

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 | 0.0832315471 | 0.9400430233 | 0.7 |
| 2 | 0.8615271223 | 0.7757295209 | 0.56219381 | 0.4953289942 | 0.7901157516 | 0.3168325007 | 0.0907134394 | 0.5927609941 | 0.4 |
| 3 | 0.9898183863 | 0.1608358235 | 0.8343072962 | 0.8000025857 | 0.2041448567 | 0.8752995632 | 0.9870388976 | 0.9585429376 | 0.1 |
| 4 | 0.4160293543 | 0.6541336076 | 0.533131226 | 0.2588331876 | 0.5838729884 | 0.5999851159 | 0.5517696308 | 0.6071065792 | 0.5 |
| 5 | 0.015721427 | 0.3214682986 | 0.7856533648 | 0.455887529 | 0.724996398 | 0.3202909518 | 0.0764947159 | 0.1158882901 | 0.3 |
| 6 | 0.6832840093 | 0.1885837403 | 0.4081572187 | 0.2237776744 | 0.3958876091 | 0.7385271892 | 0.0843965665 | 0.3581091606 | 0.7 |
| 7 | 0.2270948424 | 0.0753671026 | 0.0494126878 | 0.1149176771 | 0.5402233903 | 0.9147168584 | 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 | 0.9354177939 | 0.9258715529 | 0.1733232688 | 0.6480473382 | 0.1 |
| 11 | 0.2713187919 | 0.2857085005 | 0.903035834 | 0.9298100988 | 0.1260013632 | 0.0787446451 | 0.4399159383 | 0.2829078324 | 0.2 |
| 12 | 0.3704964726 | 0.4851600723 | 0.443642485 | 0.9373102742 | 0.7847810986 | 0.3864274416 | 0.7541396178 | 0.3867283773 | 0.0 |
| 13 | 0.9759758923 | 0.7371094346 | 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 | 0.266007051 | 0.1666468479 | 0.6403079875 | 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 | 0.8315862224 | 0.6276306058 | 0.5803535685 | 0.8411028145 | 0.3 |
| 17 | 0.9408779606 | 0.5220144195 | 0.4812654063 | 0.5331868522 | 0.1047852791 | 0.7975938418 | 0.8026251938 | 0.9925388684 | 0.9 |
| 18 | 0.1095553513 | 0.2625067541 | 0.9741186467 | 0.9848365255 | 0.2483271887 | 0.0985997624 | 0.2721076927 | 0.1593648221 | 0.0 |
| 19 | 0.4997927954 | 0.096071915 | 0.2369617911 | 0.8118898307 | 0.6460952105 | 0.8707811364 | 0.4496663761 | 0.254943708 | 0.4 |
| 20 | 0.6429012946 | 0.0803355875 | 0.5068224588 | 0.9520570131 | 0.1408375208 | 0.1400332523 | 0.9902096502 | 0.0871451208 | 0.6 |
| 21 | 0.7057224081 | 0.4630195114 | 0.9144776522 | 0.3820045677 | 0.6039446075 | 0.8557347907 | 0.9241318835 | 0.4712547997 | 0.9 |
| 22 | 0.2484984729 | 0.2377738538 | 0.3104894005 | 0.3790386592 | 0.5252200689 | 0.1912480153 | 0.8910088711 | 0.0505195535 | 0.0 |
| 23 | 0.5463155243 | 0.5161090947 | 0.086890304 | 0.506719203 | 0.0861523314 | 0.0243307806 | 0.9958105 | 0.685405273 | 0.3 |
| 24 | 0.1472119791 | 0.4978118786 | 0.3711289708 | 0.8731593913 | 0.1267880907 | 0.1878780541 | 0.9781146096 | 0.5520577445 | 0.6 |
| 25 | 0.790510121 | 0.4327639828 | 0.7940384434 | 0.0472936828 | 0.4442862018 | 0.2649208629 | 0.6887534888 | 0.1463450387 | 0.1 |
| 26 | 0.238609707 | 0.3432042131 | 0.0508109927 | 0.053046589 | 0.8862833569 | 0.4834029363 | 0.7091768233 | 0.9173126281 | 0.4 |
| 27 | 0.2727451844 | 0.7989924806 | 0.8074715317 | 0.8630560155 | 0.520064408 | 0.9395456181 | 0.7433063779 | 0.7300026792 | 0.4 |
| 28 | 0.0233547923 | 0.8923628118 | 0.8213618146 | 0.0381375295 | 0.7760785248 | 0.9095269332 | 0.4853258794 | 0.3151255539 | 0.0 |
| 29 | 0.431277638 | 0.8813339969 | 0.6464634317 | 0.1806829796 | 0.0149378479 | 0.1696688208 | 0.0184137947 | 0.4967891628 | 0.6 |
| 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 | 0.03000377 | 0.4826313867 | 0.6086405084 | 0.3349117569 | 0.7 |
| 32 | 0.8232727759 | 0.9862914453 | 0.2903483235 | 0.4740627456 | 0.5629339886 | 0.5561574434 | 0.9774761845 | 0.778319096 | 0.0 |
| 33 | 0.0751000000 | 0.1000000000 | 0.9000000000 | 0.0000000000 | 0.0000000000 | 0.1000000000 | 0.9000000000 | 0.0000000000 | 0.0 |

Here is my Data

How do I know?

Entropy



Sample \rightarrow

How do I know? Entrop

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:

Amount of **information**
encoded in variable x

$H(x)$

=

\sum_x

$p(x)$

\log

$\frac{1}{p(x)}$

Average over
all outcomes?

How many bits
should I spend
encoding this
outcome?
(more likely
outcomes get
fewer bits)

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 |

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

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 | |

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

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

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

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

$$= 1$$

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 | |

$$\begin{aligned}
 H(x) &= \sum_x p(x) \log \frac{1}{p(x)} \\
 &= p(0) \log \frac{1}{p(0)} + p(1) \log \frac{1}{p(1)} \\
 &= \frac{15}{16} \log\left(\frac{16}{15}\right) + \frac{1}{16} \log(16) \\
 &= \frac{15}{16} (0.093) + \frac{1}{16} (4) \\
 &= 0.337
 \end{aligned}$$

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 |

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

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

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

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

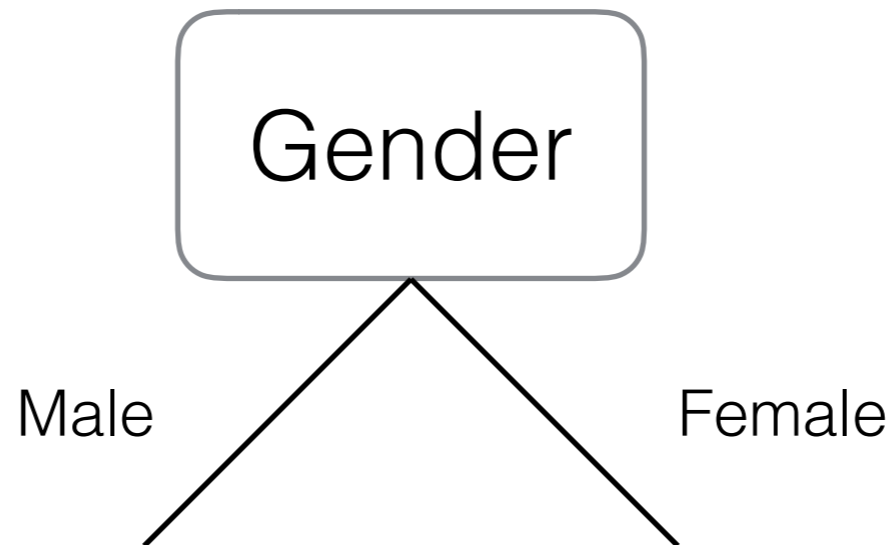
$$= 0.337$$

Info Theory & Entropy

Example

Gender

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AF | AG | AH | AI | AJ | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV | AW | AX | AY | AZ | BA | BB | BC | BD | BE | BF | BG | BH | BI | BJ | BK | BL | BM | BN | BO | BP | BQ | BR | BS | BT | BU | BV | BW | BX | BY | BZ | CA | CB | CC | CD | CE | CF | CG | CH | CI | CJ | CK | CL | CM | CN | CO | CP | CQ | CR | CS | CT | CU | CV | CW | CX | CY | CZ | DA | DB | DC | DD | DE | DF | DG | DH | DI | DJ | DK | DL | DM | DN | DO | DP | DQ | DR | DS | DT | DU | DV | DW | DX | DY | DZ | EA | EB | EC | ED | EE | EF | EG | EH | EI | EJ | EK | EL | EM | EN | EO | EP | EQ | ER | ES | ET | EU | EV | EW | EX | EY | EZ | FA | FB | FC | FD | FE | FF | FG | FH | FI | FJ | FK | FL | FM | FN | FO | FP | FQ | FR | FS | FT | FU | FV | FW | FX | FY | FZ | GA | GB | GC | GD | GE | GF | GG | GH | GI | GJ | GK | GL | GM | GN | GO | GP | GQ | GR | GS | GT | GU | GV | GW | GX | GY | GZ | HA | HB | HC | HD | HE | HF | HG | HH | HI | HJ | HK | HL | HM | HN | HO | HP | HQ | HR | HS | HT | HU | HV | HW | HX | HY | HZ | IA | IB | IC | ID | IE | IF | IG | IH | II | IJ | IK | IL | IM | IN | IO | IP | IQ | IR | IS | IT | IU | IV | IW | IX | IY | IZ | JA | JB | JC | JD | JE | JF | JG | JH | JI | JJ | JK | JL | JM | JN | JO | JP | JQ | JR | JS | JT | JU | JV | JW | JX | JY | JZ | KA | KB | KC | KD | KE | KF | KG | KH | KI | KJ | KK | KL | KM | KN | KO | KP | KQ | KR | KS | KT | KU | KV | KW | KX | KY | KZ | LA | LB | LC | LD | LE | LF | LG | LH | LI | LJ | LK | LL | LM | LN | LO | LP | LQ | LR | LS | LT | LU | LV | LW | LX | LY | LZ | MA | MB | MC | MD | ME | MF | MG | MH | MI | MJ | MK | ML | MM | MN | MO | MP | MQ | MR | MS | MT | MU | MV | MW | MX | MY | MZ | NA | NB | NC | ND | NE | NF | NG | NH | NI | NJ | NK | NL | NM | NO | NP | NQ | NR | NS | NT | NU | NV | NW | NX | NY | NZ | OA | OB | OC | OD | OE | OF | OG | OH | OI | OJ | OK | OL | OM | ON | OO | OP | OQ | OR | OS | OT | OU | OV | OW | OX | OY | OZ | PA | PB | PC | PD | PE | PF | PG | PH | PI | PJ | PK | PL | PM | PN | PO | PP | PQ | PR | PS | PT | PU | PV | PW | PX | PY | PZ | QA | QB | QC | QD | QE | QF | QG | QH | QI | QJ | QK | QL | QM | QN | QO | QP | QR | QS | QT | QU | QV | QW | QX | QY | QZ | RA | RB | RC | RD | RE | RF | RG | RH | RI | RJ | RK | RL | RM | RN | RO | RP | RQ | RR | RS | RT | RU | RV | RW | RX | RY | RZ | SA | SB | SC | SD | SE | SF | SG | SH | SI | SJ | SK | SL | SM | SN | SO | SP | SQ | SR | SS | ST | SU | SV | SW | SX | SY | SZ | TA | TB | TC | TD | TE | TF | TG | TH | TI | TJ | TK | TL | TM | TN | TO | TP | TQ | TR | TS | TU | TV | TW | TX | TY | TZ | UA | UB | UC | UD | UE | UF | UG | UH | UI | UJ | UK | UL | UM | UN | UO | UP | UQ | UR | US | UT | UU | UV | UW | UX | UY | UZ | VA | VB | VC | VD | VE | VF | VG | VH | VI | VJ | VK | VL | VM | VN | VO | VP | VQ | VR | VS | VT | VU | VV | VW | VX | VY | VZ | WA | WB | WC | WD | WE | WF | WG | WH | WI | WJ | WK | WL | WM | WN | WO | WP | WQ | WR | WS | WT | WU | WV | WW | WX | WY | WZ | XA | XB | XC | XD | XE | XF | XG | XH | XI | XJ | XK | XL | XM | XN | XO | XP | XQ | XR | XS | XT | XU | XV | XW | XX | XY | XZ | YA | YB | YC | YD | YE | YF | YG | YH | YI | YJ | YK | YL | YM | YN | YO | YP | YQ | YR | YS | YT | YU | YV | YW | YX | YY | YZ | ZA | ZB | ZC | ZD | ZE | ZF | ZG | ZH | ZI | ZJ | ZK | ZL | ZM | ZN | ZO | ZA | ZB | ZC | ZD | ZE | ZF | ZG | ZH | ZI | ZJ | ZK | ZL | ZM | ZN | ZO |
|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|



Most Informative Variable:

First Decision in my Tree

Then What?

Gender

Females Males

[illegible]

2) Split According to Most Informative Variable

3) Find Most Informative Variable in Each Subset

Gender

Age

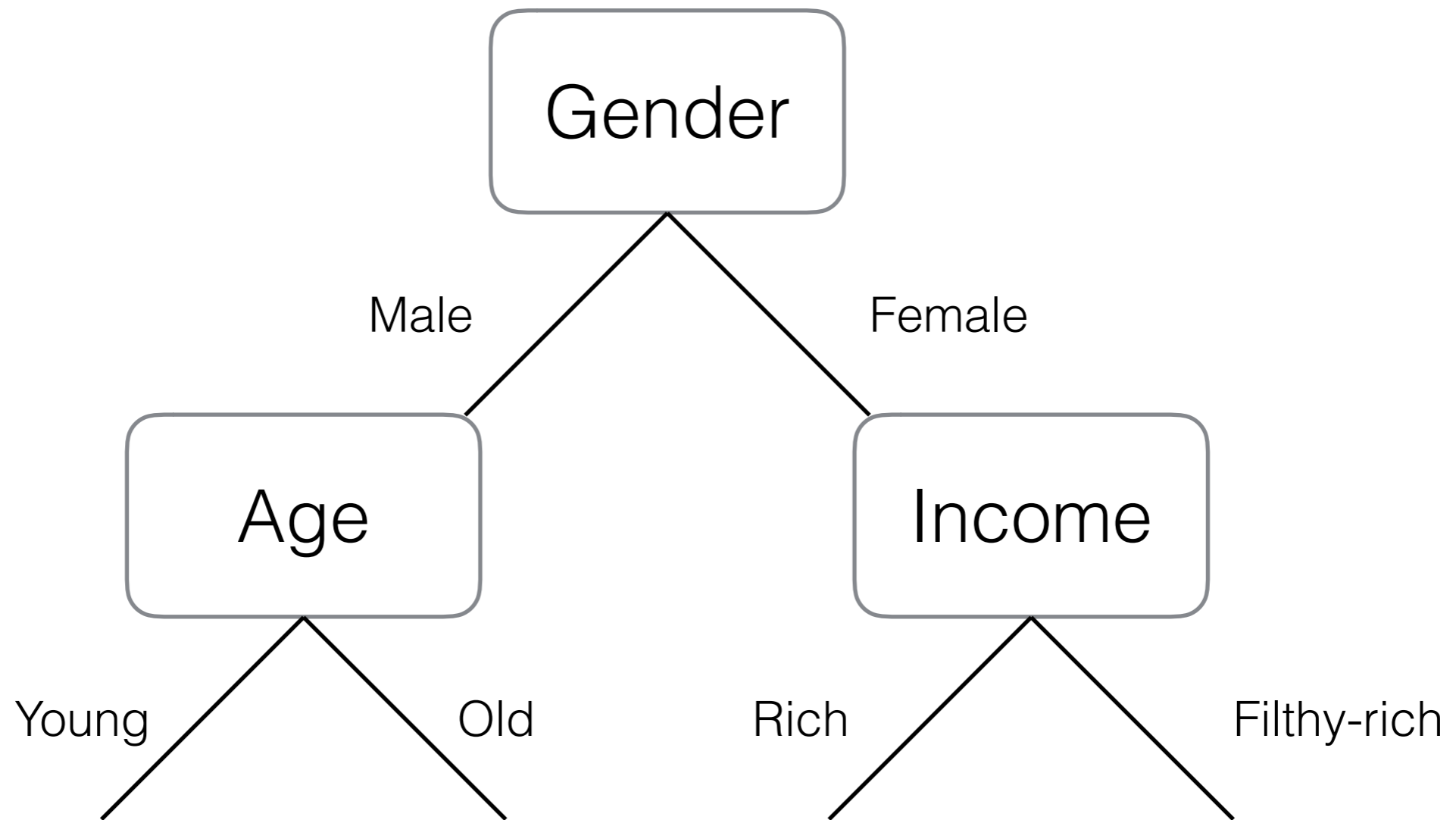
Income

Males

Females

Repeat

3) Find Most Informative Variable in Each Subset



Each Informative Variable:

One Decision in my Tree

Then What?

Gender Age

Males

und
Old

Filtv-r

23

Repeat

When do we s

- 2) Split According to Most Informative Variable
- 3) Find Most Informative Variable in Each Subset

[illegible]

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 13494504 | 13494505 | 13494506 | 13494507 | 13494508 | 13494509 | 13494510 | 13494511 | 13494512 | 13494513 | 13494514 | 13494515 | 13494516 | 13494517 | 13494518 | 13494519 | 13494520 | 13494521 | 13494522 | 13494523 | 13494524 | 13494525 | 13494526 | 13494527 | 13494528 | 13494529 | 13494530 | 13494531 | 13494532 | 13494533 | 13494534 | 13494535 | 13494536 | 13494537 | 13494538 | 13494539 | 13494540 | 13494541 | 13494542 | 13494543 | 13494544 | 13494545 | 13494546 | 13494547 | 13494548 | 13494549 | 13494550 | 13494551 | 13494552 | 13494553 | 13494554 | 13494555 | 13494556 | 13494557 | 13494558 | 13494559 | 13494560 | 13494561 | 13494562 | 13494563 | 13494564 | 13494565 | 13494566 | 13494567 | 13494568 | 13494569 | 13494570 | 13494571 | 13494572 | 13494573 | 13494574 | 13494575 | 13494576 | 13494577 | 13494578 | 13494579 | 13494580 | 13494581 | 13494582 | 13494583 | 13494584 | 13494585 | 13494586 | 13494587 | 13494588 | 13494589 | 13494590 | 13494591 | 13494592 | 13494593 | 13494594 | 13494595 | 13494596 | 13494597 | 13494598 | 13494599 | 13494600 | 13494601 | 13494602 | 13494603 | 13494604 | 13494605 | 13494606 | 13494607 | 13494608 | 13494609 | 13494610 | 13494611 | 13494612 | 13494613 | 13494614 | 13494615 | 13494616 | 13494617 | 13494618 | 13494619 | 13494620 | 13494621 | 13494622 | 13494623 | 13494624 | 13494625 | 13494626 | 13494627 | 13494628 | 13494629 | 13494630 | 13494631 | 13494632 | 13494633 | 13494634 | 13494635 | 13494636 | 13494637 | 13494638 | 13494639 | 13494640 | 13494641 | 13494642 | 13494643 | 13494644 | 13494645 | 13494646 | 13494647 | 13494648 | 13494649 | 13494650 | 13494651 | 13494652 | 13494653 | 13494654 | 13494655 | 13494656 | 13494657 | 13494658 | 13494659 | 13494660 | 13494661 | 13494662 | 13494663 | 13494664 | 13494665 | 13494666 | 13494667 | 13494668 | 13494669 | 13494670 | 13494671 | 13494672 | 13494673 | 13494674 | 13494675 | 13494676 | 13494677 | 13494678 | 13494679 | 13494680 | 13494681 | 13494682 | 13494683 | 13494684 | 13494685 | 13494686 | 13494687 | 13494688 | 13494689 | 13494690 | 13494691 | 13494692 | 13494693 | 13494694 | 13494695 | 13494696 | 13494697 | 13494698 | 13494699 | 13494700 | 13494701 | 13494702 | 13494703 | 13494704 | 13494705 | 13494706 | 13494707 | 13494708 | 13494709 | 13494710 | 13494711 | 13494712 | 13494713 | 13494714 | 13494715 | 13494716 | 13494717 | 13494718 | 13494719 | 13494720 | 13494721 | 13494722 | 13494723 | 13494724 | 13494725 | 13494726 | 13494727 | 13494728 | 13494729 | 13494730 | 13494731 | 13494732 | 13494733 | 13494734 | 13494735 | 13494736 | 13494737 | 13494738 | 13494739 | 13494740 | 13494741 | 13494742 | 13494743 | 13494744 | 13494745 | 13494746 | 13494747 | 13494748 | 13494749 | 13494750 | 13494751 | 13494752 | 13494753 | 13494754 | 13494755 | 13494756 | 13494757 | 13494758 | 13494759 | 13494760 | 13494761 | 13494762 | 13494763 | 13494764 | 13494765 | 13494766 | 13494767 | 13494768 | 13494769 | 13494770 | 13494771 | 13494772 | 13494773 | 13494774 | 13494775 | 13494776 | 13494777 | 13494778 | 13494779 | 13494780 | 13494781 | 13494782 | 13494783 | 13494784 | 13494785 | 13494786 | 13494787 | 13494788 | 13494789 | 13494790 | 13494791 | 13494792 | 13494793 | 13494794 | 13494795 | 13494796 | 13494797 | 13494798 | 13494799 | 13494800 | 13494801 | 13494802 | 13494803 | 13494804 | 13494805 | 13494806 | 13494807 | 13494808 | 13494809 | 13494810 | 13494811 | 13494812 | 13494813 | 13494814 | 13494815 | 13494816 | 13494817 | 13494818 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

[illegible]

Income

2) Split According to Most Informative Variable

3) Find Most Informative Variable in Each Subset


o 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 | |

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

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 |

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

This Variable Provides
No Information!

Info Theory & Entropy

Extreme Case

Gender Age

Old

Young

Males

Females

Fifty-rich

Rich

Income

Interest

Subset

Zero entropy

When

Variable

Interest

Information

Subset

Zero entropy

When

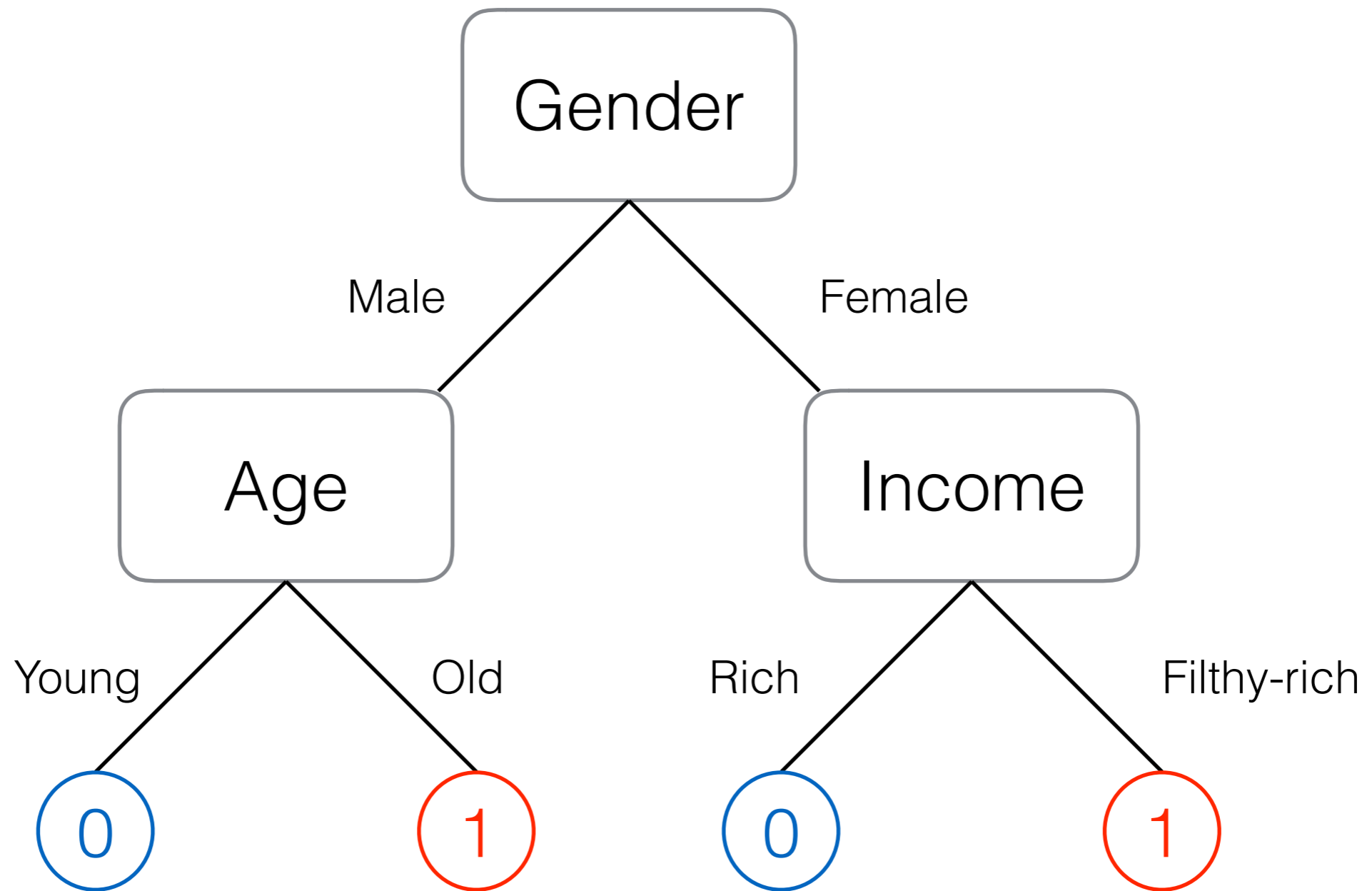
Variable

Interest

Information

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?

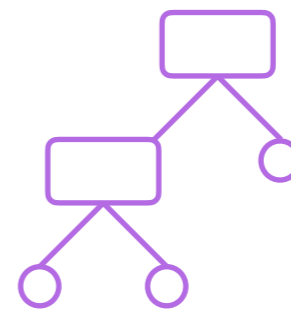
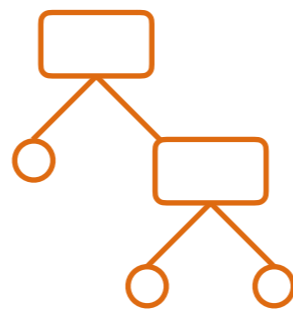
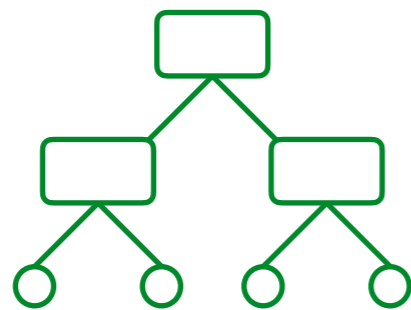
Overfitting & Bias

- **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

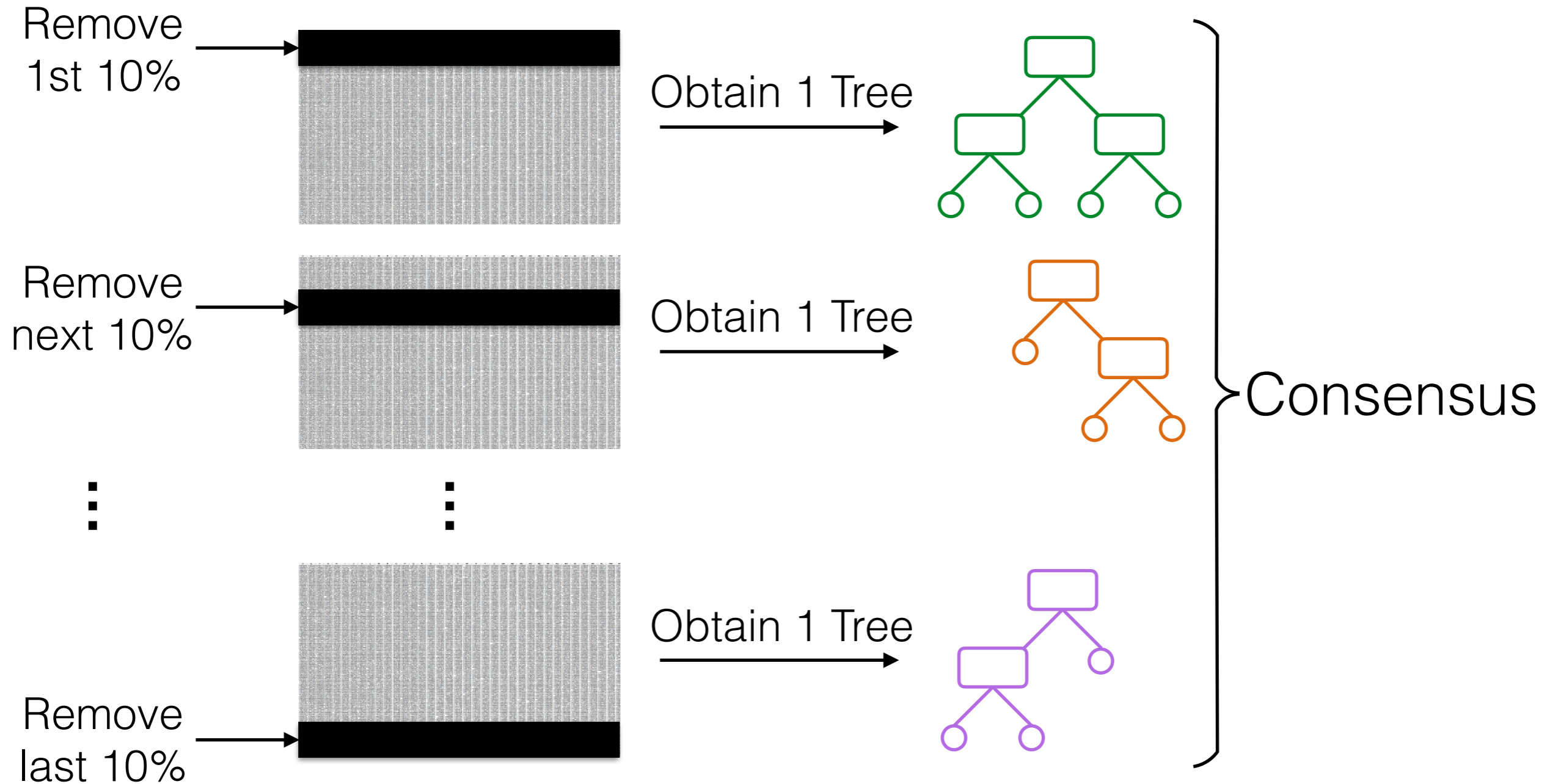
No worries:
Random Forests



Random Forests

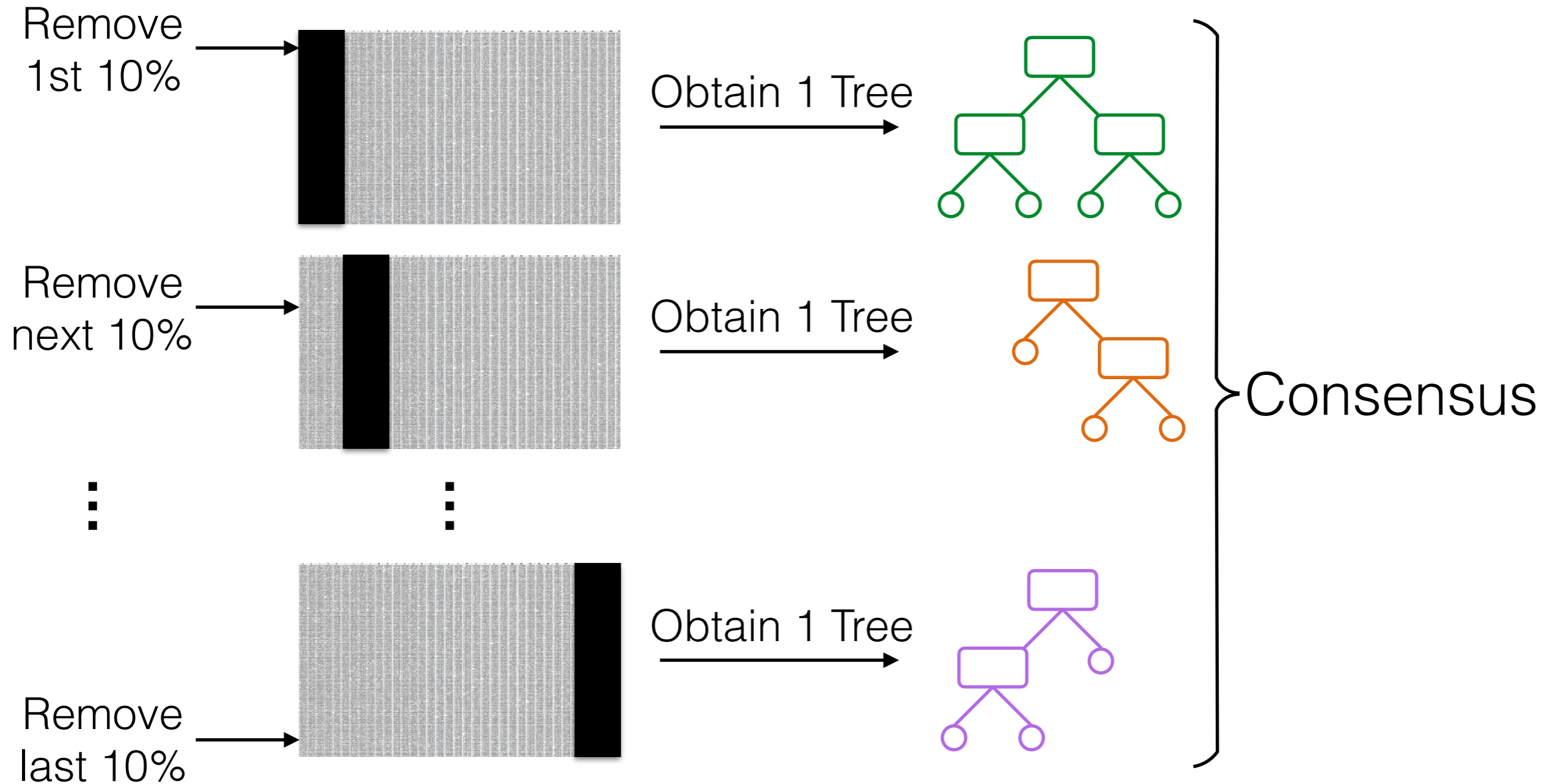
Main Idea: Do many decision trees,
each with a random subsample

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

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

Alternatives

Questions?