Things Solver

ENLIGHTEN YOUR DATA

Data Scientist's Toolbox

Data Science Tools and Skills

What is Data Science?

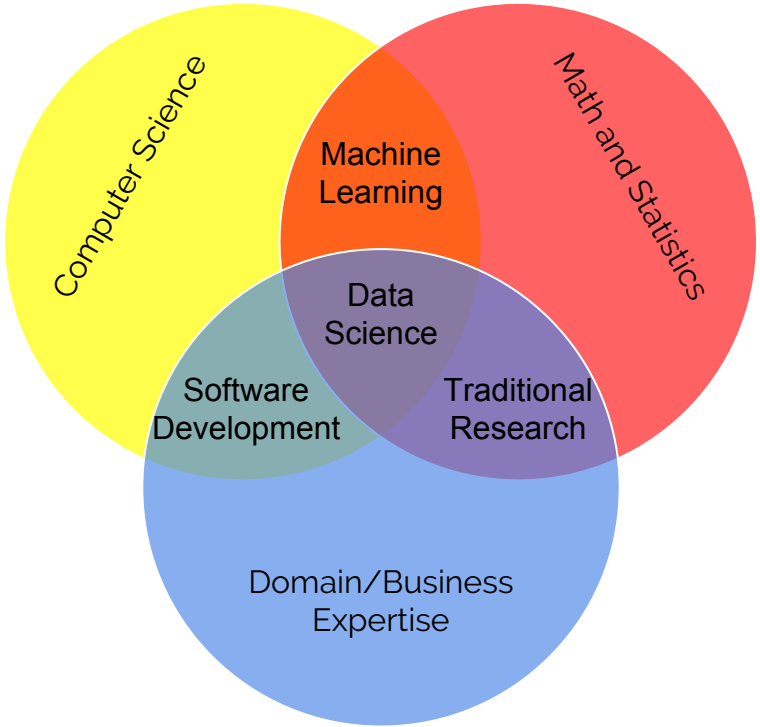
In essential, Data Science means **making sense about the world by using data**.

A buzzword typically used to describe the efforts of the companies and organizations in general to use the data to improve their performance and achieve goals.

Contrary to Data Analysts, the Data Scientists ask questions themselves driven by knowing which business goals are most important and how the data can be used to achieve certain goals for the organization. The communication is bottom up.

Data Science allows:

- Empowering management and officers to make better decisions with quantifiable, data-driven evidence, and testing these decisions.
- Directing the actions based on trends which in turn help in defining goals
- Challenging the staff to adopt best practices and focus on issues that matter.
- Identifying opportunities and target audiences.
- ...



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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

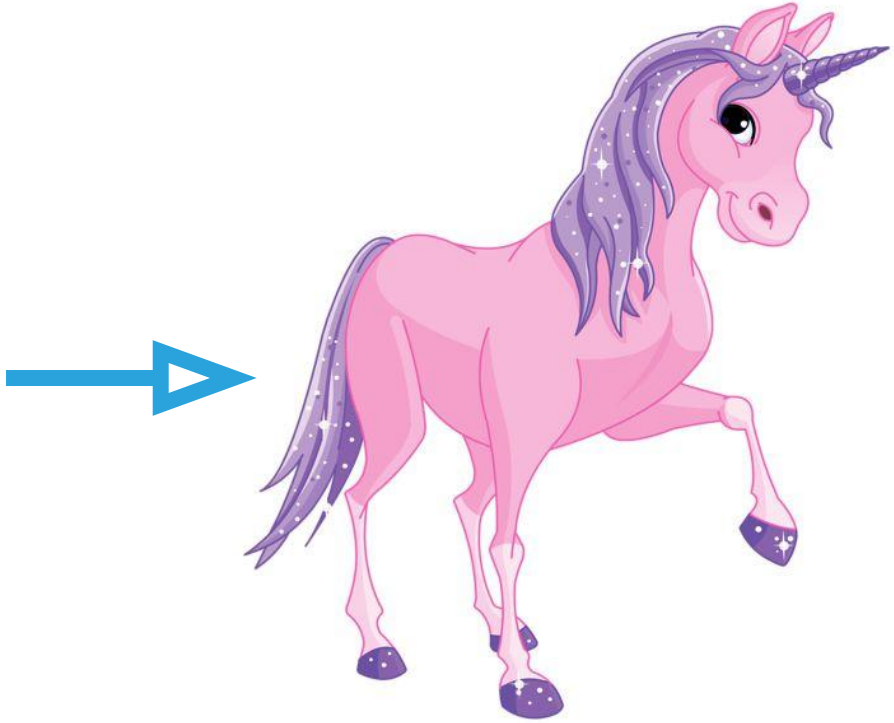
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

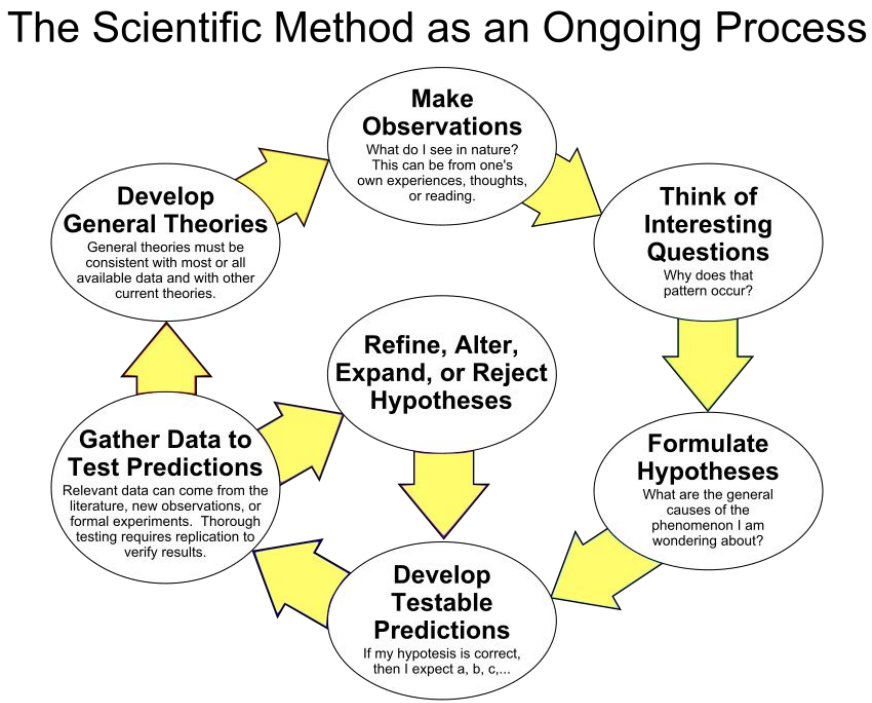
Marketing DISTILLERY



The “Science” of the Data

“Generally speaking, both traditional scientists and data scientists ask questions and/or define a problem, collect and leverage data to come up with answers or solutions, test the solution to see if the problem is solved, and iterate as needed to improve on, or finalize the solution.”

- Alex Castrounis, *What Is Data Science, and What Does a Data Scientist Do?*, KDNuggets 2017



By ArchonMagnus (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Picking the right tool for the job...

In recent years, as the popularity of Big Data tools arise, many organizations had tendencies to attack small problems with complex and powerful machinery like Hadoop and Spark.

The fact is, in most cases, you **DO NOT have a Big Data** problem (*okay, you particularly DO because you are a telecommunication provider :D, but most organizations can handle their tasks with one or two powerful machines and simple architectures*).

Besides, although there is a powerful ecosystem built around the open source community, with powerful tools and libraries, many of them are not mature enough to run at production scale.

Another issue is the lack of expertise in this domain, which makes the whole story on picking the right tools and architectures to tackle the data use cases more complex and non-trivial.

Starting and Architecture Guidelines

Input integration points. Standardized interfaces or custom integration?

Data Volume. Small or Big Data?

Data Formats. What is the structure of the data?

Data Velocity. What is the speed of the source data (batch or realtime)?

Ad-hoc Analysis. What are the demands for performing ad-hoc analysis and queries by the users.

Data Governance. Data Stewardship? Data Quality? Master Data Management? Use Cases?

Licenses. Open source or vendor tools?

Infrastructure. Where should we run the workflow - cloud or on-premise?

Know-how. Do we have internal knowhow within organization or we need to hire a team of professionals?

Production capabilities. Are there any planned changes on the source system that might break the workflow? How the implementation will affect existing systems?

Data Analytics

The role of Data Analytics is to derive actionable insights from the data and communicate the relevant information to decision makers in order to make informed decision.

Analytics refers to all the processes of working with the data to provide information, and not dealing with the infrastructure.

EXPLORATION	FEATURE ENG.	STATISTICS
ML	VISUALIZATION	HYPOTHESIS
SUMMARIZATION	CLEANSING	REPORTING

Data Engineering

Data Engineering has become important in the age of Big Data.

Data Engineer can be thought of as a type of Data Architect - they are concerned with data architecture, computing and data storage infrastructure, data flow, monitoring and production implementation, ...

SCALABILITY	ARCHITECTURE	PRODUCTION
ETL?	OPS	DATA LAKE
AVAILABILITY	BACKUPS	PIPELINES

DATA Engineer

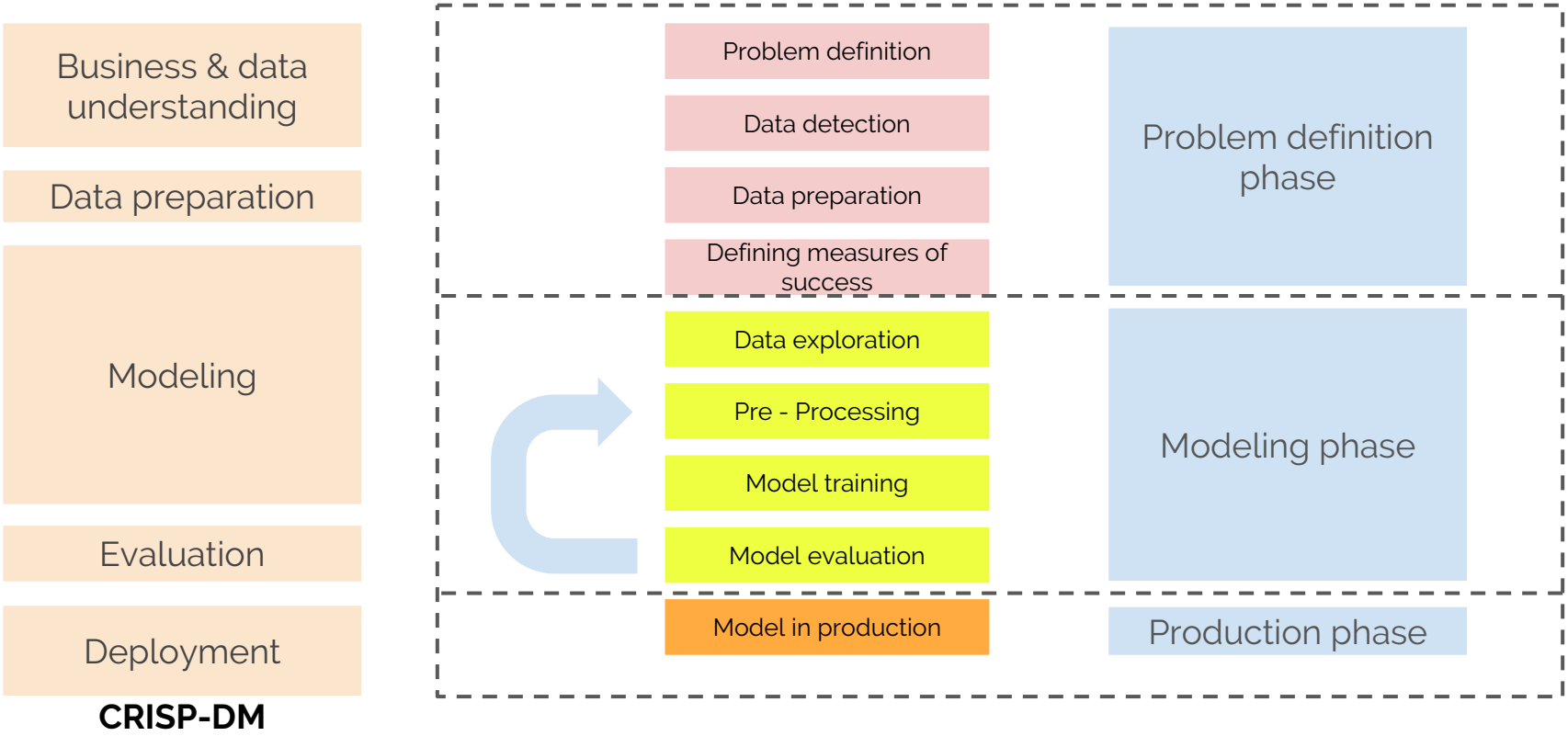
DataCamp
Learn Data Science By Doing

DATA Scientist

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Things Solver Data Science Workflow



Data Ingestion and Preparation

The first step in working with the data is **getting the input** data in a **required format**.

Data Ingestion.

Data is residing in various systems, and needs to be efficiently transformed and fed to the analytics in a required period in order to be useful.

Before integration starts, it is required to check the **capabilities** of the source systems to feed us the relevant data (and what this process might look like in the production phase).

Usually standard interfaces

Raw data available in many cases

Data quality by the source system

Initial load

Do we have the purchased licenses to use the interface?

Who is the data owner?

Will the system be able to feed us the data at production?

How frequently the system changes?

Data Ingestion and Preparation

The first step in working with the data is **getting the input** data in a **required format**.

Preparation.

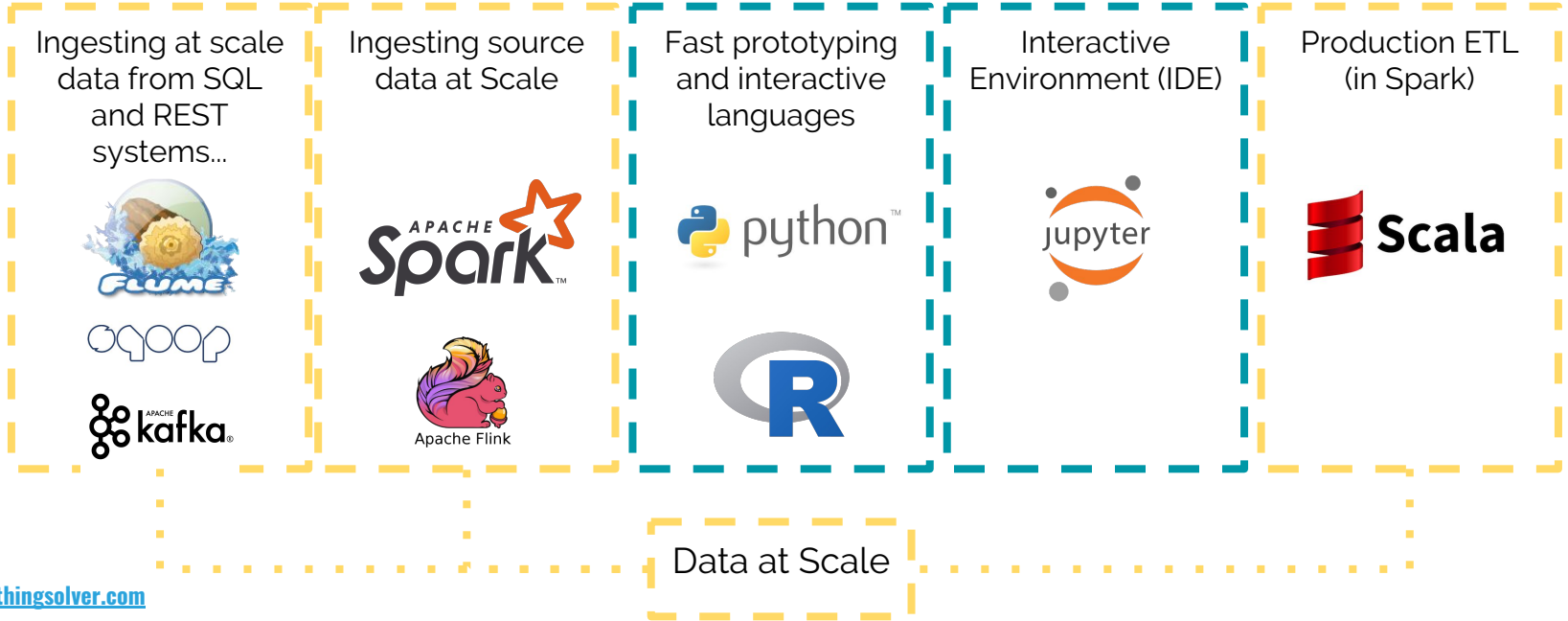
Data Preparation phase includes steps that help us learn more about the data:

- **Data structure/schema imputation.** This step helps us to identify the columns within a data, impute schema and data formats and prepare the desired output.
- **Ad-hoc queries to explore the dataset and column values.** This step helps us to learn more about the columns and their values, and the dataset itself.
- **Column standardization and preprocessing.** Calculating new columns or retrieving additional information from the existing data (unnesting collection types if necessary).

Data Integration and Preparation phase give enough knowledge to deploy the final ingestion/ETL workload.

Data Ingestion and Preparation

Due to the nature of this phase - interactivity (aka “try and fail”), there is a demand for **interactive environment and tools** that would allow us to write ad-hoc queries on top of the data. Typical toolbox includes fast prototyping languages like Python and R with belonging libraries to efficiently perform integration and preparation phase.



Data Persistence

How do we persist our data for the Analytics processes?

Operational Storage

Low latency

In-memory

Highly structured

Expensive

Storage efficiency and data integrity

Supports live operations

Usually supports updates

Relational, Key-value, document and column, ...

Analytical Storage

Higher latency

On disk

Cheap

Semi-structured

Data retrieval and summarization

Supports on-demand analysis

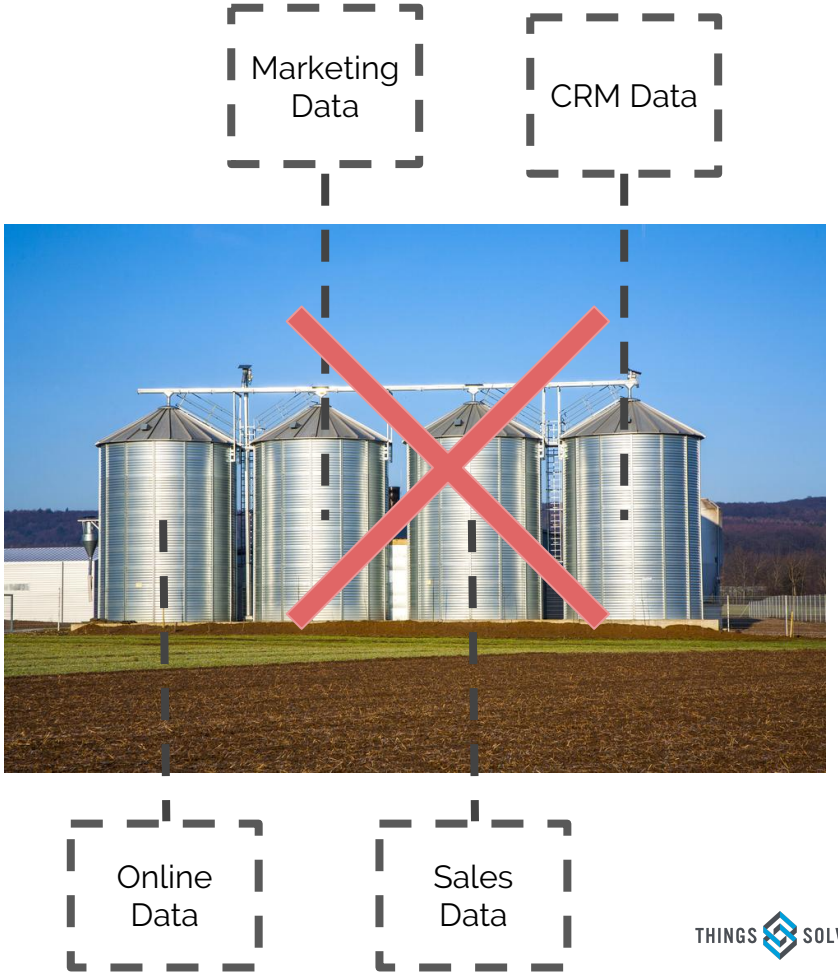
Immutable

Text formats (CSV, JSON, ORC, Parquet, ...)

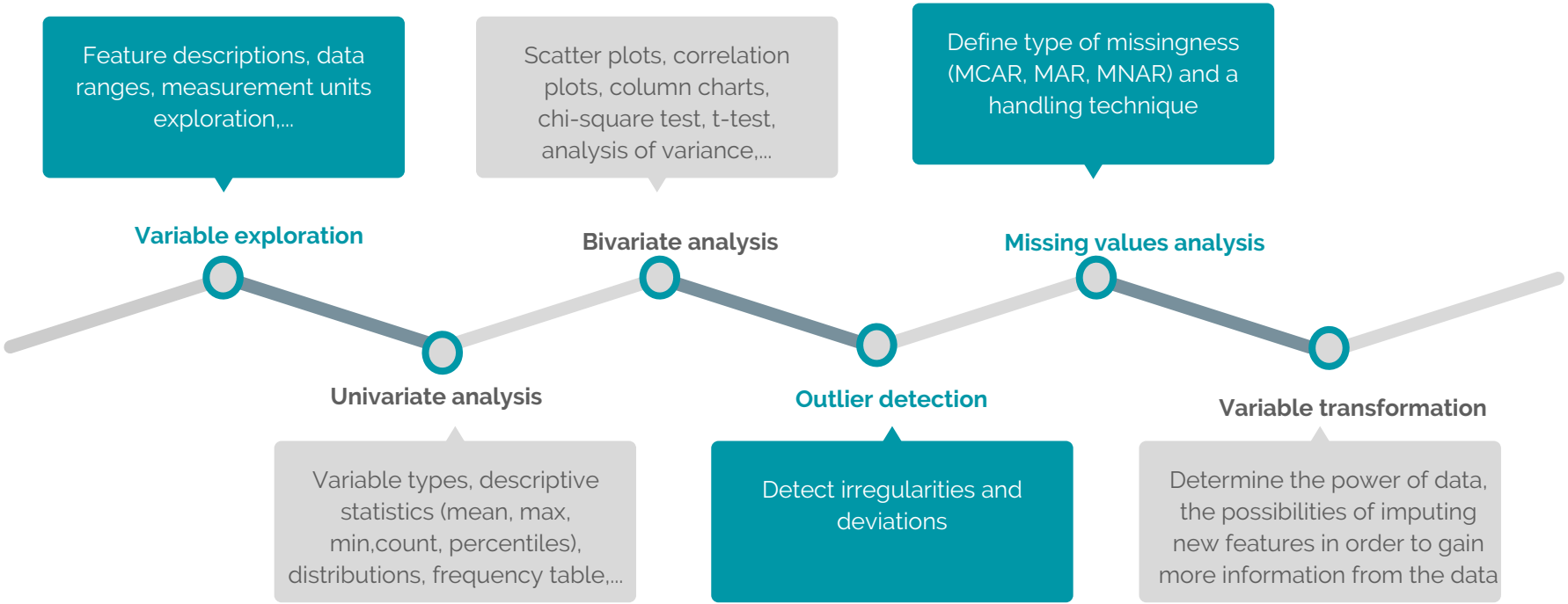
Data Persistence

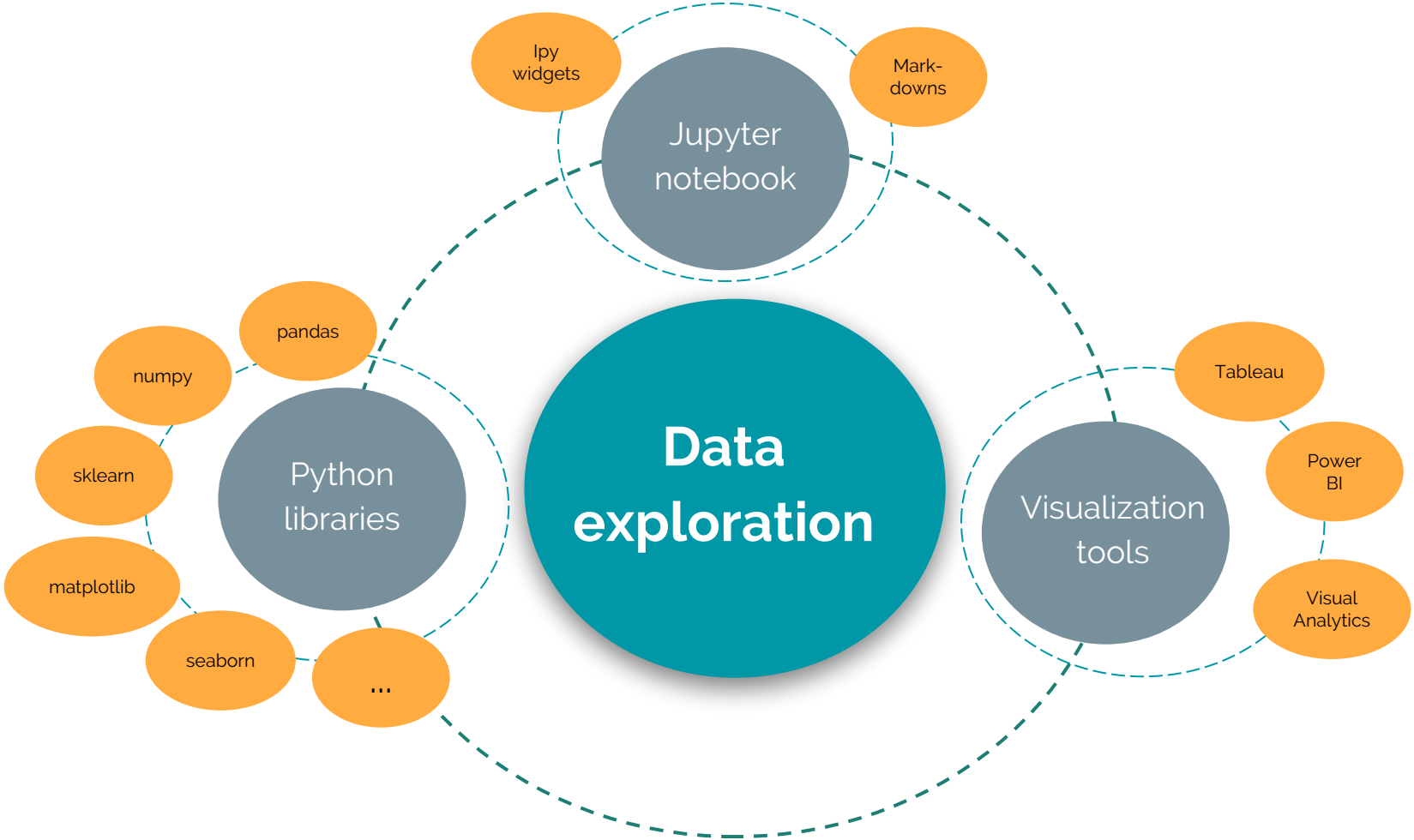
Things to have in mind while planning Data Persistence Layer:

- **Efficiency and ease of use.** Data needs to be persisted in such a way that Data Scientists can easily and efficiently deploy their pipelines.
- **Scalability.** Persistence Layer needs to support the future amount of data that will be coming from the source systems.
- **Deduplication and Integration.** Data Persistence Layer needs to be interoperable for many use cases - avoid building Data Silos and Data Ponds, instead plan a centralized approach (**Data Lake**).
- **Security and Policies.** Establish user policies and secure access to the data.

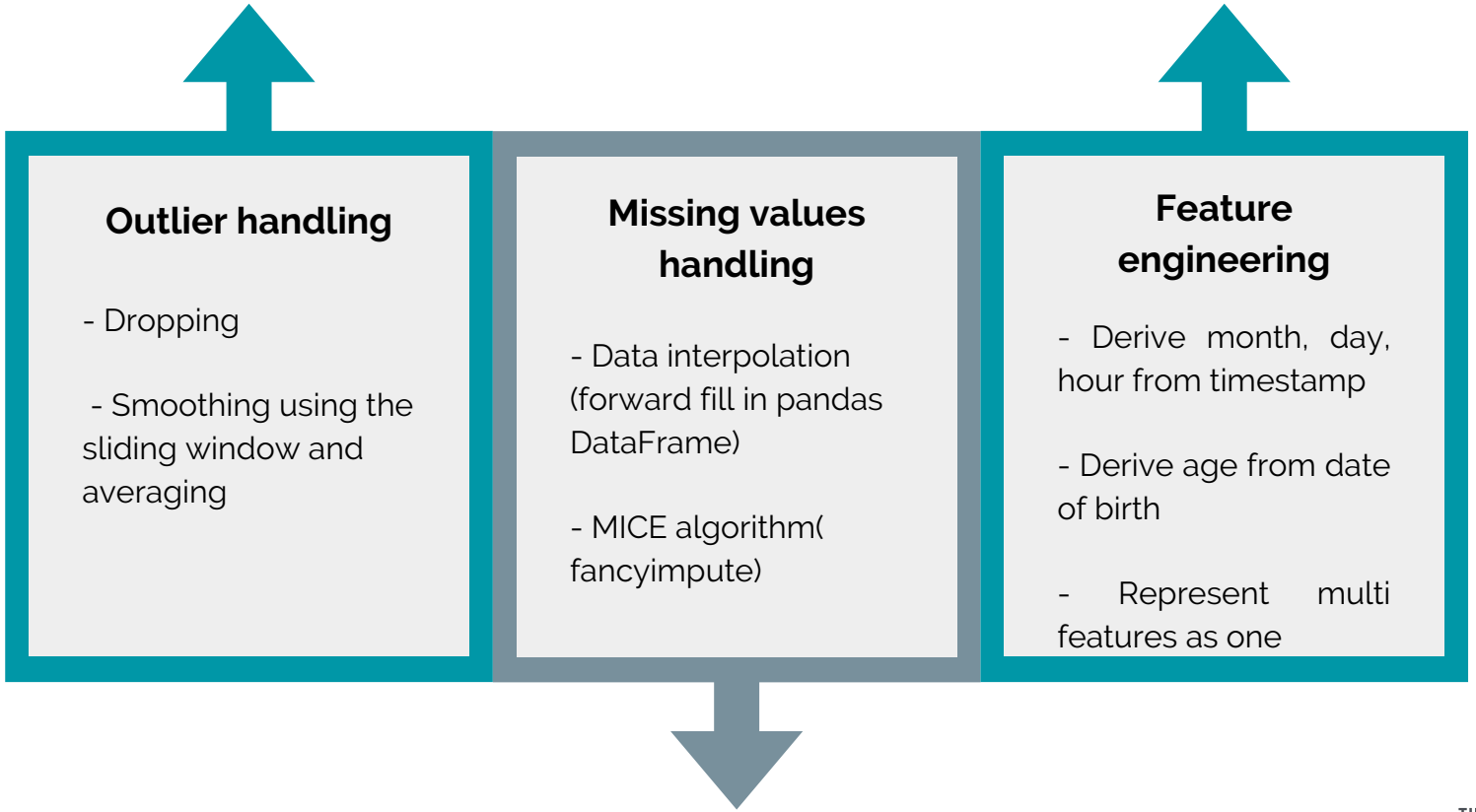


Data Exploration

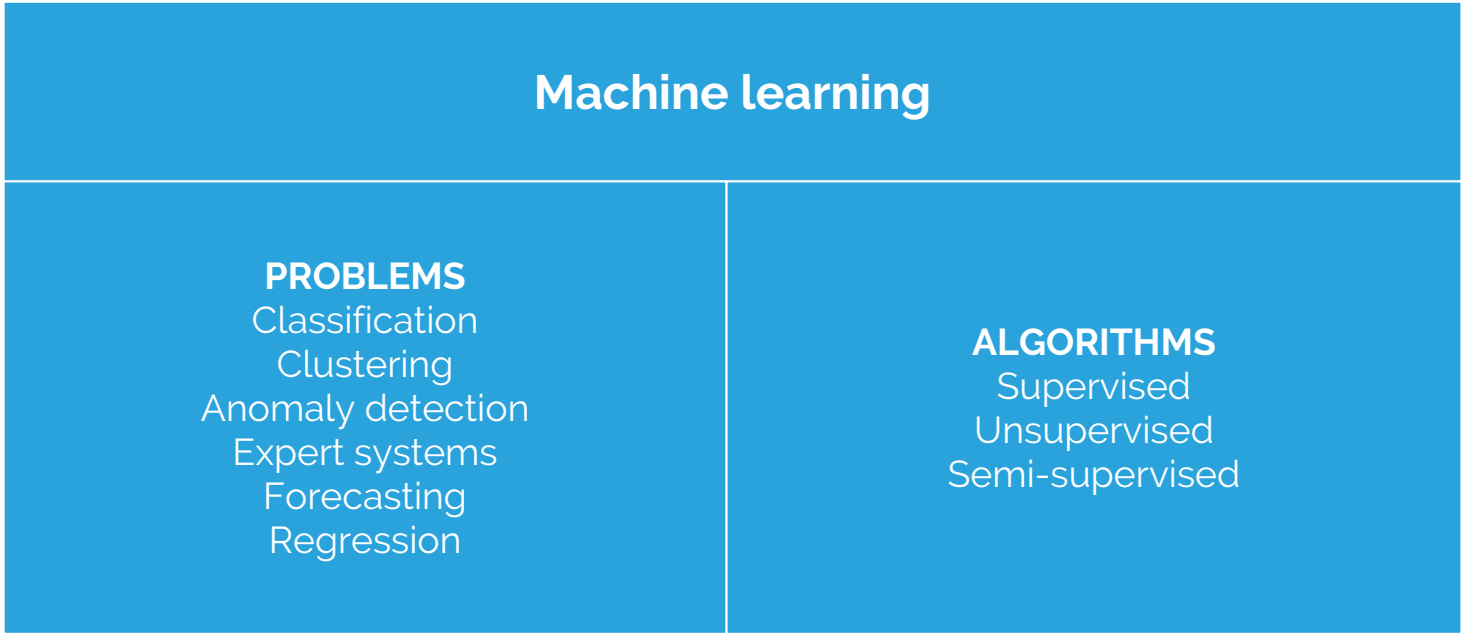



Ipy widgets
Mark-downs
Jupyter notebook
pandas
numpy
sklearn
matplotlib
seaborn
...
Python libraries
Data exploration
Tableau
Power BI
Visual Analytics
Visualization tools

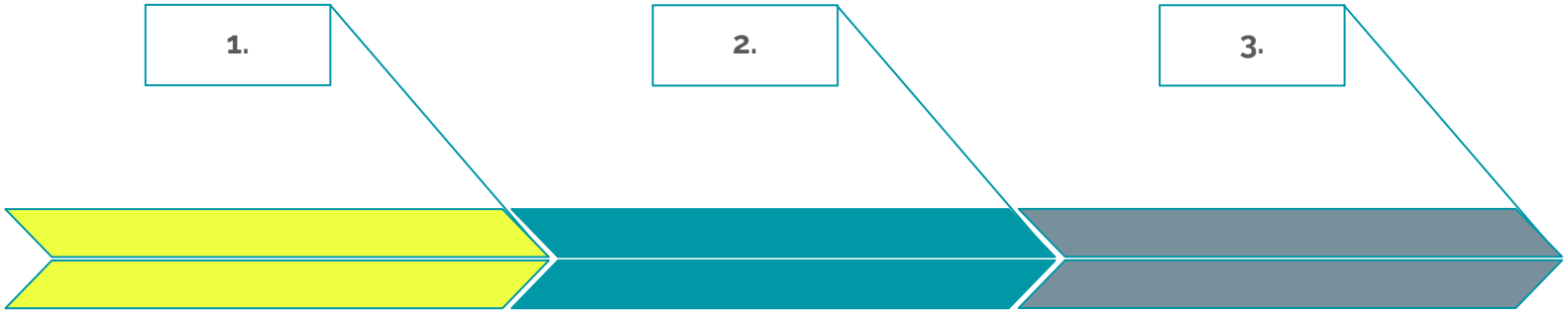
Data Preprocessing



Modeling


Machine learning
PROBLEMS
Classification
Clustering
Anomaly detection
Expert systems
Forecasting
Regression
ALGORITHMS
Supervised
Unsupervised
Semi-supervised

Modeling



Model selection

The selection of the model depends on the problem defined, as well as on the data being available.

Model training

Training includes running the model with some initial parameters and setting.

Model evaluation

Model has to be evaluated by some indicator that can be measured.

RMSE, Accuracy, Silhouette score,...

Modeling

1.

Model selection

The selection of the model depends on the problem defined, as well as on the data being available.

Step 1. Analyse the defined problem and user requirements. Interact with the user, to make sure you understood the requirements.

Step 2. Analyse data. If the problem is related to classification/regression, do you have labels? If not, can you provide it in some way? If not, can you solve the problem at all?

Step 3. Choose a corresponding model. Make sure that model assumptions are satisfied. Determine the input format for the model.

Modeling

2.

Model training

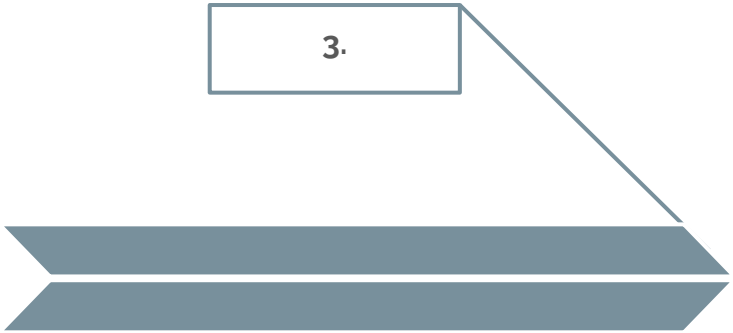
Training includes running the model with some initial parameters and setting.

Step 1. Choose a model. Try to find an implementation. In not already implemented, write it yourself, by using pure functions.

Step 2. Define a model. Analyse model parameters, and set up their initial values.

Step 3. Make sure you have the resources for running a model. Run a model. Obtain model results for testing and evaluating. Optionally, store the model, so you could use it later.

Modeling



Model evaluation

Model has to be evaluated by some indicator that can be measured.

RMSE, Accuracy, Silhouette score,...

Step 1. Define measurements of success. Measurements of success depend on the model used and data available for testing (conf.matrix, accuracy, silhouette score, RMSE,...).

Step 2. If test data is available, apply the trained model.

Step 3. Evaluate the model by using the previously defined measures. Model evaluation could be done also by domain expert, or by some external measures (sales increase,costs decrease,...)

Visualization

Data visualization as a presentation of data in a pictorial or graphical format, enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.

Interactive visualization enables a deep dive for more detail, interactively changing what data we see and looking at a certain area from different perspectives and on different levels..

Data visualization can also:

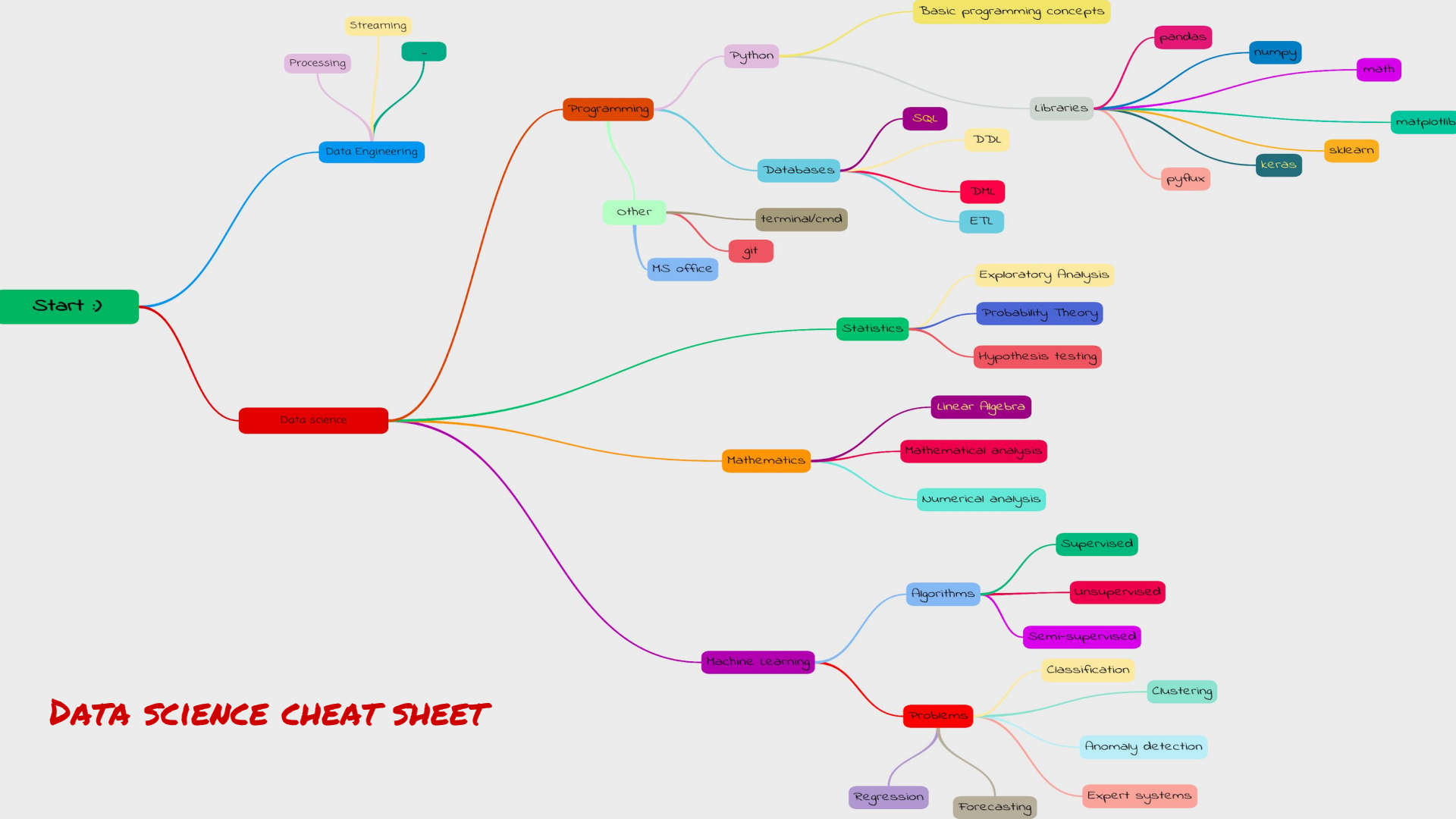
- Identify areas that need attention or improvement.
- Clarify which factors influence customer behavior.
- Help you understand which products to place where.
- Predict sales volumes.
- ...

Lack of **open source** tools in the ecosystem:

- Kibana
- Grafana (technical and infrastructure KPIs)
- Custom - JavaScript, Angular, ChartsJs, ...

Expensive? **proprietary** BI tools:

- SAS Visual Analytics
- Tableau
- Qlikview
- Power BI
- ...


Start :)
Data Engineering
Streaming
Processing
-
Data science
Programming
Python
Basic programming concepts
Libraries
pandas
numpy
math
matplotlib
sklearn
keras
pyflux
Databases
SQL
DDL
DML
ETL
other
terminal/cmd
git
MS office
Statistics
Exploratory Analysis
Probability Theory
Hypothesis testing
Mathematics
Linear Algebra
Mathematical analysis
Numerical analysis
Machine Learning
Algorithms
Supervised
Unsupervised
Semi-supervised
Problems
Classification
Clustering
Anomaly detection
Expert systems
Regression
Forecasting
DATA SCIENCE CHEAT SHEET

Production

Running the ML model in production is a very challenging task from engineering perspective:

- How to add the ML model to the existing workflows?
- What happens if the input is incorrect?
- How to still operate the workflow if the ML model is unavailable or inaccurate?
- Performance?
- Scalability?
- Security?
- How are the deployment cycles organized within an organization?
- …

Many large companies with established internal data organization are operating the following workflow within their teams:

- Data Scientist are deploying new models and testing the results. Once the new models are production ready, they are delivered to Engineering teams.
- Data Engineers are responsible to deploy new ML models developed by Data Science teams into production, since they have a better understanding of the architecture and infrastructure in general.
- A common deployment cycle for a new model is 2 weeks.

Data Engineers

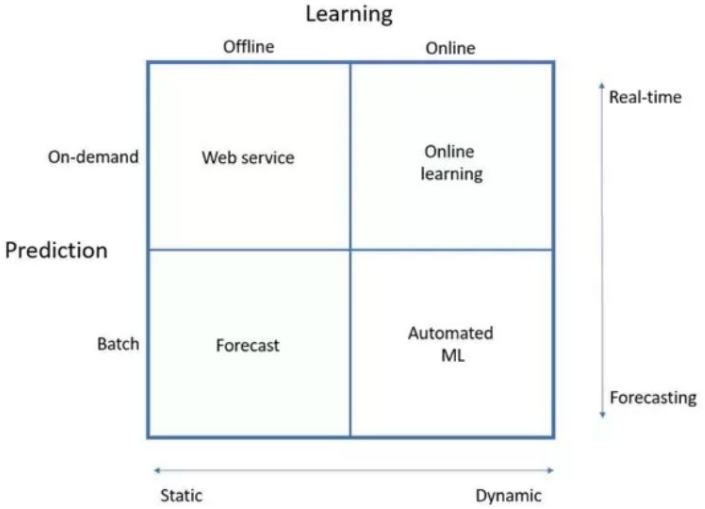
Data Scientists

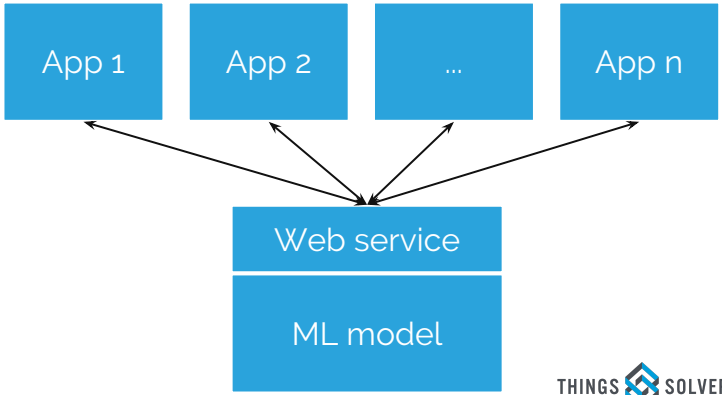
Production

A common production implementation of ML models consists of:

- Serializing a model in a usable format (Python pickle for example)
- Build a web service that either:
  - Handles a single record at a time (the data is being sent as a request parameter).
  - Triggers a batch run (with providing the batch input data location as a request parameter).
- Or run a model in a Docker container.

Main benefit of this approach is it's micro-service oriented architecture - the model can operate as an independent component in the ecosystem and would not depend on the application calling it (**ML model as a service**).



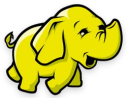


Scalability

The complexity in Data Science is increasing by the day. This complexity is driven by three fundamental factors:

- **Increased Data Generation.**
- **Low cost of data storage.**
- **Cheap computational power.**

These three characteristics enables us to benefit from using the data in it's raw format, but require using more sophisticated technologies and tools for various phases in the workflow, such as:




Apache Flink

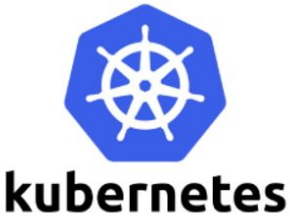
Event Stream Processing

Scheduling, Orchestrating, Workflow

In order for the whole pipeline to operate properly and in the right order proper scheduling and orchestrating mechanisms need to be implemented.

This part of the architecture is primarily directed to:

- Schedule parts of the workflow to run at required time, and in the right order
- Orchestrate aligning the business request with the applications, data, and infrastructure.
- Build automated workflows, enable provisioning and change managements.
- Maintain and operate the application workflow



Monitoring and Alerting

After building, training and deploying the models to production, it is required to have a proper monitoring systems in place. A crucial component to ensuring the success of the model is being able to measure and quantify their performance. A number of questions are worth answering in this area:

- How does my model affect the overall system performance?
- Which numbers do I measure?
- Does the model correctly handle all possible inputs and scenarios?
- Is the model taking too much time to provide the results?

Proper logging mechanisms need to be implemented in order to have monitoring and alerting capabilities.

Apache Airflow provides the mechanisms to monitor the workflow execution and notify the relevant stakeholders and trigger actions in case that something in the flow is failing. For realtime systems, solutions like ELK Stack and Splunk provide capabilities to monitor applications in realtime.

splunk>

Data Scientist's Toolbox

Data Science Tools and Skills

Things Solver

ENLIGHTEN YOUR DATA