cse6242-links-in-lecture-videos

Updated automatically every 5 minutes

- Course Introduction
    - Hi, I am Polo
        - Gartner’s definition of “data scientist” https://www.gartner.com/it-glossary/data-scientist
    - Why data and visual analytics?
    - Course goals and expectations
    - Course logistics

- Analytics Building Blocks
    - Overview
    - Example project 1: Apolo graph exploration
        - Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning https://www.cc.gatech.edu/~dchau/papers/11-chi-apolo.pdf
    - Example project 2: NetProbe auction fraud detection
        - NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks http://repository.cmu.edu/cgi/viewcontent.cgi?article=1530&context=compsci

- Data Science Buzzwords
    - Hype Cycle
        - Gartner Hype Cycle 2017 http://blogs.gartner.com/smarterwithgartner/files/2017/08/Emerging-Technology-Hype-Cycle-for-2017_Infographic_R6A.jpg
        - https://www.gartner.com/smarterwithgartner/
    - General AI vs Narrow AI
        - Self-Driving Taxis Hit the Streets of Singapore http://fortune.com/2016/08/25/self-driving-taxi-singapore/
        - Google AI beats Go world champion again to complete historic 4-1 series victory https://techcrunch.com/2016/03/15/google-ai-beats-go-world-champion-again-to-complete-historic-4-1-series-victory/
        - Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/
        - A Tragic Loss https://www.tesla.com/blog/tragic-loss
        - Preparing for The Future of Artificial Intelligence https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/N

- Data Collection
    - How to collect data?
        - Data you can download http://poloclub.gatech.edu/cse6242/2017fall/#datasets
        - Google Data API (e.g., Google Maps Directions API) https://developers.google.com/gdata/docs/directory
        - Twitter (small subset) https://dev.twitter.com/streaming/overview
        - Google Data API: GData API Directory

      https://developers.google.com/gdata/docs/directory
    - How to scrape?
        - Google Play Example

      https://play.google.com/store/apps/details?id=com.shazam.android&hl=en
        - Name any sound in seconds https://www.shazam.com/
        -

    - Selenium supports more actions http://www.discoversdk.com/blog/web-scraping-with-selenium

- SQLite
  - As simple, effective storage
    - SQLite: http://www.sqlite.org/famous.html
  - SQL refresher
    - SQL Quick Reference: https://www.w3schools.com/sql/sql_quickref.asp
  - Beware of missing indexes
    - B-Tree https://en.wikipedia.org/wiki/B-tree

- Data Cleaning
  - How dirty is real data?
  - Importance of data cleaning
    - Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#2c5226c6f637
    - For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
    - **Big Data Dirty Problem**

      http://fortune.com/2014/06/30/big-data-dirty-problem/
    - Indent Code (spacing vs tabs) https://google.github.io/styleguide/javaguide.html#s4.2-block-indentation
    - There is no way I’m going to be with someone who uses spaces over tabs http://www.businessinsider.com/tabs-vs-spaces-from-silicon-valley-2016-5
    - Trailing whitespace is evil. Don't commit evil into your repo. http://codeimpossible.com/2012/04/02/trailing-whitespace-is-evil-don-t-commit-evil-into-your-repo/
  - Data cleaners: OpenRefine & Wrangler
    - Open Refine http://openrefine.org/
    - Data Wrangler http://vis.stanford.edu/wrangler/

- Class Project Overview
  - Forming great teams
    - https://www.cs.cmu.edu/~pausch/Randy/tipoForGroups.html
  - Core project requirements
  - Project idea checklist: Heilmeier questions
    - https://en.wikipedia.org/wiki/George_H._Heilmeier
    - http://poloclub.gatech.edu/cse6242/2017spring/slides/CSE6242-999-project.pdf
  - Pay attention to software licenses early on
    - GPL(General Public License) https://en.wikipedia.org/wiki/GNU_General_Public_License

- Code Back-up & Version Control
  - Git: Overview and Benefits
    - Git is the **most popular** version control system in software development https://en.wikipedia.org/wiki/Git

        - **Dev put AWS keys on Github. Then BAD THINGS happened** **http://www.theregister.co.uk/2015/01/06/dev_blunder_shows_github_crawling_with_k**
        - **OneDrive https://ai.oit.gatech.edu/onedrive**

- Data Integration
    - Knowledge graph
        - Apple Siri https://www.apple.com/ios/siri/
        - OpenRefine (Reconcile and Match Data) https://www.youtube.com/watch?v=5tsyz3ibYzk
        - Freebase (originally by MetaWeb; acquired by Google)
            - https://en.wikipedia.org/wiki/Freebase_(database)
            - http://youtu.be/TJfrNo3Z-DU
        - The Knowledge Graph (video); Google's Knowledge Graph website is no longer available https://youtu.be/mmQl6VGvX-c
        - What does Google know about Taylor Swift? https://developers.google.com/knowledge-graph/
        - Introducing Facebook Graph Search https://www.youtube.com/watch?v=W3k1USQbq80&feature=youtu.be
            - Looks like Meta/Facebook has taken down the video, but it seems way back machine (https://archive.org/web/) took snapshots of the video!
        - [Supplemental] Mark Zuckerberg explains Facebook's new Graph Search https://youtu.be/U94DTrjAvuA
    - Data de-duplication
        - **D-Dupe: An Interactive Tool for Entity Resolution in Social Networks** **https://linqspub.soe.ucsc.edu/basilic/web/Publications/2006/bilgic:vast06/**
    - Importance of Similarity Functions
        - Distance and Similarity Measures https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures.html
        - Entity Resolution for Big Data http://legacydirs.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf
    - Example project: Firebird fire risk prediction for Atlanta https://www.cc.gatech.edu/~dchau/papers/16-kdd-firebird.pdf

- Data Analytics, Concepts and Tasks [cse6242_wk3_tasks.pptx]
    - Break complex problems into simpler ones: Part 1
        - Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking https://www.amazon.com/Data-Science-Business-data-analytic-thinking/dp/1449361323
    - Break complex problems into simpler ones: Part 2
        - How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did https://www.forbes.com/sites/kashmirhill/2012/02/16/how-

- Visualization 101
    - What is info vis and why it is important
        - http://www.infovis-wiki.net/index.php/Information_Visualization
        - Why it is importance https://www.edwardtufte.com/tufte/
            - Communication: Space Shuttle *Challenger*disaster https://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster
            - Richard Feynman: Challenger Crash O-Ring https://www.youtube.com/watch?v=6Rwcbsn19c0&feature=youtu.be
        - **The best stats you've ever seen**

          https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen/u

          next
        - Anscombe's quartet https://en.wikipedia.org/wiki/Anscombe%27s_quartet
    - Human Perception
        - Information Theory https://www.britannica.com/science/information-theory
    - Gestalt Psychology
        - https://en.wikipedia.org/wiki/Gestalt_psychology
        - **Gestalt Psychology: Definition & Principles**

          https://study.com/academy/lesson/gestalt-psychology-definition-principles-quiz.html
    - Chart Basics
        - Edward Tufte https://en.wikipedia.org/wiki/Edward_Tufte
        - Visual Business Intelligence http://www.perceptualedge.com/blog/?p=790
        - Chartjunk https://en.wikipedia.org/wiki/Chartjunk
    - Colors
        - RGB Color model https://en.wikipedia.org/wiki/RGB_color_model
        - Color Survey Results https://blog.xkcd.com/2010/05/03/color-survey-results/
        - Color Blindness https://en.wikipedia.org/wiki/Color_blindness
        - Color User Guidline for Mapping and Visualization http://www.personal.psu.edu/faculty/c/a/cab38/ColorSch/Schemes.html
        - Color Brewer for Picking Color Scales http://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3
    - Zoom + Filter
        - The eyes have it https://www.mat.ucsb.edu/g.legrady/academic/courses/11w259/schneiderman.pdf
        - Baby names popularity http://www.babynamewizard.com/voyager#prefix=&sw=both&exact=false
        - Visually https://visual.ly/community/infographic/entertainment/every-single-death-game-thrones-series
- Fixing Common Visualization Issues
    - Fixing bar charts, line charts, tables, and more
        - Blazing-fast data transfer http://www.apple.com/imac/performance/

        - http://flowingdata.com/2012/06/18/what-3-d-pie-charts-are-good-for
        - All 193% of Republicans Support Palin, Romney and Huckabee
          http://wonkette.com/412361/all-193-of-republicans-support-palin-romney-and-huckabee
        - Funniest pie chart
          http://infosthetics.com/archives/2008/09/funniest_pie_chart_ever.html
    - Applying what you've learned
        - How to fix the defaults
          https://www.darkhorseanalytics.com/blog/clear-off-the-table
    - Crown jewel, self-contained figures, more tips
        - Scene Completion Using Millions of Photographs
          http://graphics.cs.cmu.edu/projects/scene-completion/
        - Polonium: Tera-Scale Graph Mining and Inference for Malware Detection
          http://www.cs.cmu.edu/~dchau/polonium_sdm2011.pdf
        - Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning
          https://www.cc.gatech.edu/~dchau/papers/11-chi-apolo.pdf
        - Don McMillan: Life After Death by PowerPoint
          https://www.youtube.com/watch?v=lpvgfmEU2Ck&feature=player_embedded

- Data Visualization for Web (D3)
    - Why learn D3?
        - Ver4 vs ver3
          https://iros.github.io/d3-v4-whats-new/#1
        - Upgrading Ver3 code to ver4 code
          https://keithpblog.wordpress.com/2016/07/31/upgrading-d3-from-v3-to-v4/
        - Wat
          https://www.destroyallsoftware.com/talks/wat
    - Prerequisites: Javascript and SVG
        - Array map
          https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Array/map
        - Mozilla Developer Network
          https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference
        - SVG Basics
          https://en.wikipedia.org/wiki/Scalable_Vector_Graphics
        - W3C Standard
          http://www.w3.org/TR/SVG/
        - CSS Selectors
          http://www.w3schools.com/cssref/css_selectors.asp
    - D3 Overview
        - Importing a CSV into D3
          http://stackoverflow.com/questions/24473733/importing-a-csv-into-d3-cant-convert-strings-to-numbers
    - Enter-Update-Exit
        - Excellent interactive demo to explain enter-update-exit:
          http://niceone.org/examples/d3-selections/
        - Full tutorial:
          https://medium.com/@c_behrens/enter-update-exit-6cafc6014c36#.dqwkermdb
    - Attributes, Styles, Classes and Text
        - <text> elements
          http://www.w3c.org/TR/SVG/text.html
    - Scales and Axes
        - D3 Arrays
          https://github.com/d3/d3-3.x-api-reference/blob/master/Arrays.md
    - Dynamic Data and Interaction
        - Treemap

        - List of seemingly ALL the tutorials online
          https://github.com/mbostock/d3/wiki/Tutorials

- Scalable Computing: Hadoop
    - Big data is common. How to store them?
    - Why Hadoop?
        - Hadoop: The Definitive Guide
          http://shop.oreilly.com/product/0636920033448.do
    - MapReduce: overview and example
    - Example MapReduce program
    - HDFS & Recovering From Failure
        - 2003 Google File System (GFS) paper
          https://research.google.com/archive/gfs.html
        - 2004 Google MapReduce paper
          https://research.google.com/archive/mapreduce.html
    - When and how to try Hadoop?

- Scalable Computing: Pig
    - Why Pig? How to use it?
    - Example Pig program
- Scalable Computing: Hive
    - Overview, and vs Pig

- Scalable Computing: Spark
    - Overview
        - Spark
          http://spark.apache.org
        - Google dumps MapReduce
          http://www.datacenterknowledge.com/archives/2014/06/25/google-dumps-mapreduce-favor-new-hyper-scale-analytics-system/
        - The death of MapReduce at Google
          http://www.reddit.com/r/compsci/comments/296aqr/on_the_death_of_mapreduce_at_google
    - Example Spark programs
    - Spark SQL and other Spark libraries
        - MLLib
          https://spark.apache.org/docs/2.2.0/mllib-guide.html
        - Spark 2.0
          https://databricks.com/blog/2016/07/26/introducing-apache-spark-2-0.html

- Scalable Computing: HBase
    - Overview
        - HBase : The Definitive Guide
          http://shop.oreilly.com/product/0636920014348.do
    - How HBase Scales Up Storage
        - Excellent Summary
          http://blog.cloudera.com/blog/2013/04/how-scaling-really-works-in-apache-hbase/
    - How to use HBase
        - Why need to disable a table before dropping it?
          https://stackoverflow.com/questions/35441342/hbase-why-do-i-need-to-disable-a-table-before-dropping-it
    - To learn more about HBase
        - 2006 Google BigTable paper
          https://research.google.com/archive/bigtable.html
        - Bad key design
          http://hbase.apache.org/book/rowkey.design.html

- Classification
    - Overview
    - Overfitting and Cross Validation
    - K-NN
        - Elements of Statistical Learning (ESL) Book Chapter 13.3.
          https://web.stanford.edu/~hastie/ElemStatLearn/

        - Decision Tree Lectures from CMU http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15381-s06/www/DTs.pdf
        - Details on Loss Functions. http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf
        - Excellent Refresher on Information Gain http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15381-s06/www/DTs.pdf
        - Grid Search in Sklearn http://scikit-learn.org/stable/modules/grid_search.html

- Visualizing Classification
    - ROC, AUC, Confusion Matrix
        - Confusion Matrix https://en.wikipedia.org/wiki/Confusion_matrix
        - Ensemble Matrix http://research.microsoft.com/en-us/um/redmond/groups/cue/publications/CHI2009-EnsembleMatrix.pdf

- Introduction to clustering
    - K-means, hierarchical clustering, DBSCAN
        - Kmeans Demo http://tech.nitoyon.com/en/blog/2013/11/07/k-means/
        - Decide k with real data https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf
        - Time Complexity K Means http://www.cs.cmu.edu/~./dpelleg/download/kmeans.ps
        - Interactive DBScan Demo https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/
        - Dbscan http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html

    - Visualizing Clusters
        - Visualizing Topics as Matrix http://vis.stanford.edu/papers/termite
        - Visualizing Graph Communities (Convex Hulls) http://www.cc.gatech.edu/~dchau/papers/11-chi-apolo.pdf
        - Visualizing Graph Communities (Matrix) https://bost.ocks.org/mike/miserables/
        - Visualizing Graph Communities (Cross Associations) http://www.cs.cmu.edu/~christos/PUBLICATIONS/kdd04-cross-assoc.pdf
        - Graph Partitioning Tools http://glaros.dtc.umn.edu/gkhome/views/metis

- Graph analytics **(No Links in this deck)**
    - How to represent and store graphs
    - Graph power laws

- Centralities: Degree, Betweenness, Clustering Coefficient
    - PageRank and Personalized PageRank
        - Solving for steady-state probabilities https://math.stackexchange.com/questions/749145/steady-state-of-a-4-times-4-transition-matrix
            - The following two links shown in the video are no longer available; please refer to the resource above instead: ~~*Regular Markov Chains, Steady State Probability*~~ ~~https://fenix.tecnico.ulisboa.pt/downloadFile/3779579688473/6.3.pdf ;~~ ~~*Markov Chains*~~ ~~http://www.sosmath.com/matrix/markov/markov.html~~

        - http://amzn.com/0321767535
    - Interactive Graph Exploration

- Ensemble Method
    - Bagging and Random forests
    - Random Forests
        - Section 15.3.1 of http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf
        - https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr
        - http://stackoverflow.com/questions/18541923/what-is-out-of-bag-error-in-random-forests
    - PERT - Perfect Random Tree Ensembles
        - http://www.interfacesymposia.org/I01/I2001Proceedings/ACutler/ACutler.pdf
    - Extremely randomized trees
        - http://orbi.ulg.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf
    - Random forests: ESL Chapter 15
        - http://www-stat.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

- Scaling up Algorithms with Virtual Memory
    - Overview
    - Power Iteration
        - http://en.wikipedia.org/wiki/Power_iteration
    - MMap Paper
        - https://www.cc.gatech.edu/~dchau/papers/14-bigdata-mmap.pdf
    - MMap Website
        - http://poloclub.gatech.edu/mmap

- Text Analytics
    - Basics: Preprocessing, Representation, Word Importance
    - Latent Semantic Indexing (Singular Value Decomposition)
    - SVD: Dimensionality Reduction, and Other Uses
    - Text Visualization
    - Stemming
        - https://en.wikipedia.org/wiki/Stemming
    - Stopwords
        - https://en.wikipedia.org/wiki/Stop_words
    - TD-IDF Example
        - https://en.wikipedia.org/wiki/Tf–idf#Example_of_tf–idf
    - SVD vs PCA (intuitive relation)
        - https://math.stackexchange.com/questions/3869/what-is-the-intuitive-relationship-between-svd-and-pca
    - PCA Visualization
        - http://setosa.io/ev/principal-component-analysis/
    - Word bubbles
        - https://www.infocaptor.com/bubble-my-page
    - Word Tree
        - https://www.jasondavies.com/wordtree/
    - PhraseNet
        - http://hint.fm/projects/phrasenet/
    - Termite
        - http://vis.stanford.edu/papers/termite

cse6242-links-in-lecture-videos