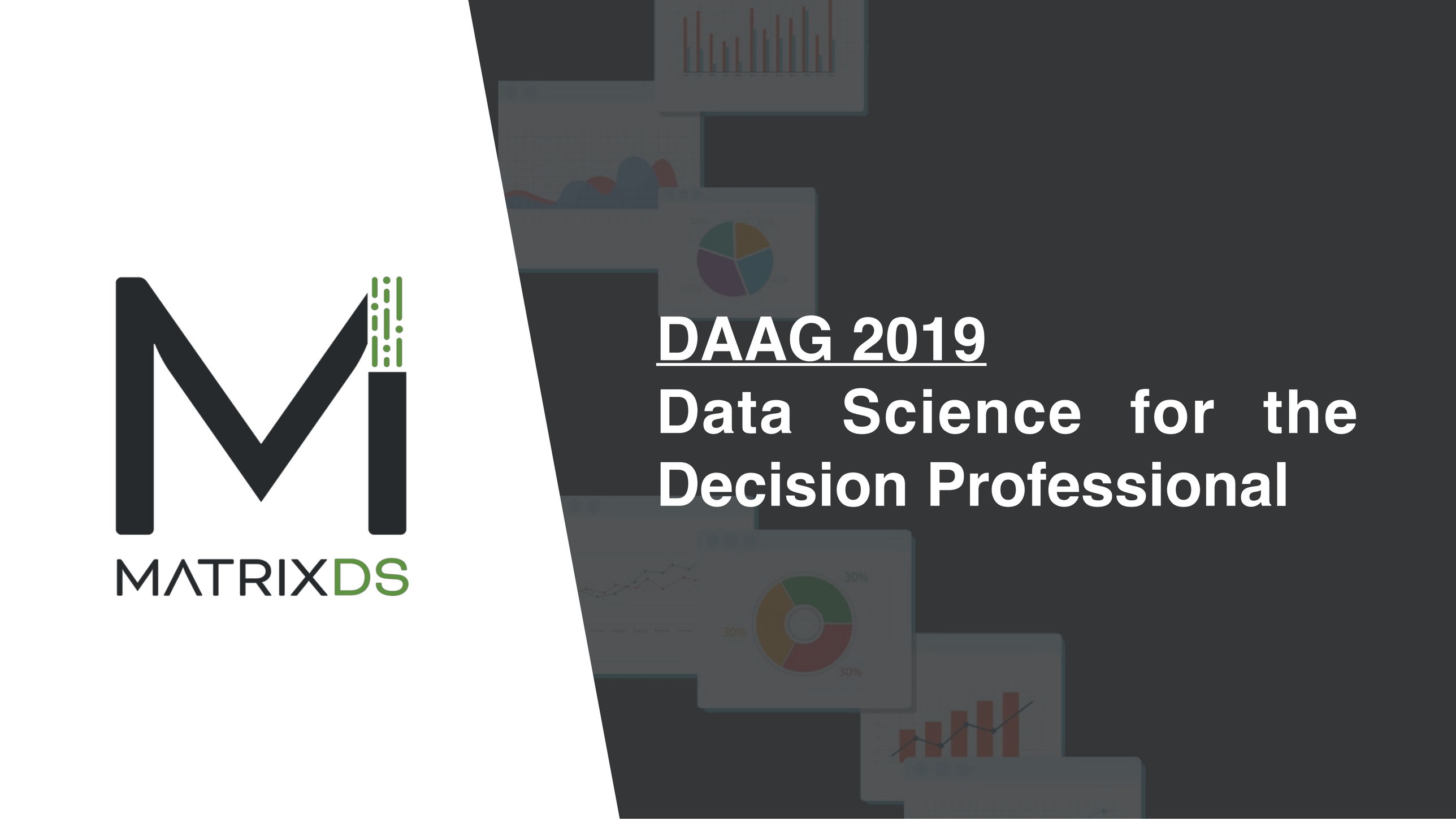


DAAG 2019

Data Science for the Decision Professional

THE INSTRUCTORS

Alejandro Martinez

CEO

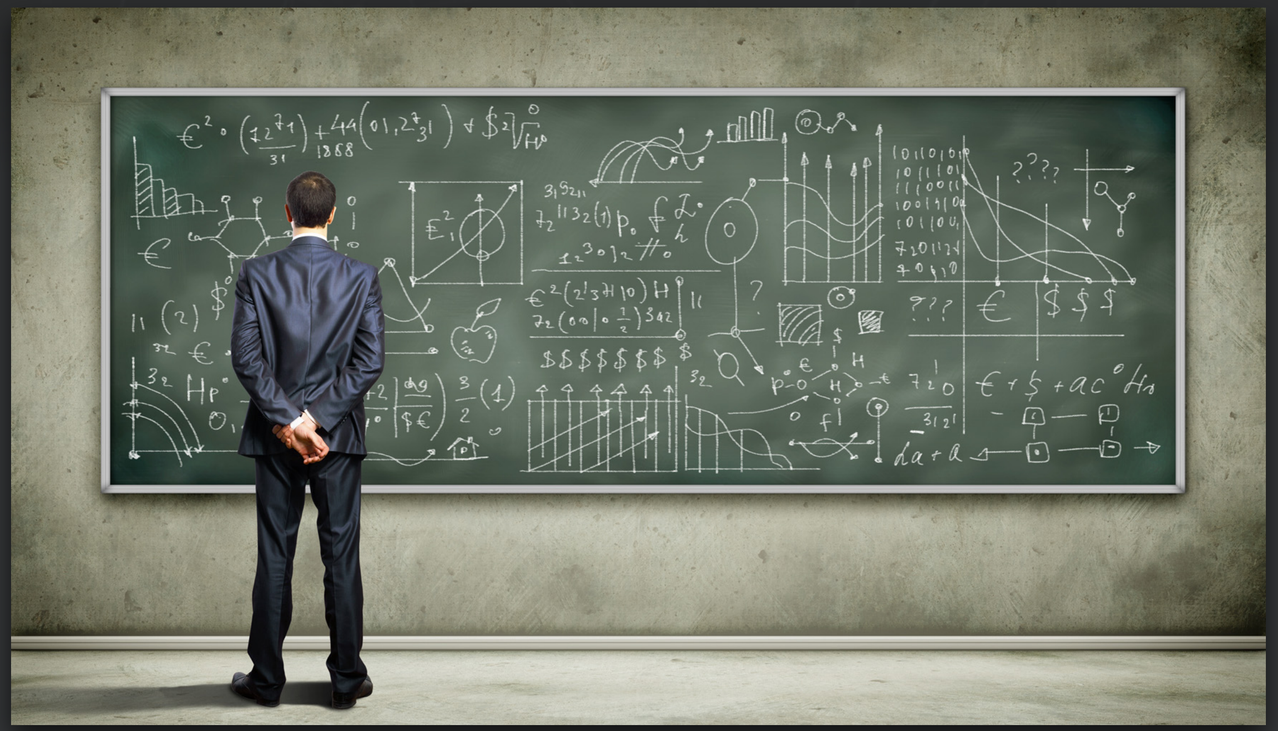
Stanford
PhD Candidate
Decision Analysis
10+ years in operations

Isaac Faber

Chief Data Scientist

Stanford/US Army
PhD Candidate
Risk Analysis
15+ years in analytics

Data Science is a field that focuses on using quantitative techniques to extract knowledge from data



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

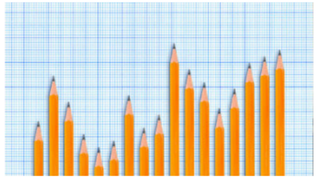
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 16 COMMENT HH TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

WHAT TO READ NEXT



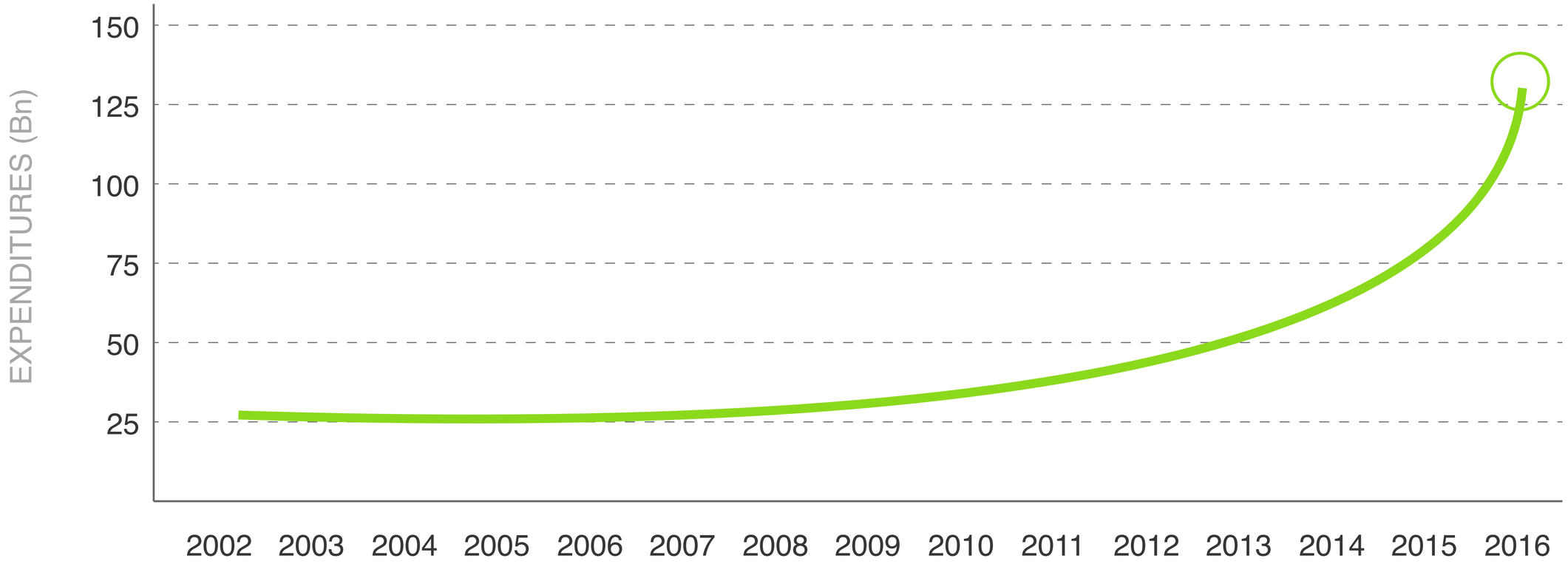
What Data Scientists Really Do, According to 35 Data Scientists

VIEW MORE FROM THE

October 2012 Issue


NETFLIX
GM
Google
airbnb
UBER
Coca-Cola

130B MARKET & GROWING



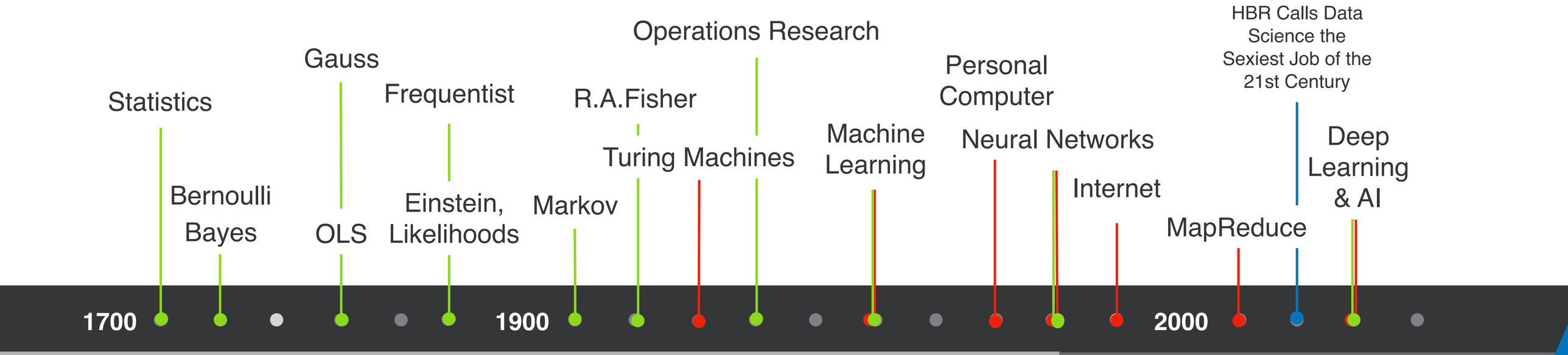
130B
in spending

3000
chief data officers

2.7M
data team positions

MATRIXDS

HISTORY OF DATA SCIENCE



Amount of Data Accessible for Analysis

1EB

1PB

1TB

1GB

- Computer Science
- Probability and Statistics
- Data

Times have changed

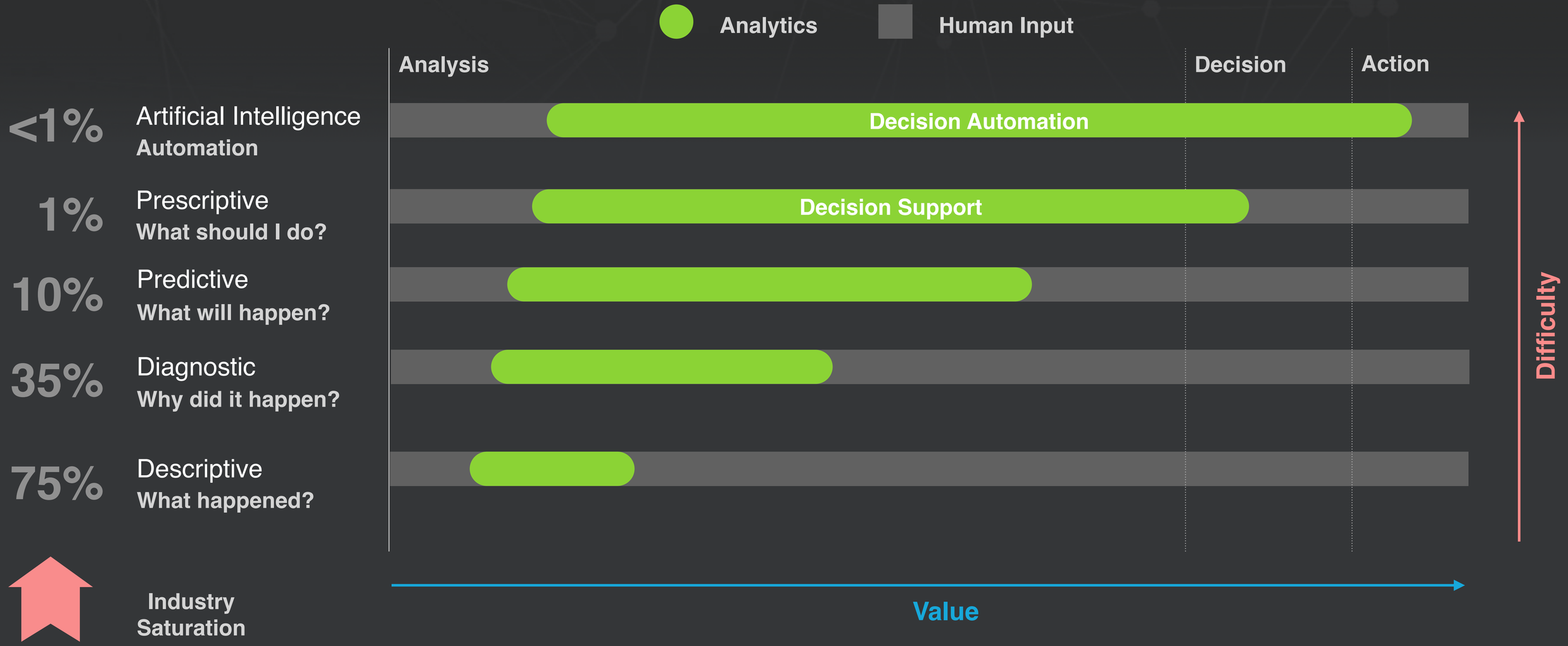
A.I. Is The New Electricity

Andrew Ng

MATRIXDS

WHO IS TEACHING DATA SCIENCE?

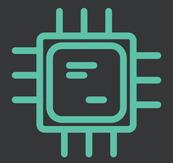
WHAT CAN WE DO WITH DATA SCIENCE



WHAT DO WE NEED TO DO DATA SCIENCE

Data

Dev Tools

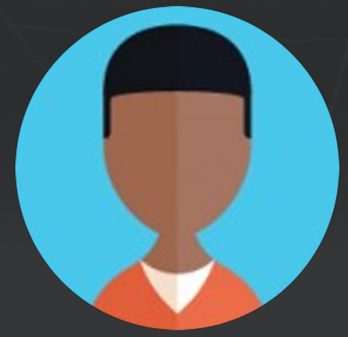


Compute

DIFFERENCE BETWEEN DS AND SD

Developer

IDE: Designed for incremental product development
Development Infrastructure: Small foot print, often just a vm on a local machine and a browser
Production Infrastructure: Can be a large high cost system with complex components that scales up based on product demand
Code Repo: The central ground truth and team management hub for all projects



Data Scientist

IDE: Designed for exploratory data analysis and iterative model building
Development Infrastructure: Can be large scale data warehouse or a large scale compute clusters including GPUs
Production Infrastructure: Light weight hosting, typically a small web dashboard or API
Code Repo: Used for end product or to build common analysis libraries when integrating with a product (done as a last step in a project)

COURSE OBJECTIVES

- Provide a basic understanding of data science and what we can achieve with it
- Compare data science to decision analysis and understand how they compliment each other
- Buzzwords explained
- Provide hands on keyboard data science coding experience
- Provide resources of next steps to continue on your data science journey

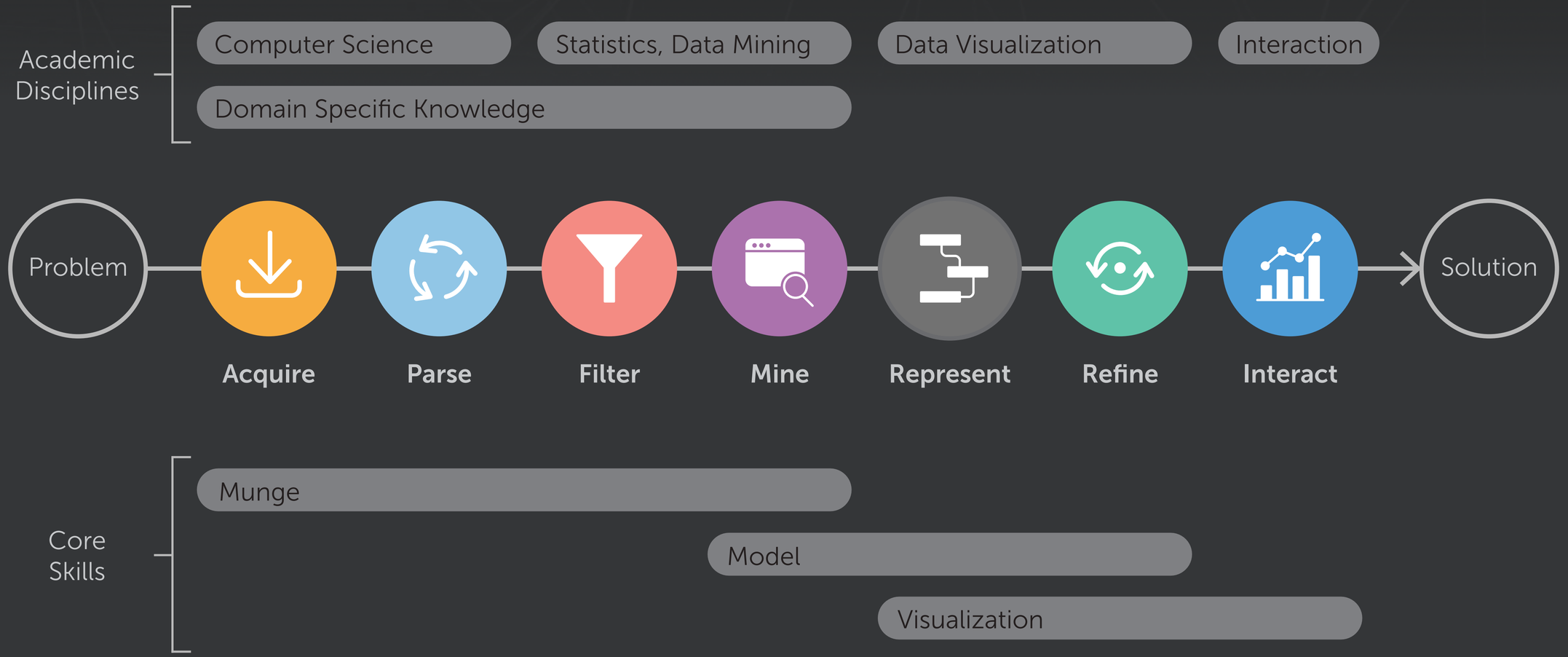
COURSE MATERIALS

- We will use a MatrixDS public project with all the code and data
- You will need access to a web browser (Safari, Chrome or Firefox preferred)

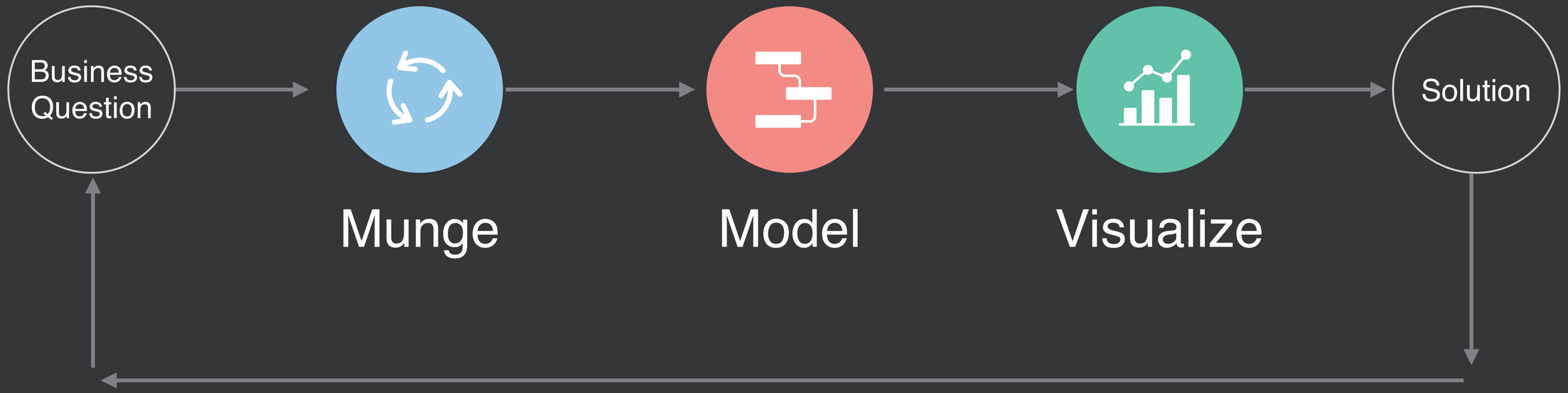
MATRIXDS

THE DATA SCIENCE PROCESS

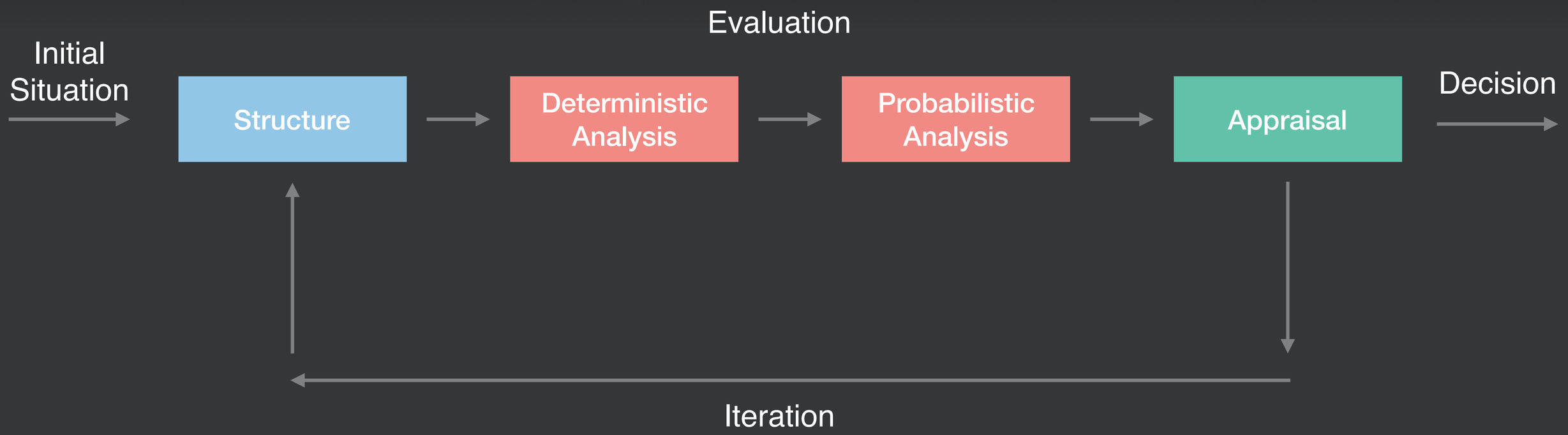
DATA SCIENCE PROCESS



THE BASIC DATA SCIENCE PROCESS



THE DECISION ANALYSIS CYCLE



THE BUSINESS PROBLEM

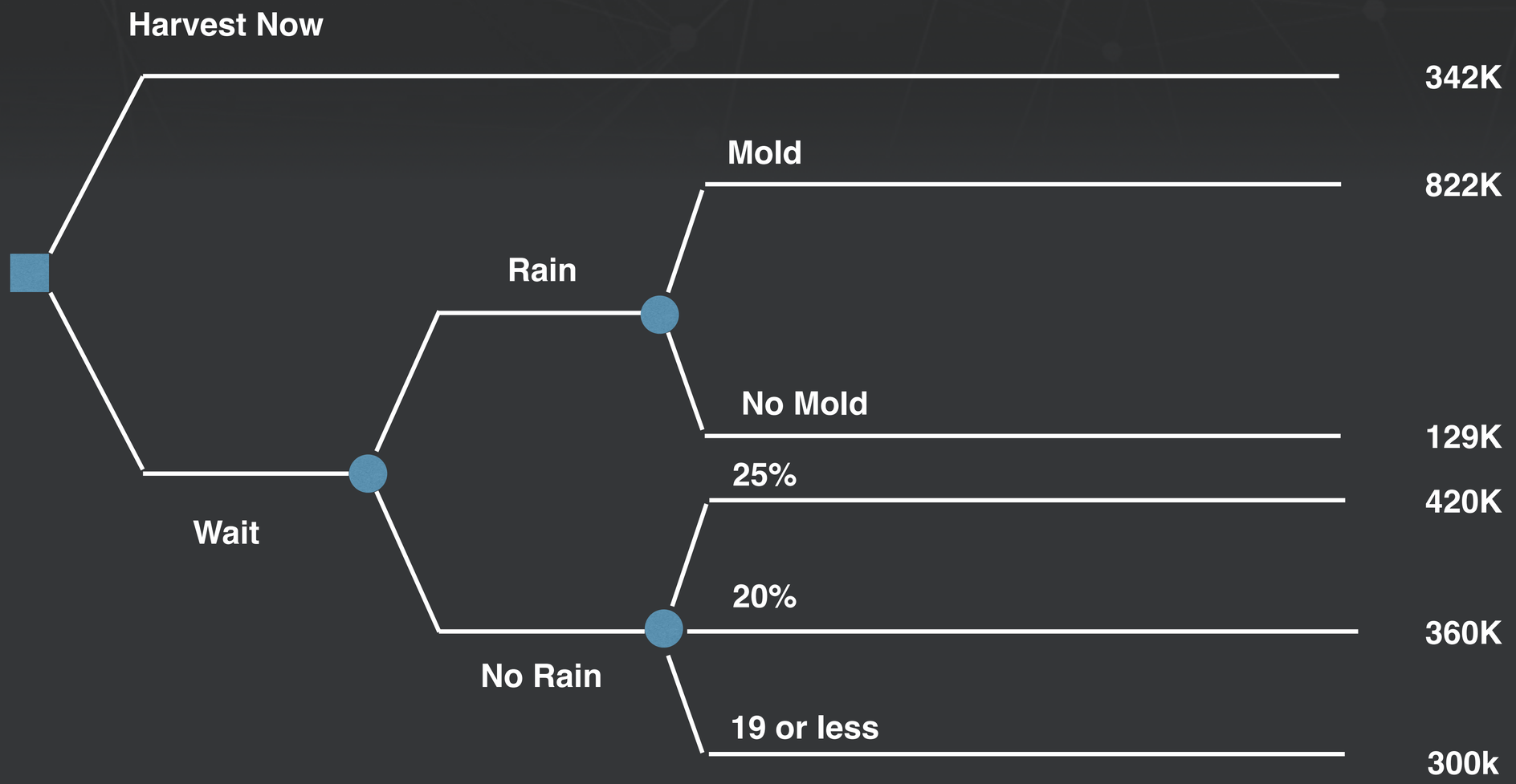
THE BUSINESS PROBLEM

HARVEST NOW OR HARVEST LATER

Assumptions:

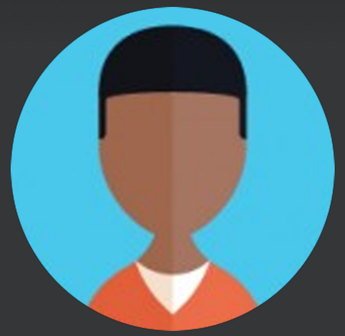
- P(Storm) = 0.667
- P(MoldlStorm) = 0.4
- P(MoldlNo_Storm) = 0
- Reputation Cost of thin wine = 250K
- 1000 cases
- Rain and no Mold causes swelling (7.5%) and thin wine
- Botrytis Rep = 150K
- 25% Sugar wine sells at $35
- 20% sugar wine sells at $30
- Lower sugar wine sells at $25
- Botrytis wine sells at $80
- Botrytis reduces juice to 70%
- Thin Wine sells at $10

DECISION TREE



MUNGE

PROGRAMMING LANGUAGES



Engineer, Scientist & Statistician

Computer Scientist & Developer

Business Analyst & Consultant

RStudio

Go to file/function | Addins | Project: (None)

userStats.R | DA HW.R | Dissertation.tex | temp | app.R | for_review.tex | sample.

Run App

```
1  #
2  # This is a Shiny web application. You can run the application by clicking
3  # the 'Run App' button above.
4  #
5  # Find out more about building applications with Shiny here:
6  #
7  #    http://shiny.rstudio.com/
8  #
9
10 ###############################################################################
11 # This points the Shiny server tool to where all your Rstudio Libraries are installed
12 # that means that any library you install on your Rstudio instance in this project,
13 # will be available to the shiny server
14 ###############################################################################
15 .libPaths( c( .libPaths(), "/srv/R/library") )
16
17 ###############################################################################
18 # Here you can call all the required libraries for your code to run
19 ###############################################################################
20 library(shiny)
21
22 ###############################################################################
```

8:2 (Top Level) R Script

Console | Terminal

```
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.

[Workspace loaded from ~/.RData]

Loading required package: lubridate

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
m	Environment
t	Environment
temp	18 obs. of 17 variables
User_info	16 obs. of 2 variables
Values	
active10	"[{\n \"$match\": {\n \"started\": { \"$gte\": \"new Date(IS…
activeUsers	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
activeUsersnow	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
columns	chr [1:2] "email" "displayName"

Files | Plots | Packages | Help | Viewer

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

RStudio

Go to file/function | Addins | Project: (None)

userStats.R | DA HW.R | Dissertation.tex | temp | app.R | for_review.tex | sample.

Run App

```
1  #
2  # This is a Shiny web application. You can run the application by clicking
3  # the 'Run App' button above.
4  #
5  # Find out more about building applications with Shiny here:
6  #
7  #    http://shiny.rstudio.com/
8  #
9
10 ################################################################################
11 # This points the Shiny server tool to where all your Rstudio Libraries are installed
12 # that means that any library you install on your Rstudio instance in this project,
13 # will be available to the shiny server
14 ################################################################################
15 .libPaths( c( .libPaths(), "/srv/R/library") )
16
17 ################################################################################
18 # Here you can call all the required libraries for your code to run
19 ################################################################################
20 library(shiny)
21
22 ################################################################################
```

8:2 (Top Level) R Script

Console | Terminal

```
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.

[Workspace loaded from ~/.RData]

Loading required package: lubridate

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
m	Environment
t	Environment
temp	18 obs. of 17 variables
User_info	16 obs. of 2 variables
Values	
active10	"[{\n \"$match\": {\n \"started\": { \"$gte\": \"new Date(IS…
activeUsers	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
activeUsersnow	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
columns	chr [1:2] "email" "displayName"

Files | Plots | Packages | Help | Viewer

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

RStudio

Go to file/function | Addins | Project: (None)

userStats.R | DA HW.R | Dissertation.tex | temp | app.R | for_review.tex | sample.

Run App

```
1  #
2  # This is a Shiny web application. You can run the application by clicking
3  # the 'Run App' button above.
4  #
5  # Find out more about building applications with Shiny here:
6  #
7  #    http://shiny.rstudio.com/
8  #
9
10 ###################################################################################
11 # This points the Shiny server tool to where all your Rstudio Libraries are installed
12 # that means that any library you install on your Rstudio instance in this project,
13 # will be available to the shiny server
14 ###################################################################################
15 .libPaths( c( .libPaths(), "/srv/R/library") )
16
17 ###################################################################################
18 # Here you can call all the required libraries for your code to run
19 ###################################################################################
20 library(shiny)
21
22 ###################################################################################
```

8:2 (Top Level) R Script

Console | Terminal

```
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.

[Workspace loaded from ~/.RData]

Loading required package: lubridate

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
m	Environment
t	Environment
temp	18 obs. of 17 variables
User_info	16 obs. of 2 variables
Values	
active10	"[{\n \"$match\": {\n \"started\": { \"$gte\": \"new Date(IS…
activeUsers	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
activeUsersnow	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
columns	chr [1:2] "email" "displayName"

Files | Plots | Packages | Help | Viewer

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

RStudio

Go to file/function | Addins | Project: (None)

userStats.R | DA HW.R | Dissertation.tex | temp | app.R | for_review.tex | sample.

Run App

```
1  #
2  # This is a Shiny web application. You can run the application by clicking
3  # the 'Run App' button above.
4  #
5  # Find out more about building applications with Shiny here:
6  #
7  #    http://shiny.rstudio.com/
8  #
9
10 ###################################################################################
11 # This points the Shiny server tool to where all your Rstudio Libraries are installed
12 # that means that any library you install on your Rstudio instance in this project,
13 # will be available to the shiny server
14 ###################################################################################
15 .libPaths( c( .libPaths(), "/srv/R/library") )
16
17 ###################################################################################
18 # Here you can call all the required libraries for your code to run
19 ###################################################################################
20 library(shiny)
21
22 ###################################################################################
```

8:2 (Top Level) R Script

Console | Terminal

```
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.

[Workspace loaded from ~/.RData]

Loading required package: lubridate

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
m	Environment
t	Environment
temp	18 obs. of 17 variables
User_info	16 obs. of 2 variables
Values	
active10	"[{\n \"$match\": {\n \"started\": { \"$gte\": \"new Date(IS…
activeUsers	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
activeUsersnow	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
columns	chr [1:2] "email" "displayName"

Files | Plots | Packages | Help | Viewer

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

RStudio

Go to file/function | Addins | Project: (None)

userStats.R | DA HW.R | Dissertation.tex | temp | app.R | for_review.tex | sample.

Run App

```
1  #
2  # This is a Shiny web application. You can run the application by clicking
3  # the 'Run App' button above.
4  #
5  # Find out more about building applications with Shiny here:
6  #
7  #    http://shiny.rstudio.com/
8  #
9
10 ###################################################################################
11 # This points the Shiny server tool to where all your Rstudio Libraries are installed
12 # that means that any library you install on your Rstudio instance in this project,
13 # will be available to the shiny server
14 ###################################################################################
15 .libPaths( c( .libPaths(), "/srv/R/library") )
16
17 ###################################################################################
18 # Here you can call all the required libraries for your code to run
19 ###################################################################################
20 library(shiny)
21
22 ###################################################################################
```

8:2 (Top Level) R Script

Console | Terminal

```
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.

[Workspace loaded from ~/.RData]

Loading required package: lubridate

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
m	Environment
t	Environment
temp	18 obs. of 17 variables
User_info	16 obs. of 2 variables
Values	
active10	"[{\n \"$match\": {\n \"started\": { \"$gte\": \"new Date(IS…
activeUsers	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
activeUsersnow	"[{\n \"$lookup\": {\n \"from\": \"tools\",\n \"localField\"…
columns	chr [1:2] "email" "displayName"

Files | Plots | Packages | Help | Viewer

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

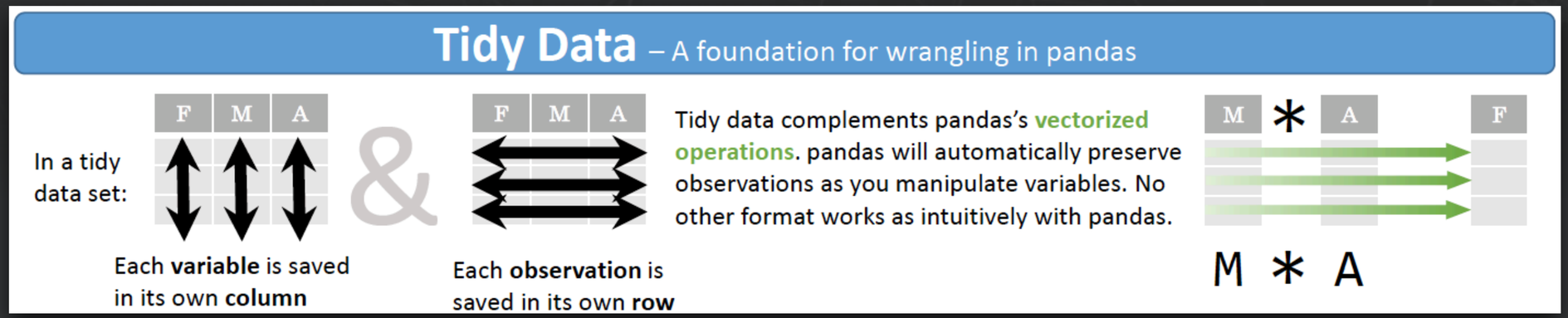
Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

TIDY DATA



- Term developed by Hadley Wickham
- Data is 'flat' only rows and columns (2D)
- Each row is one 'observation'
- Each column is one 'feature'

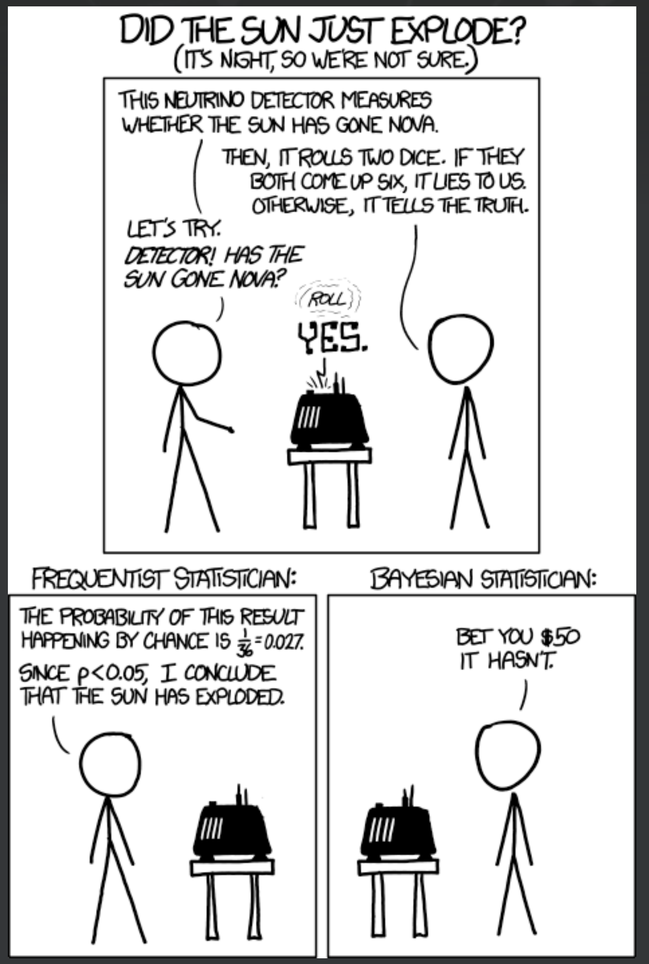
MATRIXDS

https://bit.ly/2Un1oO3

MODEL

MODEL SELECTION

- Probabilistic & Statistical Inference
  - Bayesian V Frequentist



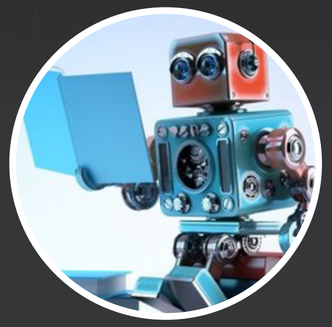
MODEL SELECTION

- Classification vs Continuous Output
  - Feature Engineering
  - Time series models
  - Regression fit models (OLS & Logit)
  - Naive Bayes
  - Decision trees
  - Random Forrest
  - Neural Networks
  - Deep Learning

MODEL SELECTION

- Clustering Models
- Ensemble Modeling (Boosting)
- Simulations
- Cost Functions (GOF)
  - Regularization
  - Overfitting
  - Training, Validation & Testing

POPULAR TOPICS



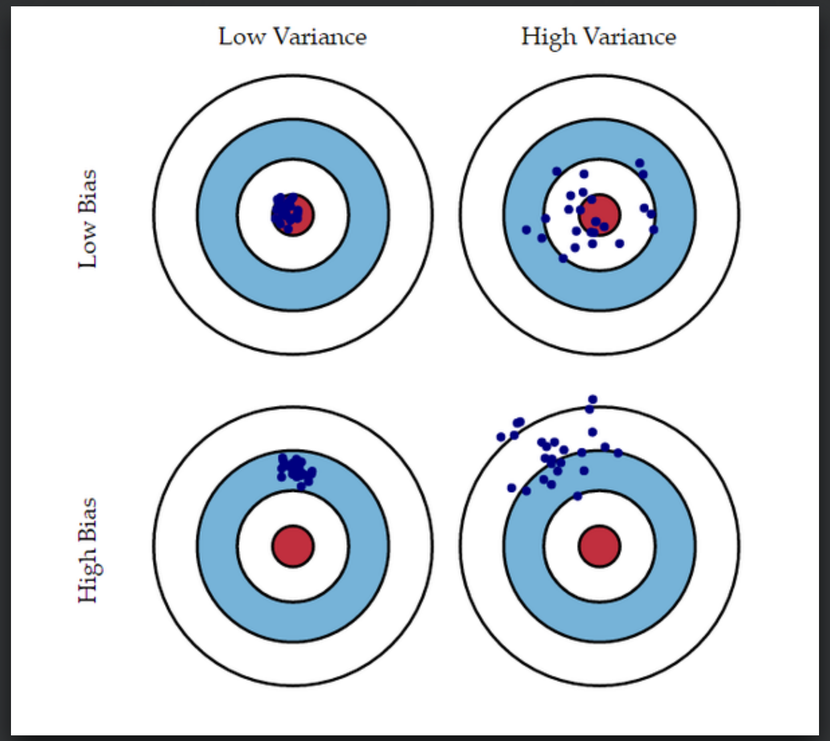
Machine Learning

Supervised / Unsupervised

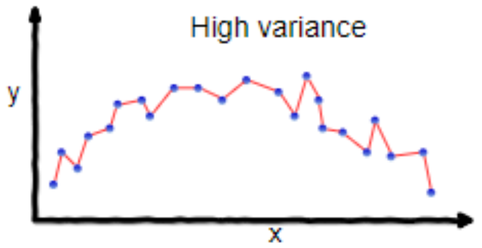
Artificial Intelligence

Big Data

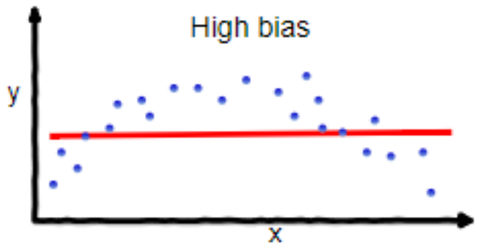
POPULAR TOPICS



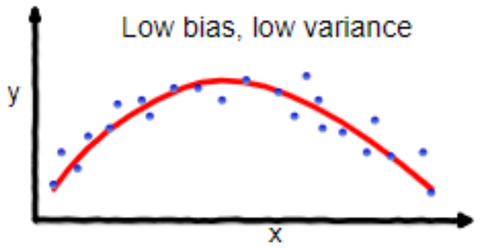
POPULAR TOPICS



overfitting

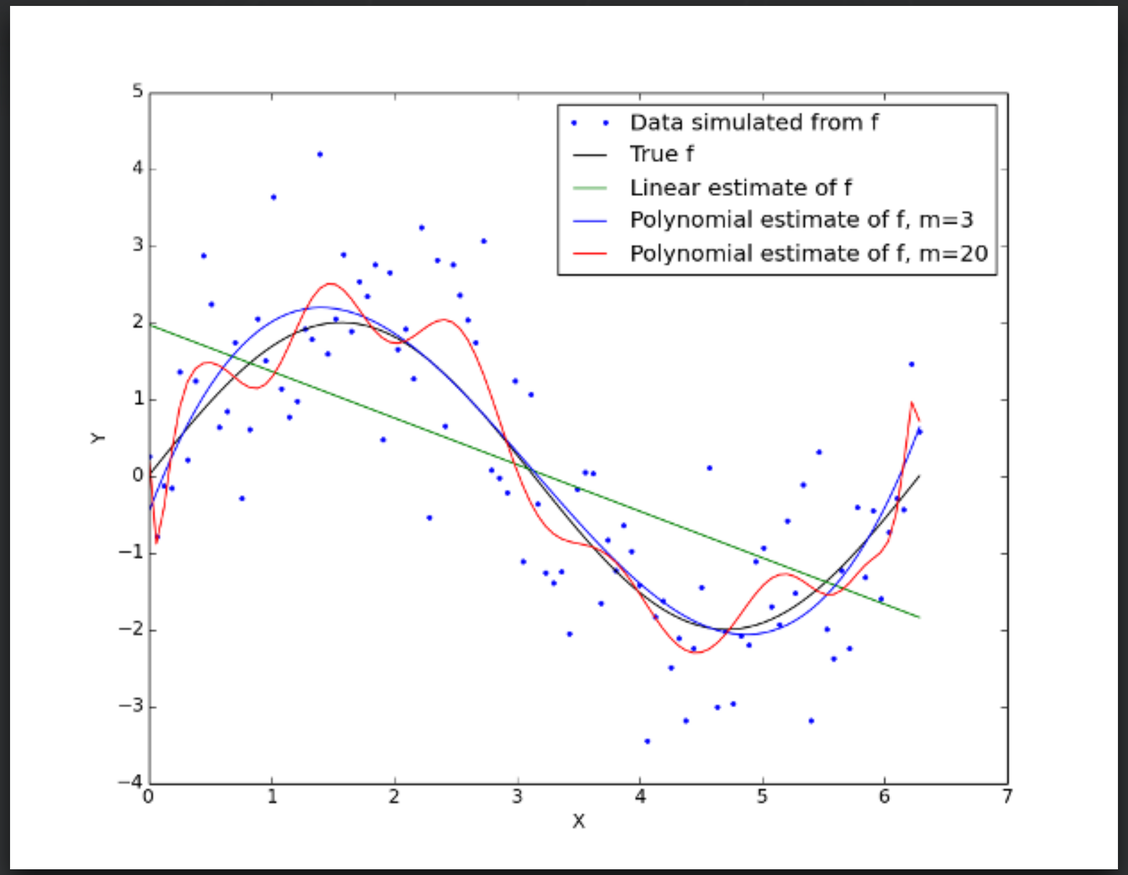


underfitting

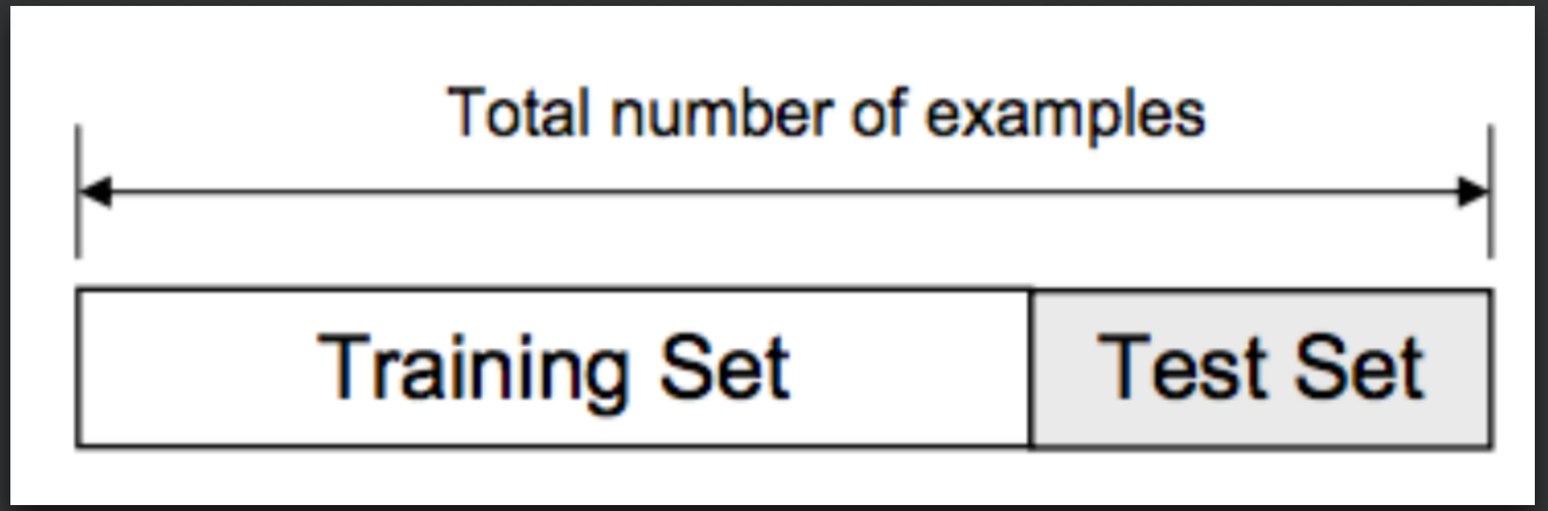


Good balance

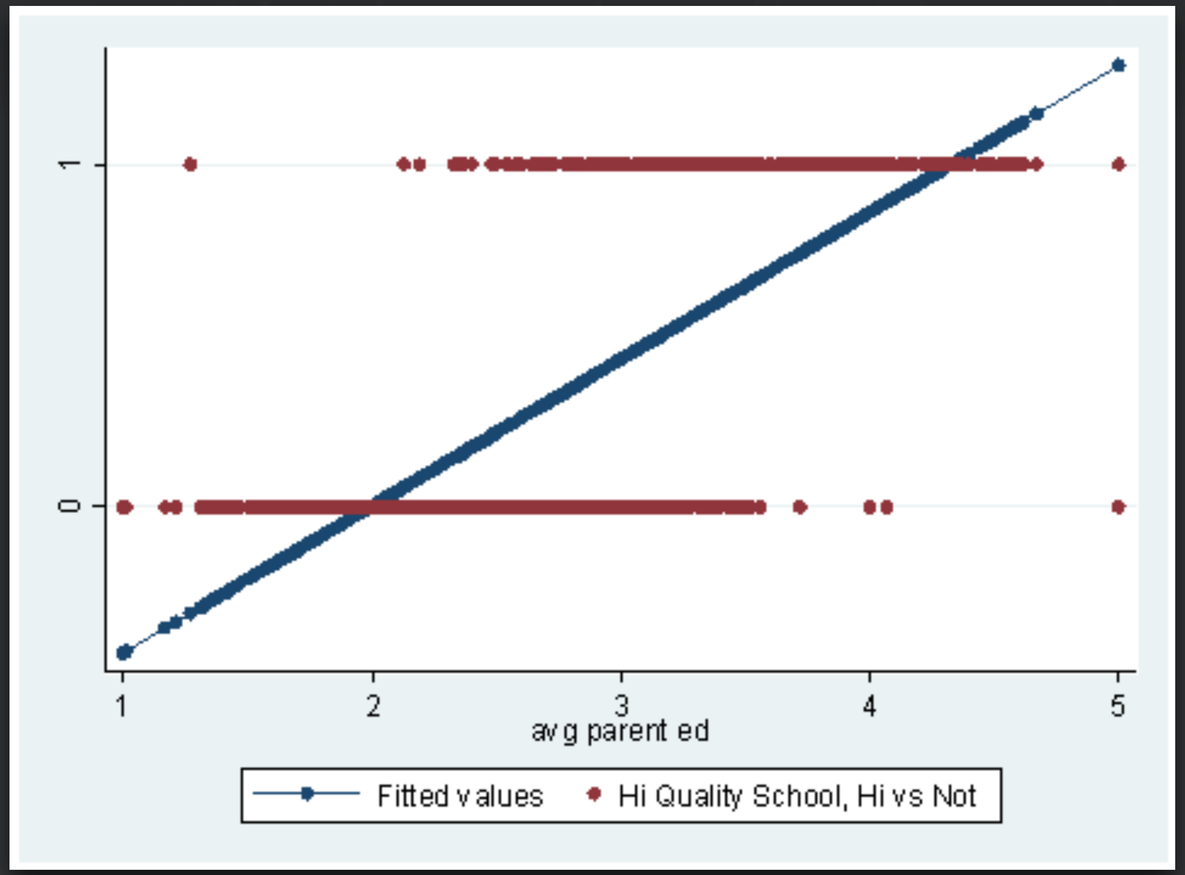
POPULAR TOPICS



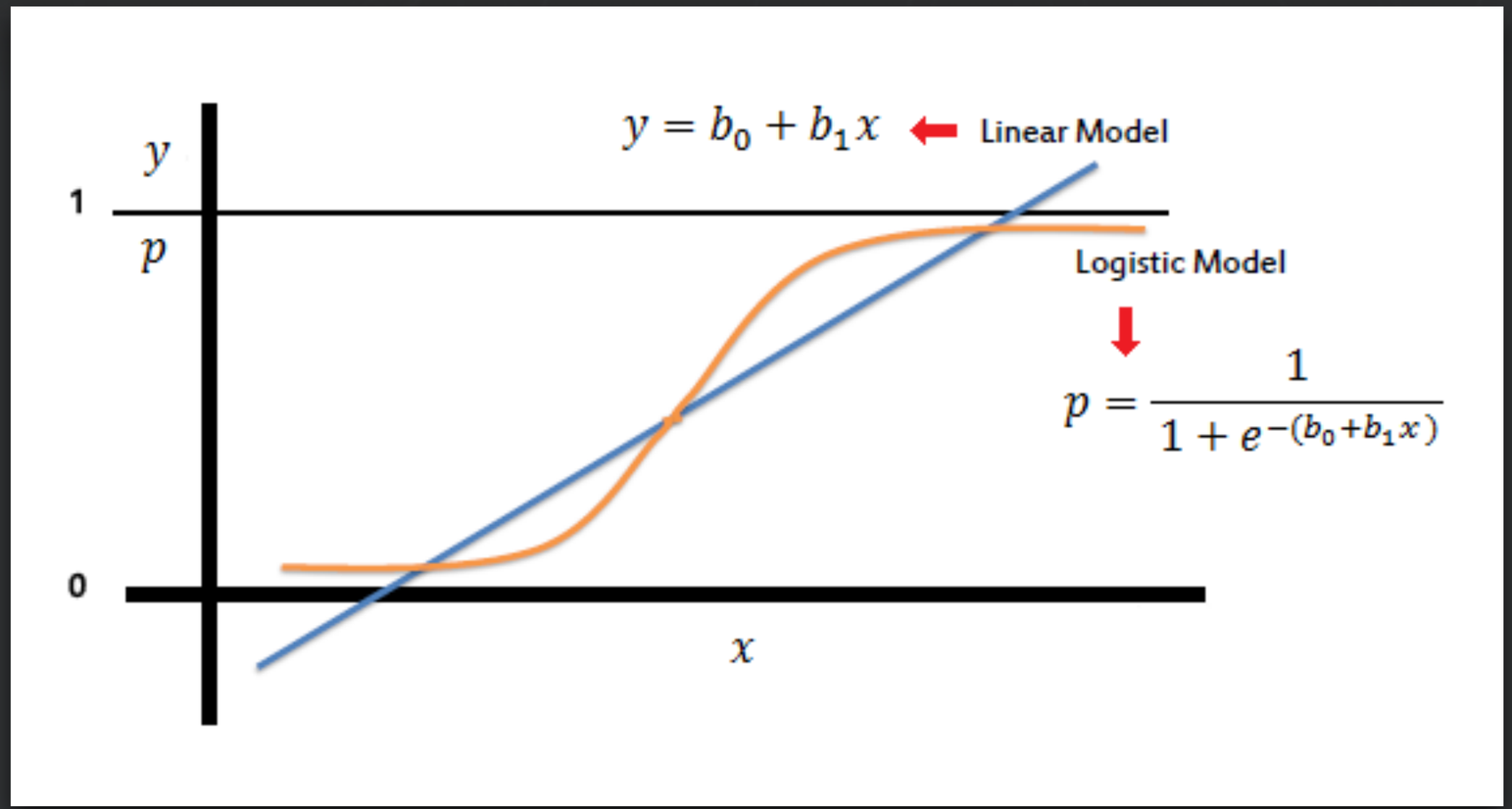
TRAINING AND TEST SET



LOGIT REGRESSION



LOGIT REGRESSION

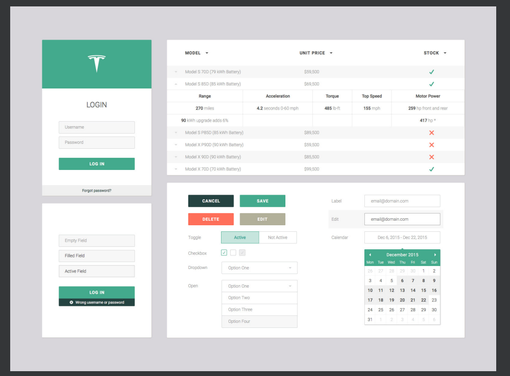


MATRIXDS

VISUALIZE

DIFFERENT FORMATS

Dashboard



Web Application

Reports

Presentations

OBJECTIVE OF VISUALIZATION

STORYTELLING

OBJECTIVE OF VISUALIZATION

Graphs are Typically about comparisons; make it easy for the reader to make them

Graphs should be made as simple as possible, but no simpler

All graphs tell a story; don’t tell a misleading one

MATRIXDS