Week 1
Data Scientist Mindset

Introduction

Adrian Spataru

Data Scientist at Know-Center

Bohdan Andrusyak

Data Scientist at Kleine Zeitung

Goal

- Understand how to start and finish data science project
- Have a finished Project for your Portfolio
- Have the skillset of a junior data scientist

Agenda

- What are the elements of data science?
- Essential Statistics & Probability Refreshers
- The importance of data in Data Science


Aha
Slides

Mat Velloso
@matvelloso

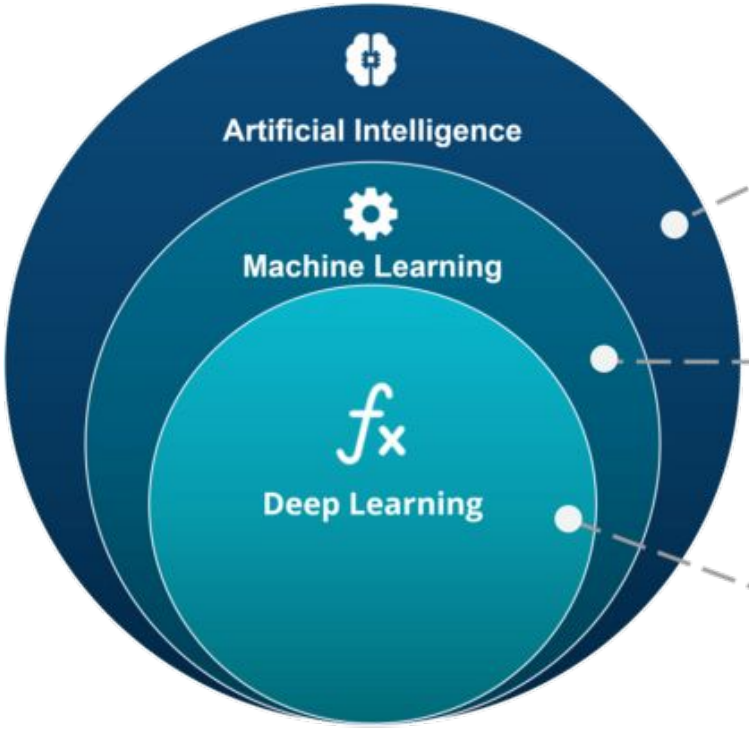
Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

22/11/18, 5:25 PM

3,514 Retweets **10.8K** Likes



ARTIFICIAL INTELLIGENCE

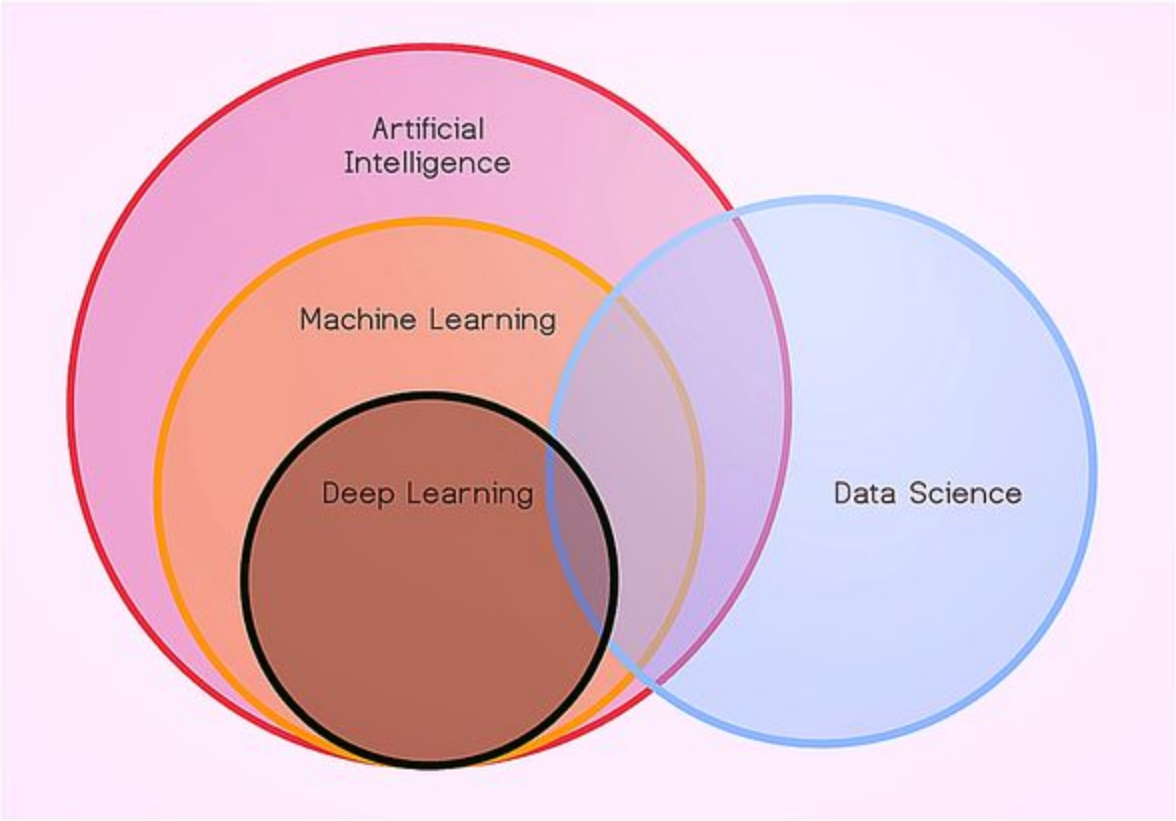
A technique which enables machines to mimic human behaviour

MACHINE LEARNING

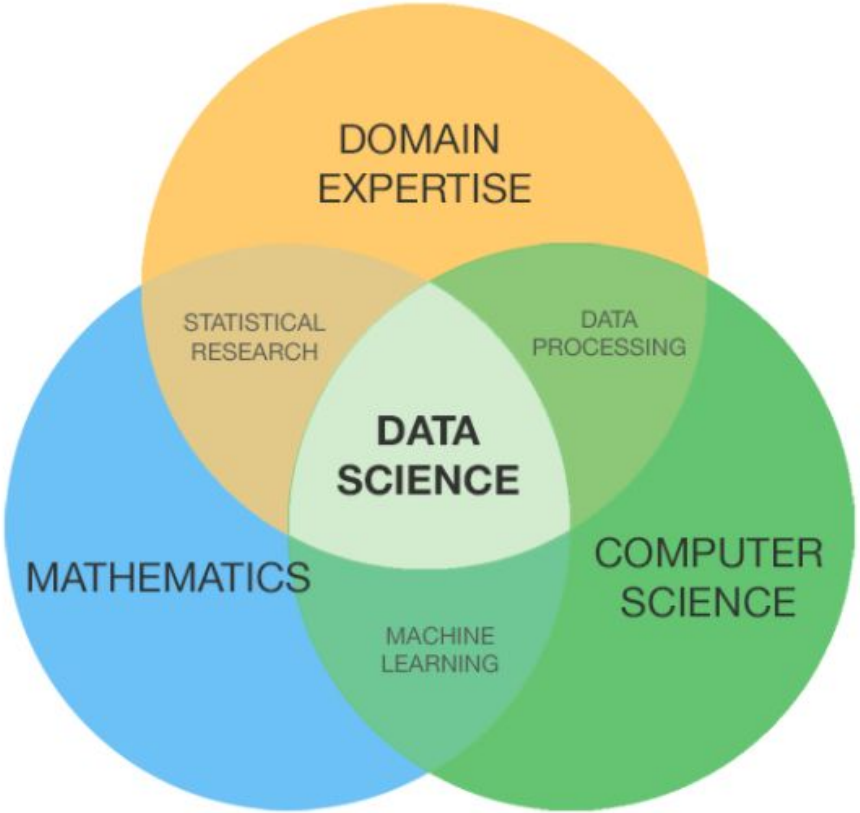
Subset of AI technique which use statistical methods to enable machines to improve with experience

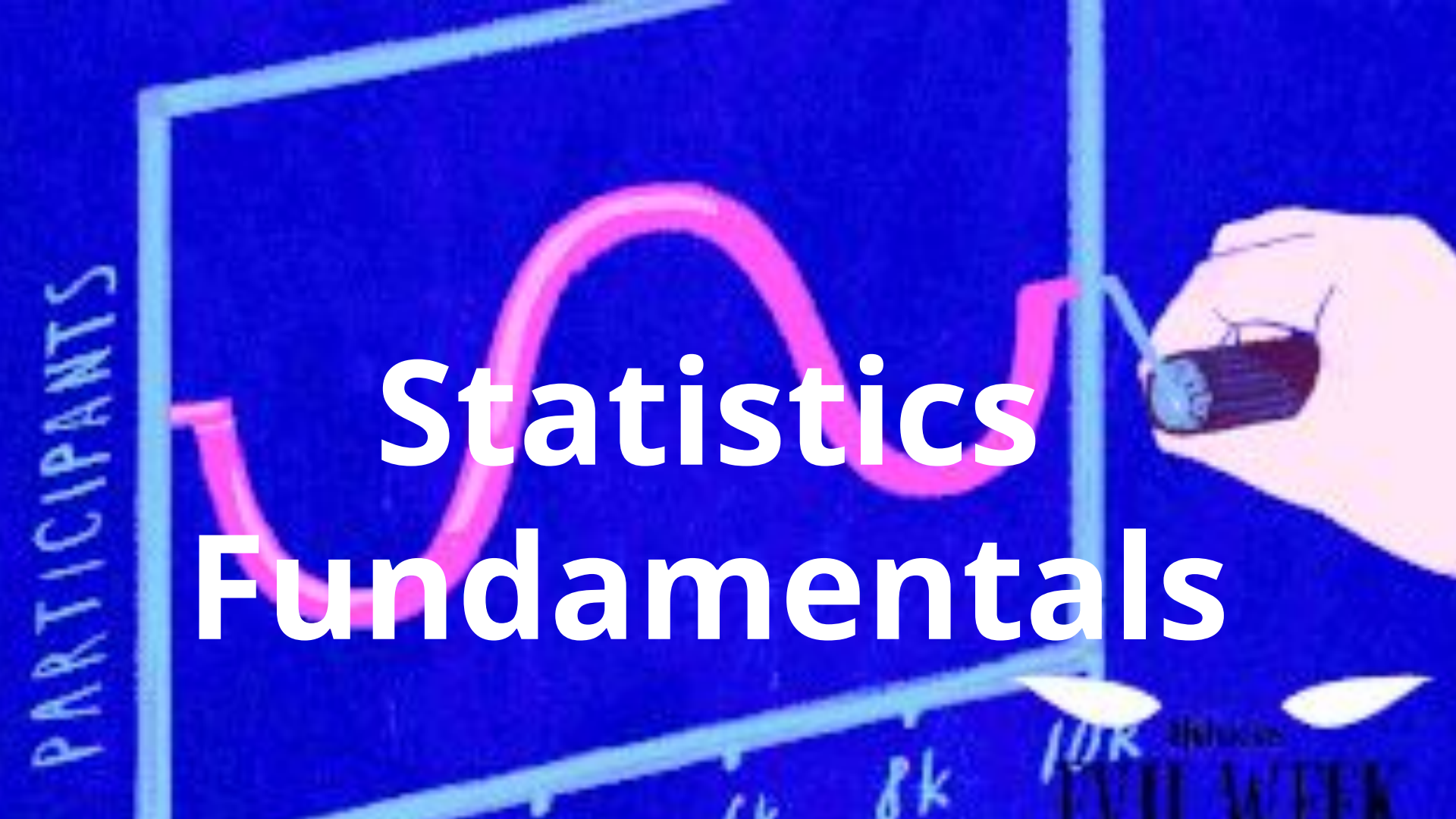
DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible


Artificial
Intelligence
Machine Learning
Deep Learning
Data Science

What has statistics to do with Data Science?





Statistics Fundamentals

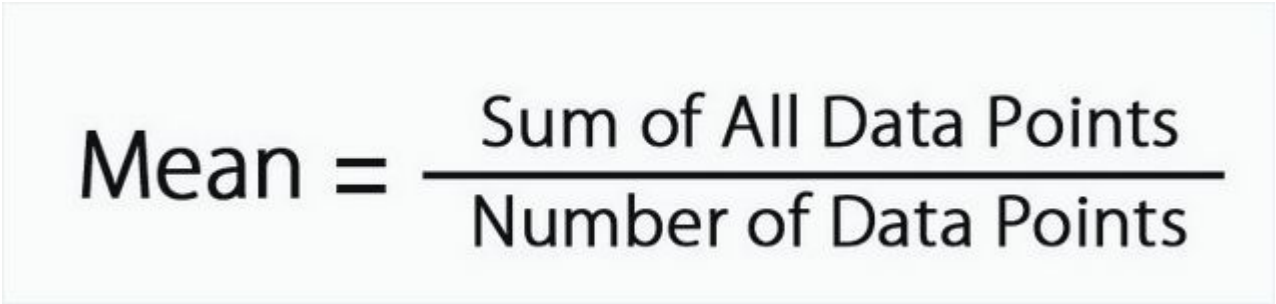
Your Grades:

3,3,1,5

Mean: (3+3+1+5)/4 = 3

Your Grades:
3,3,1,5

Mean: (3+3+1+5)/4 = 3



Your Grades:
1,1,2,2,3

Your Grades:

1,1,2,2,3

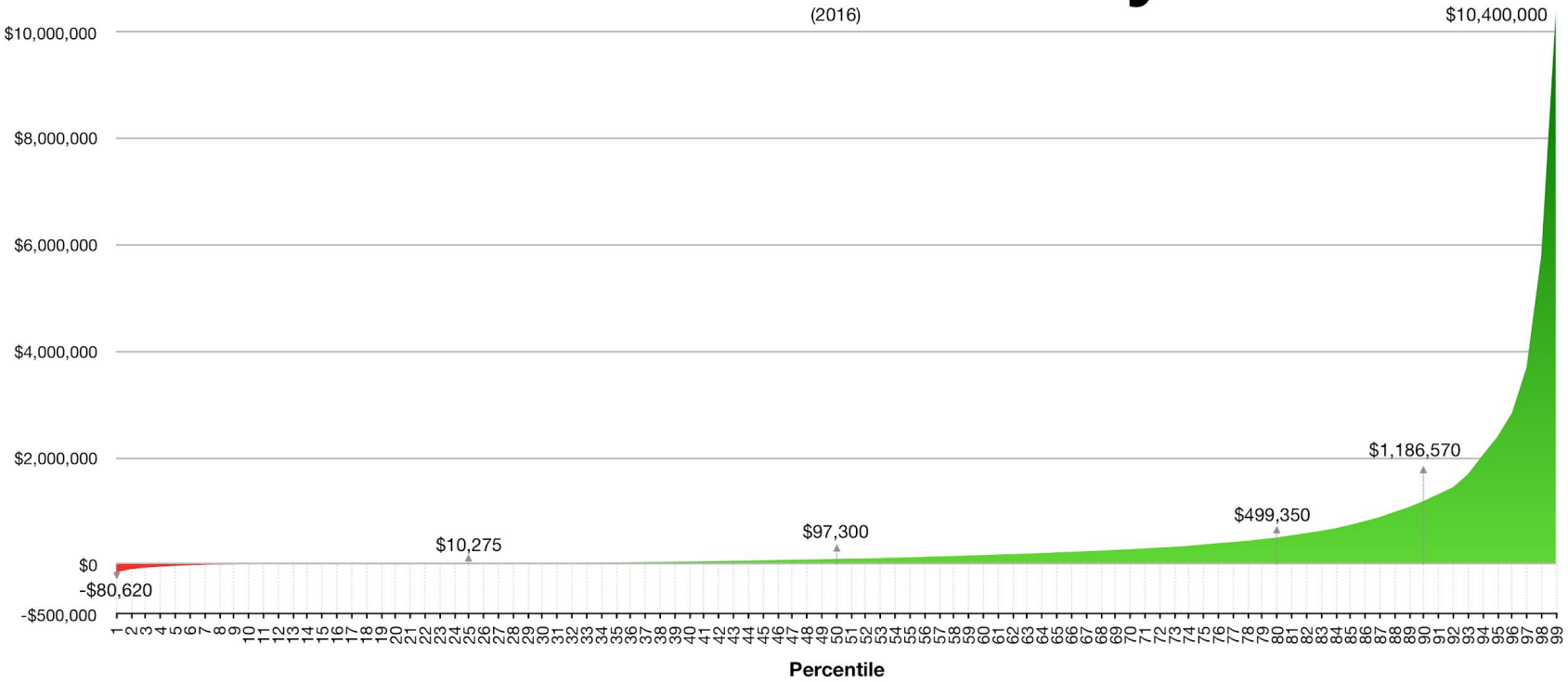
Median = 2

Your Grades:

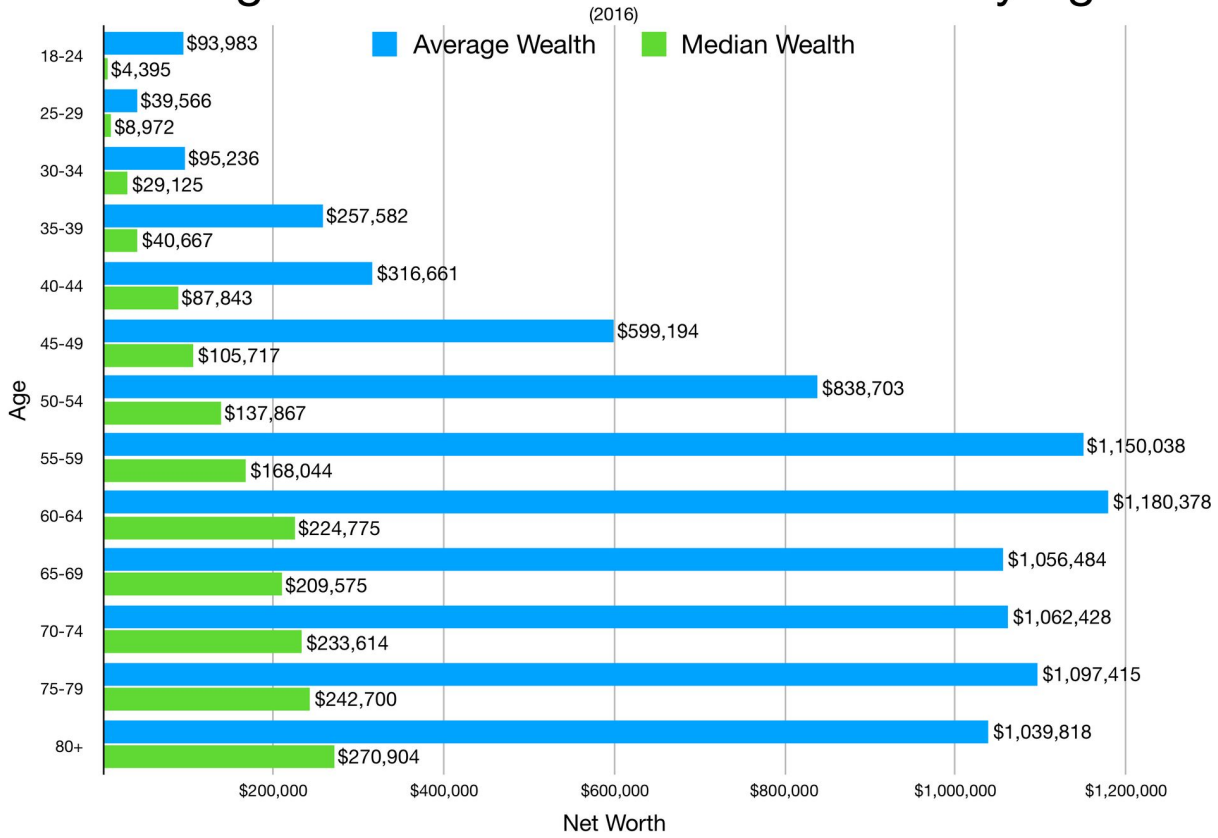
1,1,2,3,4,4

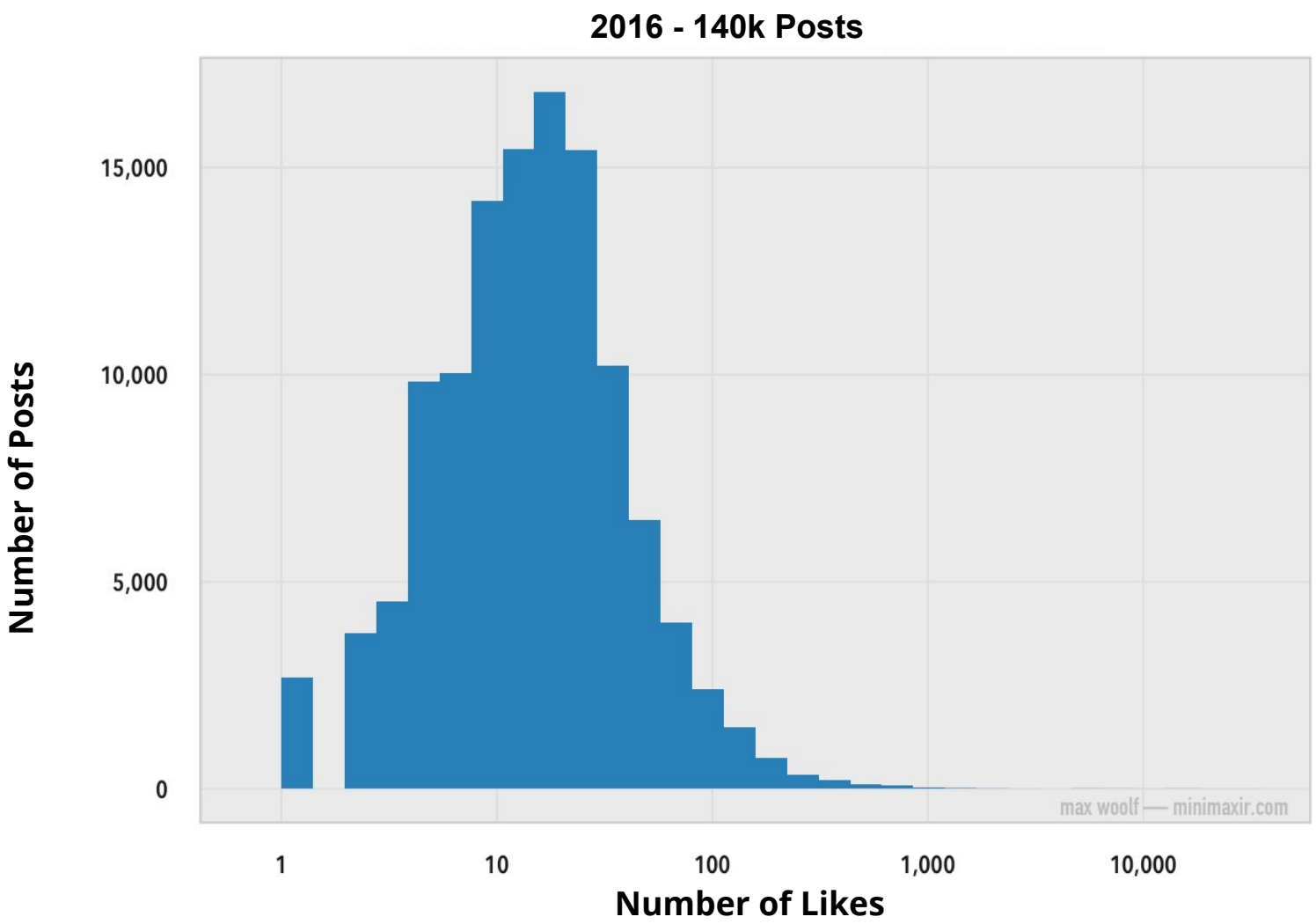
Median = (2+3)/2 = 2.5

Distribution of Family Wealth



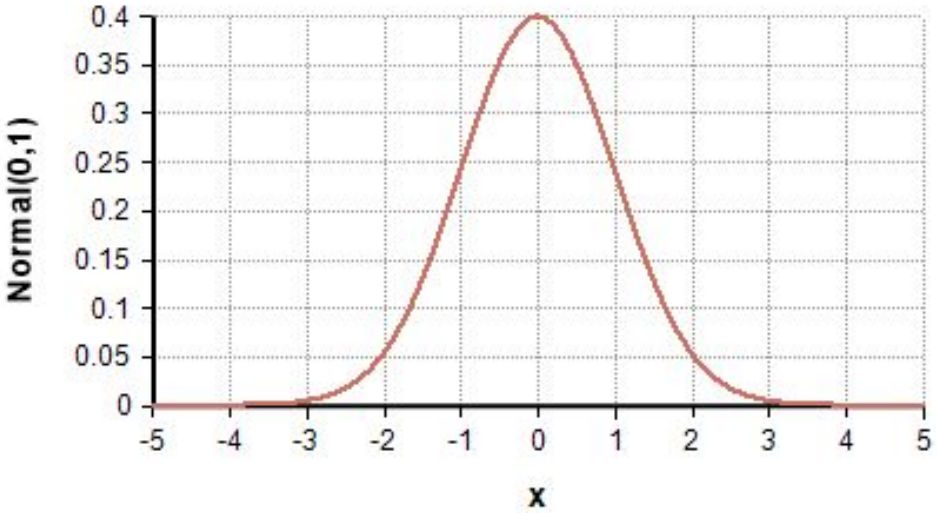
Average & Median Household Wealth by Age



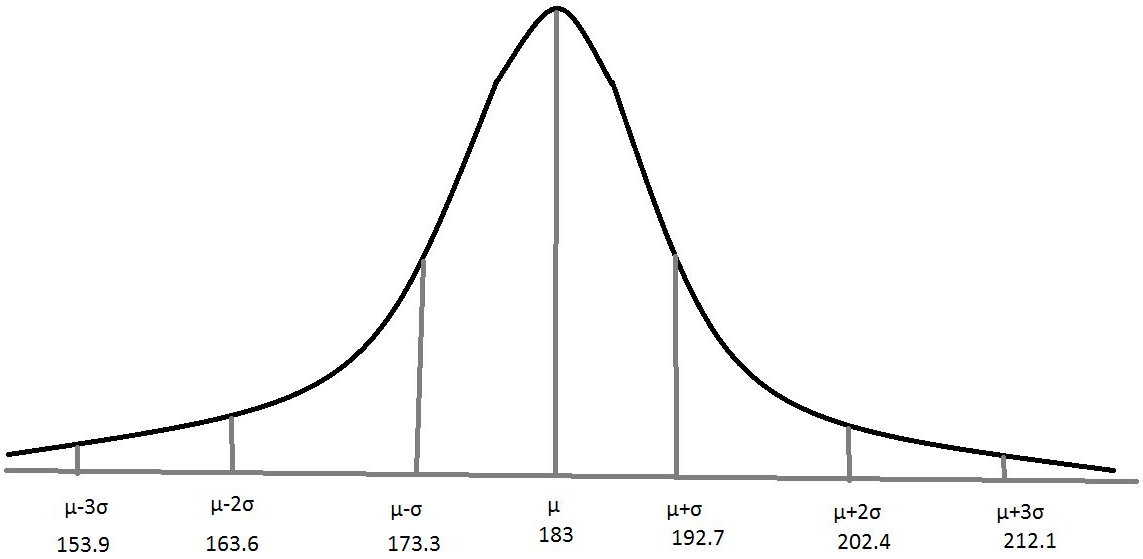

2016 - 140k Posts
Number of Posts
15,000
10,000
5,000
0
max woolf — minimaxir.com
1
10
100
1,000
10,000
Number of Likes

Median: 23 Likes
Mean: 325 Likes

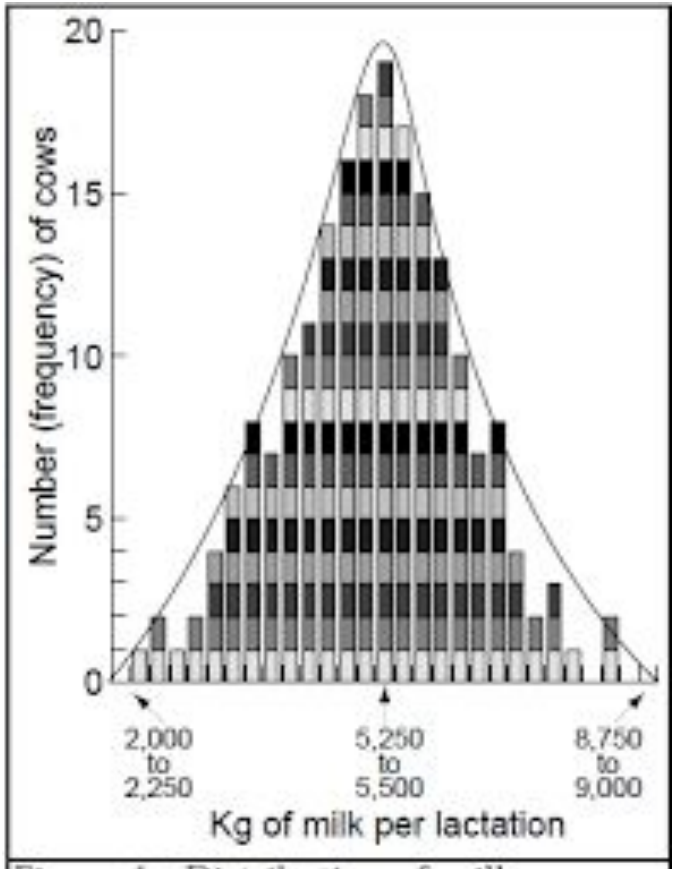
Gaussian Distribution



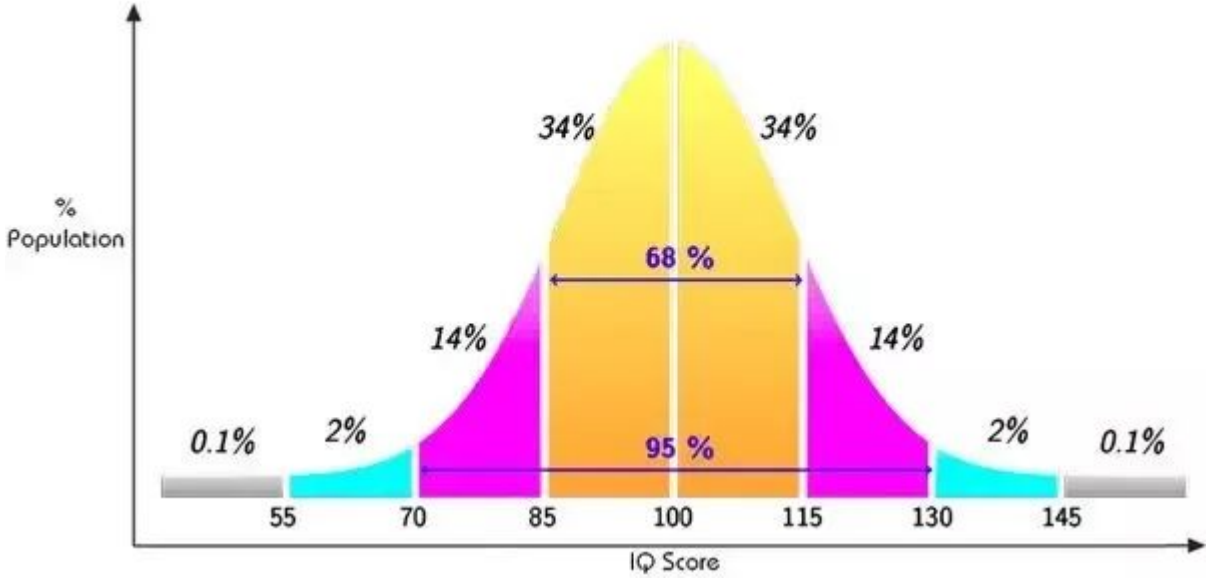
Height Distribution among Men



Volume of Milk Production From Cows



IQ Distribution

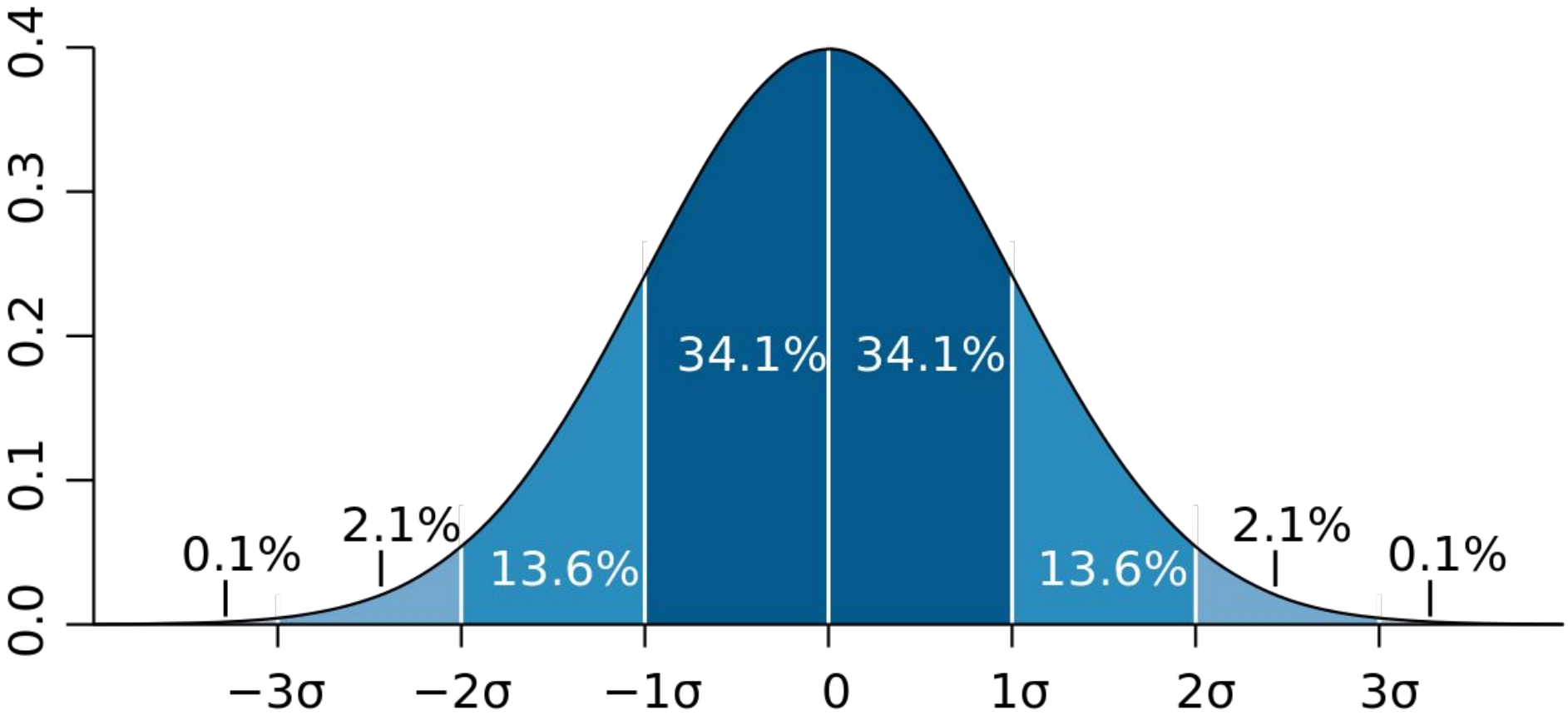


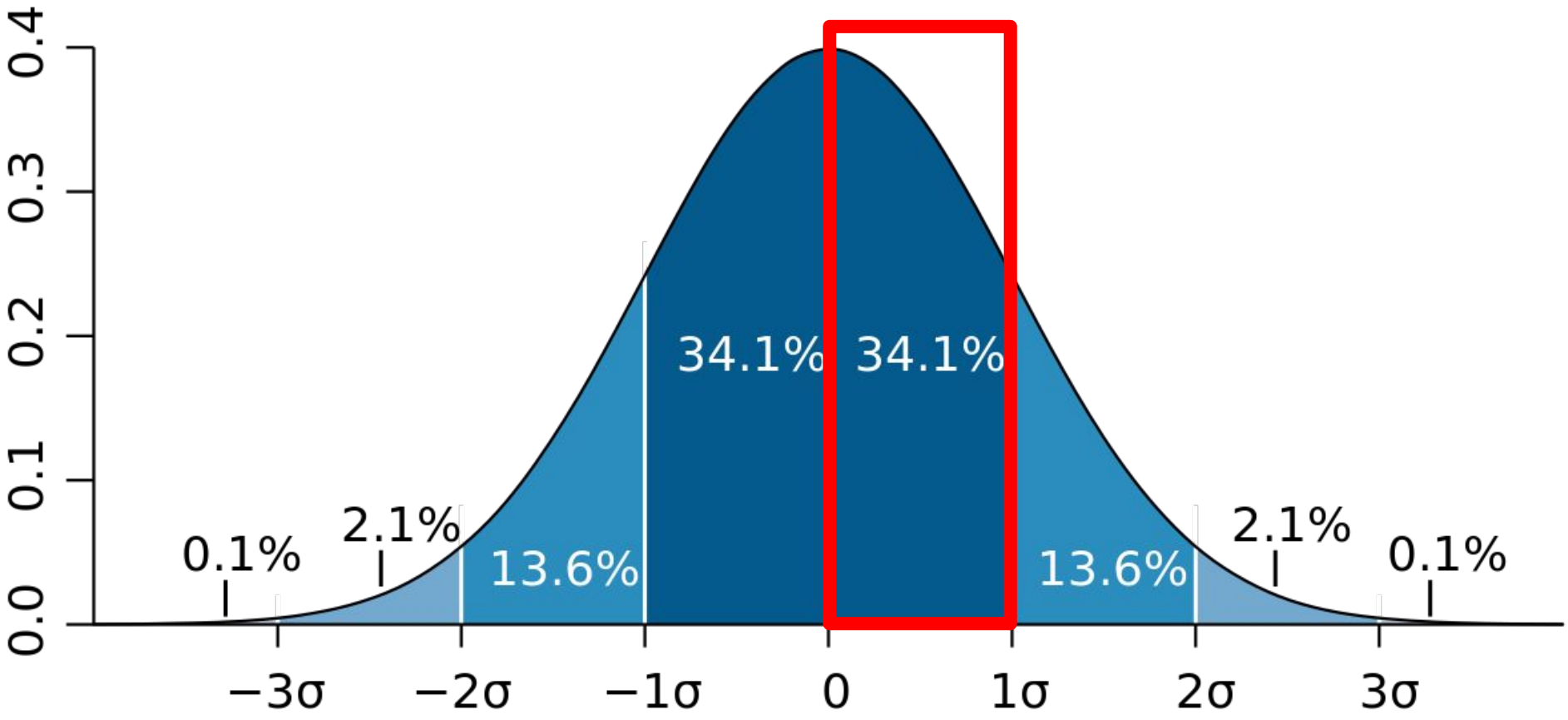
Standard Deviation

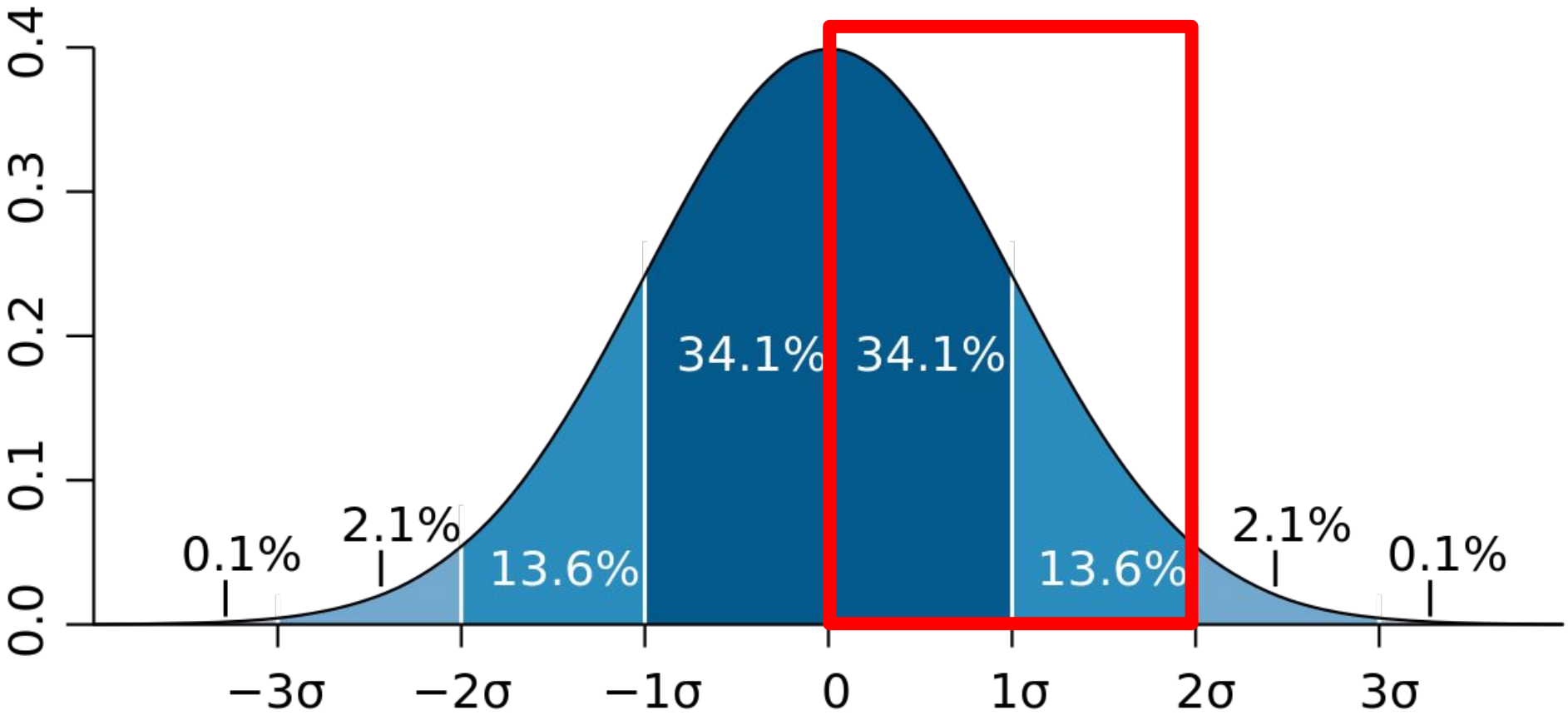
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Mean

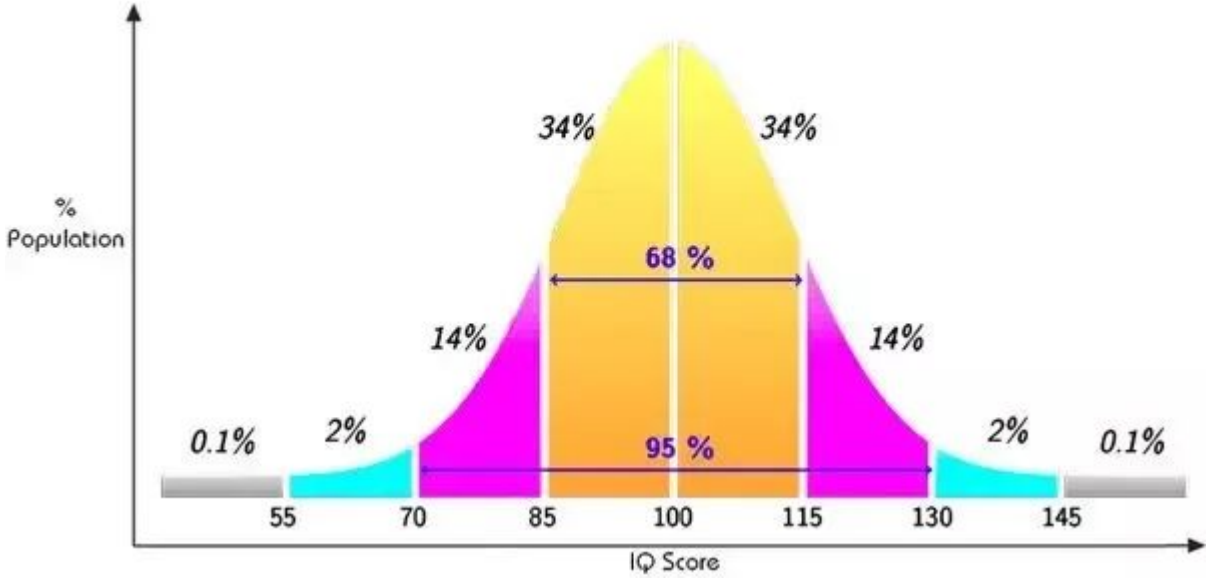
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$


0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ


0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ


0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

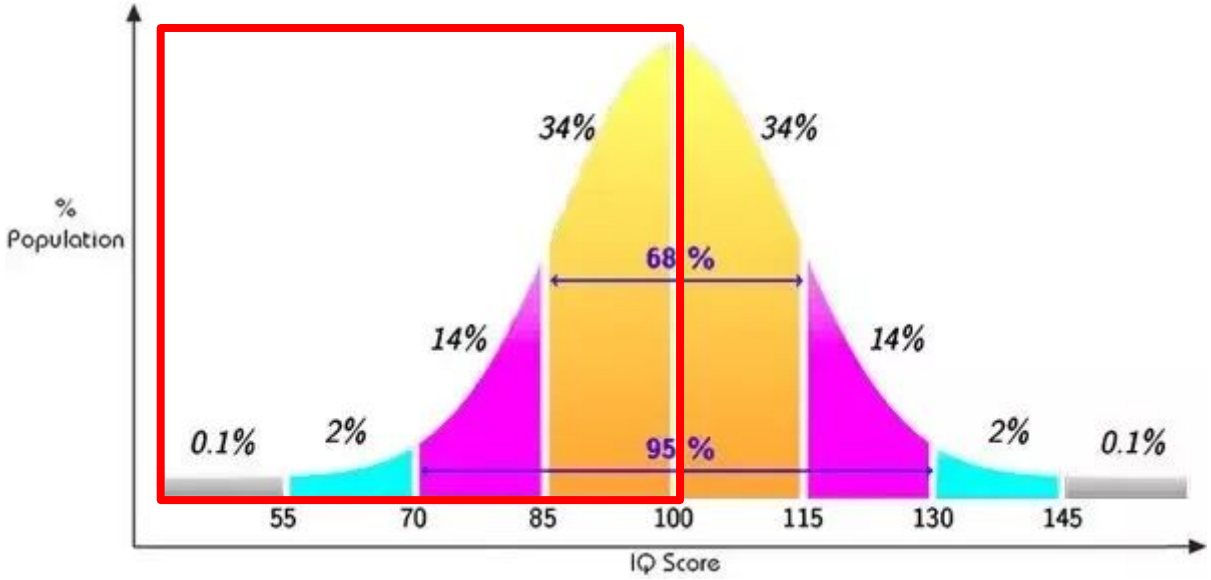
IQ Distribution



How smart are you if your IQ is 100?

How smart are you if your IQ is 100? Your IQ is average, therefore you are smarter than 50% of the people.

How smart are you if your IQ is 100? Your IQ is average, therefore you are smarter than 50% of the people.

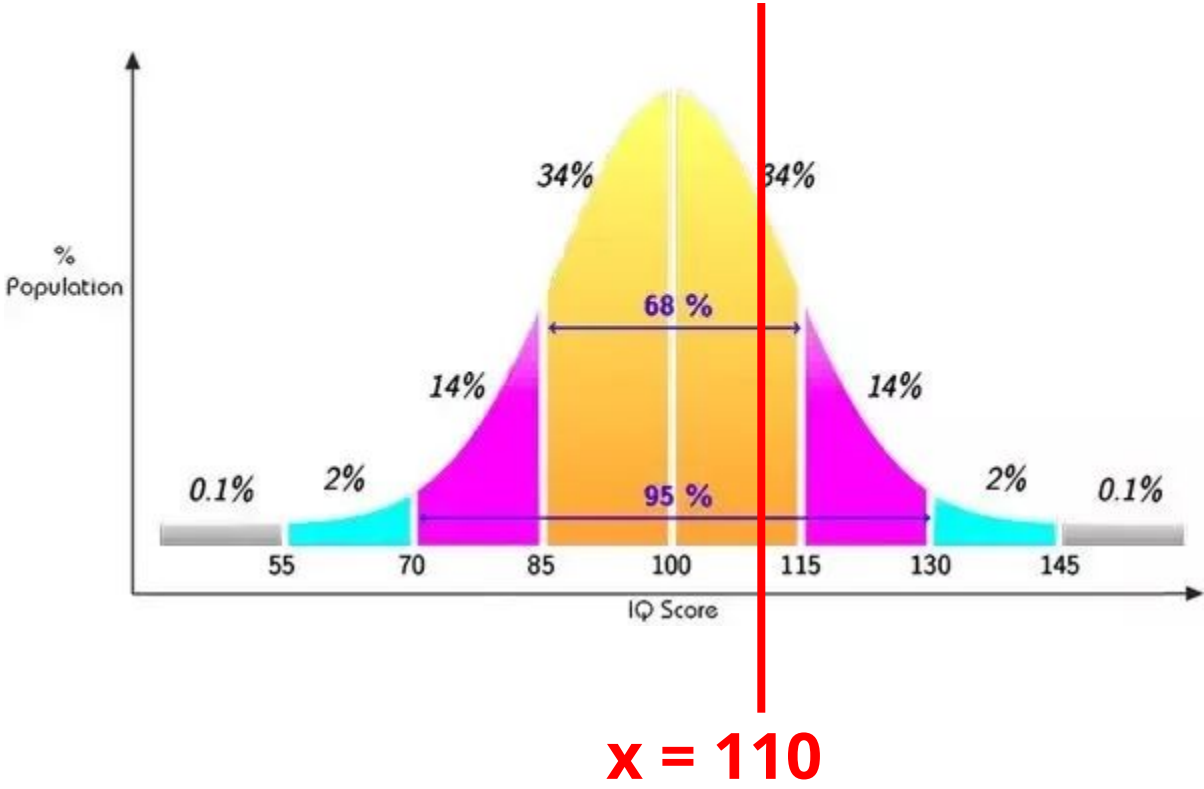


How smart are you if your IQ is 110?

How smart are you if your IQ is 110?
Z-Score!

Z-Score

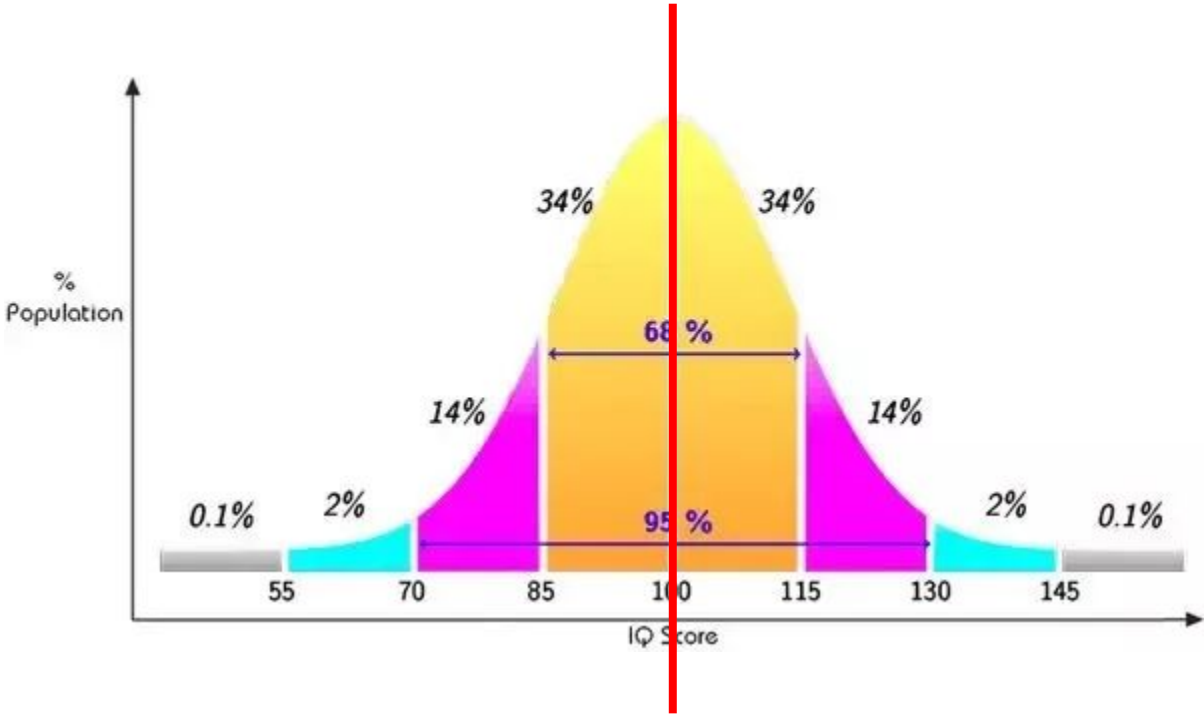
$$z = \frac{x - \mu}{\sigma}$$



Z-Score

$$z = \frac{x - \mu}{\sigma}$$

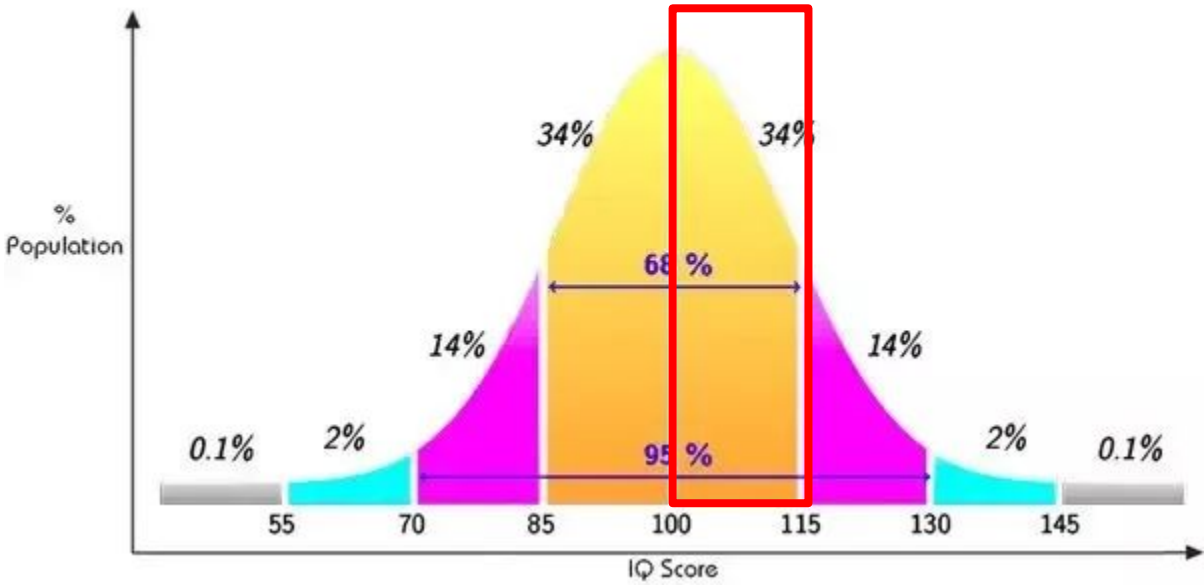
x = 110



μ = 100

Z-Score

$$z = \frac{x - \mu}{\sigma}$$



x = 110

μ = 100

σ = 15

Z-Score

$$z = \frac{x - \mu}{\sigma}$$

x = 110

μ **= 100**

σ **= 15**

Z-Score

$$z = \frac{x - \mu}{\sigma}$$

$$z = (110 - 100)/15 = 6.66$$

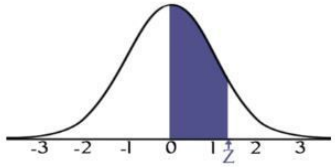
x = 110

μ **= 100**

σ **= 15**

Z-Score table

$$z = 6.66$$



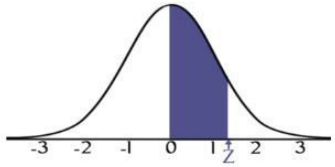
STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

Z-Score table

$$z = 6.66$$



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for *z* = 1.25 the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	[illegible]	0.2157	0.2190	0.2224
0.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	[illegible]	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

Z-Score table

0.2454

Z-Score table

0.2454 - What does that mean?

Z-Score table

0.2454 - What does that mean?

If you have a IQ of 110,
then you are 24.54% smarter than the average person (IQ 100).

Z-Score table

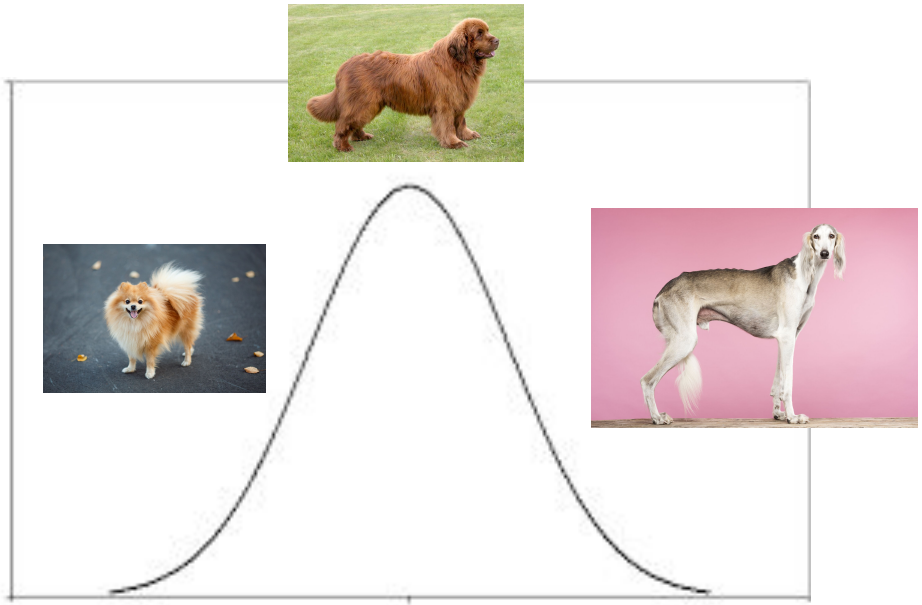
0.2454 - What does that mean?

If you have a IQ of 110,
then you are 24.54% smarter than the average person (IQ 100).

You are smarter than 74.54% of the people.
(0.5000+0.2454 = 0.7454)

Height Distribution of all dog breeds

$$\mu = 75cm$$
$$\sigma = 25cm$$



23 CM - Average Height

Great Dane

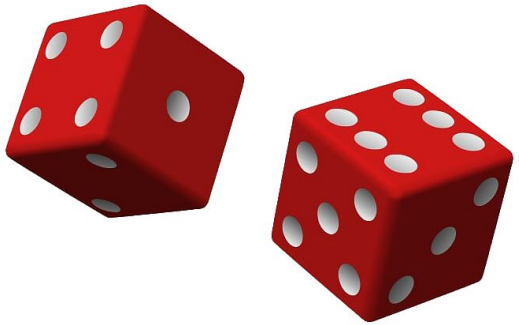
81CM - Average Height

62CM - Average Height

Probability

- Probability is a numerical description of how likely an event is to occur.

- Probability is a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.



Probability Formula

$$P(A) = \frac{x}{n}$$

P(A) - probability of event A occurring

x - number of favourable outcomes

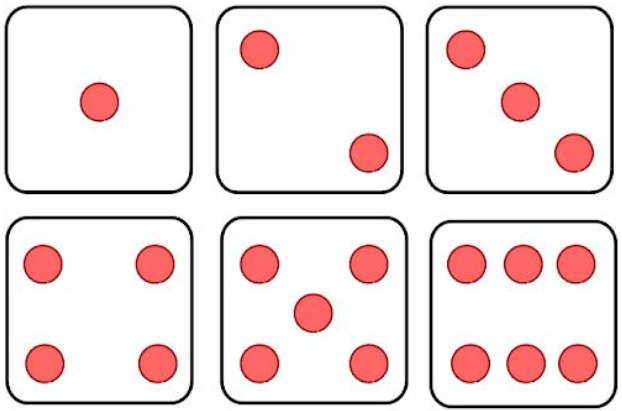
n - number of total outcomes

Coin Flip Example

$$P(Heads) = \frac{1}{2}$$

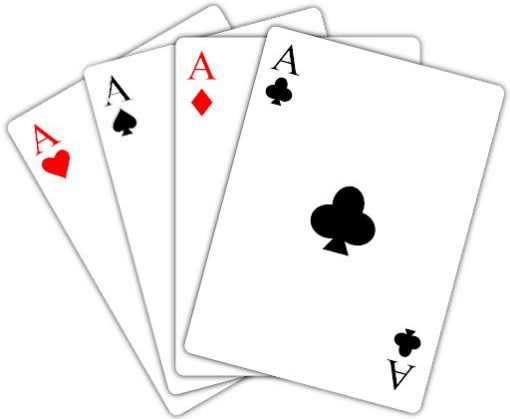
Dice Example

$$P(EvenNumber) = \frac{3}{6}$$



Cards Example

$$P(Ace) = \frac{4}{52}$$



quiz: probability number card

quiz: prob. of a certain suit

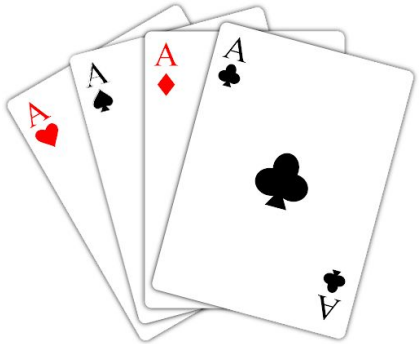
The Addition Rule

Probability of two events that are NOT mutually exclusive

$$P(A or B) = P(A) + P(B) - P(A and B)$$

Cards Example

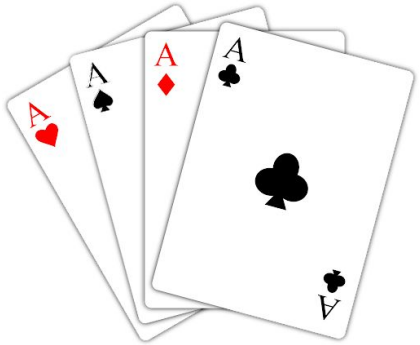
$$P(AceOrRed) = P(Ace) + P(Red) - P(AceAndRed)$$



Cards Example

$$P(AceOrRed) = P(Ace) + P(Red) - P(AceAndRed)$$

$$P(AceOrRed) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$



Independent Events

- An independent event is an event that has no connection to another event's chances of happening.
- The event has no effect on the probability of another event occurring.
- Examples:
  - Flipping multiple coins
  - Rolling multiple dices
  - Seeing black cat and failing your exam

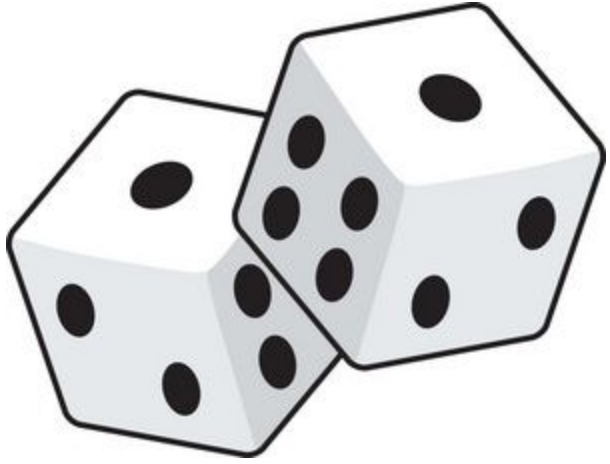
The Multiplication Rule

Probability of two independent event occurring together

$$P(AandB) = P(A)P(B)$$

Snake Eyes Example

$$P(Roll1\_Dice1 and Roll1\_Dice2) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$



Dependent Events

- When two events are dependent events, one event influences the probability of another event
- A dependent event is an event that relies on another event to happen first
- Examples:
  - Learning data science and getting well-paid job
  - Breaking quarantine and getting COVID-19
  - Buying ten lottery tickets and winning the lottery.

Conditional Probability

Probability of event A occurring given event B has occurred

$$P(A|B) = \frac{P(AandB)}{P(B)}$$

High Paid Job Example

P(A) - probability of having high paid job

High Paid Job Example

P(A) - probability of having high paid job

P(B) - probability of being Data Scientist = 1/30

High Paid Job Example

P(A) - probability of having high paid job

P(B) - probability of being Data Scientist = 1/30

P(AandB) - probability of being Data Scientist and having high paid job = 1/50

High Paid Job Example

P(A) - probability of having high paid job

P(B) - probability of being Data Scientist = 1/30

P(AandB) - probability of being Data Scientist and having high paid job = 1/50

P(A|B) - probability of Data Scientist having high paid job

High Paid Job Example

P(A) - probability of having high paid job

P(B) - probability of being Data Scientist = 1/30

P(AandB) - probability of being Data Scientist and having high paid job = 1/50

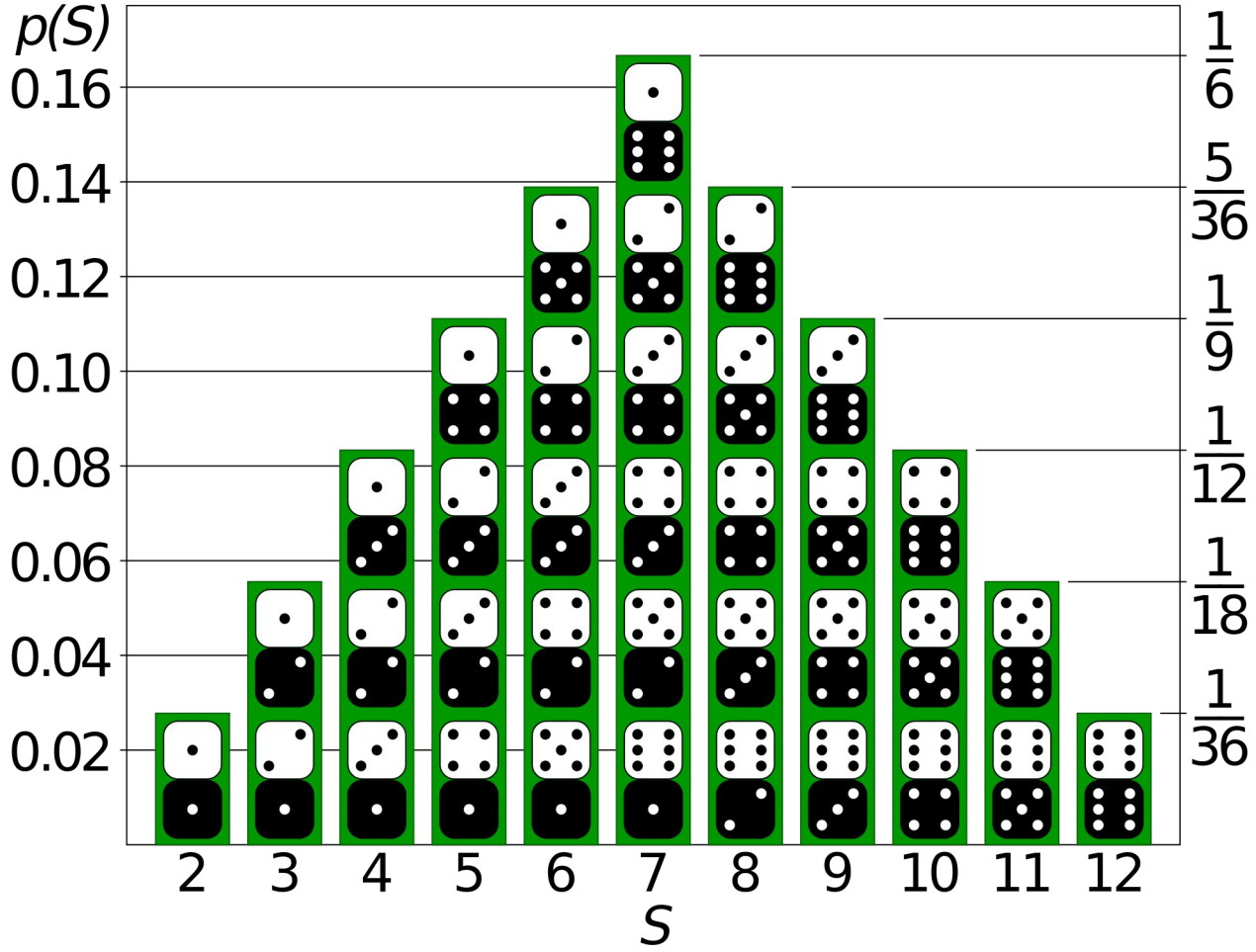
P(A|B) - probability of Data Scientist having high paid job

$$P(A|B) = \frac{\frac{1}{50}}{\frac{1}{30}} = \frac{3}{5}$$

What is the probability of 3?

What is the prob. of 7?

quiz: what prob. of 7?


p(S)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
1/6
5/36
1/9
1/12
1/18
1/36
2
3
4
5
6
7
8
9
10
11
12
S


WIN A CAR


PUNTIGAMER


PUNTIGAMER


CHOOSE A ROOM

KEEP OR SWITCH

KEEP OR SWITCH

KEEP OR SWITCH


YOU WIN THE CAR

YOU WIN BEER

50%

50%


PUNTIGAMER
PUNTIGAMER


PUNTIGAMER
PUNTIGAMER

33.3%

66.6%

33.3%

66.6%

33.3%

33.3%

33.3%

Door selection

P(Beer) = ⅔ = 66.6%

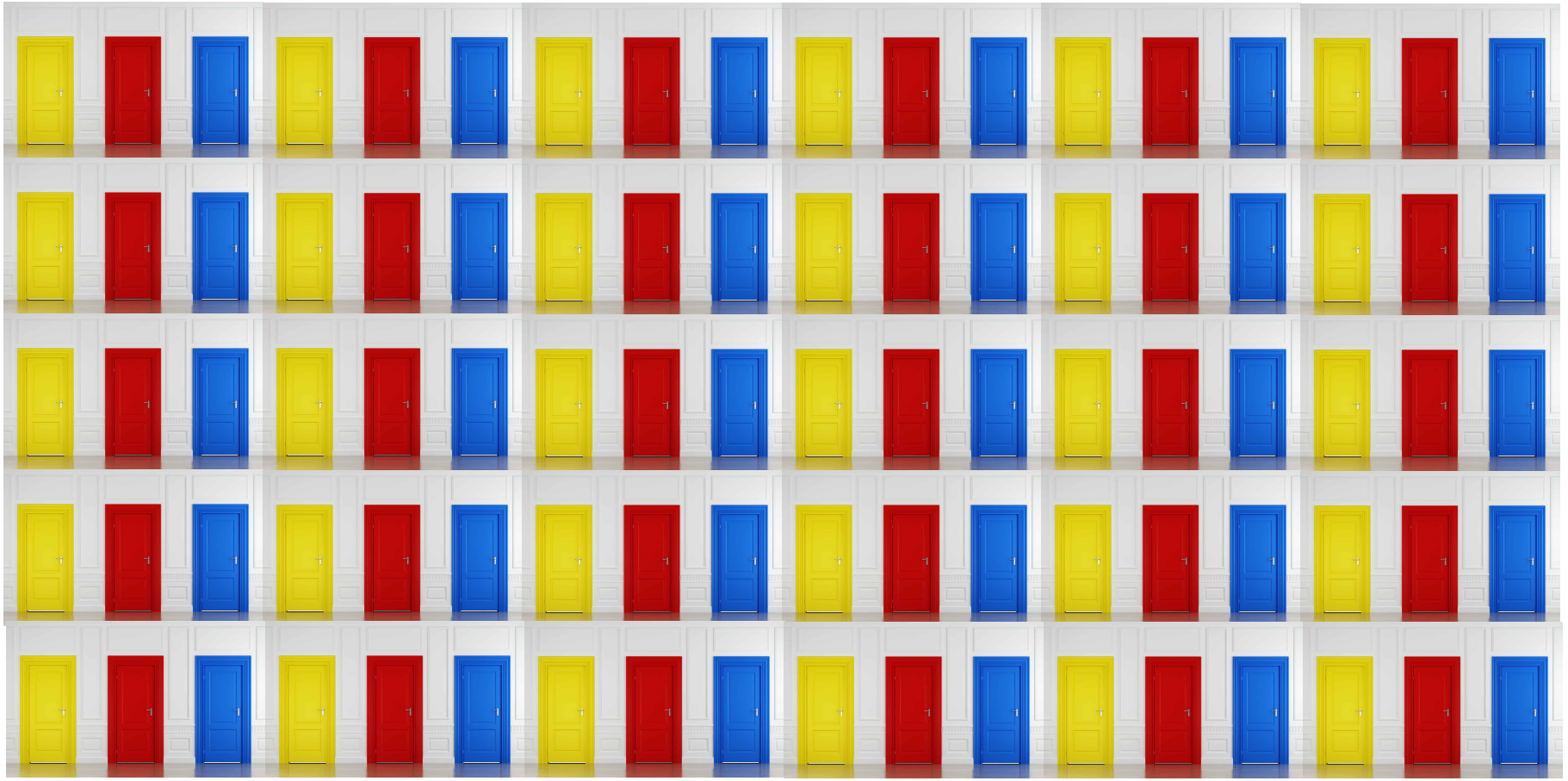
P(Car) = ⅓ = 33.3%

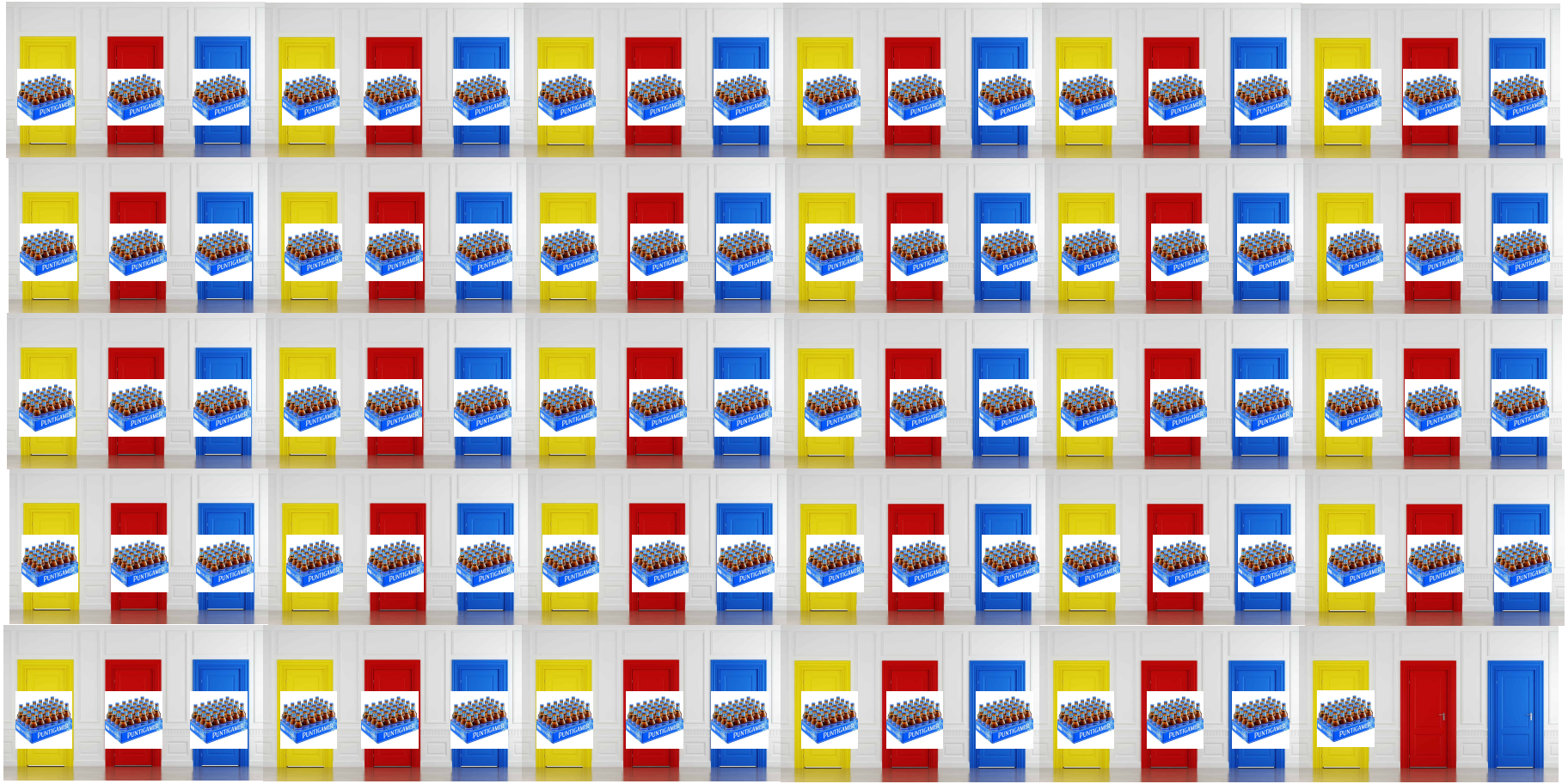
Keep or Switch?

Keeping:

$$P(\text{Car}|\text{Beer}) = 1/3 = 33.3\%$$

$$P(\text{Beer}|\text{Beer}) = 2/3 = 66.6\%$$




PUNTIGAMER

x10

x5

Probability of first marble red?

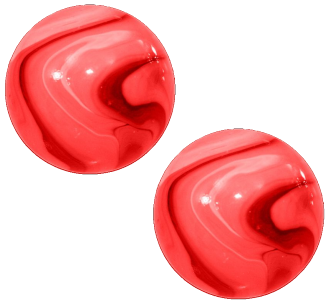
Probability of first marble red?

10/15 = 66.6%

15 Marbles

10 Desired outcomes

Probability of 2 marble red?



10/15 *
9/14 =
42.8%

15 Marbles

10 Desired outcomes

14 Marbles

9 Desired outcomes

quiz:balls examples

Last 10 Games it was:

Last 10 Games it was: ◆

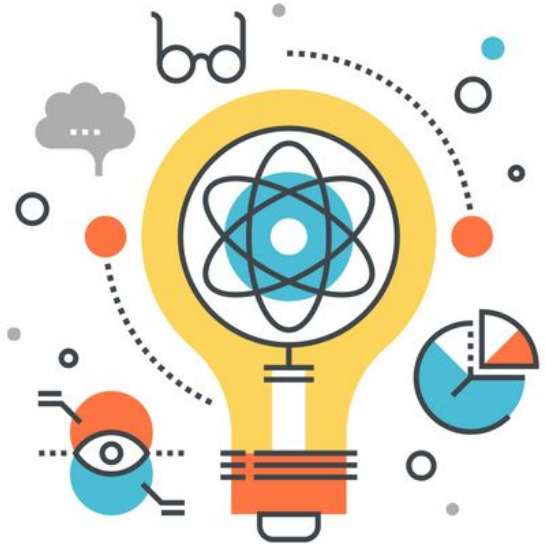
What is the probability that ◆ on the next turn?

37 Outcomes
18 Desired
= 18/37 = 48.6%

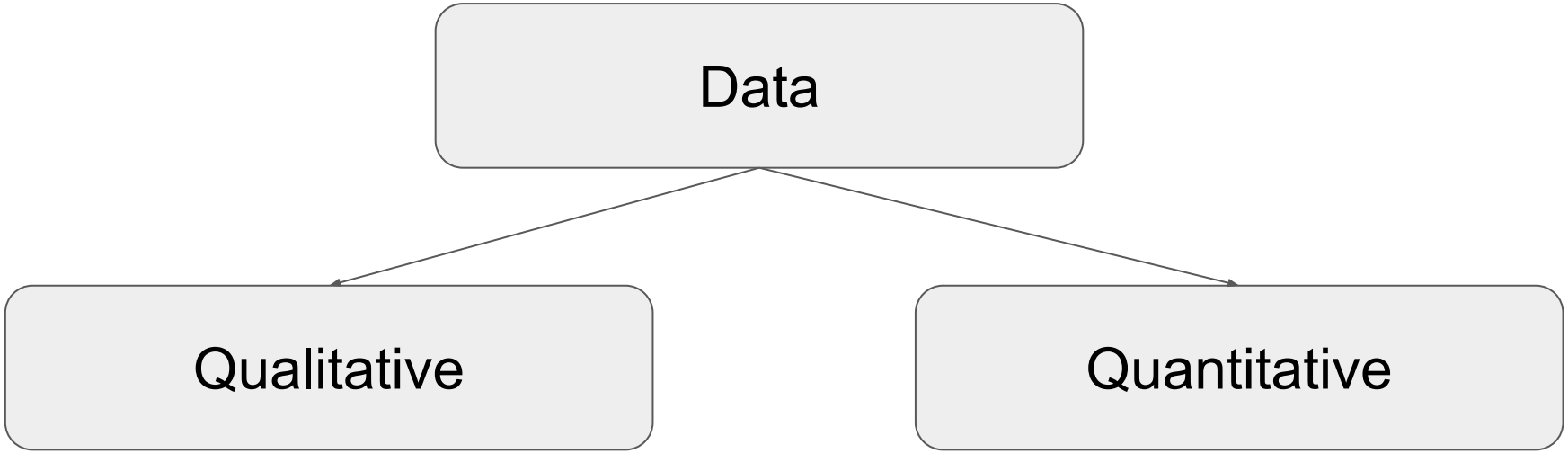
Break

What is data?

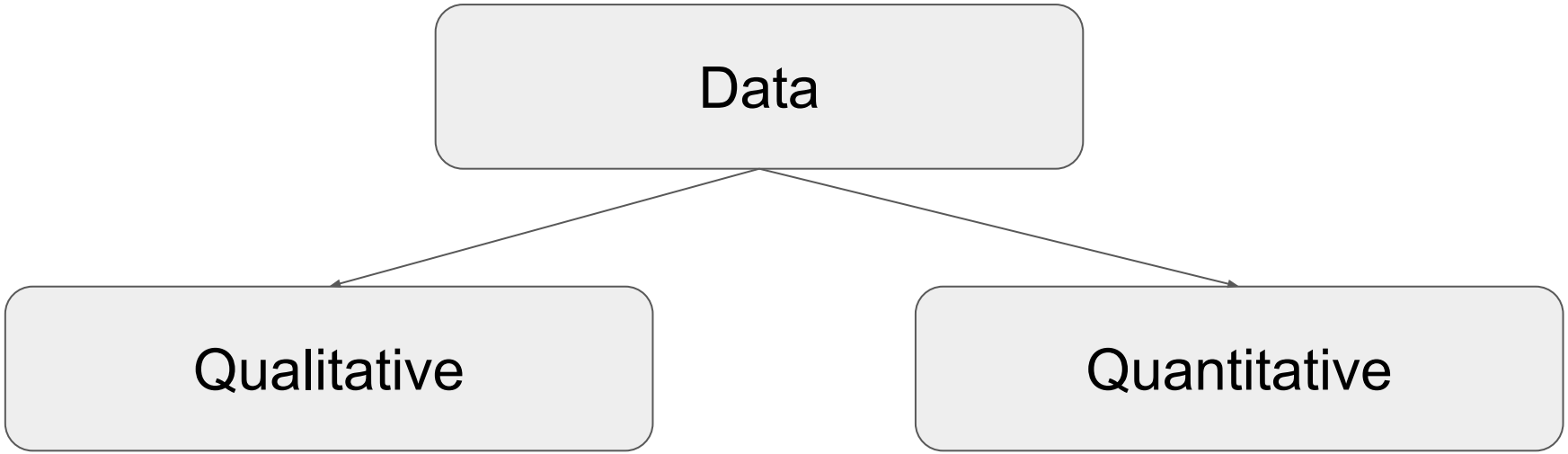
Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.



Data Types



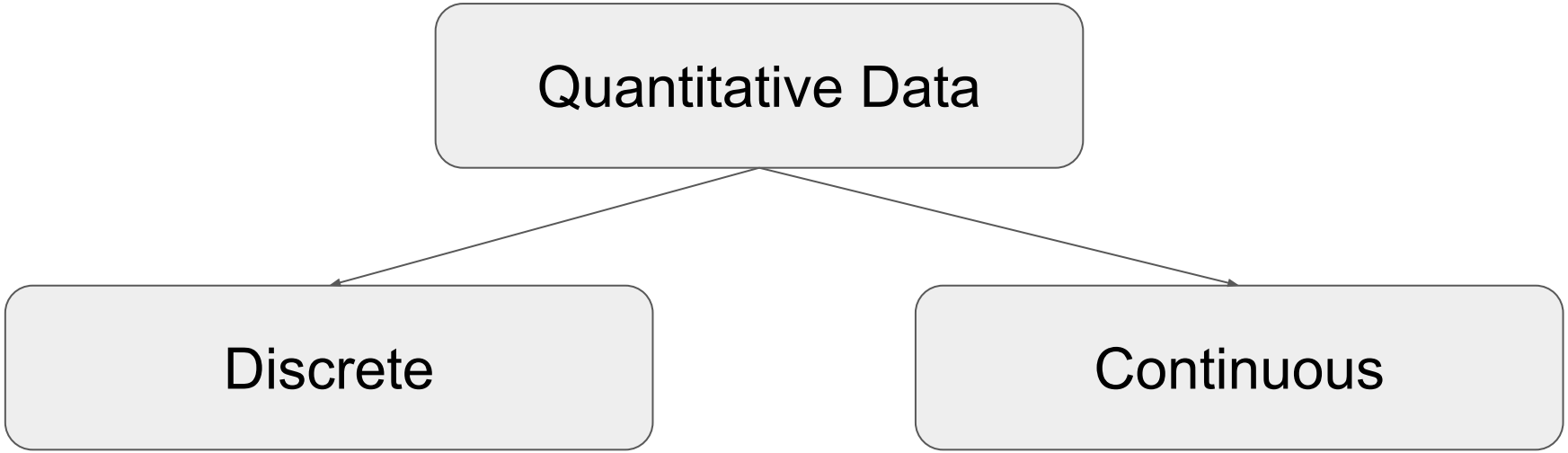
Data Types



- Excellent Data Science course
- Big soda cup
- Male

- 10
- 500 ml
- 3.625, 1.234, 5.223 ...

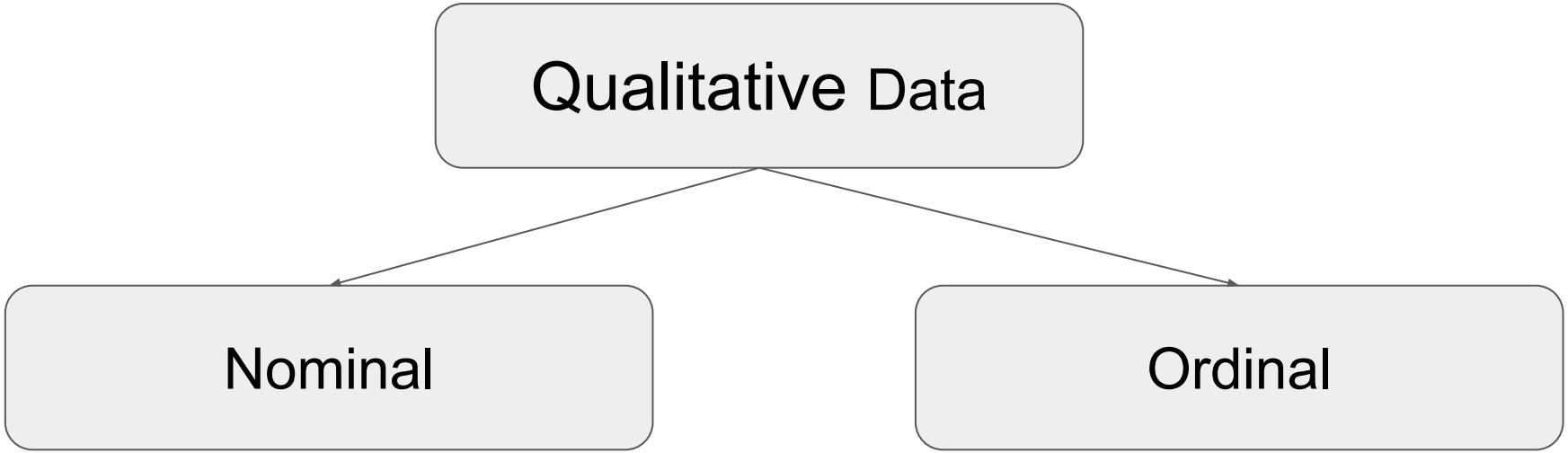
Quantitative Data Types



- Based on counting
- Finite amount values

- Based on measuring
- Can take any value (within a range)

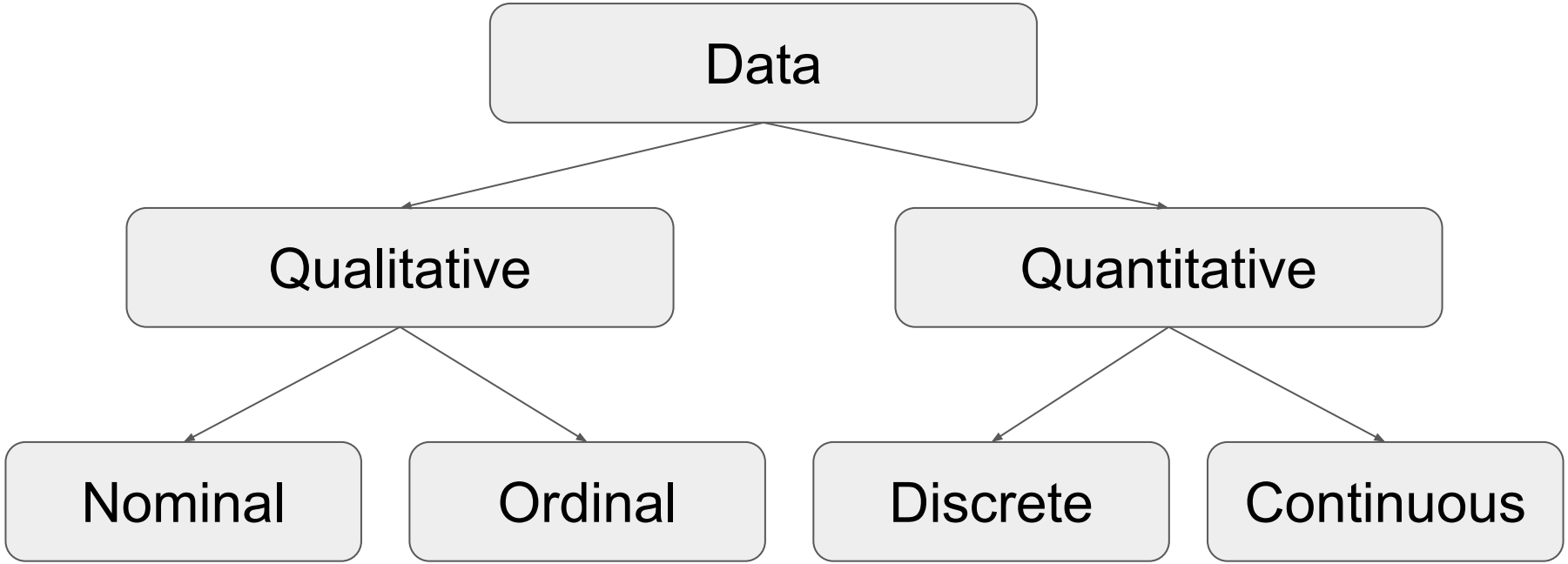
Qualitative Data Types



- Used to label variables
- Can not be ordered or measured

- Follow natural order
- Differences between the data values cannot be determined or are meaningless.

Data Types


Data
Qualitative
Quantitative
Nominal
Ordinal
Discrete
Continuous

Example Apartment Data

Street Name	Amount of Rooms	Area	Type	Price
Glacisstrasse	5	220	Luxury	2500000
Inffeldgasse	3	120	Standart	450000
Kasernstrasse	2	45	Econom	100000
Kreuzgasse	6	250	Luxury	3000000

Example Apartment Data

Street Name	Amount of Rooms	Area	Type	Price
Glacisstrasse	5	220	Luxury	2500000
Inffeldgasse	3	120	Standart	450000
Kasernstrasse	2	45	Econom	100000
Kreuzgasse	6	250	Luxury	3000000

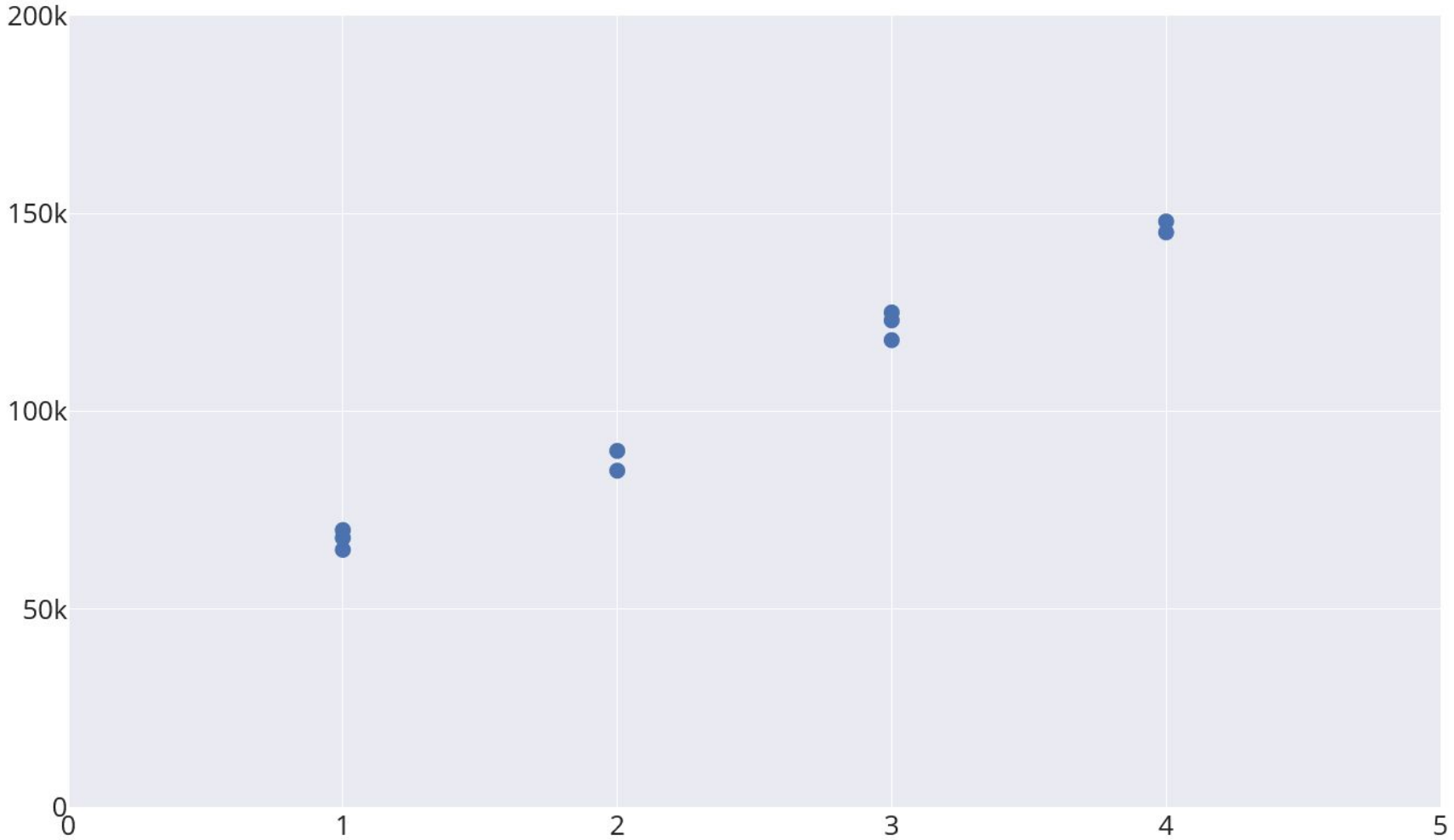
Nominal	Discrete	Continuous	Ordinal	Continuous

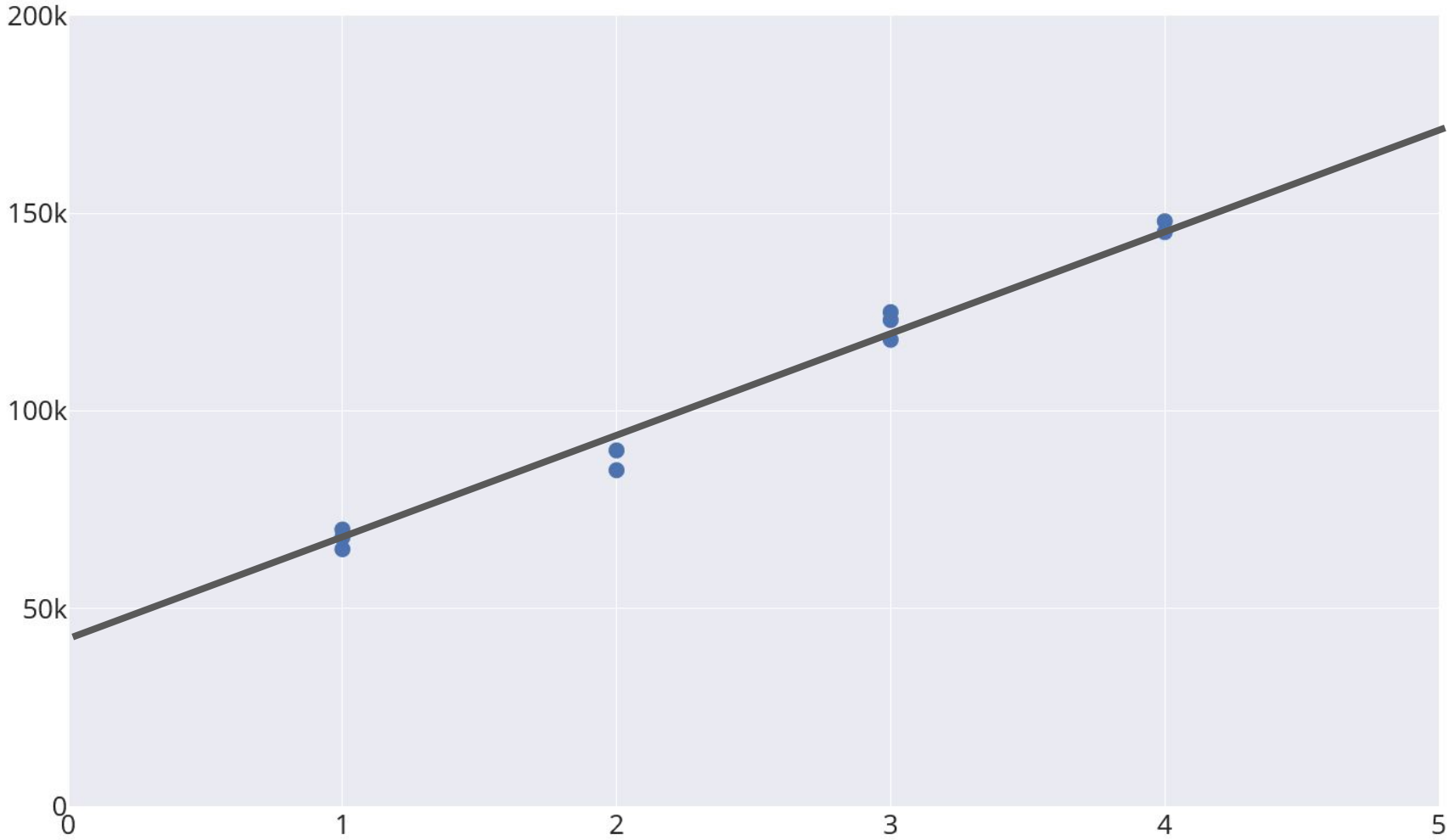
What about Postal Code?

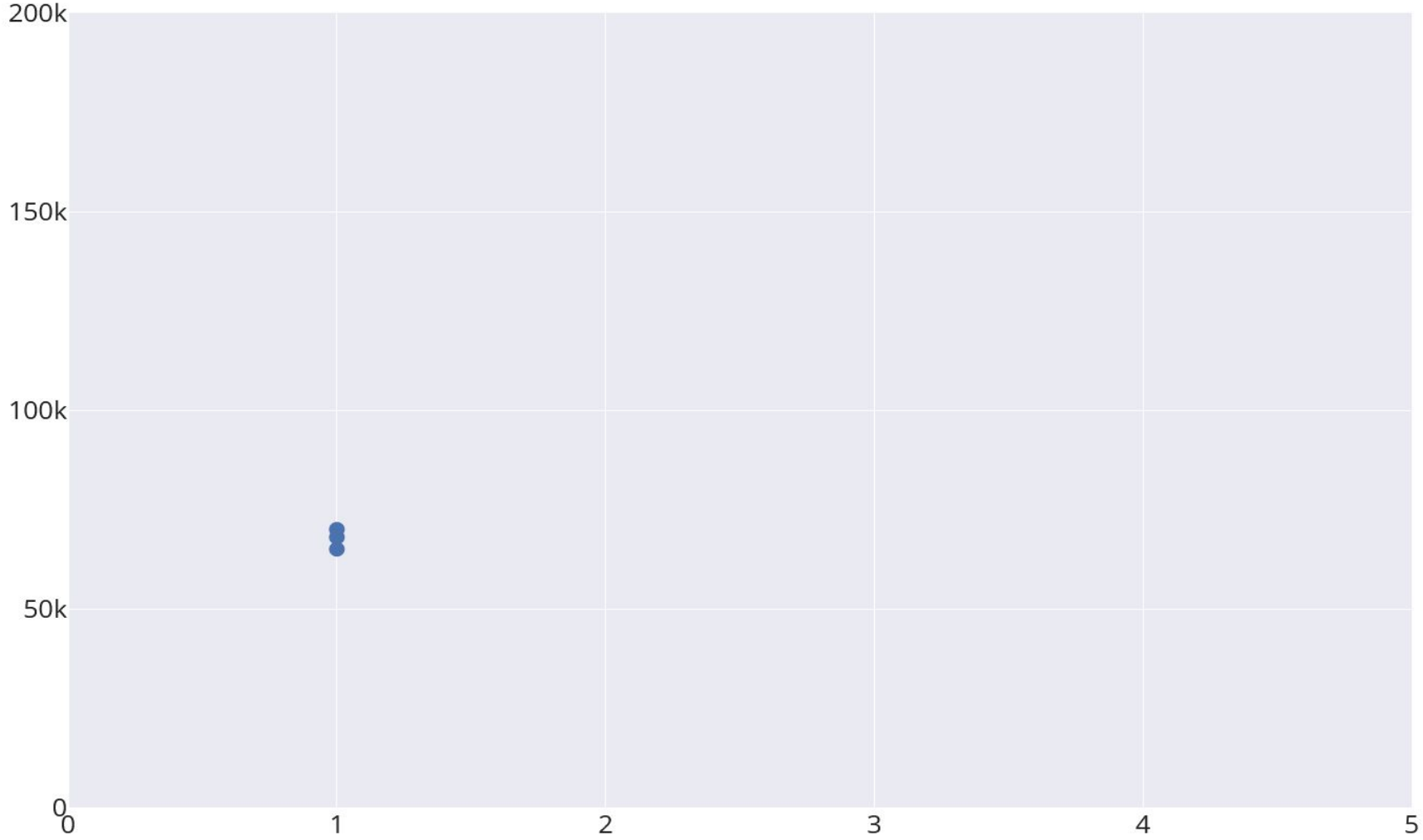
Postal Code
8010
8010
8041
8020

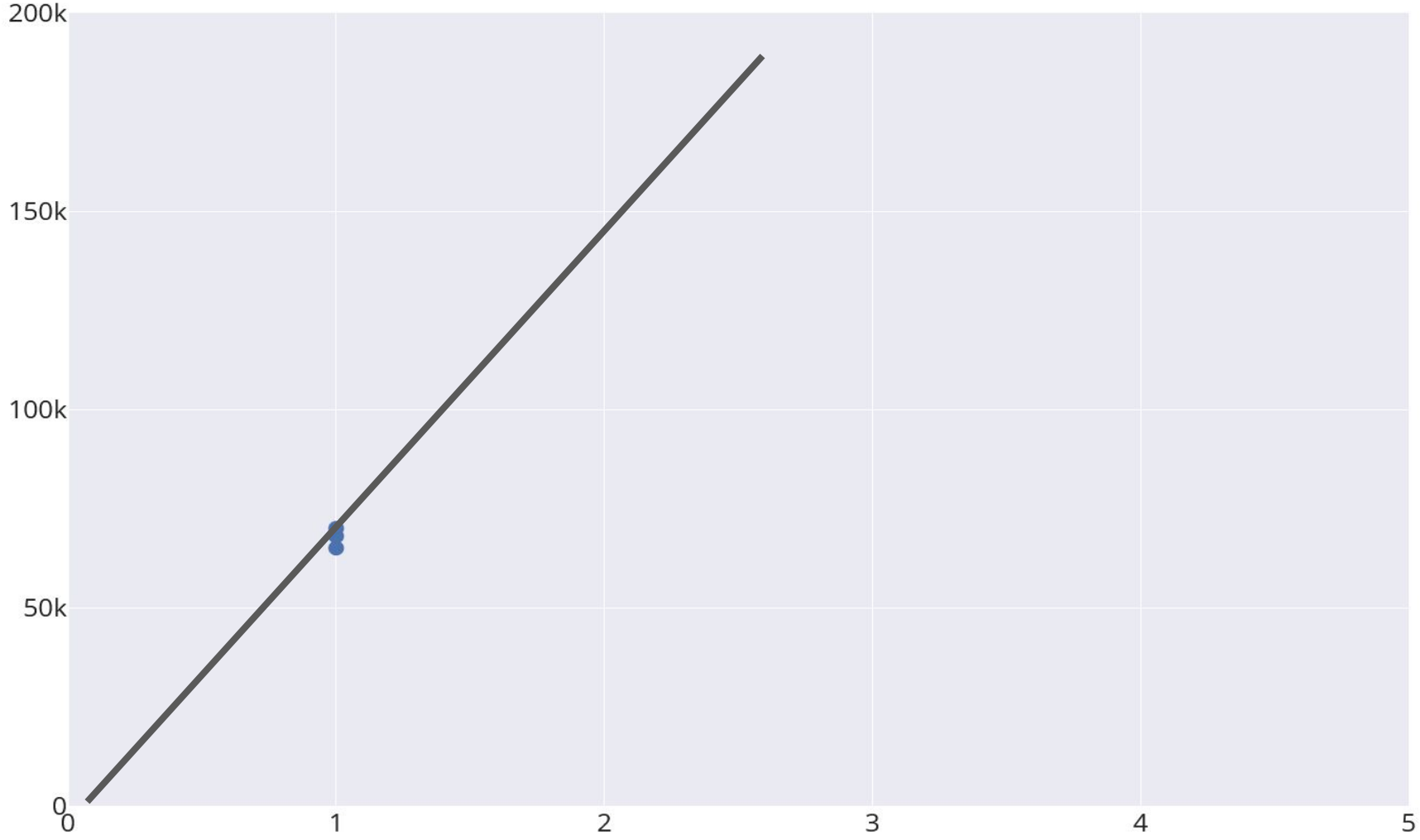
Predicting Price

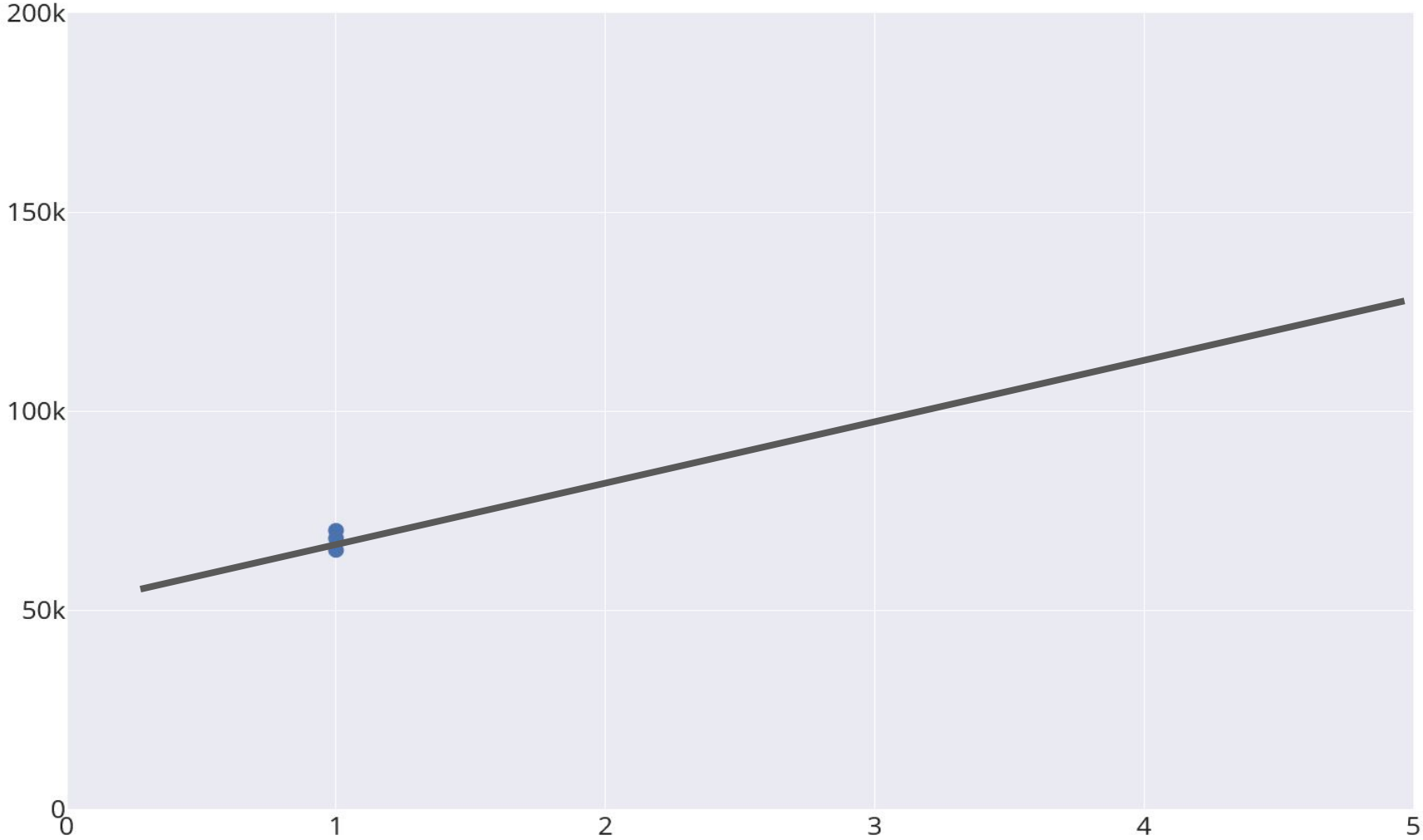
Amount of Rooms	Price
3	125000
1	52000
2	75000
4	148000

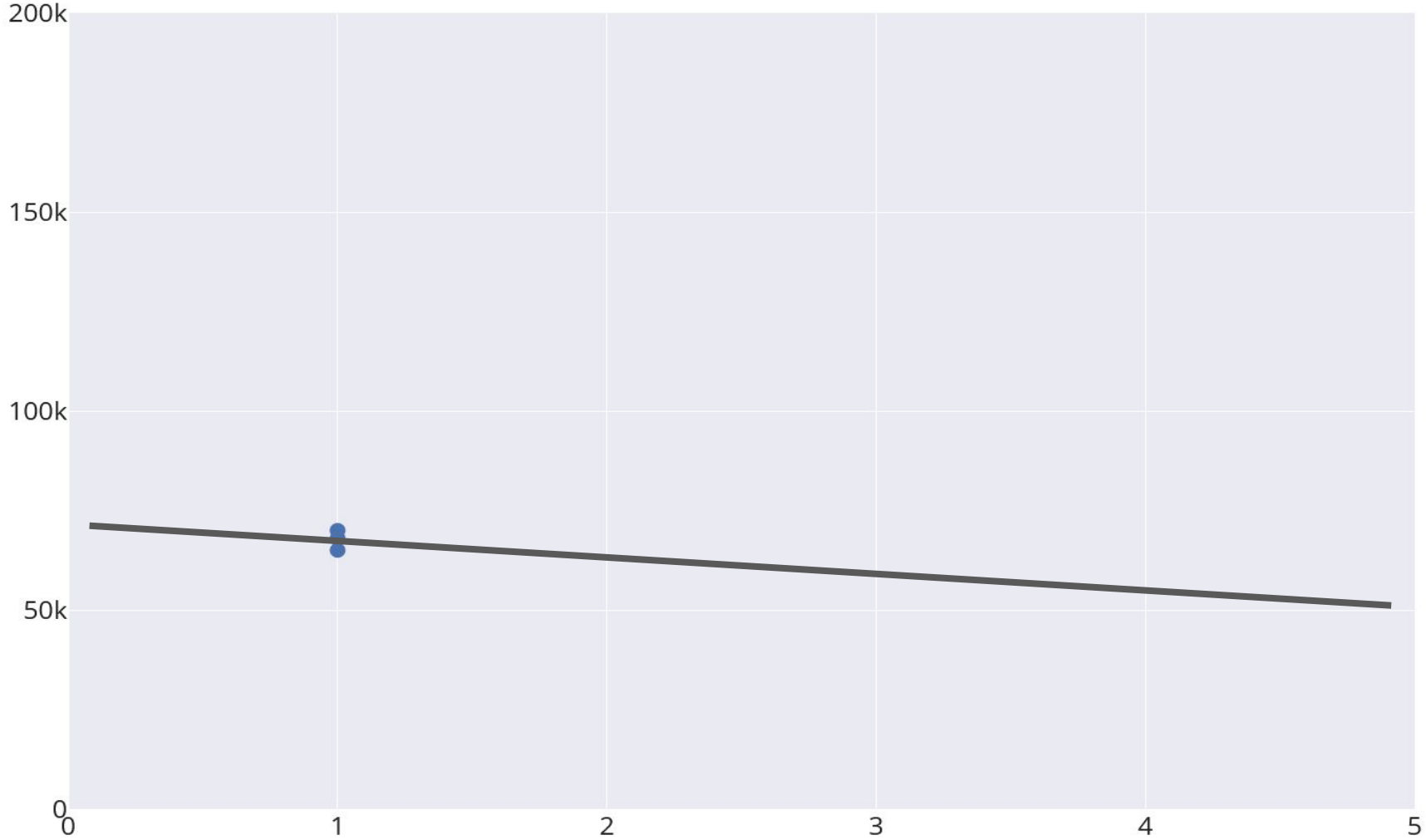

200k
150k
100k
50k
0
0
1
2
3
4
5


200k
150k
100k
50k
0
0
1
2
3
4
5


200k
150k
100k
50k
0
0
1
2
3
4
5


200k
150k
100k
50k
0
0
1
2
3
4
5


200k
150k
100k
50k
0
0
1
2
3
4
5


200k
150k
100k
50k
0
0
1
2
3
4
5

How much data do we need?

How much
data do we need?

Depends on the Problem.

https://giphy.com/videos/ConnectedMovie-sony-connected-movie-fUNQhFbC05DMRwCYPF

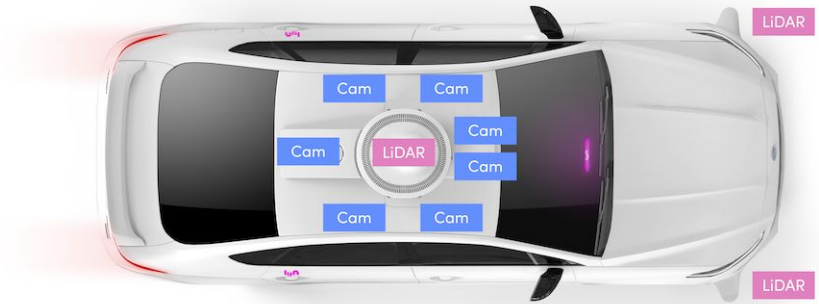
or

?

99% Dog

How much data for Image Classifier?

~ 1000 images per Class


LiDAR
Cam
Cam
Cam
Cam
LiDAR
Cam
Cam
Cam
LiDAR


lyft
Ford
California
8ANW098

VIDEO

https://level5.lyft.com/wp-content/uploads/2019/06/all_cams_example.mp4

How much data for Self-driving car?

How much data for Self-driving car?

~2080 Hours

Waymo (Self-driving Car Company)

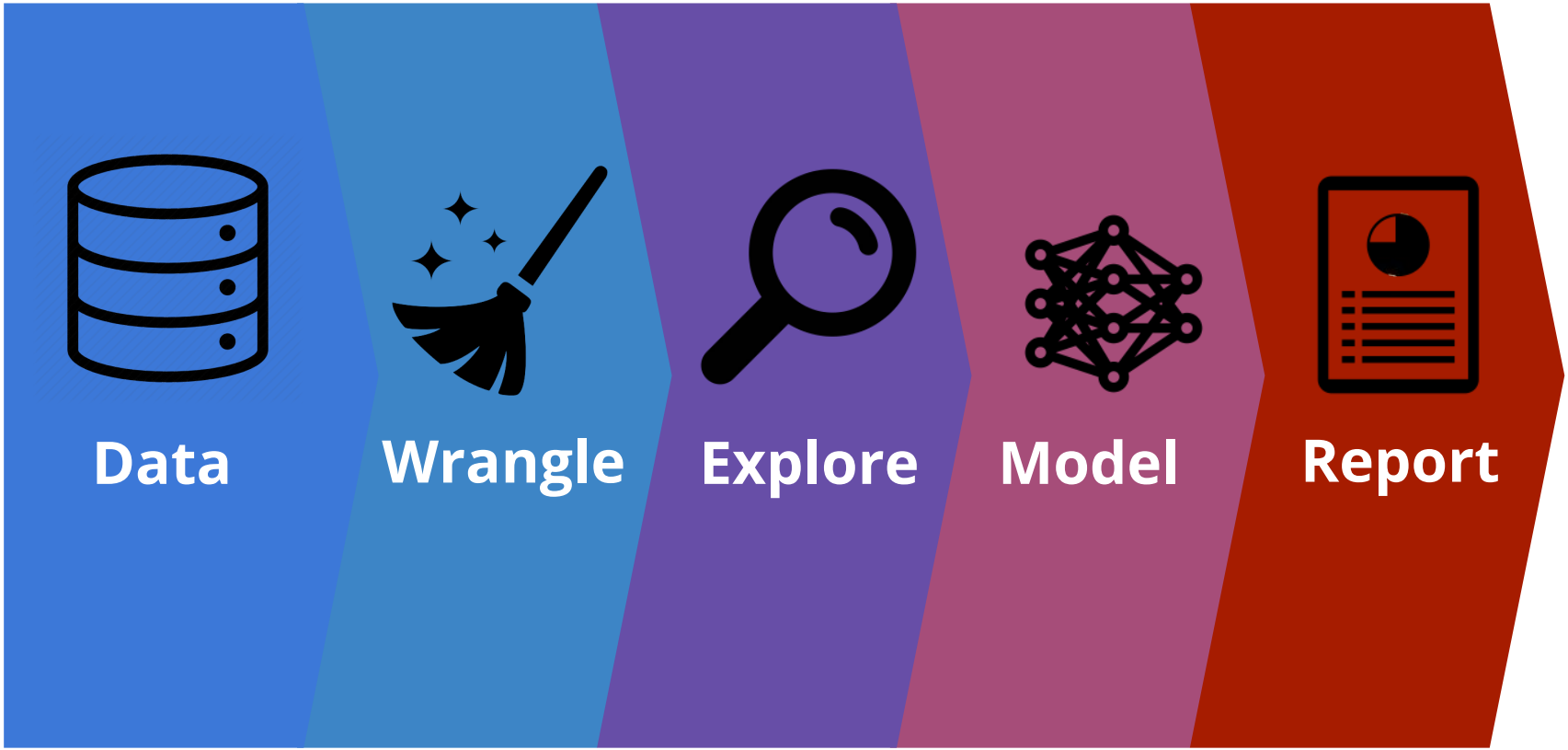
How much data for Language Translation?

How much data for Language Translation?

English - French

2,007,723 Sentences

Data Science Pipeline



Assignment

- Fill up the Quiz (will get it through mail)
- Have a Gmail Account until next week

Next week Preview

- Work on real life datasets
- Learn how to answer questions with data
- Understand the fundamentals of cleaning your data


pandas


jupyter


[adult swim]
See you guys next week!