Achieve More with Less using Active Learning & Guided Search

Data Science Summit 2018 Workshop.

Omri Allouche & Tal Perry

Active Learning

Omri Allouche

- Introduction to Active Learning
- Motivation – why I became interested in Active Learning
- Types of Active Learning
- Hands-on session
  - Image classification – MNIST
  - Text classification – SM Spam
- **https://github.com/omriallouche/active-learning**
- Wrap up – pros+cons of active learning, suggested workflow
- Coffee break

Guided Search

Tal Perry

- Introduction to Guided Search
- Introduction to Elastic Search
- Hands-on Session
- **https://github.com/LightTag/CAHLM**

Omri Allouche

Tal Perry

Inbal Horev

Eran Marom

Yonathan Guttel

Credit, Thanks & Useful Links

- Burr Settles -
  - Active Learning Survey http://burrsettles.com/pub/settles.activelearning.pdf
  - Dualist - https://github.com/burrsettles/dualist
- Tal Perry
  (check out https://github.com/LightTag/active_learning_example/blob/master/papayas.ipynb)
- Inbal Horev, Yonathan Guttel and Eran Marom
- The DSS team
- You

Dr. Omri Allouche

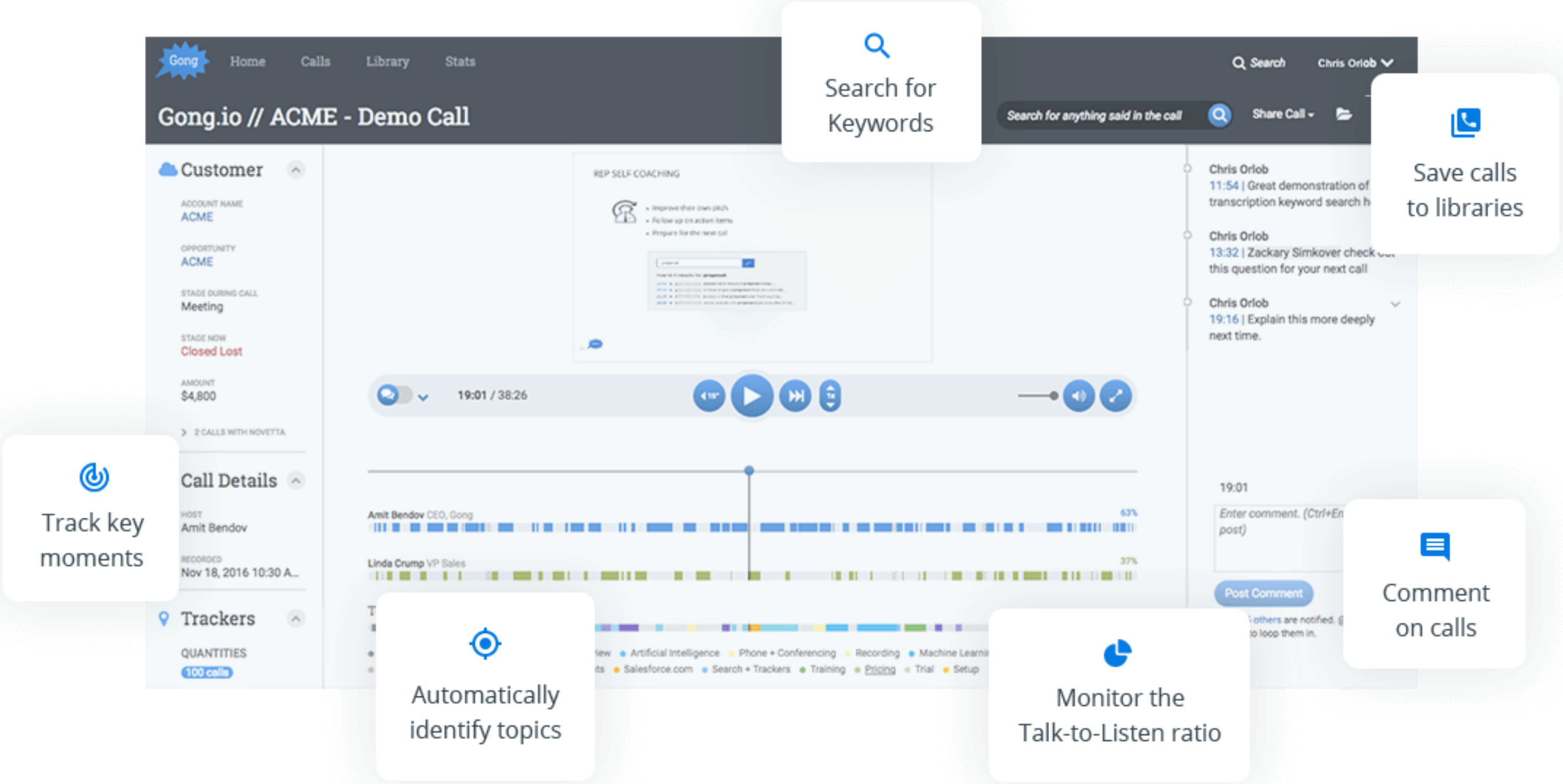
Head of Research, Gong.io

Teacher, Bar Ilan University

Record

Transcribe

Analyze


Search for Keywords
Save calls to libraries
Track key moments
Automatically identify topics
Monitor the Talk-to-Listen ratio
Comment on calls
Gong
Home
Calls
Library
Stats
Search
Chris Orlob
Gong.io // ACME - Demo Call
Search for anything said in the call
Share Call
Customer
ACCOUNT NAME
ACME
OPPORTUNITY
ACME
STAGE DURING CALL
Meeting
STAGE NOW
Closed Lost
AMOUNT
$4,800
2 CALLS WITH NOVETTA
Call Details
HOST
Amit Bendov
RECORDED
Nov 18, 2016 10:30 A...
Trackers
QUANTITIES
100 calls
REP SELF COACHING
19:01 / 38:26
Amit Bendov CEO, Gong
63%
Linda Crump VP Sales
37%
Artificial Intelligence
Phone + Conferencing
Recording
Machine Learni
Salesforce.com
Search + Trackers
Training
Pricing
Trial
Setup
Chris Orlob
11:54 | Great demonstration of
transcription keyword search h
Chris Orlob
13:32 | Zackary Simkover check
this question for your next call
Chris Orlob
19:16 | Explain this more deeply
next time.
19:01
Enter comment. (Ctrl+En
post)
Post Comment
others are notified.
o loop them in.

Teaching Machines is an Expensive Business…

- Transcribing an hour of Speech ≈ $100
- You need roughly 1,000 hours for descent performance = $100,000
- Labeling a text sentence in FigureEight is ~10 cents
- Binary Classification Tasks with ~1-5% prevalence of target class × 1,000 sentences = $10,000 per task
- Annotating Sales calls requires domain expertise – hard to come by and expensive…

Action Items Detection

I will send you an email after the call

Let's touch base next Tuesday

Let me share my screen...

Other features can indicate an Action Item:

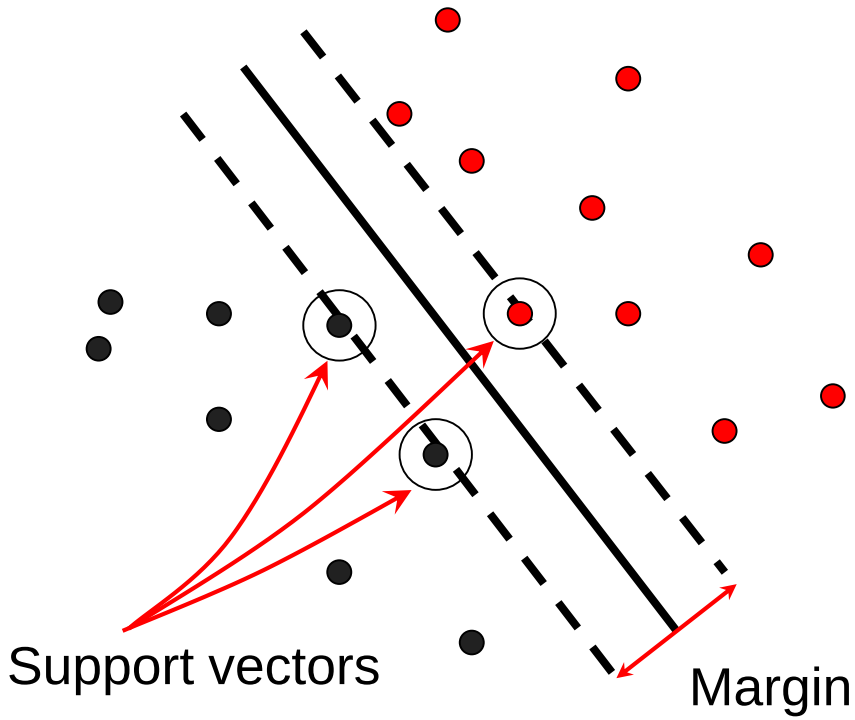
- Action items are mostly said by Sales Rep
- Action items mostly appear at the end of the call
- Action items are usually the last sentence in a paragraph

Action Items Detection

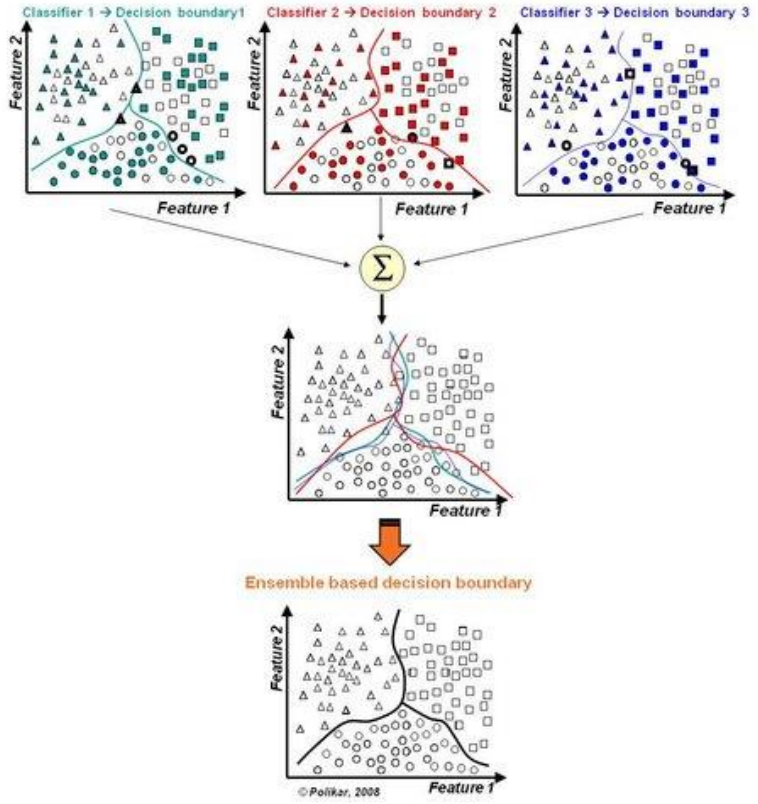
- Many sentences are obviously not an Action Item –
  - *Hey Josh. How are you?*
  - *Yeah, it's been raining all day here.*
  - *Can you hear me?*
  - *…*
- We don't want to spend precious labeling time on them
- **Claim:** Even a basic model can detect if a sentence is:
  - Definitely an Action Item
  - Definitely not an Action Item
  - Might be an Action Item

Concepts rooted in other Machine Learning concepts

SVM - Support Vector Machines –
Decision boundary is set by Support Vectors – these points are the most informative

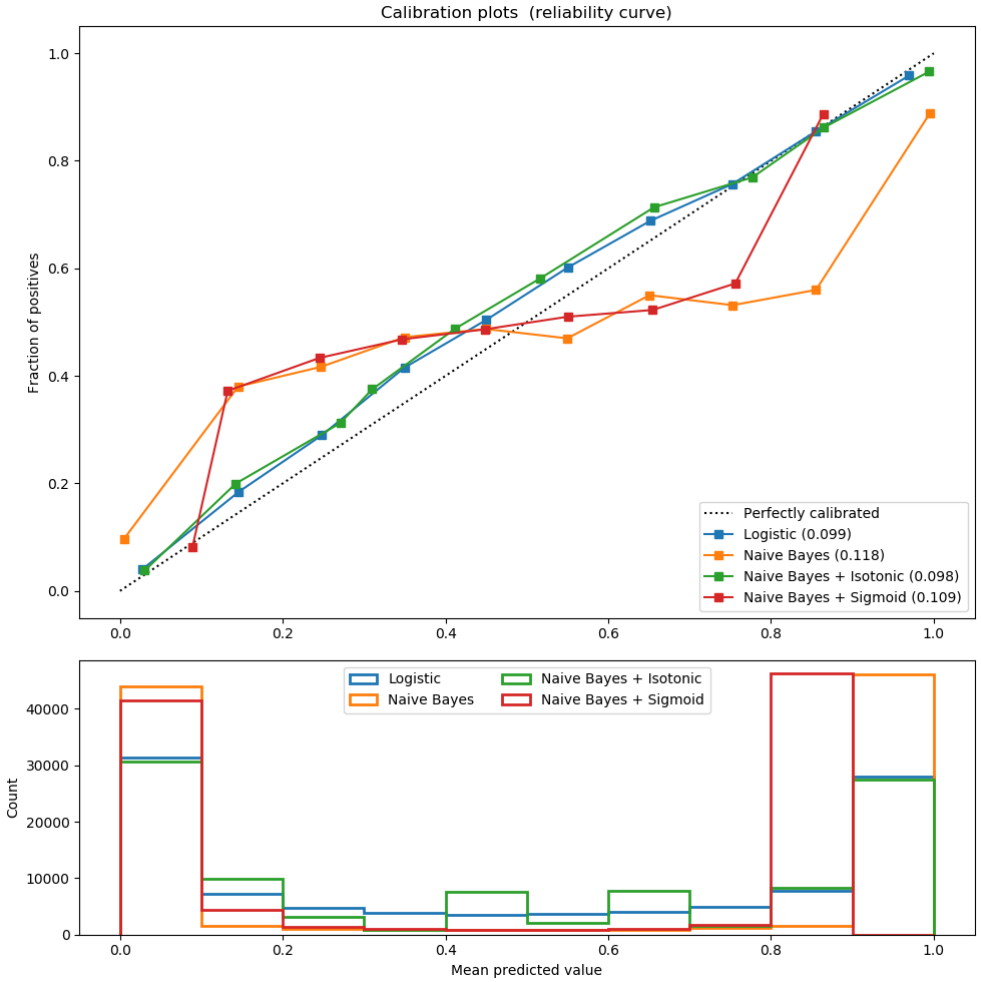


Ensemble Learning –
Combine different models for a joint prediction, with higher accuracy.



Concepts rooted in other Machine Learning concepts

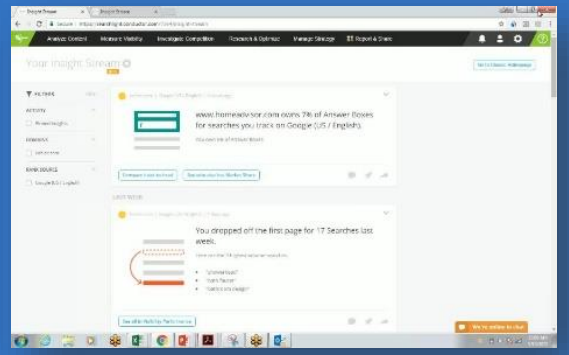
Calibration Plots –
Checking correspondence between predicted value and fraction of positives
(http://scikit-learn.org/stable/modules/calibration.html)



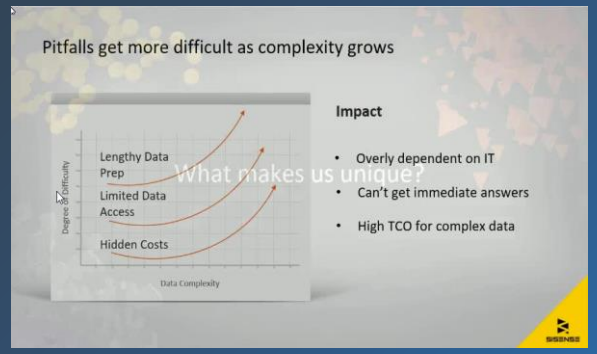
Video Frame Classification

Video recordings can be separated into 4 main categories:

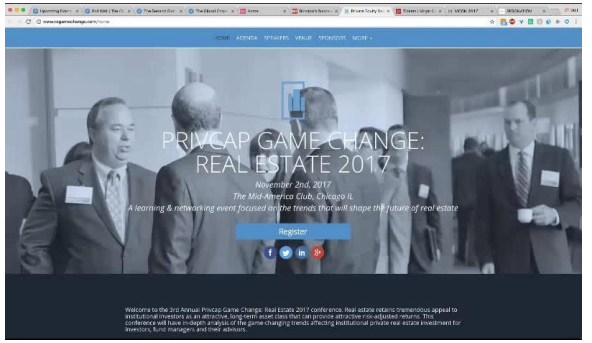
1. Web cam
2. Browser demo
3. Presentation slides (ppt / Google slides etc)
4. Conferencing provider app (Skype / Zoom / GoToMeeting etc)

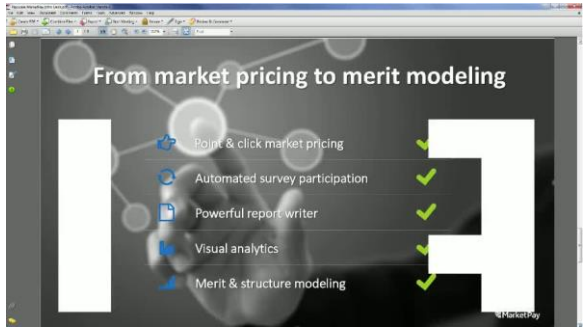


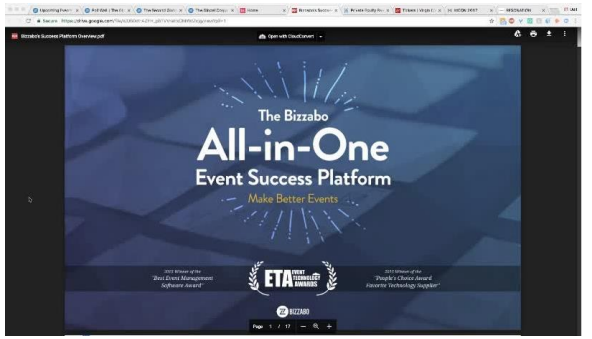
High Confidence

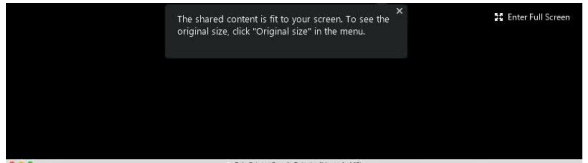


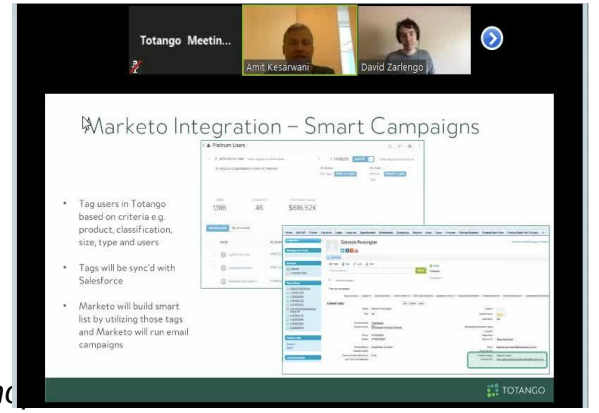
Low Confidence

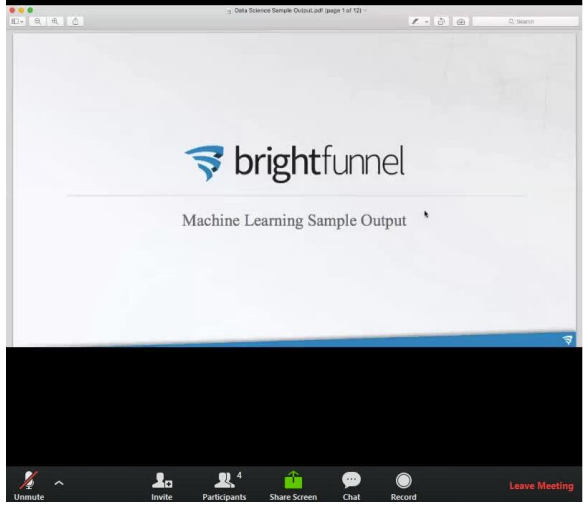












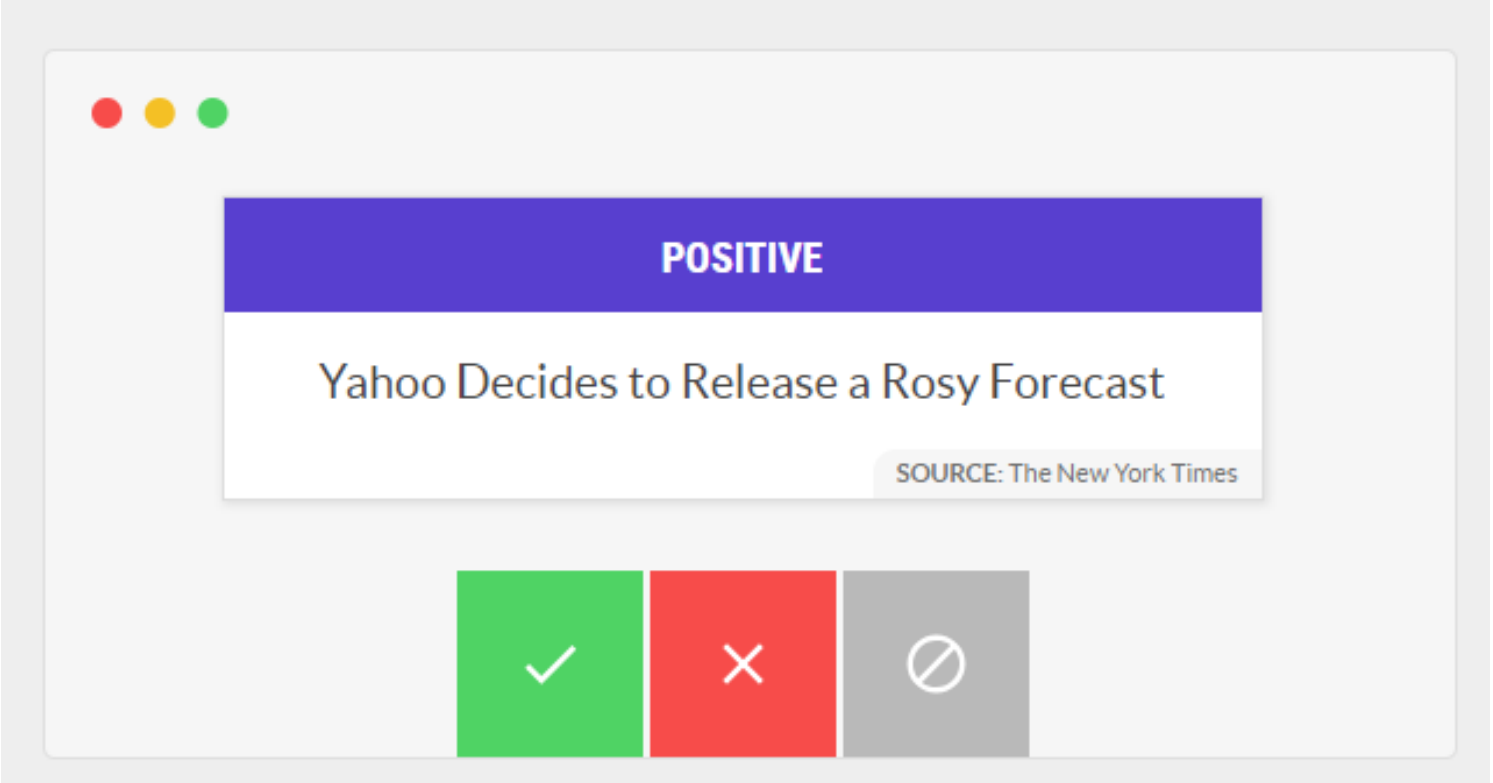
CrowdFlower changes its name to Figure Eight

> *The formal name change to Figure Eight is to better reflect the company's technology platform that combines the best of machine learning and human intelligence, rather than just using human intelligence to label data. The new name and identity brings into focus Figure Eight's core philosophy of human-in-the-loop practices being the essential ingredient to making AI work in the real world, specifically the iterative process of active learning for the training, testing, and tuning of machine learning models.*

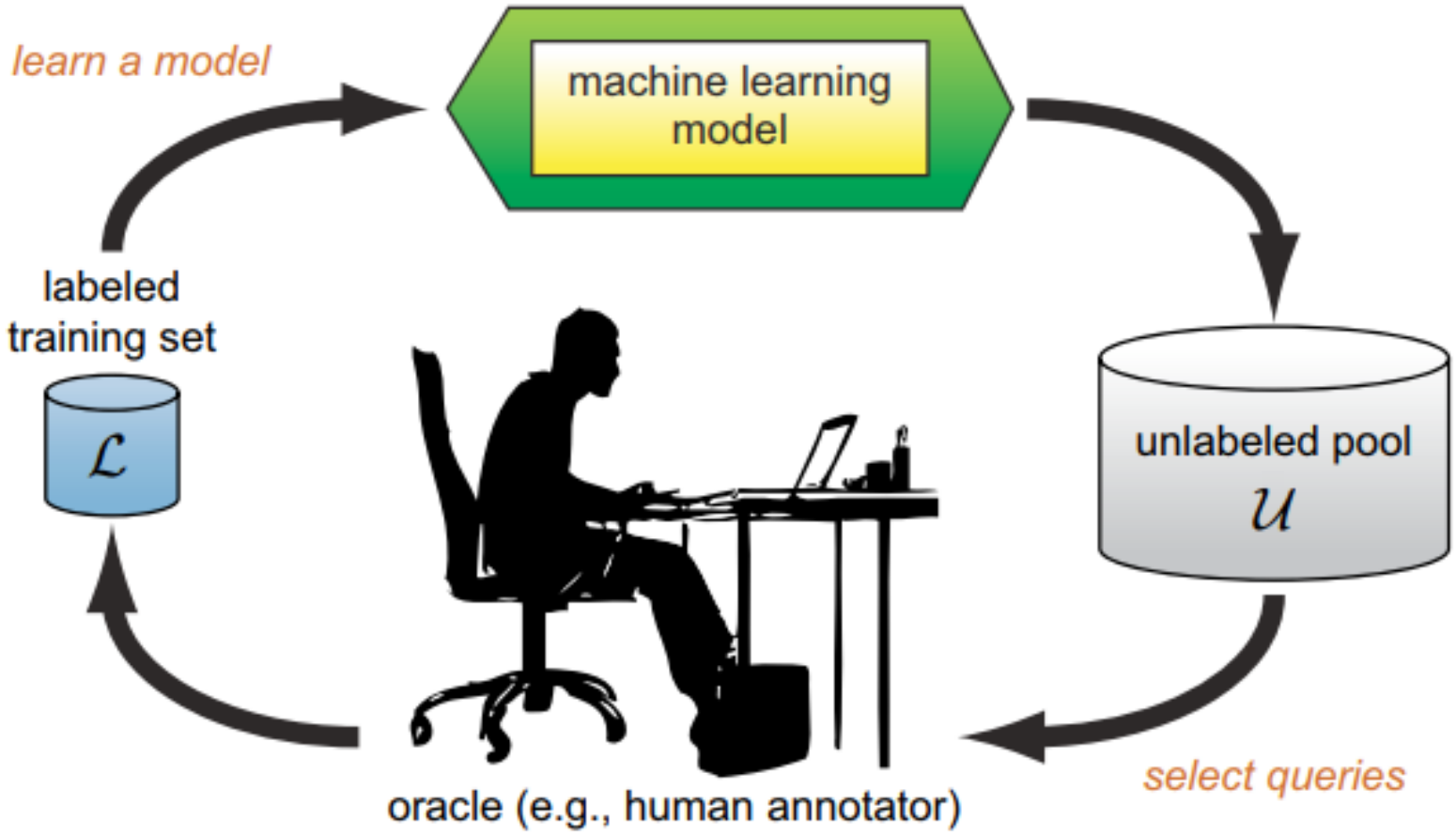
prodigy

Radically efficient machine teaching.
An annotation tool powered
by active learning.

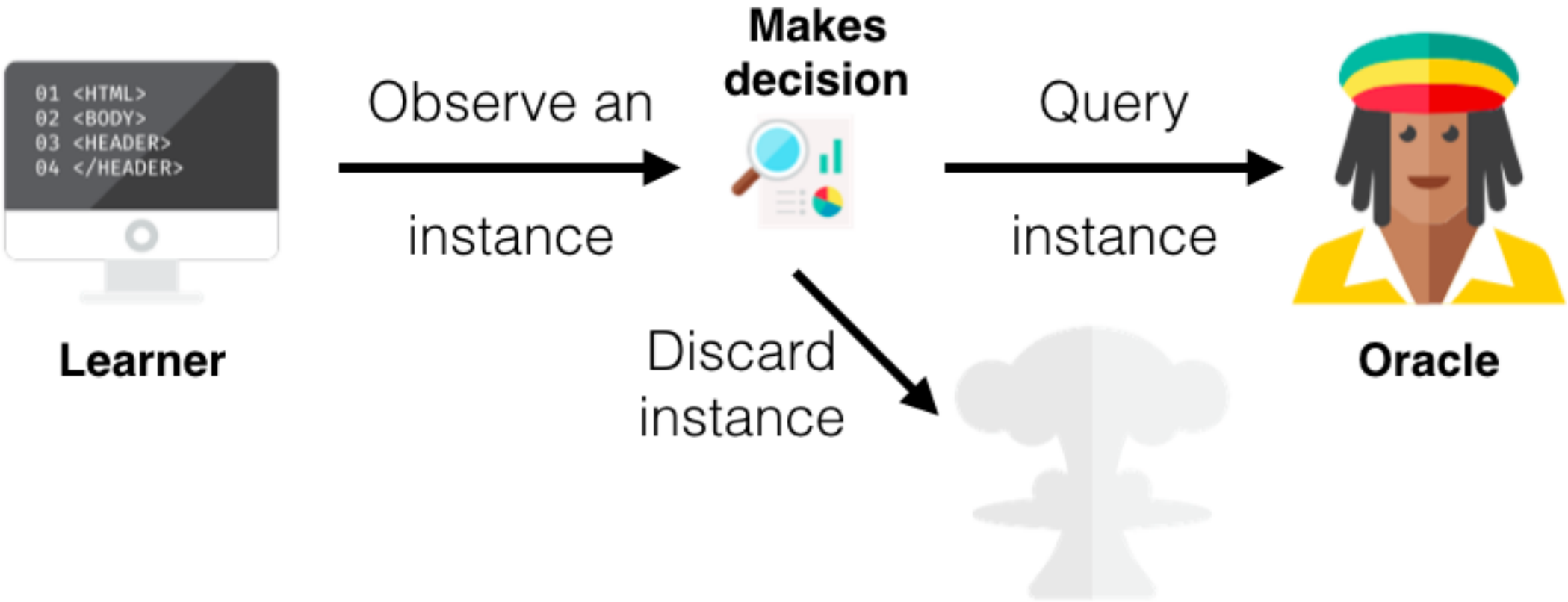
FROM THE MAKERS OF SPACY



Active Learning – Basic Framework



Stream-Based Selective Sampling



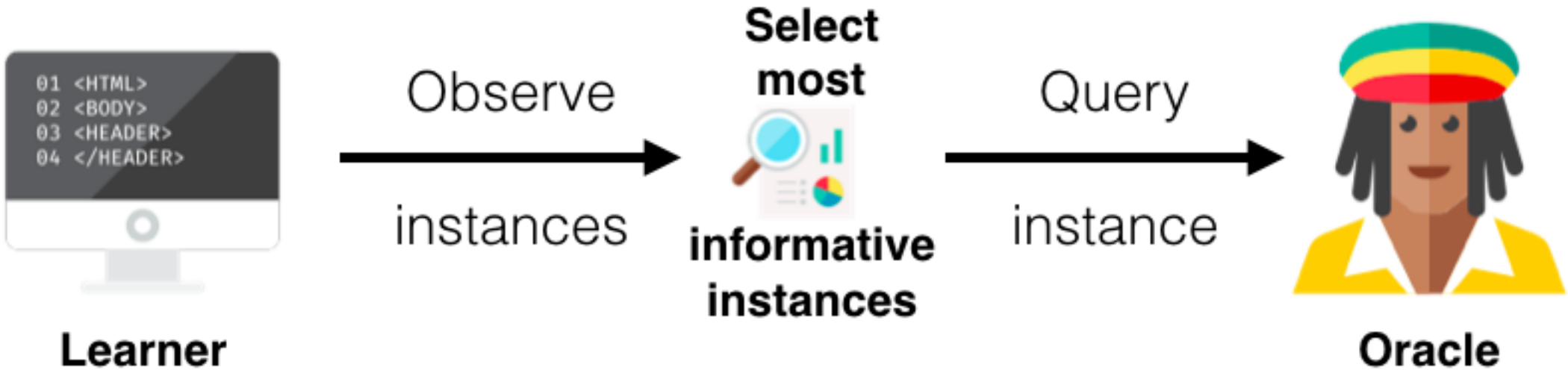
Gong

Omri Allouche. *Data Science Summit Europe 2018* ***Active Learning workshop***

Membership Query Synthesis

Omri Allouche. *Data Science Summit Europe 2018* ***Active Learning workshop***

Pool-based Sampling



Single Instance / Batch

Active Learning Methods

Uncertainty Sampling

Label points with the lowest confidence for the predicted class

$$x^*_{LC} = \underset{x}{\text{argmax}}\ 1 - P_\theta(\hat{y}|x),$$

$$\hat{y} = \text{argmax}_y\ P_\theta(y|x)$$

Margin Sampling

Label points with the lowest difference between confidence of predicted class and 2nd best class

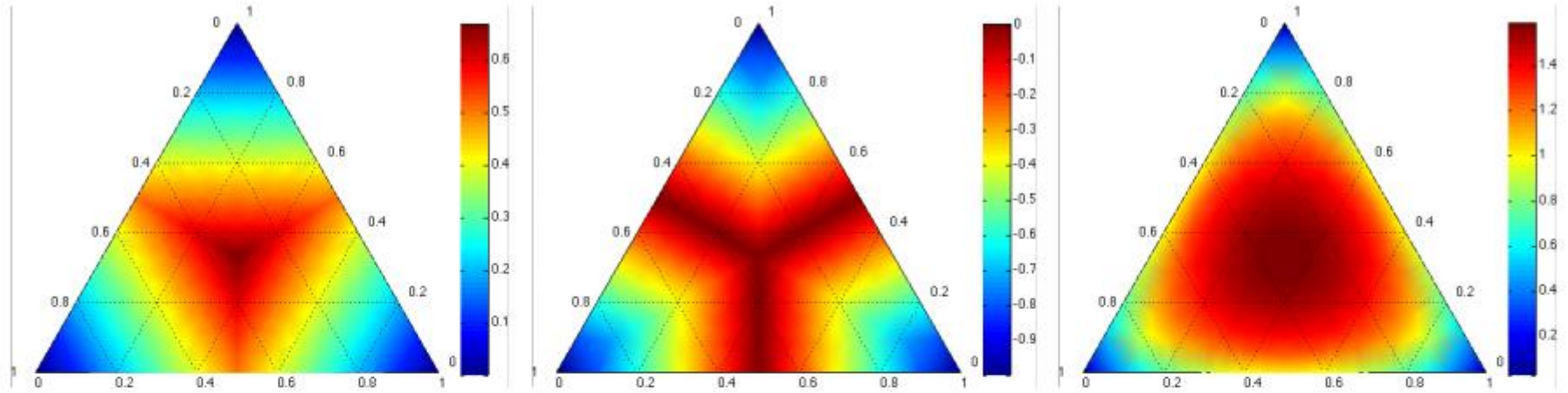
$$x^*_M = \underset{x}{\text{argmin}}\ P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x),$$

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable class labels under the model, respectively

Entropy Sampling

Label points with highest entropy for predicted classe

$$x^*_H = \underset{x}{\text{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x),$$



(a) least confident (b) margin (c) entropy

Figure 5: Heatmaps illustrating the query behavior of common uncertainty measures in a three-label classification problem. Simplex corners indicate where one label has very high probability, with the opposite edge showing the probability range for the *other* two classes when that label has very low probability. Simplex centers represent a uniform posterior distribution. The most informative query region for each strategy is shown in dark red, radiating from the centers.

Credit: Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

Active Learning Methods

Query by Committee

A variety of models are trained on the current labeled data, and vote on the output for unlabeled data

Label those points for which the "committee" disagrees the most

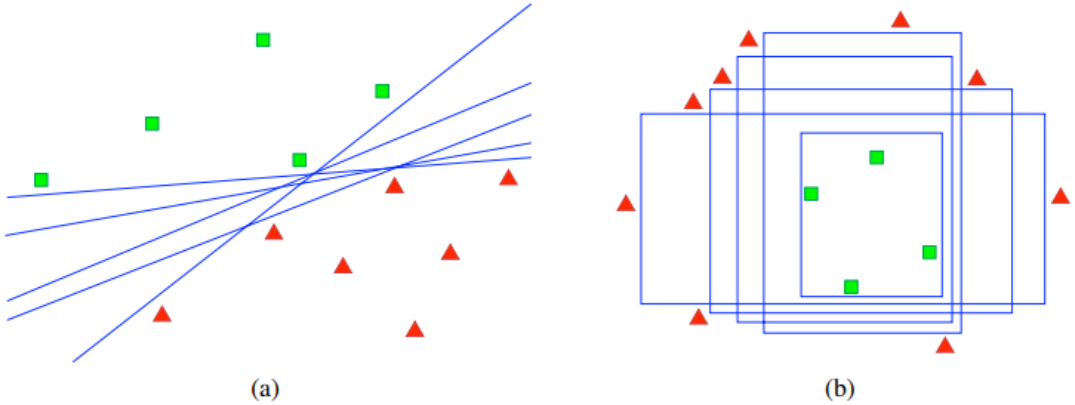


Figure 6: Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in $\mathcal{L}$ (as indicated by shaded polygons), but each represents a different model in the version space.

Using Unlabeled Data

"Believe in yourself" - Bootstrapping

- The number of labeled instances is often too small for proper classification
- Better performance are obtained when weighting labeled instances with unlabeled instances
- Running EM several times improves performance
- Biggest gain is achieved on the 1st run of EM - for performance, run EM only once

Active Learning Theory

Select instances according to

1. *Informativeness*
2. *Representativeness*
3. *Diversity*

How to select seed?

- Represent all classes
- Perhaps first cluster and select from different clusters
- Start with potentially difficult set (e.g. "I will" for Action Items)
- Consider class priors

Today's Agenda – Active Learning Part

- 2 Datasets –
  - Image - the MNIST digits dataset
  - Text – spam/ham text messages
- Train a simple Logistic Regression model
- Check if model confidence correlates with its performance
- Train several models (LR, KNN, SVM), and check when they agree/disagree
- (Finally!) Test it ourselves - can Active Learning improve our performance?

Today's Agenda – Active Learning Part

(Finally!) Test it ourselves - can Active Learning improve our performance?

1. We set an initial batch for labeling
2. We then select N new examples and ask the Oracle for their true label
3. We train a model on examples with labels
4. Repeat steps 2-3

Omri Allouche

Tal Perry

Inbal Horev

Eran Marom

Yonathan Guttel

Criticism against Active Learning

- Selected instances often tied to chosen algorithm
  - Do instances selected because of one model's confidence be helpful for a different model?
- Selected instances are a biased distribution of population
- Performance depends on the proficiency of annotator
  - Labelers tend to make mistakes for "difficult" instances

Proactive Learning

Set to relax the unrealistic assumptions of Active Learning

1. Oracle may be wrong
2. especially when question is more difficult
3. Oracle may be reluctant to answer
4. Cost of labeling increase for difficult questions

Donmez, P., Carbonell, J.G. (2008) Proactive Learning: Cost-Sensitive Active Learning with Multiple Imperfect Oracles

When Active Learning is a Bad Choice

1. Training takes very long
2. Labeling is very cheap
3. Classes are well-balanced
4. You are unsure regarding your choice of algorithm
5. There's no correspondence between model confidence and performance

Why Active Learning didn’t Take Off

- In Academia –
  - Challenges include labeled instances and focus on overall performance, without considering the number of labels used
  - When creating your own dataset, you select instances that are “interesting”
- While labeling often requires only a simple share of a file,
  Active Learning requires an interactive platform
  - UI
  - Server
- Results aren’t very robust (as our hands-on work shows), and depend on the algorithm
- There’s no foolproof “Recipe” for Active Learning

Suggested Workflow

- Collect a small set of labeled instances using “Guided Search”
  - Use domain knowledge
  - Use statistics to propose candidates that a human can label
- Build a basic model, and evaluate performance
  - Learning Curve
  - Correlation between confidence and performance
  - Review manually instances with low confidence – there are often labeling mistakes there
- Decide where to invest your efforts
  - Fix errors with a rule-based model
  - Enlarge training data – use model confidence to highlight good candidates for labeling
  - Increase training set randomly
  - …

Gong

We’re hiring!

- Full-stack Java Engineers
- Machine-learning Engineers
- Data Scientists

Time for more coffee!