Building and Managing a Data Science Team

Session 2

24 March 2018

How Airbnb hire Data Scientists

1. Resume – Basic qualifications and experience

2. Basic Data Challenge – Validate data experience described in resume

3. In-house Data Challenge with presentation – Main vetting process

4. In-person interviews – Collaboration ability and culture fit

Data Scientists Shortage in Malaysia

- Currently less than 500 experienced Data Scientists
- Many Universities have only just started their Data Science programs
- Global talent shortage leads to brain drain

Steps towards implementing Data Science into your company

- Step 1: Assemble your A-Team
- Step 2: Understand the Industry
- Step 3: Study the Data Available
- Step 4: Ask the Right Questions
- Step 5: Form an Analysis Plan
- Step 6: Partner Up
- Step 7: Application & Integration

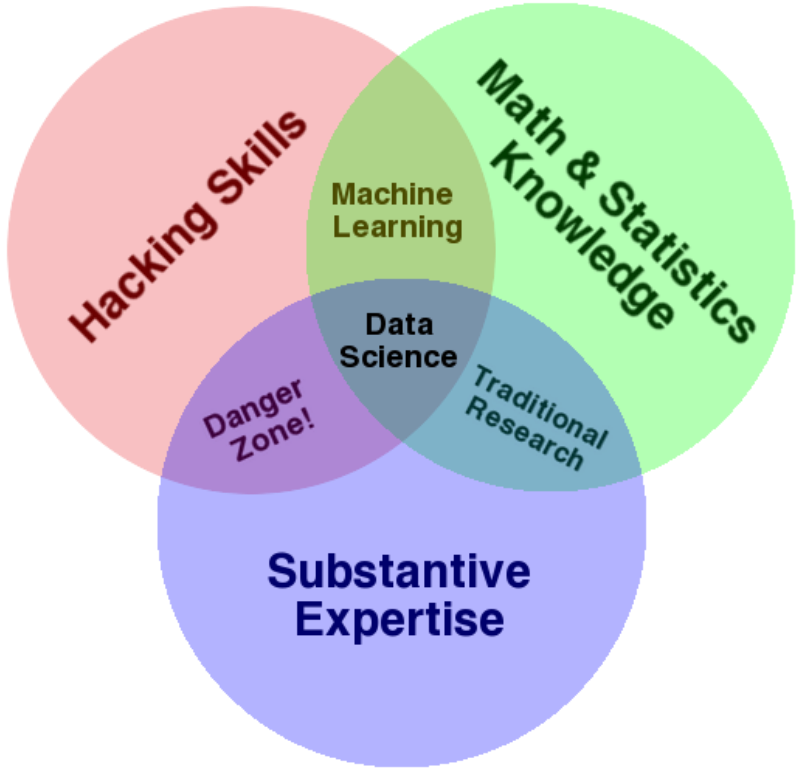
Step 1: Assemble your A-Team

- Experienced Data Scientists are currently very rare and expensive
- You need a team with the right set of skills
- This does not need to be an costly process

Data Science is a Team Sport

Compile a team of old and new staff to fill out the following roles:

1. Data Engineer
2. Coder
3. Statistician / Actuary
4. Business Analyst
5. Industry Expert



Data Engineer

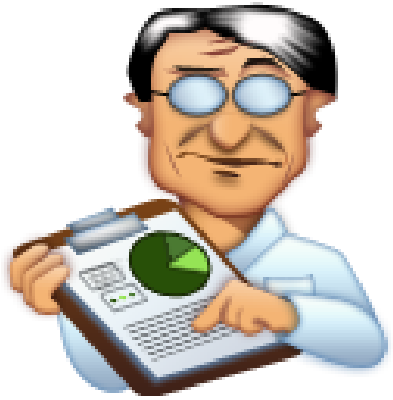
- Database background (e.g. mySQL, MongoDB)
- Focus on automation
- Understand data processing and scale
- Work closely with coders to build data processes

Coder

- Experienced with data science languages like R and Python
- Able to pull data from databases and to build models (e.g. SQL)
- Hacker mentality
- Look at portfolio instead of resume (e.g. GitHub)
- Work closely with statisticians to understand the science behind the model

Statistician / Actuary

- Mathematics and Statistics background
- Understand the scientific methods of the model used by the coder
- Willing to find answers on their own
- Not afraid of Big Data
- Most likely to be considered the “Data Scientist” of the team



Business Analyst

- Work closely with stakeholders to build use cases
- Focus on visualization and story-telling
- Must be creative and have strong communication skills
- Ensures that the project is on schedule and has intended impact

Industry Expert

- Most likely the most senior member in the team
- Most likely to be the team lead
- Many years of practical experience in that particular industry
- Able to translate personal experiences to the rest of the team who may be younger and more technically-oriented
- Needs to be open to new ideas and technologies
- Develop scope and objectives of projects
- Create project plans and share with team

Other pieces of advice…

- Your first hire and the team leader is particularly important
- If you are just starting out, your focus should be on data engineering
- Do not be overly concerned about paper qualifications
- Regular communication is especially important, given the specialized roles
- Rotate roles to explore the full potential of your staff
- Forming partnerships with academia can be useful

Step 2: Understand the Industry

- Follow and learn from world-class experts
- Andrew Ng is the former Chief Scientist at Baidu, a Stanford professor and co-founder of Coursera
- Take online courses and keep up-to-date with the latest Data Science news
- Join Data Science community groups and meet-ups

BFM 89.9 The Business Station

coursera

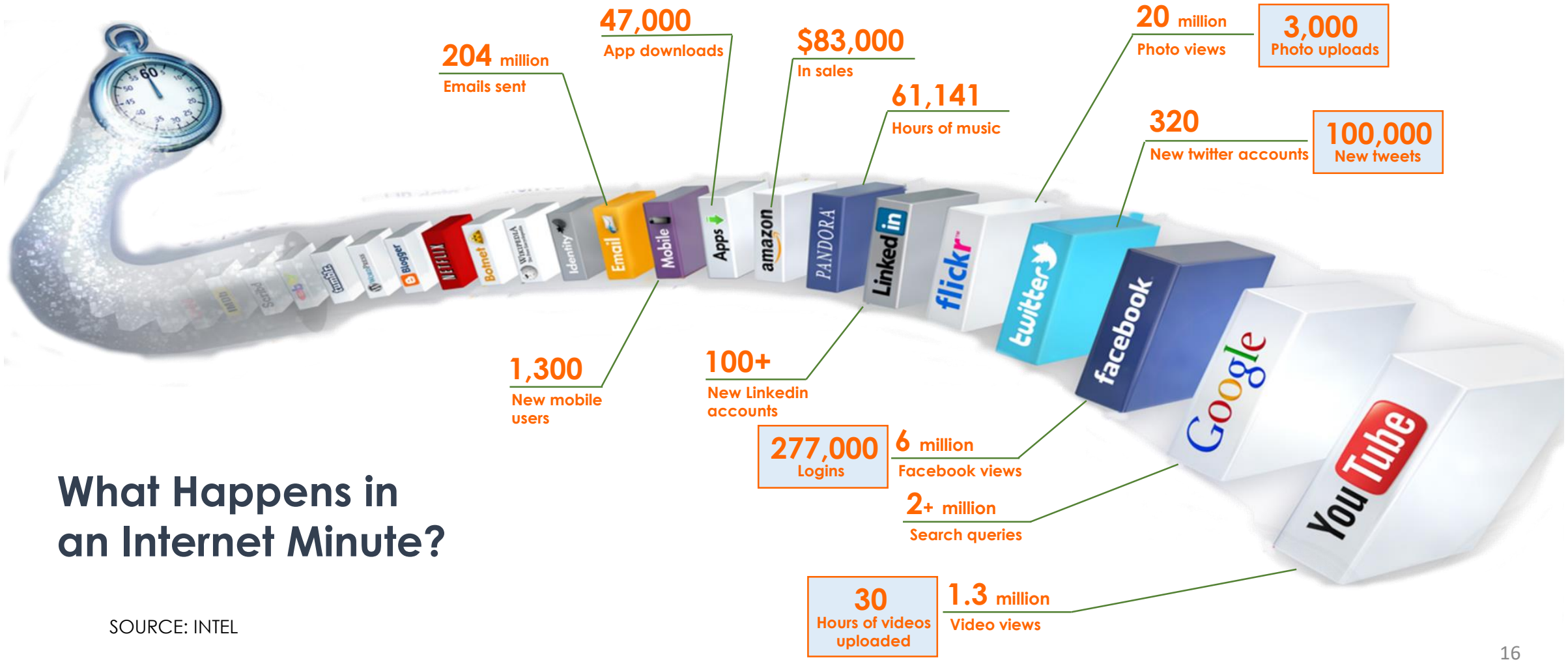
Malaysia Data Science community groups

- Artificial Intelligence Malaysia https://www.facebook.com/groups/artificialintelligencemalaysia/
- Big Data Malaysia https://www.facebook.com/groups/bigdatamy/
- Python User Group Malaysia https://www.facebook.com/groups/python.malaysia/
- TensorFlow & Deep Learning Malaysia https://www.facebook.com/groups/TensorFlowMY

Step 3: Study the Data Available

- Data are values of qualitative or quantitative variables, belonging to a set of items
- Qualitative: Gender, Race, Previous Treatments
- Quantitative: Height, Weight, Blood Pressure

Big Data Explosion

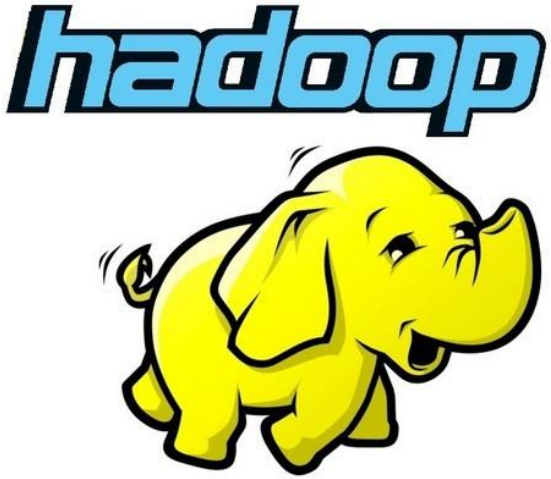


What Happens in an Internet Minute?

SOURCE: INTEL

Hadoop

- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware

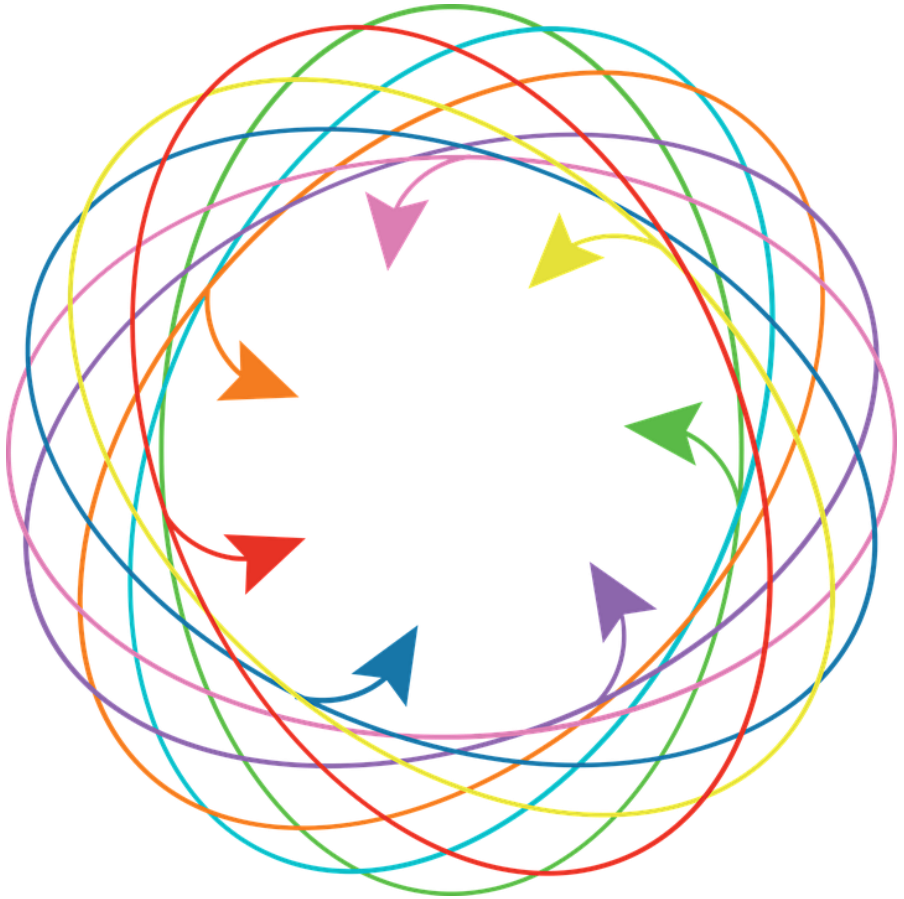


- Provides massive storage for any kind of data

- Enormous processing power to handle virtually limitless concurrent tasks or jobs

Data is the new oil

- Push out the silo mentality
- Integrate a data lake
- Obtain external data
- Combine internal and external datasets to form new insights



Open Data

- Open data is data that can be freely used, re-used and redistributed by anyone

- Open data is a key area of focus for Malaysia's Public Sector ICT Strategic Plan

- The government has set a target of being in the top 30 in the Open Data Barometer (ODB) by 2020

- As of 2016, Malaysia had an ODB rank of 53

data.gov.my

Structured Data vs Unstructured Data

Structured Data is readily usable

size of house	num of bedrooms	house price
1500 sq ft	2	$100,000
2000 sq ft	2	$180,000
3000 sq ft	4	$250,000

Majority of Data today is unstructured

Data is the second most important thing

The most important thing in Data Science is
THE QUESTION

Step 4: Ask the Right Questions

- What problem are you trying to solve?

- Hackathons are short intensive events in which everyone collaborates to discuss ideas, solve problems and develop prototypes

- This includes programmers, designers and subject-matter-experts

Malaysia Airlines Hackathon 2017

1. Internal ideas generation competition
2. Open hackathon to the public
3. External participants pitch their capabilities to internal staff
4. Form teams with both internal and external parties
5. Conduct hackathon
6. Hire new talents
7. Set up digital innovation hub

Startup Weekends

- Startup Weekend Hackathons span the course of a weekend with 60-120 participants
- Groups of programmers, business managers, marketing gurus, graphic artists and more pitch ideas, form teams around those ideas, and work to develop a working prototype, demo, or presentation
- Startup Weekend has reached more than 100 countries with over 200,000 participants

Event Name	Location	Date
Startup Weekend IIUM kuala lumpur	Kuala Lumpur, Malaysia	Mar 30 - Apr 1, 2018
Startup Weekend Petaling Jaya	Petaling Jaya, Malaysia	Mar 30 - Apr 1, 2018
Startup Weekend Kuala Lumpur Logistics Tech	Kuala Lumpur, Malaysia	Apr 6 - 8, 2018
Startup Weekend Sarikei	Sarikei, Malaysia	Apr 6 - 8, 2018
Startup Weekend Georgetown	Georgetown, Penang, Malaysia	Apr 20 - 22, 2018

Type of Data Science Questions

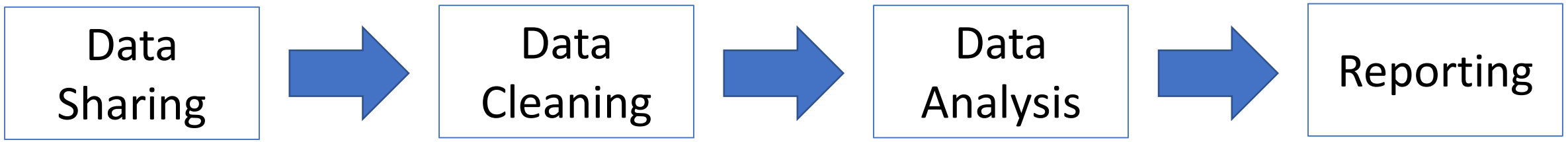
- Descriptive – focus on representation without explanation (e.g. census)
- Exploratory – discovering new projects (e.g. policyholder behaviour patterns during customer service)
- Inferential – understand the relationship between input and output (e.g. formula linking “Age”, “Occupation” and “Gender” to life expectancy)

Type of Data Science Questions

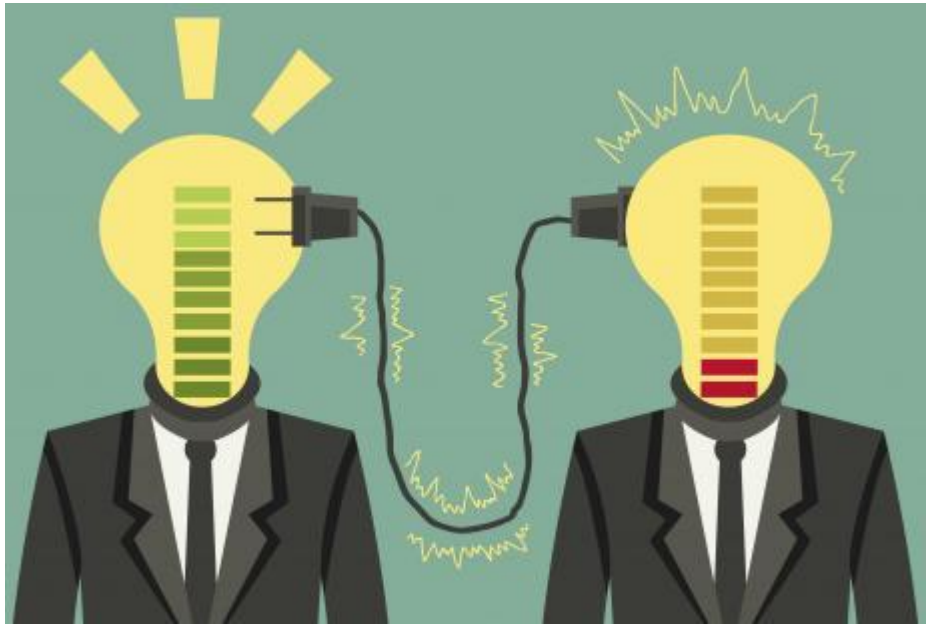
- Predictive – focus on getting the right output (e.g. lapse study)
- Causal – studying the butterfly effect (e.g. stress testing)
- Mechanistic – a more precise form of inferential and causal analysis (e.g. regulated reserve valuation)

Step 5: Form an Analysis Plan

- Covers all the details of the study
- Avoid embarrassing oversight



Data Sharing Plan



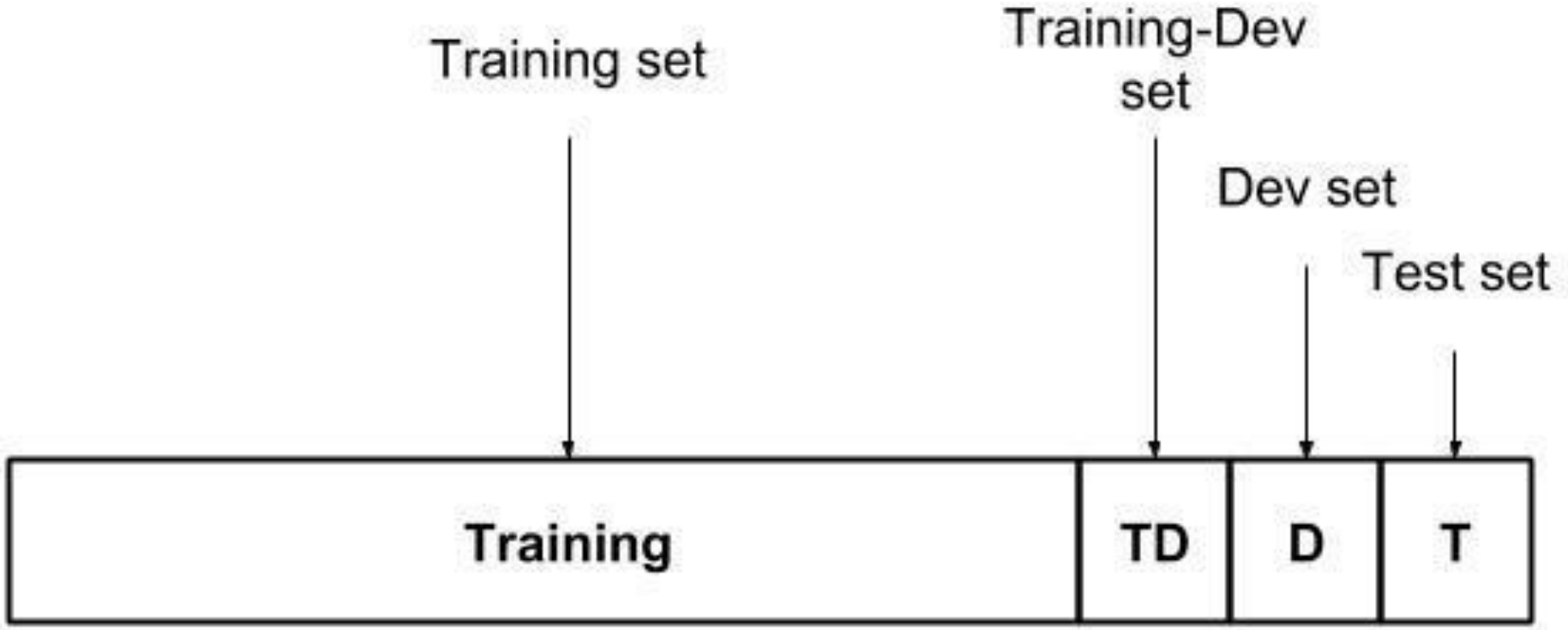
Data to be passed to the Data Scientist

1. Raw Data – no modifications
2. Tidy Data Set – labelled data
3. Data Dictionary – definitions of the provided datasets
4. Instructions – how to transform raw data into the tidy data set

Data Dredging

- Data dredging is the practice of mishandling data which leads to misleading results
- “The bed is the most dangerous place in the world”
- Correlation is not causation
- Context is important

Data Sets


Training set
Training-Dev set
Dev set
Test set
Training
TD
D
T

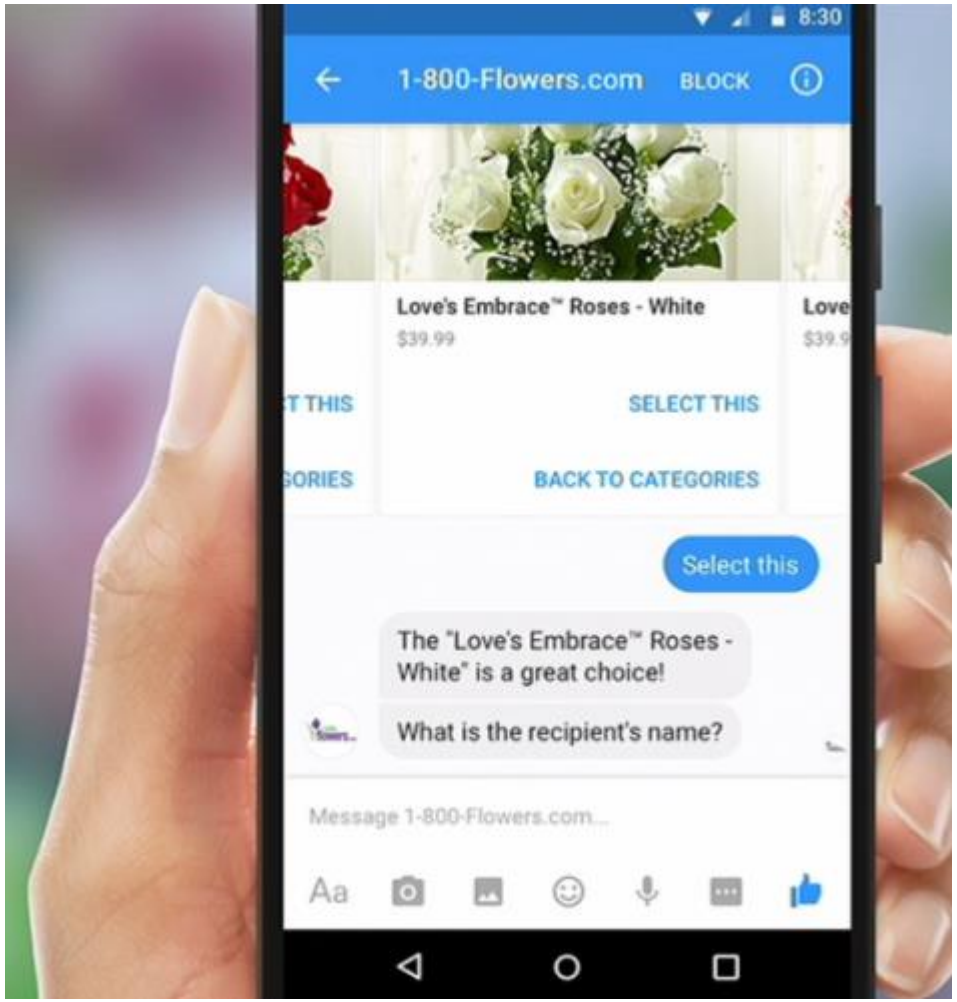
Step 6: Partner Up

Leverage on your partner’s strengths

- Do not reinvent the wheel
- Understand the tools already available
- Research which vendor has more experience in the problem that you are trying to solve
- Check that the vendor is providing actionable recommendations and not just information

Step 7: Application & Integration

- Go for quick wins at the start
- Application is unique to each industry
- Most important factor is the data storage and accessibility
- One of the simplest universal application is using a chatbot for preliminary customer service



Challenges

- Need to be clear what is possible and what is not possible
- Too focused on getting the perfect model
- Keeping a friendly environment

One more thing…

- There must be strong key stakeholders buy-in
- Alignment with long-term business drivers
- Develop a company-wide data-centric mindset
- Include self-learning within job scope
- Be open to taking risk and be not afraid to fail