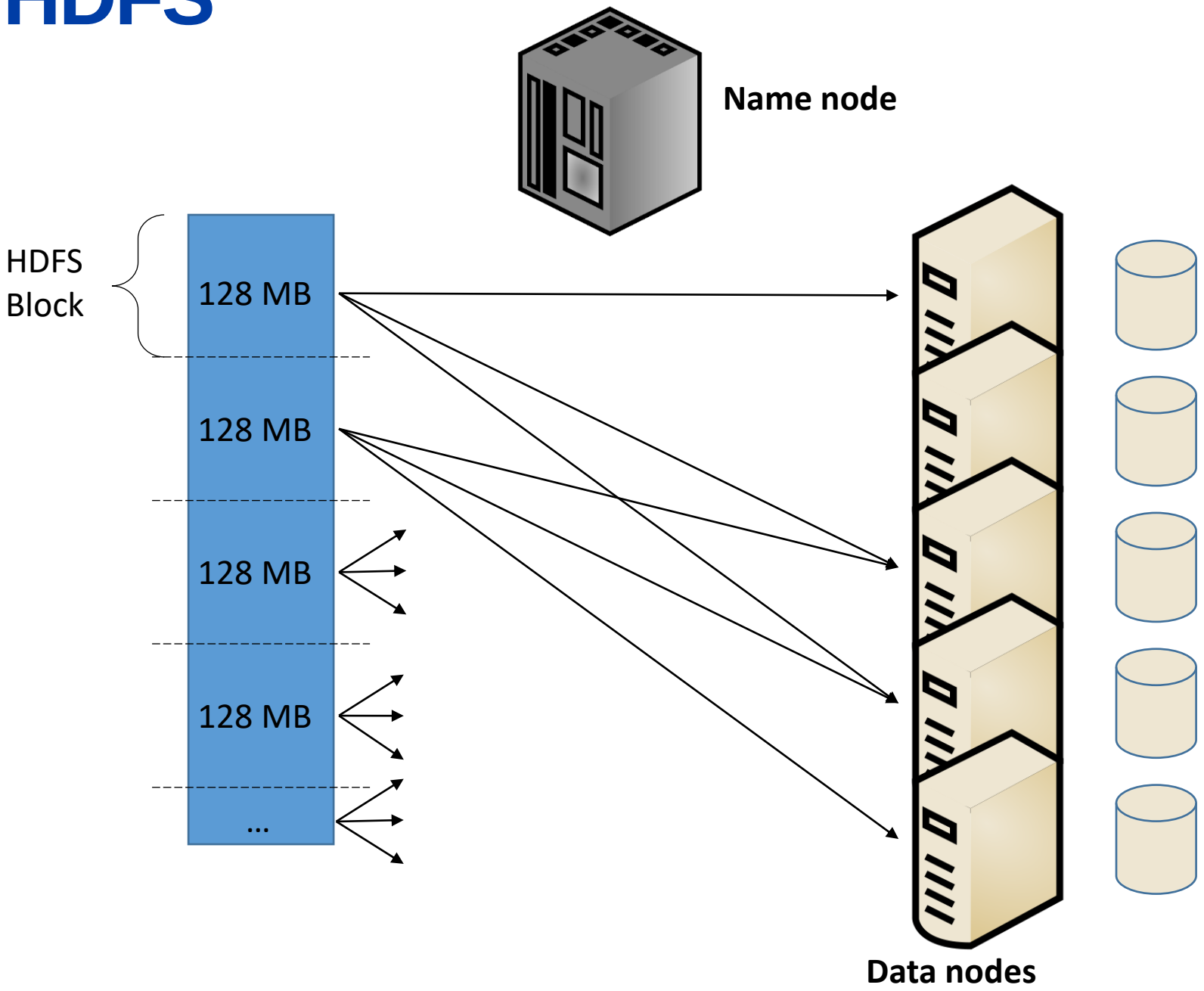
Introduction to Big-data Management

Review and next steps

What We Covered

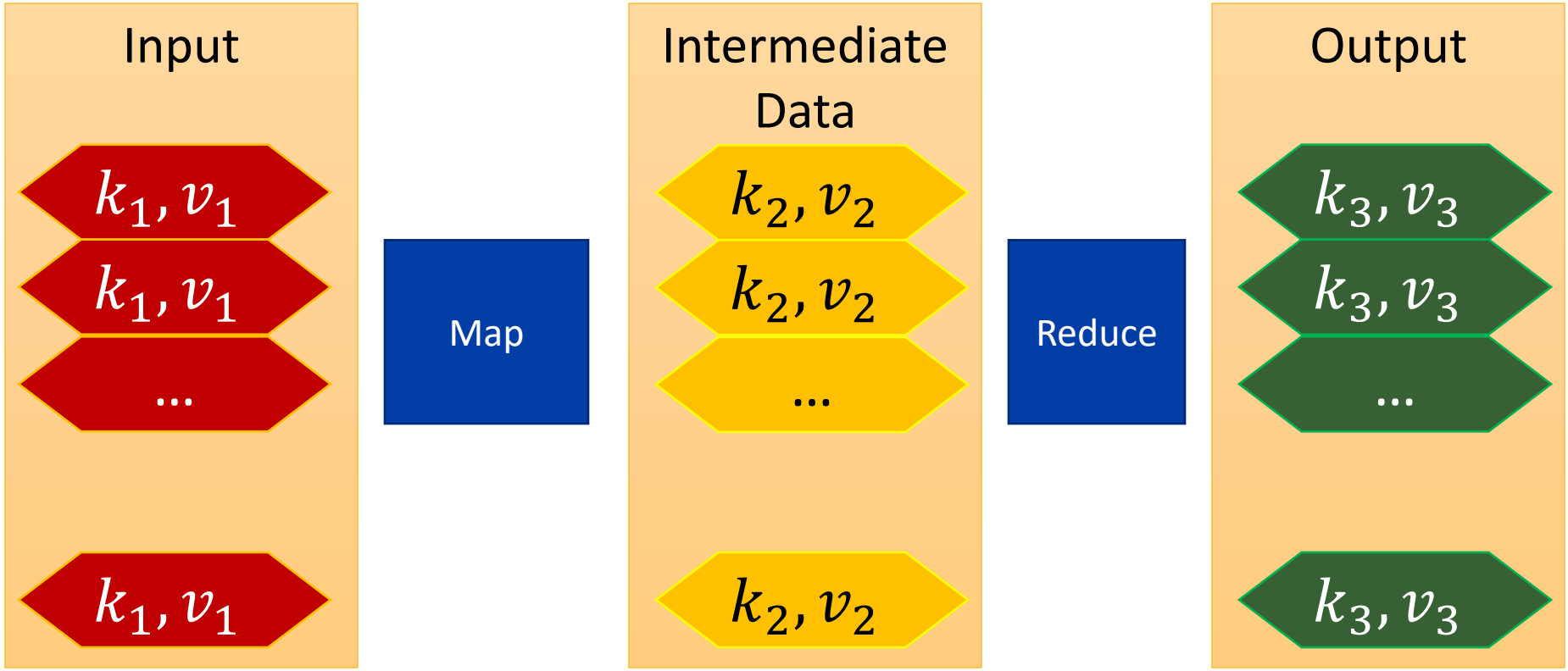
- Storage (HDFS)
- Query processing (MapReduce, RDD, Hyracks)
- Higher-level data flow engines (Pig, SparkSQL)
- Storage formats (row, column, Parquet, LSM indexing)
- Document databases (MongoDB)
- Machine learning (MLlib)

HDFS



Logical View of MapReduce

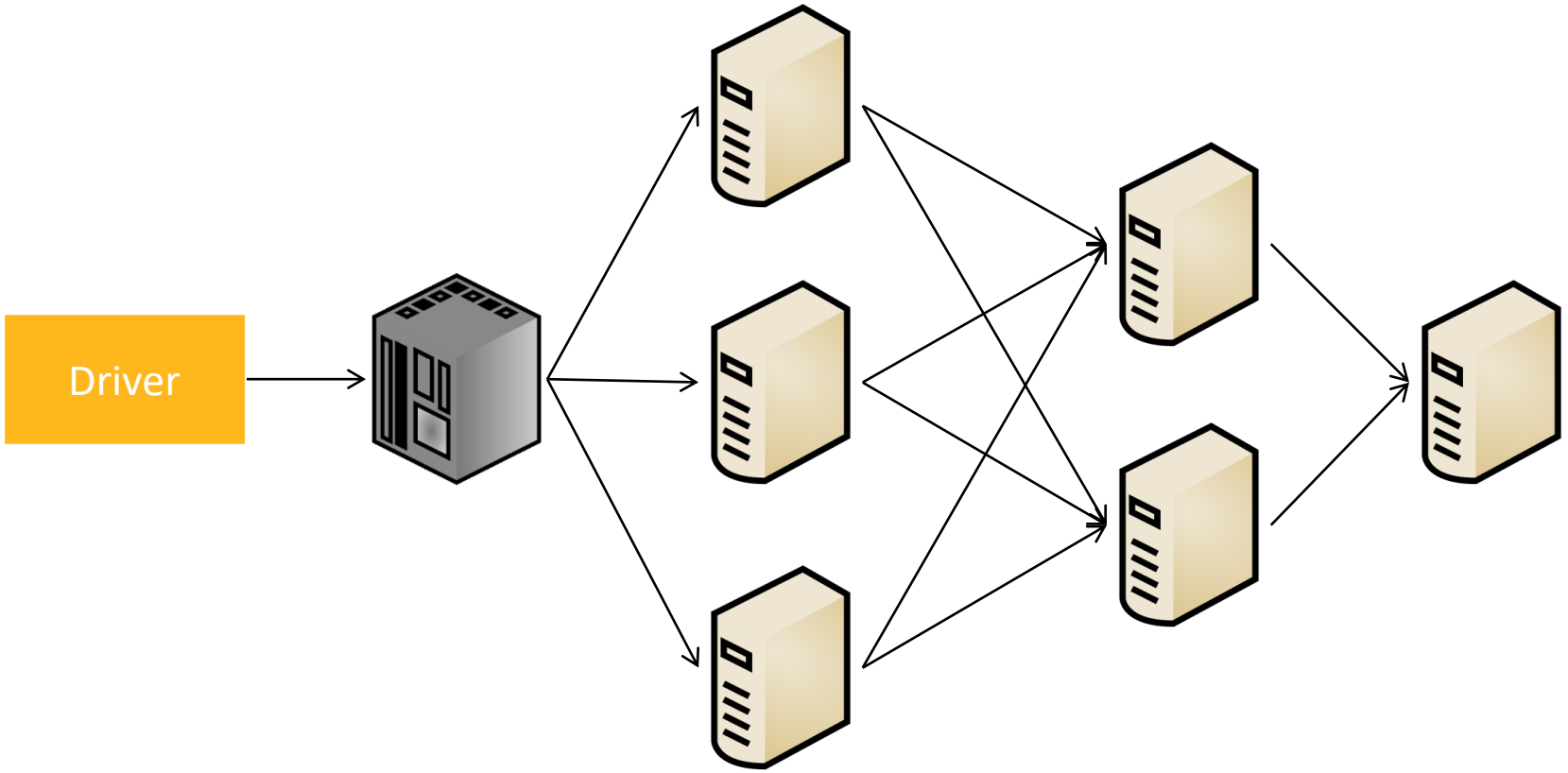
- During MapReduce, the input and output are considered a set of key-value pairs $\langle k, v \rangle$



Map and Reduce Functions

- Map Function
  - Maps a single input record to a set (possibly empty) of intermediate records
  - Map: $\langle k_1, v_1 \rangle \rightarrow \{\langle k_2, v_2 \rangle\}$
- Combine Function
  - Combine: $\langle k_2, \{v_2\} \rangle \rightarrow \{\langle k_2, v_2 \rangle\}$
- Reduce Function
  - Reduces a set of intermediate records with the same key to a set (possibly empty) of output records
  - Reduce: $\langle k_2, \{v_2\} \rangle \rightarrow \{\langle k_3, v_3 \rangle\}$

Job Execution Overview



Job submission | Job preparation | Map, Combine | Shuffle | Reduce | Cleanup

Resilient Distributed Dataset (RDD)

- RDD is a pointer to a distributed dataset
- Stores information about how to compute the data rather than where the data is
- Transformation: Converts an RDD to another RDD
- Action: Returns an answer of an operation over an RDD
- Narrow Vs wide dependencies
- How RDD operations work

SparkSQL

Dataframe (SparkSQL)

- Lazy execution
- Spark is aware of the data model
- Spark is aware of the query logic
- Can optimize the query

RDD

- Lazy execution
- The data model is hidden from Spark
- The transformations and actions are black boxes
- Cannot optimize the query

MLlib

- Main components of MLlib
  - Transformers, e.g., feature extraction
  - Estimator, e.g., clustering or regression
  - Evaluator, e.g., precision and recall calculation
  - Validator, e.g., k-fold cross validation
- Pipeline: Transformation(s) + Estimator

Big Spatial Data

- How to customize Spark for a specific domain, i.e., spatial data
- Support various file formats other than the regular text files
- Build complex query pipelines such as spatial join and visualization
- Combine spatial operations with regular Spark operations

Storage formats

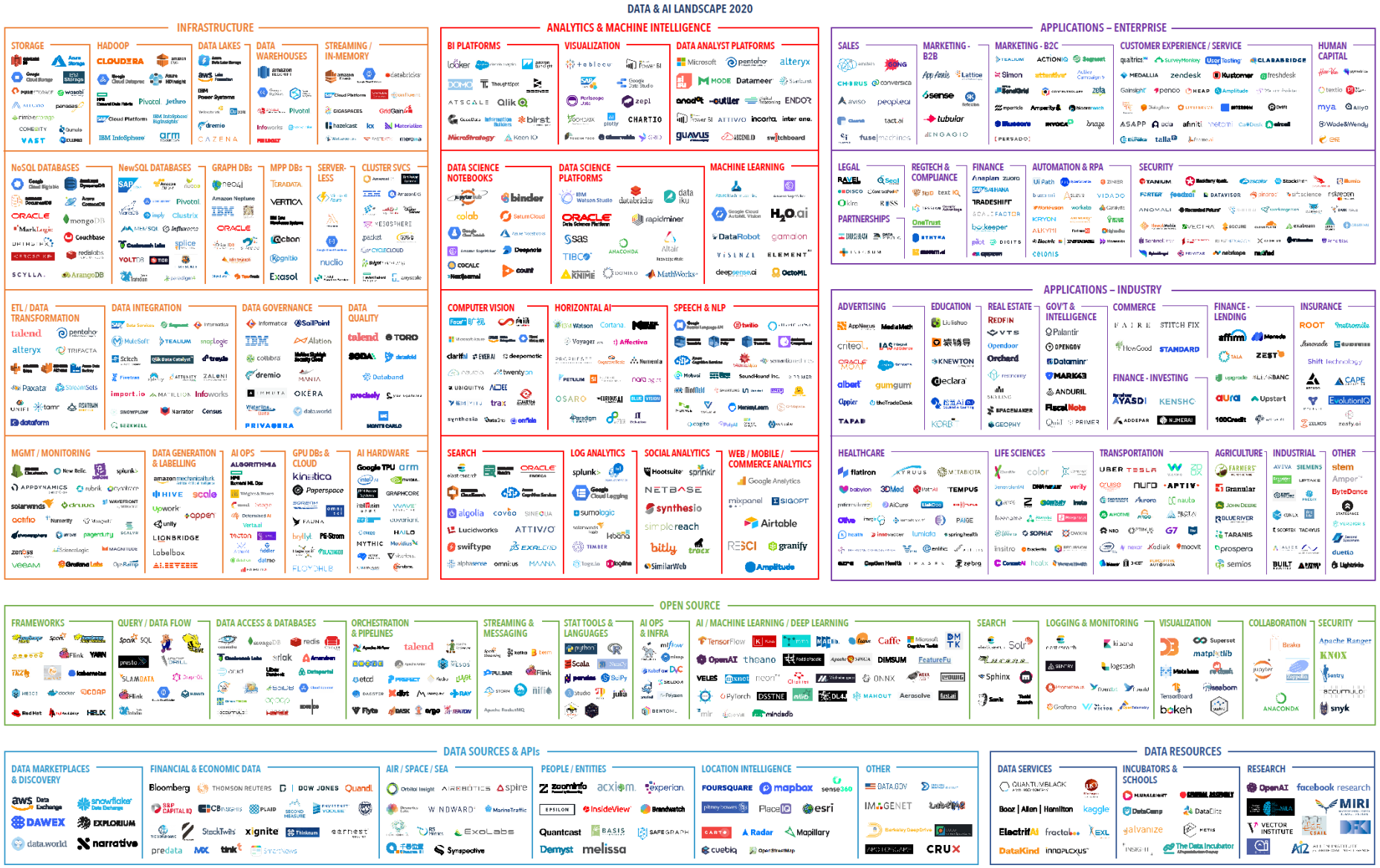
- Difference between row and column formats
  - How attributes map to disk
  - Major applications for each of them
- Parquet files
  - A column store file format
  - Handles nesting and replication
  - Schema → Maximum definition and repetition level
  - Record → Definition and repetition level for each attribute
  - Do not forget to add null (non-existent) attributes

Document databases

- How a document database compares to a relational database (RDBMS)
  - Normalization (nesting and repetition)
  - ACID compliance
- How MongoDB compares attributes
- Log-structured-merge (LSM) tree for big data indexing

Did we cover everything?

Big Data Landscape



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Topics not Covered

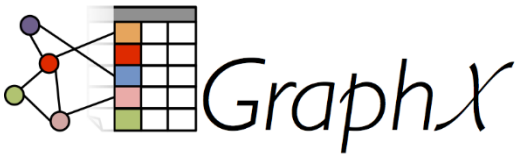
- Key-value stores
- Big graph analytics
- Visualization
- Streaming
- Coordination
- Cloud platforms

Key-value Stores

- Provide a simple API to insert/delete/update/search key-value pairs
- Records are indexed by key (typically a string)
- Internal structure is typically a Log-structured-merge tree (LSM)
- Not generally suitable for large-scale analytics

Big Graph Analytics

- Graphs are usually processed using a node-centric processing model
- Nodes and edges are both treated as first-class citizens
- Processing is normally iterative with a lot of iterations



Visualization

- Sometimes called Business Intelligence (BI)

- Focuses more on the end-user interface while producing nice graphs (e.g., bar charts and line graphs)
- Internally, the data is managed using the common big-data platforms but the systems are tuned to provide fast query response for ad-hoc queries

Streaming

- Some applications need to process data in real-time with a very small latency
- Examples: Twitter search, IoT applications, and social network trends
- Works primarily off main memory
- Keeps only the latest records to ensure real-time response

Coordination

- Most big-data systems are designed for shared-nothing large-scale analytics
- No coordination between machines is part of the design
- Coordination systems provide an easy way to coordinate the work in these distributed platforms, e.g., a catalog of information, work queue, and a global system status

Apache Zookeeper

Cloud Platforms

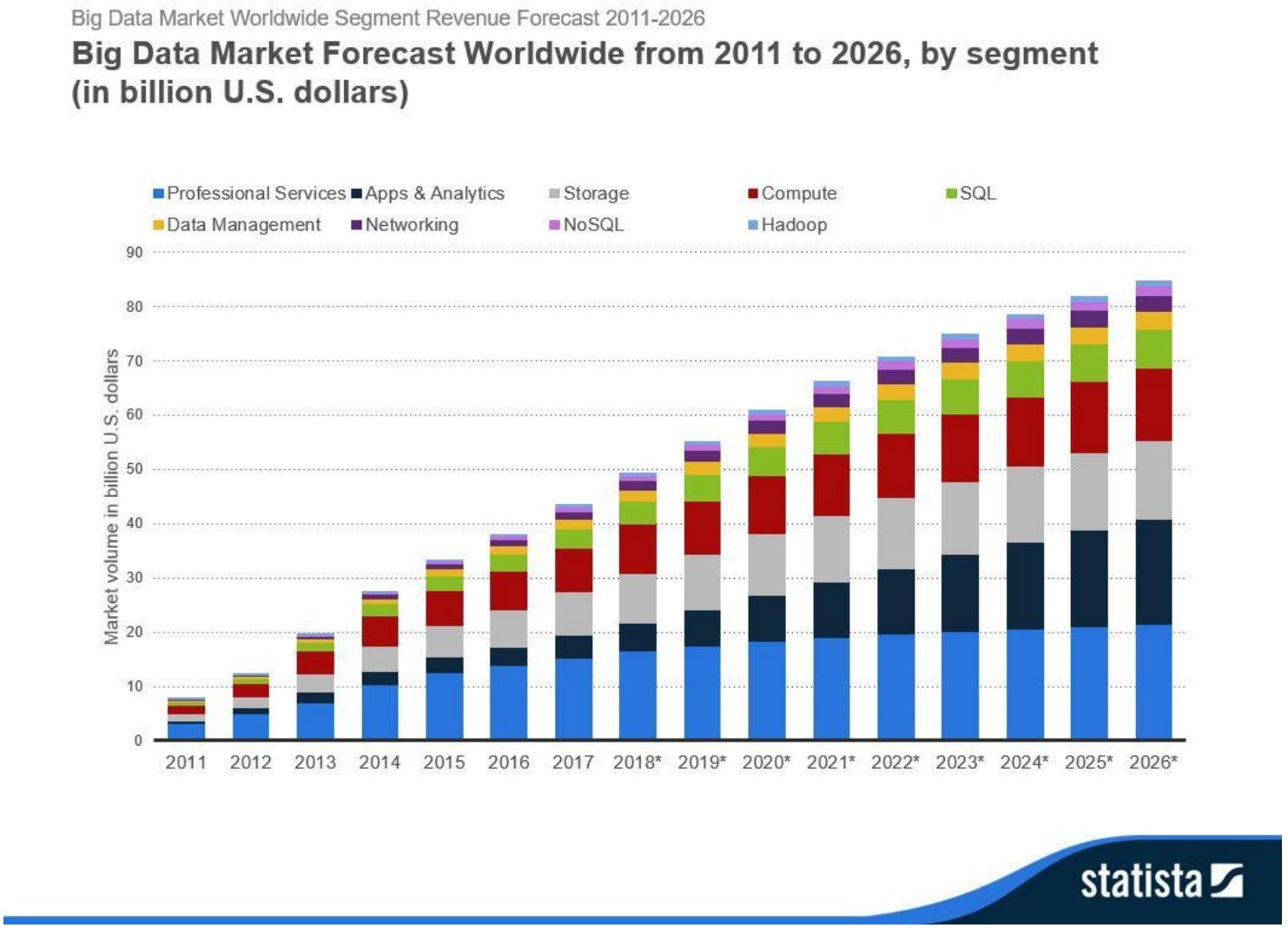
- Maintaining your own cluster is costly
- It could be underutilized most of the time
- Cloud platforms allow you to rent virtual machines to do your work and dispose them after
- They are well-integrated with big data platforms (such as Hadoop and Spark) to give the best user experience
- All you need is an internet connection and a credit card

What is next?

What is next?

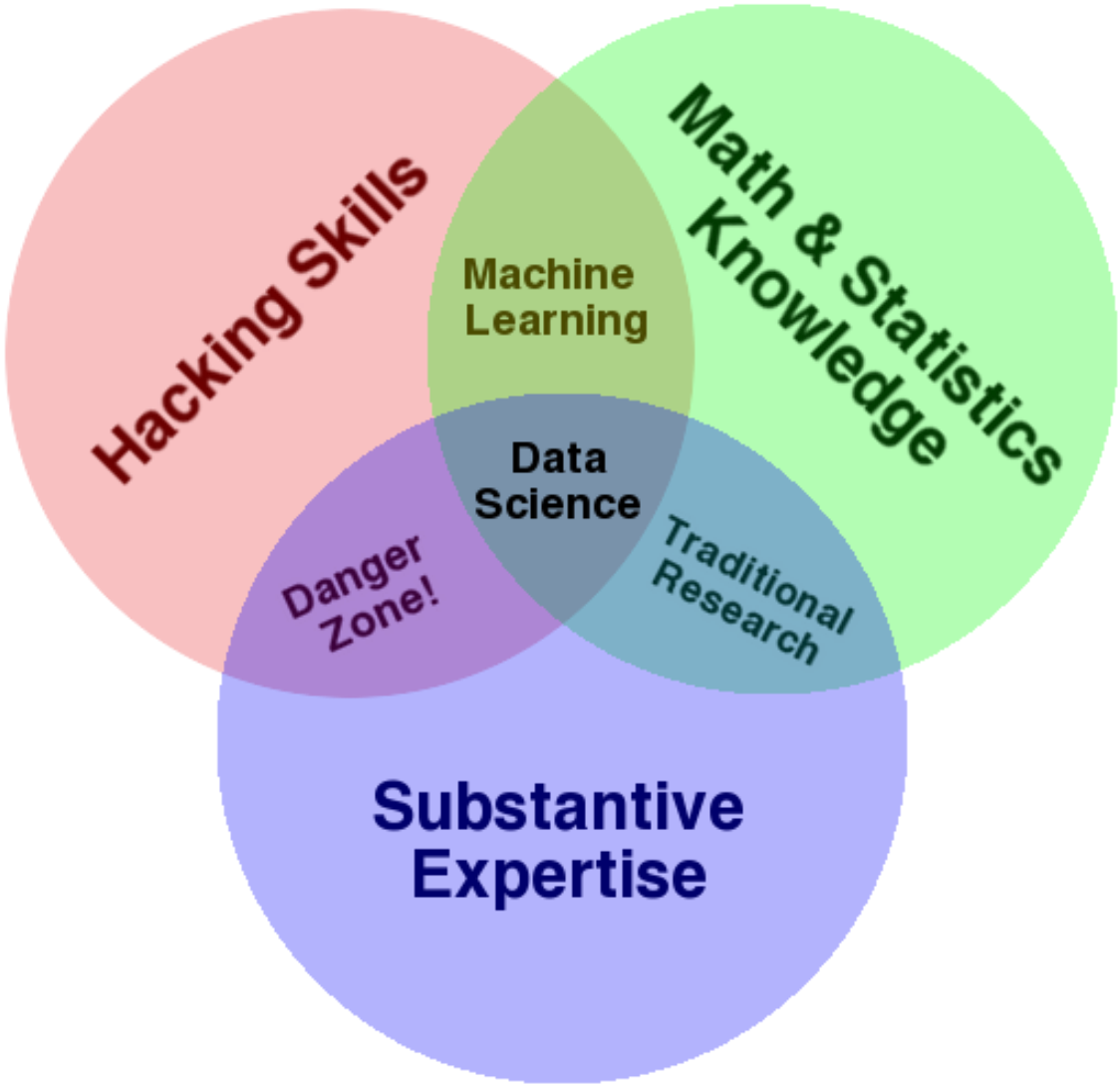
- Real big data is widely available
- Big data is like gold
- Only a few people know how to deal with it
- You're now one of them
- Applications
  - Keep your hands dirty
  - Consider using the public cloud (e.g., AWS, Google Cloud, or Microsoft Azure)

Job Market



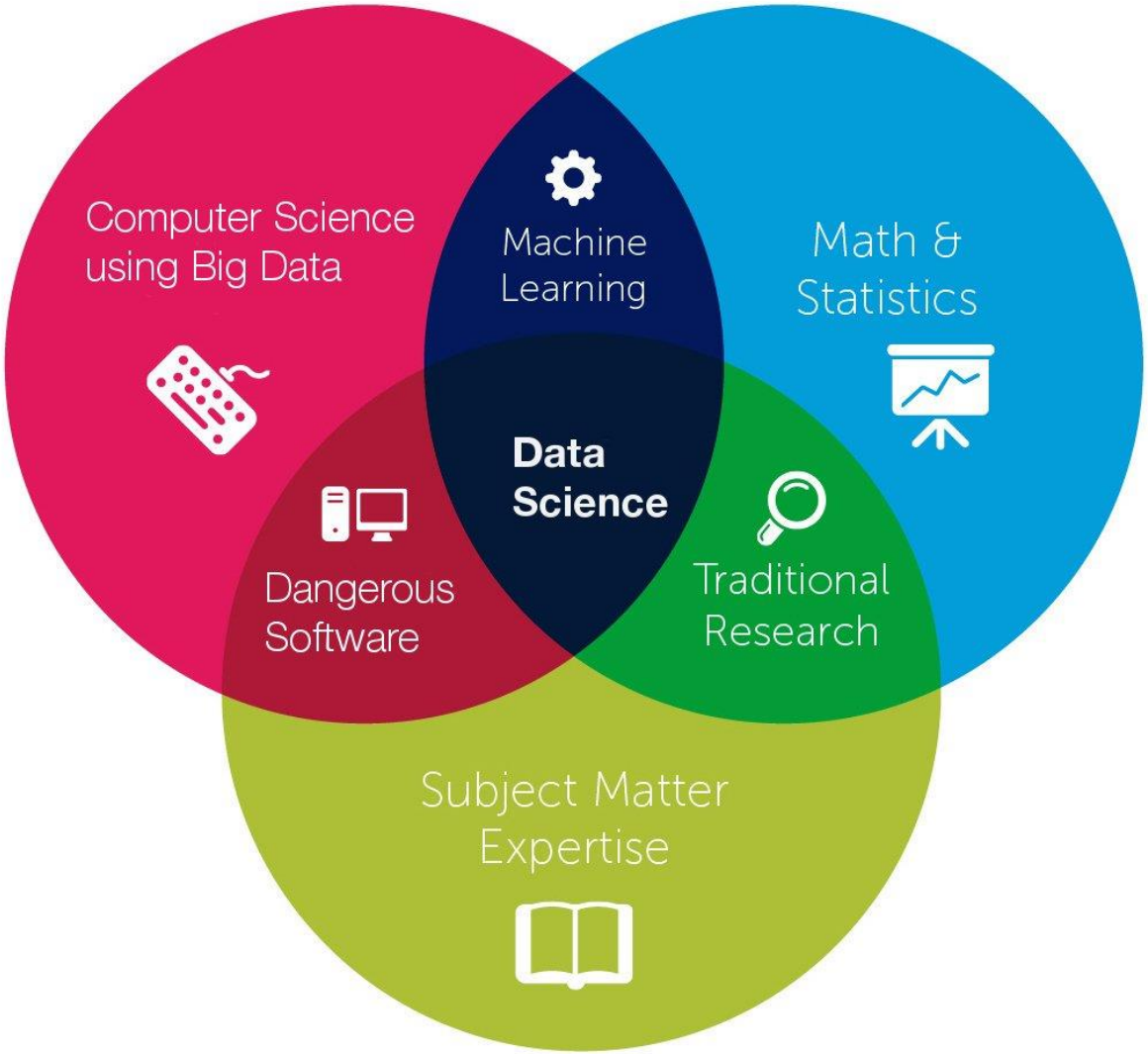
https://www.techicy.com/5-best-programming-languages-to-watch-out-in-2019-for-data-science.html

Data Science



Credits: Drew Conway

Data Science

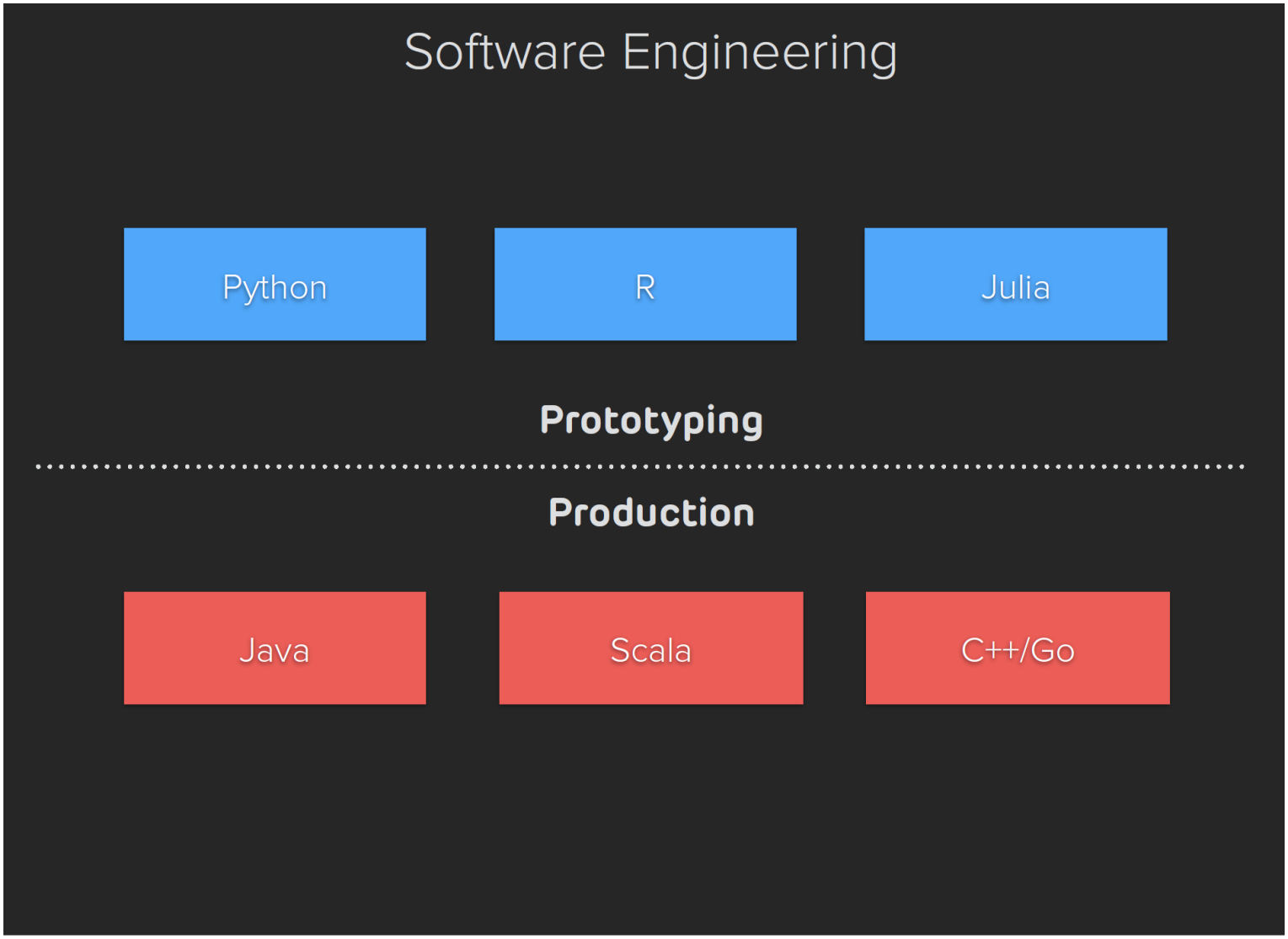


https://mashimo.wordpress.com/2016/05/28/big-data-data-science-and-machine-learning-explained/

Next Steps

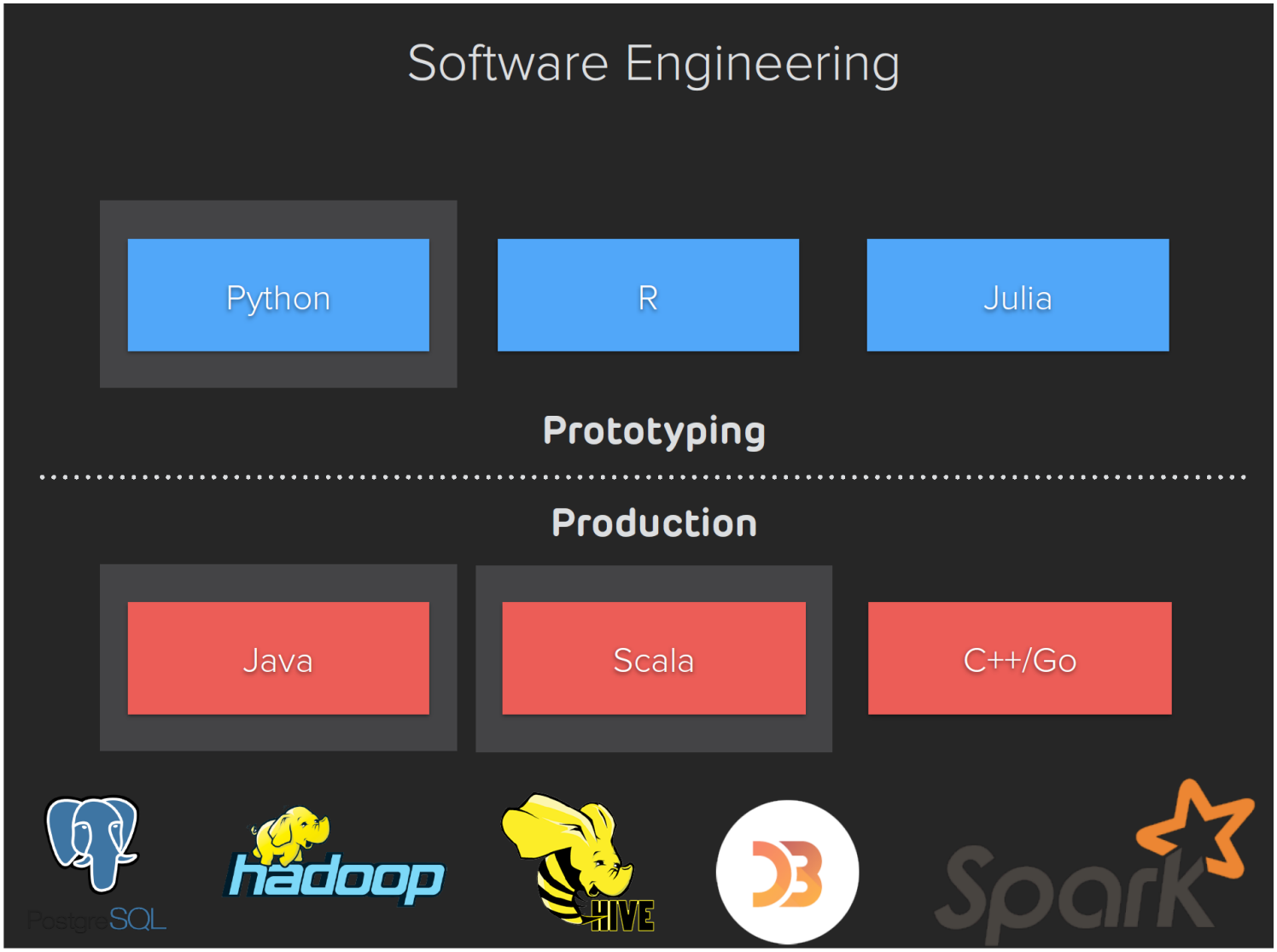
- CS
  - Big data tools
  - Python/R/Scala
- Math/Stats
  - Linear algebra
  - Correlation analysis
  - Hypothesis tests
- Collaboration with domain experts
  - Visualization
  - Prototyping

CS



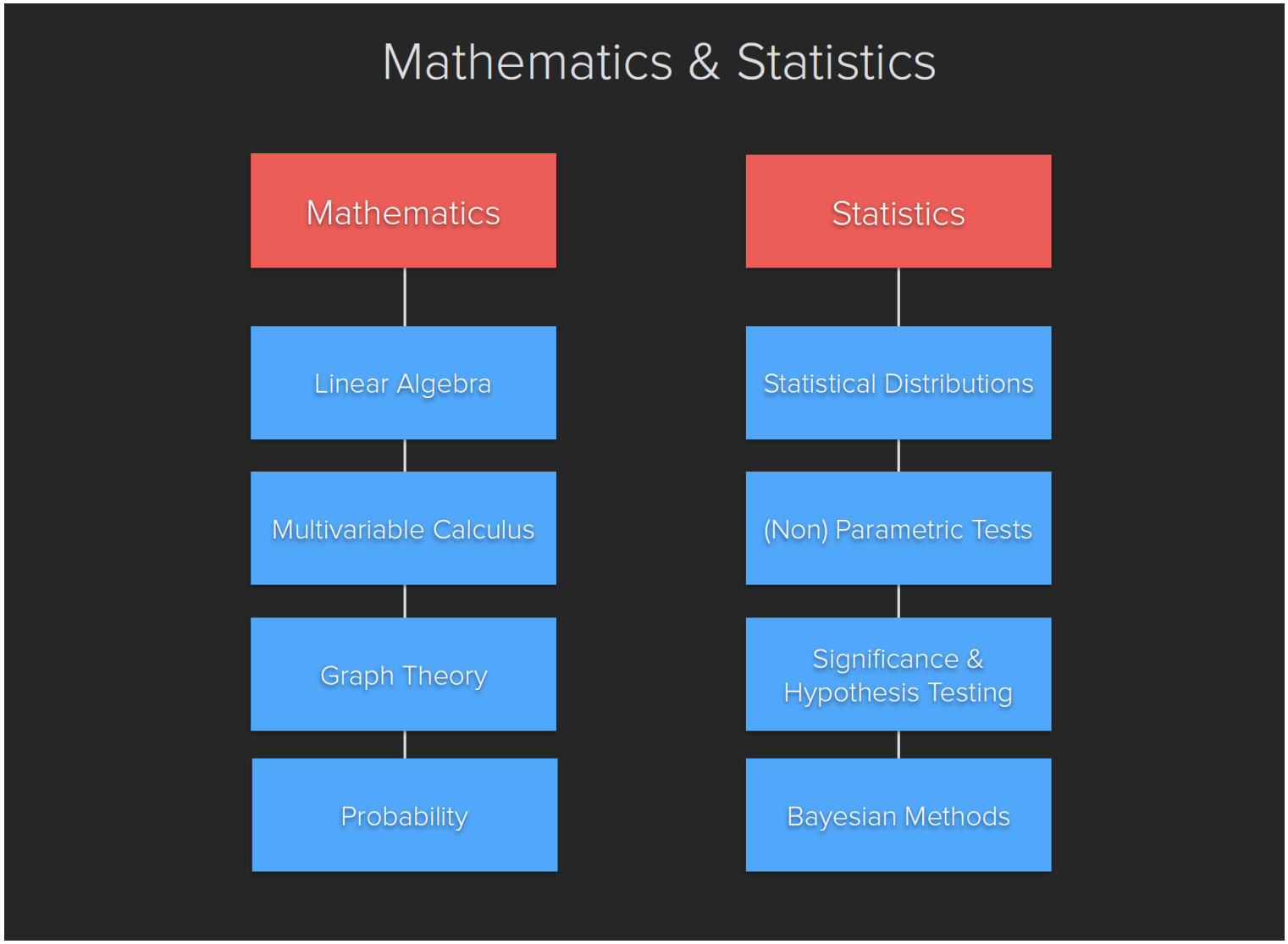
https://www.slideshare.net/galvanizeHQ/how-to-become-a-data-scientist-by-ryan-orban-vp-of-operations-and-expansion-galvanize

CS/Big Data



https://www.slideshare.net/galvanizeHQ/how-to-become-a-data-scientist-by-ryan-orban-vp-of-operations-and-expansion-galvanize

Math/Stats



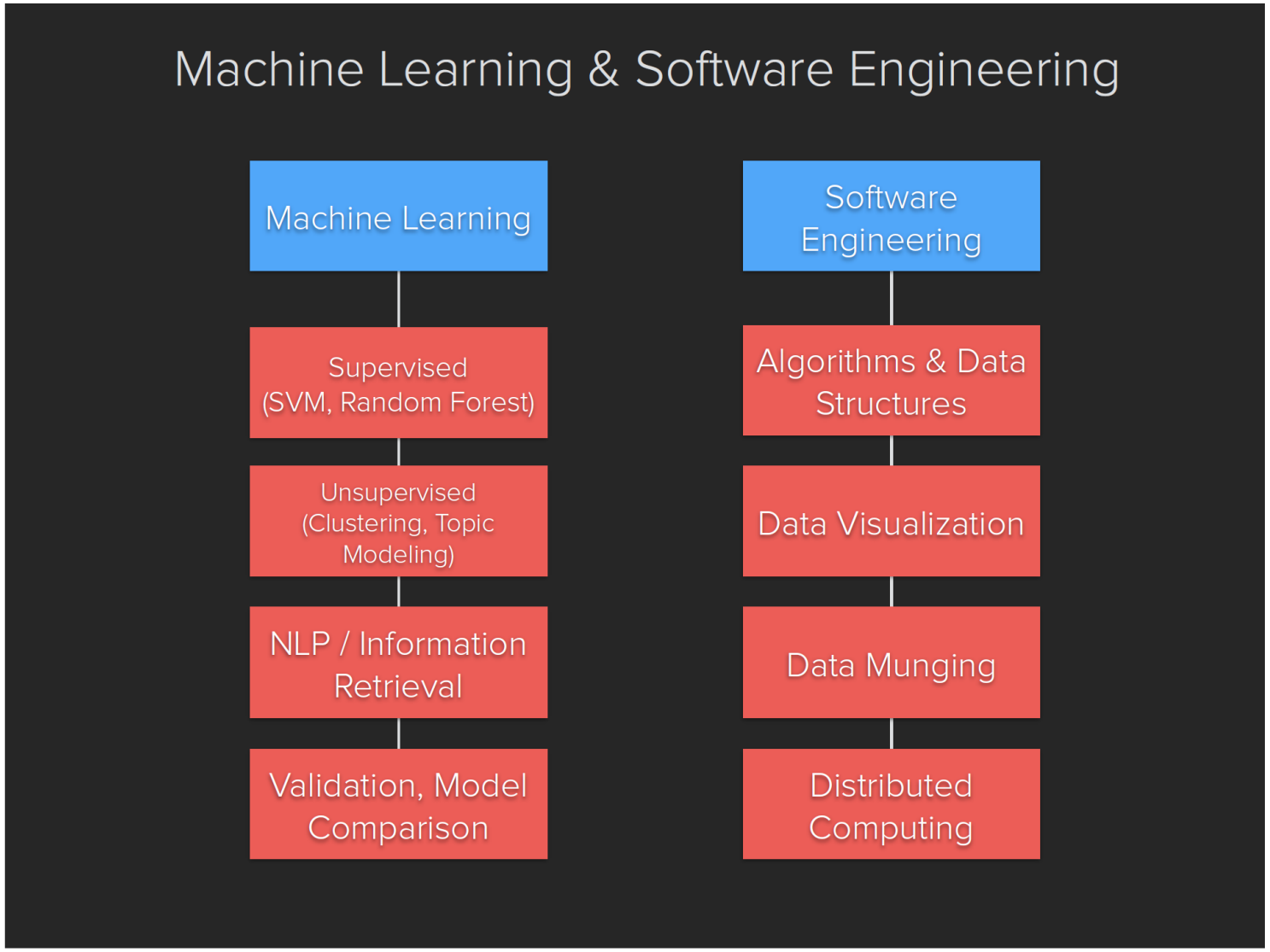
https://www.slideshare.net/galvanizeHQ/how-to-become-a-data-scientist-by-ryan-orban-vp-of-operations-and-expansion-galvanize

Online Courses

Mathematics & Statistics

https://www.slideshare.net/galvanizeHQ/how-to-become-a-data-scientist-by-ryan-orban-vp-of-operations-and-expansion-galvanize

Data Analytics



https://www.slideshare.net/galvanizeHQ/how-to-become-a-data-scientist-by-ryan-orban-vp-of-operations-and-expansion-galvanize

Thank You!

Good Luck ☺