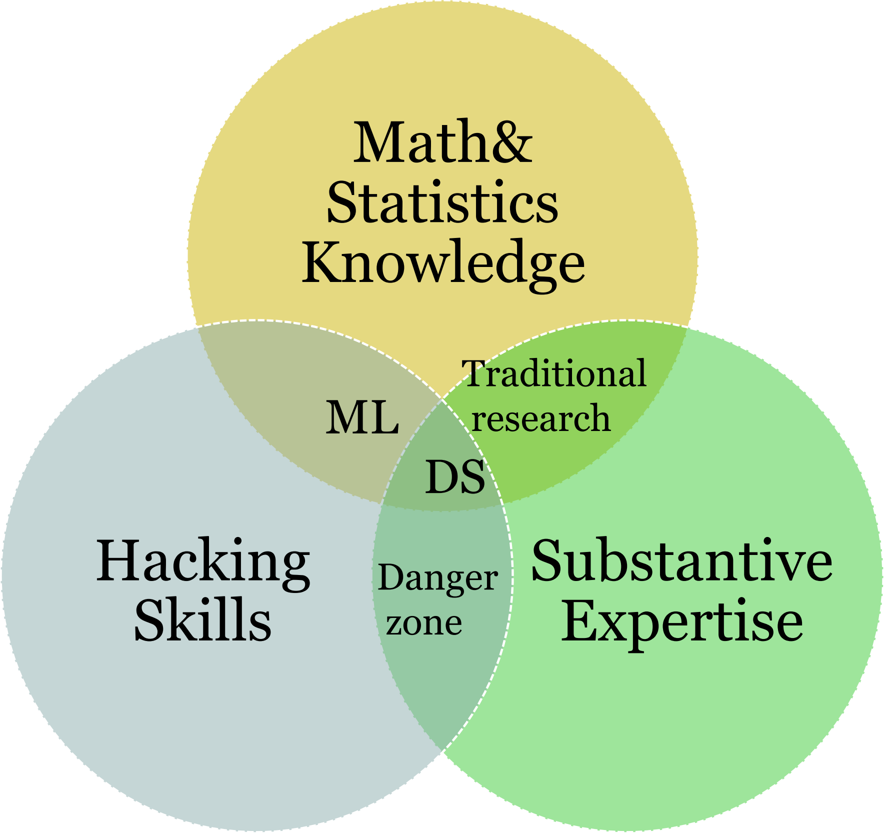
Data Science Roadmap

LECTURE 1

Learning Objectives

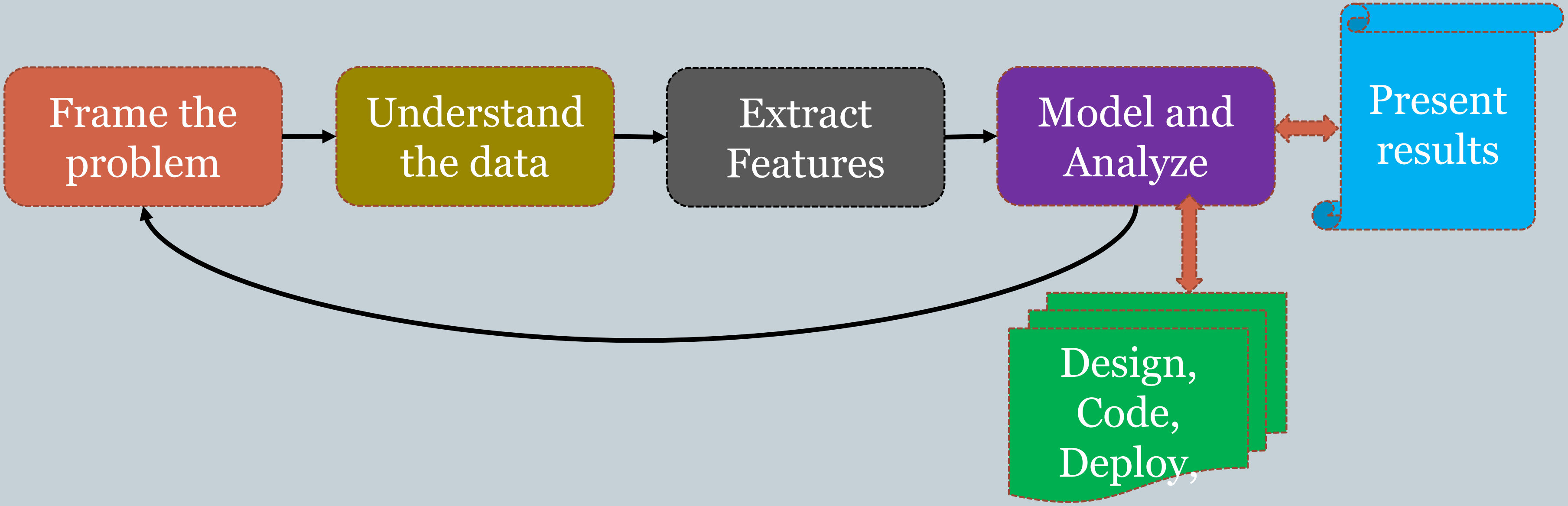
- Explain the data science process
- Illustrate the role of languages, tools and integrated development environments
  - Specifically: R, Jupyter, and Rstudio

Drew Conway's Venn Diagram on Data Science



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The DS Roadmap



Frame the problem

Understand the data

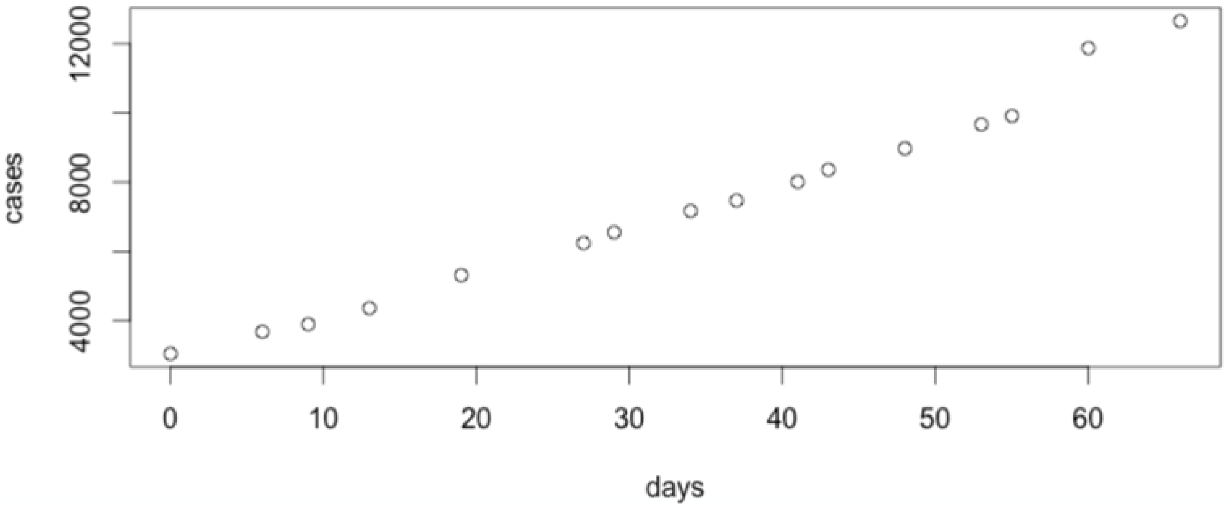
Extract Features

Model and Analyze

Present results

Design, Code, Deploy,

Bad DS: 2014 West African Ebola outbreak: This is real data collected during Ebola crisis. We have # mentions in twitter and number of cases plotted against days. Data from Twitter and CDC.gov



Chapter 1 and 2 Data Science

- Read Chapter 1 to get a perspective on “big data”
- Chapter 2: Statistical thinking in the age of big data
- You build models to understand the data and extract meaning and information from the data: statistical inference

Lets discuss the road map

1. Frame the problem: understand the use case
2. Understand the data: Exploratory data analysis
3. Extract features: what are the dependent and independent variables, cols and rows in a table data for example.
4. Model the data and analyze: big data, small data, historical, steaming, realtime etc.
5. Design, code and experiment: use tools to clean, extract, plot, view
6. Present and test results: two types of clients: humans and systems
7. Go back to any of the steps based on the insights!

Frame The problem

- Have a standard use case format (What, why, how, stakeholders, data in, info out, challenges, limitations, scope etc.)
- Refer to your software engineering course
- Statement of work (SOW): clearly state what you will accomplish

Understand Data

- Data represents the traces of the real-world processes.
  - What traces we collect depends on the sampling methods
  - You build models to understand the data and extract meaning and information from the data: statistical inference
- Two sources of randomness and uncertainty:
  - The process that generates data is random
  - The sampling process itself is random
- Your mind-set should be “statistical thinking in the age of big-data”
  - Combine statistical approach with big-data

Here are some questions to ask?

- How big is the data?
- Any outliers?
- Missing data?
- Sparse or dense?
- Collision of identifiers in different sets of data

New Kinds of Data

- Traditional: numerical, categorical, or binary
- Text: emails, tweets, NY times articles
- Records: user-level data, time-stamped event data, json formatted log files
- Geo-based location data
- Network data (How do you sample and preserve network structure?)
- Sensor data
- Images

Uncertainty and Randomness

- A mathematical model for uncertainty and randomness is offered by probability theory.
- A world/process is defined by one or more variables. The model of the world is defined by a function:
- **Model** == f(w) or f(x,y,z) (A multivariate function)
- The function is unknown→ model is unclear, at least initially. Typically our task is to come up with the model, given the data.
- **Uncertainty**: is due to lack of knowledge: this week's weather prediction (e.g. 90% confident)
- **Randomness**: is due lack of predictability: 1-6 face of when rolling a die
- Both can be expressed by probability theory

Statistical Inference

- World ➔ Collect Data➔ Capture the understanding/meaning of data through models or functions ➔ statistical estimators for predicting things about➔ world
- Development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes

Population and Sample

- Population is complete set of traces/data points
  - US population 314 Million, world population is 7 billion for example
  - All voters, all things
- Sample is a subset of the complete set (or population): how we select the sample introduces biases into the data
- See an example in http://www.sca.isr.umich.edu/
- Here out of the 314 Million US population, 250000 households are form the sample (monthly)
- Population →mathematical model → sample
- (My) big-data approach for the world population: k-nary tree (MR) of 1 billion (of the order of 7 billion) : I basically forced the big-data solution/did not sample: This is possible in the age of big-data infrastructures

Population and Sample (contd.)

- Example: Emails sent by people in the CSE dept. in a year.
- Method 1: 1/10 of all emails over the year randomly chosen
- Method 2: 1/10 of people randomly chosen; all their email over the year
- Both are reasonable sample selection method for analysis.
- However estimations pdfs (probability distribution functions) of the emails sent by a person for the two samples will be different.

Big Data vs statistical inference

- Sample size N
- For statistical inference N < All
- For big data N == All
- For some atypical big data analysis N == 1
  - World model through the eyes of a prolific twitter user
  - Followers of Ashton Kuchar: If you analyze the twitter data you may get a world view from his point of view

Big-data context

- Analysis for inference purposes you don't need all the data.
- At Google (at the originator big data algs.) people sample all the time.
- However if you want to render, you cannot sample.
- Some DNA-based search you cannot sample.
- Say we make some conclusions with samples from Twitter data we cannot extend it beyond the population that uses twitter. And this is what is happening now…be aware of biases.
- Another example is of the tweets pre- and post-hurricane Sandy..
- Yelp example..

Exploratory Data Analysis (EDA)

- You achieve two things to get you started:
  - Get an intuitive feel for the data
  - You can get a list of hypotheses
- Traditionally: histograms
- EDA is the prototype phase of ML and other sophisticated approaches;
- Basic tools of EDA are plots, graphs, and summary stats.
- It is a method for "systematically" going through data, plotting distributions, plotting time series, looking at pairwise relationships using scatter plots, generating summary stats.eg. mean, min, max, upper, lower quartiles, identifying outliers.
- Gain intuition and understand data.
- EDA is done to understand Big data before using expensive big data methodology.

Extract Features

- Data is cleaned up : Data wrangling
- Ex: remove tags from html data
- Filter out only the important fields or features, say from a json file
- Often defined by the problem analysis and use case defined.
- Example: location and temperature are the only important data in a tweet for a particular analysis

Modeling

- Abstraction of a real world process
- Lets say we have a data set with two columns x and y and y is dependent on x, we could write is as:

$y = \beta 1 + \beta 2 * x$

(linear relationship)

- How to build a model?
- Probability distribution functions (pdf) are building blocks of statistical models.
- There are many distributions possible

Probability Distributions

- Normal, uniform, Cauchy, t-, F-, Chi-square, exponential, Weibull, lognormal,..
- They are know as continuous density functions
- Any random variable x or y can be assumed to have probability distribution p(x), if it maps it to a positive real number.
- For a probability density function, if we integrate the function to find the area under the curve it is 1, allowing it to be interpreted as probability.
- Further, joint distributions, conditional distribution..

Fitting a Model

- Fitting a model means estimating the parameters of the model: what distribution, what are the values of min, max, mean, stddev, etc.
- Don’t worry R has built-in optimization algorithms that readily offer all these functionalities
- It involves algorithms such as maximum likelihood estimation (MLE) and optimization methods…
- Example: $y = \beta 1 + \beta 2 * x$ ➔ $y = 7.2 + 4.5^{*}x$

Design, code, deploy

- Design first before you code: an important principle
- Code using best practices and “Software engineering” principles
- Choose the right language and development environment
- Document within the code and outside
- Clear state the steps in deploying the code
- Provide trouble shooting tips

Present the Results

- Good annotated graphs and visuals are important explaining the results
- Annotate using text, markup and markdown
- Extras: provide ability to interact with plots and assess what-if conditions
- Explore

  d3.js : https://d3js.org/

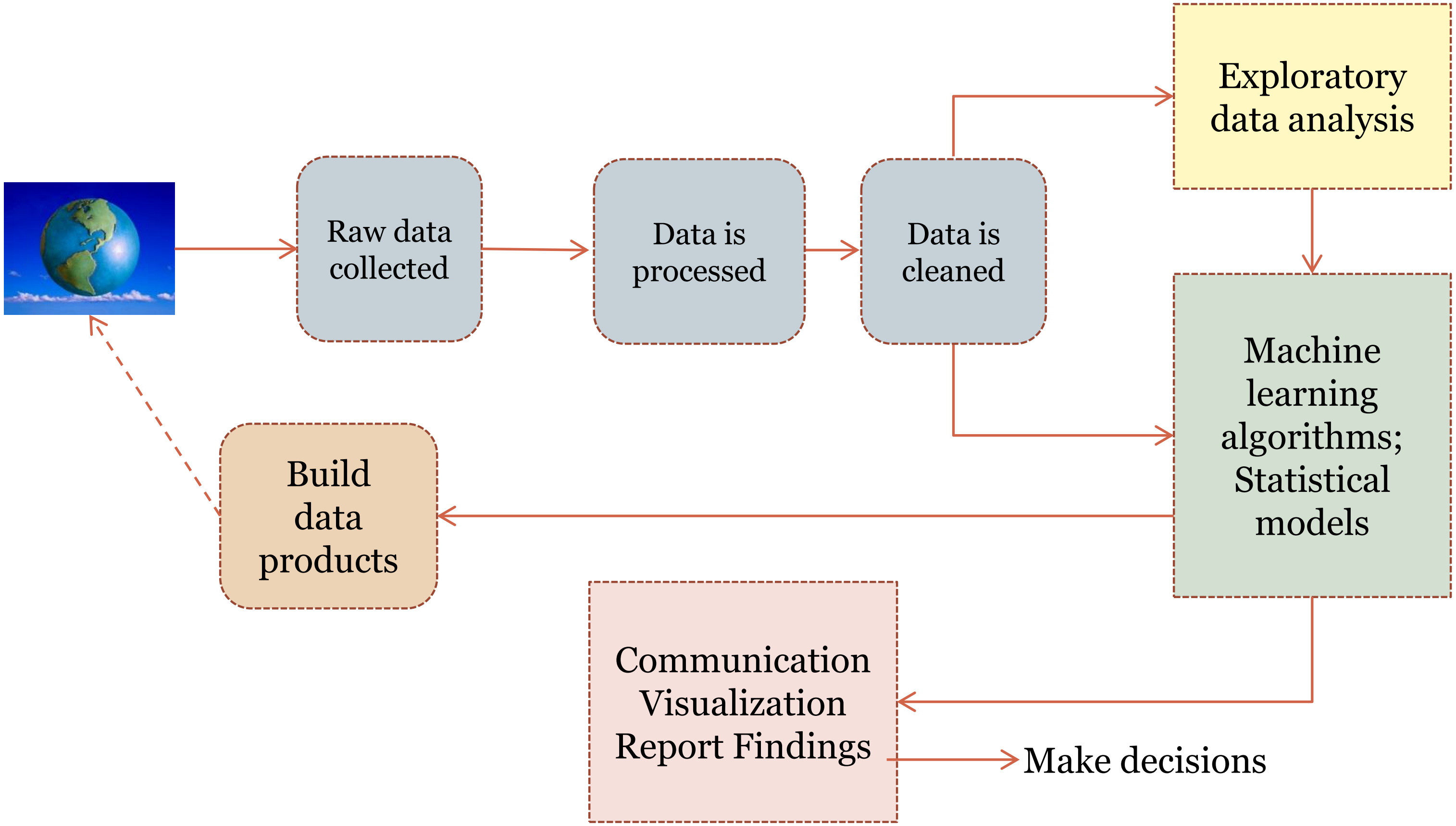
  Tableau: https://www.tableau.com/academic

  R graphs: https://www.statmethods.net/graphs/creating.html
- And a lot of creativity. Do not underestimate this: how to present your results effectively?
- Should need no explanation!

Iterate

- Iterate thru' any of steps as warranted by the feedback and the results
- Data science process is an iterative process
- Before you develop a tool or automation based on the results test the code thoroughly.
- Read Chapter 2

The Data Science Process



Raw data collected

Data is processed

Data is cleaned

Exploratory data analysis

Machine learning algorithms; Statistical models

Build data products

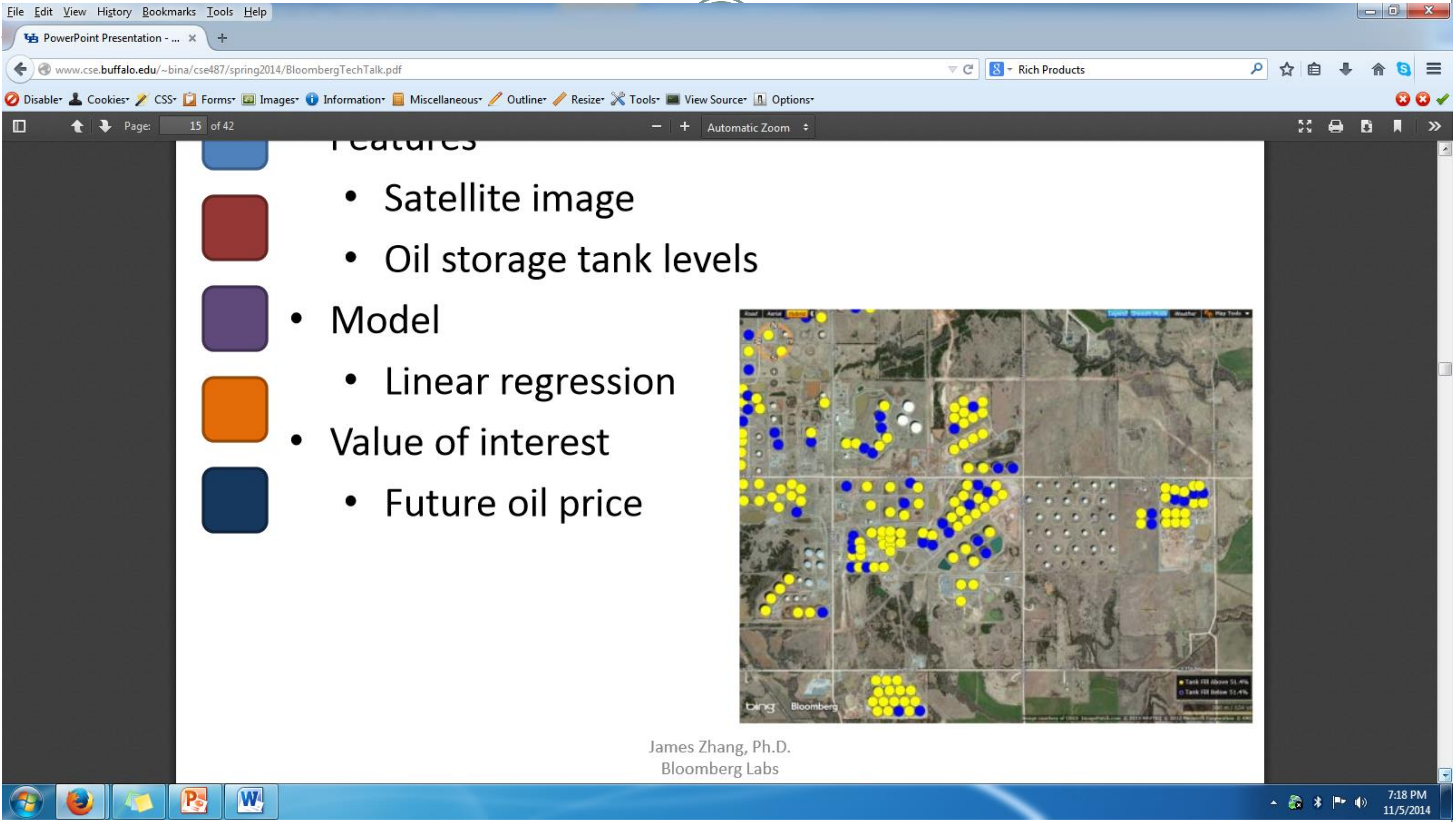
Communication Visualization Report Findings

Make decisions

Example1: Data Collection in Automobiles

- Large volumes of data is being collected the increasing number of sensors that are being added to modern automobiles.
- Traditionally this data is used for diagnostics purposes.
- How else can you use this data?
- How about predictive analytics? For example, predict the failure of a part based on the historical data and on-board data collected?
- On-board-diagnostics (OBDI) is a big thing in auto domain.
- How can we do this?

Example2: Oil Price Prediction


File Edit View History Bookmarks Tools Help
PowerPoint Presentation - ...
www.cse.buffalo.edu/~bina/cse487/spring2014/BloombergTechTalk.pdf
Rich Products
Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options
Page: 15 of 42
Automatic Zoom
bing
Bloomberg
Tank Fill Above 51.4%
Tank Fill Below 51.4%
7:18 PM
11/5/2014

Features

- Satellite image
- Oil storage tank levels

- Model
  - Linear regression
- Value of interest
  - Future oil price

James Zhang, Ph.D.
Bloomberg Labs

Our DS Environment for Lab1

- Jupyter : http://jupyter.org/index.html
  - The *Jupyter Notebook App* is a server-client application that allows editing and running notebook documents via a web browser. The *Jupyter Notebook App* can be executed on a local desktop requiring no internet access
- R Language: https://cran.r-project.org/
  - R is a free software environment for statistical computing and graphics.
- R Studio IDE: https://www.rstudio.com/
  - RStudio is an integrated development environment (IDE) for R.
- And for the platform try using AWS! You get free $100 credit per academic year as a student.

Summary

- An excellent tool supporting EDA is R
- Something to do this weekend:
  - Read chapter 2
  - Work on the sample R code in the chapter though it is not in your domain
  - Find some data sets from your work, import it to R, analyze
  - You collect or can collect a lot of data through existing channels you have.
- We will now introduce Jupyter notebook and R Studio environment that be used for understanding lab problems, designing and implementing solutions. See the handout prepared by J. Condello. We are working on updating this.
- Explore away!