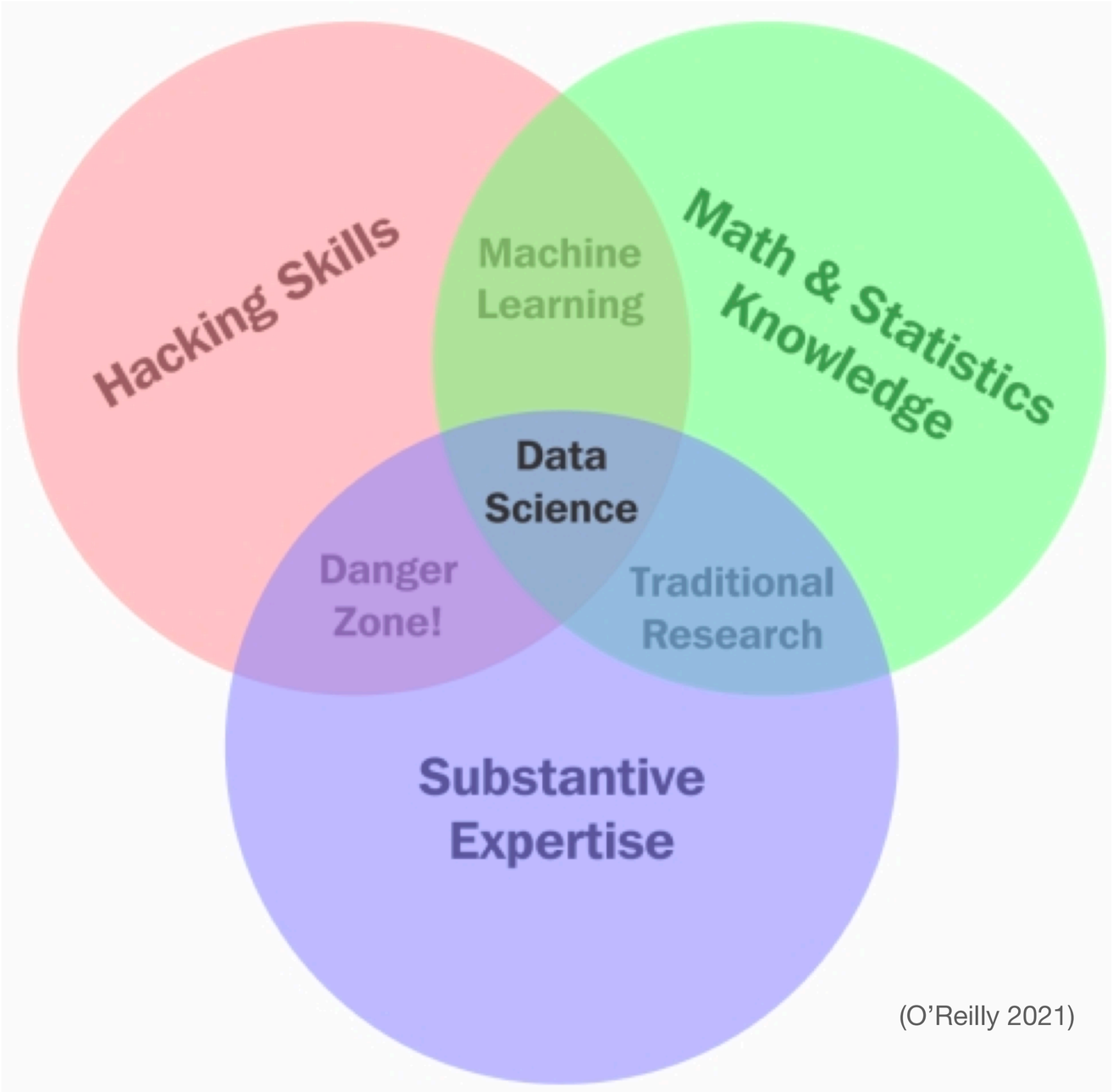
Data Science: Course Overview

Tech Frontiers

Rachel Cox and Zack Kilpatrick

July 12 & 13, 2021



(O'Reilly 2021)

Course Logistics

- virtual people should stay muted except to ask a question to prevent feedback
- if you have a short or discussion question, feel free to unmuted; ***instructor likely won't be closely monitoring the chat***
- if you have a technical/troubleshooting python question, you can private message Adam Weimerslage and he will open a Breakout Room to discuss
- Rachel Cox and Zack Kilpatrick are your instructors; while one of us is instructing, the other will be available for help for the in-room participants
- in-room participants will be able to see and hear virtual participant questions
- we will repeat in-room participant questions for virtual participants to hear

What is data?

- facts and statistics collected together for reference or analysis

- the quantities or symbols on which operations are performed by a computer, transmitted as electrical signals and recorded on magnetic, optical, or mechanical recording media.

- ***Philosophy:*** things known or assumed as facts, making the basis of reasoning or calculation.

My e-Zone | My Students | Teaching and Learning Tools | Admin | Help | Logout

e-Zgrades v1.5 — GEOG 1028 - Geography and Our World - Grade Sheet
Section #77777, Fall 2006

EXPORT PRINT E-LINK GRADES HELP LOGOUT

Add Students / Add Category / Assignment		Exams						Extra Credit				Homework																Labs				Participation				Scaling... / Letter Grade Scale...	
Name (ID, Both) [X]	LDA	E1	E2	E3	E4	E5	Sub	EX1	EX2	EX3	Sub	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	Sub	L1	L2	L3	Sub	P1	P2	P3	Sub	Calculated Total	Grade
Clark, Brian [X]	1/15	89	98	140	80	94	501/540				0/0	9	8	10	10	10	8	9	8	3	3	20	10	10	0	10	9	137/151	18	24	24	66/75	20	9	10	39/35	743/801 (92.76%)	
Clarkson, Jonathan [X]		85	78	148	54	85	450/540				0/0	8	8	8	8	9	8	8	6	2	5	17	0	9	2	10	8	116/151	25	23	23	71/75	25	9		34/35	671/801 (83.77%)	
Fran, Brianna [X]		85	78	120	54	85	422/540				0/0	8	8	8	8	9	8	8	6	9	9	17	9	9	9	9	8	142/151	25	21	21	67/75	25	9		34/35	665/801 (83.02%)	
Johnson, Amber-Ann [X]		65	72	110	75	50	372/540				0/0	7	0	7	7	7	8	6	9	2	3	0	8	8	4	10	9	95/151	0	20	25	45/75	19	8		27/35	539/801 (67.29%)	
Mann, Horrace [X]		89	88	88	75	79	419/540				0/0	7	0	8	7	9	8	4	9	2	3	0	8	8	4	10	9	96/151	0	19	24	43/75	19	8		27/35	585/801 (73.03%)	
Newton, Isaac [X]		99	98	150	89	79	515/540				0/0	7	9	7	7	7	8	6	9	2	3	20	8	8	4	10	9	124/151	25	18	20	63/75	23	8		31/35	733/801 (91.51%)	
Robinson, Krista [X]		84	70	122	68	99	443/540				0/0	8	8	8	8	6	7	10	8	3	3	19	7	8	4	10	9	126/151	23	19	19	61/75	20	8		28/35	658/801 (82.15%)	
Smith, John [X]	9-12-06	99	97	142	90	56	484/540				0/0	8	9	9	8	10	10	8	8	3	3	18	6	10	2	10	9	131/151	25	20	18	63/75	24	10		34/35	712/801 (88.89%)	
West, Mark [X]	10-14-06	60	56	0	64	0	180/540				0/0	10	9	7	10	8	9	7	0	0	0	0	0	0	5	10	9	84/151	24	0	0	24/75	24	0		24/35	312/801 (38.95%)	
Average:		83.9	81.7	113.3	72.1	69.7	420.7/540				0/0	8	6.6	8	8.1	8.3	8.2	7.3	7	2.9	3.6	12.3	6.2	7.8	3.8	9.9	8.8	116.8/151	18.3	18.2	19.3	55.9/75	22.1	7.7	10	30.9/35	624.3/801 (77.94%)	

What is *big data?*

- “Big data” refers to a large volume of data (though it may not take up much physical space) which can be statistically and computationally analyzed (e.g., using machine learning methods) to reveal patterns, trends, and associations, especially related to human behavior and interactions

- Volume: terabytes; petabytes; exabytes; or more!

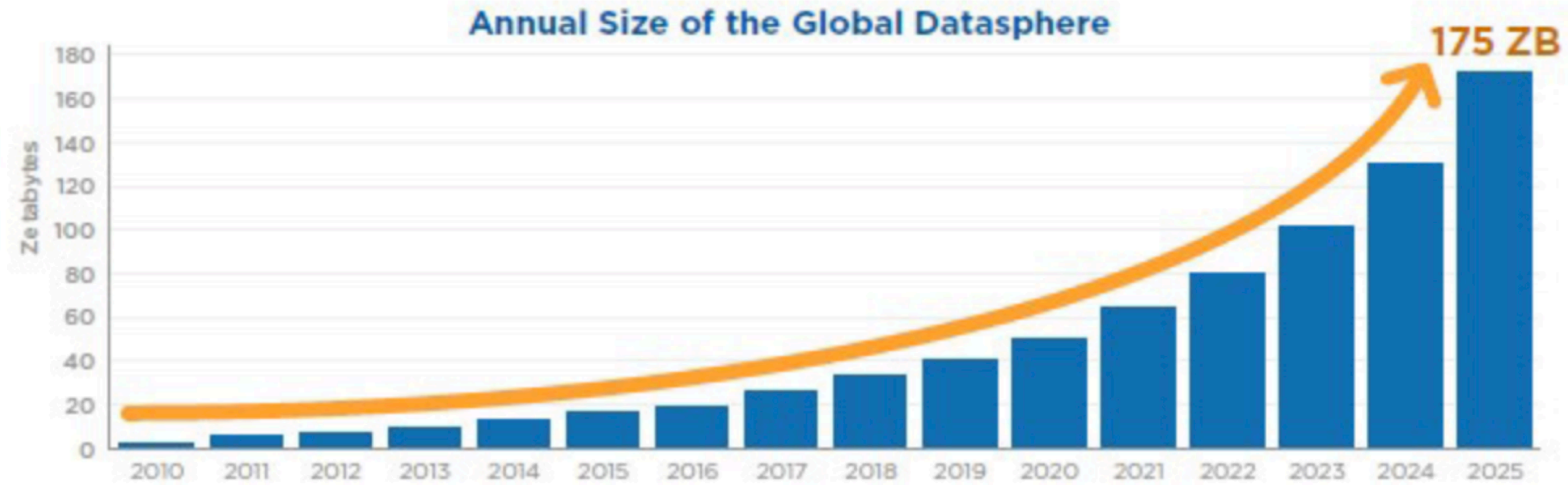
- Speed: often obtained from primary sources at a lightning pace.

- Variety: structured and unstructured formats of many different kinds

IBM Blue Gene/P supercomputer "Intrepid" at Argonne Nat’l Lab: 164,000 processor cores

Big data is the future

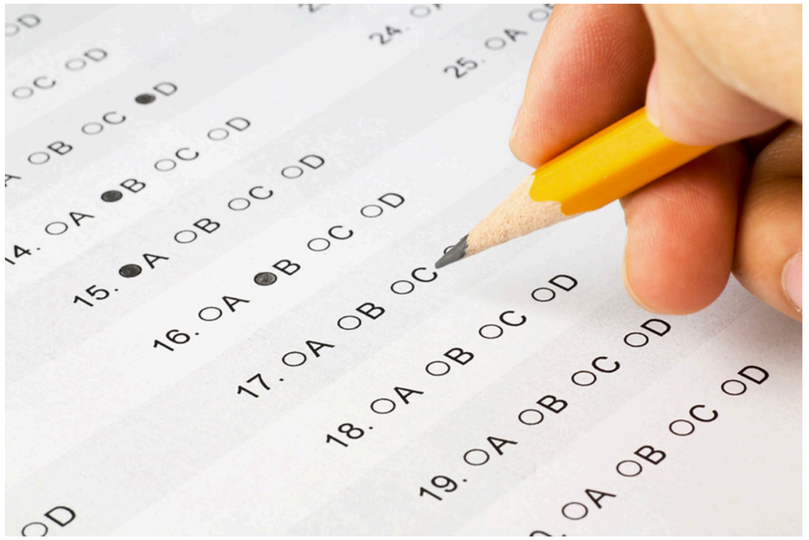
- International Data Corporation (IDC) estimates more than 59 zettabytes (10^{21} bytes or 10^{12} gigabytes) were created in 2020
- Over the next three years, more data will be created than in the past 30 years
- From the beginning of humanity to 2003, 5 exabytes (5×10^{18}) of data were created



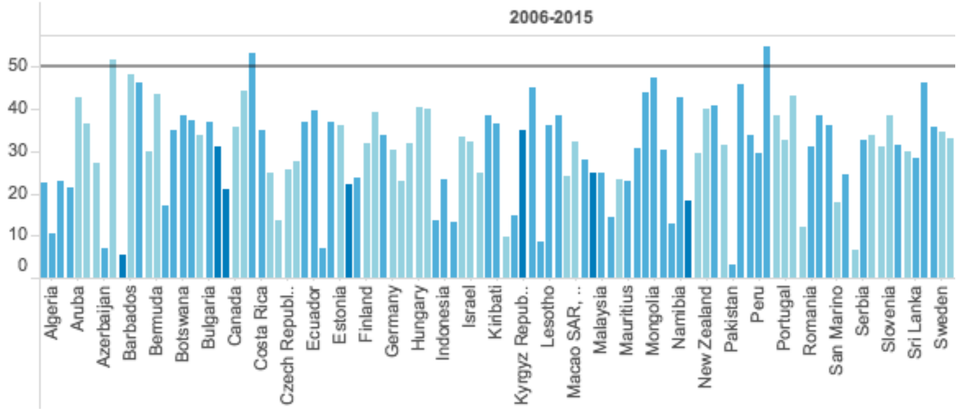
Source: Data Age 2025 report, sponsored by Seagate with data from IDC Global Datasphere, November 2018

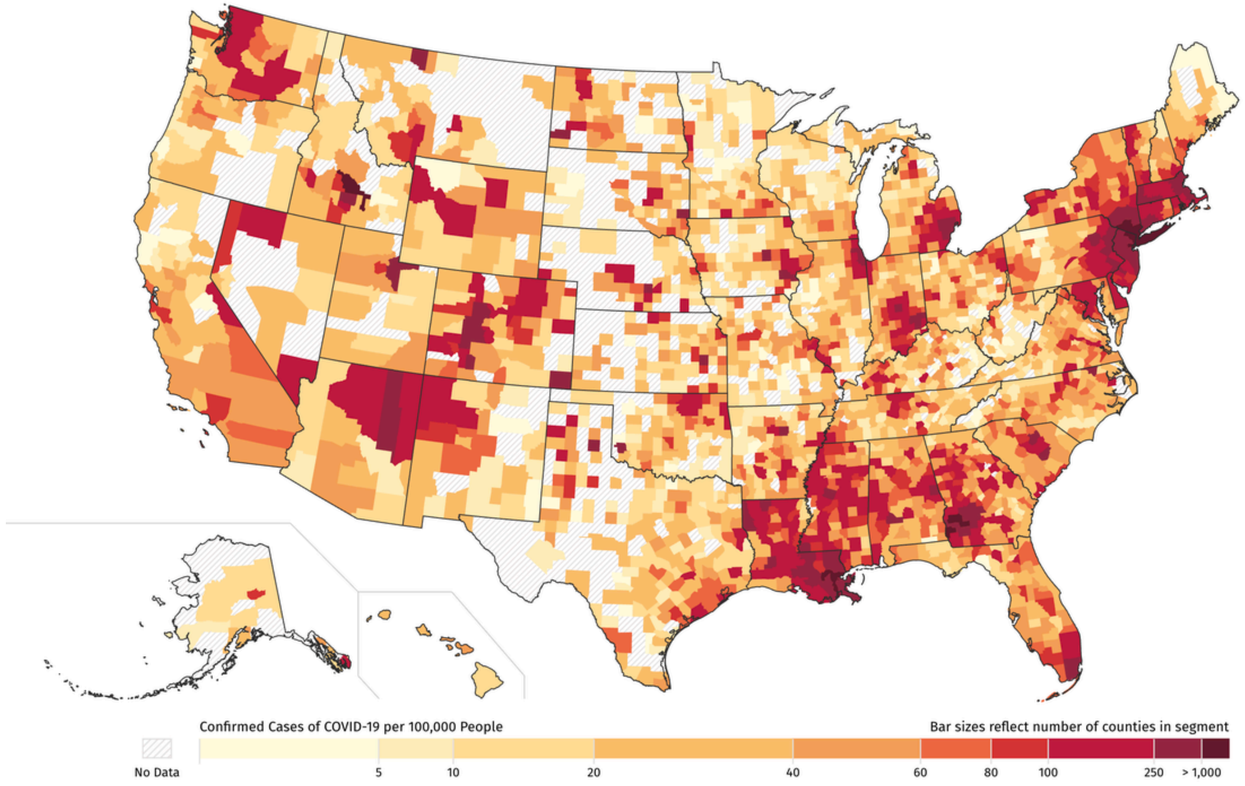
Sources of data

Some examples



- tests, interviews, surveys
- sensors, satellites
- medical intake, tracking interventions
- experiments
- open data sources: World Bank; WHO; google public data; 538; US Census; kaggle; UC Irvine ML Repository





What can using data do?

Ways to use data

- Finding a signal in the noise
  - Lada Adamic: mined data from recipe ratings to fit models and predict which recipes would rate highest — found bacon, feta, cranberries, Cool Whip all improve likelihood of recipe success

- Optimize decision-making
  - Utah Division of Wildlife Resources collected data and built models to find the ideal location for a wildlife to cross the interstate

What can using data do?

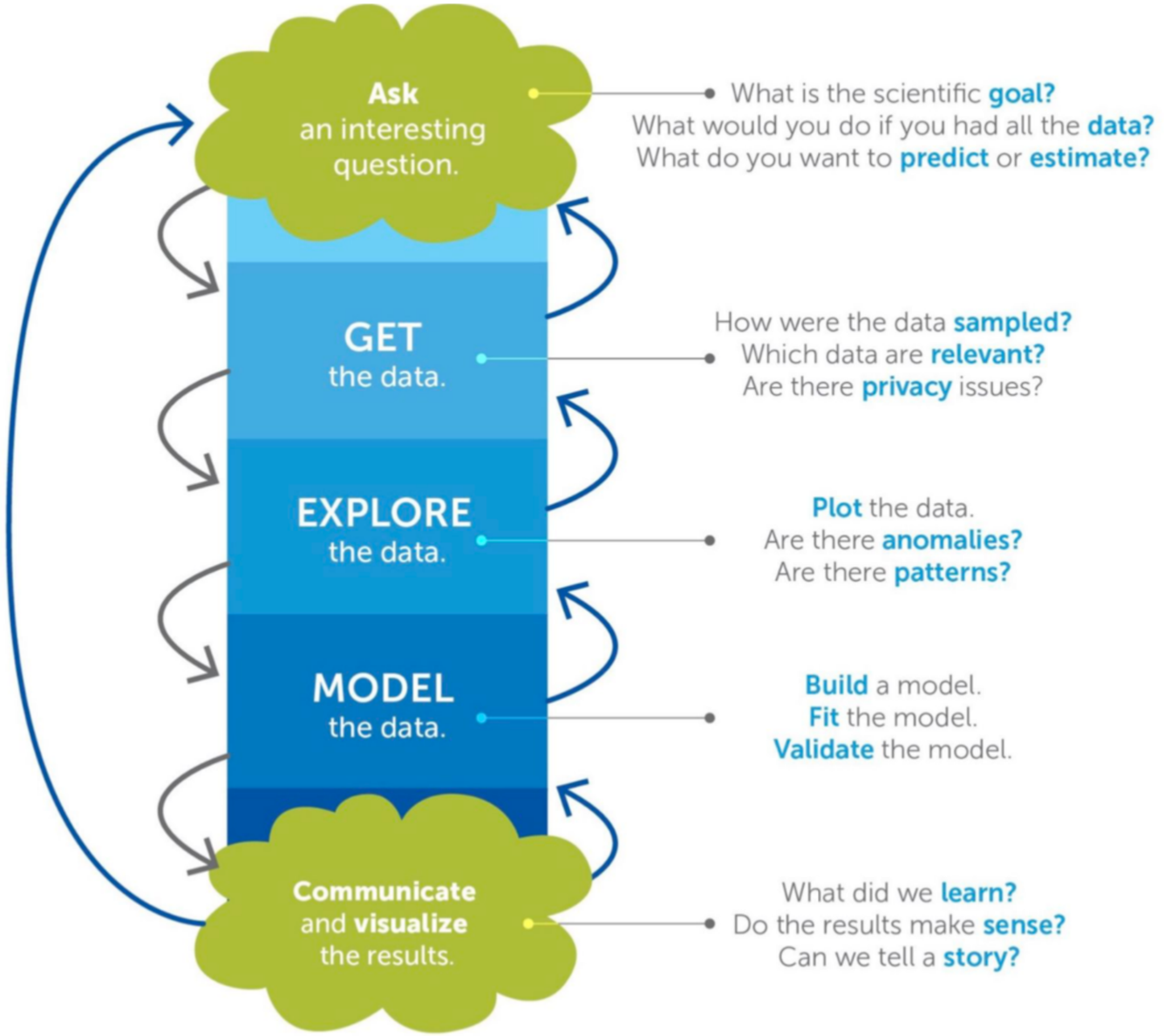
Ways to use data

- Detect patterns or aberrations
  - wearables like Fitbits and the Oura ring have been used to detect regional illness trends like onset of flu or COVID-19

- Systematizing hypothesis testing
  - nonprofit ideas42 worked with NYC courts to design and test “nudges” to reduce missed court dates: automated scheduling, text reminders, changes to ticket appearance

Data Requirements of the Scientific Method

Steps for posing questions, collecting data, testing hypotheses



What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

Plot the data.
Are there **anomalies?**
Are there **patterns?**

Build a model.
Fit the model.
Validate the model.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

Hypothesis

↓

Observations

↓

Analysis

↓

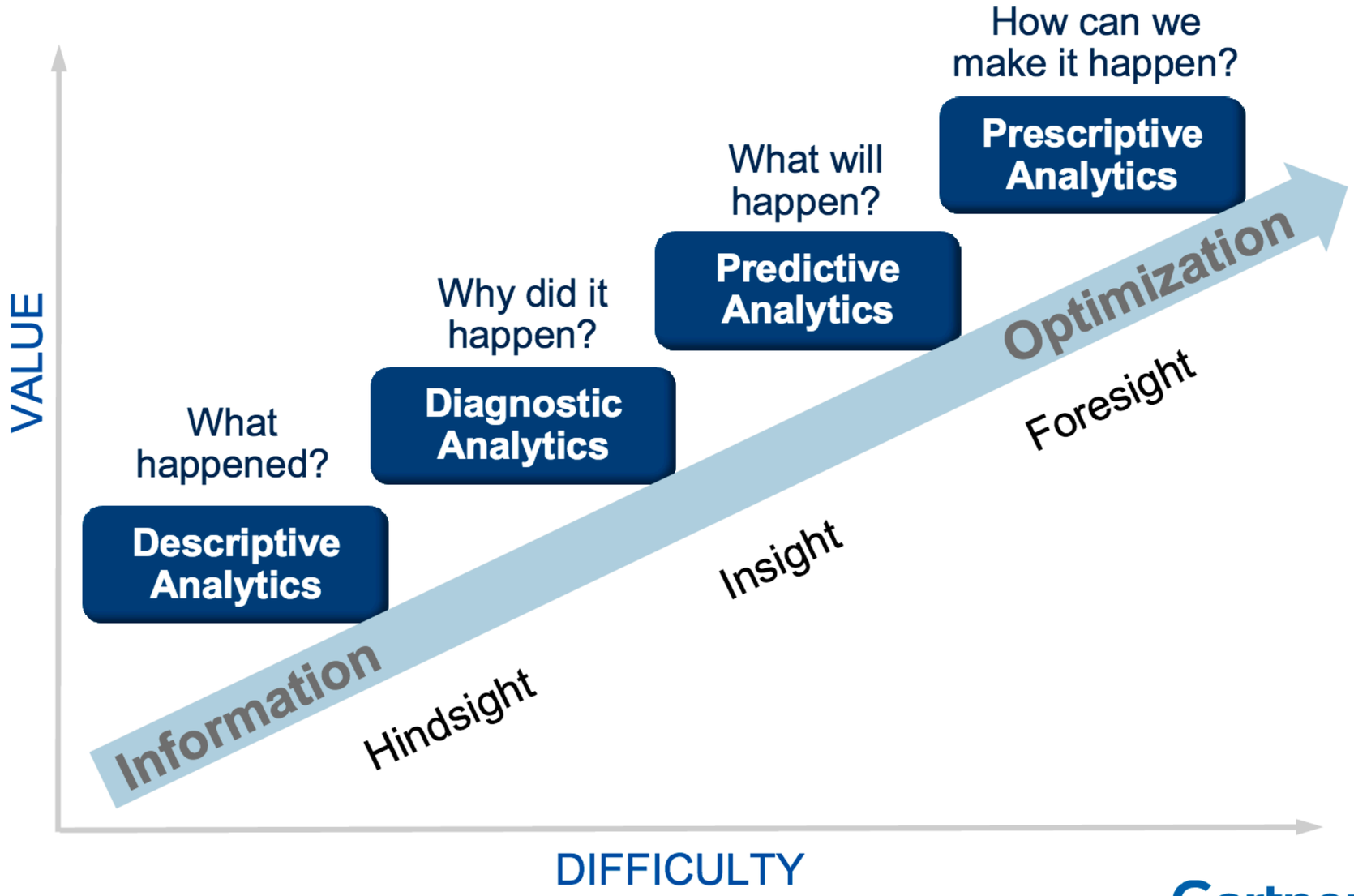
Conclusions

↓

Refinement (Do it all over again?)

Levels of data analytics

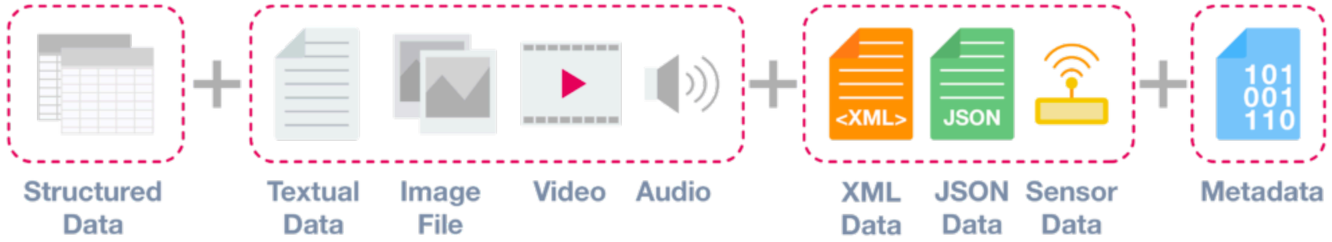
Gartner Analytic Ascendancy Model

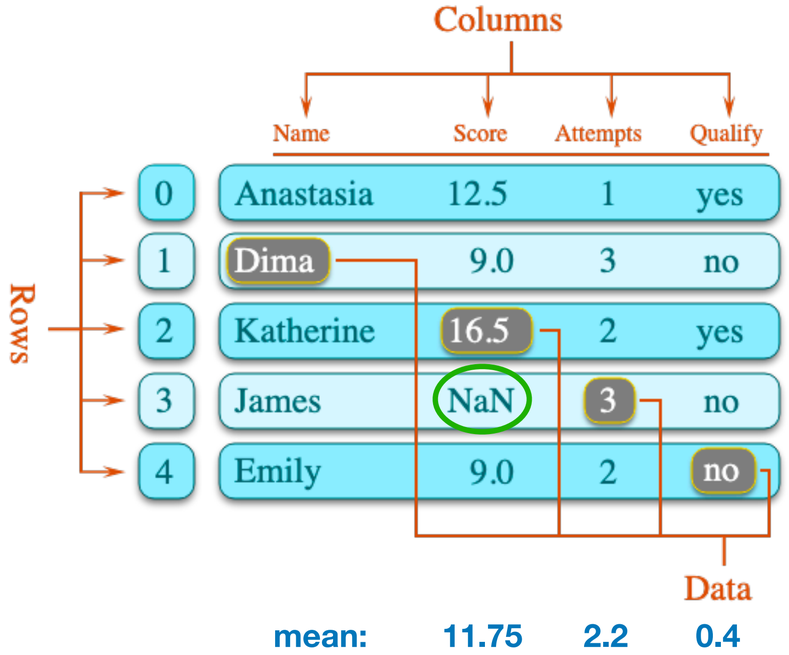


Gartner (2012)

Structuring data for analysis

In python the pandas package is very useful for this





can drop rows/columns with blank entries

can also compute descriptive statistics of each column

or append columns/rows

or count the number of nonempty/empty entries

or find ranges of values in each column

or apply functions to each column

...

Methods for data analysis

We will only learn a subset of the many methods for data analysis

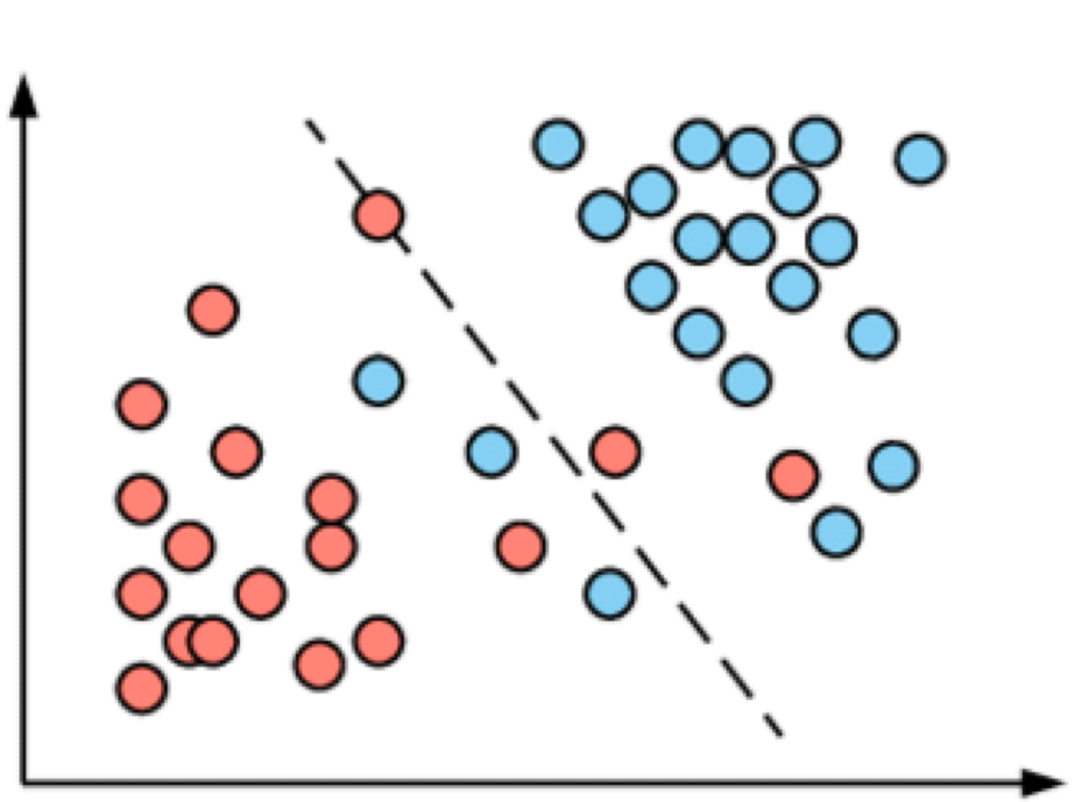
- ***supervised vs. unsupervised learning:*** knowing vs. not knowing the output/category variable

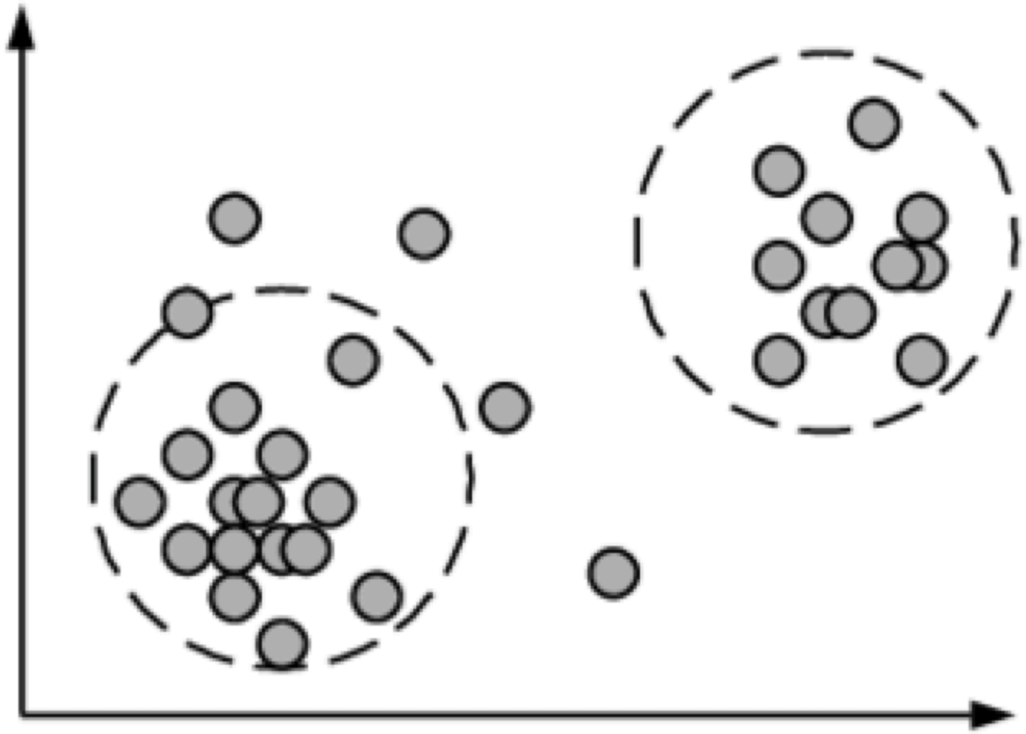
- ***regression analysis:*** linear regression, polynomial, logistic, multinomial, multilevel models, mixed-effects, nonlinear

- ***clustering:*** organizing data points into groups; used for pattern recognition, image analysis, bioinformatics, data compression

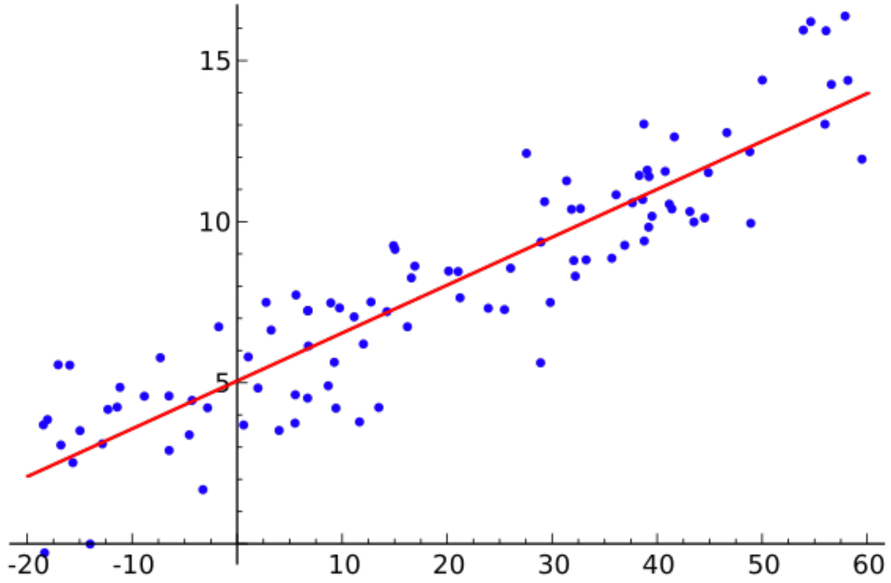
- ***classification:*** finding a mathematical mapping between features of data and known categories (supervised); binary methods, linear classifiers, neural networks, logistic regression, Bayesian methods

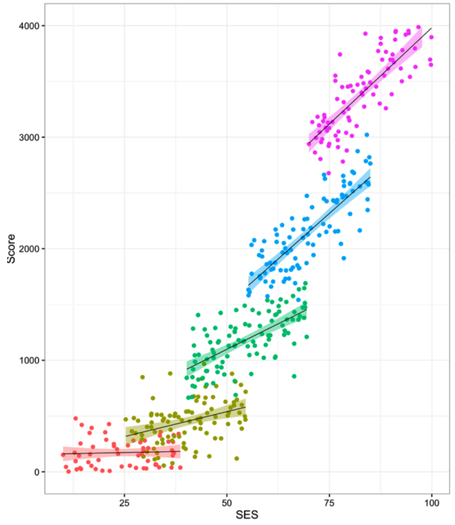
Supervised vs Unsupervised Learning

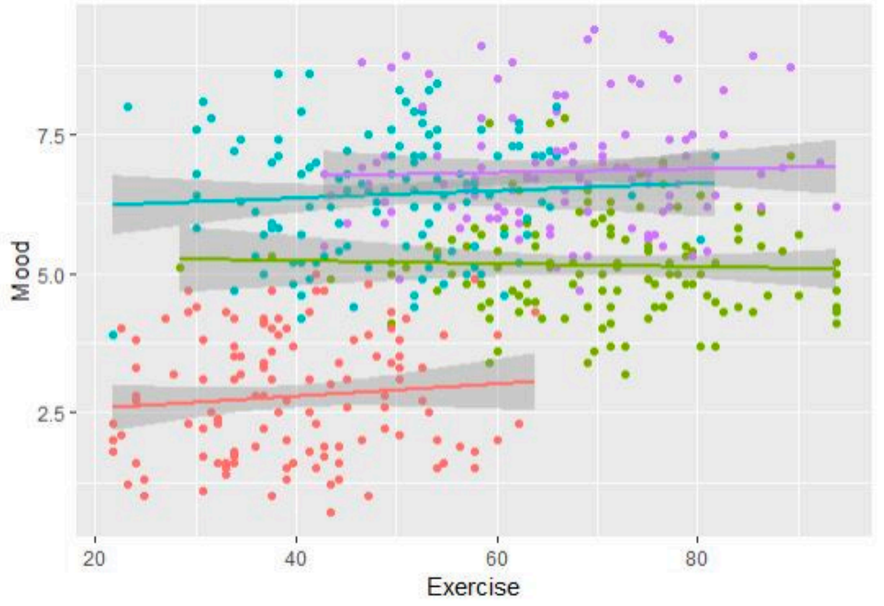


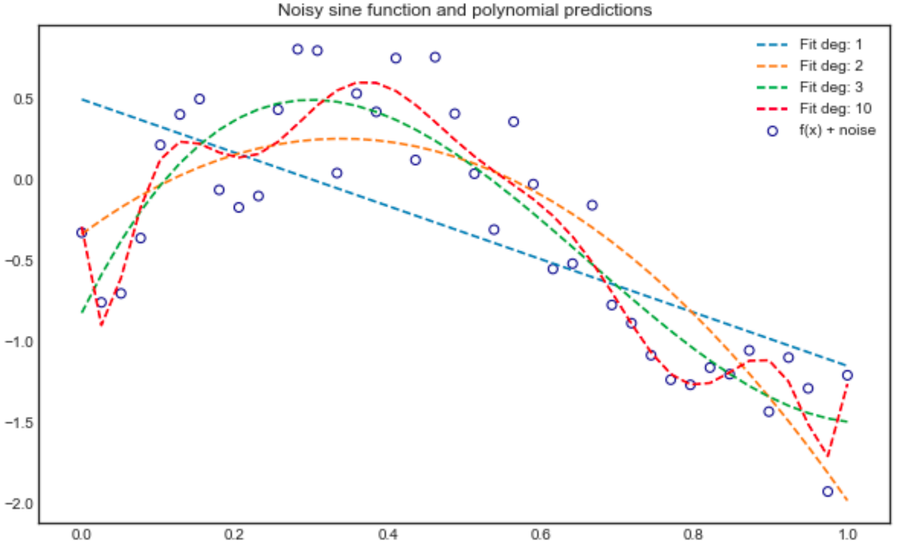


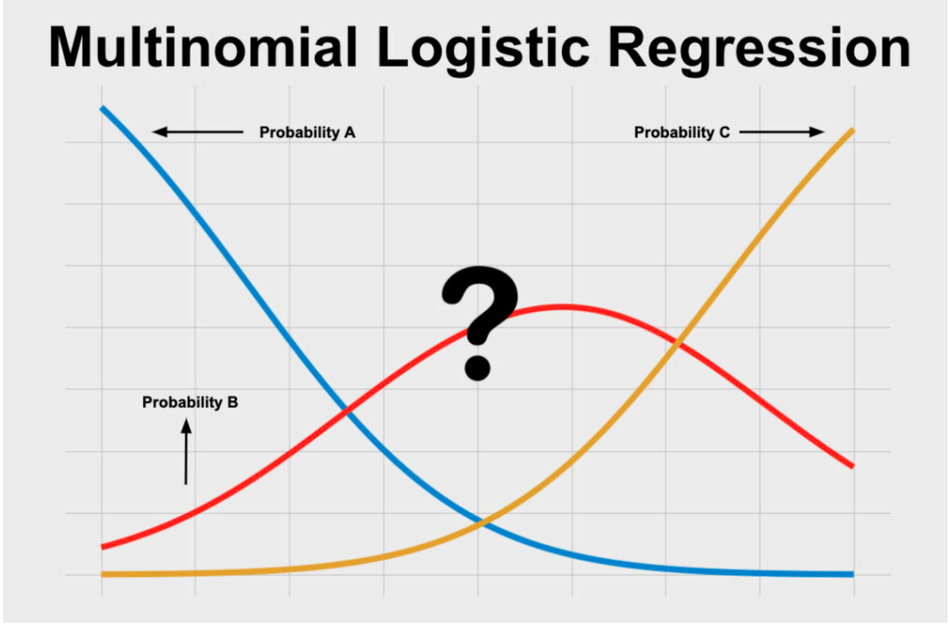
Regression analysis



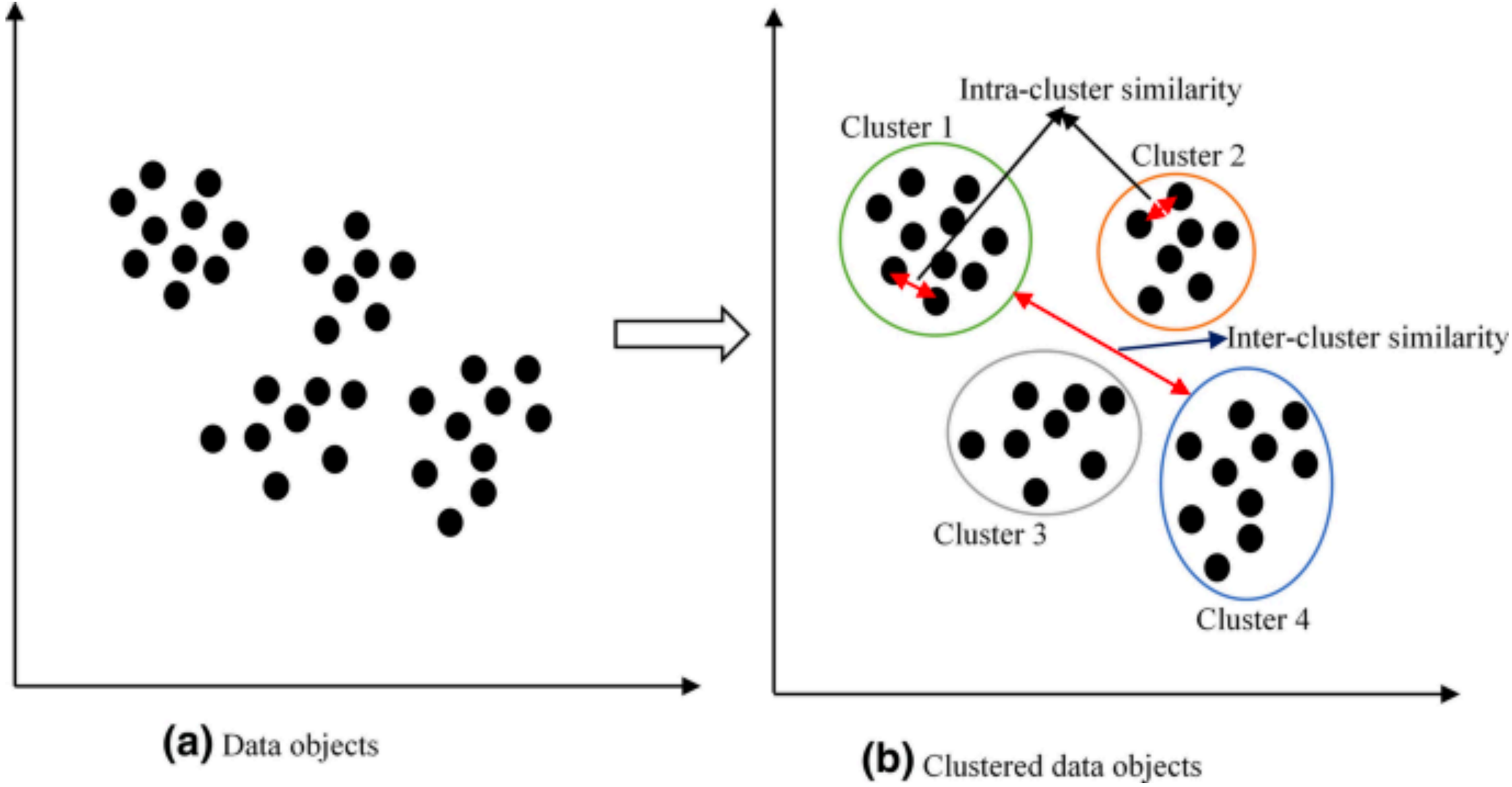








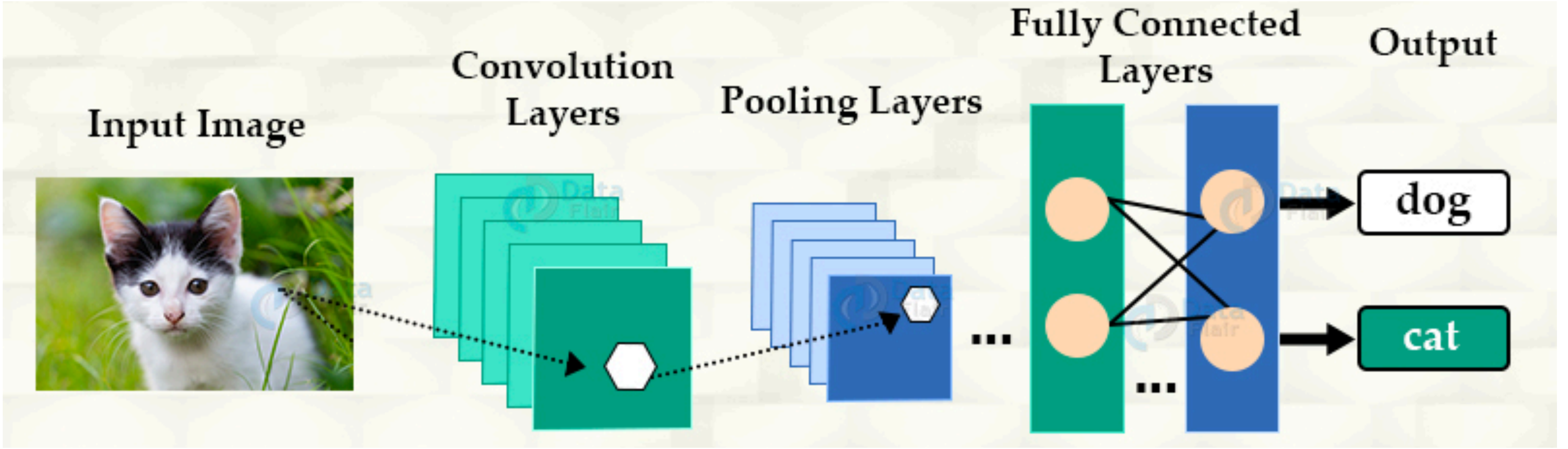
Cluster analysis

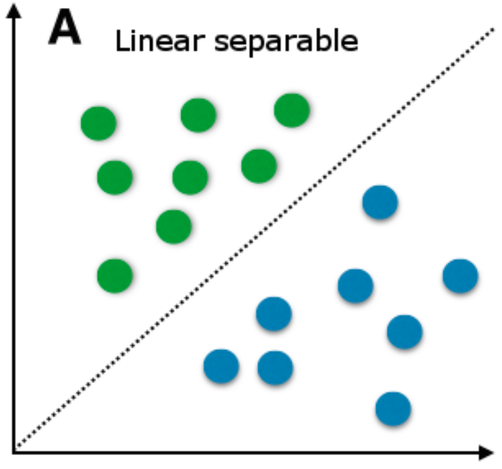


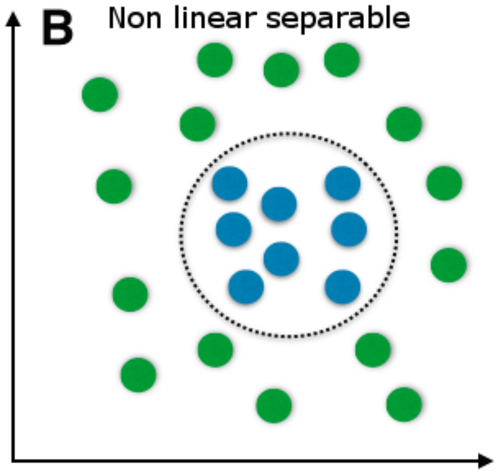
(a) Data objects

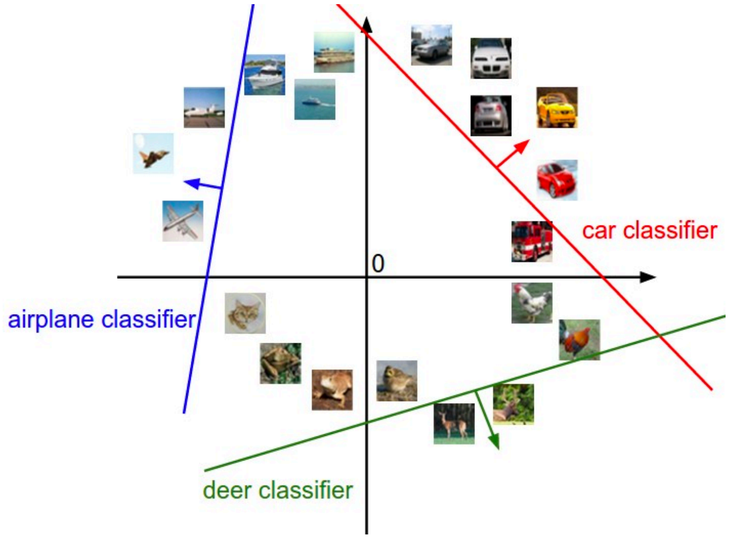
(b) Clustered data objects

Classification









Intro to Data Science Schedule

Monday July 12

9-9:30: Course Overview

9:30-10:30: Coding & Python Basics

10:30-10:45: Break

10:45-12:15: Fundamentals of Data Manipulation

12:15-1:15: Lunch

1:15-3:15: Data Visualization

3:15-3:30: Break

3:30-4:30: Ethical Issues in Data Science

4:30-5: Questions & Project Planning

5-5:30: Artificial Intelligence (AI) Institute

Tuesday July 13

9-9:30: Clustering Using K-Means

10:30-10:45: Break

10:45-12:15: Regression Methods

12:15-1:15: Lunch

1:15-4:00: Group Work on Course Project

4:00-4:30: Project Presentations and Discussion

4:30-5:00: Final Thoughts and Questions

Knowledge Check

https://forms.gle/esH9hsuZ39KnA1Yw5