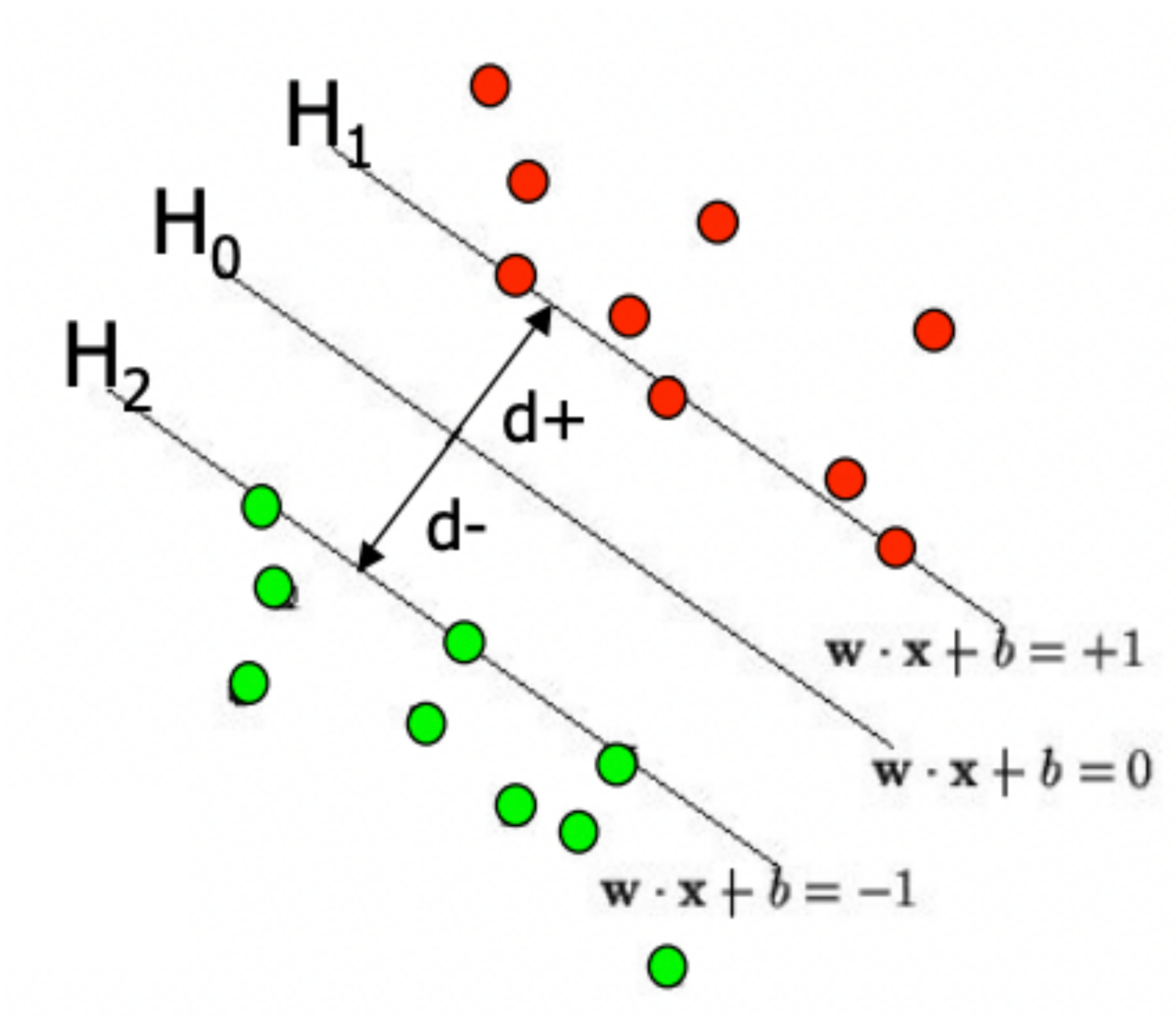
Article PDF Available

Are newspapers' news stories becoming more alike? Media content diversity in Belgium, 1983–2013

May 2017 · Journalism 20(12)
DOI:10.1177/1464884917706860

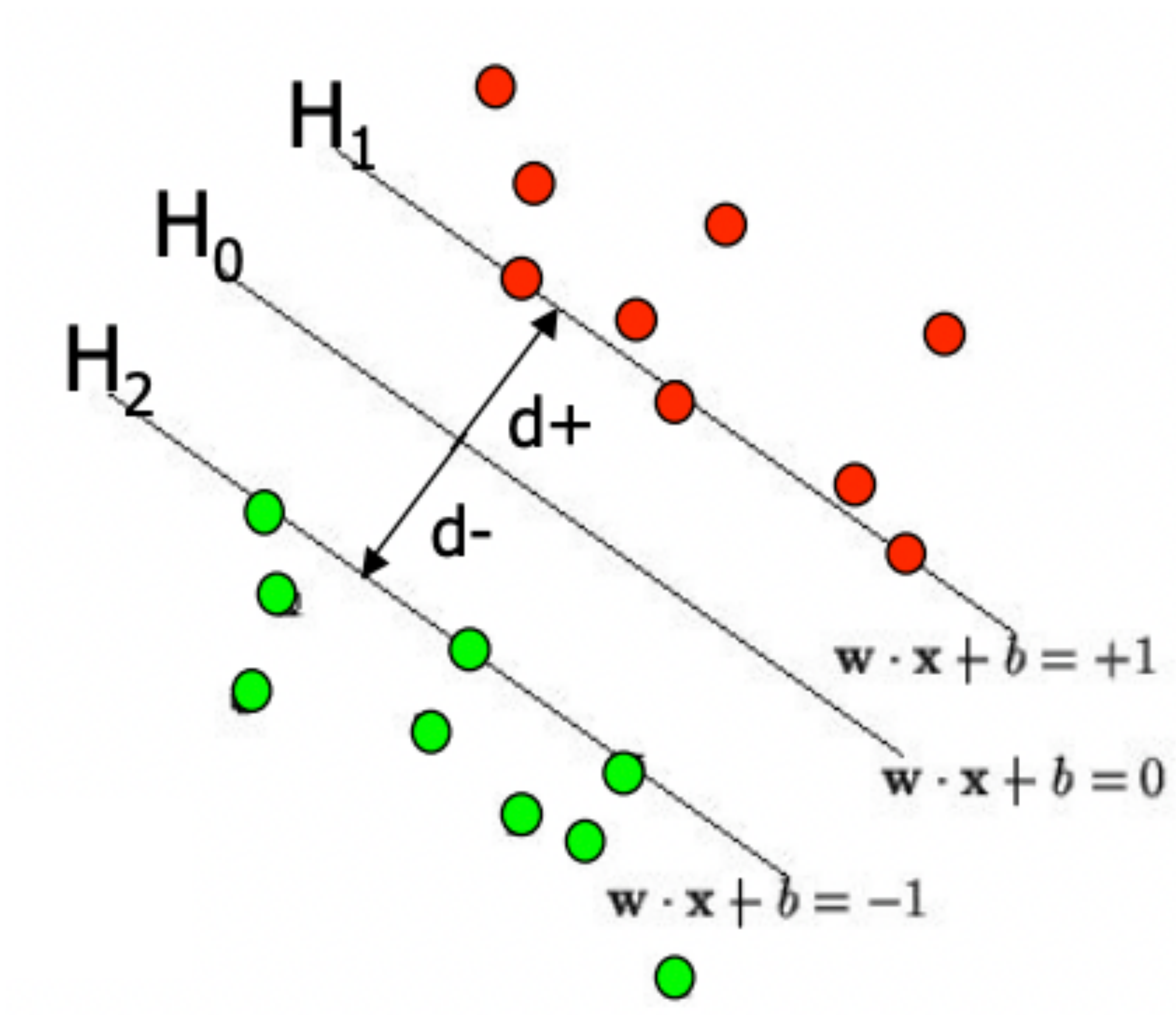
Authors:
Kathleen Beckers
University of Antwerp
Andrea Masini
Julie Sevenans
University of Antwerp
Miriam van der Burg
University of Antwerp

Recap: Classification with SVMs



Credits: R. Berwick (https://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf)

Recap: Classification with SVMs



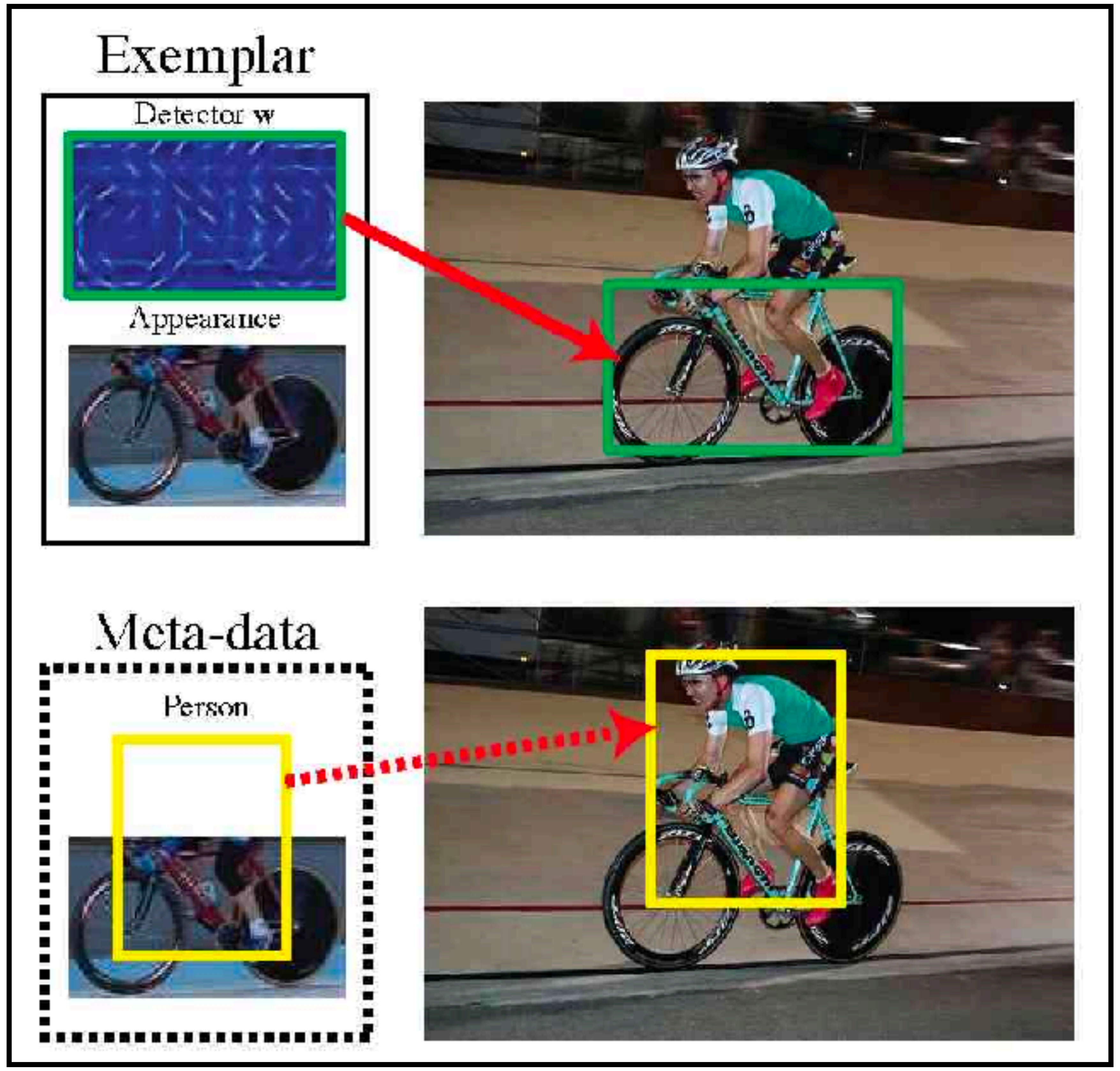
- Goal: Maximize margin / Minimize $||w||^2$
  - Also need to satisfy $y^i f(x^i) \geq 1$ for all datapoints (x^i, y^i).

$$\boxed{\min_w ||w||^2 \text{ subject to } y^i(w^T x^i + b) \geq 1}$$

- Can be solved as a quadratic optimization problem with linear constraints.

Credits: R. Berwick (https://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf)

Recap: Classification with SVMs

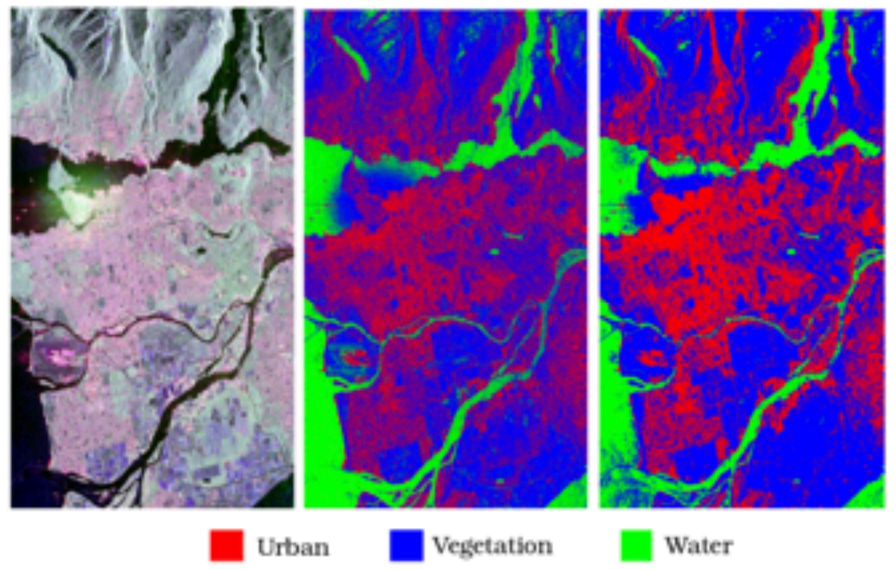


Supervised Classification of RADARSAT-2 Polarimetric Data for Different Land Features

Abhishek Maity

Abstract—The pixel percentage belonging to the user defined area that are assigned to cluster in a confusion matrix for RADARSAT-2 over Vancouver area has been analysed for classification. In this study, supervised Wishart and Support Vector Machine (SVM) classifiers over RADARSAT-2 (RS2) fine quad-pol mode Single Look Complex (SLC) product data is computed and compared. In comparison with conventional single channel or dual channel polarization, RADARSAT-2 is fully polarimetric, making it to offer better land feature contrast for classification operation.

Index Terms—Wishart, Support Vector Machine (SVM), Confusion matrix



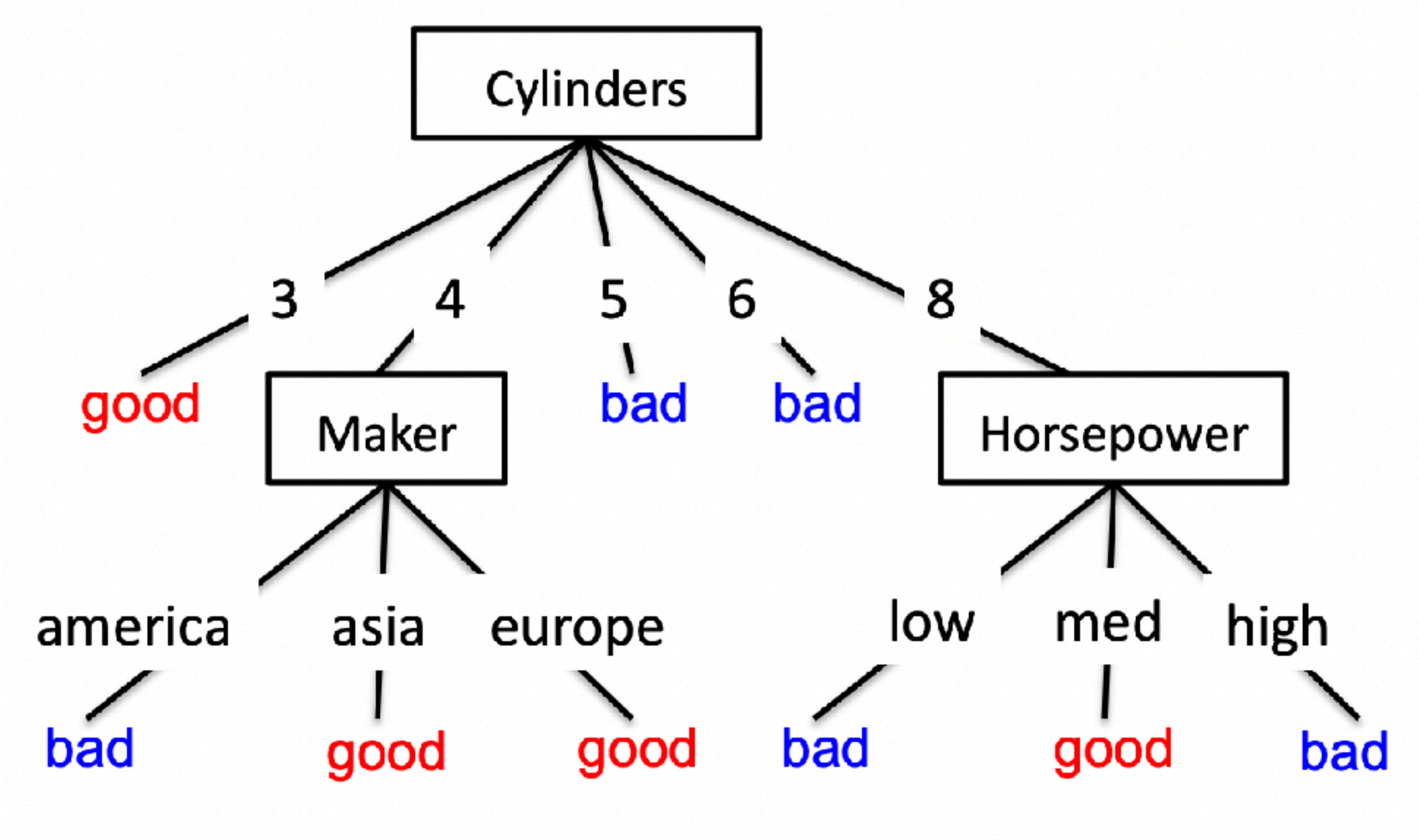
I. Introduction

Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome

Rémi Cuingnet [a,b,c,d,*,1], Charlotte Rosso [a,b,c,e], Marie Chupin [a,b,c], Stéphane Lehéricy [a,b,c,f], Didier Dormont [a,b,c,f], Habib Benali [d], Yves Samson [a,b,c,e], Olivier Colliot [a,b,c]

Recap: Classification with Decision Trees

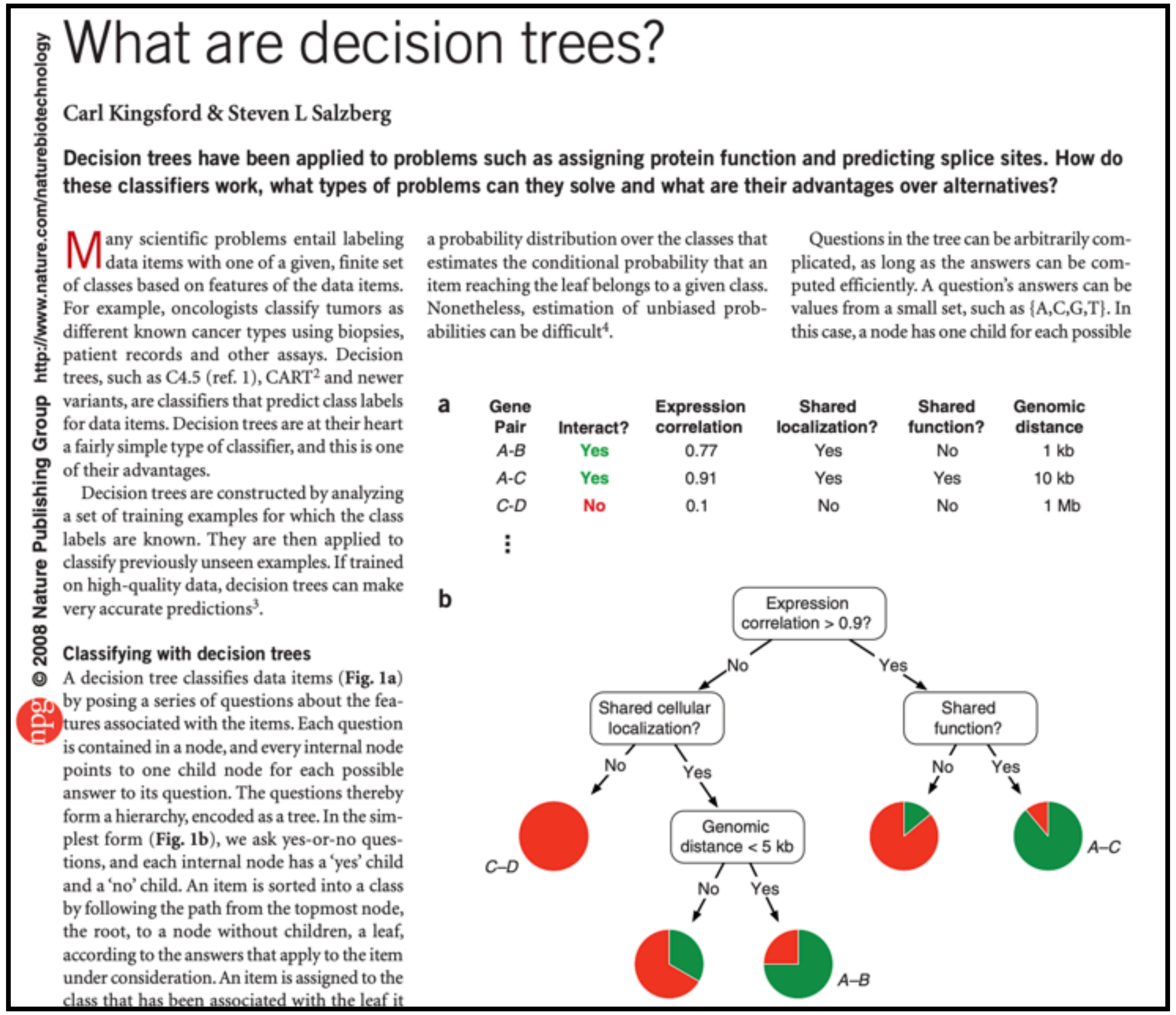
mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe



$f(x) :=$ cyl=3 ∨ (cyl=4 ∧ (maker=asia ∨ maker=europe)) ∨ …

Slide credits: David Sontag

Recap: Classification with Decision Trees



What are decision trees?

Carl Kingsford & Steven L Salzberg

Decision trees have been applied to problems such as assigning protein function and predicting splice sites. How do these classifiers work, what types of problems can they solve and what are their advantages over alternatives?

Many scientific problems entail labeling data items with one of a given, finite set of classes based on features of the data items. For example, oncologists classify tumors as different known cancer types using biopsies, patient records and other assays. Decision trees, such as C4.5 (ref. 1), CART[2] and newer variants, are classifiers that predict class labels for data items. Decision trees are at their heart a fairly simple type of classifier, and this is one of their advantages.

Decision trees are constructed by analyzing a set of training examples for which the class labels are known. They are then applied to classify previously unseen examples. If trained on high-quality data, decision trees can make very accurate predictions[3].

Classifying with decision trees

A decision tree classifies data items (**Fig. 1a**) by posing a series of questions about the features associated with the items. Each question is contained in a node, and every internal node points to one child node for each possible answer to its question. The questions thereby form a hierarchy, encoded as a tree. In the simplest form (**Fig. 1b**), we ask yes-or-no questions, and each internal node has a 'yes' child and a 'no' child. An item is sorted into a class by following the path from the topmost node, the root, to a node without children, a leaf, according to the answers that apply to the item under consideration. An item is assigned to the class that has been associated with the leaf it

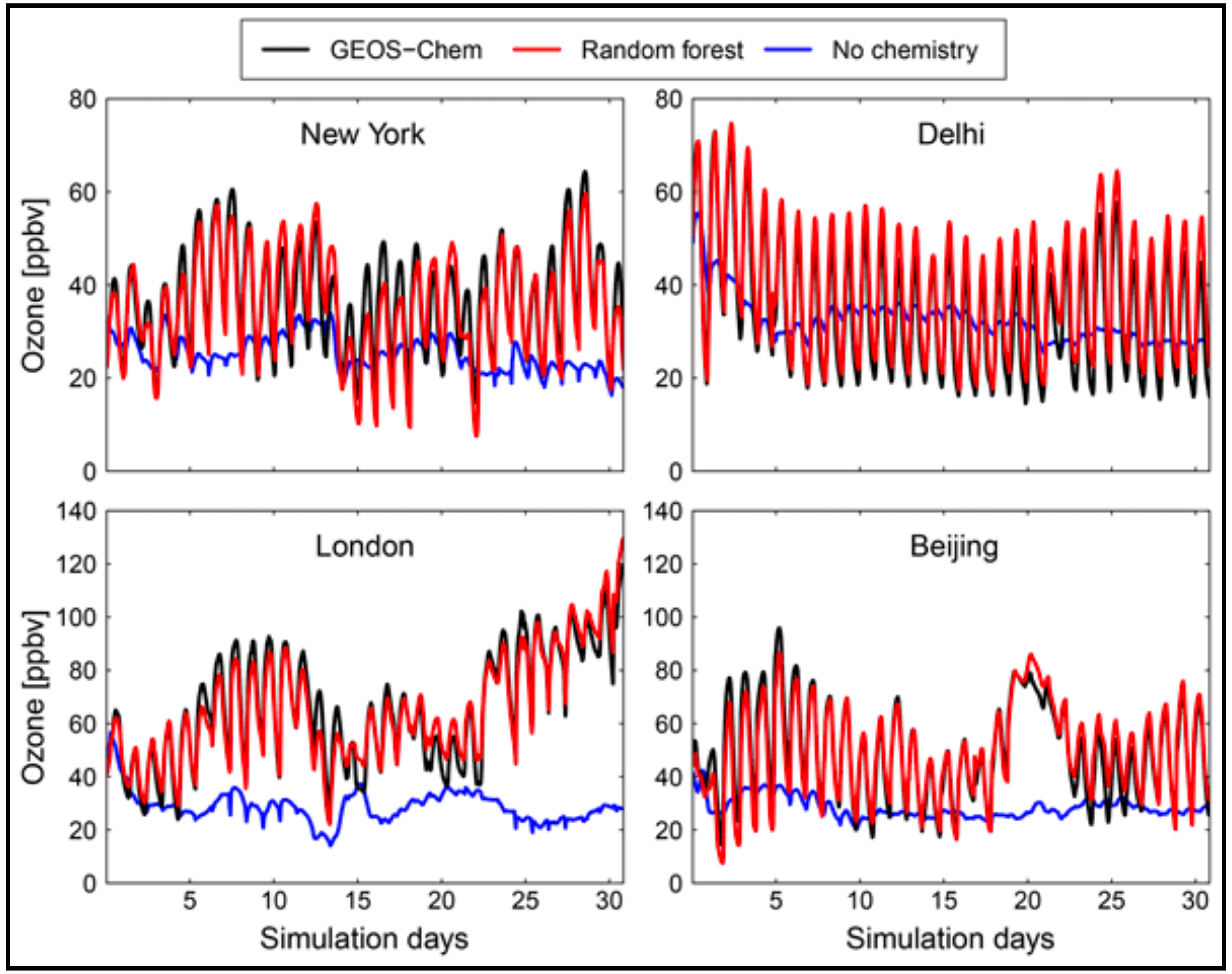
a probability distribution over the classes that estimates the conditional probability that an item reaching the leaf belongs to a given class. Nonetheless, estimation of unbiased probabilities can be difficult[4].

Questions in the tree can be arbitrarily complicated, as long as the answers can be computed efficiently. A question's answers can be values from a small set, such as {A,C,G,T}. In this case, a node has one child for each possible

a

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb
⋮					

b



Key principle so far

- Convert data to a set of equations
- Solve equations either directly (linear reg., SVM) or recursively (Decision Trees)
- What is the problem?
    - What happens when input is high-dimensional?
    - What happens when the dataset is huge (~1M-1B examples)?

Bird’s eye view of optimization

How much water to give your plant?

Plant watering as an optimization problem

Case 0: Brute-force search

Case 1: No analytic model of the plant

Case 1: No analytic model of the plant

Black-box / Derivative free optimization.

Case 2: Analytic model is available

Case 3: Analytic model is available, but difficult to directly minimize

Gradient Descent

Issues with Gradient Descent

- How to figure out the learning rate?

Newton's method in Optimization

- How to figure out the learning rate?

What about constraints?

Multiple solutions?

Is the solution optimal?

Summary for optimization

- http://www.lewissoft.com/pdf/INTRO_OPT.pdf
- Do not have derivatives? – Use black-box optimization.
- Have 'simple' cost function? – Use analytic methods, quadratic programming, convex optimization techniques.
- Have a challenging cost function? – Use derivative-based optimization.

Questions?