NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

COURSE ADMINISTRATION

- CDS, 60 5th Avenue, room 110
- Lecture: Wednesday 2-3:40pm
- Lab section: Wednesday 3:50-4:40pm

OFFICE HOURS AND CONTACT INFORMATION

- CDS, 60 5th Avenue, 6th floor, room 620
- Wednesday 12-2pm: Lecturer, Iddo Drori, idrori@nyu.edu
- Tuesday 11am-1pm: Section Leader, Datta Sainath Dwarampudi, ddattasainath@nyu.edu
- Friday 2-4pm: Grader, Samhita Damotharan, sd2941@nyu.edu
- Thursday 2-4pm: Grader, Sai Anirudh Kondaveeti, sak797@nyu.edu

SYLLABUS

- Available on NYU Classes

EXERCISES AND PROJECT

- 3 equally weighted exercises 30% individually
- Term project 25% in pairs
- Midterm 20%
- Final exam 25%

HOMEWORK 1

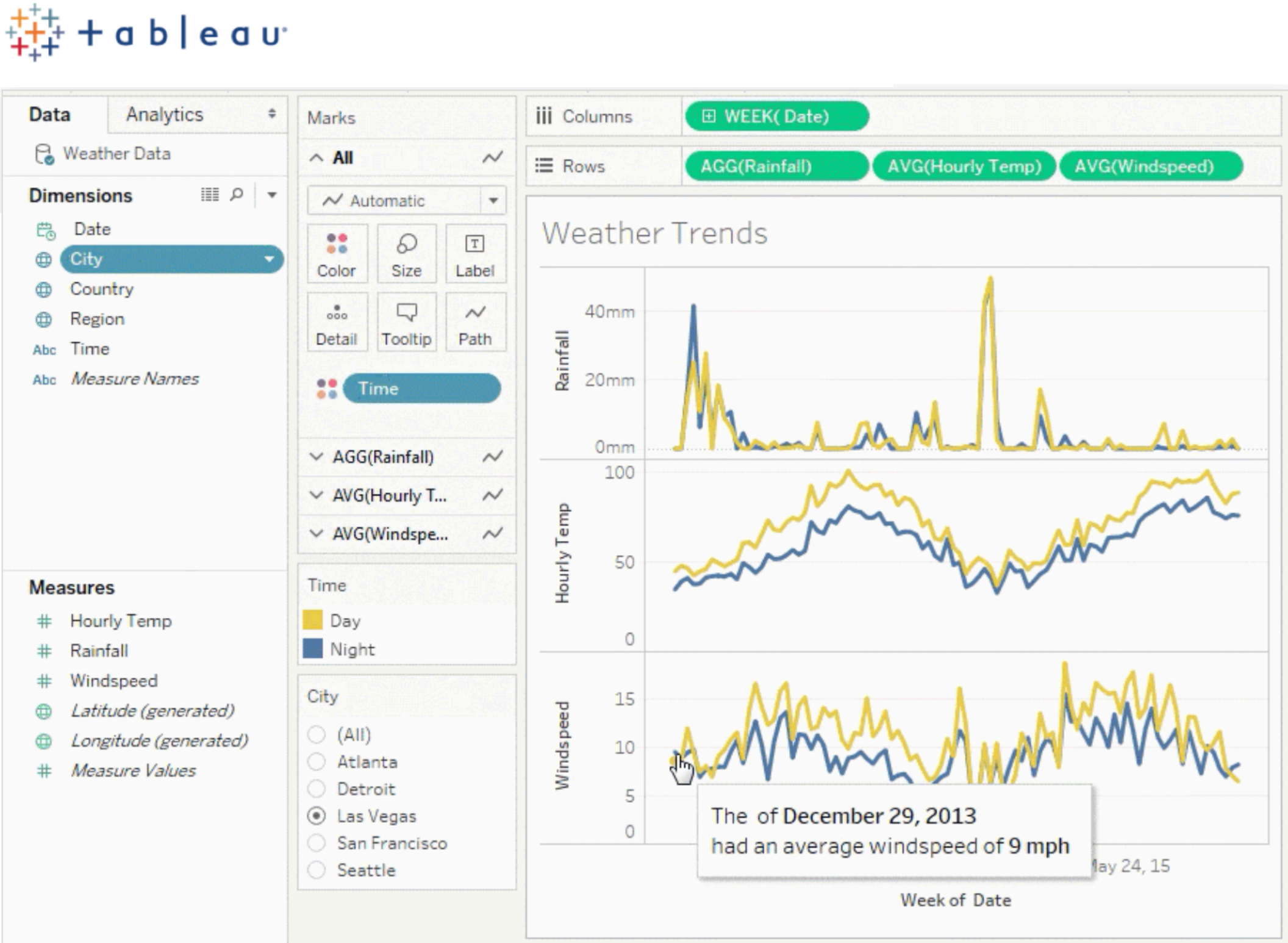
- Tableau
- 10 equally weighted questions
- Submit answers and screenshots in word/pdf
- Individually
- Due February 5th

TEXTBOOKS

- [CASI] Computer Age Statistical Inference: Algorithms, Evidence and Data Science, Bradley Efron and Trevor Hastie, Cambridge University Press, 2016.
- [DSB] Data Science for Business, Foster Provost and Tom Fawcett, O'Reilly Media, 2013.
- Introduction to Computation and Programming using Python, with Application to Understanding Data, John Guttag, 2nd Edition, MIT Press, 2016.

EXAMPLES

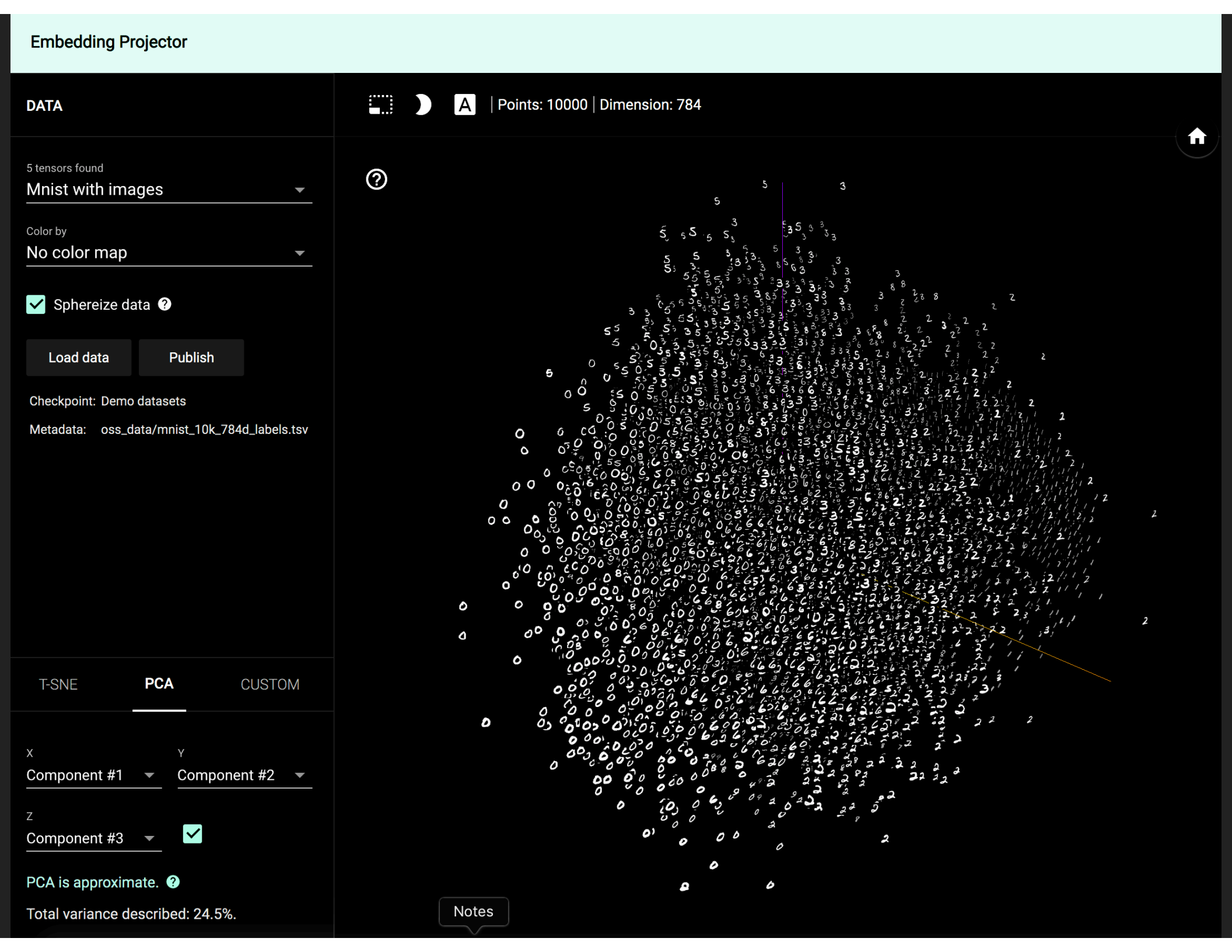
- Tableau
- Google experiments


Data
Analytics
Weather Data
Dimensions
Date
City
Country
Region
Time
Measure Names
Measures
Hourly Temp
Rainfall
Windspeed
Latitude (generated)
Longitude (generated)
Measure Values
Marks
All
Automatic
Color
Size
Label
Detail
Tooltip
Path
Time
AGG(Rainfall)
AVG(Hourly T...
AVG(Windspe...
Time
Day
Night
City
(All)
Atlanta
Detroit
Las Vegas
San Francisco
Seattle
Columns
WEEK(Date)
Rows
AGG(Rainfall)
AVG(Hourly Temp)
AVG(Windspeed)
Weather Trends
Rainfall
40mm
20mm
0mm
Hourly Temp
100
50
0
Windspeed
15
10
5
0
The of December 29, 2013
had an average windspeed of 9 mph
May 24, 15
Week of Date

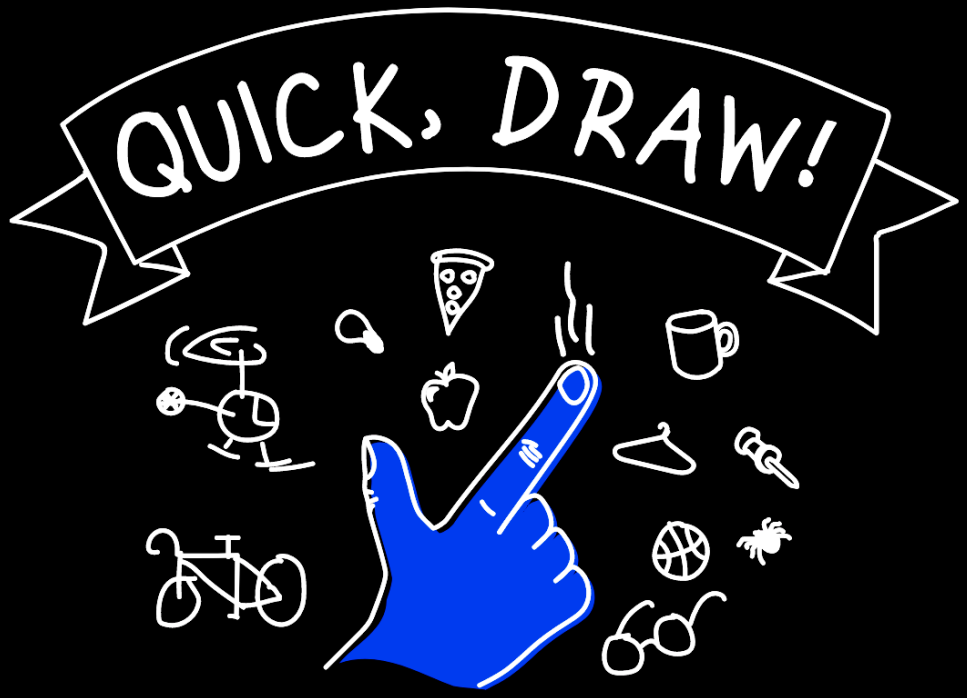
Area-wide solar potential

Search for a city, state, or zip code to see solar potential and impact across entire geographic areas. We currently have solar data for portions of 50 states and Washington DC. See if we've got you covered.

EXPLORE YOUR AREA


Embedding Projector
DATA
Points: 10000 | Dimension: 784
5 tensors found
Mnist with images
Color by
No color map
Sphereize data
Load data
Publish
Checkpoint: Demo datasets
Metadata: oss_data/mnist_10k_784d_labels.tsv
T-SNE
PCA
CUSTOM
X
Component #1
Y
Component #2
Z
Component #3
PCA is approximate.
Total variance described: 24.5%.
Notes

?



Can a neural network learn to recognize doodling?

Help teach it by adding your drawings to the world's largest doodling data set, shared publicly to help with machine learning research.

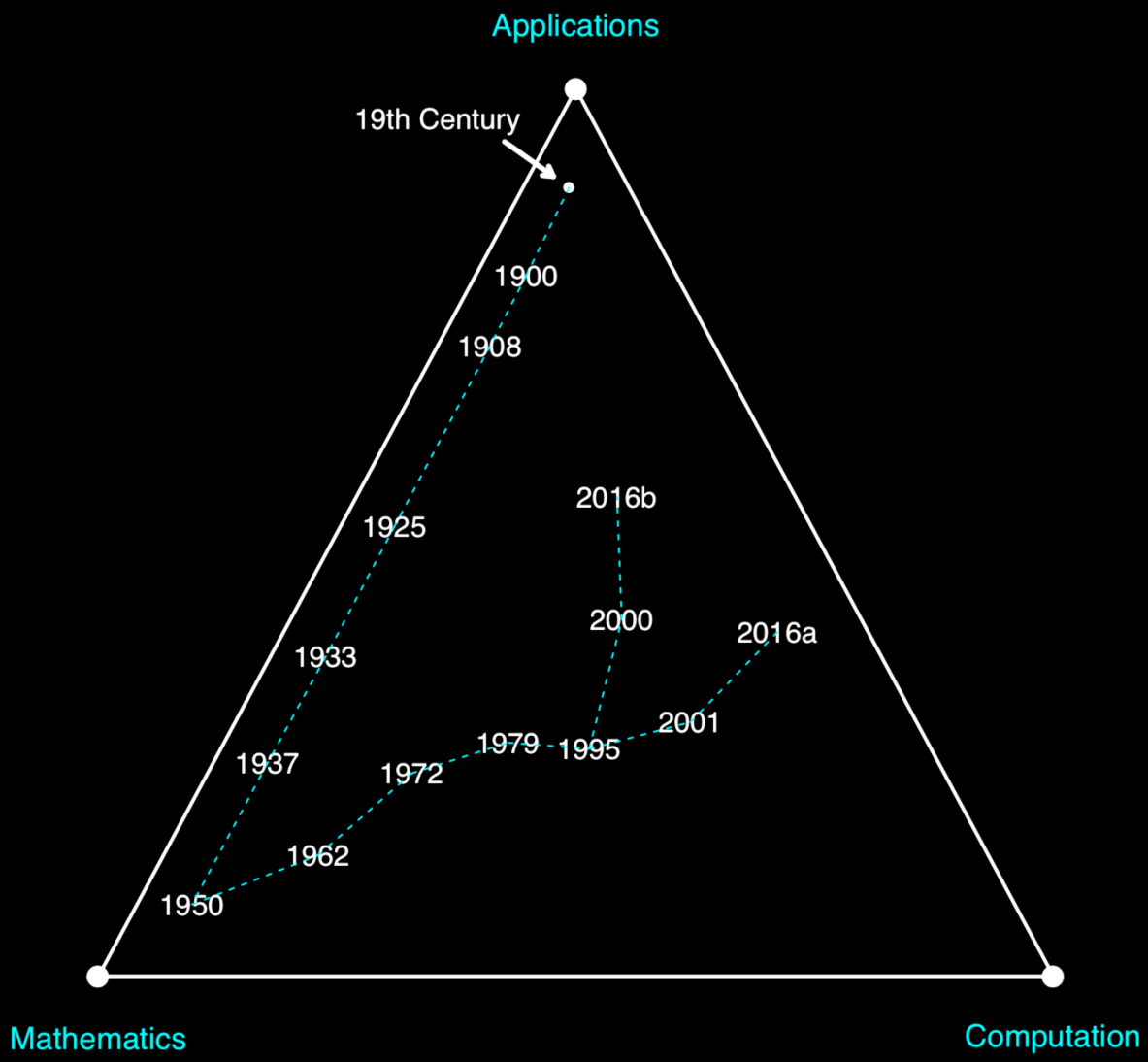
Let's Draw!

This is an A.I. Experiment

Made with some friends from Google

DATA SCIENCE DEFINITIONS

- CASI Epilogue.
- DSB Ch 1-2.
- Science and data science, David M. Blei and Padhraic Smyth, PNAS 2017.
- 50 years of data science, David Donoho, 2015.



1950

The publication of Wald's *Statistical Decision Functions*. Decision theory completed the full mathematization of statistical inference. This date can also stand for Savage's and de Finetti's decision-theoretic formulation of Bayesian inference. We are as far as possible from the Applications corner of the triangle now, and it is fair to describe the 1950s as a nadir of the influence of the statistics discipline on scientific applications.

1962

The arrival of electronic computation in the mid 1950s began the process of stirring statistics out of its inward-gazing preoccupation with mathematical structure. Tukey's paper "The future of data analysis" argued for a more application- and computation-oriented discipline. Mosteller and Tukey later suggested changing the field's name to *data analysis*, a prescient hint of today's *data science*.

1972

Cox's proportional hazards paper. Immensely useful in its own right, it signaled a growing interest in biostatistical applications and particularly survival analysis, which was to assert its scientific importance in the analysis of AIDS epidemic data.

1979

The bootstrap, and later the widespread use of MCMC: electronic computation used for the extension of classic statistical inference.

1995

This stands for false-discovery rates and, a year later, the lasso.[6] Both are computer-intensive algorithms, firmly rooted in the ethos of statistical inference. They lead, however, in different directions, as indicated by the split in the diagram.

2000

Microarray technology inspires enormous interest in large-scale inference, both in theory and as applied to the analysis of microbiological data.

2001

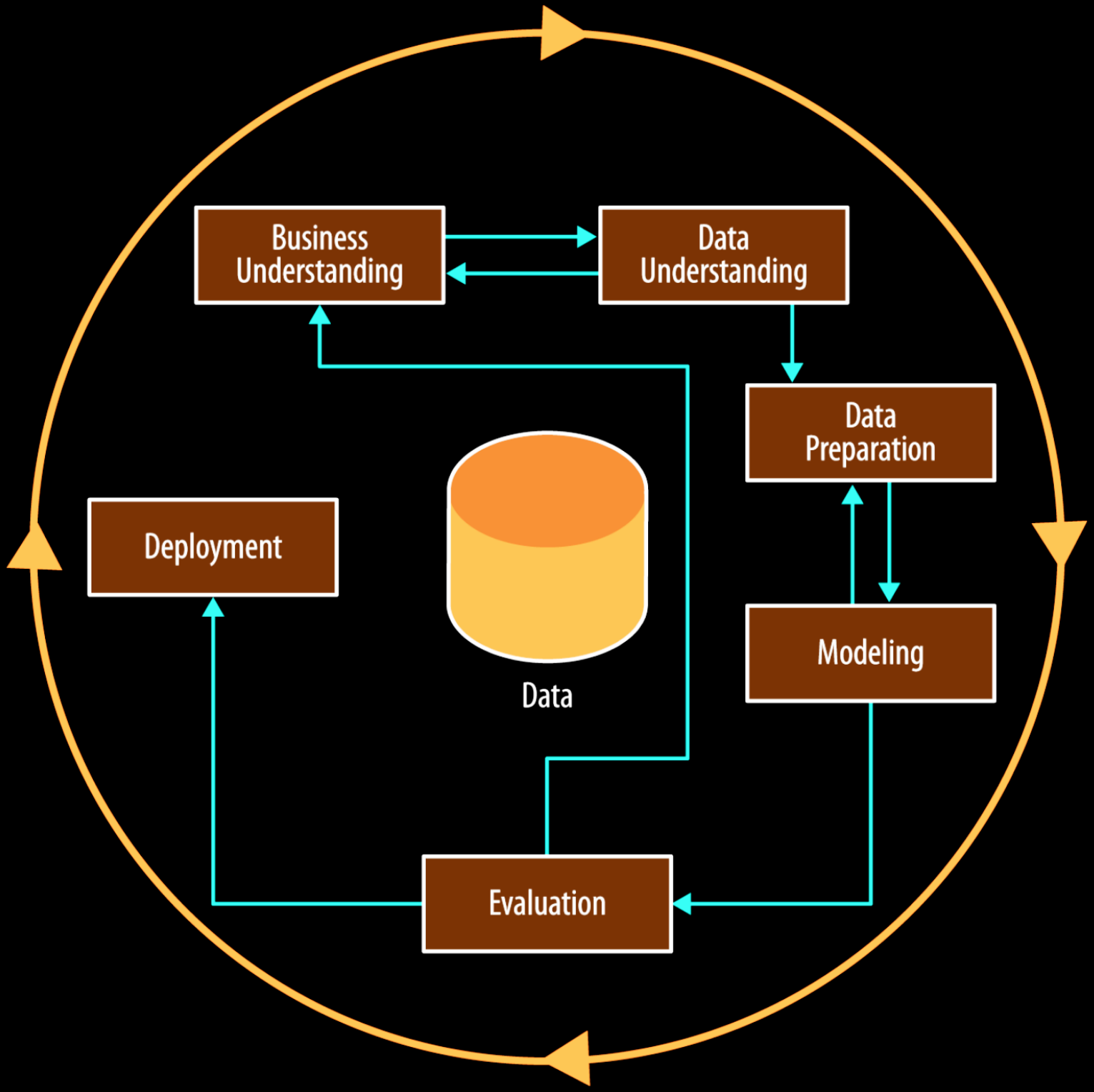
Random forests; it joins boosting[7] and the resurgence of neural nets in the ranks of *machine learning* prediction algorithms.

2016a

Data science: a more popular successor to Tukey and Mosteller's "data analysis," at one extreme it seems to represent a statistics discipline without parametric probability models or formal inference. The Data Science Association defines a practitioner as one who "…uses scientific methods to liberate and create meaning from raw data." In practice the emphasis is on the algorithmic processing of large data sets for the extraction of useful information, with the prediction algorithms as exemplars.

2016b

This represents the traditional line of statistical thinking, of the kind that could be located within Figure 14.1, but now energized with a renewed focus on applications. Of particular applied interest are biology and genetics. Genome-wide association studies (GWAS) show a different face of big data. Prediction is important here,[8] but not sufficient for the scientific understanding of disease.

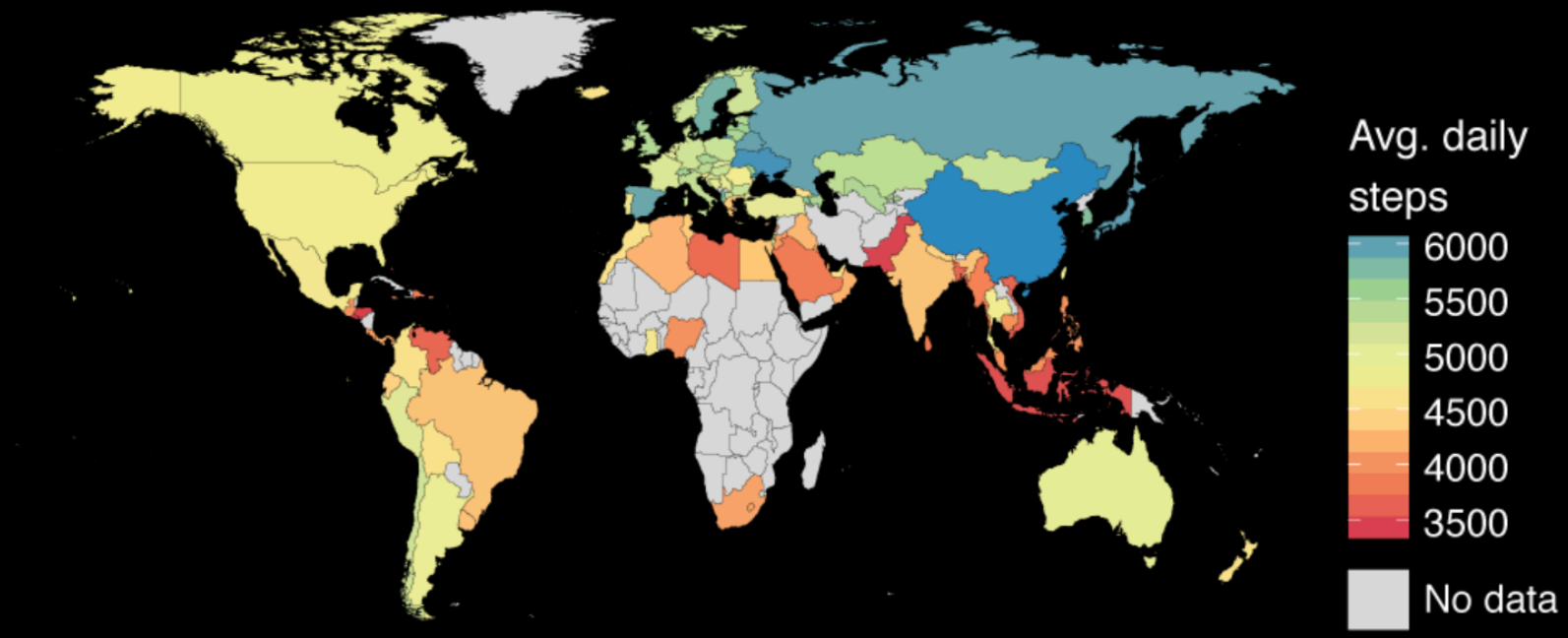

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

DATA SCIENCE PROJECT EXAMPLE

- Large-scale physical activity data reveal worldwide activity inequality, Tim Althoff, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp, and Jure Leskovec, Nature 2017.

ACTIVITY INEQUALITY PROJECT

Smartphone data from over 68 million days of activity by 717,527 individuals across 111 countries reveal variability in physical activity across the world.



TASKS

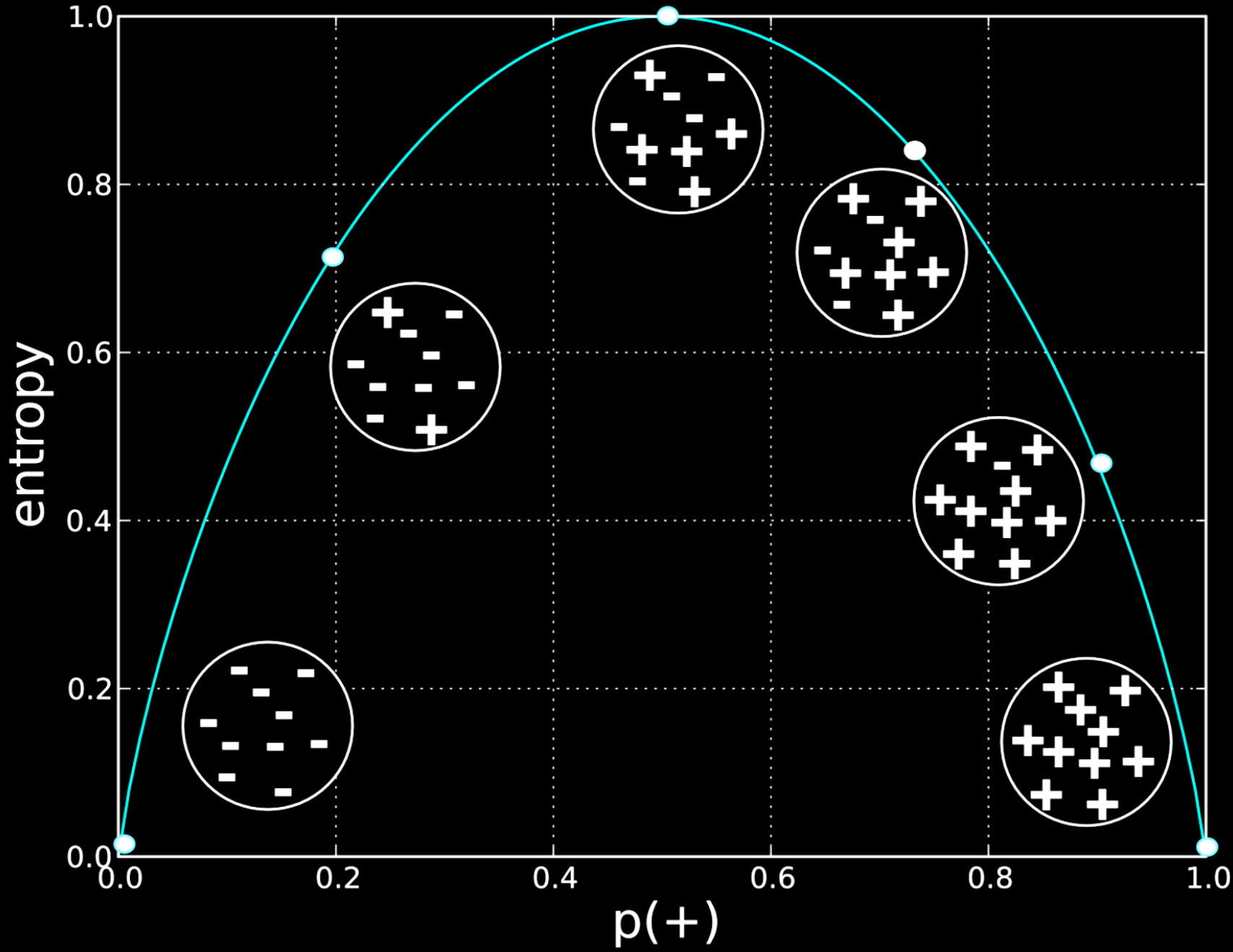
- Classification
- Regression
- Similarity matching
- Clustering
- Co-occurence grouping
- Profiling
- Link prediction
- Data reduction
- Causal modeling

TASK TYPES

- Supervised
- Unsupervised
- Reinforcement

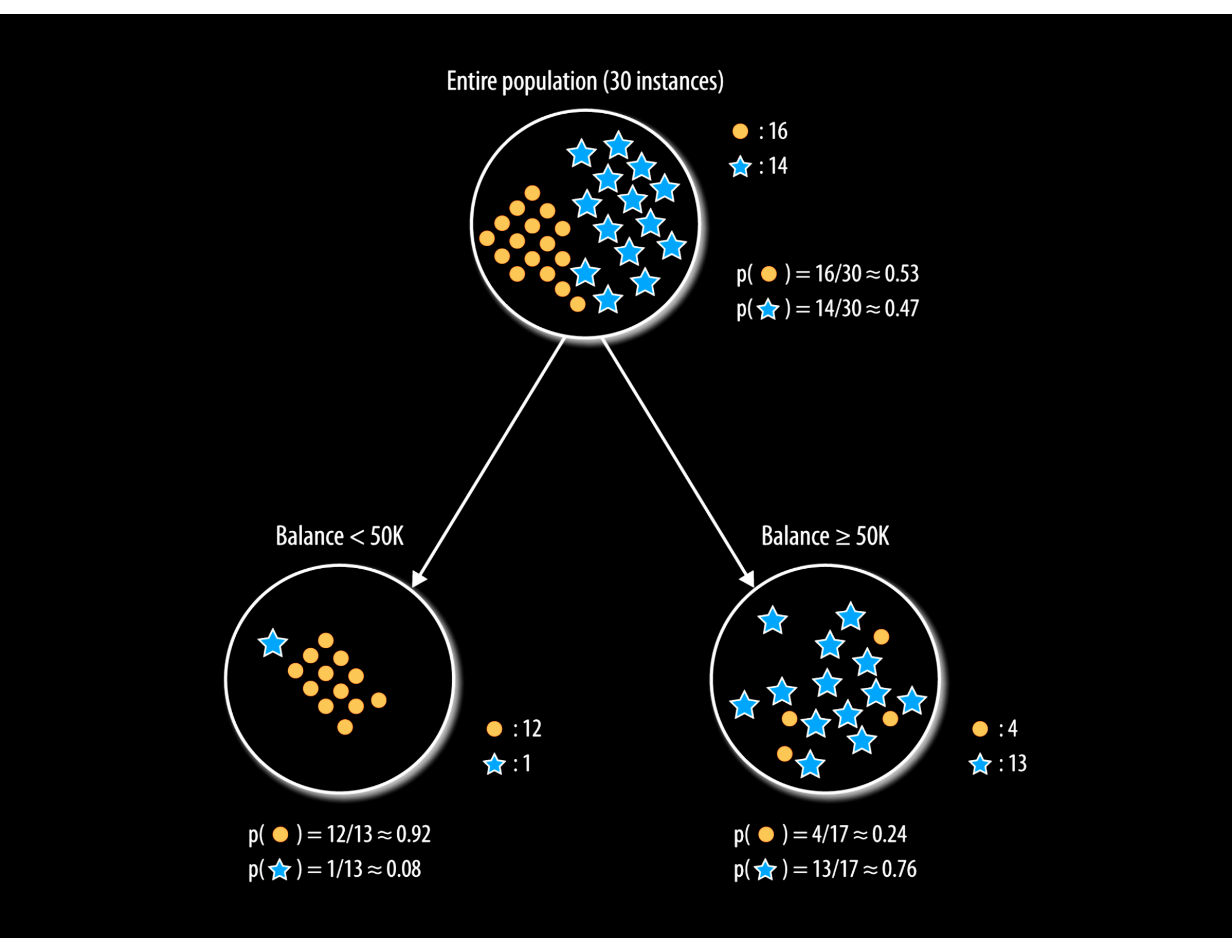
INFORMATION

- DSB Ch 3


1.0
0.8
0.6
0.4
0.2
0.0
entropy
0.0
0.2
0.4
0.6
0.8
1.0
p(+)

$$
\begin{aligned}
entropy(S) &= -[0.7 \times \log_2\ (0.7) + 0.3 \times \log_2\ (0.3)] \\
&\approx -[0.7 \times\ -0.51 + 0.3 \times\ -1.74] \\
&\approx 0.88
\end{aligned}
$$

$$
\begin{aligned}
IG(parent, children) = \ & entropy(parent) - \\
& [\, p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \ \cdots \,]
\end{aligned}
$$

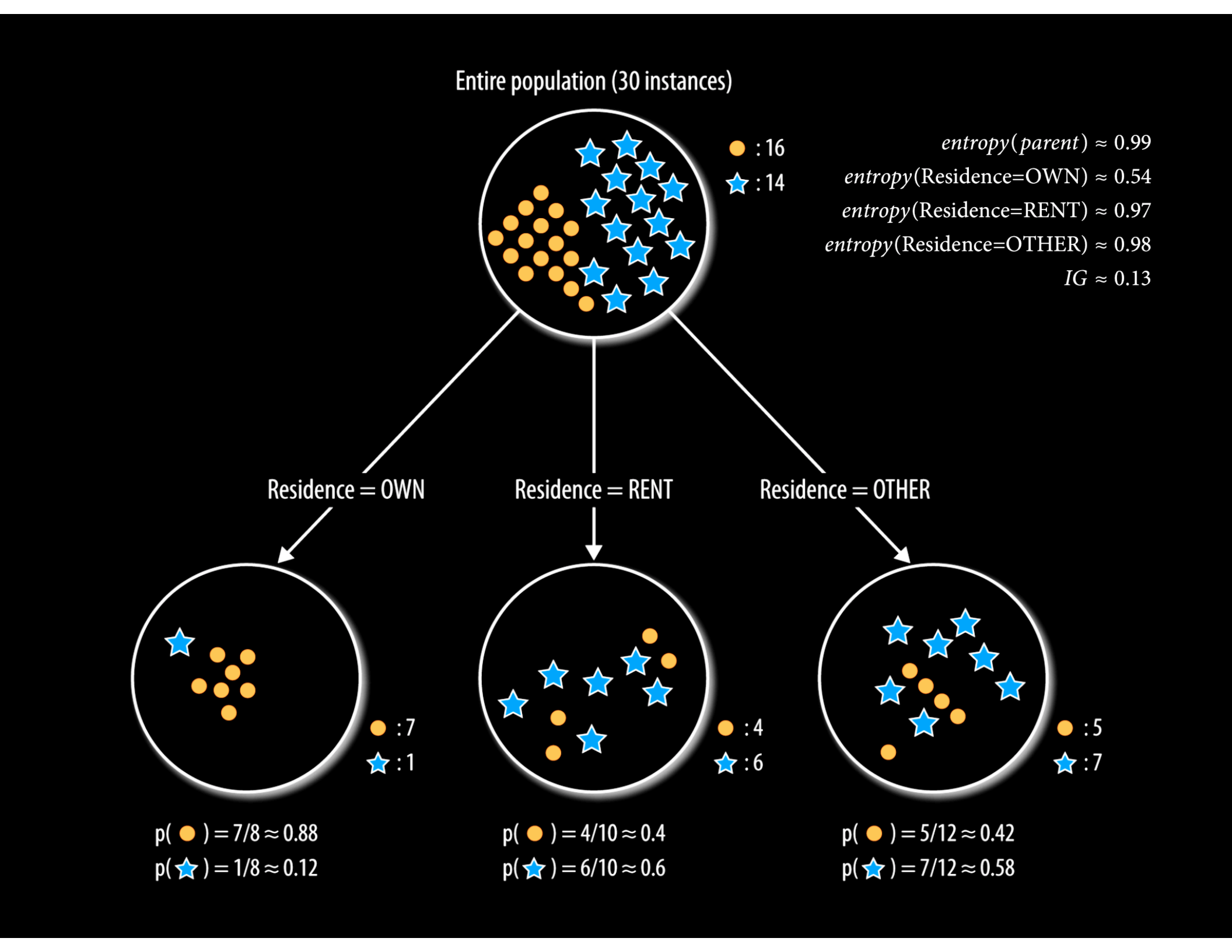

Entire population (30 instances)
● : 16
★ : 14
p(●) = 16/30 ≈ 0.53
p(★) = 14/30 ≈ 0.47
Balance < 50K
Balance ≥ 50K
● : 12
★ : 1
p(●) = 12/13 ≈ 0.92
p(★) = 1/13 ≈ 0.08
● : 4
★ : 13
p(●) = 4/17 ≈ 0.24
p(★) = 13/17 ≈ 0.76

$$
\begin{aligned}
entropy(parent) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.53 \times -0.9 + 0.47 \times -1.1] \\
&\approx 0.99 \quad \text{(very impure)}
\end{aligned}
$$

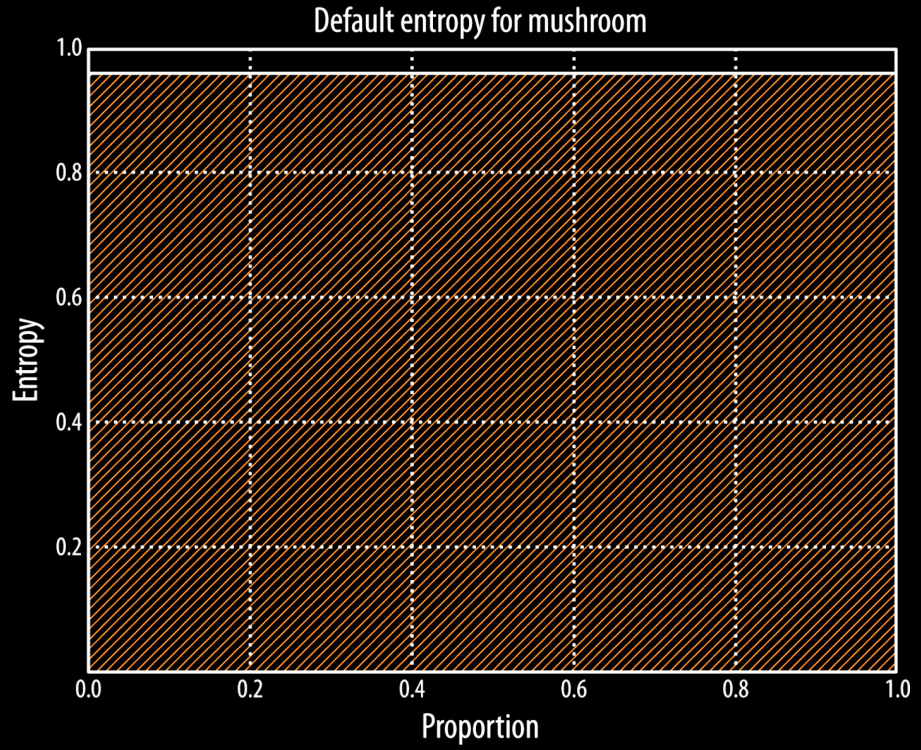
$$
\begin{aligned}
entropy(Balance < 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\
&\approx 0.39
\end{aligned}
$$

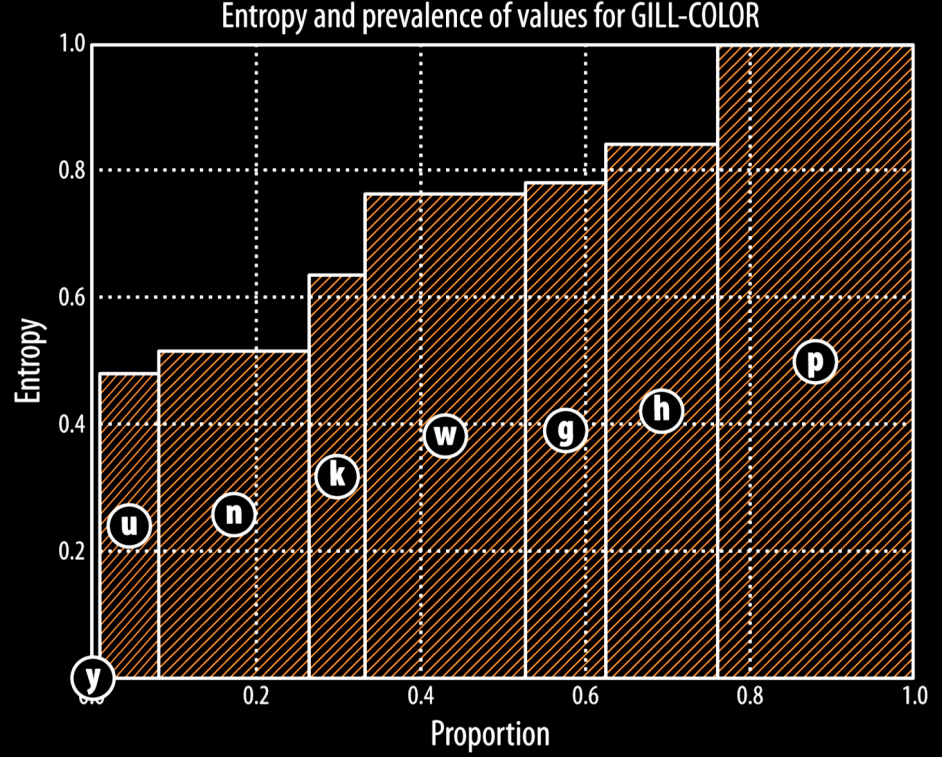
$$
\begin{aligned}
entropy(Balance \geq 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.24 \times (-2.1) + 0.76 \times (-0.39)] \\
&\approx 0.79
\end{aligned}
$$

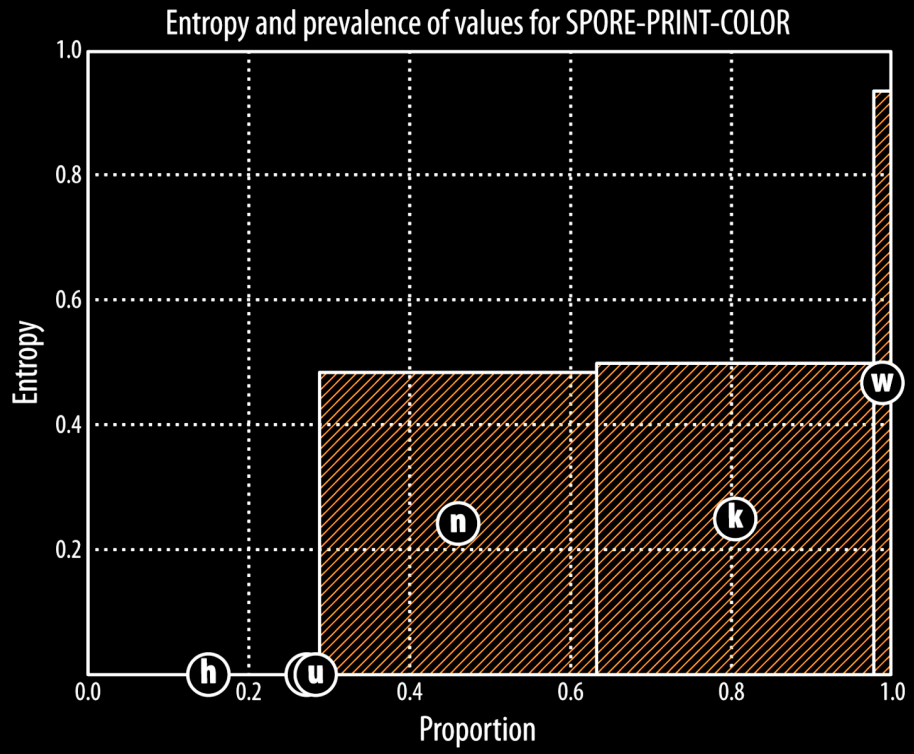
$$
\begin{aligned}
IG &= entropy(parent) - [p(\text{Balance} < 50\text{K}) \times entropy(\text{Balance} < 50\text{K}) \\
&\qquad + p(\text{Balance} \geq 50\text{K}) \times entropy(\text{Balance} \geq 50\text{K})] \\
&\approx 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79] \\
&\approx 0.37
\end{aligned}
$$

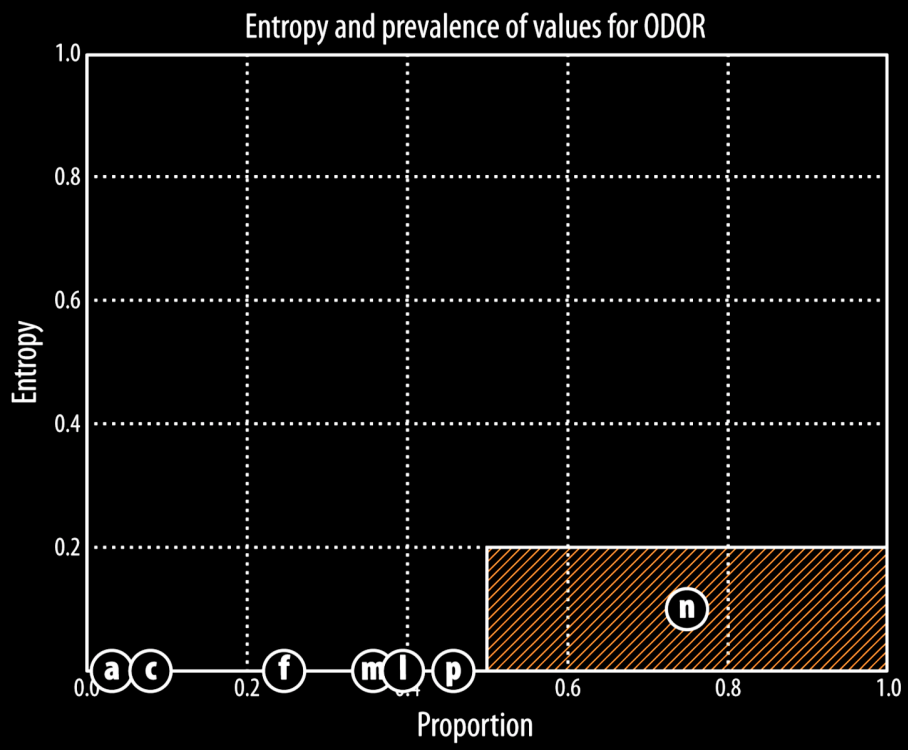

Entire population (30 instances)
: 16
: 14
entropy(parent) ≈ 0.99
entropy(Residence=OWN) ≈ 0.54
entropy(Residence=RENT) ≈ 0.97
entropy(Residence=OTHER) ≈ 0.98
IG ≈ 0.13
Residence = OWN
Residence = RENT
Residence = OTHER
: 7
: 1
: 4
: 6
: 5
: 7
p(●) = 7/8 ≈ 0.88
p(★) = 1/8 ≈ 0.12
p(●) = 4/10 ≈ 0.4
p(★) = 6/10 ≈ 0.6
p(●) = 5/12 ≈ 0.42
p(★) = 7/12 ≈ 0.58

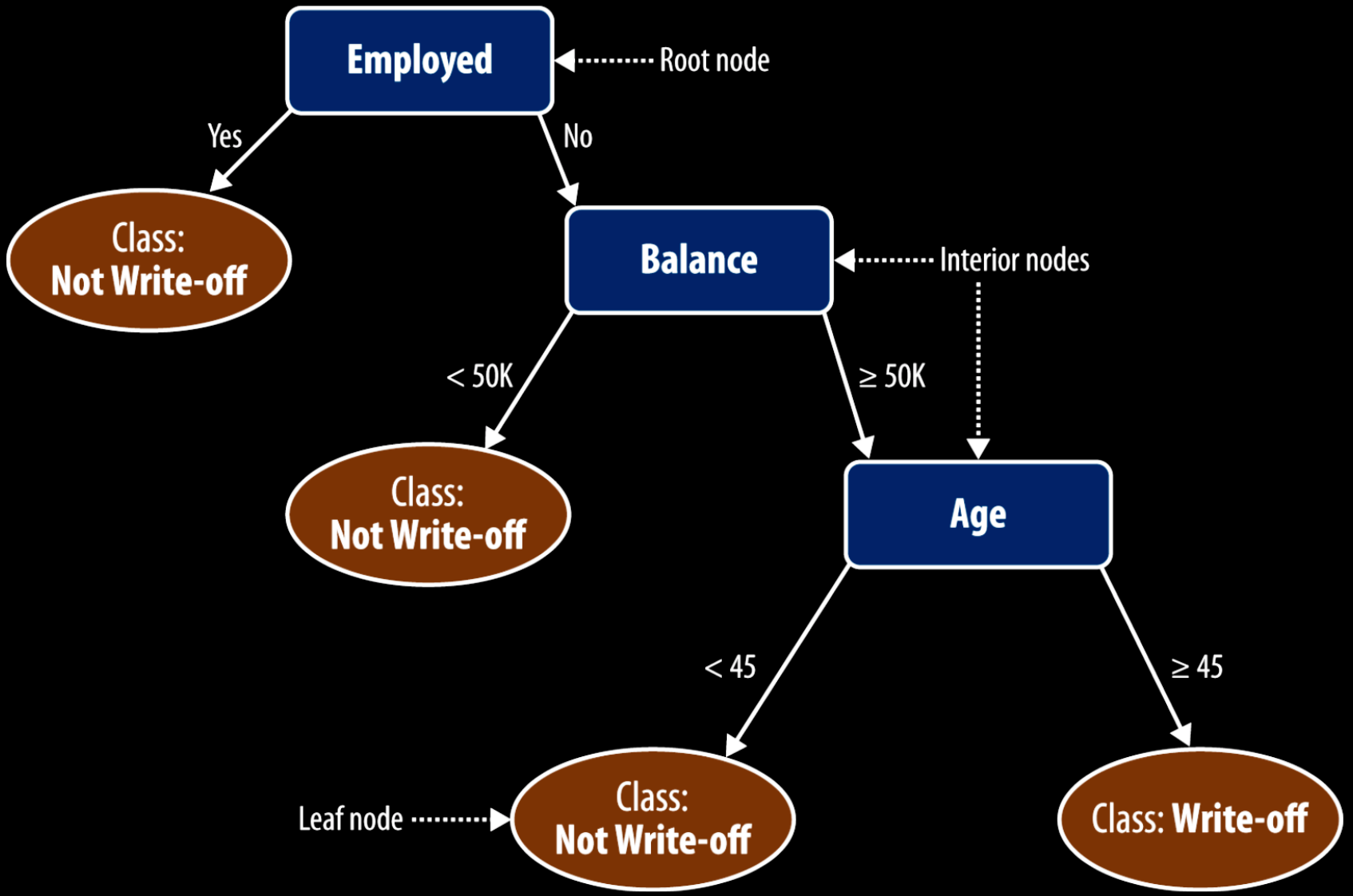
Attribute name	Possible values
CAP-SHAPE	bell, conical, convex, flat, knobbed, sunken
CAP-SURFACE	fibrous, grooves, scaly, smooth
CAP-COLOR	brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow
BRUISES?	yes, no
ODOR	almond, anise, creosote, fishy, foul, musty, none, pungent, spicy
GILL-ATTACHMENT	attached, descending, free, notched
GILL-SPACING	close, crowded, distant
GILL-SIZE	broad, narrow
GILL-COLOR	black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow
STALK-SHAPE	enlarging, tapering
STALK-ROOT	bulbous, club, cup, equal, rhizomorphs, rooted, missing
STALK-SURFACE-ABOVE-RING	fibrous, scaly, silky, smooth
STALK-SURFACE-BELOW-RING	fibrous, scaly, silky, smooth
STALK-COLOR-ABOVE-RING	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
STALK-COLOR-BELOW-RING	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
VEIL-TYPE	partial, universal
VEIL-COLOR	brown, orange, white, yellow

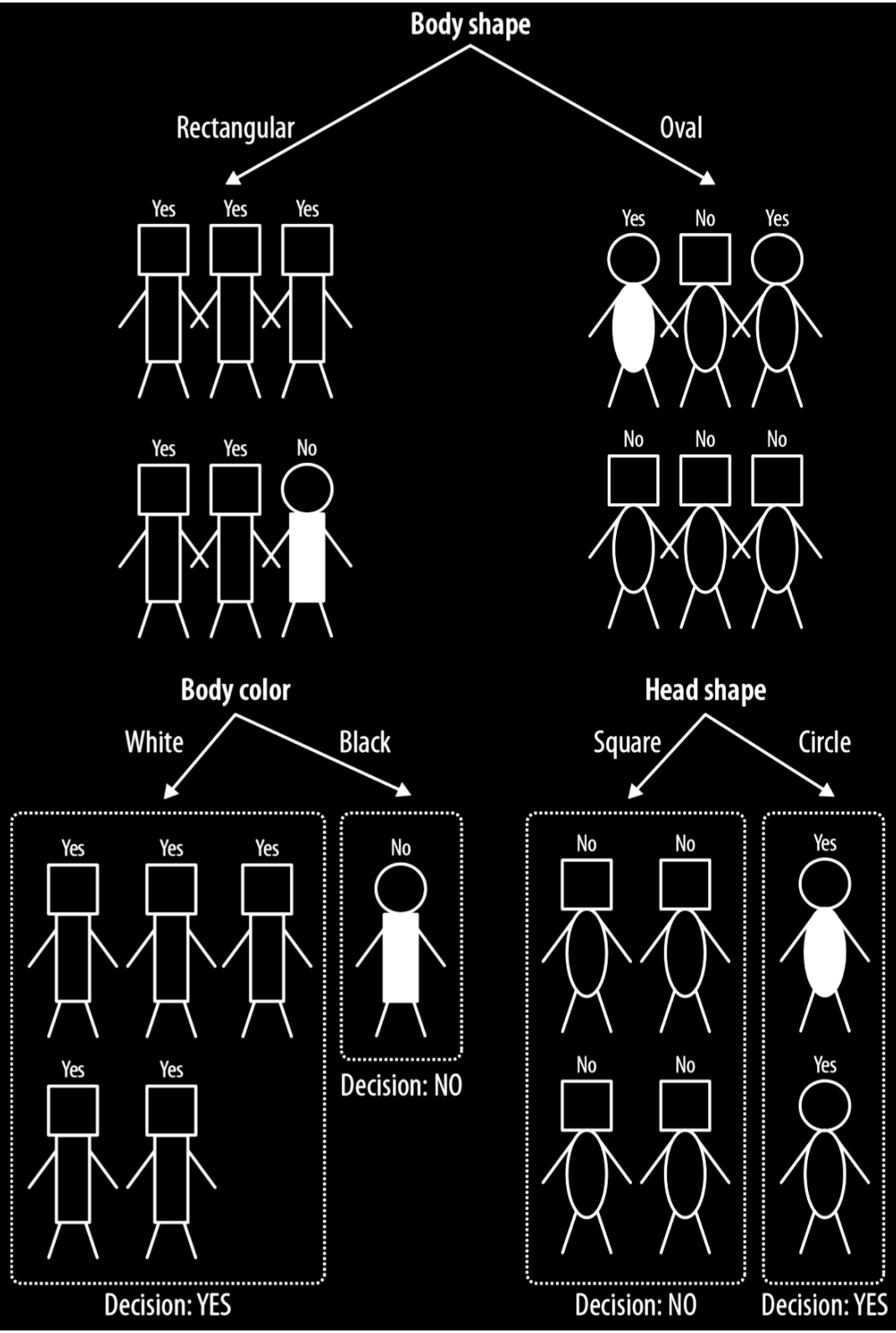

Default entropy for mushroom
Entropy
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0

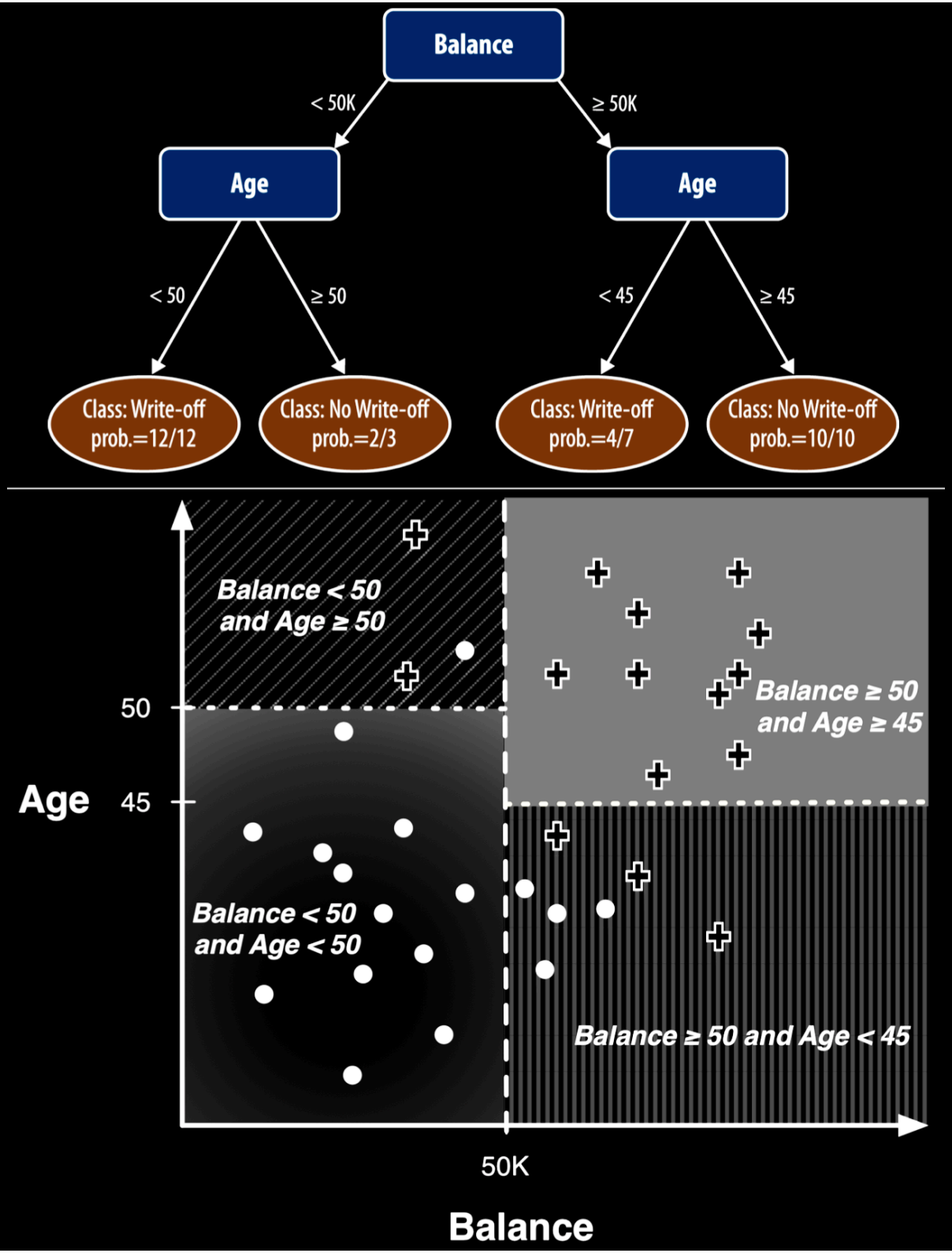

Entropy and prevalence of values for GILL-COLOR
Entropy
Proportion
y
u
n
k
w
g
h
p


Entropy and prevalence of values for SPORE-PRINT-COLOR
Entropy
Proportion
h
u
n
k
w


Entropy and prevalence of values for ODOR
Entropy
Proportion
a
c
f
m
l
p
n

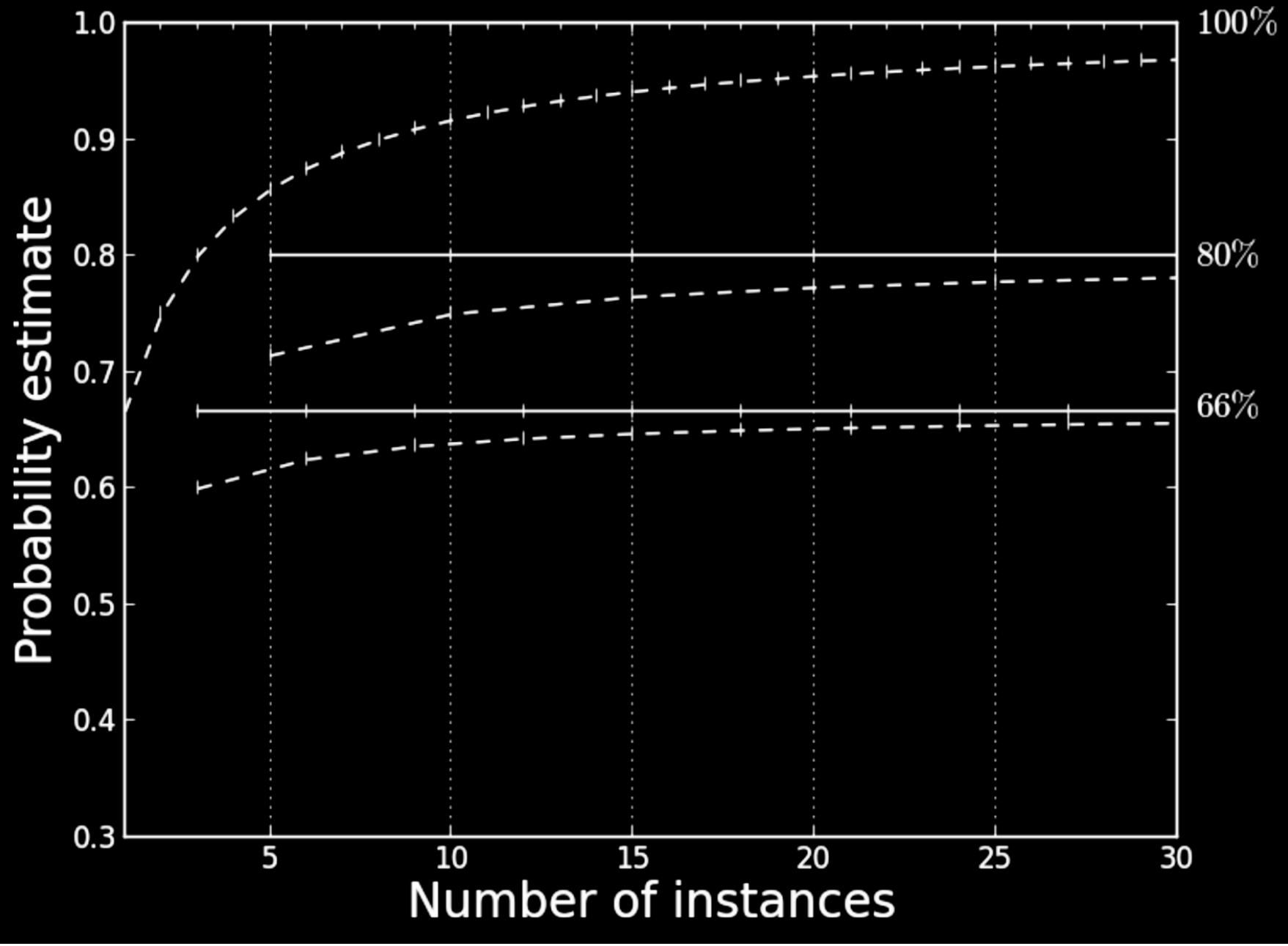

Employed
Root node
Yes
No
Class:
Not Write-off
Balance
Interior nodes
< 50K
≥ 50K
Class:
Not Write-off
Age
< 45
≥ 45
Leaf node
Class:
Not Write-off
Class: Write-off


Body shape
Rectangular
Oval
Yes Yes Yes
Yes No Yes
Yes Yes No
No No No
Body color
White
Black
Yes Yes Yes
Yes Yes
No
Decision: NO
Decision: YES
Head shape
Square
Circle
No No
No No
Yes
Yes
Decision: NO
Decision: YES

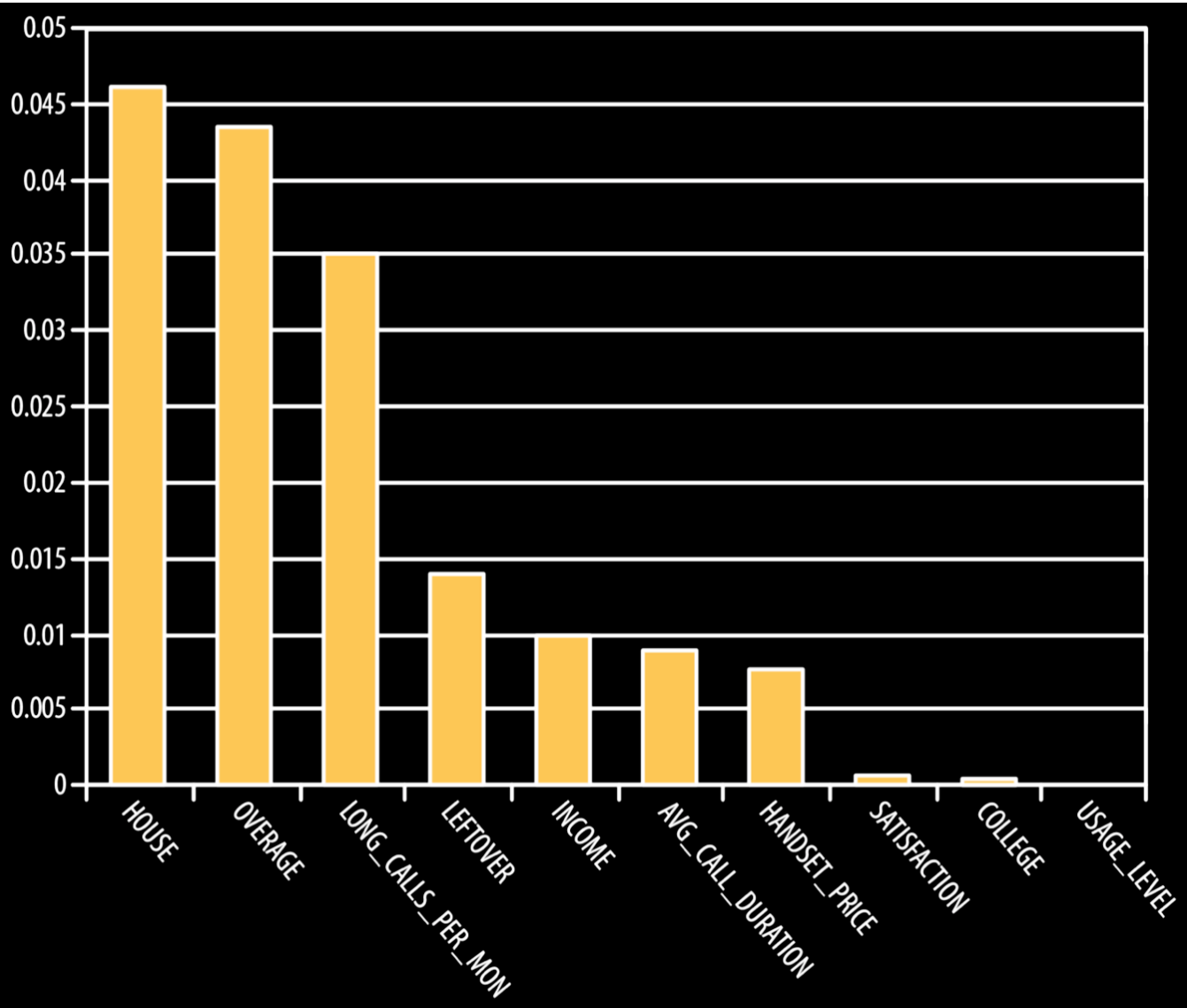

Balance
< 50K
≥ 50K
Age
Age
< 50
≥ 50
< 45
≥ 45
Class: Write-off
prob.=12/12
Class: No Write-off
prob.=2/3
Class: Write-off
prob.=4/7
Class: No Write-off
prob.=10/10
Balance < 50
and Age ≥ 50
Balance ≥ 50
and Age ≥ 45
Balance < 50
and Age < 50
Balance ≥ 50 and Age < 45
50
45
Age
50K
Balance

```
IF (Balance < 50K) AND (Age < 50) THEN Class=Write-off
IF (Balance < 50K) AND (Age ≥ 50) THEN Class=No Write-off
IF (Balance ≥ 50K) AND (Age < 45) THEN Class=Write-off
IF (Balance ≥ 50K) AND (Age < 45) THEN Class=No Write-off
```

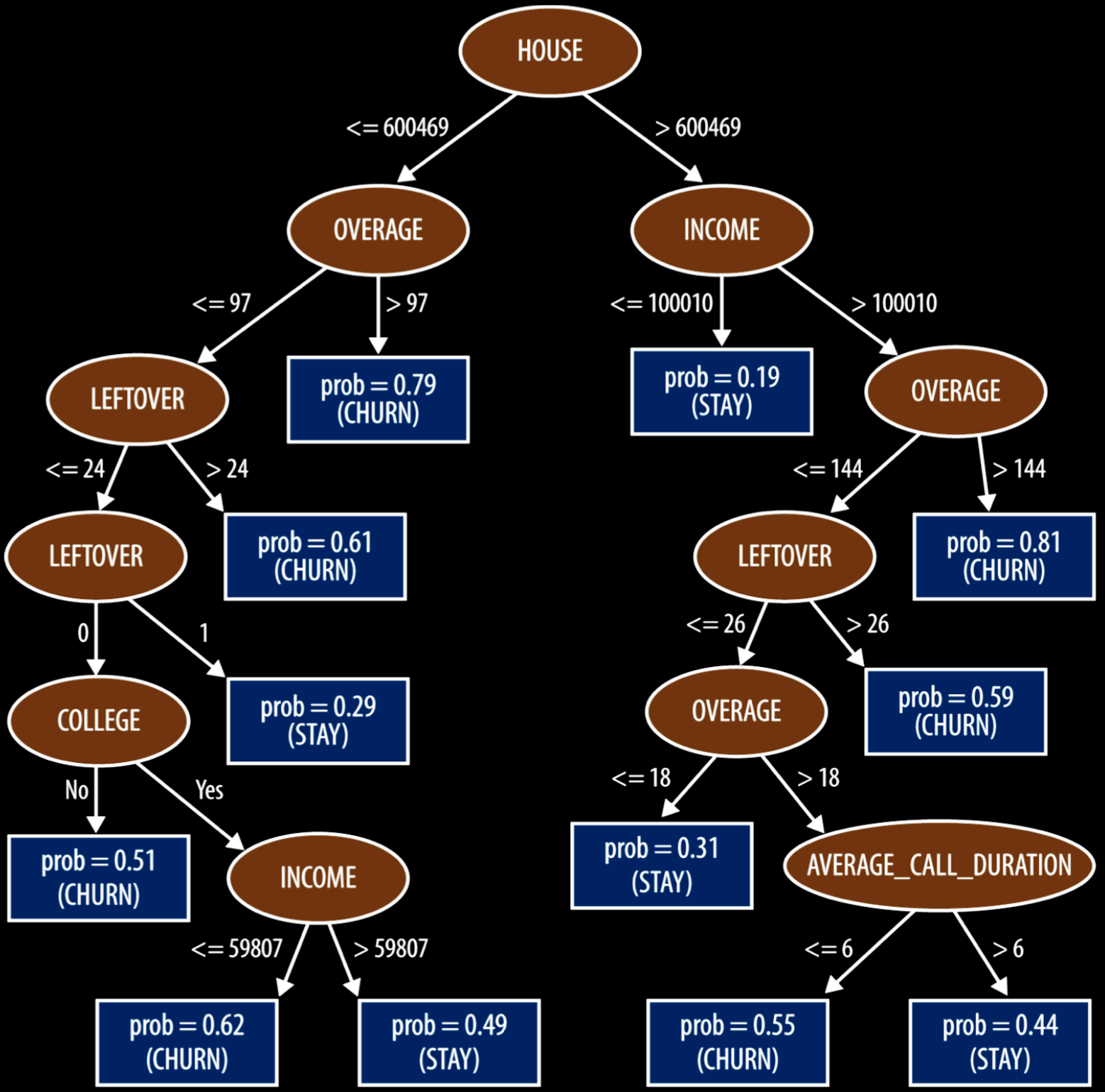
$$p(c) = \frac{n+1}{n+m+2}$$



Variable	Explanation
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
LONG_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE *(Target variable)*	Did the customer stay or leave (churn)?



Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.000	COLLEGE
10	0.000	USAGE_LEVEL


HOUSE
<= 600469
> 600469
OVERAGE
INCOME
<= 97
> 97
<= 100010
> 100010
LEFTOVER
prob = 0.79 (CHURN)
prob = 0.19 (STAY)
OVERAGE
<= 24
> 24
<= 144
> 144
LEFTOVER
prob = 0.61 (CHURN)
LEFTOVER
prob = 0.81 (CHURN)
0
1
<= 26
> 26
COLLEGE
prob = 0.29 (STAY)
OVERAGE
prob = 0.59 (CHURN)
No
Yes
<= 18
> 18
prob = 0.51 (CHURN)
INCOME
prob = 0.31 (STAY)
AVERAGE_CALL_DURATION
<= 59807
> 59807
<= 6
> 6
prob = 0.62 (CHURN)
prob = 0.49 (STAY)
prob = 0.55 (CHURN)
prob = 0.44 (STAY)

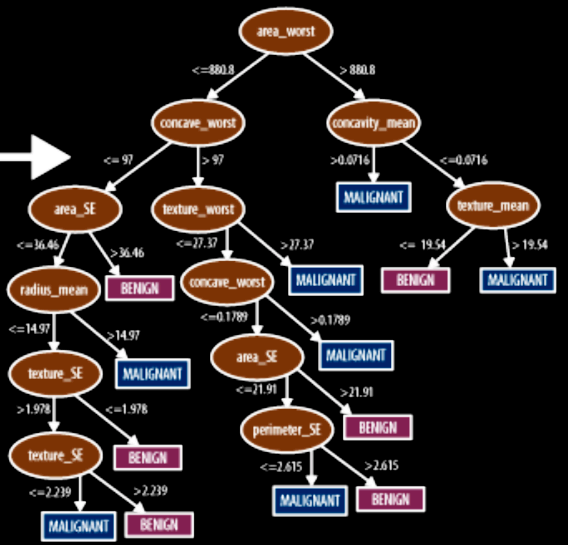
Historical Data

x	y	z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Training data have all values specified

Data mining

Model



Model is deployed

Mining

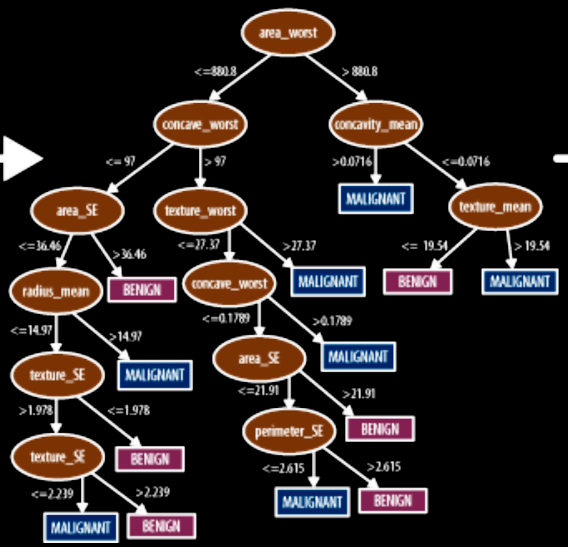
Use

New data item

x	y	z	class
30	false	Red	?

New data item has class value unknown (e.g. will customer accept?)

Model



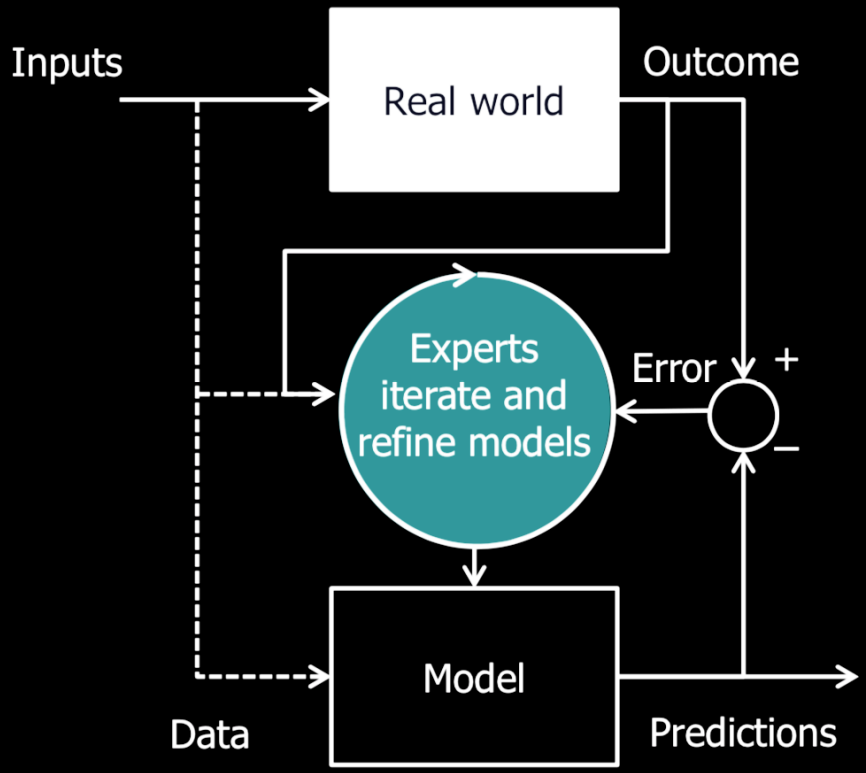
Class: accepted, Probability: 0.88

AUTOMATING DATA SCIENCE EXAMPLE

- DARPA Data Driven Discovery of Models (D3M)

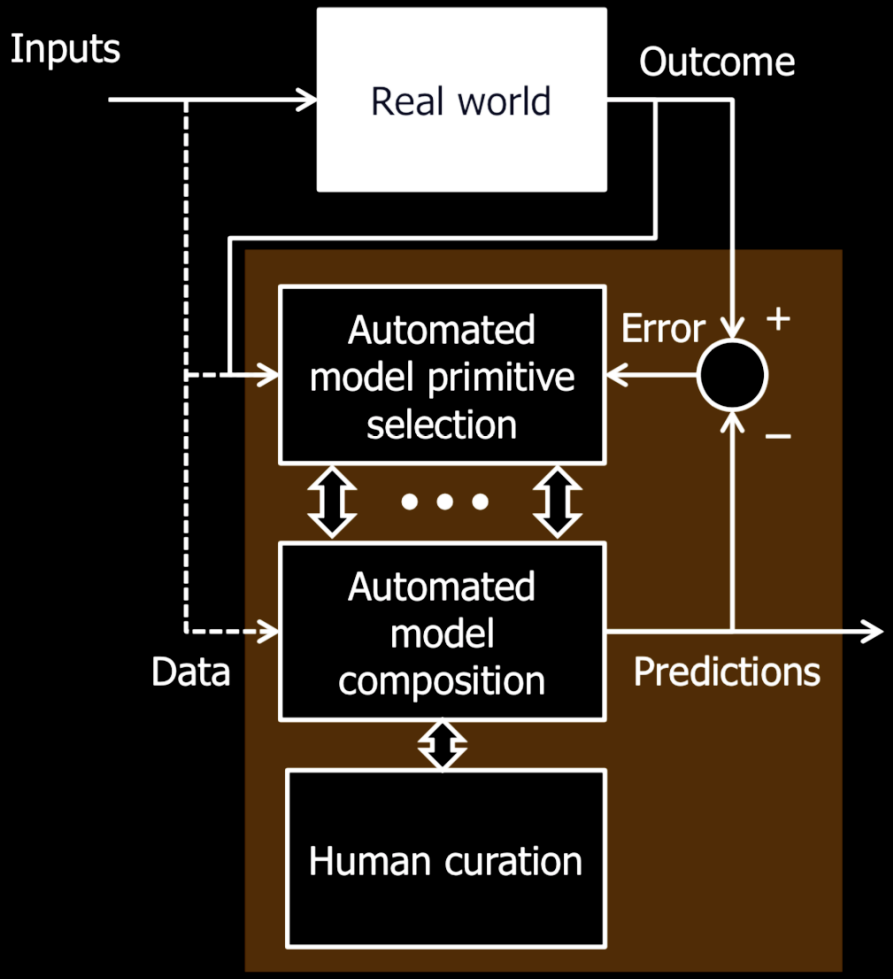
D³M: Data-driven discovery of models

Today: Manual



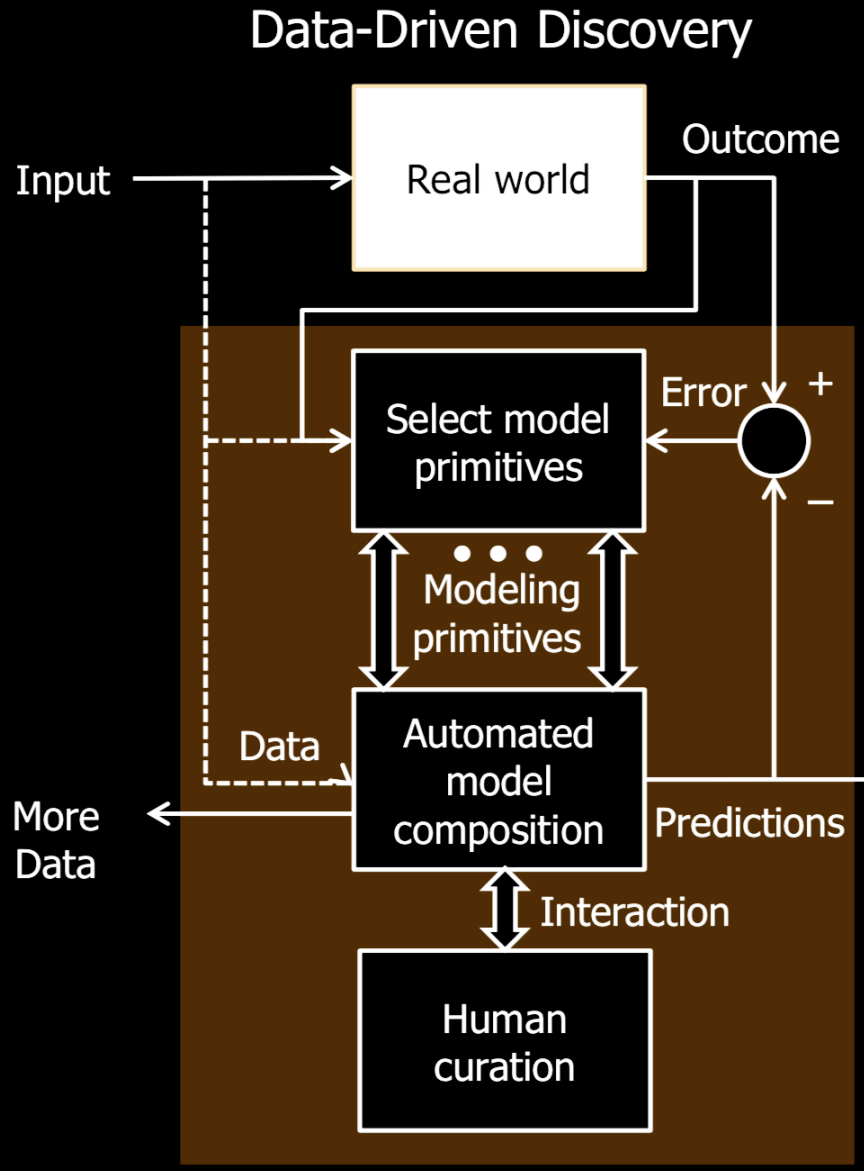
- Model: representation of a real-world system
  - Examples
    - Inferring locations of images
    - Prediction of election outcomes
    - Estimation model for disease outbreaks
- Manual process: 10-1000s of person-years
- Teams of experts required to develop the model

Tomorrow: Automated



- Automatically select problem-specific model primitives
  - Extend the library of modeling primitives
- Automatically compose complex models from primitives
- Facilitate user interaction with composed models

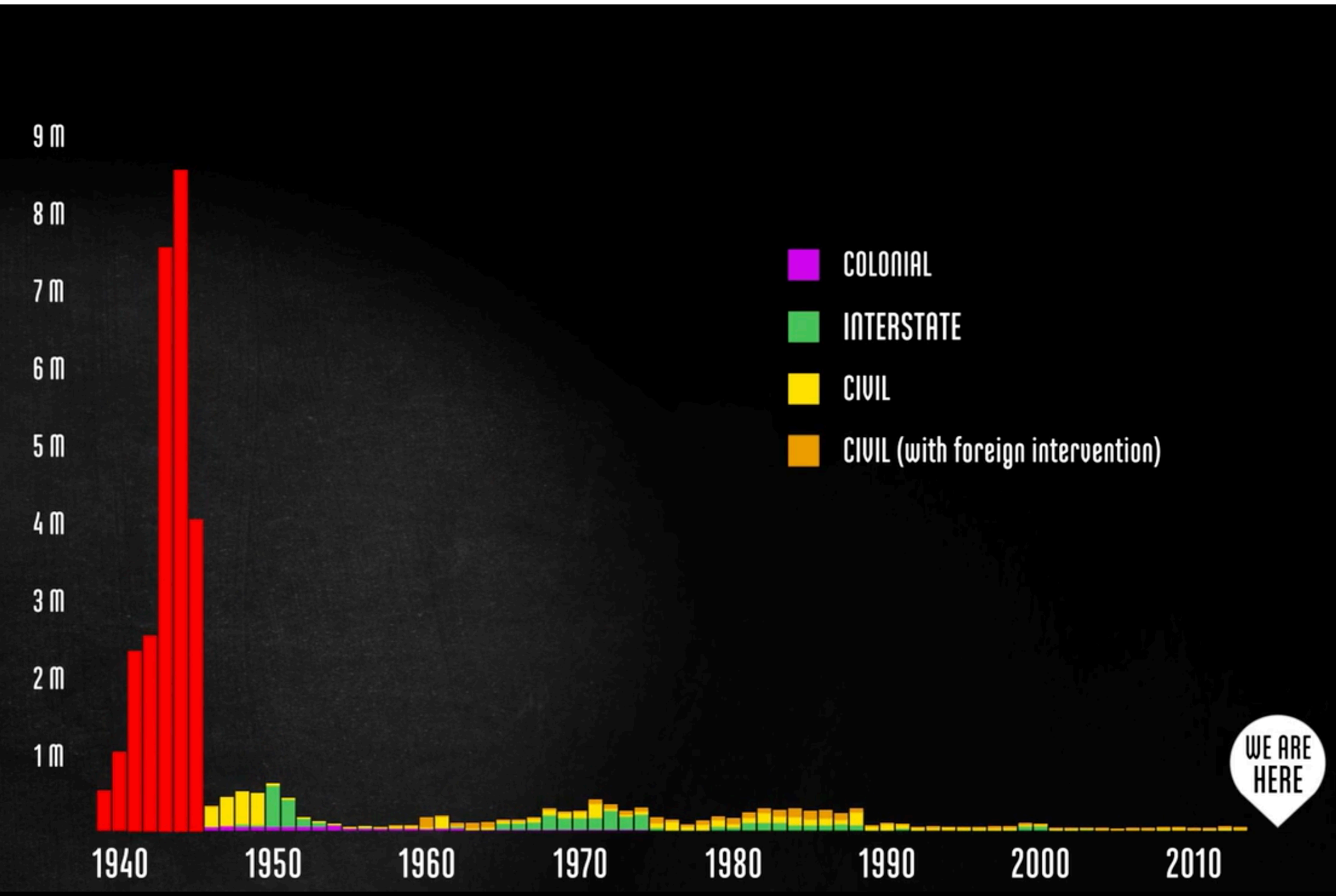
Automated discovery of complex models with non-expert curation



- TA1: Discover and develop model primitives
  - Create a "vocabulary" of modeling primitives
  - Make primitives automatically discoverable
- TA2: Automatically compose complex models
  - Mine corpora of complex models to learn the "syntax" of primitive composition
  - Find optimal compositions
  - Predict additional data requirements
- TA3: Curation of models by non-experts
  - Decompose questions
  - Explain data and models to enable selection and editing

DATA VISUALIZATION EXAMPLE

- The Fallen, www.fallen.io/ww2, Niel Halloran


9 M
8 M
7 M
6 M
5 M
4 M
3 M
2 M
1 M
COLONIAL
INTERSTATE
CIVIL
CIVIL (with foreign intervention)
WE ARE HERE
1940
1950
1960
1970
1980
1990
2000
2010