

Introduction to Data Science

Gerard de Melo

<http://gerard.demelo.org>

Rutgers University



Instructor



Gerard de Melo

Instructor



Gerard de Melo

Doctoral Degree at
Max Planck Institute for Informatics,
Germany

Instructor



max planck institut
informatik

**Computer Science branch
of Max Planck Society, which
has had 33 Nobel Prize winners**

Image: weinbrenner.single.arabzadeh. Architektenwerkgemeinschaft
<http://www.wsa-nt.de>

Instructor



Gerard de Melo

Doctoral Degree at
Max Planck Institute for Informatics,
Germany

Post-Doctoral Researcher at
ICSI/UC Berkeley

Instructor



Instructor



Instructor



Instructor



Gerard de Melo

Doctoral Degree at
Max Planck Institute for Informatics,
Germany

Post-Doctoral Researcher at
ICSI/UC Berkeley

Assistant Professor at
IIS, Tsinghua University (2013–2016)

Instructor



Image: https://commons.wikimedia.org/wiki/File:Rutgers_building_on_stilts_Livingston_campus.JPG

Instructor



Gerard de Melo

Doctoral Degree at
Max Planck Institute for Informatics,
Germany

Post-Doctoral Researcher at
ICSI/UC Berkeley

Assistant Professor at
IIS, Tsinghua University (2013–2016)

Assistant Professor at
Rutgers University (since 2017-01)

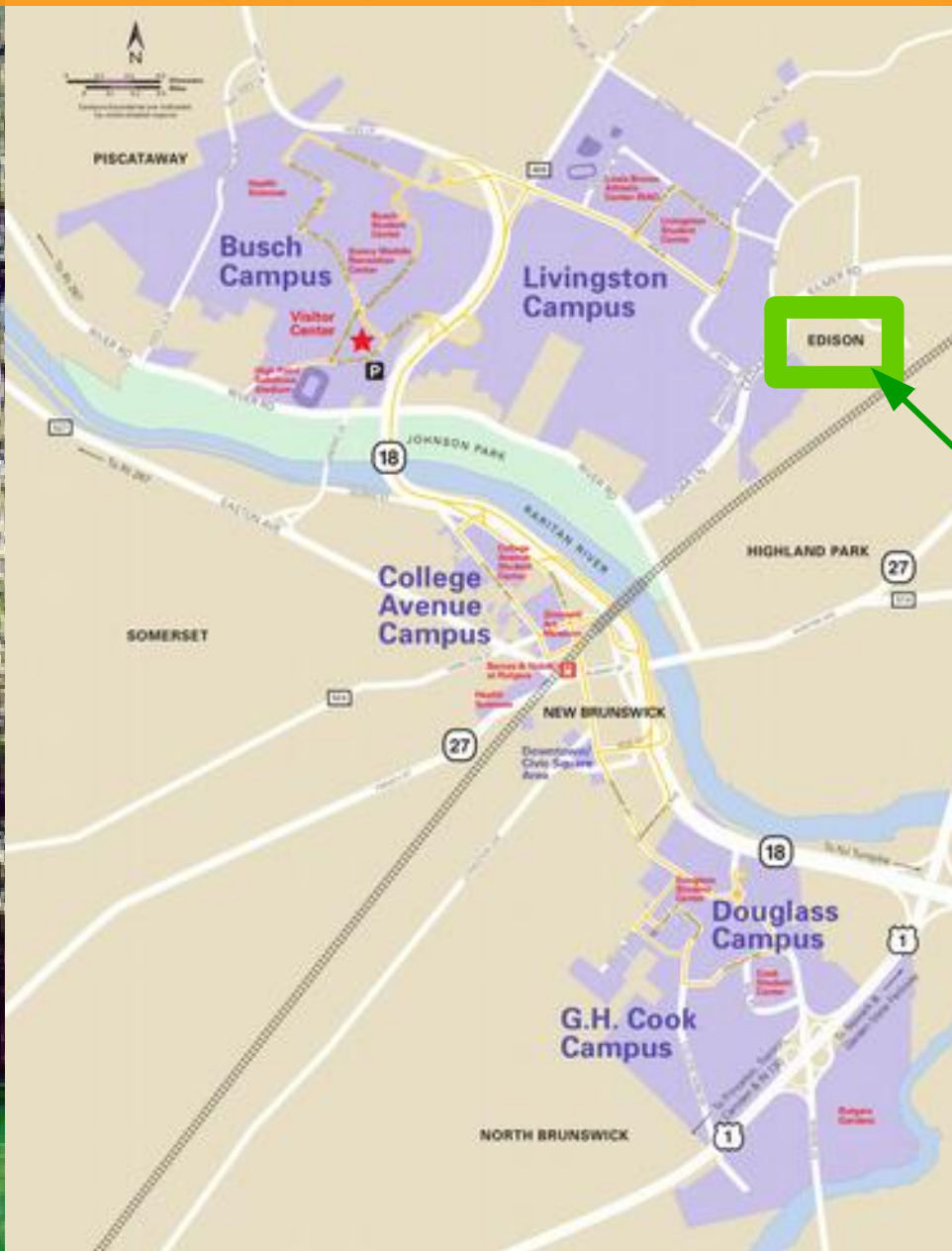
Rutgers University

**Founded in 1766
(older than the United States)**

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY



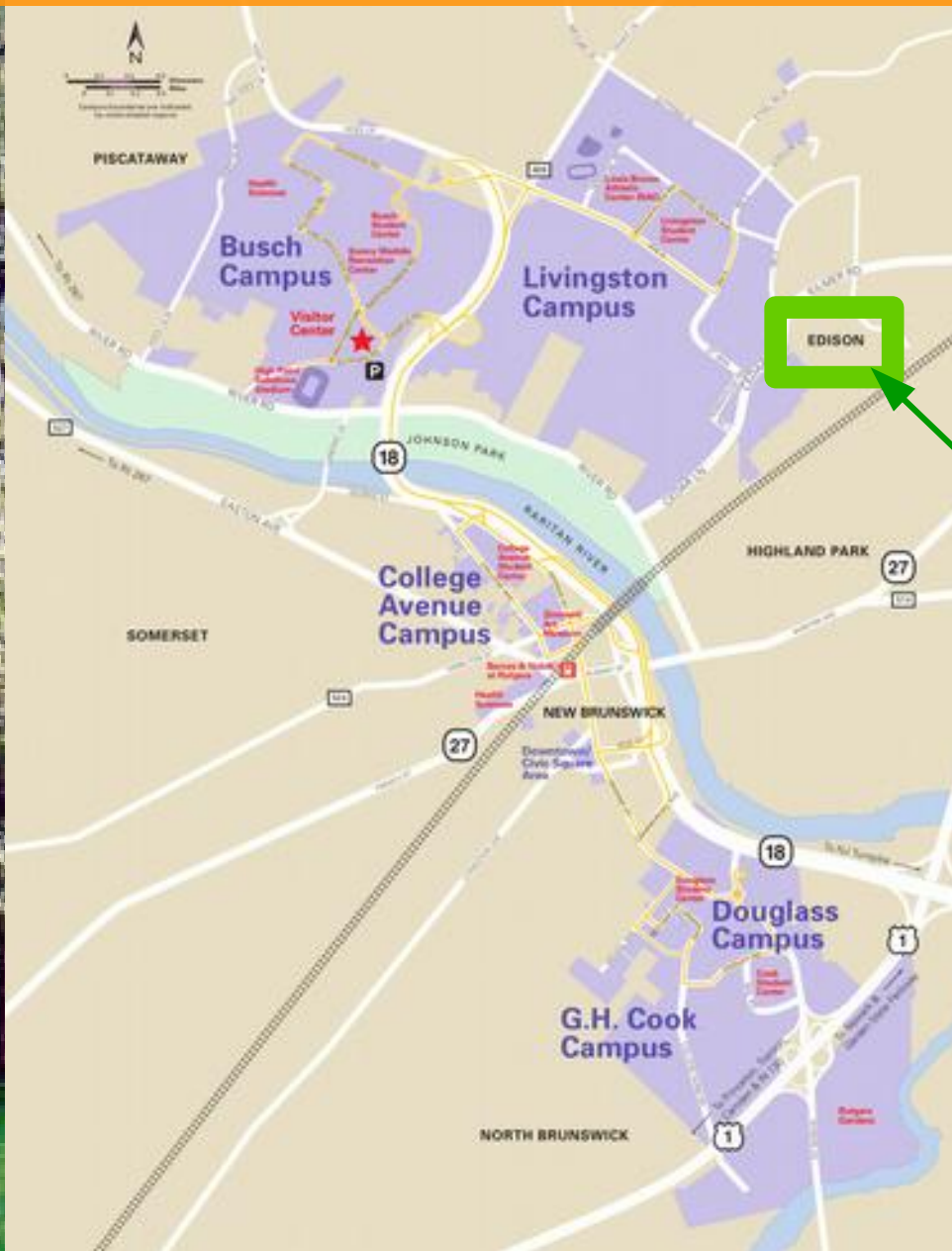
Rutgers University



RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Named after Thomas Edison,
who in this area invented
the lightbulb, sound recording,
etc.

Rutgers University



RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

**Named after Thomas Edison,
who in this area invented
the lightbulb, sound recording,
etc.**

**Later: Bell Labs
(transistor, information theory,
Unix, C/C++, digital photography),
→ AT&T**

Rutgers University

Top 30 on csrankings.org

Strengths:

- Theoretical CS
(many Gödel Prizes)
- AI, e.g. Computer Vision
- Information Science (Top 10)
- RUCCS, Top 3 in Philosophy
- Etc.



Logistics



Course Home Page

Rutgers 16:198:439

INTRODUCTION TO DATA SCIENCE

INSTRUCTOR

Gerard de Melo

CBIM 8, Dept. of Computer Science

TEACHING ASSISTANTS

TIME AND LOCATION

Linked from Sakai

<http://gerard.demelo.org/teaching/datascience/>

Teaching Assistants



Abu Shoeb

`as2352@scarletmail.rutgers.edu`

Office hours: Thu 4:30–6 PM in CoRE 246



Shahab Raji

`shahab.raji@cs.rutgers.edu`

Office hours: Tue 11AM-1PM in CBIM Cubicle H

Getting in Touch

Option 1

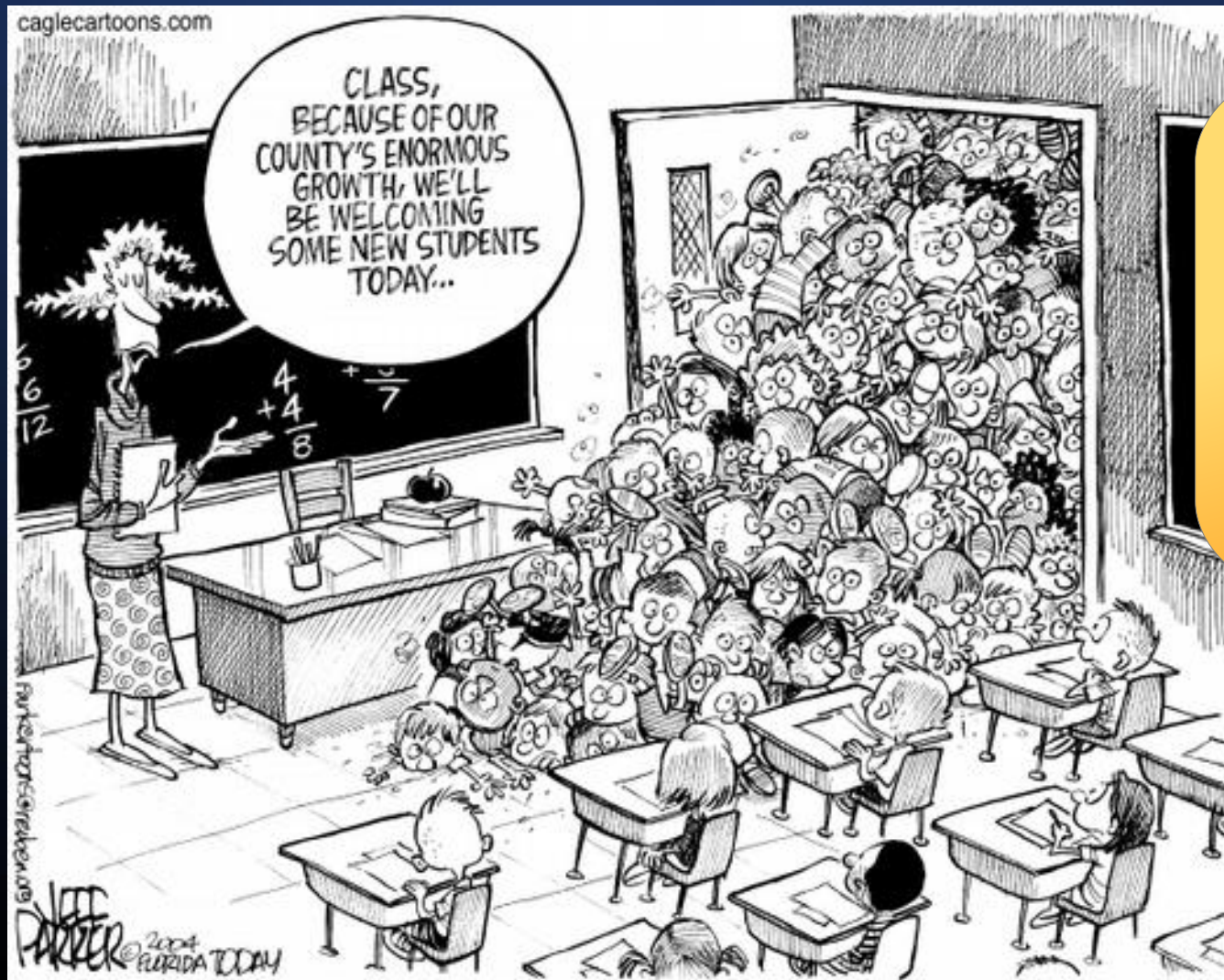
Email to TA with “[CS439]” in subject

Option 2

Office: CBIM 08

Office Hours: Wednesdays, 6-7PM
(but first check course home page for announcements)

Enrollment



**Course is in very high demand
100+ already enrolled
many SP requests**

Enrollment

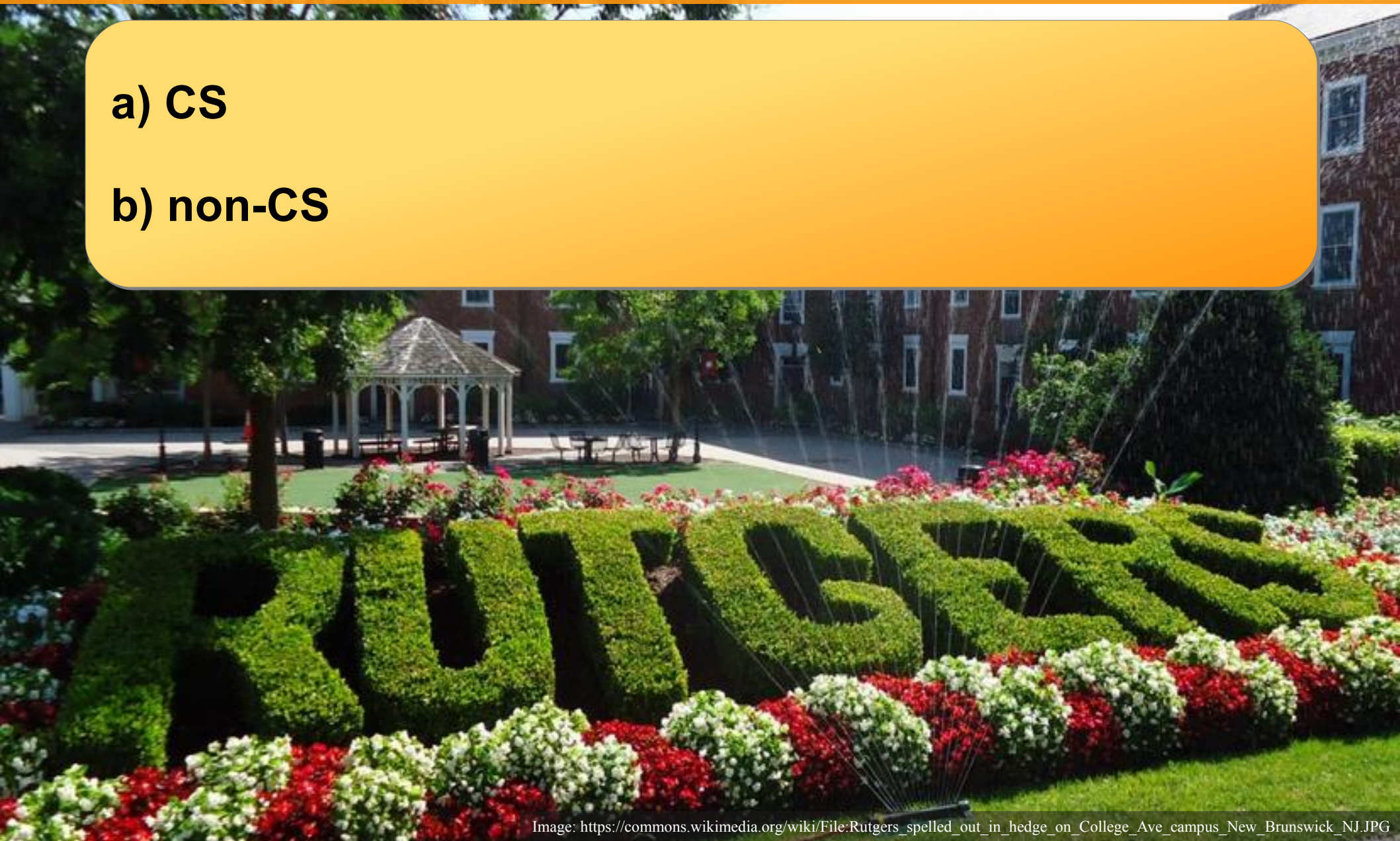


Data 8 at Berkeley
(Fall 2018, Day 1)

What about you?

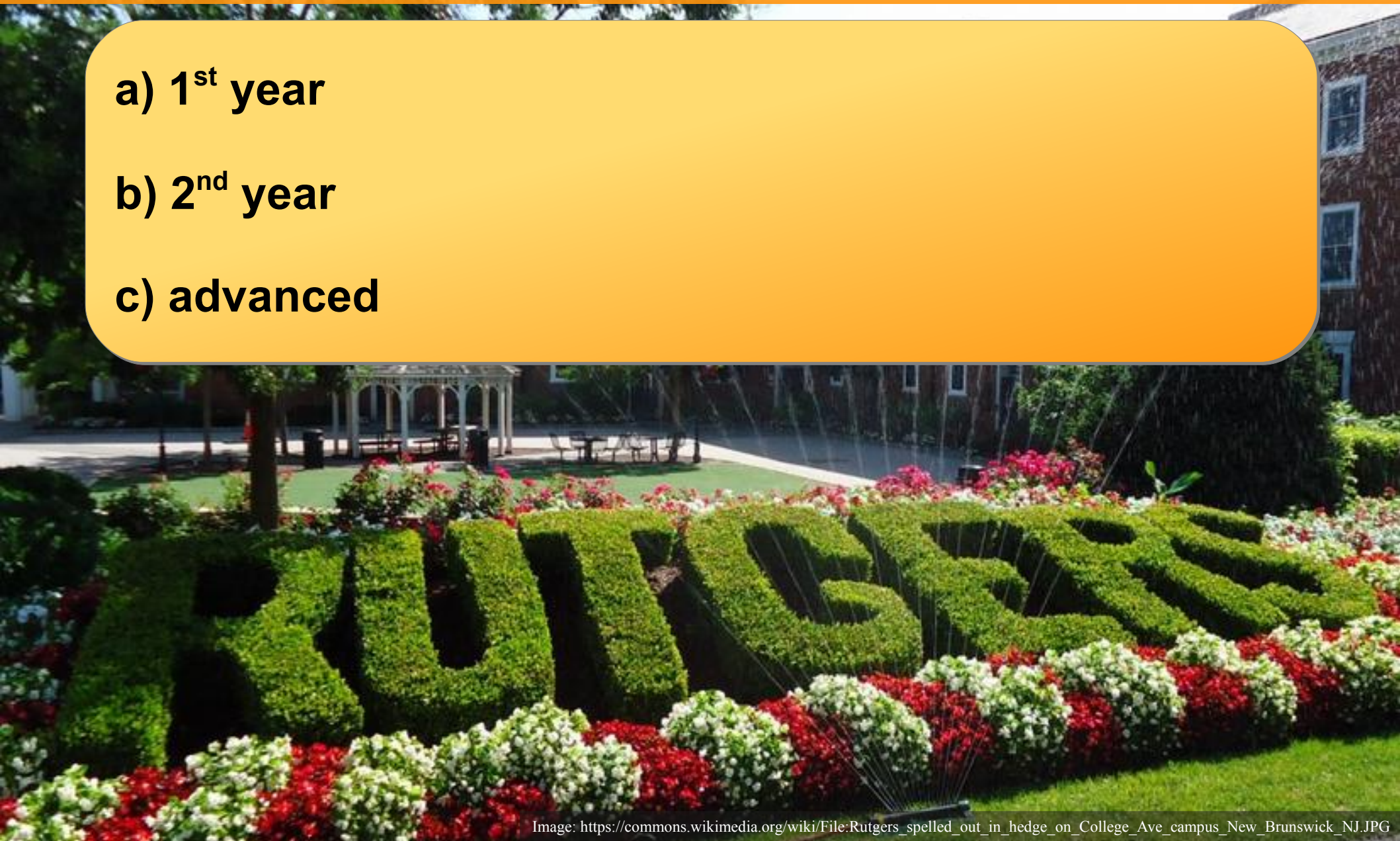
a) CS

b) non-CS



What about you?

- a) 1st year
- b) 2nd year
- c) advanced



Do you know any

a) Java

b) Python



Do you know:

- a) R**
- b) Python Data Science Stack**
- c) Machine Learning Tools**



Grades

Grading:

- **30% Assignments**
- **30% Course Project**
- **20% Mid-Term Exam**
- **20% Final Exam**

We will sample attendance in the recitations. Up to 10% Bonus if you attend and participate regularly.



Course Project

Sep

Oct

Nov

Dec

- Project proposal
- Intermediate Report/Assignments
- Presentation (some)
- Final report

Topics

January	Introduction
	Data Collection and Analysis
February	Working with Big Data
	Textual Data, Social Networks, Clustering
March	Clustering, Machine Learning
	Mid-Term, Spring Break
April	Data Mining, Applications
	Presentations, Final Exam

What Questions Do You Have?



The Growing Importance of Data (Science)



Data in the Past

**Cave paintings
and
petroglyphs**



Data in the Past



**Recording
information
via Egyptian
Hieroglyphs**

Data in the Past

心作道情六門休歇不勞形有緣不是
無用雙眉却弟兄
迷同未悟人無心勝負自安神從前古德稱
同此門中有幾人
大法眼禪師因僧看經頌
古教不免心中鬧欲免心中鬧但知看
古德頌曰照溫皆空處深行般若時
苦厄決定證無生
見正性先摧我相亡形容何更有六尤本無
雲明性儼然世界通
不出還燒木智因情起却除情正心親妻
心僧替你入苦舌
和尚抄錄佛祖直指心體要義
宣光七年丁巳七月日清
守鑄字印施

**Movable Type
Printing**
(China, then Korea,
then Gutenberg in
Germany)

Data in the Past

STATE Ohio COUNTY Cuyahoga TOWNSHIP OR OTHER DIVISION OF COUNTY Bedford Township NAME OF INSTITUTION Y NAME OF INCORPORATED PLACE Bedford City WARD OF CITY 1 SUPERVISOR'S DISTRICT NO. 19 SHEET NO. 11 (191-478) ENUMERATION DISTRICT NO. 566 7387

FOURTEENTH CENSUS OF THE UNITED STATES: 1920—POPULATION

ENUMERATED BY ME ON THE 9th DAY OF January, 1920. Robert T. Fox ENUMERATOR.

PLACE OF ABODE.	NAME of each person whose place of abode on January 1, 1920, was in this family.	RELATION.	SEX.	RACE.	BIRTH DATE.	BIRTH PLACE.	CITIZENSHIP.	EDUCATION.	NATIVITY AND MOTHER TONGUE.		OCCUPATION.
									Place of birth.	Mother tongue.	
1187 229 205	Caro, Caroline D.	Daughter	F	W	35	Pa		High School	English	English	Stenographer
	Caro, William F.	Brother	M	W	38	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	38	Pa		High School	English	English	Stenographer
	Caro, John C.	Brother	M	W	38	Pa		High School	English	English	Stenographer
	Caro, John D.	Brother	M	W	38	Pa		High School	English	English	Stenographer
	Caro, John E.	Brother	M	W	38	Pa		High School	English	English	Stenographer
111 230 281	Caro, Robert C.	Brother	M	W	23	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	38	Pa		High School	English	English	Stenographer
222 131 247	Caro, John F.	Brother	M	W	60	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	60	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	60	Pa		High School	English	English	Stenographer
250 271 250	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
1260 133 24	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
170 234 250	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
227 235 253	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
226 236 269	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
125 237 253	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
127 238 253	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer
	Caro, John F.	Brother	M	W	37	Pa		High School	English	English	Stenographer

https://commons.wikimedia.org/wiki/File:Cuyahoga_County_US_Census_Form-Herbert_Birch_Kingston_1920.jpg

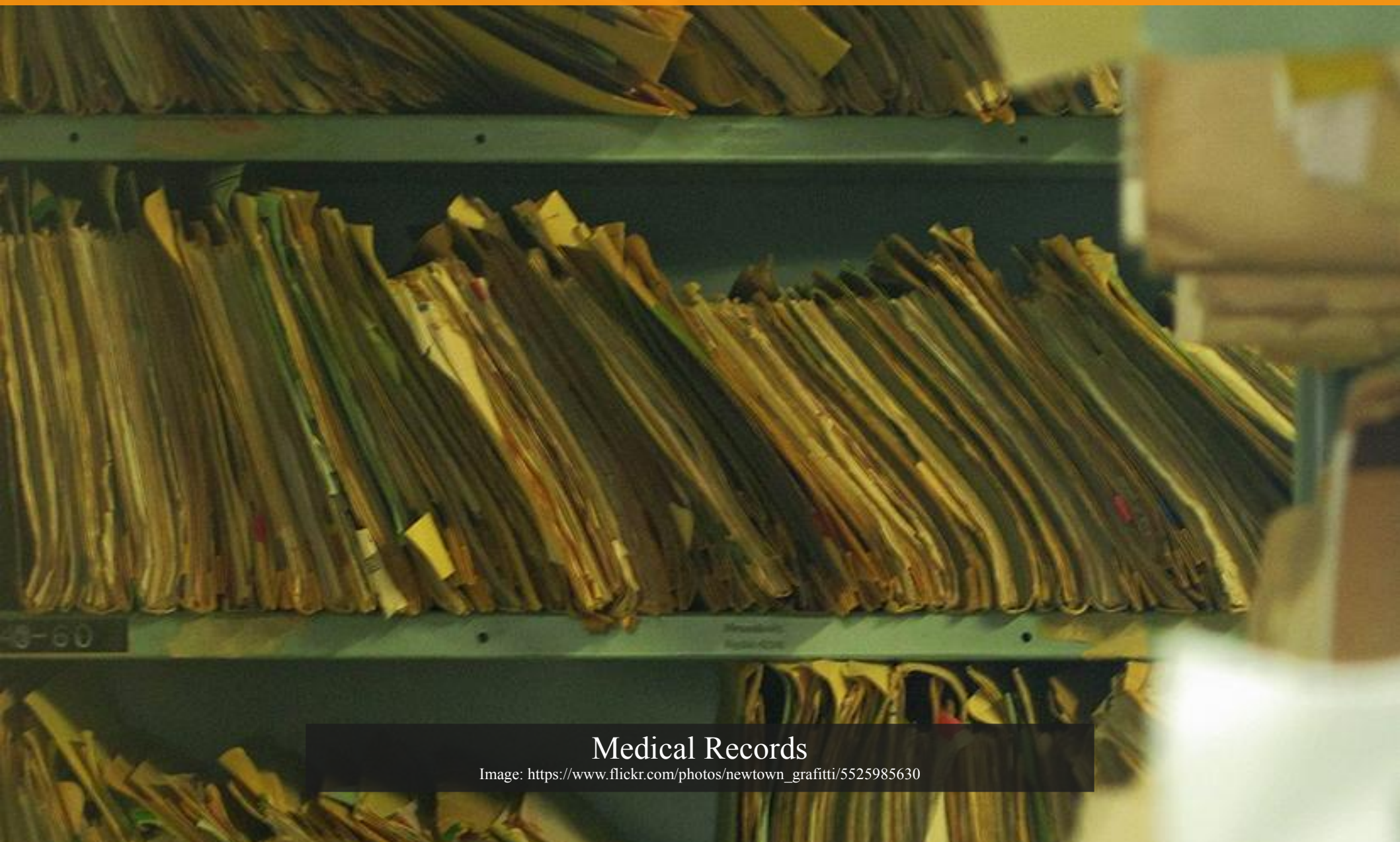
Printing books was a big step for spreading information, but recording it often still via hand-writing

Data in the Past



Cincinnati Old Main Library

Data in the Past



Medical Records

Image: https://www.flickr.com/photos/newtown_graffiti/5525985630

Data in the Past



Archives of the Sierra-Pambley Foundation

https://commons.wikimedia.org/wiki/File:Fondos_archivo.jpg

Data in the Past

**Microfiche
to store
information
in a more
compact
form**



Data in the Past

**"This CD-ROM can
hold more information
than all the paper
that's here below me"**

**Bill Gates, 1994
(unconfirmed)**



Data Today

Max. capacity?



Data Today

**1TB (Feb. 2019):
more than
1400 CD-ROMs**

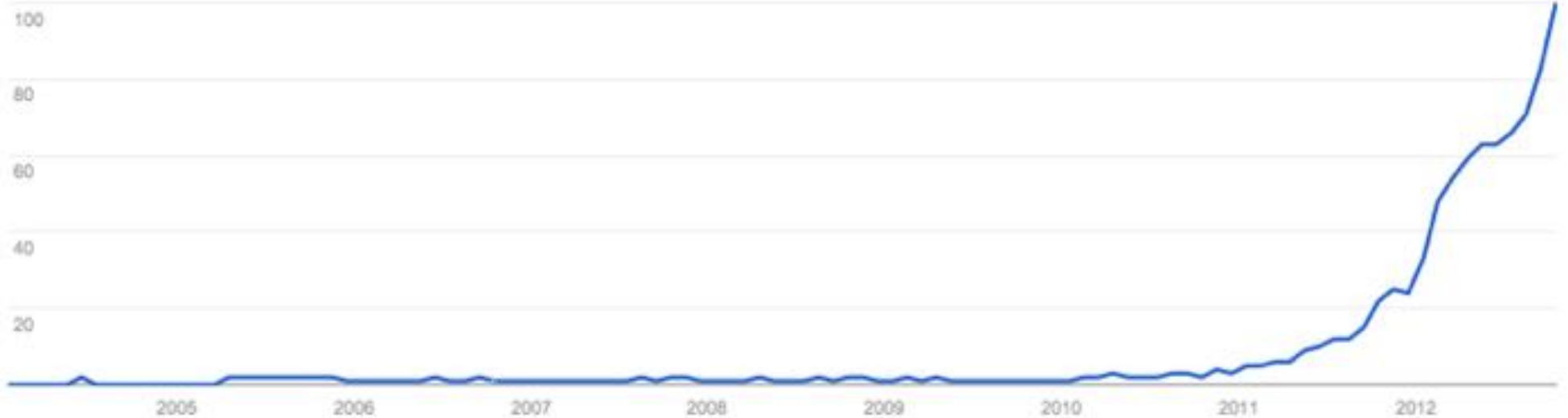


Big Data

Interest over time ?

The number 100 represents the peak search volume

☐ News headlines ☐ Forecast ?



<http://www.google.com/trends/explore#q=%22big%20data%22>

Why Big Data?

**Why do we have
so much big data?**

Why Big Data?

1. Extensive Logging and Capturing



Why Big Data?

1. Extensive Logging and Capturing



Why Big Data?

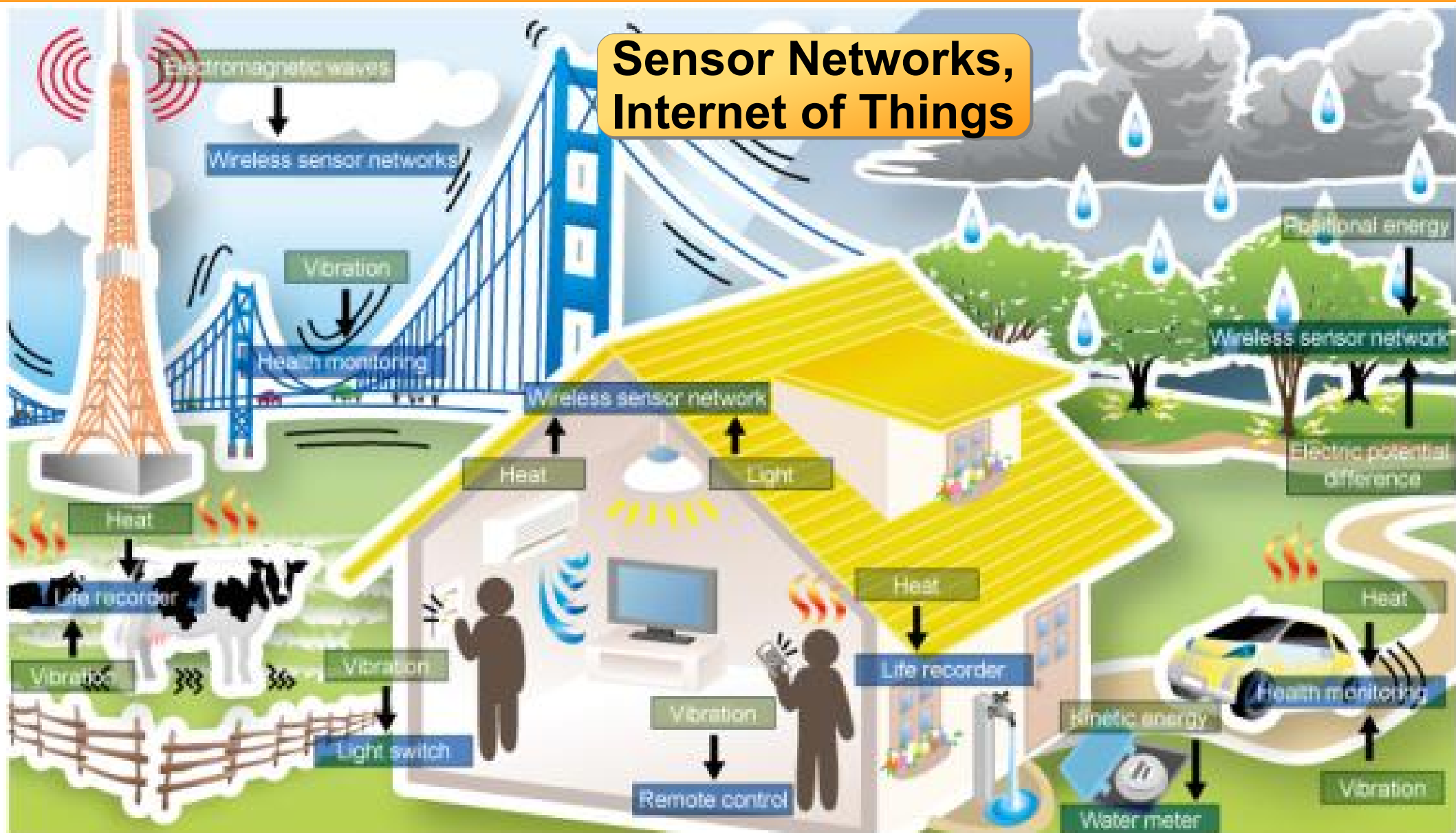
2. Connectivity and Sharing

**Data that used to be isolated is
now being shared
Example: Internet**



Why Big Data?

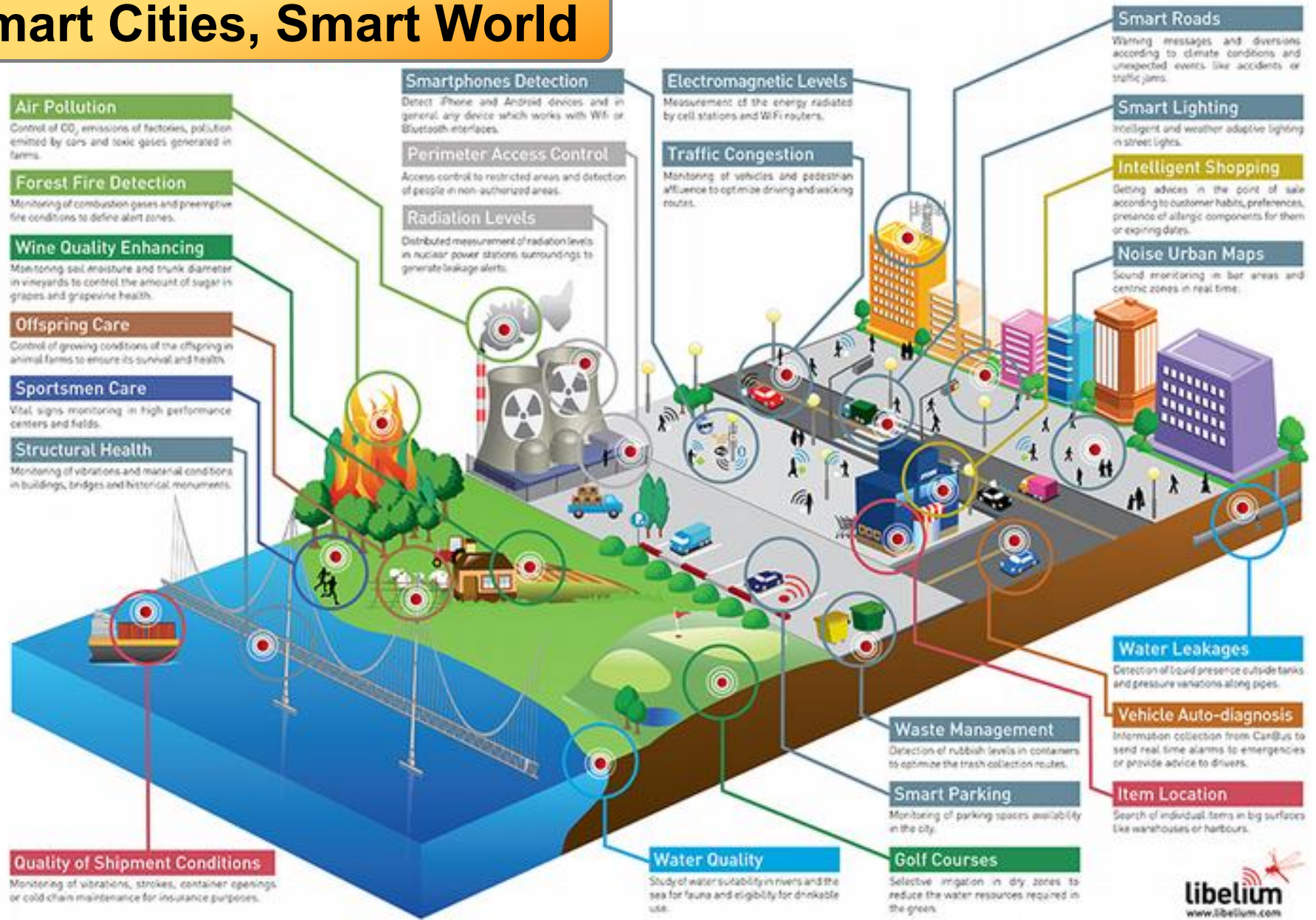
2. Connectivity and Sharing



Why Big Data?

2. Connectivity and Sharing

Smart Cities, Smart World



Big Data: Examples

1 BILLION
FACEBOOK USERS



1.13 TRILLION
FACEBOOK LIKES



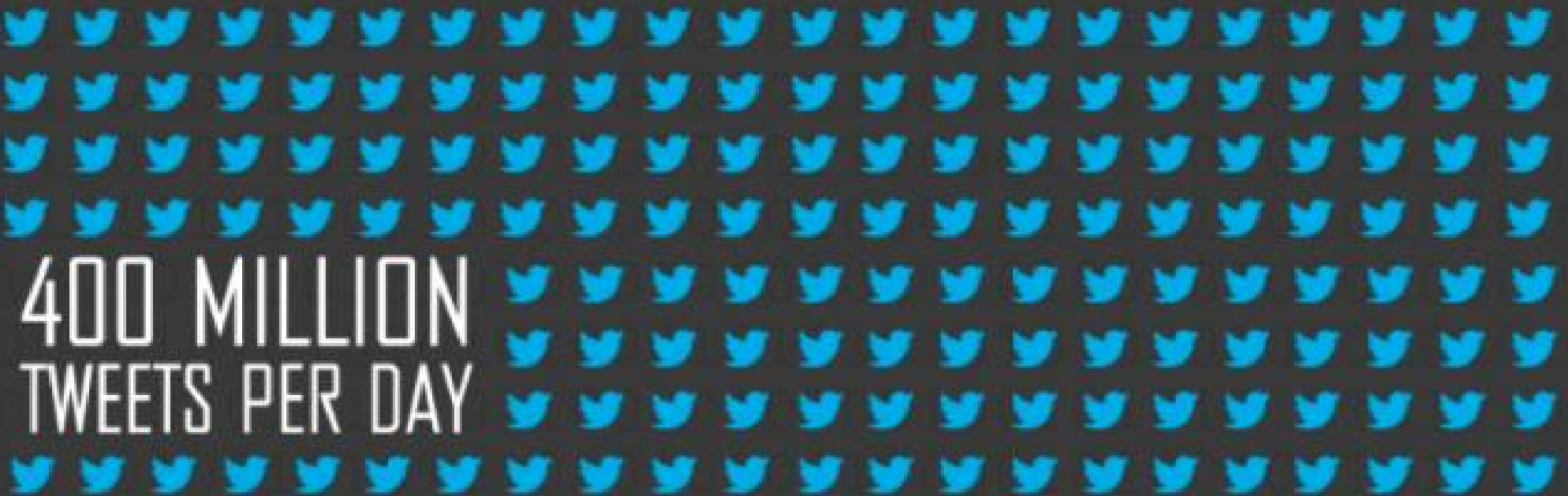
140.3 BILLION
FRIENDS



219 BILLION
SHARED PHOTOS

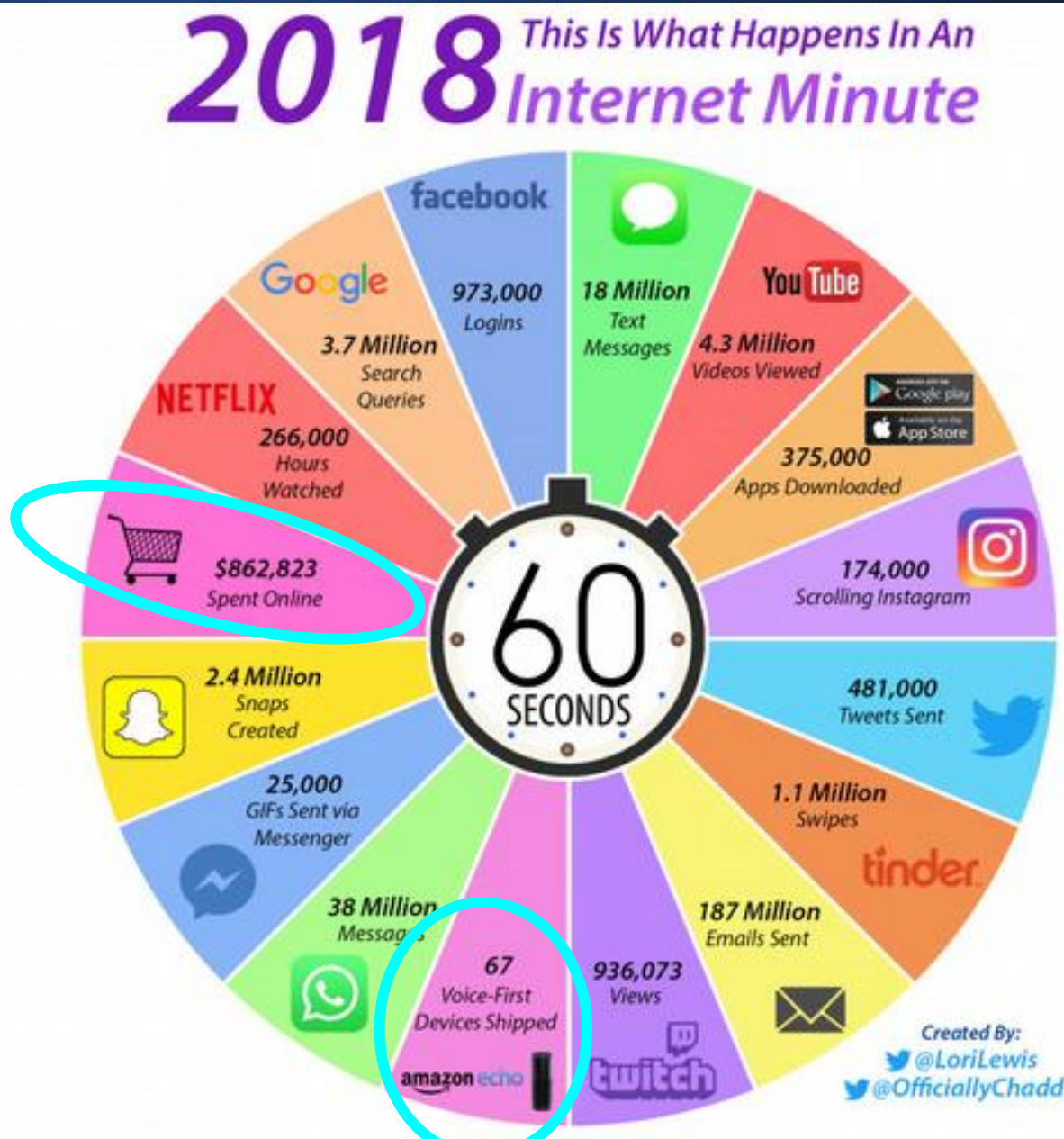
Big Data: Examples

400 MILLION
TWEETS PER DAY



72 HOURS OF VIDEO UPLOADED PER MINUTE

Big Data: Examples



Why Study Data Science?



Data Science



2.3M

Projected jobs requiring data science skills by 2018

2.9M

As our world is increasingly driven by data, there is a huge need for people with data-related skills

Notes: US data only.

Source: Burning Glass Technologies analysis of 28.9 million US job postings from 2015. McKinsey Global Institute,

Big Data: The next frontier for innovation, competition, and productivity (June 2011).

Image: <https://www.pwc.com/us/en/publications/data-science-and-analytics.html>

Data Science

glassdoor Jobs Company Reviews Salaries Interviews Know Your Worth Sign In Write Review For Employers Post Jobs Free

Q Job Title, Keywords, or Company Location Jobs Search

50 Best Jobs in America

Awards

- Best Places to Work
- Highest Rated CEOs
- Best Places to Interview

Lists

- Best Jobs
- Best Cities for Jobs
- Highest Paying Jobs
- Oddball Interview Questions


Trends

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? [Find out how.](#)

United States 2017 12k Shares f t in

1 Data Scientist



4.8 / 5 Job Score	4.4 / 5 Job Satisfaction
\$110,000 Median Base Salary	4,184 Job Openings

[View Jobs](#)

2 DevOps Engineer

“Glassdoor's 50 Best Jobs in America report identifies specific jobs with the highest overall Glassdoor Job Score. The Glassdoor Job Score is determined by weighing three factors equally: earning potential (median annual base salary), overall job satisfaction rating, and number of job openings.”

Data Science

glassdoor Jobs Company Reviews Salaries Interviews Salary Calculator Sign In Write Review For Employers Post Jobs Free

Job Title, Keywords, or Company Jobs Location Search

50 Best Jobs in America for 2019

Best Jobs 2019 United States Share

Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1 Data Scientist	\$108,000	4.3/5	6,510	View Jobs
#2 Nursing Manager	\$83,000	4/5	13,931	View Jobs
#3 Marketing Manager	\$82,000	4.2/5	7,395	View Jobs
#4 Occupational Therapist	\$74,000	4/5	17,701	View Jobs
#5 Product Manager	\$115,000	3.8/5	11,884	View Jobs
#6 Devops Engineer	\$106,000	4.1/5	4,657	View Jobs
#7 Program Manager	\$87,000	3.9/5	14,753	View Jobs
#8 Data Engineer	\$100,000	3.9/5	4,729	View Jobs
#9 HR Manager	\$85,000	4.2/5	3,908	View Jobs
#10 Software Engineer	\$104,000	3.6/5	49,007	View Jobs


Data Science

TechRepublic. SEARCH 5G Developer Top DaaS providers More Newsletters Forums Resource Library TR Premium

Why data scientist is the most promising job of 2019

by **Allison DeNisco Rayome** in **CXO** on January 10, 2019, 6:00 AM PST

Data scientists saw a 56% increase in job openings in the US over the past year, according to LinkedIn.



WHITE PAPERS, WEBCASTS, AND DOWNLOADS

- Okta + ICSynergy: Hybrid IT Got You Down? There's A Better Way with Identity**
Webcasts from Okta
[WATCH NOW](#)
- Free Trial: Identify and Track What's Breaking the Internet**
Downloads from SolarWinds
[DOWNLOAD NOW](#)
- Are Your Access Rights Putting You at Risk?**
Downloads from SolarWinds
[DOWNLOAD NOW](#)

Getting these jobs has become harder than in the past, but the overall future still seems very promising

Data Analytics



Image: Corbis



Image: Caixin

Alibaba: 31 million orders
per day! (2014)

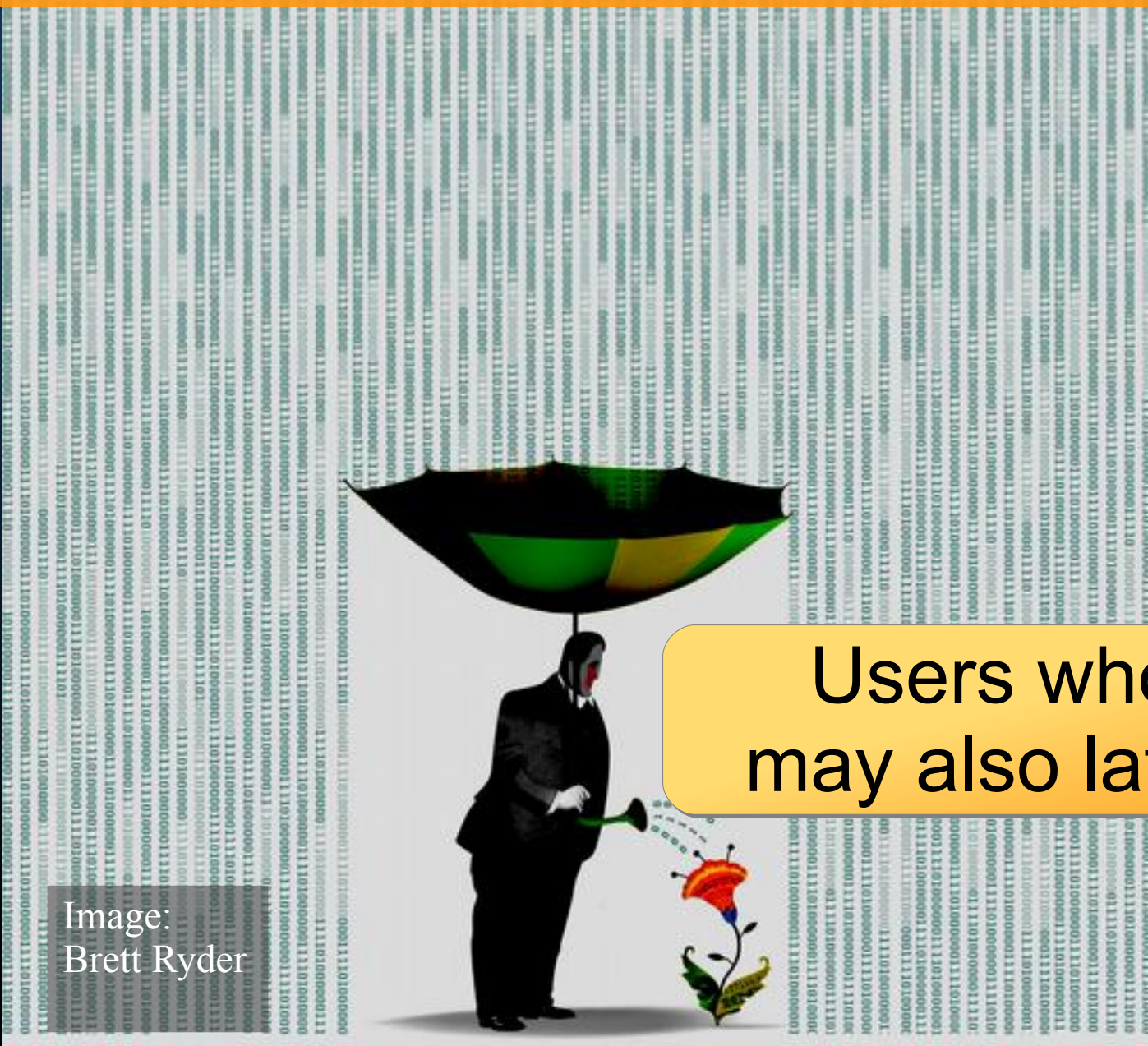
Data Analytics

Alibaba: 31 million orders
per day! (2014)

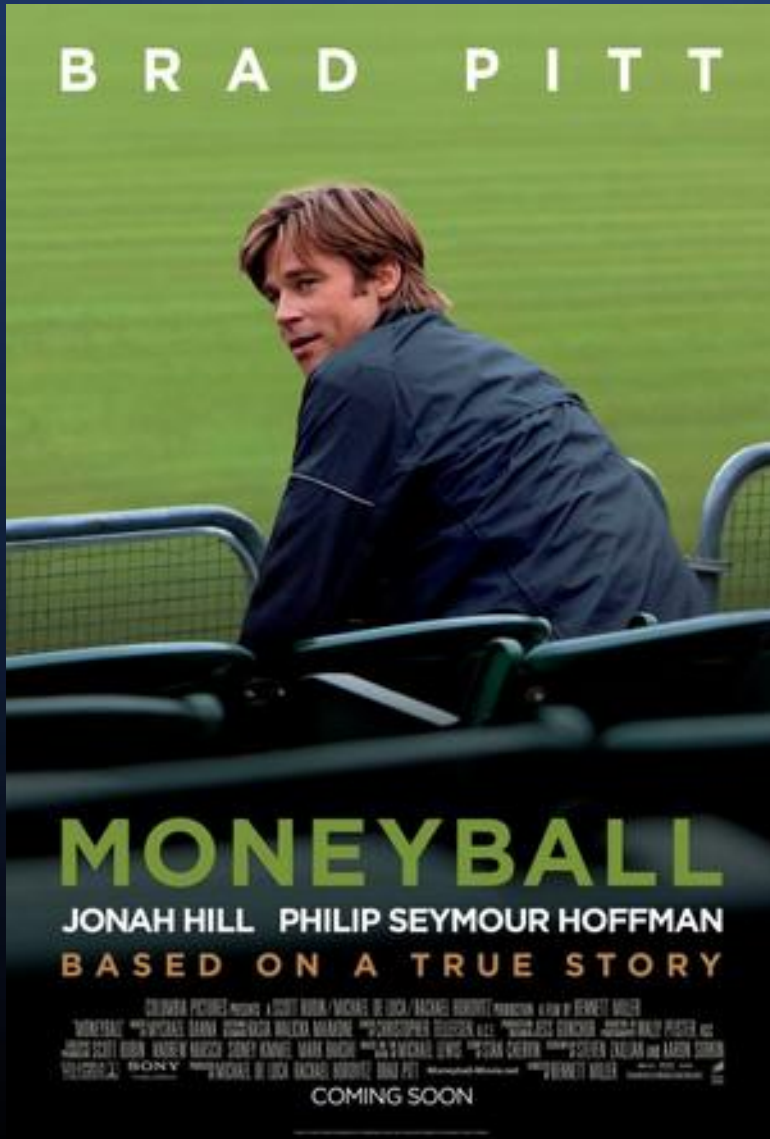
Data Analytics

Users who buy X
may also later buy Y.

Image:
Brett Ryder



Data Analytics



True Story

Oakland A's had no money for top players, but used data mining to predict performance of less known players.

Won
20 games
in a row.



Data Analytics

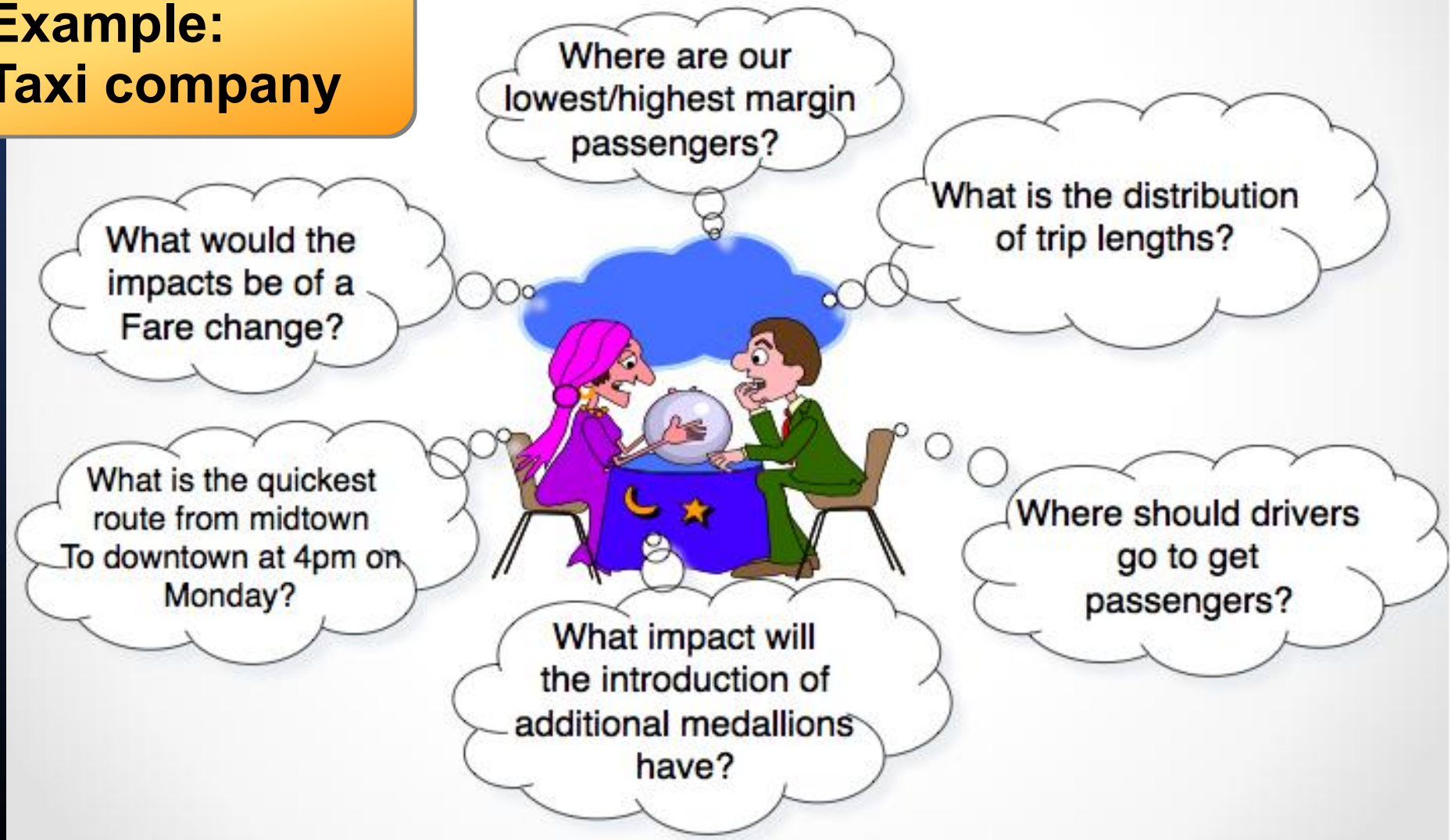
In the future, every decision that mankind makes is going to be informed by a cognitive system [...] and our lives will be better for it.

Ginni Rometty, IBM CEO



Data Science Questions

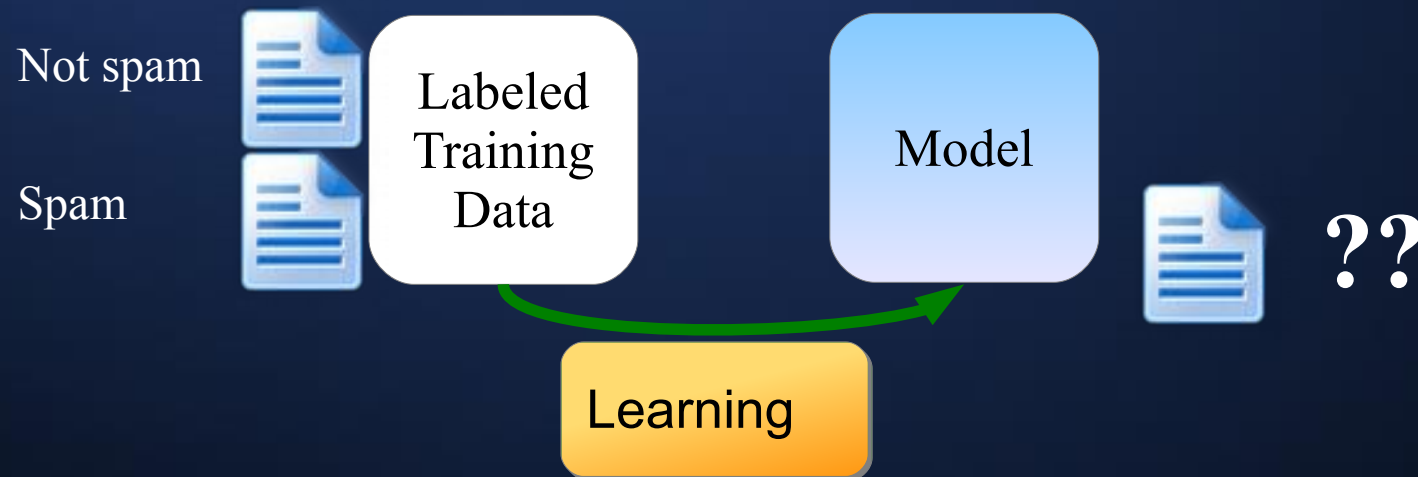
Example: Taxi company



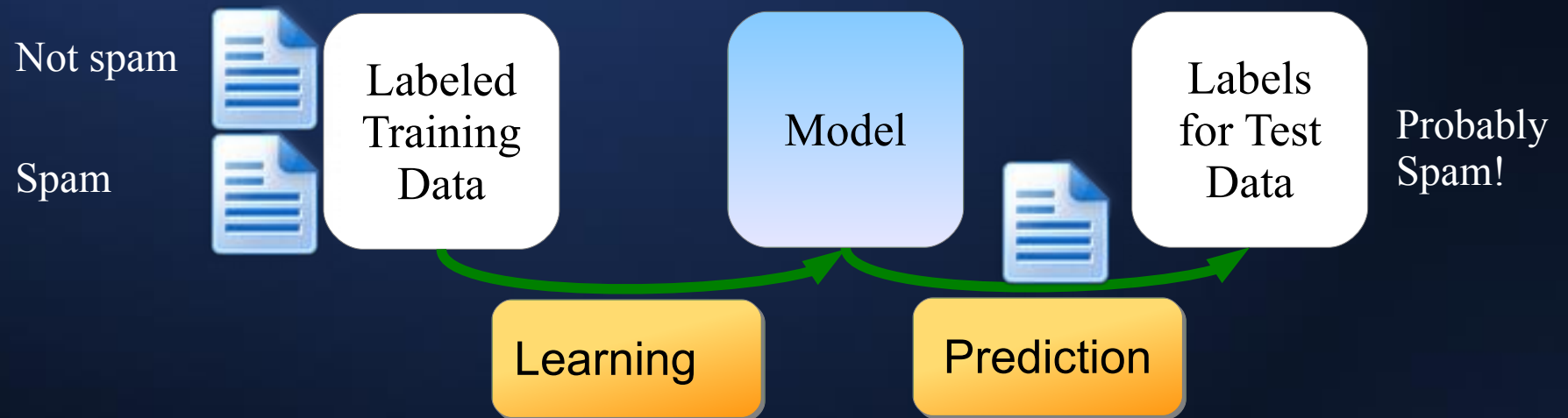
What Exactly is Data Science?



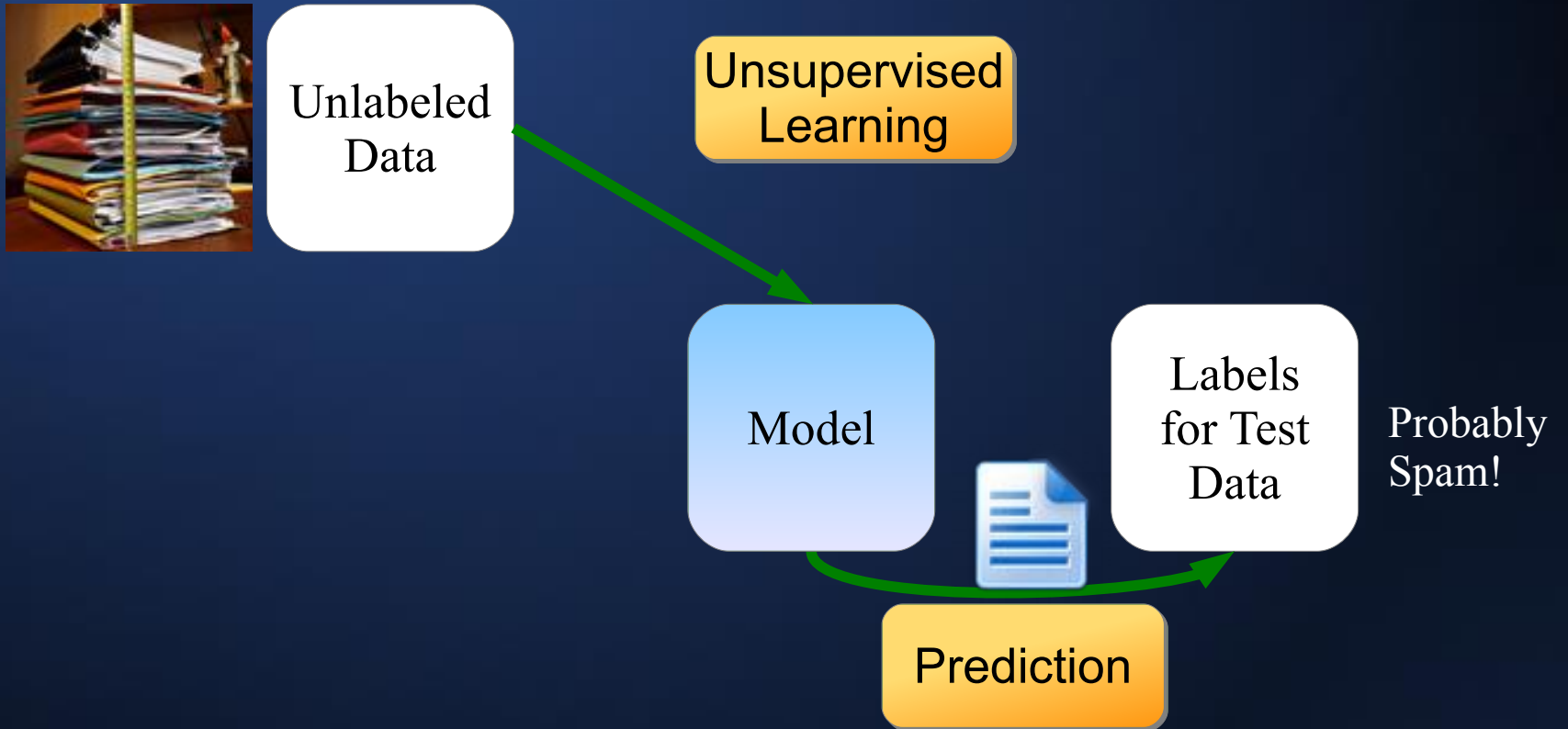
Supervised Machine Learning



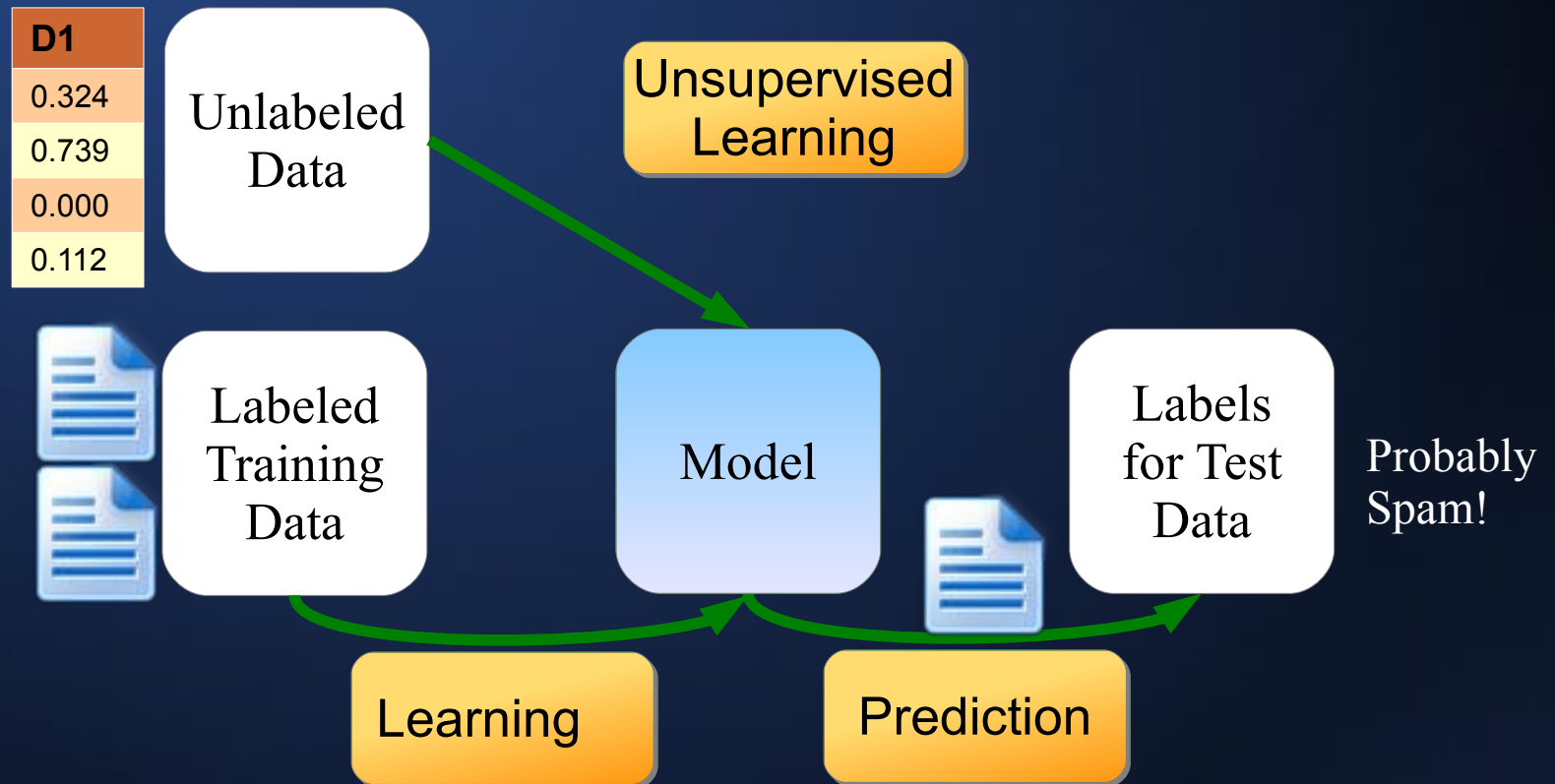
Supervised Machine Learning



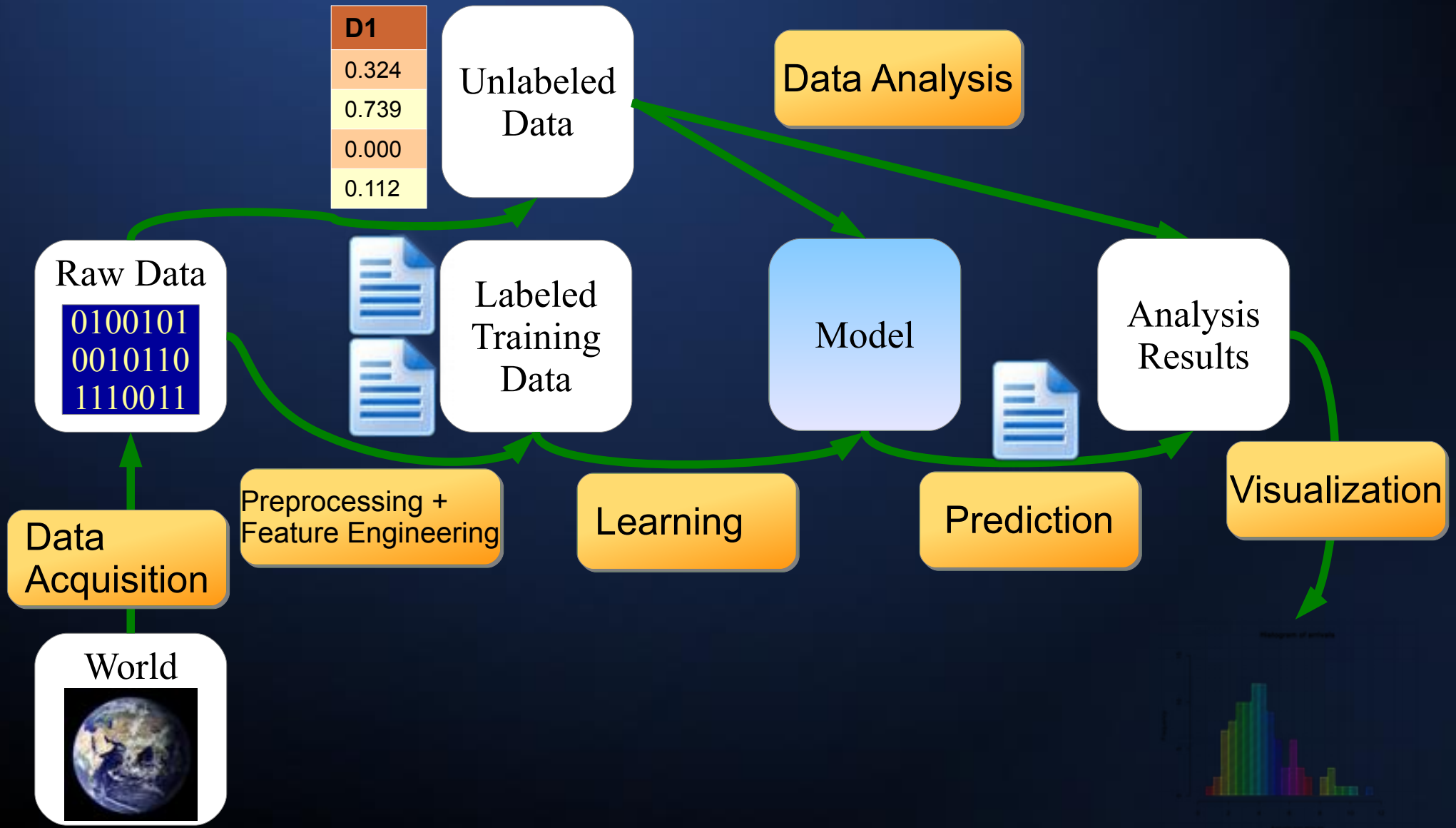
Unsupervised Learning



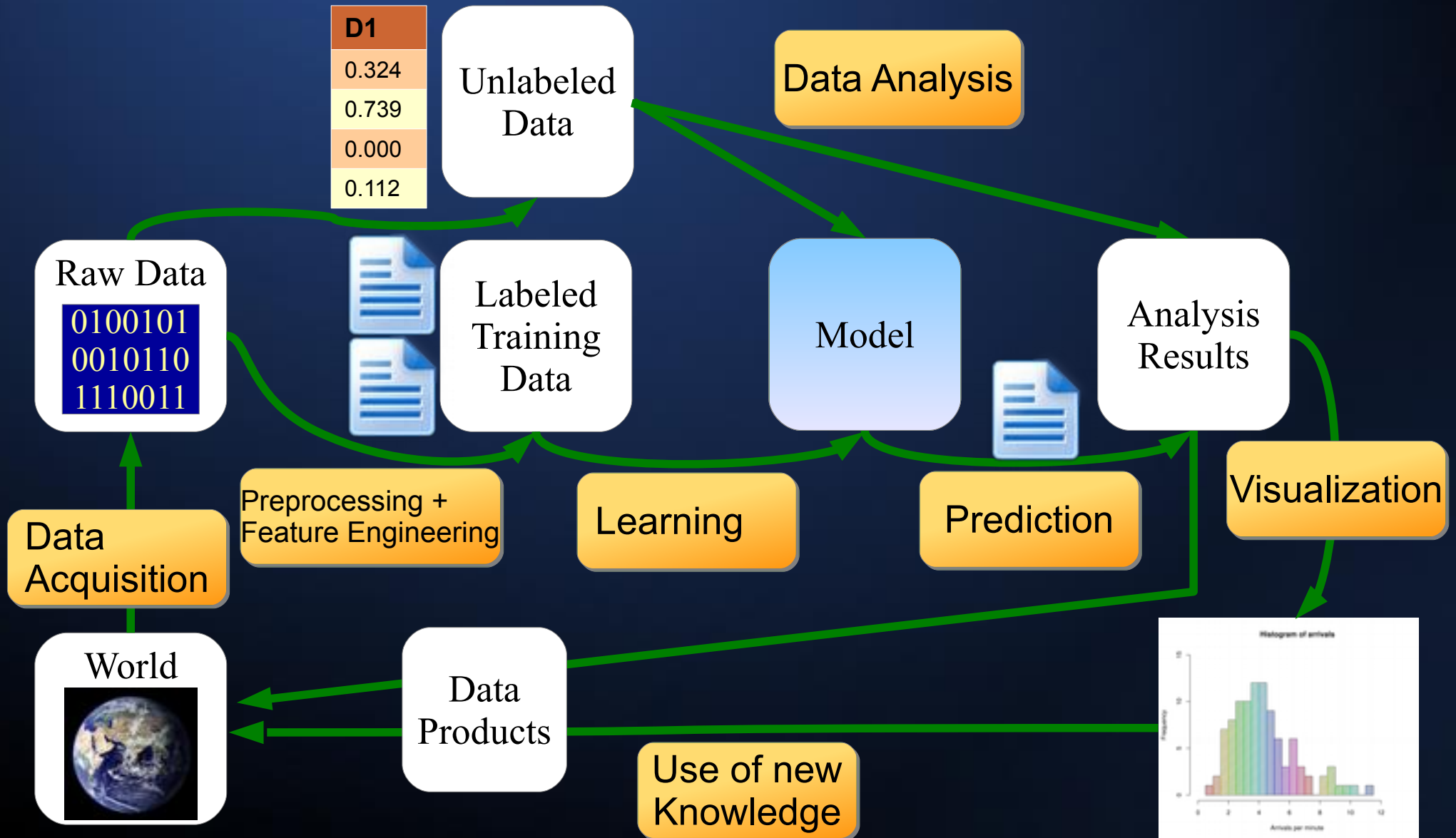
Machine Learning



Data Mining



Data Science



Data Science vs. Data Engineering, Data Analytics

Data Engineering

Data Management,
Database
Administration,
Big Data Management

Data Pipelines,
Distributed Processing

Data Integration

Data Engineers build
the software infrastructure
that allows data to go
where it needs to be,
enabling company operations
as well as Data Science



Have you hugged your on-call engineer today? Or said thank you? (disclaimer, don't know the person but they're a freaking hero! 🙌)

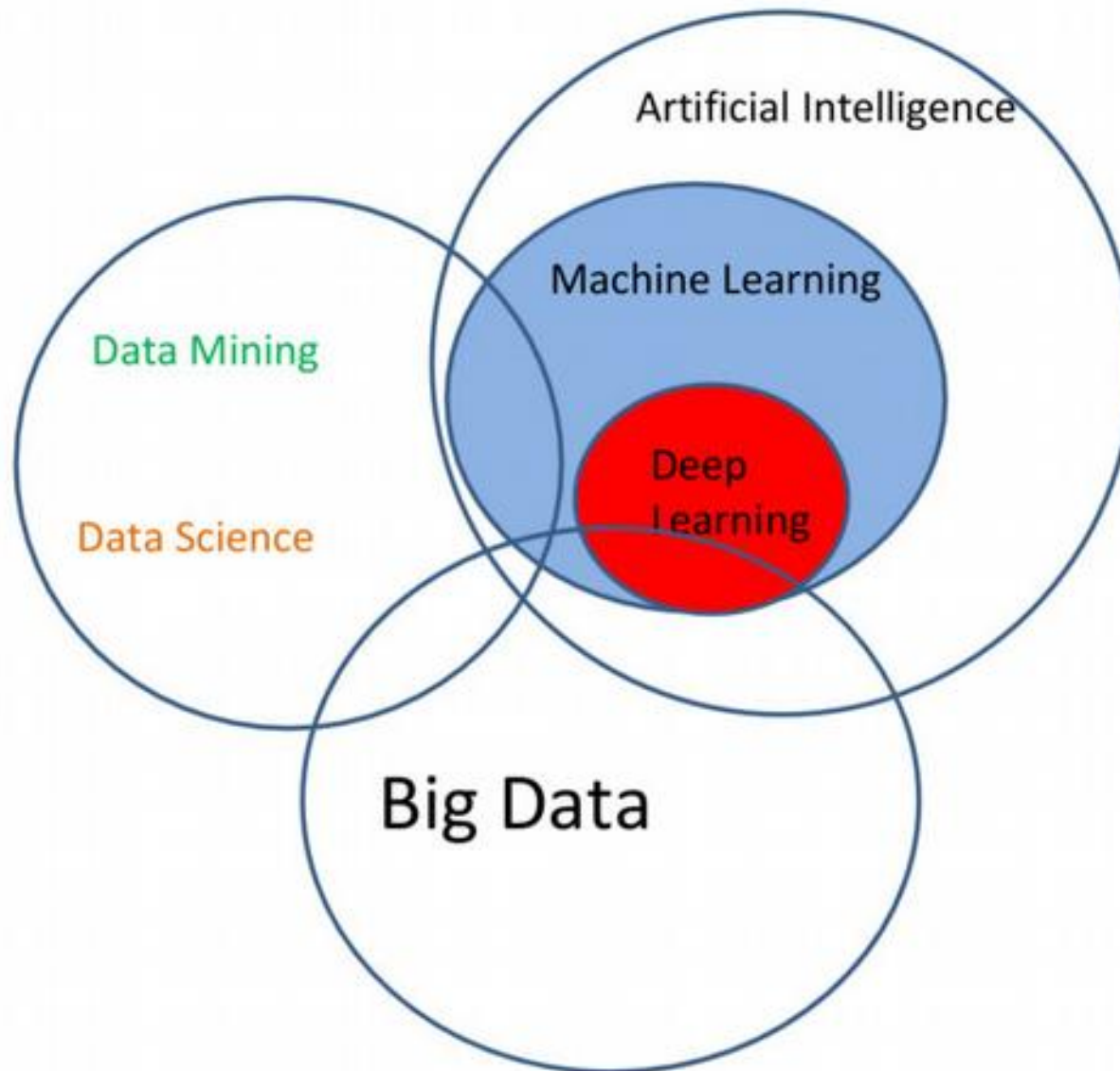


<https://twitter.com/opinionatedpie/status/976937160547930112>

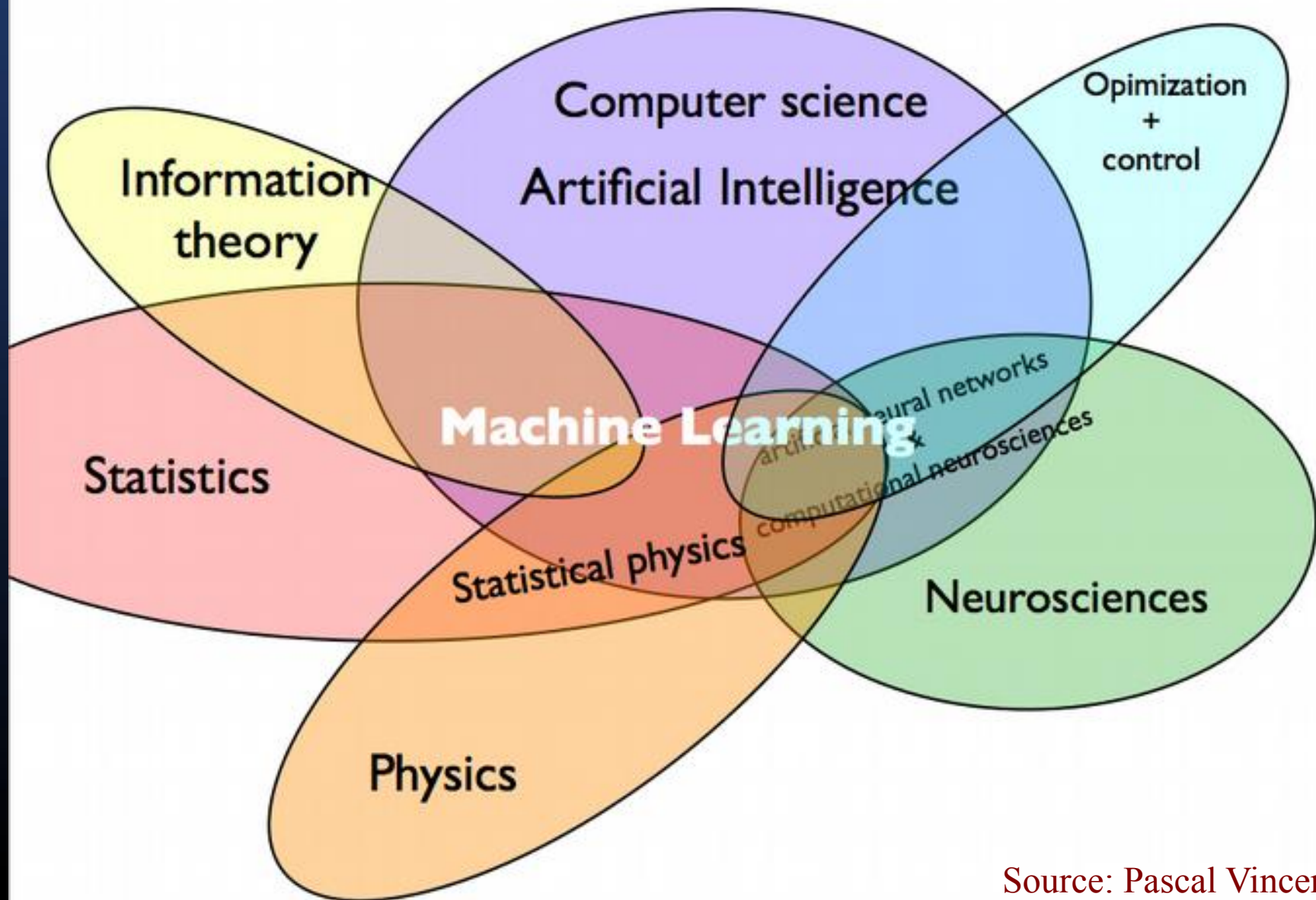
Data Science vs. Data Engineering, Data Analytics

Data Engineering	Data Science	Data Analytics
Data Management, Database Administration, Big Data Management	Write code (sometimes interactively in Jupyter, R, etc.)	SQL queries, Excel, etc.
Data Pipelines, Distributed Processing	Develop models using Data Mining/ML algorithms	Computing statistics
Data Integration	Open-ended, often predictive	Focus on Business metrics (e.g. Key Performance Indicators)
	Create new data products	Create reports, visualizations (charts etc.)

Artificial Intelligence



Other Fields



Conclusion: Need a Mix of Different Skills

The Data Science Cake



Ingredients:

50g statistics
120g linear algebra
200g programming
1kg visualisation
300g software engineering

twitter.com/jensdittrich

Additional skills:

creativity
out of the box thinking
grit
team spirit

Opportunities

**The best way
to predict
the future
is to
invent it.**

~ Alan Kay ~

Questions?



Summary

■ Data Keeps Growing

- ▶ Increasingly digital world, sensors

■ Data Science is Among the Top Jobs

- ▶ High demand
- ▶ Appealing job

■ What Exactly is Data Science?

- ▶ Greater focus on real-world usage than machine learning, data mining, etc.
- ▶ But requires more technical knowledge than Data Analytics