Introduction & Model

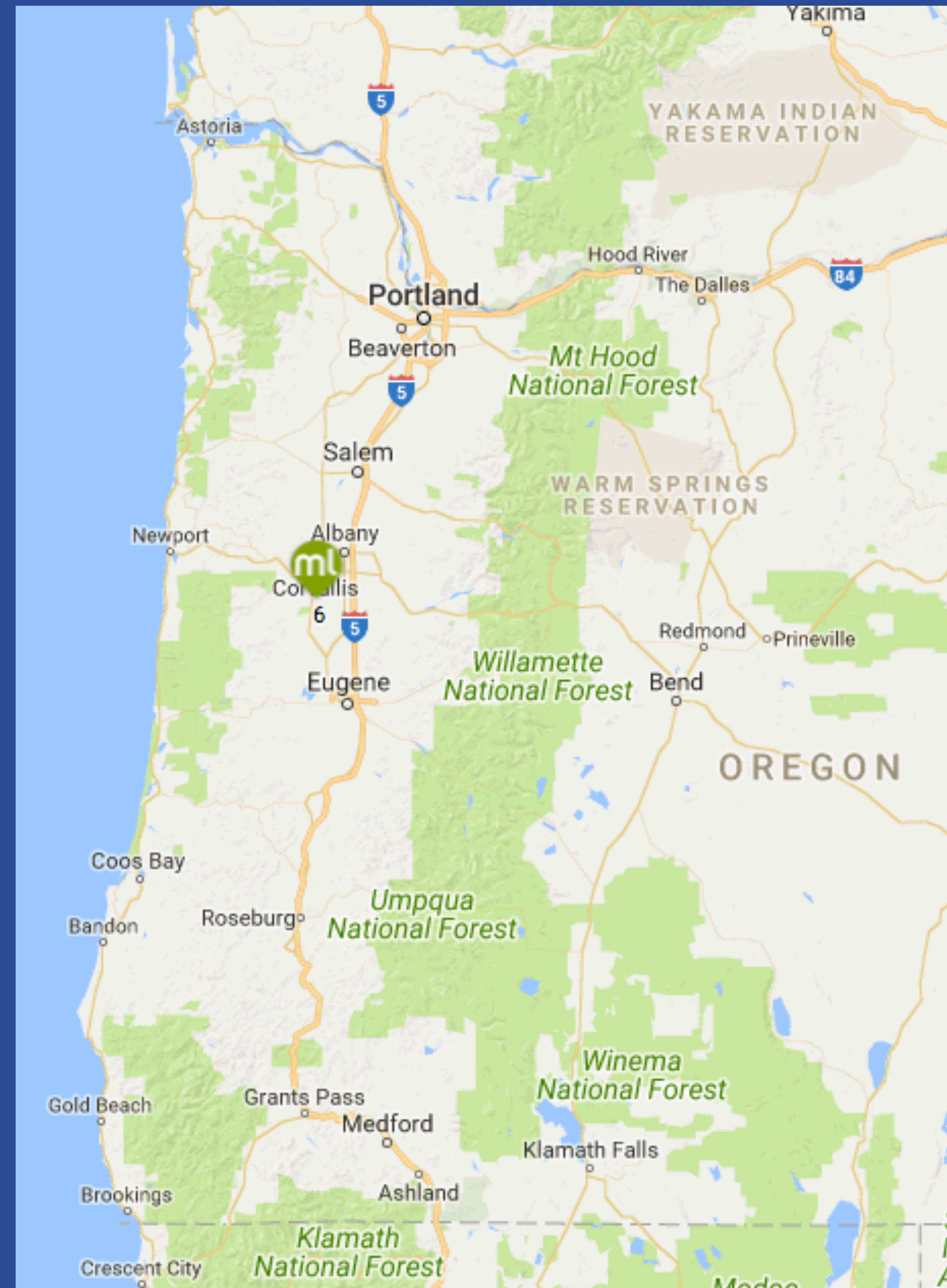
Introduction to Machine Learning & Models

Poul Petersen

CIO, BigML, Inc

A Brief History of BigML

- BigML Mission: To make Machine Learning ***Beautifully Simple***
- BigML Founded in **Corvallis**, Oregon in 2011 - long before ML was "cool"
- You’ve never heard of it?
- Most innovative city in the United States!



A Brief History of BigML

www.gazettetimes.com/news/local/statistics-show-corvallis-no-for-patents/article_e2de00a4-2858-11e0-8594-001cc4c03286.html

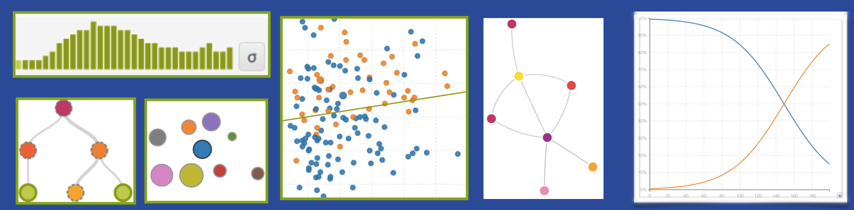
MENU Corvallis Gazette-Times News Sports Photo/Video Buy & Sell Obits Opinion 56°

Statistics show Corvallis No. 1 for patents

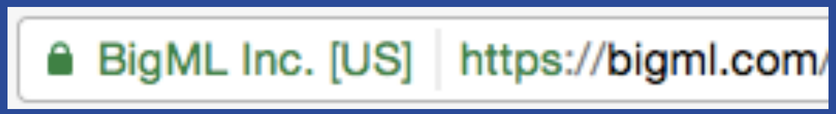
By BENNETT HALL, Gazette-Times reporter Jan 25, 2011 0

BigML Platform

Visualizations

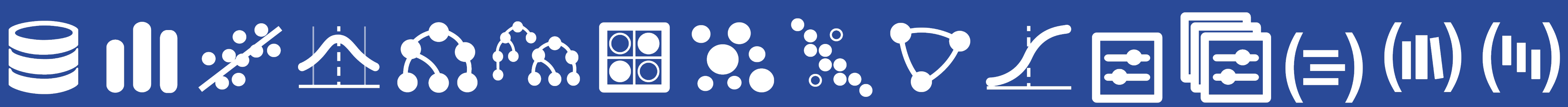


Web-based Frontend

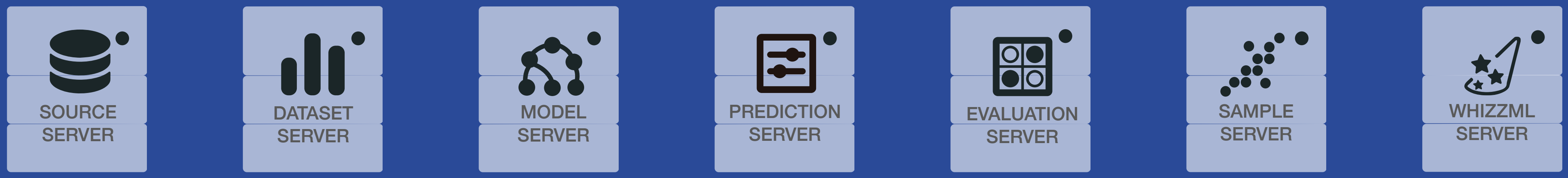


Tools - https://bigml.com/tools

REST API - https://bigml.com/api

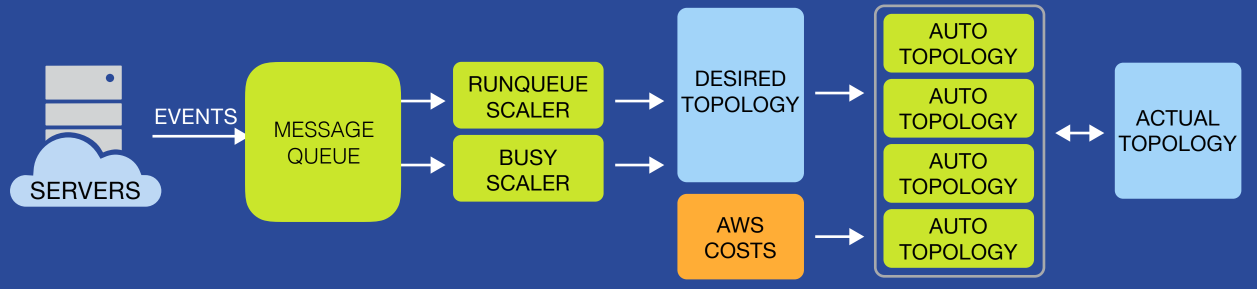


Distributed Machine Learning Backend



Smart Infrastructure

(auto-deployable, auto-scalable)

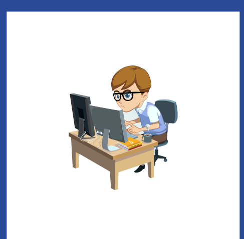


Who am "I"?

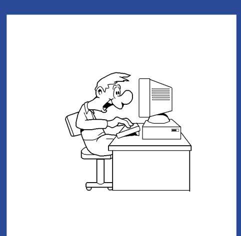
- Poul Petersen, BigML, CIO since Nov 2011
- Background in Mathematics, Physics, Engineering
- Wrote "Sauron"
  - Fully automated bigml.com, bigml.com.au
  - Ported Sauron for private deployments
- Created:
  - BigML for Alexa (& house recommender)
  - Original PredictServer
  - First "scriptify", called "reifier"
- Demos, Training, and Sales Support

Who are you?

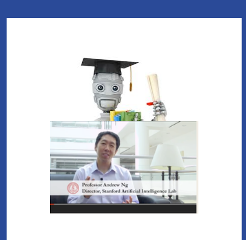
Expert: Published papers at KDD, ICML, NIPS, etc or developed own ML algorithms used at large scale



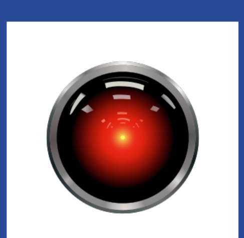
Aficionado: Understands pros/cons of different techniques and/or can tweak algorithms as needed



Practitioner: Very familiar with ML packages (Weka, Scikit, BigML, etc.)



Newbie: Just taking Coursera ML class or reading an introductory book to ML



Absolute beginner: ML sounds like science fiction

Before we Begin…

Pacing Goal

Reality

Before we Begin…

- All course materials including slides, CSVs and related videos are hosted on the **Training website:**

http://training.bigml.com

- Each module will have learning exercises to be performed in-session, so…
- You need a laptop and a BigML account: **Get signed up**

https://bigml.com/account/register

Machine Learning Motivation

Imagine:

- You are looking to buy a house
- Recently found a house you like
- Is the asking price fair?

What Next?

Machine Learning Motivation

Why not ask an expert?

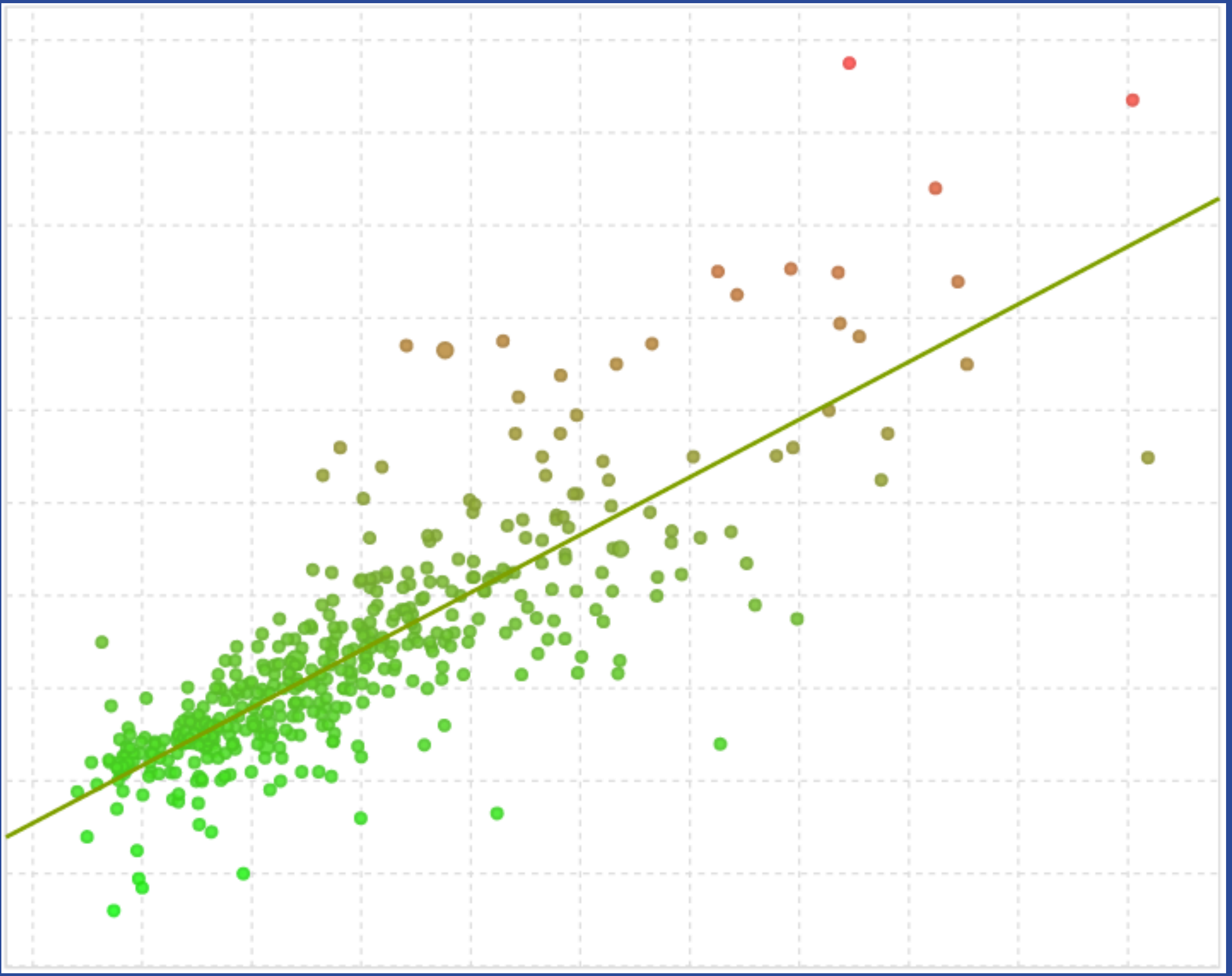
- Experts can be rare / expensive
- Hard to validate experience:
  - Experience with similar properties?
  - Do they consider all relevant variables?
  - Knowledge of market up to date?
- Hard to validate answer:
  - How many times expert right / wrong?
  - Probably can't explain decision in detail
- Humans are not good at intuitive statistics

Data vs Expert

Replace the expert with data?

- Intuition: square footage relates to price.
- Collect data from past sales

SQFT	SOLD	PREDICT
2424	360000	400262
1785	307500	320195
1003	185000	222211
4135	600000	614651
1676	328500	306538
1012	247000	223339
3352	420000	516541
2825	435350	450508



PRICE = 125.3*SQFT + 96535

Data vs Expert

Replace the expert scorecard

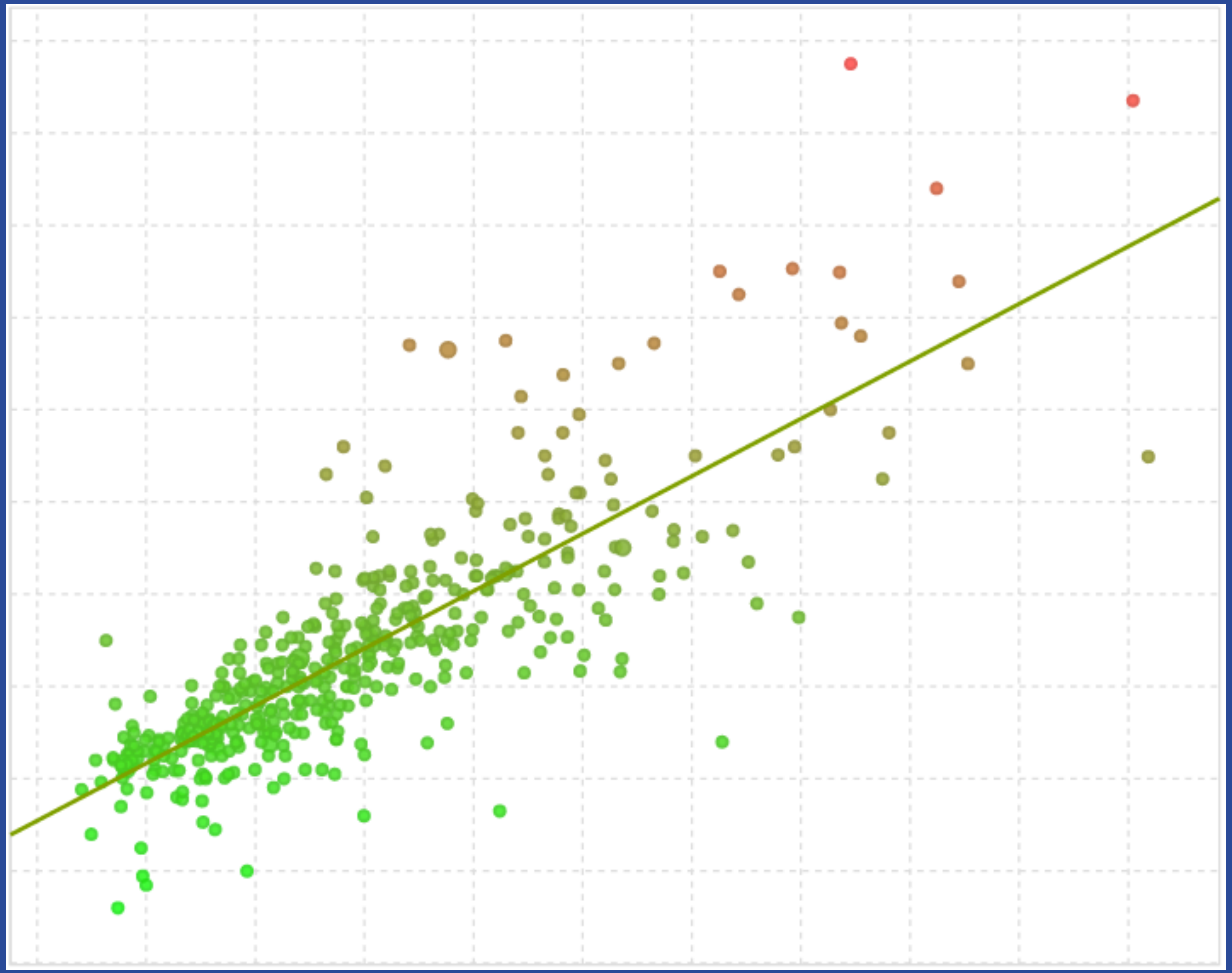
- ~~Experts can be rare / expensive~~
- Hard to validate experience:
    - ~~Experience with similar properties~~?
    - Do they consider all relevant variables?
    - ~~Knowledge of market up to date~~?
- Hard to validate answer:
    - ~~How many times expert right / wrong~~?
    - ~~Probably can't explain decision in detail~~
- Humans are not good at intuitive statistics

Data vs Expert

Replace the expert with data

- Intuition: square footage relates to price.
- Collect data from past sales

SQFT	SOLD
2424	360000
1785	307500
1003	185000
4135	600000
1676	328500
1012	247000
3352	420000
2825	435350



PRICE = 125.3*SQFT + 96535

More Data!

SQFT	BEDS	BATHS	ADDRESS	LOCATION	LOT SIZE	YEAR BUILT	PARKING SPOTS	LATITUDE	LONGITUDE	SOLD
2424	4	3	1522 NW Jonquil	Timberhill SE 2nd	5227	1991	2	44.594828	-123.269328	360000
1785	3	2	7360 NW Valley Vw	Country Estates	25700	1979	2	44.643876	-123.238189	307500
1003	2	1	2620 NW Chinaberry	Tamarack Village	4792	1978	2	44.593704	-123.295424	185000
4135	5	3.5	4748 NW Veronica	Suncrest	6098	2004	3	44.5929659	-123.306916	600000
1676	3	2	2842 NW Monterey	Corvallis	8712	1975	2	44.5945279	-123.291523	328500
1012	3	1	2320 NW Highland	Corvallis	9583	1959	2	44.591476	-123.262841	247000
3352	4	3	1205 NW Ridgewood	Ridgewood 2	60113	1975	2	44.579439	-123.333888	420000
2825		3	411 NW 16th	Wilkins Addition	4792	1938	1	44.570883	-123.272113	435350

Uhhhh........

- Can we still fit a line to 10 variables? (well, yes)
- Will fitting a line give good results? (unlikely)
- What about those text fields and categorical values?

Linear Regression

What Just Happened?

- We started with Housing data as a CSV from Redfin
- We uploaded the CSV to create Source
- Then we created a Dataset from the Source and reviewed the summary statistics and scatter plot
- With 1-click we build a Linear Regression which can predict home prices based on all the housing features
- We explored the Model and used it to make a Prediction of a home with SQFT=1000 and LOT SIZE=0
- We noticed that it was impossible to disable some fields like LOCATION in the prediction form.

Aside: BigML Resources

- Everything created on BigML is a *resource*
- Resources are:
  - Immutable: Ensures consistency, repeatability
  - Assigned a canonical ID
  - Traceable: How the resource is created is part of resource
  - Always available via both the API and the UI
  - Sharable with a secret link, api key, or organization
- Working with BigML is a process of creating resources
- Resources can be grouped into Projects
  - Projects can be private or part of an organization
  - Organizations allow real-time sharing of all resources in a project

Your Turn!

- Create a new project called “Training”
- Build a Linear Regression for the PDX dataset
    - The CSV is in http://training.bigml.com
- What is the predicted price of:
    - SQFT = 1,000, LOT SIZE=0
    - All other fields at default
    - What inputs can you not disable?
- Bonus:
    - Share your model with someone…

More Data!

SQFT	BEDS	BATHS	ADDRESS	LOCATION	LOT SIZE	YEAR BUILT	PARKING SPOTS	LATITUDE	LONGITUDE	SOLD
2424	4	3	1522 NW Jonquil	Timberhill SE 2nd	5227	1991	2	44.594828	-123.269328	360000
1785	3	2	7360 NW Valley Vw	Country Estates	25700	1979	2	44.643876	-123.238189	307500
1003	2	1	2620 NW Chinaberry	Tamarack Village	4792	1978	2	44.593704	-123.295424	185000
4135	5	3.5	4748 NW Veronica	Suncrest	6098	2004	3	44.5929659	-123.306916	600000
1676	3	2	2842 NW Monterey	Corvallis	8712	1975	2	44.5945279	-123.291523	328500
1012	3	1	2320 NW Highland	Corvallis	9583	1959	2	44.591476	-123.262841	247000
3352	4	3	1205 NW Ridgewood	Ridgewood 2	60113	1975	2	44.579439	-123.333888	420000
2825		3	411 NW 16th	Wilkins Addition	4792	1938	1	44.570883	-123.272113	435350

Uhhhh........

- Can we still fit a line to 10 variables? (well, yes)
- Will fitting a line give good results? (unlikely)
- What about those text fields and categorical values?

Mythical ML Model?

- High representational power
  - Fitting a line is an example of low
  - Deep neural networks is an example of high
- High Ease-of-use
  - Easy to configure - relatively few parameters
  - Easy to interpret - how are decisions made?
  - Easy to put into production
- Ability to work with real-world data
  - Mixed data types: numeric, categorical, text, etc
  - Handle missing values
  - Resilient to outliers
- There are actually hundreds of possible choices…

Model Choices

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

Editor: Russ Greiner

Abstract

We evaluate **179 classifiers** arising from **17 families** (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods), implemented in Weka, R (with and without the caret package), C and Matlab, including all the relevant classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. **The classifiers most likely to be the bests are the random forest** (RF)

A churn problem…

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Hey! Those are not numbers!

Supervised Learning

Regression

label(s)

animal	state	…	proximity	min_kmh
tiger	hungry	…	close	70
hippo	angry	…	far	10
…	….	…	…	…

Classification

animal	state	…	proximity	action
tiger	hungry	…	close	run
elephant	happy	…	far	take picture
…	…	…	…	…

Multi-Label Classification

animal	state	…	proximity	action1	action2
tiger	hungry	…	close	run	look untasty
elephant	happy	…	far	take picture	call friends
…	…	…	…	…	…

Question: What would Unsupervised Learning look like?

ML: Two Methods

Supervised

- Requires labelled data
- Goal is to predict the label often called the objective
- Can be evaluated against the label
- Algorithms:
    - Models/Ensembles
    - Logistic Regression
    - Deepnets
    - Time Series

Unsupervised

- Does not require labelled data
- Goal is “discovery”, with algorithms focused on type
- Each algorithm has it’s own quality measures
- Algorithms:
    - Clustering
    - Anomaly Detection
    - Association Discovery
    - Topic Modeling

Back to the Data…

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Minutes Used > 200

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Website Visits > 0

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Last Bill > $180

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Last Bill > $180, Support Calls > 0

Minutes Used	Last Month's Bill	Calls To Support	Website Visits	Churn?
104	103.60	0	0	No
124	56.33	1	0	No
56	214.60	2	0	Yes
2410	305.60	0	5	No
536	145.70	0	0	No
234	122.09	0	1	No
201	185.76	1	7	Yes
111	83.60	3	2	No

Models

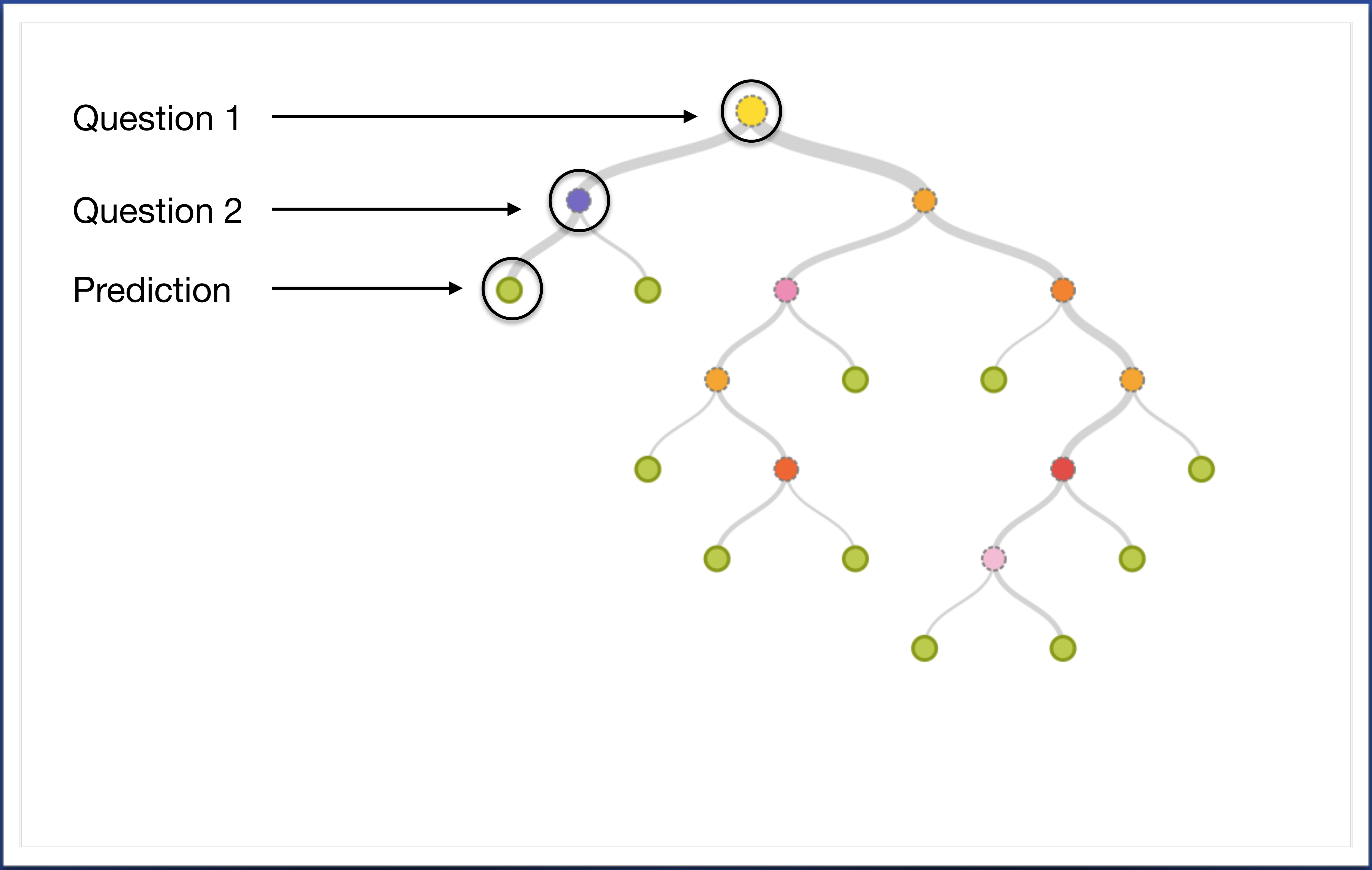
What Just Happened?

- We started with Churn data as a CSV
- We uploaded the CSV to create **Source**
- Then we created a **Dataset** from the **Source** and reviewed the summary statistics
- With 1-click we build a **Model** which can predict which customers will churn.
- We explored the **Model** and used it to make a **Prediction**

Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time

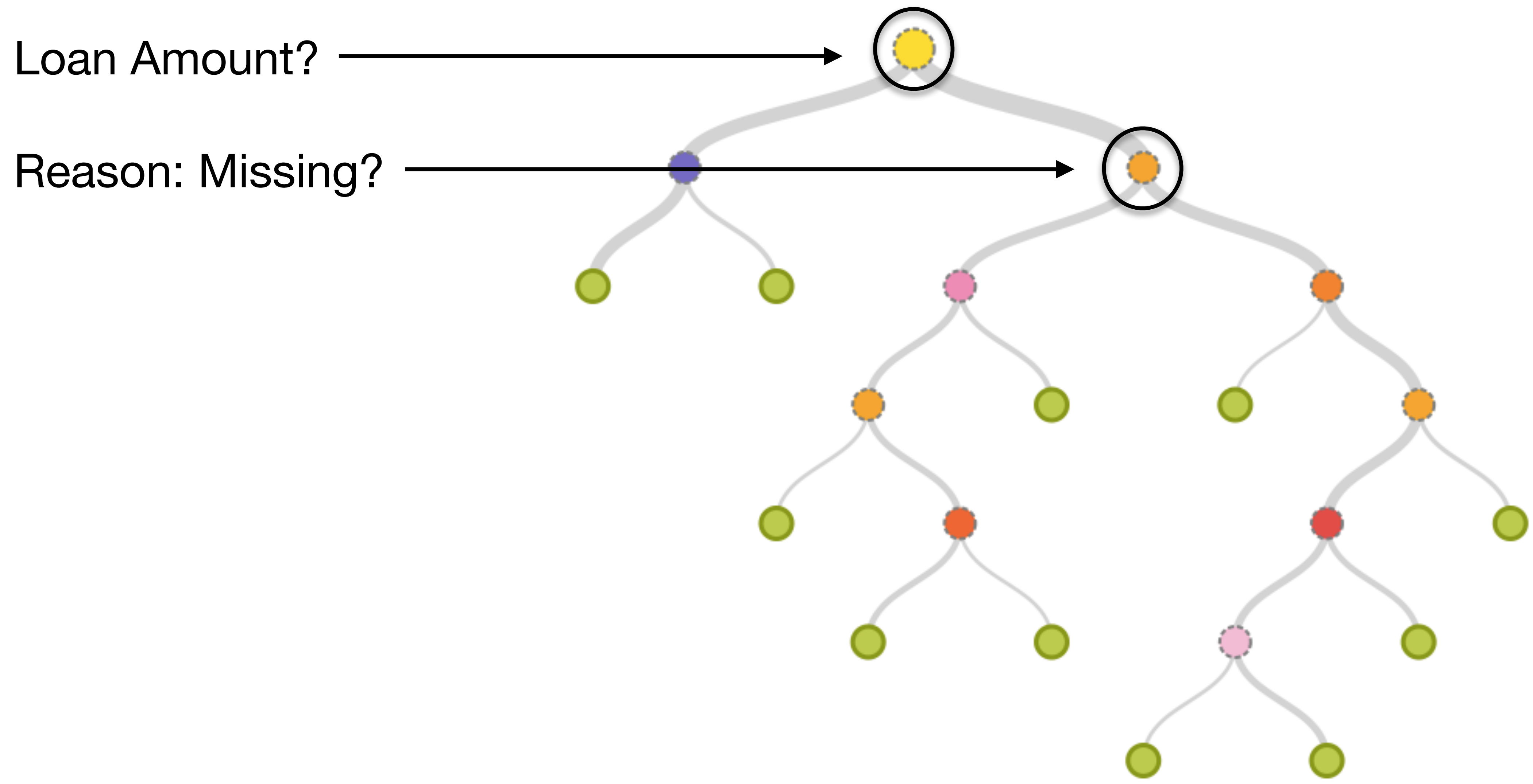
DT Predictions



Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
    - Useless features are automatically ignored
    - Works with un-normalized data
    - Works with missing data at Training

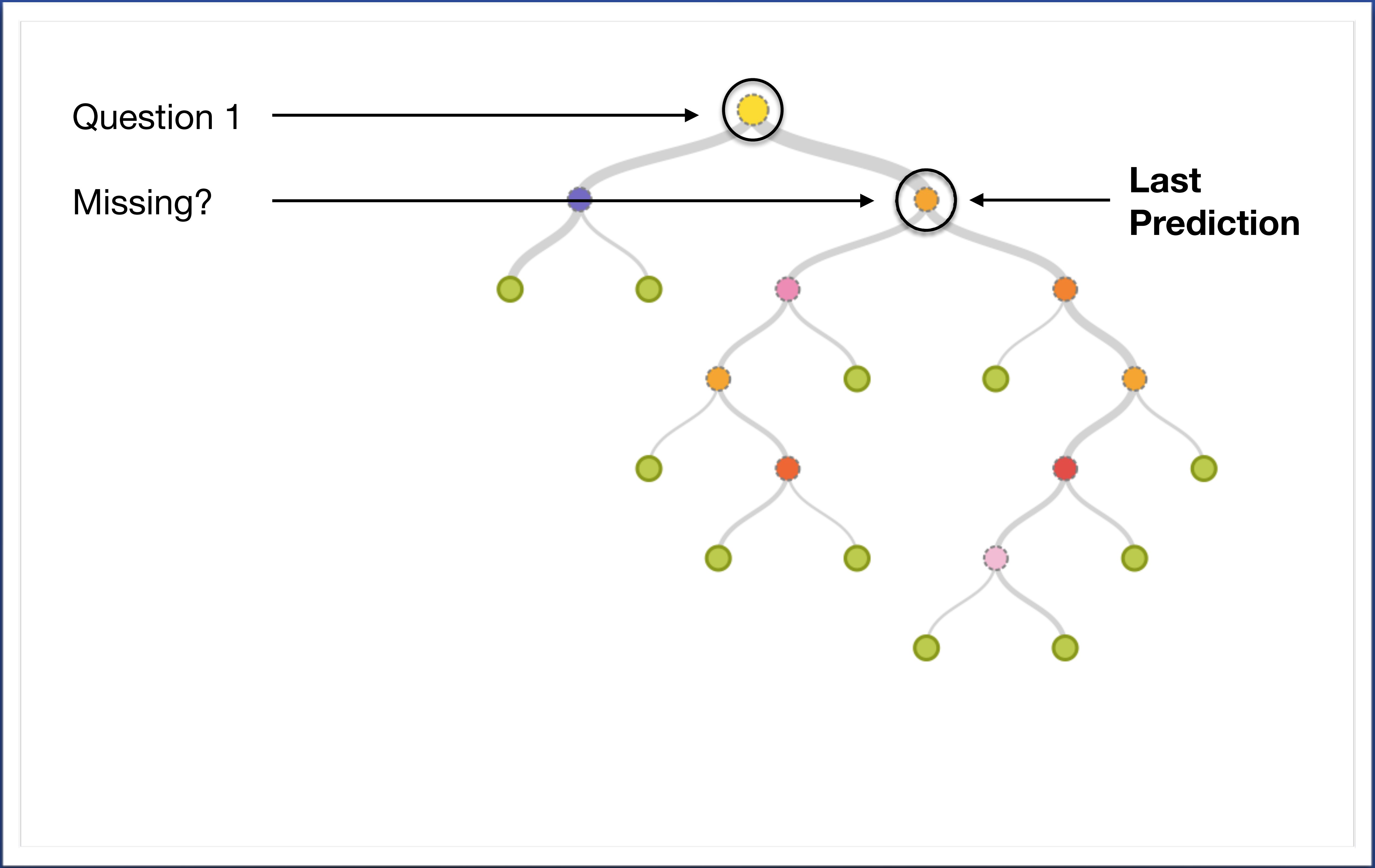
Training with Missing



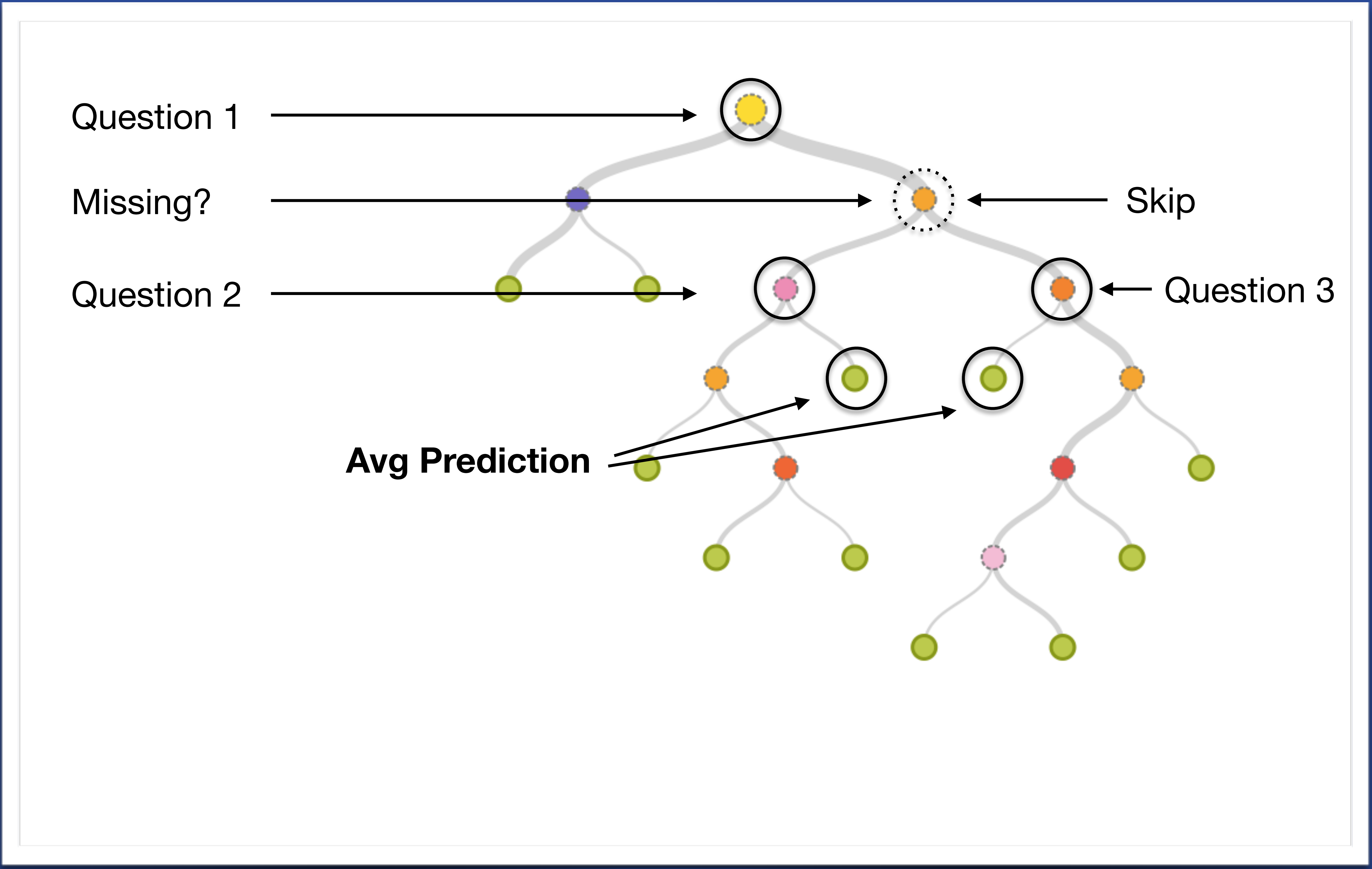
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
    - Useless features are automatically ignored
    - Works with un-normalized data
    - Works with missing data at Training **& Prediction**

Predictions with Missing



Predictions with Missing



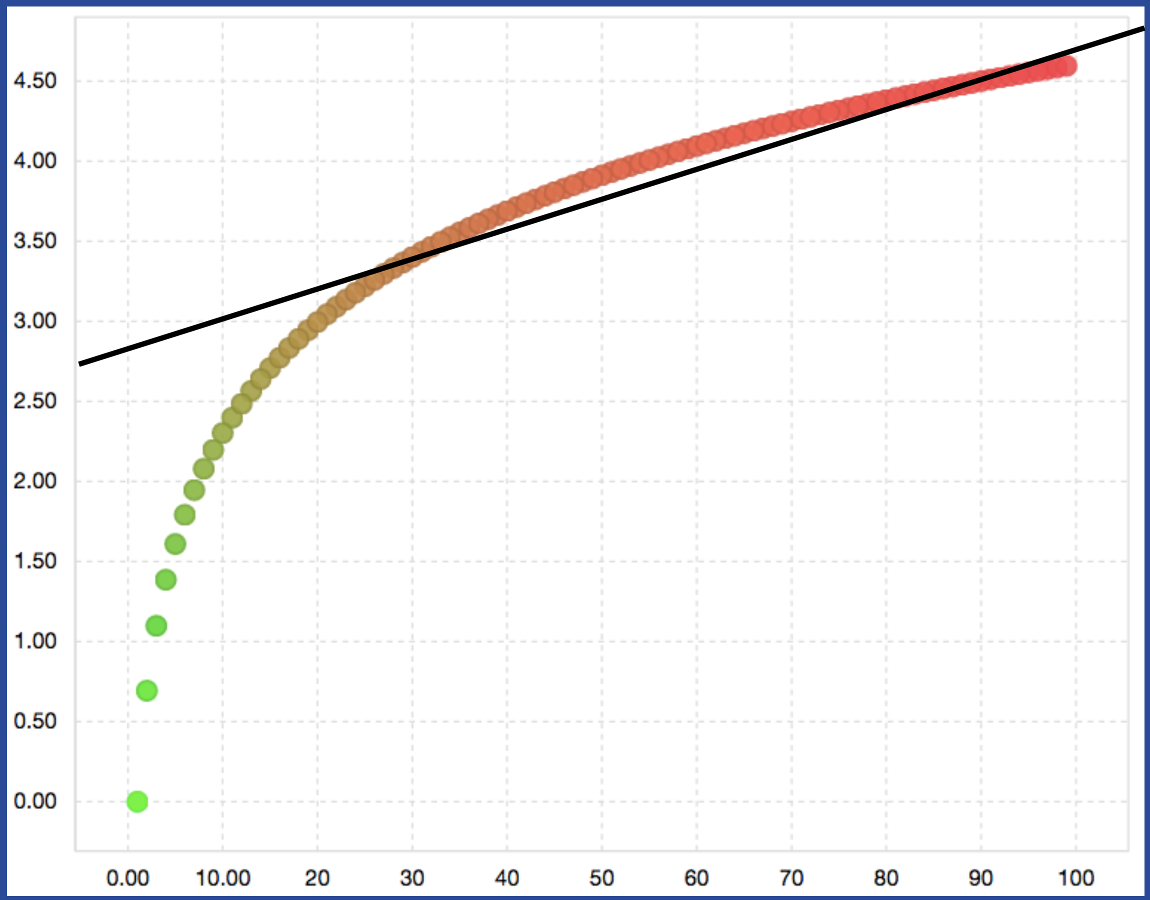
Why Decision Trees

- Works for classification or regression
- Easy to understand: splits are features and values
- Lightweight and super fast at prediction time
- Relatively parameter free
- Data can be messy
    - Useless features are automatically ignored
    - Works with un-normalized data
    - Works with missing data at Training & Prediction
    - Resilient to outliers
- High representational power
- Works easily with mixed data types

Why Not Decision Trees

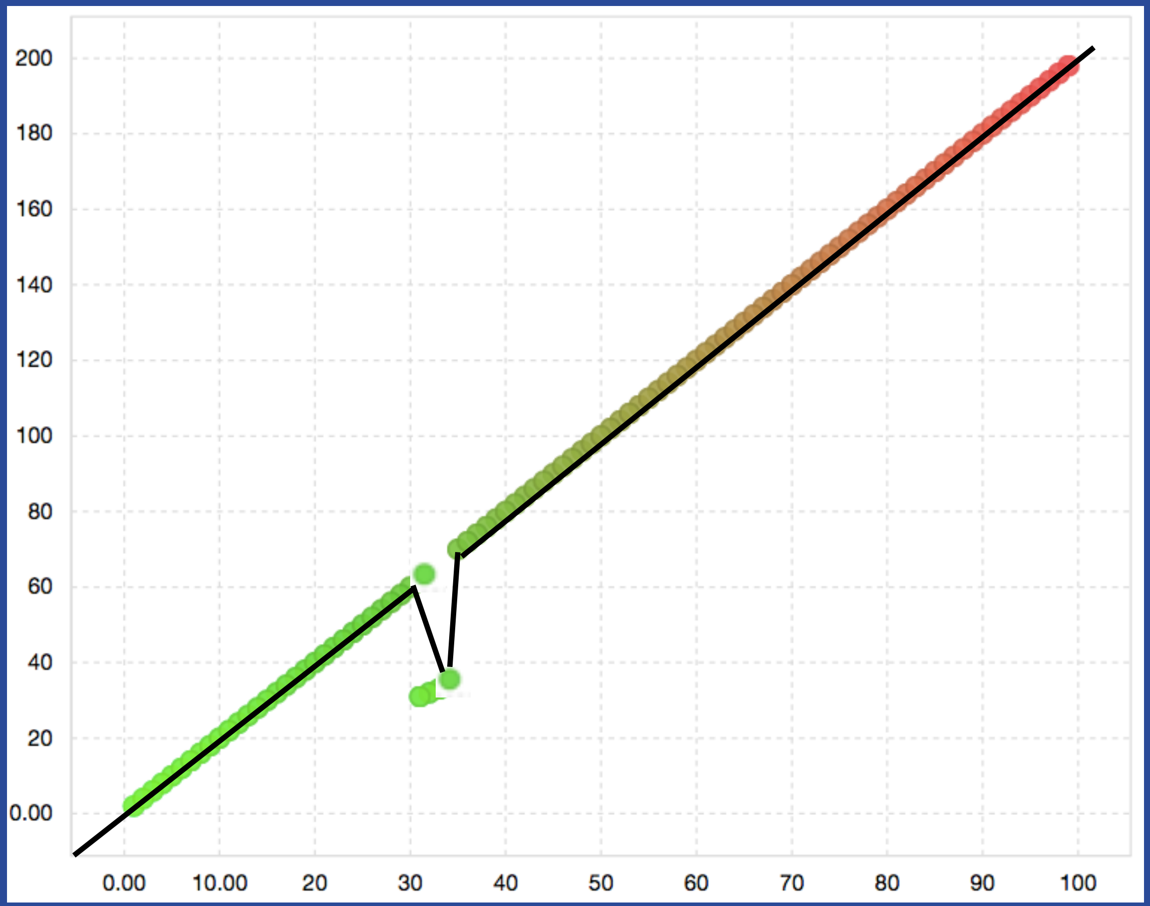
- Slightly prone to over-fitting. (wait: what is that?)

Learning Problems (fit)



Under-fitting

- Model does not fit well enough
- Does not capture the underlying trend of the data
- Change algorithm or features



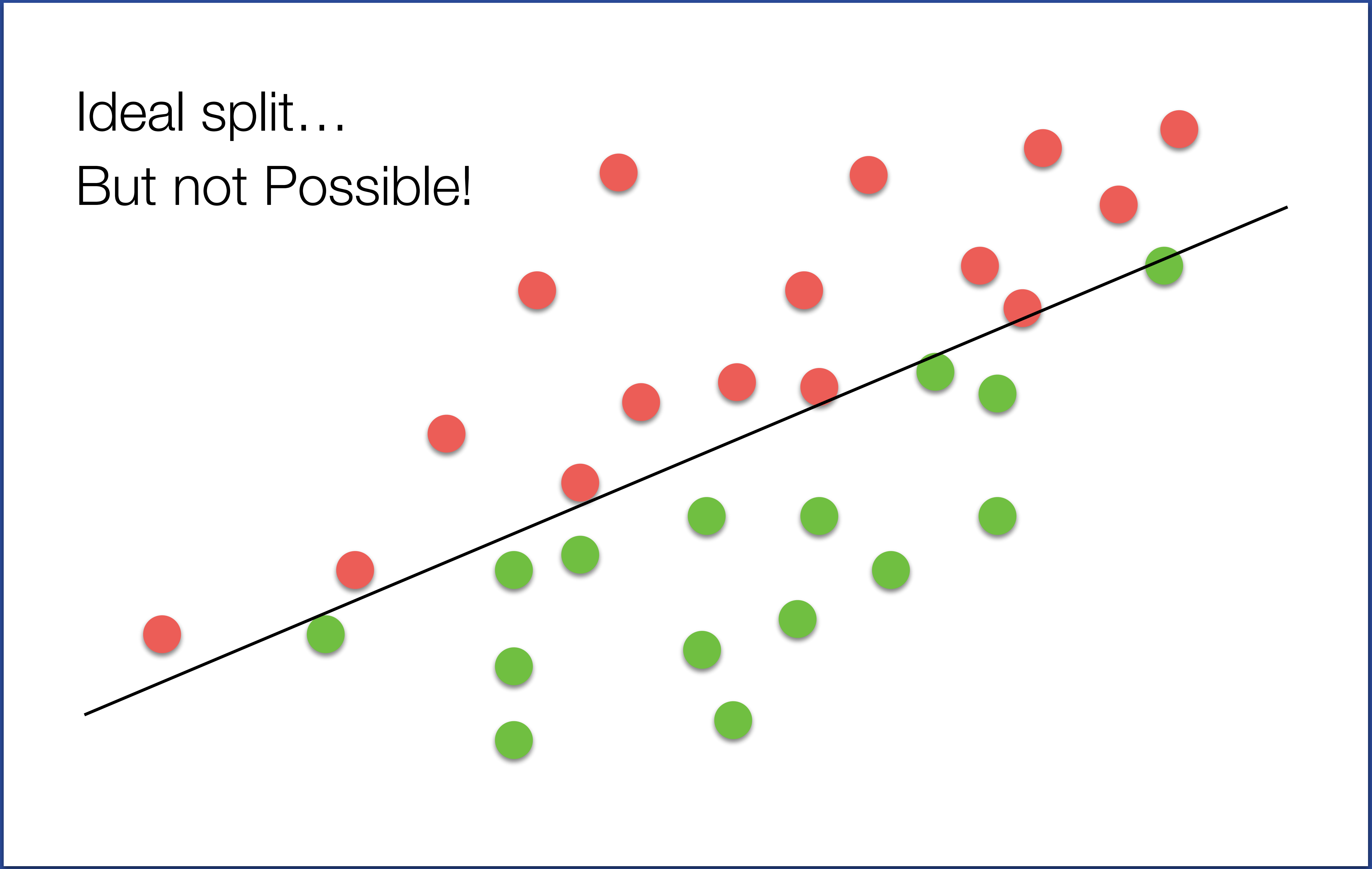
Over-fitting

- Model fits too well does not “generalize”
- Captures the noise or outliers of the data
- Change algorithm or filter outliers

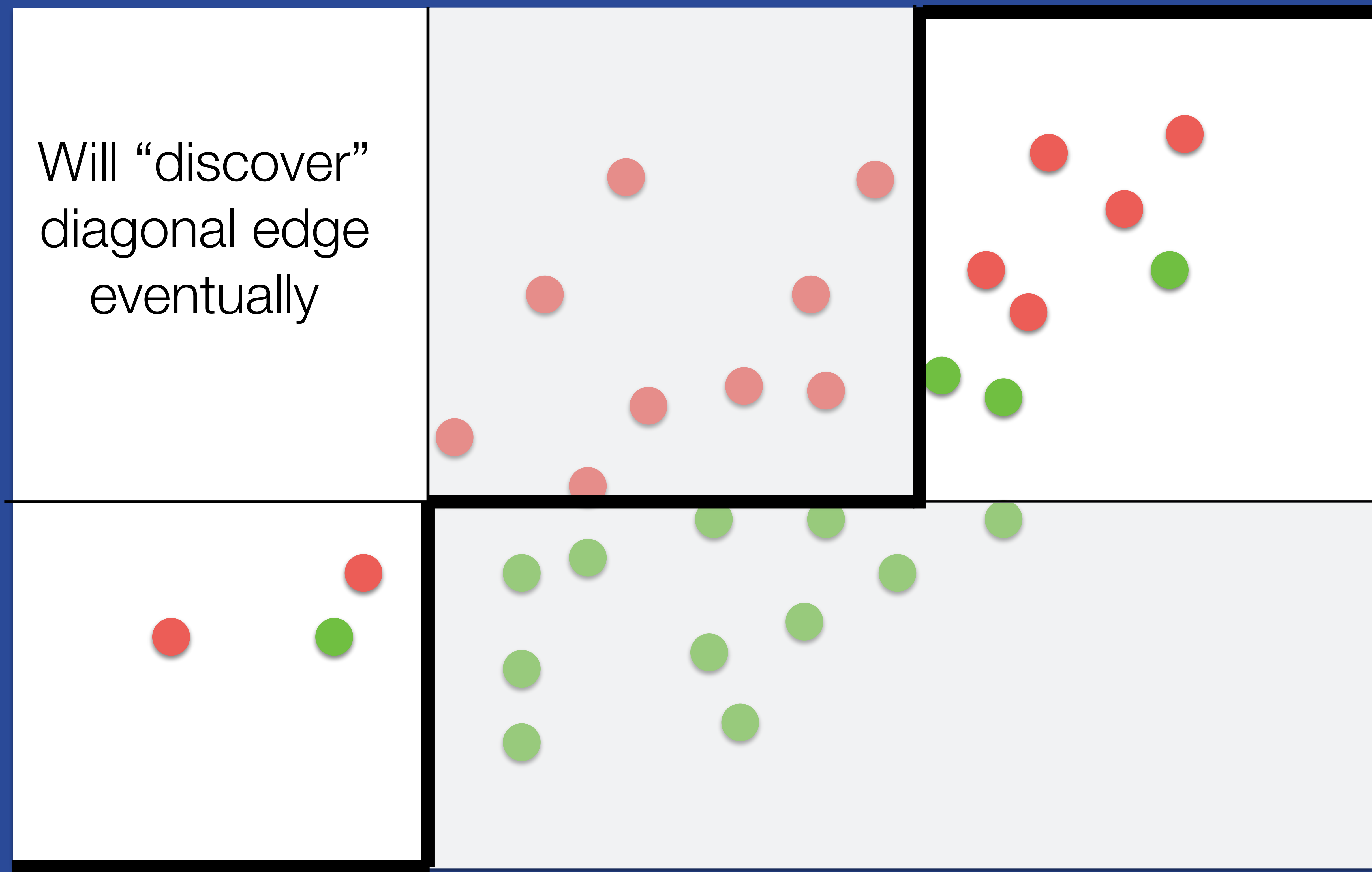
Why Not Decision Trees

- Slightly prone to over-fitting
    - But we'll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes

Splits Parallel to Axis



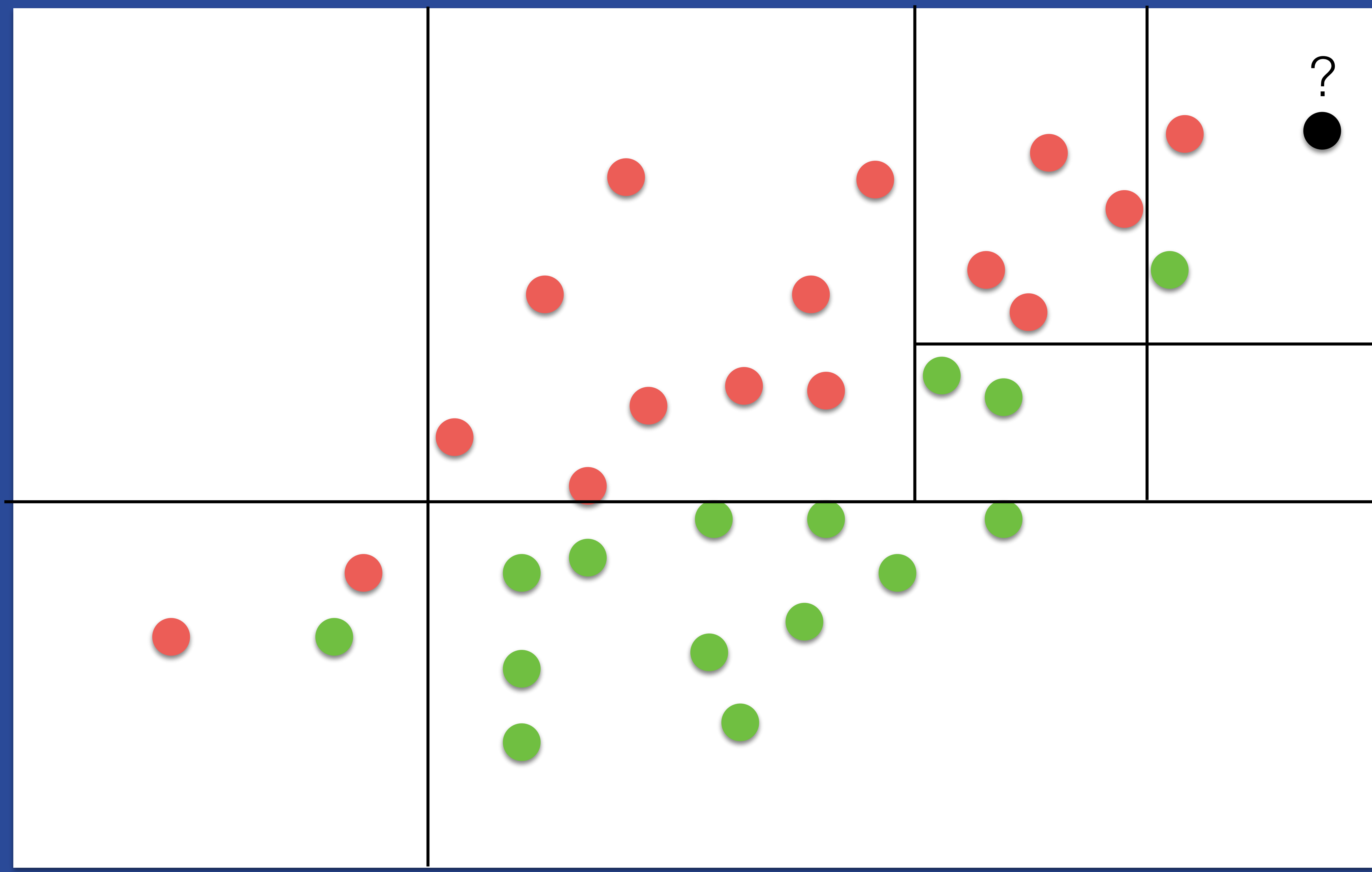
Splits Parallel to Axis



Why Not Decision Trees

- Slightly prone to over-fitting
    - But we'll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
    - More data!
- Predictions outside training data can be problematic

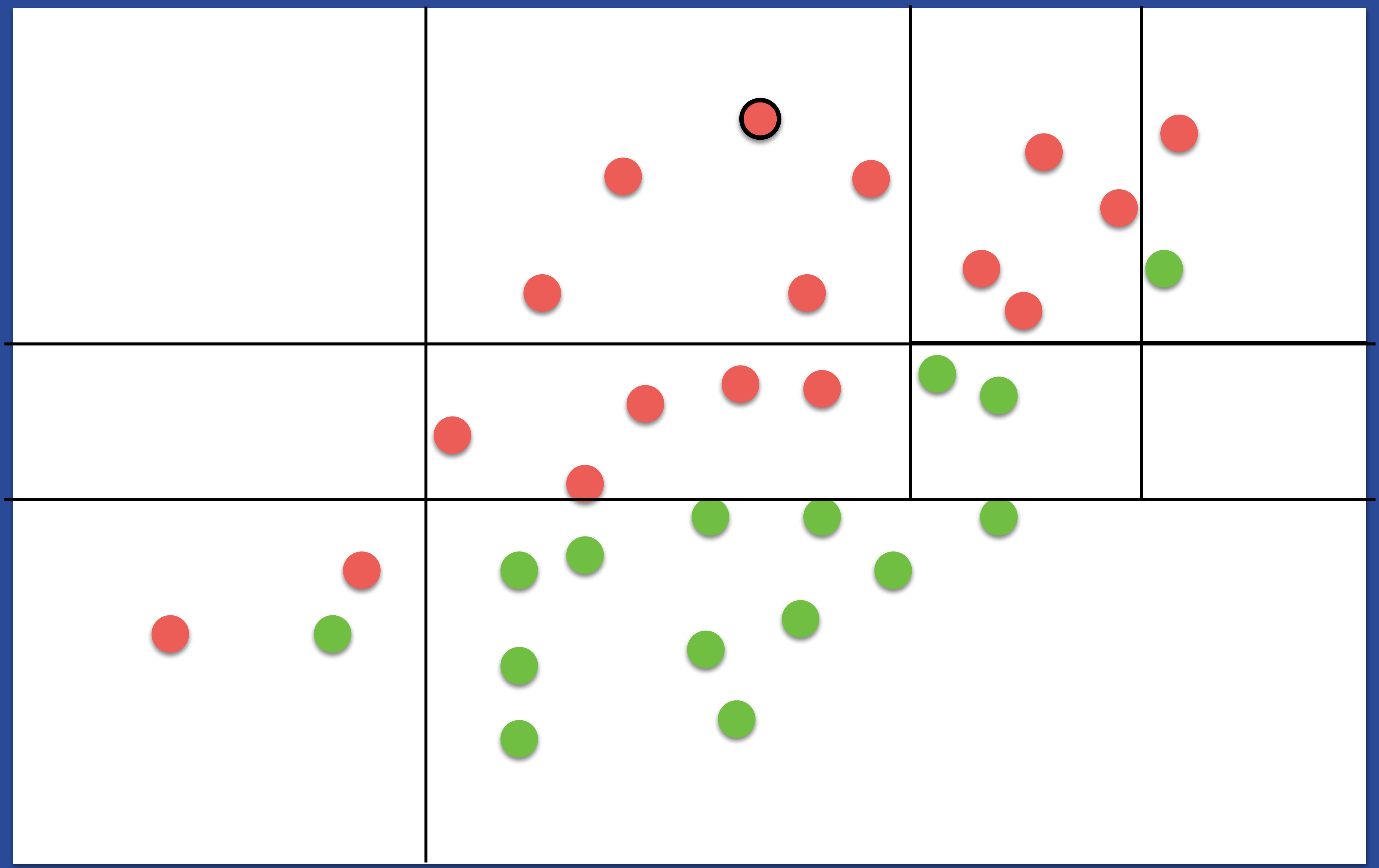
Outlier Predictions



Why Not Decision Trees

- Slightly prone to over-fitting
    - But we'll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
    - More data!
- Predictions outside training data can be problematic
    - We can catch this with model competence
- Can be sensitive to small changes in training data

Outlier Predictions



Why Not Decision Trees

- Slightly prone to over-fitting
    - But we'll fix this with ensembles
- Splitting prefers decision boundaries that are parallel to feature axes
    - More data!
- Predictions outside training data can be problematic
    - We can catch this with model competence
- Can be sensitive to small changes in training data

Questions:

- What other models can we try?
- And how will we know which one works best?

Your turn!

- Build a **Model** for the PDX dataset
    - Use the same dataset from before!
    - This will be a regression
- What is the most important feature?
- What is the predicted price of:
    - SQFT = 1,000, LOT SIZE=0
    - All other fields at default
- Can you disable all fields? What prediction is that?
- How does the prediction change with Missing Strategy: Last versus Proportional

bigml®