# PivotalR: A Package for Machine Learning on Big Data

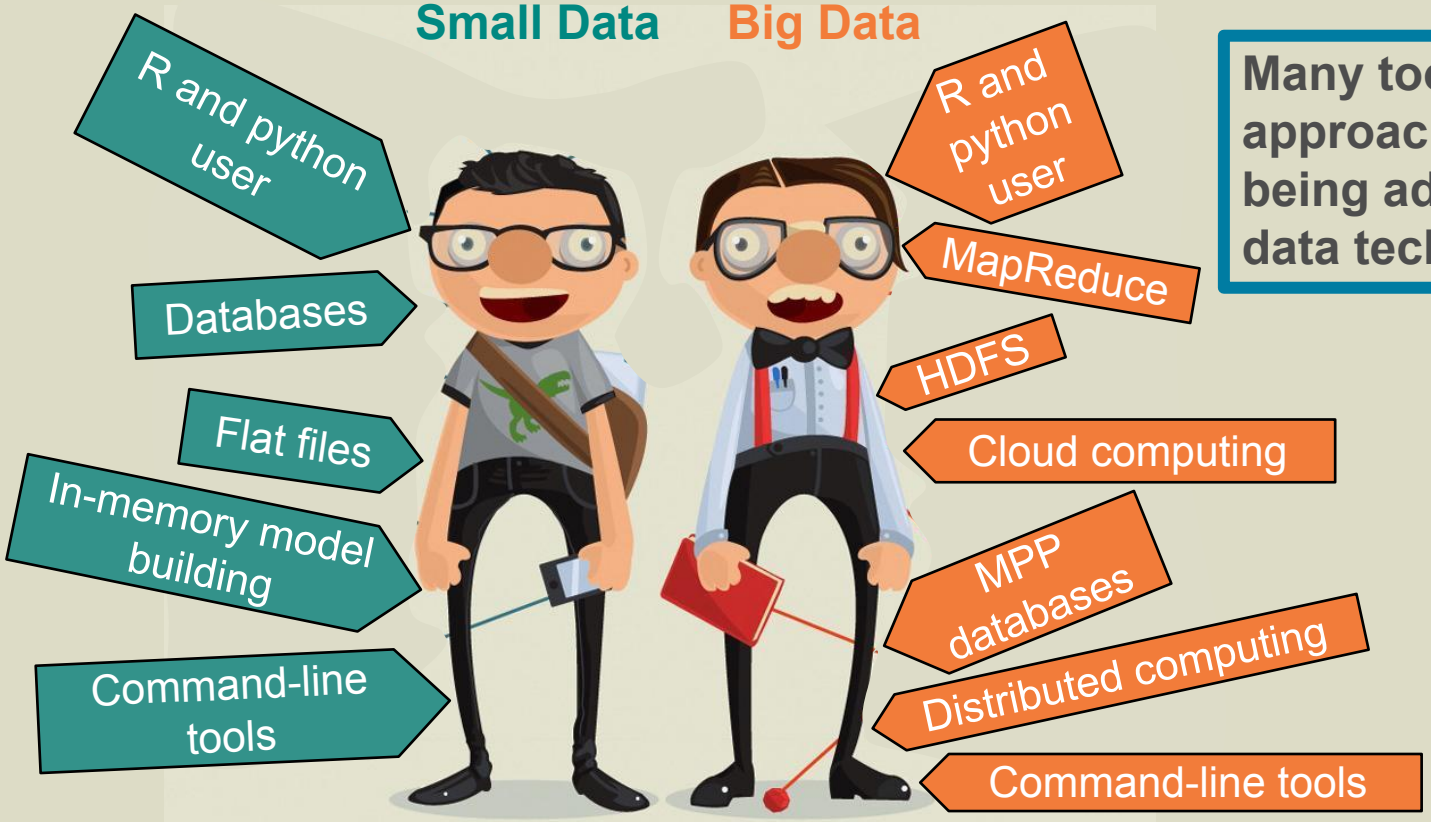
Hai Qian

Predictive Analytics Team, Pivotal Inc.

madlib@gopivotal.com

Pivotal

# What Can “Small Data” Scientists Bring on Their “Big Data” Journey?

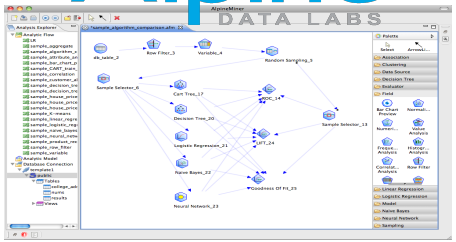



**Many tools and approaches are being adapted to big data technologies**

# Tools for Data Scientists

## COMMERCIAL

Alpine DATA LABS



sas

MATLAB

+ a b l e a u SOFTWARE

## OPEN SOURCE (OR FREE)

GraphLab

MADlib

Spark MLlib

mahout™

PMML Predictive Model Markup Language

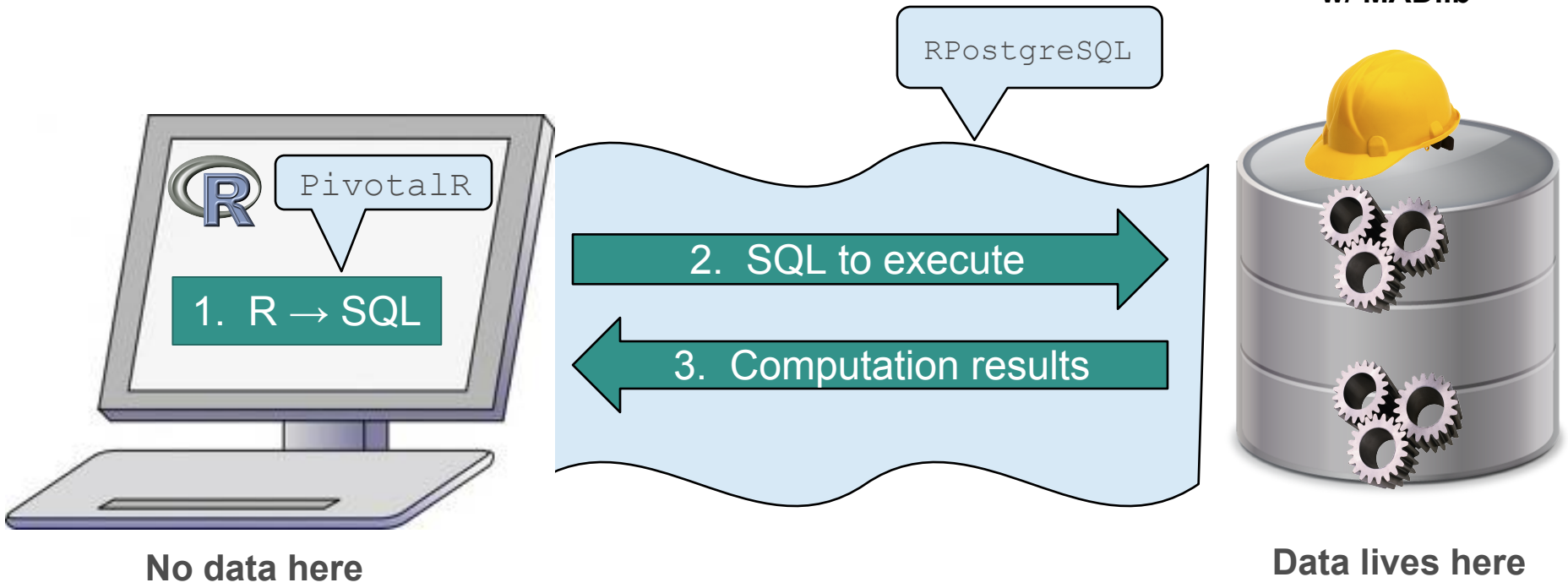
python

PL/R, PL/Python PL/Java

Gephi

Pivotal

# PivotalR Design Overview



# Data Operators

```
crossprod, scale, sample

Arith, Logical, Cast

Extraction: x[,-2], x[,1:3], x[,c("rings", "sex")], x$arr[,
1]

Replacement: x[x$sex == "I",] <- NA
```

- Support array columns
- Easy to construct complicated SQL queries. For example, filter NULL values

```
for (i in 1:ncol(w)) w <- w[!is.na(w[i]),]
```

# Machine learning

- Current MADlib wrapper functions:

  `madlib.lm, madlib.glm, madlib.summary, madlib.arima, madlib.elnet`

- Related functions:

  `generic.cv, generic.bagging, margins, predict`

- Support for formula `y ~ . - x[1:2] - z + factor(w)`

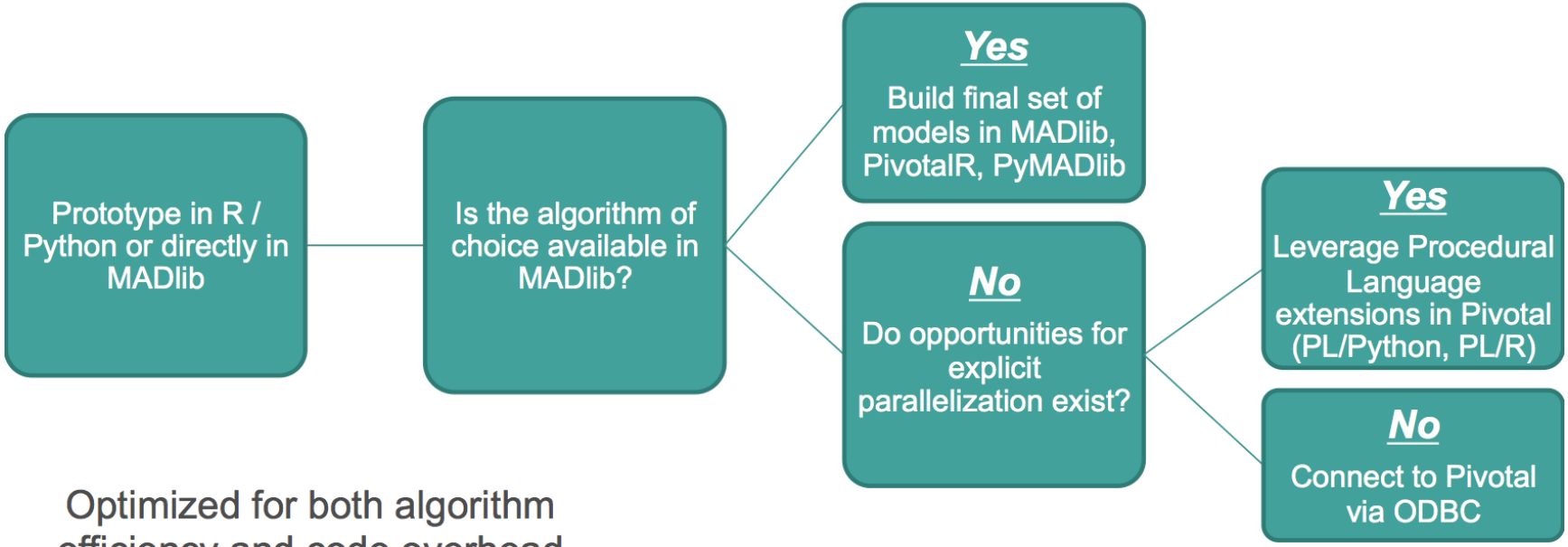
- Support for categorical variables `as.factor, relevel, predict`

Pivotal

# MADlib

## In-database Machine Learning Library

# How Pivotal Data Scientists Select Which Pivotal Tool to Use



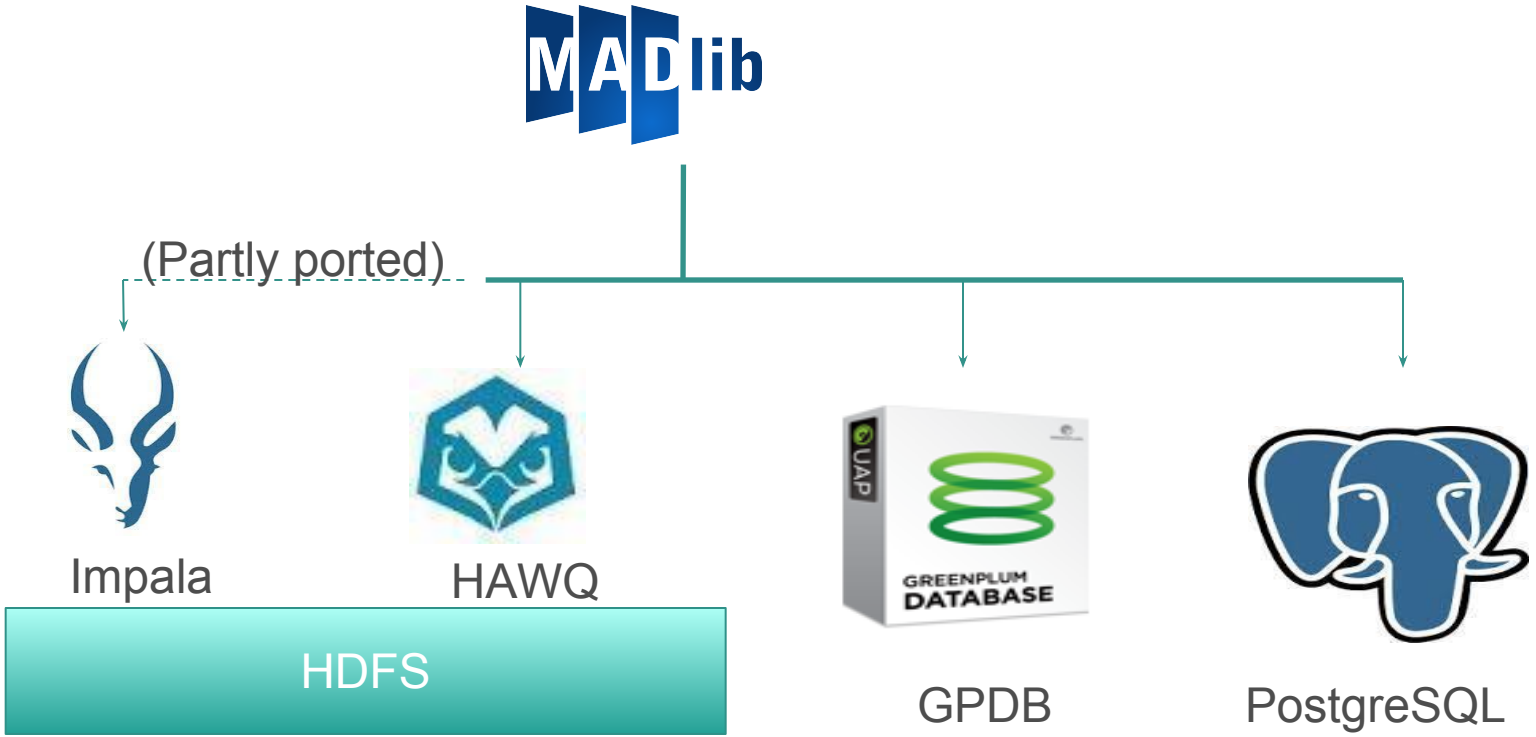

Optimized for both algorithm efficiency and code overhead

Pivotal

# MADlib: Toolkit for Advanced Big Data Analytics

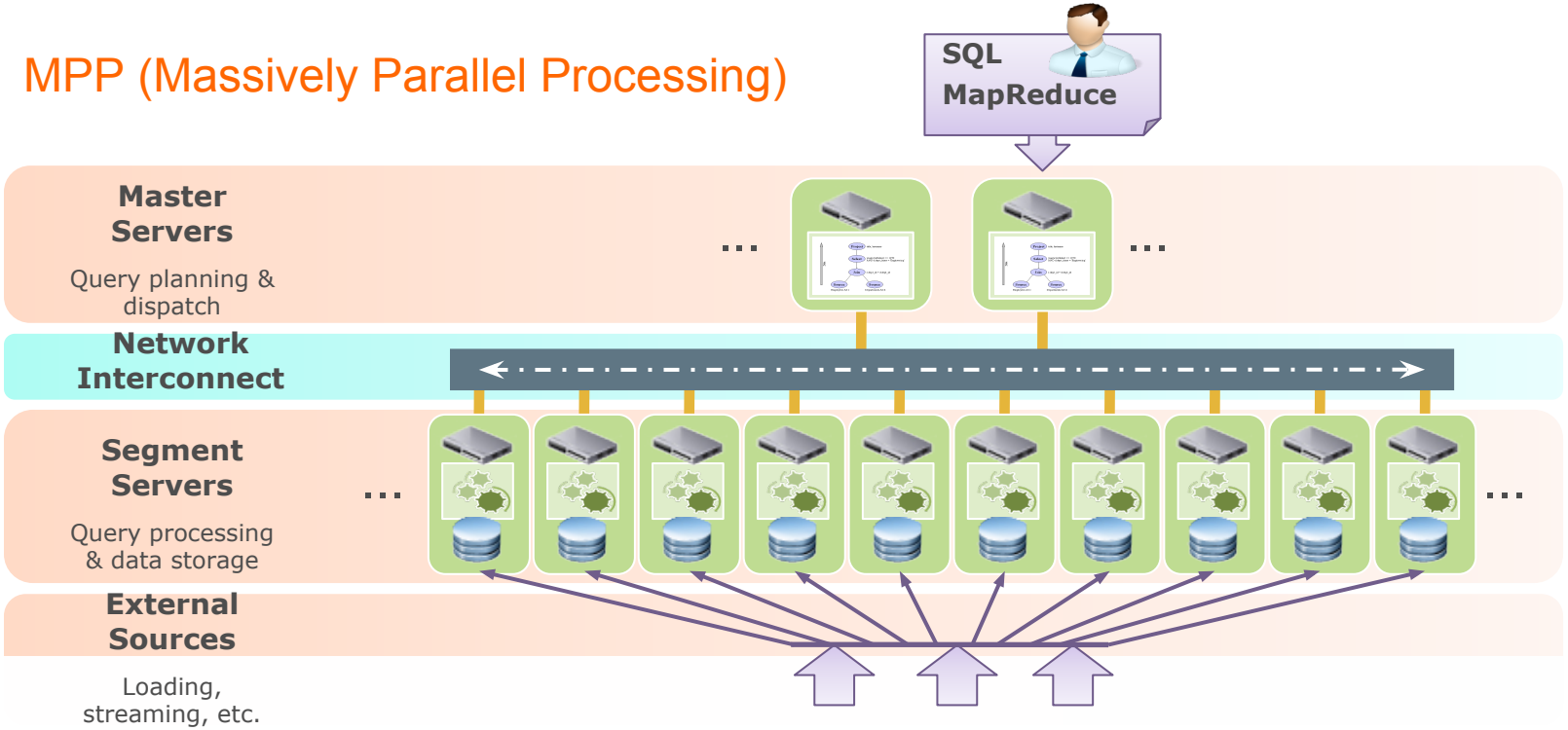

- Better Parallelism
  - Algorithms designed to leverage MPP or Hadoop architecture
- Better Scalability
  - Algorithms scale as your data set scales
  - No data movement
- Better Predictive Accuracy
  - Using all data, not a sample, may improve accuracy
- Open Source
  - Available for customization and optimization by user

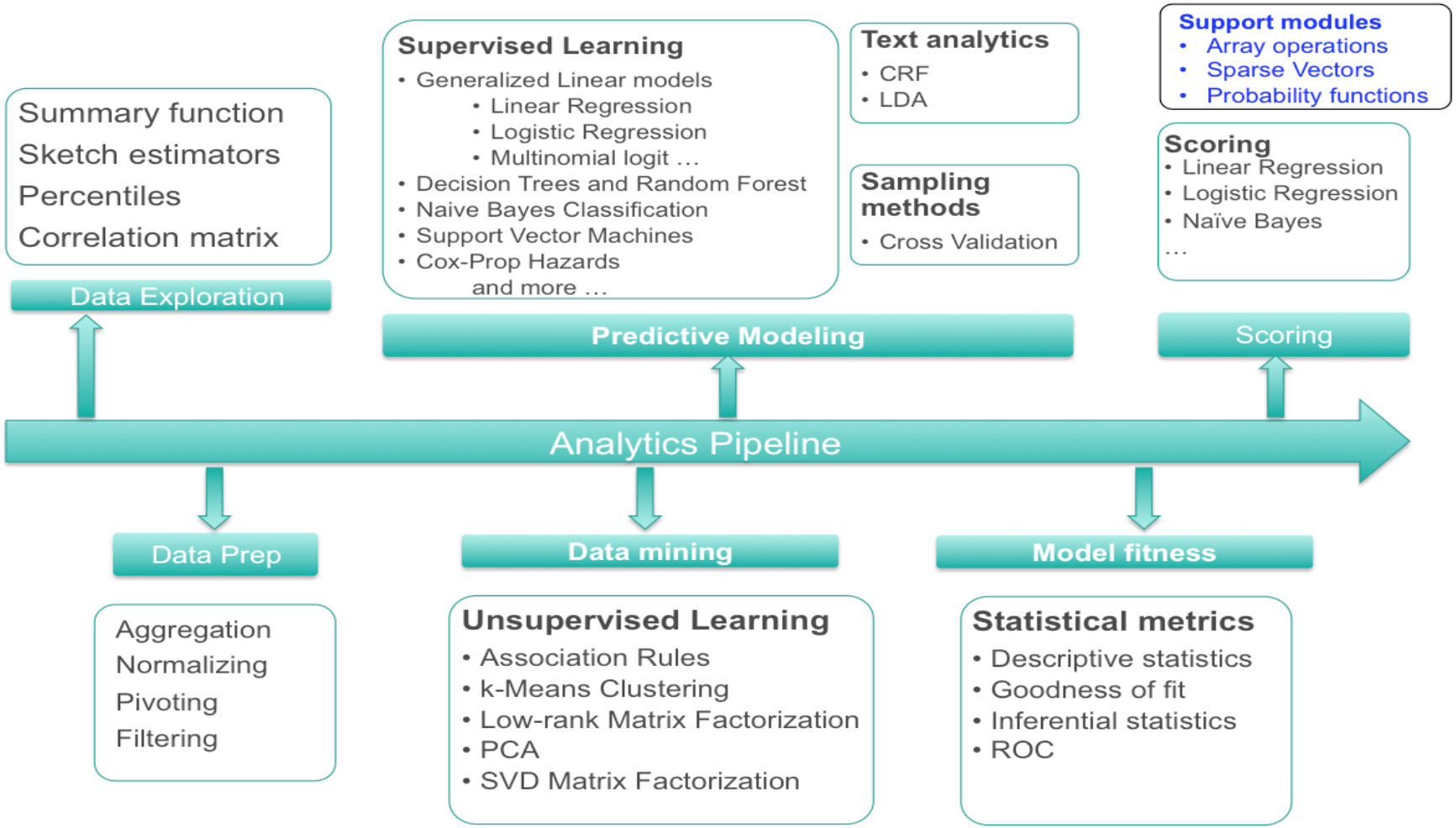
# Which platforms does it run on?



# Shared-Nothing Database Architecture

MPP (Massively Parallel Processing)




Supervised Learning
• Generalized Linear models
• Linear Regression
• Logistic Regression
• Multinomial logit …
• Decision Trees and Random Forest
• Naive Bayes Classification
• Support Vector Machines
• Cox-Prop Hazards
and more …
Text analytics
• CRF
• LDA
Support modules
• Array operations
• Sparse Vectors
• Probability functions
Summary function
Sketch estimators
Percentiles
Correlation matrix
Sampling methods
• Cross Validation
Scoring
• Linear Regression
• Logistic Regression
• Naïve Bayes
…
Data Exploration
Predictive Modeling
Scoring
Analytics Pipeline
Data Prep
Data mining
Model fitness
Aggregation
Normalizing
Pivoting
Filtering
Unsupervised Learning
• Association Rules
• k-Means Clustering
• Low-rank Matrix Factorization
• PCA
• SVD Matrix Factorization
Statistical metrics
• Descriptive statistics
• Goodness of fit
• Inferential statistics
• ROC

# Example usage

Train a model

```
SELECT madlib.linregr_train('houses',                          --- Input table
                            'houses_out',                      --- Output table
                            'price',                           --- Variable to predict
                            'ARRAY[1, tax, bath, size]',       --- Features in data
                            'bedroom'                          --- Group data to create
                                                               ---    multiple models
                           )
```

Predict for new data

```
SELECT houses.*,
    madlib.linregr_predict(ARRAY[1, tax, bath, size],    --- Use same features
                           model.coef)as predict
FROM houses_test, houses_out as model;                   --- Combine test data
                                                         ---    and model table
```

# But not all Data Scientists speak SQL …

## Accessing Scalability through R

# PivotalR: A familiar R interface

Current version 0.1.16.12

Pivotal R

```
d <- db.data.frame("houses")
houses_linregr <- madlib.lm(
     price ~ tax + bath + size, data=d)
```

SQL Code

```
SELECT madlib.linregr_train( 'houses',
                             'houses_linregr',
                                     'price',
                    'ARRAY[1, tax, bath, size]');
```

Pivotal

# Machine learning

- Something that MADlib cannot do
  - `generic.cv`
  - `margins`
- Not easy to do in MADlib on the server side
  - `as.factor` and `relevel`
  - `step`

# Quick Prototype

Examples (see the script):

- Linear regression
- PCA
- Poisson regression
- Left inverse of a matrix
- AdaBoost

**Portable**

Same code on all supported platforms

# Sending R code into Databases

- Write R scripts that be sent into the database
  - No translation to SQL
  - Any R functions

Pivotal

# Testing Framework

```
R CMD INSTALL --install-tests PivotalR_0.1.16.2.tar.gz
```

```
> PivotalR:::test(reporter = 'tap',
+ env.vars=list(.port=5333, .dbname='madlib'),
+ run = 'test')

Running tests -----------------------
1..39
# Context Examples that show how to write tests
ok 1 Examples of speed test
ok 2 Examples of class attributes
ok 3 Examples of class attributes
ok 4 Examples of value equivalent
ok 5 Examples of value equivalent
ok 6 Examples of value equivalent
ok 7 Examples of testing TRUE or FALSE
ok 8 Examples of testing TRUE or FALSE
ok 9 Example of identical
ok 10 Examples of testing string existence
```

```
> PivotalR:::test(reporter = 'summary',
+ env.vars=list(.port=5333, .dbname='madlib'),
+ run = 'example')

Running examples in the user doc ------
Doc example in abalone.Rd : .
Doc example in aggregate-methods.Rd : .
Doc example in aic.Rd : .
Doc example in arith-methods.Rd : .
Doc example in array.len.Rd : .
Doc example in arraydb.to.arrayr.Rd : .
Doc example in as.db.data.frame-methods.Rd : .
Doc example in as.factor-methods.Rd : .
Doc example in by-methods.Rd : .
Doc example in cbind2.Rd : .
Doc example in clean.madlib.temp.Rd :
```

```
> PivotalR:::test(reporter = 'summary',
+ env.vars=list(.port=5333, .dbname='madlib'),
+ run = 'test')

Running tests -----------------------
Examples that show how to write tests : ..................................,,,,,..
```

```
> PivotalR:::test(reporter = 'summary',
+ env.vars=list(.port=5333, .dbname='madlib'),
+ run = 'test', tests.path = "~/workspace/rwrapper/PivotalR/tests/")

Running tests -----------------------
Examples that show how to write tests : ..................................,,,,,..1


1. Failure(@test-examples.r#271): Test MADlib SQL ----------------------------
res not equal to as.numeric(fit$coefficients)
Mean relative difference: 1.308246

Error: Test failures
```

- Based on `testthat`

# Future Work

- Better support of PL/R
- Better graphics support
- Support more platforms

Pivotal

# Additional References

- MADlib
  - http://madlib.net/
  - http://doc.madlib.net/latest/
- PivotalR
  - http://cran.r-project.org/web/packages/PivotalR/PivotalR.pdf
  - https://github.com/gopivotal/PivotalR
  - Video Demo

Pivotal