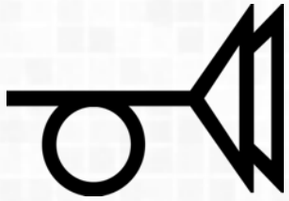
apache spark for everyone

amcasari + deb siegel

2016 May 05 - Seattle Spark Meetup

who: @amcasari  @ CONCURLABS @ CONCUR

@dsiegel  @ W

what: @SparkSeattle

where: @Concur

why: @ApacheSpark

(now we can be found)

.....

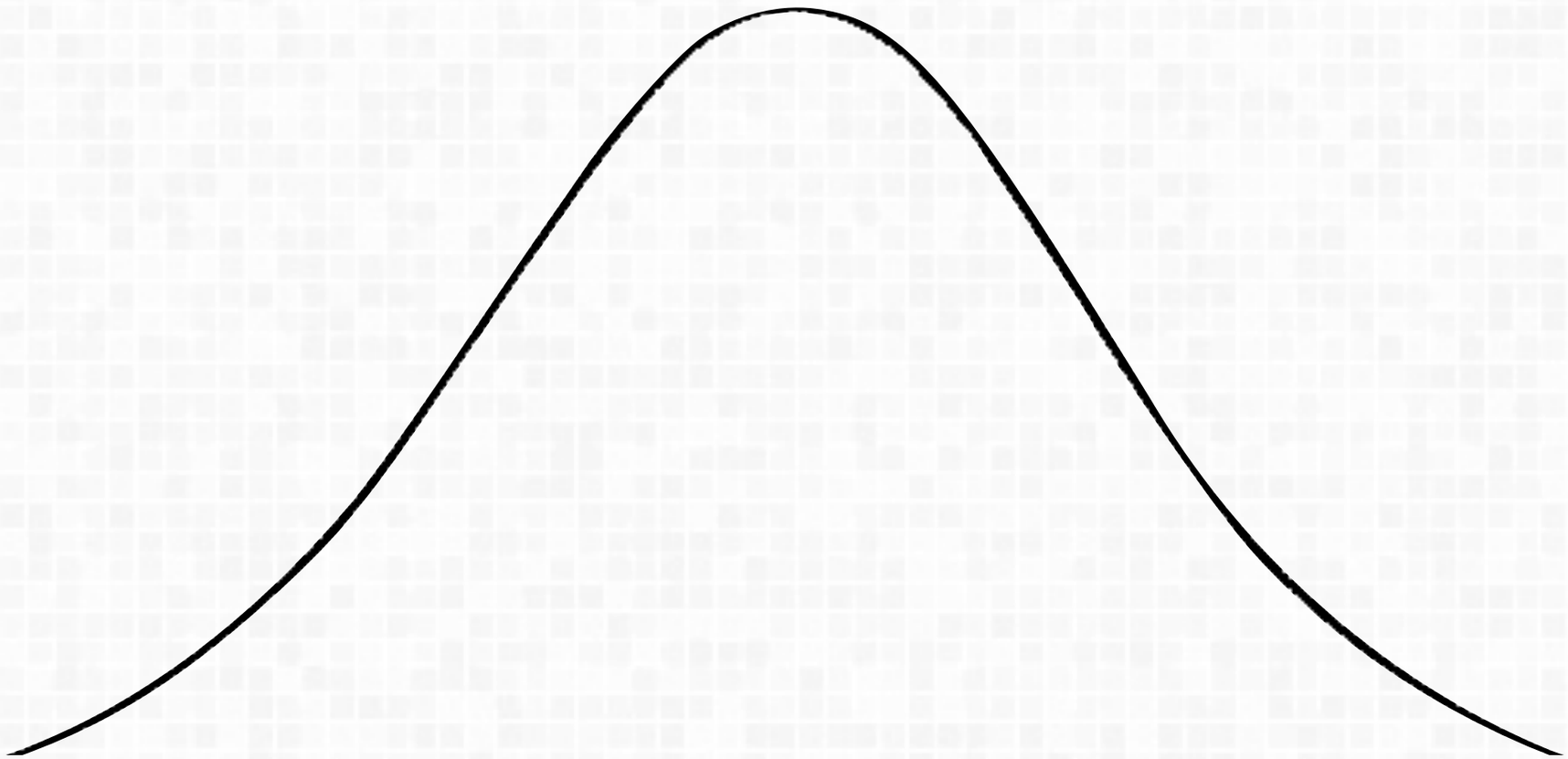
COORDINATES

https://github.com/morningc/wwconnect-2016-spark4everyone

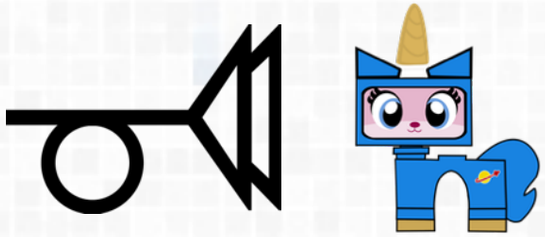
{now you are safe take a nap….}

courtesy YouTube

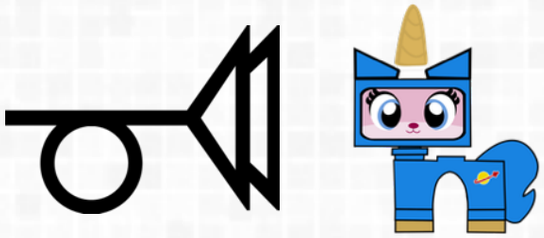
don’t worry about this….



you

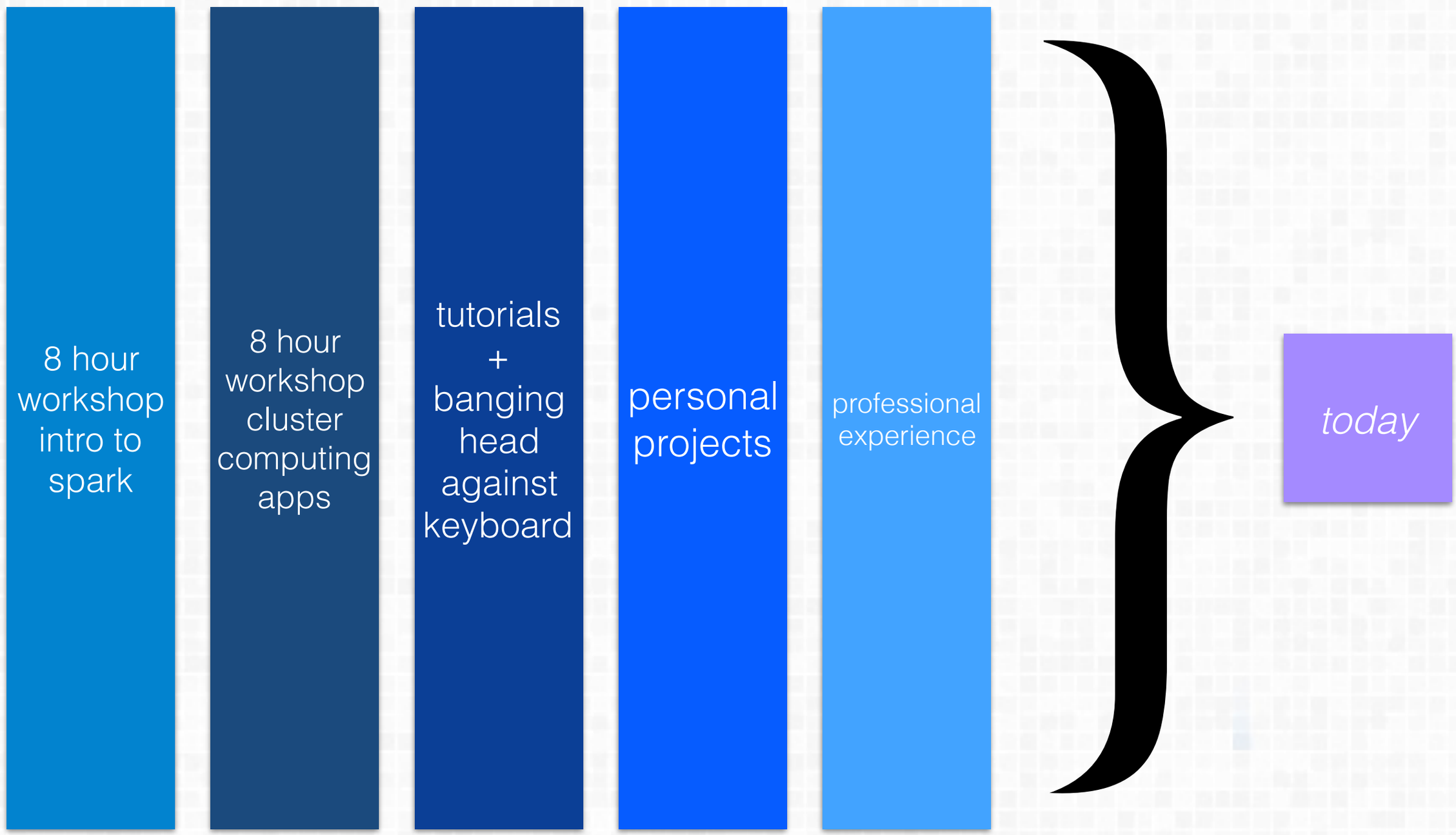


because we feel the same way…
we are all learning!



you

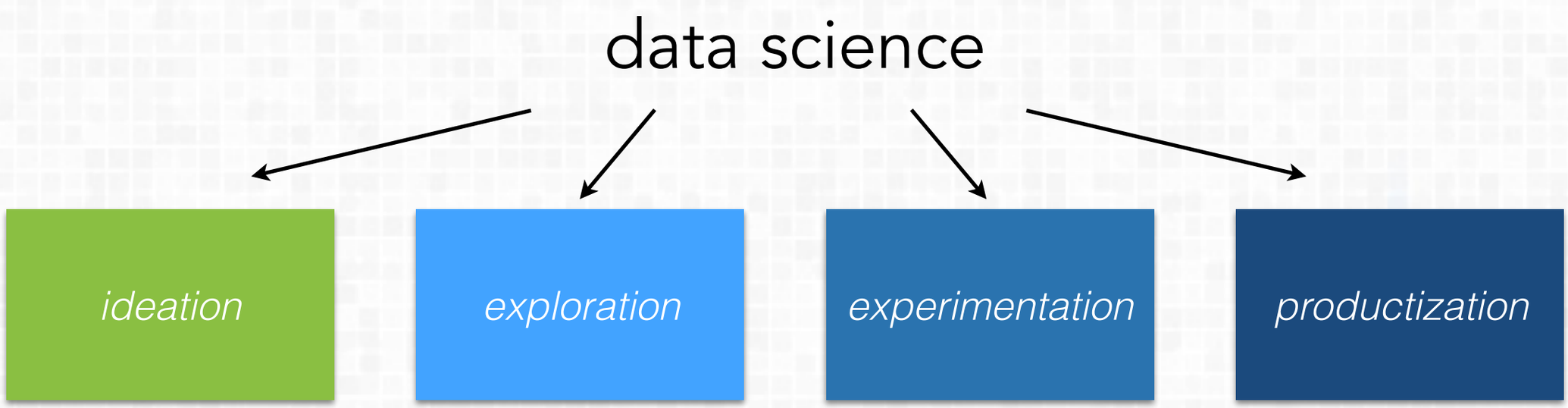
we might be a wee bit ambitious…



why do we care about spark?

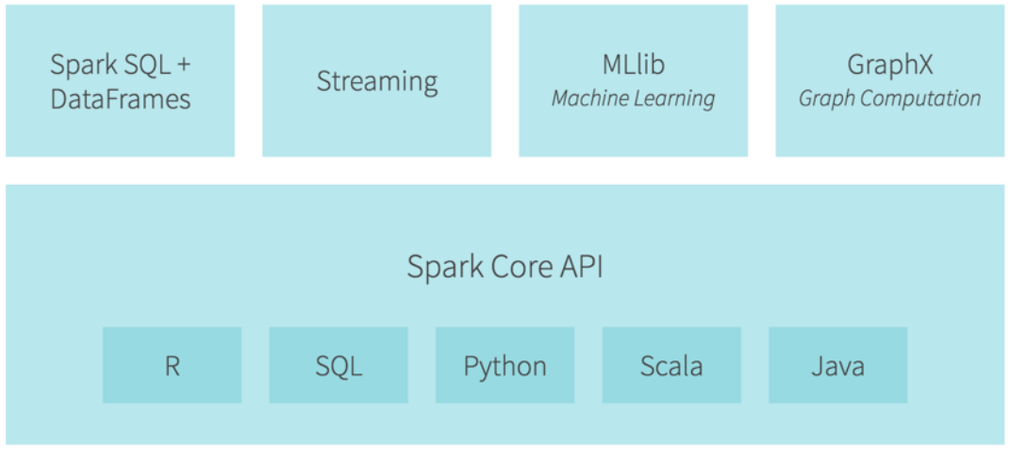
what we data people do all day:

- lots of data collection, curation + storage
- lots and lots of data engineering
- product development with machine learning algorithms!



what is spark?

- "fast and general-purpose cluster computing system"
- advanced cyclic data flow and in-memory computing
  > runs 10x-100x faster than Hadoop MR
- interactive shells in several languages (incl. SQL)
- performant + scalable



courtesy databricks

WARNING: THINGS CHANGE IN SPARK ALL THE TIME. SOME THINGS MIGHT BE HIDDEN, NO LONGER ACCESSIBLE. LIKE SPARK.UNICORNS()

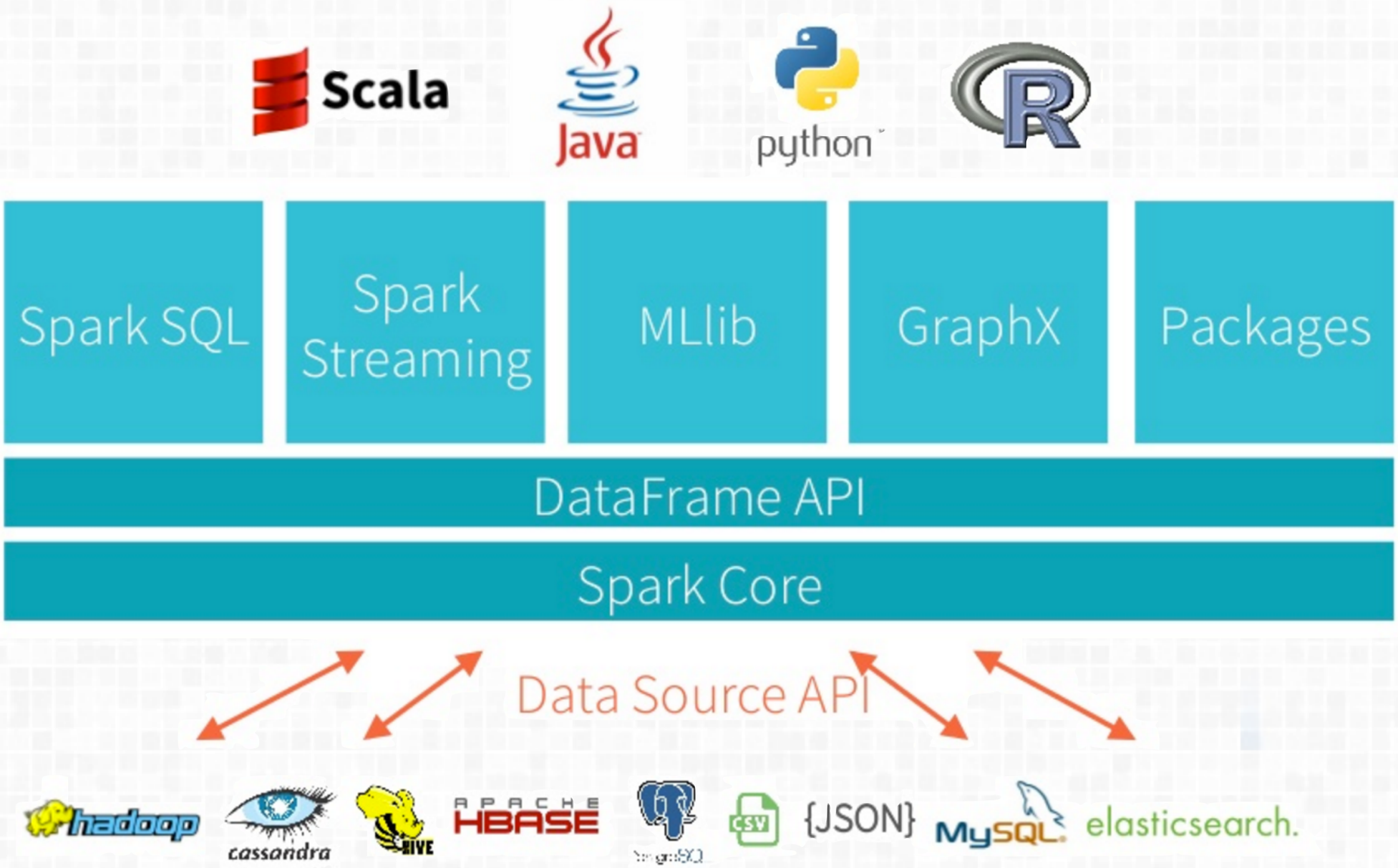
what is spark?

n.b.> it will not solve *every* problem for *everyone*

- not an all-in-one cluster management + admin tool. utilizes other resource managers (YARN, Mesos, Amazon EC2)
- quickly changing updates (major release every 3 months) sometimes requires additional work for backwards compatibility
- for small and medium sized data: not necessary for performant analysis, data science + ML apps
- learning curve is broad for designing cluster applications @ scale

what is spark?

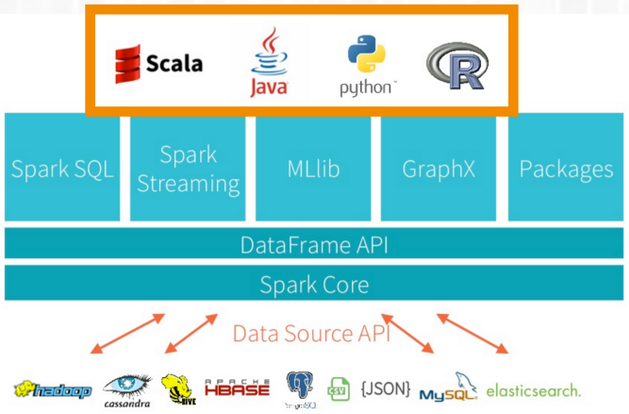
Spark Overview: *Spark Components*



courtesy databricks

what is spark?

- multi-language APIs give many different users the ability to work with Spark
- gateway into Spark but you must still run Spark!
- current languages supported (with various levels of depth): Scala, Python, Java, R
- moving beyond the shell + text edit



courtesy databricks

what is spark?

- how can we continue to approach every data science product with scale + performance as top priority?

MLlib + GraphX on Apache Spark

- Classification and regression
  - linear models (SVMs, logistic regression, linear regression)
  - naive Bayes
  - decision trees
  - ensembles of trees (Random Forests and Gradient-Boosted Trees)
  - Isotonic regression
- Collaborative filtering
  - alternating least squares (ALS)
- Clustering
  - k-means
  - Gaussian mixture
  - Power iteration clustering (PIC)
  - Latent Dirichlet allocation (LDA)
  - streaming k-means
- Dimensionality reduction
  - singular value decomposition (SVD)
  - principal component analysis (PCA)
- Optimization (developer)
  - stochastic gradient descent
  - limited-memory BFGS (L-BFGS)
- Graph analytics

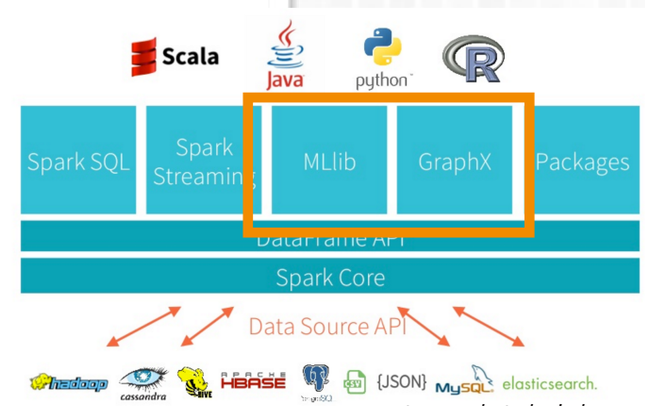
v1.3 -> v1.6

spark.mllib: data types, algorithms, and utilities

- Data types
- Basic statistics
  - summary statistics
  - correlations
  - stratified sampling
  - hypothesis testing
  - streaming significance testing
  - random data generation
- Classification and regression
  - linear models (SVMs, logistic regression, linear regression)
  - naive Bayes
  - decision trees
  - ensembles of trees (Random Forests and Gradient-Boosted Trees)
  - isotonic regression
- Collaborative filtering
  - alternating least squares (ALS)
- Clustering
  - k-means
  - Gaussian mixture
  - power iteration clustering (PIC)
  - latent Dirichlet allocation (LDA)
  - bisecting k-means
  - streaming k-means
- Dimensionality reduction
  - singular value decomposition (SVD)
  - principal component analysis (PCA)
- Feature extraction and transformation
- Frequent pattern mining
  - FP-growth
  - association rules
  - PrefixSpan
- Evaluation metrics
- PMML model export
- Optimization (developer)
  - stochastic gradient descent
  - limited-memory BFGS (L-BFGS)

spark.ml: high-level APIs for ML

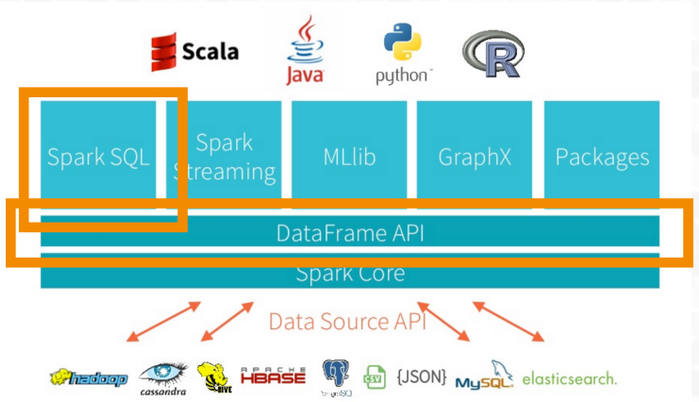
- Overview: estimators, transformers and pipelines
- Extracting, transforming and selecting features
- Classification and regression
- Clustering
- Advanced topics



courtesy databricks

what is spark?

- Spark SQL allows you to query structured data in Spark programs either using SQL or DataFrames API
- can be used in applications + iterative workflows from a shell or notebook
- DataFrames API conceptually similar to a table in a relational database or data frame in R/Python
- preserves schema of original data for many file formats, including Parquet
- highly optimized, distributed collection of data
- Datasets: experimental interface (Scala + Java)



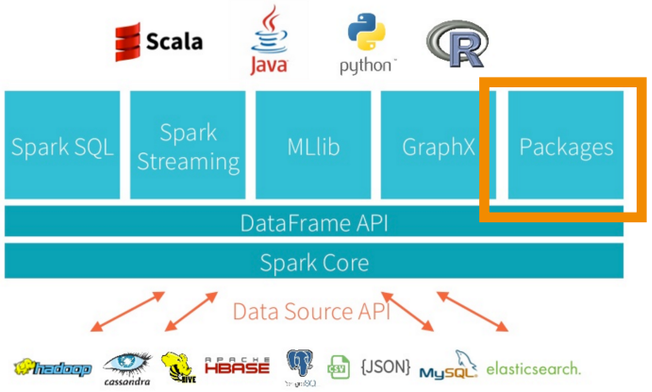
courtesy databricks

what is spark?

- spark-packages is a hosted module resource center for packages developed by the Spark community

- extends functionality + integration options for current Spark releases

- examples: spark-csv, spark-testing-base



courtesy databricks

you are not alone…

courtesy xkcd