The Boundary Forest Algorithm for Fast Online Learning of Large Datasets

Nate Derbinsky

Computer Science & Networking

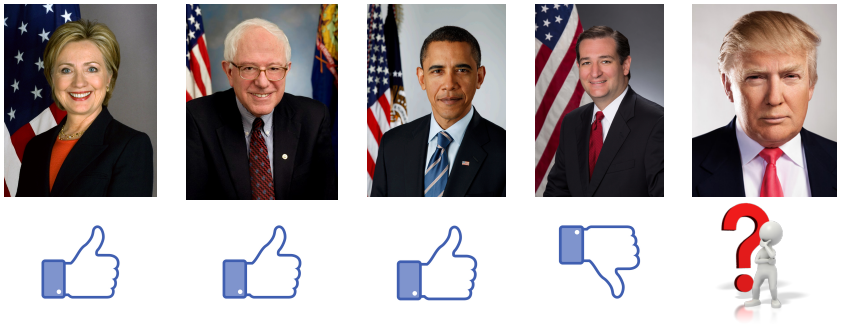
Machine Learning!?

The development and application of computer systems that can *learn* from *data*.

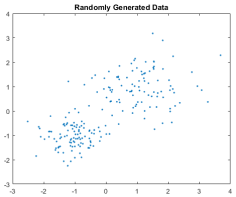
We judge that *learning* has taken place when "performance" on some task improves after exposure to data.

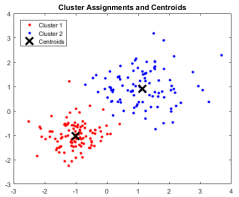
Kinds of Learning

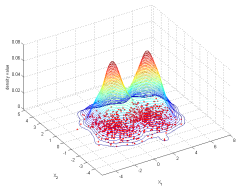
- **Supervised**. Given "labeled" data, generalize to new inputs

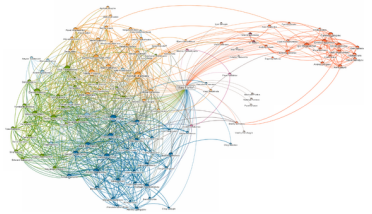


- **Unsupervised**. Given data, find "interesting" patterns









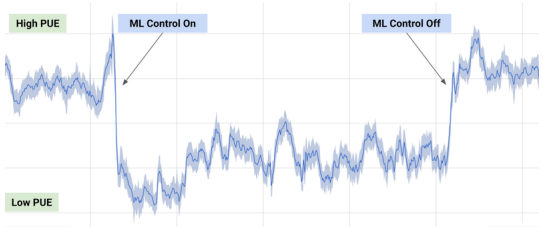
- **Reinforcement**. Learn actions given experience + reward

AlphaGo

ML is Saving the World

Google uses DeepMind AI to cut data center energy bills

The AI successfully reduced power consumption by 15 percent overall



Saving a few percentage points of electricity usage means major financial gains for Google. Typical electricity prices companies pay in the U.S. range from about $25 to $40 per MWh, according to data from the U.S. Energy Information Administration. (Prices in different regions range from a few dollars to more than $100). Either way, saving 10 percent on data center power consumption, for instance, could translate to hundreds of millions of dollars in savings for Google over multiple years.

Computers will require more energy than the world generates by 2040

Moore's Law is about to hit a wall.

PETER DOCKRILL 26 JUL 2016

ML is Changing the World

Tesla car mangled in fatal crash was on Autopilot and speeding, NTSB says

Amazon robots close to replacing the rest of warehouse workers

By David Cardinal on July 5, 2016 at 12:31 pm | 34 Comments

571 shares

ML is in Demand!

Position	Salary*
Data Scientist	**$113,436**
Machine Learning Engineer	**$114,826**
Software Engineer	$95,195

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."
– Josh Blumenstock (UW)

"Data Scientist = statistician + programmer + coach + storyteller + artist"
– Shlomo Aragmon (Ill. Inst. of Tech)

*glassdoor.com, National Avg as of July 27, 2016

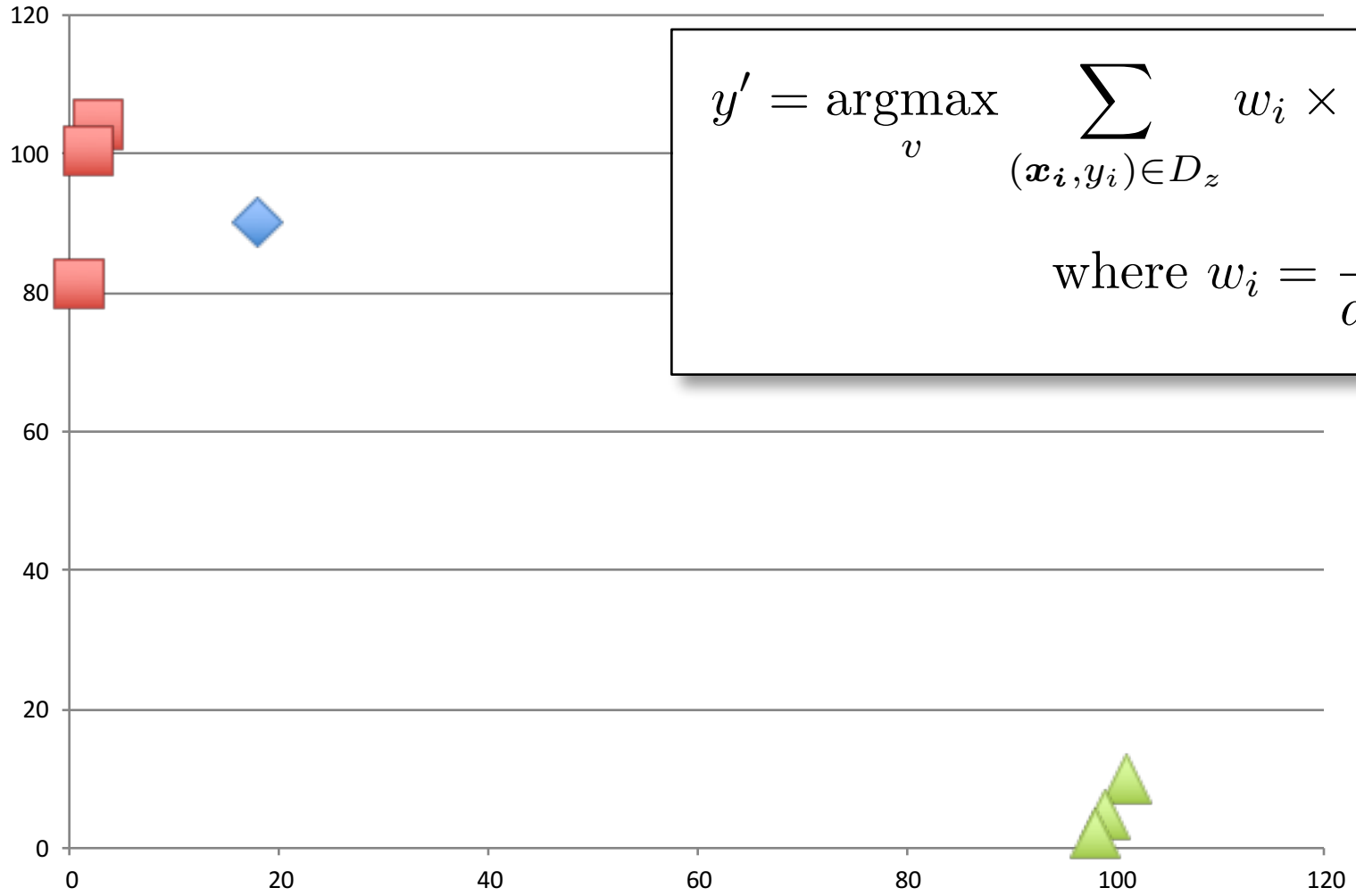
Your 1st ML Algorithm

k-Nearest Neighbors

Movie Title	# of Kicks	# of Kisses	Type of Movie
California Man	3	104	Romance
He's Not Really into Dudes	2	100	Romance
Beautiful Woman	1	81	Romance
Kevin Longblade	101	10	Action
Robo Slayer 3000	99	5	Action
Amped II	98	2	Action
Foo Bar Baz	**18**	**90**	**?**

Your 1st ML Algorithm

k-Nearest Neighbors

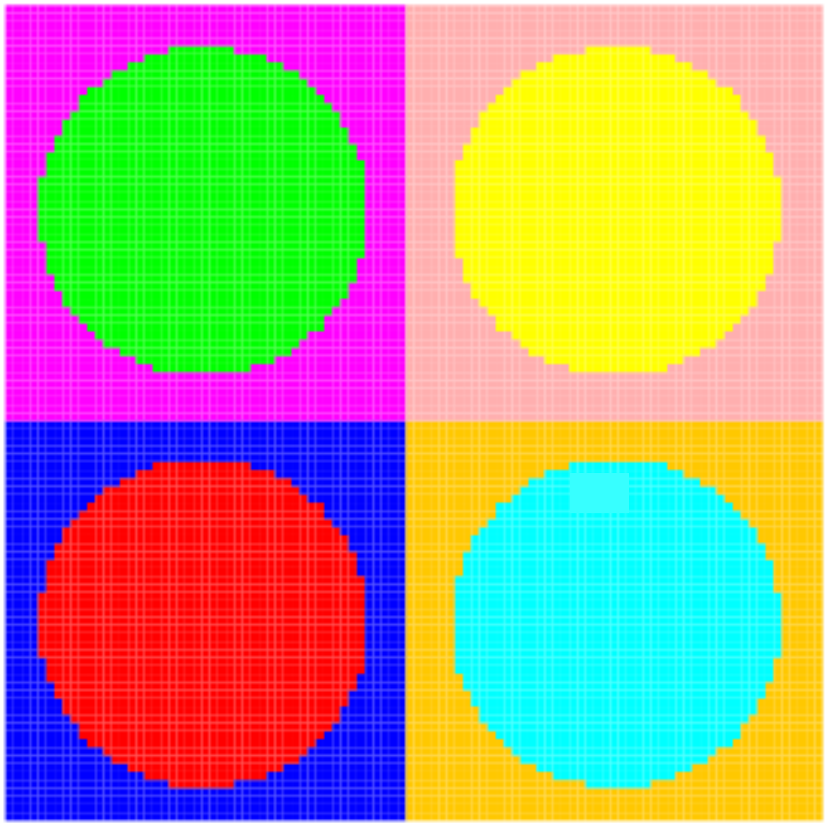


$$y' = \underset{v}{\operatorname{argmax}} \sum_{(\boldsymbol{x_i}, y_i) \in D_z} w_i \times I(v = y_i)$$

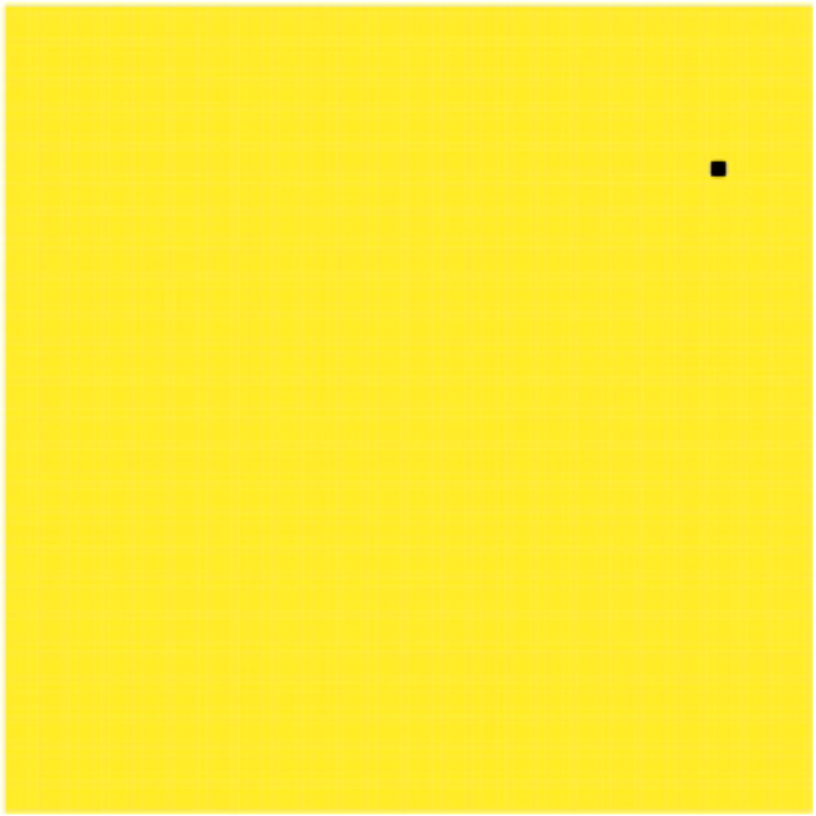
$$\text{where } w_i = \frac{1}{d(\boldsymbol{x'}, \boldsymbol{x}_i)^2}$$

Learning Simple Shapes

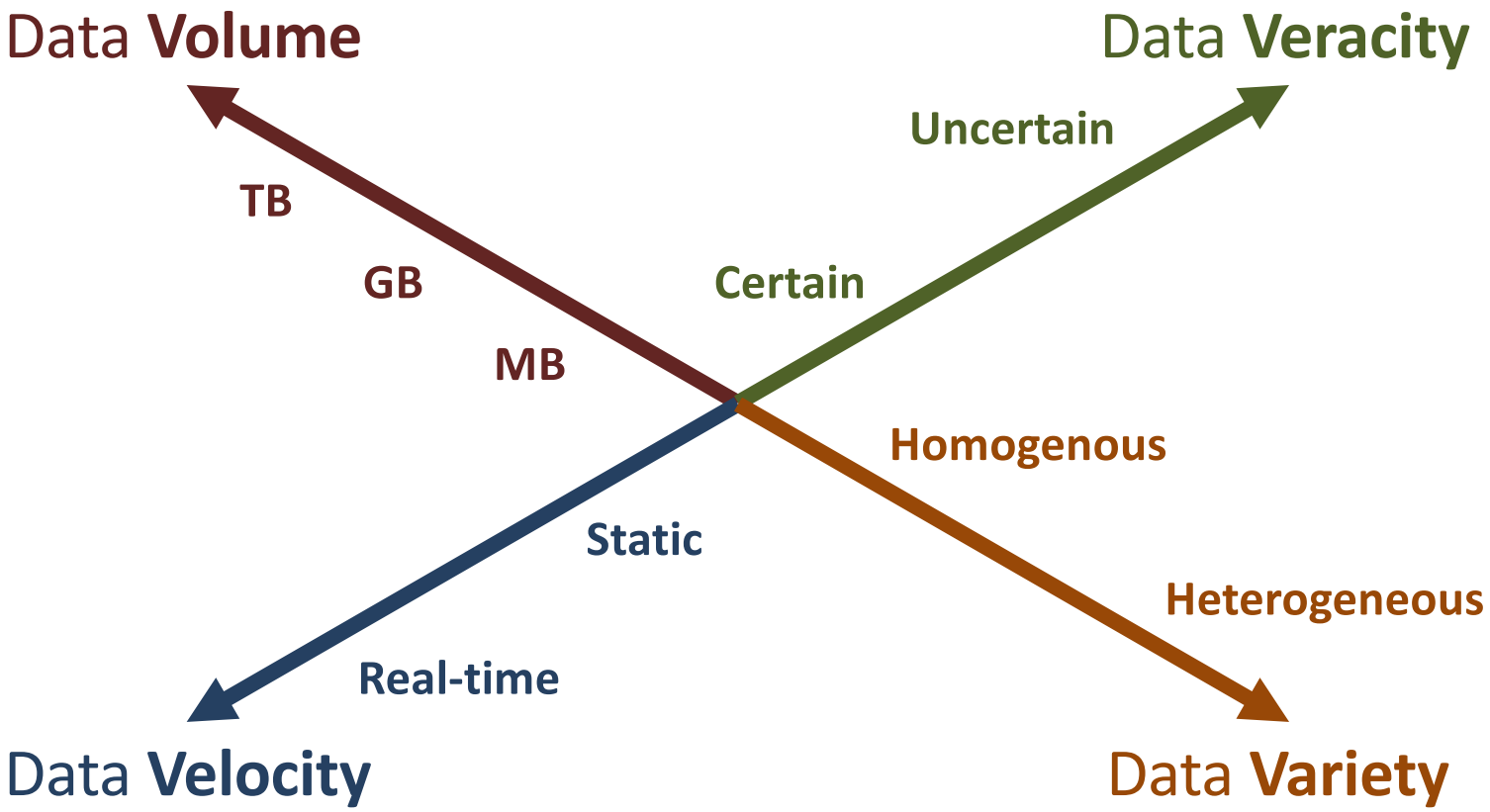
Ground Truth



1-Nearest Neighbor



What is "Big Data"?



Our goal: learning that gets more accurate, but not slower, with *more data*

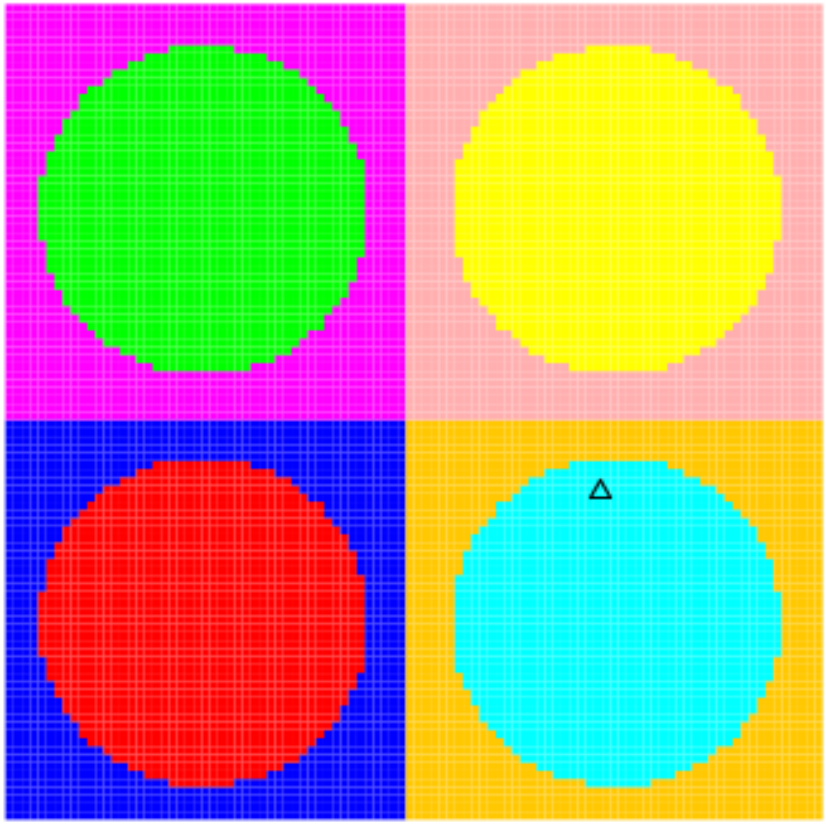
The Boundary Forest Algorithm

- The promise
  - About as accurate as kNN
  - Scales *logarithmically*
    - +1000x data = ~3x slower

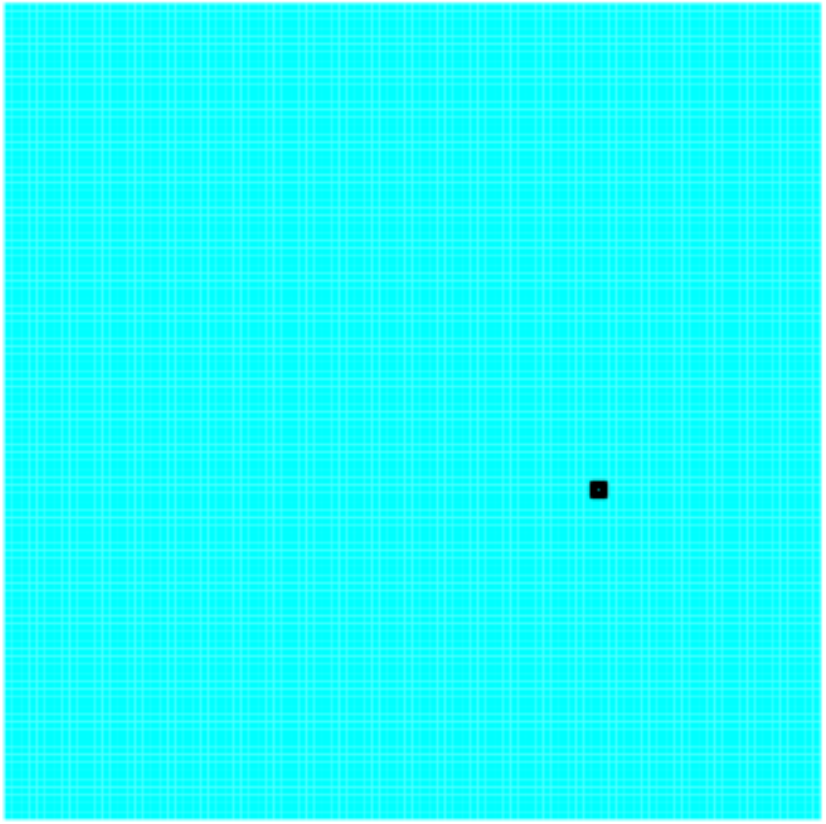
- Basic idea
  - Only store examples at “boundaries”
  - Learns a “forest” of search trees

Boundary Forest "Training"

Ground Truth

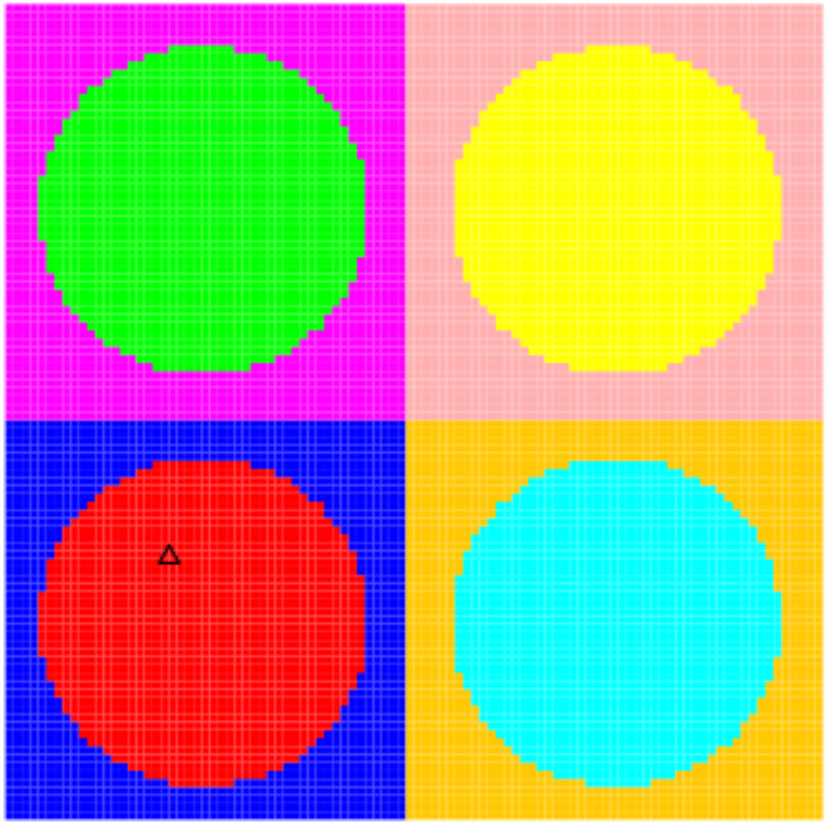


Boundary Tree

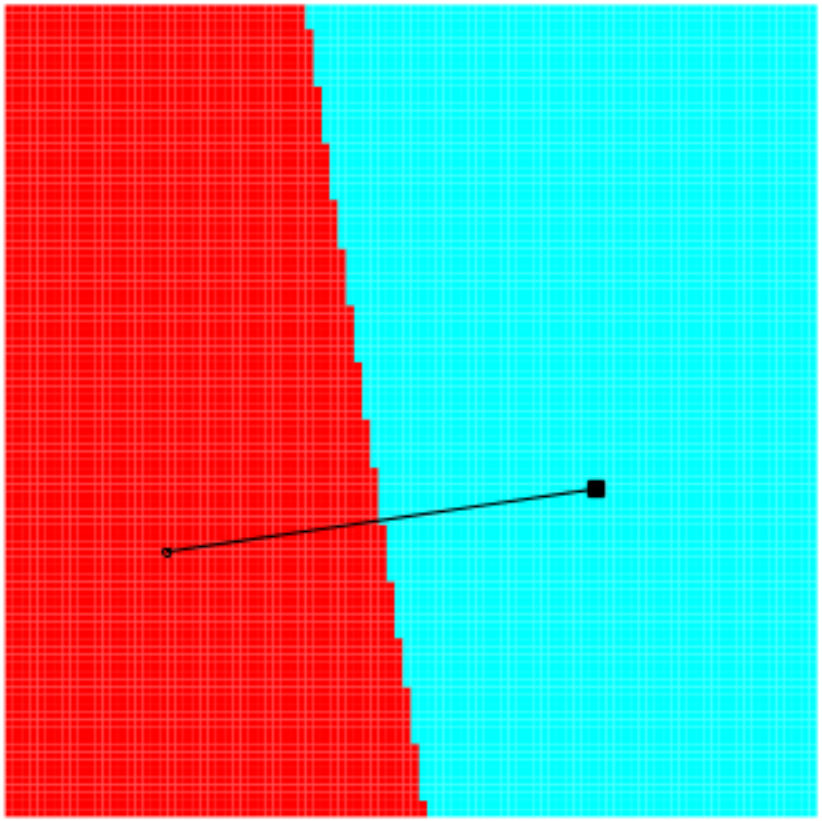


Boundary Forest "Training"

Ground Truth

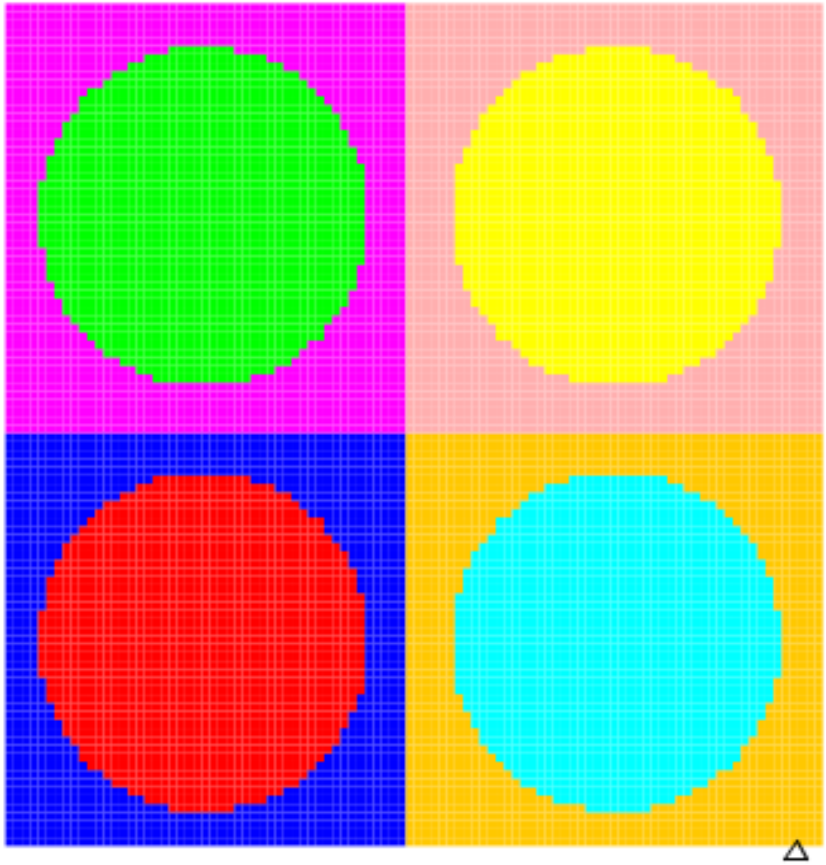


Boundary Tree

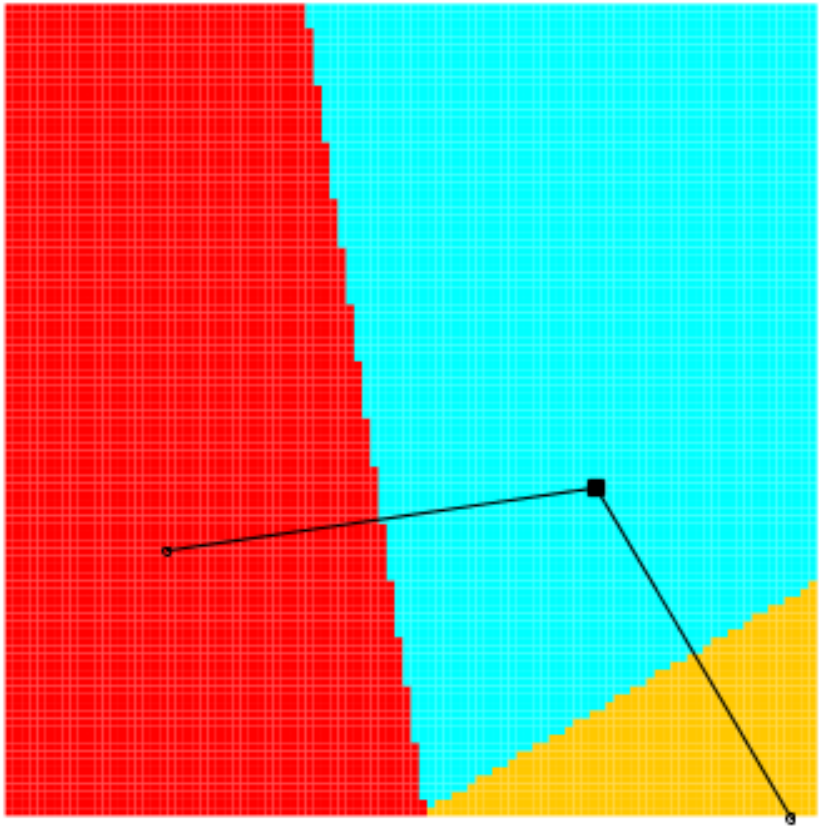


Boundary Forest "Training"

Ground Truth

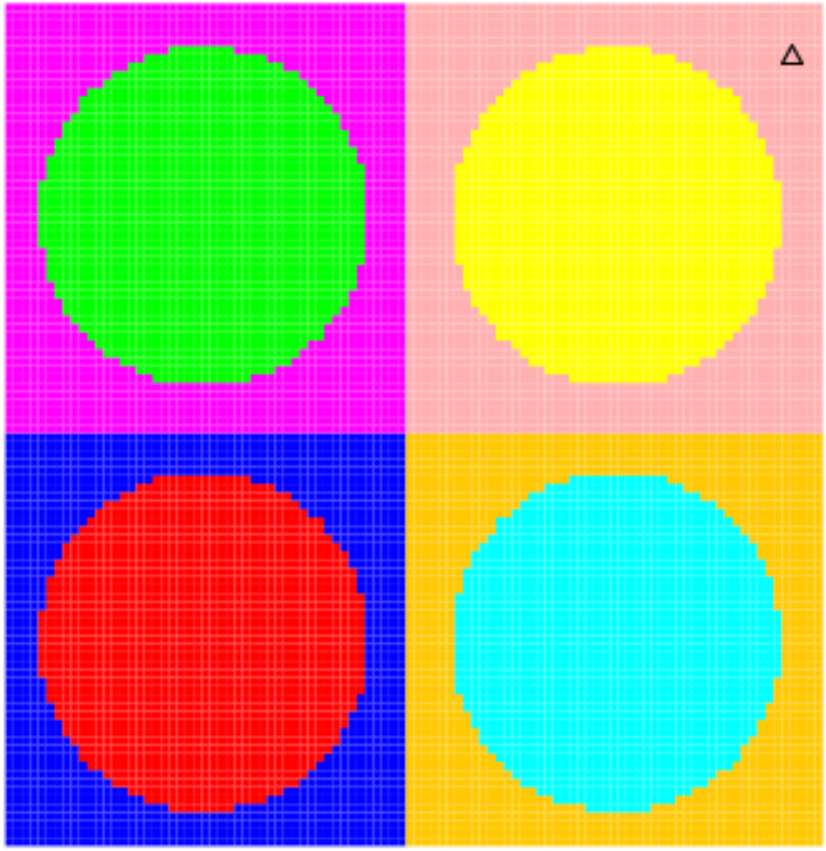


Boundary Tree

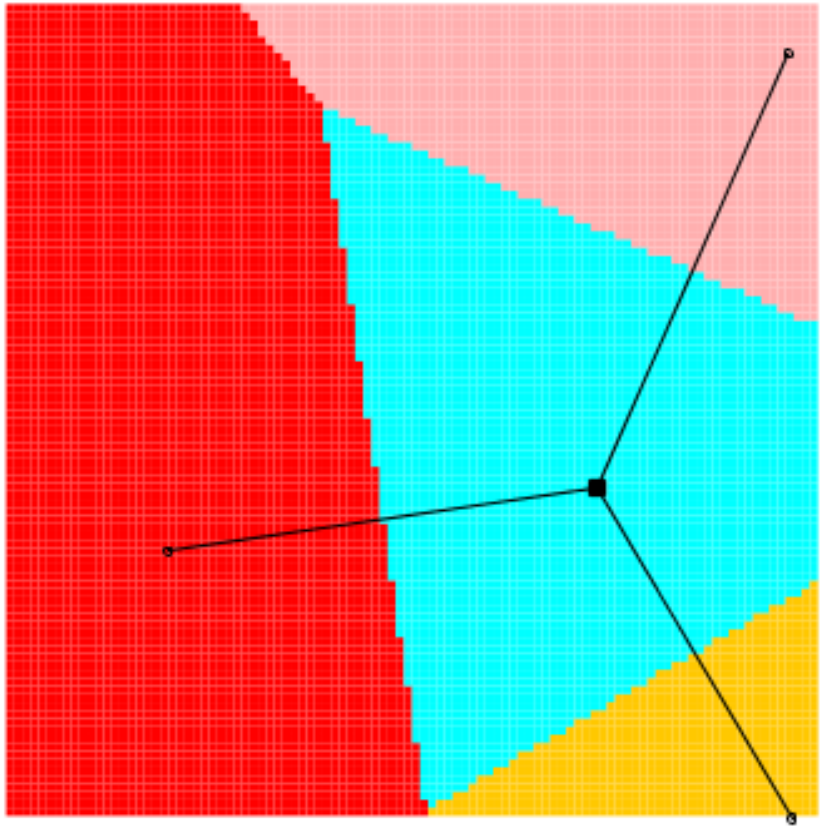


Boundary Forest "Training"

Ground Truth

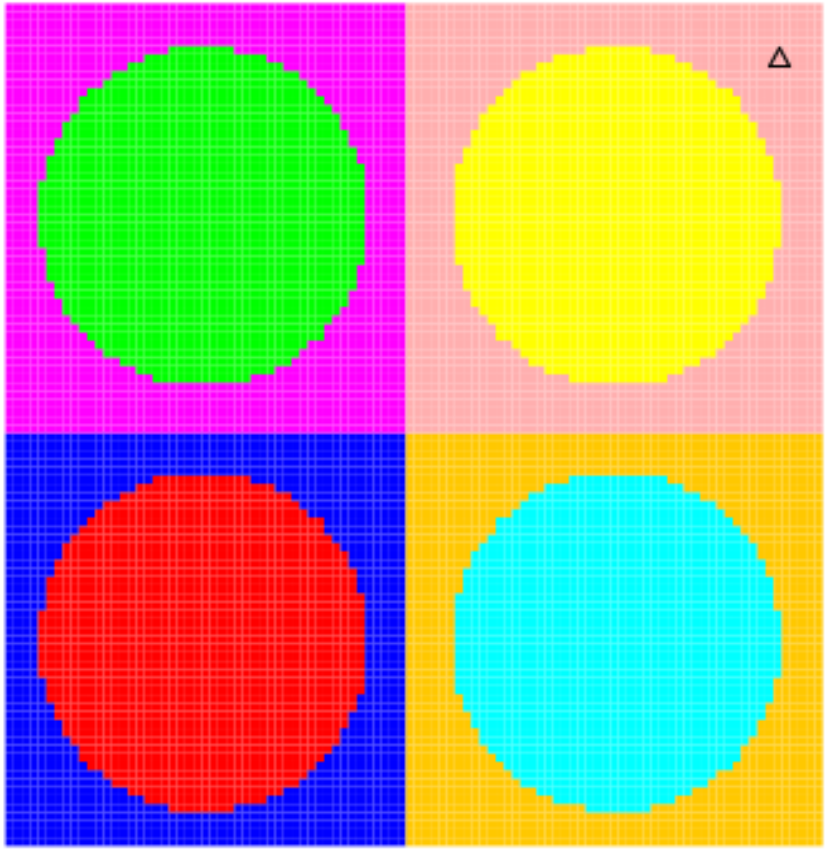


Boundary Tree

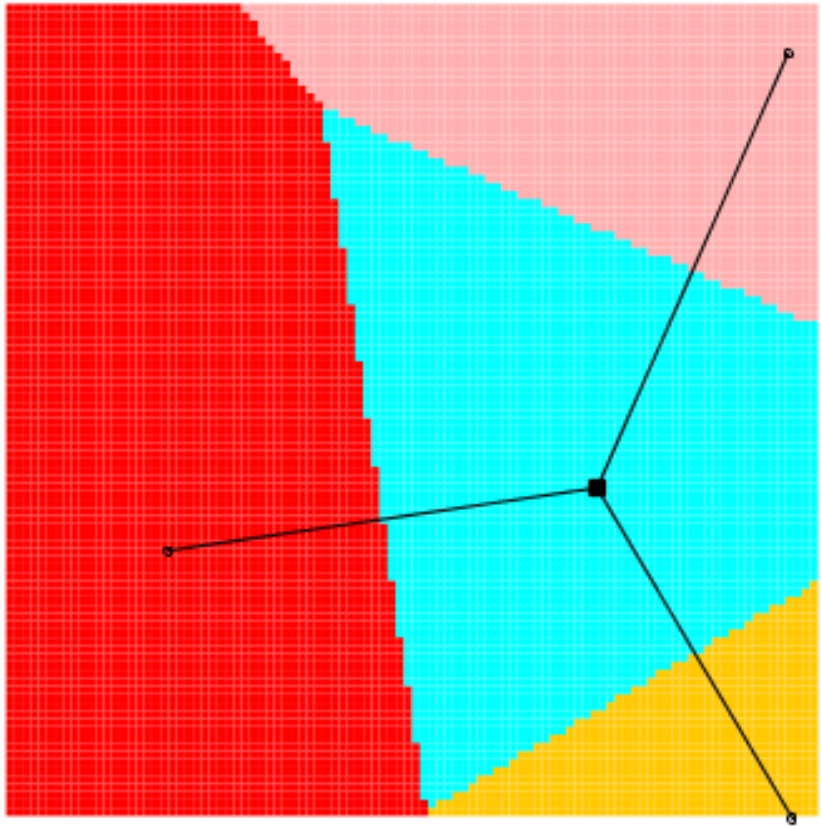


Boundary Forest "Training"

Ground Truth

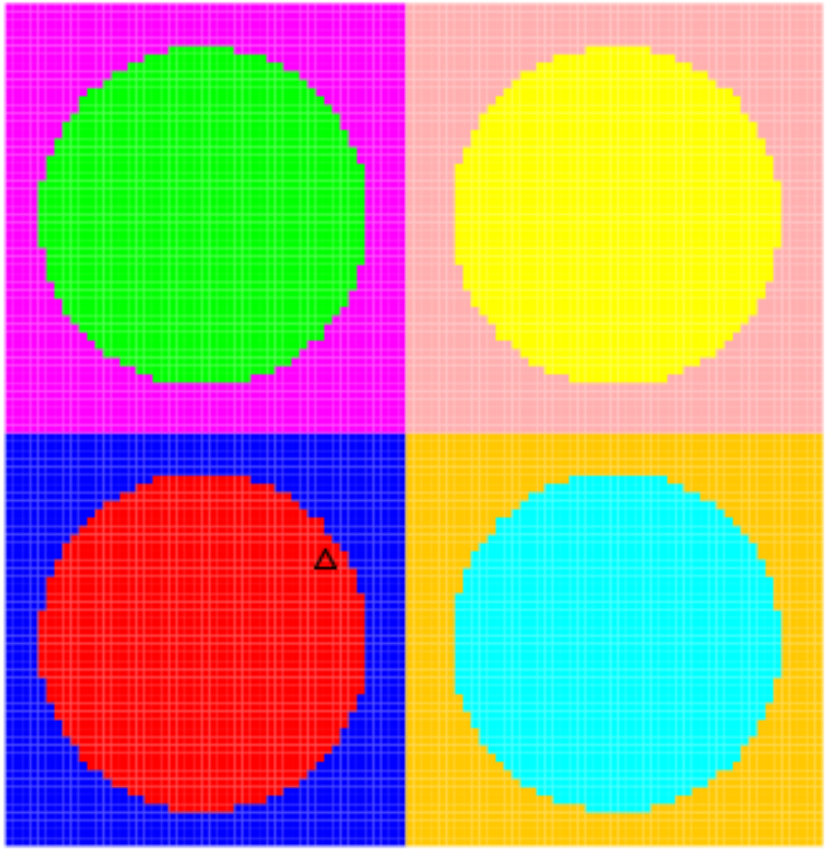


Boundary Tree

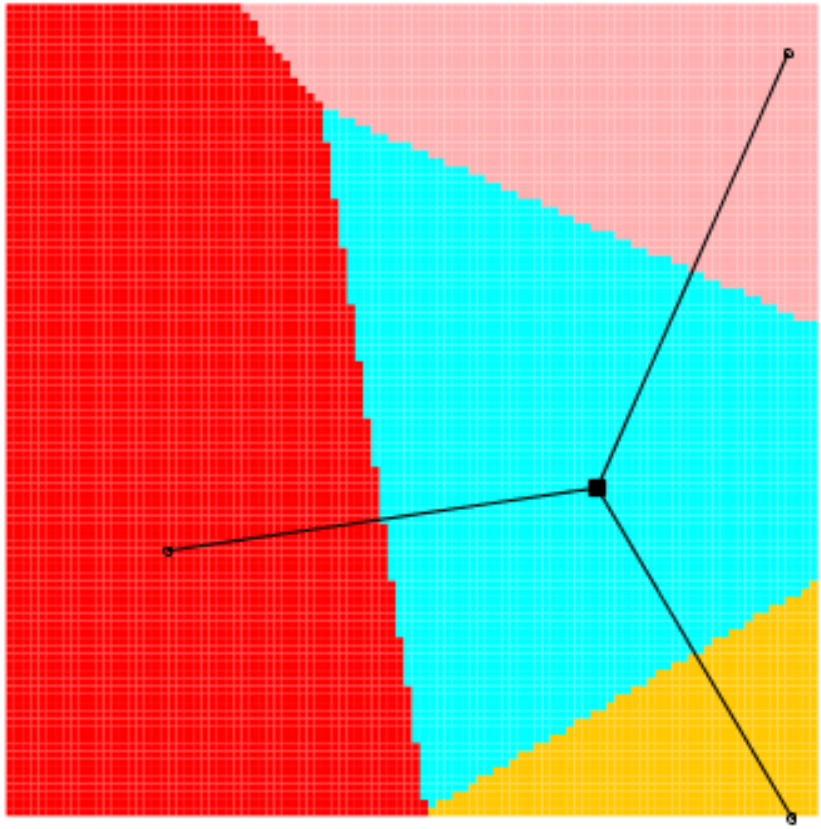


Boundary Forest "Training"

Ground Truth

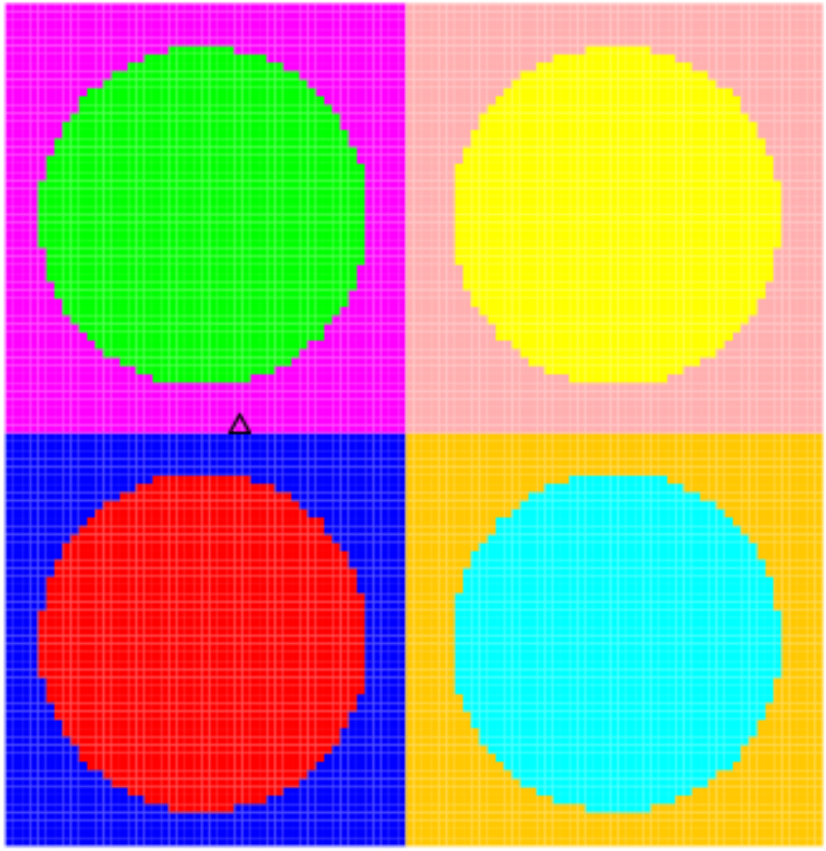


Boundary Tree

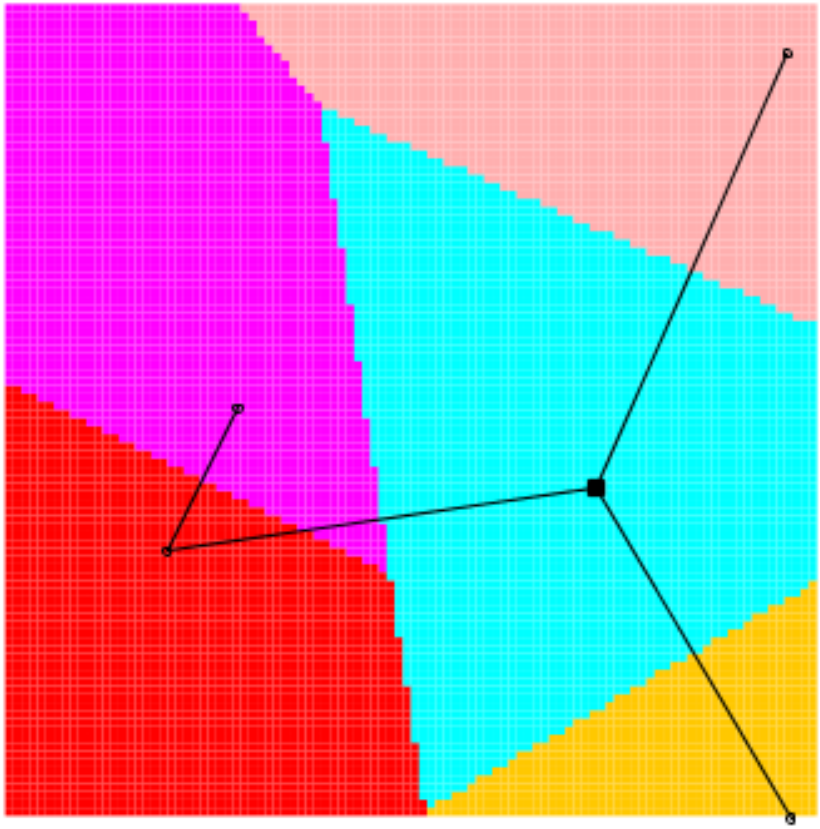


Boundary Forest "Training"

Ground Truth

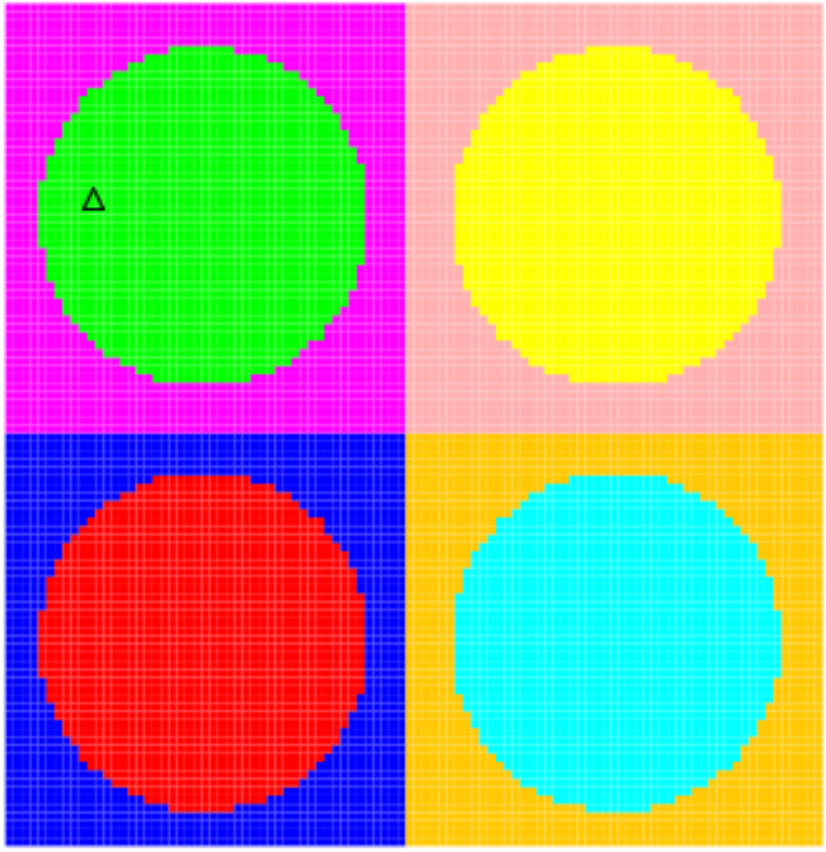


Boundary Tree

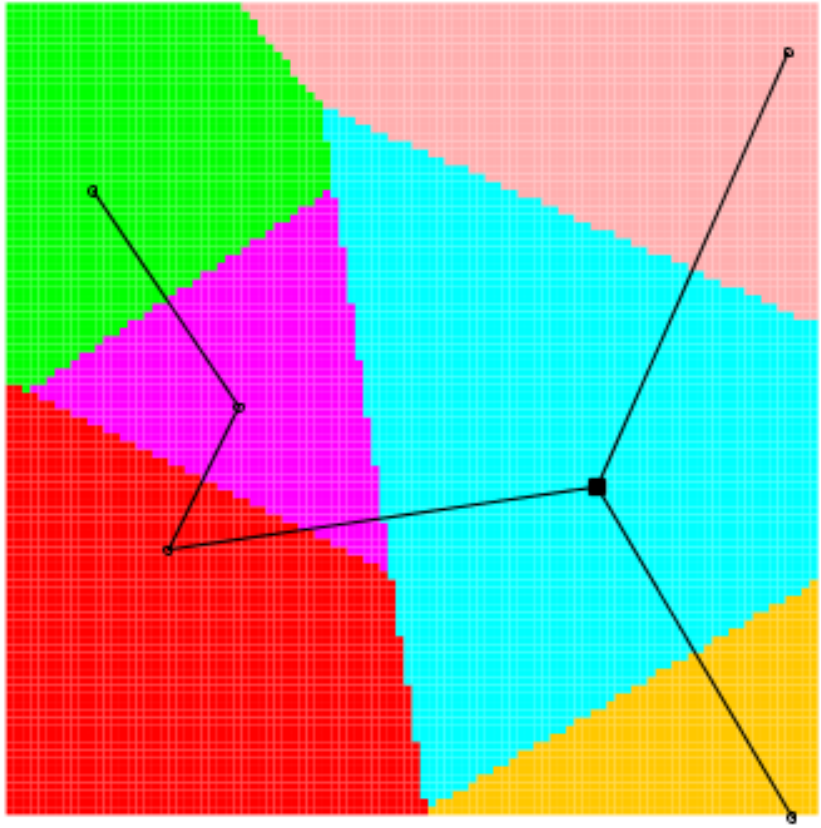


Boundary Forest "Training"

Ground Truth

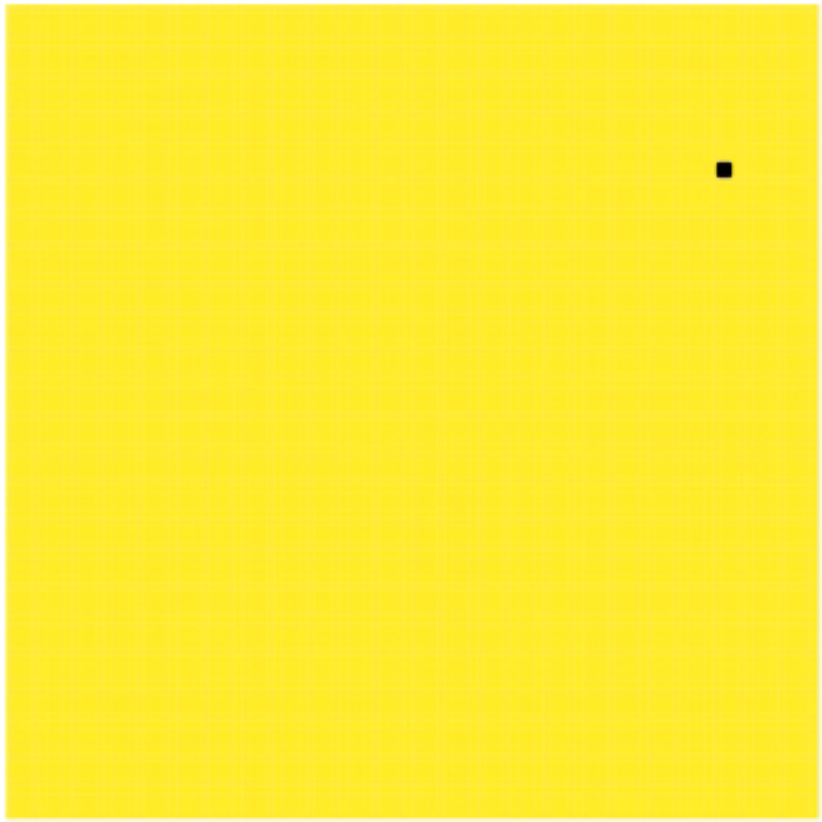


Boundary Tree

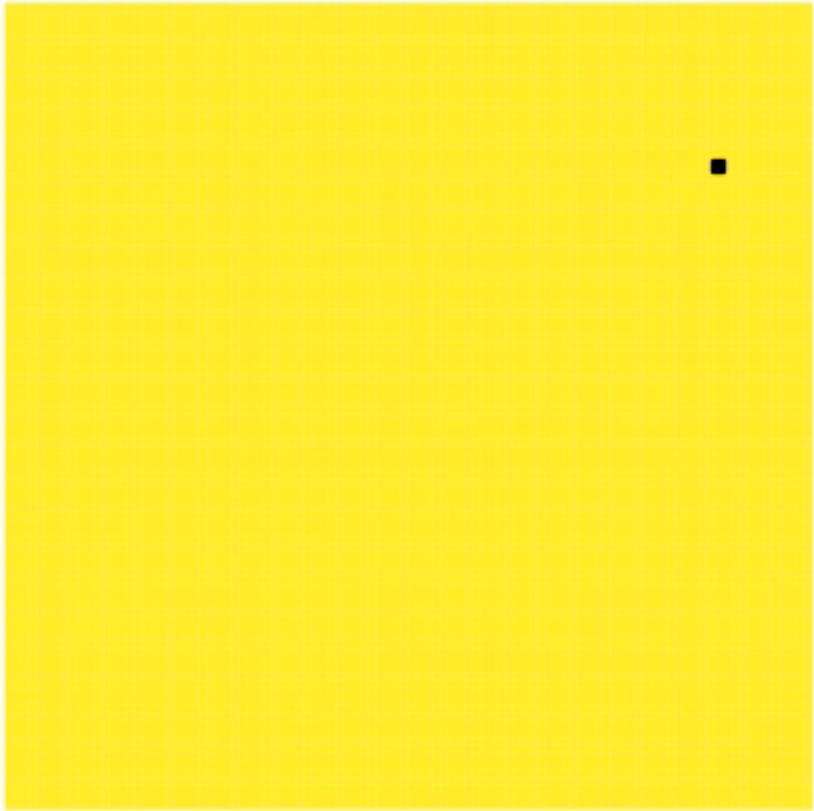


Performance & Scaling

Boundary Tree

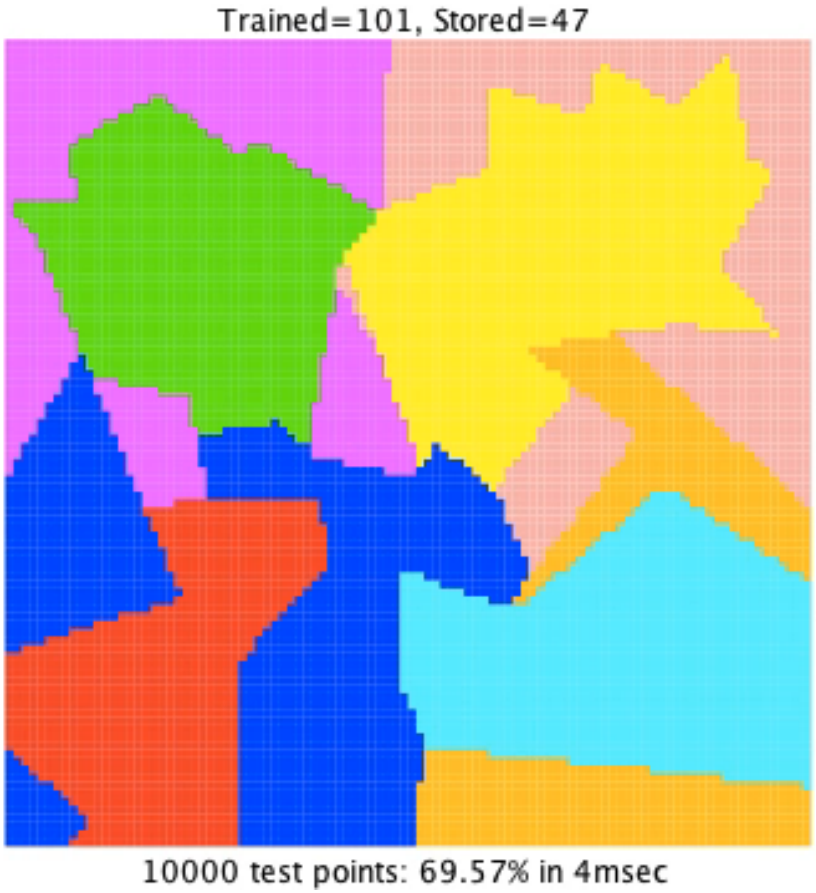


1-NN

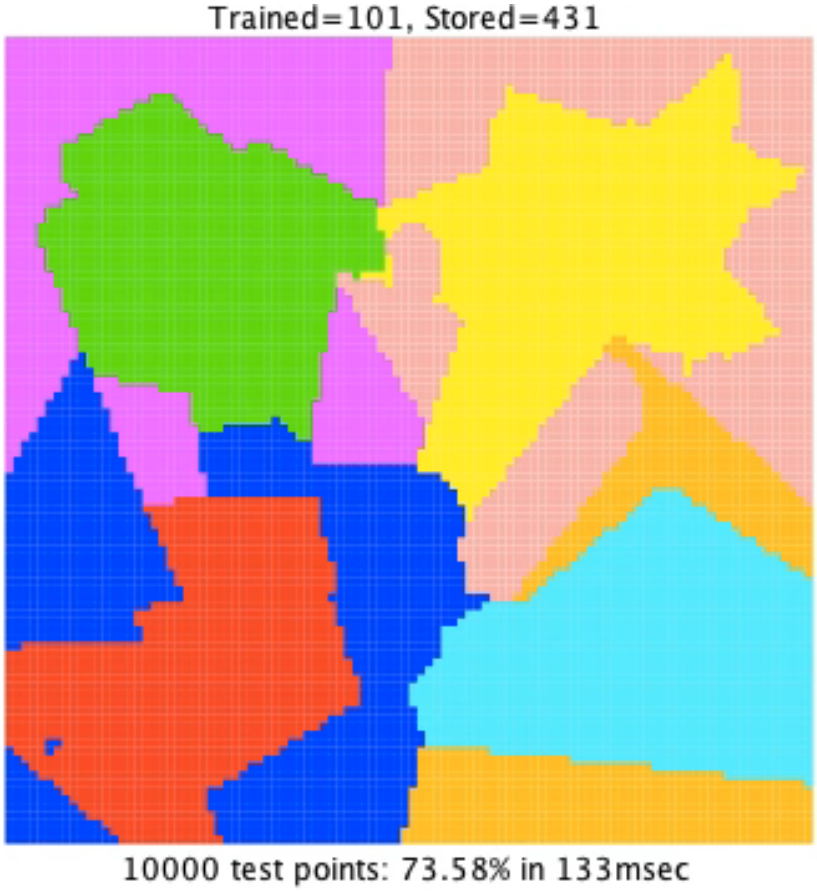


Performance & Scaling

1 Tree



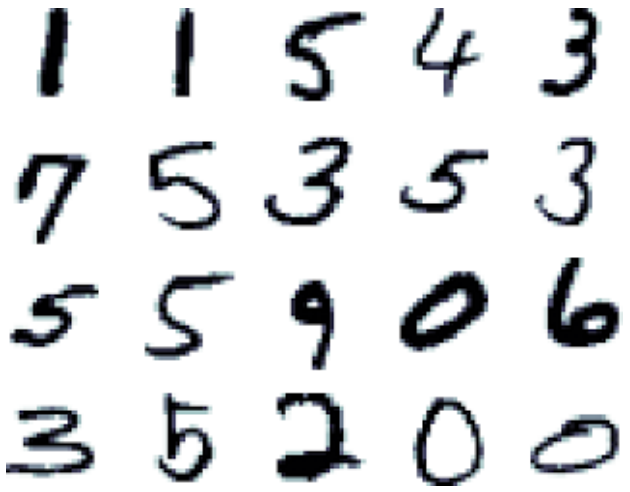
10 Trees



Works Well on Other Problems!

Handwritten Digits

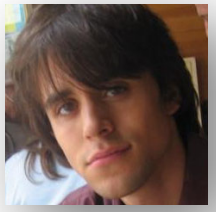
- Dataset: 60k images
- Accuracy: > 97%
- Speed: < 3 ms/image



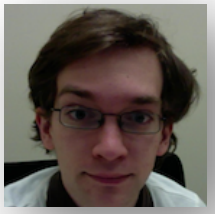
Take-Home Points

- Machine Learning is an exciting field!
  - Systems that get better with data
  - Can be applied to many problems
- The increasing availability of digital systems and cloud computing offers many opportunities for applying Big Data to solve many world problems, *with scalable ML*
- k-Nearest Neighbors learns to predict outcomes by comparing to "close" examples
- The Boundary Forest algorithm achieves accuracy similar to kNN but scales much better

The End :)



Charles Mathy
Machine Learning Scientist
Analog Devices, Lyric Labs



Jonathan Rosenthal
Lab Associate
Disney Research, Boston

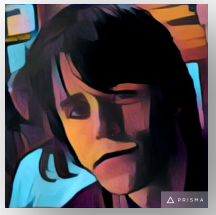
José Bento
Assistant Professor
Computer Science, Boston College

Jonathan Yedidia
Director of AI Research
Analog Devices, Lyric Labs

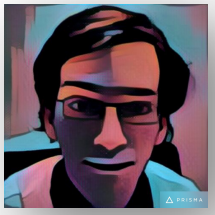
Disney Research

ANALOG DEVICES | Lyric Labs

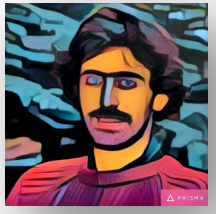
The End :)



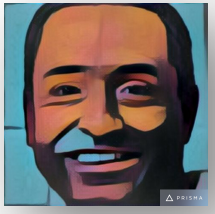
Charles Mathy
Machine Learning Scientist
Analog Devices, Lyric Labs



Jonathan Rosenthal
Lab Associate
Disney Research, Boston



José Bento
Assistant Professor
Computer Science, Boston College



Jonathan Yedidia
Director of AI Research
Analog Devices, Lyric Labs

Disney Research

ANALOG DEVICES | Lyric Labs