The Big Picture of Data Science Approaches and Tools

AMARNATH GUPTA

Three Definitions

Data Intensive Computing

a class of parallel computing applications which use a data parallel approach to processing large volumes of data

Big Data

data sets so large, diverse, complex or fast that traditional ingestion strategies, data processing and computing algorithms, and system infrastructure are inadequate

Data Science

almost everything that has something to do with data: collecting, analyzing, modeling to model, predict and infer information for domain applications

International Statistical Review (2001), 69, 1, 21–26, Printed in ...

Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA
E-mail: wsc@research.bell-labs.com

Summary

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department, and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Key words: Future; Applications; Computing; Methods; Models; Theory.

1 Summary of the Plan

This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science".

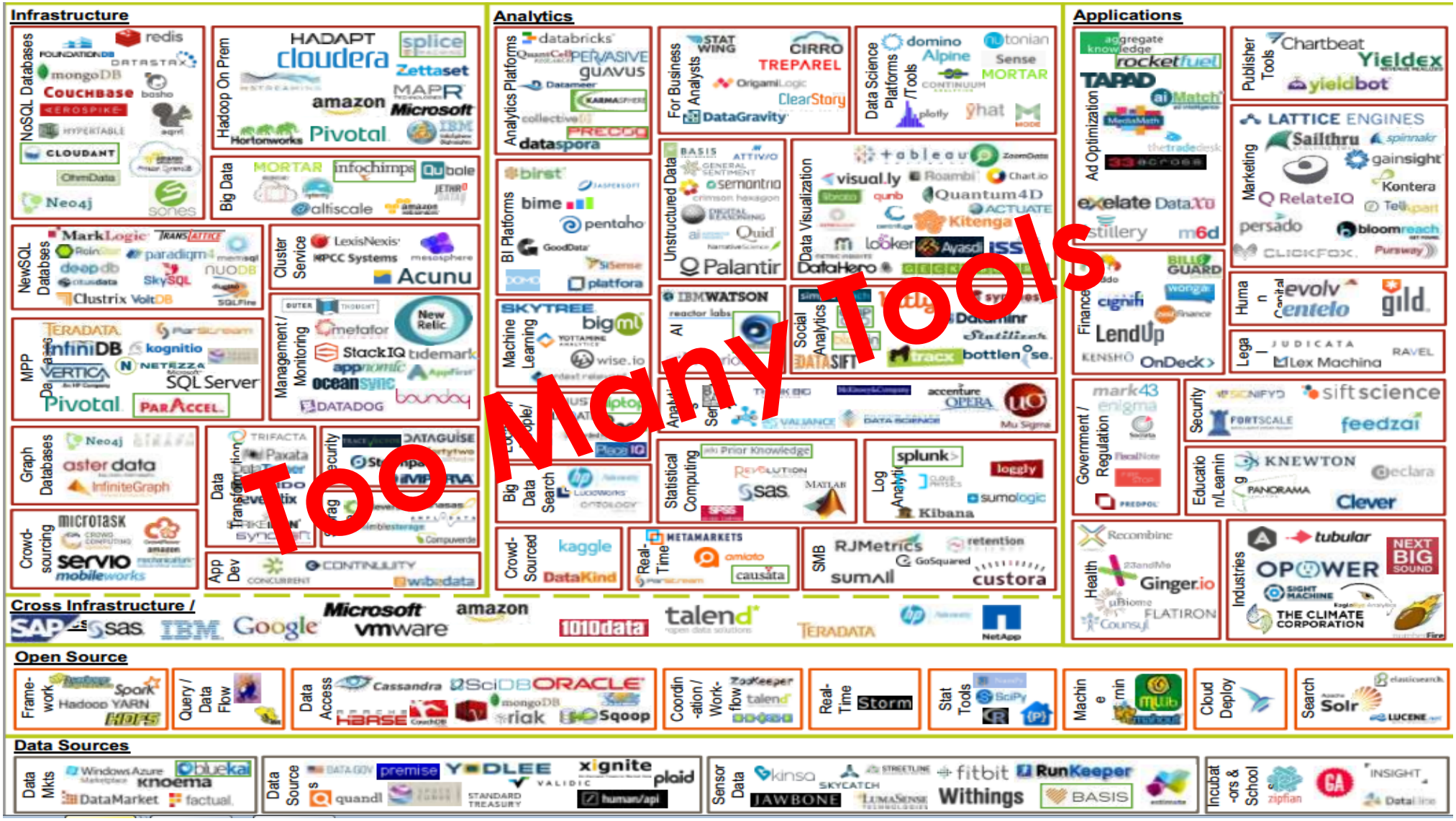
The focus of the plan is the practicing data analyst. A basic premise is that technical areas of data science should be judged by the extent to which they enable the analyst to learn from data. The benefit of an area can be direct or indirect. Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit. A broad successful theory can have a wide-ranging benefit, affecting data analysis in a fundamental way. For example, the Bayesian theory of statistics affects all methods of estimation and distribution.

The plan sets out six technical areas for a university department, and advocates a specific allocation of resources to research and development in each area as a percent of the total resources that are

The "official" start of Data Science

(25%) Multidisciplinary Investigations: data analysis collaborations in a collection of subject matter areas.
(20%) Models and Methods for Data: statistical models; methods of model building; methods of estimation and distribution based on probabilistic inference.
(15%) Computing with Data: hardware systems; software systems; computational algorithms.
(15%) Pedagogy: curriculum planning and approaches to teaching for elementary school, secondary school, college, graduate school, continuing education, and corporate training.
(5%) Tool Evaluation: surveys of tools in use in practice, surveys of perceived needs for new tools, and studies of the processes for developing new tools.
(20%) Theory: foundations of data science; general approaches to models and methods, computing with data, teaching, and tool evaluation; mathematical investigations of models and methods, computing with data, teaching, and evaluation.

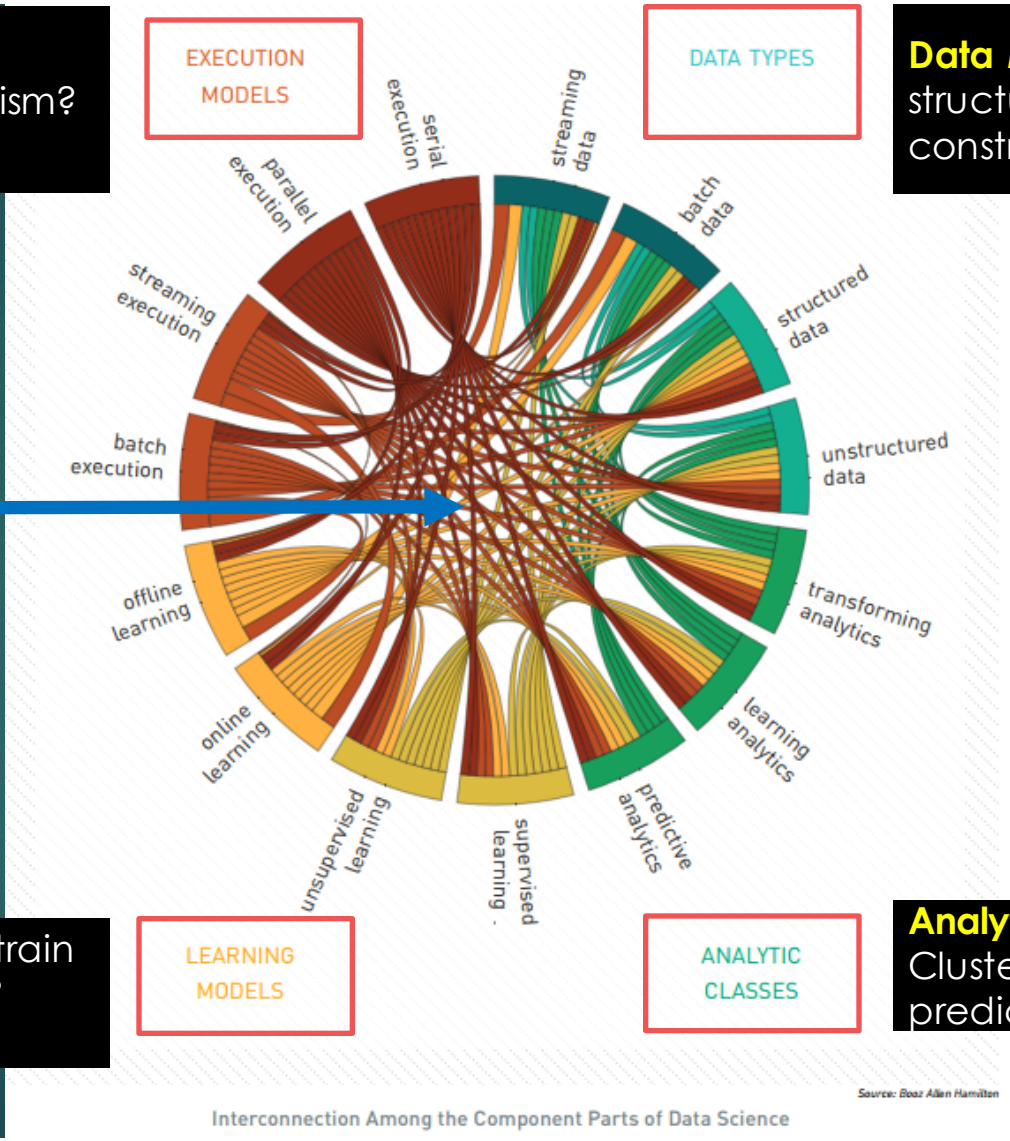
Domain Knowledge + Analytical Methods + Algorithms+ Databases


Infrastructure
NoSQL Databases
redis
FOUNDATIONDB
DATASTAX
mongoDB
Couchbase
basho
AEROSPIKE
HYPERTABLE
aerospike
CLOUDANT
OrientDB
Neo4j
sones
Hadoop On Prem
HADAPT
cloudera
splice machine
Zettaset
amazon
MAPR
Microsoft
Hortonworks
Pivotal
IBM
Big Data
MORTAR
infochimps
Qubole
JETHRO
altiscale
amazon web services
NewSQL Databases
MarkLogic
TRANSLATTICE
RAIN
paradigm4
memsql
deepdb
nuoDB
citusdata
SkySQL
Clustrix
VoltDB
SQLFire
Cluster Service
LexisNexis
HPCC Systems
mesosphere
Acunu
MPP Databases
TERADATA
ParStream
InfiniDB
kognitio
VERTICA
NETEZZA
SQL Server
Pivotal
ParAccel
Management / Monitoring
OUTER THOUGHT
metafor
New Relic
StackIQ
tidemark
appnomic
AppFirst
oceansync
DATADOG
boundary
Graph Databases
Neo4j
aster data
InfiniteGraph
Data Transformation
TRIFACTA
Paxata
Security
DATAGUISE
Crowd-sourcing
microtask
CROWD COMPUTING
CrowdFlower
amazon mechanical turk
servio
mobileworks
App Dev
CONCURRENT
CONTINUITY
wibidata
Analytics
Analytics Platforms
databricks
QuantumCell
PERVASIVE
guavus
Datameer
KARMASPHERE
collective
PRECOG
dataspora
For Business Analysts
STAT WING
CIRRO
TREPAREL
OrigamiLogic
ClearStory
DataGravity
Data Science Platforms /Tools
domino
tonian
Alpine
Sense
MORTAR
CONTINUUM
plotly
yhat
BI Platforms
birst
JASPERSOFT
bime
pentaho
GoodData
SiSense
platfora
Unstructured Data
BASIS
ATTIVIO
GENERAL SENTIMENT
semantria
DIGITAL REASONING
ai
Quid
Palantir
Data Visualization
tableau
ZoomData
visual.ly
Roambi
Chart.io
qunb
Quantum4D
ACTUATE
Kitenga
looker
Ayasdi
ISS
DataHero
Machine Learning
SKYTREE
bigml
YOTTAMINE ANALYTICS
wise.io
IBM WATSON
reactor labs
AI
Social Analytics
Datasift
trackur
bottlenose
Statistical Computing
Prior Knowledge
REVOLUTION
SAS
MATLAB
SPSS
Log Analytics
splunk
loggly
sumologic
Kibana
Big Data Search
Lucidworks
ONTOLOGY
Crowd-Sourced
kaggle
DataKind
Real-Time
METAMARKETS
metamarkets
amiato
causata
SMB
RJMetrics
GoSquared
retention
sumall
custora
Cross Infrastructure /
SAP
sas
IBM
Google
Microsoft
vmware
amazon
1010data
talend
TERADATA
NetApp
Too Many Tools
Applications
Ad Optimization
aggregate knowledge
rocketfuel
TAPAD
aiMatch
MediaMath
thetradedesk
33across
exelate
DataXu
stillery
m6d
Finance
BILLGUARD
cignifi
wonga
LendUp
KENSHO
OnDeck
Government / Regulation
mark43
enigma
FiscalNote
PREDPOL
Health
Recombine
23andMe
Ginger.io
uBiome
FLATIRON
Counsyl
Publisher Tools
Chartbeat
Yieldex
yieldbot
Marketing
LATTICE ENGINES
Sailthru
spinnaker
gainsight
RelateIQ
Kontera
persado
bloomreach
CLICKFOX
Pursway
Human Capital
evolv
entelo
gild
Legal
JUDICATA
Lex Machina
RAVEL
Security
NIFTY5
siftscience
FORTSCALE
feedzai
Education/Learning
KNEWTON
declara
PANORAMA
Clever
Industries
tubular
NEXT BIG SOUND
OPOWER
SIGHT MACHINE
THE CLIMATE CORPORATION
Open Source
Framework
Spark
Hadoop YARN
HDFS
Query / Data Flow
Data Access
Cassandra
SciDB
ORACLE
HBase
CouchDB
mongoDB
riak
Sqoop
Coordination / Workflow
ZooKeeper
talend
Real-Time
Storm
Stat Tools
SciPy
R
{P}
Machine
mahout
Cloud Deploy
Search
elasticsearch
Solr
LUCENE
Data Sources
Data Mkts
Windows Azure Marketplace
kaggle
knoema
DataMarket
factual
Data Sources
DATA.GOV
premise
YODLEE
xignite
VALIDIC
plaid
quandl
STANDARD TREASURY
human/api
Sensor Data
kinsa
SKYCATCH
STREETLINE
fitbit
RunKeeper
JAWBONE
LUMASENSE TECHNOLOGIES
Withings
BASIS
estimote
Incubators & School
zipfian
GA
INSIGHT
Dataline

Execution Model: Data parallelism? Task parallelism? Stream processing?

Data Model: specifies the structure, operations and constraints on data

There are tools for each of these combinations



Interconnection Among the Component Parts of Data Science

Partitioning the Problem

Learning Model: How to train data? Model of training? Algorithm of learning?

Analytic Class: Classification? Clustering? Time-series based prediction? Recommendation?

Some Data Science Problems

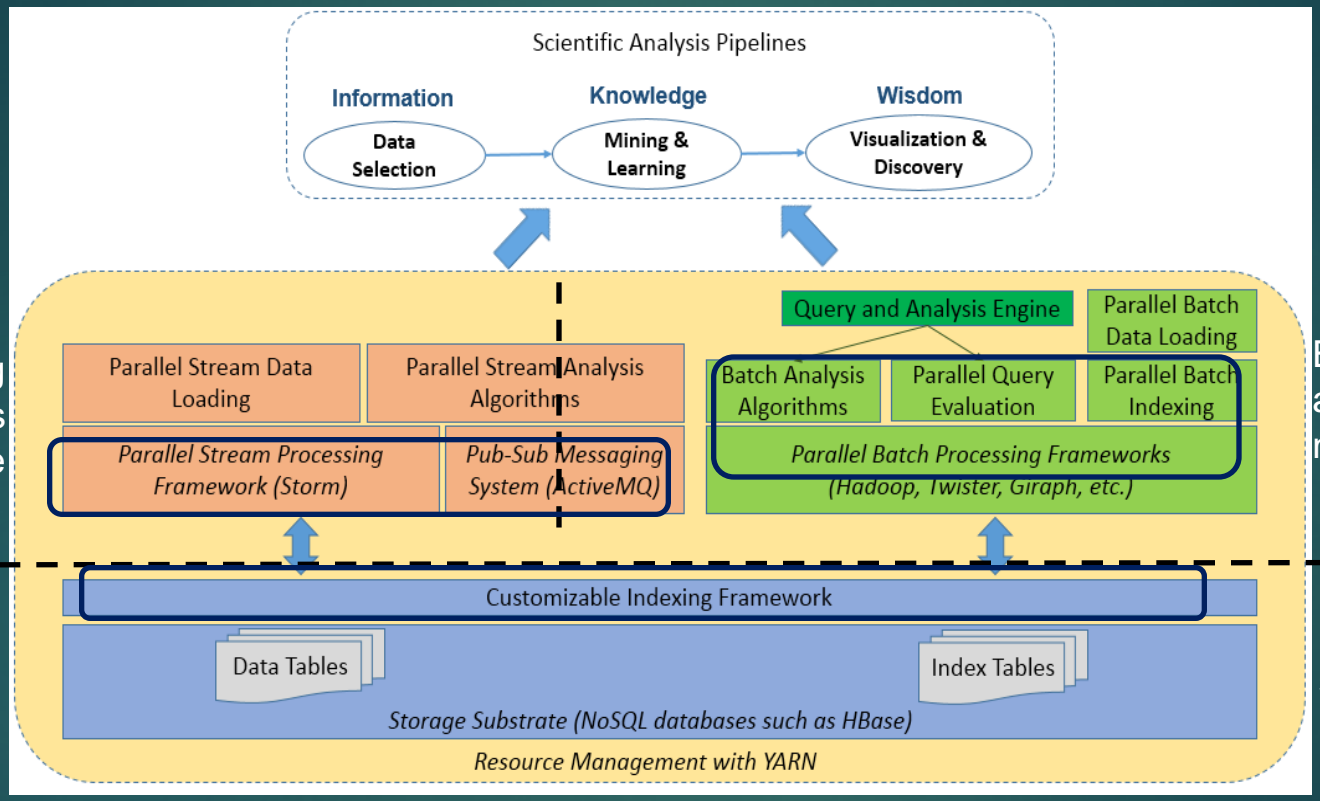
- Looking at public tweets, is it possible to
  - Predict how the political leanings of the tweeting populations are changing?
  - Predict the success or failure of a political campaign

The answer may evolve with time

- Given
  - Historical information
  - Current weather data
  - Current state of wildfires
  - Simulation engines for wind patterns and fire perimeters
  - Can we predict if the fire is likely to reach *this area* in the next 10 minutes?

Has a time-limit of 10 minutes

A General Architecture



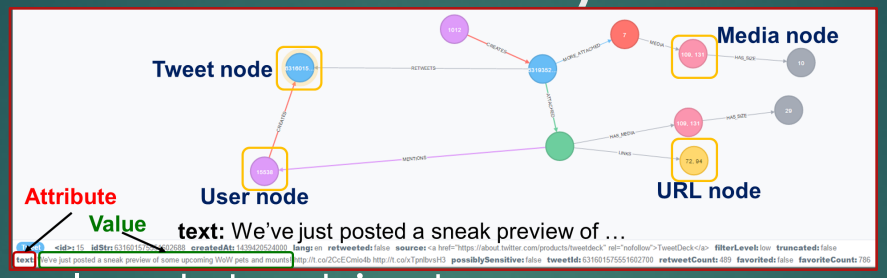
Streaming analysis module

Batch analysis module

Storage substrate

Ideally, must run on a (virtual) cluster of machines to offer horizontal scalability

Data Genres (Data following different data models) - I



- A tweet is a complex data object
  - Data format – JSON — MongoDB?
  - Metadata – relational (i.e., table like) — Any relational DBMS that scales?
  - Tweet body – text — Index with Apache Solr?
  - Inter-tweet connections – network
  - Inter-user connections – network

  Neo4J? Spark/GraphX?
- Number of tweets we collect – 4-8GB/day, >1 TB over 5 months

Any Data Mgmnt System that works well on all data genres?

Not really!!

Analytical Query Processing

- Do we need
  - More custom data processing, or — Hadoop or Spark-based Solutions
  - Store data and perform arbitrary queries, or — Vertica (relational), AsterixDB (JSON, XML), VoltDB (in-memory)…
  - Both at the same time — SPARK-SQL ??
- For our data science example
  - Store in AsterixDB
  - Network queries in Neo4J
  - Compute in SPARK
  - Move data with a message bus (e.g., RabbitMQ, Apache Ignite, …)

Multi-System Solution

Analytics Models

- ~ 200 Million Tweets
- We need
  - Classifiers to identify only the political tweets
    - Train to eliminate ads, personal tweets, spams
      - Size of training set – 20k
  - Topic Model construction over time windows
    - Vocabulary size – 30,000
    - Document count – 170 million
  - Multivariable Regression

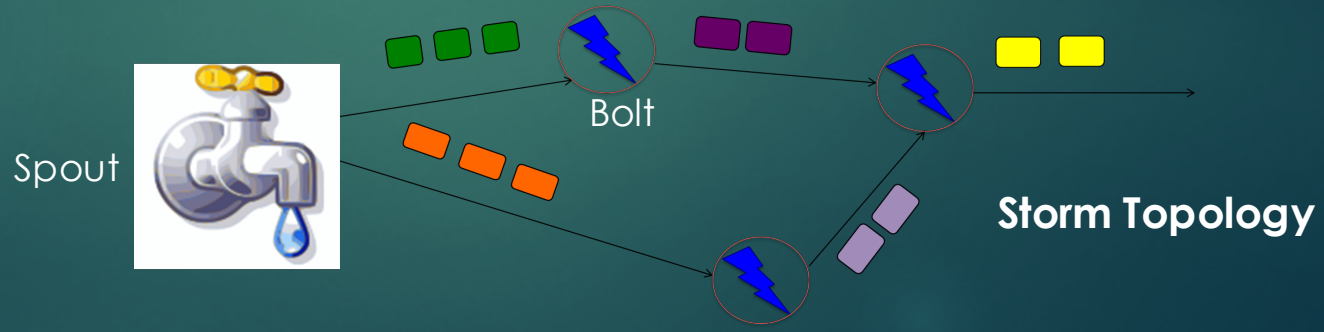
Spark with MLLib running over a virtual cluster of machines

SparkR to use the R statistical analysis system over Spark

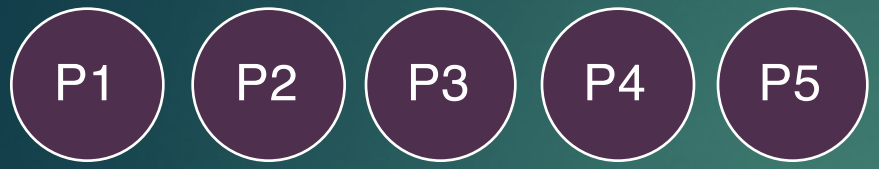
Why Spark?
Iterative Memory-intensive Data Parallel Computation

Execution Model – Stream Processing

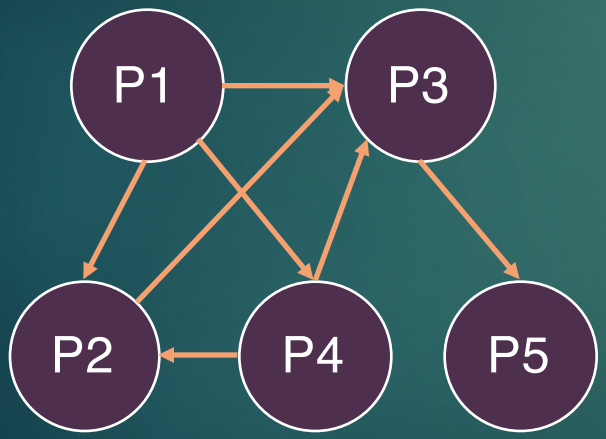
- Apache Storm
  - Operations that can be performed
    - **Filter**: forward only tuples which satisfy a condition
    - **Joins**: When receiving two streams A and B, output all pairs (A,B) which satisfy a condition
    - **Apply/transform**: Modify each tuple according to a function



Execution Model – Network Processing with Bulk Synchronous Parallelism



- Processors
  - Have local memory
  - Can perform some computation



- Processors can communicate pairwise
- Communication can overlap with another node's computation

Barrier Synchronization

Apache Giraph

Spark/GraphX

Conclusion

- No universal big picture solution for large-scale data science problem
- Tool choice based on data types, analytical model, execution model and learning model
- A glimpse of tools to consider within the context of one data science problem