Decision Trees & Random Forests

BUGS Meeting

Daniel Pimentel-Alarcón
Computer Science, GSU

Decision Trees

Goal: Predict

- Will I get El Cáncer?
- Will I develop Diabetes?
- Is my boyfriend/girlfriend cheating on me?
- Will my Bacteria develop Antibiotic Resistance?

	A	B	C	D	E	F	G	H	
1	0.166390734	0.0739022542	0.9380422919	0.4197909476	0.199372394	0.8773471711	0.083215471	0.9400430233	0.7
2	0.8615271223	0.7757295209	0.56219381	0.4953289942	0.7901157516	0.3168325007	0.0901134394	0.5927609941	0.4
3	0.9898183863	0.1608356235	0.8343072962	0.6000025857	0.2041446567	0.8752995632	0.9870366976	0.9585429376	0.1
4	0.4160293543	0.6541336075	0.533131226	0.2588331876	0.5838729884	0.599851159	0.5517696308	0.6071065792	0.5
5	0.015721427	0.3214682985	0.7856533648	0.4555887529	0.724996398	0.3202909518	0.0764947159	0.1158882901	0.3
6	0.6832840093	0.1835837403	0.4081572187	0.2237776744	0.395887609	0.7385271892	0.0843965665	0.3581091606	0.7
7	0.2270948424	0.0753671025	0.0494126878	0.1149176771	0.5402233903	0.9147158584	0.8683215391	0.4956815843	0.7
8	0.559706779	0.3547628012	0.6737842499	0.4400304139	0.3052613081	0.9529025748	0.908458523	0.7750171428	0.9
9	0.0639803573	0.8327802755	0.754443808	0.6734893338	0.8825760537	0.5428943436	0.0870754241	0.725444772	0.9
10	0.8509917245	0.1453037842	0.969216953	0.9897150856	0.9354477939	0.9258715529	0.1733232688	0.6480473382	0.1
11	0.2713187919	0.2857065005	0.903035834	0.9298100968	0.1260013632	0.0767446451	0.4399159383	0.2829076324	0.2
12	0.3704984726	0.4851600723	0.443642485	0.9373102742	0.7847810986	0.3864274416	0.7541396178	0.3867283773	0.0
13	0.9759758923	0.7371094345	0.0455157282	0.8346701767	0.4383846298	0.6298020934	0.0993899375	0.5817343793	0.7
14	0.6351593339	0.3977721159	0.5993215961	0.2587002467	0.1375043502	0.2660007051	0.1686486479	0.8403079675	0.0
15	0.0716924155	0.6319312733	0.3823427393	0.6420220807	0.2364369316	0.3835056194	0.5611884128	0.3820433638	0.7
16	0.8422852685	0.542874712	0.5883257678	0.760320741	0.8315882224	0.6276306058	0.5803535685	0.8411028145	0.3
17	0.9408779606	0.5220144195	0.4812654063	0.5331868522	0.1047852791	0.7975936416	0.8026251936	0.9925389684	0.9
18	0.1095553513	0.2625067541	0.9741186467	0.9848365255	0.2483271887	0.0985997624	0.2721076927	0.1593645221	0.0
19	0.4997927954	0.096071915	0.2369617911	0.8118898307	0.6460952105	0.8707811364	0.4498663761	0.254943708	0.4
20	0.6429012946	0.0800355875	0.5068224588	0.9520570131	0.1408375208	0.1400302523	0.9902096502	0.0871451208	0.8
21	0.7057224081	0.4630195114	0.9144776522	0.3820045677	0.6039446075	0.8557347907	0.9241318835	0.4712547997	0.9
22	0.2484984729	0.2377738535	0.3104894005	0.3790386592	0.5252200689	0.1912480153	0.8910088711	0.0505195535	0.
23	0.5463155243	0.5161090947	0.086890304	0.5067192O3	0.0861523314	0.0243307806	0.9956105	0.685405273	0.3
24	0.1472119791	0.4978118786	0.3711289708	0.8731593913	0.1267880907	0.1878780541	0.9781146096	0.5520577445	0.8
25	0.790510121	0.4327639828	0.7940384434	0.0472936828	0.4442862018	0.2649206629	0.6867534886	0.1463450387	0.1
26	0.238609707	0.3432042131	0.0508109627	0.053046589	0.8862833569	0.4834029363	0.7091768233	0.9173125281	0.4
27	0.2727451844	0.7989924805	0.8074715317	0.8630560155	0.520064408	0.9395456181	0.7433063779	0.7300025792	0.4
28	0.0233547923	0.8923628118	0.8213618146	0.0381375295	0.7760785248	0.9095269332	0.4853258794	0.3151255539	0.
29	0.431277638	0.8813339959	0.6464634317	0.1806829795	0.0149378479	0.1696698208	0.0184137947	0.4957891628	0.8
30	0.9908212749	0.394212218	0.7033778366	0.671705693	0.3933258315	0.6299235066	0.3324489791	0.7547388354	0.2
31	0.3143209	0.6560834548	0.4499266401	0.4371305082	0.0300377	0.4826313867	0.6088405084	0.3349117569	0.7
32	0.8232727759	0.9862914153	0.2903183235	0.4740527456	0.5629339886	0.5561574434	0.9774761845	0.778319096	0.

Here is my Data

How do I know? Entropy

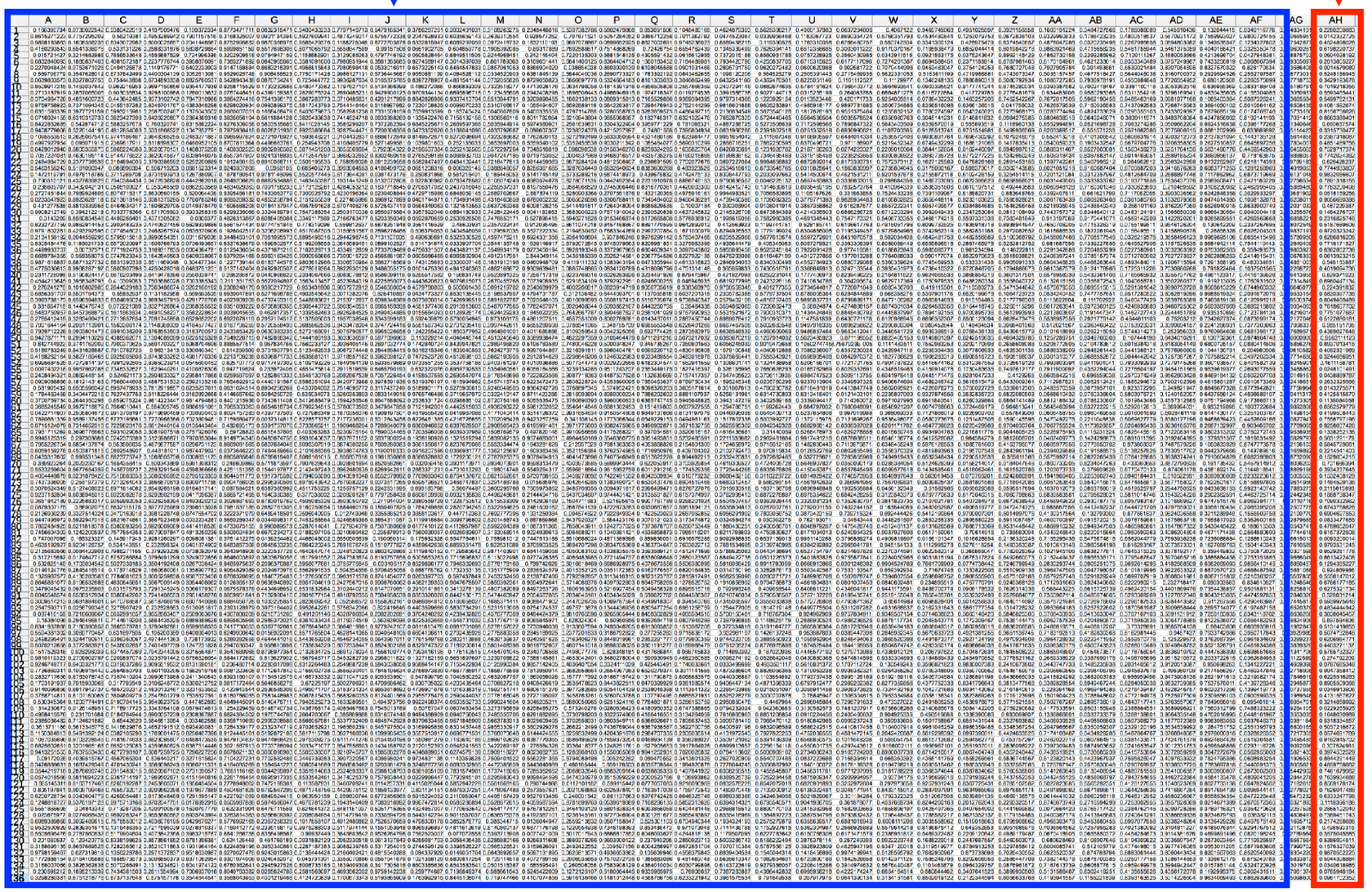
Other Variables

(age, gender, etc.)

Variable of Interest

(cancer, diabetes, cheater)

Samples



Here is my Data

How do I know? Entropy

THE FOLLOWING **PREVIEW** HAS BEEN APPROVED FOR

ALL AUDIENCES

BY THE MOTION PICTURE ASSOCIATION OF AMERICA INC.

THE FILM ADVERTISED HAS BEEN RATED

R	RESTRICTED
	UNDER 17 REQUIRES ACCOMPANYING PARENT OR GUARDIAN
PARTIAL NUDITY & INFO THEORY	

www.filmratings.com

www.mpaa.org

Horse	1	2	3	4	5	6	7	8	Length
P(winning)	1/2	1/4	1/8	1/16	1/32	1/32	1/32	1/32	
Message	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1	**3 bits**

Message

Info Theory & Entropy

A Quick Detour

Horse	1	2	3	4	5	6	7	8	E[Length]
P(winning)	1/2	1/4	1/8	1/16	1/32	1/32	1/32	1/32	
Message	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1	**3 bits**
Optimal	0	10	110	1110	1111 00	1111 01	1111 10	1111 11	**2 bits**

Message

Info Theory & Entropy

A Quick Detour

This is what you need to remember:

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

How frequent is this outcome?

How many bits should I spend encoding this outcome?

(more likely outcomes get fewer bits)

Amount of **information** encoded in variable x

Average over all outcomes?

Info Theory & Entropy

A Quick Detour

	Gender	PhD?
1	0	0
2	1	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	1	0
9	1	0
10	1	1
11	0	0
12	0	0
13	0	0
14	1	0
15	0	0
16	1	0

$$H(x) = \sum_{x} p(x) \log \frac{1}{p(x)}$$

Info Theory & Entropy

Example

	Gender	PhD?
1	0	0
2	1	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	1	0
9	1	0
10	1	1
11	0	0
12	0	0
13	0	0
14	1	0
15	0	0
16	1	0
p(0)	**1/2**	**15/16**
p(1)	**1/2**	**1/16**

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)}$$

$$= \frac{1}{2} \log(2) + \frac{1}{2} \log(2)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$= 1$$

Info Theory & Entropy

Example

	Gender	PhD?
1	0	0
2	1	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	1	0
9	1	0
10	1	1
11	0	0
12	0	0
13	0	0
14	1	0
15	0	0
16	1	0
p(0)	**1/2**	**15/16**
p(1)	**1/2**	**1/16**
H(x)	**1**	

$$
\begin{aligned}
H(x) &= \sum_x p(x) \log \frac{1}{p(x)} \\
&= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)} \\
&= \frac{1}{2} \log(2) + \frac{1}{2} \log(2) \\
&= \frac{1}{2} + \frac{1}{2} \\
&= 1
\end{aligned}
$$

Info Theory & Entropy

Example

	Gender	PhD?
1	0	0
2	1	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	1	0
9	1	0
10	1	1
11	0	0
12	0	0
13	0	0
14	1	0
15	0	0
16	1	0
p(0)	**1/2**	**15/16**
p(1)	**1/2**	**1/16**
H(x)	**1**	

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)}$$

$$= \frac{15}{16} \log(\frac{16}{15}) + \frac{1}{16} \log(16)$$

$$= \frac{15}{16}(0.093) + \frac{1}{16}(4)$$

$$= 0.337$$

Info Theory & Entropy

Example

Most informative!

	Gender	PhD?
1	0	0
2	1	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	1	0
9	1	0
10	1	1
11	0	0
12	0	0
13	0	0
14	1	0
15	0	0
16	1	0
p(0)	**1/2**	**15/16**
p(1)	**1/2**	**1/16**
H(x)	**1**	**0.337**

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)}$$

$$= \frac{15}{16} \log(\frac{16}{15}) + \frac{1}{16} \log(16)$$

$$= \frac{15}{16}(0.093) + \frac{1}{16}(4)$$

$$= 0.337$$

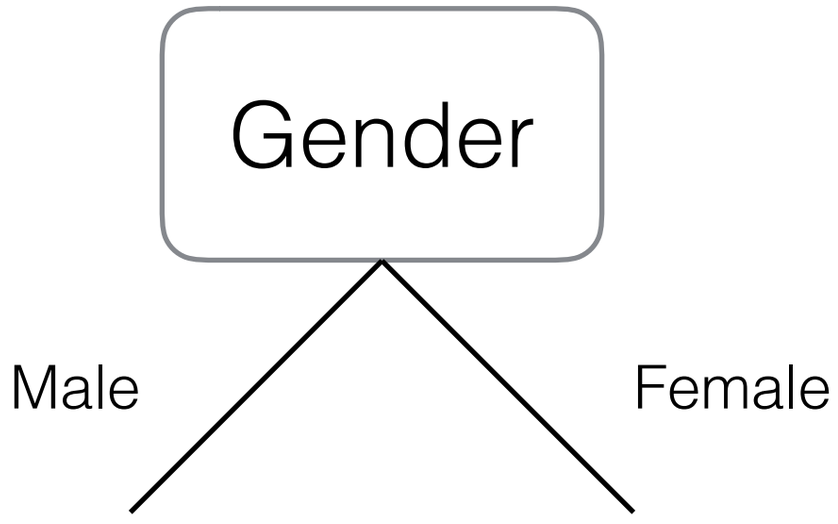
Info Theory & Entropy

Example

Gender

Here is my Data

1) Find Most Informative Variable



Most Informative Variable:

First Decision in my Tree

Then What?

Gender

Females

Males

2) Split According to Most Informative Variable

3) Find Most Informative Variable in Each Subset

Gender

Age

Income

Females

Males

Repeat

2) Split According to Most Informative Variable

3) Find Most Informative Variable in Each Subset

Gender

Age

Females

Males

Filthy-rich

Old

Young

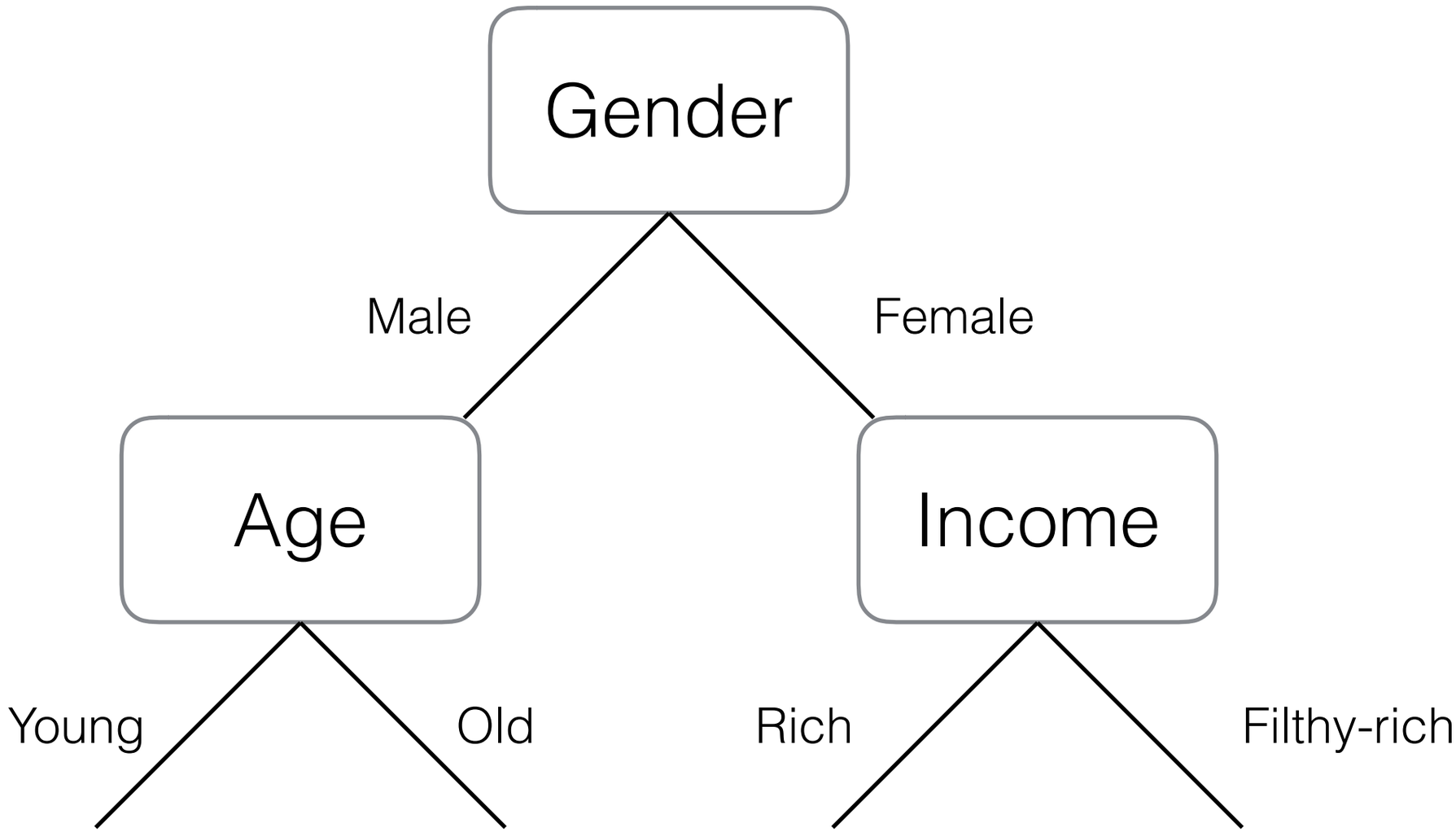
Rich

Income

Repeat

2) Split According to Most Informative Variable

3) Find Most Informative Variable in Each Subset



Each Informative Variable:

One Decision in my Tree

Then What?

Gender Age

Males Old

Females Filthy-rich

Young

Rich

Income

2) Split According to Most Informative Variable

Repeat

3) Find Most Informative Variable in Each Subset

When do we stop?

	Gender	PhD?	Will Die?
1	0	0	1
2	1	0	1
3	1	0	1
4	0	0	1
5	0	0	1
6	0	0	1
7	1	0	1
8	1	0	1
9	1	0	1
10	1	1	1
11	0	0	1
12	0	0	1
13	0	0	1
14	1	0	1
15	0	0	1
16	1	0	1
p(0)	**1/2**	**15/16**	**0**
p(1)	**1/2**	**1/16**	**1**
H(x)	**1**	**0.337**	

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)}$$

$$= 0 \log(\frac{1}{0}) + 1 \log(1)$$

$$= 0$$

Info Theory & Entropy

Extreme Case

	Gender	PhD?	Will Die?
1	0	0	1
2	1	0	1
3	1	0	1
4	0	0	1
5	0	0	1
6	0	0	1
7	1	0	1
8	1	0	1
9	1	0	1
10	1	1	1
11	0	0	1
12	0	0	1
13	0	0	1
14	1	0	1
15	0	0	1
16	1	0	1
p(0)	**1/2**	**15/16**	**0**
p(1)	**1/2**	**1/16**	**1**
H(x)	**1**	**0.337**	**0**

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)}$$

$$= 0 \log(\frac{1}{0}) + 1 \log(1)$$

$$= 0$$

This Variable Provides No Information!

Info Theory & Entropy

Extreme Case

Gender Age

Females Males

Filthy-rich Old

Rich Young

Income

1

0

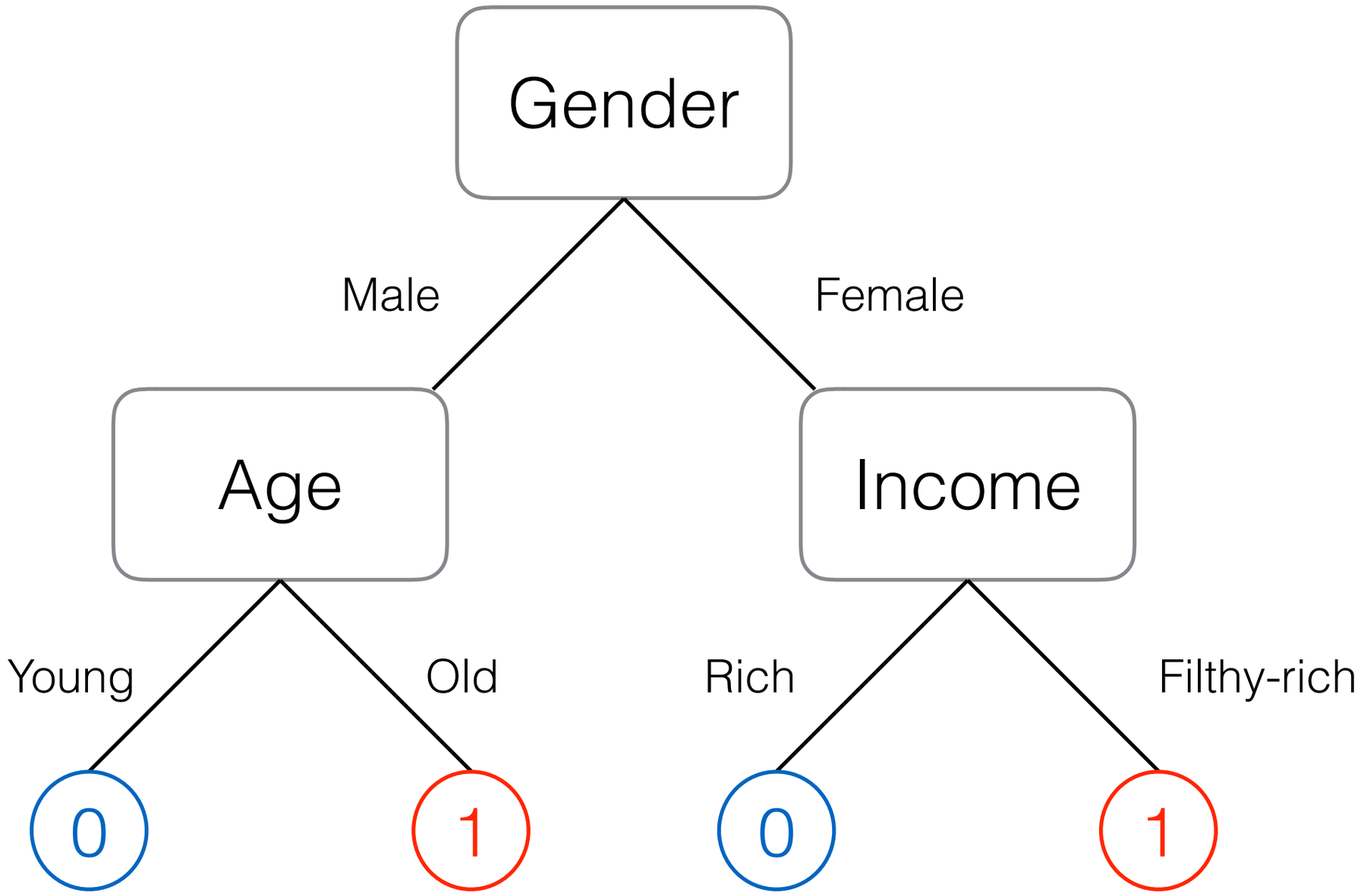
1

0

When do we stop?

When **Variable of Interest** is **Uninformative** in Subset

(~zero entropy)



Finally…

Put Result on my Decision Tree

And enjoy! (start predicting)

(Didn’t I promise partial nudity?)

What could possibly go wrong?

Overfitting & Bias

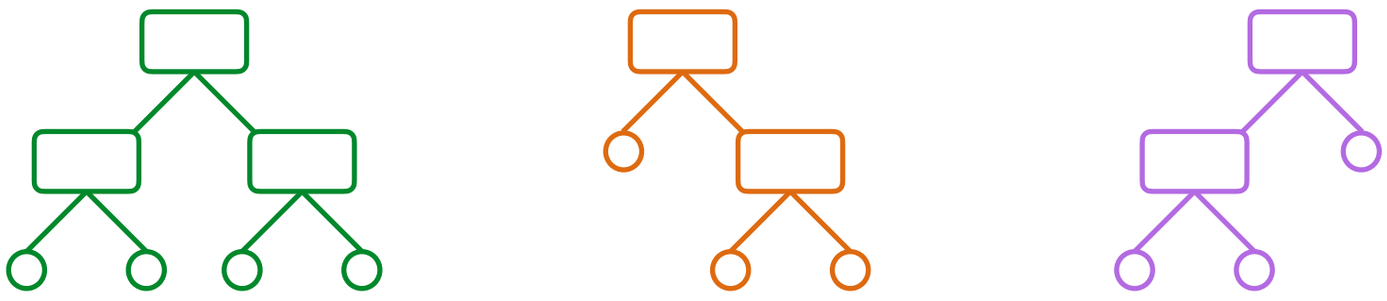
- **Overfitting.** My tree is accurate, but only for my given data (for which I already know the answer)
    - Not a lot of “predictive power”.
- **Bias.** It may heavily depend on my particular sample.
    - If I add/remove a few people, the result may be very different!

What could possibly go wrong?

Overfitting & Bias

No worries:

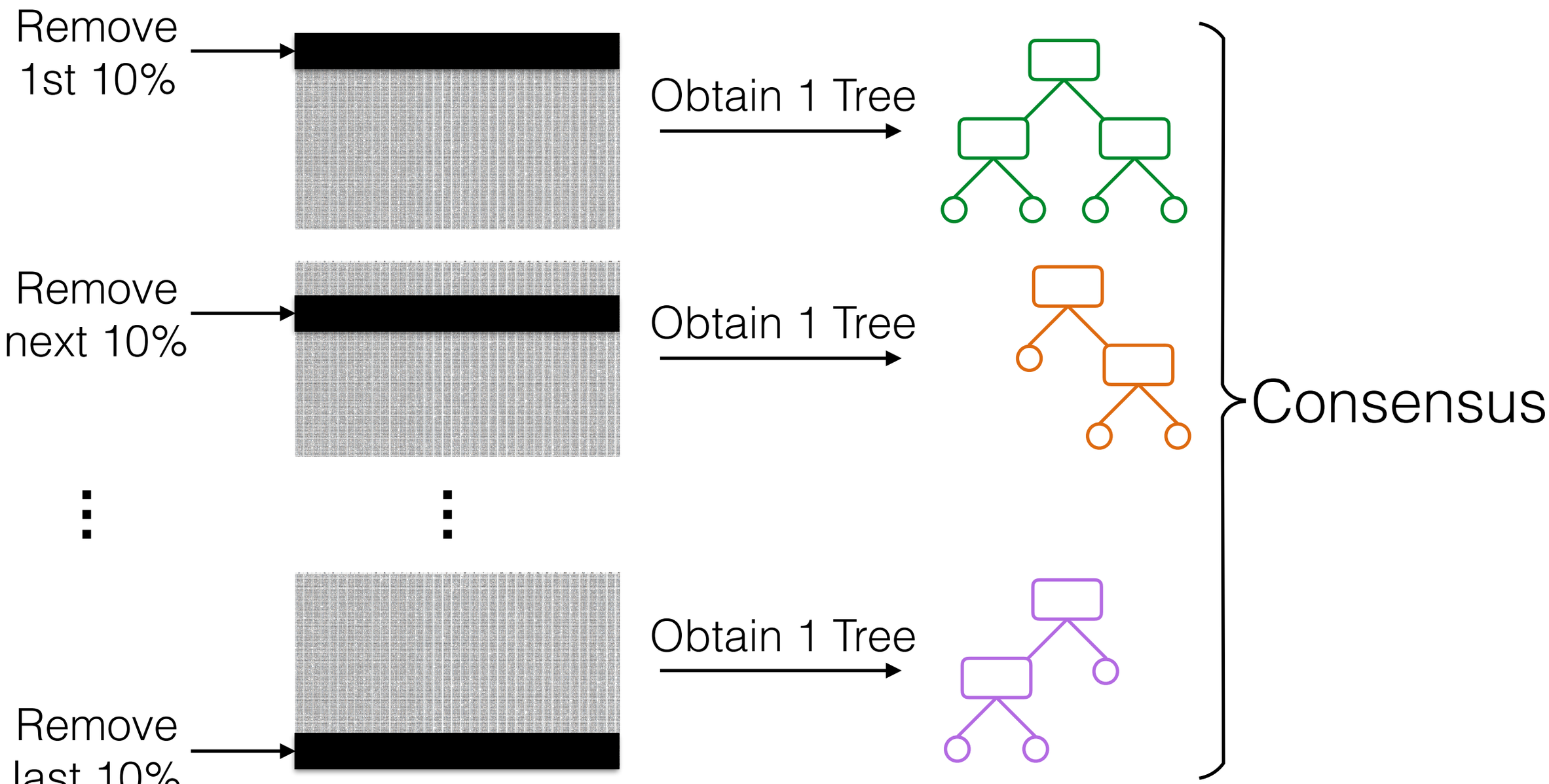
Random Forests



Random Forests

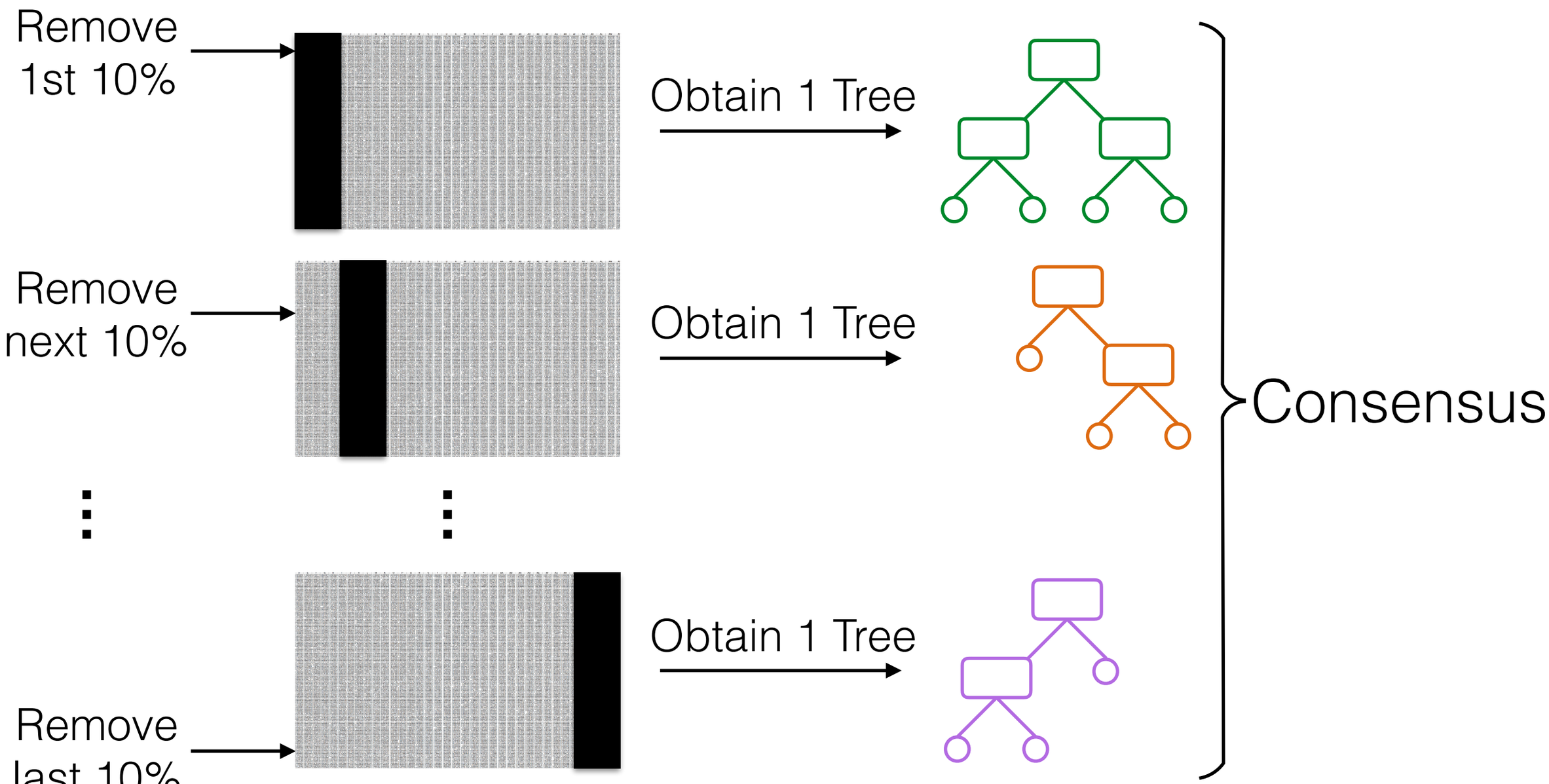
Main Idea: Do many decision trees,
each with a random subsample

aka
Bootstrap
Bagging



Random Forests

Main Idea: Do many decision trees,
each with a random subsample



Random Forests

Main Idea: Do many decision trees,
each with a random subsample

- Good for **Prediction**, but bad for **Description**.
- Fast to train, but **slow** to predict.
- Poor performance on **unbalanced** data.

Advantages/Disadvantages

Alternatives

- Neural Networks
- Regression (Linear, Logistic, Polynomial)
- Other clustering methods (e.g., subspaces)

Questions?