Data Science Pipeline
From raw data to insights

Renata Borovica-Gajic & Goce Ristanoski

Data is invaluable

Bloomberg, Stock market°

Atlas experiment, CERN*

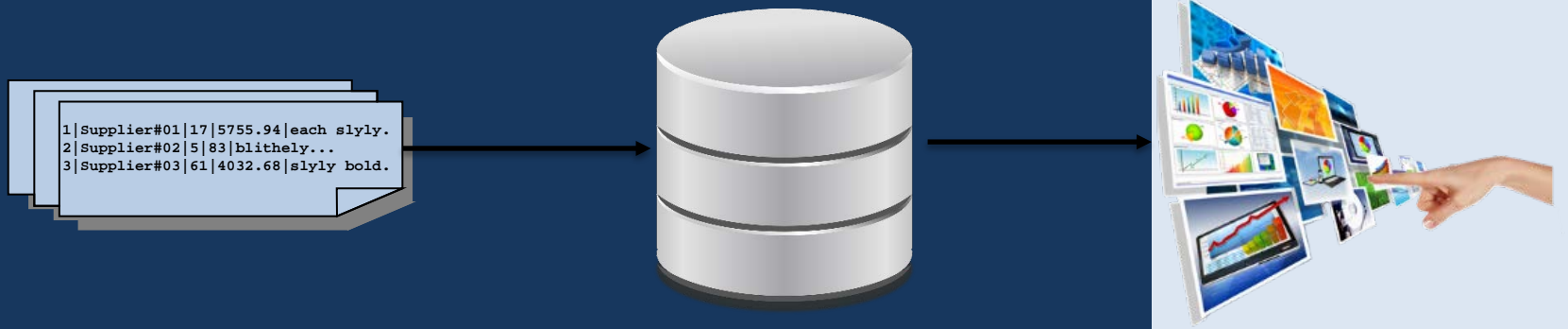**Fuels decision-making and reveals insights …**

Finding hidden gold is hard

And expensive
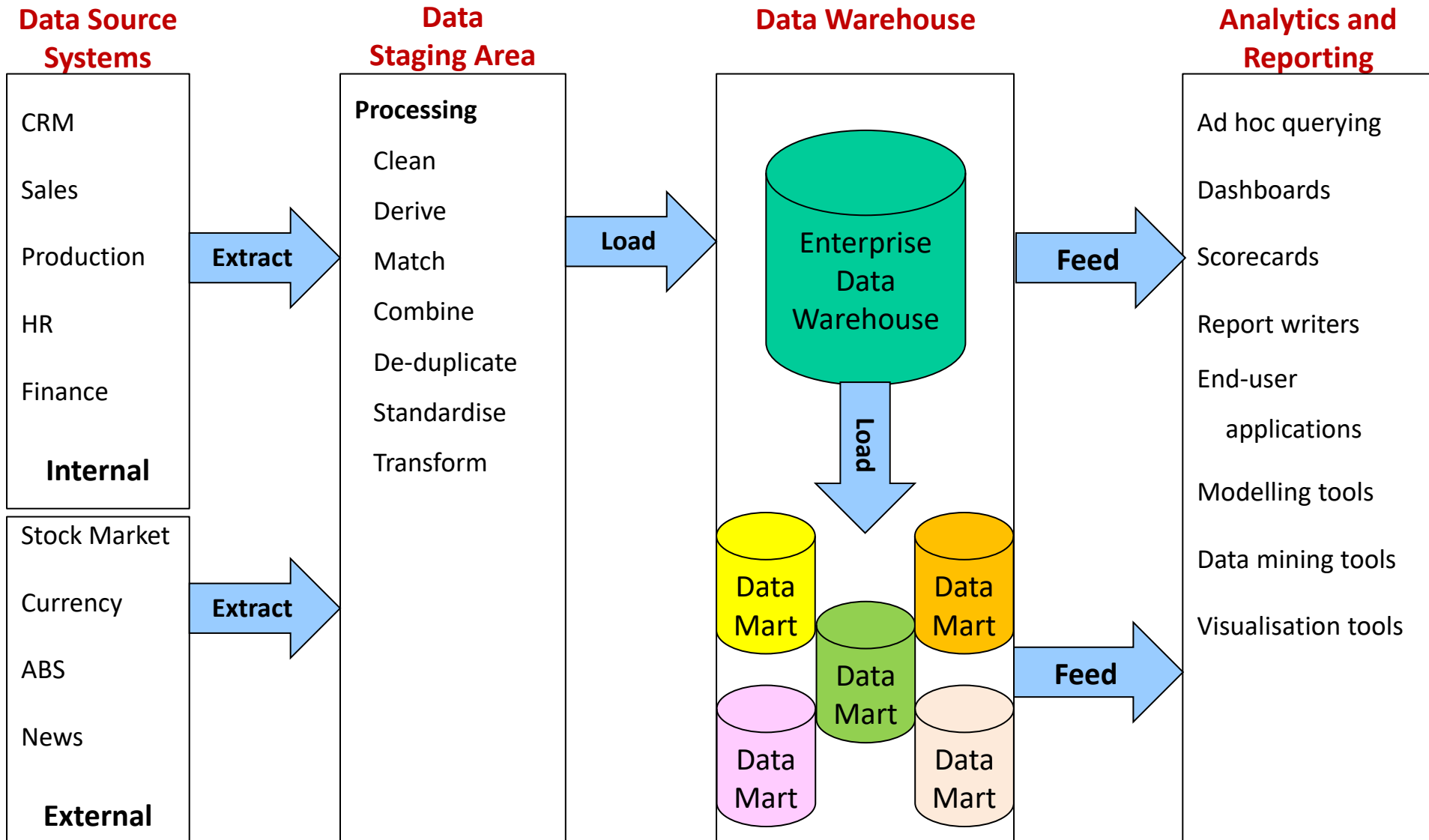
Data analysis pipeline: long and cumbersome

```
1|Supplier#01|17|5755.94|each slyly.
2|Supplier#02|5|83|blithely...
3|Supplier#03|61|4032.68|slyly bold.
```
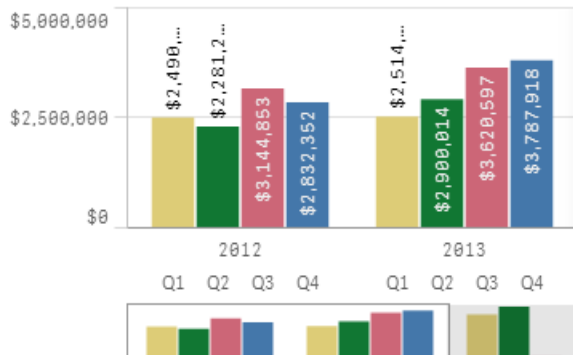
| Data Acquisition | Data cleaning | Record linkage (ETL) | DBMS tuning | Feature engineering | Predictive analytics | Validation & deployment |

Data engineering

Machine learning

Pipeline within a company

Business Intelligence Dashboards
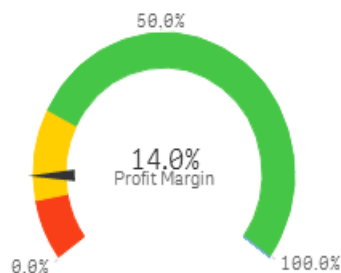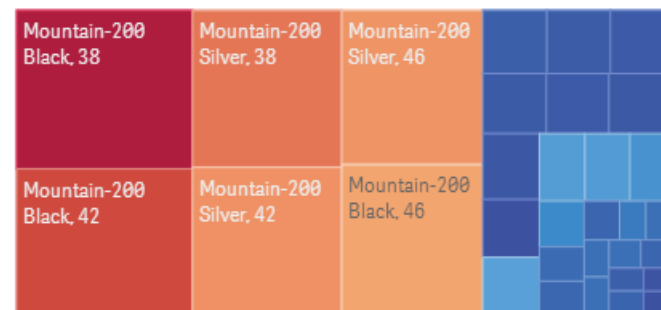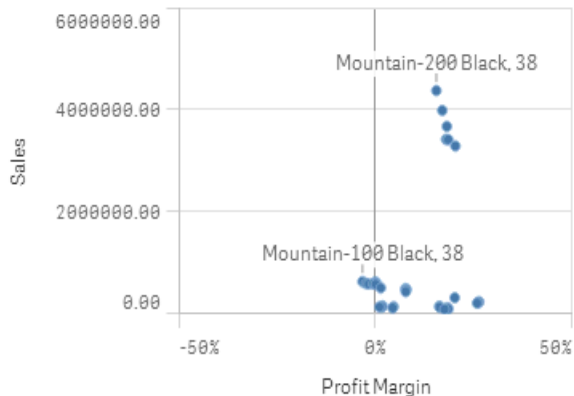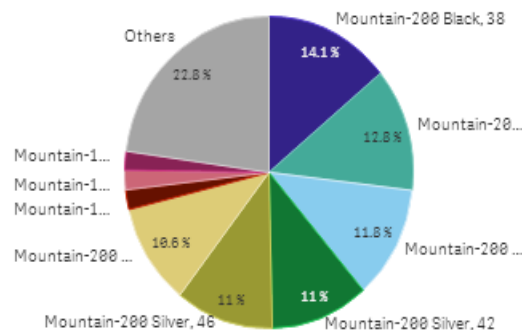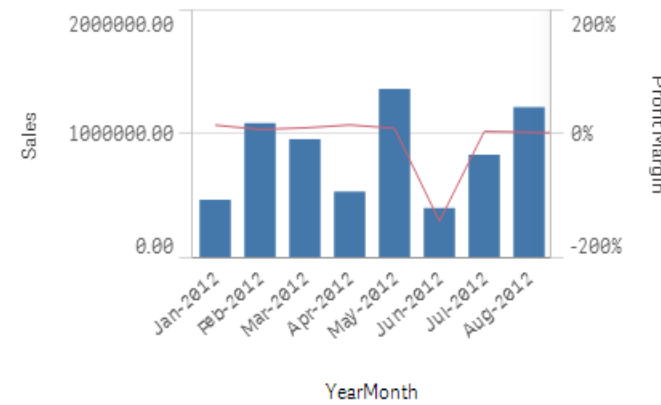
Advanced (Predictive) Analytics

SUPERVISED LEARNING

Reliance on algorithm trained by human input, reduced expenditure on manual review for relevance and coding

Raw Data — Sample Data, Code and test new sample data - Feedback — Algorithm — Product of trained algorithm — Manual Verification — Production

E-Discovery Concepts: Machine Learning

Original unclustered data

Clustered data
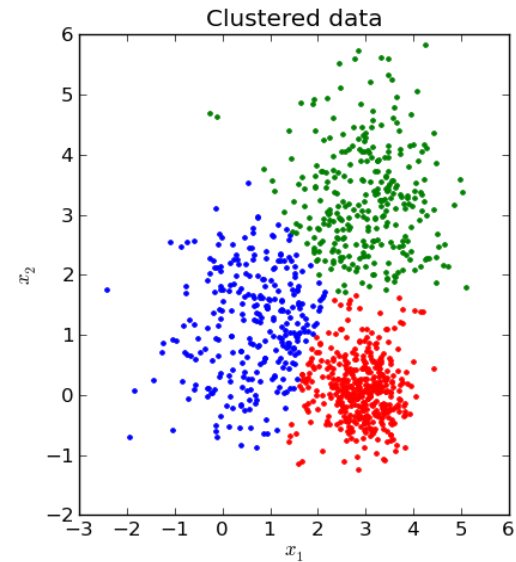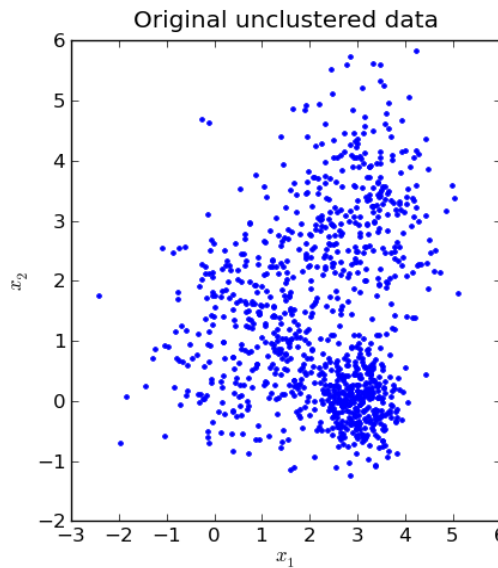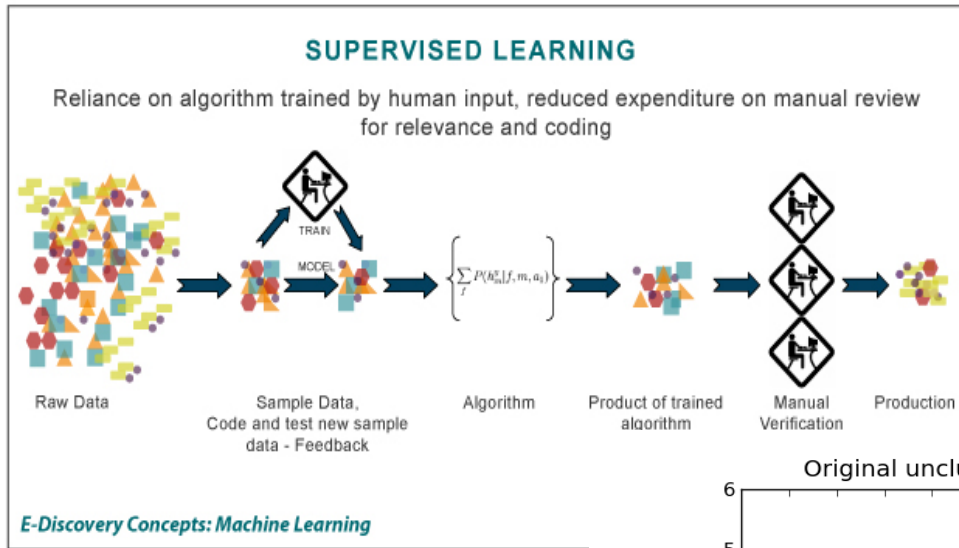
AI, Data Engineering and Data Science

- Setting up a well defined project:
  - Project definition: what is the opportunity of interest/hypothesis we are investigating;
  - Data Engineering: Linking the data sources and preparing them for analysis;
  - Data Science: testing variables, building models, delivering insights:
  - Project deliverable: actioning the AI in our work, increasing capabilities, adding value and commercializing AI products.
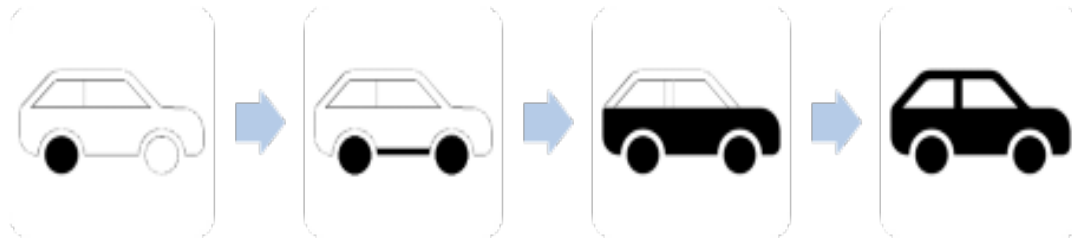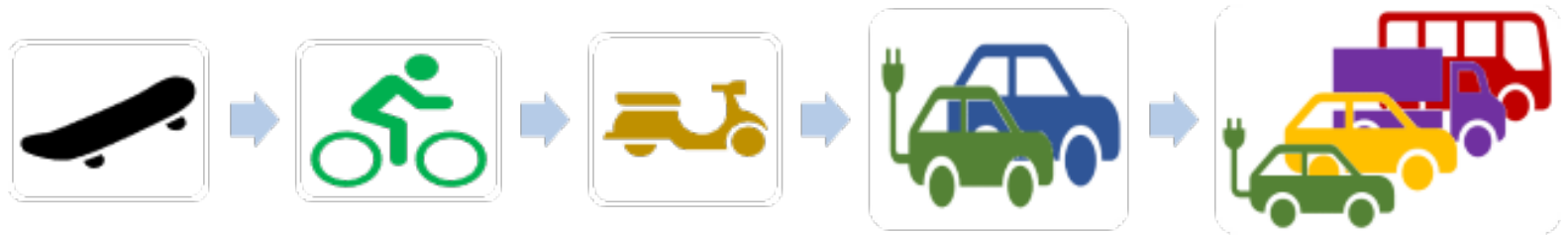
Data Science upward trajectory
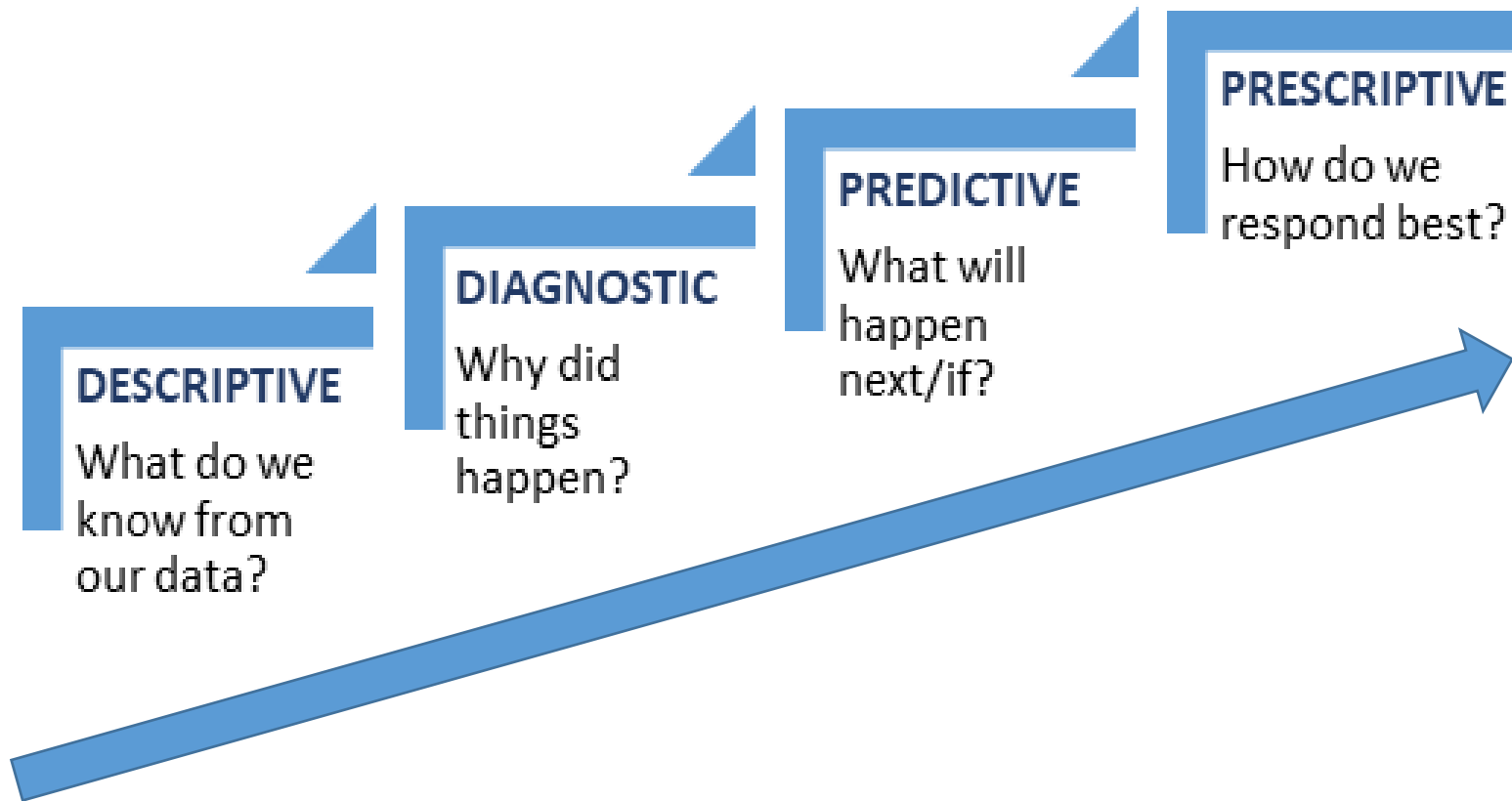
Bridge between Data Science and stakeholders
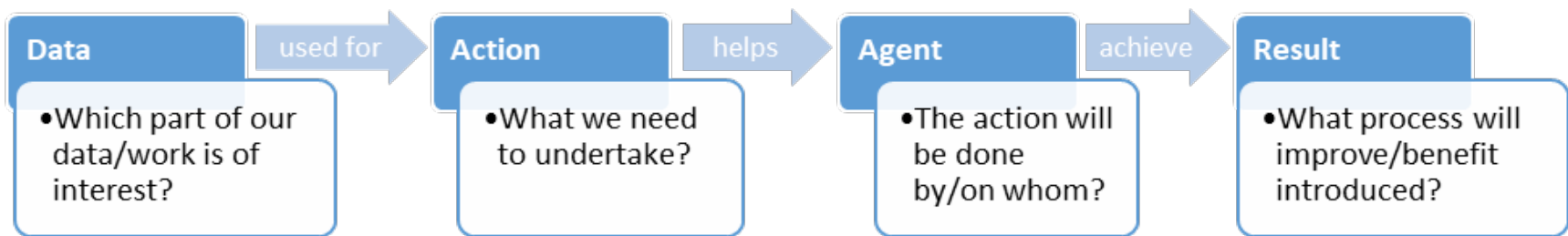
One day at a time …

VS

Some pointers on skills required

- **A Beginner's Guide to the Data Science Pipeline**

https://towardsdatascience.com/a-beginners-guide-to-the-data-science-pipeline-a4904b2d8ad3

- **Data, structure, and the data science pipeline**

https://developer.ibm.com/technologies/data-science/articles/ba-intro-data-science-1/