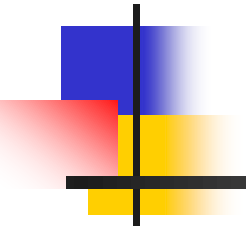


Big Data

Chapter 1 from "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data"
1st Edition by EMC Education Services

Ka-Chun Wong, Department of Computer Science City University of Hong Kong
Charles Tappert Seidenberg, School of CSIS, Pace University

Introductory Example 1

An interesting example of data analytics

A grocery store in the USA found that **when men bought diapers on Thursdays and Saturdays, they also had a strong tendency to buy beer**.

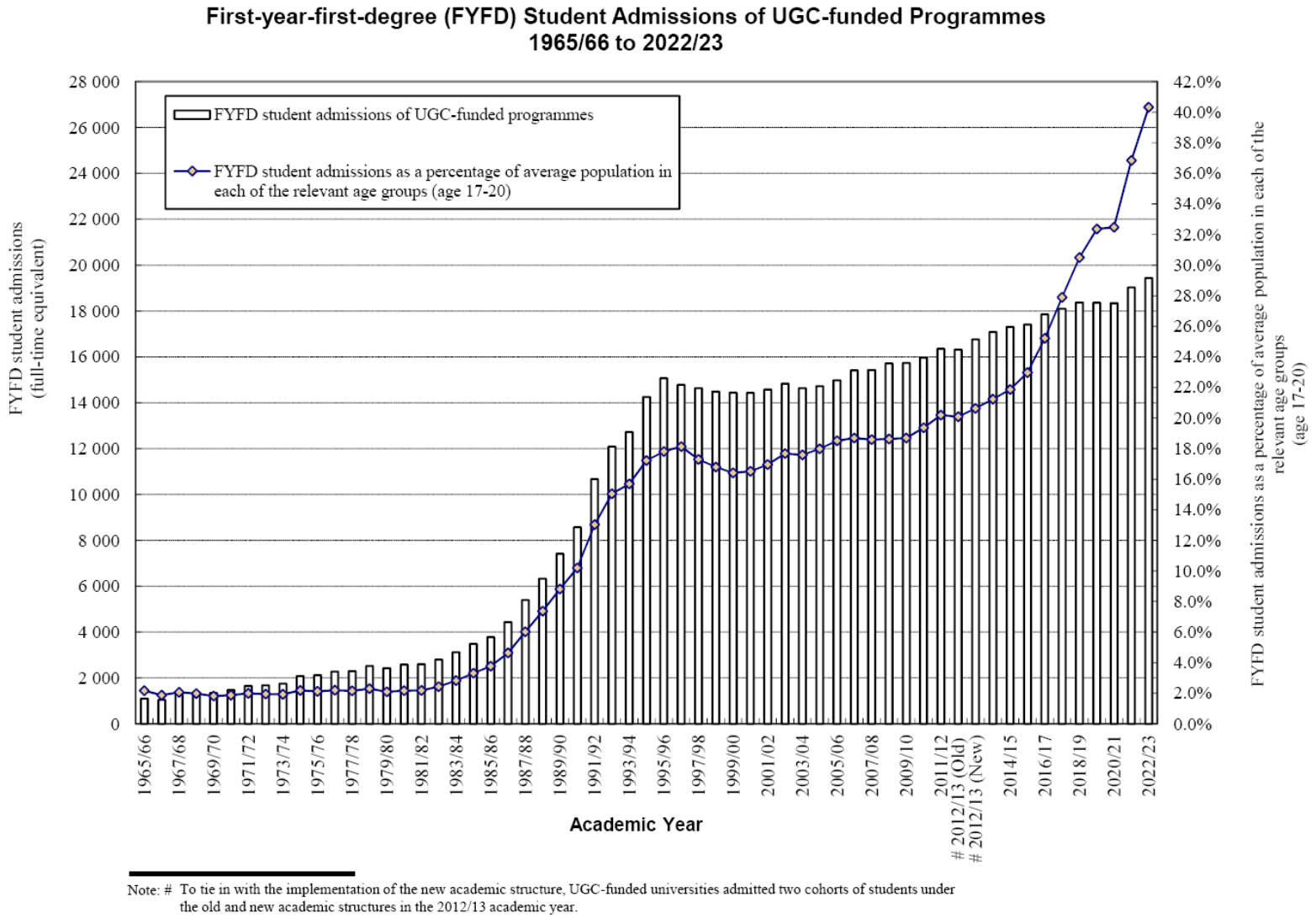
The grocery store could use this valuable information:

-- move the beer display closer to the _____.

-- made sure not to give any discounts on _____ on Thursdays and Saturdays.

http://www.programmerinterview.com/index.php/database-sql/data-mining-vs-warehousing/

Introductory Example 2



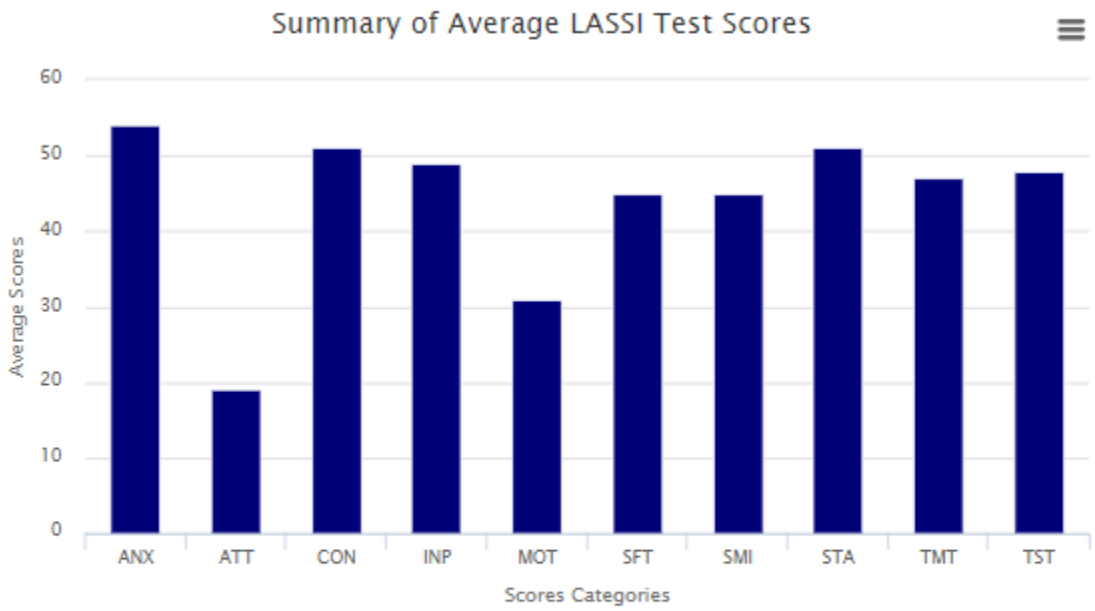
Last updated: June 2023

https://cdcf.ugc.edu.hk/cdcf/statSiteIndex.action#

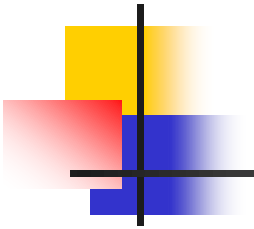
Introductory Example 3

LASSI Test Scores

For explanatory notes of LASSI, please click here.



ANX	Anxiety	**SFT**	Self Testing
ATT	Attitude	**SMI**	Selecting Main Ideas
CON	Concentration	**STA**	Study Aids Scale
INP	Information Processing Scale	**TMT**	Time Management
MOT	Motivation	**TST**	Test Strategies



1.1 Overview

- Industries that gather and exploit data
  - Credit card companies monitor purchase
    - Good at identifying fraudulent purchases
  - Mobile phone companies analyze calling patterns – e.g., even on rival networks
    - Look for customers might switch providers
  - Social networks data is growing
    - Intrinsic value increases as data grows

Introductory Example 4

rthk.hk 香港電台網站 ENGLISH NEWS

Menu

2022.06.02 Thursday 29°C 78%

Search...

LATEST NEWS

Home ▸ Latest News ▸ Local

MIRROR tickets still hard to come by: fans

2022-05-31 HKT 17:20

Recommend 0 Share this story

Getting your hands on a ticket to see the popular boy band MIRROR was never

watsons

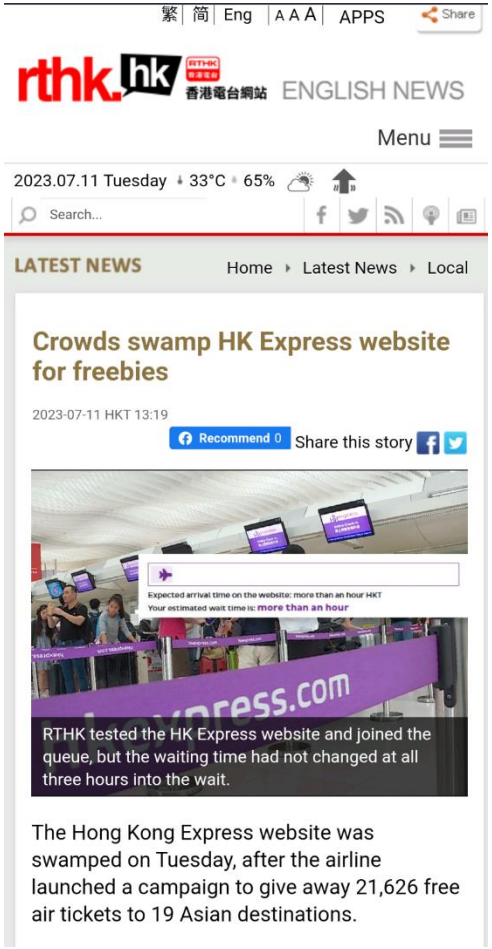
You are now in line

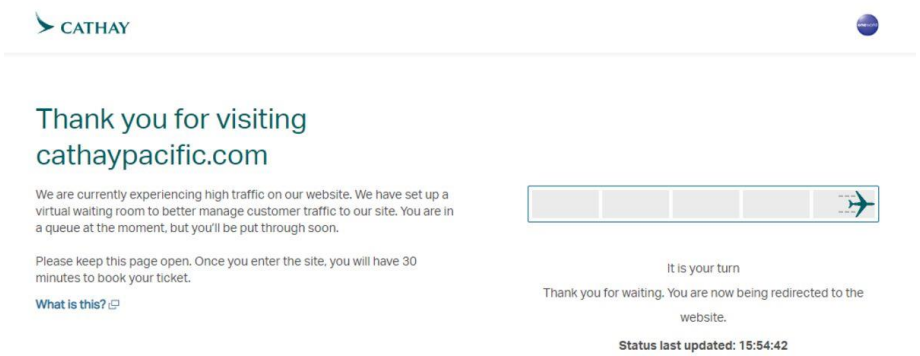
You are in line for Watsons Hong Kong online store. Your order is almost here for you. (Don't close this current page, you will soon enter the website)

What is this?

Expected arrival time: **more than an hour**
Your estimated wait time: **more than an hour**

Introductory Example 5





Attributes Defining Big Data Characteristics

- Huge volume of data
  - Not just thousands/millions, but billions of items
- Complexity of data types and structures
  - Varity of sources, formats, structures
- Speed of new data creation and grow
  - High velocity, rapid ingestion, fast analysis

What is Big Data --

[Li01] Slides #23

Let's look at

Big Data

in a different way.

DevOps Borat

@DEVOPS_BORAT

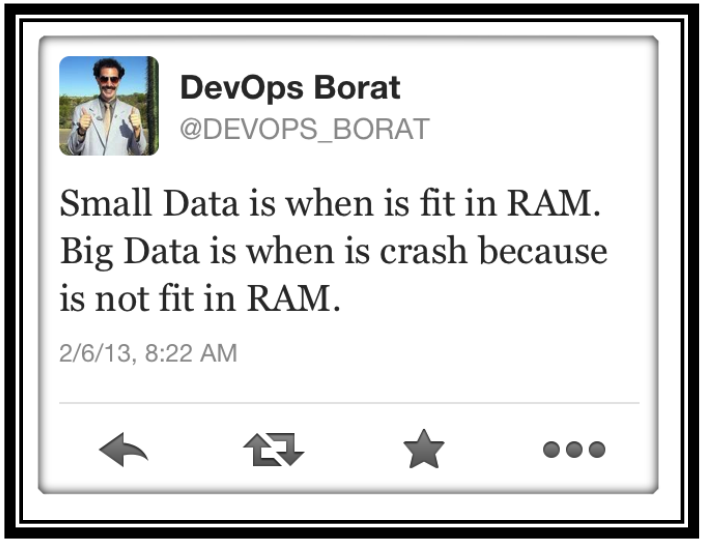
Small Data is when is fit in RAM. Big Data is when is crash because is not fit in RAM.

2/6/13, 8:22 AM




What is Big Data --

[Li01] Slides #23-43

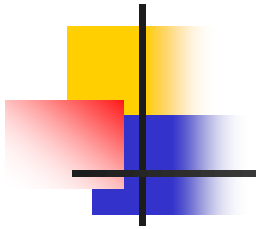


A different way to look at Big Data :

Byte	: one grain of rice	Hobbyist
Kilobyte	: cup of rice	
Megabyte	: 8 bags of rice	Desktop
Gigabyte	: 3 Semi trucks	
Terabyte	: 2 Container Ships	Internet
Petabyte	: Blankets Manhattan	
Exabyte	: Blankets west coast states	Big Data
Zettabyte	: Fills the Pacific Ocean	
Yottabyte	: A EARTH SIZE RICE BALL!	*The Future?*

Sources of Big Data Deluge

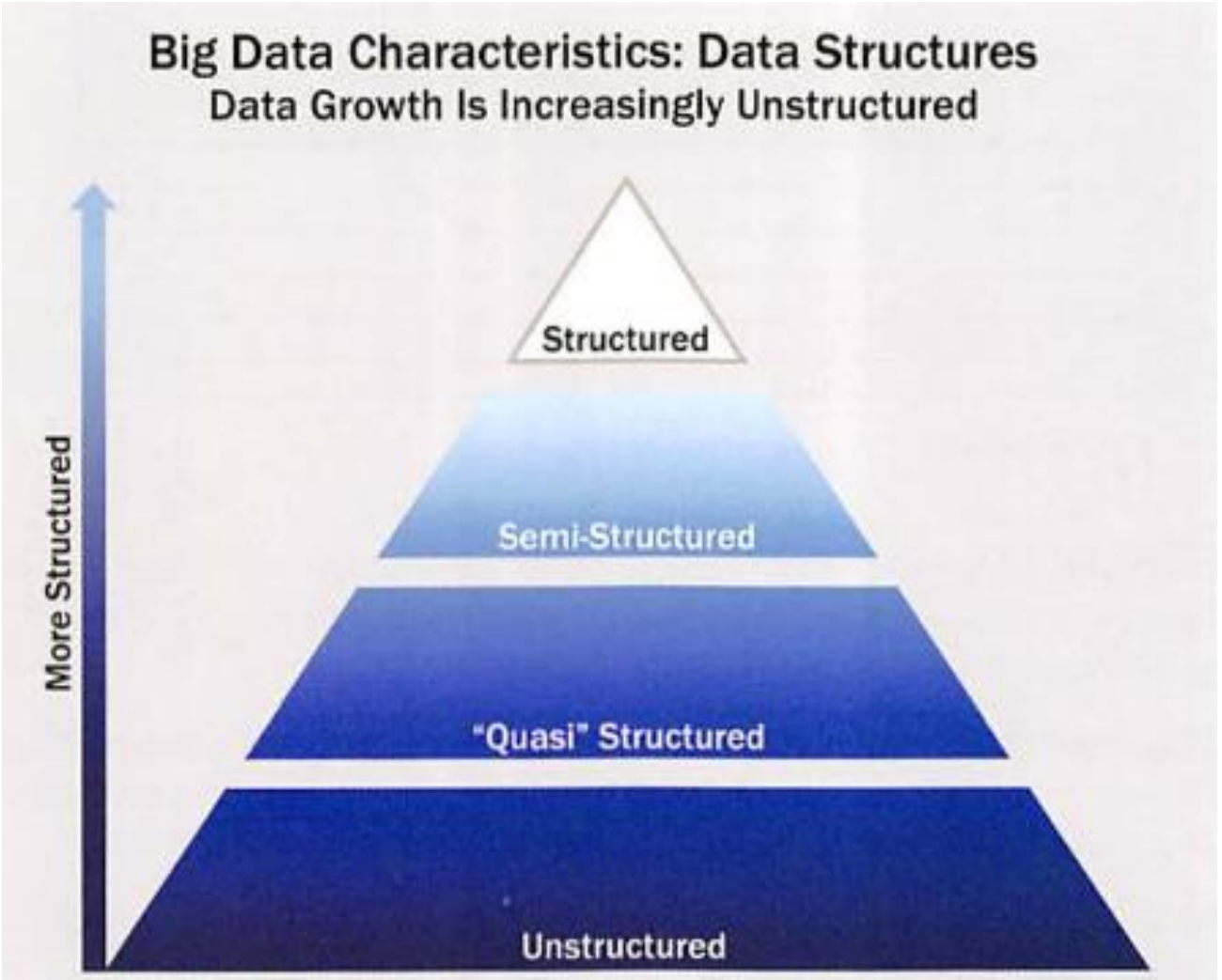
Sources of Big Data Deluge



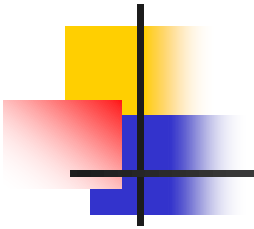
- Mobile sensors – GPS, accelerometer, …..
- Social media – personal updates every sec, ……
- Video surveillance – street cameras, stores, ……
- Video rendering – signal processing, streaming, ……
- Smart grids – intelligent controls, optimization, ……
- Geophysical exploration – oil, gas, coral, metals, ……
- Medical imaging – high-resolution images, ……
- Molecular Sequencing – genomics, transcriptomics, ……
- ……

Example: Genotyping from 23andme.com

1.1.1 Data Structures: Characteristics of Big Data



Data Structures: Characteristics of Big Data

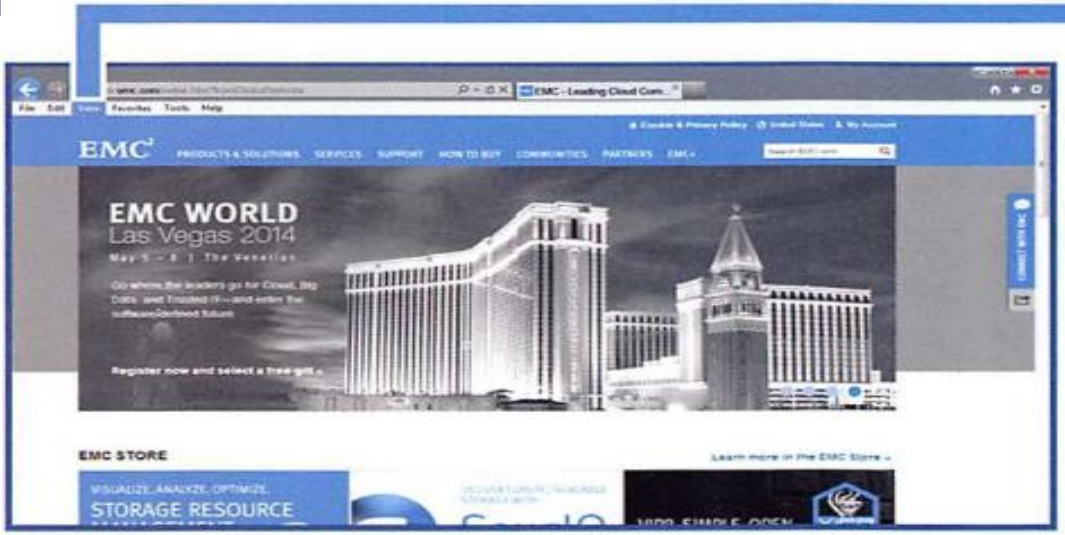


- Structured – defined data type, format, structure
  - Transactional data, OLAP cubes, RDBMS, CSV files, spreadsheets
- Semi-structured
  - Text data with discernable patterns – e.g. XML data
- Quasi-structured
  - Text data with erratic data formats – e.g. clickstream data
- Unstructured
  - Data with no inherent structure – text docs, PDF's, images, video

Example of Structured Data

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

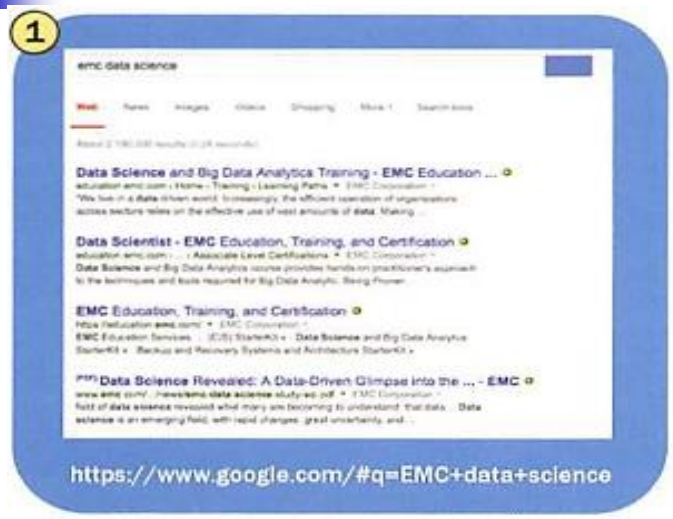
Example of Semi-Structured Data




```
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

<meta name="description" content="EMC is a leading provider of IT storage hardware solutions to promote dat
cloud computing.">
name="keywords" content="emc,network storage,data recovery,information management,backup software,nas storage

<meta name="viewport" content="width=device-width, initial-scale=1">

<link href="/_admin/css/html-layout-css-includes-combined-min.css" rel="stylesheet">
<script src="/_admin/js/jquery.js"></script>
<link rel="stylesheet" href="/R1/assets/css/common/normalize.css">
<link rel="stylesheet" href="/R1/assets/css/homepage/main.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-header.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-footer.css">


<script type="text/javascript" src="//platform.twitter.com/widgets.js"></script>
<script src="/R1/assets/js/common/modernizr-2.6.2.min.js"></script>
```

Example of Quasi-Structured Data

visiting 3 websites adds 3 URLs to user's log files

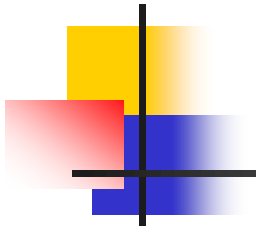


Example of Unstructured Data
Video about Antarctica Expedition

1.1.2 Types of Data Repositories from an Analyst Perspective

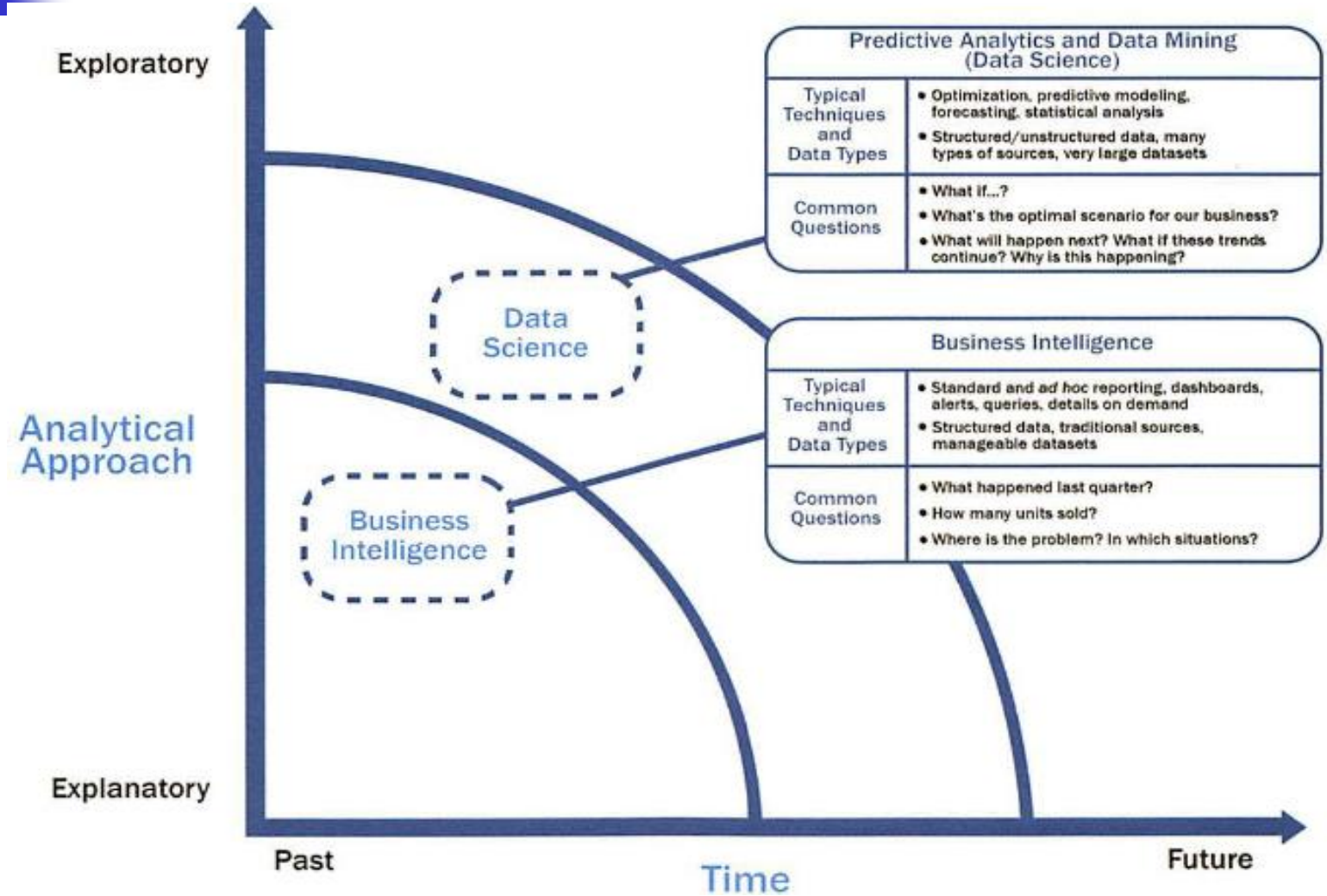
Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

1.2 State of the Practice in Big Data



- BI vs Data Science
  - BI = Business Intelligence
- Current Analytical Architecture
- Drivers of Big Data
- Emerging Big Data Ecosystem and a New Approach to Analytics

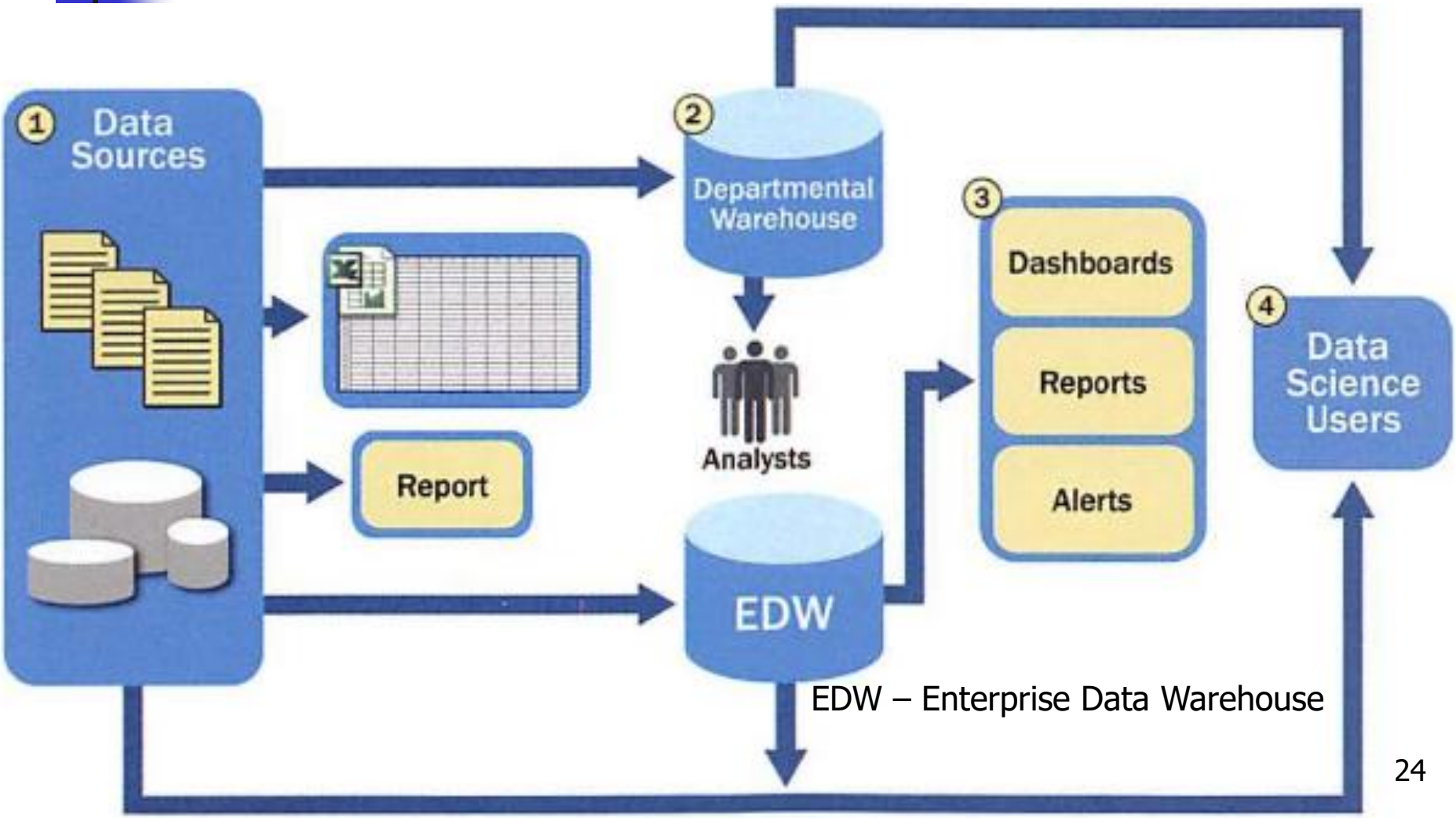
1.2.1 Business Intelligence (BI) vs. Data Science



Business Drivers for Big Data Analytics

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

1.2.2 Current Analytical Architecture



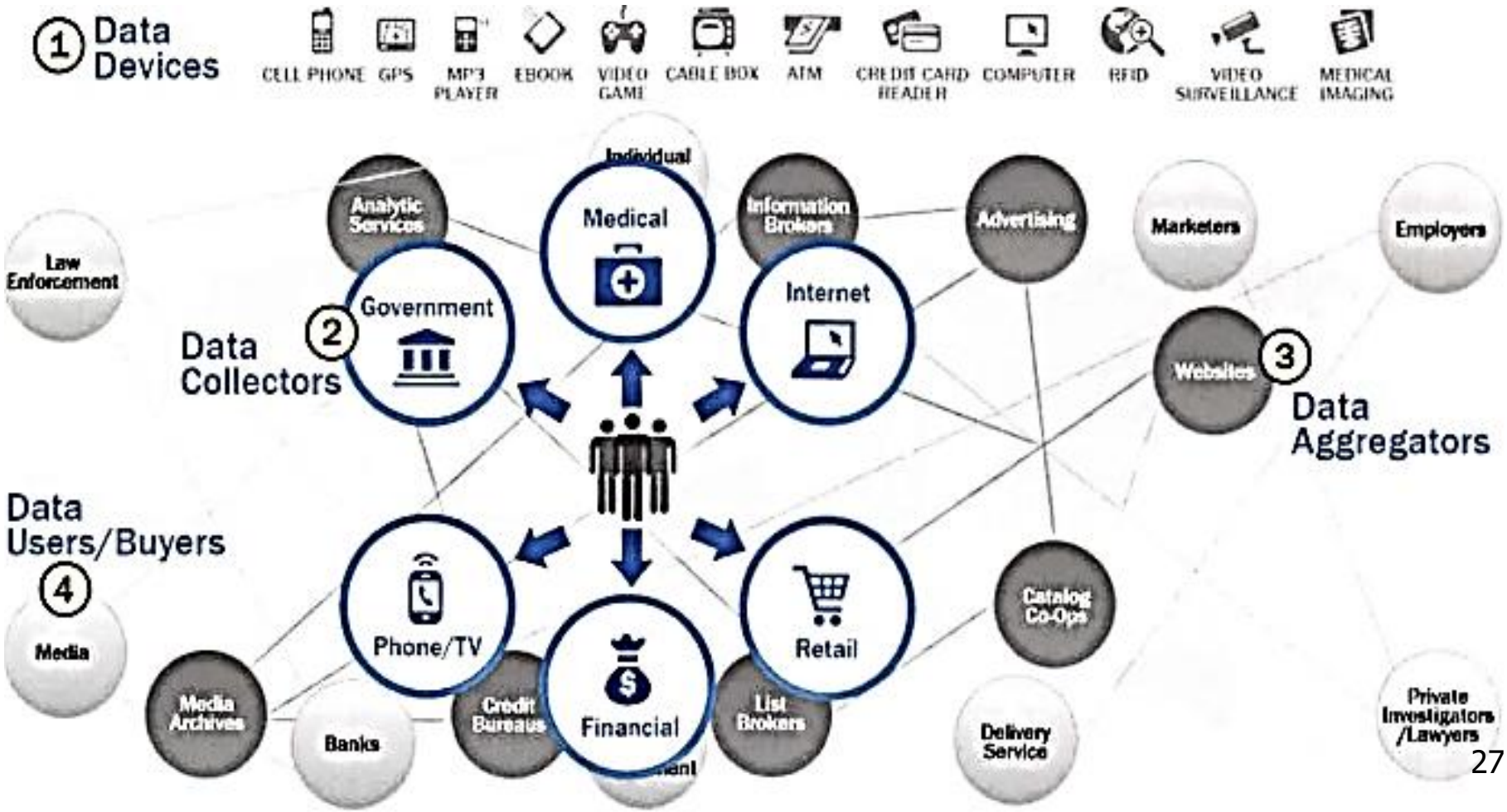
1.2.3 Drivers of Big Data

Data Evolution & Rise of Big Data Sources

1.2.4 Emerging Big Data Ecosystem

- Four main groups of players
  - Data devices
    - Smartphones, Computers, IofT, etc.
  - Data collectors
    - Telecom. Companies, Internet, Gov't, etc.
  - Data aggregators – make sense of data
    - Websites, Credit bureaus, Media archives, etc.
  - Data users and buyers
    - Banks, Law Enforcement, Marketers, Employers, etc.

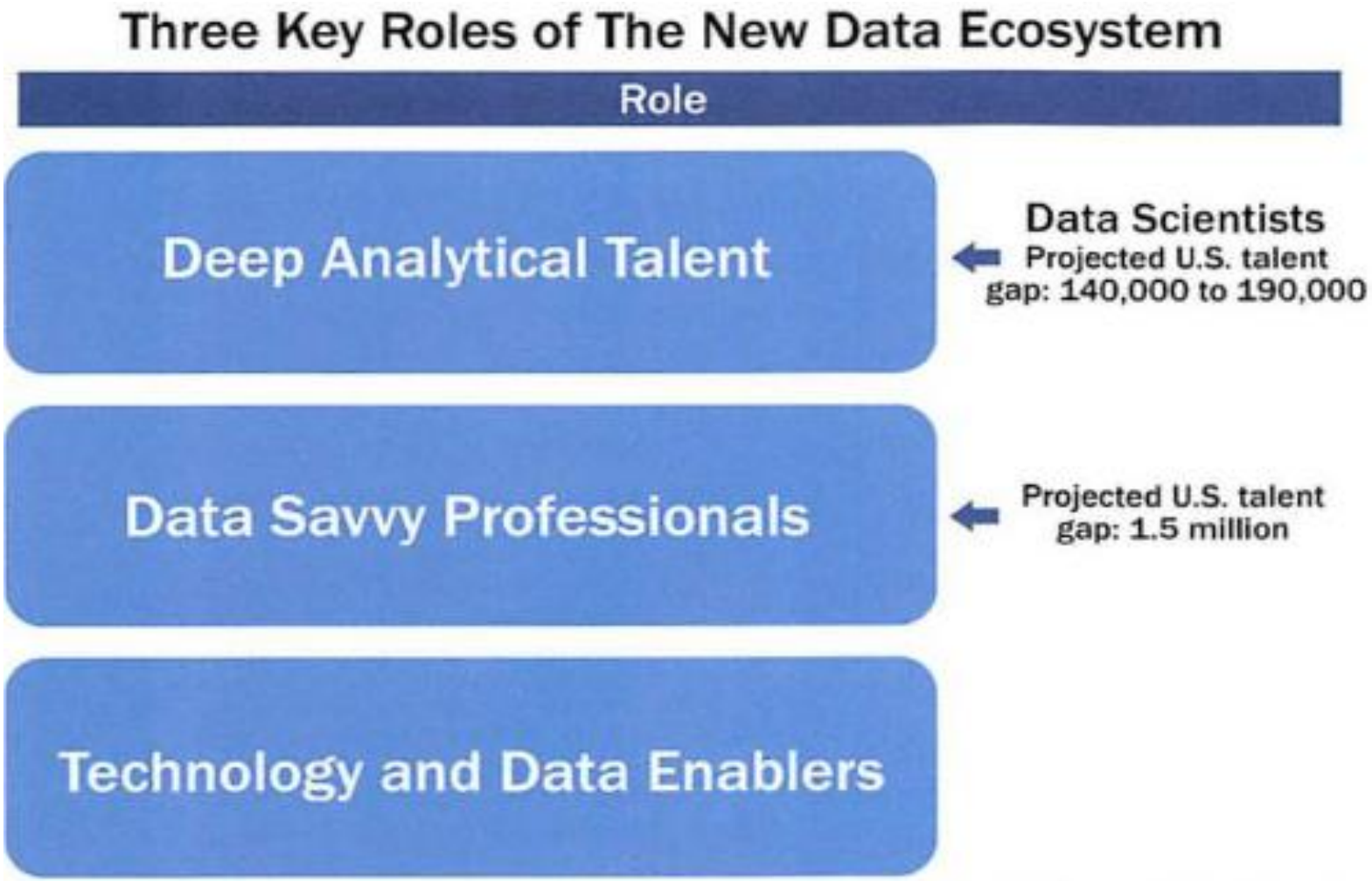
1.2.4 Emerging Big Data Ecosystem



1.3 Key Roles for the New Big Data Ecosystem

1. Deep Analytical Talent
   - Advanced training in quantitative disciplines – e.g., math, statistics, machine learning
2. Data Savvy Professionals
   - Savvy but less technical than group 1
3. Technology and data enablers
   - Technical Support
     - e.g. DB admins, programmers, etc.

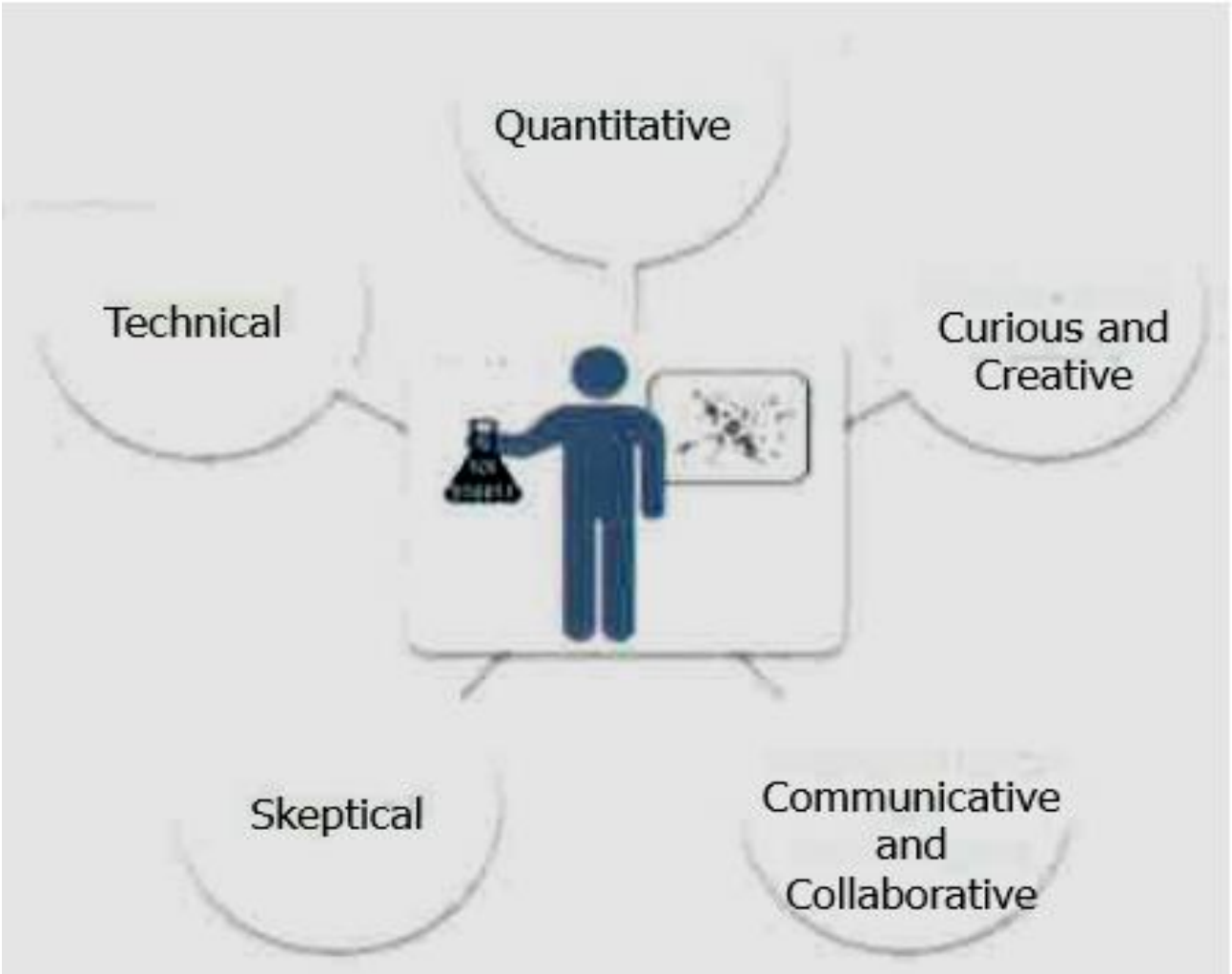
Three Key Roles of the New Big Data Ecosystem



Three Recurring Data Scientist Activities

1. Reframe business challenges as analytics challenges
2. Design, implement, and deploy statistical models and data mining techniques on big data
3. Develop insights that lead to actionable recommendations

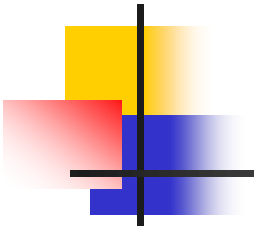
Profile of Data Scientist Five Main Sets of Skills



Profile of Data Scientist
Five Main Sets of Skills

- Quantitative skill – e.g., math, statistics
- Technical aptitude – e.g., software engineering, programming
- Skeptical mindset and critical thinking – ability to examine work critically
- Curious and creative – passionate about data and finding creative solutions
- Communicative and collaborative – can articulate ideas, can work with others

1.4 Keyword Examples of Big Data Analytics



- Retailer Targeting
  - Uses life events: marriage, divorce, pregnancy
- Apache Hadoop
  - Open source Big Data infrastructure innovation
  - MapReduce paradigm, ideal for many projects
- Social Media Company (e.g. LinkedIn)
  - Social network for working professionals
  - Can graph a user's professional network
  - 700 million users in 2020

Case Study WWII

http://geekool.pbworks.com/w/page/93674231/Imitation%20Game

Germany at the height of WWII success (September to December 1942)

http://knowyourmeme.com/memes/people/adolf-hitler
https://en.wikipedia.org/wiki/Nazi_Germany
https://en.wikipedia.org/wiki/File:Bundesarchiv_Bild_183-L05487,_Paris,_Avenue_Foch,_Siegesparade.jpg

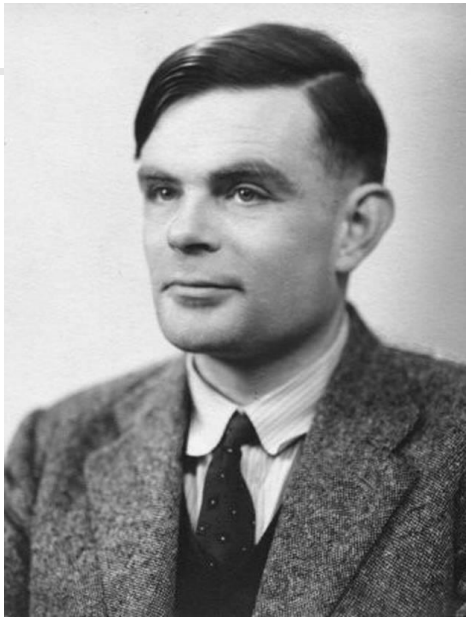
Nazi Enigma encryption machine

- An **Enigma machine** was a series of electro-mechanical rotor cipher machines developed and used in the early to early-mid twentieth century for commercial and military usage.
- Enigma was invented by the German engineer Arthur Scherbius at the end of World War I.[1] Early models were used commercially from the early 1920s, and adopted by military and government services of several countries, most notably Nazi Germany before and during World War II.[2]

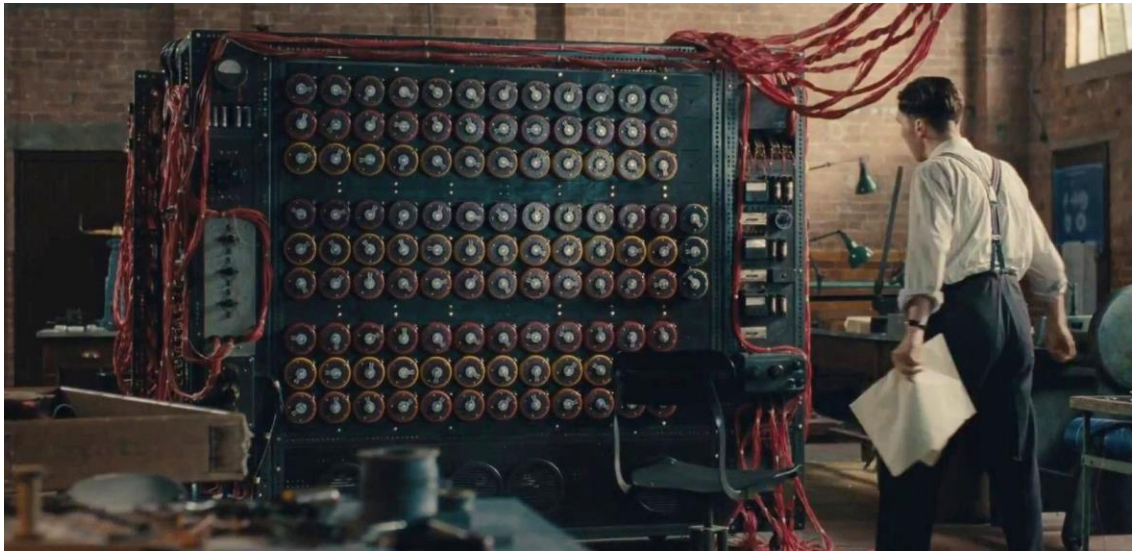
A Nazi Enigma encryption machine is displayed at the World War II Museum in Natick, Mass.

https://en.wikipedia.org/wiki/Enigma_machine#Breaking_Enigma
http://phys.org/news/2015-02-imitation-game-wwii-codebreakers-audiences.html

Alan Turing (1912-1954)



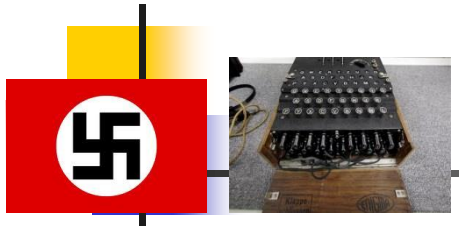
http://geekool.pbworks.com/w/page/93674231/Imitation%20Game

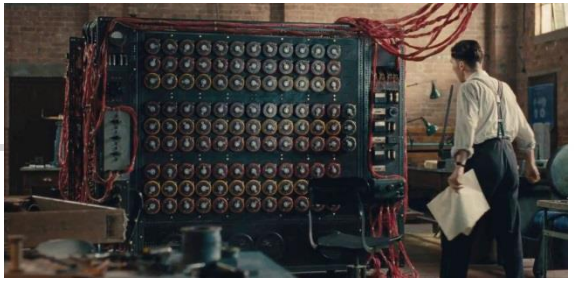


http://www.independent.co.uk/news/people/news/alan-turing-gets-royal-pardon-for-gross-indecency--61-years-after-he-poisoned-himself-9023116.html

https://vickster51corner.files.wordpress.com/2014/12/the-imitation-game-the-machine.jpg

Cryptanalysis of the Enigma (WWII)

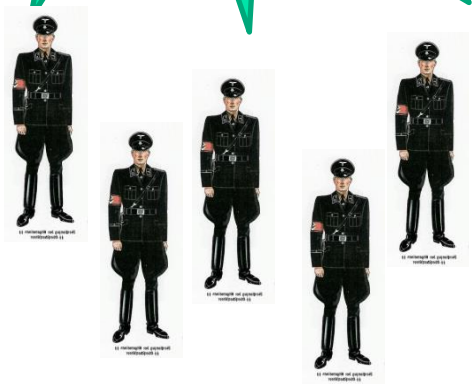




Encrypted Data

Intercept

Decrypted Data
(Military Intelligence)



Normandy landings

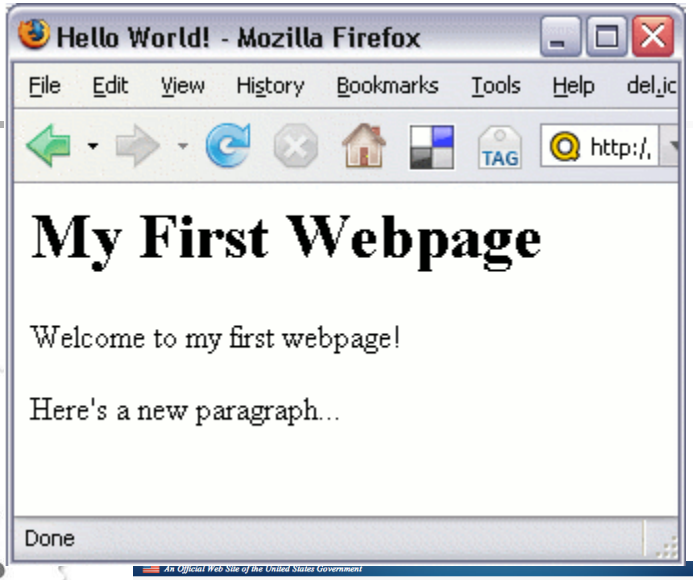
Sinks of U-Boats

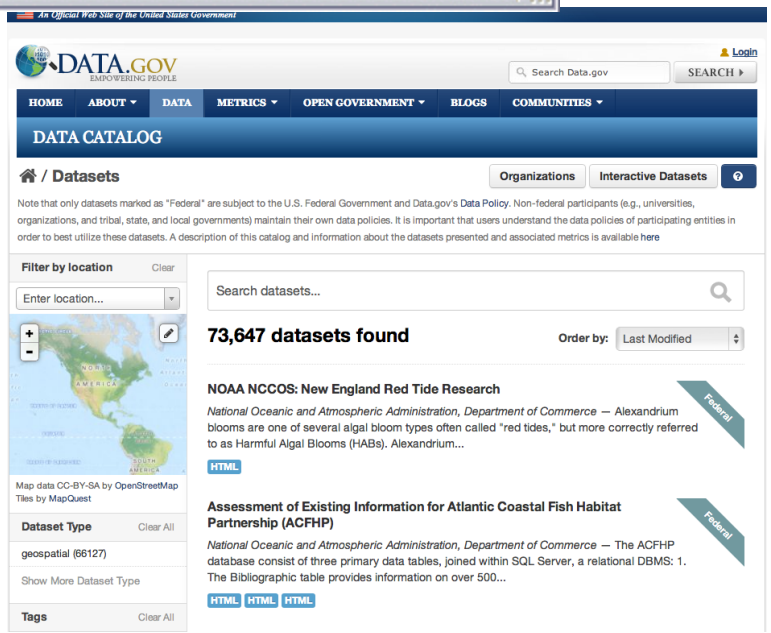
https://en.wikipedia.org/wiki/Cryptanalysis_of_the_Enigma
https://information2share.wordpress.com/2011/08/02/freedom-shall-prevail/
http://www.uboatarchive.net/208592.htm https://en.wikipedia.org/wiki/Normandy_landings

Case Study

http://www.telegraph.co.uk/technology/internet/8204092/Data-mapping-visualisations-the-best-on-the-web.html

Web Data





http://www.ipfw.edu/departments/dcs/depts/corporate/social-media-professional-certificate.html
https://opendatahk.com/2015/03/data-gov-hk-site/
http://ckan.org/2013/05/23/data-gov-relaunch-on-ckan/
http://www.quackit.com/make-your-own-website/

Internet Archive

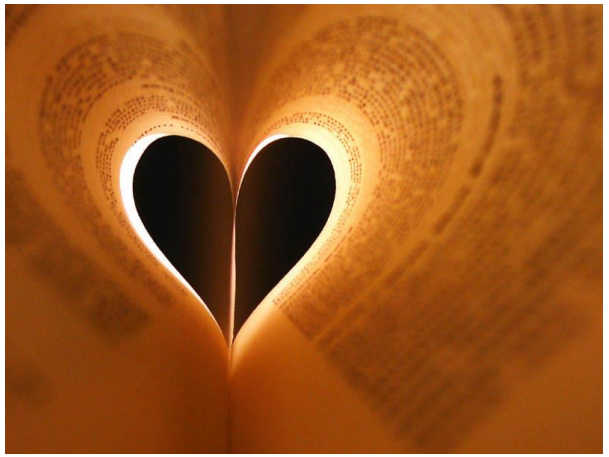
- A non-profit digital library with the stated mission of "universal access to all knowledge".[2][3]
- As of October 2012, its collection topped 10 petabytes
- https://archive.org/

DEMO

https://en.wikipedia.org/wiki/Internet_Archive

Use of Web Data (Positive)

- Government Planning
- Academic Studies
- Market Research
- Business Analysis
- Leisure Reading
- Mass Media
- Information Freedom and Exchange
- …



https://hovelingi.wordpress.com/2011/01/06/its-a-brand-new-year-brand-new-books-i-love-reading/

Use of Web Data (Negative)

http://socialjusticeland.tumblr.com/
http://www.hkgolden.com
http://thelowell.org/

http://www.sloansportsconference.com/?p=8387

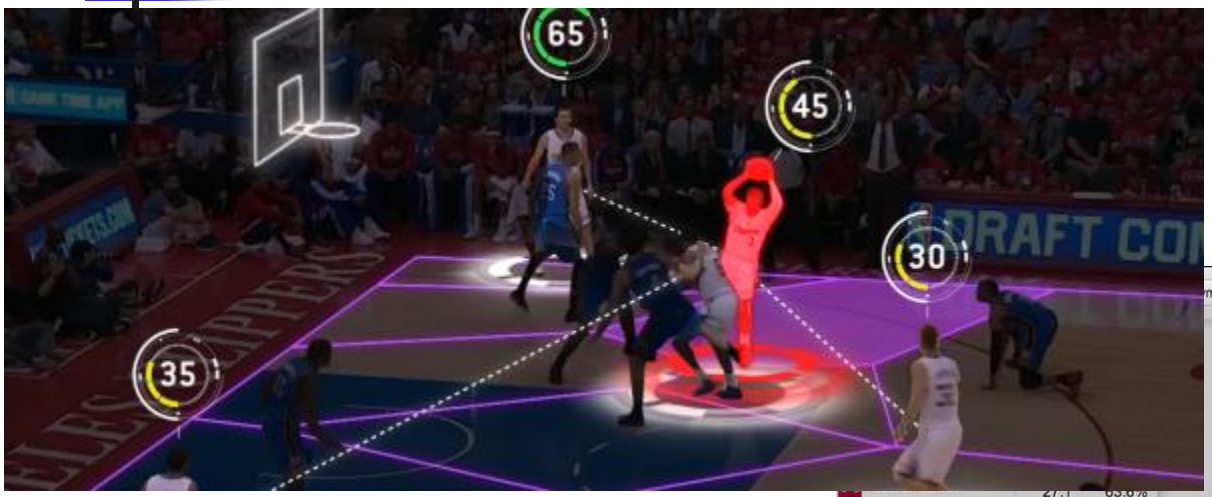
Volleyball

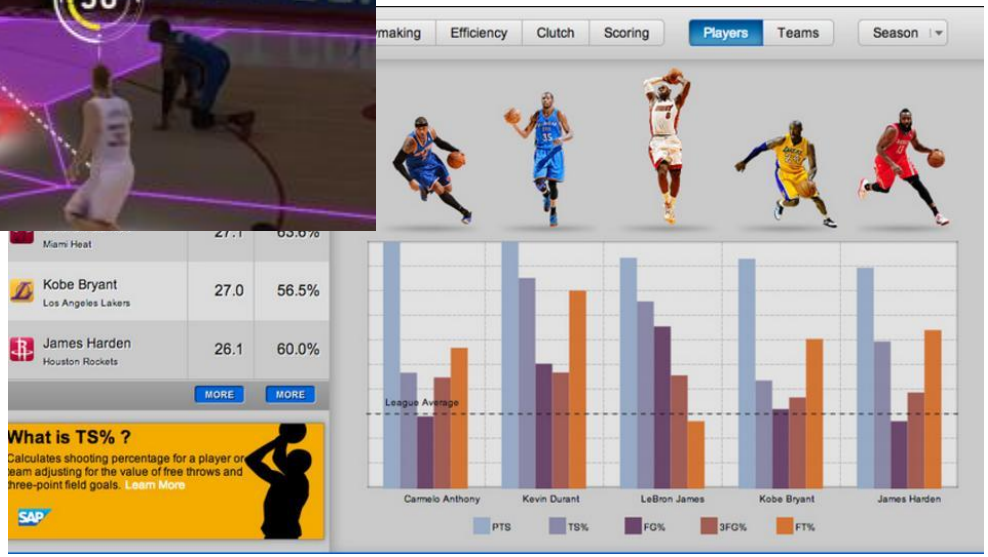
http://www.dataproject.com/Products/EN/en/Volleyball/DataVolley

Soccer

http://www.movdata.net/soccer-stats.html
http://www.trumedianetworks.com/soccer-analytics/

Basketball

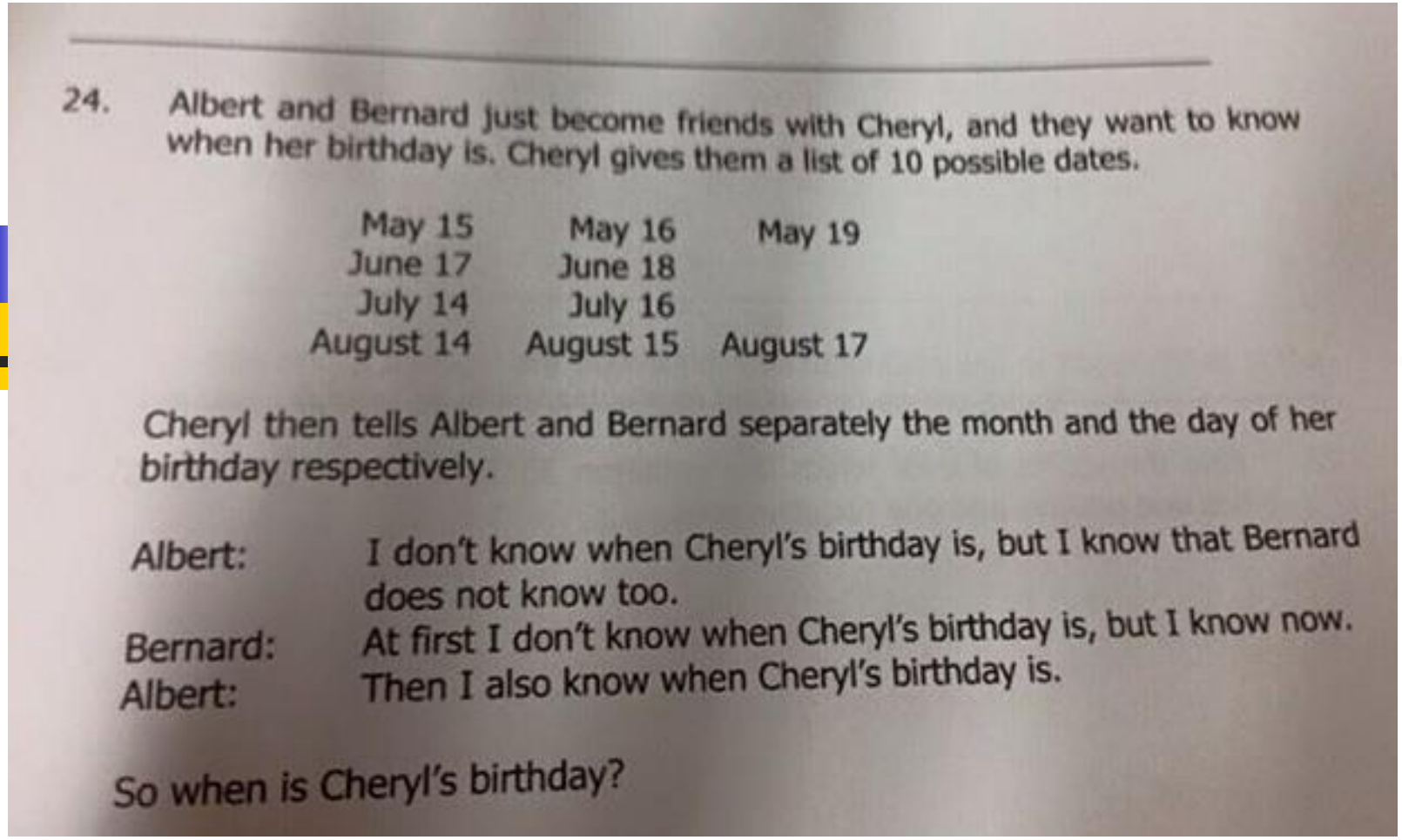




http://dataconomy.com/next-gen-sports-analytics-rolls-out-as-second-spectrums-proprietary-offering-debuted-at-2014-nba-playoff/
http://blogswithballs.com/2013/02/3858/

Case Study

School Question from Singapore



24. Albert and Bernard just become friends with Cheryl, and they want to know when her birthday is. Cheryl gives them a list of 10 possible dates.

May 15	May 16	May 19
June 17	June 18	
July 14	July 16	
August 14	August 15	August 17

Cheryl then tells Albert and Bernard separately the month and the day of her birthday respectively.

Albert: I don't know when Cheryl's birthday is, but I know that Bernard does not know too.

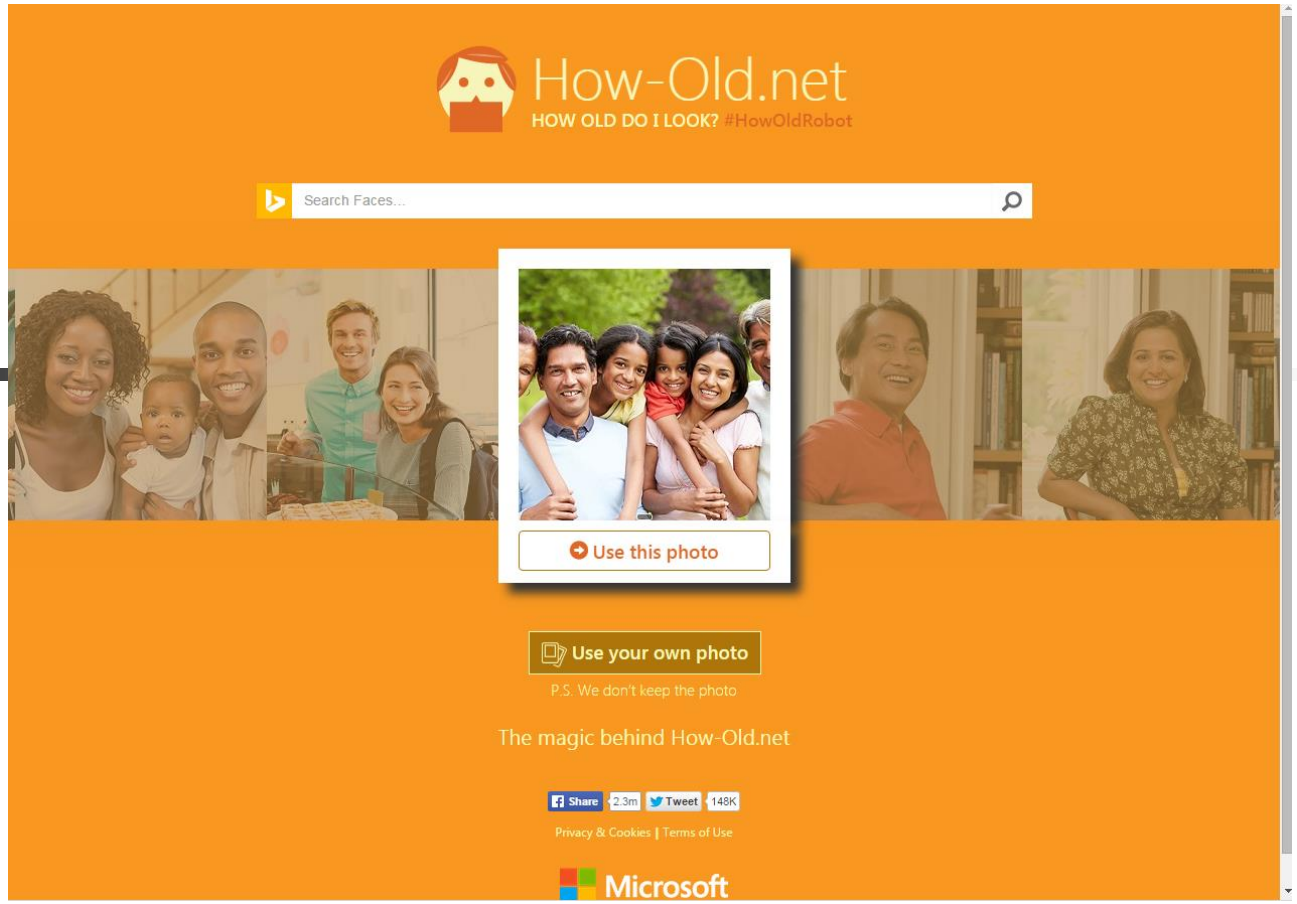
Bernard: At first I don't know when Cheryl's birthday is, but I know now.

Albert: Then I also know when Cheryl's birthday is.

So when is Cheryl's birthday?

http://www.theguardian.com/science/alexs-adventures-in-numberland/2015/apr/15/why-the-cheryl-birthday-problem-turned-into-the-maths-version-of-thatdress?CMP=fb_gu

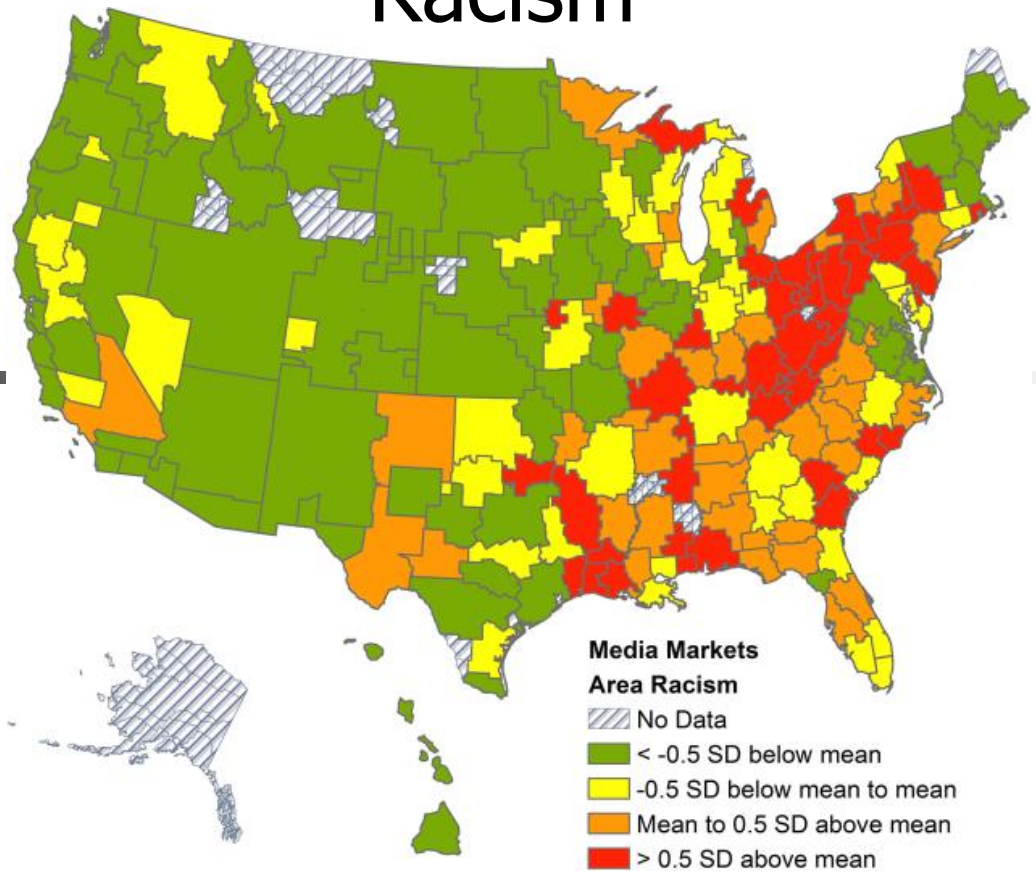
Case Study
Applied Machine Learning from Data



https://www.reddit.com/r/microsoft/comments/nivlou/what_happened_to_howoldnet/

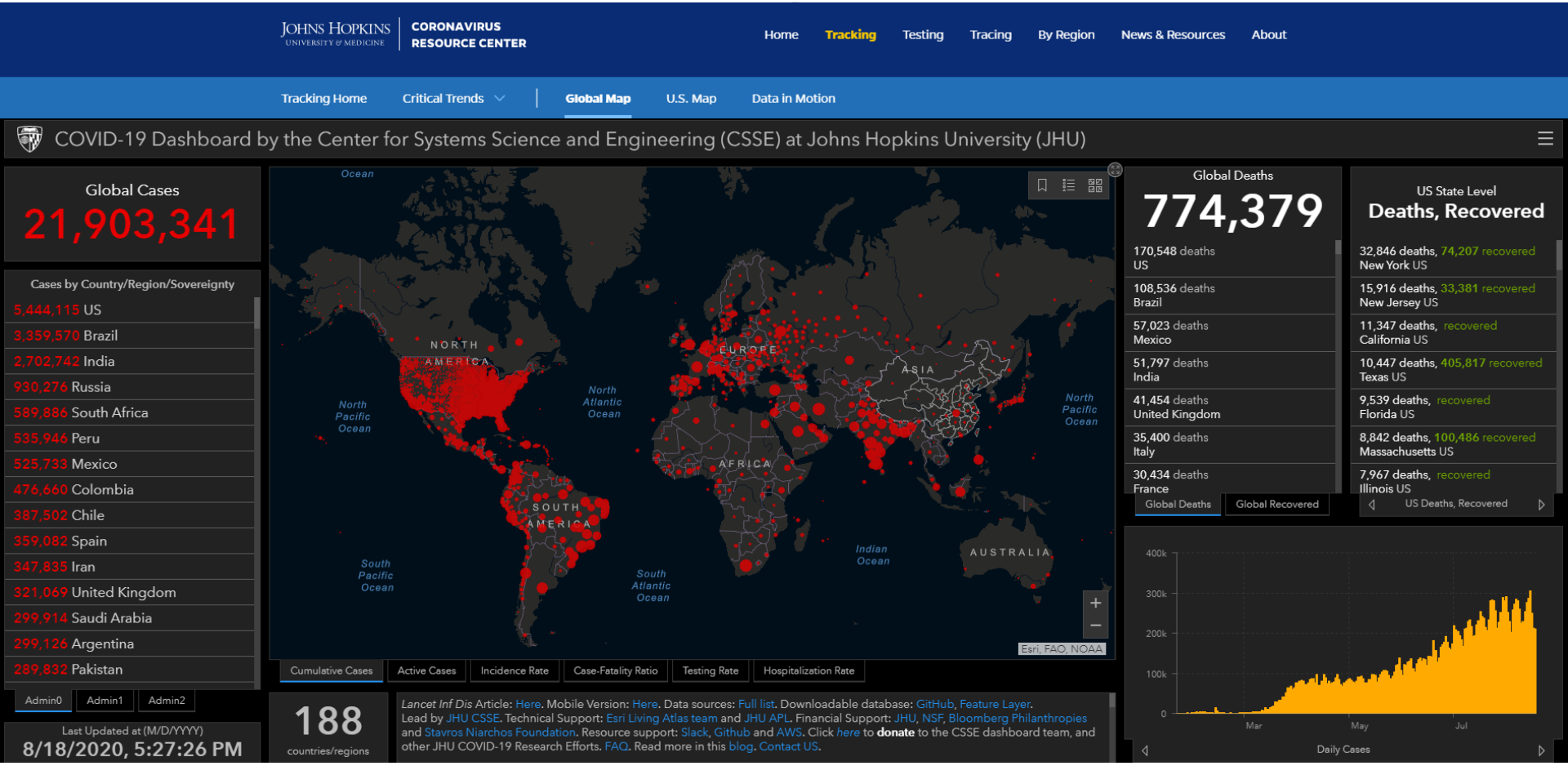
Case Study

Racism



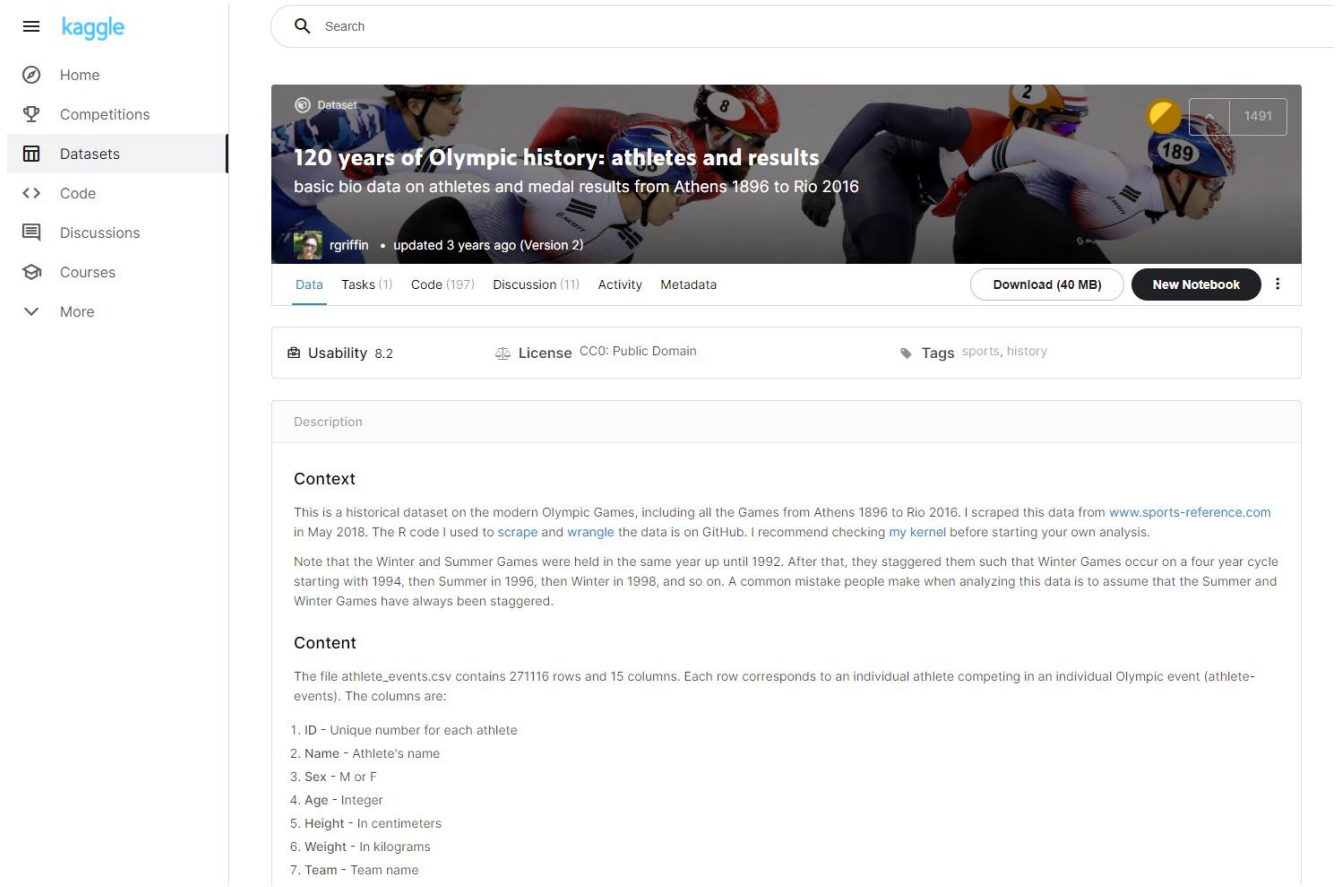
http://gizmodo.com/use-google-searches-to-figure-out-how-racist-your-neigh-1709200937
https://doi.org/10.1371/journal.pone.0122963

Case Study



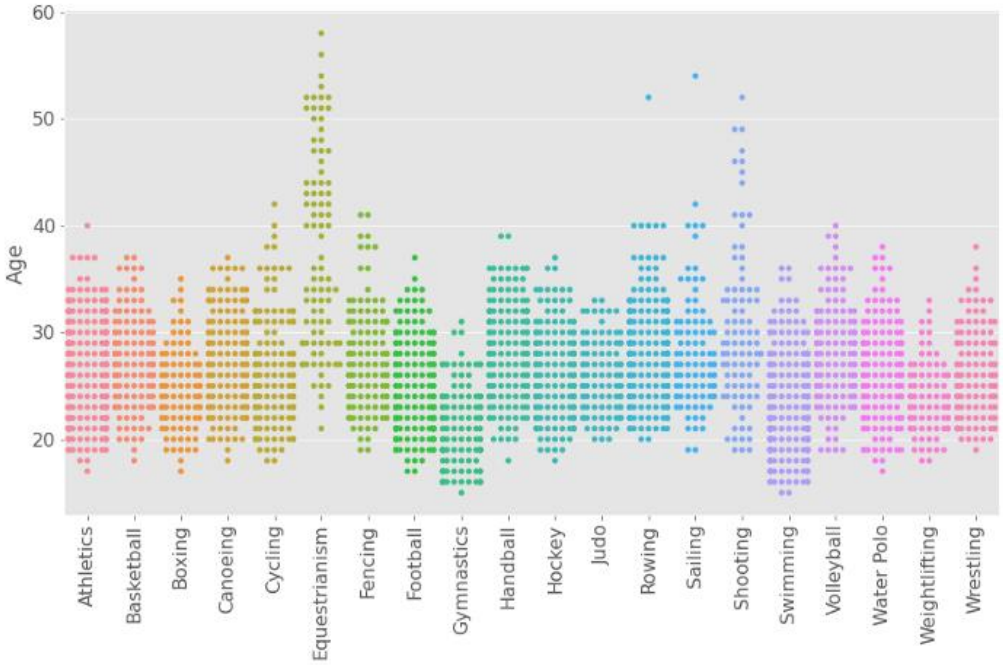
https://coronavirus.jhu.edu/map.html

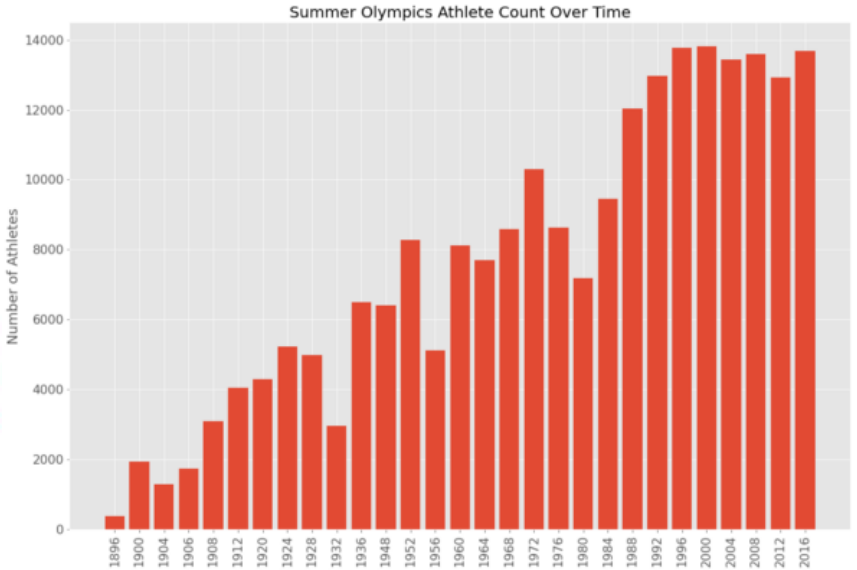
Case Study



https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

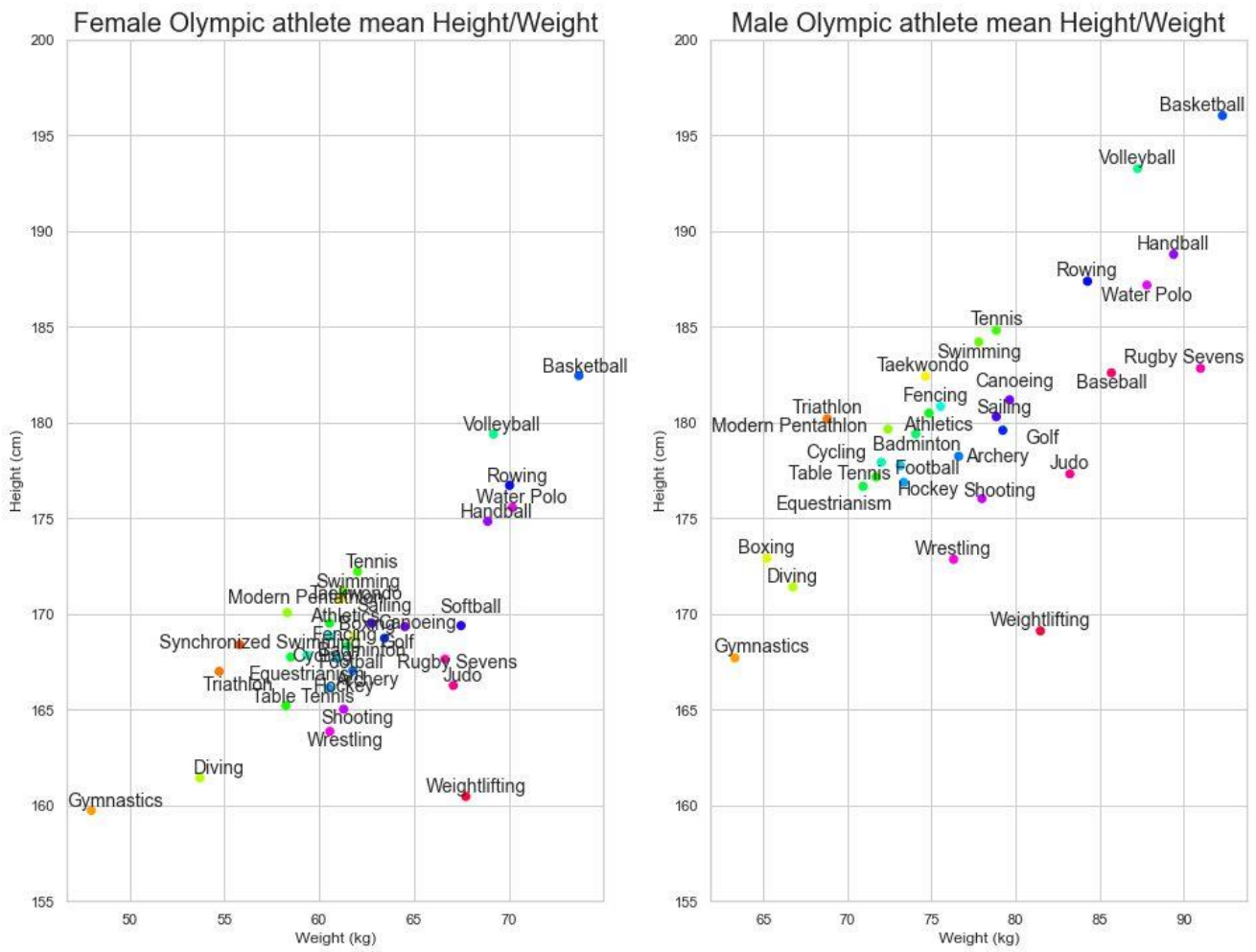
Case Study





https://towardsdatascience.com/studying-up-for-the-tokyo-2021-olympics-with-sql-719a0ae3779b

Case Study



https://www.actuaries.digital/2021/07/30/the-olympics-by-numbers-for-people-who-love-data-and-sports-but-mainly-data/

Case Study

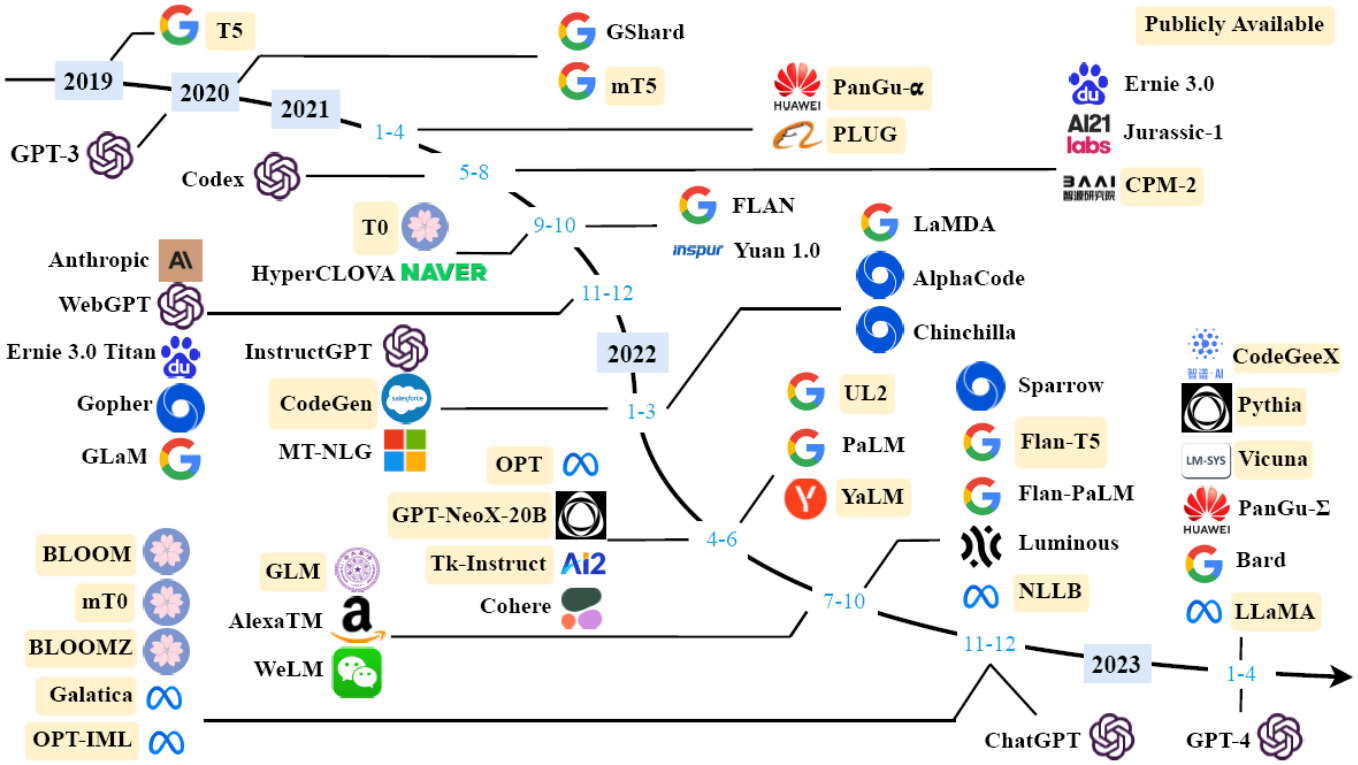
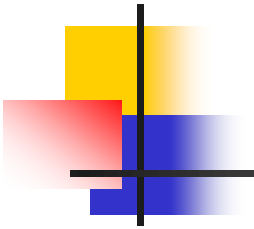


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.*, the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

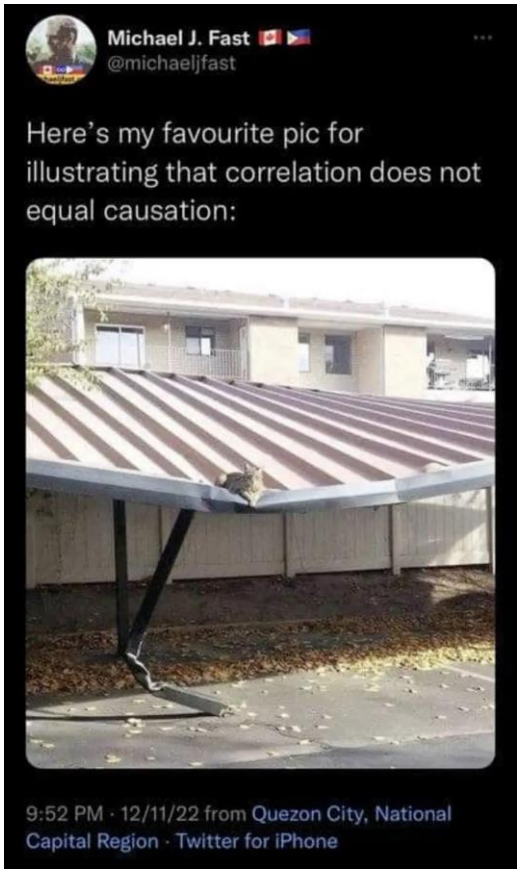
https://arxiv.org/pdf/2303.18223.pdf

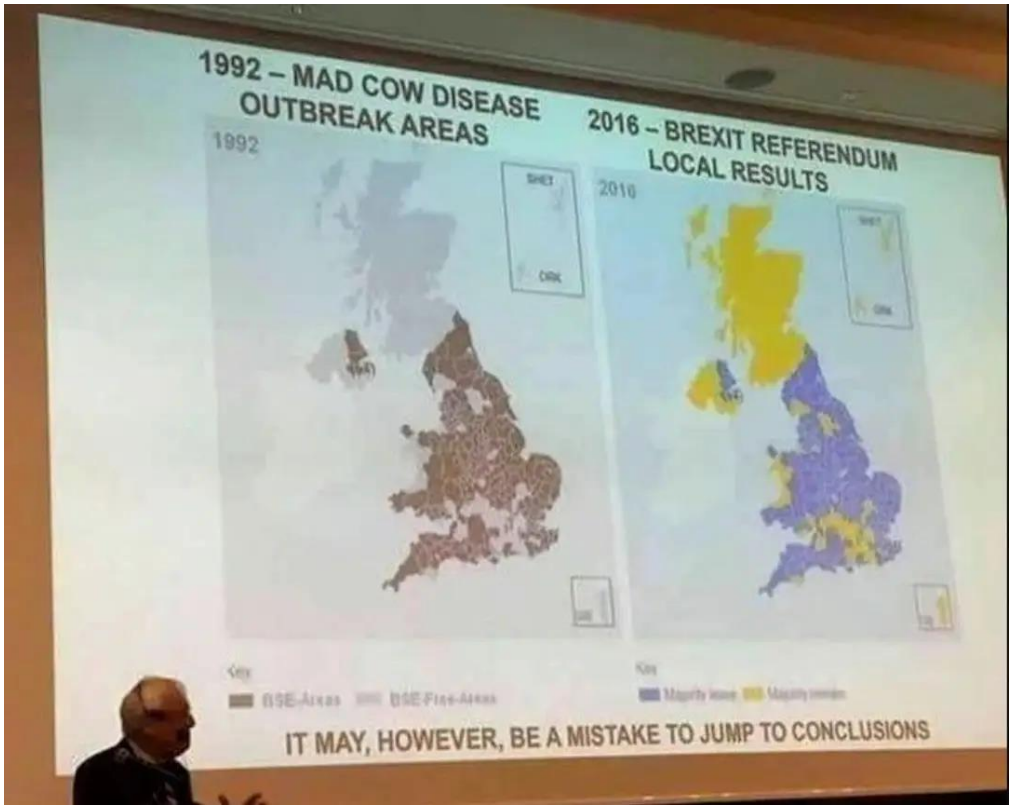


Summary

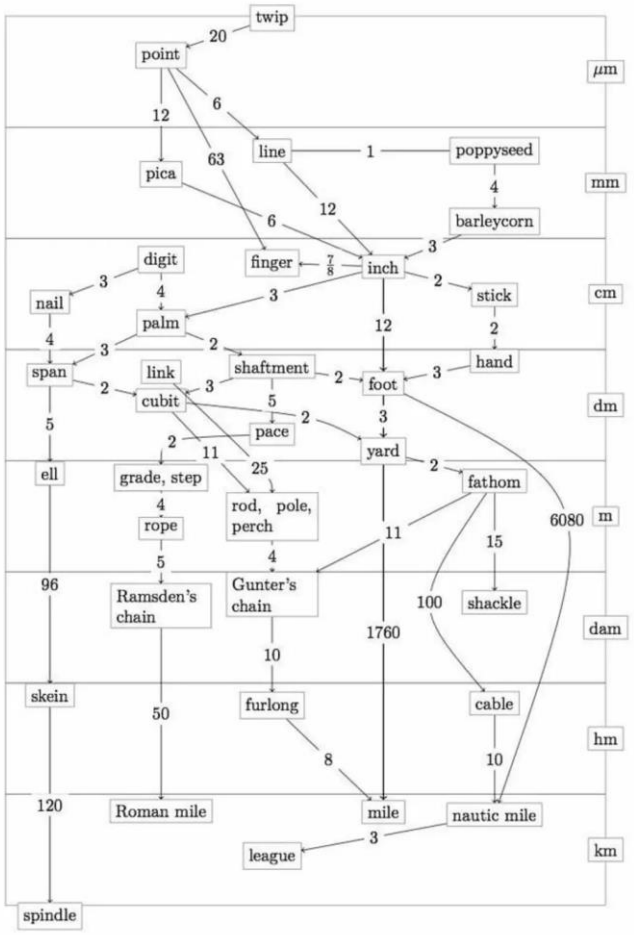
- Big Data comes from myriad sources
  - Social media, sensors, IoT, video surveillance, and other new and emerging technologies.
- People are finding creative and novel data computing
- Exploiting Big Data opportunities requires
  - Novel data processing architectures
  - Mature machine learning algorithms
  - People with open-minded skill sets

Remarks


Michael J. Fast
@michaeljfast
Here's my favourite pic for illustrating that correlation does not equal causation:
9:52 PM · 12/11/22 from Quezon City, National Capital Region · Twitter for iPhone


1992 – MAD COW DISEASE OUTBREAK AREAS
1992
2016 – BREXIT REFERENDUM LOCAL RESULTS
2016
IT MAY, HOWEVER, BE A MISTAKE TO JUMP TO CONCLUSIONS

Remarks



Data Units matter Feature Scales.

https://commons.wikimedia.org/wiki/File:English_length_units_graph.png

Answer

School Question from Singapore Kids

1. Albert knows that Bernard doesn't know. (Maybe Cheryl told him as much).
2. Albert deduces Bernard can't have a unique date such as 18 or 19.
3. Albert, smugly taunts Bernard, announcing Bernard doesn't know. This is the first statement of the problem.
4. Bernard realises what Albert has realised, which is that Bernard does not have 18 or 19. Now if Albert was holding June he would know the answer, because there is only one remaining date in June, namely June 17. So Bernard deduces it is not June.
5. Bernard announces he knows the answer. This is the second statement of the problem.
6. If Bernard is so confident, he must have a unique date. We know it's not 18 or 19. What other unique date can it be? There are two 14s, two 15s, two 16s and two 17s - but Bernard has eliminated June 17 - leaving him with August 17 only. That's how he worked it out.
7. Albert is furious Bernard beat him to the answer. Albert puts himself in Bernard's shoes, running through the six steps above. Finally Albert reaches the same conclusion we have, Bernard must have 17. Albert announces he knows the answer too.

So August 17 is a valid answer.

http://www.theguardian.com/science/alexs-adventures-in-numberland/2015/apr/15/why-the-cheryl-birthday-problem-turned-into-the-maths-version-of-thatdress?CMP=fb_gu