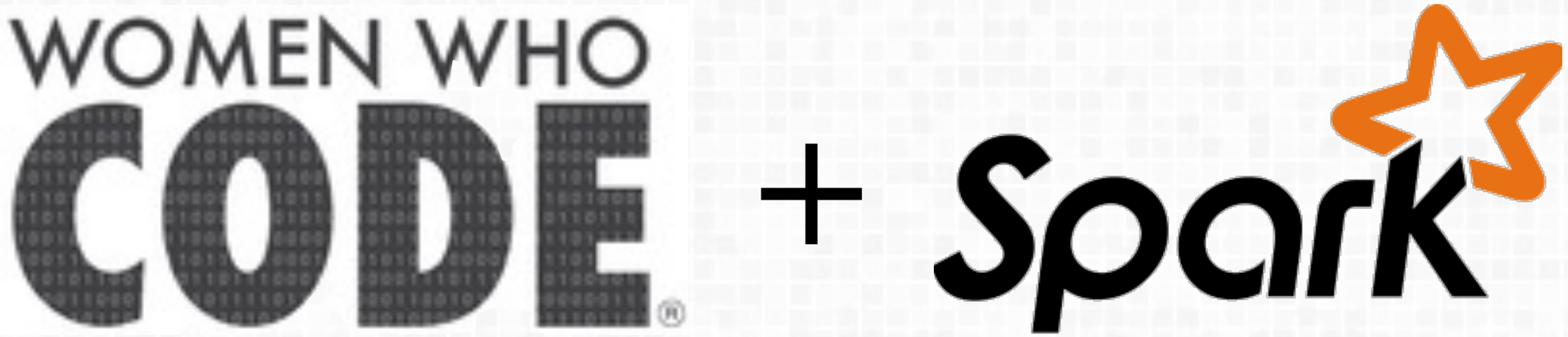
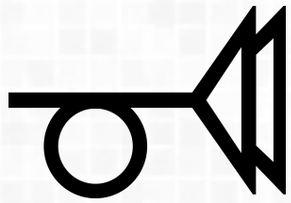


apache spark for everyone

amcasari + deb siegel

WWConnect 2016, Seattle

who: @amcasari  @ CONCUR

@dsiegel  @ W

2016 CONNECT by WWCODE

what: #WWConnect2016

where: @galvanizeSEA

why: @ApacheSpark

(now we can be found)

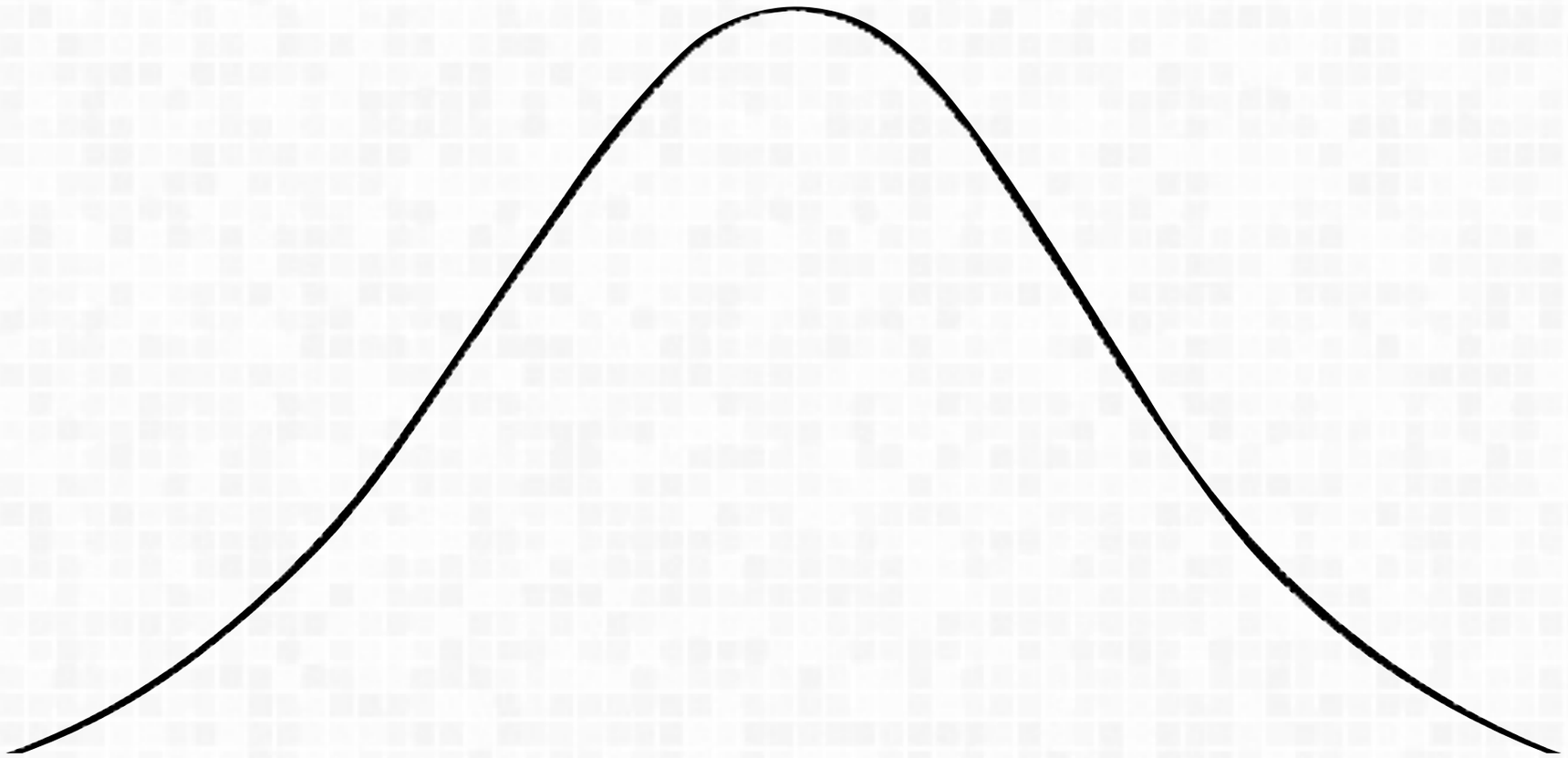
.....

COORDINATES

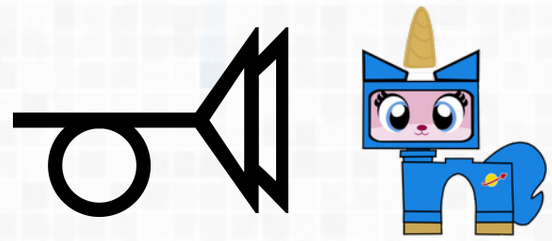
galvanize

Spark

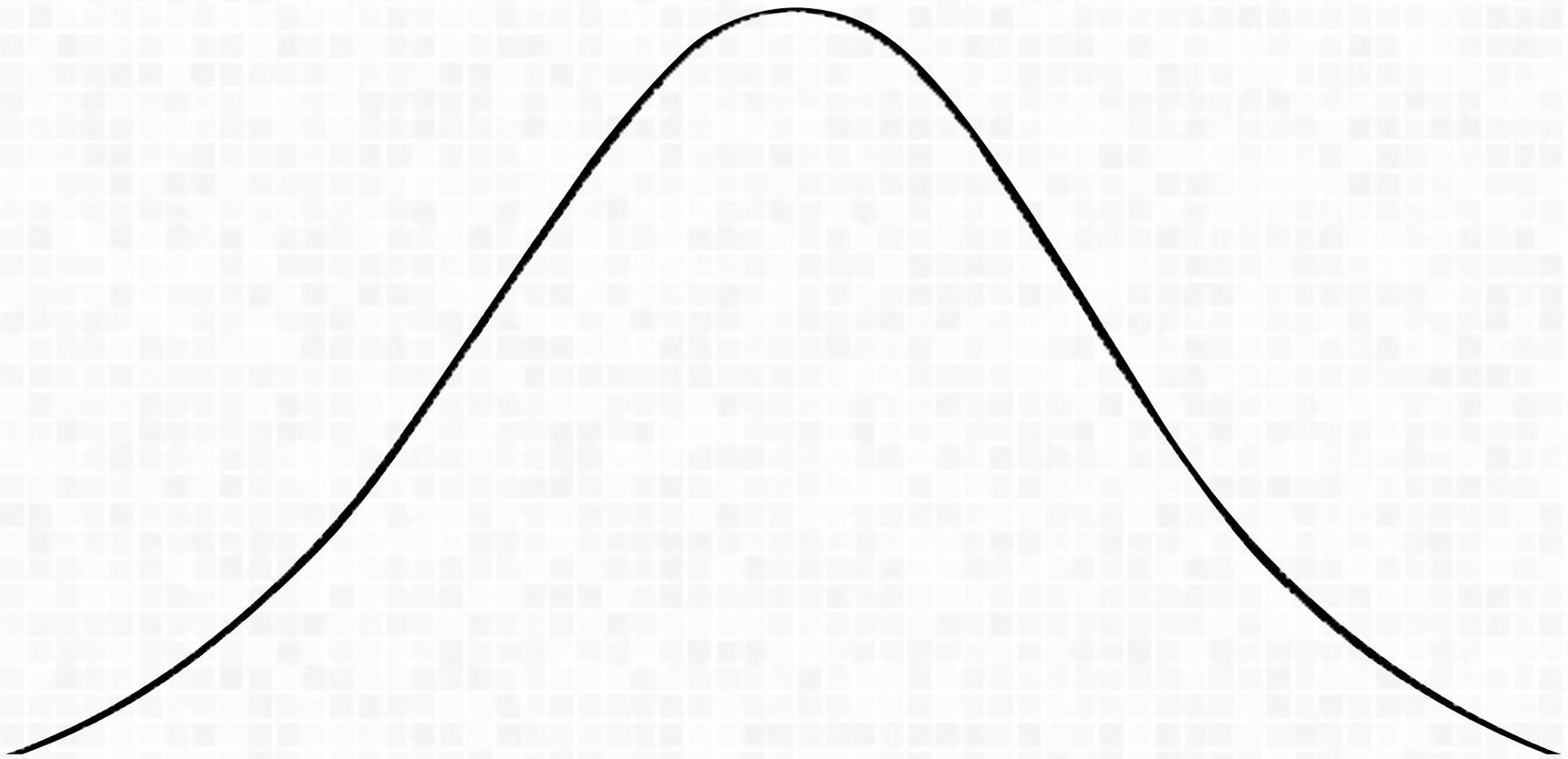
don’t worry about this….

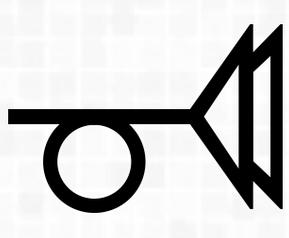


you



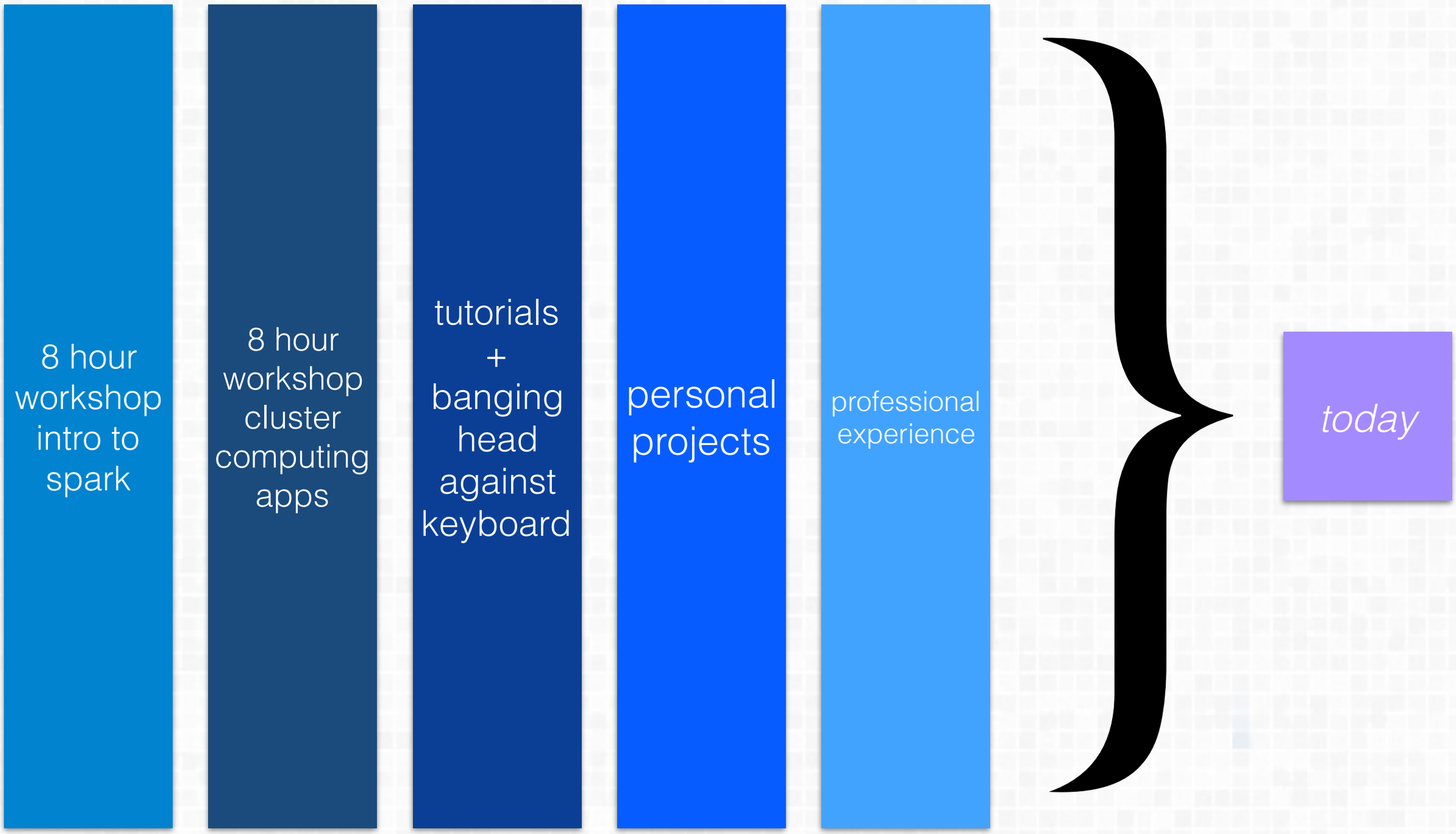
because we feel the same way…
we are all learning!





you

we might be a wee bit ambitious…



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

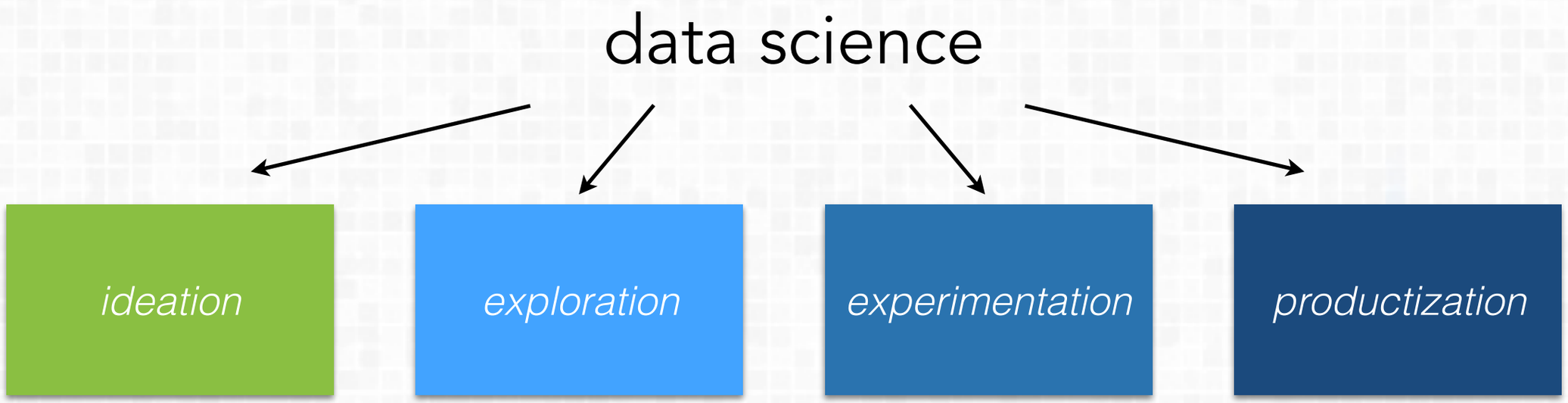
{now you are safe take a nap….}

courtesy YouTube

why do we care about spark?

what we data people do all day:

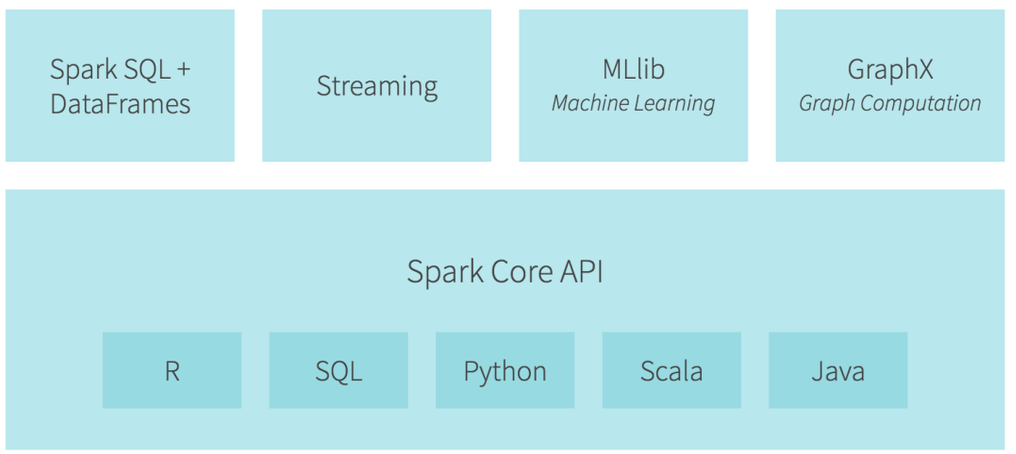
- lots of data collection, curation + storage
- lots and lots of data engineering
- product development with machine learning algorithms!



WOMEN WHO CODE + Spark

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

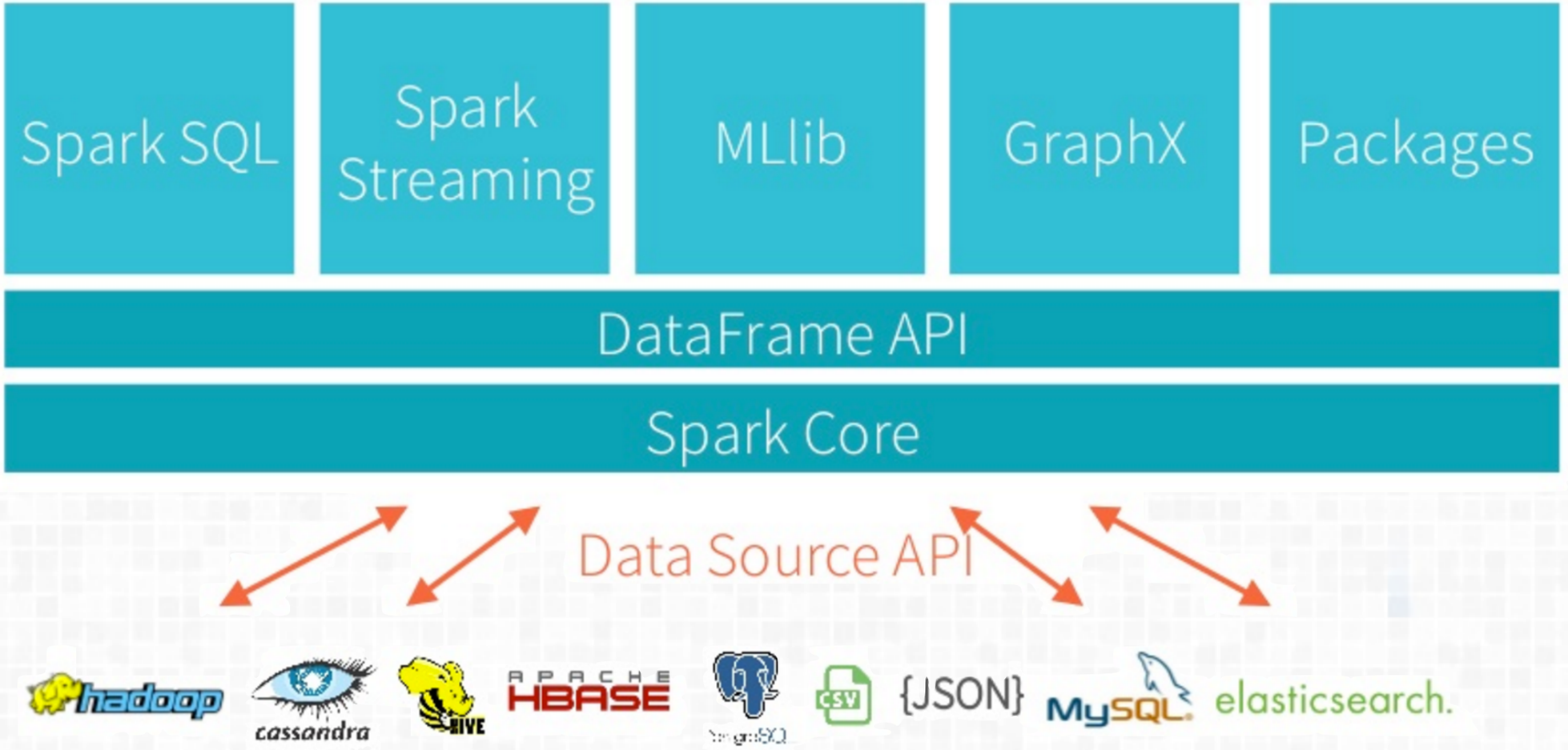
WARNING: THINGS CHANGE IN SPARK ALL THE TIME. SOME THINGS MIGHT BE HIDDEN, NO LONGER ACCESSIBLE.
LIKE SPARK.UNICORNS()

WOMEN WHO CODE + Spark

what is spark?

Spark Overview: *Spark Components*

Scala Java python R



courtesy databricks

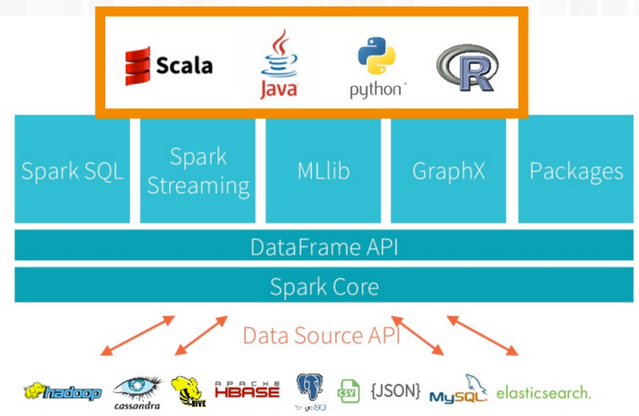
what is spark?

- 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

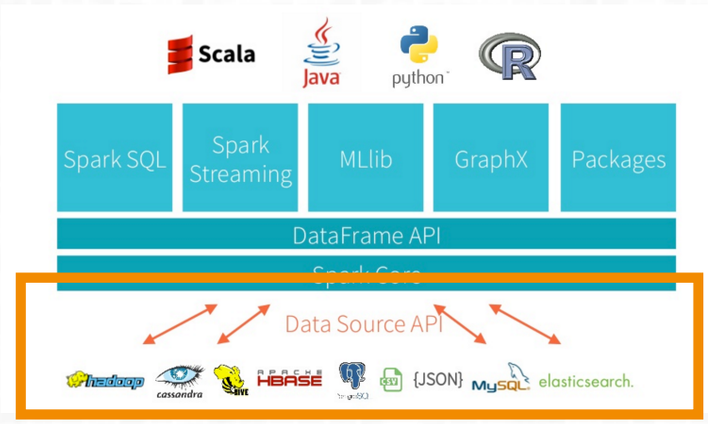
what is spark?

- Data Sources API provides a "pluggable mechanism for accessing structured data through Spark SQL"

- changes the question from "where to store the data" to "how can we access + work w/ the data"

- not exposed to users, for Spark devs

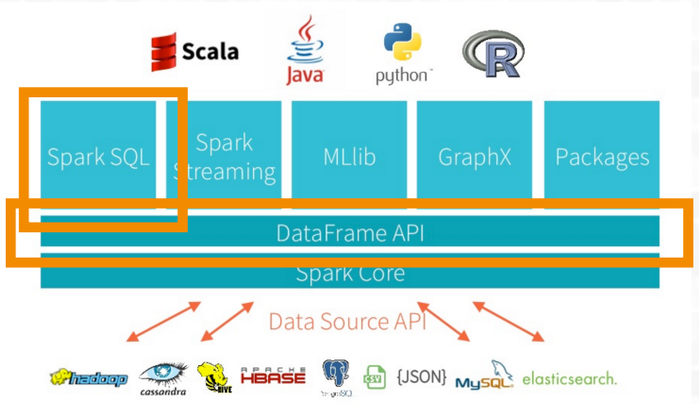
- supplemented by spark-packages



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

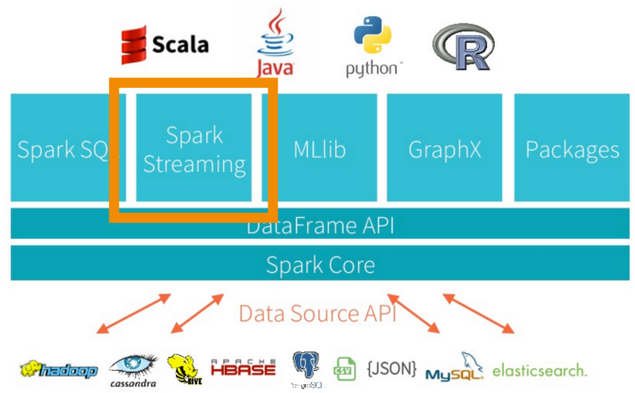
what is spark?

- Spark Streaming allows for discretized event-stream processing

- why would we be excited about it?

- single platform/analysis pipeline for batch + real-time analysis

- on-line learning for ML applications



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

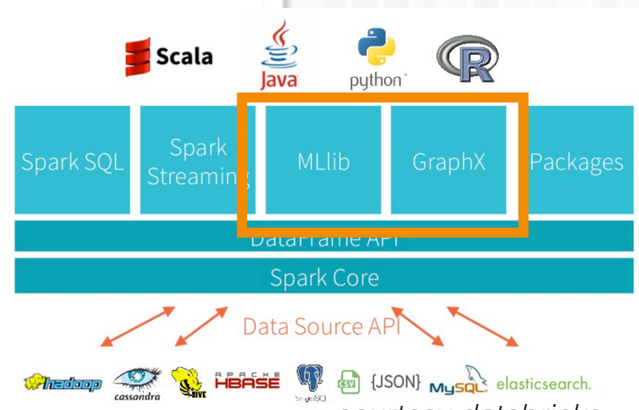
v1.3 -> v1.6

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



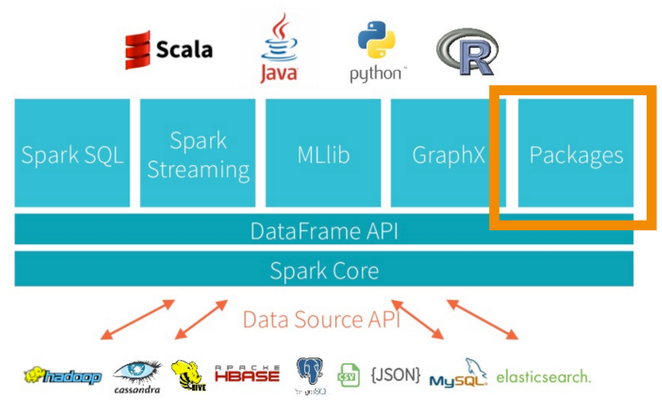
courtesy databricks

what is spark?

- spark-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

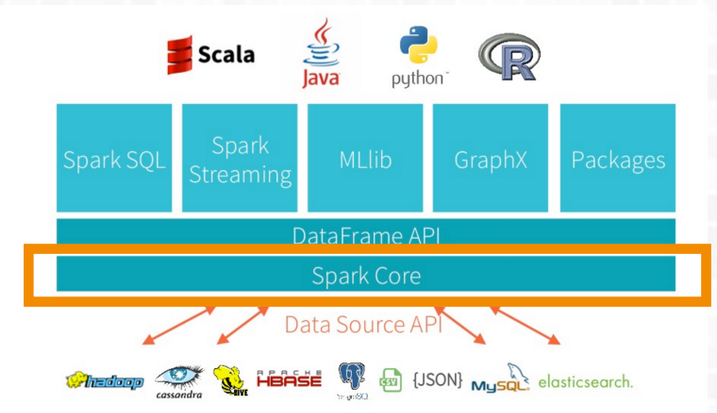
what is spark?

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

how does spark work?

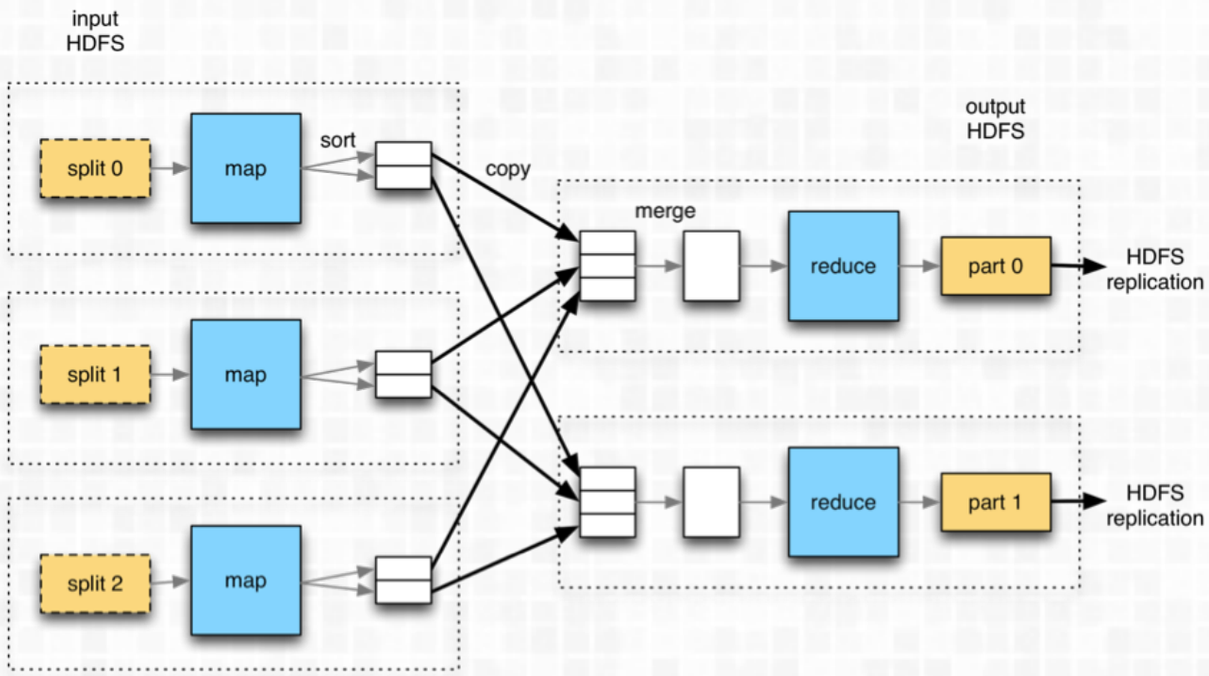
- basic abstraction: Resilient Distributed Dataset (RDD)
- items distributed across many compute nodes that can be manipulated in parallel
- we primarily use core functionality to quickly build applications for production, including data processing for the DataScience Web API
- UDFs expand core functionality



courtesy databricks

how does spark work?

spark != mapreduce



Hadoop MR

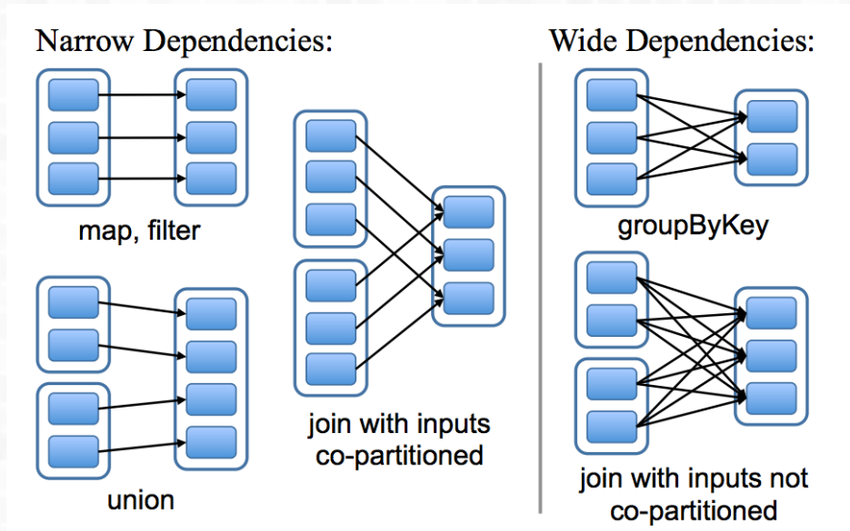
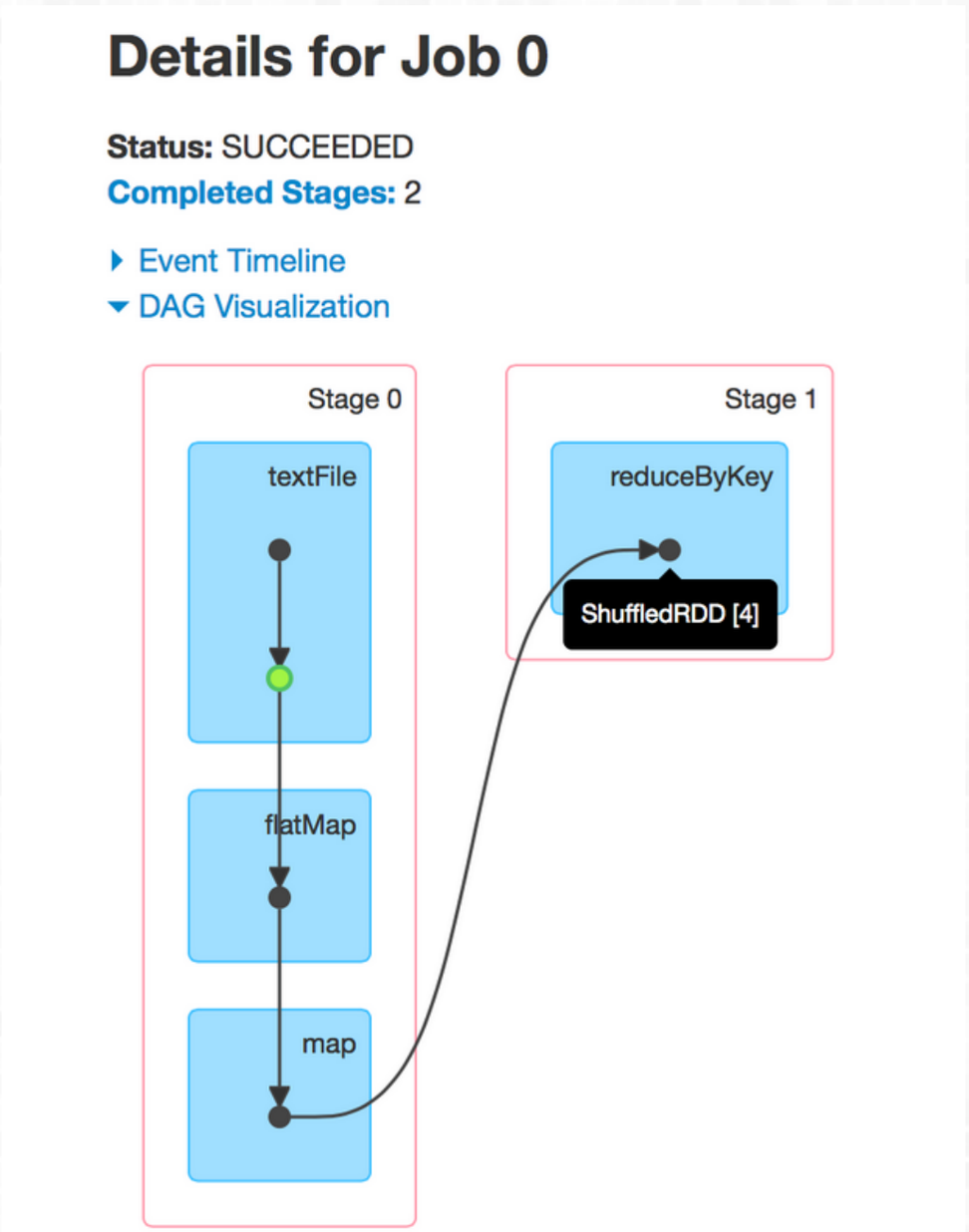


Figure 4: Examples of narrow and wide dependencies. Each box is an RDD, with partitions shown as shaded rectangles.

Spark

read more here from @pacoid

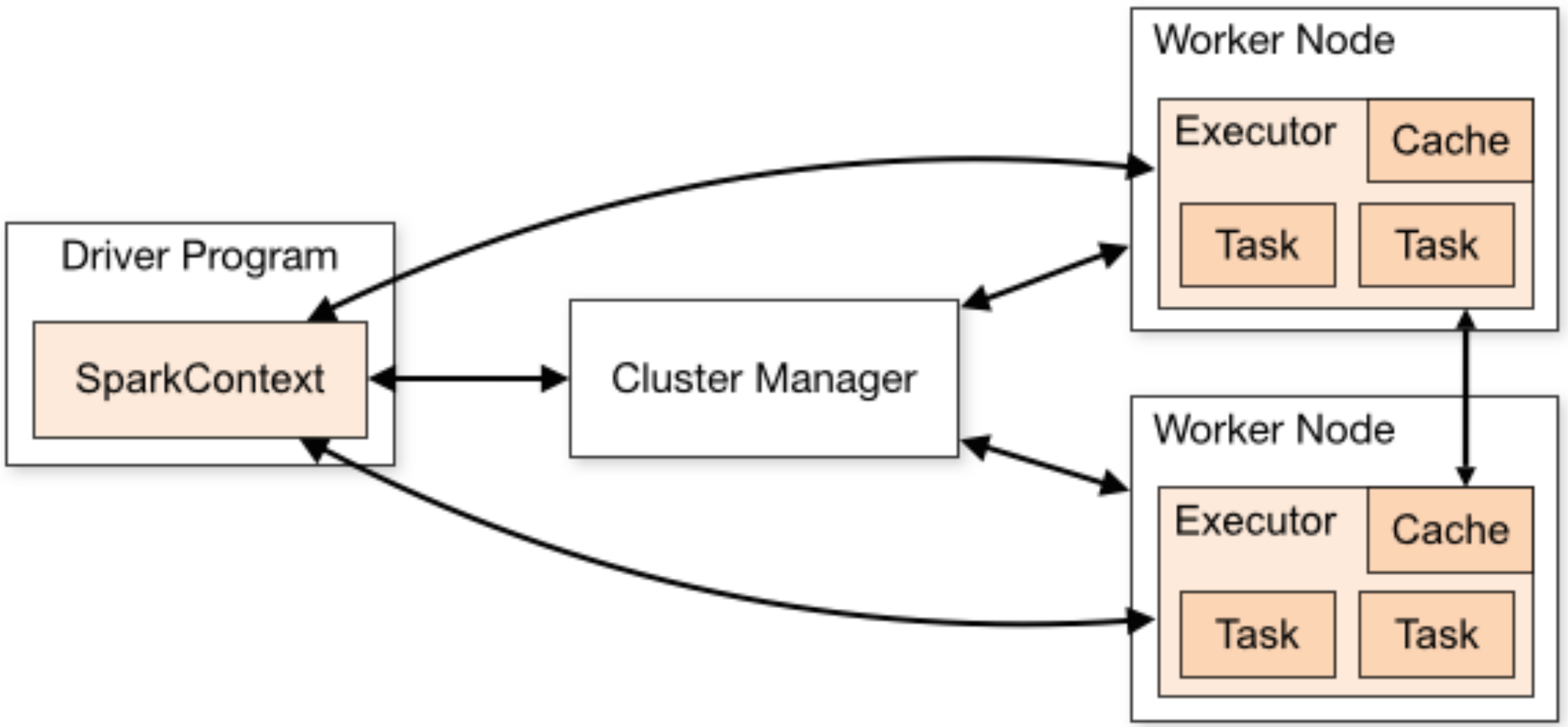
how does spark work?



transformation, actions, laziness, DAGs

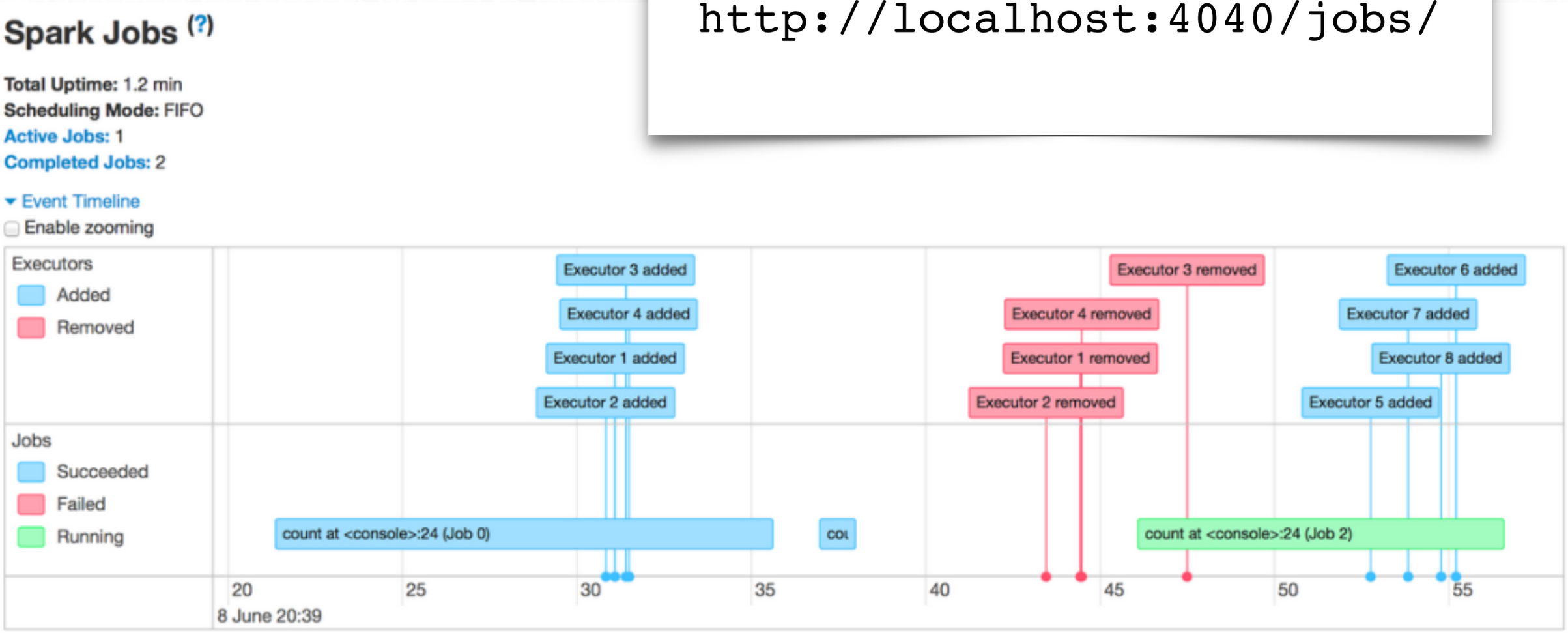
how does spark work?

- fault-tolerant cluster computing framework for most clusters



how does spark work on a cluster?

`http://localhost:4040/jobs/`



courtesy databricks

tasks, jobs, stages, oh my!

spark + notebooks (today)

- Scala via Databricks Community Edition
- Python via Jupyter
- R via RStudio
- ~~Java via~~ *well, we can't fit in everything….*

Jupyter	R, Scala, Python, Java
Zepplin	Scala, Python, Java, SQL
RStudio	R
Databricks	Scala, Python, Java, SQL, R

WOMEN WHO CODE + Spark

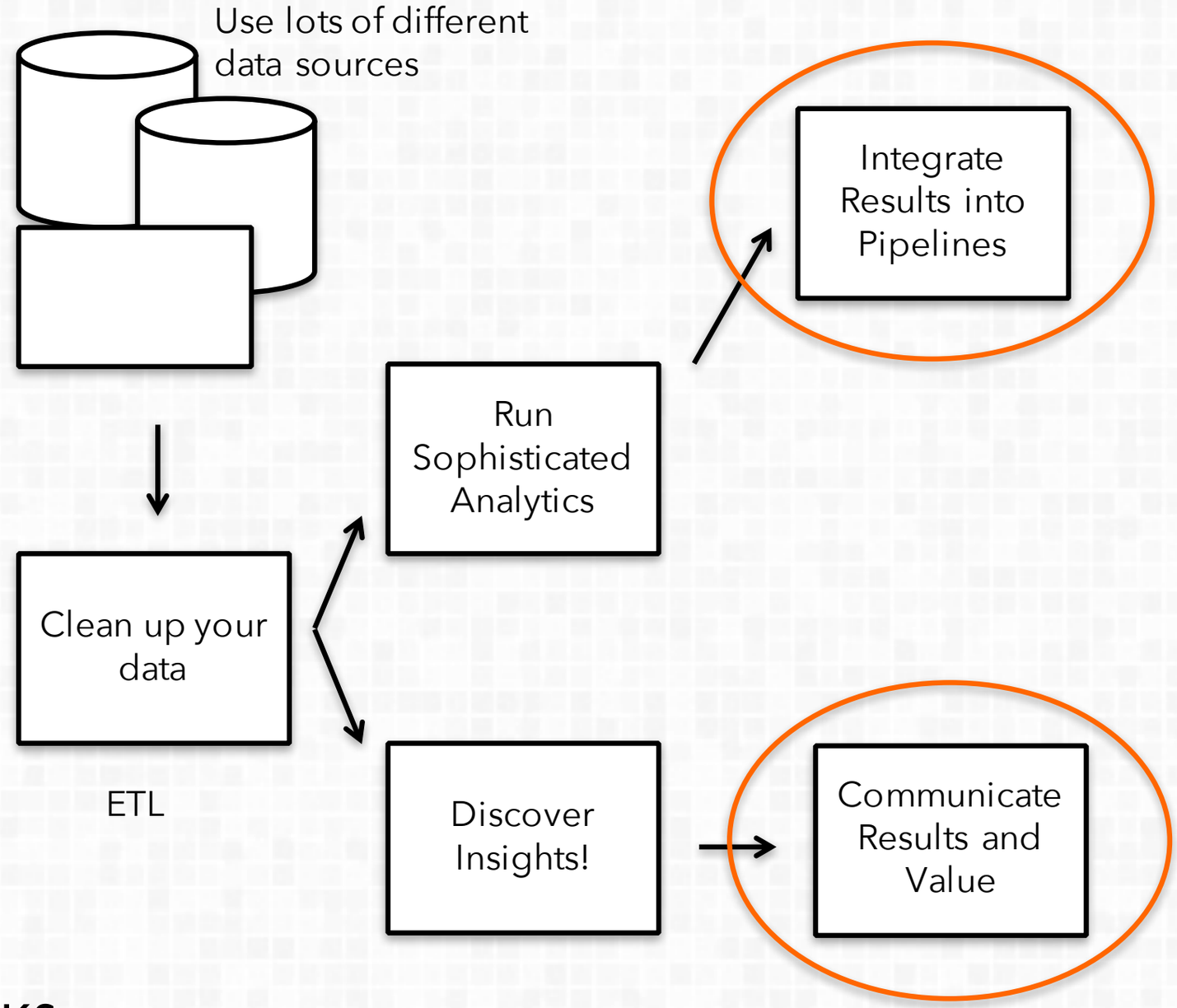
why notebooks?

problem formulation -> tool chain construction

Moving down the data pipeline from raw to results

How best to quickly move through pipeline to:

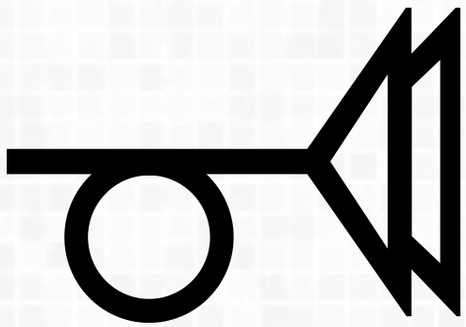
1. Show value of work
2. Communicate results
3. Move models into production pipeline



…NOTEBOOKS

scala + databricks community edition

deb siegel



@ W

WOMEN WHO CODE + Spark

scala for spark on two slides!

- Spark is written in scala - no extra language API libraries or wrappers needed

- Runs on the Java Virtual Machine. Interoperates with JAVA and is converted to JAVA.

- Can use in REPL/shell and Notebook (ie. Jupyter, Zeppelin, DBCE).

scala for spark on two slides!

- Integrated Object Oriented and Functional Programming
  - Data is generally immutable
  - Great for spark programming model!
  - Anonymous functions

```
def yummify(x: String): String = x + " & chocolate"
val yummyRDD = plainRDD.map(x => yummify(x))
```

- Statically Typed
  - Types checked at compile-time
  - Type Inference
- Implicit Type Conversion
  - If a method is not found for your object, it will be converted to an object which does have that method.
- Case Class

  Serializable object with override methods such as equals(), hash() and toString().

scala + databricks community edition demo

We have some DBCE accounts to give away.
Most of the code can otherwise be used on Apache Zeppelin.

pyspark + python + jupyter

amanda casari

@

what is python?

- google: “python is a high-level general-purpose programming language”

- python.org: “Python is a programming language that lets you work more quickly and integrate your systems more effectively.”

- most important: python is open source!

- python is friendly for beginners: programmers + non-yet-programmers

what is pyspark?

- pyspark programming API: you can talk to spark using python

- spark python API docs fairly well maintained

- you can work on pyspark objects using pyspark or on python objects using python

  - some pyspark functions return python objects

- very popular, well developed on + optimized for cluster performance

WOMEN WHO CODE + Spark

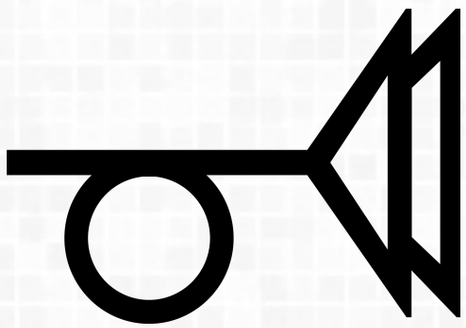
pyspark + jupyter

- easy python distribution: anaconda
  - installs + manages python packages + soo much more
- creating a PySpark kernel for Jupyter notebook
- running spark from any directory: create a symlink
- starting pySpark + Jupyter from command line:

```
$ IPYTHON_OPTS="notebook" pyspark
```

pyspark + python + jupyter demo

sparkR + RStudio

deb siegel  @ W

what is R?

- R is a language & environment for statistical computing and graphics

- Many specialized packages available in CRAN (Comprehensive R Archive Network)

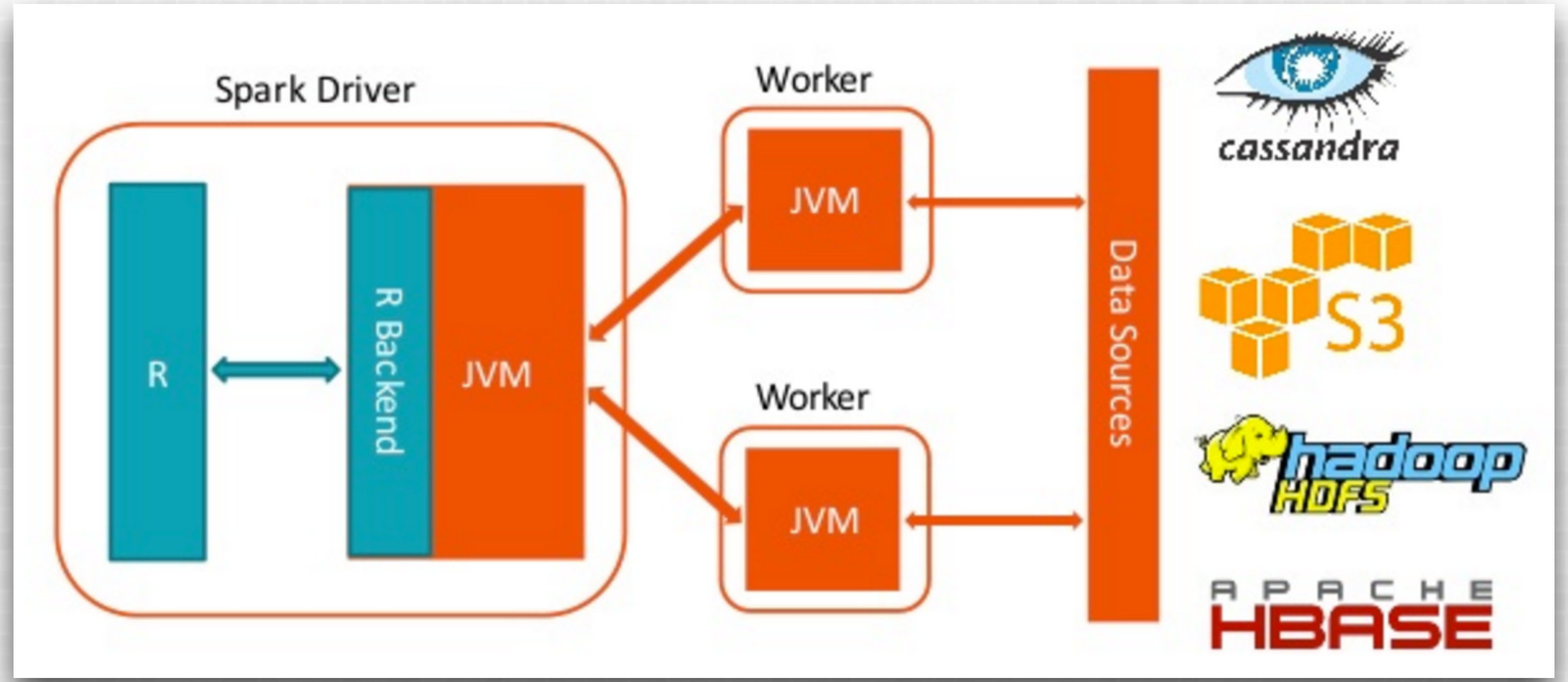
- RStudio is an open source R notebook / IDE

what is sparkR?

- Spark DataFrame is the main data structure. (You can still use R data.frame but it is not Sparky)
  - Translated to RDD[Row] behind the scenes. Columns have names and data type.
- All local R functions and libraries available (locally on the driver only for now).
- Stage of maturity = evolving: Use for filtering, selecting, and aggregating your BIG data down to small or medium data which you can work with locally.
- Some machine learning algorithms and stats are available on the BIG data, more will be added as the SparkR project grows.
- Can work in REPL/shell or Notebook (eg. RStudio, Jupyter, DBCE, Zeppelin).

WOMEN WHO CODE + Spark

what is sparkR?



courtesy databricks

sparkR + RStudio

```
> Sys.setenv(SPARK_HOME="/yourpathto/spark")

> .libPaths(c(file.path(Sys.getenv("SPARK_HOME")
, "R", "lib"), .libPaths()))

> library(SparkR)

> sc <-
sparkR.init(master="local[2]",appName="SparkR-
for-
everyone",sparkPackages="com.databricks:spark-
csv_2.11:1.2.0")

>sqlContext <- sparkRSQL.init(sc)
```

sparkR + RStudio

demo

now your turn…workshoppy bit!

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

Our possibly overly ambitious goals:

- Get you set up w/ Spark + a notebook
- Show you where to find help
- Get you started on a notebook

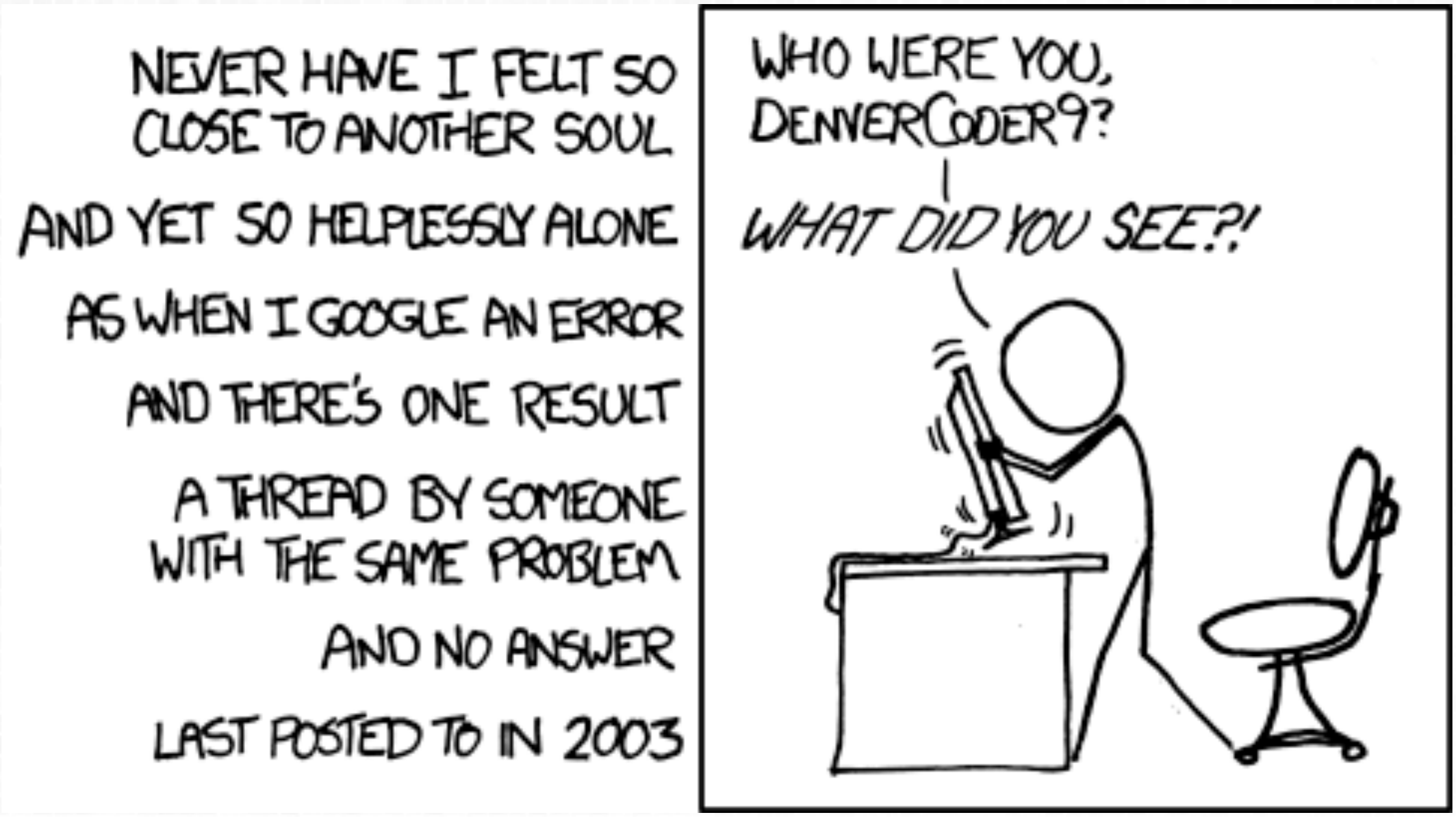
/all the spark for windows

/all the spark for mac os x

~~/all the spark for linux~~

Spark runs on Java 7+, Python 2.6+ and R 3.1+, Scala 2.10.x

you are not alone…



courtesy xkcd