

Introduction to Data Science

Lecture 02; April 6th, 2015

Ernst Henle

ErnstHe@UW.edu

Skype: ernst.predixion

Agenda

- Social Interactions: LinkedIn going strong!
 - Help classmates
 - with homework
 - with issues like: Getting Octave to run and downloading the VM.
 - Discuss long-term employment plans
- Reminder: Optional class on programming in R on April 11th 2015 from 9:00 AM to 12:00 noon. Use the following link:
<http://uweoconnect.extn.washington.edu/datasci250>
- Review
 - Class Structure
 - DFD
- Data Sharing
- Basics of R (Very slow intro)
- Quiz 02a: Use R and Octave
- Data Preparation
- Break
- Data Preparation in R
- Break
- Data Science Business Perspective
- GNU-Octave (Time permitting)

Review

- Class Structure
 - Class Prerequisites: Review last week's slide deck for required software, attendance, participation, and assignments
 - Last week was slow and non-technical. This week will be slow and technical. Starting next week, we will pick up the pace.
 - Review of Assignments are sometimes individual and sometimes at the beginning of the next lecture.
- DFD
 - Individual feedback provided. Main problems were those described in the lecture notes and Quiz 01b
 - Starting lecture 03 (next week), I will use DFDs to present overviews and concepts in data transformations and data movement.

Data Sharing (watch these videos)

- <http://www.youtube.com/watch?v=RVZbk3GEVSw>
 - It's all there.
 - Just follow that link. You don't need the data.
 - I've already analyzed the data.
 - I can't find my data.
- <http://www.youtube.com/watch?v=RtSv0gSbCP8>
 - This is my only copy
 - The data format is unusable.
- <http://www.youtube.com/watch?v=-MIH8PkuUo4>
 - The attributes (headings) are self-evident
 - Somebody else knows how that column was calculated.
 - You can figure out for yourself what that column means.

Some Sample Archives

- <http://data.bls.gov/cgi-bin/surveymost?bls>
 - Bureau of Labor Statistics provided monthly datasets on labor since 2003. These data sets are somewhat short as they only provide a single number in a single year. You could extract a lot of data, but it would need to be compiled together as this website separates every attribute of the data.
- <http://socialcomputing.asu.edu/datasets/Twitter>
- <http://socialcomputing.asu.edu/uploads/1296759055/Twitter-dataset.zip>
 - Link provides a .zip file with GB data of twitter data
- <http://archive.ics.uci.edu/ml/datasets.html>
 - **Status:** Well-formed and existing
 - **Description:** UCI's machine learning repository, A collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.
- <http://www.ourairports.com/data/>
 - **Status:** Well-formed and existing CSV file
 - **Description:** Provide 6 CSV files for airport information, including type, name, location, and frequency of airports, and also containing website URL for different regions etc.
- <http://www.quandl.com/>
 - **Status:** Well-formed and existing CSV, Excel files
 - **Description:** Give share's price and volume for Microsoft, Oracle, IBM, HP, Dell, Cisco, Apple, and Google in every business day from September 1997 to 2013, nowadays. Updated every day.
- Archive of archives: DataSets.doc

R

- Open in R Studio: DataScience02a.R, DataScience02b.R, and DataScience02c.R

Quiz 02

- <https://catalyst.uw.edu/webq/survey/ernsthe/266667>
- You should use R and Octave during the Quiz.

Data Preparation

Data Preparation (0)

- Later we will consider these techniques using DataScience02d.R

Data Preparation (1)

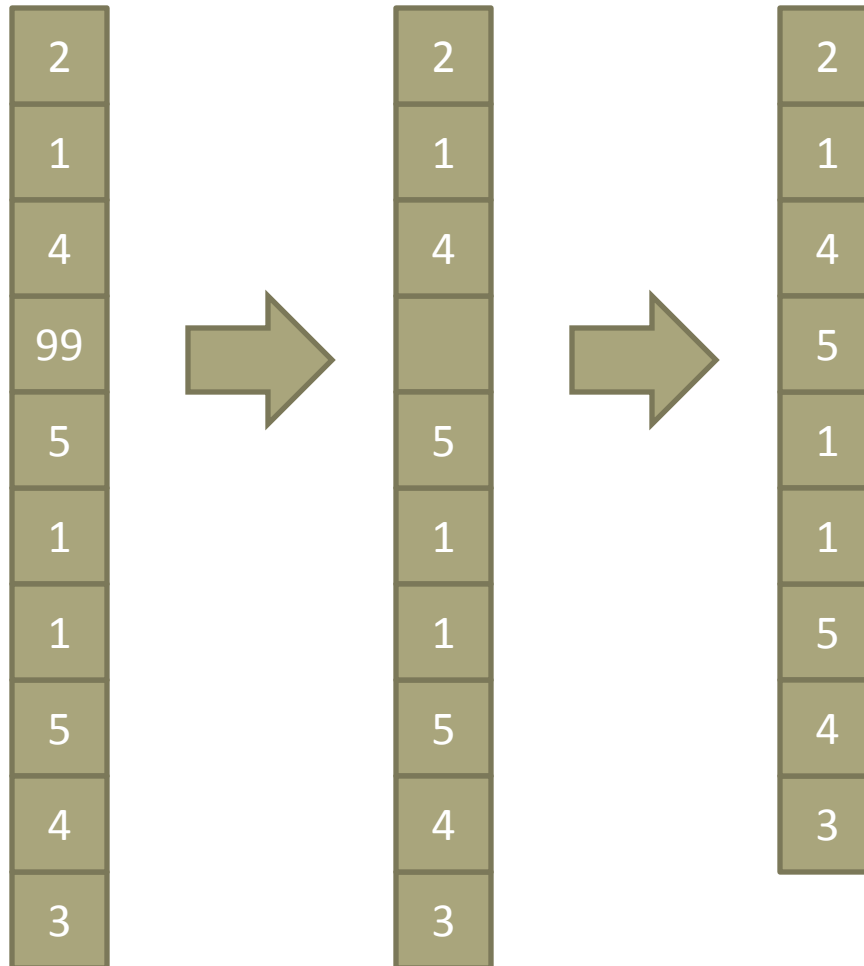
- Schematize (Shape Data)
 - Create Tables
 - Flatten Data
 - Star http://en.wikipedia.org/wiki/Star_schema
 - Snowflake http://en.wikipedia.org/wiki/Snowflake_schema
 - Specify Input vs. Target
 - Specify attributes that are neither Input nor Target
- Clean Data (Today's topic)

Data Preparation (2)

- Clean Data
 - Outlier Removal
 - Numeric
 - Remove data beyond 3 standard deviations (1, 2, 2, 3, 3, 3, 4, 4, 5, 99)
 - `x <- x[x < 10]`
 - Categorical
 - Categories: Remove cases that have less than 1% support (20 X A, 20 X B, 1 X C, 20 X D, 20 X E, 20 X F)

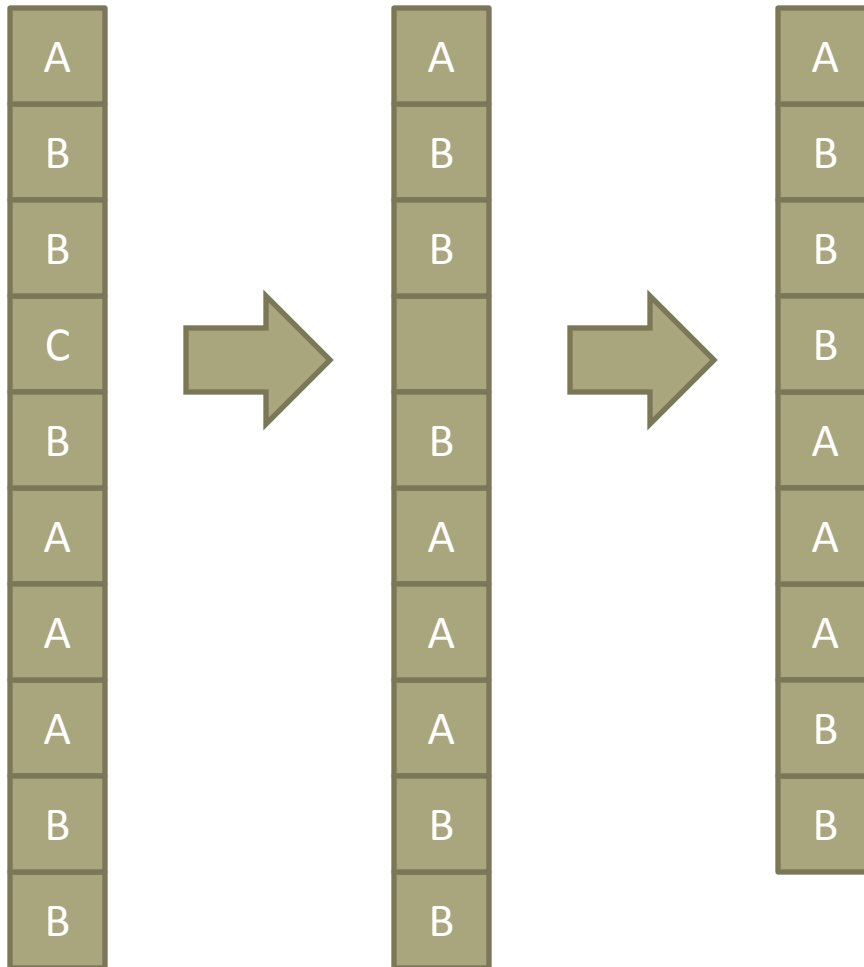
Data Preparation (3)

Outlier Removal (Numeric)



Data Preparation (4)

Outlier Removal (Category)

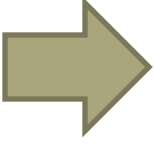


Data Preparation (5)

- Clean Data
 - Relabeling
 - Simplify (e.g. all 4 year degrees, like Bachelors, A.B. BSc, etc. as BS)
 - Example:
 - Vehicle: (Car, Automobile, Bike, Truck, Bicycle, Sedan, Coupe, Cycle, Truck, Velo, Automobile, Bike)
 - Car, Automobile, Sedan, Coupe -> Car
 - Bike, Bicycle, Cycle, Velo -> Bike
 - Truck -> Truck
 - Vehicle: (Car, Car, Bike, Truck, Bike, Car, Car, Bike, Truck, Bike, Car, Bike)
 - De-code (numbers to categories)
 - Example1: Origin: (3, 1, 2, 1, 1, 2)
 - 1 -> USA
 - 2 -> Europe
 - 3 -> Japan
 - Origin: (Japan, USA, Europe, USA, USA, Europe)
 - Example2: Origin: `x <- as.character(3, 1, 2, 1, 1, 2)`

Data Preparation (6)

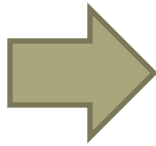
Relabeling (Simplify)

Vehicle		Vehicle
Car		Car
Bike		Bike
Velo		Bike
Truck		Truck
Bicycle		Bike
Sedan		Car
Coupe		Car
Auto		Car
Lorry		Truck
Truck		Truck

Data Preparation (7)

Relabeling (Decode)

Vehicle
1
2
2
3
2
1
1
1
3
3



Vehicle
Car
Bike
Bike
Truck
Bike
Car
Car
Car
Truck
Truck

Code	Item
1	Car
2	Bike
3	Truck

Data Preparation (8)

- Clean Data (continued)
 - Casting
 - Characters to Numbers: ("4", "-7", "X", "3") -> (4, -7, NA, 3)
 - Numbers to Characters : (4, -7, NA, 3) -> ("4", "-7", NA, "3")
 - Normalize
 - Normalize (Linear)
 - offset and multiplier $y = a + bx$ or $y = (x - c)/d$; Where: $a = -c/d$; $b = 1/d$
 - Min-Max where: $c = \min$; $d = \max - \min$
 - Z-score: where $c = \text{mean}$; $d = \text{sigma}$
 - MAD (http://en.wikipedia.org/wiki/Median_absolute_deviation) where $c = \text{median}$; $d = \text{median of differences to median}$
 - Normalize (Non-Linear)
 - Log-normalization: $y = \text{Log}(x)$ or similar
 - Equalization

Data Preparation (9)

Normalization

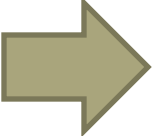
Orig		MM	Z
2		.3	-0.23
-1		0	-1.09
0		.1	-0.8
1		.2	-0.52
7		.8	1.2
9		1	1.78
7		.8	1.2
1		.2	-0.52
1		.2	-0.52
1		.2	-0.52

Data Preparation (10)

- Clean Data (continued)
 - Binarization Categorical to Numerical (Binary)
 - 1 column of Colors (Red, Green, Blue) -> three columns called isRed, isGreen, and isBlue
 - Color: Red, Green, Blue, Blue, Red, Red ->
 - -> isRed: 1, 0, 0, 0, 1, 1
 - -> isGreen: 0, 1, 0, 0, 0, 0
 - -> isBlue: 0, 0, 1, 1, 0, 0
 - Discretization
 - Age: (10, 23, 11, 55, 60, 32, 99, 4, 32, 33, 0) ->
 - Equal Range (0 – 33) (34 – 66) (67 – 99) -> (Low, Low, Low, Med, Med, Low, High, Low, Low, Low, Low)
 - Equal Area (0 - 11) (23 - 33) (55 - 99) -> (Low, Med, Low, High, High, Med, High, Low, Med, Med, Low)
 - Null Handling
 - (4, -7, NA, 3) ->
 - Value removal or Row Removal -> (4, -7, 3)
 - value substitution -> (4, -7, 0, 3)

Data Preparation (11)

Binarization

Vehicle		Car	Bike	Truck
Car		1	0	0
Bike		0	1	0
Bike		0	1	0
Truck		0	0	1
Bike		0	1	0
Car		1	0	0
Car		1	0	0
Car		1	0	0
Truck		0	0	1
Truck		0	0	1

Data Preparation

Break

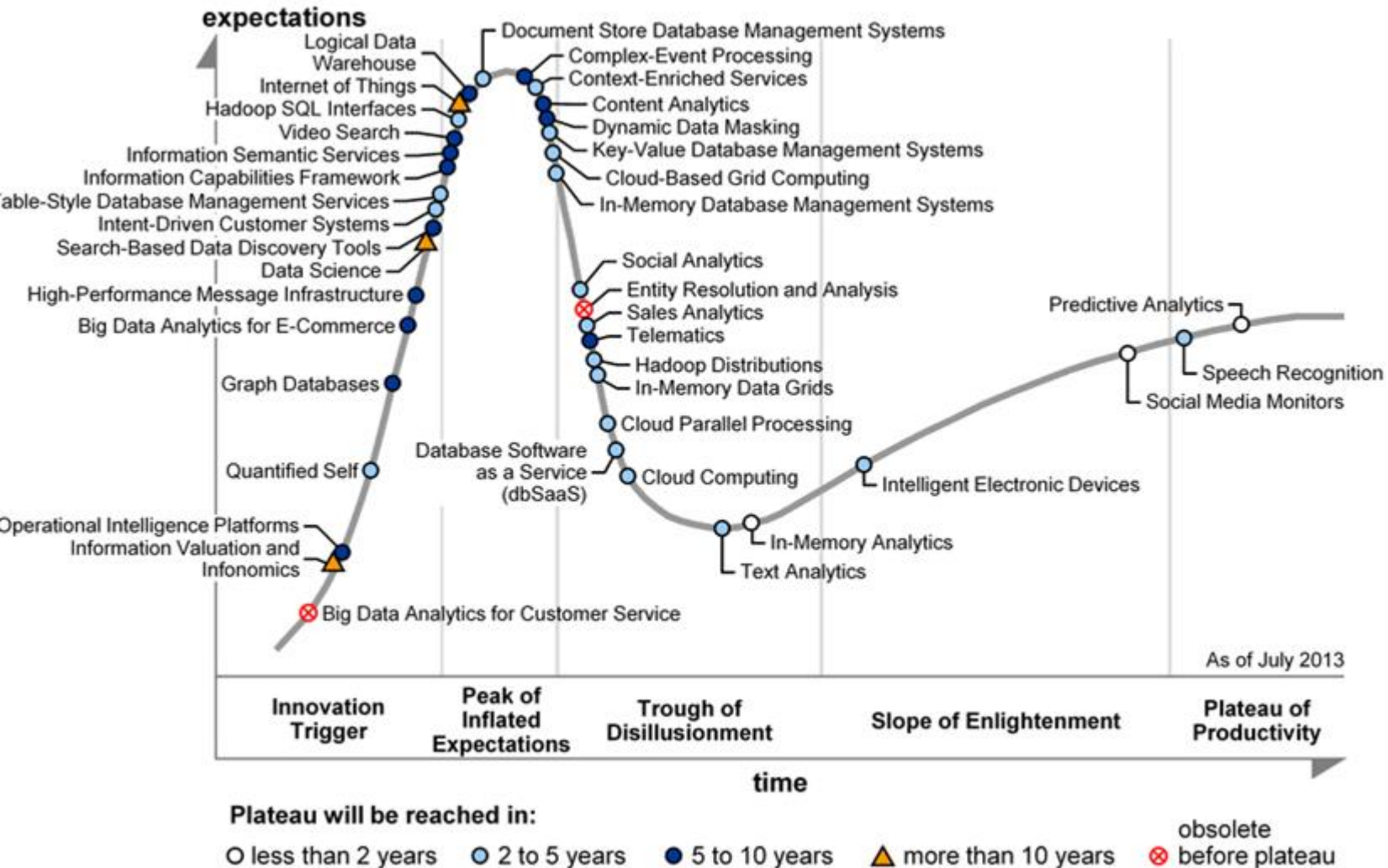
Data Preparation in R

- Open in R Studio: DataScience02d.R

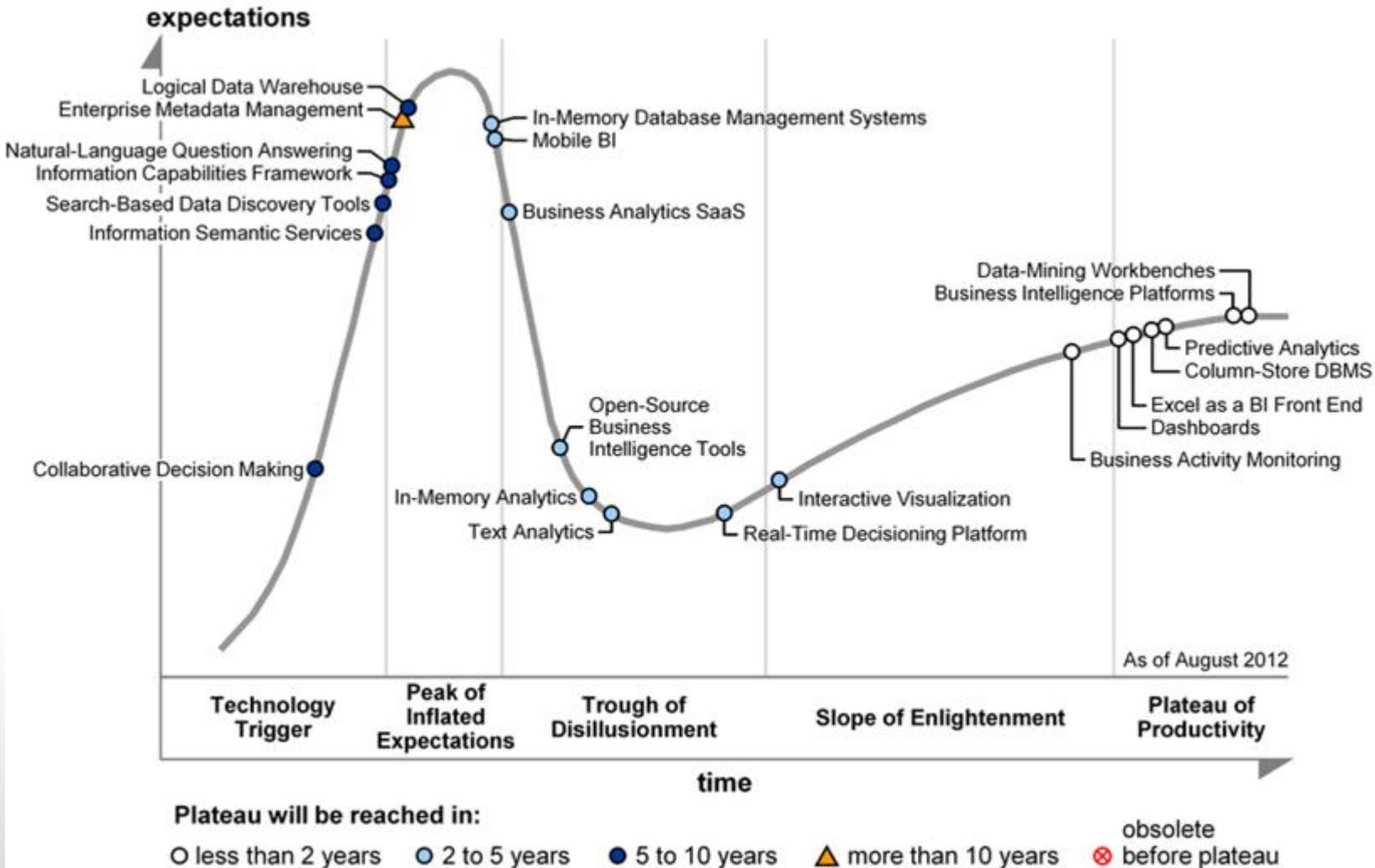
Break

Data Science – Business Perspective

Hype Curve Big Data 2013

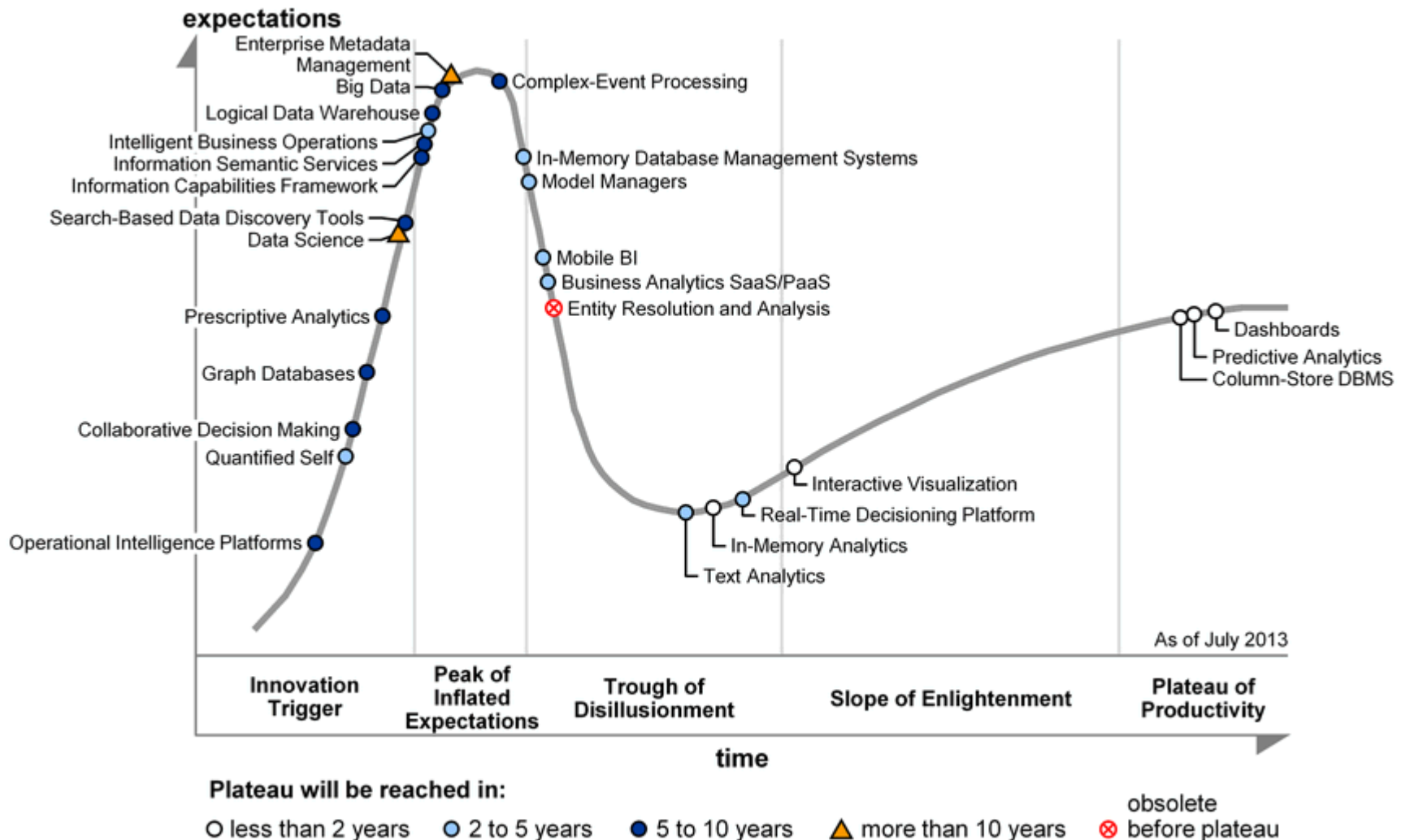


Hype Curve Business Analytics 2012



Hype Curve BI and Analytics 2013

Figure 1. Hype Cycle for Business Intelligence and Analytics, 2013



Benefit vs. Years To Adoption

benefit	years to mainstream adoption			
	less than 2 years	2 to 5 years	5 to 10 years	more than 10 years
transformational	Column-Store DBMS	In-Memory Database Management Systems Intelligent Business Operations	Big Data Collaborative Decision Making Complex-Event Processing Graph Databases Information Capabilities Framework	
high	Dashboards In-Memory Analytics Interactive Visualization Predictive Analytics	Quantified Self Real-Time Decisioning Platform Text Analytics	Logical Data Warehouse Operational Intelligence Platforms Prescriptive Analytics Search-Based Data Discovery Tools	Data Science Enterprise Metadata Management
moderate		Business Analytics SaaS/PaaS Mobile BI Model Managers	Information Semantic Services	
low				

As of July 2013

Data Scientists in the Job Market

- <http://www.kdnuggets.com/2015/03/salary-analytics-data-science-poll-well-compensated.html>
- <http://www.hadoop360.com/blog/salaries-for-hadoop-professionals>
- <http://www.analyticbridge.com/group/salary-trends-and-reports/forum/topics/salary-trends-for-data-science-professionals>
- <http://www.analyticbridge.com/group/salary-trends-and-reports/forum/topics/the-10-highest-paying-jobs-for-math-geeks>

Data Science – Business Perspective

Assignment (1)

For all R assignment items, use the patterns described in DataScience02b.R

1. Using R: Get Data

- a. Get Indian Liver Patient Dataset from the UCI machine learning repository:
 - a. `url <- "http://archive.ics.uci.edu/ml/machine-learning-databases/00225/Indian%20Liver%20Patient%20Dataset%20\(ILPD\).csv"`
 - b. `ILPD <- read.csv(url, header=FALSE, stringsAsFactors=FALSE)`
- b. Get the 11 column headers from this page:
[http://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)#](http://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)#)
- c. Manually construct a vector of column using
 - a. `headers <- c(<name1>, <name2>, ...) # Each column has a name`
- d. Associate names with the dataframe
 - a. `names(<dataframe>) <- headers`

Assignment (2)

2. Using R: Data Exploration

- a. Use **head(ILPD)** to view the first 6 rows.
- b. Determine the **mean**, **median**, and standard deviation (**sd**) of each column.
- c. What does **na.rm = TRUE** do in `sd(x, na.rm = TRUE)`?
- d. Create Histograms (**hist**) for each column where possible.
- e. Use the **plot(ILPD)** function on this data frame to present a general overview of the data. You want to see a matrix of many plots. Your efforts may be thwarted because the Gender column is not numeric. You can skip the Gender column, or you can turn the gender column into a numeric column. You might need help from a fellow student, the LinkedIn group, or me. Look at the plots and answer:
 - a. What can you say about the data?
 - b. How can you tell if a vector contains continuous numbers or binary data?
 - c. How can you tell if two vectors are correlated?

Assignment (3)

3. Using Data Preparation concepts from Lecture, write the code and the results for data preparation in R:
 - a. Remove Outliers: `c(-1, 1, 5, 1, 1, 17, -3, 1, 1, 3)`
 - b. Relabel: `c('BS', 'MS', 'PhD', 'HS', 'Bachelors', 'Masters', 'High School', 'BS', 'MS', 'MS')`
 - c. Normalize: `c(-1, 1, 5, 1, 1, 17, -3, 1, 1, 3)`
 - a. Min-Max Normalization
 - b. Z-score normalization
 - d. Binarize: `c('Red', 'Green', 'Blue', 'Blue', 'Blue', 'Blue', 'Blue', 'Red', 'Green', 'Blue')`
 - e. Discretize: `c(3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 12, 23, 23, 25, 81)`
 - a. 3 Bins of equal range
 - b. 3 Bins Equal of near equal amounts (Code is optional, but present the results. Writing equalization code is tricky)

Assignment (4)

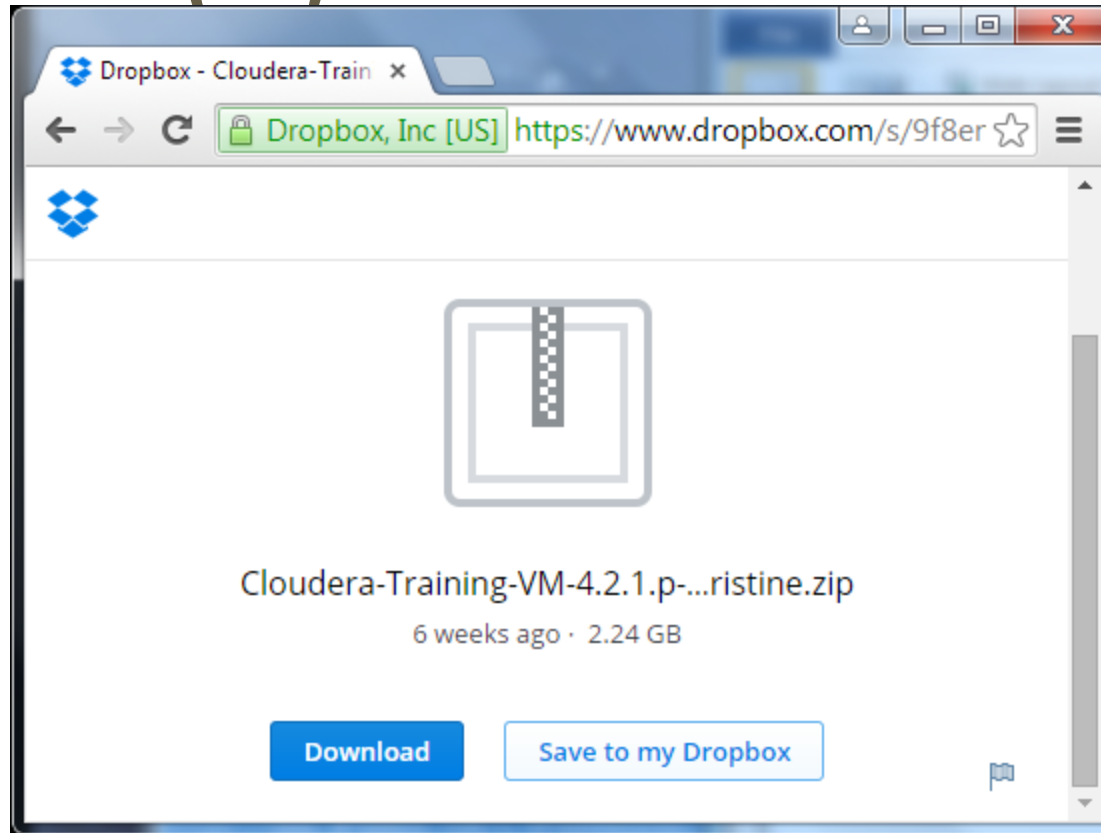
4. Combine the assignment items 1, 2, and 3 into a single R file. Submit the file by Saturday 11:00 PM to the homework submission site on catalyst. If you cannot submit the assignment on time, please notify me before the deadline at ErnstHe@UW.edu.
5. Reading assignment:
 - http://en.wikipedia.org/wiki/Cluster_analysis
 - http://en.wikipedia.org/wiki/K-means_clustering
 - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
 - <http://www.sqlserverdatamining.com/ArtOfClustering/default.aspx>
6. If you haven't downloaded the VM or Octave, please download it now. Note that the VM has a new download url.
7. Optional: Take a look at Lecture_03_Preview.pdf and Quiz03_Preview.txt. These files are available in the catalyst site for class resources under Lecture_02
8. Optional quizzes
 - Quiz 02b <https://catalyst.uw.edu/webq/survey/ernsthe/266668>
 - Quiz 03a (Available Tuesday night)
<https://catalyst.uw.edu/webq/survey/ernsthe/266790>
9. Optional class on programming in R on April 11th 2015 9:00 AM to 12:00 noon. Use this link: <http://uweoconnect.extn.washington.edu/datasci250>

Download Hadoop VM from Dropbox

Download Hadoop VM from Dropbox (1)

- If you have not been able to download the VM image, then please try this link by April 9th 2015:
- https://www.dropbox.com/s/4wn01pqax3kw32x/Cloudera-Training-VM-4.2.1.p-vmware_prist2.zip?dl=0
- Paste the url in a browser (I use chrome). Make sure that the url didn't change (pdfs often introduce changes).
- You should see the file (Cloudera-Training-VM ...) associated with a download button (see picture in next slide). If you are prompted to sign in or create an account, then just close that dialog. The download button will be underneath that dialog.
- The password is (without the quotes): "**Data Science UW 2015**"
- You do not need to open or unzip the file, yet. I just wanted you to download the file so that we can use the file in week. I wanted us to have these problems (if any) now and not in 6 weeks from now when we need it.

Download Hadoop VM from Dropbox (2)



Click on the Download button and save the file to a convenient location

Download Hadoop VM from Dropbox

Introduction to Data Science