mahout

Java as a fundamental working tool of the Data Scientist

Speaker : Alexey Zinoviev

ТАМТЭК

About

- I am a <graph theory, machine learning, traffic jams prediction, BigData algorythms> scientist
- But I'm a <Java, JavaScript, Android, NoSQL, Hadoop, Spark> programmer

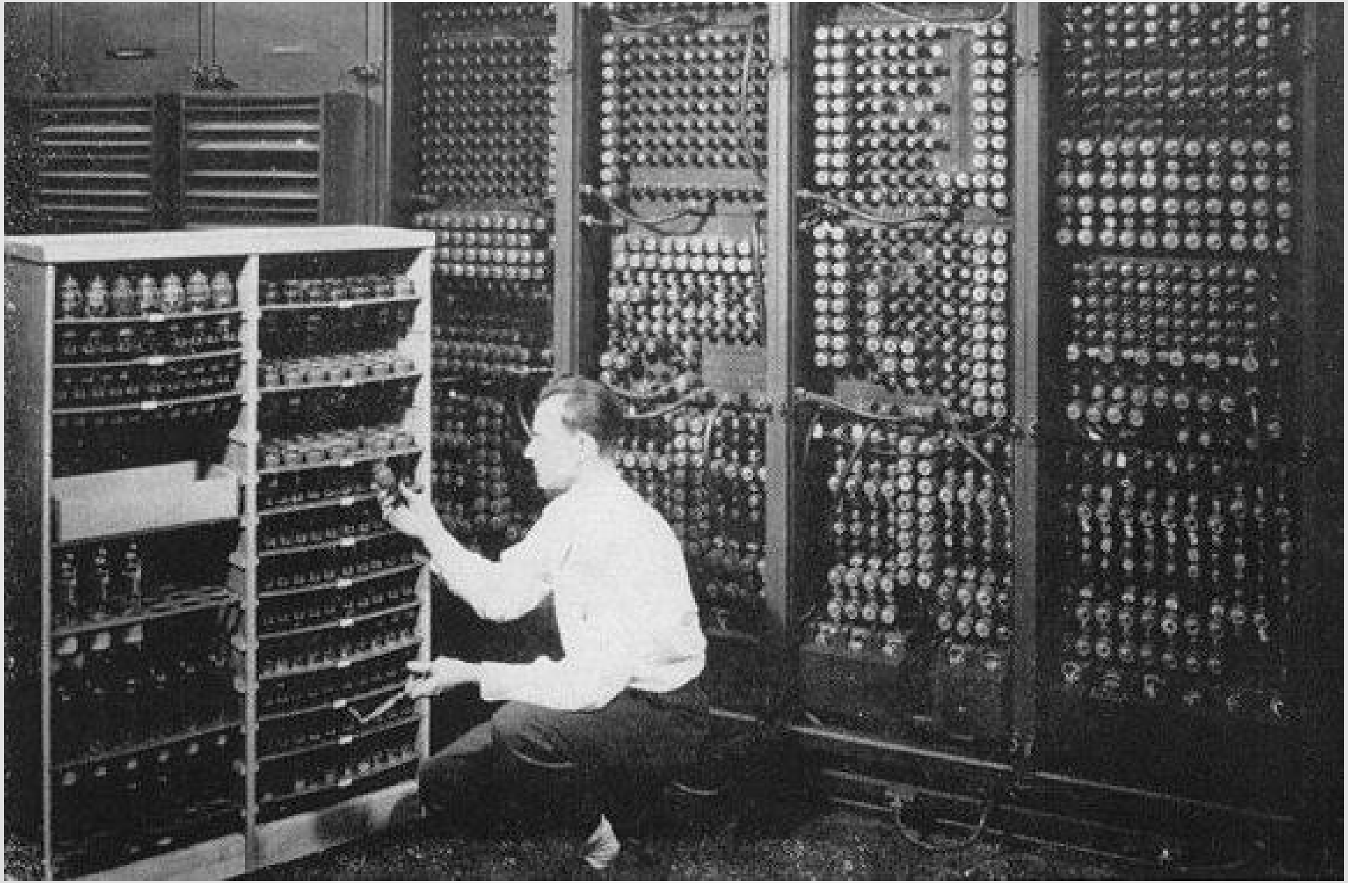
Omsk 2014

Java Days

One of these fine days…

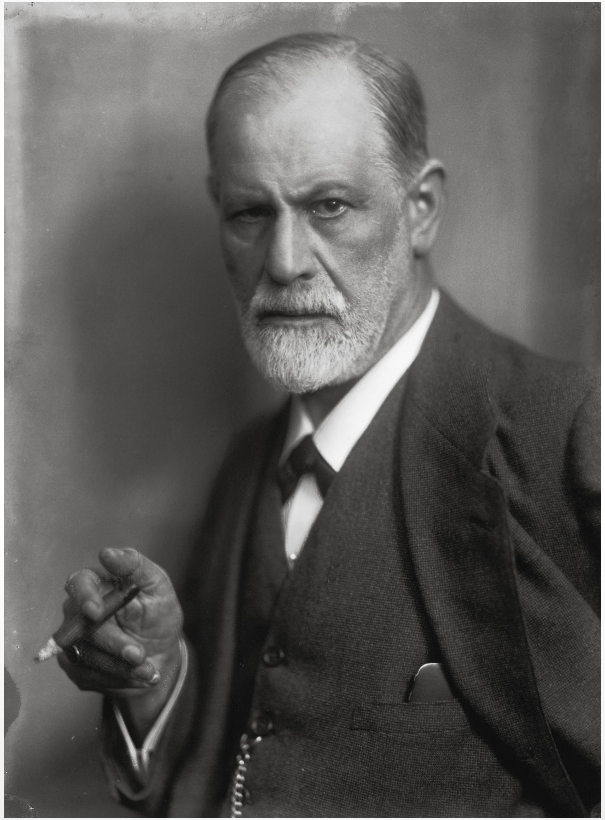
We need in Python dev 'cause Data Mining

You're a programmer, not an analyst



Write your backends!

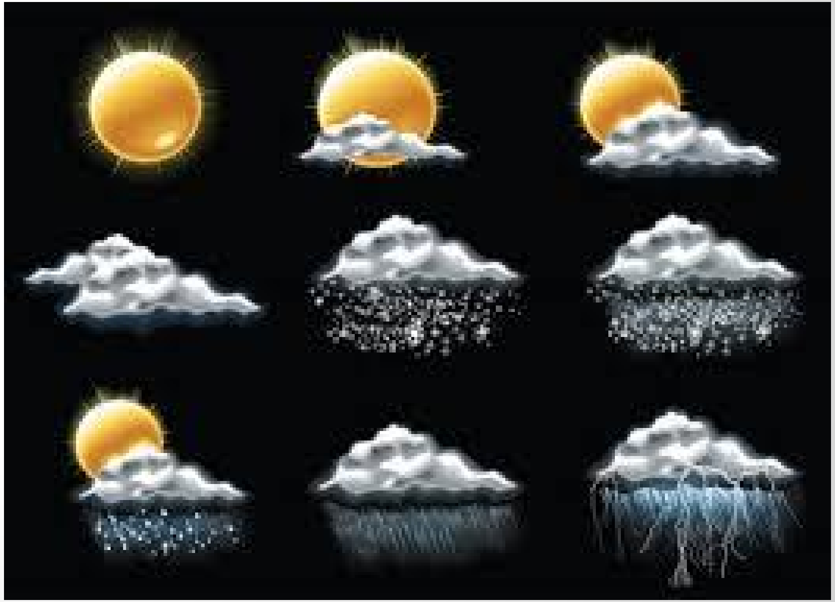
Let’s talk about it, Java-boy…



Data mining

Mining coal
in your data

Hey, man, predict me something!



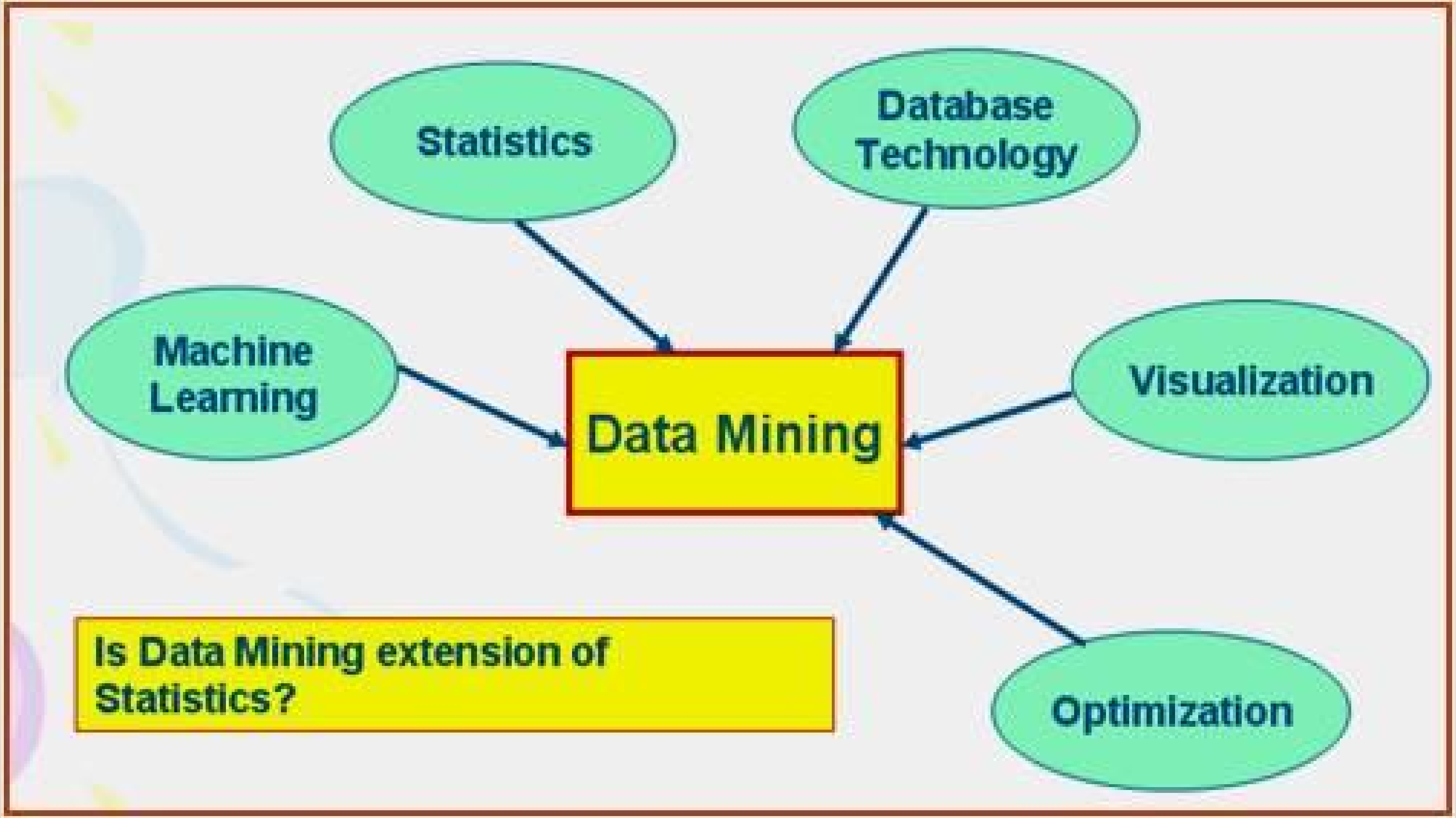
Man or sofa?

Typical questions for DM

- Which loan applicants are high-risk?
- How do we detect phone card fraud?
- Which customers do prefer product A over product B?
- What is the revenue prediction for next year?

What is Data Mining?

Statistics?

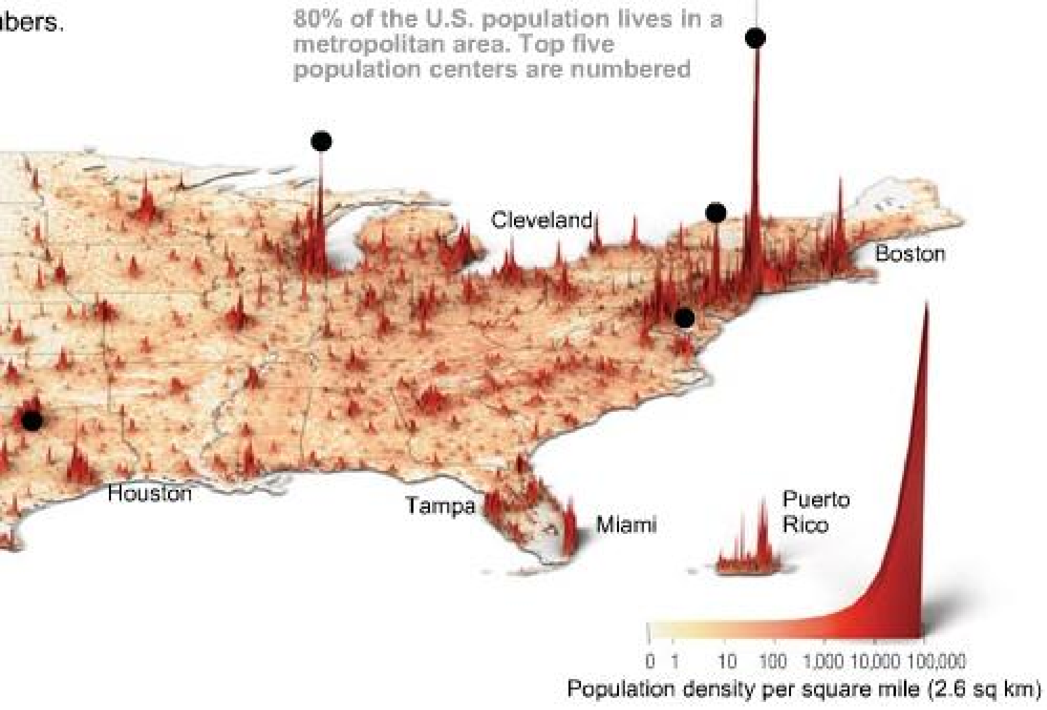


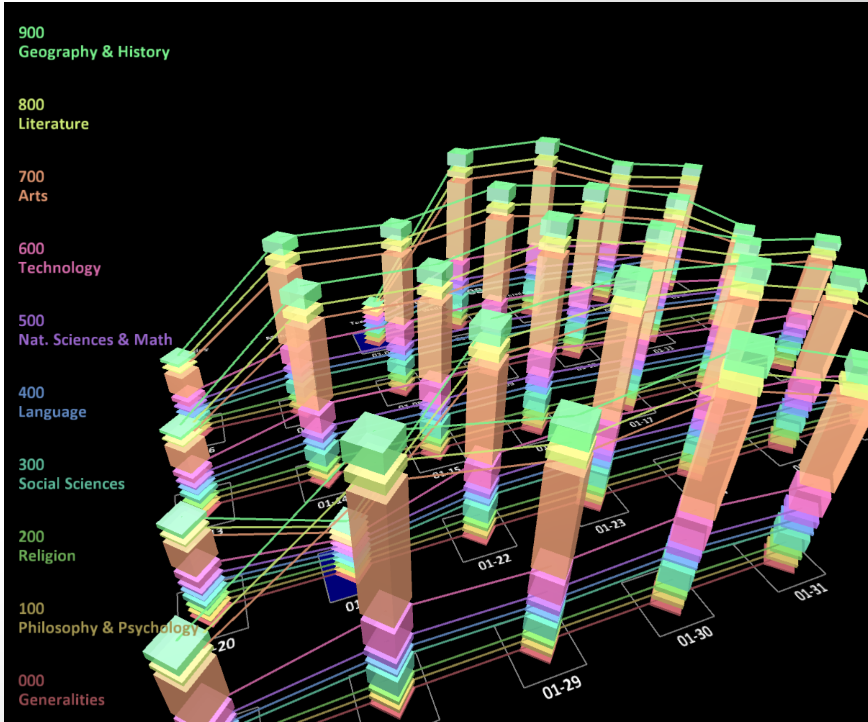
Is Data Mining extension of Statistics?

Tag cloud?

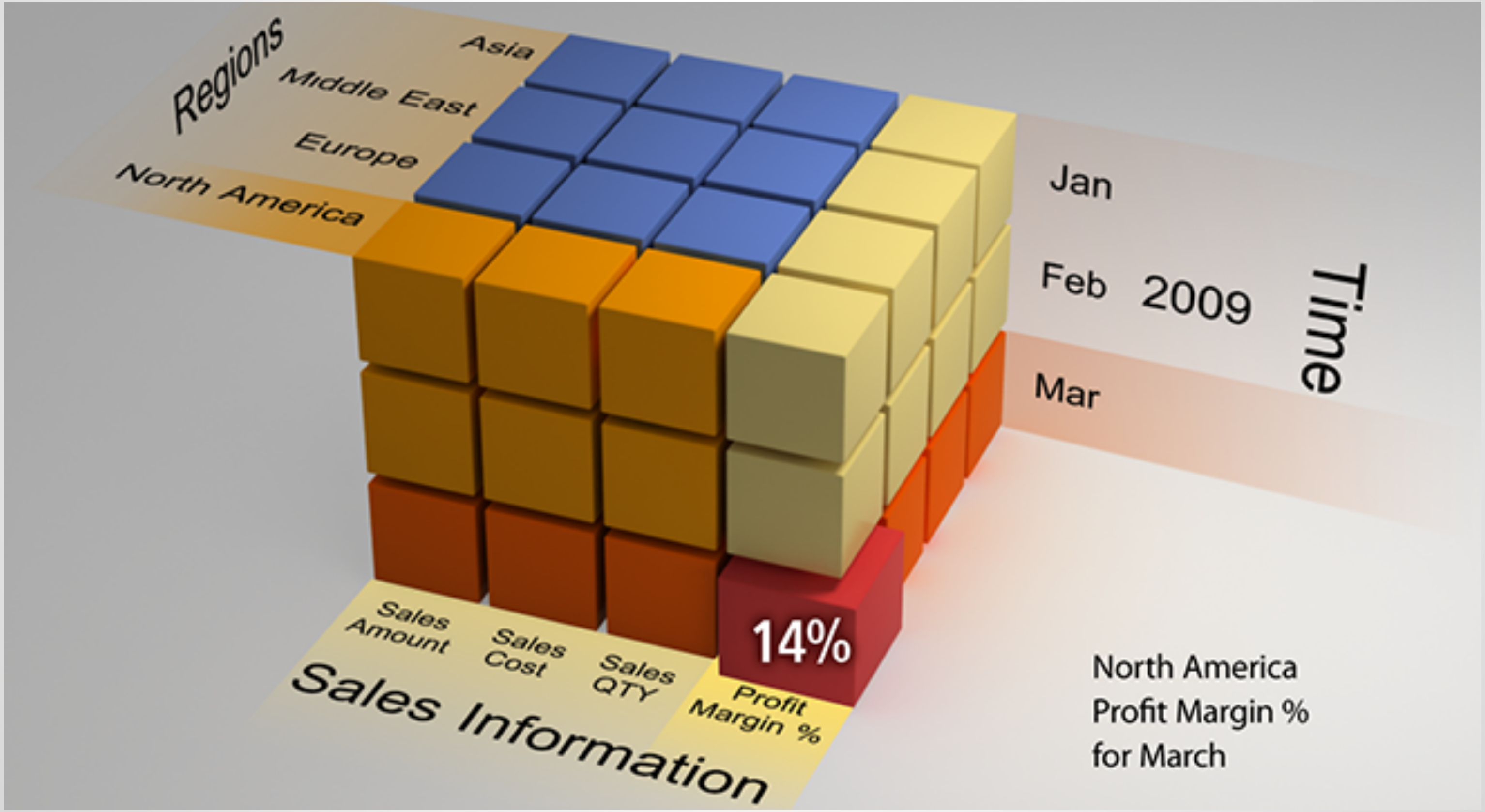
history
amount
uncover
Exploring
applications
pre-existing
marketing
different
held
supermarket
stored
find
transactions
warehousing
data
pattern
digging
depends
search
processing
businesses
behavior
statistical
mining
gathered
preferences
implies
establish
class
relationship
summarizing
computerized
meaning
analyze
look
practice
learn
habits
group
trends
Web
understand
important
detect
way
dicing
relationships
databases
raw
information
predict
order
algorithms
hidden
management
increase
part
automatically
decrease
huge
done
large
strategies
database
used
sorting
use
gathering
business
patterns
preexisting
process
new
turn
one
programs
identify
perspectives
searching
using
amounts
companies
manually
useful
via
develop
better
well
collecting
computer
within
effective
analyzing
collection
activity
slicing
sophisticated
becomes
sales
customers
customer
costs
CRM
batches
tons
discover
contained
detailed
shopping
capabilities
future
correlations
fraud
software
obvious
wade

Data visualization?


bers.
80% of the U.S. population lives in a metropolitan area. Top five population centers are numbered
Cleveland
Boston
Houston
Tampa
Miami
Puerto Rico
0 1 10 100 1,000 10,000 100,000
Population density per square mile (2.6 sq km)

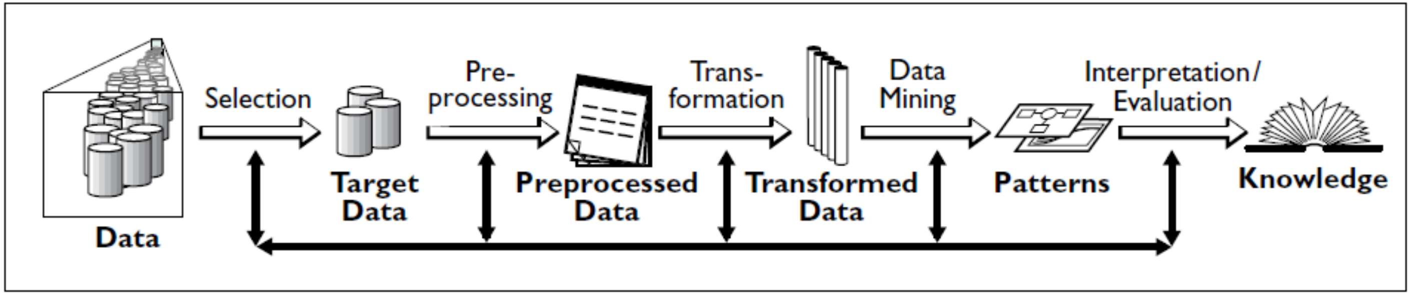


Not OLAP, 100%


Regions
Asia
Middle East
Europe
North America
Jan
Feb
2009
Mar
Time
Sales Amount
Sales Cost
Sales QTY
Profit Margin %
Sales Information
14%
North America
Profit Margin %
for March

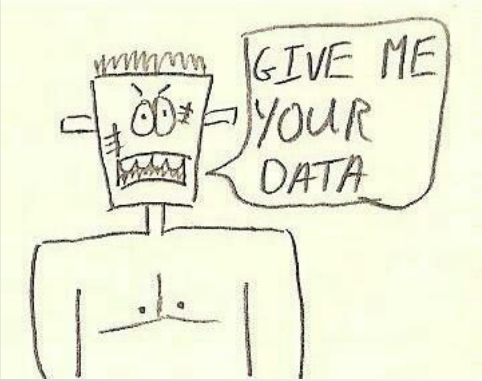
Magic part of KDD (Knowledge Discovery in Databases)

Figure 1. Overview of the steps constituting the KDD process



1. Selection
2. Pre-processing
3. Transformation
4. *Data Mining*
5. Interpretation/Evaluation

How it really works



1. Share your date with us
2. Our magic manipulations
3. Building an answering machine
4. PROFIT!!!

Data

Data examples

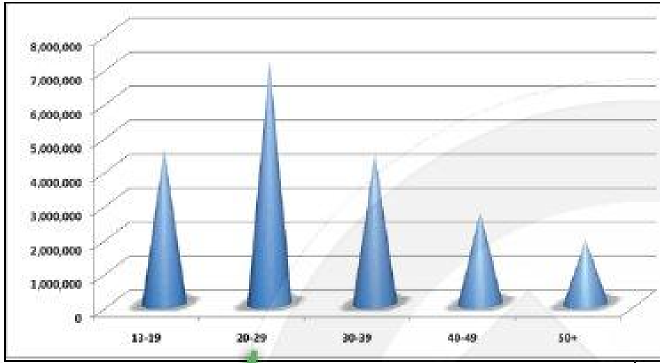
- Facebook users, tweets
- Weather
- Sea routes
- Trade transactions
- Goverment
- Medicine (genomic data)
- Telecommuncations (phone call records)

Data sources

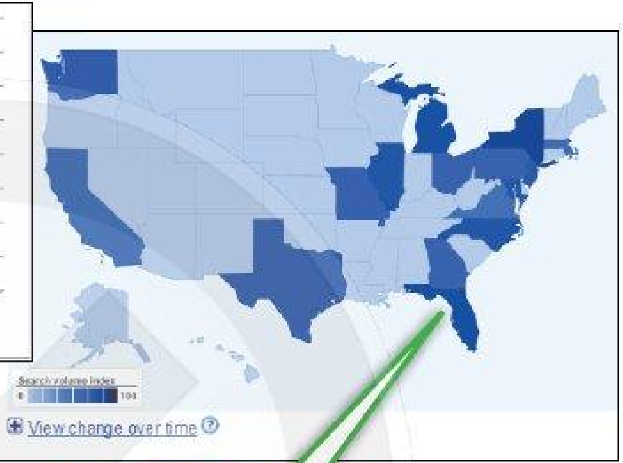
+
-
DO NOT REVERSE POLARITY
IOXUS
IOXUS www.ioxus.com 2.3V-800F RHE2R3807SR
IOXUS www.ioxus.com 2.3V-1000F
IOXUS www.ioxus.com 2.7V-100F
IOXUS www.ioxus.com 2.3V-220F
IOXUS www.ioxus.com 2.7V-350F RSC2R7357SR

- Relational Databases
- Data warehouses (Historical data)
- Files in CSV or in binary format
- Internet or electronic mails
- Scientific, research (R, Octave, Matlab)

Target Data & Personal Data



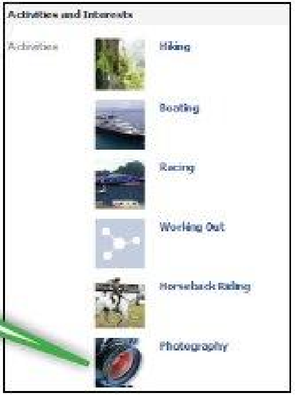
Target By Age & Gender



Target By Geographical Location

Target Customer "Likes"

Target Activities & Interests


Activities and Interests

Activities

Hiking

Boating

Racing

Working Out

Horseback Riding

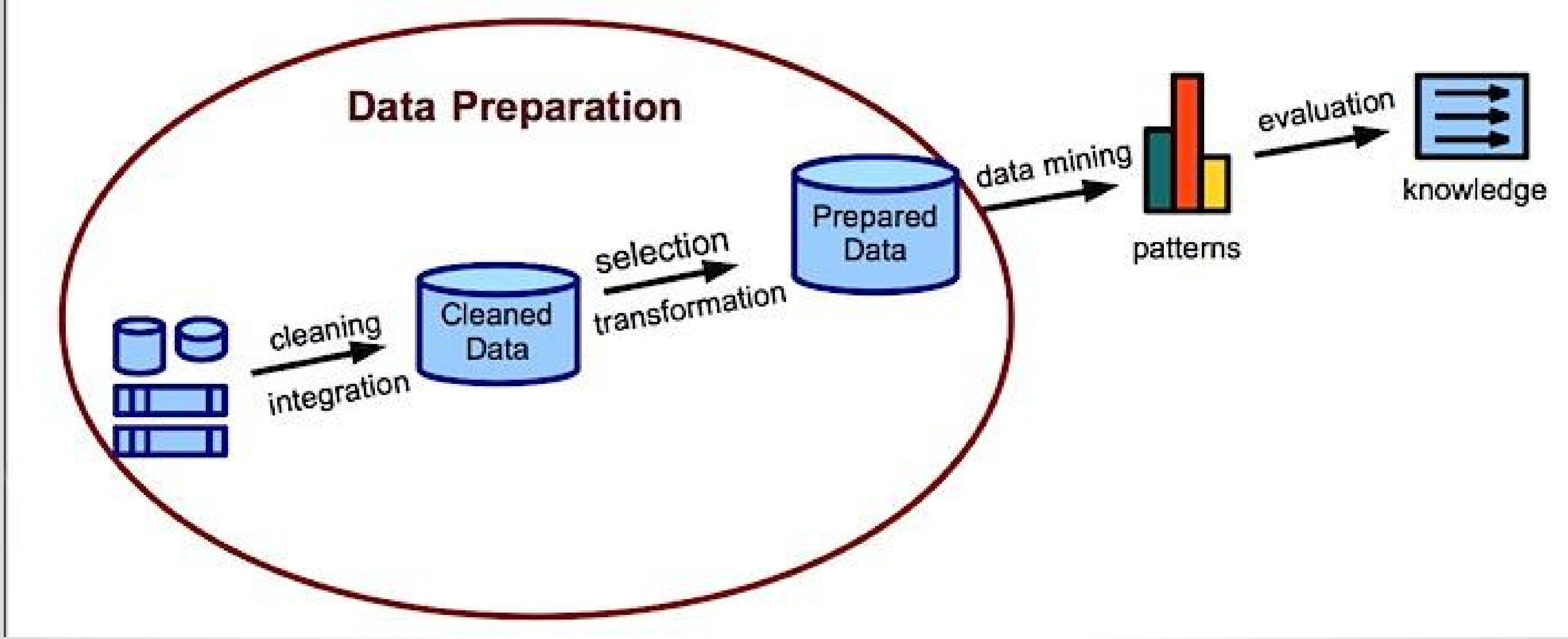
Photography

Pay with your personal data

MARCH 21, 2011
A tale of two Libyas
Plus: Why the U.S. can't sit on the sidelines BY FAREED ZAKARIA
The GOP's misinformation campaign BY JOE KLEIN
Could your baby be depressed?
THE CULTURE Word up: A dictionary of slang
TIME
YOUR DATA FOR SALE
Everything about you is being tracked—get over it
BY JOEL STEIN
Owns a laptop
Age: 38-39
Household income: $100,000+
Likes: online news
Politically active
Religion: Jewish
hockey
Married
Smart-phone user
had LASIK surgery
Likes: restaurants
Dislikes: autos & vehicles
What data-mining companies think they know about Joel Stein
© POST TYPOGRAPHY ⟵ PLEASE CREDIT AND LINK TO ⟶ POSTTYPOGRAPHY.COM
www.time.com

- All your personal data (PD) are being deeply mined
- The industry of collecting, aggregating, and brokering PD is “database marketing.”
- 1.1 billion browser cookies, 200 million mobile profiles, and an average of 1,500 pieces of data per consumer in Acxiom

Preprocessing



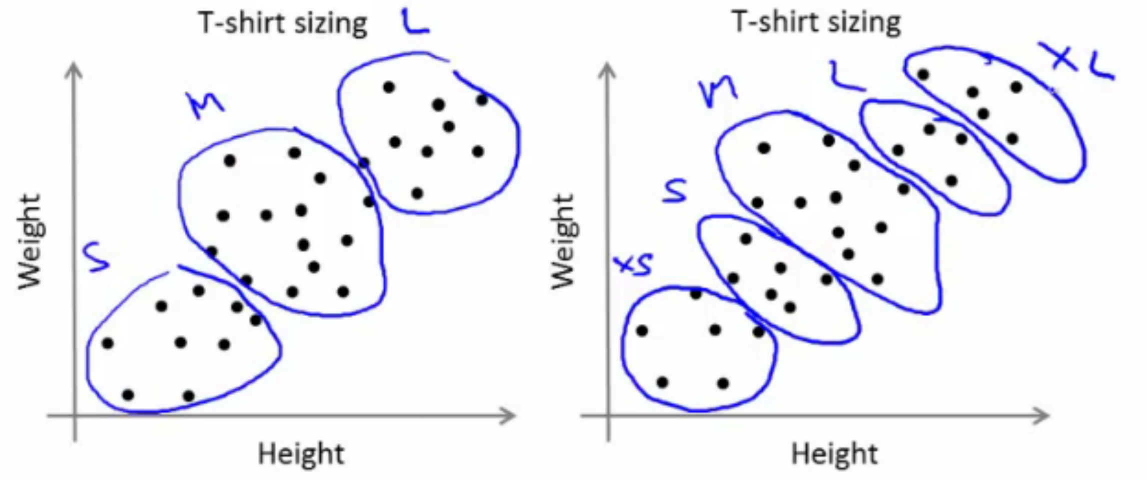
- Select small pieces
- Define default values for missed data
- Remove strange signals from data
- Merge some tables in one if required

Pattern mining

28/72

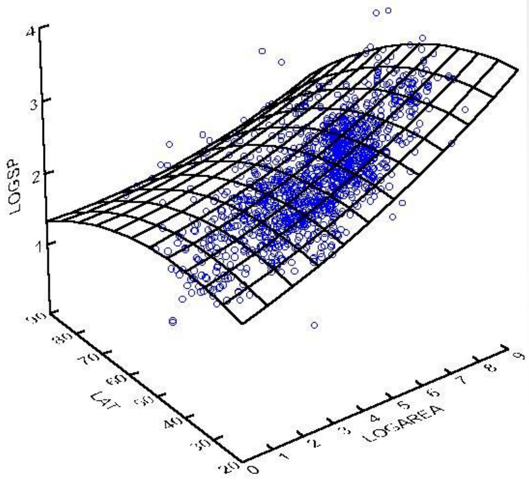
Association rule learning

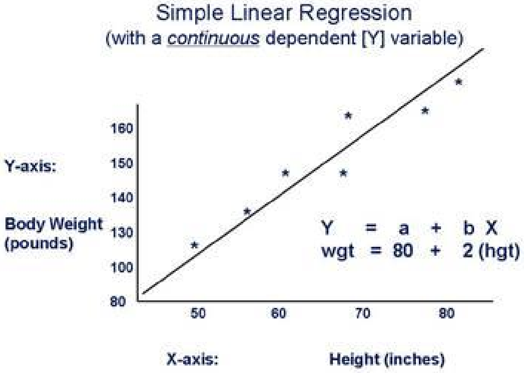
What is Cluster Analysis?



It is the process of finding model of function that describes and distinguishes data class to predict the class of objects whose class label is unknown.

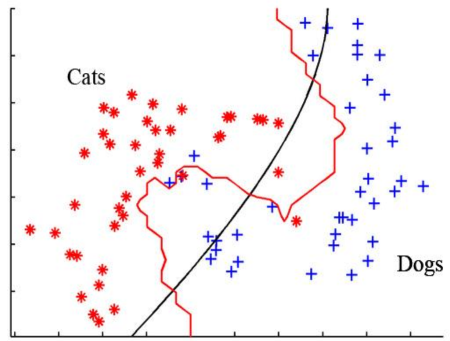
Regression

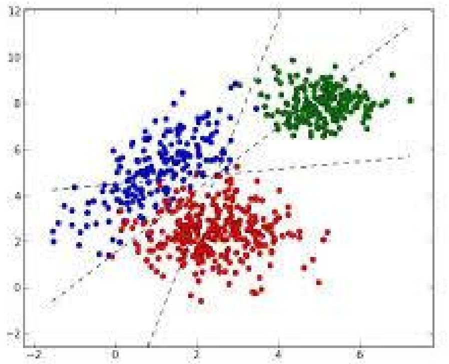




- Statistical process for estimating the relationships among variables
- The estimation target is function (it can be probability distribution)
- Can be linear, polynomial, nonlinear and etc.

Classification

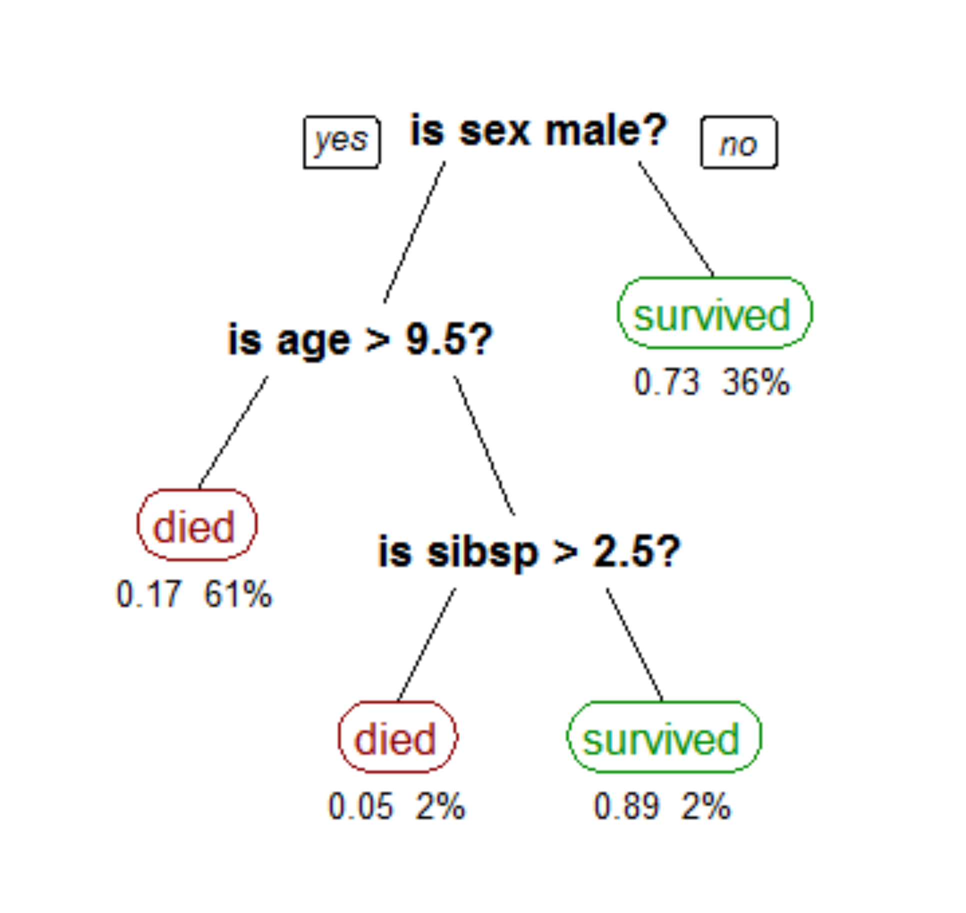


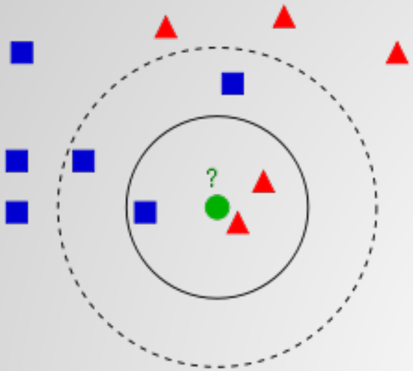


- Training set of classified examples (supervised learning)
- Test set of non-classified items
- Main goal: find a function (classifier) that maps input data to a category
- Computer vision, drug discovery, speech recognition, biometric indentification, credit scoring

Decision trees

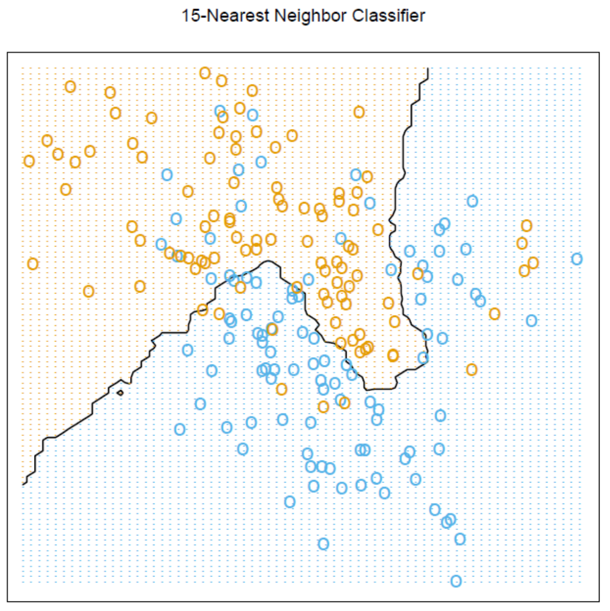
Decision trees





kNN (k-nearest neighbor)

- There are two classes of objects A & B
- Define the class of new object, based on information about its neighbors
- Changing the boundaries of an new object area, we form a set of neighbors.
- New object is B becuase majority of the neighbors is a B.



Skills & Tools

Big Data Landscape

Vertical Apps

- PREDICTIVE POLICING
- bloomreach GET FOUND.
- MYRRIX

Ad/Media Apps

- rocketfuel
- collective[i]
- bluefin
- Recorded Future
- LuckySort
- Media Science
- TURN
- DataXu Data. Insight. Action.

Log Data Apps

- splunk>
- loggly
- sumologic

Data As A Service

- kaggle
- factual.
- knoema beta
- GNIP
- DATASIFT
- Windows Azure Marketplace
- INRIX
- LexisNexis
- SPACE CURVE
- LOQATE Everything Location

Business Intelligence

- ORACLE | Hyperion
- SAP
- Business Objects
- RJMetrics
- Microsoft | Business Intelligence
- IBM
- COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- Chart.io
- GoodData

Analytics and Visualization

- tableau
- Palantir
- OPERA
- metaLayer
- dataspora
- METAMARKETS
- centrifuge
- TERADATA ASTER
- sas
- TIBCO
- KARMASPHERE
- panopticon Real-Time Visual Data Analysis
- pentaho
- Datameer
- ClearStory
- CIRRO
- platfora
- alteryx
- visual.ly
- AYATA

Analytics Infrastructure

- Hortonworks
- VERTICA An HP Company
- MAPR
- cloudera
- INFOBRIGHT
- PARACCEL
- EMC²
- GREENPLUM
- NETEZZA
- kognitio
- DATASTAX
- EXASOL
- calpont

Operational Infrastructure

- Couchbase
- 10gen | MongoDB company
- TERADATA
- HADAPT
- TERRACOTTA
- VoltDB
- MarkLogic
- INFORMATICA

Infrastructure As A Service

- amazon web services
- Windows Azure
- infochimps
- Google BigQuery

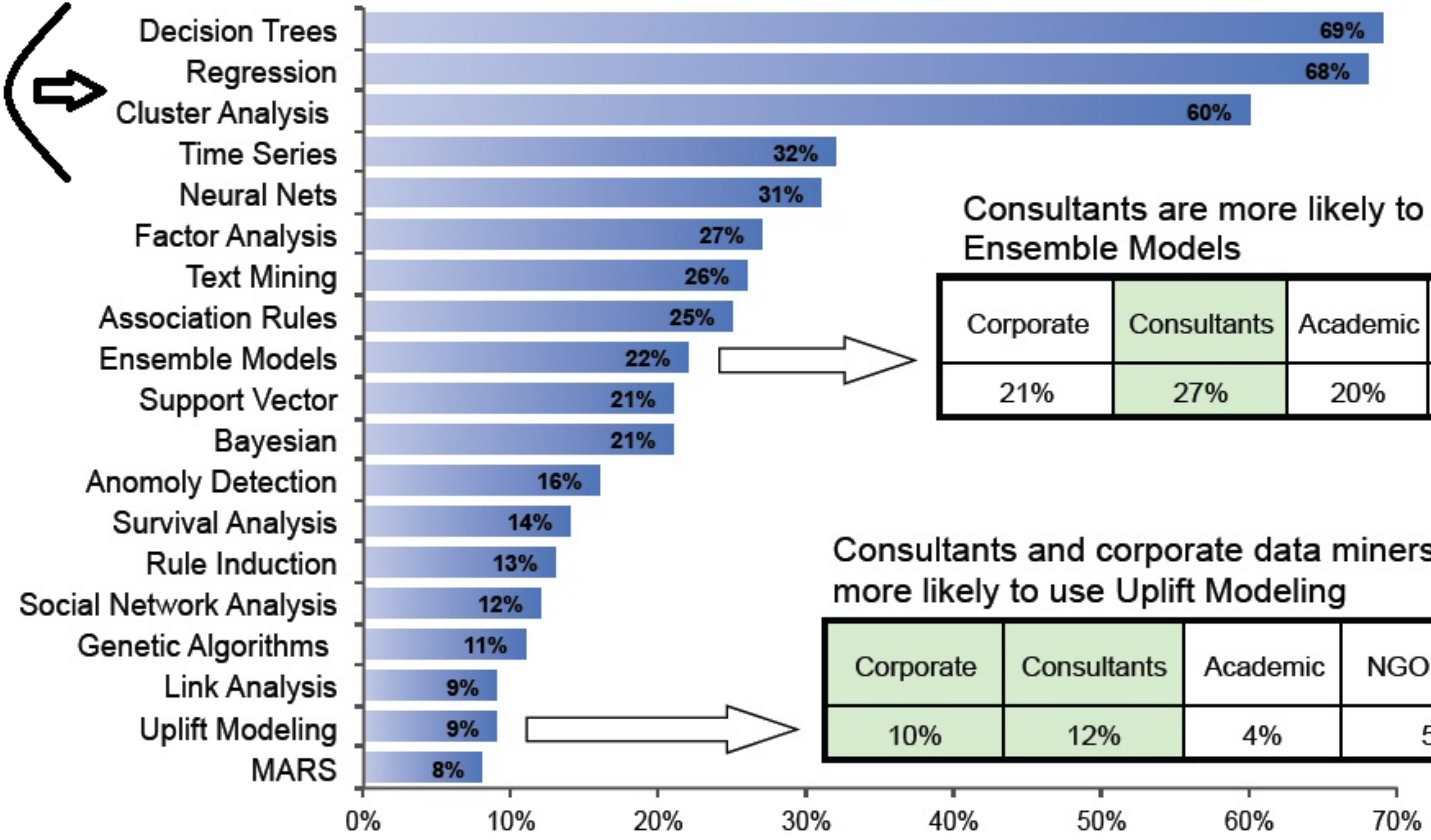
Structured Databases

- ORACLE
- MySQL
- Microsoft SQL Server
- PostgreSQL
- IBM DB2
- SYBASE
- memsql

Technologies

- hadoop
- hadoop MapReduce
- mahout
- APACHE HBASE
- Cassandra

dave@vcdave.com blogs.forbes.com/davefeinleib



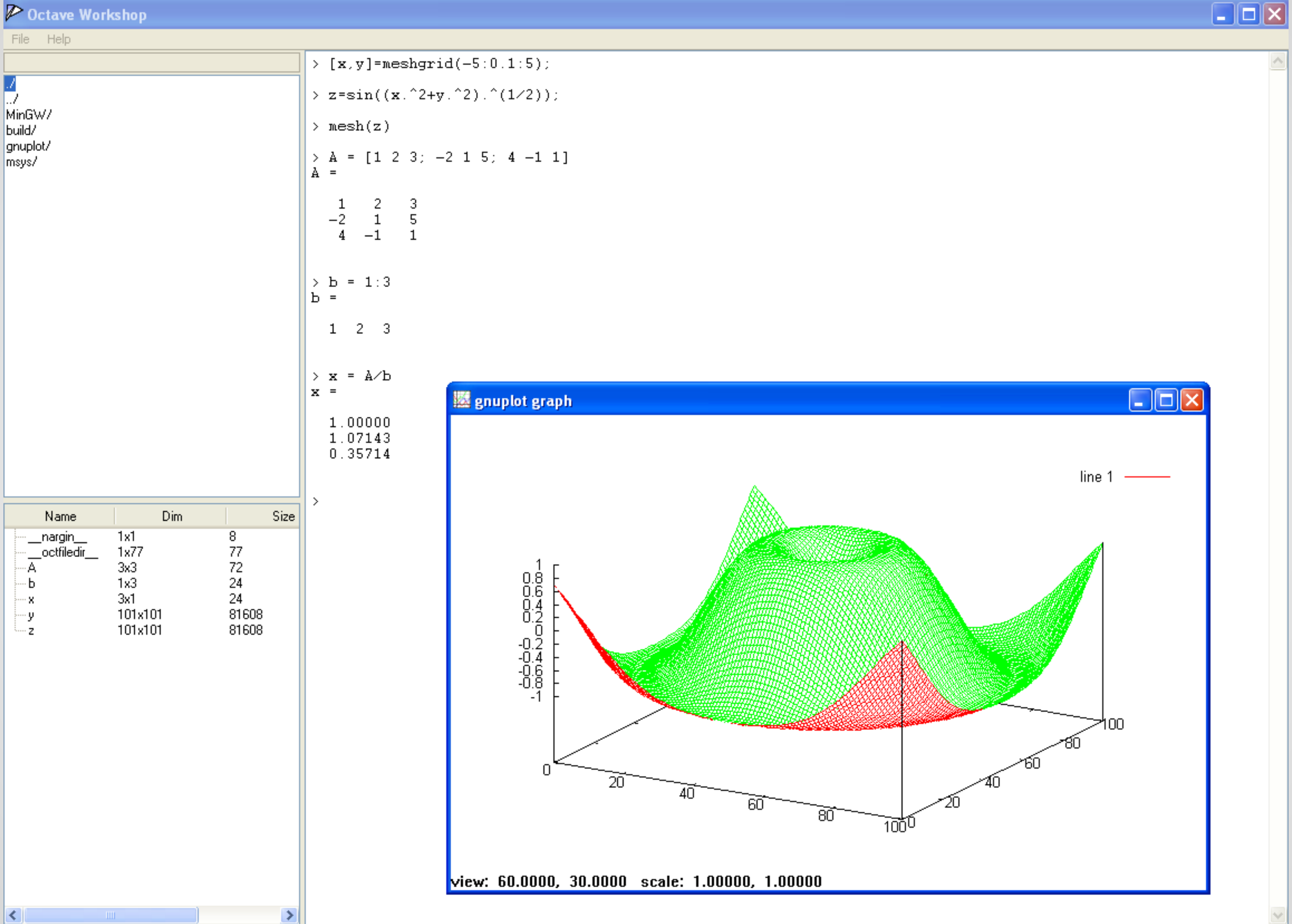
Consultants are more likely to use Ensemble Models

Corporate	Consultants	Academic	NGO / Gov't
21%	27%	20%	18%

Consultants and corporate data miners are more likely to use Uplift Modeling

Corporate	Consultants	Academic	NGO / Gov't
10%	12%	4%	5%

Fashion Languages



./
../
MinGW/
build/
gnuplot/
msys/

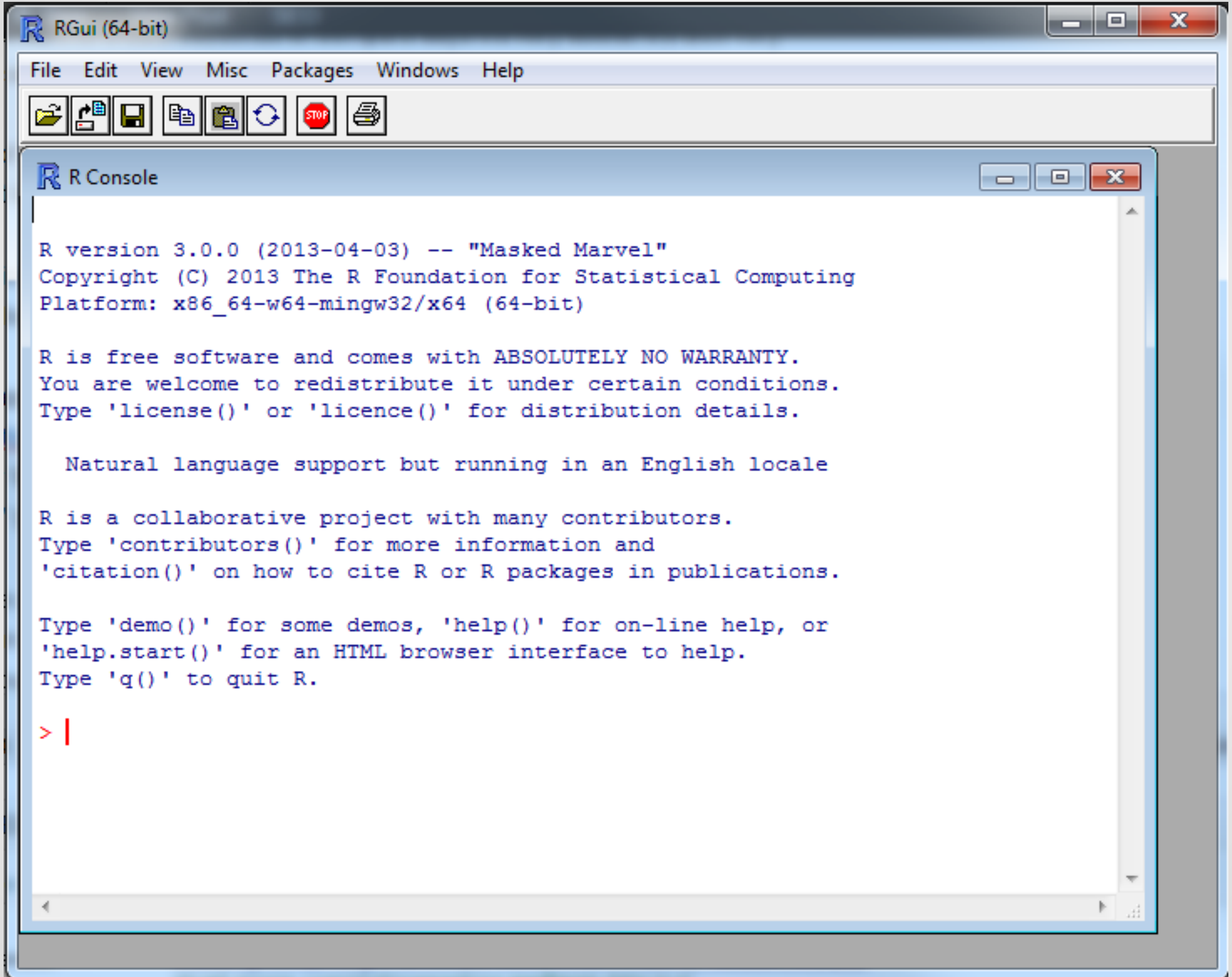
Name	Dim	Size
__nargin__	1x1	8
__octfiledir__	1x77	77
A	3x3	72
b	1x3	24
x	3x1	24
y	101x101	81608
z	101x101	81608

```
> [x,y]=meshgrid(-5:0.1:5);

> z=sin((x.^2+y.^2).^(1/2));

> mesh(z)

> A = [1 2 3; -2 1 5; 4 -1 1]
A =

   1   2   3
  -2   1   5
   4  -1   1


> b = 1:3
b =

  1  2  3


> x = A/b
x =

  1.00000
  1.07143
  0.35714


>
```

view: 60.0000, 30.0000   scale: 1.00000, 1.00000

Why not Octave?

- It’s free
- Not full implemented stack of ML algorythms
- All your matrix are belong to us!
- Single thread model
- Java support



```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

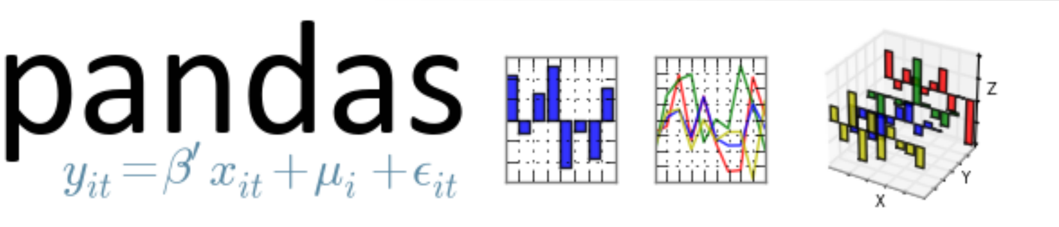
Why not R?

- 25% of R packages are written in Java
- Syntax is too sweet
- You should read 1000 lines in docs to write 1 line of code
- Single thread model for 95% algorythms

Why not Python?

- Now Python is an idol for young scientists due to the low barrier to entry
- We are not Python developers
- High-level language
- Have you ever heard about a Jython?
- Long long way to real Highload production

DM libraries in Python



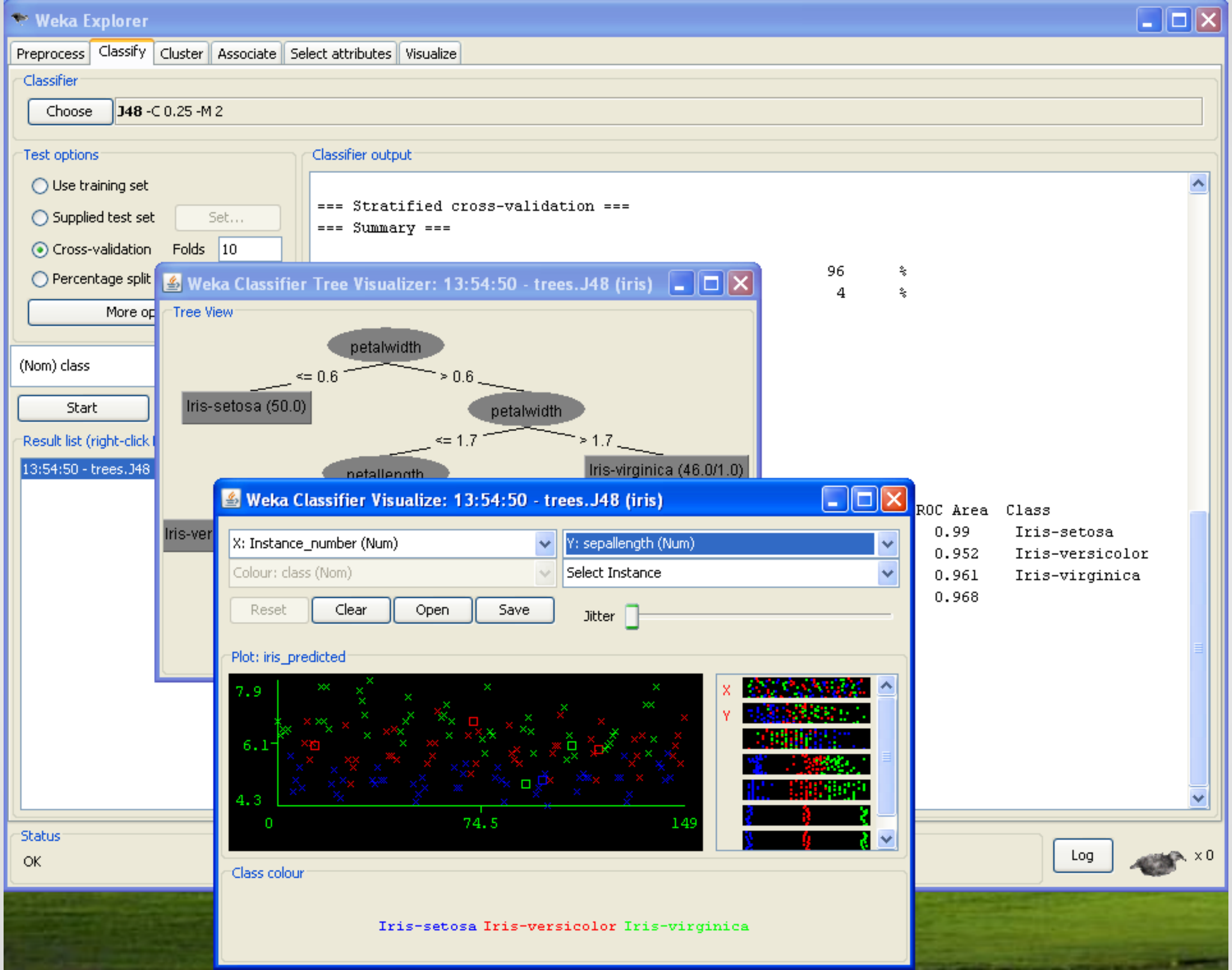
Java ecosystem

JDM

- Java API for Data mining, JSR 73 and JSR 247
- **javax.datamining.supervised** defines the supervised function-related interfaces
- **javax.datamining.algorithm** contains all mining algorithm subclass packages
- JDM 2.0 adds Text Mining, Time series and so on..

Weka

- Connectors to R, Octave, Matlab, Hadoop, NoSQL/SQL databases
- Source code of all algorythms in Java
- Preprocessing tools: discretization, normalization, resampling, attribute selection, transforming and combining



J48 -C 0.25 -M 2

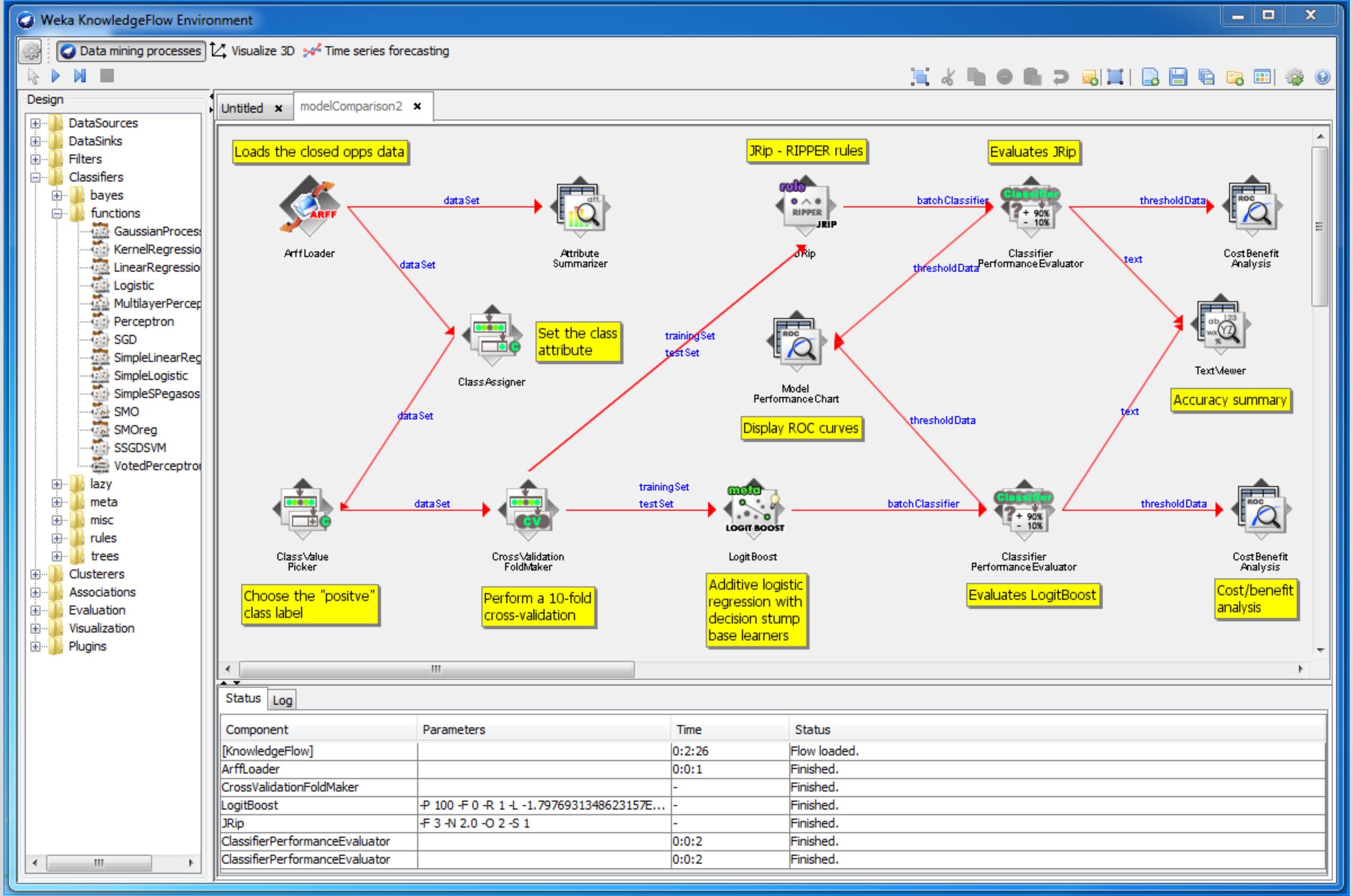
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split

OK

```
=== Stratified cross-validation ===
=== Summary ===

                                   96      %
                                    4      %

ROC Area  Class
  0.99     Iris-setosa
  0.952    Iris-versicolor
  0.961    Iris-virginica
  0.968
```

Weka Classifier Tree Visualizer: 13:54:50 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
Iris-ver
Weka Classifier Visualize: 13:54:50 - trees.J48 (iris)

X: Instance_number (Num)
Y: sepallength (Num)
Colour: class (Nom)
Select Instance



Loads the closed opps data

JRip - RIPPER rules

Evaluates JRip

Set the class attribute

Accuracy summary

Display ROC curves

Choose the "positve" class label

Perform a 10-fold cross-validation

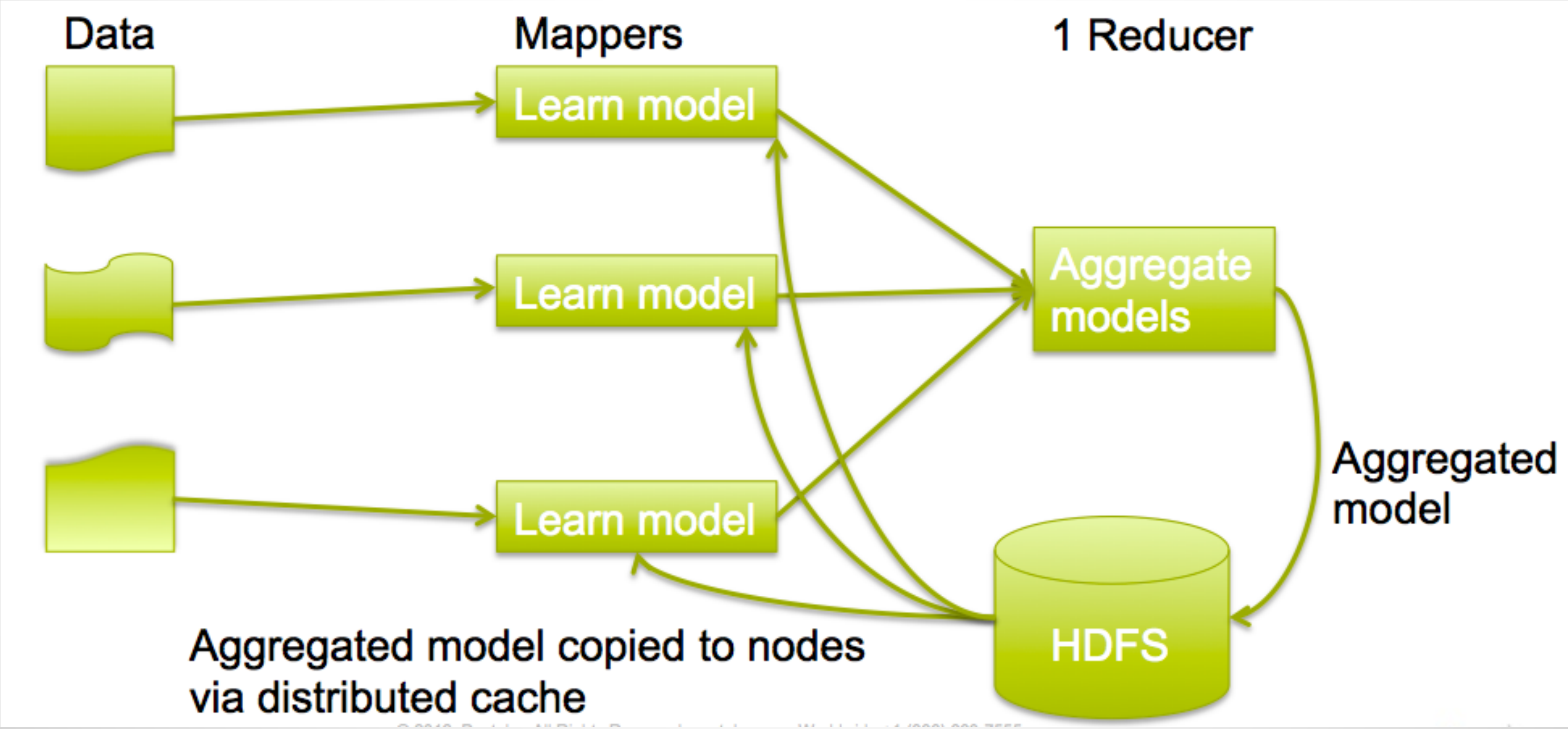
Additive logistic regression with decision stump base learners

Evaluates LogitBoost

Cost/benefit analysis

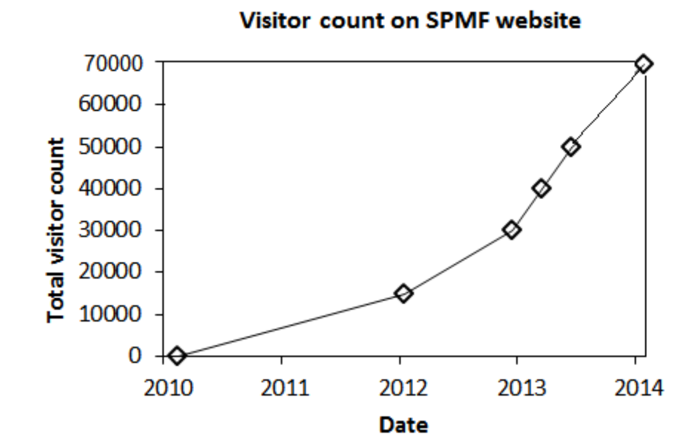
Component	Parameters	Time	Status
[KnowledgeFlow]		0:2:26	Flow loaded.
ArffLoader		0:0:1	Finished.
CrossValidationFoldMaker		-	Finished.
LogitBoost	-P 100 -F 0 -R 1 -L -1.7976931348623157E...	-	Finished.
JRip	-F 3 -N 2.0 -O 2 -S 1	-	Finished.
ClassifierPerformanceEvaluator		0:0:2	Finished.
ClassifierPerformanceEvaluator		0:0:2	Finished.

Weka + Hadoop



Aggregated model copied to nodes via distributed cache

SPMF

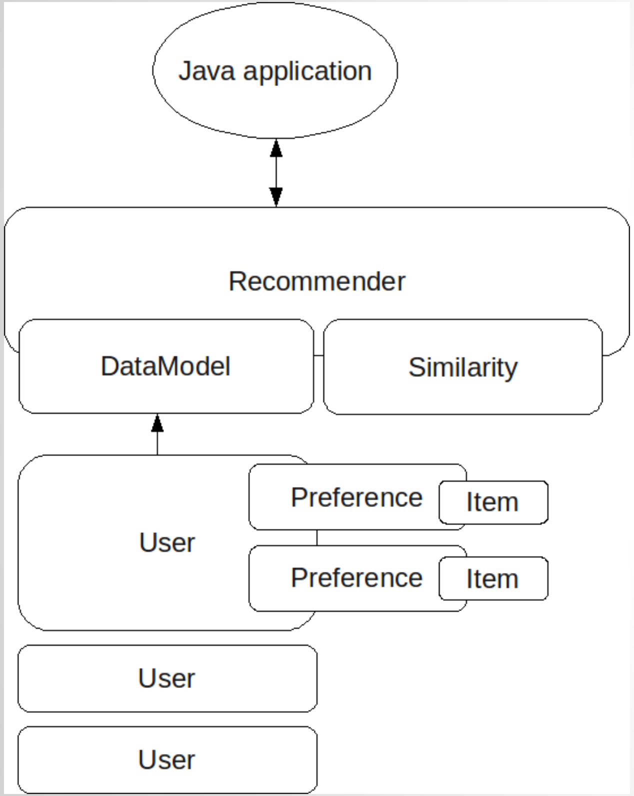


- It’s codebase of algorythms in pattern mining field
- It has cool examples and implementation of 78 algorythms
- Cool performance results in specific area
- Codebase grows very fast

Mahout

- Driven by Ng et al.'s paper "MapReduce for Machine Learning on Multicore"
- Next algorythms were adopted: Locally Weighted Linear Regression(LWLR), Naive Bayes (NB), k-means, Logistic Regression, Neural Network (NN), Principal Components Analysis (PCA), Support Vector Machine (SVM) and so on..
- The complexity was reduced in *n* times for *n* processors.

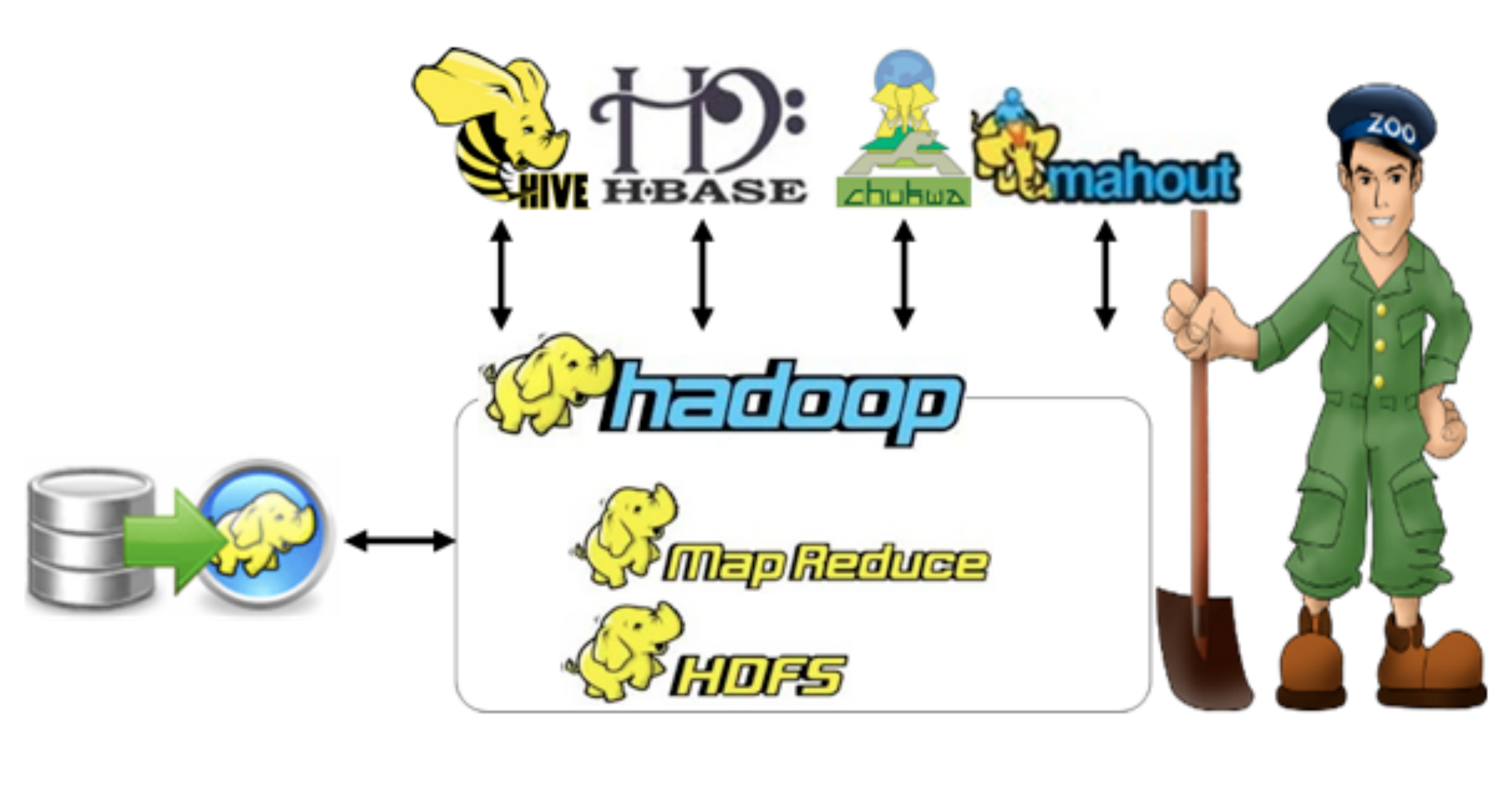
Mahout



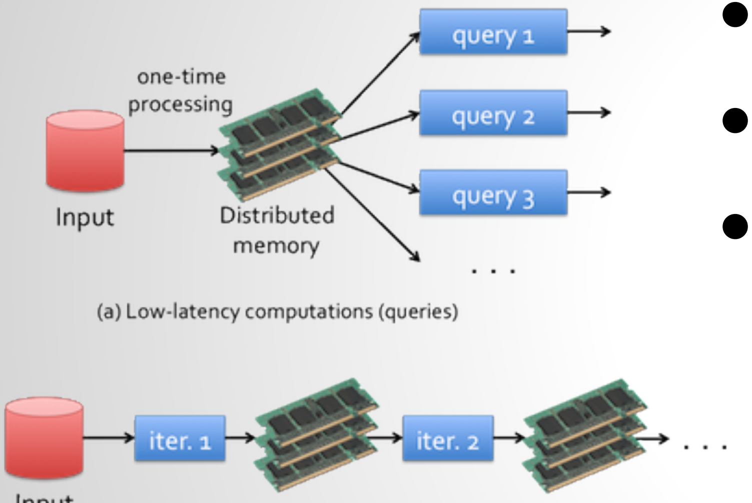
- DataModel (File, MySQL, PostgreSQL, Mongo, Cassandra)
- UserSimilarity
- ItemSimilarity
- UserNeighborhood
- Recommender

Mahout

- Advanced Implementations of Java’s Collections Framework for better Performance.
- Very close to Apache Giraph
- New algorythms will build on Spark platform
- Spark shell
- Spring + Mahout demo
- Collaborative Filtering, Classification, Clustering, Dimensionality Reduction, Miscellaneous are supported


HIVE
H·BASE
chukwa
mahout
ZOO
hadoop
Map Reduce
HDFS

Spark

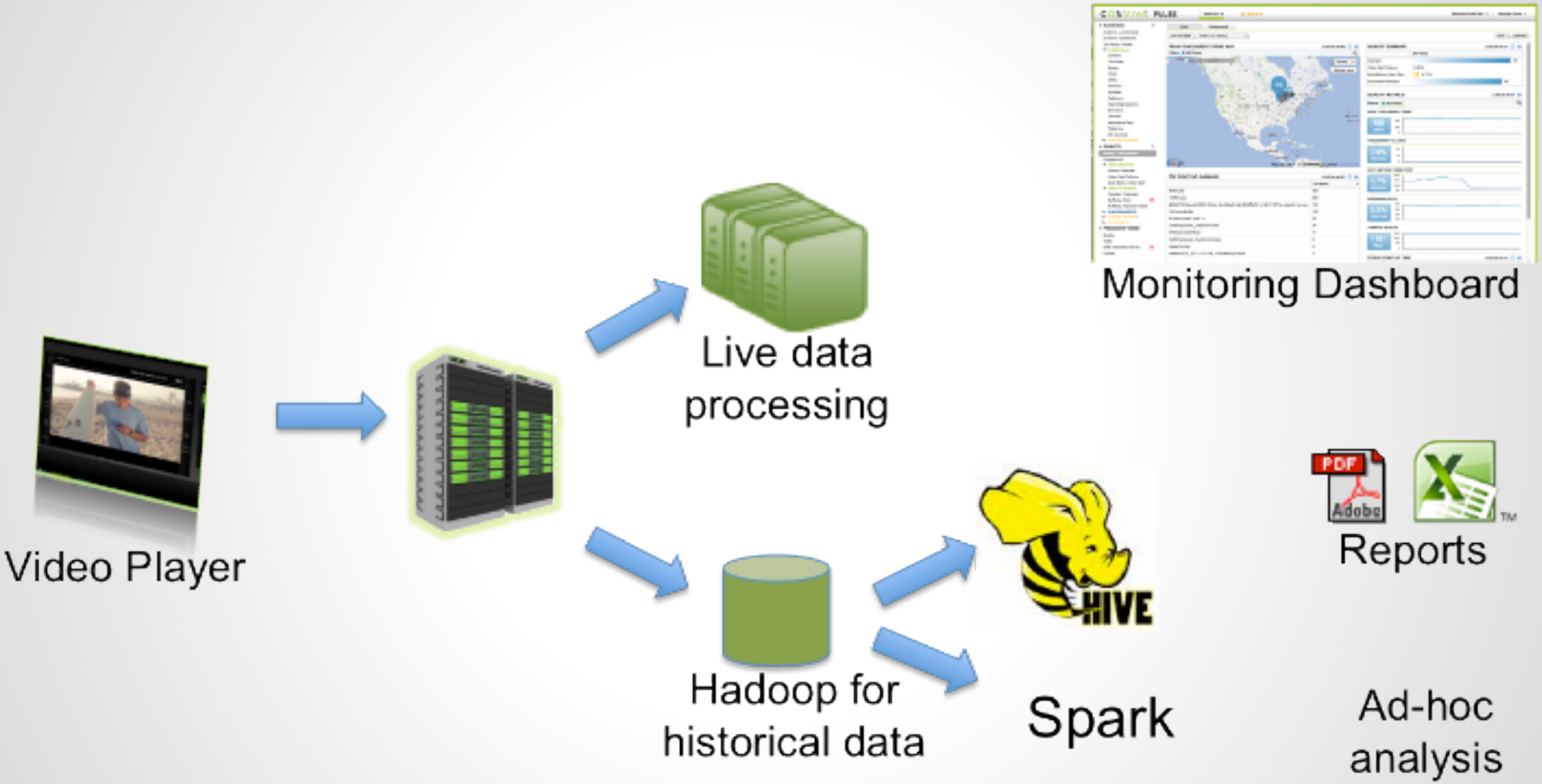


(a) Low-latency computations (queries)

(b) Iterative computations

- MapReduce in memory
- Up to 50x faster than Hadoop
- Support for Shark (like Hive), MLlib (Machine learning), GraphX (graph processing)
- RDD is a basic building block (immutable distributed collections of objects)

Spark



Monitoring Dashboard

Live data processing

Video Player

Reports

Hadoop for historical data

Spark

Ad-hoc analysis

Spark + Java 8

Java 7 search example:

```
JavaRDD<String> lines = sc.textFile("hdfs://log.txt").filter(
  new Function<String, Boolean>() {
    public Boolean call(String s) {
      return s.contains("Tomcat");
    }
});
long numErrors = lines.count();
```

Java 8 search example:

```
JavaRDD<String> lines = sc.textFile("hdfs://log.txt")
                          .filter(s -> s.contains("Tomcat"));
long numErrors = lines.count();
```

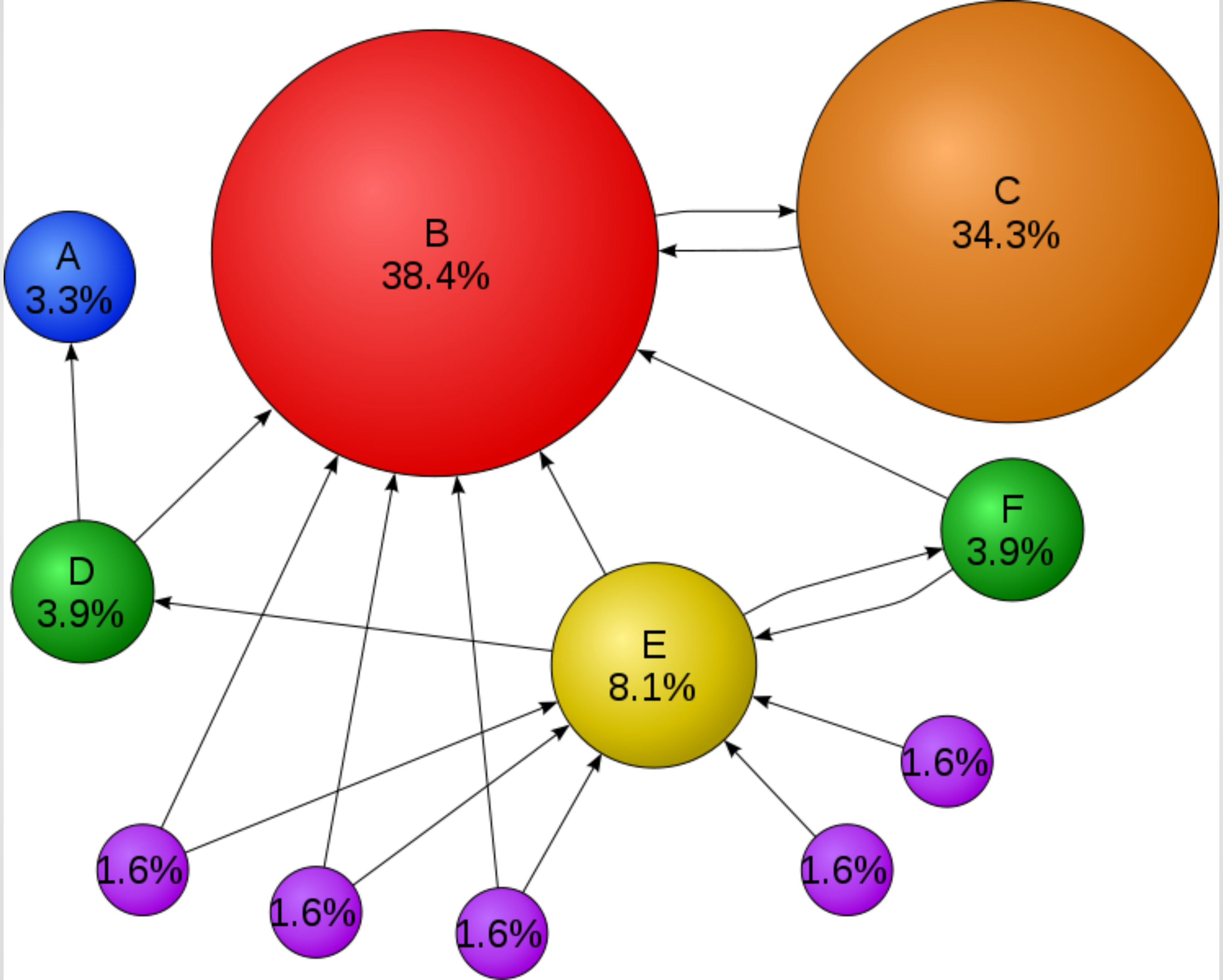
Mahout’s killer

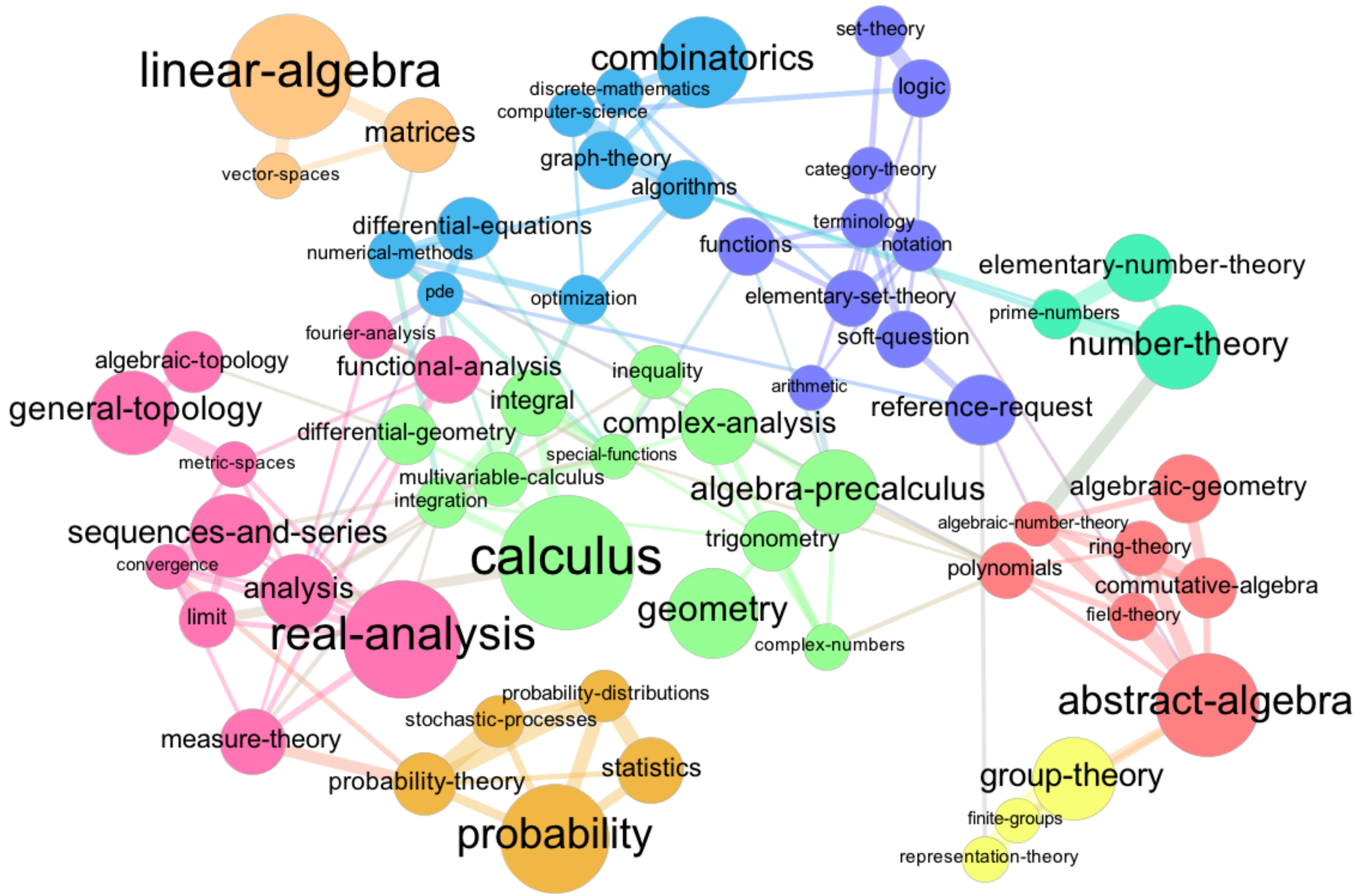
MLlib

- Classification and regression. collaborative filtering and clustering, Dimensionality reduction and Optimization are supported
- It extends scikit-learn (Python lib) and Mahout and run on Spark
- Well documented and integrated with many Java solutions

Size	**Classification**	**Tools**
Lines Sample Data	Analysis and Visualization	Whiteboard, bash
KBs - low MBs Prototype Data	Analysis and Visualization	Matlab, Octave, R
MBs - low GBs Online Data	Storage	MySQL (DBs)
MBs - low GBs Online Data	Analysis	NumPy, SciPy, Weka, BLAS/LAPACK
GBs - TBs - PBs BigData	Storage	HDFS, HBase, Cassandra
GBs - TBs - PBs Big Data	Analysis	Hive, Mahout, Hama, Giraph,MLlib

Large graph processing tools


A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%


linear-algebra
combinatorics
set-theory
logic
discrete-mathematics
computer-science
matrices
vector-spaces
graph-theory
algorithms
category-theory
differential-equations
terminology
functions
notation
numerical-methods
elementary-number-theory
elementary-set-theory
pde
optimization
prime-numbers
fourier-analysis
soft-question
number-theory
algebraic-topology
functional-analysis
inequality
arithmetic
general-topology
integral
complex-analysis
reference-request
differential-geometry
special-functions
metric-spaces
multivariable-calculus
algebra-precalculus
algebraic-geometry
integration
algebraic-number-theory
sequences-and-series
trigonometry
ring-theory
calculus
convergence
polynomials
analysis
commutative-algebra
field-theory
limit
geometry
real-analysis
complex-numbers
probability-distributions
abstract-algebra
stochastic-processes
measure-theory
statistics
probability-theory
group-theory
finite-groups
probability
representation-theory

Graph	Number of vertexes	Number of edges	Volume	Data/per day
Web-graph	1,5 * 10^12	1,2 * 10^13	100 PB	300 TB
Facebook (friends graph)	1,1 * 10^9	160 * 10^9	1 PB	15 TB
Road graph of EU	18 * 10^6	42 * 10^6	20 GB	50 MB
Road graph of this city	250 000	460 000	500 MB	100 KB

MapReduce for iterative calculations

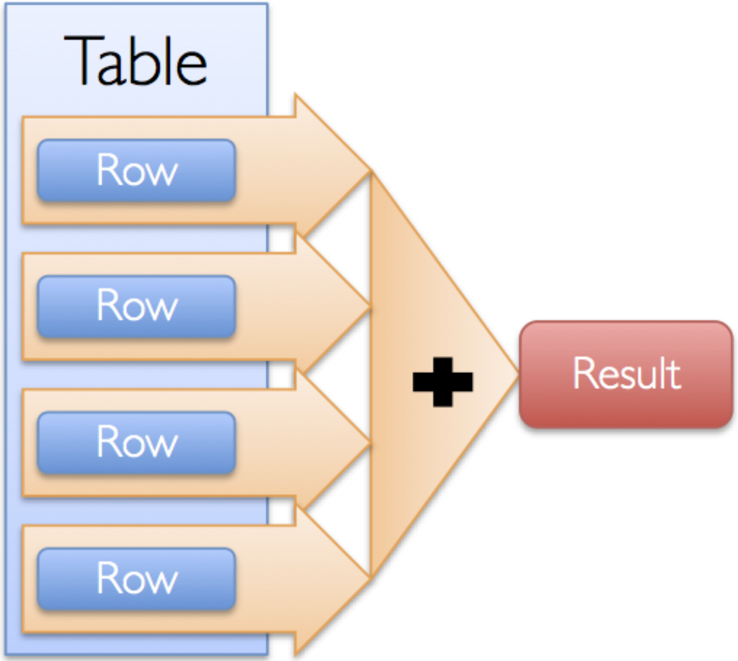
- High complexity of graph problem reduction to key-value model
- Iteration algorythms, but multiple chained jobs in M/R with full saving and reading of each state

Think like a vertex…

Data-Parallel

hadoop

Spark



Table

Row

Row

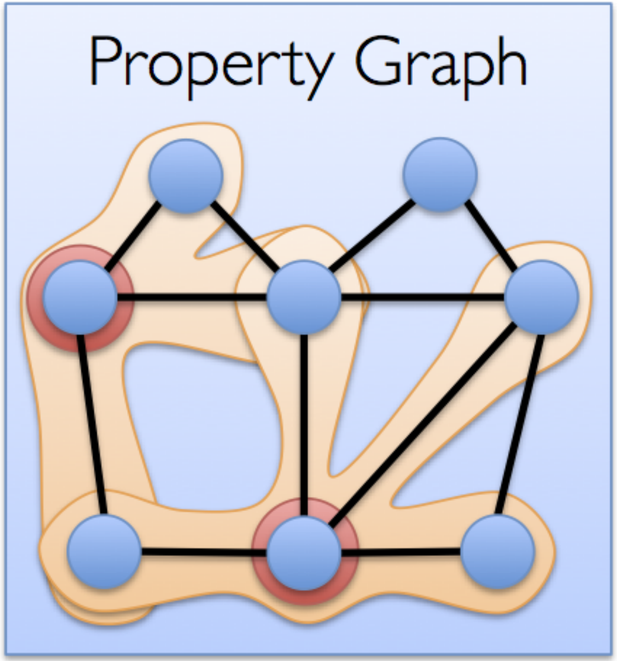
Row

Row

Result

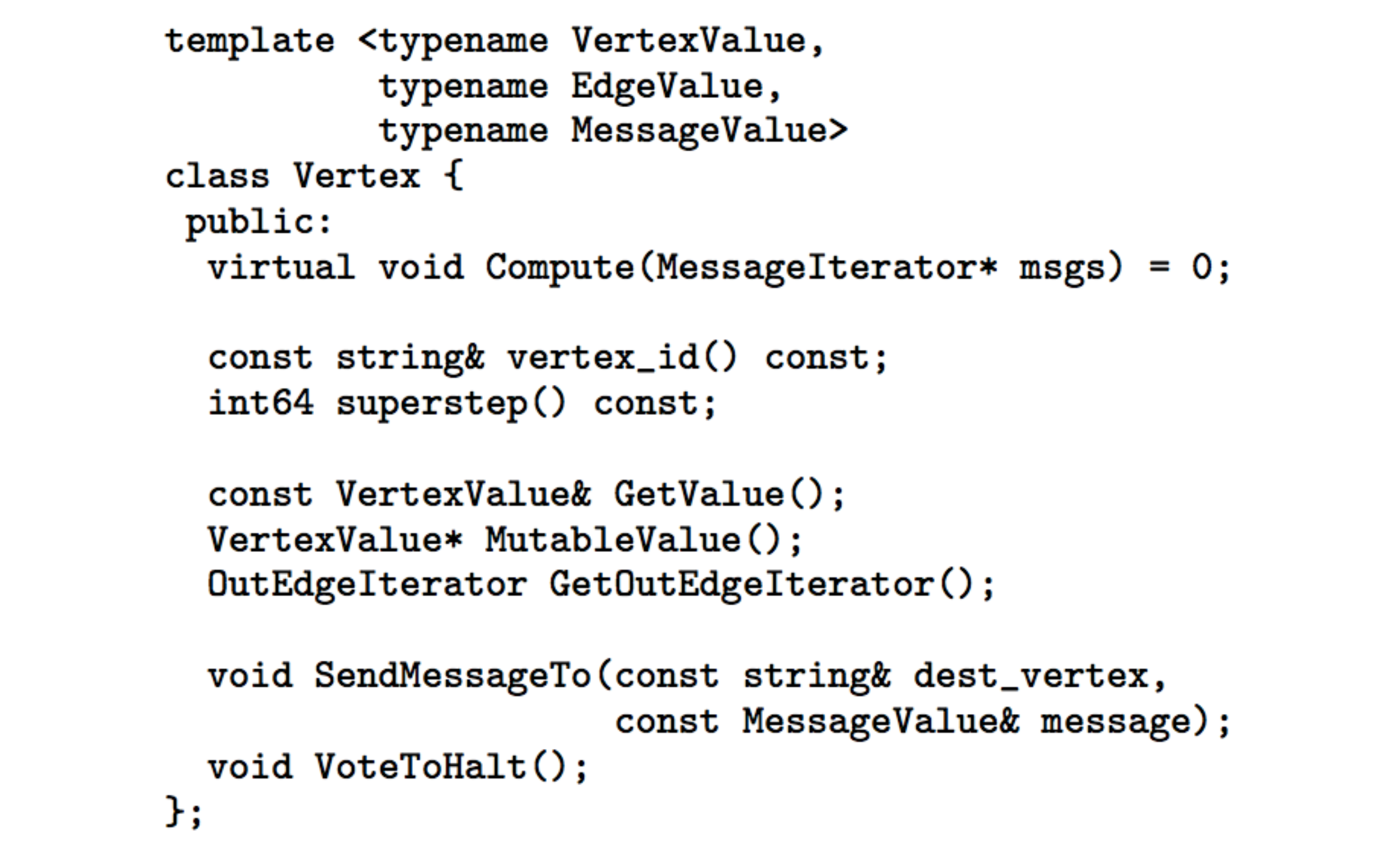
Graph-Parallel

Pregel GraphLab APACHE GIRAPH

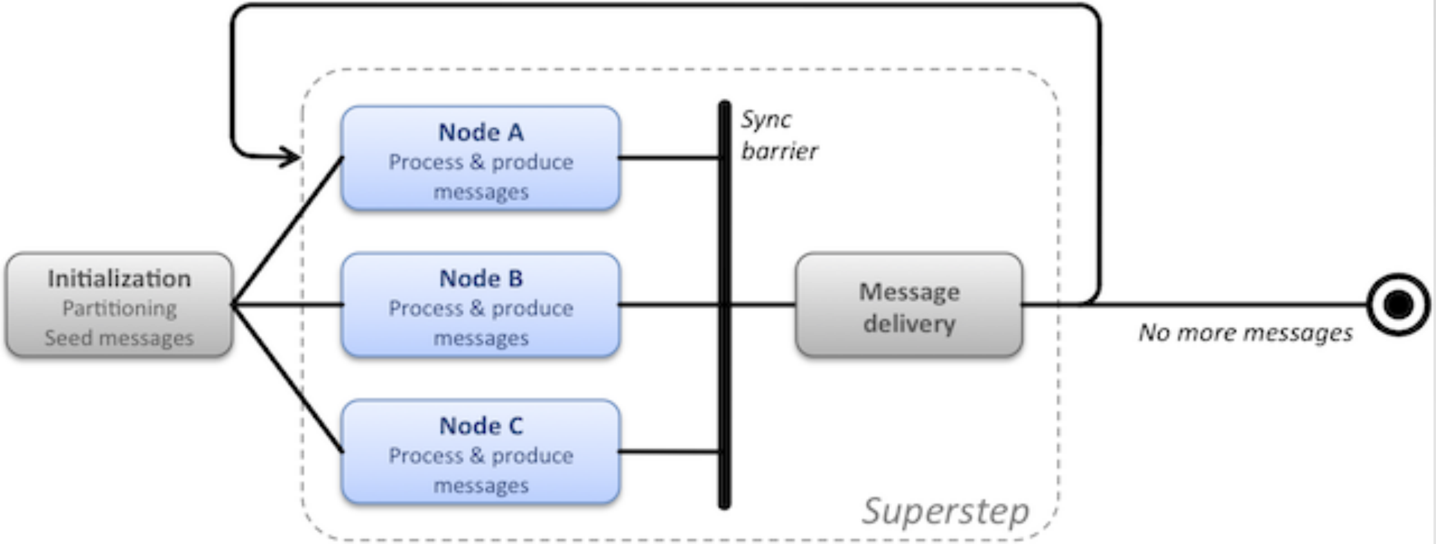


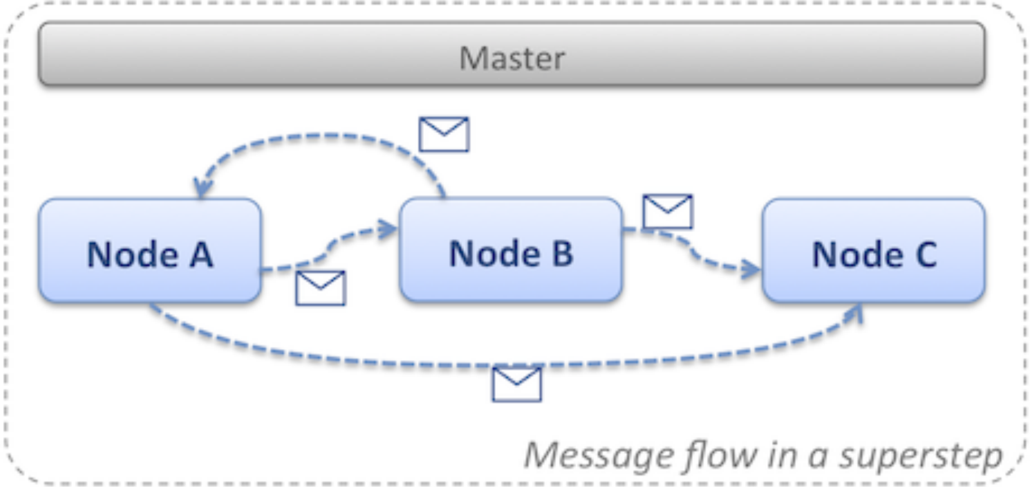
Property Graph

C++ API



```
template <typename VertexValue,
          typename EdgeValue,
          typename MessageValue>
class Vertex {
 public:
  virtual void Compute(MessageIterator* msgs) = 0;

  const string& vertex_id() const;
  int64 superstep() const;

  const VertexValue& GetValue();
  VertexValue* MutableValue();
  OutEdgeIterator GetOutEdgeIterator();

  void SendMessageTo(const string& dest_vertex,
                     const MessageValue& message);
  void VoteToHalt();
};
```


Initialization
Partitioning
Seed messages
Node A
Process & produce messages
Node B
Process & produce messages
Node C
Process & produce messages
Sync barrier
Message delivery
No more messages
Superstep


Master
Node A
Node B
Node C
Message flow in a superstep

Books and papers

- “Mahout in Action”, Owen et. al., Manning Pub.
- “Pattern Recognition and Machine Learning”, Christopher Bishop, Springer Pub.
- “Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Hastie et. al., Springer Pub.
- “Collective Intelligence in Action” Satnam Alag et. al., Manning Pub.

Your questions?