Introduction to Active Learning

Rob Zinkov

January 18th, 2011

Outline

What is Active Learning?

Active Learning is a form of semi-supervised learning
Allows classification with exponentially fewer labelled instances

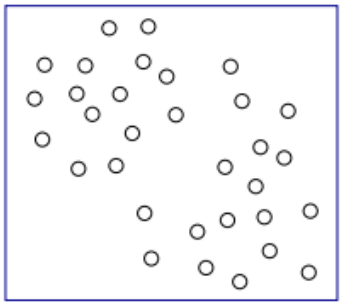
Goal of this talk

- Explain on a high-level Active Learning
- Show that Active Learning is a paradigm worth exploring

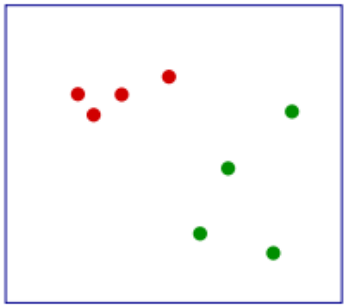
Caveats

- This talk won't prove the bounds it claims
- Active Learning best on well-defined hypothesis spaces
- We assume a low-noise setting where classifying is cheap and labeling is expensive

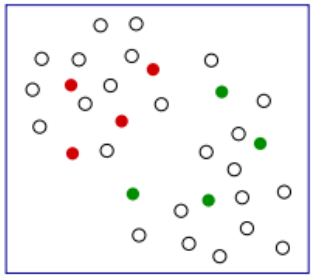
Defintions



Unlabeled points



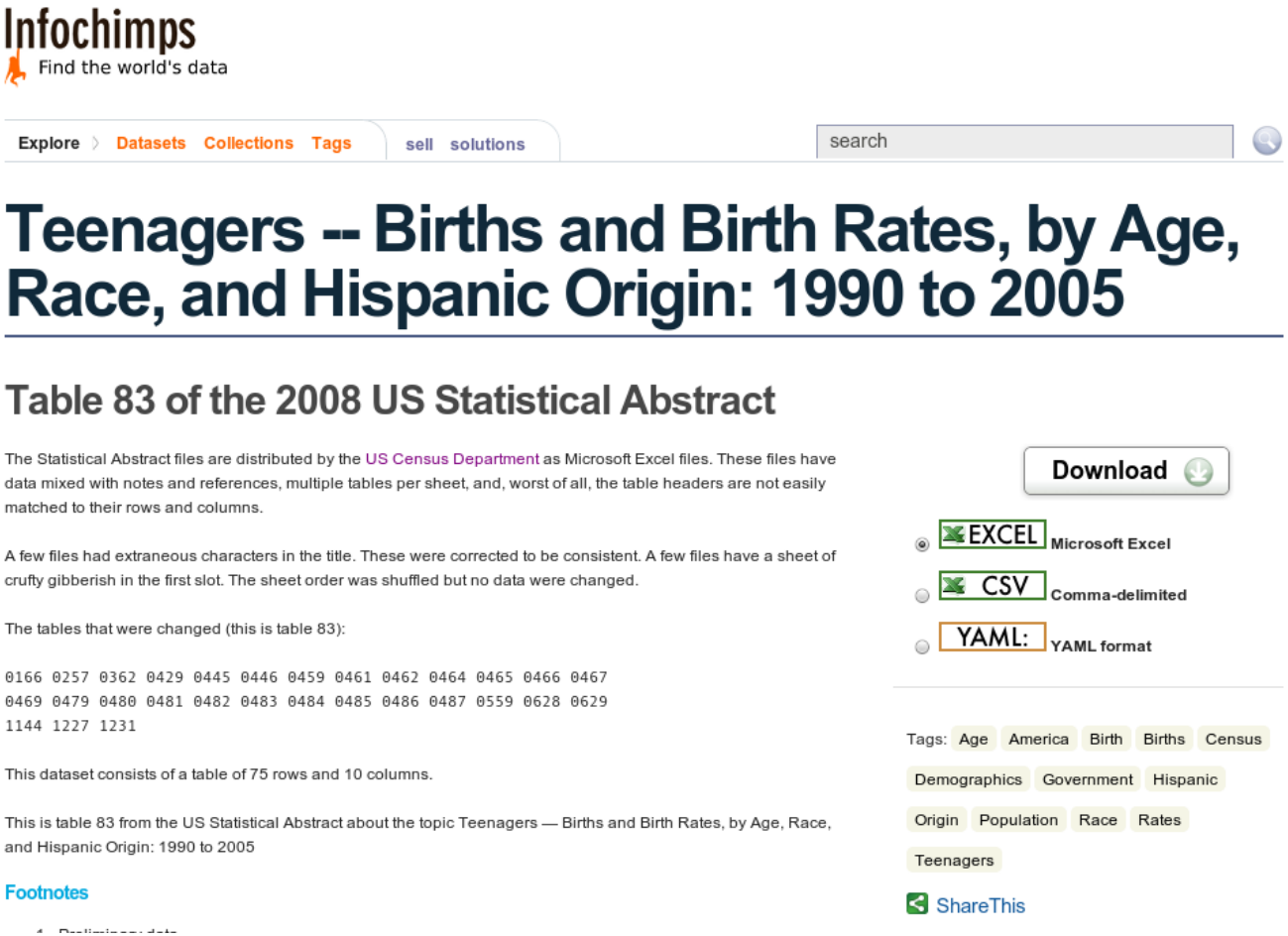
Supervised learning



Semisupervised and active learning

Traditional Supervised Learning

Collect some data



Infochimps
Find the world's data

Explore › Datasets Collections Tags sell solutions search

Teenagers -- Births and Birth Rates, by Age, Race, and Hispanic Origin: 1990 to 2005

Table 83 of the 2008 US Statistical Abstract

The Statistical Abstract files are distributed by the US Census Department as Microsoft Excel files. These files have data mixed with notes and references, multiple tables per sheet, and, worst of all, the table headers are not easily matched to their rows and columns.

A few files had extraneous characters in the title. These were corrected to be consistent. A few files have a sheet of crufty gibberish in the first slot. The sheet order was shuffled but no data were changed.

The tables that were changed (this is table 83):

0166 0257 0362 0429 0445 0446 0459 0461 0462 0464 0465 0466 0467
0469 0479 0480 0481 0482 0483 0484 0485 0486 0487 0559 0628 0629
1144 1227 1231

This dataset consists of a table of 75 rows and 10 columns.

This is table 83 from the US Statistical Abstract about the topic Teenagers — Births and Birth Rates, by Age, Race, and Hispanic Origin: 1990 to 2005

Footnotes

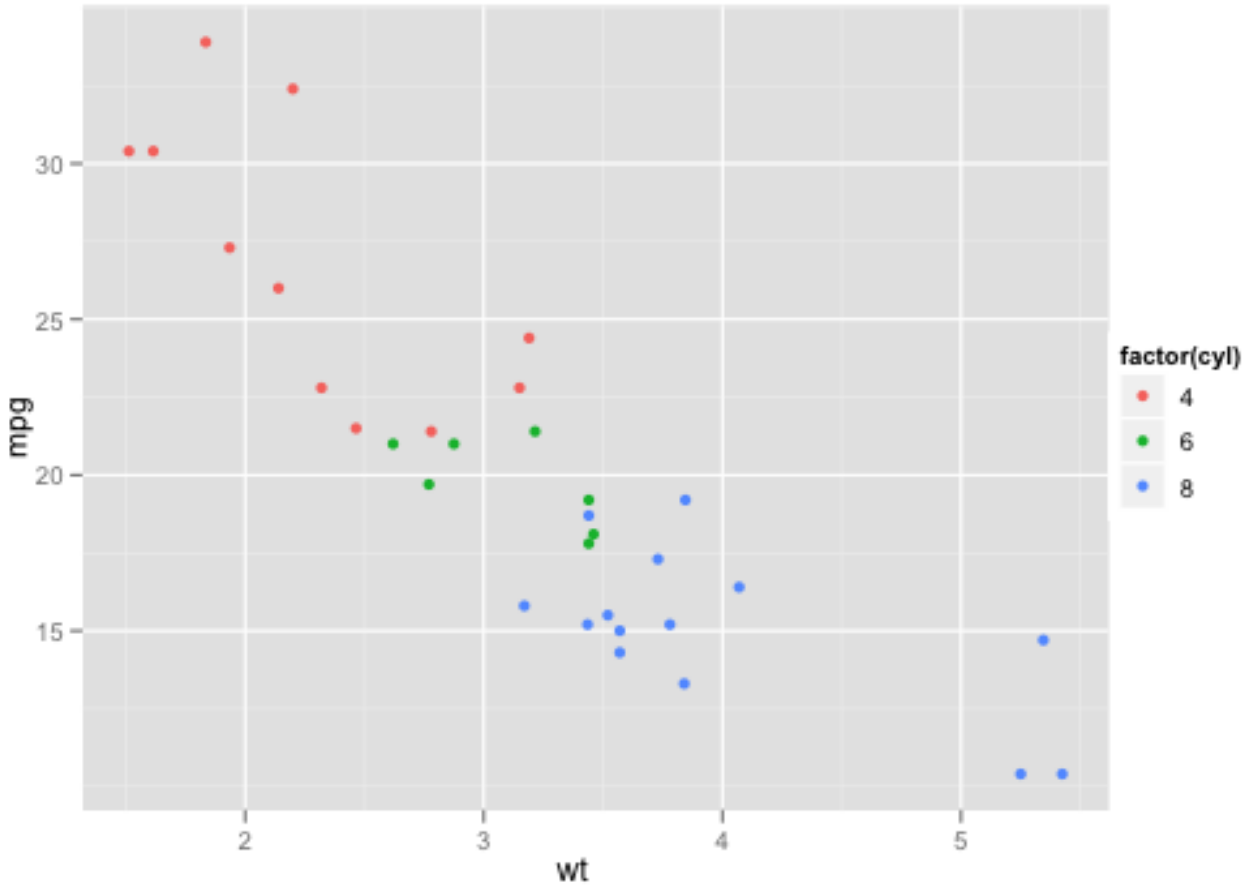
Download

EXCEL Microsoft Excel
CSV Comma-delimited
YAML: YAML format

Tags: Age America Birth Births Census Demographics Government Hispanic Origin Population Race Rates Teenagers

ShareThis

Eyeball it



Pick a loss function

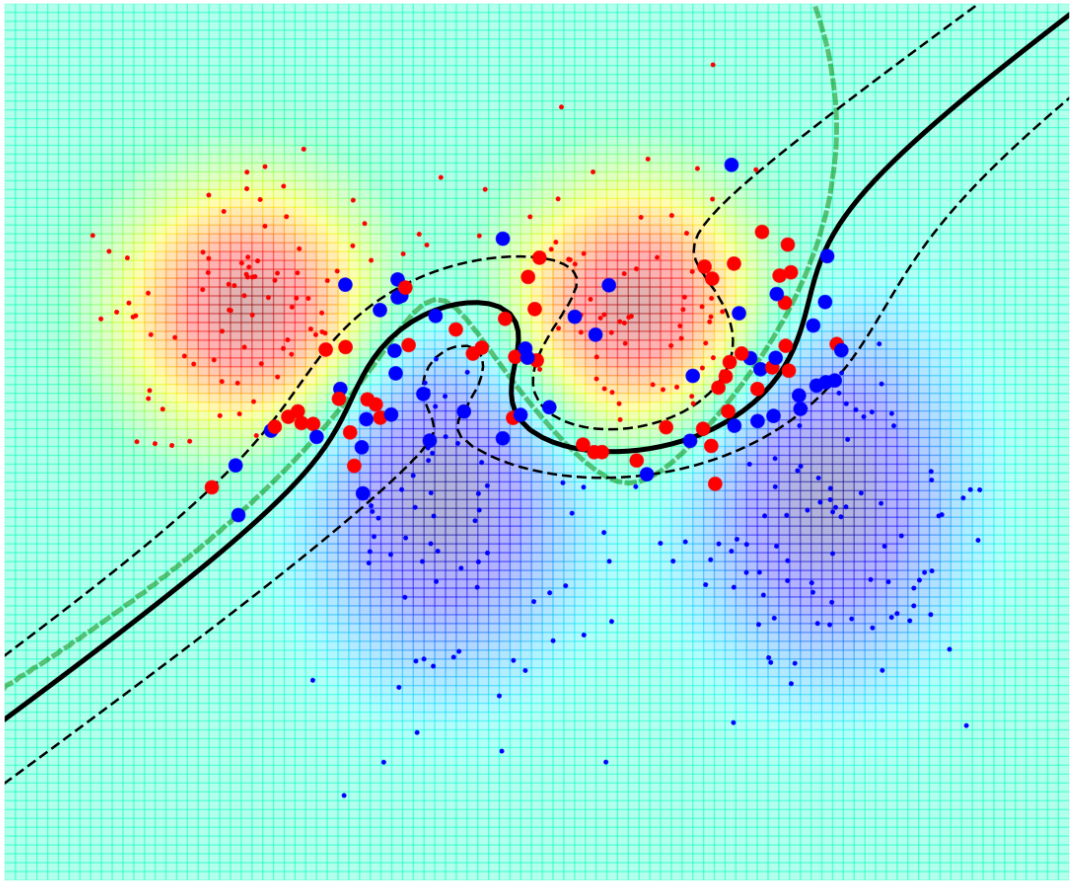
$$\mathcal{L}(x, y, \hat{x}) = \begin{cases} 0 \text{ if } y = \hat{x} \\ 1 \text{ otherwise} \end{cases}$$

$$\mathcal{L}(x, y, \hat{x}) = (x - \hat{x})^2$$

$$\mathcal{L}(x, y, \hat{x}) = |x - \hat{x}|$$

$$\mathcal{L}(x, y, \hat{x}) = \max\{0,1 - y\hat{x}\}$$

Solve it



Reality

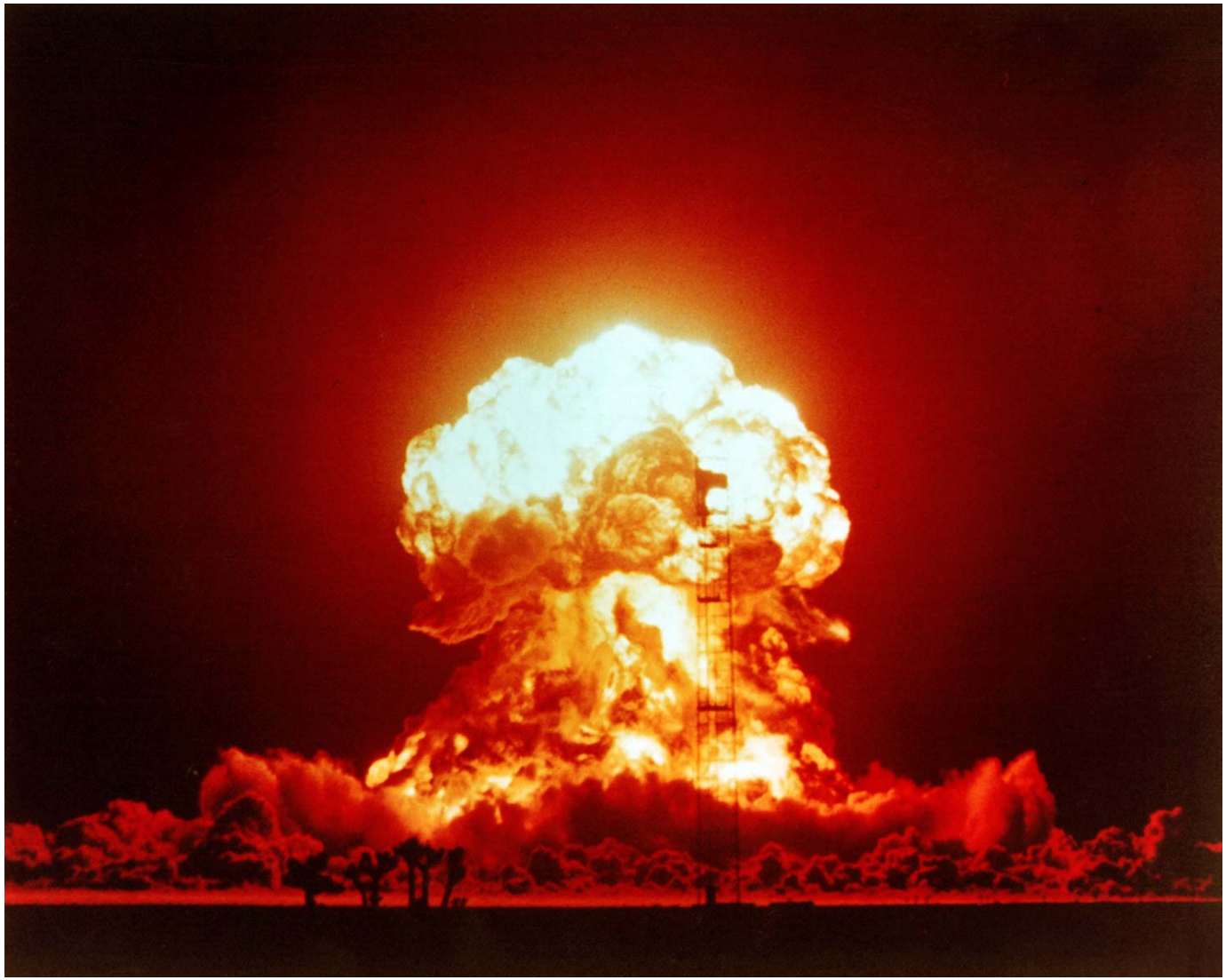
Problem

Nearly no one solves their problem like this

While data collection is cheap

Labeling data is usually pricey

Applications



Sentiment Tagging

Labeling words with their sentiment.

7 of 7 people found the following review helpful:

★★★★★ **This Milk Changed My Life**, August 8, 2010

By **Robert D. Queen "itcbob"** (Springfield, VA) - See all my reviews

REAL NAME™

This review is from: Tuscan Whole Milk, 1 Gallon, 128 fl oz (Misc.)

The Tuscan whole milk is the most amazing drink I have ever had. I used to be an alcoholic, but after one drink of this amazing milk, alcohol has never touched my lips again. Why drink bourbon when this amazing milk from the hills of Tuscany is now available to us all. Nothing short of the Second Coming compares to the sight of Tuscan Milk. Less than $100 per gallon is a steal. Don't miss out on the amazing opportunity to experience Tuscany as its finest.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comment

14 of 19 people found the following review helpful:

★☆☆☆☆ **No Protection at All**, August 8, 2007

By **J. McArthur** - See all my reviews

REAL NAME™

This review is from: JL421 Badonkadonk Land Cruiser/Tank

My wife and kids were playing in my JL421, and I thought I would give them a bit of a scare as a joke, so a shot a few rounds at the side with a rather large gun that I have and the bullets penetrated right through and killed them all! I am so disappointed with the quality of this land cruiser. I called the manufacturer and they said it wouldn't be covered under warranty because I did it intentionally. I'm never buying from this company again.

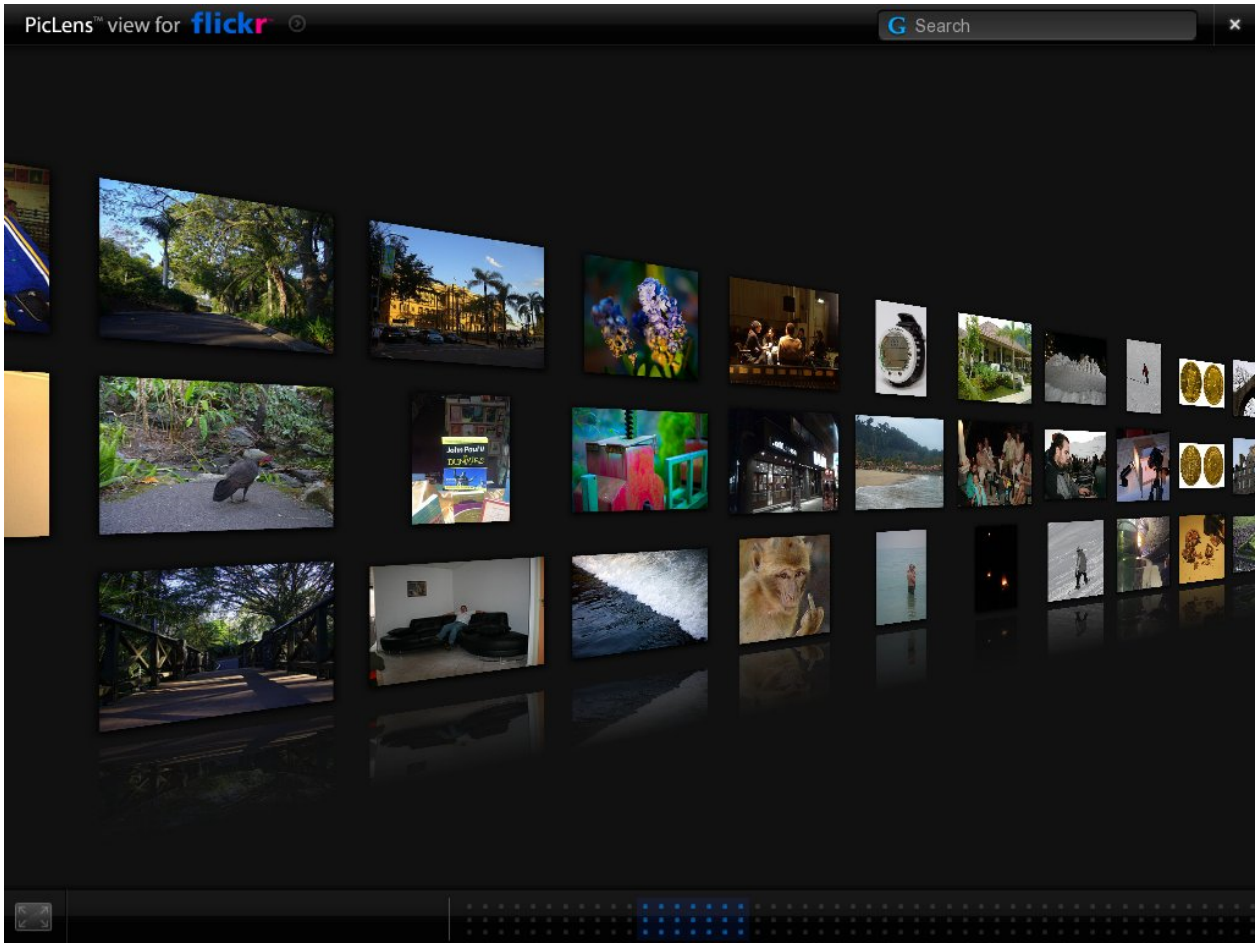
Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comment

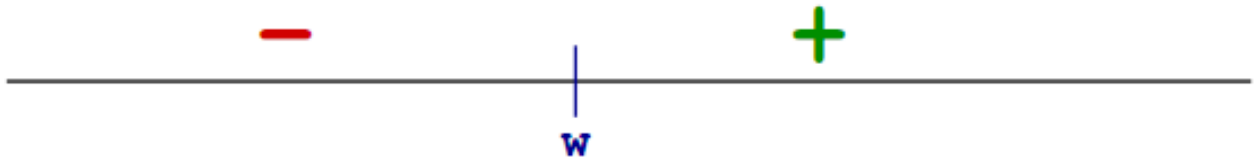
Image Labeling



"99% of the data doesn't matter" – Ted Dunning

1D classification task

$$H = \{h_w : w \in \mathbb{R}\}$$
$$h_w(x) = 1(x \geq w)$$

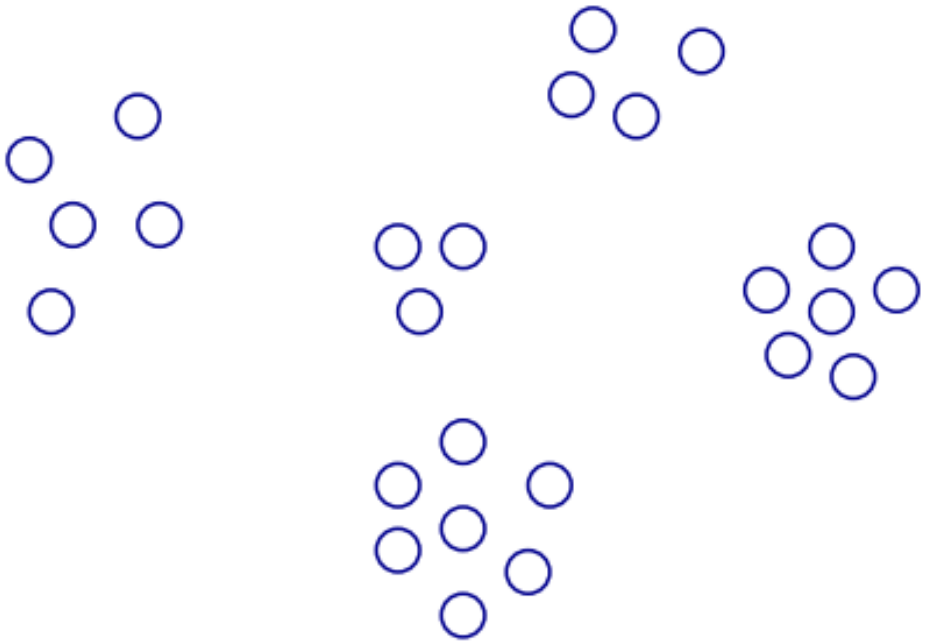


1D classification task

Active learning: instead, start with $1/\epsilon$ *unlabeled* points.

Binary search: need just $\log 1/\epsilon$ labels, from which the rest can be inferred. *Exponential improvement in label complexity!*

Clustered classification task

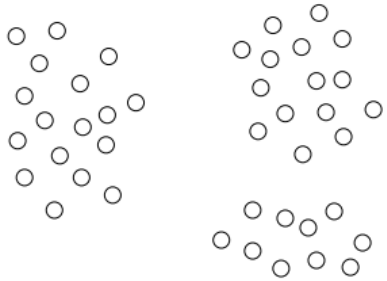


Clustered classification task

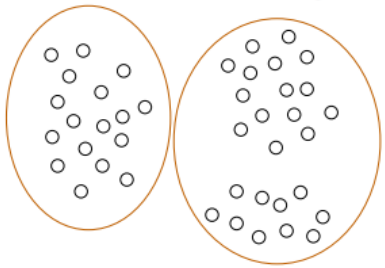
- Find clusters
- Sample points from clusters
- label rest of points with most popular label
- Else label data point, retrain predictors

Clustered classification task

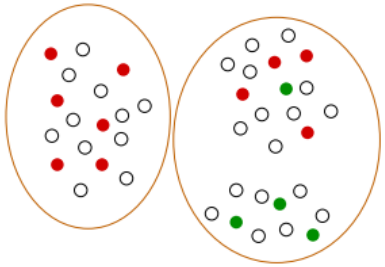
Unlabeled data



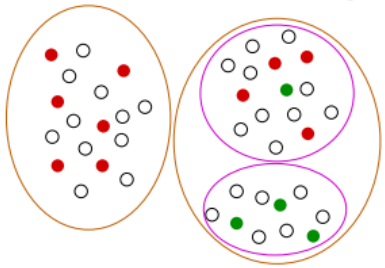
Find a clustering



Ask for some labels



Refine the clustering



Now what?

Selective Sampling

- Sample unexplored regions of space
- If predictors agree, ignore
- Else label data point, retrain predictors

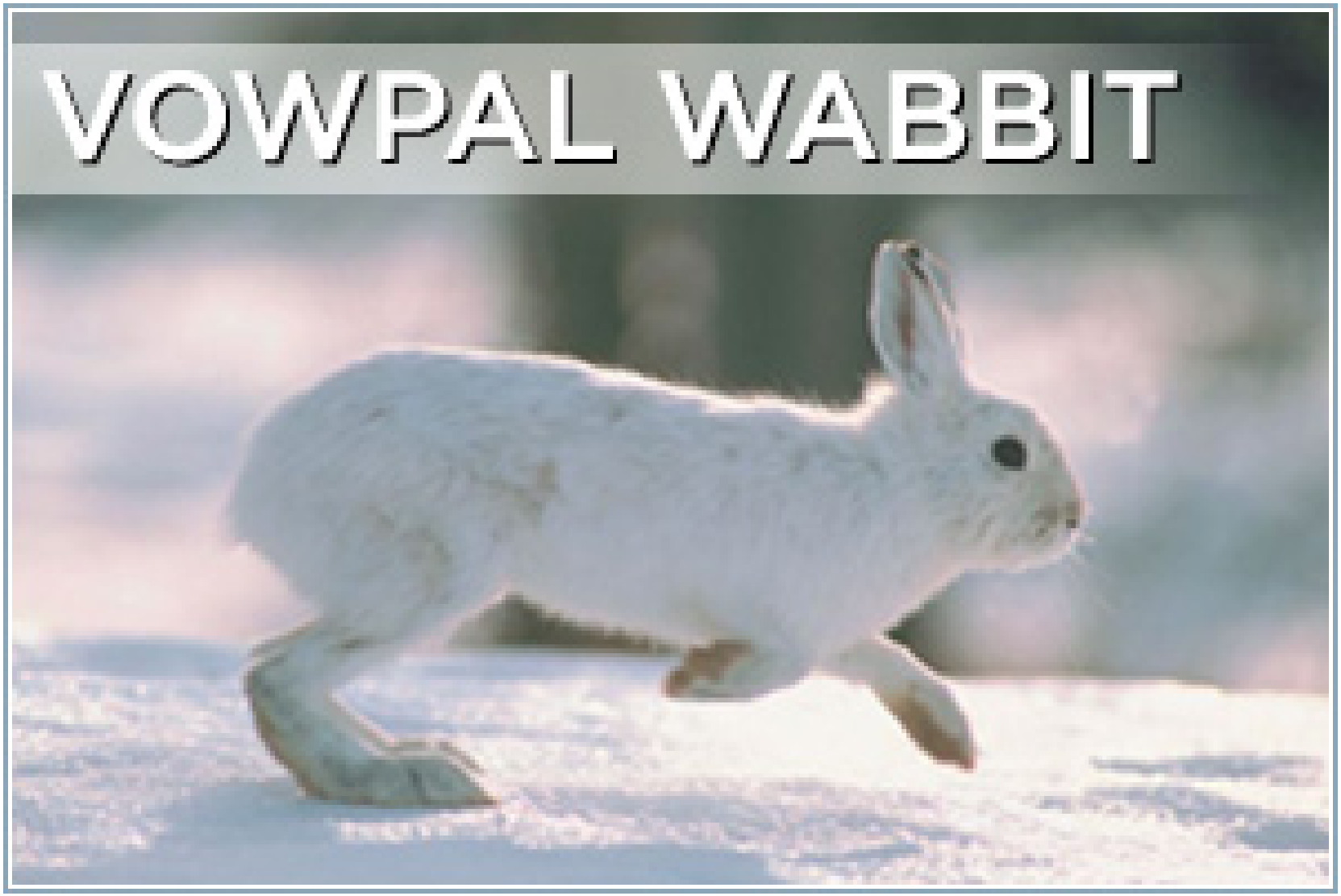
Agnostic Learning

- Track regions hypotheses disagree within
- Eliminate region they agree
- Sample labeled point in contentious region
- Place bounds on error, and eliminate hypotheses not within them
- Repeat

Labeled points needed for ϵ optimality

$$O(\lg \frac{1}{\epsilon})$$

Vowpal Wabbit



Demo

Demo!

Where Active Learning wins

- Low-noise settings
- Easy to collect unlabeled data
- Expensive to label data
- Comprehendable hypothesis classes

Conclusions

Active learning is an exciting development

Check out more http://hunch.net/~active_learning/

Questions?