Lecture 0:
Course Overview

BIOMEDICAL DATA SCIENCE

Instructor:

Parisa Rashidi

FALL 2019

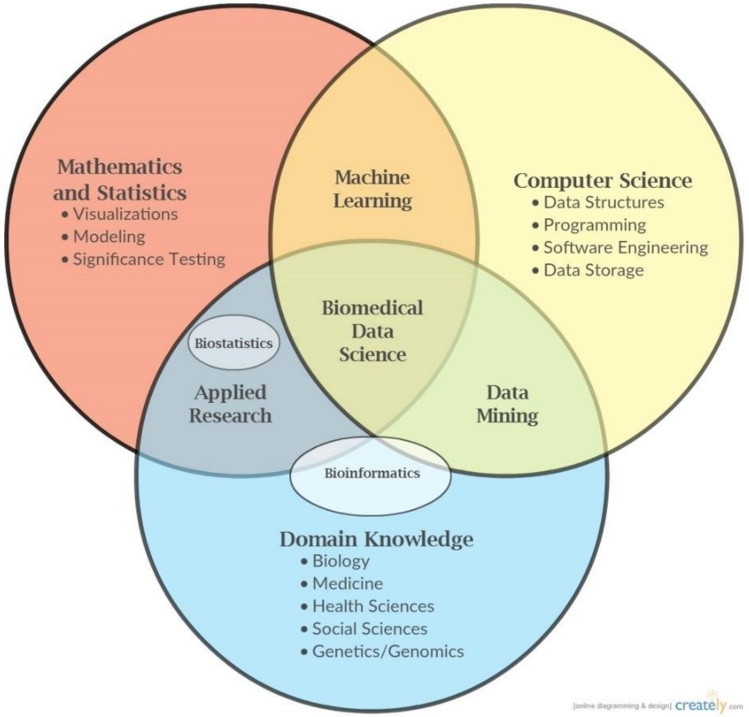
Agenda

- Short Introduction
  - What you will learn in this course
- Logistics
  - What is expected of you in this course

Introduction

Biomedical Data Science

- Data science techniques applied to biomedical science problems



Mathematics and Statistics
- Visualizations
- Modeling
- Significance Testing

Machine Learning

Computer Science
- Data Structures
- Programming
- Software Engineering
- Data Storage

Biostatistics

Biomedical Data Science

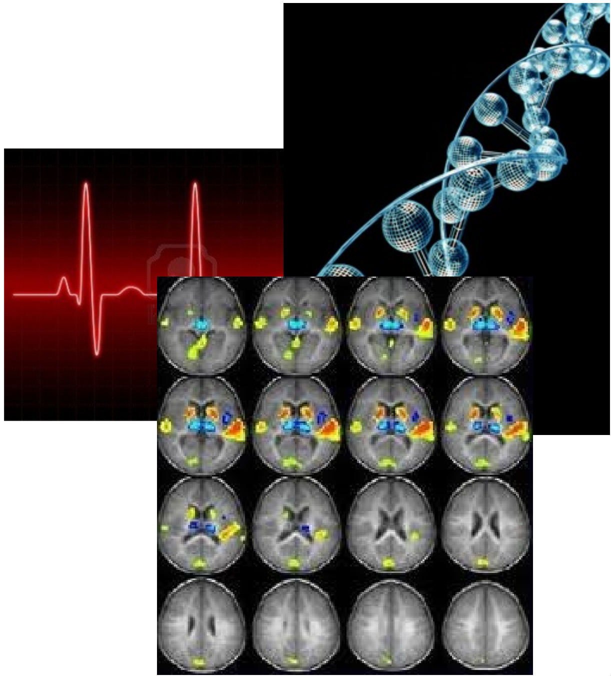
Applied Research

Data Mining

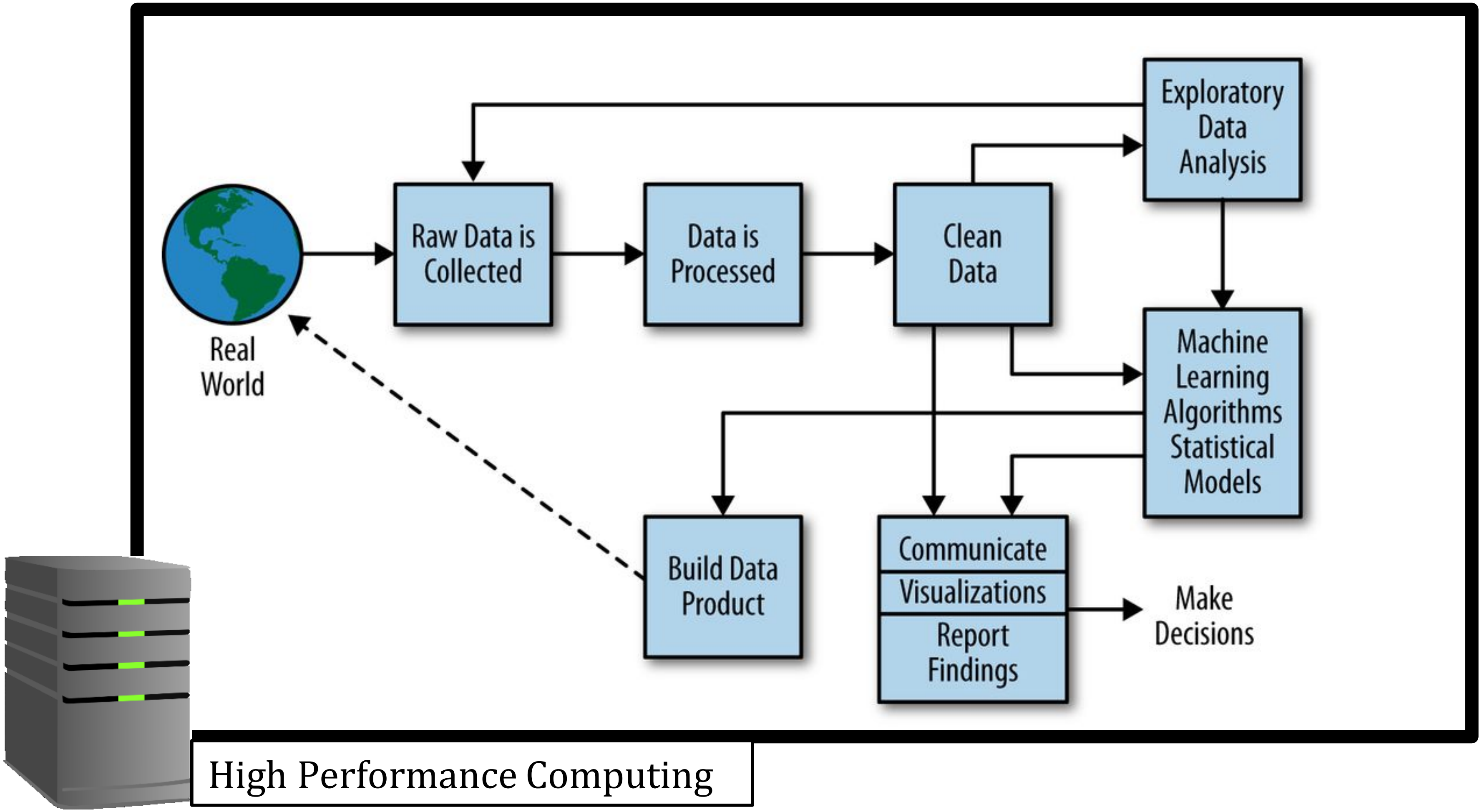
Bioinformatics

Domain Knowledge
- Biology
- Medicine
- Health Sciences
- Social Sciences
- Genetics/Genomics

[online diagramming & design] creately.com

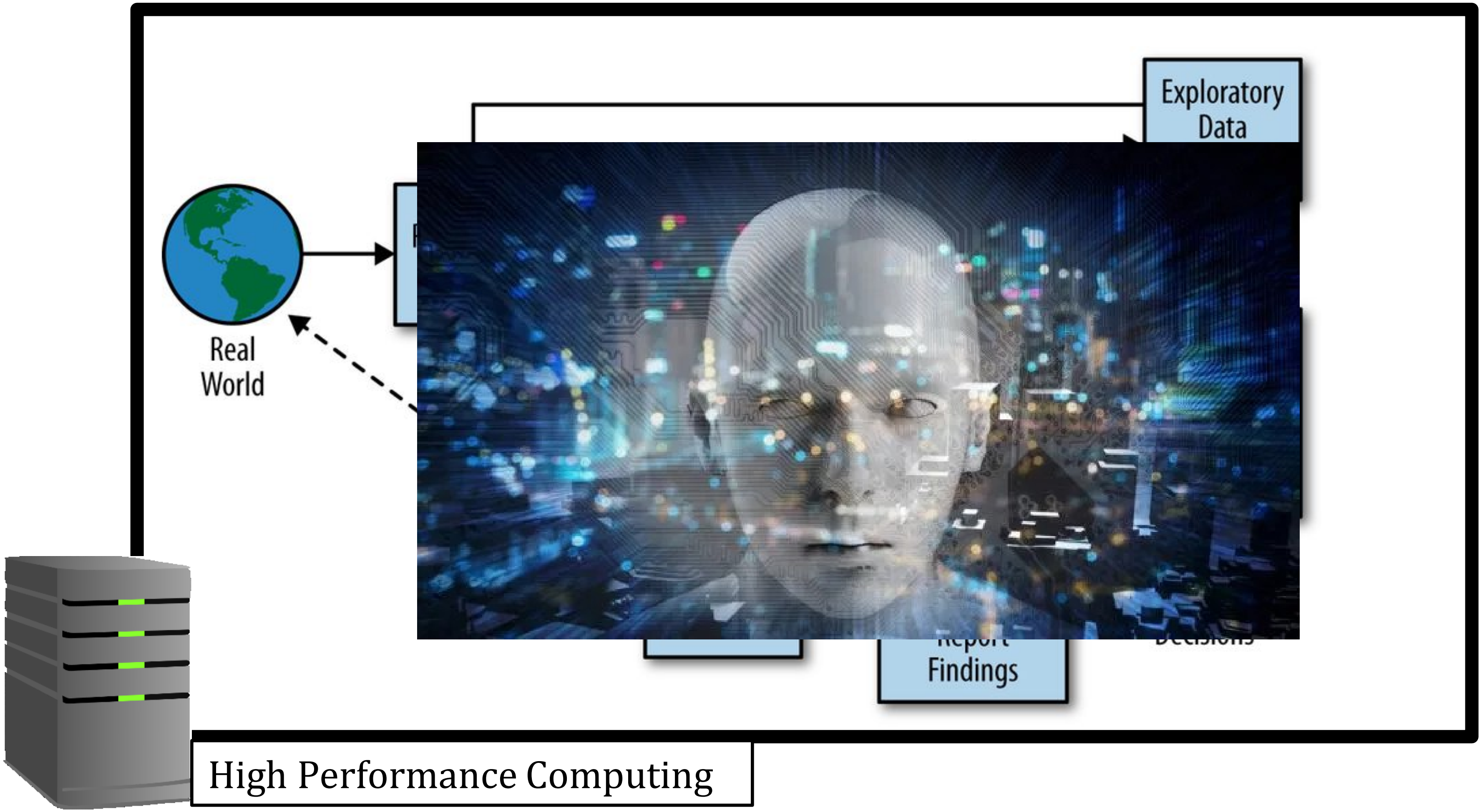


What is Data Science?



High Performance Computing

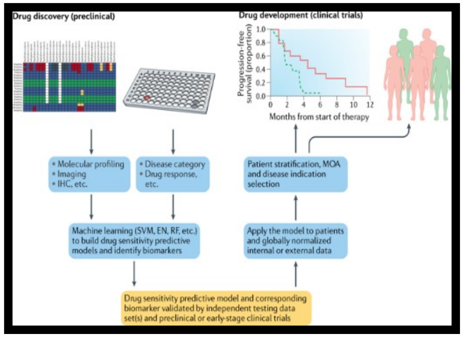
Beware of the AI Hype



High Performance Computing

Why Biomedical Data Science?

Why Study Data Science?


Drug discovery (preclinical)
Drug development (clinical trials)
Progression-free survival (proportion)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
Months from start of therapy

- Molecular profiling
- Imaging
- IHC, etc.

- Disease category
- Drug response, etc.

Machine learning (SVM, EN, RF, etc.) to build drug sensitivity predictive models and identify biomarkers

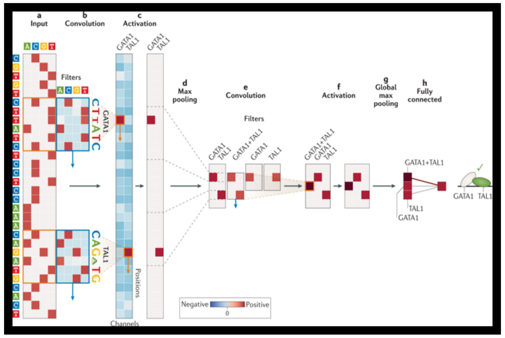
Drug sensitivity predictive model and corresponding biomarker validated by independent testing data set(s) and preclinical or early-stage clinical trials

Apply the model to patients and globally normalized internal or external data

Patient stratification, MOA and disease indication selection

ML FOR DRUG DISCOVERY

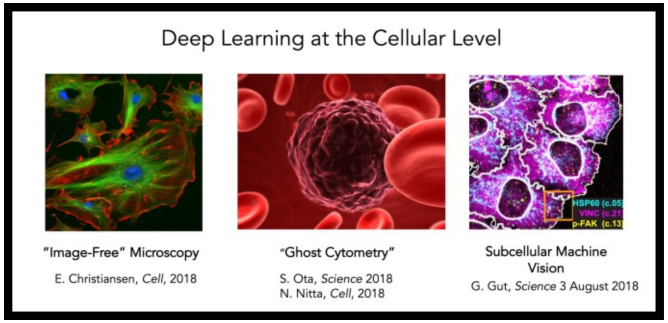
(*NATURE REVIEWS-DRUG DISCOVERY, 2019*)



MODELLING TRANSCRIPTION FACTOR BINDING SITES

(*NATURE REVIEW -GENTICS, 2019*)

nature
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images
Science
NOLOGIES
MING BIOLOGY
nature medicine
Medicine in the Digital Age



Deep Learning at the Cellular Level

"Image-Free" Microscopy

E. Christiansen, *Cell*, 2018

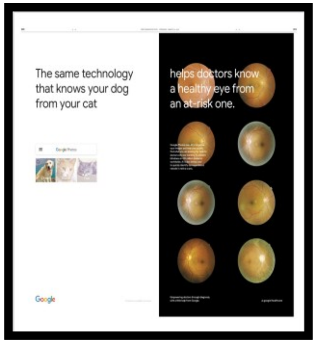
"Ghost Cytometry"

S. Ota, *Science* 2018

N. Nitta, *Cell*, 2018

Subcellular Machine Vision

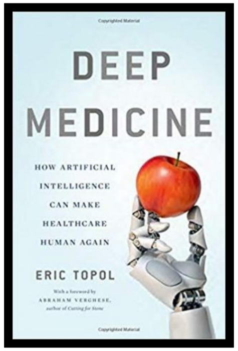
G. Gut, *Science* 3 August 2018



The same technology that knows your dog from your cat

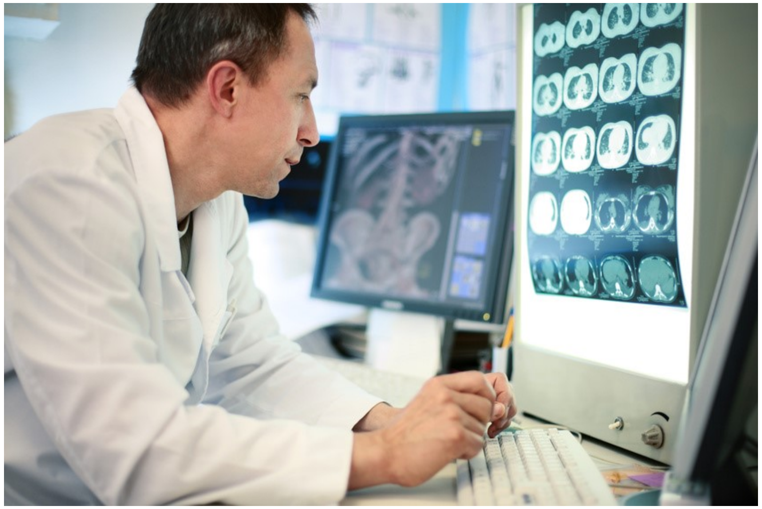
helps doctors know a healthy eye from an at-risk one.

AI.GOOGLE/HEALTHCARE


DEEP MEDICINE
HOW ARTIFICIAL INTELLIGENCE CAN MAKE HEALTHCARE HUMAN AGAIN
ERIC TOPOL
With a foreword by ABRAHAM VERGHESE, author of Cutting for Stone

Current Healthcare Themes (1): Electronic Health Records (EHR)

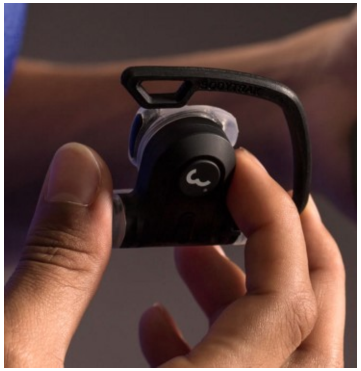
- **Moving beyond paper-based records**
  - Regional and national health data integration facilitating
    - Patient care
    - Administrative & financial
    - Research
    - Scholarly information
    - Office automation



Current Healthcare Themes (2): Smart and Connect Health

- Smart and connected health (Quantified Self)

Polar: embedded with heart rate-monitoring sensors, a motion-tracking sensor to track speed, distance and acceleration



Bodytrak: a pair of earbuds equipped with an in-ear thermometer to measure core body temperature

Flow, a smart air quality tracker

Current Healthcare Themes (3): Omics, Imaging, …

- Complex heterogeneous types of data
- Past: single experiment, small set of results
- Today: using sophisticated instruments, we can generate very large datasets, e.g. gene sequencing.

Link

Current Healthcare Themes (4): Information Access

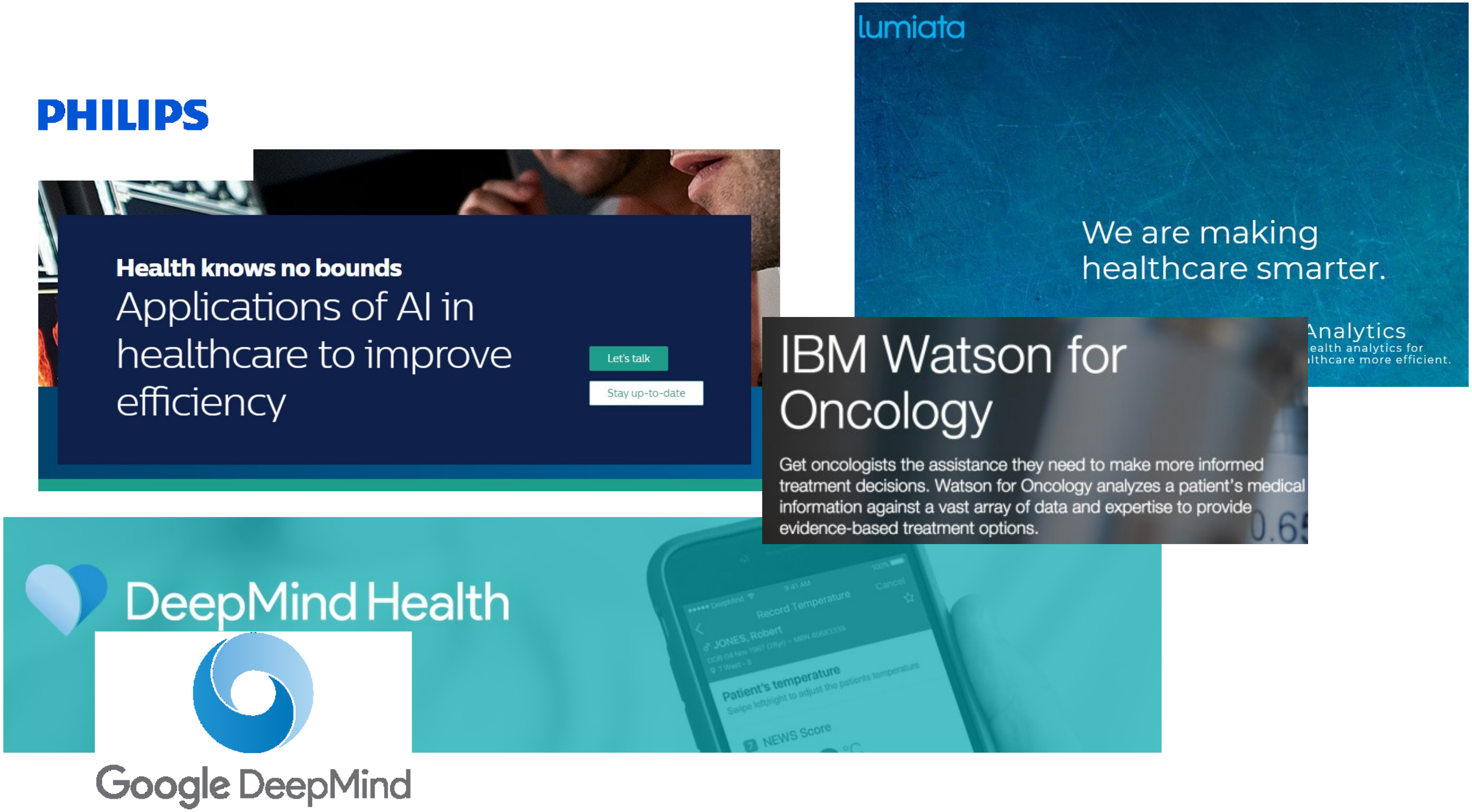
- **Information access**
  - Health-related searches among the most popular
    - Most Web information are anecdotal

WebMD

PubMed

ttp://wholeheartedlyhealthy.com/2012/06/helping-you-find-healthy-2.html

Industry & Healthcare



PHILIPS

Health knows no bounds

Applications of AI in healthcare to improve efficiency

Let's talk

Stay up-to-date

lumiata

We are making healthcare smarter.

Analytics
ealth analytics for
lthcare more efficient.

IBM Watson for Oncology

Get oncologists the assistance they need to make more informed treatment decisions. Watson for Oncology analyzes a patient's medical information against a vast array of data and expertise to provide evidence-based treatment options.

DeepMind Health

Google DeepMind

Healthcare Costs

- 2019: $3.8 trillion, $10K+ per person
- 1960: $27.2 billion, $146 per person

Top 10 most valuable companies combined

EXXON Google BERKSHIRE HATHAWAY Johnson&Johnson WELLS FARGO Roche Walmart

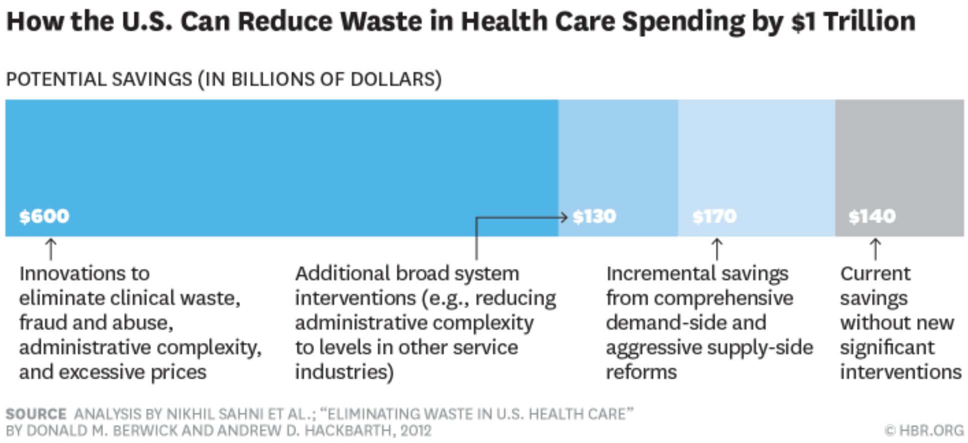
Net worth of

$72 billion $58 billion

Yan Liu, Jimeng Sun, Deep Learning Models for Health Care - Challenges and Solutions, 2017

Healthcare is Broken

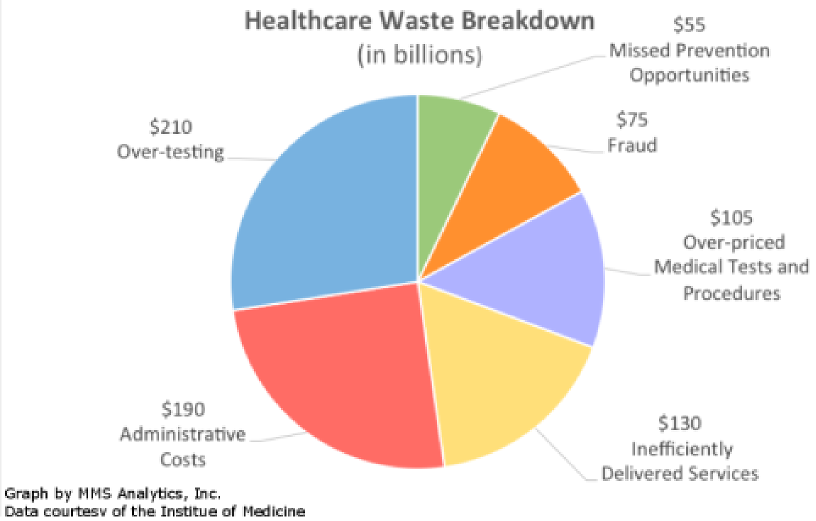
How the U.S. Can Reduce Waste in Health Care Spending by $1 Trillion



SOURCE ANALYSIS BY NIKHIL SAHNI ET AL.; "ELIMINATING WASTE IN U.S. HEALTH CARE" BY DONALD M. BERWICK AND ANDREW D. HACKBARTH, 2012

© HBR.ORG

$765 billions = 50 years budget NASA



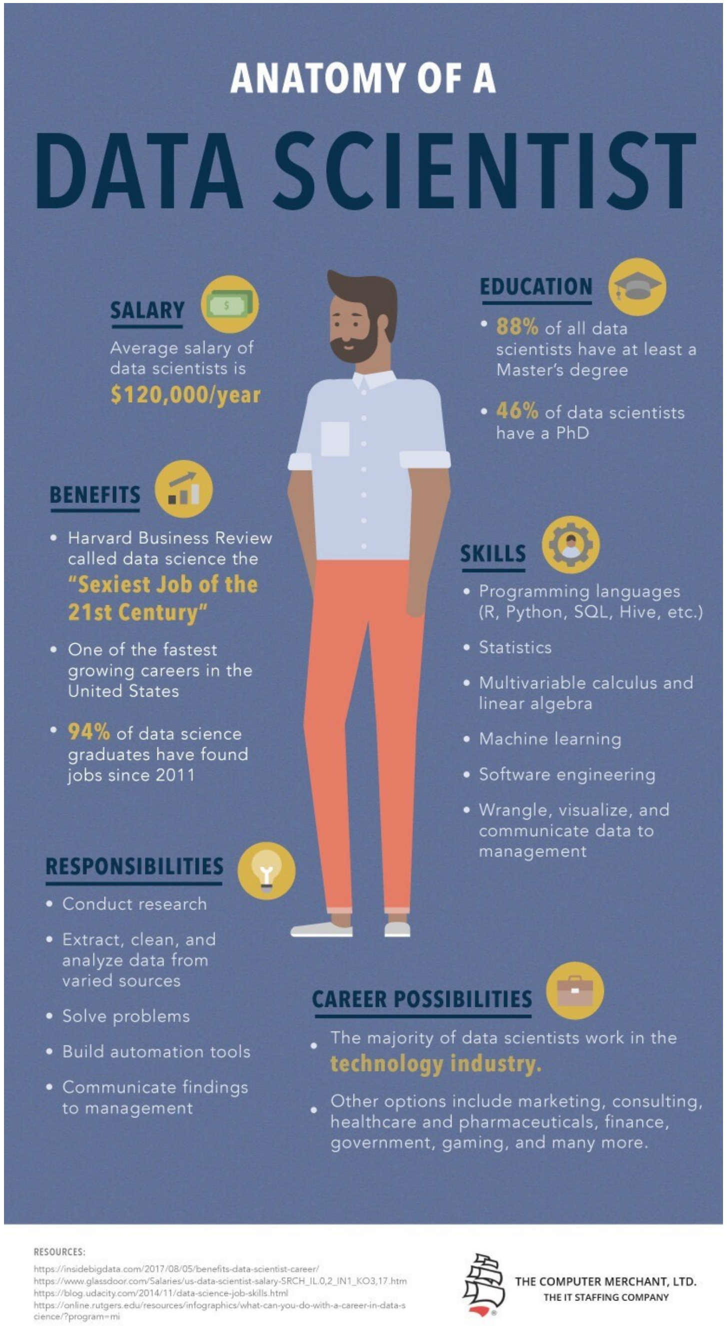
Graph by MMS Analytics, Inc.
Data courtesv of the Institue of Medicine

Yan Liu, Jimeng Sun, Deep Learning Models for Health Care - Challenges and Solutions, 2017

Data Revolution

- All the data processing we did in the last 2 years is more than all the data processing we did in the last three thousand years.
- We are now being exposed to as much information in a single day as our 15th century ancestors were exposed to in their entire lifetime.
- Every two days the human race is now generating as much data as were generated from the dawn of humanity through the year 2003.

http://pressroom.pbs.org/Programs/h/The-Human-Face-of-Big-Data.aspx7



ANATOMY OF A DATA SCIENTIST

SALARY

Average salary of data scientists is **$120,000/year**

EDUCATION

- **88%** of all data scientists have at least a Master's degree
- **46%** of data scientists have a PhD

BENEFITS

- Harvard Business Review called data science the **"Sexiest Job of the 21st Century"**
- One of the fastest growing careers in the United States
- **94%** of data science graduates have found jobs since 2011

SKILLS

- Programming languages (R, Python, SQL, Hive, etc.)
- Statistics
- Multivariable calculus and linear algebra
- Machine learning
- Software engineering
- Wrangle, visualize, and communicate data to management

RESPONSIBILITIES

- Conduct research
- Extract, clean, and analyze data from varied sources
- Solve problems
- Build automation tools
- Communicate findings to management

CAREER POSSIBILITIES

- The majority of data scientists work in the **technology industry.**
- Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.

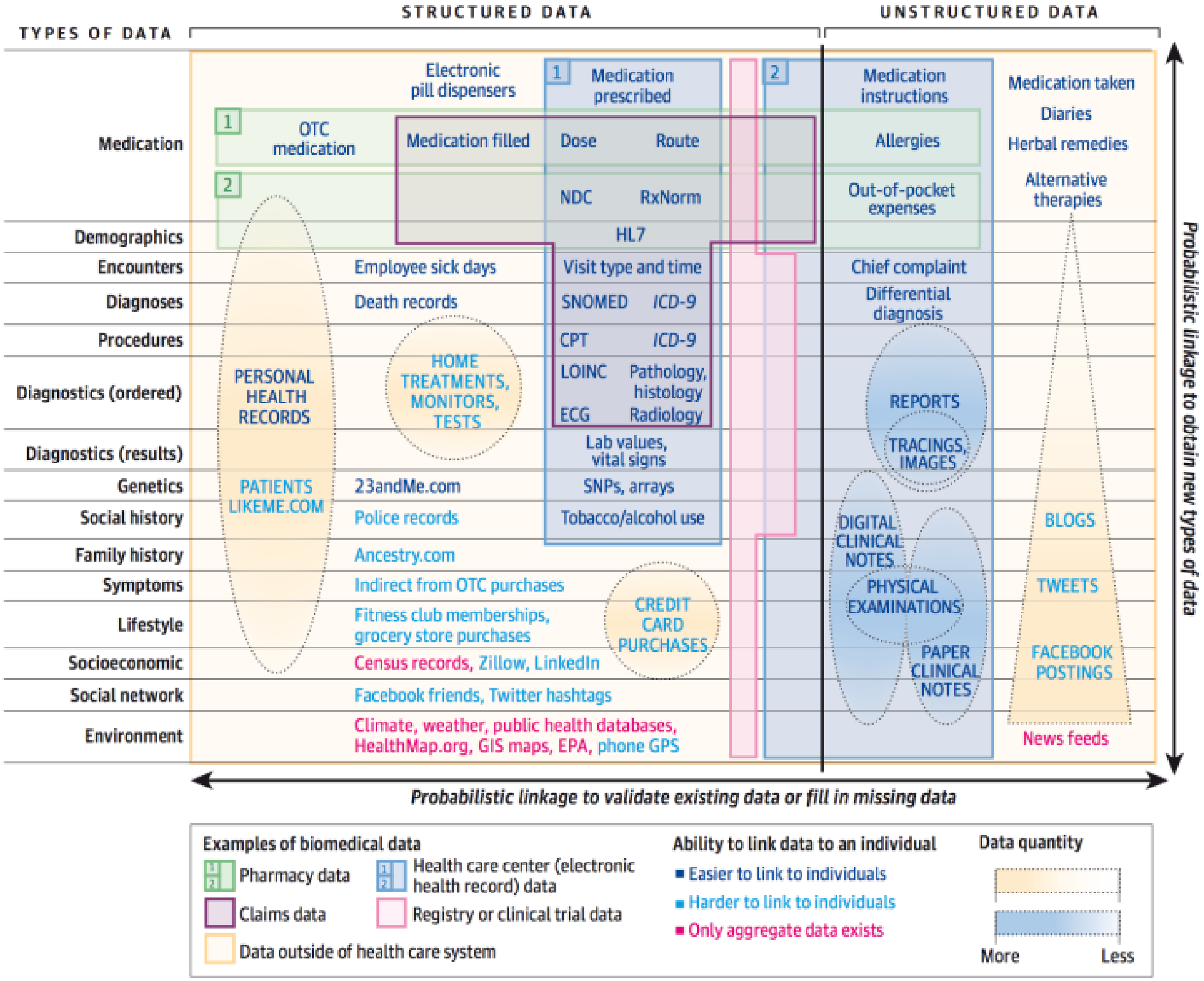
RESOURCES:

https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
https://online.rutgers.edu/resources/infographics/what-can-you-do-with-a-career-in-data-science/?program=mi

THE COMPUTER MERCHANT, LTD.
THE IT STAFFING COMPANY

Image credit: The IT staffing company

Healthcare data sources

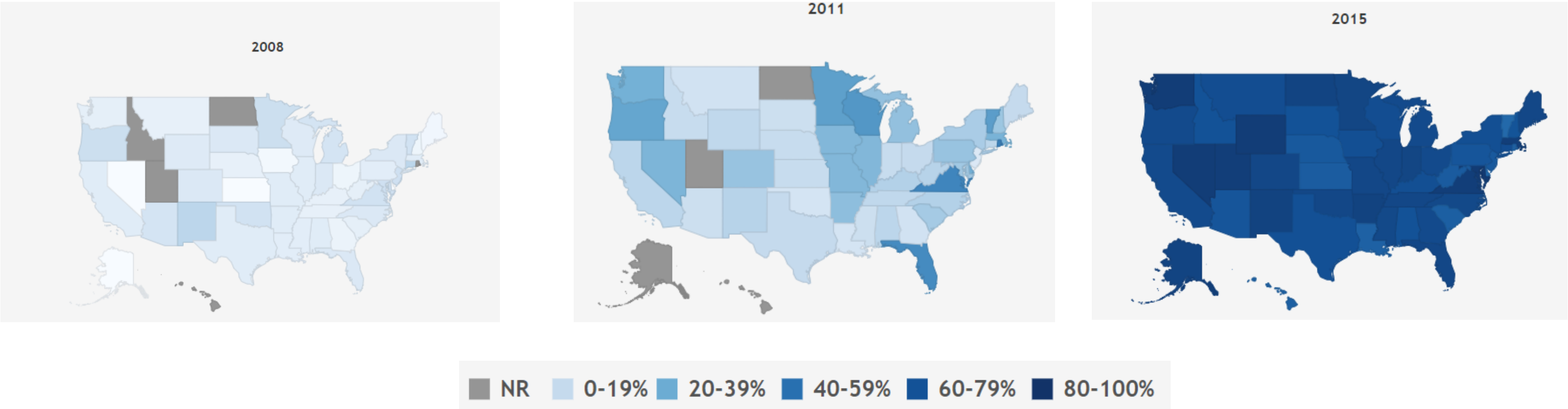


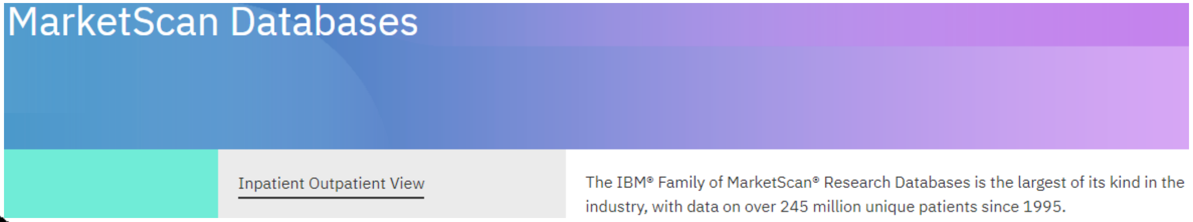
Weber, Griffin M., Kenneth D. Mandl, and Isaac S. Kohane. 2014. "Finding the Missing Link for Big Biomedical Data." *JAMA: The Journal of the American Medical Association* 311 (24): 2479–80.

Clinical Data

- Adoption of Electronic Health record (EHR) data

% non-federal acute care hospitals with EHR adoption





MarketScan Databases

Inpatient Outpatient View

The IBM® Family of MarketScan® Research Databases is the largest of its kind in the industry, with data on over 245 million unique patients since 1995.

Example: MarketScan dataset
245 million patients

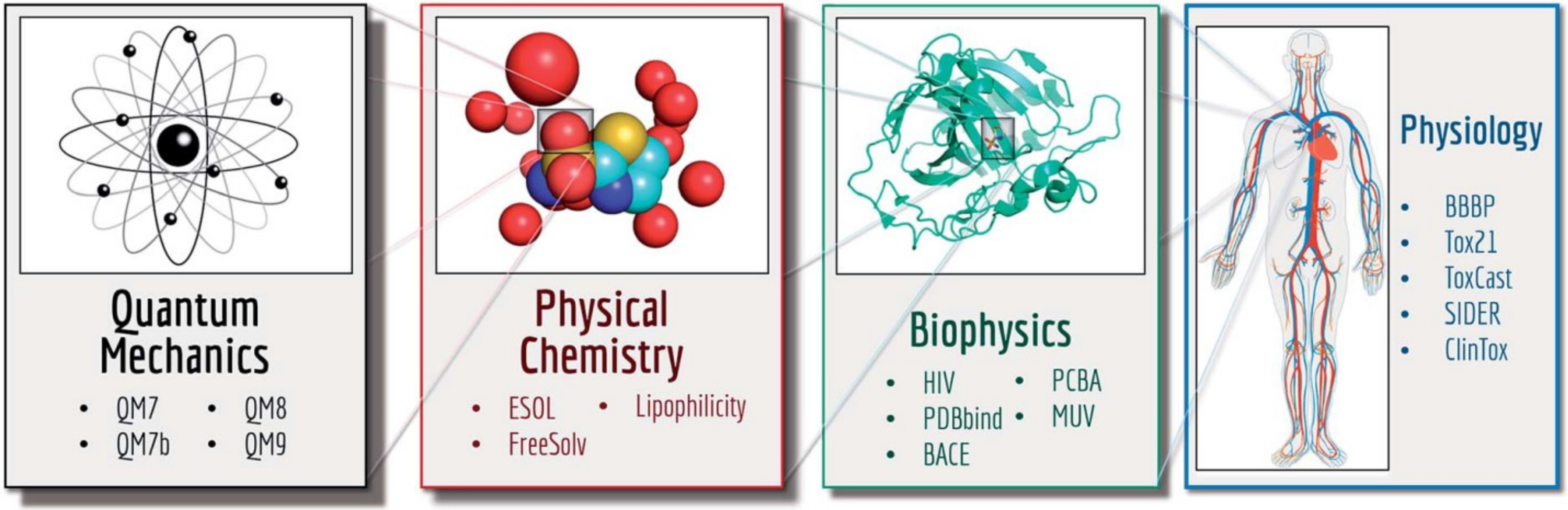
Physiological Data

If you use MIMIC data or code in your work, please cite the following publication:

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

MoleculeNet

- A large set of dataset useful for molecular machine learning



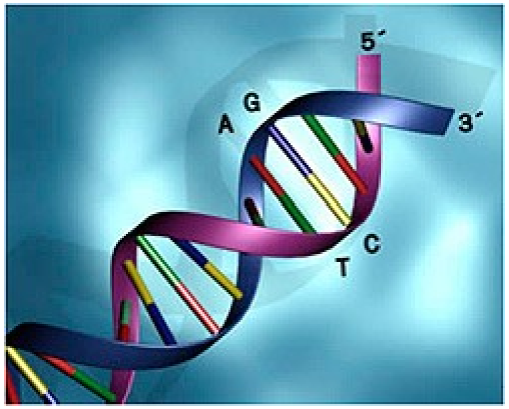
Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Many Other Sources of Data

Biomedical Images

Social Media

Wearables



Genomics

Lab Tests

Course Logistics

Main Topics Covered

- It can be subject to some change:
    - Python Programming
    - Biomedical Ontologies (e.g. ICD-9, SNOMED, …)
    - Conventional Machine Learning Techniques
        - kNN, SVM, RF
    - Deep Learning
        - Images, Time series, Text data
    - Useful machine learning and deep learning libraries
        - Scikit-learn, Pandas, Keras, DeepChem
    - Introduction to Natural Language Processing (NLP)

Week	Date	Tuesday Lecture	Thursday Lecture	Code Examples	Paper Reading	HW
1	08/19 - 08/23	Course Overview	Data Sci. Introduction, Python Programming	• Intro. To Python • Setting up Google Colab	-	-
2	08/26- 08/30 (Add/Drop)	EHR, NumPy	ML Metrics	• NumPy Intro • Toxicity Prediction • Solubility Prediction	Watson Oncology	
3	09/02 - 09/06	Pandas, Preprocessing	Version Control	• Pandas Intro • Preprocessing (5) • Setting up Git		HW1
4	09/09 - 09/13	RF, kNN	Biophysical Modeling	• kNN • RF • Biophysical Modeling	Septic Shock Prediction	
5	09/16 - 09/20	Evaluation	SVM	• Evaluation		HW2
6	09/23 - 09/27	Neural Network	Neural Network, Deep Learning	• FC Networks	Melanoma Detection	
7	09/30 - 10/04	Review	Exam 1			
8	10/07 - 10/11	ConvNet	Convnet (cont.)	• CNN	Adversarial Attacks	
9	10/14 – 10/18	CNN Architectures	Microscopy	• Cell Counting • Cell Segmentation		HW3
10	10/21 – 10/25	RNN	RNN	• Physiological Signals (MIMIC)	Drug Discovery	
11	10/28 – 11/01	Intro to NLP	NLP Representations	• Clinical Notes		HW4
12	11/04 – 11/08	Generative Models	Genomics	• Generating Drug Molecules • TF Binding • RNA Interference • Chromatic Accessibility		
13	11/11 – 11/15	Feature Selection	Guest Lecture			HW5
14	11/18 – 11/22	Review	Exam 2			
15	11/25 – 11/29	Ethical AI	Holiday			
16	12/02 – 12/06		Reading Day			

Course Objectives

- At the end of the semester, students are expected to be able to:
    - Understand what biomedical data science is,
    - Identify different techniques used to solve biomedical data science problems,
    - Identify when and why a certain library or platform should be used,
    - Demonstrate the ability to apply methods from each of the major domains to solve practical problems.

Course Website

- Everything will be available on Canvas
- Canvas should be considered as the reference
  - All announcements, project postings, schedule changes, ..
  - Check your email!
  - Upload your photo

Class

- We will meet on the following days:
  - Tuesdays
    - One session: 1:55 – 2:45 pm
  - Thursdays
    - First session: 1:55 - 2:45 pm
    - 15 minutes break
    - Second session: 3:00 – 3:50 pm

Office Hours

- Office location: NEB 459
- Office hours: by appointment
- E-mail address: parisa.rashidi@ufl.edu
- NOTE: When contacting by email include "Course – 4760/6938" in the subject line to ensure delivery.

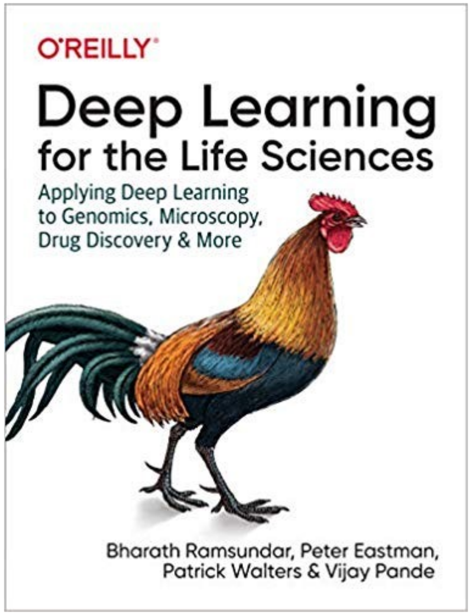
Supervised Teaching Student

- Subhash Nerella, subhashnerella@ufl.edu,
- office location: TBA, office hours: TBA

- Programming/HW questions should be directed to Subhash.

Textbook

- *Recommended*, not required
- Your main source: Lecture Notes,
  - You still have to take your own notes!



Undergraduate Grading

- Final grade is calculated according to:
  - Homework (5) = 35%
    - Equal weight
  - Exams (2) = 45%
    - Equal weight
  - Paper Discussions (5) = 10%
    - Undergraduate students are not asked to present in class, but are expected to contribute to discussions in class and write up a paragraph on each paper (strengths, weakness, suggestions).
  - Quiz (5) = 10%
    - Equal weight

Graduate Grading

- Final grade is calculated according to:
  - Homework (5) = 35%
    - Equal weight
  - Exams (2) = 40%
    - Equal weight
  - Paper Presentation (1) = 10%
    - Graduate students are expected to present in class
  - Paper Discussions (5) = 5%
    - Graduate students are also expected to contribute to discussions in class and write up a paragraph on each paper (strengths, weakness, suggestions).
  - Quiz (5) = 10%
    - Equal weight

Exam

- Exam 1: Thursday, Oct. 3
- held during regular class hours

- Exam 2: Thursday, November 21
- held during regular class hours

Quick Poll

- Programming experience
  - Python
  - Matlab
  - R
  - Other
  - No experience
- Prior machine learning experience

- Introduce Yourself
    - Your name
    - Your major
    - Your research topic if a graduate student
    - Why you enrolled in this class