

Introduction to Data Science

Lecture 01; March 30th, 2015

Ernst Henle

ErnstHe@UW.edu

Skype: ernst.predixion

Agenda



- Social: Introductions
- Break
- Class Structure
- Data Science:
 - Today we will ease into this series by providing a very high-level overview of data science. The overview is philosophical in nature. The purpose of today's lecture is get a feel for data science.
- Quiz 01a
- Data Flow
- Break
- Tools and Concepts
- Business Perspective

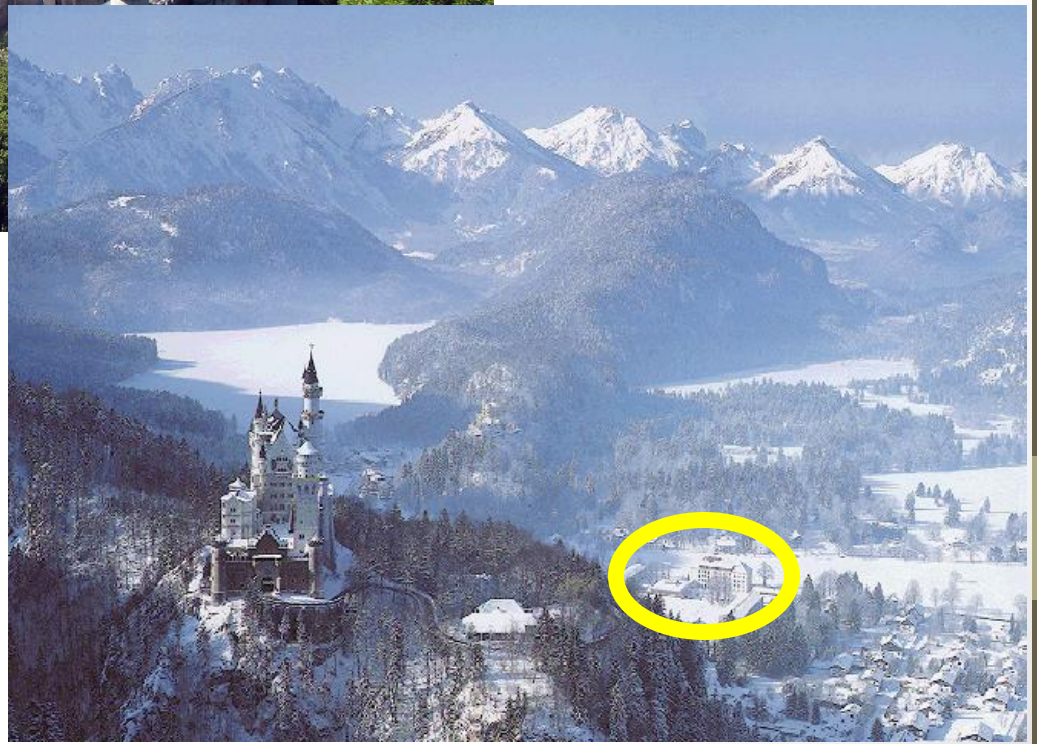
Introductions (1)

Welcome to Data Science

- A major component of this certificate program is the social component.
- Let's introduce ourselves.
 - Name
 - Professional interests and interest in data science
 - Something personal

Introductions (2) My Background

- Ernst Henle
 - Email: ErnstHe@UW.edu
 - Skype: ernst.predixion
- My interests
 - I use Physics, Math, and chemistry to solve problems in Medicine
 - As an Experimental Scientist
 - My Ph.D. is in Biophysics. I studied: Ageing, Radiation Therapy, DNA Damage, and Free Radicals
 - As a "Data Scientist"
 - Drug discovery using Genomics, Proteomics, and Metabolomics (In drug discovery and medicine, these -omics are the equivalence of "data science")
 - Adapt BI tools for medicine: Dataflow and Predictive Analytics
- Personal
 - I grew up in Bavaria and lived close to the castle that was the model for Disneyland's castle.



Introductions (4) Scientists as Data Scientists



Source: Reproduced by permission of Charles Peattie and Russell Taylor, www.alexcartoon.com.

Introductions (5)

- Name
- Professional interests and interest in data science
- Something personal

Introductions (6) Welcome to Data Science

- A major component of this certificate program is the social component.
- Visit the LinkedIn group: **Data Science UW 2015**. Comment on an ongoing discussion or create a new discussion (<http://www.linkedin.com/groups/Data-Science-UW-2015-8269996>)



Break

Class Structure (0)

Class Structure (1)

- Attendance
 - Attend at least 8 of the 10 Lectures
 - Attendance is registered by quizzes for all students.
 - Attendance is registered by chat participation for on-line students
 - Attendance means that you are present for all of the class.
- Assignments
 - You must complete at least 6 of the 8 assignments.
 - A late assignment counts for maximally half. If you cannot hand in your assignment on time, then please notify me before the deadline.
 - Assignments are due by Saturday 11:00 PM for full credit.
 - Assignments should be done **collaboratively**. You will get more out of your assignment if you do it with a fellow student.
 - Assignments must be submitted individually by each student to the catalyst web site.

Class Structure (2)

- Class Prerequisites include:
 - UW NetID
 - Access to the Internet
 - Access to Catalyst
 - Ability to use R/R Studio, Octave, and other programs on your Computer
 - Ability to run VMWare or VirtualBox on your Computer
 - Ability to participate in the LinkedIn group called **Data Science UW 2015** (<http://www.linkedin.com/groups/Data-Science-UW-2015-8269996>)
 - Access to Catalyst (<https://catalyst.uw.edu/>)
 - Ability to submit homework to the Catalyst drop box called "**Data Science UW 2015 Homework Submission**" (<https://catalyst.uw.edu/collectit/dropbox/ernsthe/35087>)
 - Ability to get class resources before each class from: "**Data Science UW 2015 Resources**" (<https://catalyst.uw.edu/workspace/ernsthe/49742/>)
 - Ability to take quizzes in class on catalyst like: "**Data Science UW 2015 Quiz 01a**"

Class Structure (3)

- Misc
 - Optional class on programming in R on Saturday April 11th 2015 from 9:00 AM to noon. Use this link:
(<http://uweoconnect.extn.washington.edu/datasci250>)
 - Office hours by appointment. If requested, I can also establish office hours at a specific time each week. I use Skype.

Approximate Course Agenda

	Lecture Dates	Homework Due	Topic
1	March 30, 2015	April 4, 2015	What is Data Science?; Basic Terminology; Data Movement; Tools
2	April 6, 2015	April 11, 2015	Data Preparation; Data Curation; R and MATLAB
3	April 13, 2015	April 18, 2015	Machine Learning; Normalization
4	April 20, 2015	April 25, 2015	Data Structures; K-Means in R
5	April 27, 2015	May 2, 2015	Real World Predictive Analytics; Statistics; Sampling
6	May 4, 2015	May 9, 2015	Statistics Evaluation; RDBMS with Relation Algebra
7	May 11, 2015	May 16, 2015	Relational Algebra cntd; Data Storage Concepts; Cloud Computing (AWS or Azure)
8	May 18, 2015	May 30, 2015	CAP Theorem; No SQL; Scalability; Sparse Matrices and EAV
9	June 1, 2015		Hadoop Sqoop Hive MapReduce Hue
10	June 8, 2015		Hadoop; MapReduce cntd; Graph Data, SPARQL, Page Rank

Optional class on R: April 11th 2015 9:00 AM to noon

Data Science

What is Data Science?

- My answer:
 - Data science is the generalization of the scientific method
- Why is data science a “new” discipline?
 - An abundance of data outside of the traditional sciences
 - Tools to investigate these data

Science Past vs. Future

- Past
 - Science used to be about devising the best experiment to verify a specific hypothesis.
 - Data acquisition was coupled to a specific hypothesis

Science Past vs. Future

- Past
 - Science used to be about devising the best experiment to verify a specific hypothesis.
 - Data acquisition was coupled to a specific hypothesis
- Future
 - The abundance of data has led to a new paradigm
 - Data is ubiquitous
 - We need methods to sift through the data and extract meaning.
 - Data acquisition will support many hypotheses

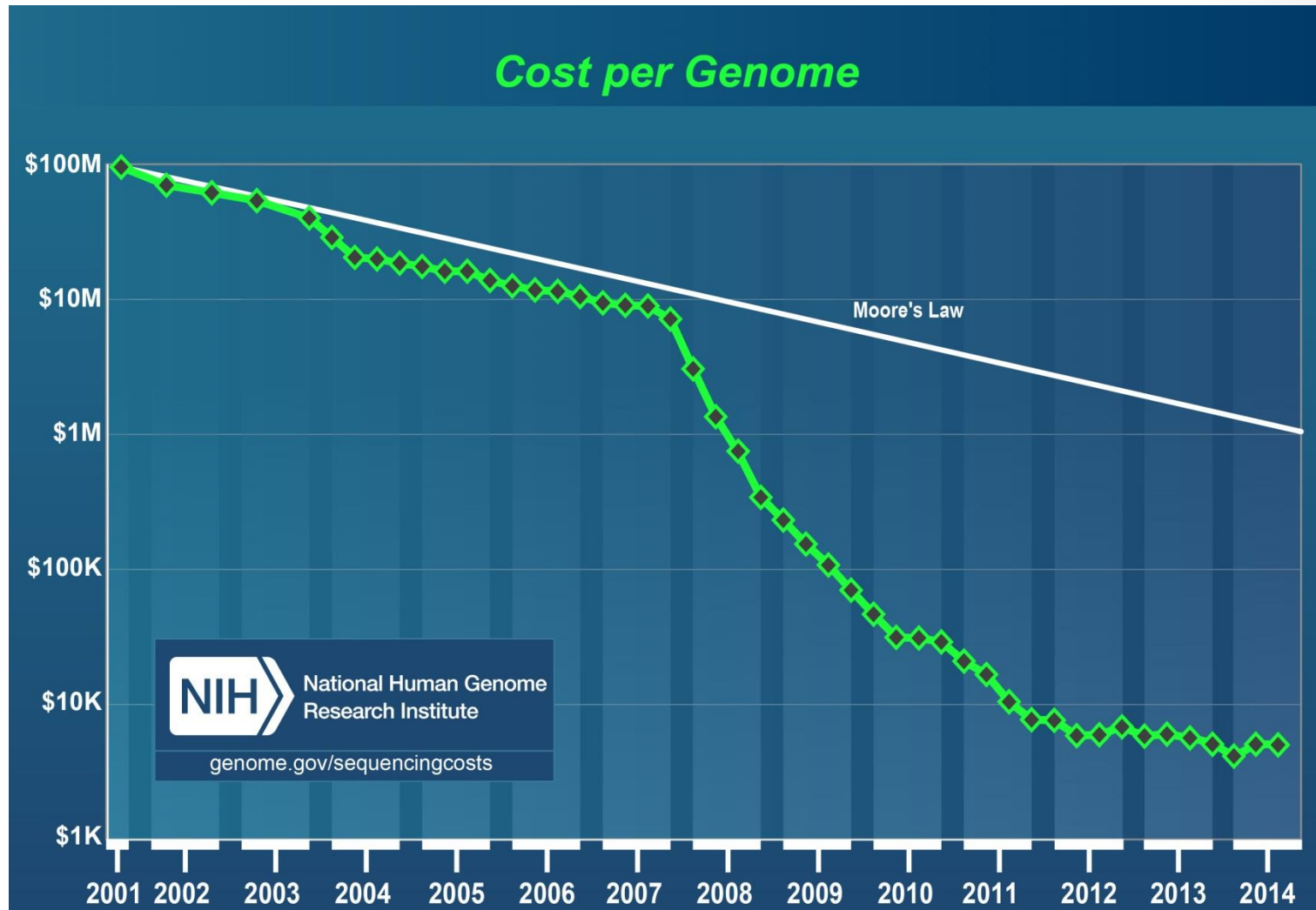
Example: Genomics

- First Human Genome
 - Time: 12 years (1990 – 2002) to sequence first genome
 - Cost: 3 Billion \$
 - Use in Science: Too expensive to devise an experiment that requires whole-genome sequencing

Example: Genomics

- First Human Genome
 - Time: 12 years (1990 – 2002) to sequence first genome
 - Cost: 3 Billion \$
 - Use in Science: Too expensive to devise an experiment that requires whole-genome sequencing
- Today
 - Time: 24 hours
 - Cost: \$1000.
 - Use in Science: Most investigators do not even need to pay for sequencing since enough sequences already exist

Example Genomics



The Scientific Method

The Scientific Method

1. A hypothesis is formulated that explains observations


The Scientific Method

1. A hypothesis is formulated that explains observations
2. The hypothesis is tested
 - A verified hypothesis constitutes a theory

The Scientific Method

1. A hypothesis is formulated that explains observations
2. The hypothesis is tested
 - A verified hypothesis constitutes a theory
3. New observations are considered in light of the theory

The Scientific Method

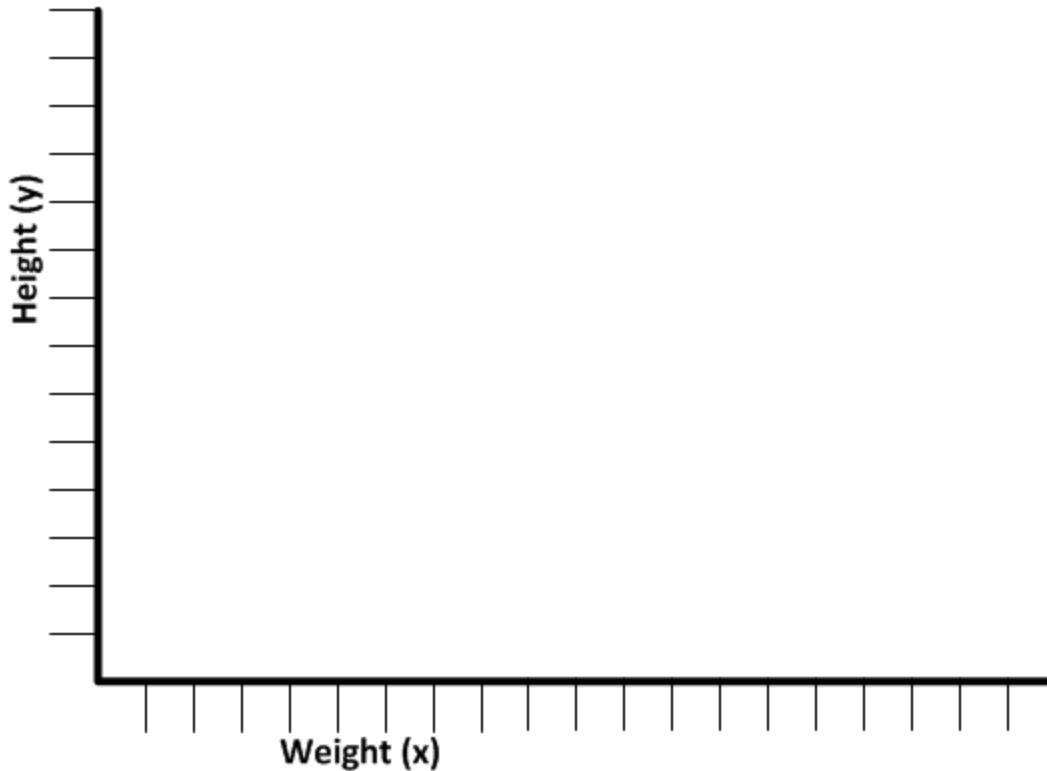
- 
1. A hypothesis is formulated that explains observations
 2. The hypothesis is tested
 - A verified hypothesis constitutes a theory
 3. New observations are considered in light of the theory
 - When the theory is insufficient to account for the new observations, then we formulate a new hypothesis and cycle starts again.

The Scientific Method in Data Science

1. A data scientist wants to answer a business question.
 - E.g. "Can we predict hospital readmissions based on patient data?"
 - Question as a task like: Please, predict hospital readmissions!
 - Question as hypothesis: Inpatient hospital readmissions at Mercy General can be predicted with more than 85% certainty using decision trees based on available patient data.
2. The hypothesis is tested
 - Data transformations
 - Transformations are verified by accuracy assessments
3. If the hypothesis is verified ask more questions like:
 - Why are patients being readmitted?
 - Can we prevent readmissions?
 - How can we increase our accuracy from 85% to 90%.

Model as a Hypothesis

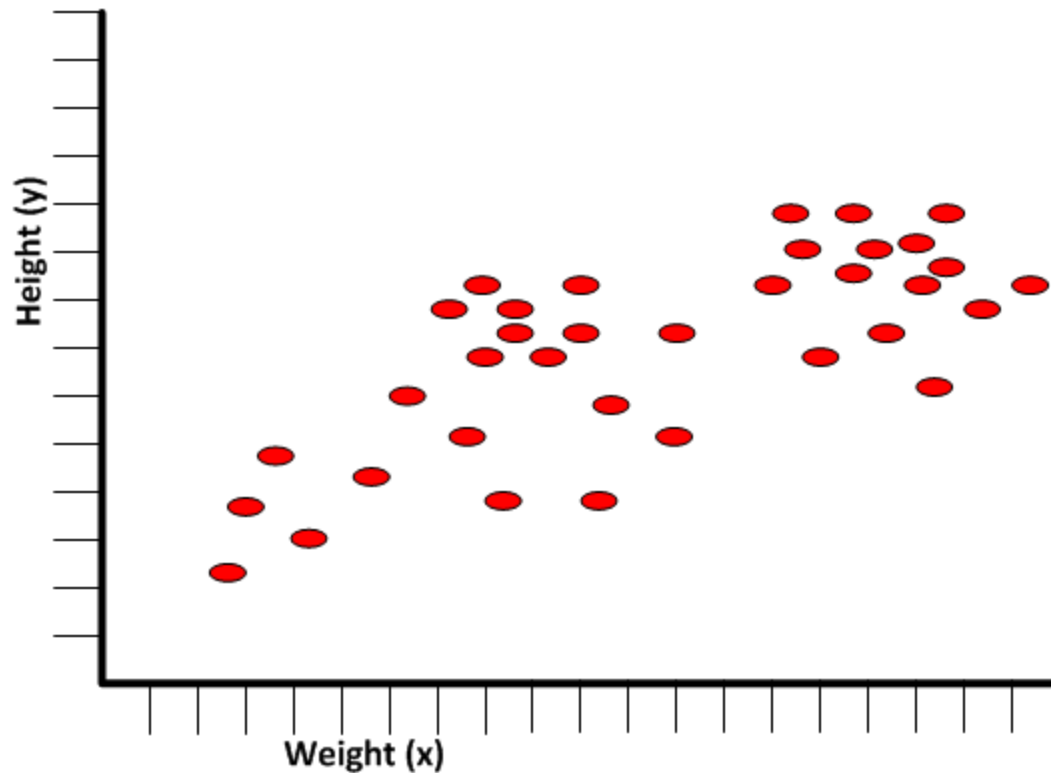
Model as a Hypothesis



Schema, Space

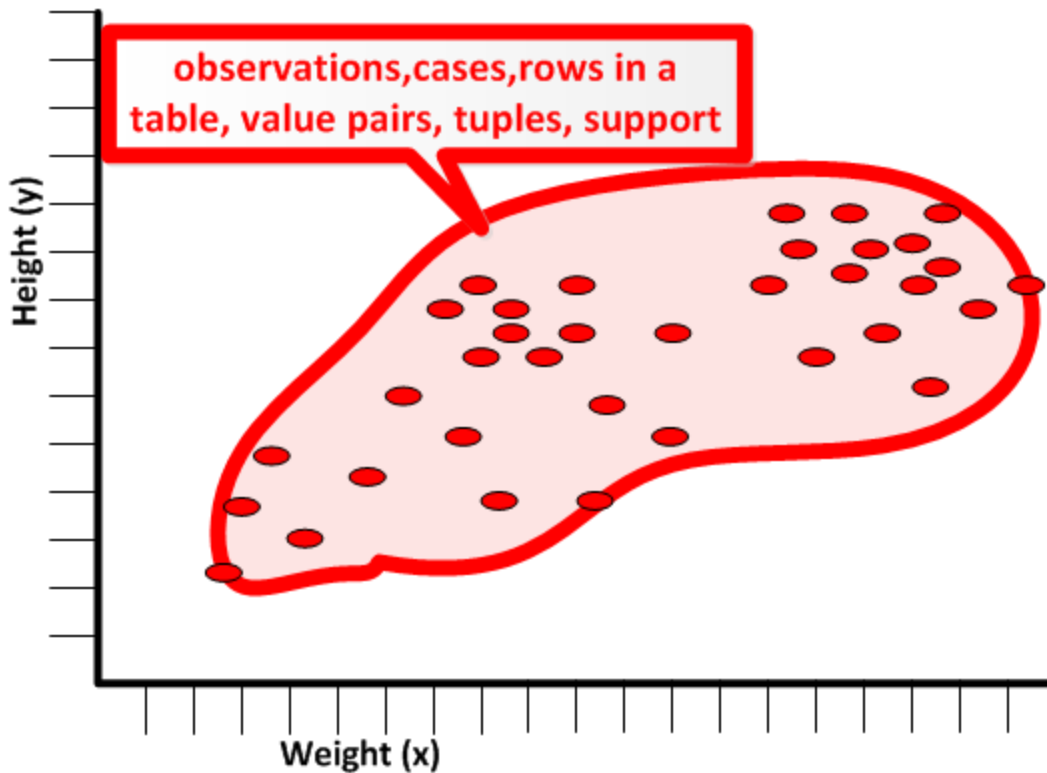
This is a two dimensional space of
weight and height

Model as a Hypothesis



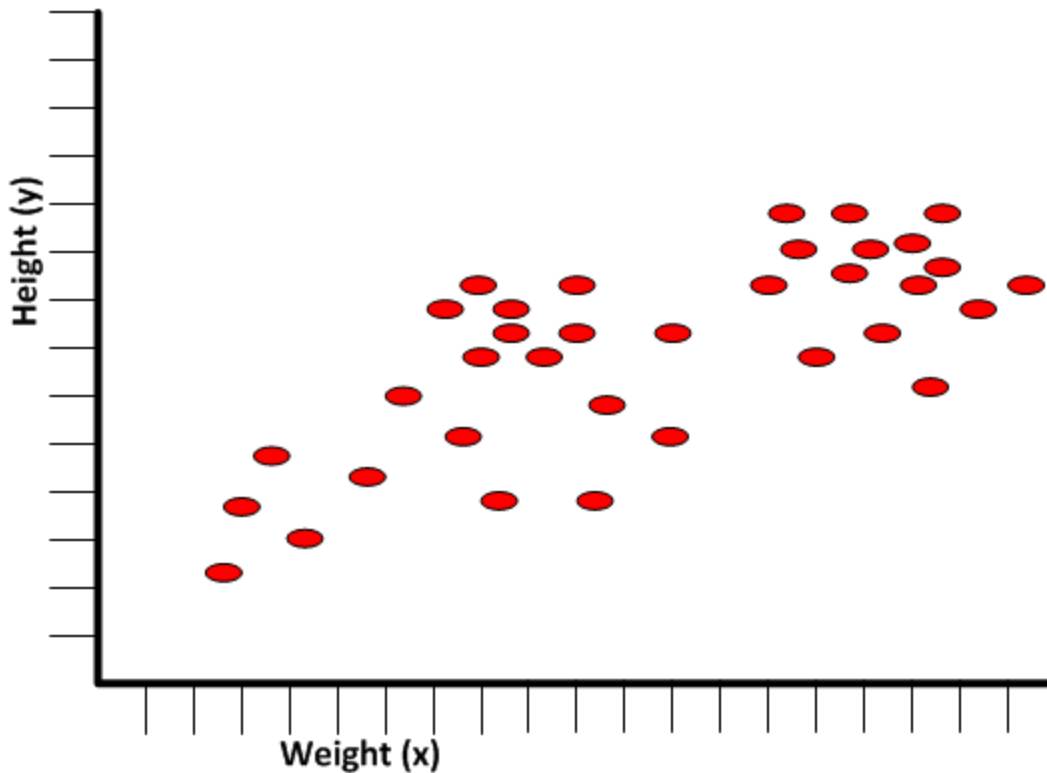
Data: These data represent observed people.

Model as a Hypothesis



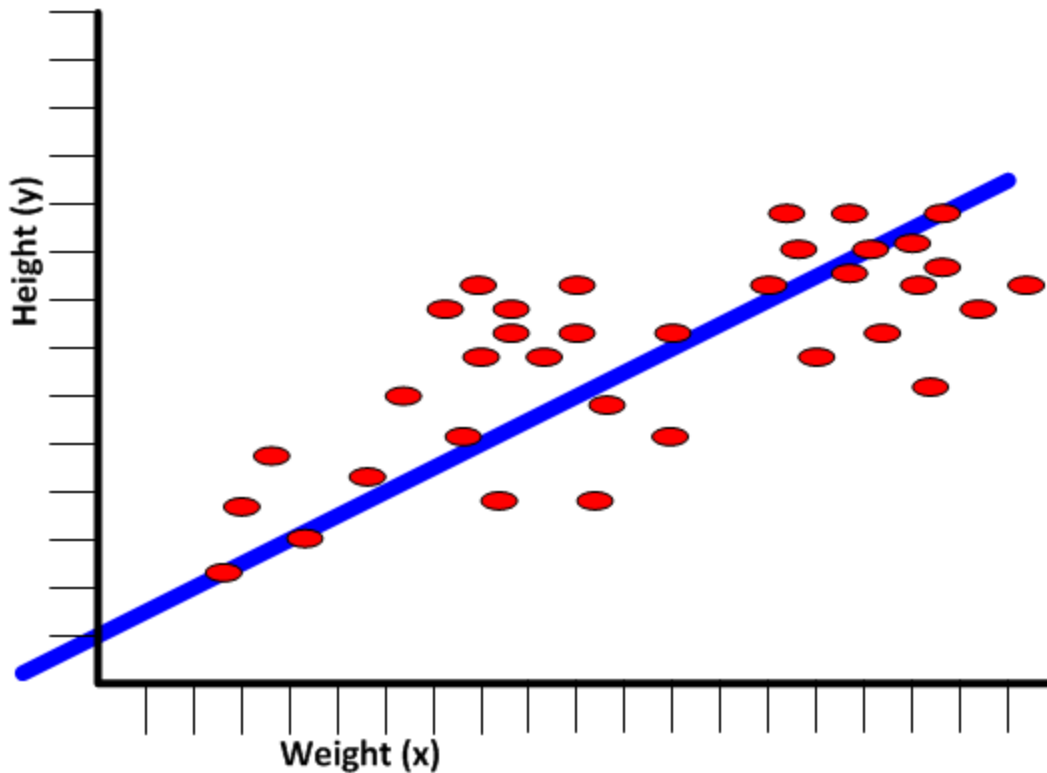
Data: These data represent observed people. The data can be represented as a point in two-dimensional space

Model as a Hypothesis



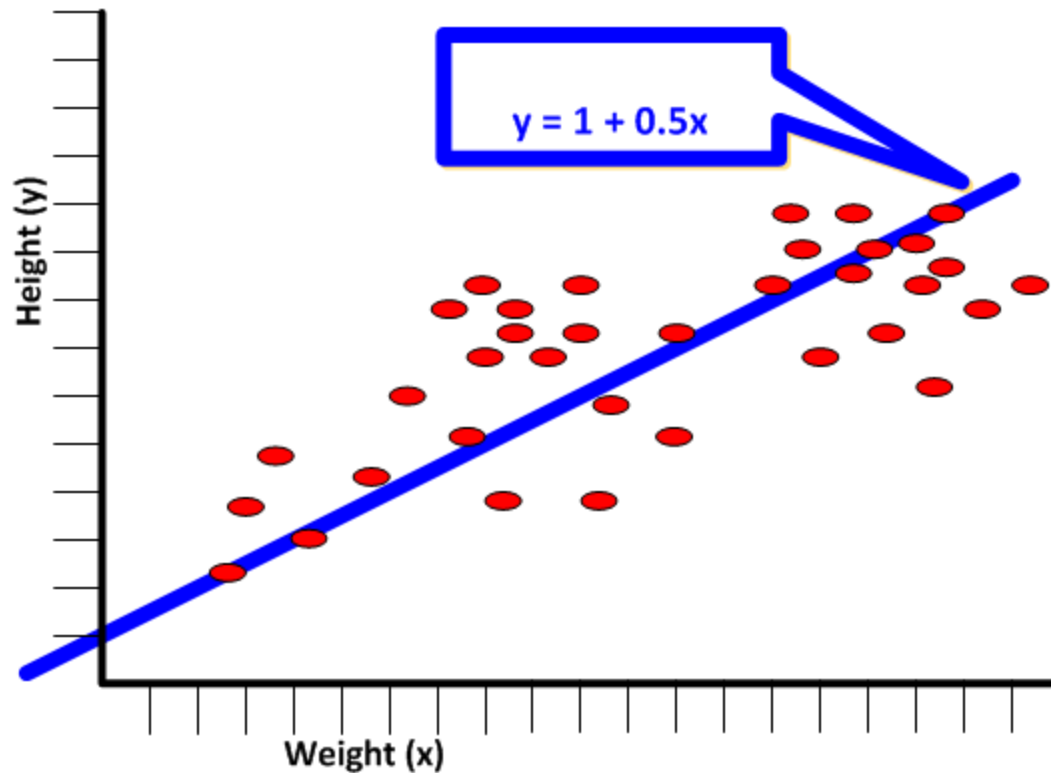
Use Data to Create Hypothesis

Model as a Hypothesis



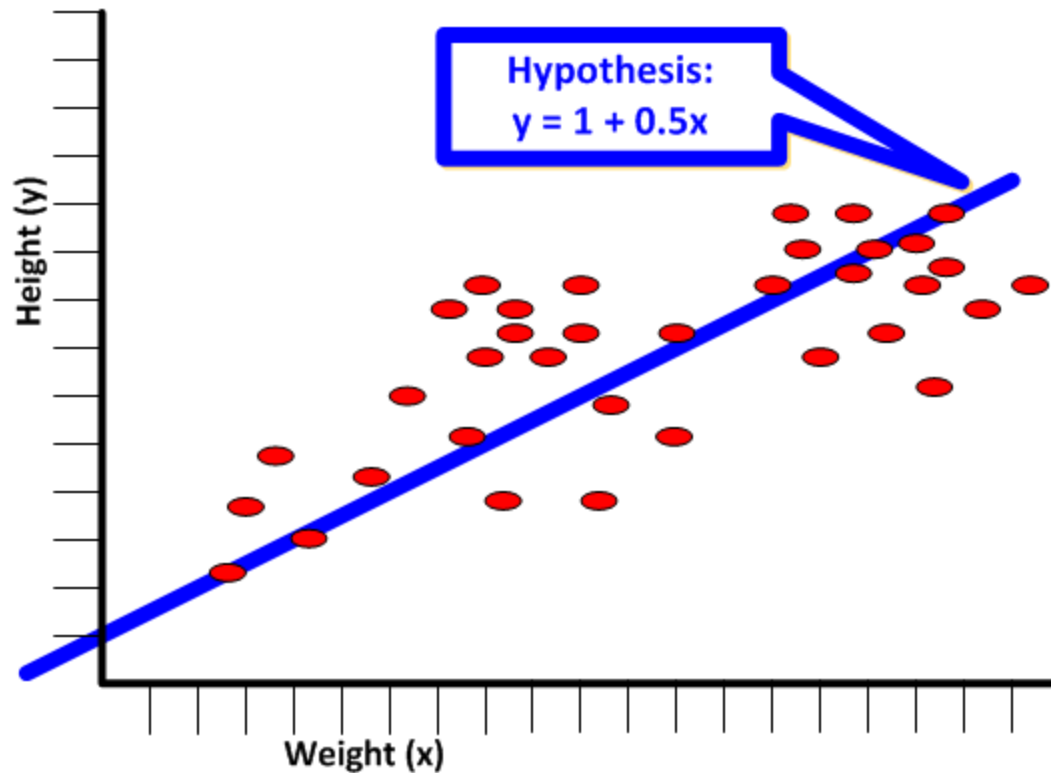
The thick blue line represents the best fit through these data.

Model as a Hypothesis



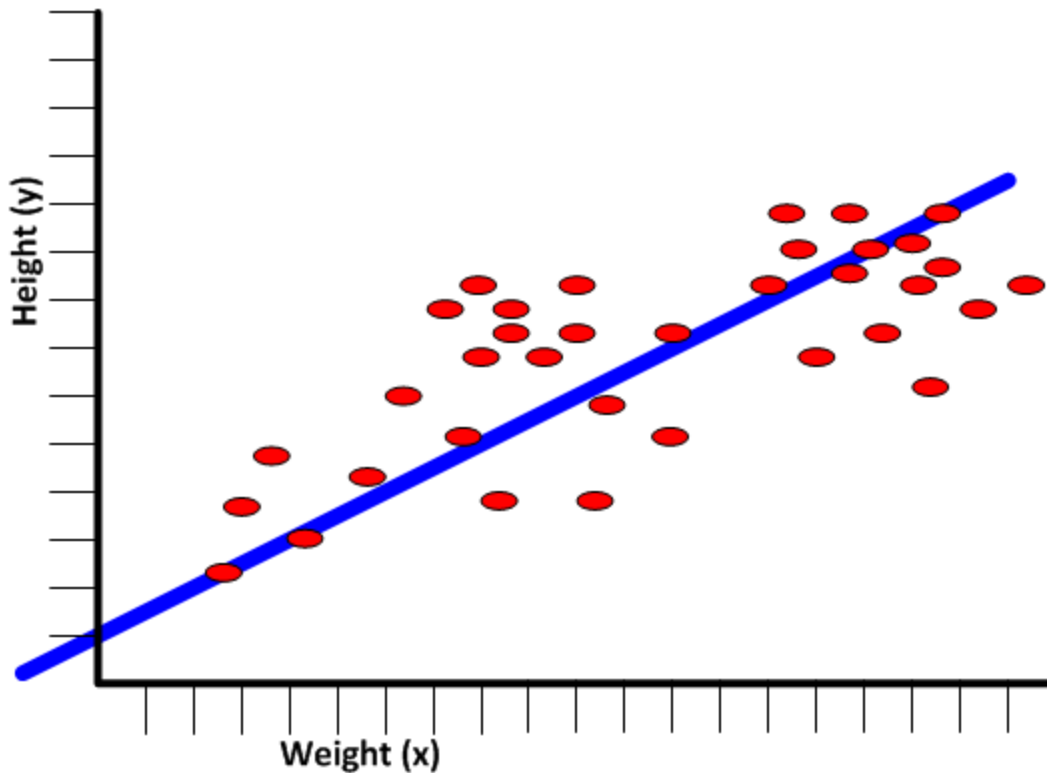
The line can be represented algebraically.

Model as a Hypothesis



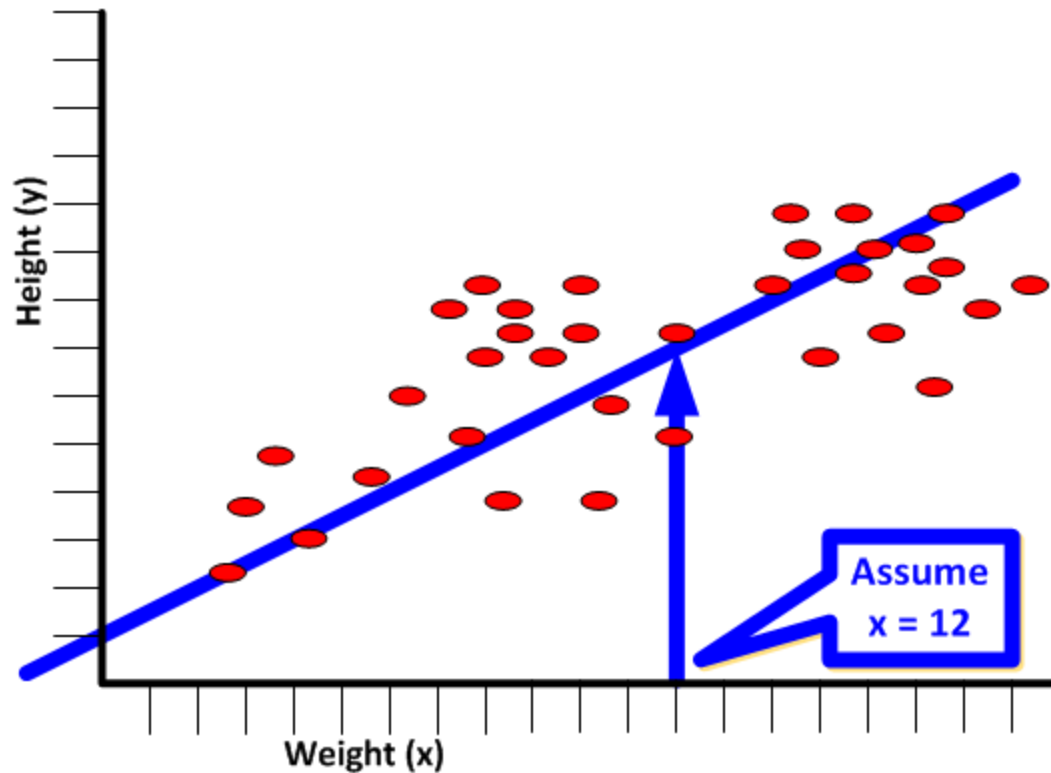
The hypothesis can be represented algebraically. The hypothesis is the best fit line
 $y = 1 + 0.5x$

Model as a Hypothesis



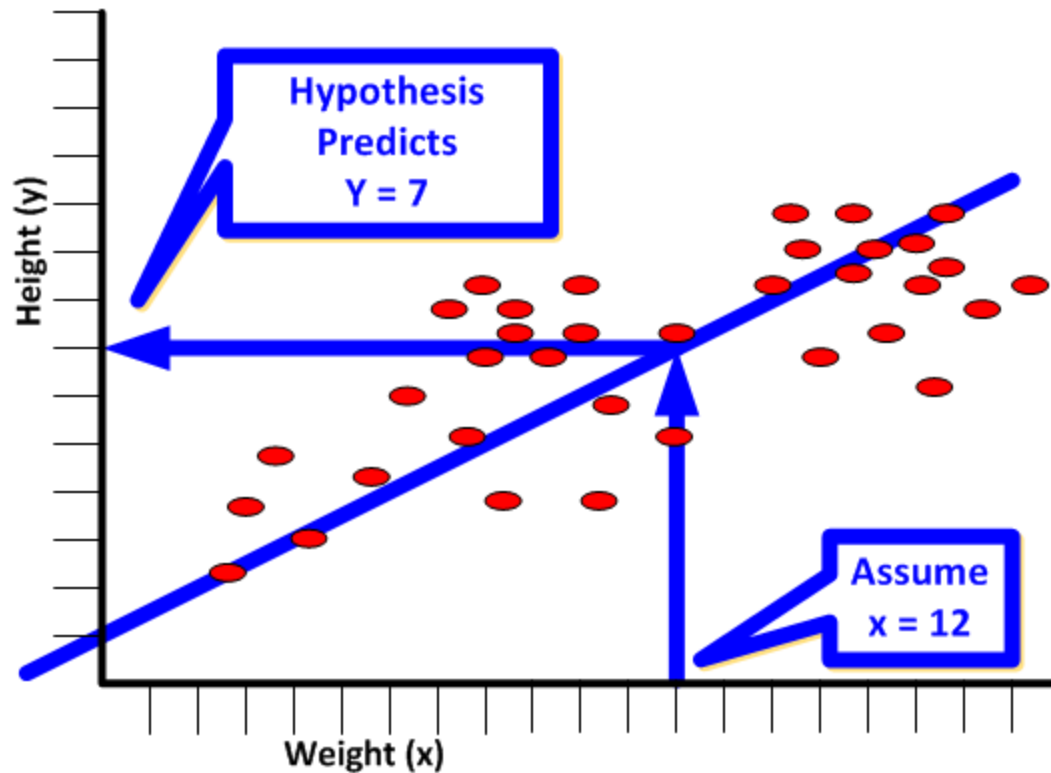
Use Hypothesis can be used to predict

Model as a Hypothesis



Use Hypothesis can be used to predict. A prediction requires an input.

Model as a Hypothesis



The prediction is an output.
Given an x value we can predict a
 y value. If $x = 12$, then $y = 7$.

Some Terminology

- What is a model?
 - A model is a hypothesis based on data and a method (algorithm)
- What is a hypothesis? (<http://en.wikipedia.org/wiki/Hypothesis>)
 - Wikipedia: A proposed explanation for an observation that can be tested
 - A hypothesis is an explanation for the organization of a data set that allows a prediction
- Falsification (<http://en.wikipedia.org/wiki/Falsifiability>)
 - Falsification is the process that attempts to disprove a hypothesis
 - If a hypothesis is not falsifiable then it is not really a hypothesis
- What is a Theory?
 - A fact-based explanation for an observation (a well-tested hypothesis)
- What is a Law?
 - A prediction with no exceptions
 - A law does not attempt to explain the predictions as a theory would

What is Data?

- Data are observations that are put into context
- Given that every observation has a context, an observation is a datum
 - Not Data:
 - 1 (one)
 - Chair
 - Diabetes
 - Data:
 - I see a chair
 - The patient has diabetes

Unstructured Data?

- Unstructured data does not exist.
 - The essence of data is that they are structured.
 - The context is what makes data.
- What is meant by unstructured data?
 - Answer: Insufficiently structured data
- For example: A list of tweets is often used as an example of unstructured data.
 - But they are well structured:
 - The tweets are organized into a list
 - Any single tweet comes from one source at one time
 - A tweet is a text with a length constraint.
- Such lists may be insufficiently structured to easily derive conclusions.
 - Need to process the data (parsing, etc.)

Data Structure leads to Data Types (1)

- For example: a list of tweets
 - The data type is “a list of tweets”
 - A list inherits many characteristics of lists in general
 - A tweet inherits many characteristics from the data type text
- Typing makes Data
 - Typing is the context
- A list of tweets can be represented as a table
 - The column header provides context
 - The table structure states that all column values have comparable structures

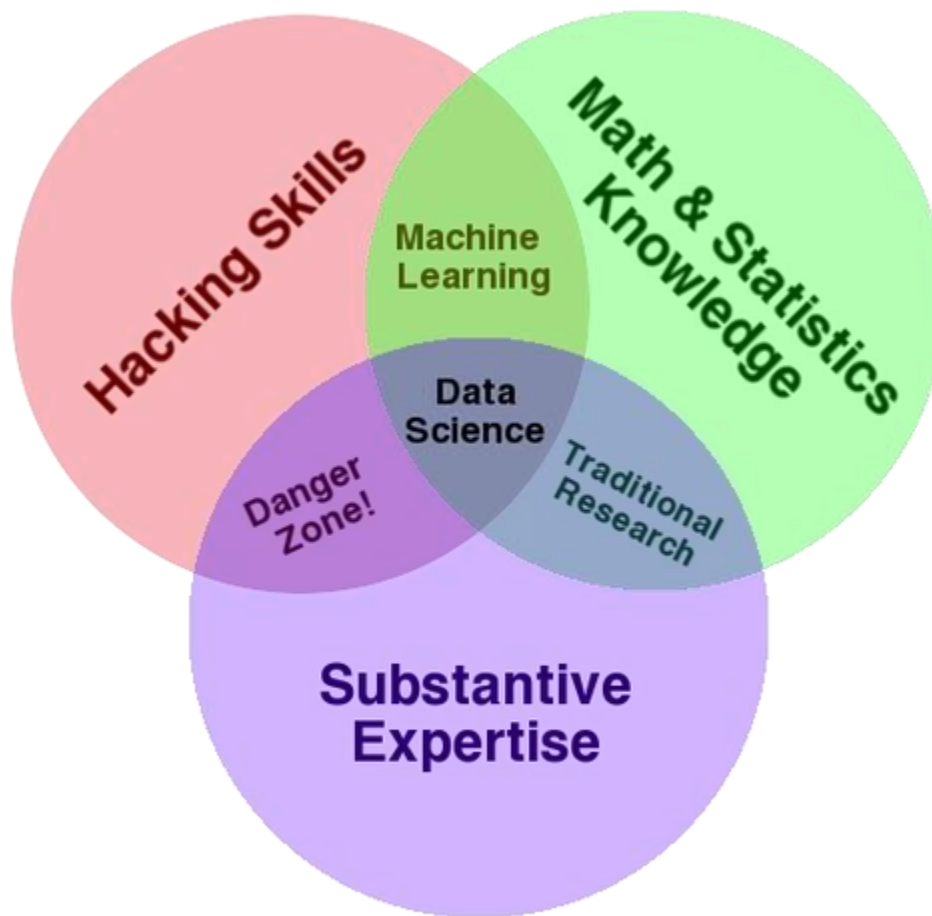
Data Structure leads to Data Types (2)

- Physics
 - A data type is called a unit
 - Data are well-typed and can be universally converted
- Computer Science
 - Typing may demand structure: Strong Typing
 - http://en.wikipedia.org/wiki/Strong_typing
 - Structure and context may determine typing: Weak typing
 - http://en.wikipedia.org/wiki/Weak_typing

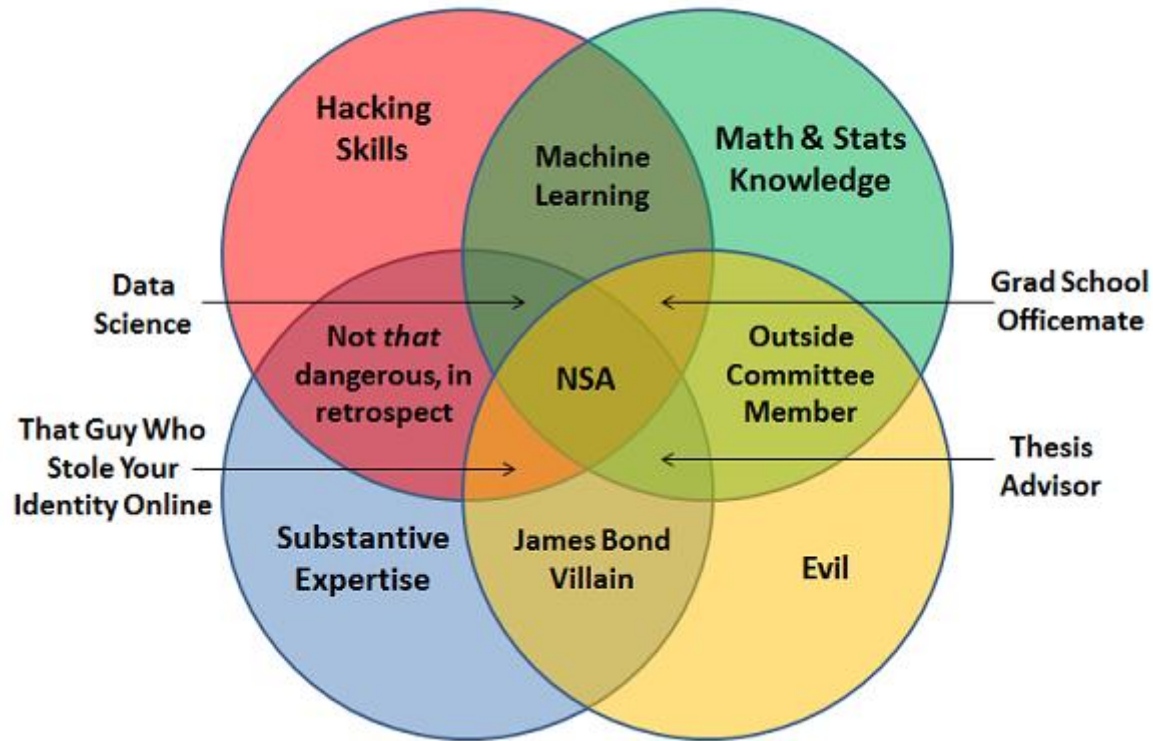
Data Science

- What is Data Science?
- Data Science is made of two words: **Data** + **Science**
 - **Data** and their structures are well explained in computer science
 - The **Science** part of data science is explained by the scientific method.
 - The synthesis of these two disciplines allows
 - Data Visualization
 - Data Extraction
 - Data Processing / Transformation
 - Hypothesis Verification or Falsification

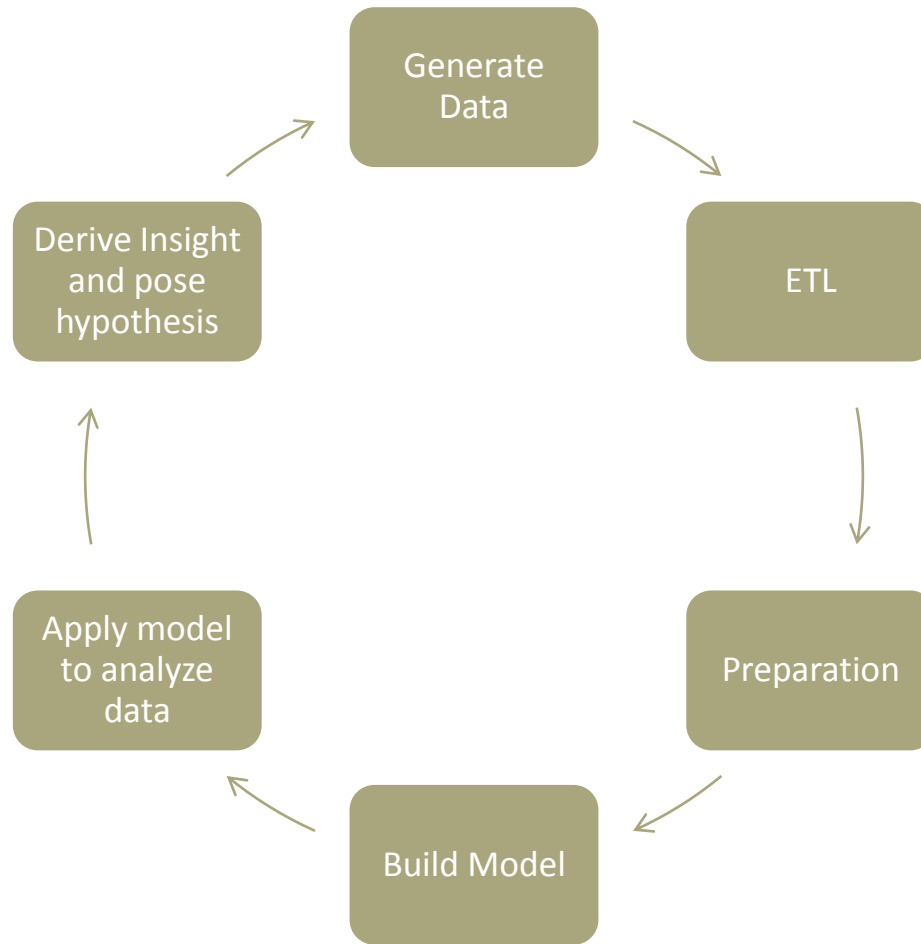
Drew Conway's Venn Diagram



Another Venn Diagram



Data Science Cycle



Data Science Today and Tomorrow

	Today	Tomorrow
Generate Data	Survey or create physical experiment. Ask: How do I generate data from which I can derive an insight?	Take the data from a web source. Ask: How do I find the data that is interesting to me from the data I already have access to?
ETL	Data is structured due to data generation design	Data are poorly structured since purpose of data was not anticipated when data were generated.
Preparation	Find ways to properly format poorly prepared data	Disregard poorly prepared data, because there is adequate well prepared data
Build Model	Classify and associate data via complex models.	Model complexity is traded in for model scale (parallelization, etc.)
Apply model to analyze data	Model is used to score local data.	Apply the model to data that is not local
Derive insight and pose hypothesis	Insight based on limited amounts of data and thus conclusions have limited validity.	Insights are based on large amounts of data and thus conclusions are generally valid.

Data Science Links

- <http://www-01.ibm.com/software/data/infosphere/data-scientist/>
- <https://datajobs.com/what-is-data-science>

Quiz 01a

- Quiz URL:
<https://catalyst.uw.edu/webq/survey/ernsthe/266065>
- **If you cannot access catalyst, then please send me an email now (within 5 min). I will send the quiz to you later.**

In-Class Assignment

- In "Data Science UW 2015 Homework Submission" there is a submission site for assignment "Lecture 00". Submit a text file named test.txt that contains your name. The link is:
<https://catalyst.uw.edu/collectit/dropbox/ernsthe/35087>

Data Flow (0)

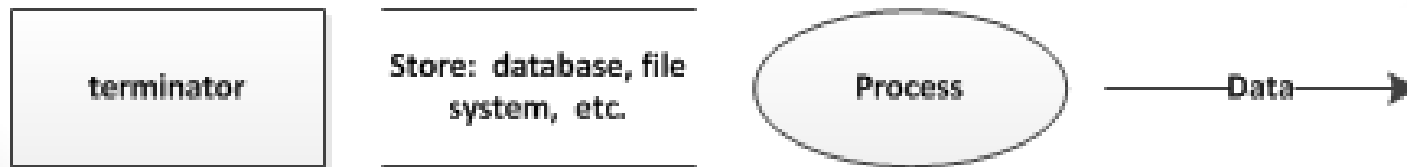
Data Flow (1)

- Data Flow is required for Data Processing
- SSADM specifies Data Flow Diagrams (DFD)
 - http://en.wikipedia.org/wiki/Data_flow_diagram
- Four components of a DFD:
 - Terminator
 - Store
 - Process
 - Data Flow



Data Flow (2) DFD

- These diagrams are called Data Flow Diagrams (DFD). A benefit of DFDs are that they are a defined language in the structured systems analysis and design method (<http://en.wikipedia.org/wiki/SSADM>). DFD is a particularly good language for describing processes that involve movement and transformation of data. Dataflow diagrams (DFD, http://en.wikipedia.org/wiki/Data_flow_diagram) define processes and do not necessarily represent components. DFDs processes are easily related to development tasks.
- An Ellipse represents a process that transforms data.
- An arrow represents data.
- The rectangles without sides are stores, like databases.
- The complete rectangles are starting or terminating processes that either generate or consume data.



Data Flow (3)

Image Aggregation Story

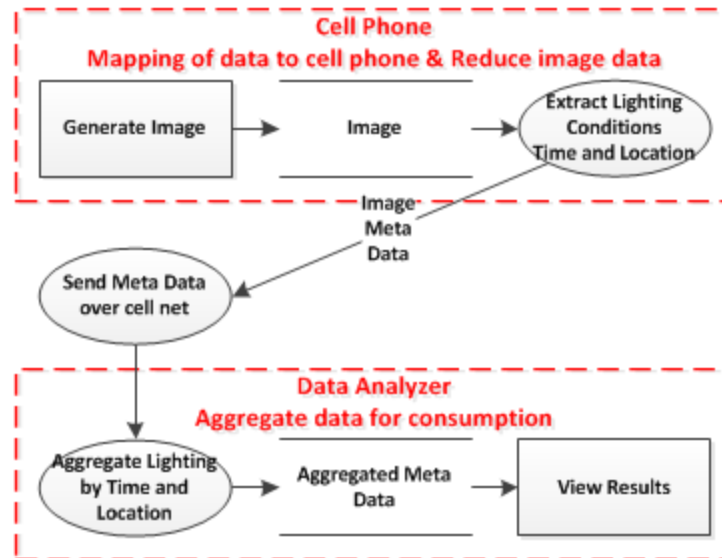
1. Describe, in a few sentences, a data science task that interests you. The following is one that interests me:
 1. Data are extracted and processed from images on cell phones
 2. The processed data are combined
 3. The combined data are used to derive meaning, like: Which are the popular tourist locations?
2. Construct a data flow diagram that depicts the data processing that is required to complete the task in item 1

Data Flow (4)

Image Aggregation Steps

- Collect and aggregate cell phone camera images
 - The image is taken (Image is mapped to cell phone)
 - Image is associated with cell location and time
 - The image data is extracted (Data Reduction)
 - The data (Image characteristics, time, and location) are sent
 - The data are collected and aggregated by location and time
 - The data are viewed

Data Flow (5): Image Aggregation DFD



Data Flow (6): DFD Arrow

- An arrow represents data or data flow. The arrow is labeled by the name of the data. Example:



- An arrow is necessary to connect the other data flow components. Every data flow component must have at least one arrow.

Data Flow (7): DFD Arrow

- Which example is correct?

—————Eat—————→

—————Lunch—————→

—————Eat Lunch—————→

—————If Hungry
Eat Lunch—————→

Data Flow (8): DFD Arrow

- Which example is correct?

———— Eat —————>

———— Lunch —————>

———— Eat Lunch —————>

———— If Hungry
Eat Lunch —————>

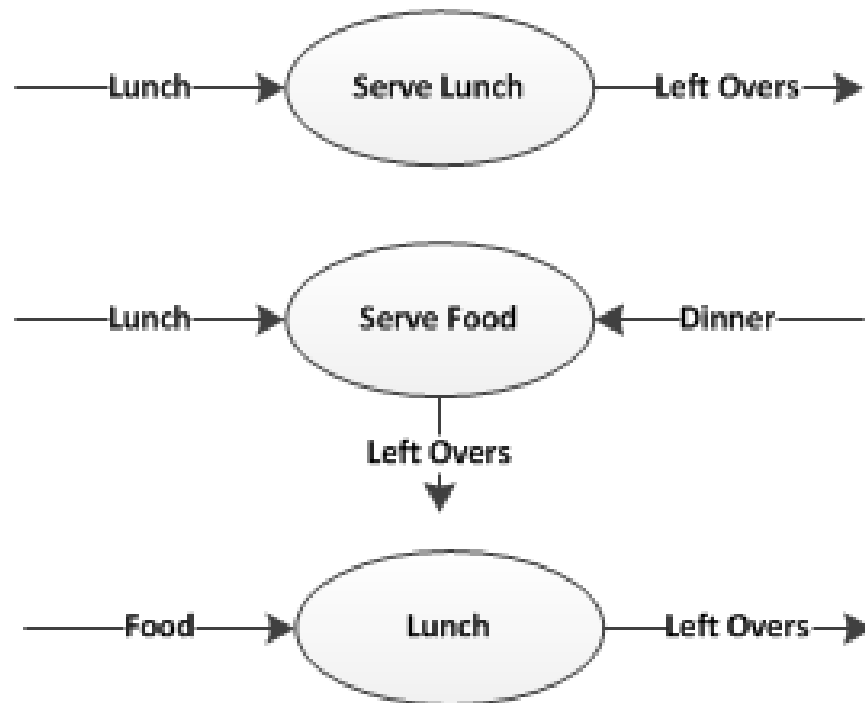
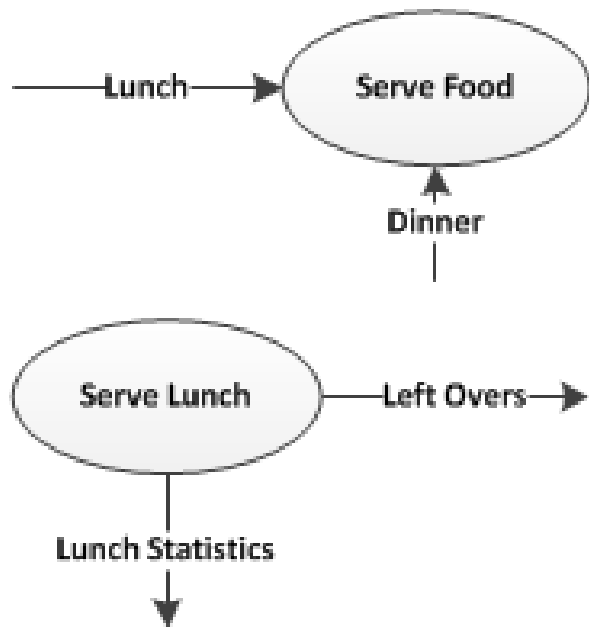
Data Flow (9): DFD Process

- A process is represented by an ellipse
- A process takes in data from one or more data sources, transforms the data, and then outputs the data.
- A process must have at least one input arrow
- A process must have at least one output arrow.
- A process is labeled with a verb, like “Brighten”
- Example:



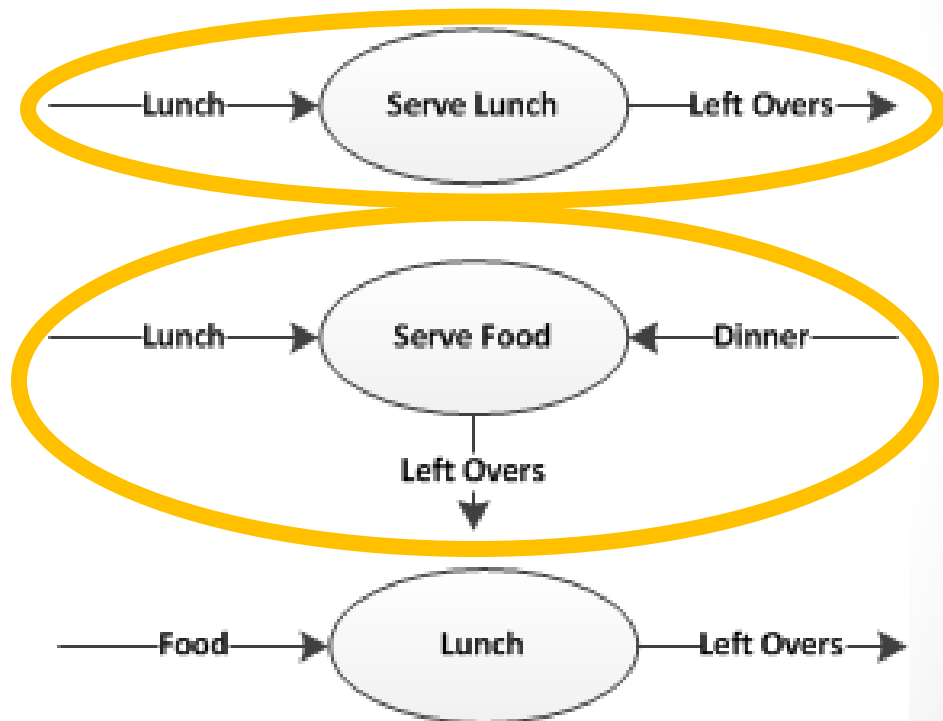
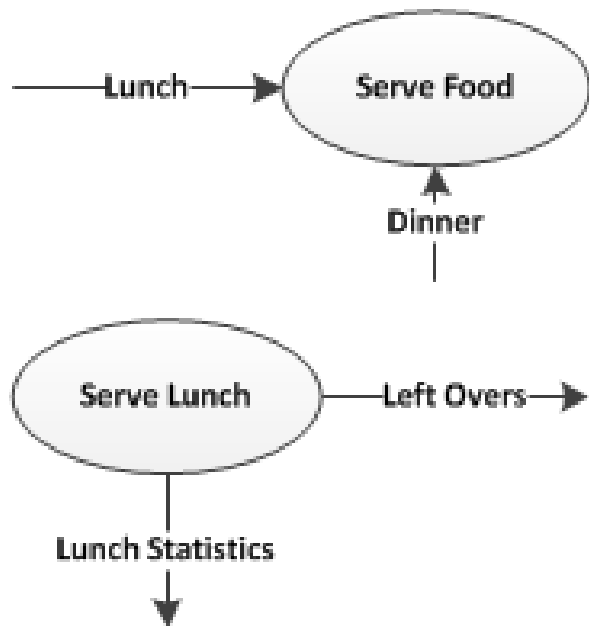
Data Flow (10): DFD Process

- Which of these are correct?



Data Flow (11): DFD Process

- Which of these are correct?



Data Flow (12):

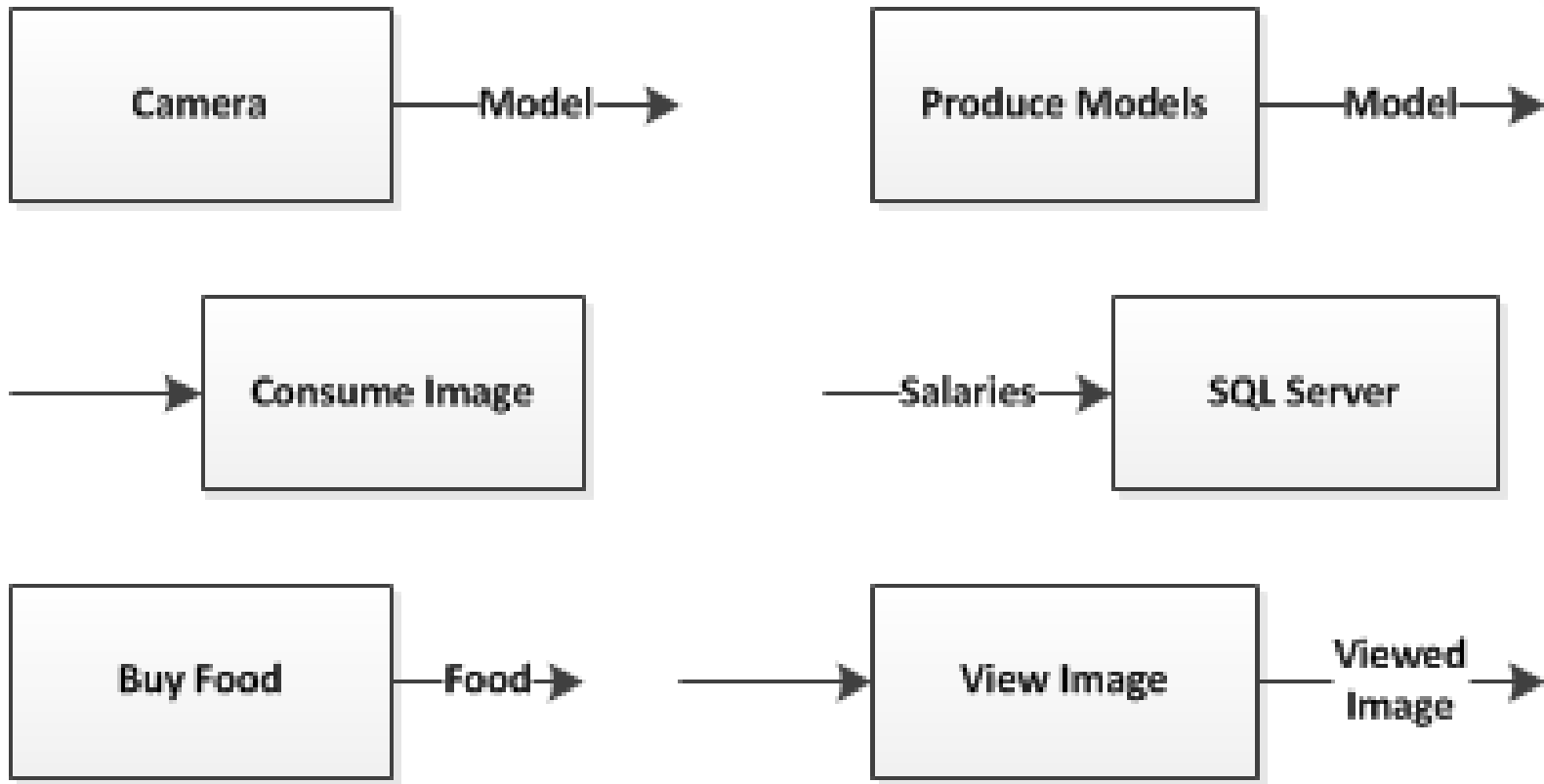
DFD Terminator

- A terminator is represented by a rectangle with all four sides drawn.
- A terminator is a process that either generates or consumes data. This process may reference a component like: Get data from Internet or View data in Monitor
- Example:



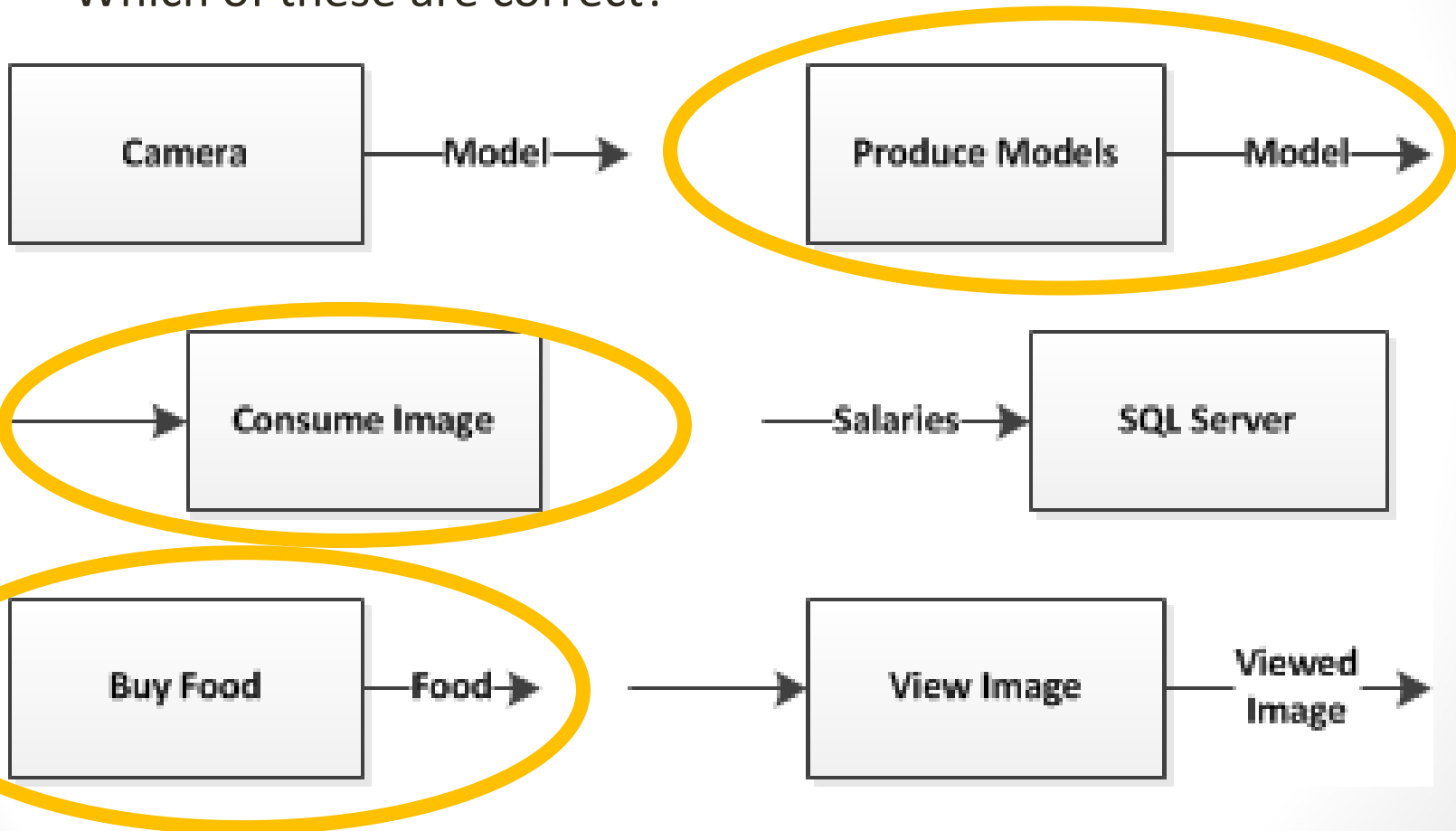
Data Flow (13): DFD Terminator

- Which of these are correct?



Data Flow (14): DFD Terminator

- Which of these are correct?



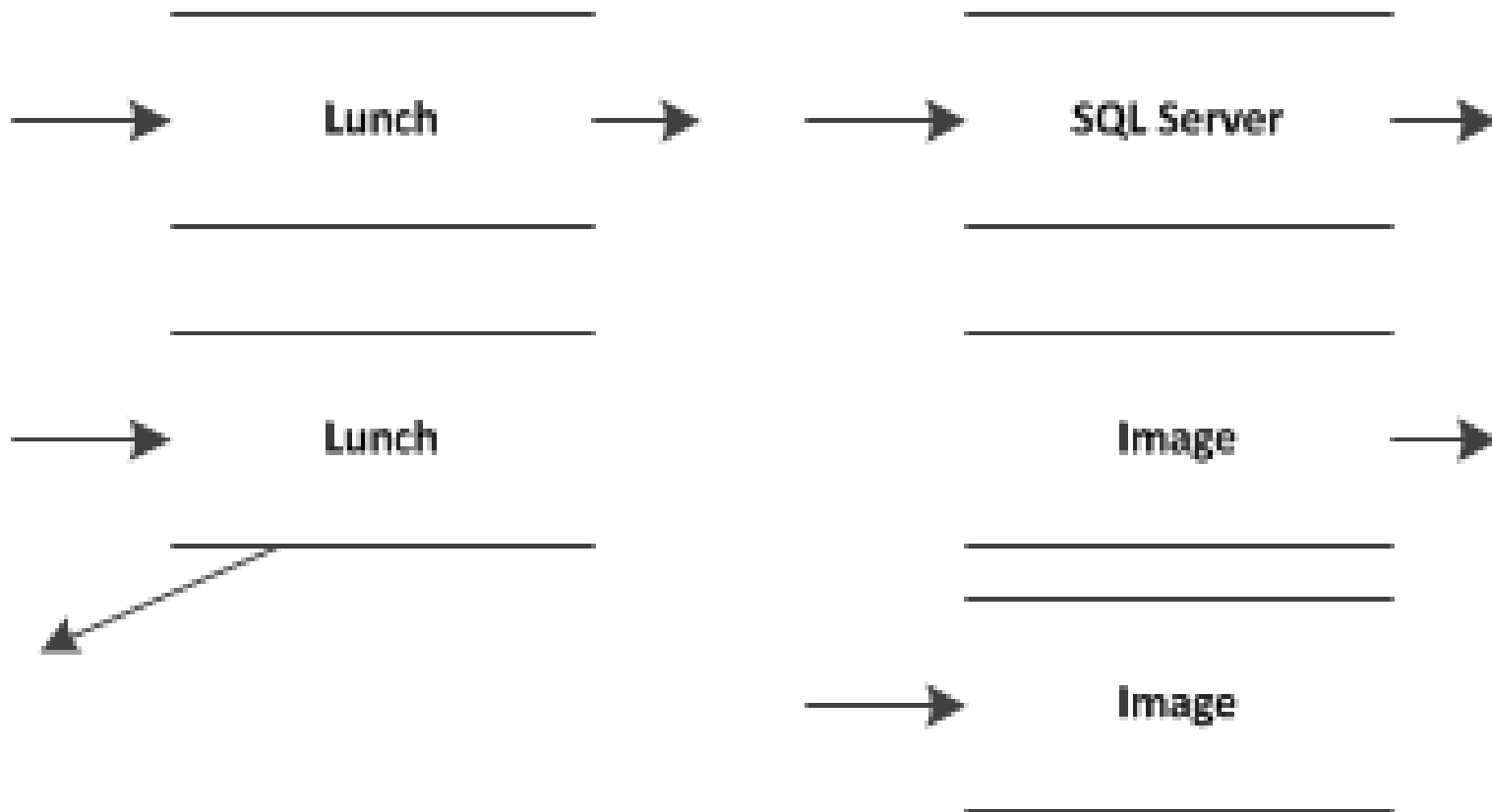
Data Flow (15): DFD Store

- Stores are represented by a rectangle that is missing the right-hand side or both the right- and left-hand sides.
- A store is a place where the data is persisted. Typical stores are text files, websites, and relational data bases.
- A store has at least one input arrow
- A store has at least one output arrow
- Typically, the input and output arrows are not labeled.
- The name of the store describes the nature of the data (not the nature of the data base)
- Example:



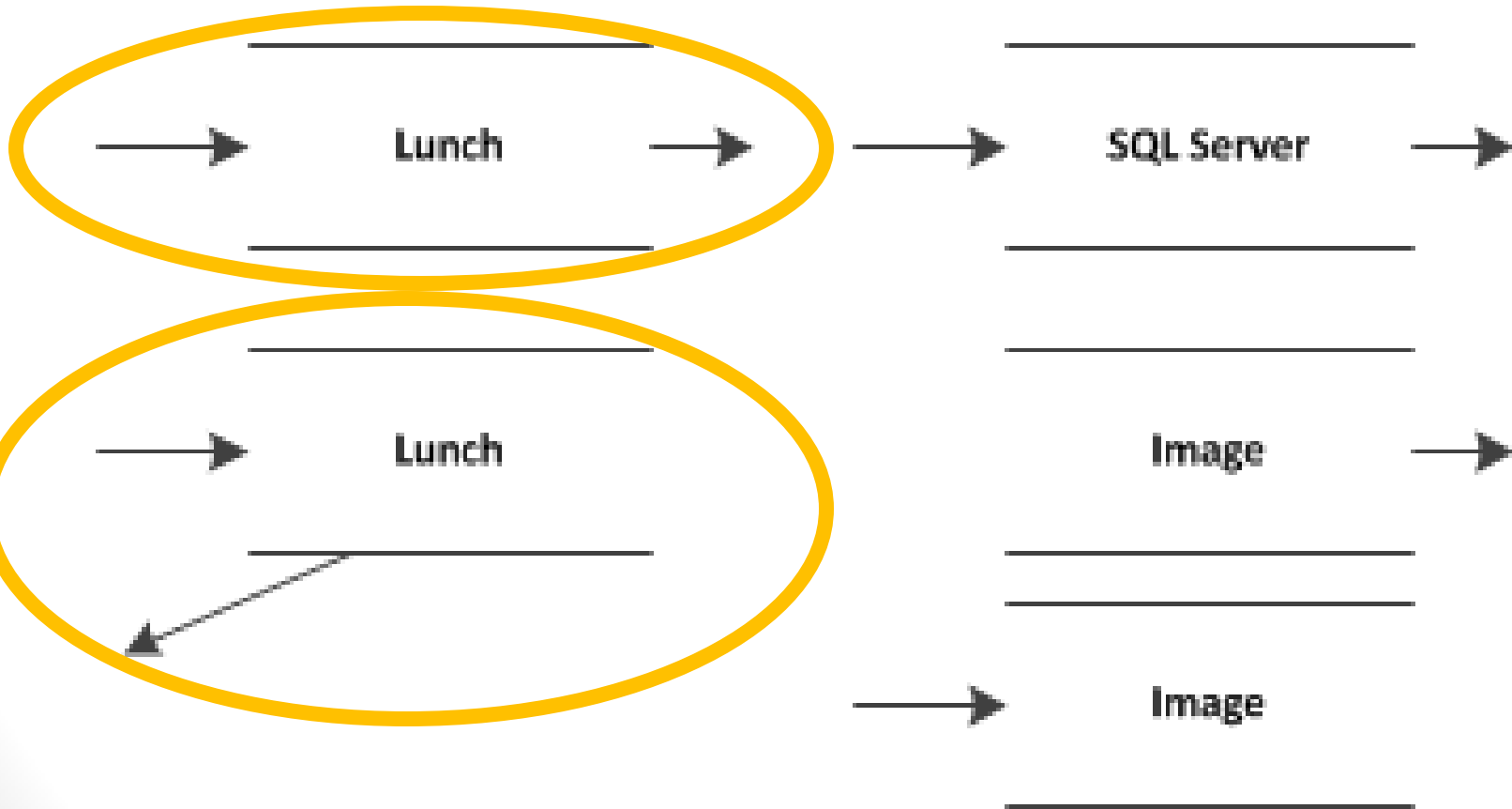
Data Flow (16): DFD Store

- Which are correct?



Data Flow (17): DFD Store

- Which are correct?



DFD: Digital Pathology (0)

- An Example

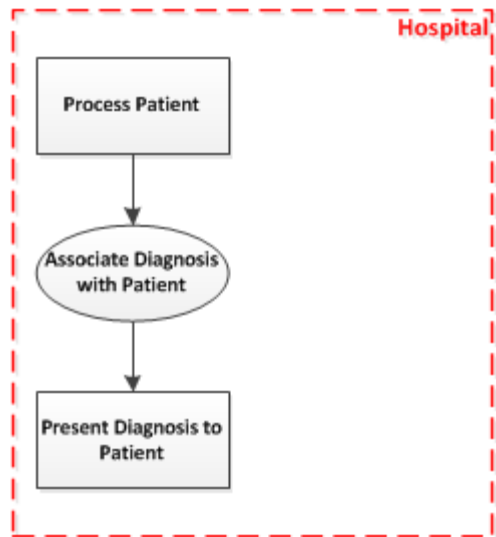
DFD: Digital Pathology (1)

Digital Pathology

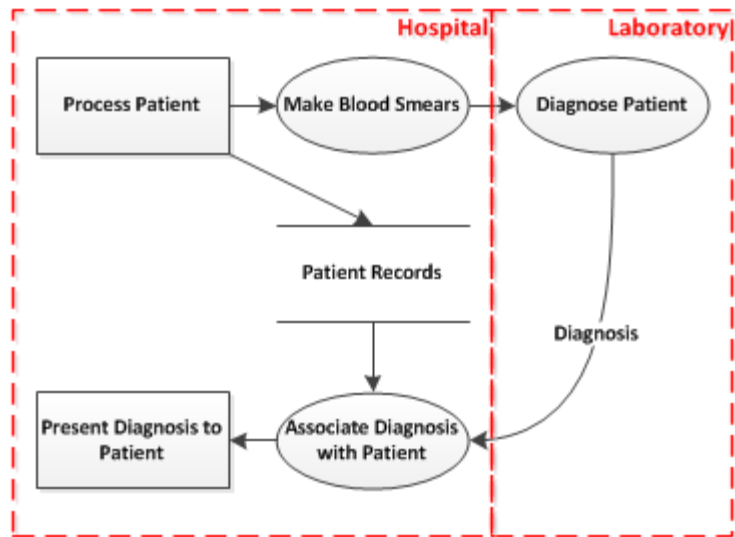
Many blood disorders manifest themselves through easily recognizable morphological changes, but the affected cells may be as few as one in a hundred thousand. Given the scarcity and cost of pathologists, it is not possible to routinely screen for these blood disorders. We would like to find an automated way of diagnosing such disorders.

We use a pathologist to score aberrant cells and correlate these findings with shape characteristics determined by image segmentation.

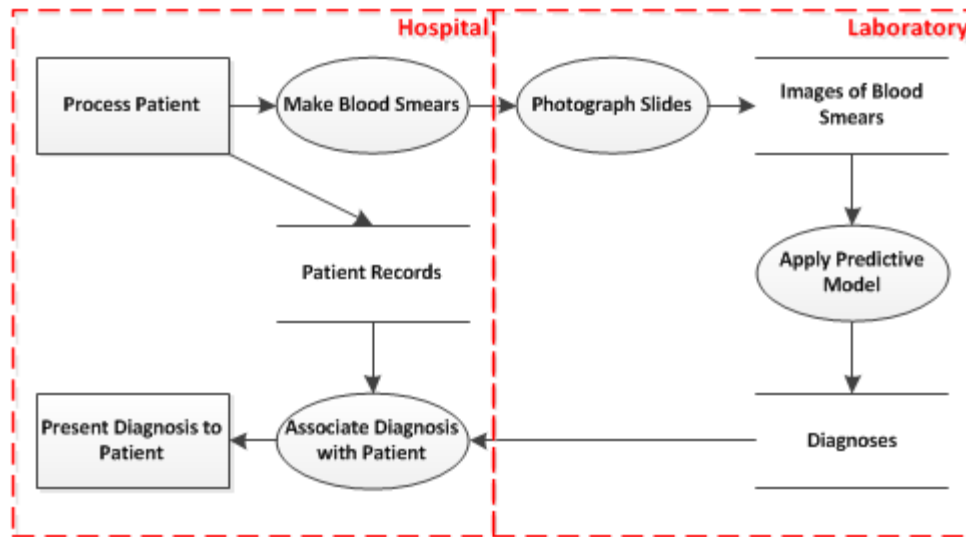
DFD: Digital Pathology (2)



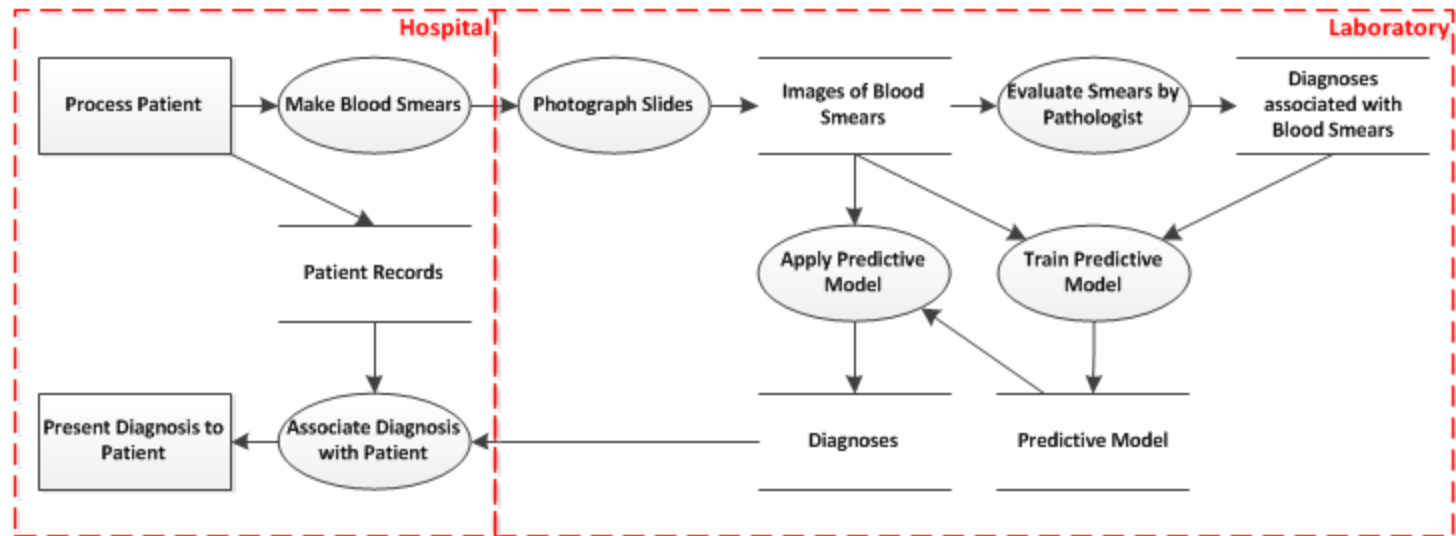
DFD: Digital Pathology (3)



DFD: Digital Pathology (4)



DFD: Digital Pathology (5)



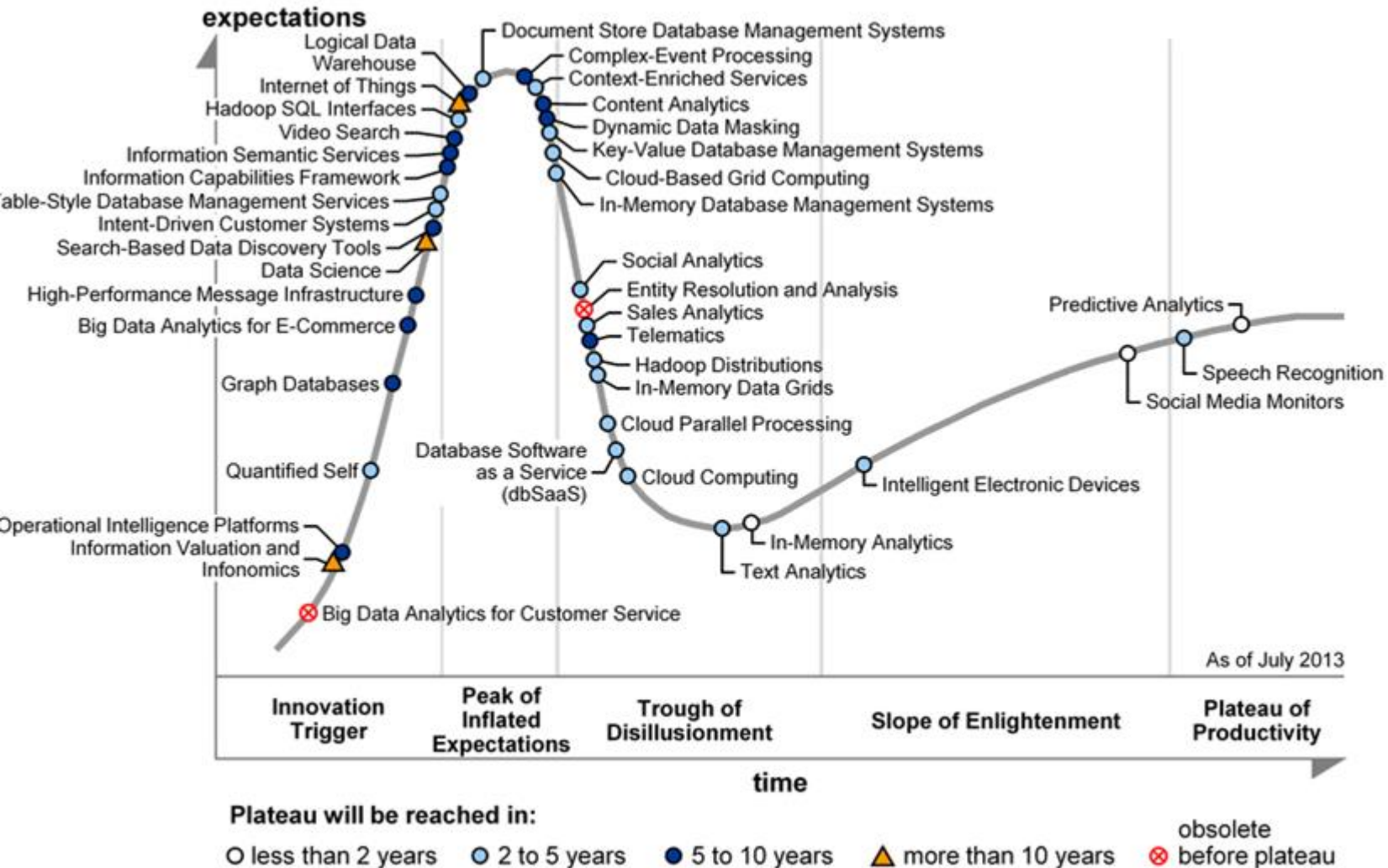
In-Class Exercise:

Diagram a DFD

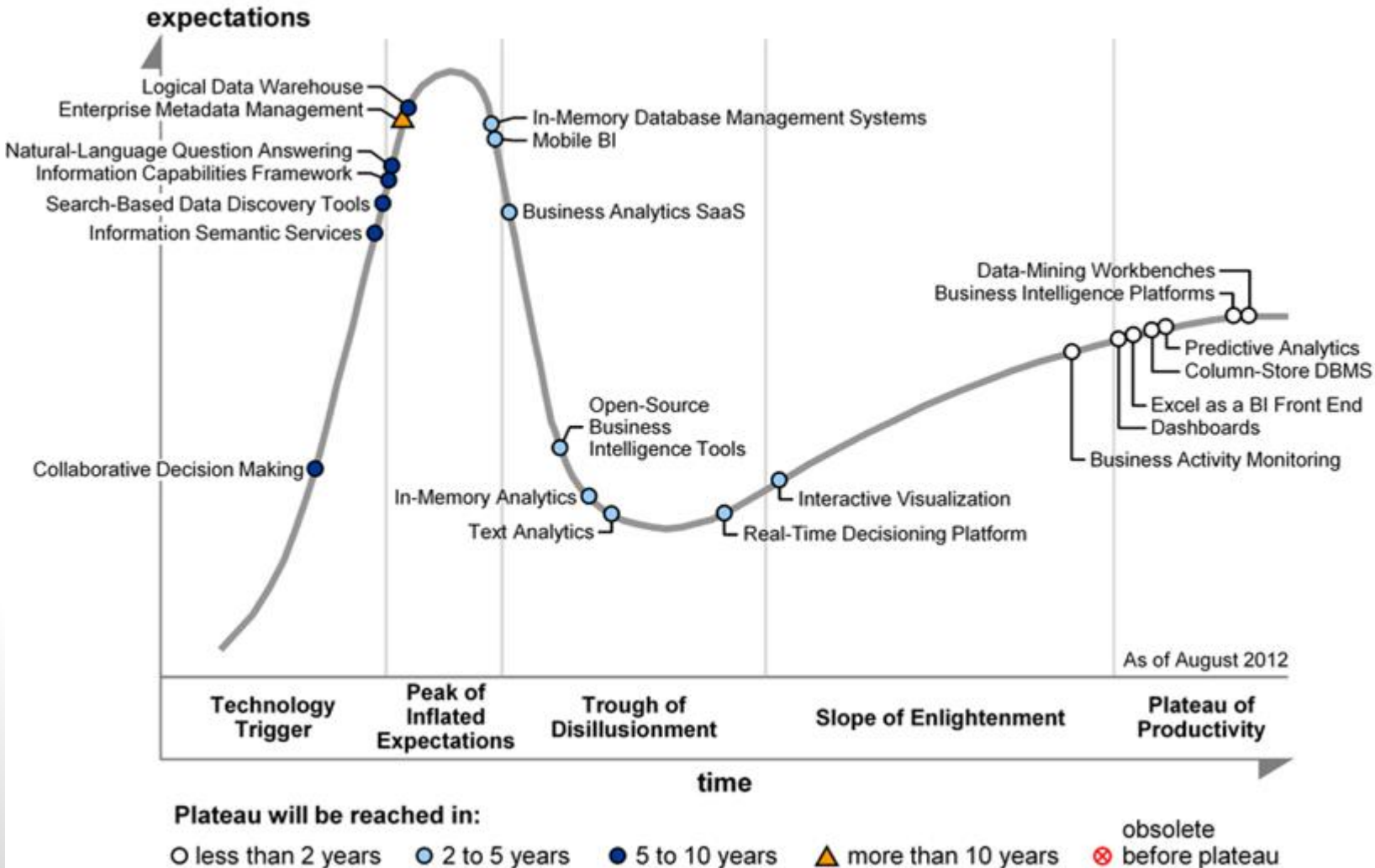
1. Describe, in a few sentences, a data science task that interests you.
2. Construct a data flow diagram that depicts the data processing that is required to complete the task in item 1

Data Science – Business Perspective

Hype Curve Big Data 2013

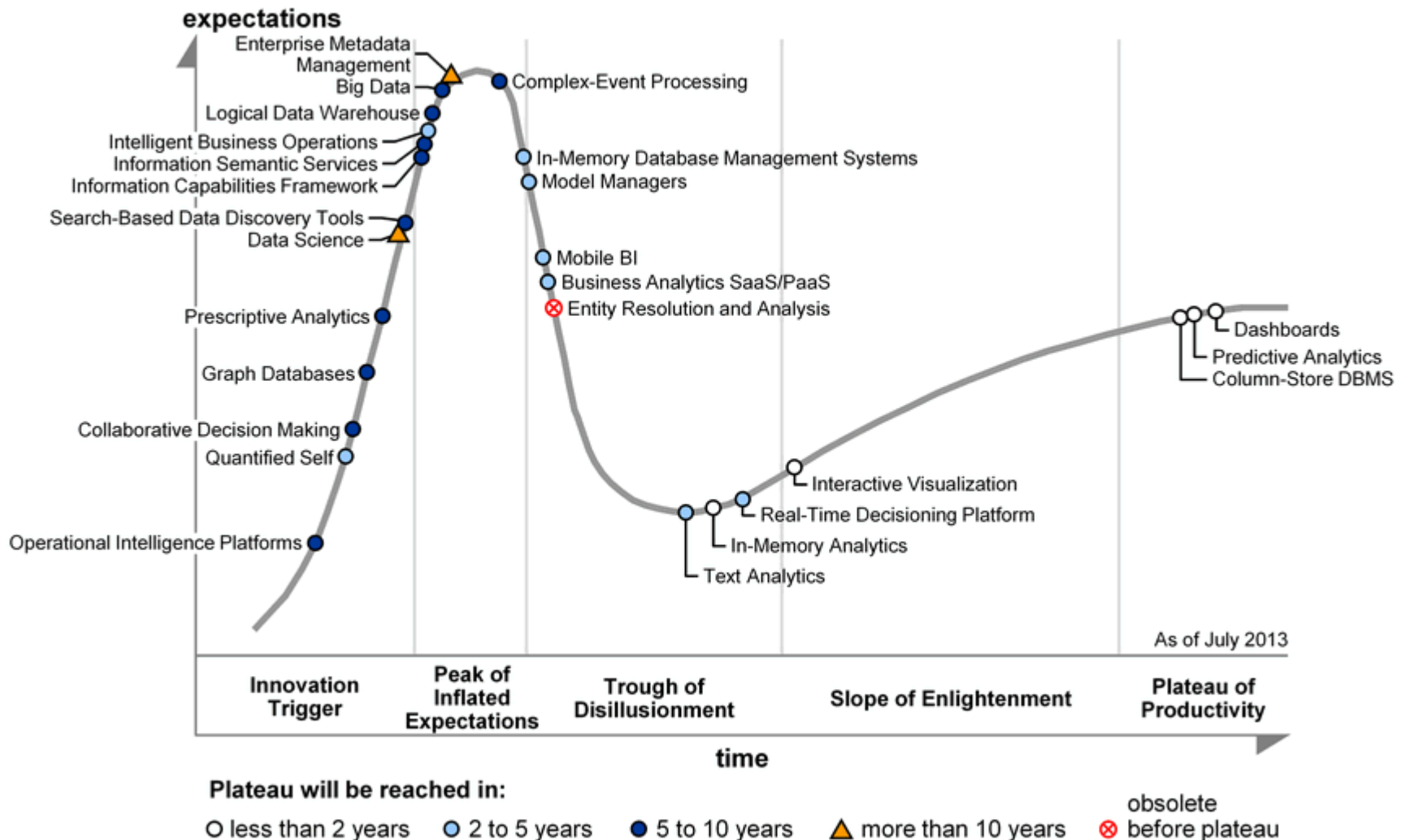


Hype Curve Business Analytics 2012



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>

We will become acquainted with these tools and concepts

- Some tools:
 - R
 - MATLAB
 - SQL Server/MySQL/Hadoop
 - SPARQL/SQL/HIVE/Hue/Impala
 - Predixion Insight
- Some Concepts:
 - DFD
 - Data Types
 - Graph Analytics
 - Relational Algebra
 - Database structures
 - Predictive Modeling
 - NoSQL
 - CAP Theorem
 - MapReduce

Some Links(1)

- Data Science
 - http://sqlblog.com/blogs/buck_woody/archive/2012/10/16/is-data-science-science.aspx
 - <http://radar.oreilly.com/2010/06/what-is-data-science.html>
 - http://en.wikipedia.org/wiki/Data_science
- Data flow Diagram
 - http://en.wikipedia.org/wiki/Data_flow_diagram

Some Links(2)

- Data Mining (Mining of Massive Datasets)
- <http://infolab.stanford.edu/~ullman/mmds.html>
- Predictive Analytics
- <http://www.predixionsoftware.com/predixion/download>
- <http://en.wikipedia.org/wiki/Weka> (machine learning)

Some Links(3)

- R
 - <http://cran.r-project.org/bin/windows/base/>
 - <http://cran.r-project.org/bin/macosx/>
 - <http://cran.r-project.org/bin/linux/>
 - <http://www.r-project.org/>
 - <http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
 - <http://www.rstudio.com/ide/download/desktop>
 - http://en.wikipedia.org/wiki/R_%28programming_language%29
-
- MATLAB / GNU Octave
 - <http://en.wikipedia.org/wiki/MATLAB>
 - http://en.wikipedia.org/wiki/GNU_Octave
 - <http://sourceforge.net/projects/octave/files/latest/download?source=files>
 - <http://sourceforge.net/projects/octave/files/Octave%20Windows%20binaries/Octave%203.6.2%20for%20Windows%20MinGW%20installer/>

Assignment (1)

1. Describe, in one sentence or one paragraph, a data science task that interests you.
2. Complete Quiz 01b before you make or submit your data flow diagram (DFD). Quiz 01b will become available on Wednesday April 1st 2015.
3. Construct a DFD that depicts the data processing of the data science task in item 1.
4. Download and Install R. Then download and install R studio. Calculate $2 + 3$ in R studio . Take a screenshot.
5. Download and Install GNU Octave. Calculate $2 + 3$ in GNU Octave. Take a screenshot.
6. Download the Hadoop VM. Take a screenshot of the file on your operating system:
https://www.dropbox.com/s/9f8enhk5z0xv7kw/Cloudera-Training-VM-4.2.1.p-vmware_pristine.zip?dl=0

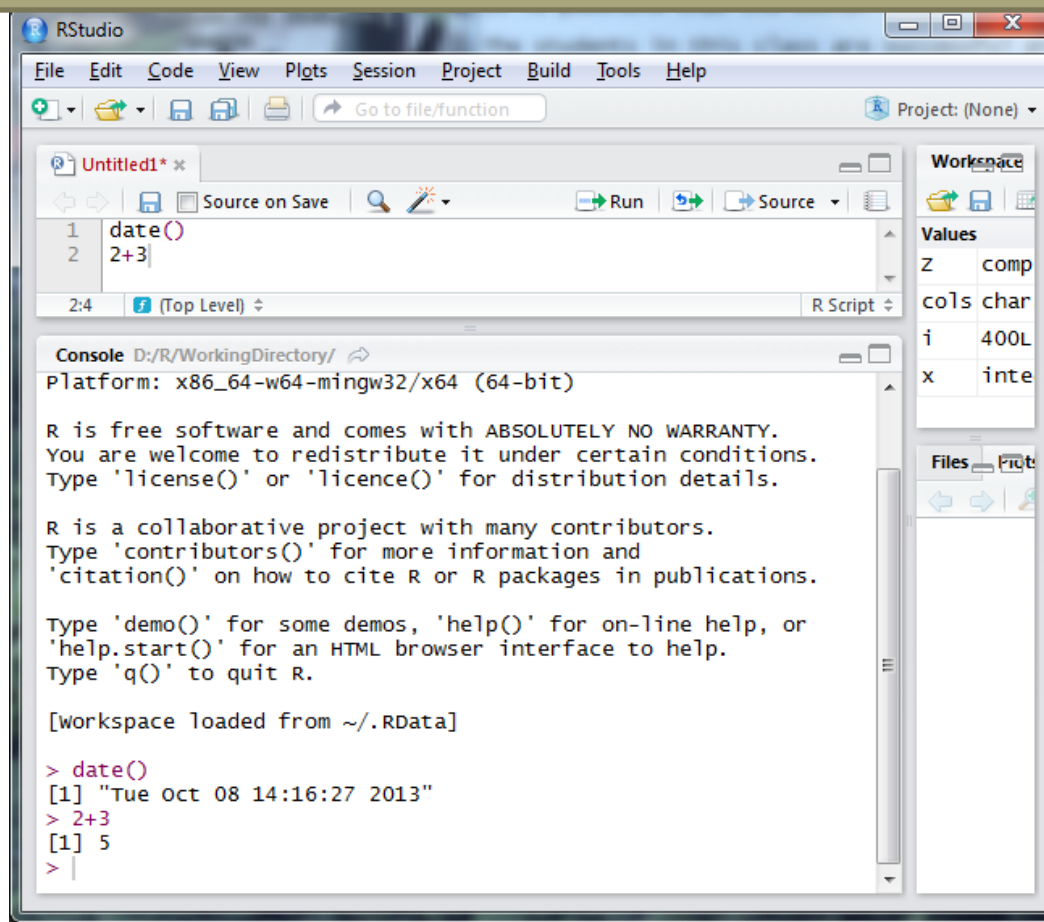
Assignment (2)

7. Find on the web a publically available dataset that is free of charge and that can be downloaded as a rectangular dataset. Describe the dataset in one sentence. Note the URL. I will compile a list of datasets for future reference.
8. Join the LinkedIn group. Introduce yourself, start a discussion, or make a comment on an existing discussion.
9. Combine the assignment items 1, 3, 4, 5, 6 and 7 in a single doc (If it isn't code, I prefer pdf). Note in this document whether you completed item 8. Submit this document 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.

Download and Installs

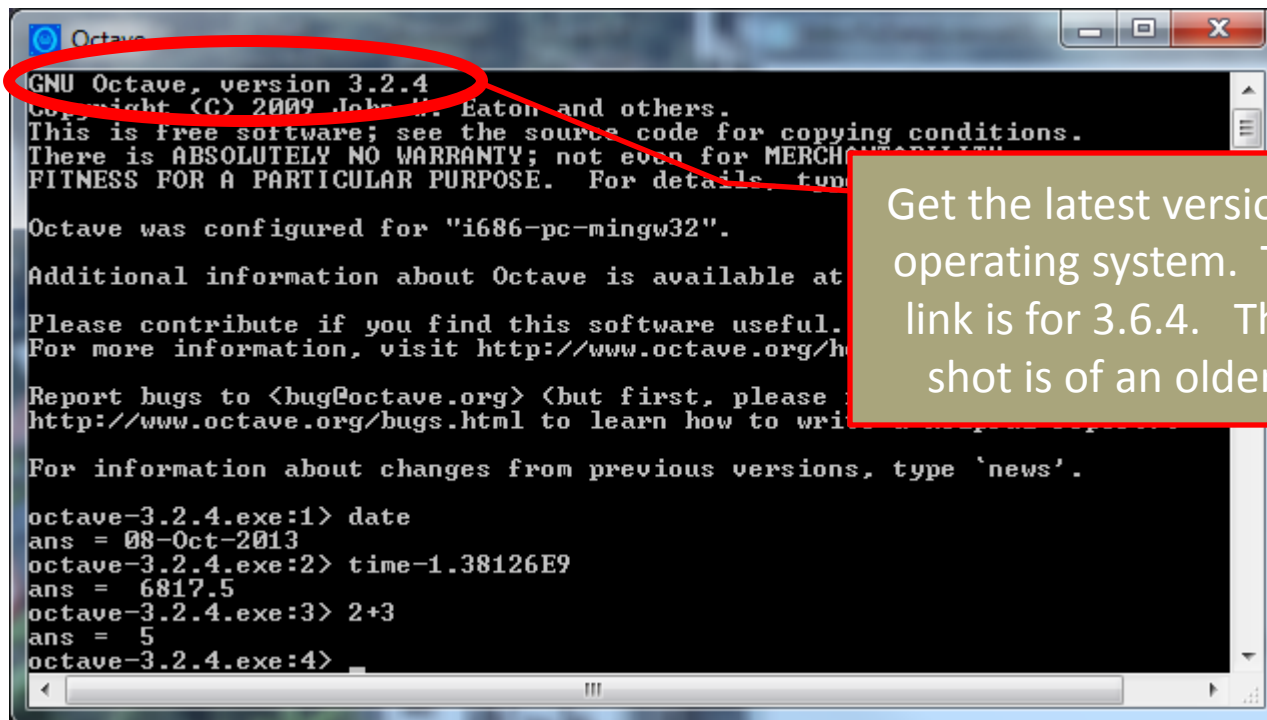
Download (1) R and R studio

- <http://cran.r-project.org/bin/windows/base/>
- <http://www.rstudio.com/ide/download/>
- Get the latest version of R and the latest version of R studio. This screen shot is of an older version



Download (2) GNU Octave for Windows

<http://sourceforge.net/projects/octave/files/Octave%20Windows%20binaries/Octave%203.6.4%20for%20Windows%20MinGW%20installer/>
http://sourceforge.net/projects/octave/files/Octave%20Windows%20binaries/Octave%203.6.4%20for%20Windows%20MinGW%20installer/Octave3.6.4_gcc4.6.2_20130408.7z/download



```
GNU Octave, version 3.2.4
Copyright (C) 2009 John M. Eaton and others.
This is free software; see the source code for copying conditions.
There is ABSOLUTELY NO WARRANTY; not even for MERCHANTABILITY
or FITNESS FOR A PARTICULAR PURPOSE. For details, type 'warranty'.

Octave was configured for "i686-pc-mingw32".

Additional information about Octave is available at
http://www.octave.org.

Please contribute if you find this software useful.
For more information, visit http://www.octave.org/h
Report bugs to <bug@octave.org> (but first, please
http://www.octave.org/bugs.html to learn how to write a good bug report.

For information about changes from previous versions, type 'news'.

octave-3.2.4.exe:1> date
ans = 08-Oct-2013
octave-3.2.4.exe:2> time-1.38126E9
ans = 6817.5
octave-3.2.4.exe:3> 2+3
ans = 5
octave-3.2.4.exe:4>
```

Get the latest version for your operating system. The above link is for 3.6.4. This screen shot is of an older version

Download (3) GNU Octave for Windows









Octave Forge - Browse /Oc... x +

sourceforge.net/projects/octave/files/Octave Windows binaries/Octave 3.6.4 for Windows MinGW installer/

Summary Files Reviews Support Mailing Lists Code package releases Discussion Me

Looking for the latest version? [Download optim-1.3.0.tar.gz \(193.7 kB\)](#)

Home / Octave Windows binaries / Octave 3.6.4 for Windows MinGW installer

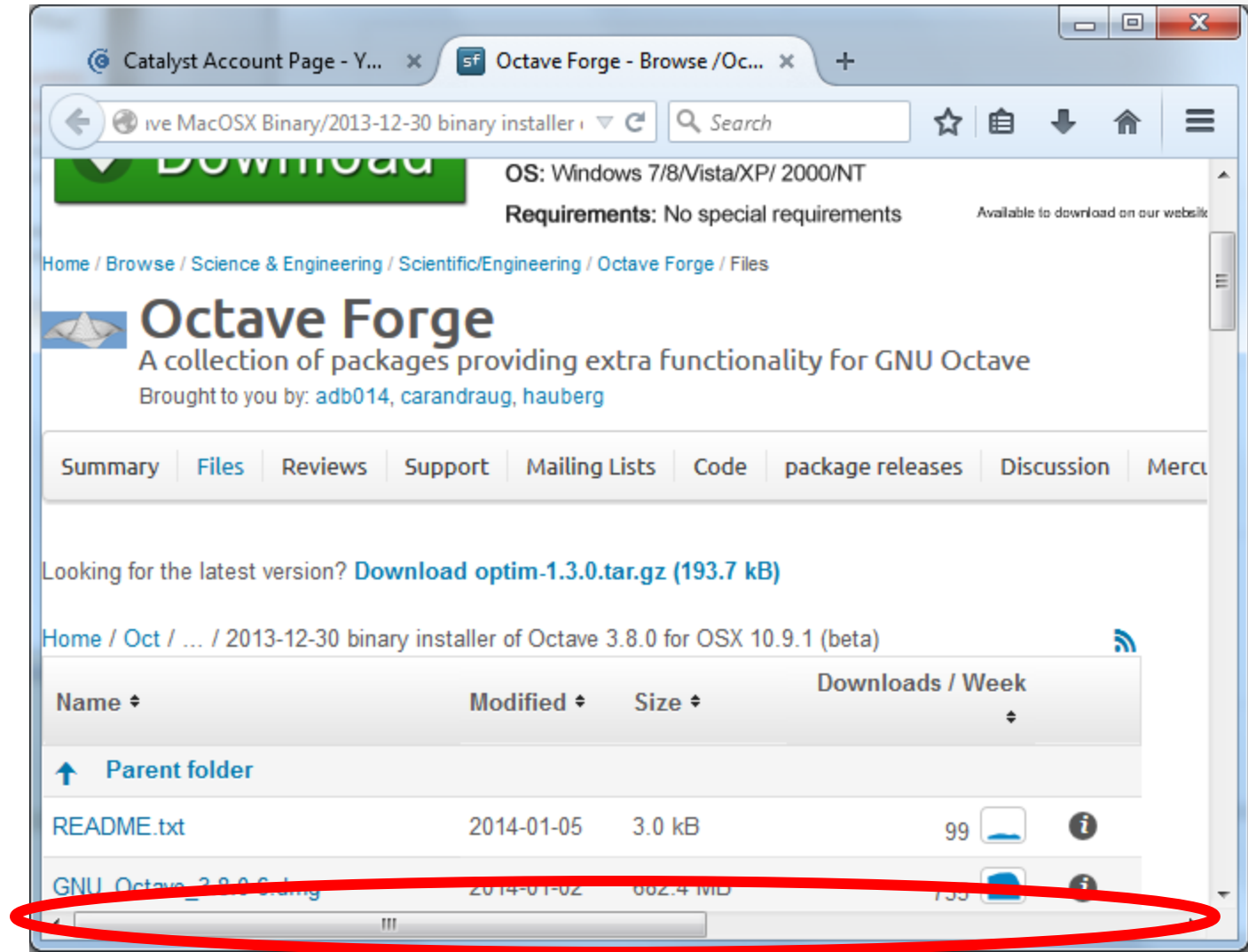
Name ↕	Modified ↕	Size ↕	Downloads / Week ↕
↑ Parent folder			
README	2013-04-10	8.0 kB	235  
Octave3.6.4_gcc4.6.2_sourceforge_r...	2013-04-10	8.2 kB	262  
Octave3.6.4_gcc4.6.2_pkgs_201304...	2013-04-10	53.1 MB	601  
Octave3.6.4_gcc4.6.2_20130408.7z	2013-04-10	156.9 MB	1,226  
Totals: 4 items		210.0 MB	2,324

Octave-3.6.4-mingw + octaveforge pkgs

Download (4) GNU Octave for Mac

- I googled: "octave for mac"
- My first link led me to: [http://wiki.octave.org/Octave for MacOS X](http://wiki.octave.org/Octave_for_MacOS_X)
- In that wiki page there was link titled "binary installer"
- <http://sourceforge.net/projects/octave/files/Octave%20MacOSX%20Binary/2013-12-30%20binary%20installer%20of%20Octave%203.8.0%20for%20OSX%2010.9.1%20%28beta%29/>
- Then I found myself on a sourceforge page that had the following link: GNU_Octave_3.8.0-6.dmg
- That link downloaded a 682.4 MB file called GNU_Octave_3.8.0-6.dmg
- I saved the file
- I double-clicked on GNU_Octave_3.8.0-6.dmg
- I control-clicked on Octave-3.8.0-6.mpkg" and selected Open from the context
- I clicked the Open button in the dialog titled: "Octave-3.8.0-6.mpkg" is from an unidentified developer. ...

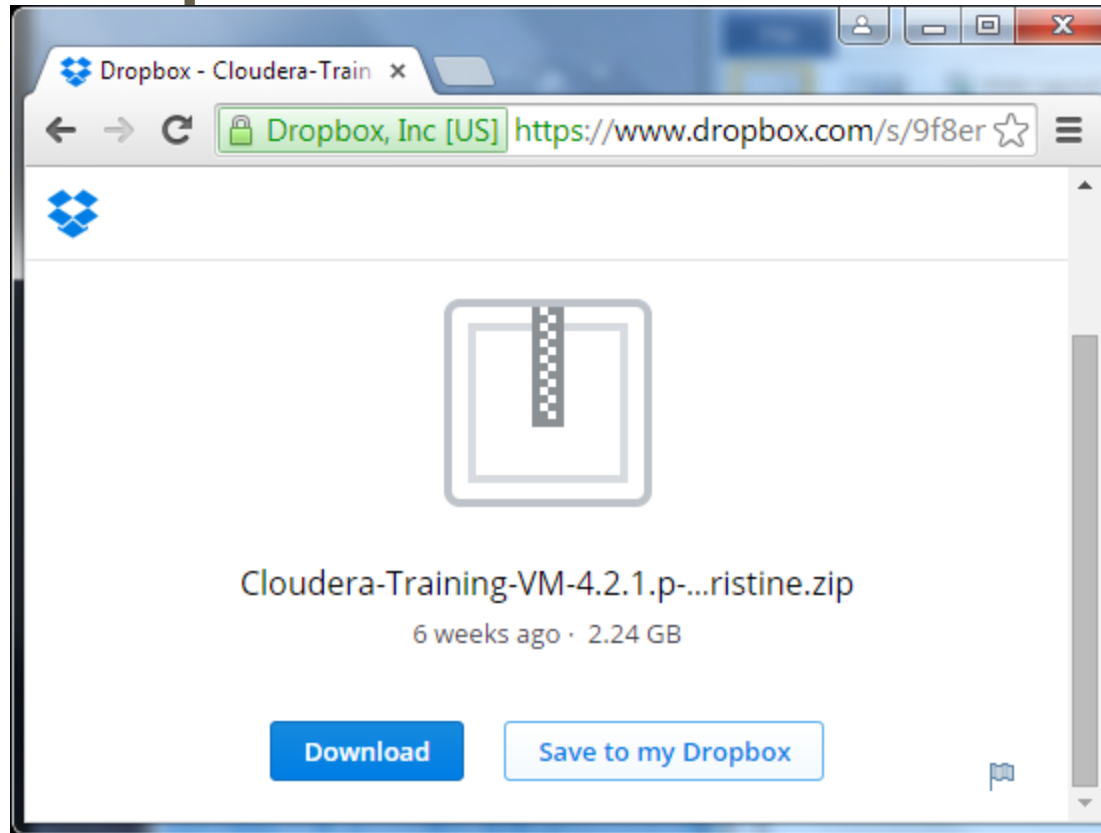
Download (5) GNU Octave for Mac



Download (6) Hadoop VM from Dropbox

- https://www.dropbox.com/s/9f8enhk5z0xv7kw/Cloudera-Training-VM-4.2.1.p-vmware_pristine.zip?dl=0 That url will lead you to the VM.
- Paste the url in a browser (I use chrome)
- You should see the file (Cloudera-Training-VM ...) associated with a download button (see picture below). If you are prompted to sign in or create an account, then just close that dialog. The file and download button will be underneath that dialog.

Download (7) Hadoop VM from Dropbox



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

Download and Installs

Introduction to Data Science