PREDICT POKER HANDS

AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop

SIMPLE TRAINING DATA SET

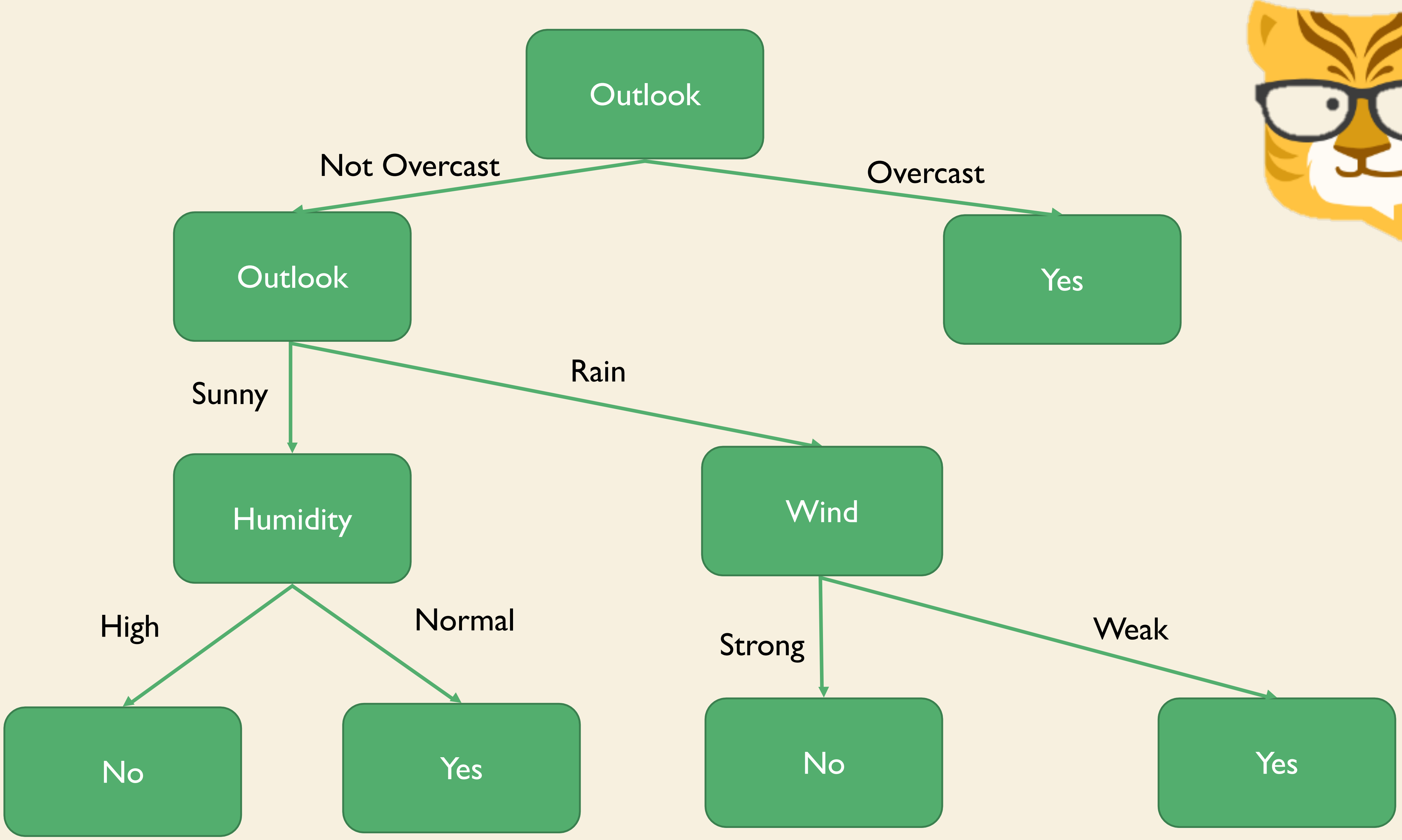
Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Week	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop

COMMON CLASSIFICATION ALGORITHMS

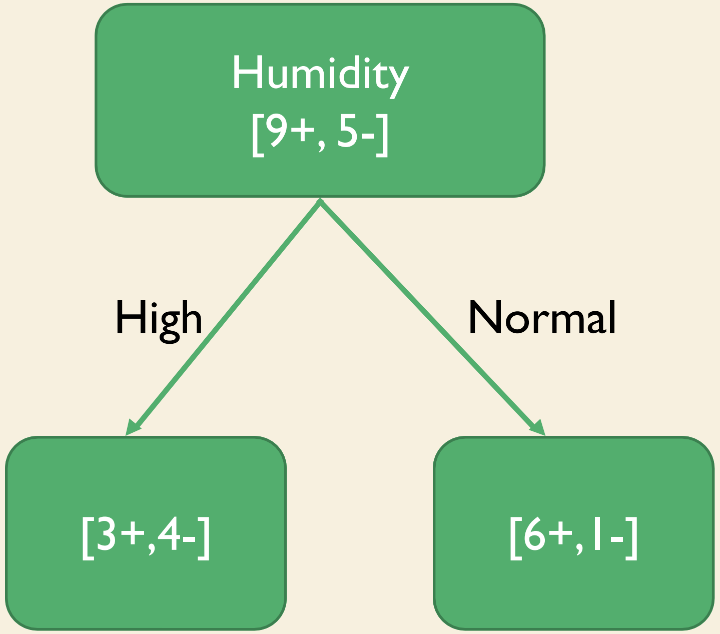
- Logistic regression
- Naive Bayes classifier
- Perceptron
- Support vector machines
- Decision trees
- Random forests
- Neural networks


Outlook
Not Overcast
Overcast
Outlook
Yes
Sunny
Rain
Humidity
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes

PYTHON DATA SCIENCE PACKAGES

- Pandas
  - Easy-to-use data structures
  - Data analysis
- Numpy
  - A powerful N-dimensional array object
  - Useful functions for number processing
- Matplotlib
  - Produces high quality figures in a variety of formats
- Sckit-learn
  - Simple and efficient tools for data mining and data analysis

SEPARATE TWO DATASETS



Humidity
[9+, 5-]

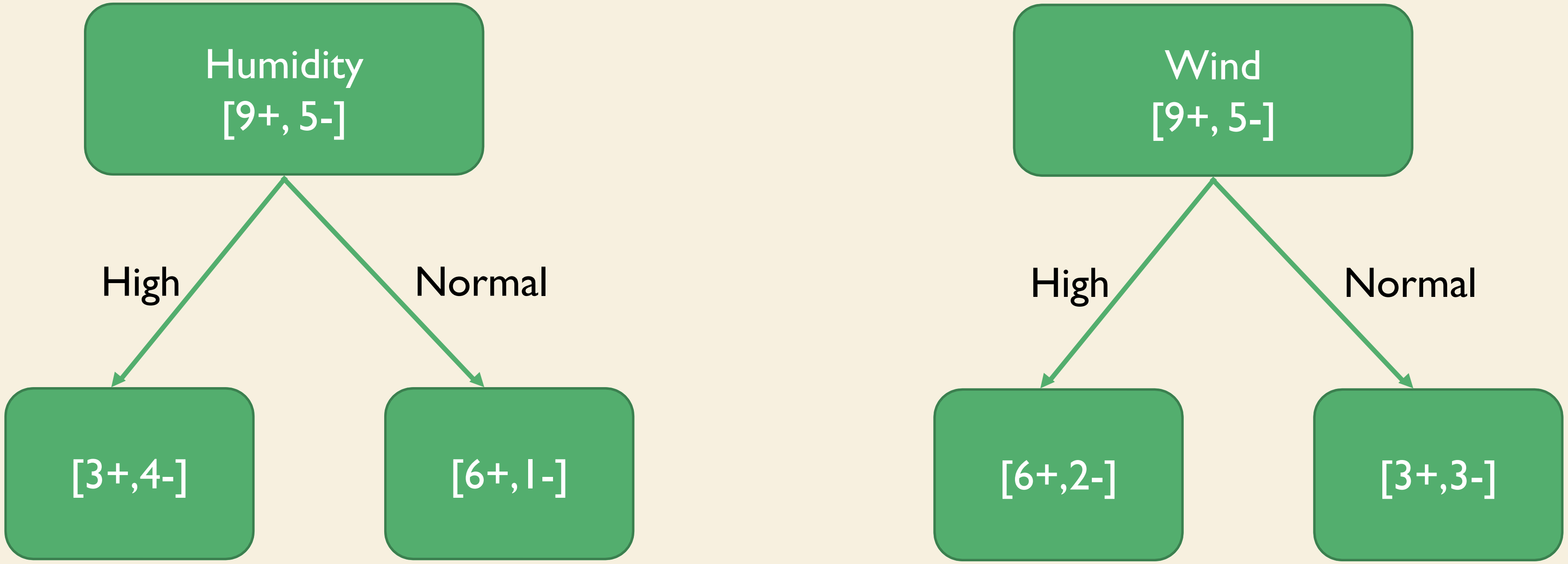
High

Normal

[3+,4-]

[6+,1-]

WHICH ONE IS A BETTER CLASSIFIER?



Humidity
[9+, 5-]

High

Normal

[3+,4-]

[6+,1-]

Wind
[9+, 5-]

High

Normal

[6+,2-]

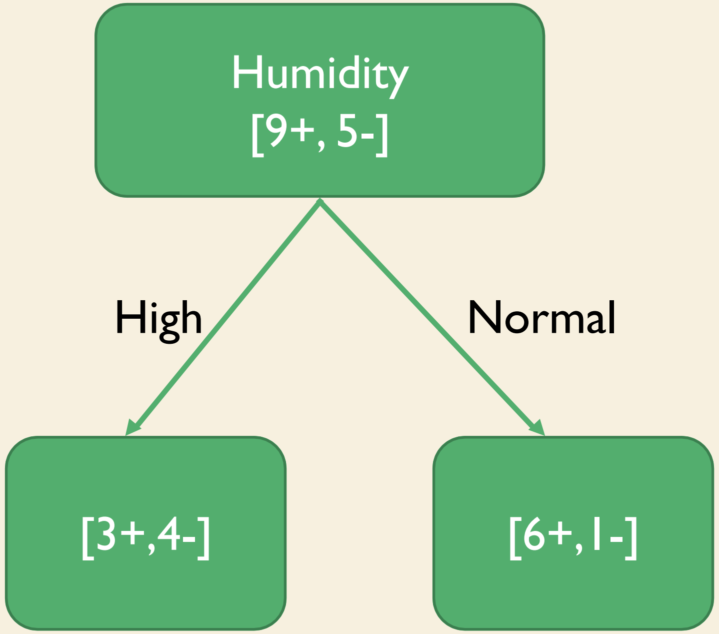
[3+,3-]

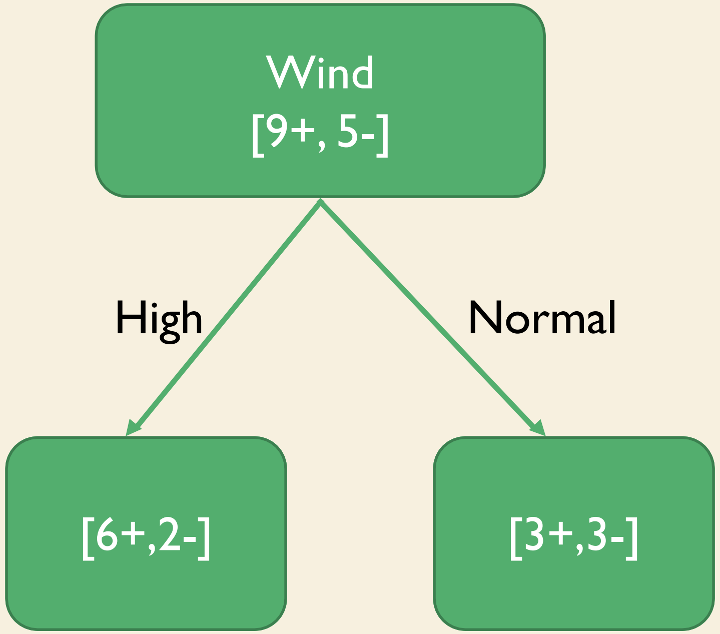
MEASURE IMPURITY

- What kind of property should it have
  - Lower the better
  - If all the same, impurity is 0%, If half and half, impurity is 100%
- Entropy
  - X: sample space contains *n* event
  - P(X=i): probability of X being *i*th event
  - Entropy is the sum of the probability of each label times the log probability of that same label

$$\boldsymbol{H(X) = -\sum_{i=1}^{n} P(X=i)\, log_2 P(X=i)}$$

$$H(X) = -\sum_{i=1}^{n} P(X = i)\, \mathrm{LOG}_2\, P(X = i)$$





Calculate Entropy Yourself

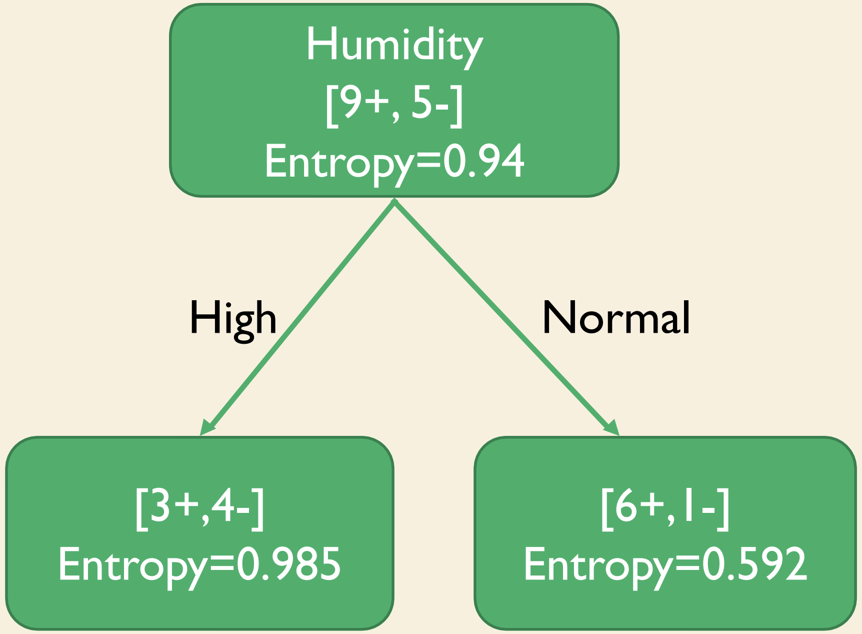
INFORMATION GAIN

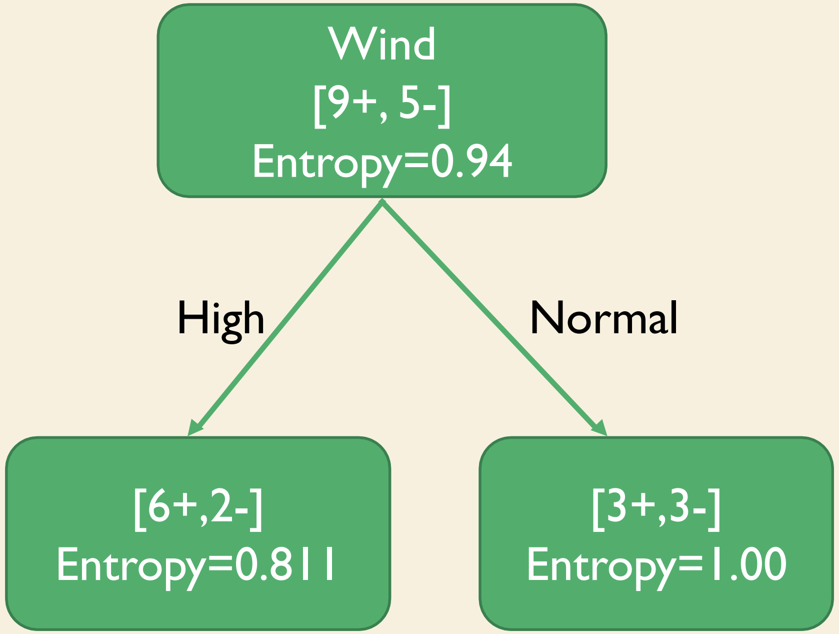
- Input attribute A
- Target variable Y
- Sample S
- Larger the better

Weighted

$$Gain(S,A) = H_S(Y) - H_S(Y|A)$$

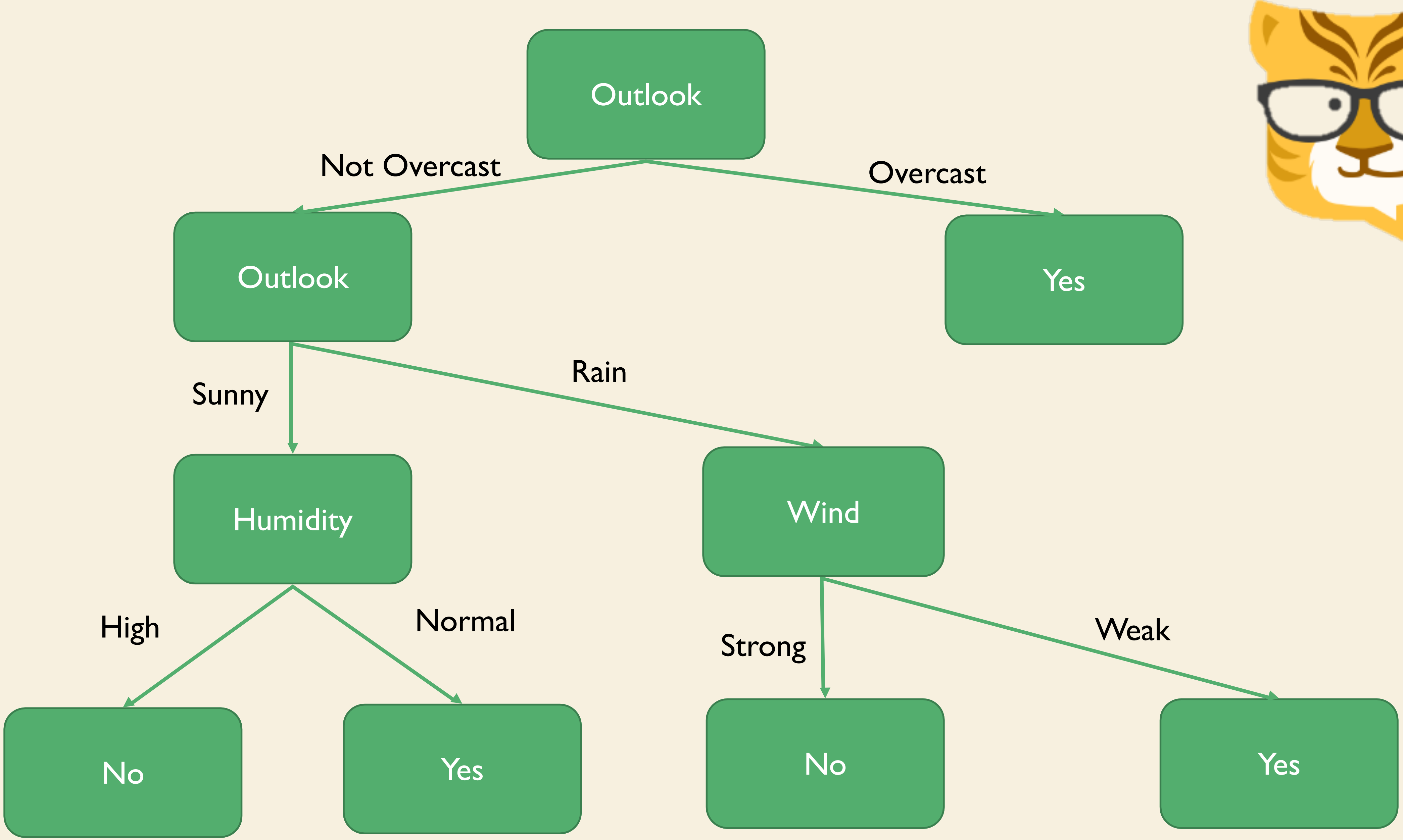
$$Gain(S,A) = H_S(Y) - H_S(Y|A)$$





Gain = 0.94 – ((7/14)*0.985 + (7/14)*0.592)
= 0.151

Gain = 0.94 – ((8/14)*0.811 + (6/14)*1.00)
= 0.048



Outlook

Not Overcast

Overcast

Outlook

Yes

Sunny

Rain

Humidity

Wind

High

Normal

Strong

Weak

No

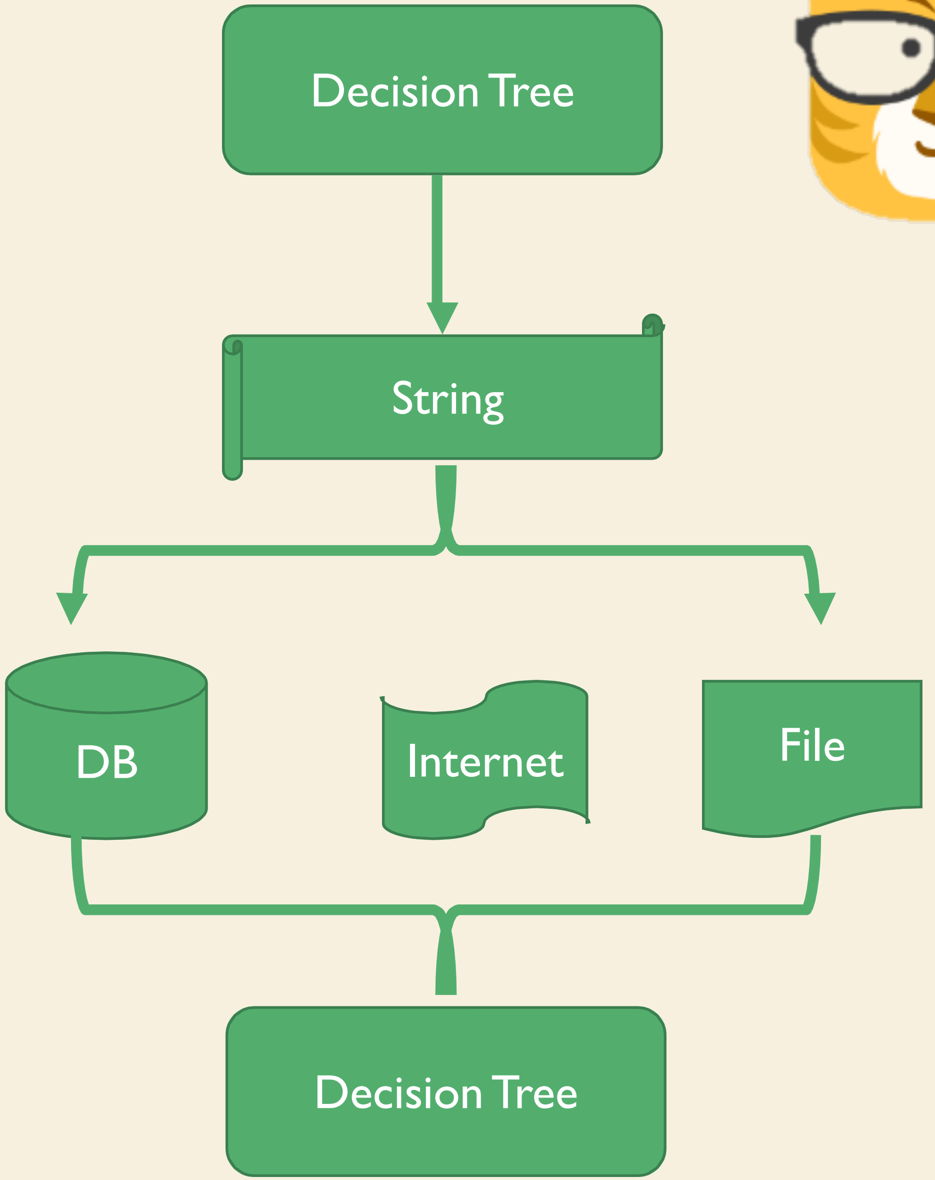
Yes

No

Yes

SERIALIZATION

- Convert an object into string
- Easy to store or transmit
- Save the state of an object in order to be able to recreate it when needed.
- JSON
- Pickle



AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop

IRIS DATA SET

- Generated in 1936
- 150 samples
- 50 for each of the species
- Sepal length
- Sepal width
- Petal length
- Petal width

CONFUSION MATRIX

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Precision	$\frac{tp}{tp + fp}$
Recall	$\frac{tp}{tp + fn}$
F1	$\frac{2pr}{p + r}$

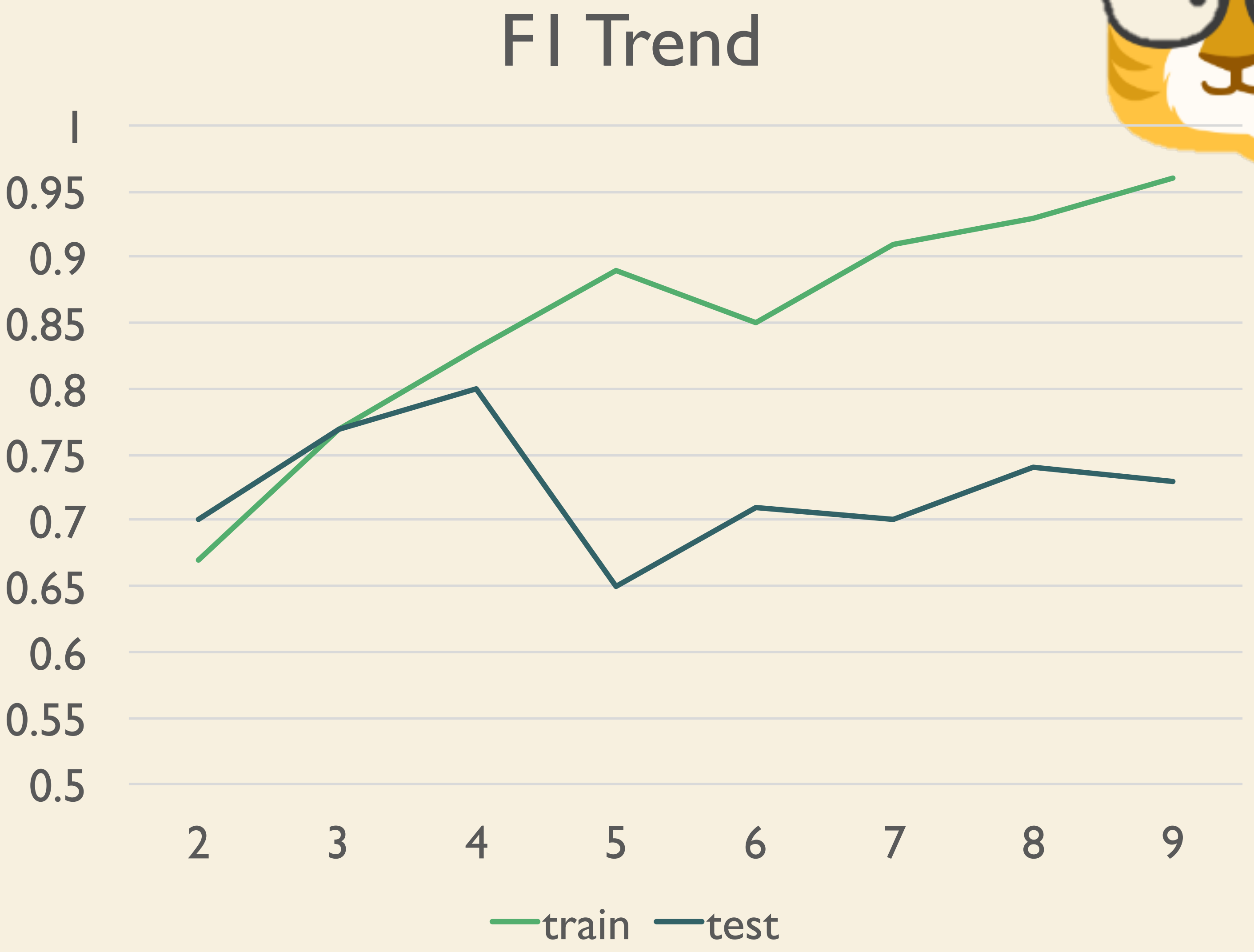
Precision	$\frac{tp}{tp+fp}$
Recall	$\frac{tp}{tp+fn}$
F1	$\frac{2pr}{p+r}$

	Play	Not Play
Play	7	2
Not Play	0	3

	Play	Not Play
Play	5	0
Not Play	2	5

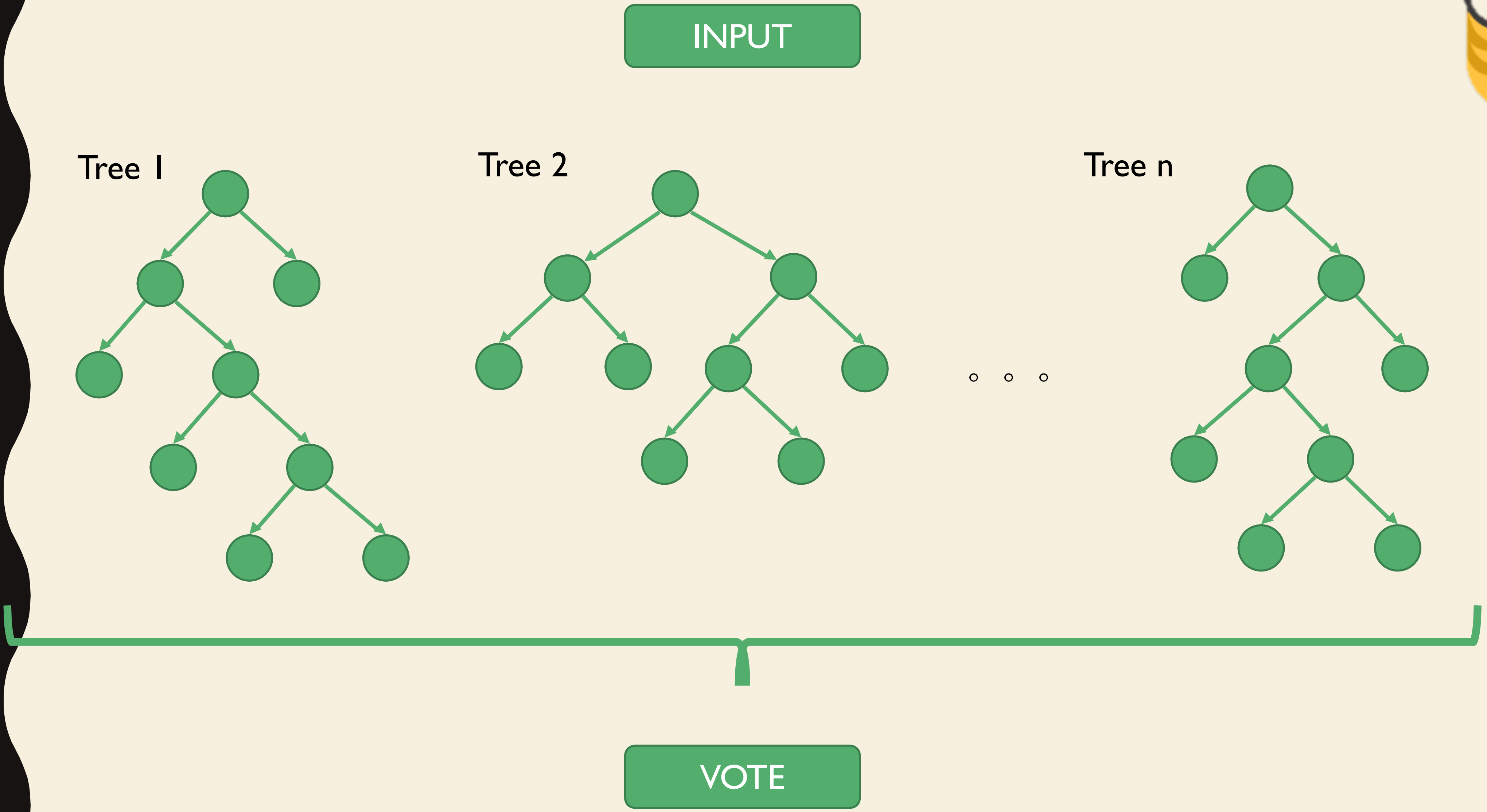
	Play	Not Play
Play	5	2
Not Play	2	3

Train	Test
0.67	0.70
0.77	0.77
0.83	0.80
0.89	0.65
0.85	0.71
0.91	0.70
0.93	0.74
0.96	0.73



AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop



INPUT

Tree 1

Tree 2

∘ ∘ ∘

Tree n

VOTE

RANDOM SELECTION

- Select SQRT(nFeature) of feature
- Select nRecord/3 of records

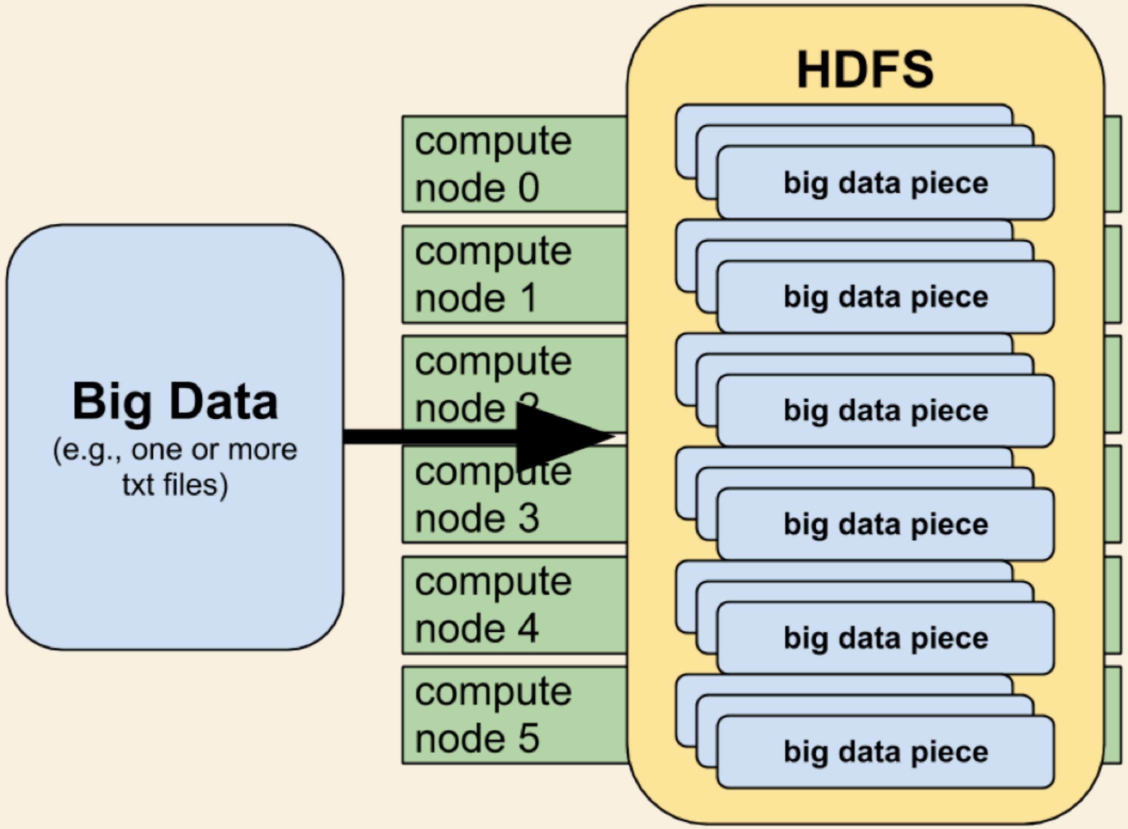
	A	B	C	D	E	F	G	H	I
1	▨	🟩	▨	▨	▨	▨	🟩	▨	🟩
2		▨					▨		▨
3	▨	🟩	▨	▨	▨	▨	🟩	▨	🟩
4		▨					▨		▨
5		▨					▨		▨
6	▨	🟩	▨	▨	▨	▨	🟩	▨	🟩

AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop

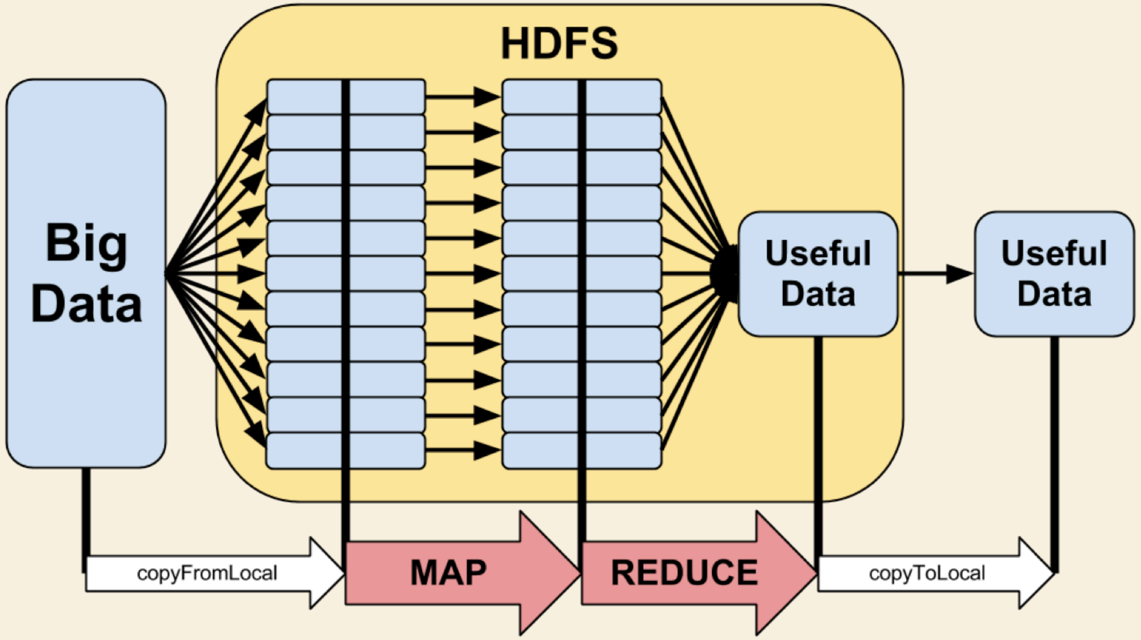
EXTRA: HADOOP

- HDFS
- MapReduce

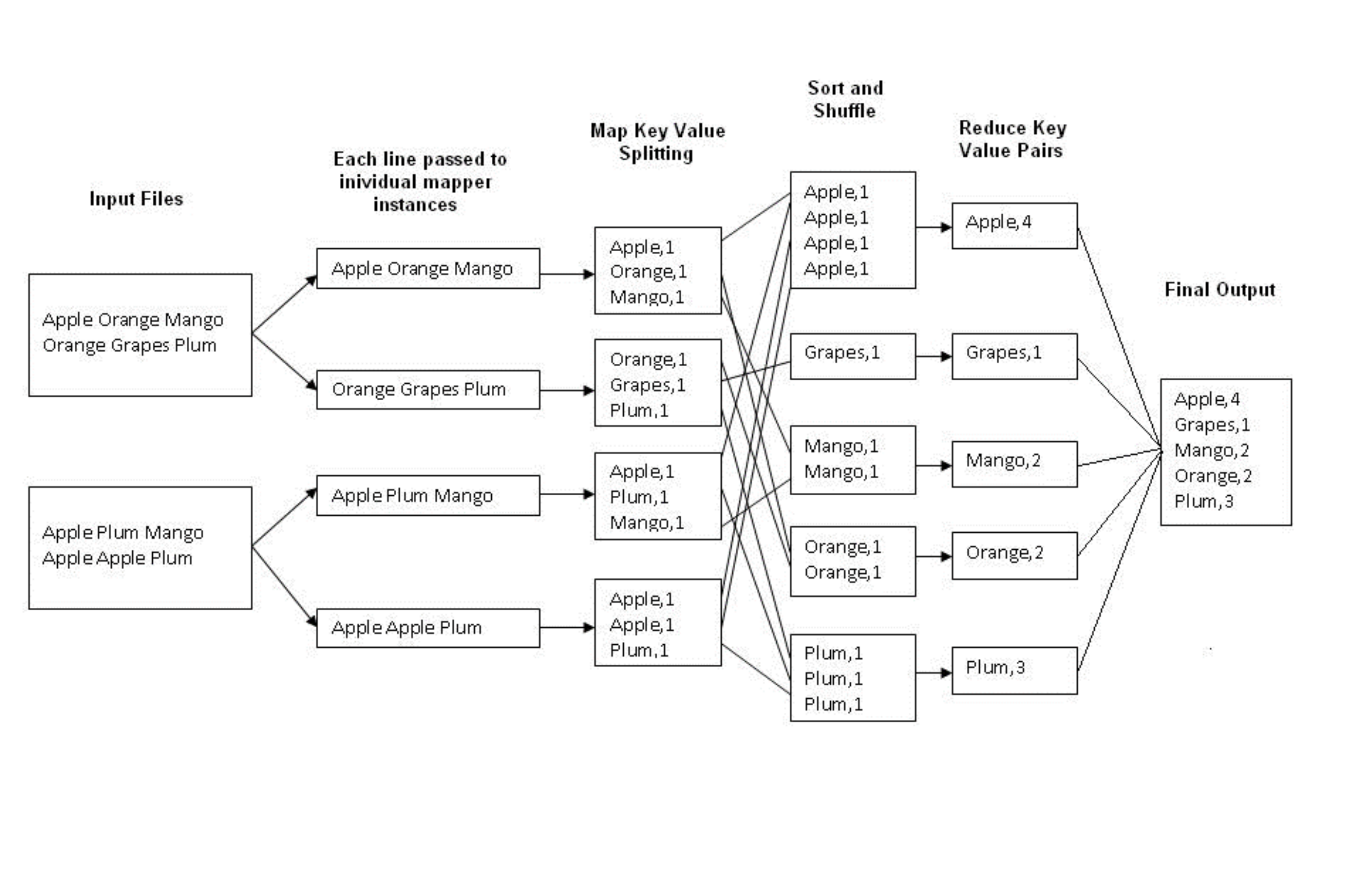


EXTRA: HADOOP

- HDFS
- MapReduce
- WordCount

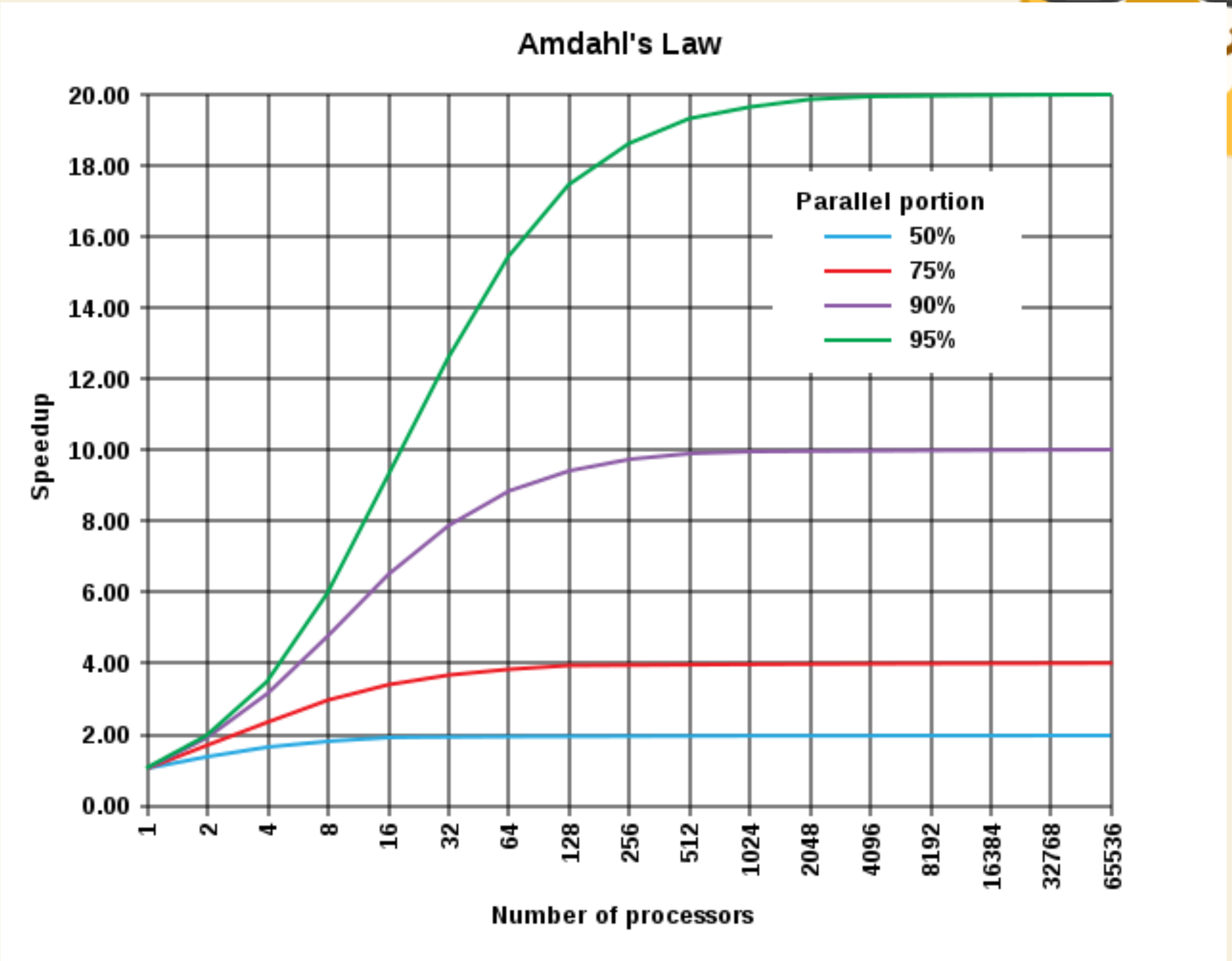


WORD COUNT

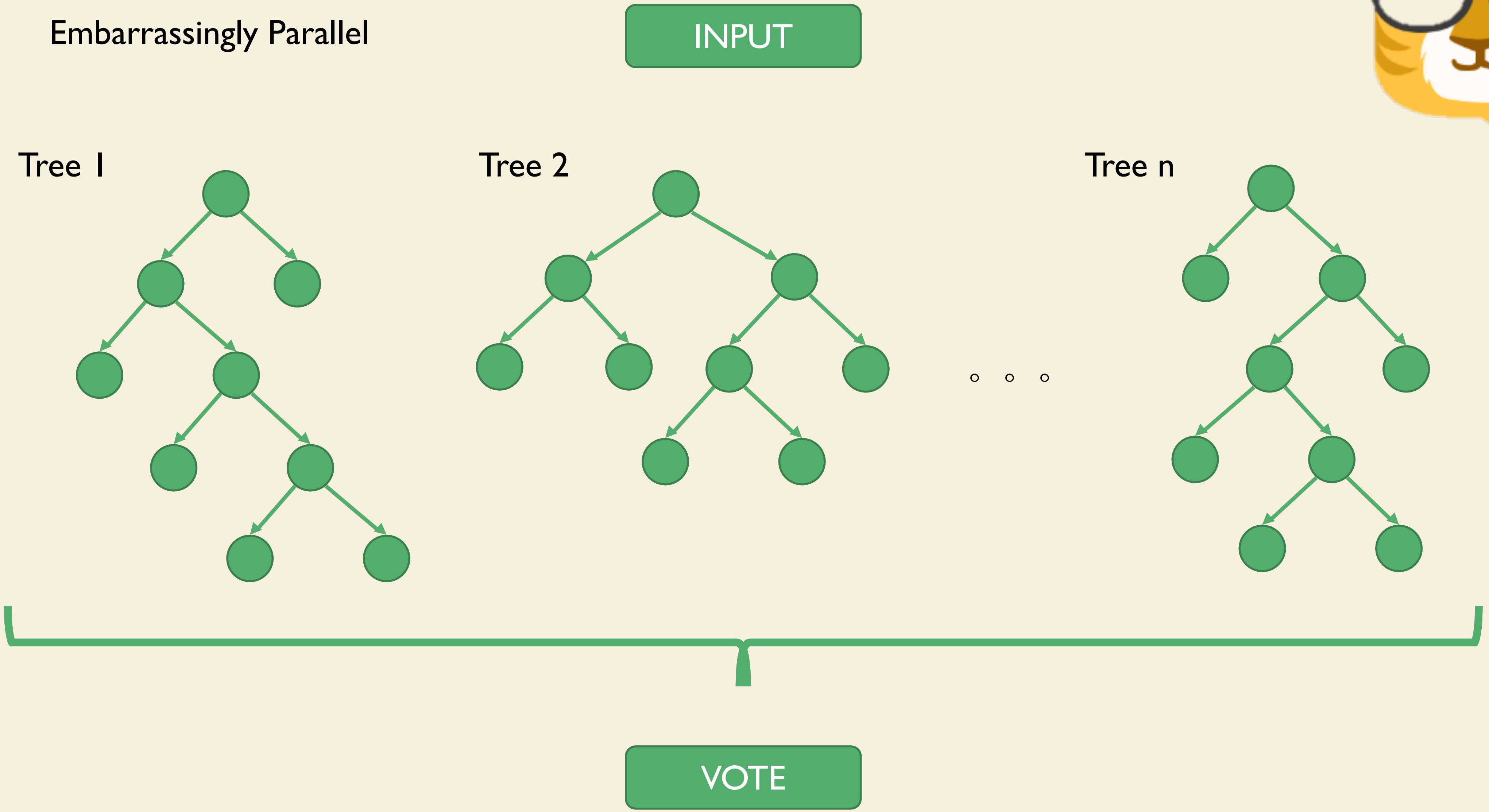

Input Files
Apple Orange Mango
Orange Grapes Plum
Apple Plum Mango
Apple Apple Plum
Each line passed to
inividual mapper
instances
Apple Orange Mango
Orange Grapes Plum
Apple Plum Mango
Apple Apple Plum
Map Key Value
Splitting
Apple,1
Orange,1
Mango,1
Orange,1
Grapes,1
Plum,1
Apple,1
Plum,1
Mango,1
Apple,1
Apple,1
Plum,1
Sort and
Shuffle
Apple,1
Apple,1
Apple,1
Apple,1
Grapes,1
Mango,1
Mango,1
Orange,1
Orange,1
Plum,1
Plum,1
Plum,1
Reduce Key
Value Pairs
Apple,4
Grapes,1
Mango,2
Orange,2
Plum,3
Final Output
Apple,4
Grapes,1
Mango,2
Orange,2
Plum,3

PARALLELIZATION

- Fine-grained parallelism
- Coarse-grained parallelism
- Embarrassing parallelism
- Amdahl's law
- $S_{latency}(s) = \frac{1}{1-p+\frac{p}{s}}$
- S is the speedup of the parallelable part
- p is the percentage of the parallelable part



Embarrassingly Parallel



AGENDA

- Day 1
  - Play Tennis Data Set
  - Decision Tree
  - Iris Data Set
  - Random Forest
- Day 2
  - Poker Hands Data Set
  - Data Exploration
- Extra: Hadoop