THE UNIVERSITY OF MEMPHIS

COMP7/8150:
Fundamentals of
Data Science

Week 1: The Big Picture

Max H. Garzon

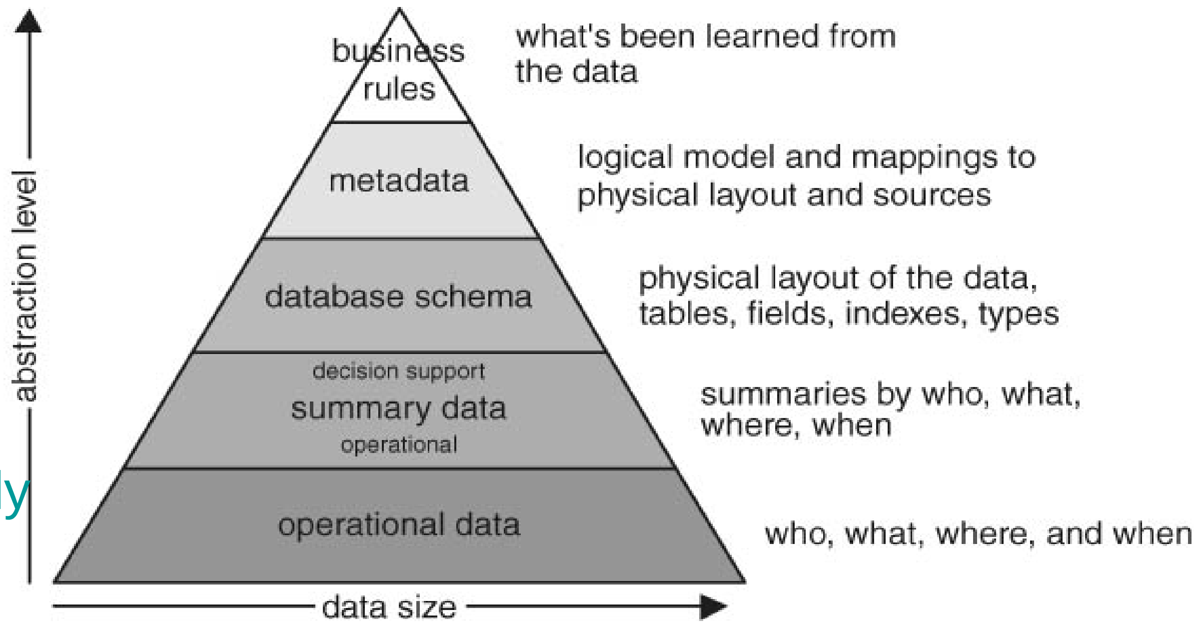
Learning Objectives

- **To characterize key stages in the data life cycle**
  - Problem definition ("business" goal)
  - pre-processing (identify, gather, cleanse, org/store)
  - Processing (patterns, trends: data => info => knowledge)
  - Presentation and deployment (back to the real world)
- Characterize **"**Data**" > "**Information**" > "**Knowledge**"**
- Characterize/define Data Science
- To identify basic features in a computing environment:
  - data capture, visualization and understanding
  - Analysis (statistical, clustering, mining, forecasting)
  - graphical presentation and visualization for results/deployment

What is DATA?

- IFIP: "[Data is] a representation of facts or ideas in a formalized manner capable of being communicated and manipulated by some process." [2]
- Relation to world is other sciences' business
- Data is produced at nearly exabytes **daily** (weather, transactional, financial, social, research, archival, ..)
- Data is ..

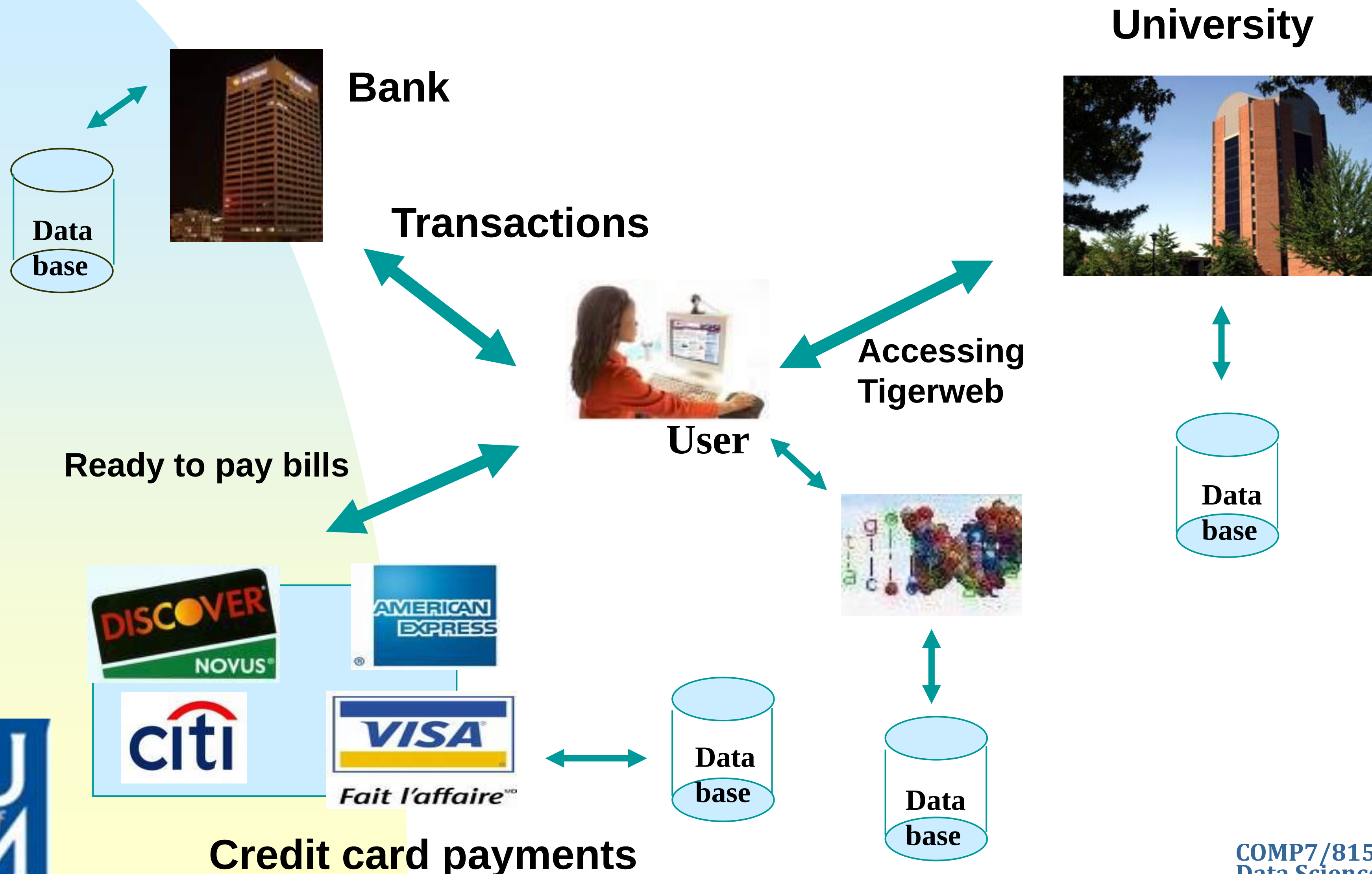
everything/everywhere!



Data is an event of some energy pattern in the physical world hitting a sensor and being recorded somewhere.

[1] G.J. Myatt, W.P. Johnson (2014). **Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining**. John Wiley & Sons.
[2] I. H Gould (1971). *IFIP Guide to Concepts and Terms in Data Processing. John Wiley & Sons.*

Real World Examples



.. What is DATA?

- **Examples**
  - Financial Systems (NYSE, TIAA/CREF, ..)
  - Weather Systems
  - Banking/credit systems
  - Airline Reservation Systems
  - Telecomm systems (GPS, ..)
  - The web
  - Services
    CRM, ERM, Search, Business Intelligence
  - Scientific Research Data
  - Many others

History of Data Science (DS) [3]

1300s 2017

- Many examples in history of science
  - Physics: T. Brahe / J. Kepler (1400s) / Galileo (1600s) / ..
  - Probability and Statistics (1700s)
  - Computer Science (1950-80s)
    Peter Naur/IFCS use of “Data Science” to describe CS

    J. Wu call to rename Statistics Data Science
  - Internet / web (late 1900s) / Gene banks [4]

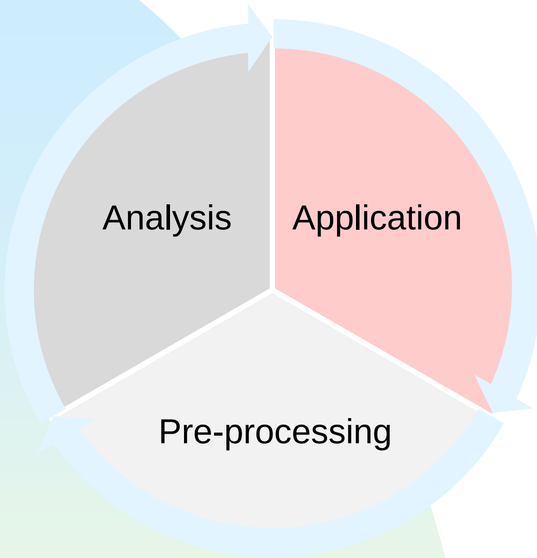
    (Emergence of first massive data repositories)
    By 2025, genomics is expected to become the largest producer of data (exceeding astronomy, twitter, social networks)
  - Amazon, Google, Microsoft, Facebook .. (Big data outfits)

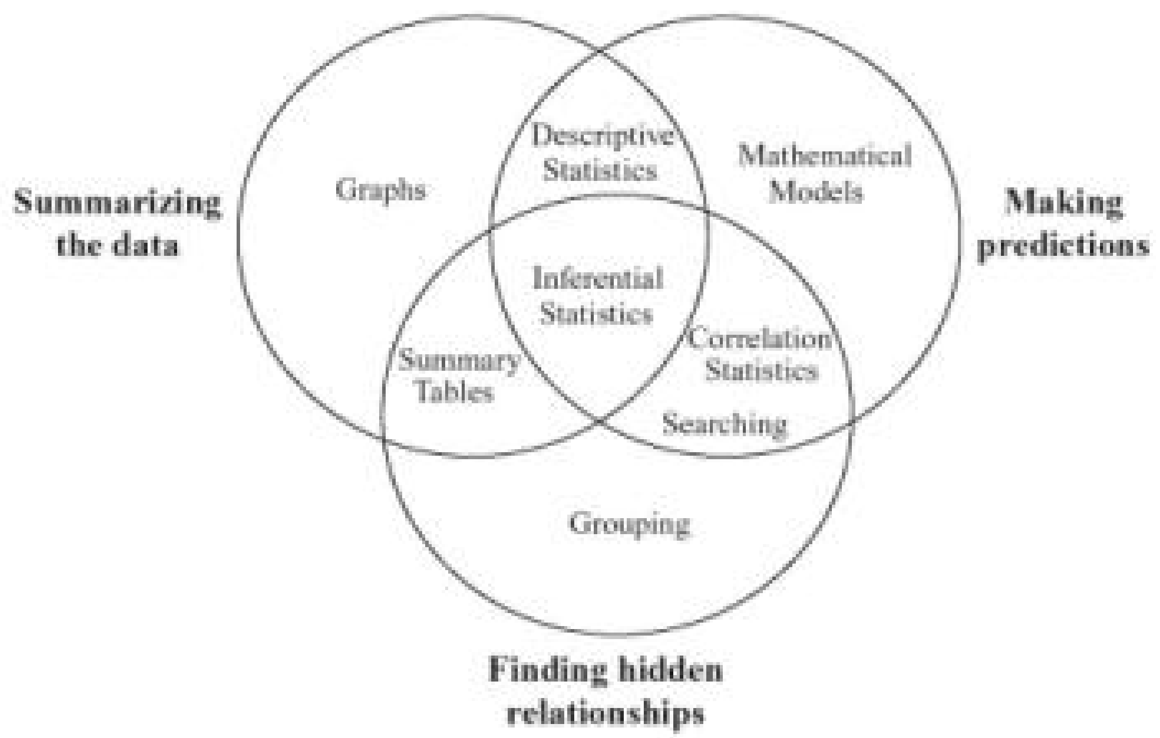
[3] G. Press, A Very Short History of Data Science (2013). Forbes Magazine/

Standard Methodology

- **Problem Definition/goal**
  - Identify/specify goals of the data analysis
  - commit to specific deliverables
- **Data pre-processing**
  - Identify appropriate data
  - Acquire data (gather, lookup, understand) Necessary/Relevant? Sufficient? Cleansed?
- **Data processing**
  - Identify methods (gather, cleanse, store)
  - Carry out the analysis (patterns, trends, predictions?)
- **Data post-processing**
  - Visualize and present
  - Deploy and evaluate. Iterate, if necessary

Data Life Cycle





What is Data Science (DS)?

- Fundamental goal of DS is
  to turn data into information and knowledge
  - Information is money (1980-90s)
  - Knowledge is power (21^{st} century)
  - Data Science = Data + Analysis + Hacking + Problem Solving
    Child of Statistics (mature groom) and Computer Science (young bride)
- DS is interdisciplinary/requires a lot of help to be effective:
  - Computer Science
    (data mining, machine learning, analytics, HCI, simulations, ..)
  - Probability/Statistics (inferential, models, Bayesian, ..)
  - Mathematics
  - Business / management (Project management, OLAP, ..)

Major Kinds of Problems in DS

Most problems in DS reduce to three fundamental kinds of problems, all of them computational problems

- **Classification (Π)**
  - Instance: an element x from partition of $\Omega = U\ \Pi_i$
  - Question: Which category Π_i in Π does x belong in?
- **Prediction (f)**
  - Given an unknown numerical function f: $\Omega \rightarrow$ R and x in Ω
  - What is the value of f(x)?
- **Clustering (Ω)**
  - Given a similarity metric d= |*,*|: $\Omega \rightarrow$ **R** on a sample space Ω
  - *What's a partition of Ω into disjoint categories so that points in the same category (cluster) are more similar to each other than to points in any other category?*

.. Major Kinds of Problems in DS

Examples are

- **Classification (Π)**

  Iris Flower classification ({Setosa,Versicolor, Virginica})

  - Instance: a 4D feature vector (sl,sw,pl,pw) of an iris flower
  - Question: Which category Π_i in Π does x belong in?

.. Major Kinds of Problems in DS

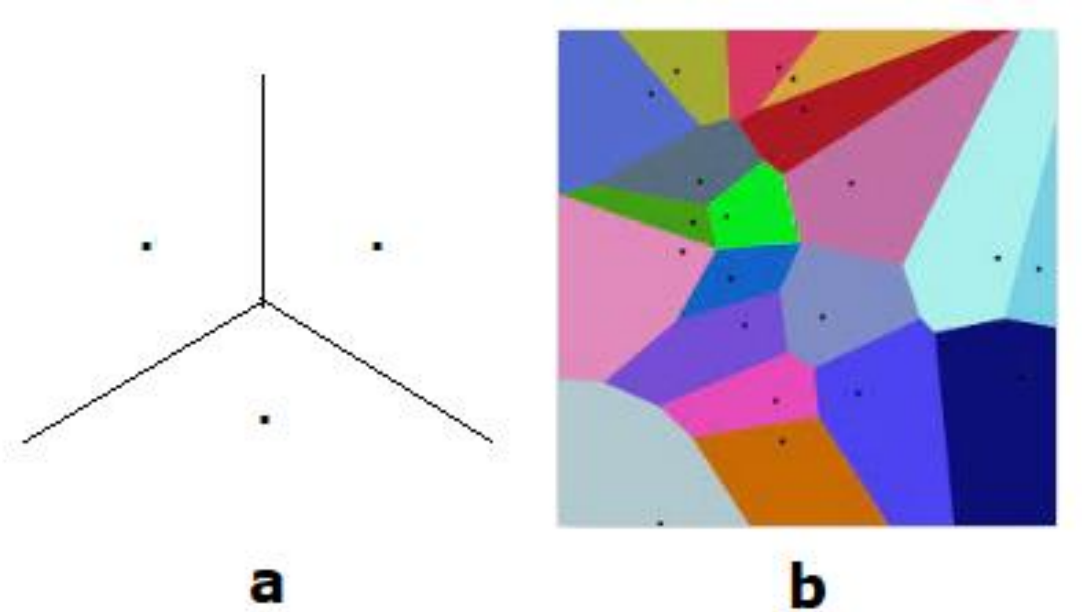
Examples are

- **Prediction (f)**
  Prediction (apotome's area)
  - Given DNA genes (e.g., COI, COII, COIII) of a backfly
  - What is the area of the larva's apotome?

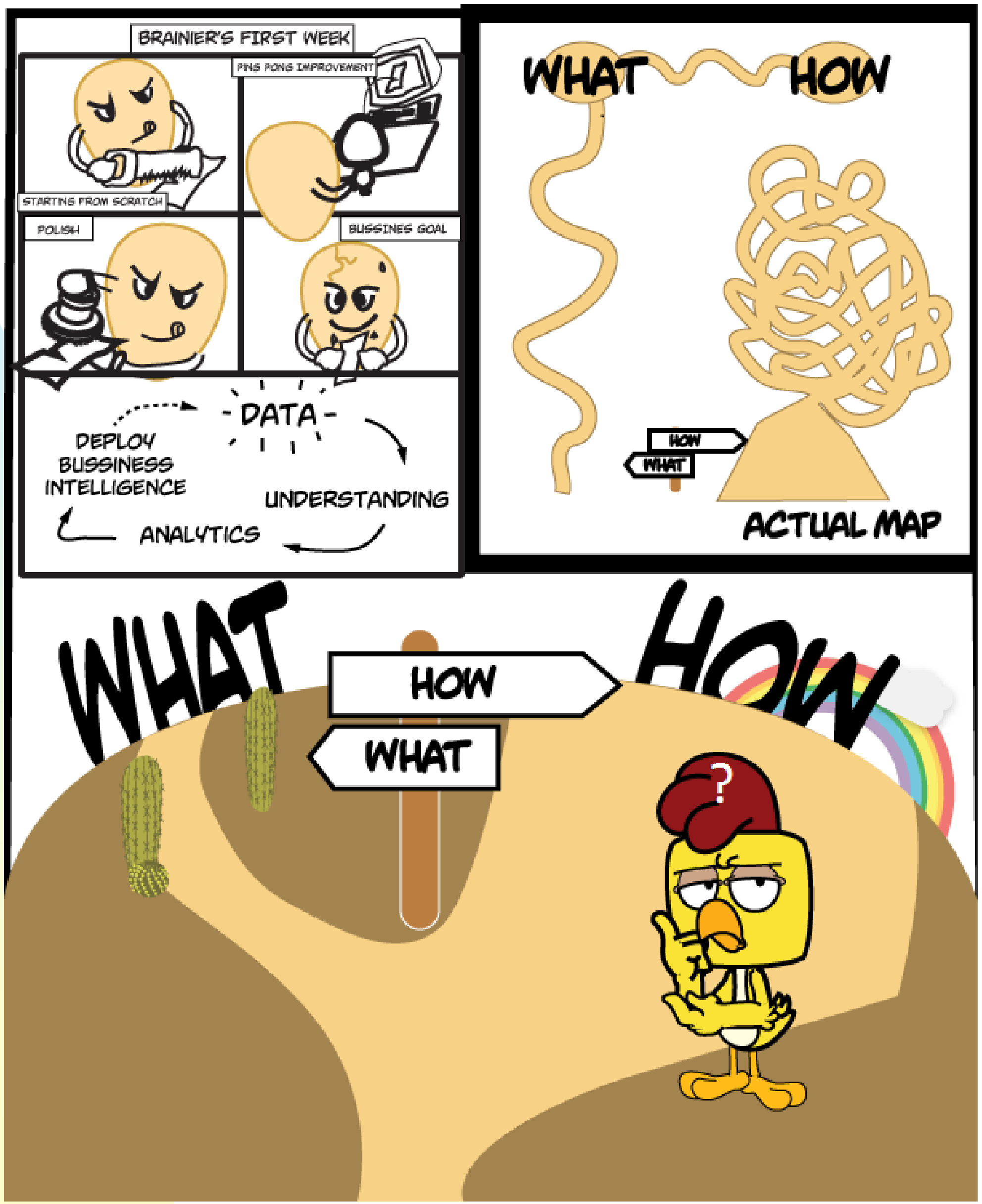
.. Major Kinds of Problems in DS

Examples are

- **Clustering (Ω)**
  Clustering (Memphis area)
  - Given 3 fire stations in the Memphis metropolitan area
  - *What's a partition of the houses into three disjoint categories so that they can take care of all fires in the area?*



.. What is DS?


BRAINIER'S FIRST WEEK
PING PONG IMPROVEMENT
STARTING FROM SCRATCH
POLISH
BUSSINES GOAL
DATA
UNDERSTANDING
ANALYTICS
DEPLOY
BUSSINESS
INTELLIGENCE
WHAT
HOW
HOW
WHAT
ACTUAL MAP
WHAT
HOW
HOW
WHAT
?

Case Study: Health Case Systems

- Questions to be addressed for HCSs:
  - What is the overarching goal of the project?
    Provide health care to this group of policy holders
  - What data are relevant to this goal? Where is it?
    Biographic/metric data; Genomic data; legal regulations concerning medical records (e.g., HIPPA, malpractice); actuarial data; latest advances in treatment of certain diseases; …
  - What techniques will transform the data into guidelines to improve quality of care and costs?
    DBs, clustering, hypothesis testing, …
  - How can these guidelines be implemented in practice?
    ParTNers in Health, Preferred premium groups, ..

What is your ***data corpus***?

- *A data point* is an nD vector of features ,
  - Each value is alphabetic or numeric
  - Usually d >> 1 (d is the dimensionality of the points)
  - The size of the table is n=(the # of records stored in it).
- Data points of the same dim can be clustered into data tables
- A data corpus consists of a finite number of data tables, possible of various dimensions, comprising all the data for the project
- There is a variety of means to store the data
  - eFiles
  - Data bases (relational and nonrelational)
  - Multimedia
  - Paper records …

Case Study: Biomedicine

- Genome Projects (human and others)
- Leroy Hood's P4 Medicine (Predictive, Personalized, Preventive, Participatory) [4], Now P5 = P4 + Precision [4b]
- Goal: what information is there in genomic data about a person's future health?
- Data
  - What has his family suffered of in the past?
  - What is his/her OMICS profile?
- Methods to determine
  - what might hw/she be suffering from in the future?
  - OMICS sciences

[4] L. Hood, R. Balling, C. Auffray (2012). Revolutionizing medicine in the 21st century through systems approaches. Biotechnol J. 7(8), 992-1001.
[4b] Big Data to Knowledge Initiative , NIH, 2015.

Case Study: Business

- Goal: What do customers want/How to deliver it?
- Data
  - What has s/he purchased in the past?
  - What is his/her profile?
  - 5Vs: Volume (large), Variety (multiple sources), Velocity (rapidly changing), Veracity (quality), Value (usefulness).
- Methods
  - What might his/her be looking for?
  - Targeted advertising
  - Anticipate needs translates into big savings
- Back to reality
  - Business Intelligence (BI): what decision to make?
  - Stay ahead of the competition

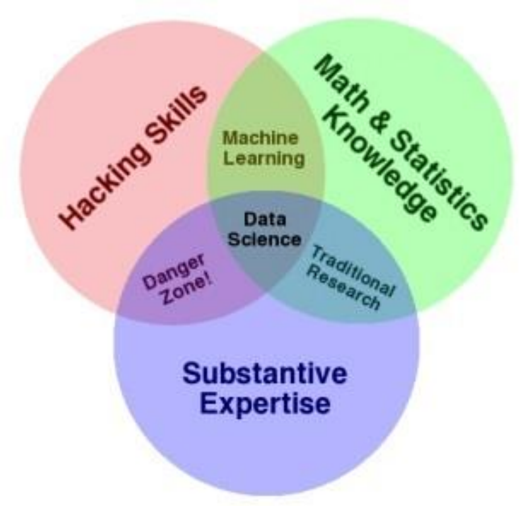
Case Studies: Journalism [5]

- Goal: What do media consumers want and how to deliver it? What should a journalist be?
- Data
  - What is the story is about?
  - What are the facts informing it?
  - What is at stake morally/societally?
  - What are the experts' opinions?
- Methods: google, interview experts, crunch data
- Back
  - What is the take-home?
  - BI: what to take away from the story?

[5] K. Kirkcpatrick (2015), Putting the Data Science into Journalism. Comm ACM, 58:5, 15-17.

Profile of a Data scientist [6]

- What exactly is a data scientist?
  - Not a statistician or CSt (but child therefrom), or Business Analyst, or Data Analyst, or .. just Scientist
  - Someone who extracts value from data
  - Someone who manages DBs ? [Data Science J., 1990s]
  - Someone who applies data ? [J. Data Science, 1990s]
  - Goes by data analyst, dataologist, "person crucial to the success of data science"[NSF'05]
  - A jack of all trades, Life-long learner [Astrophysics-09] and master of reality



Source: Drew Conway (2010)

[6] T. Davenport, D. Patil (2021). Data Scientist: The Sexiest Job of the 21st Century Harvard Business Review.

Areas of Immediate Impact [6]

- Ongoing projects with immediate DS impact:
  - Biomedicine and human health
  - Climate Science
  - Material Science
  - High-Energy Physics
  - Text and Data Mining
  - […]

- Common themes
  - Complex spatio-temporal heterogeneous data
  - Predictive *simulations* requiring models/HPC
  - New research needs and job opportunities

[6] T. Davenport, D. Patil (2021). Data Scientist: The Sexiest Job of the 21st Century Harvard Business Review.

Big DS Initiatives [7]

- **European Union**
  - CREST (Collaborative Research into Exascale Systemware, Tools and Applications)
  - DEEP (Dynamical Exascale Entry Platform)
  - Mont-Blanc (green design for exascale HPCs)
- **North America**
  - Supercomputing: HPCS (ORNL, National Labs, Titan, ..)
  - Corporate: Google, Amazon, Microsoft's Big Data outfits
- **Japan**
  - RIKEN (Collaborative for Exascale system by 2020)
- **China**
  - Tianhe (world's fastest HPC as of 2014) and TH-Express2

[7] D. Reed, J. Dongarra (2015), Exascale Computing with Big Data. Comm ACM, 58:7, 56-68.

Big Challenges from DSs [7]

- Scalability **is a huge issue**
  - Cell phone today is faster/better than Cray 1 (first HPC)
  - GHz is no longer fast. HPCs are at petaflop ranges
  - Sticks hold terabytes; cloud data centers hold petabytes
  - Extreme scales - Computing ecosystems
- **Other technical challenges**
  - Hw/Sw to handle 5Vs: Volume (large), Variety (multiple sources), Velocity (rapidly changing), Veracity (quality) of Value (usefulness).
  - Minimize data movement, robust apps, precision+recall
  - Refactoring problems and solutions in CS/Sci Computing
  - Diverging scientific computing and big-data ecosystems in programming models and tools

[7] D. Reed, J. Dongarra (2015), Exascale Computing with Big Data. Comm ACM, 58:7, 56-68.

.. Big Challenges from DSs

- Computing Ecosystems (costing $1B+)
  - System power consumption and green cooling
  - Metadata and data ontology management
  - Data storage/preservation/fusion/fault-recovery
  - Commoditization of software/PLs/methods/toolkits (Hadoop+Map Reduce, Mahout, Giraph, …)
  - **Economic/political power** will be measured by speed of HPCs/ranking in the supercomputer world

Supercomputer Rankings [8]

Rank	Rmax Rpeak (PFLOPS)	Name	Model	Processor	Interconnect	Vendor	Site country, year	Operating system
1 —	148.600 200.795	Summit	IBM Power System AC922	POWER9, Tesla V100	InfiniBand EDR	IBM	Oak Ridge National Laboratory United States, 2018	Linux (RHEL)
2 —	94.640 125.712	Sierra	IBM Power System S922LC	POWER9, Tesla V100	InfiniBand EDR	IBM	Lawrence Livermore National Laboratory United States, 2018	Linux (RHEL)
3 —	93.015 125.436	Sunway TaihuLight	Sunway MPP	SW26010	Sunway[22]	NRCPC	National Supercomputing Center in Wuxi China, 2016[22]	Linux (Raise)
4 —	61.445 100.679	Tianhe-2A	TH-IVB-FEP	Xeon E5–2692 v2, Matrix-2000[23]	TH Express-2	NUDT	National Supercomputing Center in Guangzhou China, 2013	Linux (Kylin)
5 ▲	23.516 38.746	Frontera	Dell C6420	Xeon Platinum 8280	InfiniBand HDR	Dell EMC	Texas Advanced Computing Center United States, 2019	Linux (CentOS)
6 ▼	21.230 27.154	Piz Daint	Cray XC50	Xeon E5-2690 v3, Tesla P100	Aries	Cray	Swiss National Supercomputing Centre Switzerland, 2016	Linux (CLE)
7 ▼	20.159 41.461	Trinity	Cray XC40	Xeon E5–2698 v3, Xeon Phi 7250	Aries	Cray	Los Alamos National Laboratory United States, 2015	Linux (CLE)
8 ▼	19.880 32.577	AI Bridging Cloud Infrastructure[24]	PRIMERGY CX2550 M4	Xeon Gold 6148, Tesla V100	InfiniBand EDR	Fujitsu	National Institute of Advanced Industrial Science and Technology Japan, 2018	Linux
9 ▼	19.477 26.874	SuperMUC[25]	ThinkSystem SD530	Xeon Platinum 8174 (plus e.g. 32 cloud GPU nodes with Tesla V100[26])	Intel Omni-Path	Lenovo	Leibniz Supercomputing Centre Germany, 2018	Linux (SLES)
10 ▲	18.200 23.047	Lassen	IBM Power System S922LC	POWER9, Tesla V100	InfiniBand EDR	IBM	Lawrence Livermore National Laboratory United States, 2018	Linux (RHEL)

Legend:

- Rank – Position within the TOP500 ranking. In the TOP500 list table, the computers are ordered first by their Rmax value. In the case of equal performances (Rmax value) for different computers, the order is by Rpeak. For sites that have the same computer, the order is by memory size and then alphabetically.

[8] (2018) TOP500 Supercomputers by Country at https://en.wikipedia.org/wiki/TOP500.

World's Fastest Supercomputer

[9] (2018) TOP 2018 - Summit (+ SIerra)
https://en.wikipedia.org/wiki/Summit_(supercomputer)#/media/File:Summit_(supercomputer).jpg.

Move Over, China: U.S. Is Again Home to World's Speediest Supercomputer **Summit**

[9] (2018) TOP 2018 - Summit (+ SIerra)
https://www.nytimes.com/2018/06/08/technology/supercomputer-china-us.html.