BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Unstructured Data Analytics – Vegetation Response Prediction

**Presenters:**

- **Andy Lathrop**, Principal Consultant

- **Colby Ford**, Data Scientist

---

**Contributors:**

- **Josh Fennessy**, Principal Architect

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Purpose

Based on published work:

**Statistical forecasting of vegetation response to manage natural risks in forested areas.**

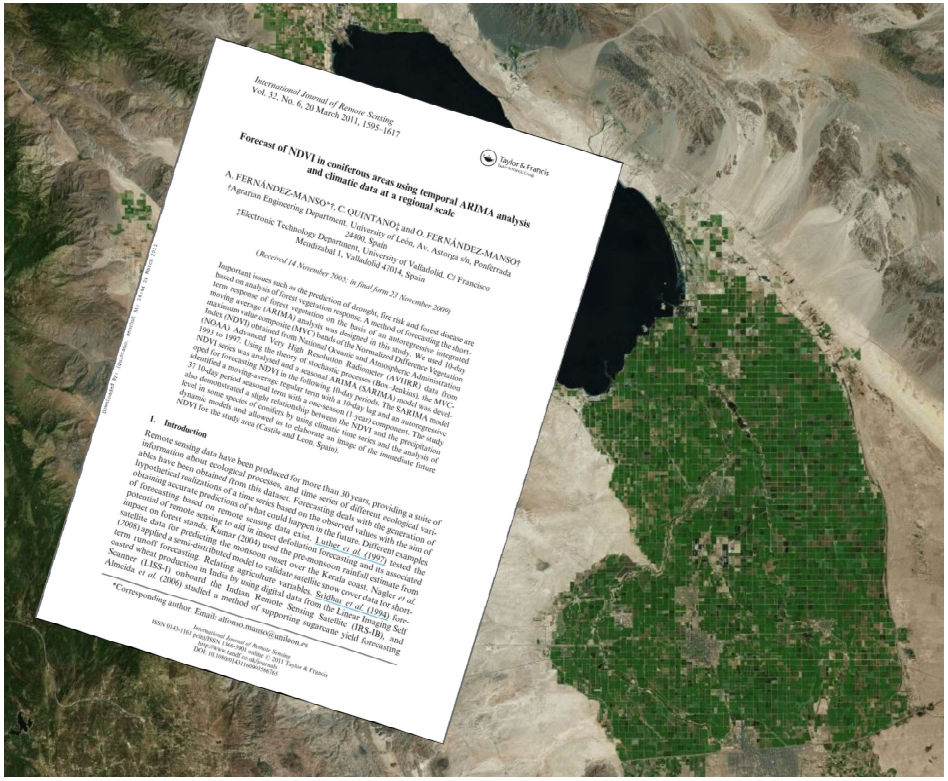
**Our Project: similar … but different…**

- Big Data:

  Data Shaping

  Modeling

- Development process for distributed team




**"Forecast of NDVI in coniferous areas using temporal ARIMA analysis and climatic data at a regional scale"**
*- International Journal of Remote Sensing,*
March 2011

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

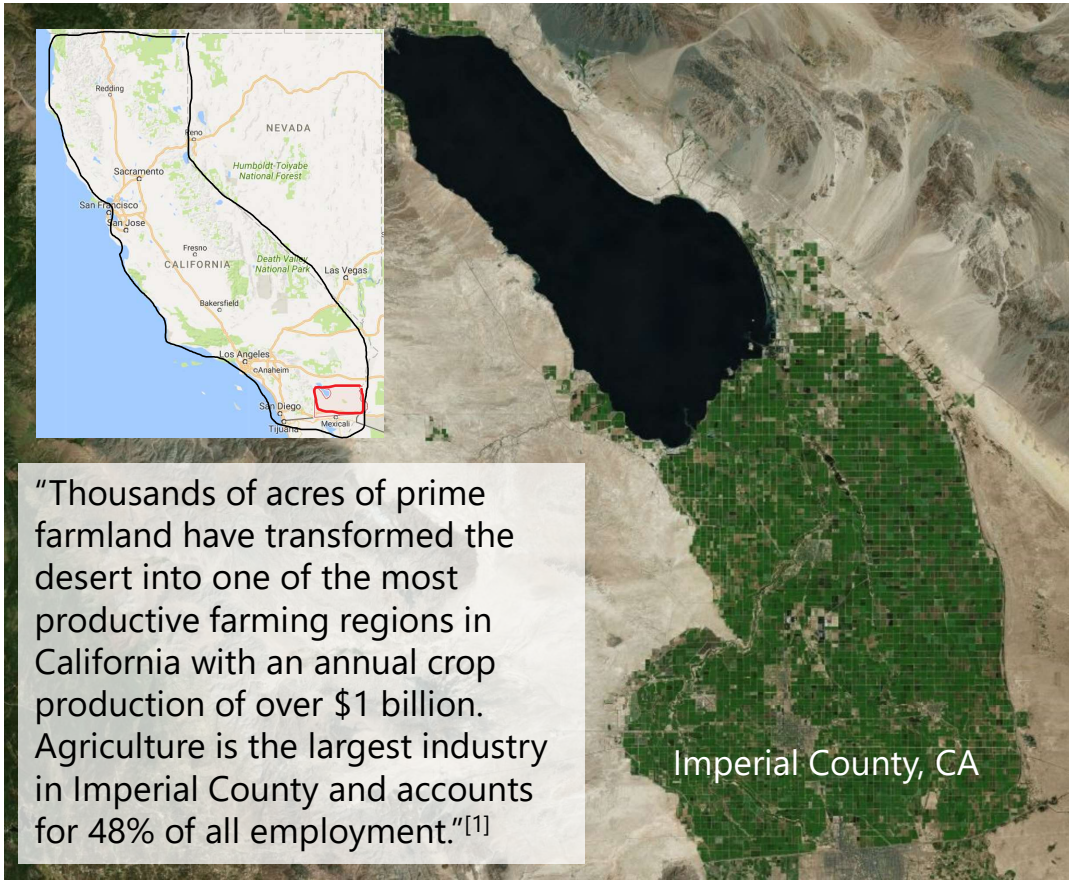
# Social Good

## Agricultural Economics

- The Agriculture industry provides jobs for **787,000** individuals in the US.[2]
- "**2.2 million** farms dot America's rural landscape. About 97 percent of U.S. farms are operated by families – individuals, family partnerships or family corporations."[3]
- Contributed **$985 billion** to the U.S. gross domestic product in 2014.[4]

## Environmental and Risk Management

- Better natural resource management by understanding relationship between crop cycles and climate
- Identify risk situations due to water stress of vegetation





[1] Din, F. A. (2010, May 19). Welcome to El Centro, California. Retrieved February 07, 2017, from http://activerain.com/blogsview/99241/welcome-to-el-centro--California

[2] US Department of Agriculture. (2016, September 27). USDA: Farm Labor - Background. Retrieved February 10, 2017, from https://www.ers.usda.gov/topics/farm-economy/farm-labor/background.aspx

[3] American Farm Bureau Federation. (2017). Fast Facts About Agriculture. Retrieved February 10, 2017, from http://www.fb.org/newsroom/fast-facts

[4] USDA - Economic Research Service. (2016, October 14). Ag and Food Sectors and the Economy. Retrieved February 10, 2017, from https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/ag-and-food-sectors-and-the-economy.aspx

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Social Good, continued

**Forecasting with remote sensing data**

- Insect defoliation and associated impact on forest stands
- Predicting precipitation coverage from monsoon onset / snow cover
- Forecast crop production (wheat and sugar cane)

**Other Environmental Applications of Time series forecasting / Multivariate analysis**

- Anomaly detection to determine fire risk or drought conditions
- Patterns in climate change
- Predict vegetation status for crops and grasslands

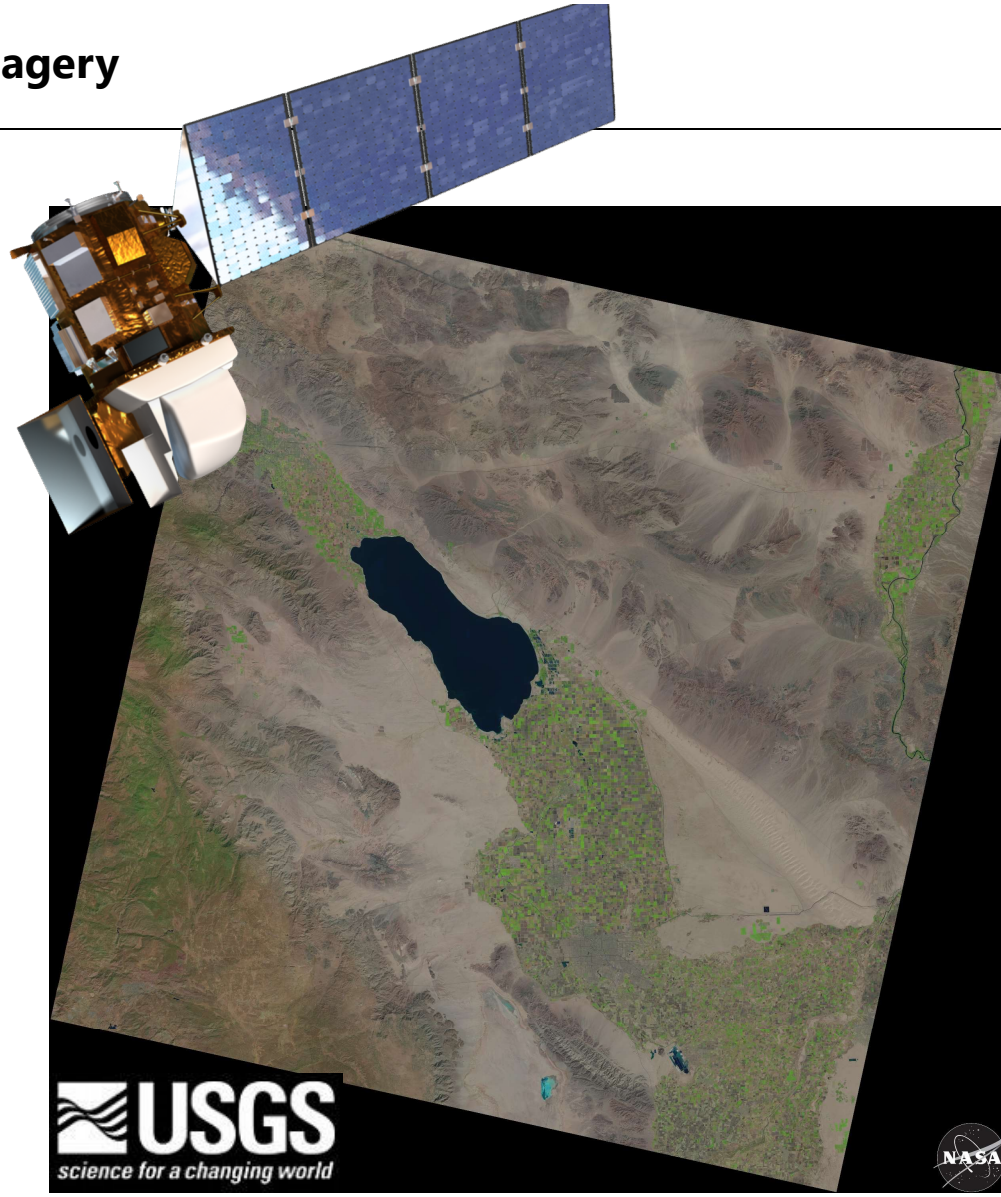
BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Satellite Imagery

## LandSat8:

- Images the entire Earth every 16 days
- Captures 11 bands of images
- ~7600px$^2$ or 1km$^2$ sections
- ~1GB compressed tarball files containing .GeoTiff images and a metadata text file

| Band Number | µm | Description |
|---|---|---|
| 1 | 0.433–0.453 | Coastal Aerosol |
| 2 | 0.450–0.515 | Blue |
| 3 | 0.525–0.600 | Green |
| 4 | 0.630–0.680 | Red |
| 5 | 0.845–0.885 | Near Infrared |
| 6 | 1.560–1.660 | SWIR 1 |
| 7 | 2.100–2.300 | SWIR 2 |
| 8 | 0.500–0.680 | Panchromatic |
| 9 | 1.360–1.390 | Cirrus |
| 10 | 10.6-11.2 | Thermal Infrared |
| 11 | 11.5-12.5 | Thermal Infrared |



BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

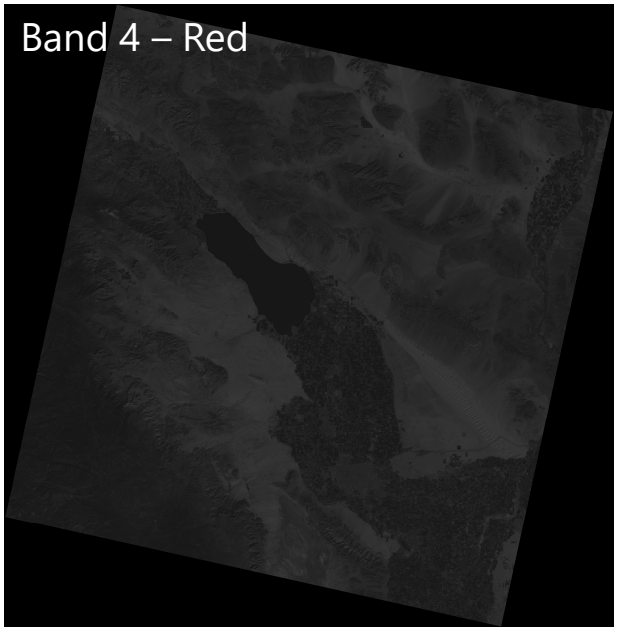
# Calculating NDVI

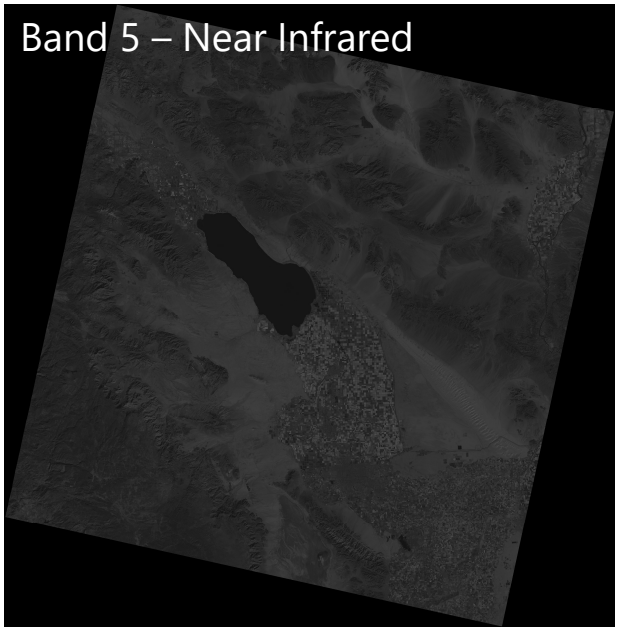
$$NDVI = \frac{(B_5 - B_4)}{(B_5 + B_4)}$$

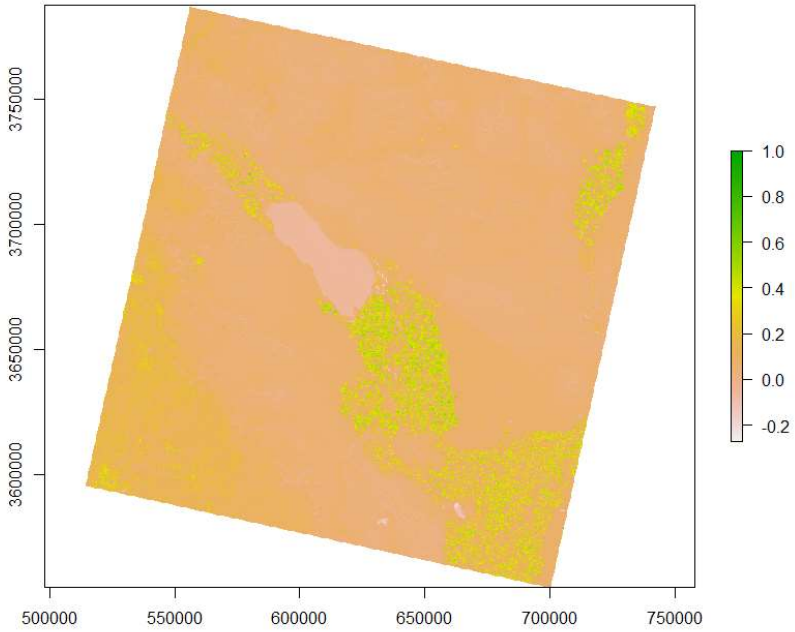
## Normalized Difference Vegetation Index:

- Developed by researchers at NASA
- Ratio of two bands of the satellite images
- Measures light absorption of the chlorophyll in vegetation
- Range: -1 to +1

| Ground Cover | Red | Near Infrared | NDVI |
|---|---|---|---|
| **Dense vegetation** | 0.1 | 0.5 | **0.7** |
| **Dry Bare soil** | 0.269 | 0.283 | **0.025** |
| **Clouds** | 0.227 | 0.228 | **0.002** |
| **Snow and ice** | 0.375 | 0.342 | **-0.046** |
| **Water** | 0.022 | 0.013 | **-0.257** |



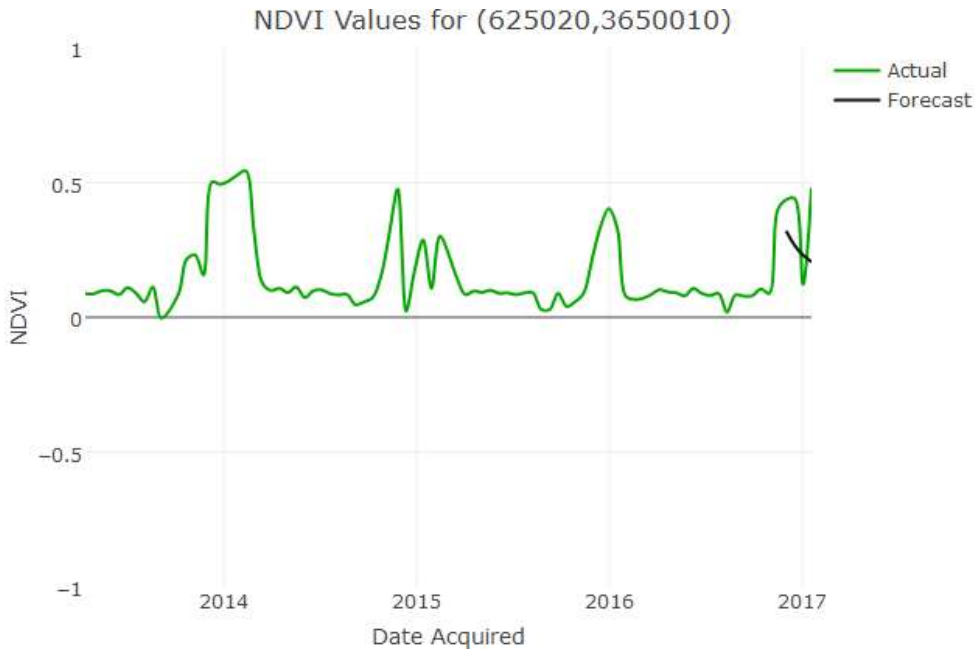



BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Predicting NDVI using ARIMA

## Autoregressive Integrated Moving Average:

- Algorithm to forecast time series data
- Requires two variables: a time series variable and a numerical value variable
- Developed in the 1990's for economics




| Pixel: (625020,3650010) | |
|---|---|
| DATE_ACQUIRED | NDVI |
| 2013-04-13 | 0.08826414 |
| 2013-04-29 | 0.08710080 |
| … | … |
| 2017-01-02 | 0.02909486 |
| 2017-01-18 | 0.02919372 |

**Current Performance:**

**~49.7%**

**Mean Absolute Percentage Error**

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Fast Facts

1.7GB each

**85** Image Sets

**3 Files**
Only use Bands 4 & 5 + Metadata (~250MB)

7500x7500 TIF Files
= 56,250,000 Pixels of data

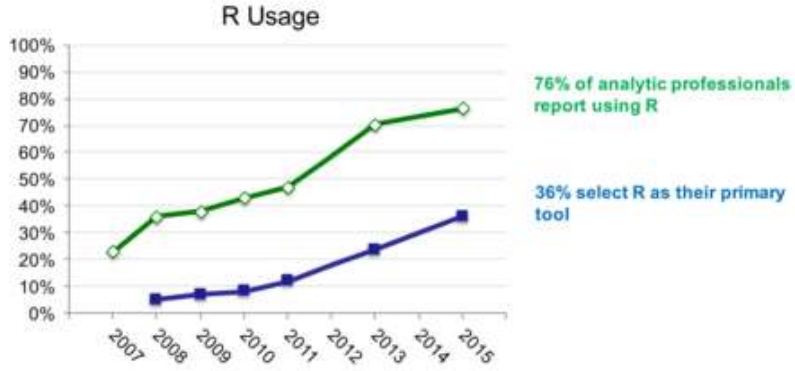
× 85 ≈ **4.8 BILLION Records**

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

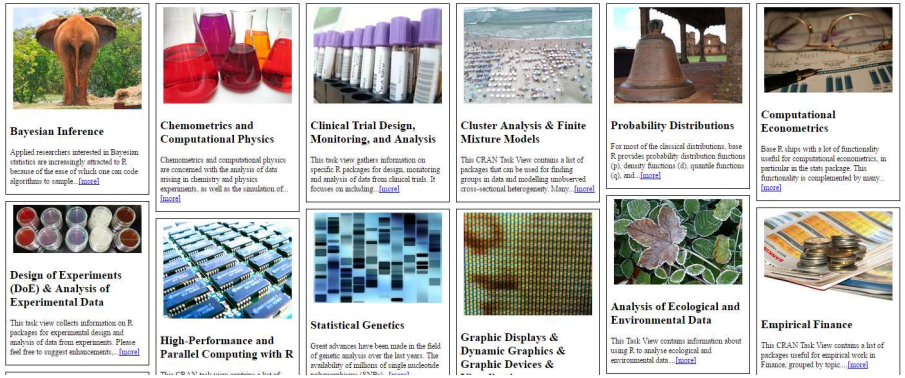
# Introduction to R

- 1993: Created by Ross Ihaka and Robert Gentlemen in Auckland, NZ as an open-source implementation of the S programming language
- Most widely used data analysis software
  #1 for data science; #6 general purpose (IEEE Spectrum Rankings)
- Common programming language of analytics and statistical computing
- Unique and immersive data visualizations
- Open-source, extensible, scalable
  library of 7500+ add-on packages; community of millions of users

## R Popularity is Growing!




## CRAN Packages



BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

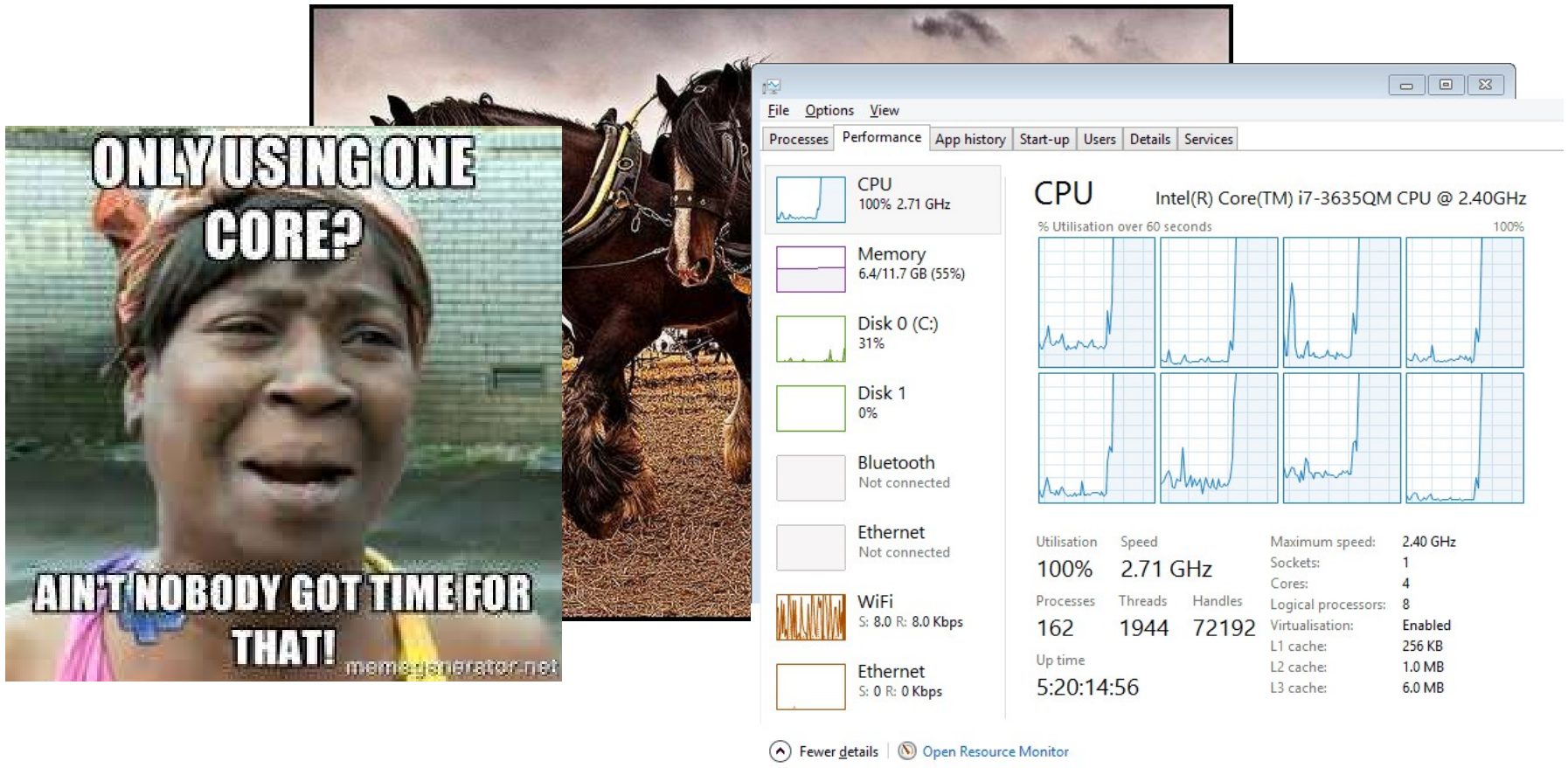
# Microsoft R Product Family

## CRAN, MRO, MRS Comparison

| |  |  Microsoft R Open |  Microsoft R Server |
|---|---|---|---|
| **Datasize** | In-memory | In-memory | **In-Memory or Disk Based** |
| **Speed of Analysis** | Single threaded | Multi-threaded | **Multi-threaded, parallel processing 1:N servers** |
| **Support** | Community | Community | **Community + Commercial** |
| **Analytic Breadth & Depth** | 7500+ innovative analytic packages | 7500+ innovative analytic packages | **7500+ innovative packages + commercial parallel high-speed functions** |
| **License** | Open Source | Open Source | **Commercial license. Supported release with indemnity** |

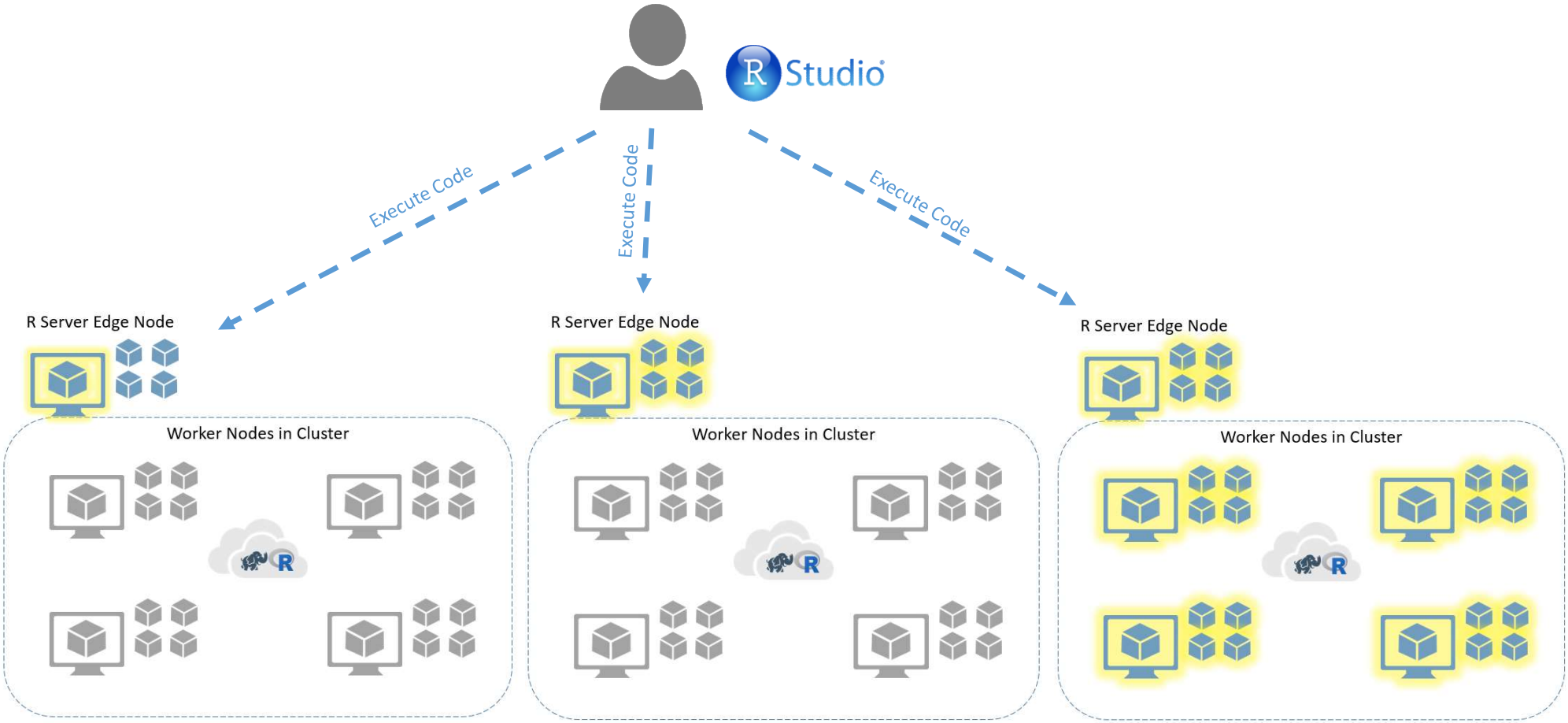
BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Parallel Computing with R

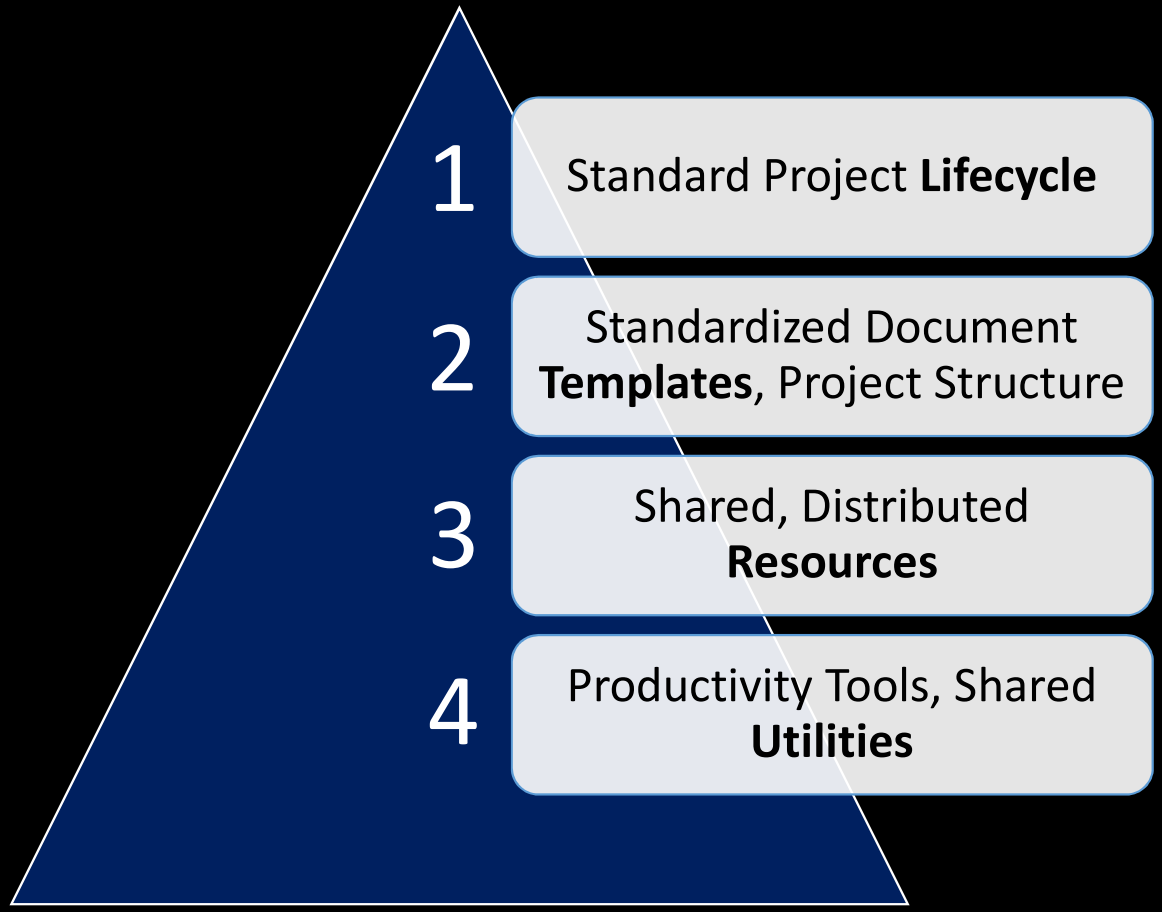


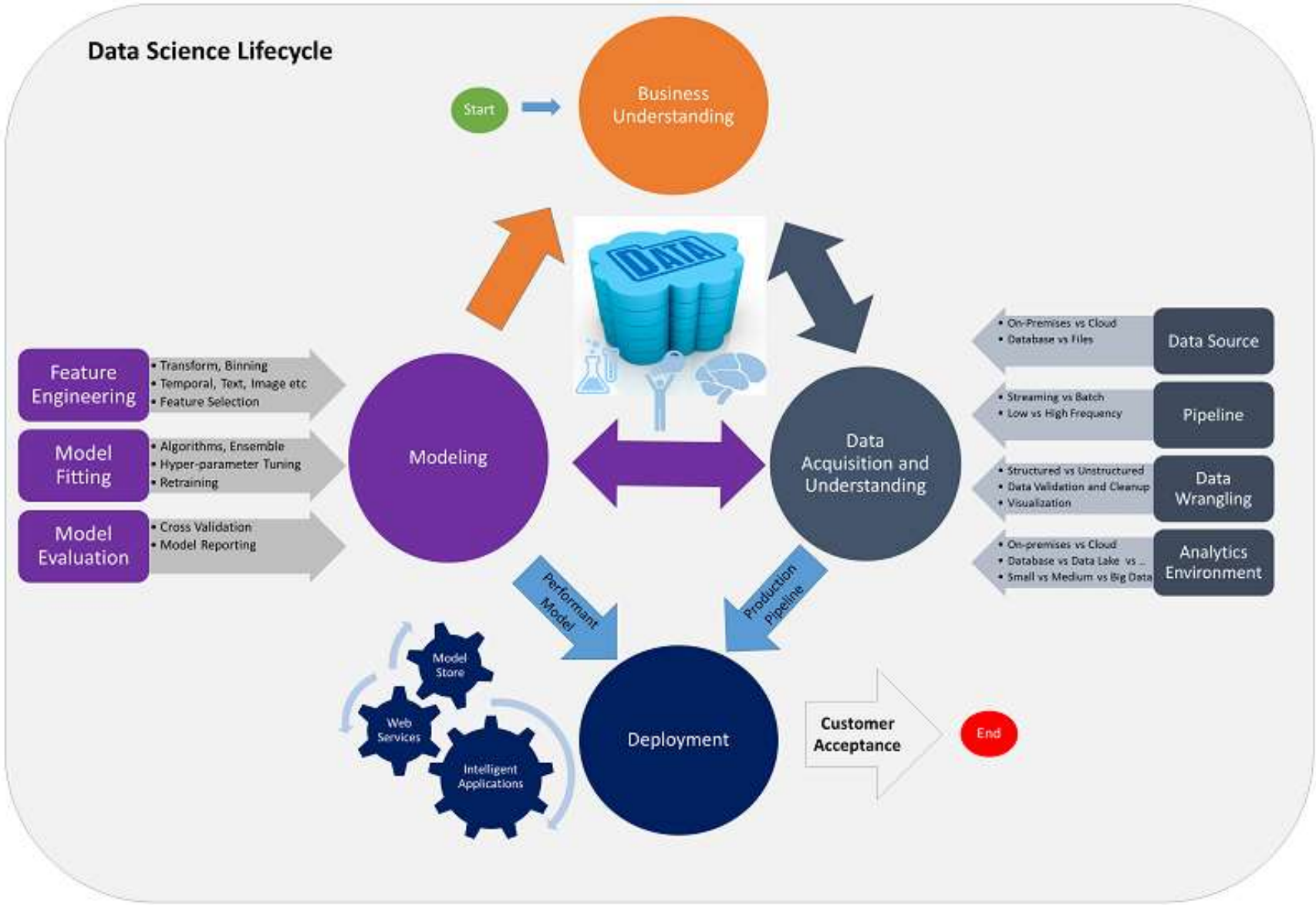
BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Execution Contexts in HDInsight with R Server



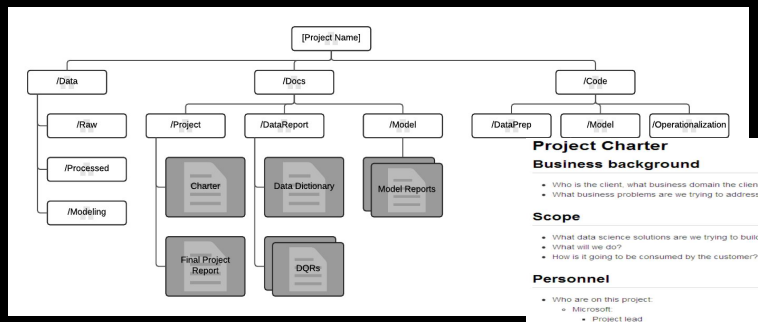
# 4 Pillars of Our Team Data Science Process

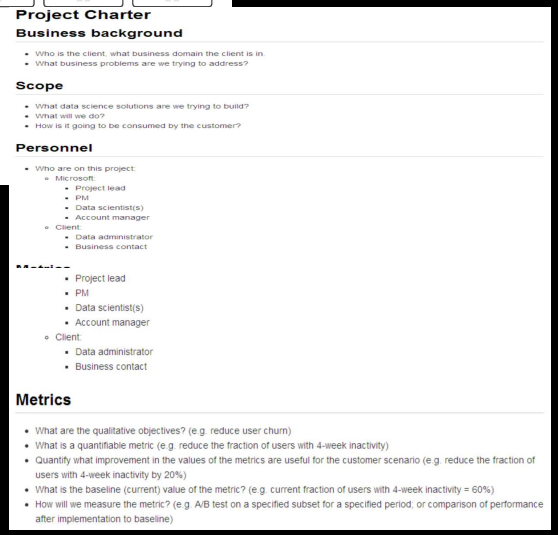



Data Science Lifecycle
Start
Business Understanding
DATA
Feature Engineering
• Transform, Binning
• Temporal, Text, Image etc
• Feature Selection
Model Fitting
• Algorithms, Ensemble
• Hyper-parameter Tuning
• Retraining
Model Evaluation
• Cross Validation
• Model Reporting
Modeling
Data Acquisition and Understanding
• On-Premises vs Cloud
• Database vs Files
Data Source
• Streaming vs Batch
• Low vs High Frequency
Pipeline
• Structured vs Unstructured
• Data Validation and Cleanup
• Visualization
Data Wrangling
• On-premises vs Cloud
• Database vs Data Lake vs ..
• Small vs Medium vs Big Data
Analytics Environment
Performant Model
Production Pipeline
Model Store
Web Services
Intelligent Applications
Deployment
Customer Acceptance
End

# Standardized Artifacts, Shared Productivity Utilities

## Templates



**Project Charter**

**Business background**

- Who is the client, what business domain the client is in.
- What business problems are we trying to address?

**Scope**

- What data science solutions are we trying to build?
- What will we do?
- How is it going to be consumed by the customer?

**Personnel**

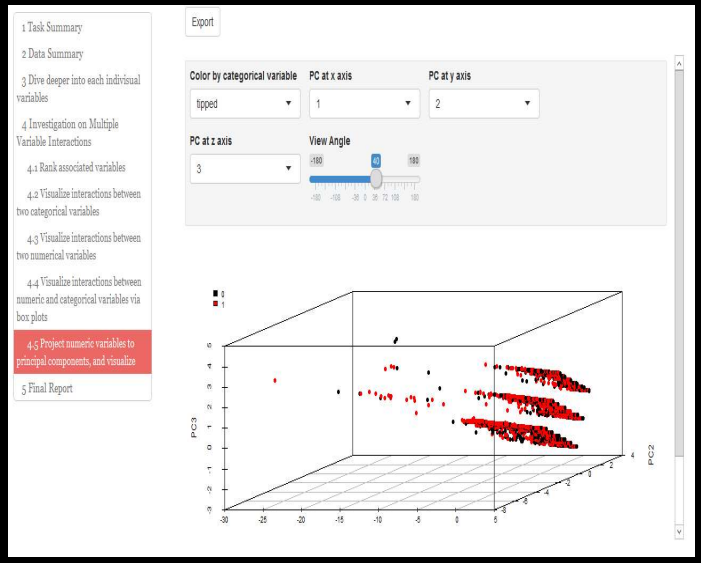
- Who are on this project:
  - Microsoft:
    - Project lead
    - PM
    - Data scientist(s)
    - Account manager
  - Client:
    - Data administrator
    - Business contact

- Project lead
- PM
- Data scientist(s)
- Account manager
- Client:
  - Data administrator
  - Business contact

**Metrics**

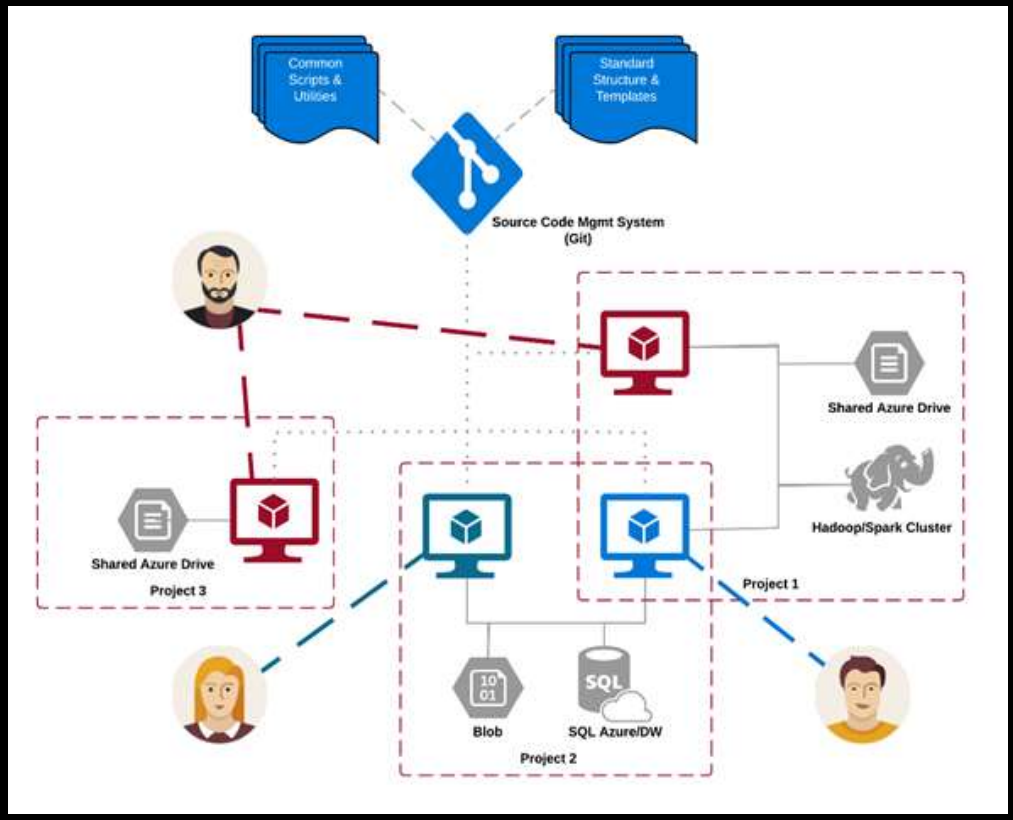
- What are the qualitative objectives? (e.g. reduce user churn)
- What is a quantifiable metric (e.g. reduce the fraction of users with 4-week inactivity)
- Quantify what improvement in the values of the metrics are useful for the customer scenario (e.g. reduce the fraction of users with 4-week inactivity by 20%)
- What is the baseline (current) value of the metric? (e.g. current fraction of users with 4-week inactivity = 60%)
- How will we measure the metric? (e.g. A/B test on a specified subset for a specified period; or comparison of performance after implementation to baseline)

## Utilities



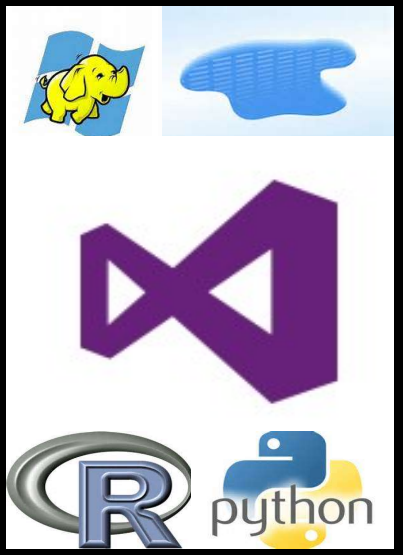
# Team Data Science Process – Implementation Overview

- Data science virtual machines (DSVMs) as the fundamental development platform on cloud
- Use Visual Studio Team Services (VSTS)
  - Work item tracking and scrum planning
  - Git repositories
- Every project has a single git repository with standardized directory structure and standardized document templates
- Shared data science utilities in Git repository



# Data Science Virtual Machine: What's included?








BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Microsoft Teams

## Microsoft Teams

- Messaging and Code Snippet Sharing
- Git Repository Updates
- Checklists
- Document Collaboration
- Most Importantly, *memes*…

Visual Studio Team Services 2/7 12:11 AM

**CODE** | Andy Lathrop pushed a commit on refs/heads/master branch

Repository: Unstructured Data Demo

Commit id 1f520feda6a7f0d55fa030dcc8c4ee08a68aaab5

Commit comment updates to copy and processing scripts to validate matching metadata, band 4, and band 5 files. Include parallel code in processing script. switching to 20 cores!

View commit

Reply

Andy Lathrop 2/7 1:47 AM Edited

**Colby** heads up there are a handful of TIF files that won't convert - I'm guessing the're corrupt or something. Can you be prepared to do some missing value substitution for the time series? I should be able to get you the full (85 minus a few missing values) filtered data set tomorrow a.m. I upgraded to a 20-core DSVM so things are flying!

10 replies from you and Andy

Colby Ford 2/7 4:17 PM 1

Andy Lathrop I've re-downloaded these files and they are now in the Lake. Let me know if you run into any more that need to be downloaded again!

Visual Studio Team Services 2/8 5:23 PM

CF

**CODE** | Colby Ford pushed a commit on refs/heads/master branch

Repository: Unstructured Data Demo

Commit id d7142ec6e3c782a1e2fa1ebc7628708537b66a74

Commit comment Rewrote ARIMA functions to work with xdf's, but without using dplyr.

View commit

Reply


High level data science tasks

- [ ] Gather data - GeoTIFF and weather - ba...
- [ ] Data ingestion - Azure data lake and M...
- [ ] EDA - Tidy, transform, visualize
- [ ] Model - ARIMA using MRS on HDInsight
- [ ] Visualize Data and Results

0 / 5

## Research

GeoLocation of Image:
Lat: 33.25
Long: -115.5

**NVDI - Normalized Differential Vegetation Index**

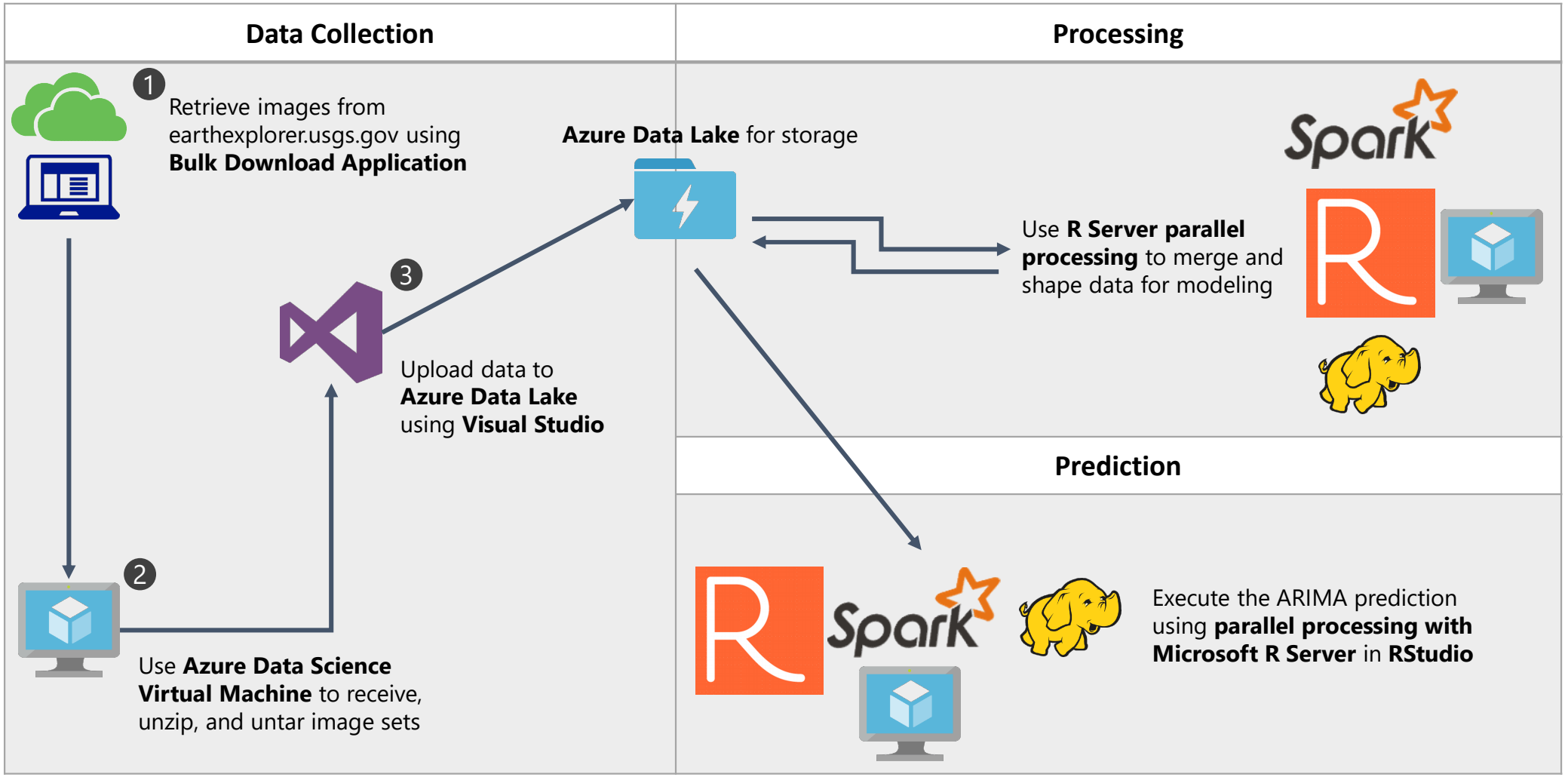
NDVI is a measurement of the amount of vegetation visible in an image – see here for cal
(http://grindgis.com/blog/vegetation-indices-arcgis#3)

**NDVI Values close to 0 means low vegetation, close to 1 means high**

Output of the NDVI method creates a single-band dataset that only shows greenery. Valu
rock and bare soil and negative values represent water, snow and clouds. Taking ratio or d

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Analysis Pipeline



BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Fast Facts

1.7GB each

**85** Image Sets

**3 Files**
Only use Bands 4 & 5 + Metadata (~250MB)

7500x7500 TIF Files
= 56,250,000 Pixels of data

× 85 ≈

**4.8 BILLION Records**

Single core

**13.7 hours**

Multi-core (20)

**3.7 hours**

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Resources

**Software**

- Microsoft R Core
- Microsoft R Client
- RStudio

**Students**

- Microsoft Azure for Students
- Microsoft Software for Students

Microsoft Imagine

**Presentation Materials & Sample Code**

GitHub.com/BlueGranite/Unstructured-Data-Demo

**BlueGranite - Blue-Granite.com**

- Hands-On Labs
- Advanced Analytics Blog Posts
- Microsoft R and Revolution Analytics Webinar


**Andy Lathrop**
Principal Consultant

- Linkedin.com/in/ALathrop

**Colby Ford**
Data Scientist

- Linkedin.com/in/ColbyFord
- ColbyFord.com

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Additional Information

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Microsoft R: Write Once, Deploy Anywhere

| Available in these Platforms | |
|---|---|
| **R Server for Red Hat Linux** | 64-Bit Red Hat Enterprise Linux (or CentOS) 5.x or 6.x |
| **R Server for SUSE Linux** | 64-Bit SUSE Linux Enterprise Server 11 SP2 or SP3. |
| **R Server for Hadoop on Red Hat** | **Hadoop Distributions / Operating Systems:**<br>Cloudera CDH 5.0-5.4 on RHEL 6.x<br>Hortonworks HDP 2.0-2.3 on RHEL 6.x<br>MapR M3/5/7 3.x, 4.0-4.1 on RHEL 6.x |
| **R Server for Teradata DB** | Teradata Database 14.10, 15.00, 15.10 on SLES 10.x or 11.x |

Microsoft R Server





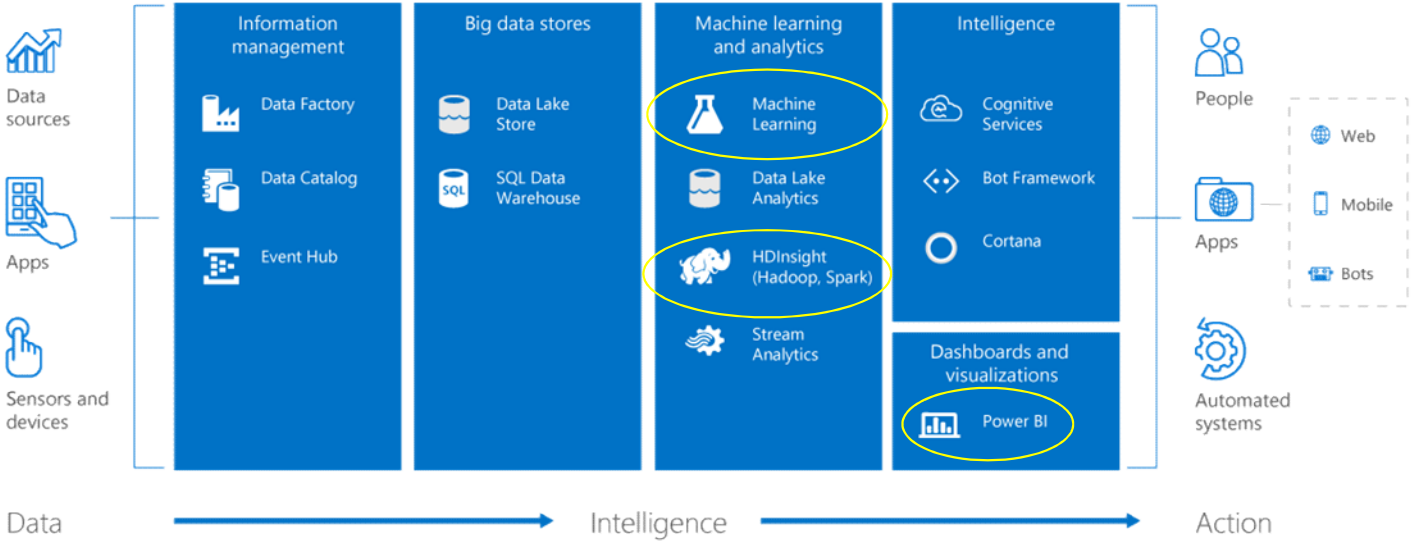
BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Deployment Environments

## Cloud

### Cortana Intelligence Suite

A suite of products that allow you to **Predict Outcomes, Prescribe Actions** and **Automate Decisions** for **Operationalized Solutions**




## On-Premises

Microsoft SQL Server 2016

| Available in these Platforms | |
|---|---|
| **R Server for Red Hat Linux** | 64-Bit Red Hat Enterprise Linux (or CentOS) 5.x or 6.x |
| **R Server for SUSE Linux** | 64-Bit SUSE Linux Enterprise Server 11 SP2 or SP3. |
| **R Server for Hadoop on Red Hat** | **Hadoop Distributions / Operating Systems:**<br>Cloudera CDH 5.0-5.4 on RHEL 6.x<br>Hortonworks HDP 2.0-2.3 on RHEL 6.x<br>MapR M3/5/7 3.x, 4.0-4.1 on RHEL 6.x |
| **R Server for Teradata DB** | Teradata Database 14.10, 15.00, 15.10 on SLES 10.x or 11.x |

Microsoft R Server

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# SQL Server R Services

## Development

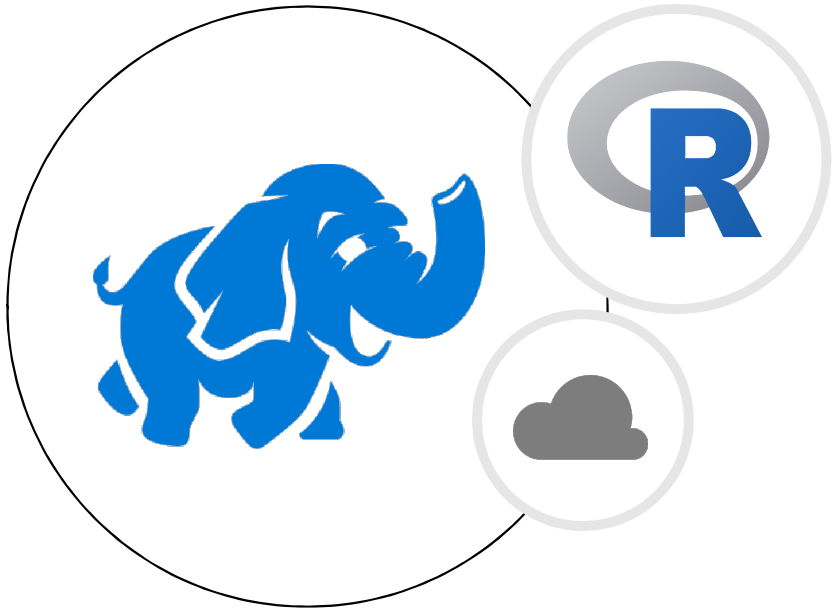
```
sql <- RxInSqlServer([SQL connection])
rxSetComputeContext(sql)

[…]
I/O → RxSqlServerData(), etc.
Stats → rxSummary(), etc.
Models → rxLogit(), etc.
Plots → rxRocCurve(), etc.
```


## Deployment

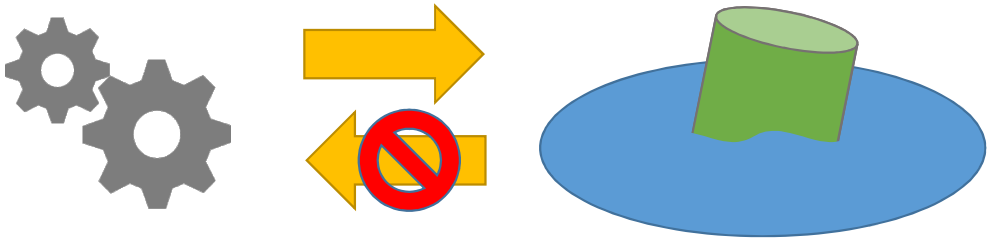
```
EXEC sp_execute_external_script
   @language = N'R',
   @script = N'[R code goes here]',
   @input_data_1 = N'[SQL input]'
   [ , @input_data_1_name = N'InputDataSet' ]
   [ , @output_data_1_name = N'OutputDataSet' ]
   [ , @params = N'parameter' ]
WITH RESULT SETS ((SQL output]));
```

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

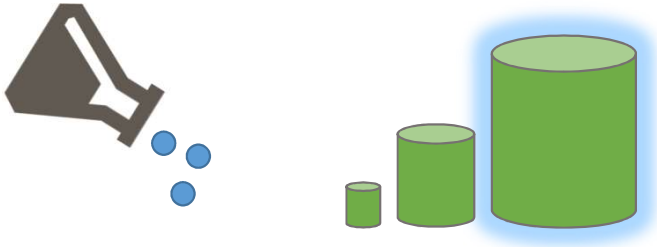
# R Server for HDInsight



**HDInsight with R Server**



Bring your compute to the data in the data lake,
not the other way around!



Machine Learning on Terabytes of Data

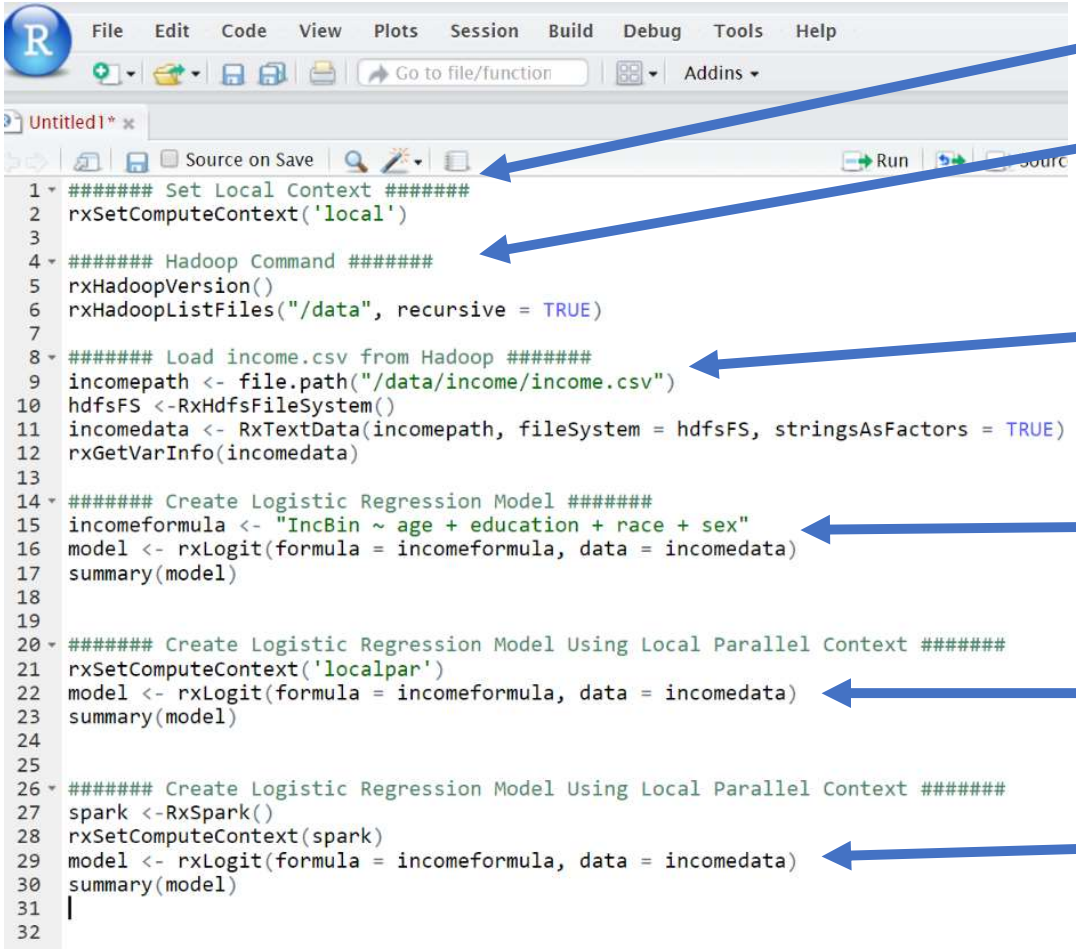
RStudio

Use familiar IDEs like R Studio


No Hadoop experience? No Problem!

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Quick HDInsight with R Server Example




```
1  ####### Set Local Context #######
2  rxSetComputeContext('local')
3
4  ####### Hadoop Command #######
5  rxHadoopVersion()
6  rxHadoopListFiles("/data", recursive = TRUE)
7
8  ####### Load income.csv from Hadoop #######
9  incomepath <- file.path("/data/income/income.csv")
10 hdfsFS <-RxHdfsFileSystem()
11 incomedata <- RxTextData(incomepath, fileSystem = hdfsFS, stringsAsFactors = TRUE)
12 rxGetVarInfo(incomedata)
13
14 ####### Create Logistic Regression Model #######
15 incomeformula <- "IncBin ~ age + education + race + sex"
16 model <- rxLogit(formula = incomeformula, data = incomedata)
17 summary(model)
18
19
20 ####### Create Logistic Regression Model Using Local Parallel Context #######
21 rxSetComputeContext('localpar')
22 model <- rxLogit(formula = incomeformula, data = incomedata)
23 summary(model)
24
25
26 ####### Create Logistic Regression Model Using Local Parallel Context #######
27 spark <-RxSpark()
28 rxSetComputeContext(spark)
29 model <- rxLogit(formula = incomeformula, data = incomedata)
30 summary(model)
31
32
```

Set the compute context to 'local'.

Check Hadoop version and explore the "data" folder in HDFS.

Define path to income.csv file, create a data source using the path specifying the HDFS file system, and view the variables for the data source.

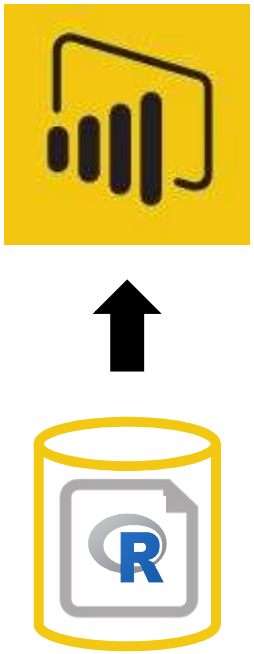
Create a Logistic Regression model for the binary income variable using age, education, race, and sex as features.

Change the compute context to 'localpar' and create the same Logistic Regression model

Change the compute context to RxSpark() and create the same Logistic Regression model

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# R in Power BI

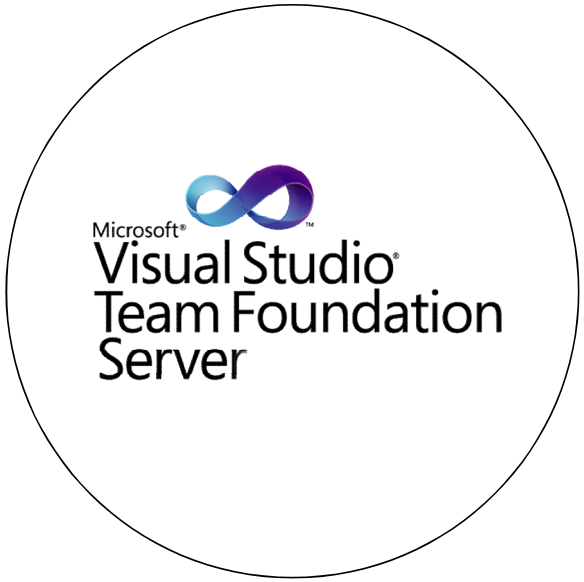


**Use an R Script as a Data Source**


**Use an R Script to Create a Visualization**

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Project Management and Analytics Governance




Source control, reporting, and project management capabilities


Share code, track work, and ship software


Distributed version control system


Hosted public and private repositories; integrated with Visual Studio

BLUE GRANITE
BUSINESS INSIGHT. DELIVERED.

# Microsoft R Parallelized Algorithms and Functions

## Scale R – Parallelized Algorithms & Functions

### Data Preparation

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

### Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

### Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

### Sampling

- Subsample (observations & variables)
- Random Sampling

### Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

### Variable Selection

- Stepwise Regression

### Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

### Cluster Analysis

- K-Means

### Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

### Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms