Introduction to Data Science
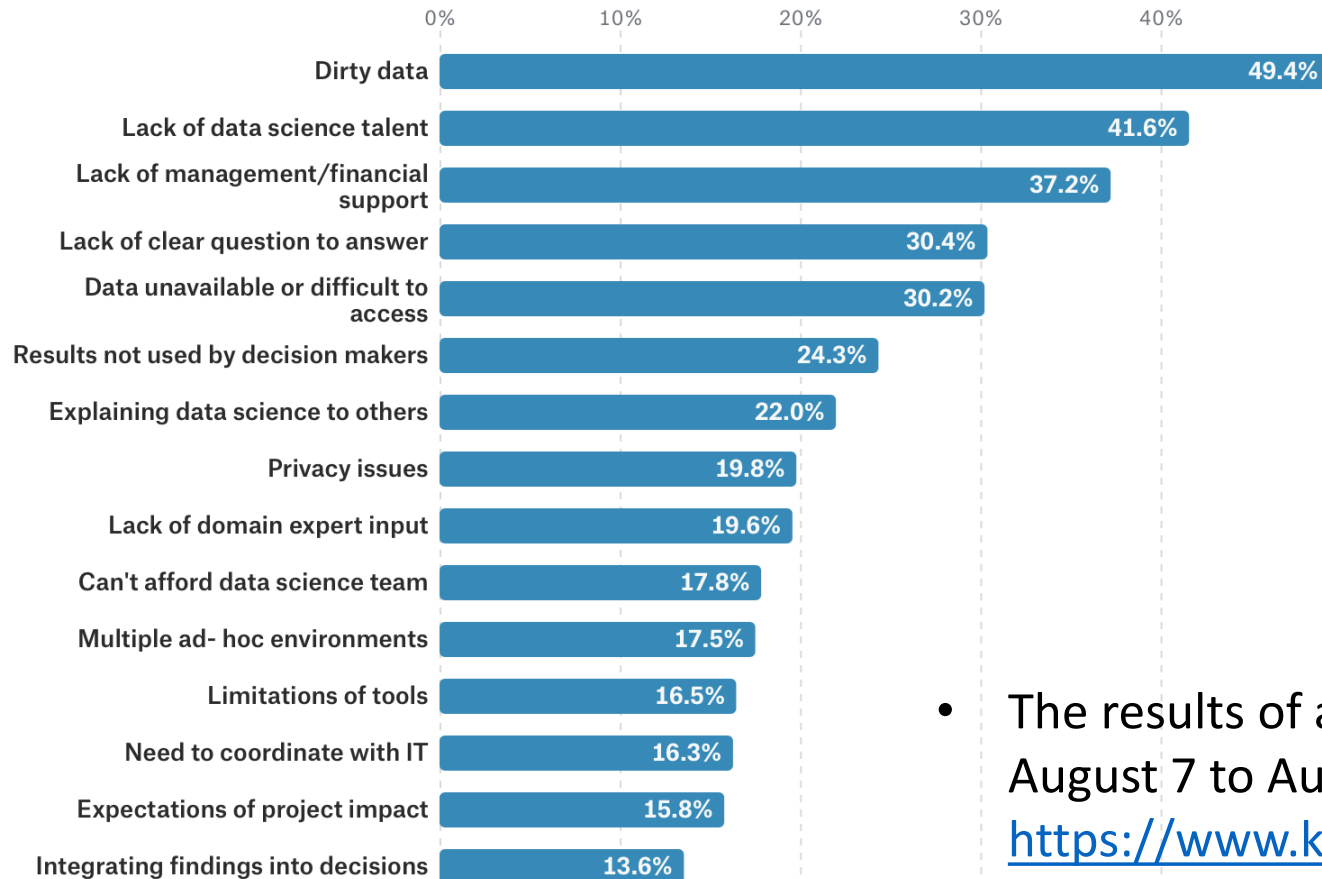
Chapter 7
Challenging Issues in Data Science

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science

Chiang Mai University

Challenging Issues in Data Science

What barriers are faced at work?

Barrier	Percentage
Dirty data	49.4%
Lack of data science talent	41.6%
Lack of management/financial support	37.2%
Lack of clear question to answer	30.4%
Data unavailable or difficult to access	30.2%
Results not used by decision makers	24.3%
Explaining data science to others	22.0%
Privacy issues	19.8%
Lack of domain expert input	19.6%
Can't afford data science team	17.8%
Multiple ad- hoc environments	17.5%
Limitations of tools	16.5%
Need to coordinate with IT	16.3%
Expectations of project impact	15.8%
Integrating findings into decisions	13.6%

- The results of a survey conducted by Kaggle, from August 7 to August 25, 2017. https://www.kaggle.com/surveys/2017

7,376 responses
Only displaying the top 15 answers. There are 7 answers not shown.

Challenging Issues in Data Science

What barriers are faced at work?

- Dirty Data
- Data unavailable or difficult to access
- Privacy issues
- Lack of clear question to answer
- Lack of domain expert input
- Results not used by decision makers
- Explaining data science to others
- Need to coordinate with IT

- Lack of data science talent
- Integration finding into decisions
- Multiple ad-hoc environments
- Limitations of tools
- Lack of management/financial support
- Can't afford data science team
- Expectation of project impact
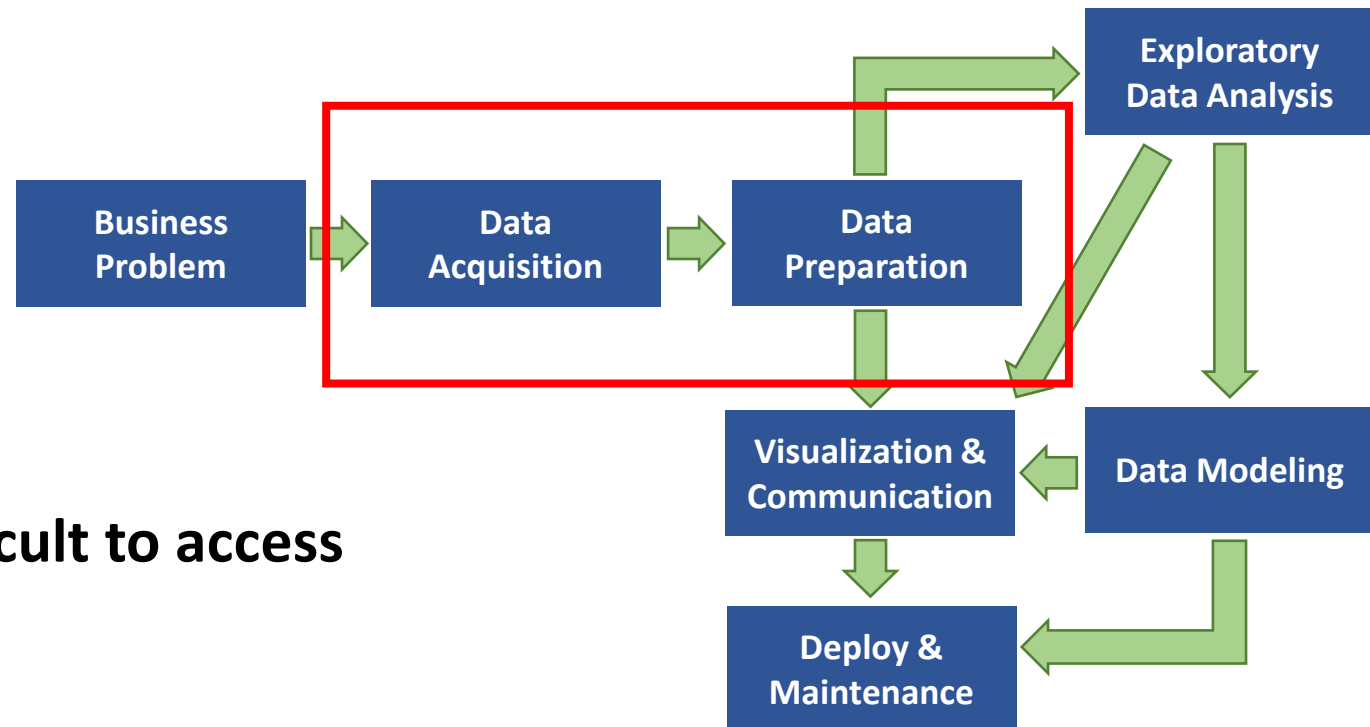
Challenging Issues in Data Science

Data is the biggest challenge in data science

1. **Dirty Data**
   - Inaccurate
   - Incomplete
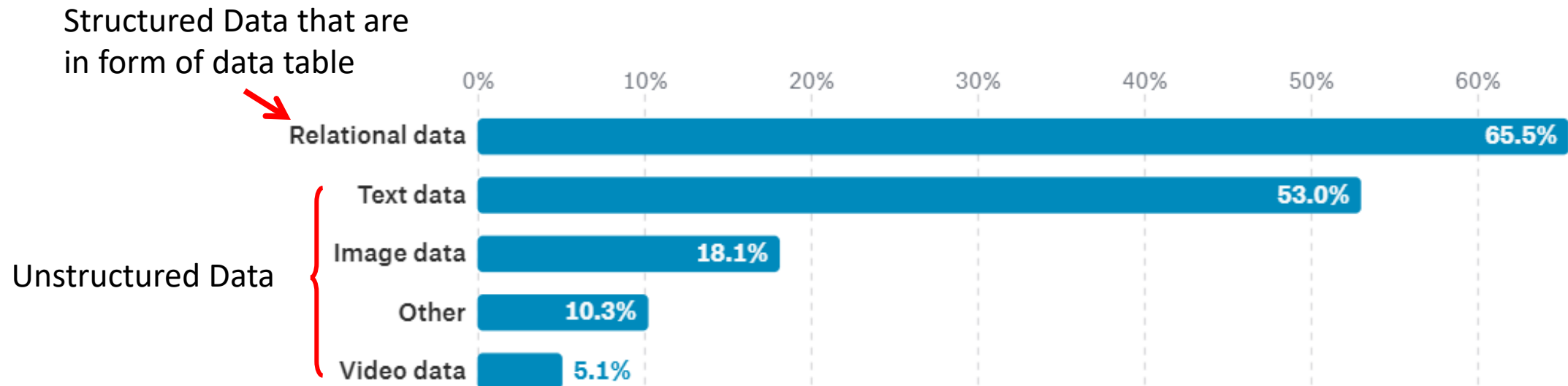   - Inconsistency
   - Invalid
   - Nonuniform
   - Duplicate
2. **Data unavailable or difficult to access**
3. **Privacy issues**

Challenging Issues in Data Science

Types of data used at work

Structured Data that are
in form of data table

Unstructured Data

Relational data	65.5%
Text data	53.0%
Image data	18.1%
Other	10.3%
Video data	5.1%

A survey conducted by Kaggle, from August 7 to August 25, 2017.
https://www.kaggle.com/surveys/2017

Challenging Issues in Data Science

Lack of clear question to answer
- Understanding the business problem is important
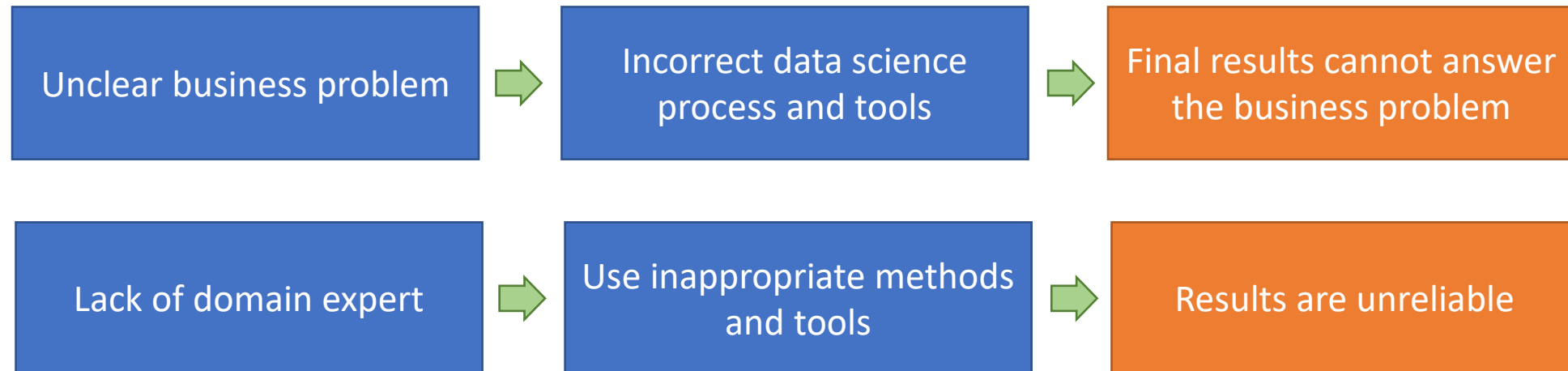- If the business problem is not clear, the next processes will fail.

Challenging Issues in Data Science

Lack of domain expert input
- Domain knowledge is required in many steps of data science process.
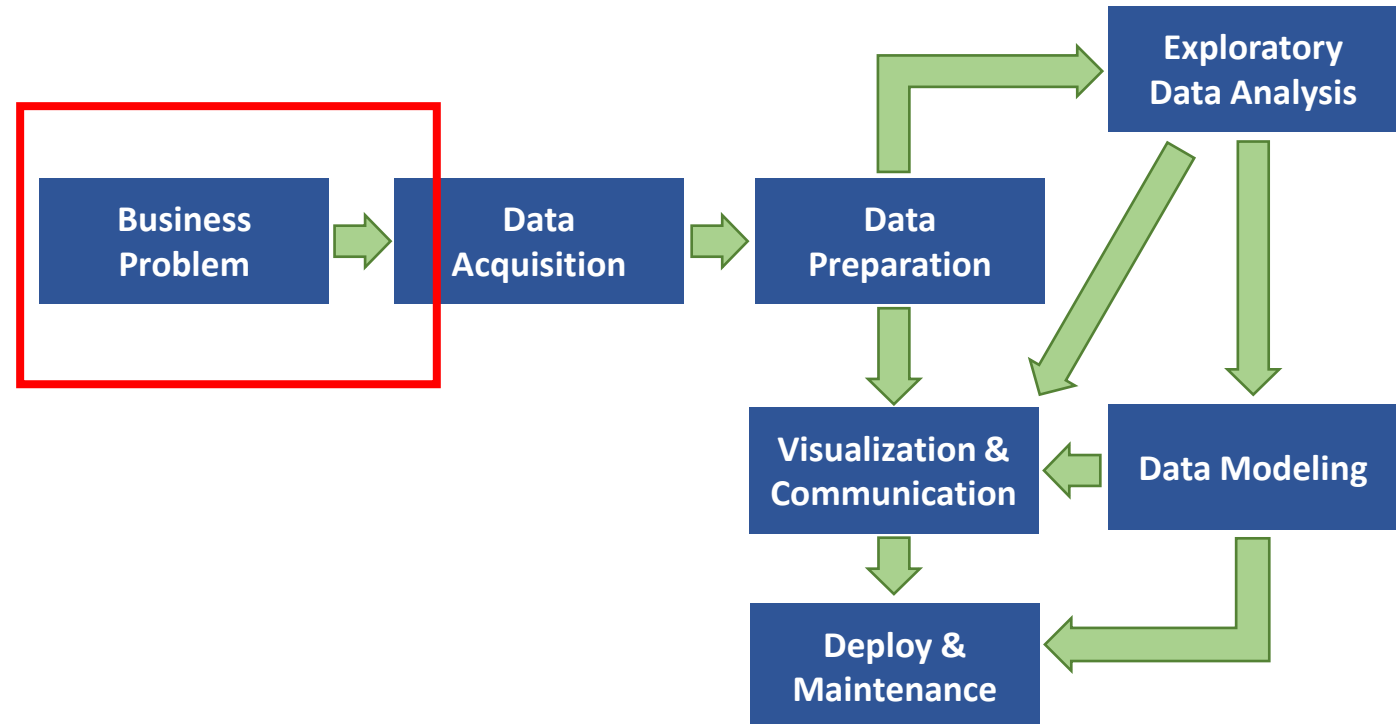
Challenging Issues in Data Science

Results not used by decision makers

This issue may be caused by

- Lack of clear question to answer
- Lack of domain expert input

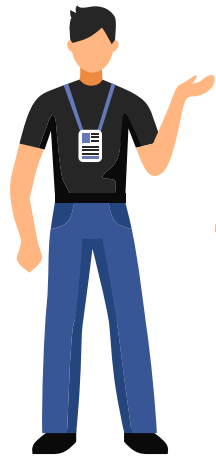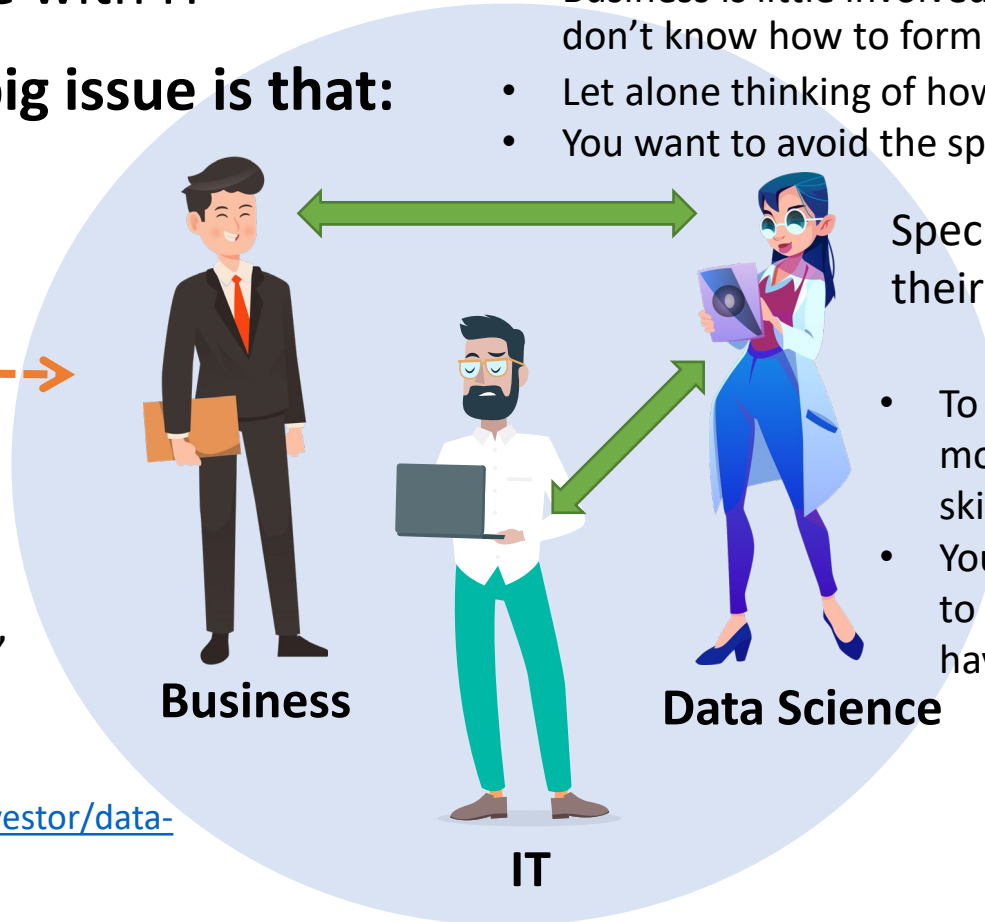|		
Unclear business problem	Incorrect data science process and tools	Final results cannot answer the business problem
Lack of domain expert	Use inappropriate methods and tools	Results are unreliable

Challenging Issues in Data Science

Collaboration and organizational issues

1. Explaining data science to others
2. Need to coordinate with IT

The reason it is a big issue is that:

- Business is little involved because they don't know how to formulate the problem.
- Let alone thinking of how to solve it.
- You want to avoid the specialist bias.

Specialists tend to have their own language

- To minimize the time required to put a model in production, good software design skills are required.
- You do not want your development team to have to rewrite the code data scientists have produced.

Business

IT

Data Science

Need to add to the mix generalists, who understand business, data science and IT.

Source: https://medium.com/datadriveninvestor/data-science-challenges-b7622b85b807

Challenging Issues in Data Science

Talent

1. Lack of data science talent
   - The generalist talent is required to do the initial problem solving
2. Integration finding into decisions
   - Need more experiences.

Challenging Issues in Data Science

Tools

1. Multiple ad-hoc environments

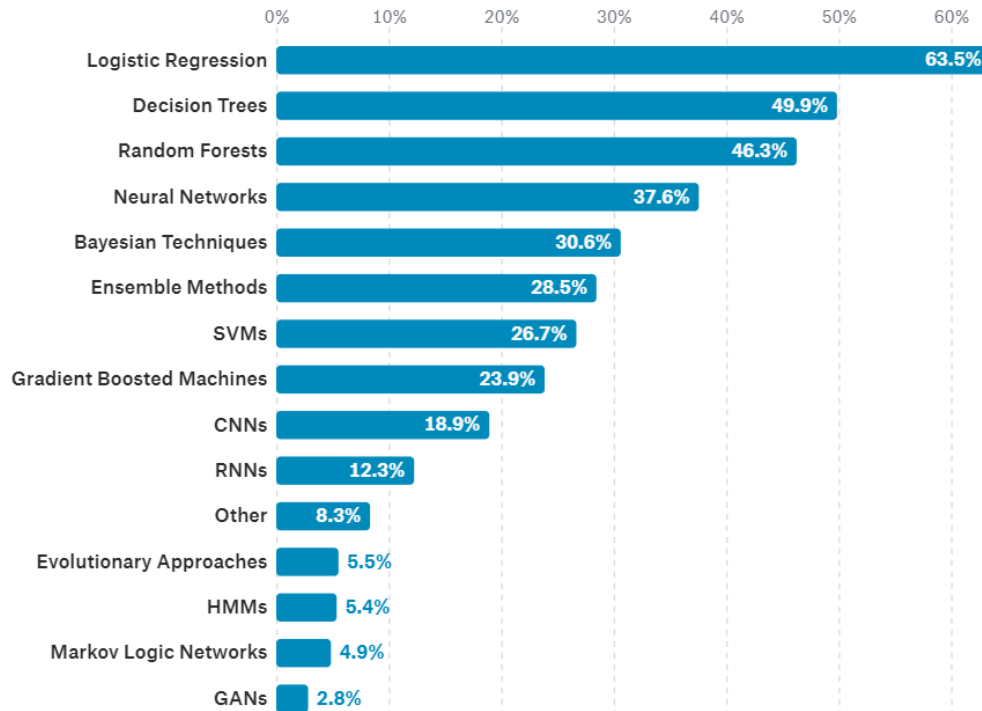   Ad-hoc code needs to be written to optimize the computing time required to run the models.

2. Limitations of tools

   - Limitation of software
     - A software has both advantages and drawbacks.
     - We need to use more than one software to perform a data science project.
   - Limitation of hardware
     - Hight computing power
     - Expensive technology stack
   - Limitation of algorithms/methods
     - Simpler algorithms might limit your findings.
     - Most advanced algorithms do not scale well with the size of the datasets.
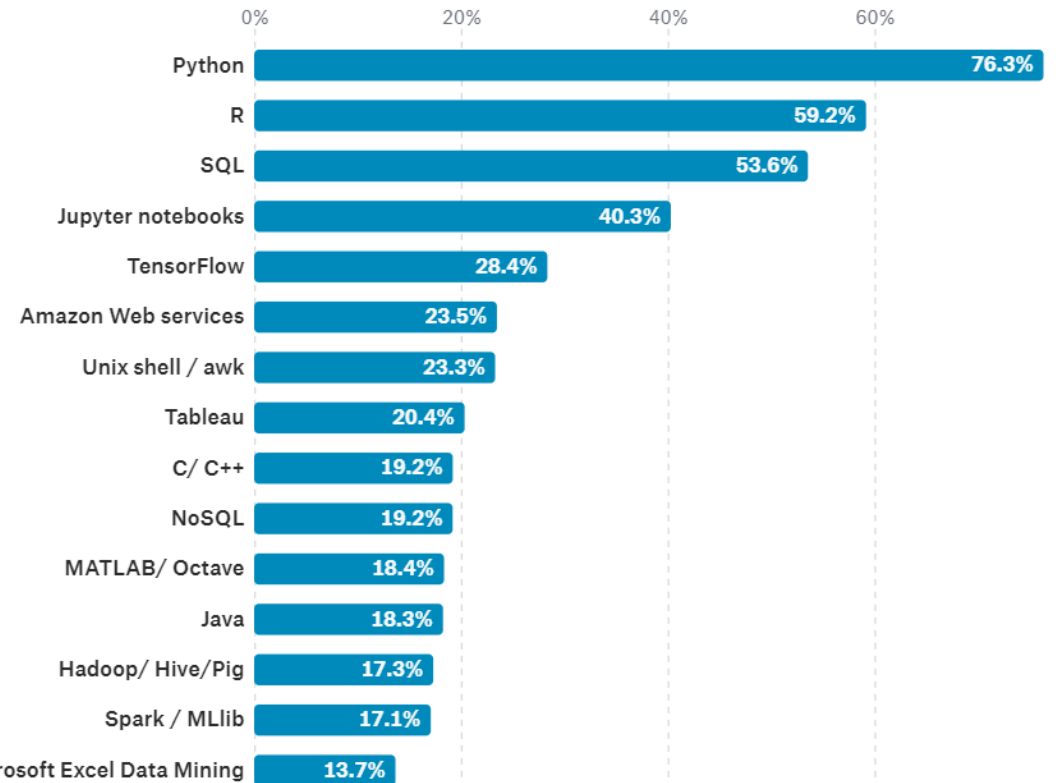
Challenging Issues in Data Science

Tools

Most commonly data science method used at work

Method	%
Logistic Regression	63.5%
Decision Trees	49.9%
Random Forests	46.3%
Neural Networks	37.6%
Bayesian Techniques	30.6%
Ensemble Methods	28.5%
SVMs	26.7%
Gradient Boosted Machines	23.9%
CNNs	18.9%
RNNs	12.3%
Other	8.3%
Evolutionary Approaches	5.5%
HMMs	5.4%
Markov Logic Networks	4.9%
GANs	2.8%

Most commonly used data analysis tool

Tool	%
Python	76.3%
R	59.2%
SQL	53.6%
Jupyter notebooks	40.3%
TensorFlow	28.4%
Amazon Web services	23.5%
Unix shell / awk	23.3%
Tableau	20.4%
C/ C++	19.2%
NoSQL	19.2%
MATLAB/ Octave	18.4%
Java	18.3%
Hadoop/ Hive/Pig	17.3%
Spark / MLlib	17.1%
Microsoft Excel Data Mining	13.7%

A survey conducted by Kaggle, from August 7 to August 25, 2017.
https://www.kaggle.com/surveys/2017

Challenging Issues in Data Science

Budget

1. Lack of management/financial support
2. Can't afford data science team
3. Expectation of project impact
   - If the <u>costs are higher than the expected outcome</u> of the project, then there is no point doing this project.

Further Study

Website:

- https://www.kaggle.com/surveys/2017
- https://medium.com/datadriveninvestor/data-science-challenges-b7622b85b807