HOW DID I BECOME A DATA SCIENTIST?

PARVANEH SHAFIEI

EY, 3 MAY 2018

WHO I AM

- ✓ Bachelor: Software engineering, Iran
- ✓ Past experience: Web & Software developer
- ✓ Master: Computer science in Polimi

Research assistant

Data scientist

Data scientist

Senior
Data Scientist

Founder

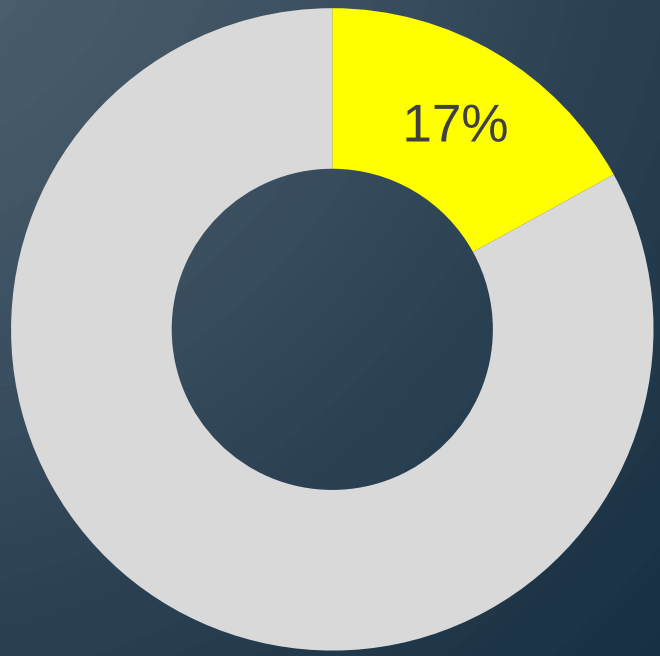
RLADIES?
WHO THEY ARE?

+70 CHAPTERS

+20 COUNTRIES

- Founded July 2017
- 309 Rladiers

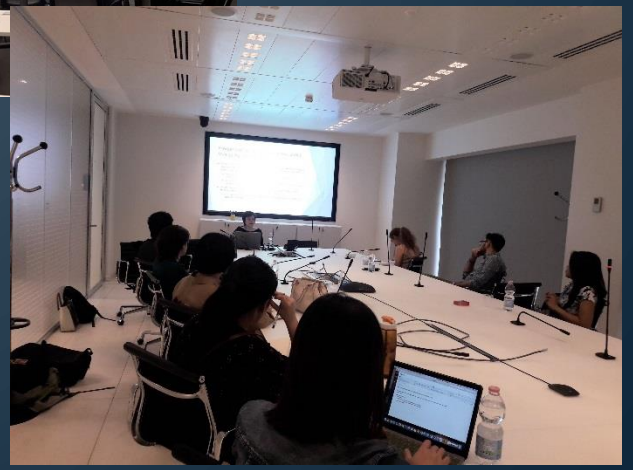
In 2017, Tech startups across the world founded by women



kaggle

Survey by Kaggle in 2017, about data professional

RLADIES MILAN EVENTS



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

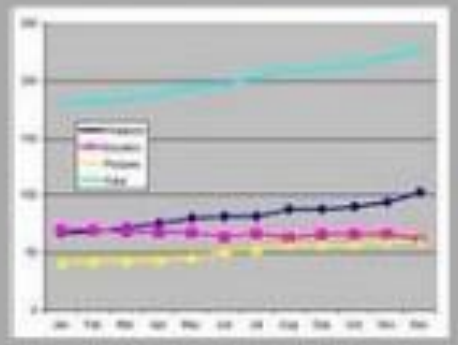
"A data scientist is better at statistic than a software engineer and is better at software engineering than a statistician"

DATA SCIENTIST

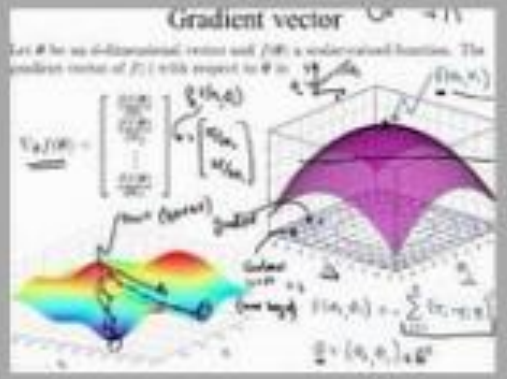
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

What I actually do

Thesis

Predict decision making by analyzing brain signals

Security

- Anomaly network detection

Insurance

- Customer segmentation
- Incident prediction

HR

- Turnover analysis & prediction
- Who are at the risk of leaving
- Root- cause analysis

Finance

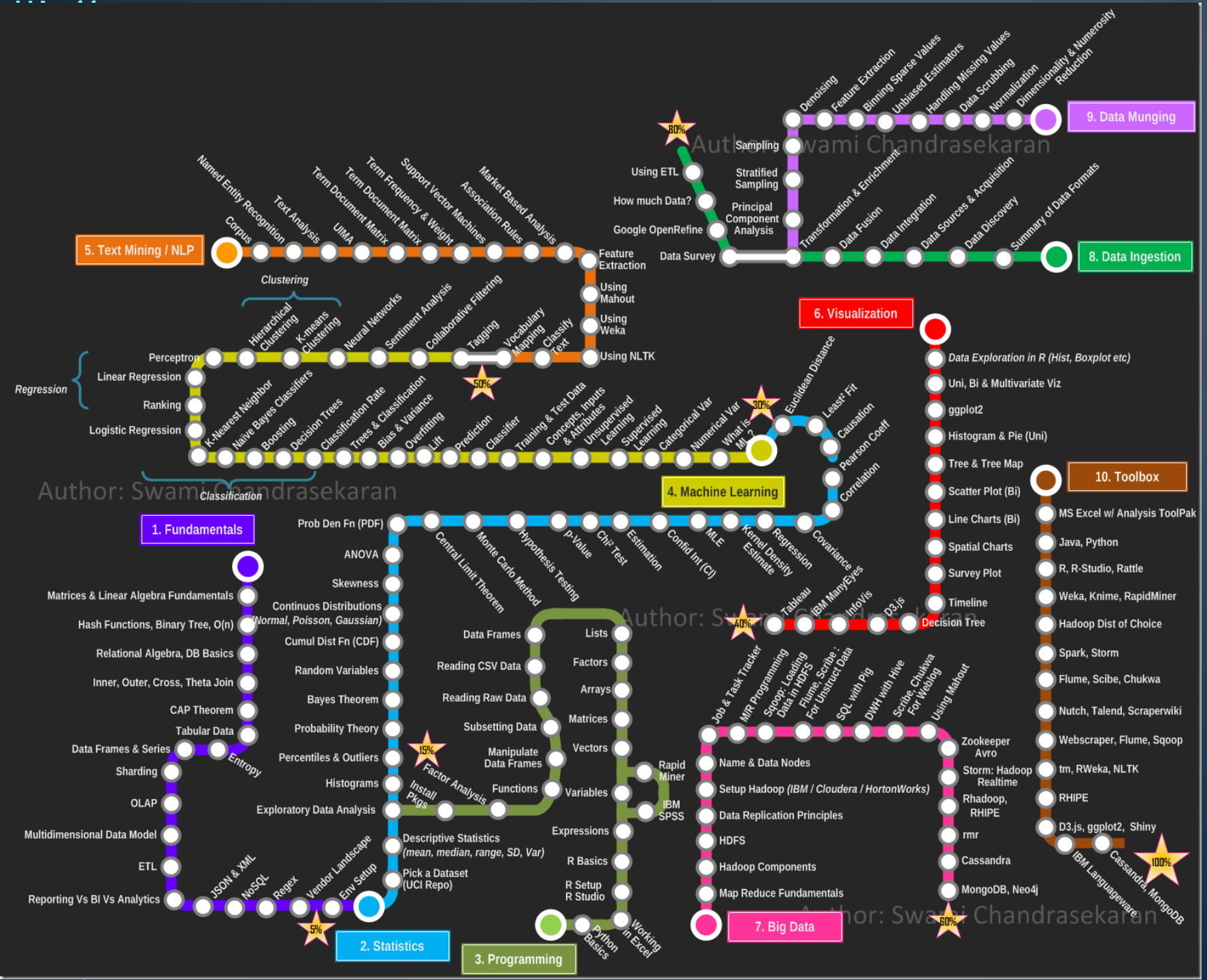
- Data mining & pattern analysis
- Customer segmentation & sentiment analysis

Marketing

- Customer segmentation
- Identify who is best for next offer

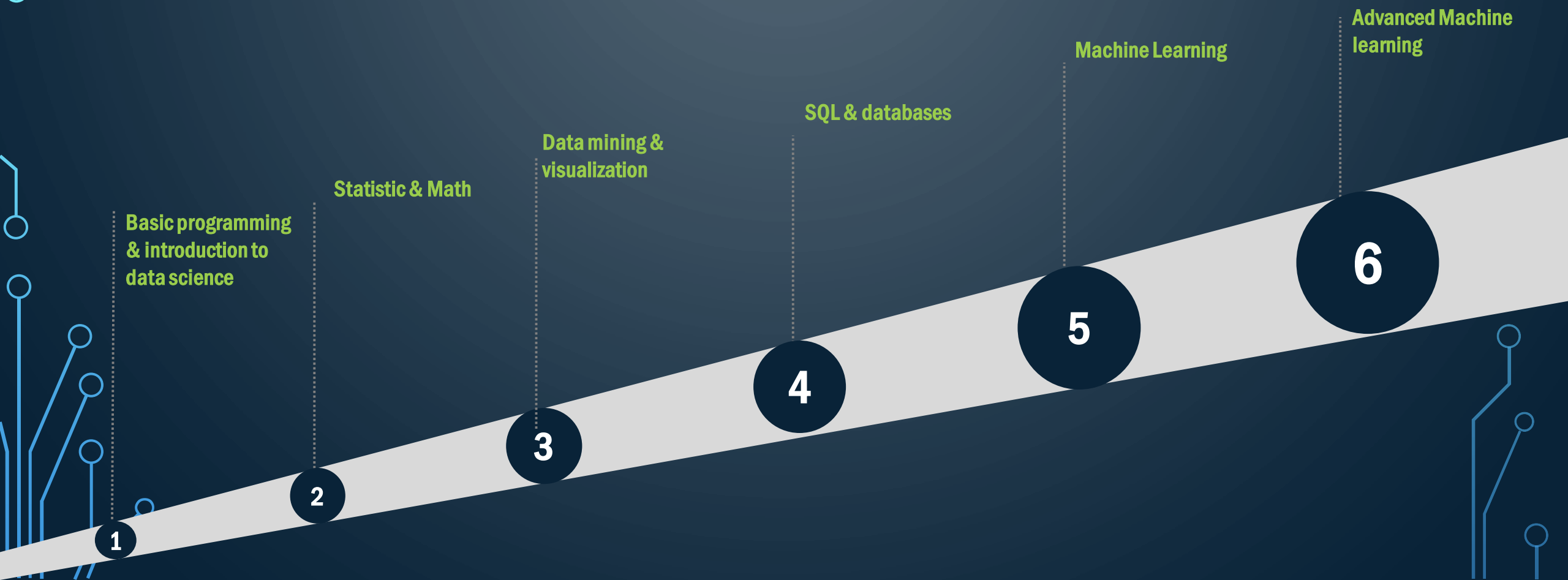
Publishing

- Products order forecasting & planning
- Understand trends

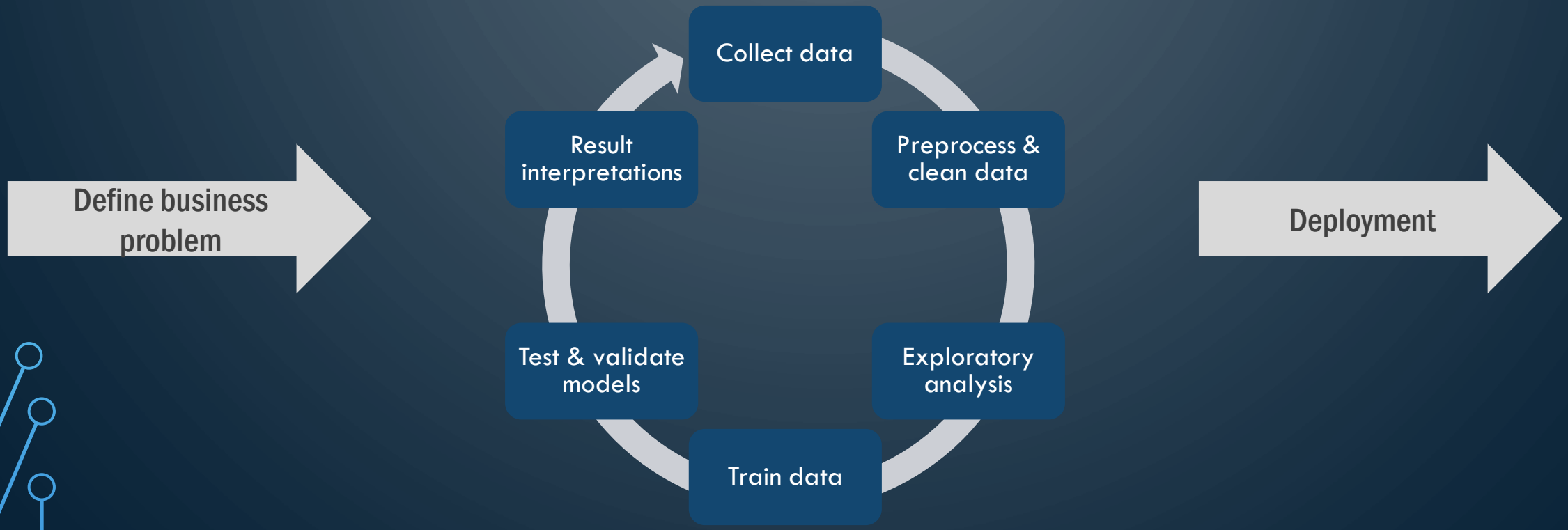


Be ready for the journey!

THERE IS A STARTING POINT BUT NO END ROAD!



OK, I'M ALL SET, BUT HOW CAN I TACKLE A REAL PROBLEM?



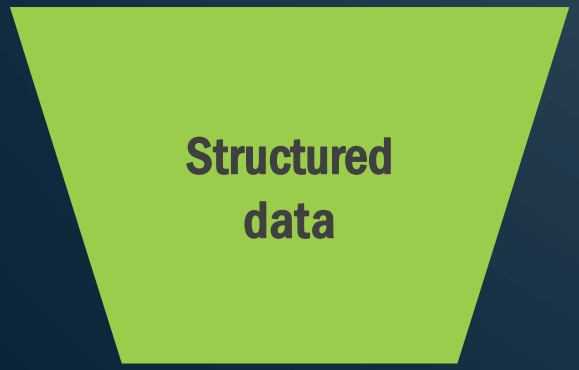
DEFINE THE BUSINESS PROBLEM

- Listen to business's need
- You must be a good listener

- The business you are facing do not know their problem yet? That is even fine too!
- A lot of time, they don't have any idea where to start

COLLECT DATA

Data are in a structured format such as database, csv file,...



Data are in a unstructured format such as text, images, audio and video

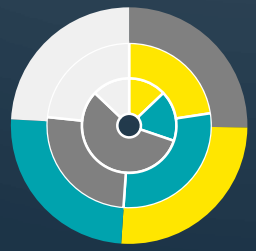
Near 80% of data in organization are unstructured

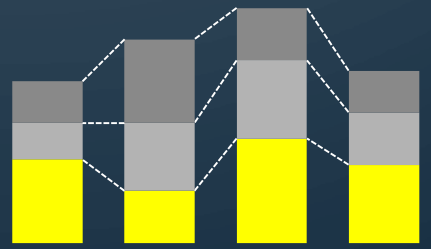
PREPROCESSING & CLEANING DATA

- **Never trust** your data
- **Check consistency & structure** of the data to be sure about its quality
- Fill **missing values**
- **Be aware of noisy data** and treat them well!
- Check **outliers** & remove them
- Apply **data transformation** such as normalization or aggregation

EXPLORATORY ANALYSIS

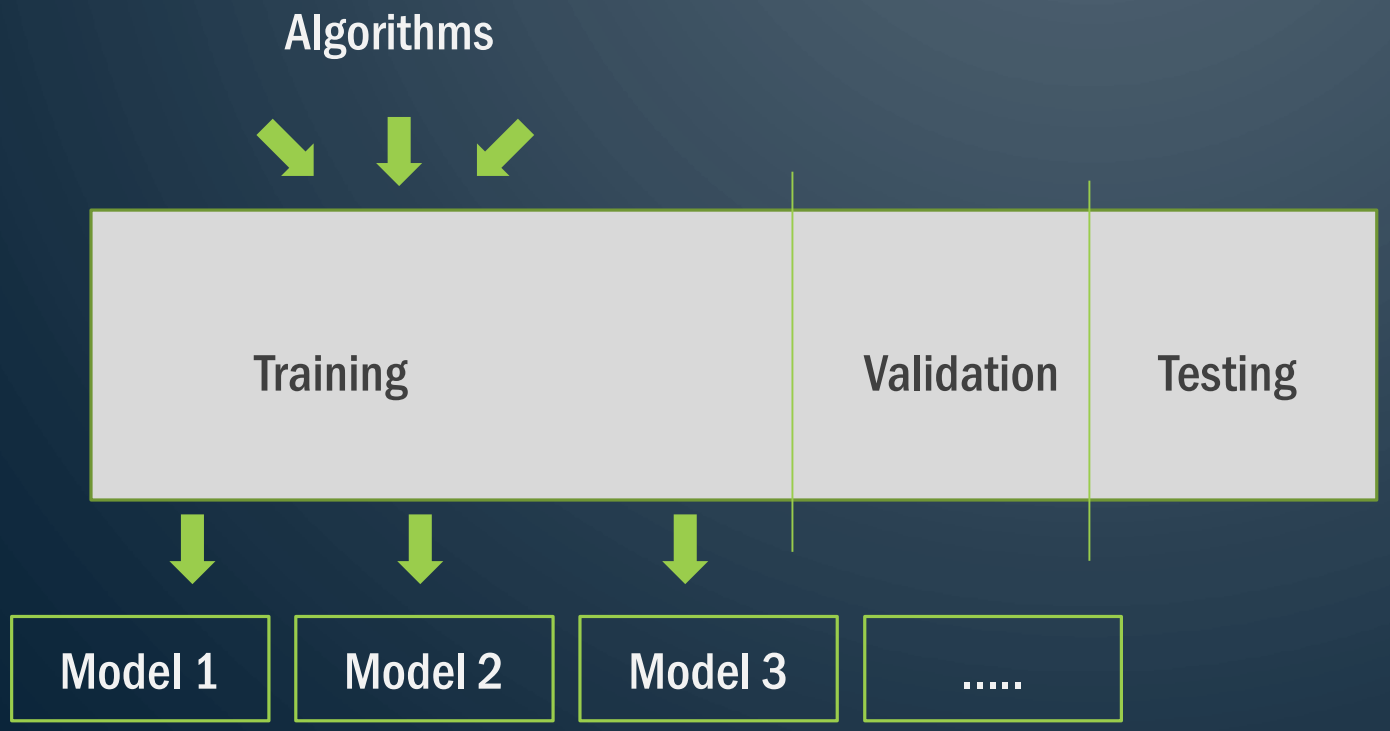
- **Understand and summarize the** data
- **Explore** for insights & hidden patterns
- Investigate the correlations among various variables
- Create some **hypothesis** based on the findings
- **Understand which model & technique to** use for analyzing the data
- **Feature engineering** can happen in this stage too





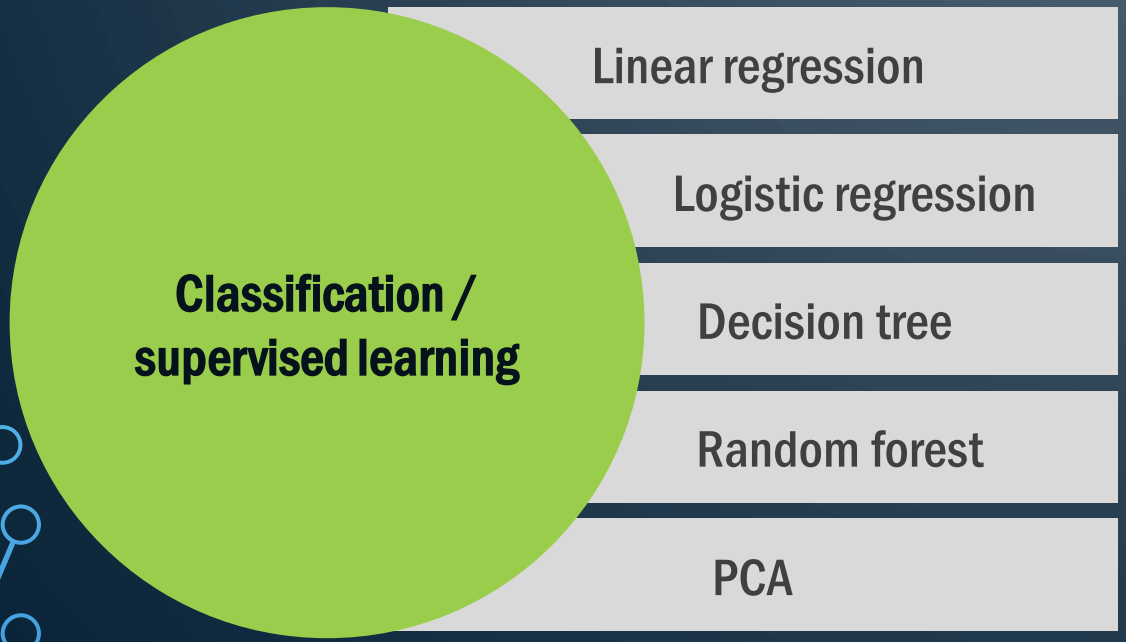
Think about the questions that you are going to answer

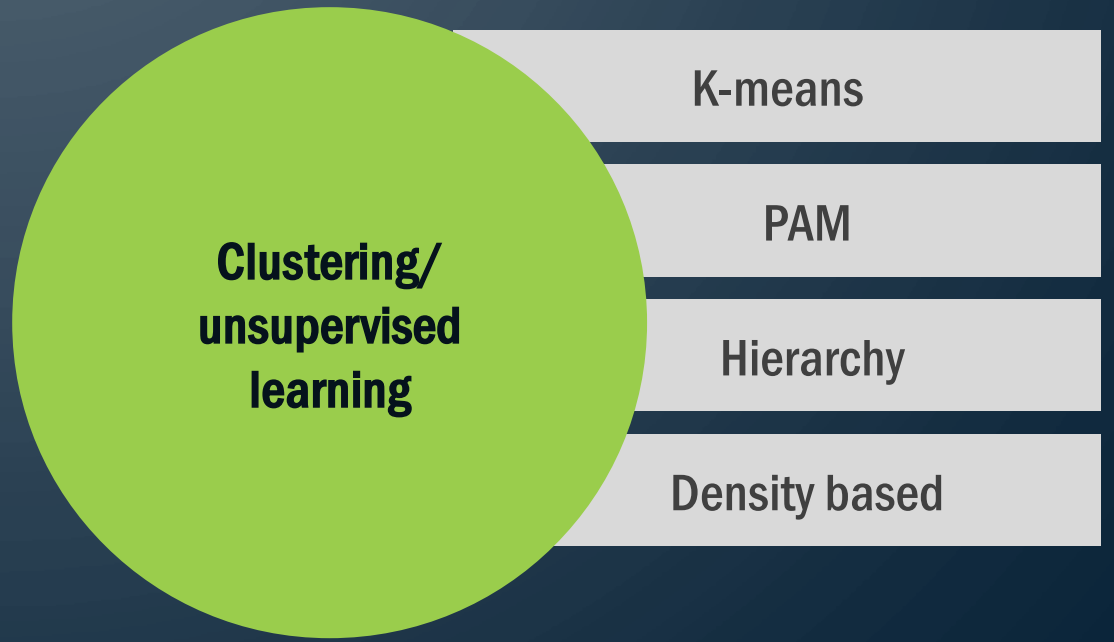
MODEL TRAINING, TESTING & VALIDATION



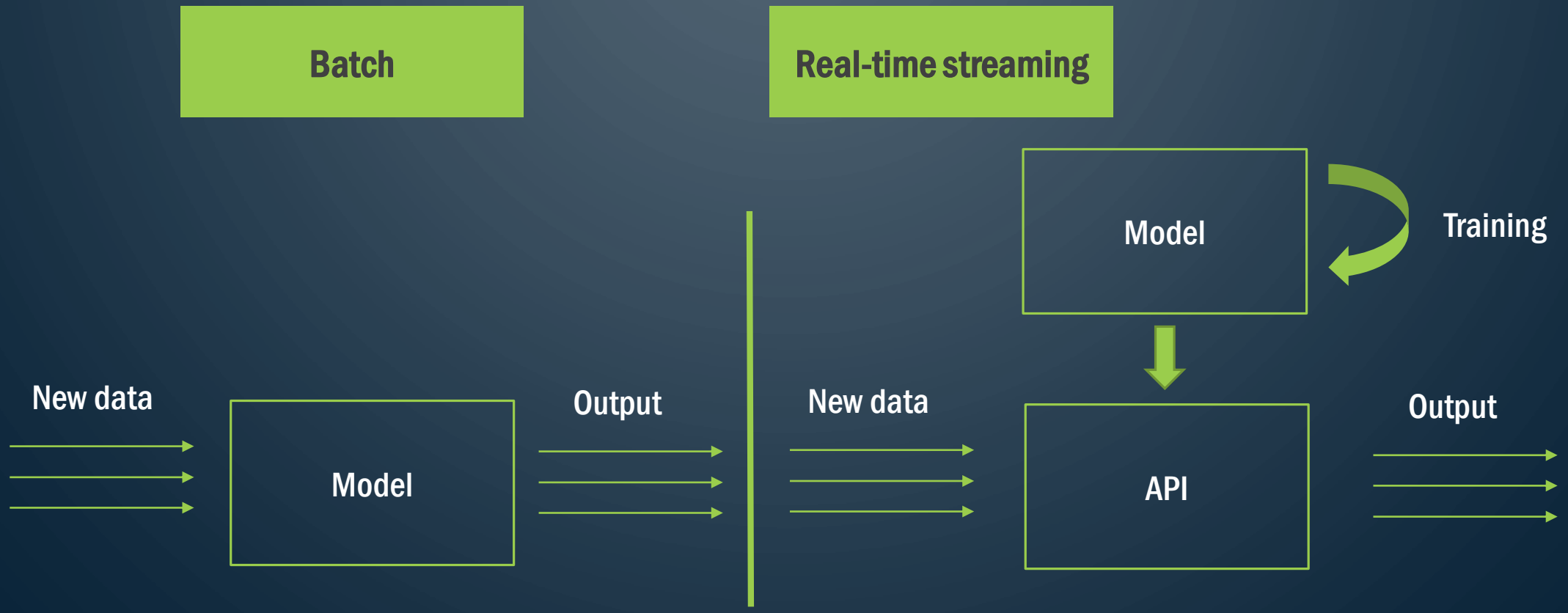
60%, 20%, 20%

ALGORITHMS?


Classification / supervised learning
Linear regression
Logistic regression
Decision tree
Random forest
PCA


Clustering/ unsupervised learning
K-means
PAM
Hierarchy
Density based

DEPLOYMENT


Batch
Real-time streaming
Model
Training
New data
Model
Output
New data
API
Output

"MORE DATA BEATS BETTER MODELS. BETTER DATA BEATS MORE DATA."

RILEY NEWMAN