Introduction to HiDAV Summer Boot Camp

Project ENABLE

May 20, 2019

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Wecome to HiDAV (Health Informatics Data Analytics and Visualization Summer School)

- Extensible Network-Accessible Biomedical & Health Informatics Lifelong Learning Environment (ENABLE)
- Established to encourage students and professionals from diverse backgrounds to explore careers or advanced training in biomedical and health informatics
- The program has two components: ENABLE HiDAV Summer Program and Online Master's program.

Introduction to ENABLE Project

Who are Your Instructors?

- Heejun Kim (https://heejunkim.web.unc.edu/)
- Ph.D. in Information and Library Sciences at the UNC-CH.
- MS in Geography and BS in Electrical Engineering.
- Dissertation: Credibility assessment of health information on social media: Discovering credibility factors, operationalization, and prediction.
- Research interests: Text mining, machine learning, information retrieval, health informatics, human-computer interaction, and geographic information science.

Who are Your Instructors?

- Michael S. Ortiz: LinkedIn Page.
- BA in Biology from UNC-CH in 2015.
- 3rd Year Ph.D. Student at Carolina Health Informatics Program.
- National Library of Medicine Pre-Doctoral Fellow 2019-2021.
- Research interests: Cyberinfrastructure, text mining, machine learning, information retrieval, health informatics, genomics.

Who are Your Instructors?

- Mika Wang
- Second-year PhD student at Carolina Health Informatics Program.
- MA in Linguistics.
- Research Interest: Computational Linguistics, Natural Language Processing, Health Informatics, Machine Learning, Information Retrieval.

Who are Your Instructors?

- Eric Cui
- Incoming master's student for biomedical and health informatics at UNC-CH.
- Between my 2nd and 3rd year of medical school at UNC-CH.
- BS Biology from Indiana University.
- Current research topics: NLP, quality improvement, EMS, epidemiology, big data.

Introduction

- Introduce about yourself

  - *Name, school, and degree program*
  - *How did you find ENABLE summer boot camp interesting?*
  - *Any special interest in data analytics?*
  - *One interesting fact… about yourself*

Introduction to Biomedical and Health Informatics and Data Analytics

Definition of Biomedical Informatics

- Biomedical informatics (BMI) is “the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health.” (AMIA)

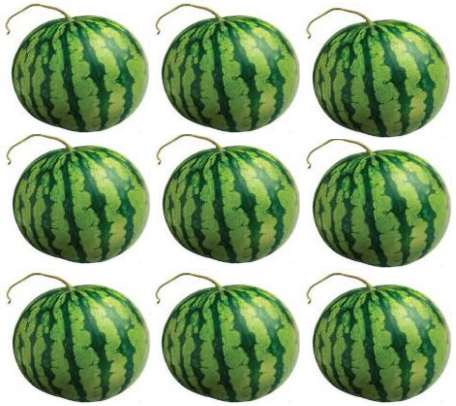
Definition of Health Informatics

- Health informatics is “the interdisciplinary study of the design, development, adoption, and application of IT-based innovations in healthcare services delivery, management, and planning.” (NLM)

What is “Data Analytics”? (An Analogy)

Data Analytics Process (an Analogy)

Watermelons



Cues

Stripe Thickness
Sound
Grocery level
....

Model

Linear
Decision Tree
Deep Learning
....

Selection & Tasting

Reinforcement

Why Data Analytics?

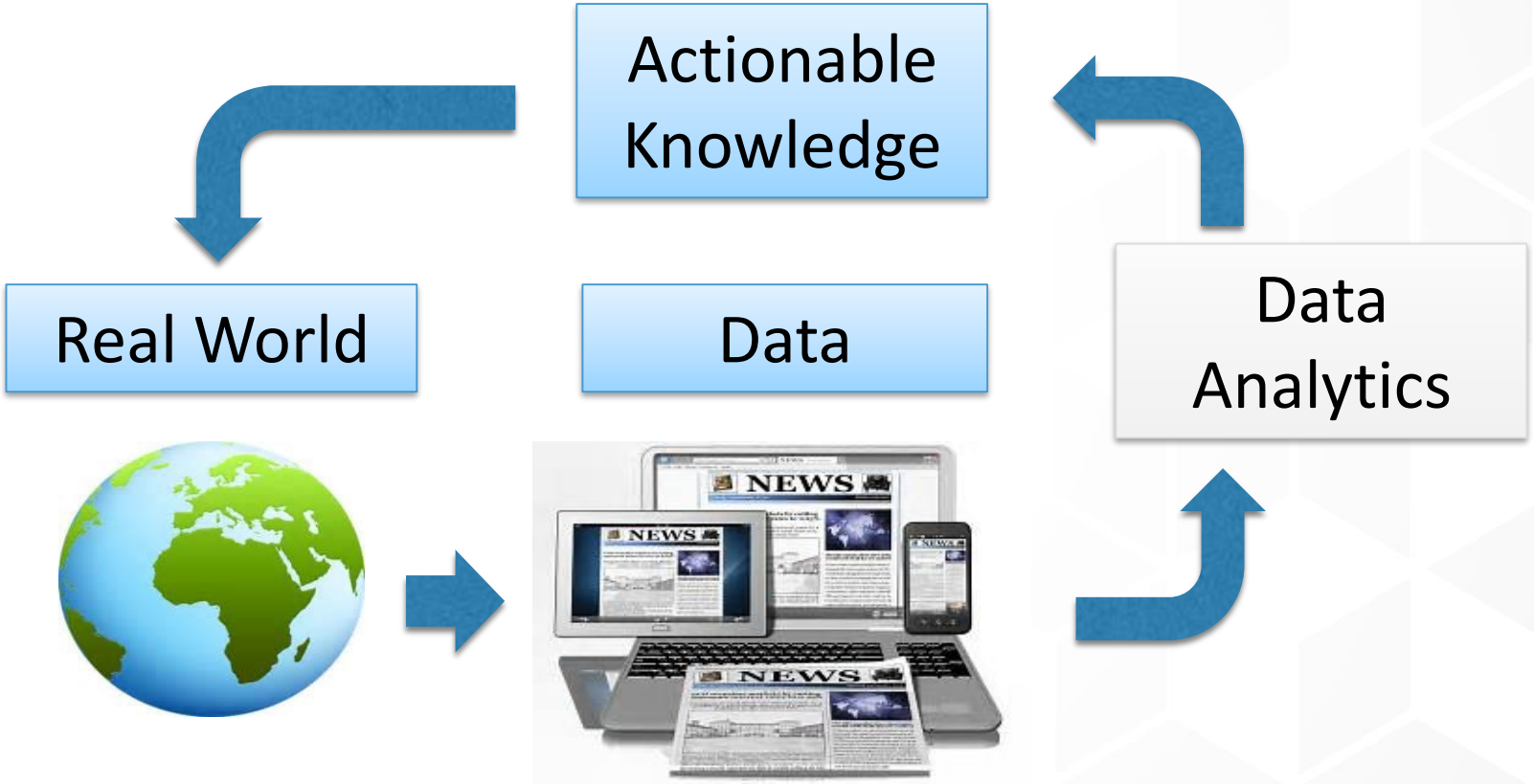






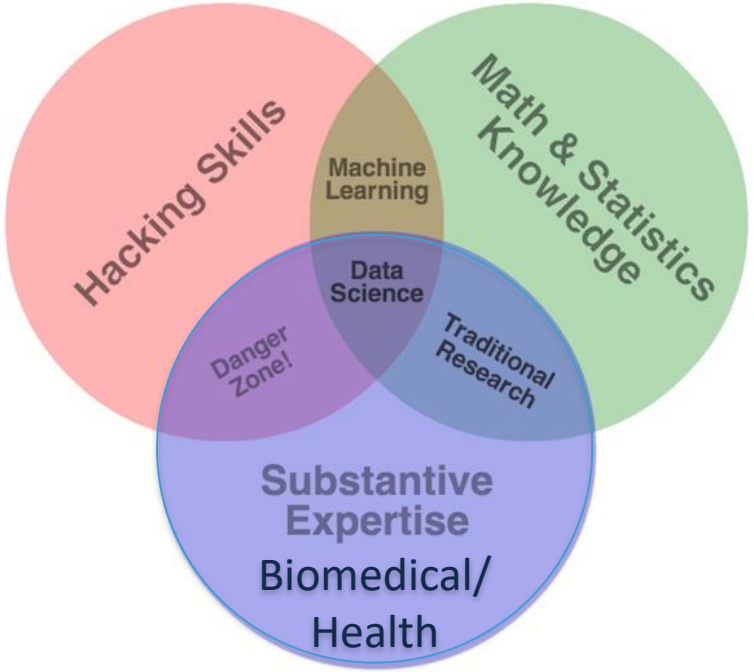



What is Data Analytics?



Related Field – Data Science

Data processing



Model

Knowledge

Related Field – Data Science

Link

Related Fields

- **Statistics**: developing methods for the interpretation of data and experimental outcomes in reaching conclusions with a certain degree of confidence.
- **Machine Learning**: developing computer programs that improve their performance with "experience."
- **Data Mining**: developing methods that discover patterns within large structured datasets.
- **Natural Language Processing**: developing computer programs to understand human speech as it is spoken.

Introduction to Two Modules of HiDAV Summer Boot Camp

- Text Mining
  - Instructor: Michael Ortiz
  - Goal: discover useful knowledge or insight in collections of natural language text by using statistical models and computer tools
- Data Mining
  - Instructor: Heejun Kim
  - Goal: discover useful knowledge or insight in collections of structured and/or unstructured data by using statistical models and computer tools

Class Schedule

- Day-by-day schedule is going to be available at: https://enable.unc.edu/boot-camp-day-by-day-schedule/
- Classroom: HSL 329
- One day is for text mining and the other day is for data mining.
- Field trip is on every Wednesday.
- One hour lecture and then one hour hand-on activity.
- Friday morning, we are expected to update the progress of your project and present paper(s).

NIH U.S. National Library of Medicine

A Platform for Biomedical Discovery and Data-Powered Health

Strategic Plan 2017–2027

G1: Accelerate discovery and advance health by providing the tools for data-driven research

- Connect the resources of a digital research enterprise
- Advance research and development in biomedical informatics and data science
- Foster open science policies and practices
- Create a sustainable institutional, physical, and computational infrastructure

G3: Build a workforce for data-driven research and health

- Expand and enhance research training for biomedical informatics and data science
- Assure data science and open science proficiency
- Increase workforce diversity
- Engage the next generation and promote data literacy

The View on the Data Analytics of NLM Director

Link

Expectation of the HiDAV Summer Boot Camp

Project Expectation

- Form a group with 2 or 3 people.
- Select one topic from two modules.
- Select a particular task that interest your group members. (previous projects)
- Find and select a dataset.
- Develop one or more hypotheses.
- Use an existing toolkit (https://healthit.unc.edu:8000/) or design a new program
- Do error analysis and report your findings.

Data Mining Projects

- Objective: create predictive models to identify patients most likely to require high healthcare costs regarding diabetes and hypertension.
- Data: Medical Expenditure Panel Survey (MEPS) which is a large-scale surveys of families and individuals, medical providers, and employers across the US.
- Tools: R scripts integrated with Jupyter Notebook will be available at: https://healthit.unc.edu:8000

Text Mining Projects

- Objective: use predictive and/or exploratory models to identify relationships within or across sources of text. For example:
  - Exploratory (generative) modeling:
    - What biomedical publications in a particular sub-field are highly related? How and why?
    - What network of genes involved in various diseases are highly related? How and why?
    - What appear to be to the core genes involved in a particular disease? How and why?
    - How many different molecular functions can a set of genes have for a particular disease? What are they? Did we expect this outcome?
  - Predictive modeling:
    - Based on drug mechanism, patient side effects, and patient review of medication, can we accurately predict what the benefit would be for another patient? Define a "good precision", what does this mean? Can we improve it (error analysis)?
    - Based on gene and molecular function, can we predict what the annotation for a gene should be? Should a genetic mutation be classified as benign, malignant, etc.?
- Data: National Center for Biotechnology Information (NCBI):
  - PubMed
  - Clinical Variations of Disease (ClinVar)
  - Gene DB
  - Swiss Uniprot
- Tools: Python scripts integrated with Jupyter Notebook will be available at: https://healthit.unc.edu:8000:
  - Python is a programming language used throughout the analytics community, VERY popular, simple, and powerful.
  - Jupyter Noteook is somewhat of an electronic laboratory notebook, except the experiments are for coding and data analysis instead of "wet-lab" (bio, chem, etc.) experiments.
  - You will be introduced to them both!

Project Deliverable

- Poster
  - Research poster
  - Samples will be shared

- Presentation
  - About 30 minutes per group
  - Introduction, overview, description of the approaches tested, methods, results, discussion, and conclusion

Project Timeline

- First week (the week of May 24th)
  - Topic exploration
  - Form groups and decide a topic
  - Brief project proposal due
- Second week (the week of May 31st)
  - Search data and explore tools to use
  - Finalize topic and official project proposal due
- Third and fourth week
  - Develop one or more hypotheses and investigate, test, and improve appropriate approaches
  - Work with your instructors

Project Timeline

- Fifth week (the week of June 21st)
  - Start to have results
  - Revise your approach based on error analysis
  - Work closely with your instructors
- Sixth week (the week of June 28th)
  - Wrap-up
  - Create a poster and presentation slides
  - UHF Board Meeting on June 25th
  - Poster due on June 25th
  - The presentation will be on June 28th

Rules and Policies

- Certificate award will be determined by following criteria:
  - Active participation in class
  - Completion of hands-on practices
  - Preparedness of weekly project update talk (and/or paper review)
  - Successful completion of term project

Any Questions?

Introduction to Data Mining

Next Class

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL