Navigating the Data Science Landscape

Half-day workshop at

International Conference on Applied Statistics, ICAS 2019

Institute of Statistical Research and Training (ISRT)

University of Dhaka, Dhaka, Bangladesh

December 26, 2019

Enayetur Raheem, PhD
enayetur.raheem@brfbd.org

Part 2
Organizational Hierarchy

How works get done

Background

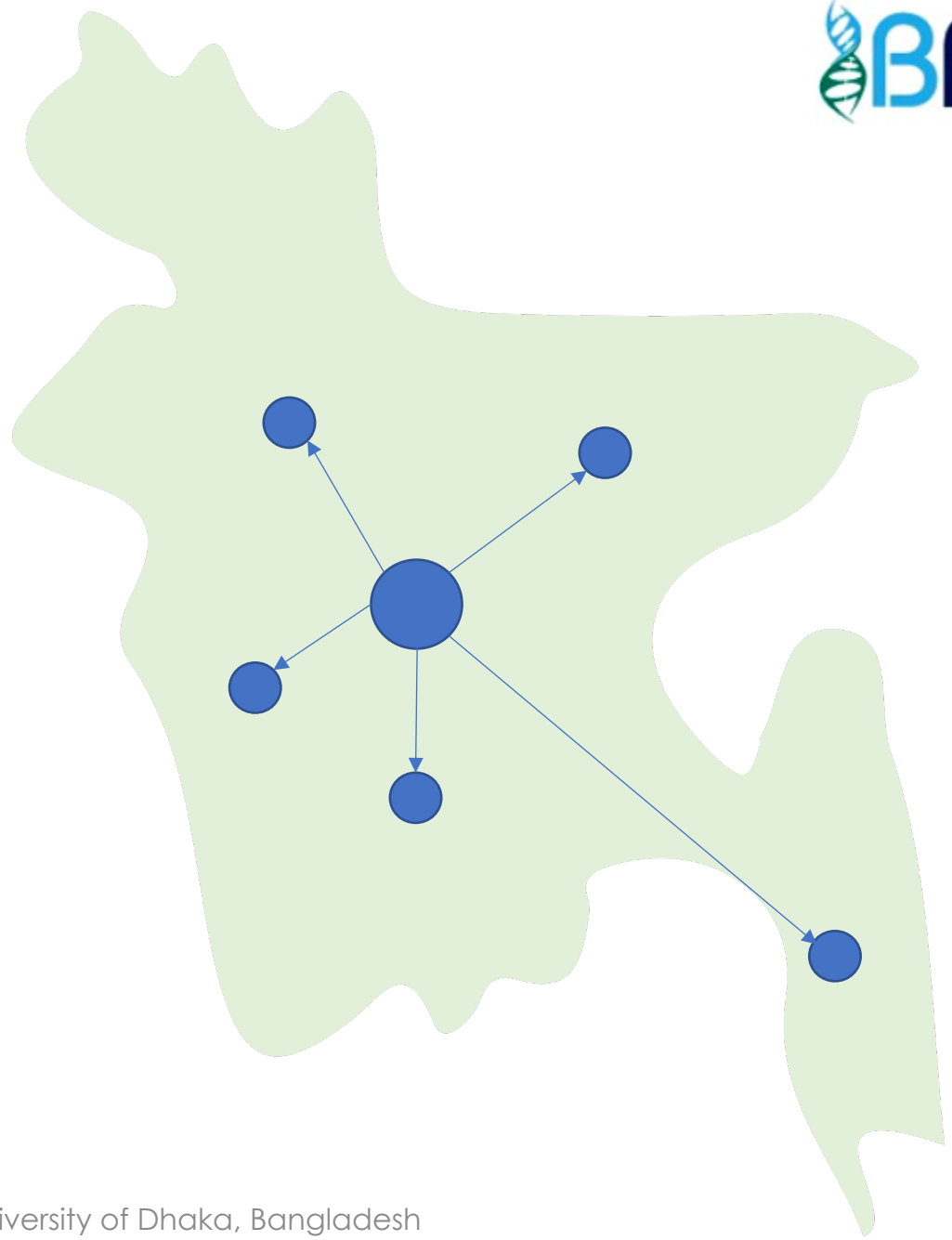
- This is a boring but important section
- Understanding the workplace is crucial to your success which most of us fail to realize
- New graduates have no way of knowing this until …

If you want to be successful in getting a job, learn the organization first

Centralized

Vs

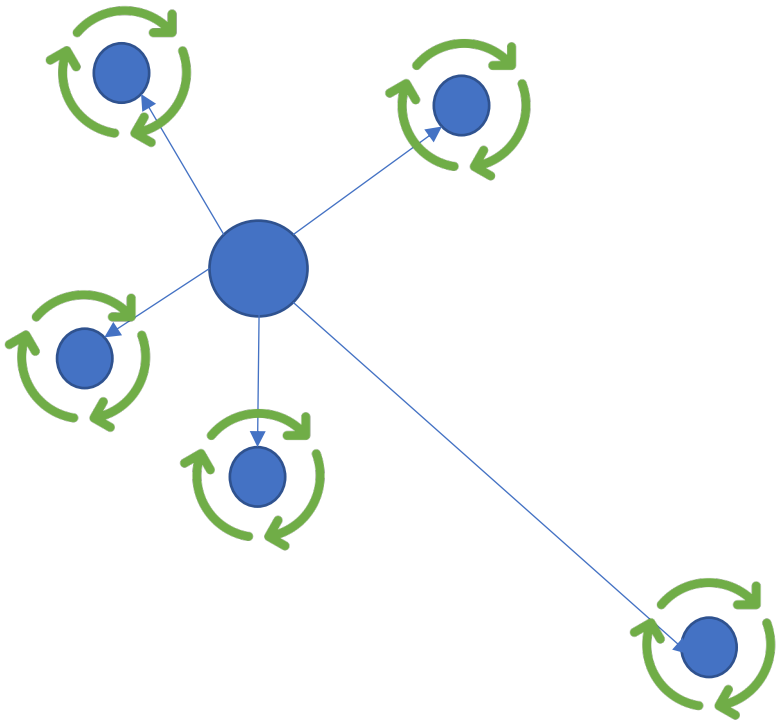
De- centralized



Centralized

Vs

De- centralized

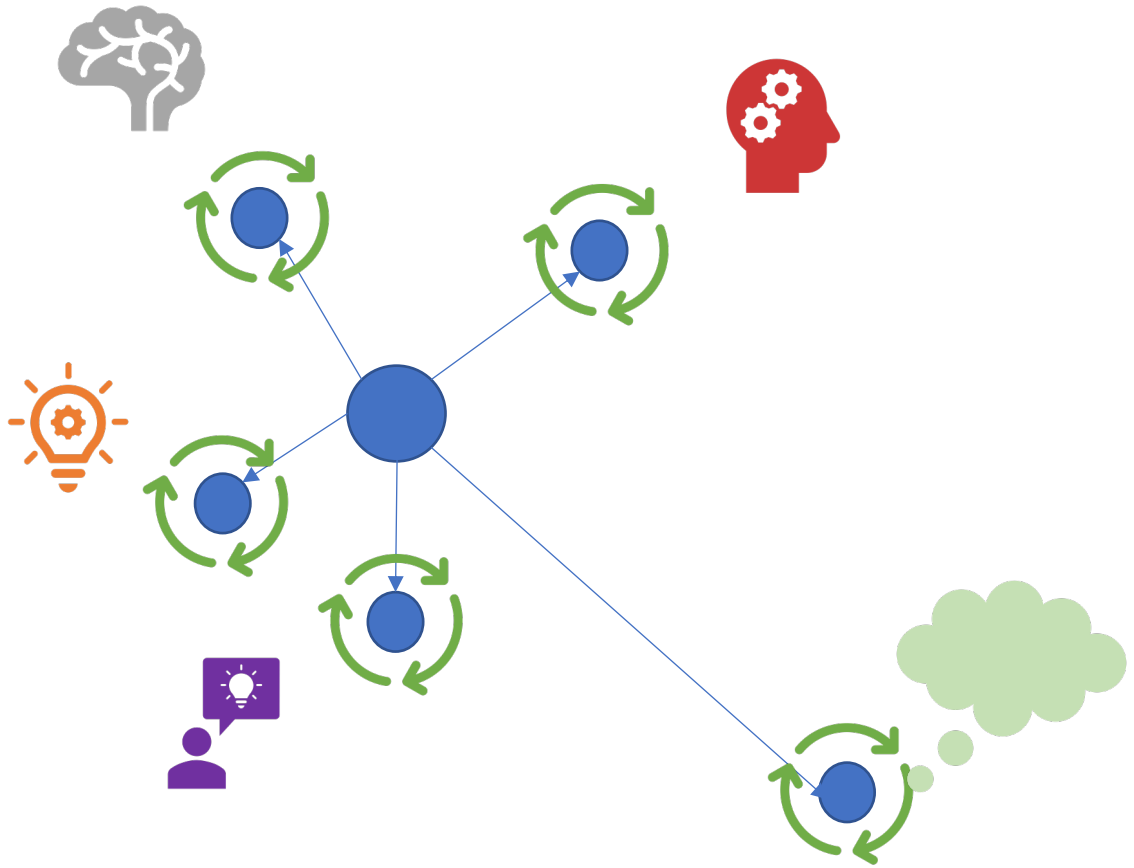


Centralized

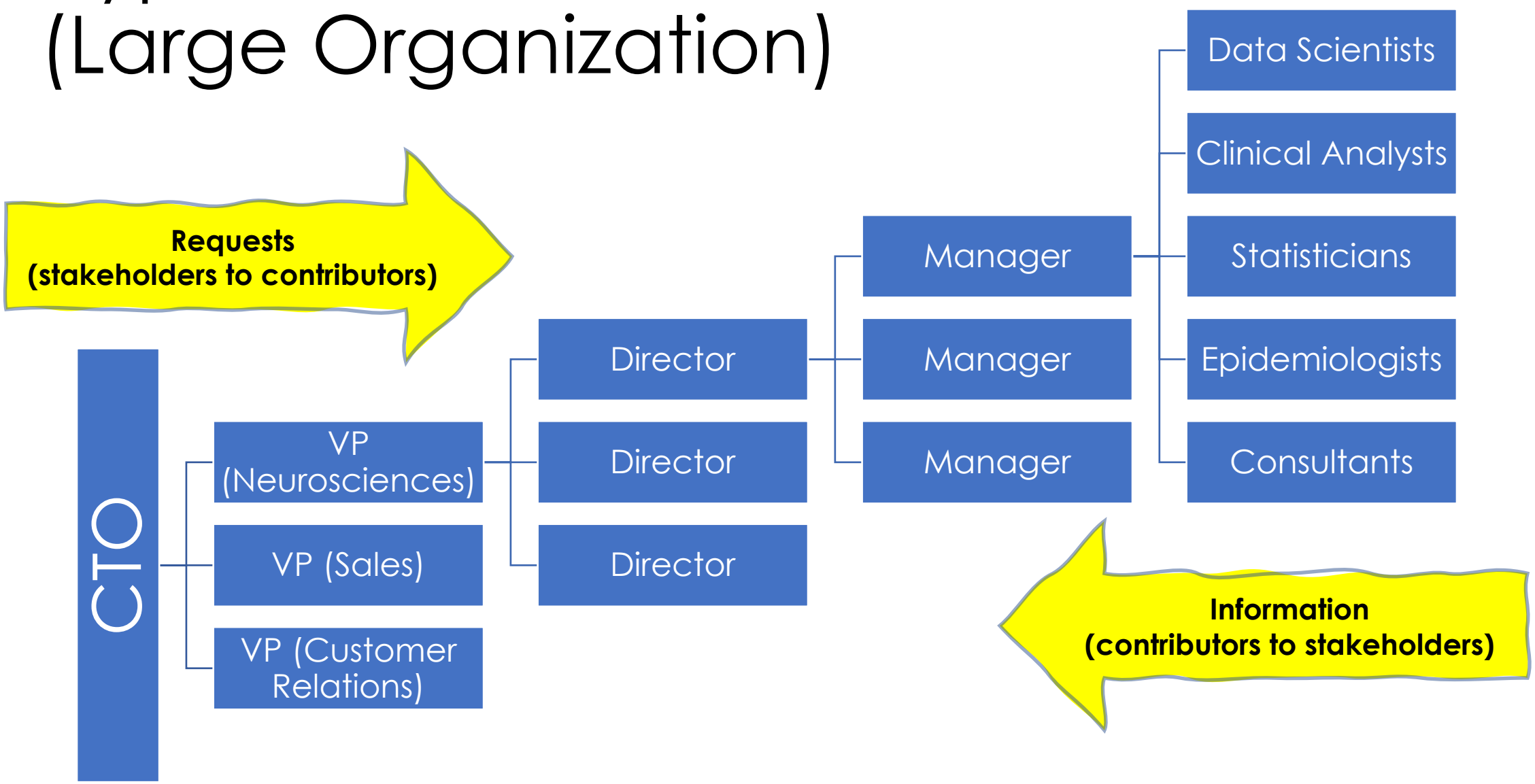
Vs

De- centralized

Each unit becomes
- center of innovation
- semi autonomous
- focused towards common goal



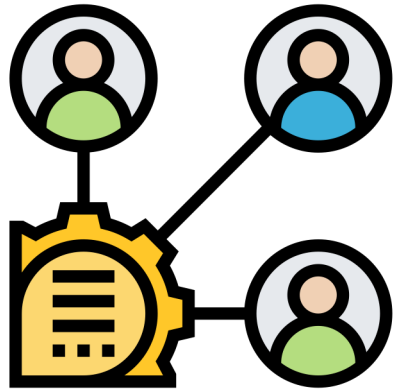
Hypothetical Structure (Large Organization)



Stakeholders

A stakeholder is a party that has an interest in a company and can either affect or be affected by the business. The primary stakeholders in a typical corporation are its

- investors
- employees
- customers and suppliers



Data Scientists provide the information that stakeholders use to make decision that affects the company (growth/loss)

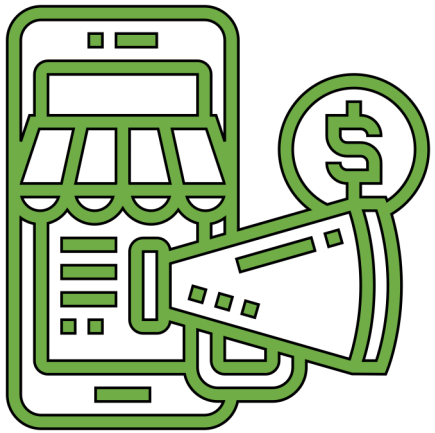
https://www.investopedia.com/terms/s/stakeholder.asp

Example Scenario

- A company is preparing their marketing campaign for the upcoming year
- Several teams are working to deliver the market proof point

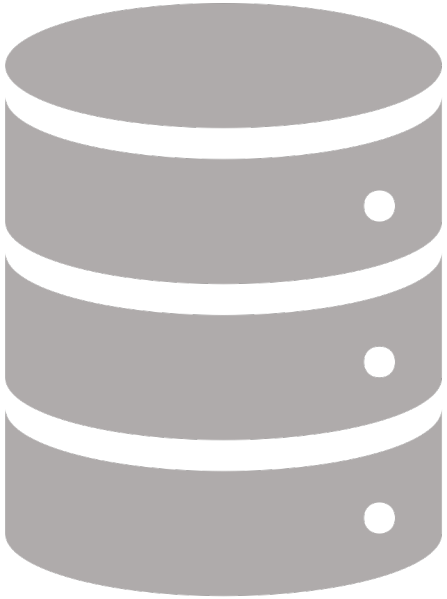
A marketing campaign can't just claim their product is the best. They must show evidence with data.

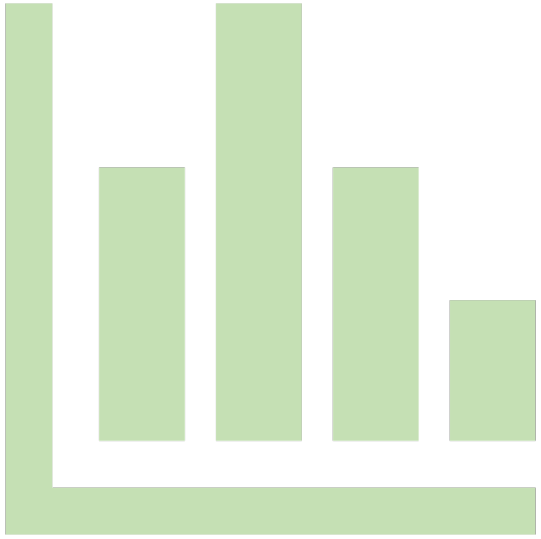
Market Proof Point



Marketplace

Question is, where do you get the data from and how do the Data Scientists work?






Data Lake

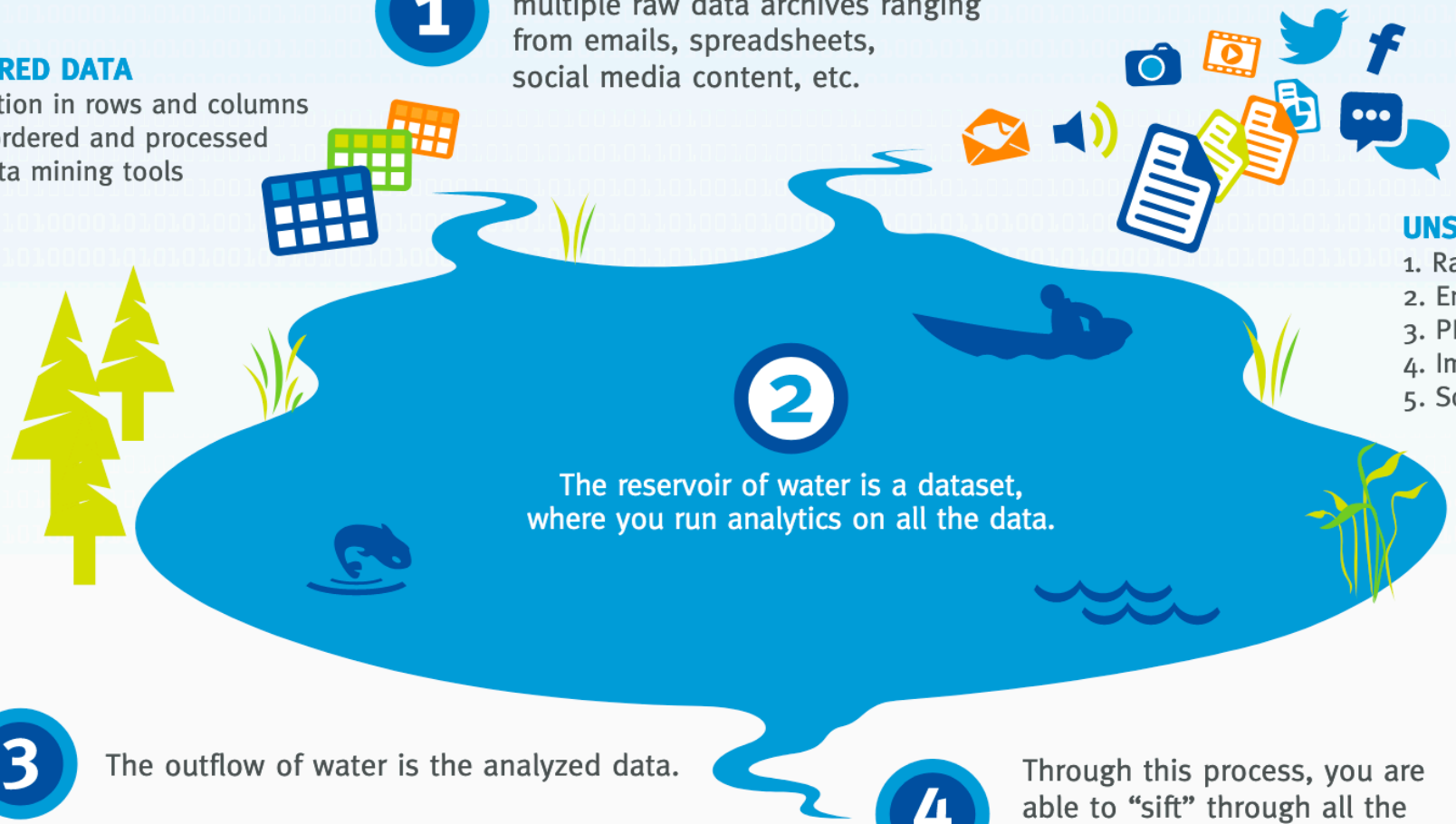
HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

1. The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools



UNSTRUCTURED DATA

1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools

2. The reservoir of water is a dataset, where you run analytics on all the data.

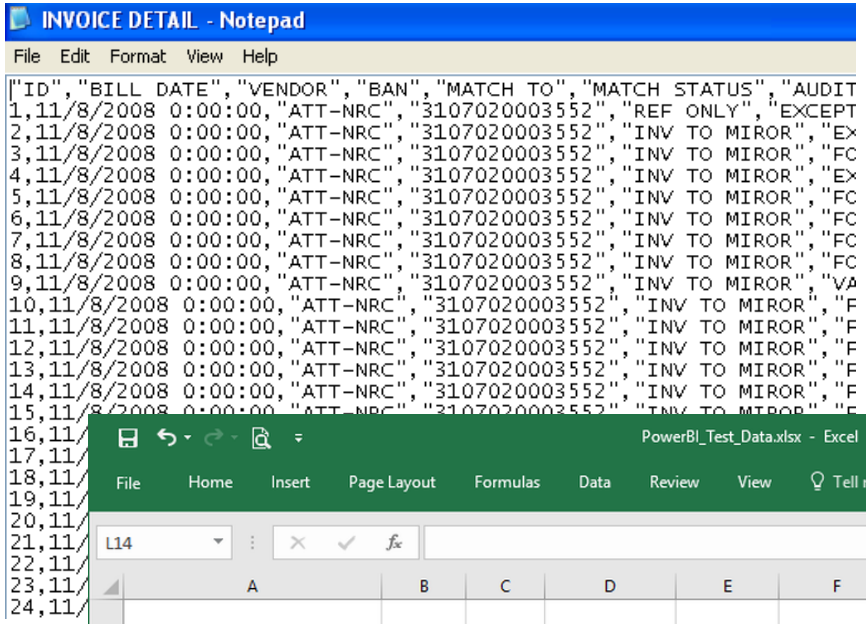
3. The outflow of water is the analyzed data.

4. Through this process, you are able to "sift" through all the data quickly to gain key business insights.

https://www.dellemc.com/en-us/index.htm

Tabular Data

- Most of the times statisticians work with almost finished data
  - in flat files or spreadsheets
- For a data scientist, this is often a deliverable for downstream use

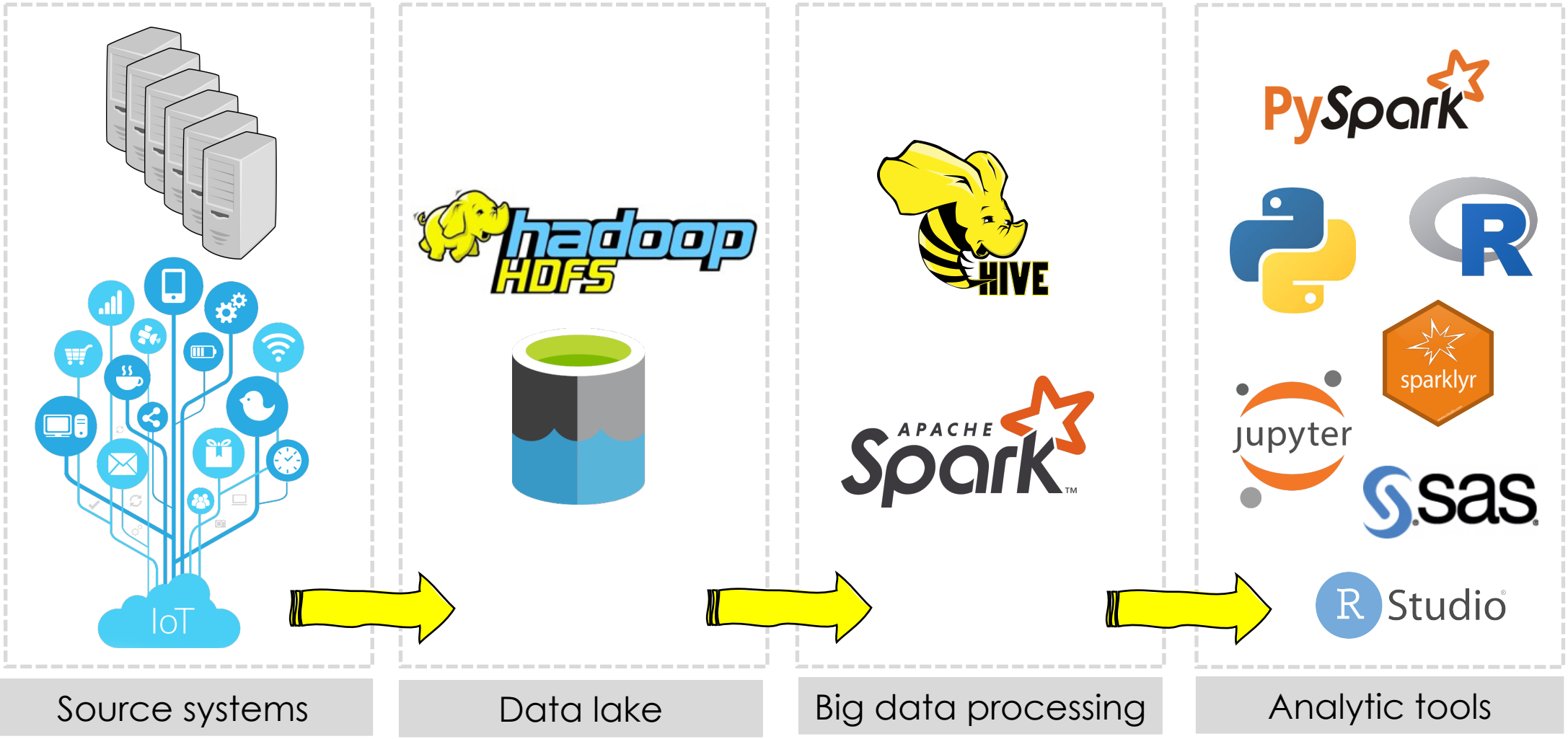


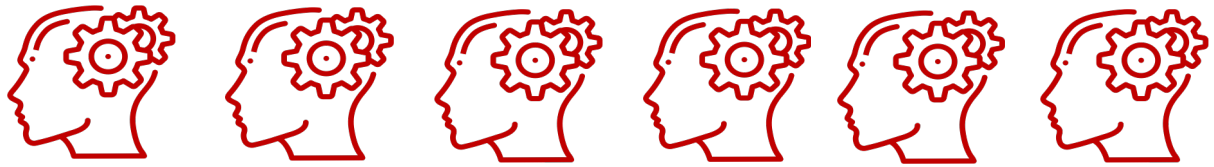
	A	B	C	D	E	F	G	H	I	J
1	Stock Name	Symbol	Shares	Purchase Price	Cost Basis	Current Price	Market Value	Gain/Loss	Dividend/share	Annual Yield
2	Apple	AAPL	100	$90.00	$9,000.00	$144.13	$14,413.27	$14,269.14	$2.28	1.58%
3	Microsoft	MSFT	200	$32.00	$6,400.00	$65.57	$13,114.14	$13,048.57	$1.56	2.38%
4	Salesforce	CRM	150	$25.00	$3,750.00	$82.57	$12,385.50	$12,302.93	$0.00	0.00%
5	Oracle	ORCL	250	$50.00	$12,500.00	$44.56	$11,138.75	$11,094.20	$0.64	1.44%
6	Hewlett Packard Enterprise	HPE	500	$18.00	$9,000.00	$17.69	$8,842.50	$8,824.82	$0.26	1.47%
7	Alphabet	GOOG	100	$225.00	$22,500.00	$833.36	$83,336.00	$82,502.64	$0.00	0.00%
8	Intel	INTC	200	$22.00	$4,400.00	$36.07	$7,213.00	$7,176.94	$1.09	3.02%
9	Cisco	CSCO	225	$18.00	$4,050.00	$33.24	$7,478.78	$7,445.54	$1.16	3.49%
10	Qualcomm	QCOM	185	$65.00	$12,025.00	$56.48	$10,447.88	$10,391.40	$2.12	3.75%
11	Amazon	AMZN	50	$800.00	$40,000.00	$897.64	$44,882.00	$43,984.36	$0.00	0.00%
12	Redhat	RHT	100	$95.00	$9,500.00	$86.26	$8,626.00	$8,539.74	$0.00	0.00%
13	Facebook	FB	1000	$17.00	$17,000.00	$141.64	$141,640.00	$141,498.36	$0.00	0.00%
14	Twitter	TWTR	500	$45.00	$22,500.00	$14.61	$7,302.55	$7,287.94	$0.00	0.00%
15										

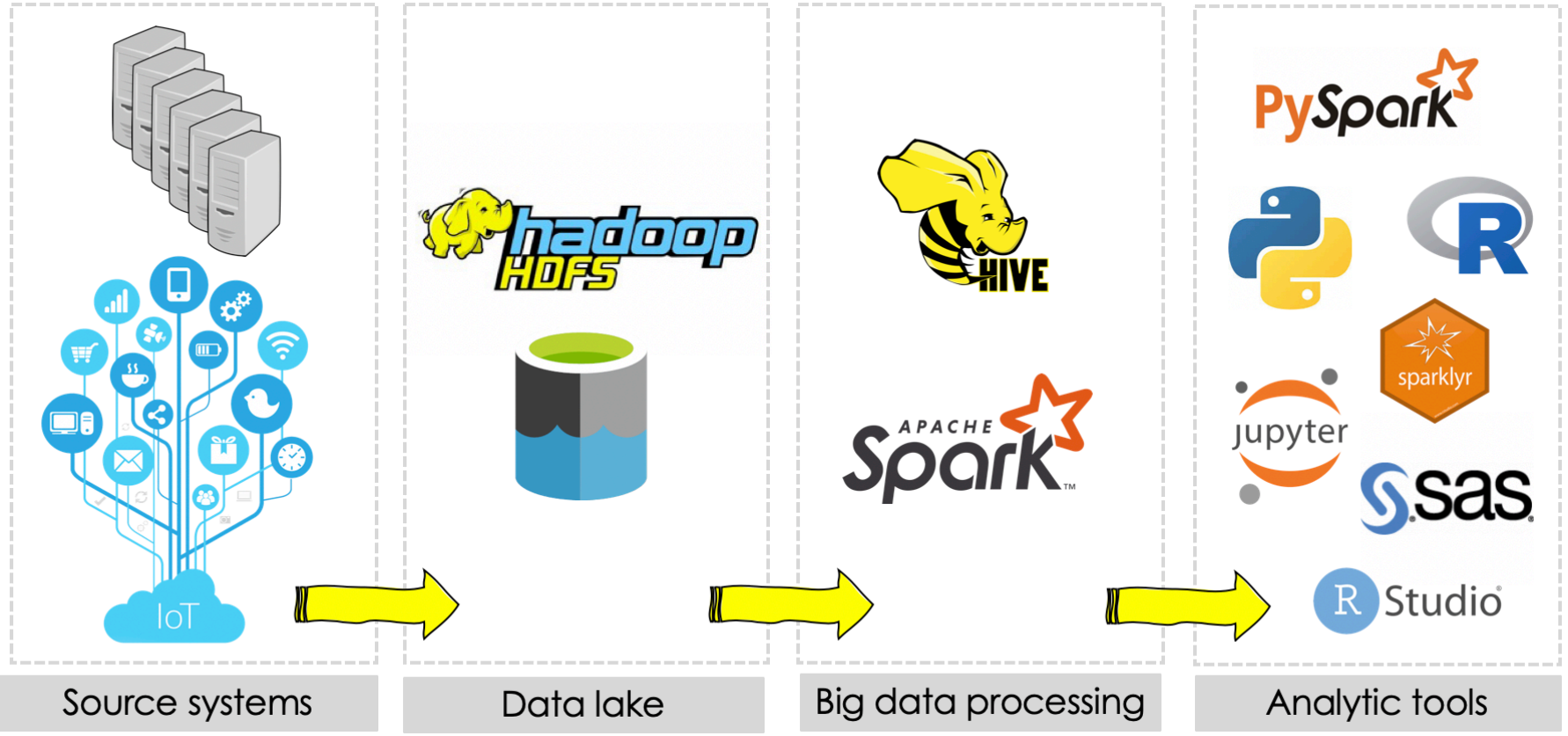
Image sources:

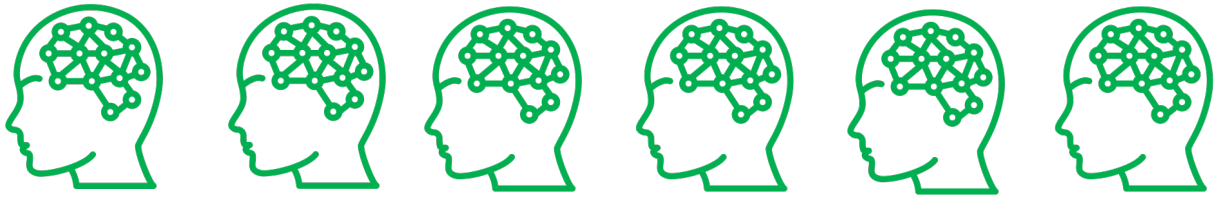
https://bit.ly/2X51fRF
https://tek.io/33A5Pd5

Data Architecture





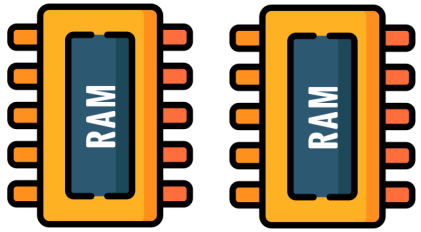

IoT
hadoop
HDFS
HIVE
APACHE
Spark
PySpark
sparklyr
jupyter
sas
R Studio
Source systems
Data lake
Big data processing
Analytic tools



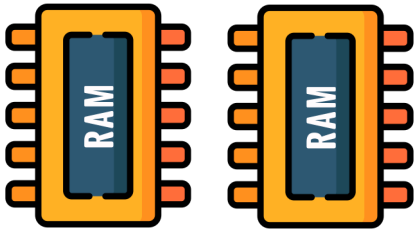
Data Science Tools

Data Shape/Dimension

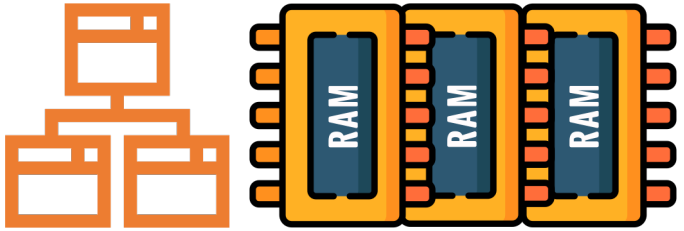
```
df.shape
```



```
dim(df)
```




```
df.count()
len(df.columns)
```



Data Size

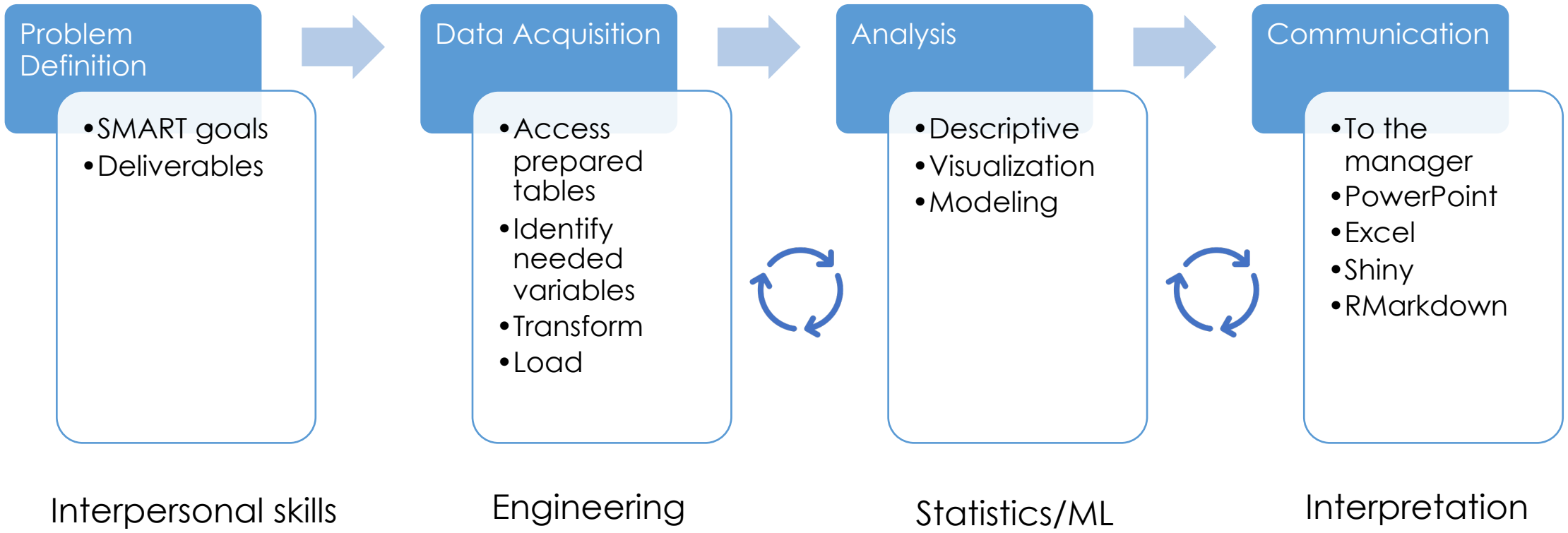
Unprocessed Data	Processed Data	Working Data
In Hadoop/data lake	- In RDBMS - Hive tables - Impala tables	- CSV - Parquet
Size on disk: 100s of GB to Terabytes	100s GBs	10s of GBs
		- Several hundred to a few thousand columns - Several millions to billions of rows

Challenges with Big Data

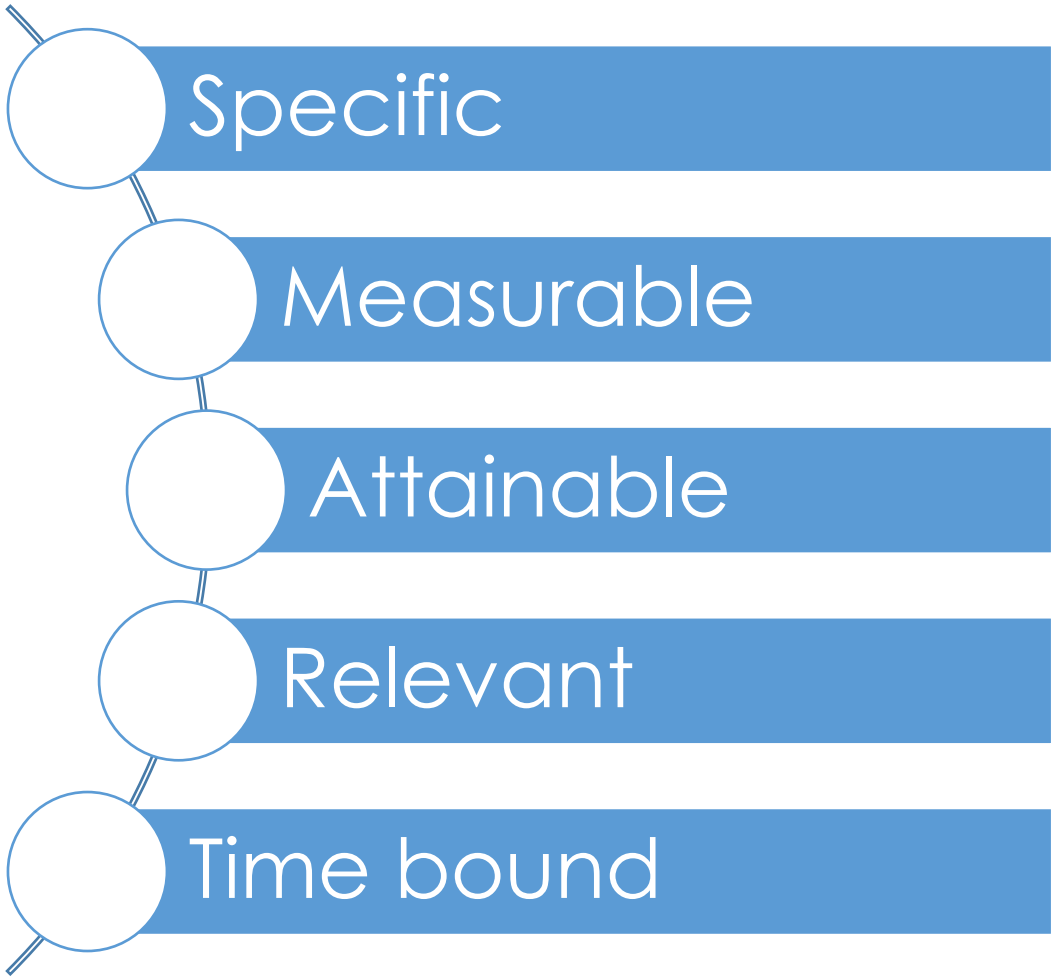
- As unbelievable as it may sound, it often takes several days (run time) to bring data to a usable shape

- From that stage, another 6-12 hours (run time) to get the data to a final tabular format to do any statistics on it

Data Science Workflow Schematic



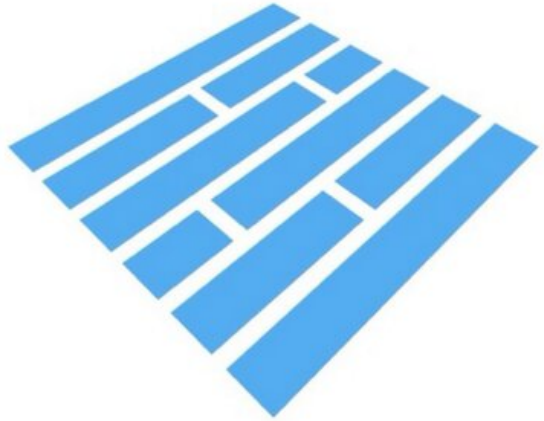
SMART Goals



Pitfalls
You have no specific outcome to measure
Haven't decided how to measure your outcome
Challenging but not attainable due to lack of data
Trying to solve a problem that has no business justification
Without deadline, your work will unlikely to be accomplished

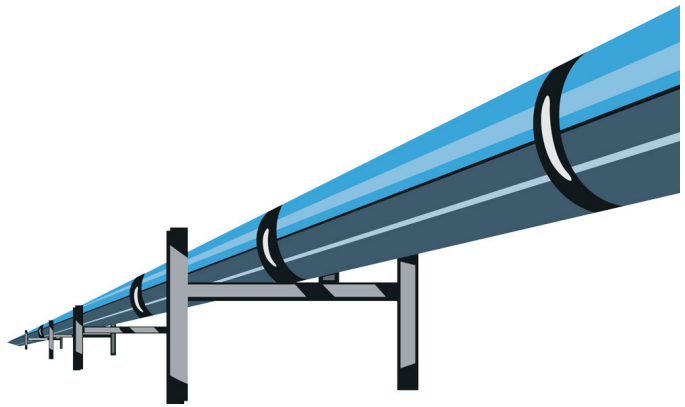
Data Acquisition

Use Parquet



- Space saving: 6x vs CSV
- I/O gain: 3x vs CSV

Use Pipeline

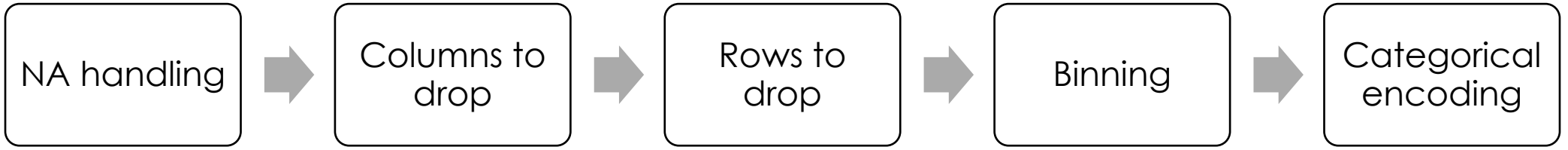


- Pandas: `pdpipe`
- R: `pipeliner`

Use Spark

- Distributed computing
- Prediction is the goal

What is Pipeline?




```
import pdpipe as pdp
```

Create the Stages

```
drop_name = pdp.ColDrop("Name")
binar_label = pdp.OneHotEncode("Label")
map_job = pdp.MapColVals("Job", {"Part": True, "Full":True, "No": False})
```

Create the Pipeline

```
pipeline = pdp.PdPipeline([drop_name, binar_label, map_job])
```

Apply Pipeline

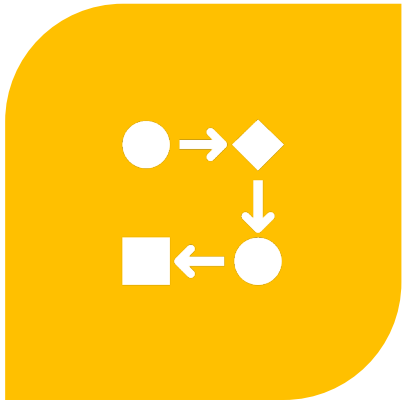
```
df = pipeline(df)
```

https://github.com/shaypal5/pdpipe

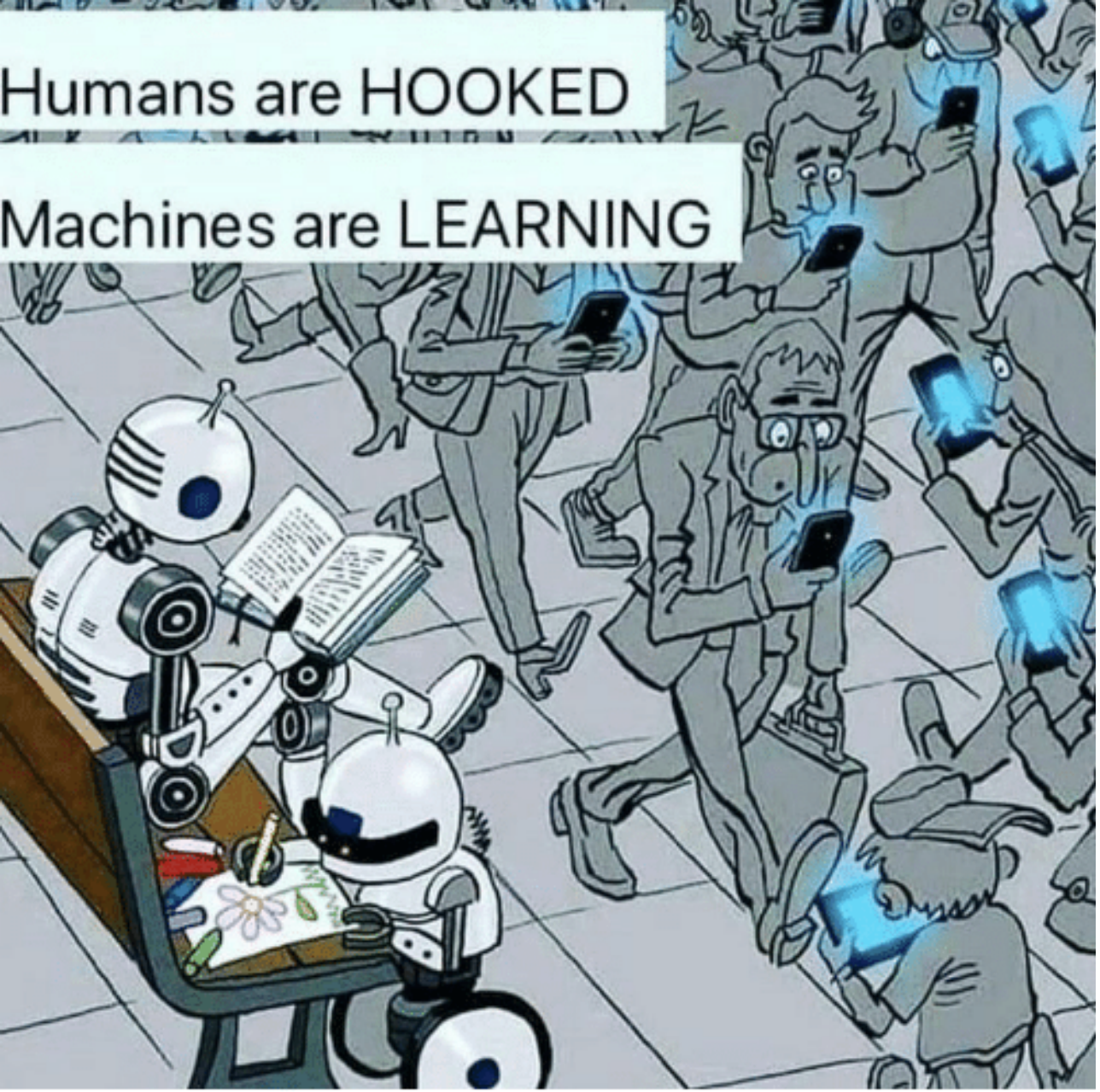
Analysis, Commutation, Iteration

ANALYZE WITH APPROPRIATE METHODS AND MODELS

COMMUNICATE RESULTS TO MANAGER OR STAKEHOLDERS



ITERATION AND REPETITION OF ANALYSIS + COMMUNICATION



SL, ML, DL

Statistical Learning

Machine Learning

Deep Learning

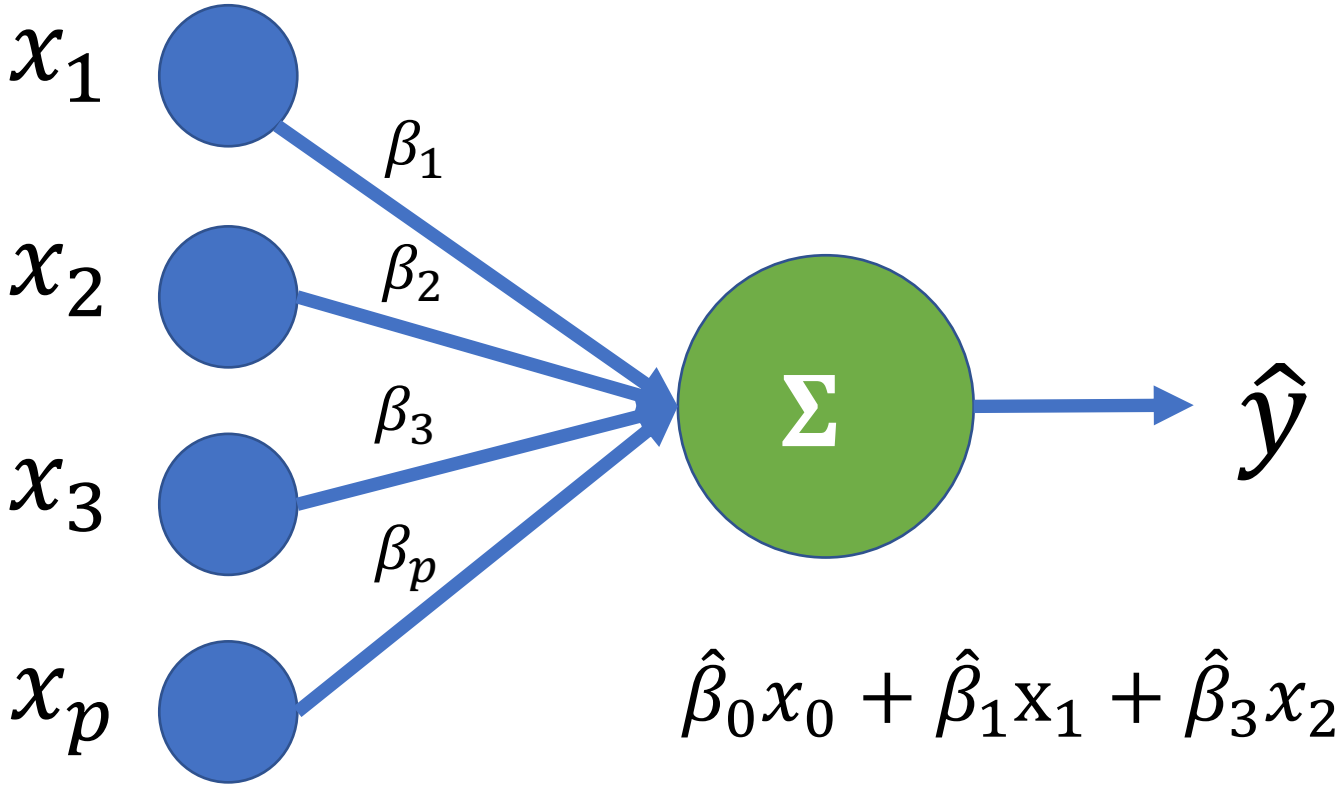
SL vs ML

SL	ML
• Primarily inferential • Operates on assumptions • Developed mostly for smaller data sets • Explicit reliance on relationship between variables and their interactions	• Primarily for prediction • No explicit assumptions • Generally suitable for big data sets • Mostly ignores such relations—automatically figures them out


CAN I KNOW WHAT IS
DEEP LEARNING?


WHAT IF I TOLD YOU IT IS
REGRESSION

Statistical Learning (Regression)



$$\Sigma(y - \hat{y})^2$$

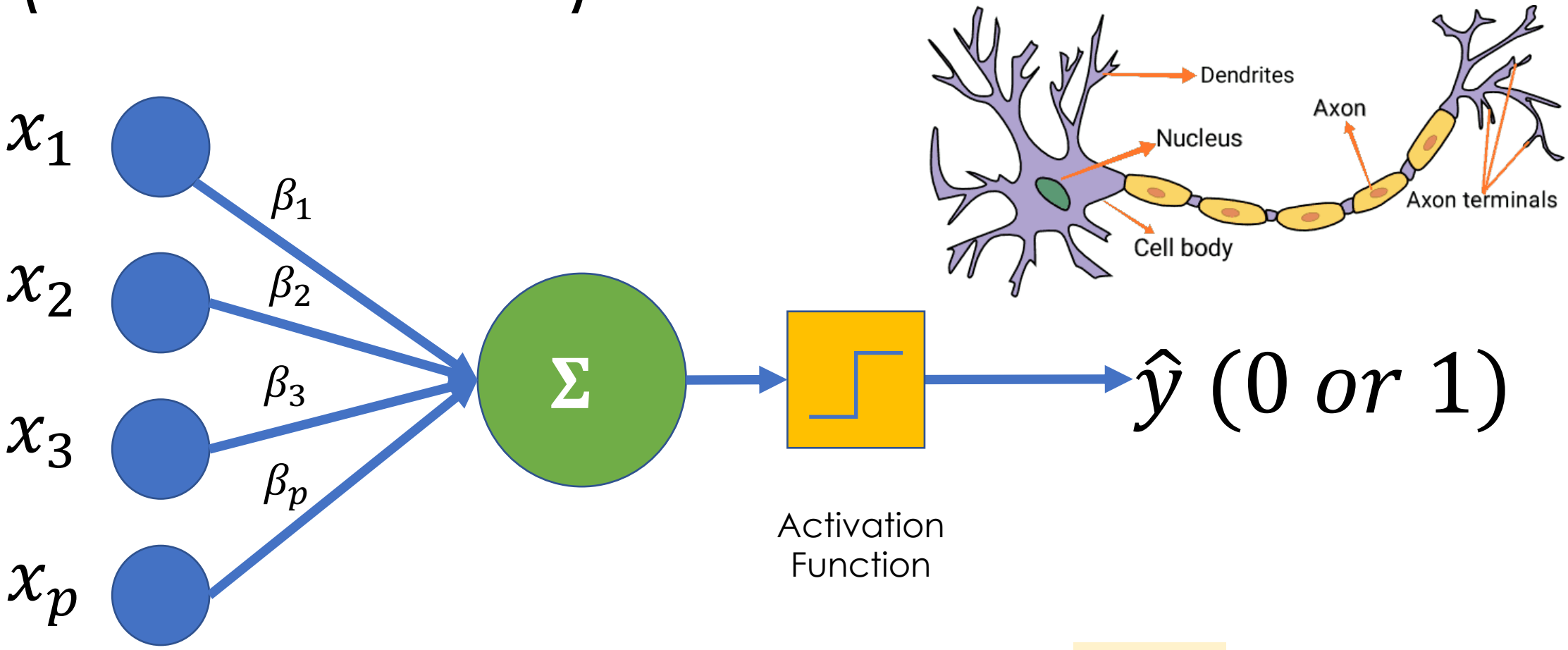
Minimize the loss function

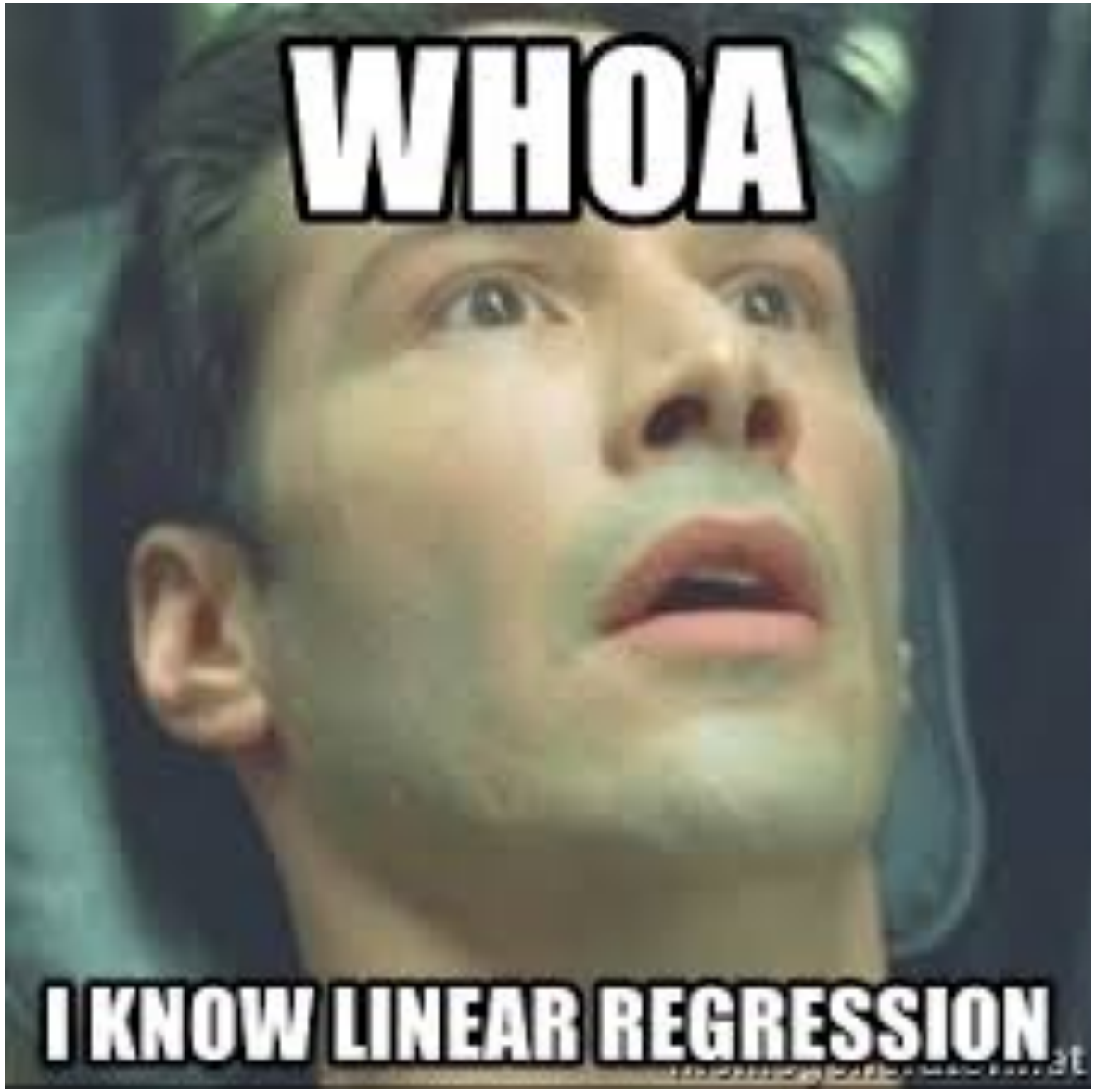
$$\hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_2 + .. + \hat{\beta}_p x_p$$

Input

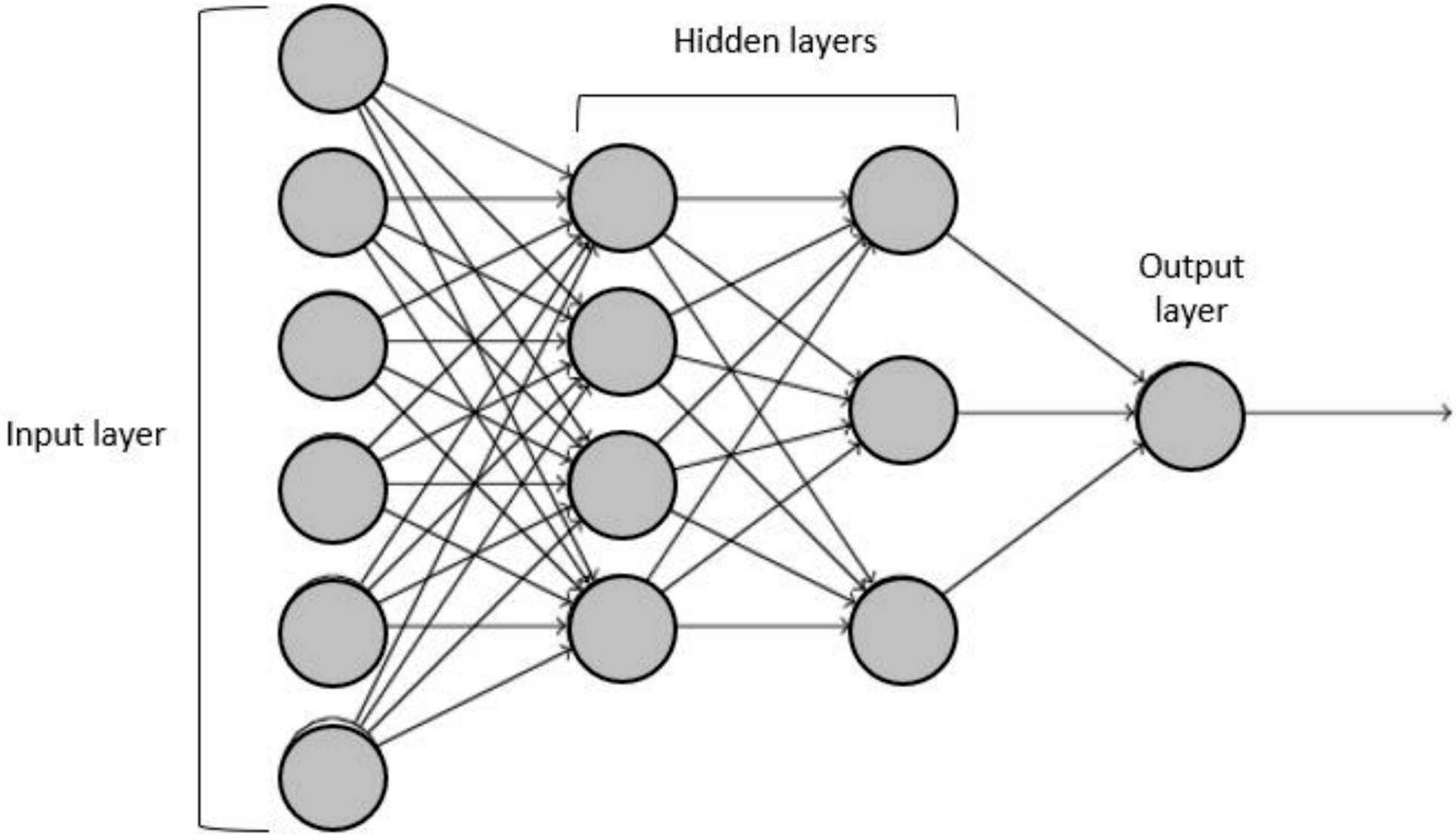
Output

Artificial Neural Network (Classification)



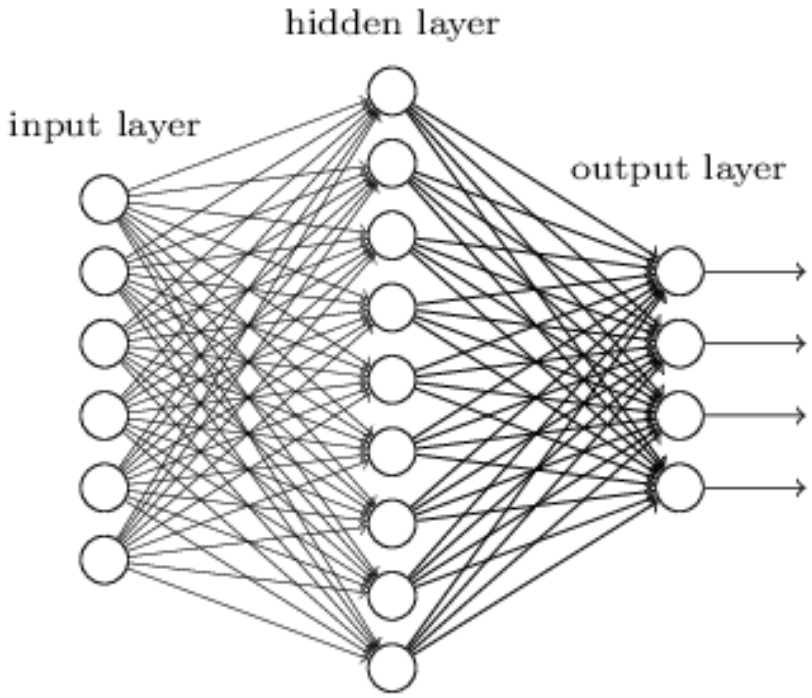


Just dip it into many neurons and through many layers

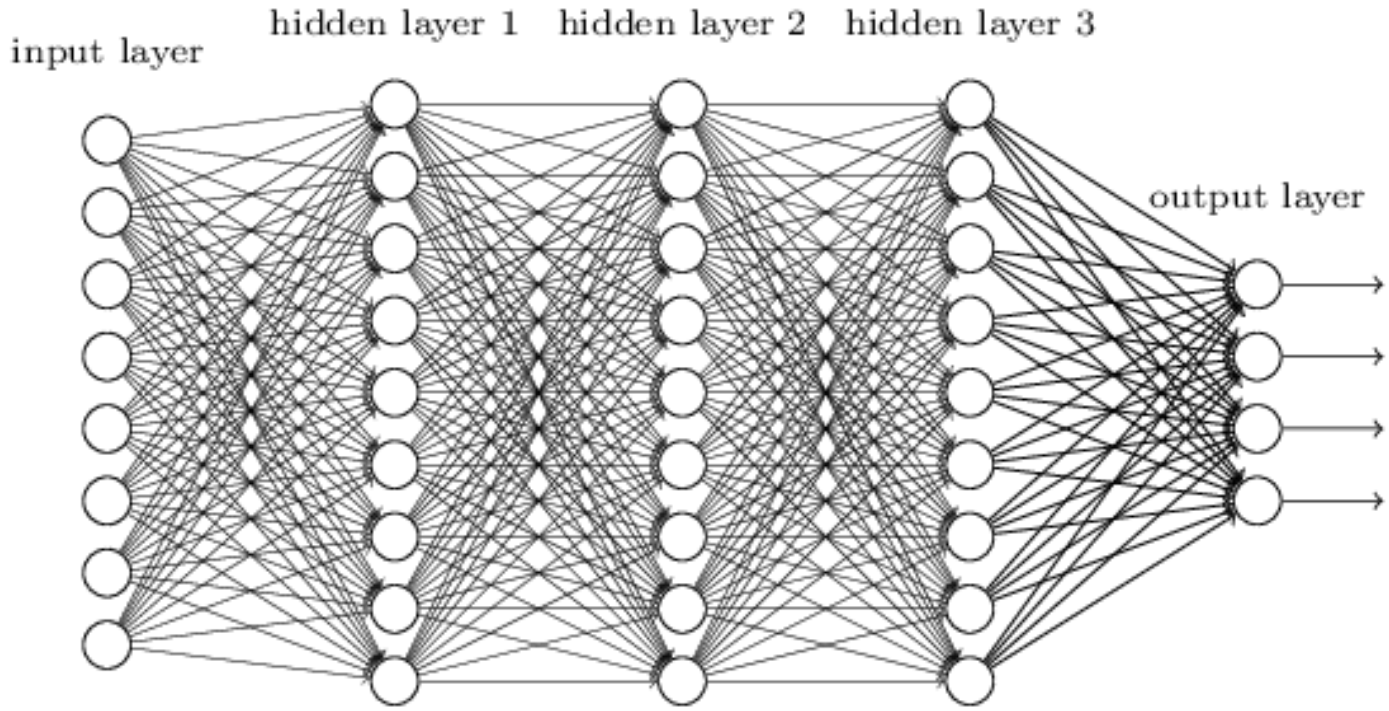


https://images.app.goo.gl/ehAQZFTj9SHGEK4A7

"Non-deep" feedforward neural network



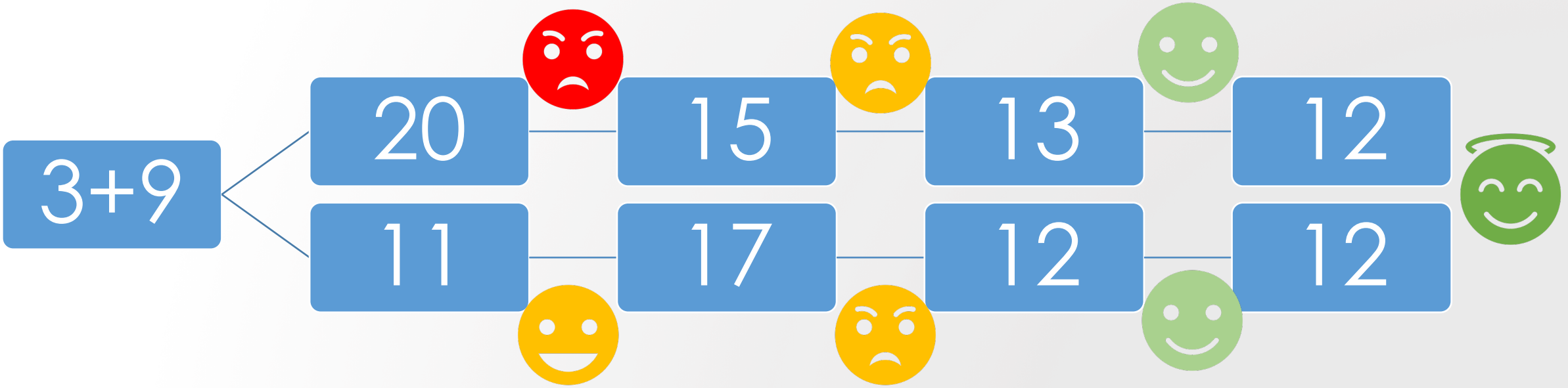
Deep neural network



https://images.app.goo.gl/WtLX8s5byqXGBUGM8

Question is, how do you estimate the parameters of a neural network?

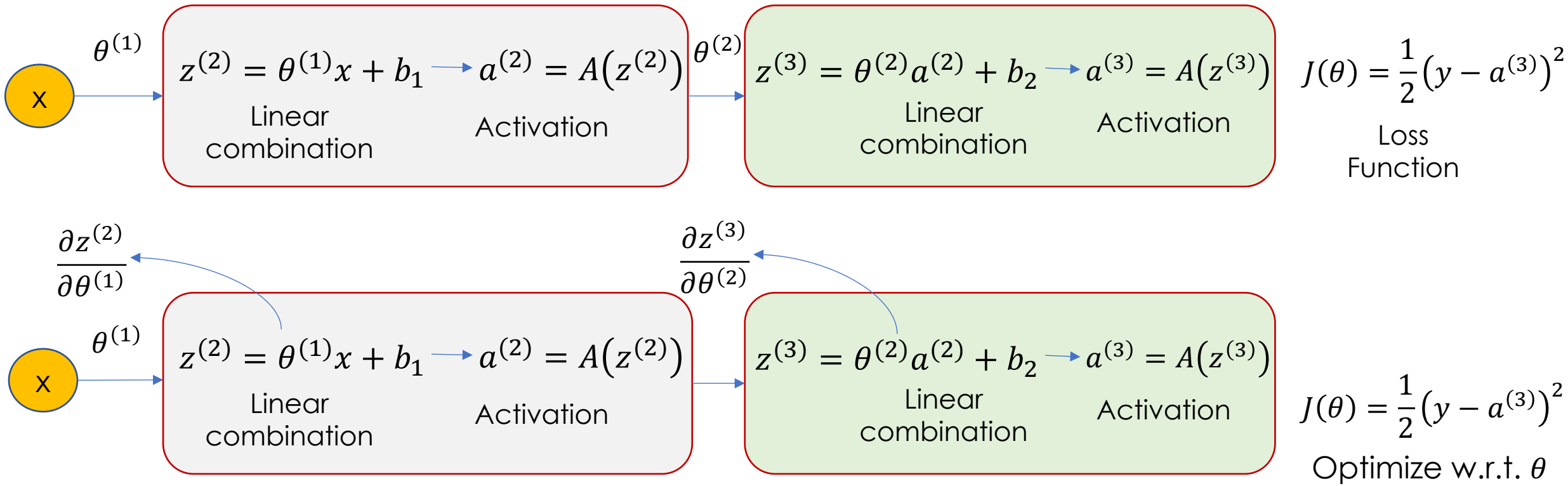
When your mom asks you…



Knowing the representation Vs figuring it out

when you start machine learning without calculus

INPUT



$$\frac{\partial a^{(2)}}{\partial z^{(2)}} \leftarrow \frac{\partial z^{(3)}}{\partial a^{(2)}} \leftarrow \frac{\partial a^{(3)}}{\partial z^{(3)}} \leftarrow \frac{\partial J(\theta)}{\partial a^{(3)}}$$

Updating the Weights

For a given learning rate λ

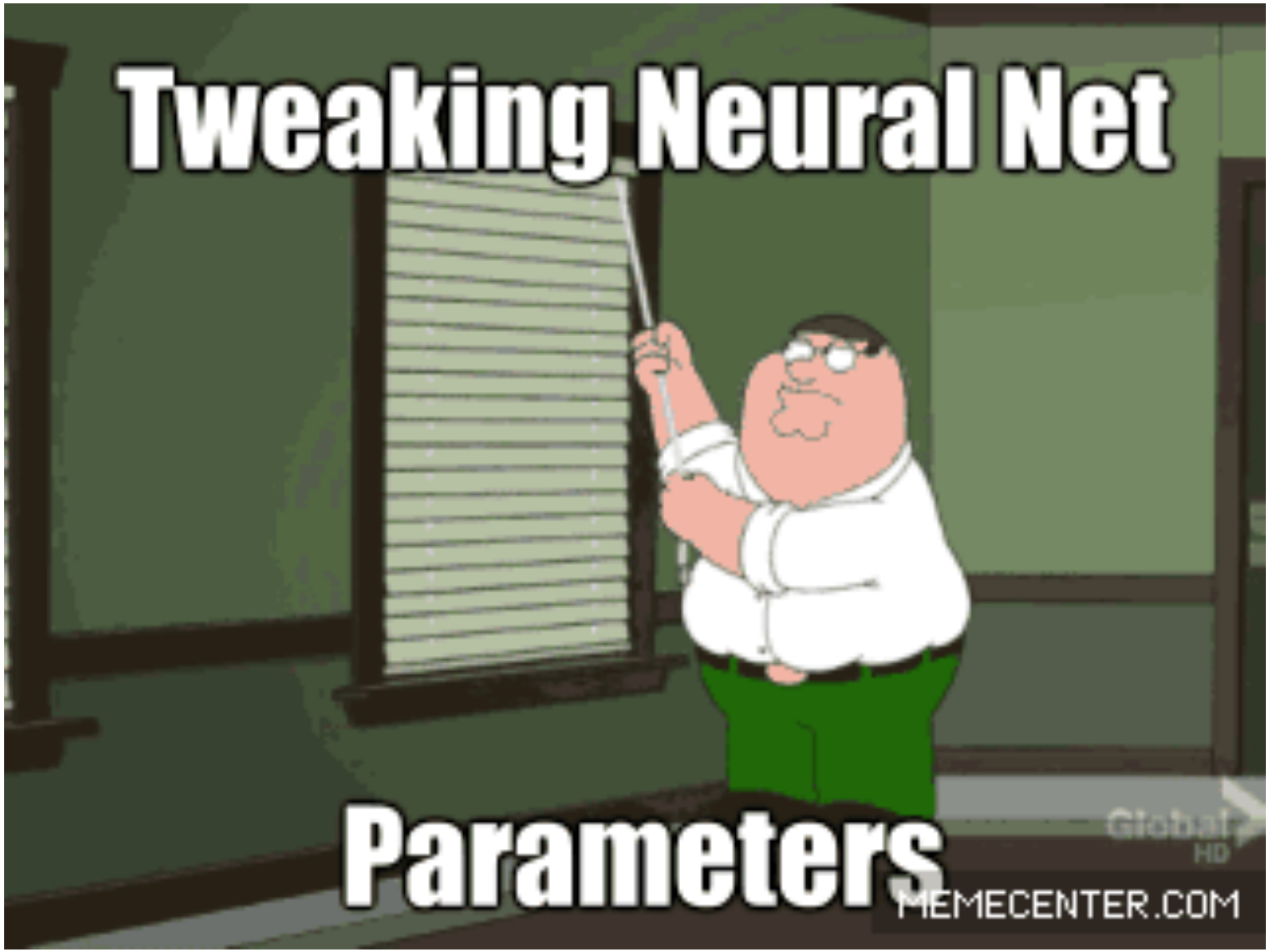
Initialize $\boldsymbol{\theta}$

While stopping criterion not met

Forward pass through the network and compute gradient
$\hat{\boldsymbol{g}} = \frac{1}{m}\nabla_{\theta}\,(Loss\ Function)$

Update parameter $\boldsymbol{\theta}_{new} \leftarrow \boldsymbol{\theta}_{old} - \lambda * \hat{\boldsymbol{g}}$

End while

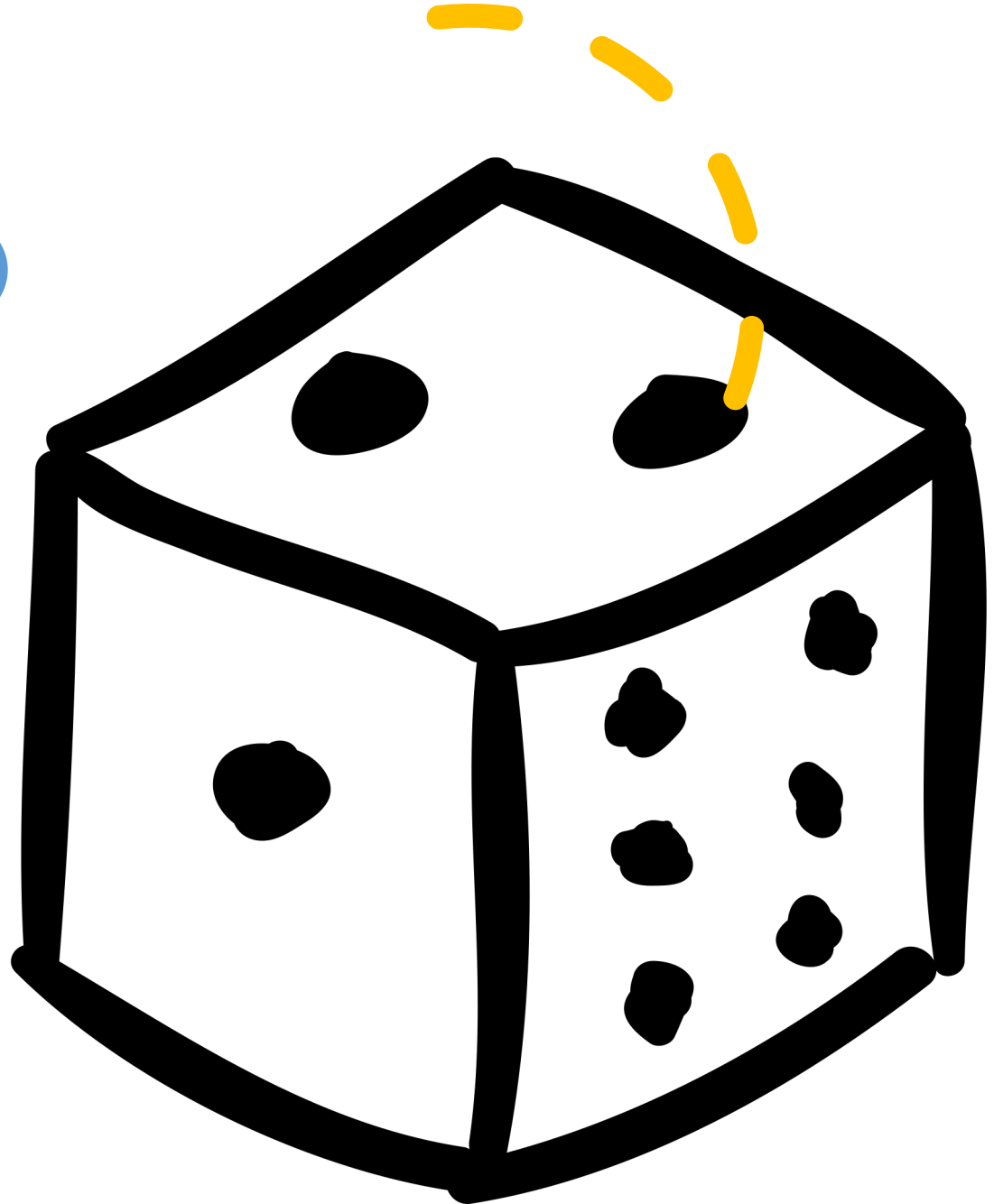

Tweaking Neural Net
Parameters
MEMECENTER.COM

A little about Loss Function

- Maximum likelihood principle– the mother of all loss functions
- Concretely, choose the parameters of the model such that the likelihood of the observed data is maximized
- In other words, what would have been the true parameters that led to the observed data

Dice Example

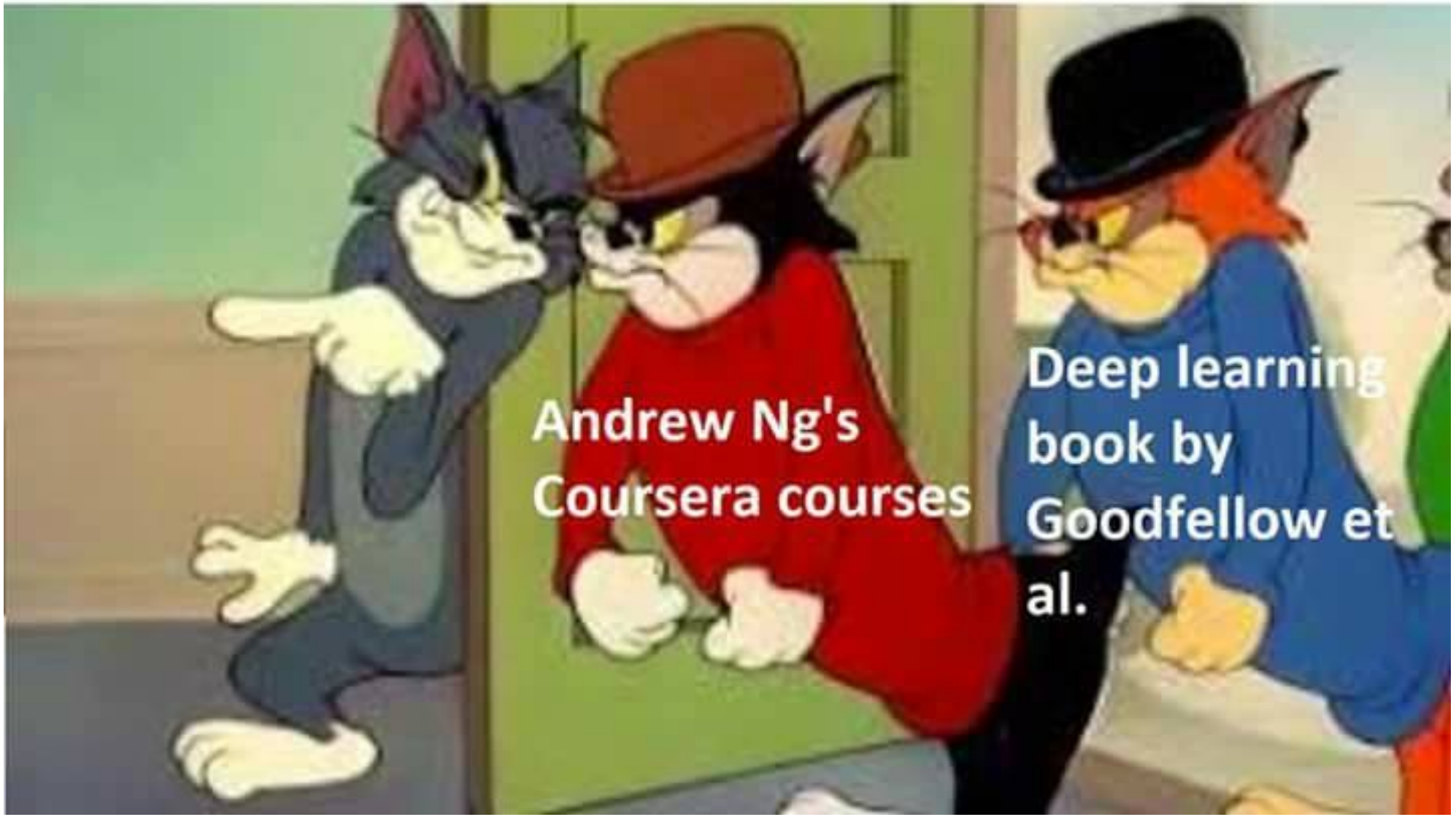
- Consider a custom-made 6-sided dice
  - You do not know how many 6s are there
- Someone rolled it 10 times and 2 sixes appeared
- The question is– how many 6-signs are there in that dice?

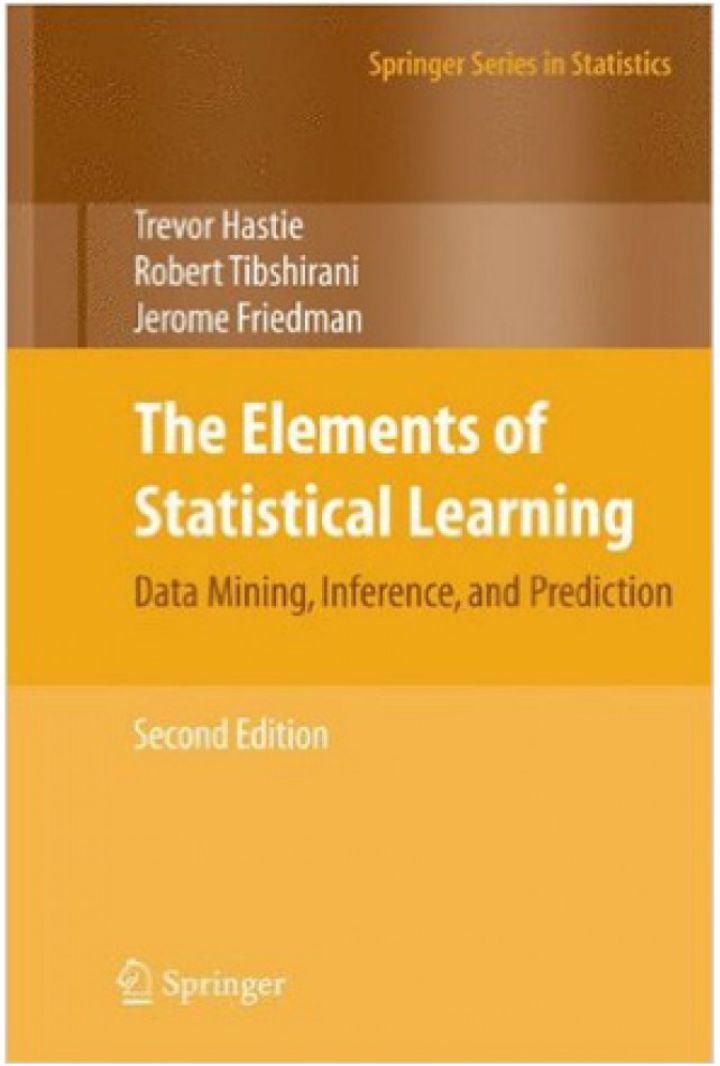


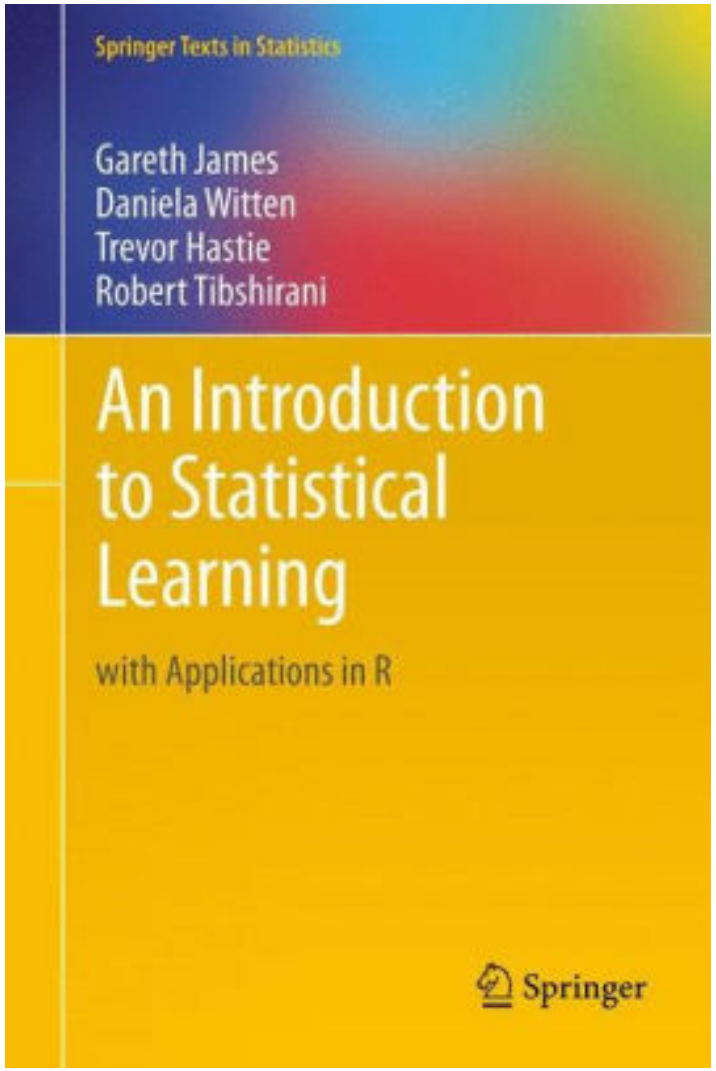
- What are the possible values of number of 6s?
  - 1, 2, 3, 4, 5
  - Why not 0 and 6?
- Lets calculate the following probabilities

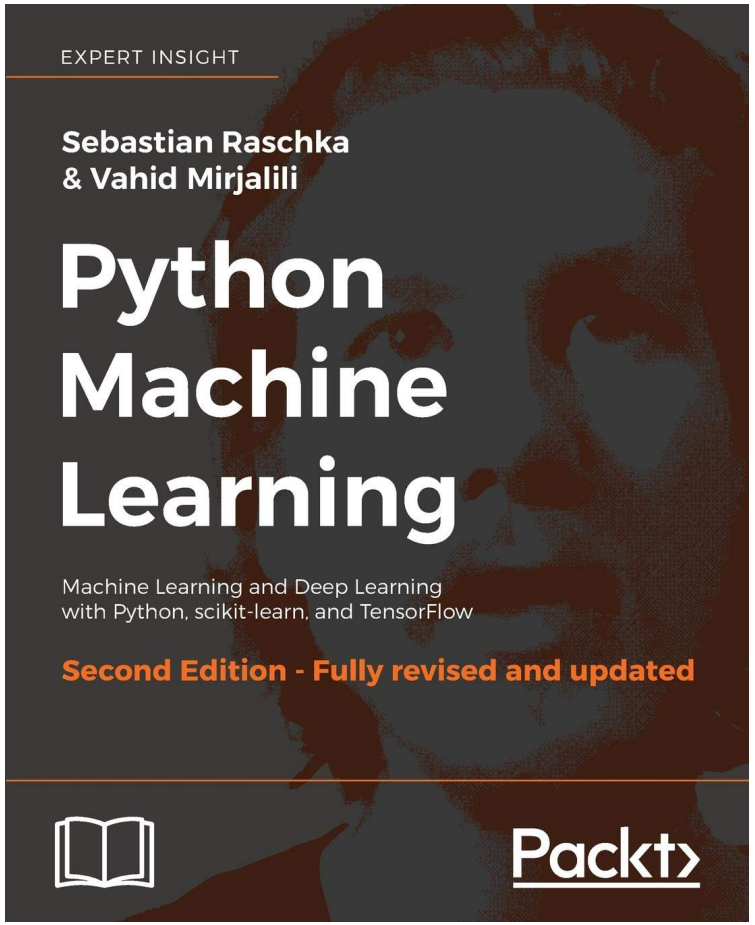
	Prob(Success)	Prob (getting 2 sixes in 10 rolls)
If there is no 6	0/6	0
If there is one 6	1/6	0.291
If there are two 6	2/6	0.195
If there are three 6s	3/6	0.044
If there are four 6s	4/6	0.003
If there are five 6s	5/6	~0.001
If there are six 6s	6/6	0

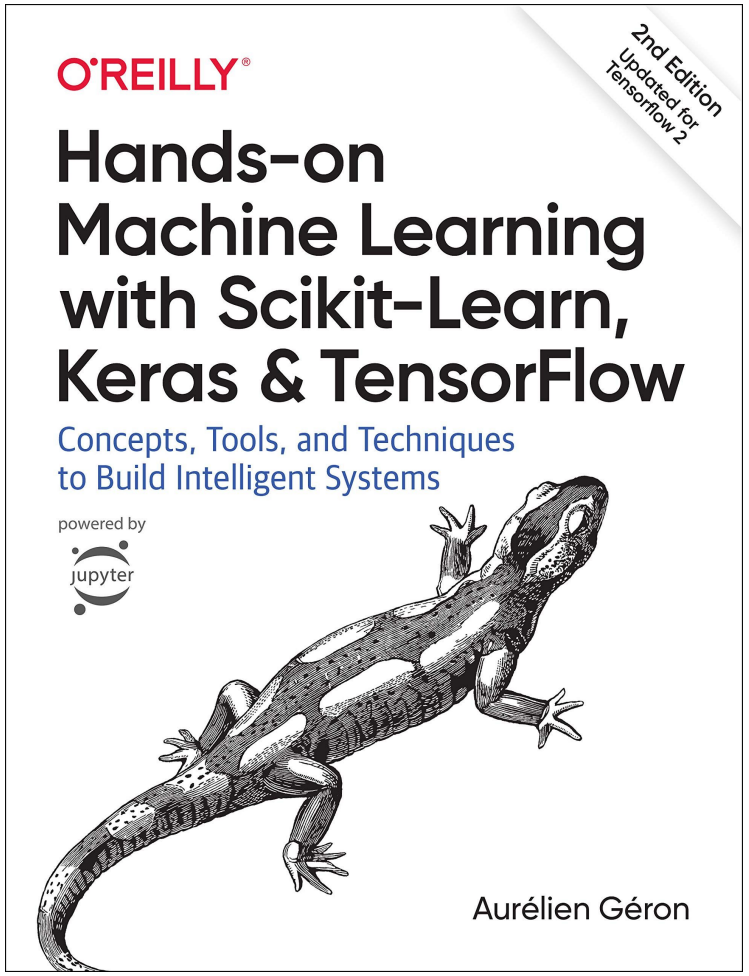
When a beginner asks for recommendations for studying machine learning

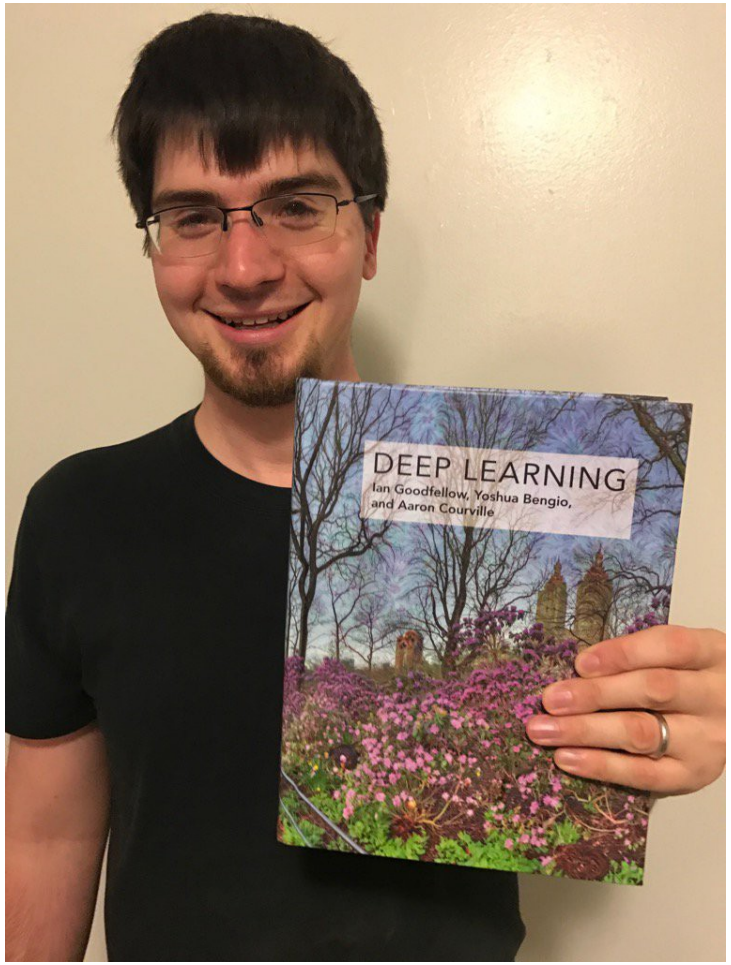



Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

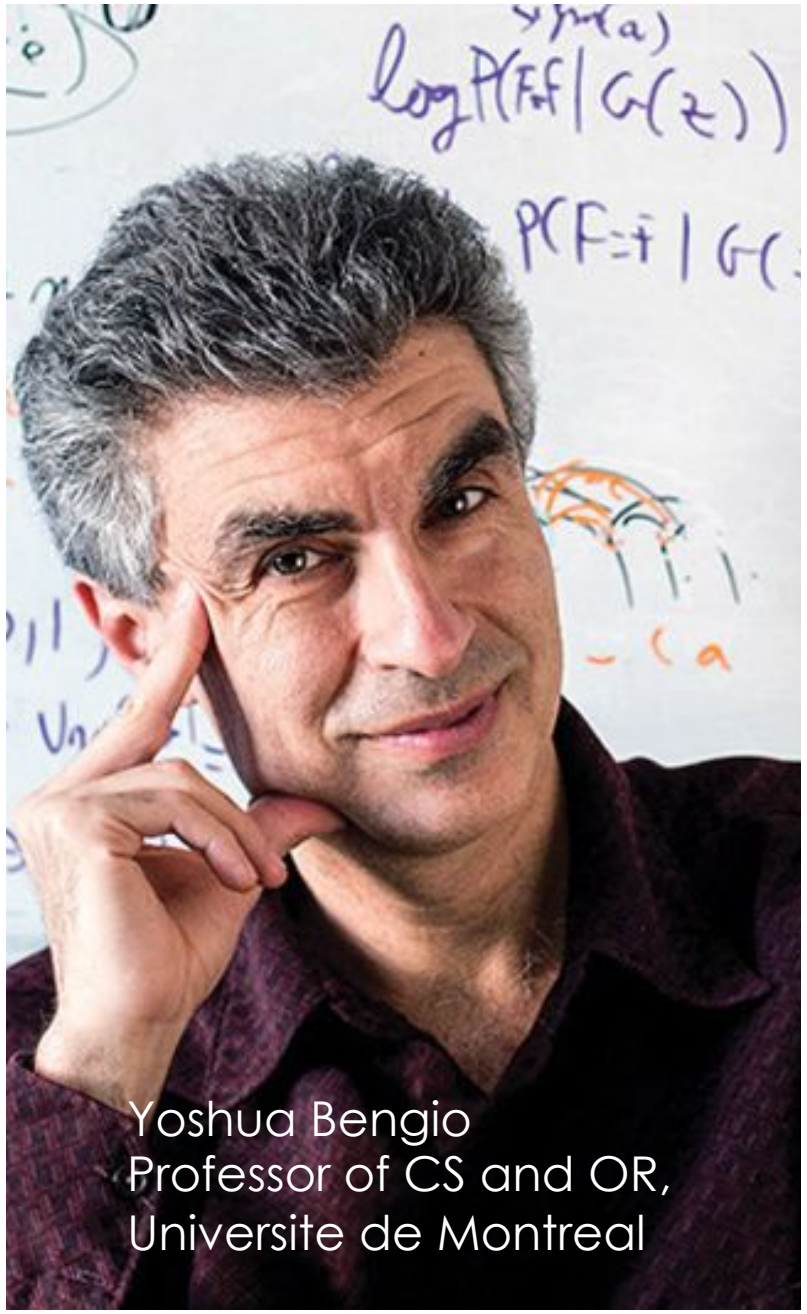

EXPERT INSIGHT
Sebastian Raschka
& Vahid Mirjalili
Python Machine Learning
Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow
Second Edition - Fully revised and updated
Packt


O'REILLY
2nd Edition
Updated for Tensorflow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville


Geoffrey Hinton
Professor of CS, U of T

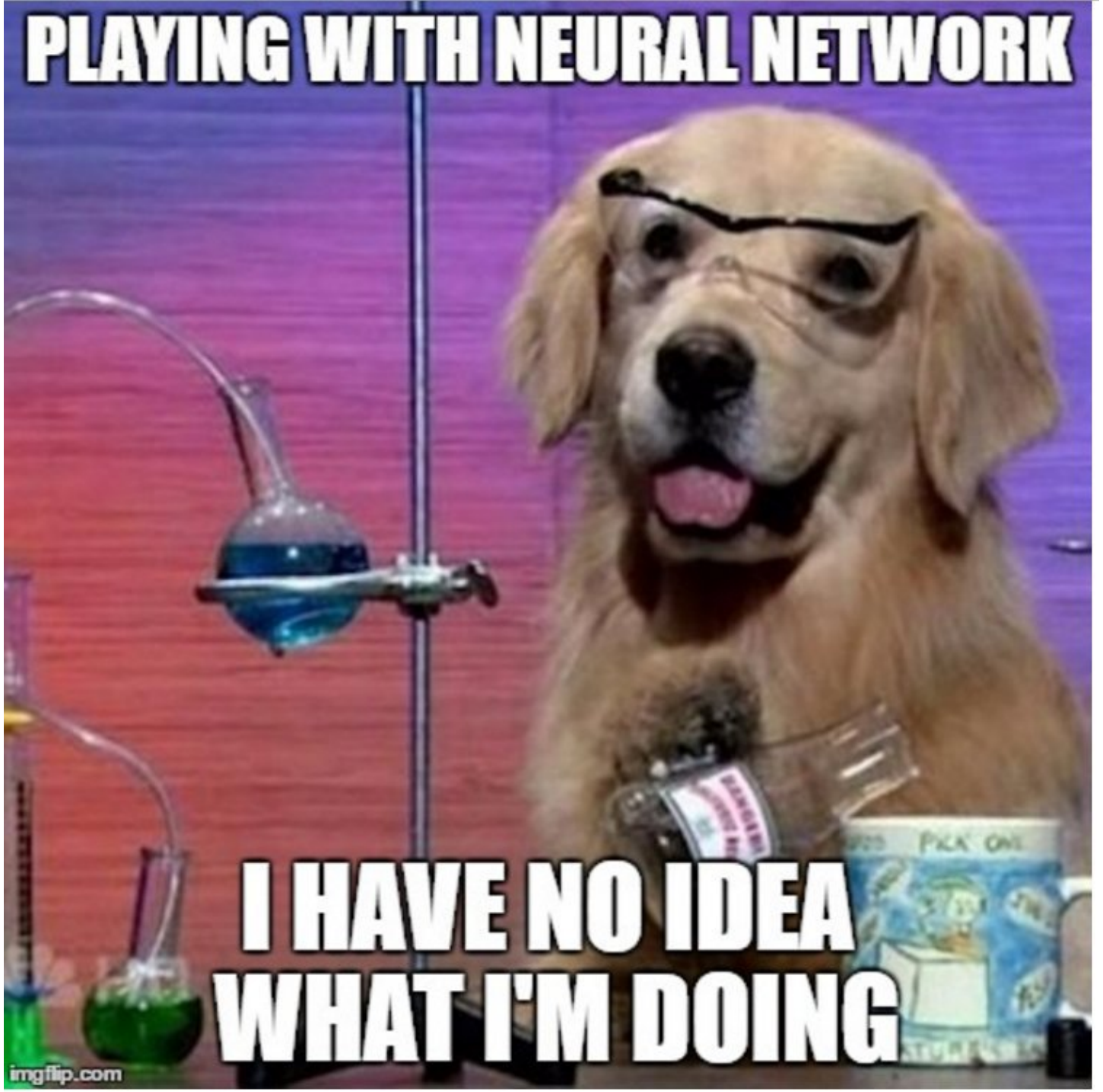

Yann LeCun
Chief AI Scientist, Facebook


Yoshua Bengio
Professor of CS and OR,
Universite de Montreal

Deep Learning Demonstration

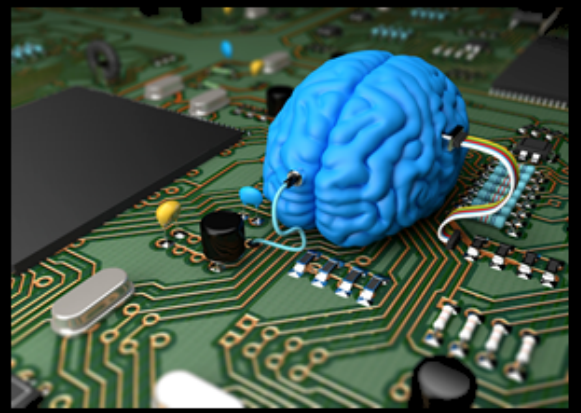
TensorFlow

https://playground.tensorflow.org


PLAYING WITH NEURAL NETWORK
I HAVE NO IDEA
WHAT I'M DOING
imgflip.com

Deep Learning

What society thinks I do



What my friends think I do

What other computer scientists think I do

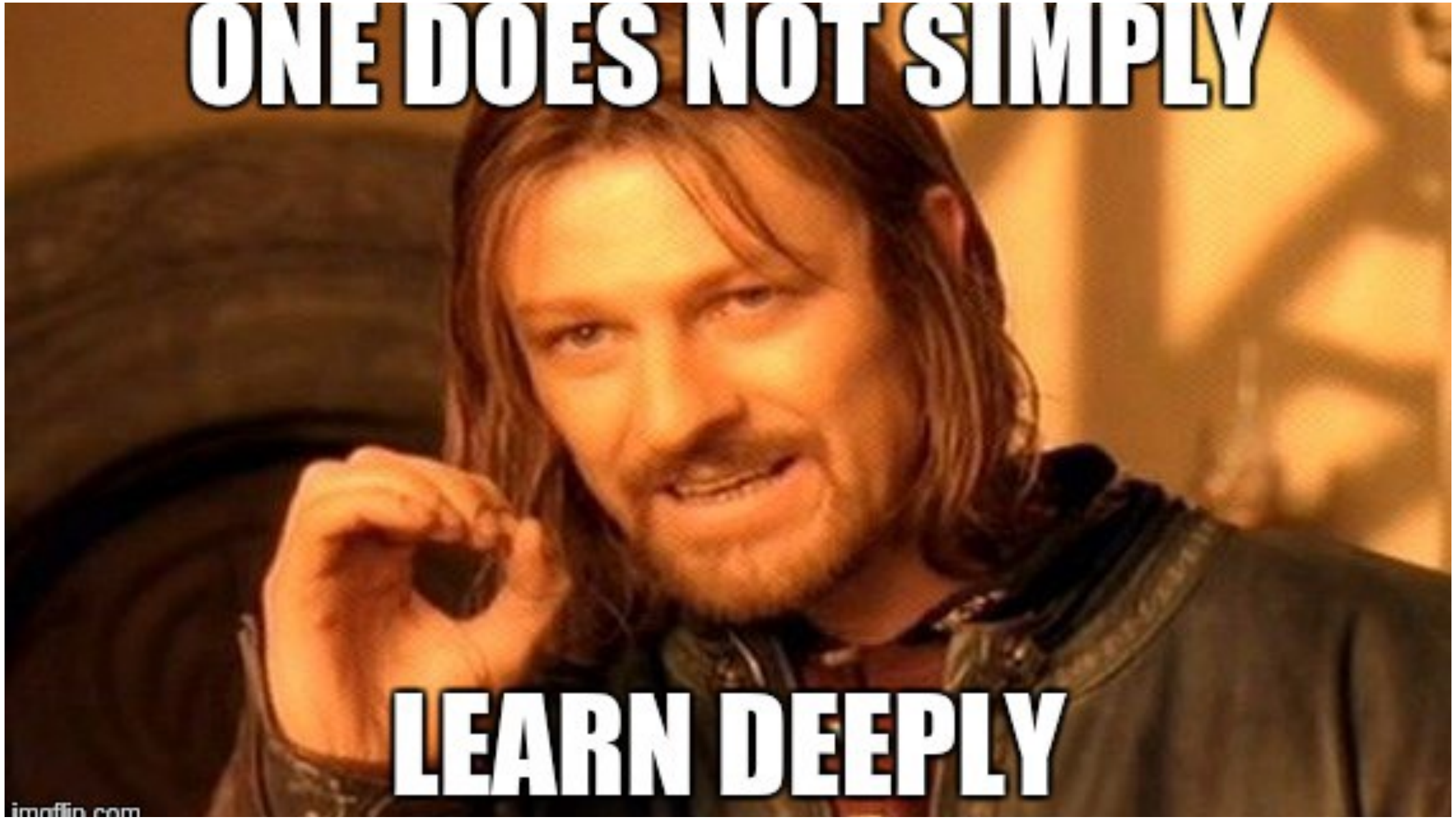
What mathematicians think I do

What I think I do

```
In [1]:
import keras
Using TensorFlow backend.
```

What I actually do

মশা মারতে কামান দাগা


ONE DOES NOT SIMPLY
LEARN DEEPLY
imgflip.com


SAY DEEP LEARNING
ONE MORE TIME

Credits

- Some Icons are made by Freepik https://www.flaticon.com/authors/freepik
- Data Lake Image Source: https://blog.equinix.com/blog/2016/11/10/why-companies-are-jumping-into-data-lakes/
- Stepped wedge trial schematic picture from https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0760-6
- Pipeline picture from https://images.app.goo.gl/aMJMxDqTfETDESzM6

- Apache Spark logo image from https://images.app.goo.gl/moGwHgHrAYfPiXjn7
- Back propagation schema adapted from https://medium.com/usf-msds/deep-learning-best-practices-1-weight-initialization-14e5c0295b94
- Biological neuron figure from https://images.app.goo.gl/bGgAjyBq2hkM86V76