Computer Science Principles

Unit 2: Data and Information

LECTURE 3: DATA REPRESENTATION (U9)

DR. ERIC CHOU

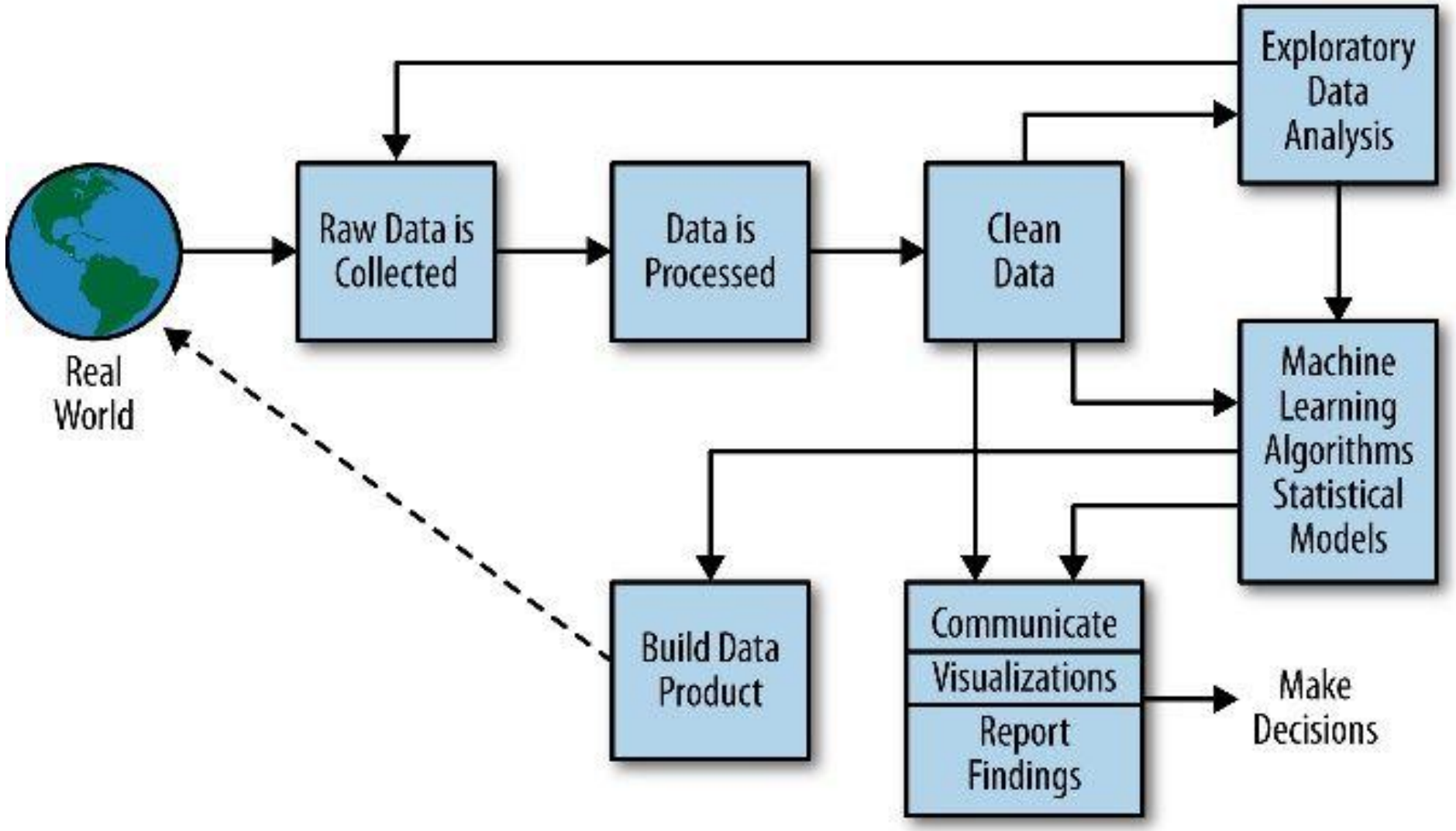
IEEE SENIOR MEMBER

Topics

- Overview of Data Science
- Big Data
- Data Cleansing
- Data Processing
- Data Visualization
- Data Base
- Data Model

Objectives

- Big data is a big deal right now, both in the field of computer science and more broadly across fields and industries.


Real World
Raw Data is Collected
Data is Processed
Clean Data
Exploratory Data Analysis
Machine Learning Algorithms Statistical Models
Build Data Product
Communicate
Visualizations
Report Findings
Make Decisions

Overview of Data Science

Extracting Information from Data

LECTURE 1

Big Data

- One of the neat things about data is that we can get information from it! By examining it closely, we can identify trends, make connections and address problems. This is true in almost every field, from science to history.

Big Data

- One or a few data points may not be enough to create a coherent conclusion—you could be dealing with an outlier, and it's difficult to see trends. With larger data sets (often very large data sets, known as **big data**), you can establish more comprehensive patterns. As the world gets bigger and more interconnected, a vast amount of data becomes more accessible and needs to be tracked.

Take a look at **this visual** of global shipping in the year 2012!

Server Farms

- To create this graphic, researchers had to track millions of shipments over months. This is just one example of the rapidly-growing need for big data processing.
- As the data sets get larger and larger, computers become a necessary tool to help us process it. They can process data faster and with less error than humans. (imagine sorting through all of that shipping data by hand...). At larger scales, you may even need multiple computers or parallel systems to process all the data involved.

Server Farms

- This demand has even led to the creation of **server farms**, areas where many large computers are housed for the purpose of meeting intense processing needs such as dealing with big data sets. Server farms are often located in large **data centers**, but they can also be stored in much smaller rooms as well.

imgIX
imgIX
imgIX
imgIX

Scalability

- When you're working with data sets in a system, you need to consider how scalable the system is. **Scalability** is the ability of a resource to adapt as the scale of the data it uses increases (or decreases). Although a scalable resource might require, for example, more servers or access points, it won't have to change the basics of how it operates. The more scalable a system is, the more data you can process and store within it.

Scalability

- In terms of data processing, you're only as good as the tools you have to work with. The more powerful your computer is, the better the data processing you'll be able to do.
- One of the ways to make processing data easier is through the use of **metadata**.

Metadata

- **Metadata** is data about data.
- It's like the packing label on a box in the mail or the tags on a piece of clothing: it gives you information about the item it's attached to. For example, the metadata for a YouTube video could include the title, creator, description, and tags for the video, as well as when it was uploaded and how large it is.

Metadata

This file contains additional information, probably added from the digital camera or scanner used to create or digitize it. If the file has been modified from its original state, some details may not fully reflect the modified image.

Camera manufacturer	Canon
Camera model	Canon EOS 30D
Author	unknown
Exposure time	1/160 sec (0.00625)
F Number	f/5.6
Date and time of data generation	20:21, 20 January 2005
Lens focal length	400 mm
Show extended details	

Metadata

- If you change or delete the metadata, it doesn't affect the data itself. For example, if you change the description of a video, you don't affect the video itself.
- Metadata is used to help find and organize data: you can use it to sort and group it. It can also provide additional information to help you use your data more effectively. For example, metadata that tells you when a video was uploaded or a post was made can help you decide whether or not the information you're looking at is outdated.

Problems Collecting and Processing Data

- Regardless of how large they are, data sets can be challenging to deal with. Fortunately, computers can help us deal with some of these issues.
- For example, data may not be uniform due to its collection process.
- Imagine you made a Google Survey to find out what class people like the most at your school. You create a form that looks like this:

Favorite Class Questionnaire

What's your favorite class?

Your answer

Submit

Problems Collecting and Processing Data

- Looking at the results, you see that a lot of people say AP Computer Science Principles is their favorite class! However, they don't all choose to say it the same way.

AP CSP
AP Comp Sci
AP Computer Science Principles
ap csp
Ap Csp
Miss. Lovelace's 5th period

Problems Collecting and Processing Data

- ...the list goes on. This lack of uniformity could make it very difficult to work with your data, especially if you're potentially looking at hundreds of entries.
- You might also run into this uniformity issue if you're compiling data from many different resources, where formatting standards may be a little different. For example, let's say you're working with a friend to track what time of day people find the most productive, but you're writing down results using 12 hour time while they're using 24 hour time. (Probably just to be difficult.)

Problems Collecting and Processing Data

- The way that computers deal with this is through a process known as cleaning data. This process makes data uniform by eliminating all of these differences. **Cleaning data** can also help flag or remove invalid and incomplete data

Data Basics

Data sets may be biased for a variety of reasons.

Taking the favorite class questionnaire as an example, there are many ways in which bias could sneak in:

- People have to take the initiative to fill out this form, which means they're more likely to have strong opinions about the topic when compared to the rest of the population. People who don't really have a favorite class or don't care about school in general won't be all that represented in the study, skewing results for your school overall.
- This survey just looks at one school, which could also bias results. Your AP Comp Sci teacher might just be a really good teacher, for example, so people might like the class for the teacher rather than the subject.

Data Basics

- The context of your survey might also impact results. The results you get if you post this survey in your AP Comp Sci class could be completely different from the results you'd get if you posted it in a Basketball group chat.
- Data can also be biased on societal lines such as race and gender.

Just collecting more data won't make this problem go away by itself; you need to identify potential biases in your data and take steps to correct them, such as surveying people from different schools or classes.

Overview of Data Science

Using Programs with Data

LECTURE 2

Data Mining

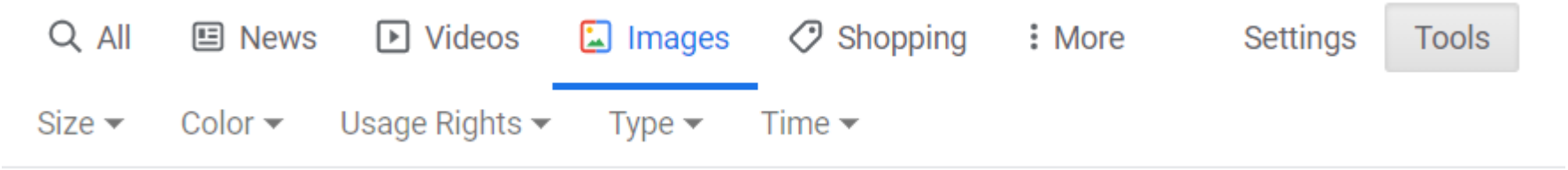
- With the current demand for data processing, it's no wonder that so many computer programs exist for that very purpose. In fact, there's even a name for it: the process of examining very large data sets to find useful information, such as patterns or relationships, is known as **data mining**.
- A common example is a spreadsheet program, such as Google Sheets or Microsoft Excel. You can use these programs to record, modify, and organize data. If you're using numbers, you can write equations and perform operations on your data as well.

Text Analysis

- You can also process text data using text analysis (or text mining) tools. **Text analysis** looks for patterns within a written piece (anywhere in length from a clause to a novel and beyond) to categorize or classify it. If you've ever had a program tell you what the tone of your writing was, you've seen text analysis at work. Text analysis can be used to sort product reviews, detect trends in public opinion and **identify anonymous authors.**
- Data processing programs can also allow you to make tables and diagrams, such as line or bar graphs, to visualize your data. Creating visualizations of data allows you to convey what the data means and to make trends apparent. It's much easier to see positive or negative trends from a line chart, for instance, than when the data's sitting in a table. This is especially true when a lot of data is involved.

Search Tool

- Other examples of data processing programs include **search tools**, like the ones that Google uses for images:



- You can use these to make finding information easier and faster and to specify what it is that you're looking for. For example, if you want an image in a certain color for a mood-board, you can find it using the color filter. If you want images taken before or after a certain date, you can use the time filter.

Search Tool

- Different engines have different search tools based on what the search engine is used for. The search tools for an online academic journal, for instance, differ from the search tools that Google Images uses.

Data Filtering Capabilities

- Some programs also have **data filtering capabilities**, which means that they can create and extract different subsets of data for users to work with. These subsets can be based on time (like only looking at results from the winter) or value (like only looking at values below 30 or only positive values).

Transforming Data

- One of the cool things that programs can do with data is to **transform it**! This is when you edit or modify data in some way to extract more information from it.

Data Transformation Examples:

- Modifying every element of a data set. This can be an arithmetic modification, although it isn't necessarily one.
  - Ex. Multiplying each number by some constant value (Like if you wanted to convert a list of measurements from liters to millilitres.)
  - Another non-arithmetic example is adding a grade level or class rank to a list of student records.
- Filtering a data set by category, as mentioned above.
  - Besides time or value, data sets can also be filtered by *quality*, such as which extracurricular activities a group of students are in.

Data Transformation Examples:

- Combining or comparing data in some way.
  - Ex. Comparing the average SAT score of students going to all the colleges in one state and combining that data with average scores from other states.
- Creating data visualization tools.
  - Ex. graphs, charts, and word-bubbles.

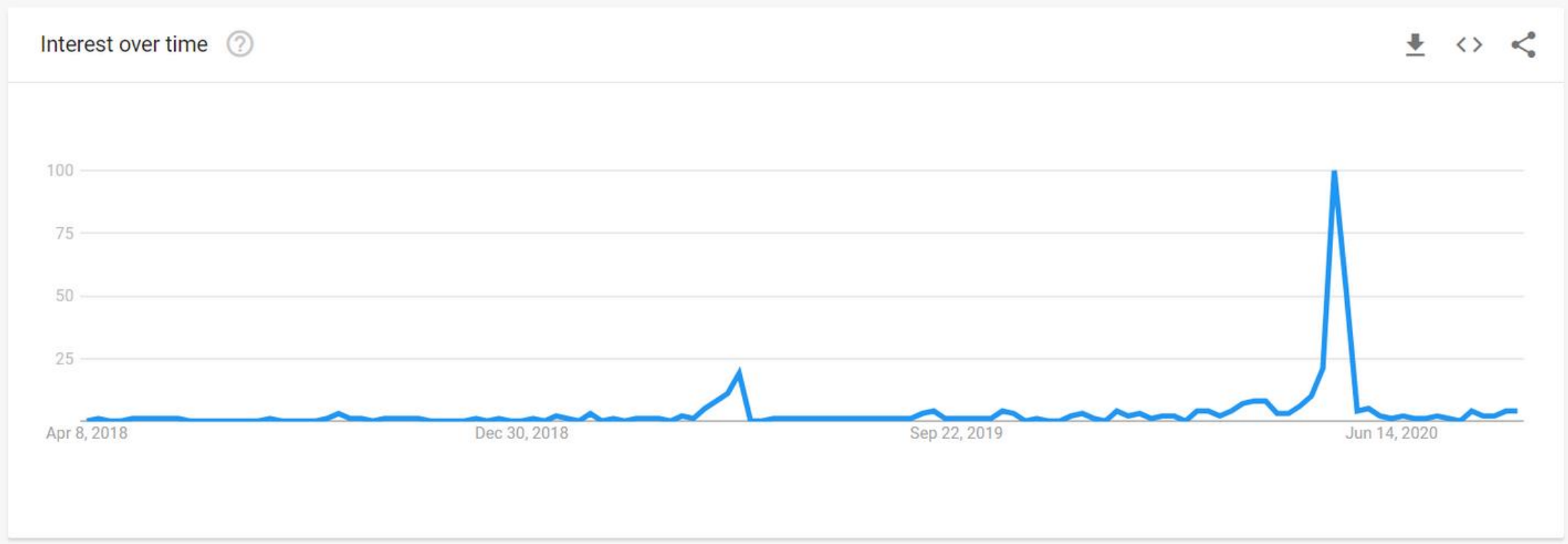
Data Transformation

- These tools are often used in an iterative and interactive process by users. You get to choose what filtering tools you want to use or what subsets you want to look at. You can also run data through data processing programs multiple times, depending on what information you want to look for. For example, you can look at data by the date it was collected, then sort it again by the location where it was collected from.
- Manipulating data by combining, clustering or classifying it can bring out new information and patterns previously unseen in the raw data, making it a helpful tool for data analysis.

Data Analysis Discoveries

Some of the things we can discover by analyzing data are:

- Patterns
  - What repeats again and again?
  - Ex: Does a product sell particularly well over one season or in one area over many years?
- Trends
  - Is there a steadily rising trend? A falling trend? Any fluctuations, or variances, in the data?
  - Ex: Here's a graph of the interest trends over time for Fiveable since its founding in 2018. (Note the two spikes around AP exam season.)


Interest over time
100
75
50
25
Apr 8, 2018
Dec 30, 2018
Sep 22, 2019
Jun 14, 2020

Data Analysis Discoveries

- Correlations, or relationships
  - Ex: Is there a relationship between extracurricular choice and favorite subject? Time of day and performance on tests? Driving experience and number of traffic tickets?
  - Is there a positive correlation? A negative one? None at all?
  - Remember that **correlation does not equal causation**! Just because two things correlate with each other doesn't mean that one caused the other.
- **Outliers**
  - Are there any outliers? Why is there an outlier?

In conclusion...

- That's Big Idea 2: Data for you! Our next Big Idea Guide is a crash course to algorithms and (basic) programming.

What is Big Data?

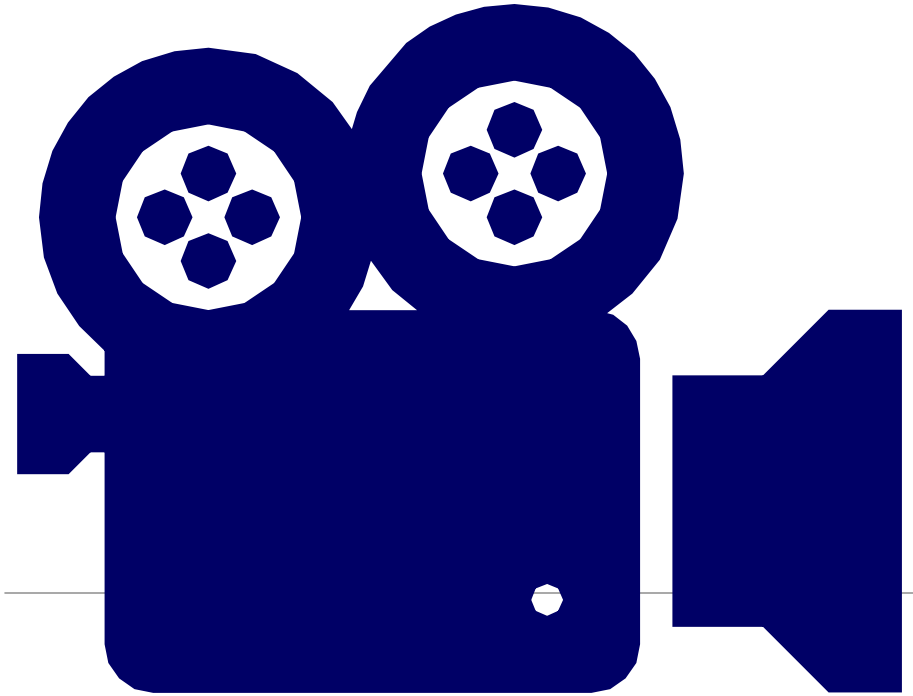
LECTURE 3

Vocabulary

- **Big Data** - a broad term for datasets so large or complex that traditional data processing applications are inadequate.
- **Moore's Law** - a predication made by Gordon Moore in 1965 that computing power will double every 1.5-2 years, it has remained more or less true ever since.

What is Big Data?

- Where does Big Data come from, what makes it “big,” and how people use big data to solve problems?
- How much of your lives are “datafield” or could be? Who knows what about you?



You should Learn to Program

VIDEO

Video

Watch Video (15 MINUTES):

Video: Big data is better data - TED talk - Video

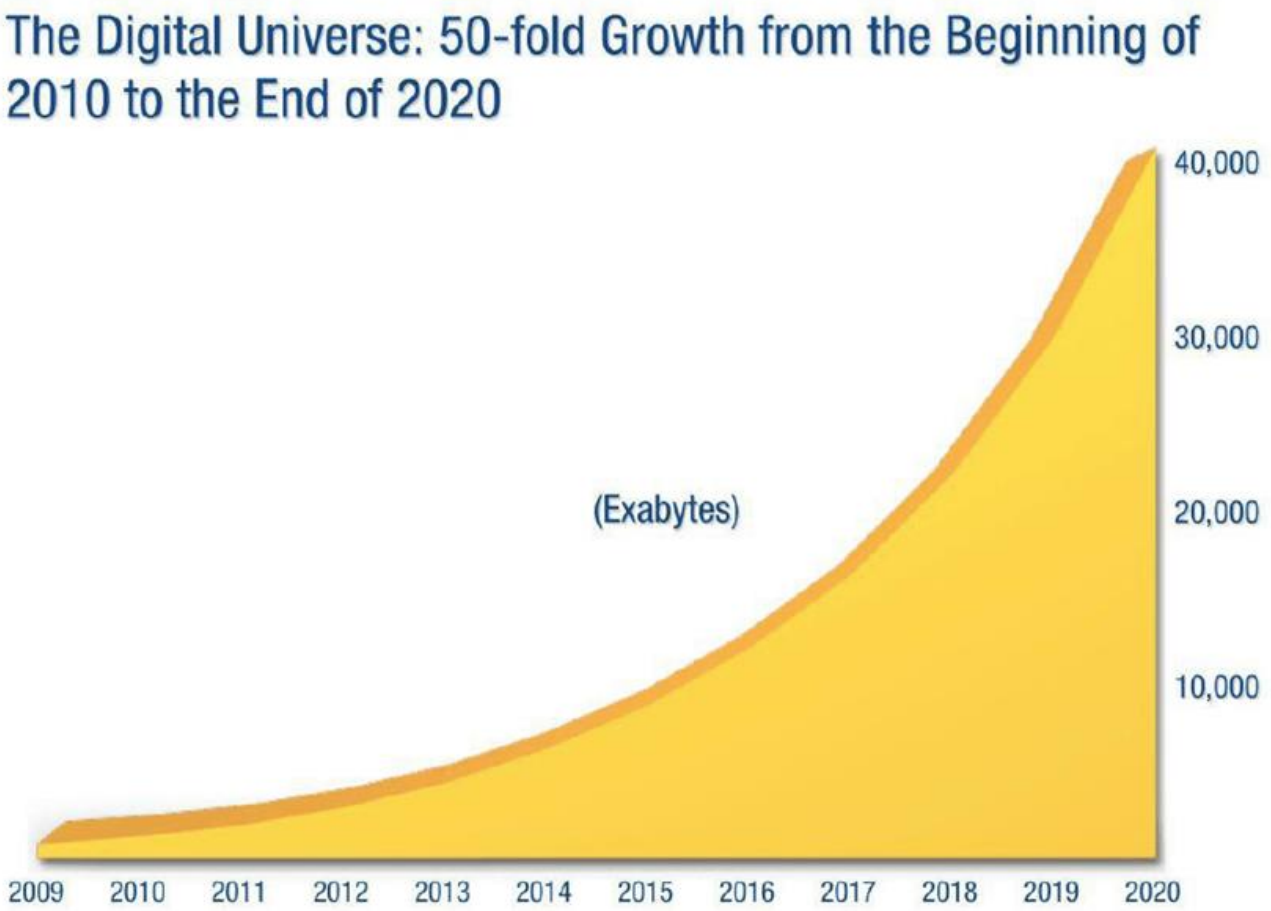
What is Big Data?

- Big data means different things, at different times, to different people.
- It can mean devices that are constantly collecting data.
- It can mean digitizing data that’s been around for a long time (e.g., every book ever written).
- It can mean machine learning and artificial intelligence.

Code Studio

- The IDC's "Digital Universe" is described as "a measure of all the digital data created, replicated, and consumed in a single year." source:

http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020



This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Big Data

- Part of what contributes to data being "big" is the sheer growth of the amount of data in the world.
- Let’s have a look at a graph that shows us just how large big data is.
- As you can see from the chart, the amount of data flying around is growing exponentially, doubling every two years or so. Here’s a way to think about how fast this is: The world will produce as much digital data over the next 2 years, as currently existed in all of humanity prior to that. And it will do the same the 2 years after that. And so on. That’s a lot!

BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or instgrammed?

Key Technologies that enable Big Data Analytics for Business

- Predictive Analytics
- NoSQL Databases
- Knowledge Discovery Tools
- Stream Analytics
- In-memory Data Fabric
- Distributed Storage
- Data Virtualization
- Data Integration
- Data Preprocessing
- Data Quality

Vertical Apps
PREDICTIVE POLICING, bloomreach, MYRRIX

Log Data Apps
splunk>, loggly, sumologic

Ad/Media Apps
rocketfuel, collective[i], bluefin, Recorded Future, LuckySort, Media Science, TURN, DataXu

Business Intelligence
ORACLE | Hyperion, SAP Business Objects, RJMetrics, Microsoft Business Intelligence, IBM COGNOS, birst, Autonomy, MicroStrategy, QlikView, bime, Chart.io, GoodData

Analytics and Visualization
tableau, Palantir, OPERA, metaLayer, METAMARKETS, dataspora, TERADATA ASTER, centrifuge, sas, TIBCO, KARMASPHERE, panopticon, pentaho, Datameer, platfora, ClearStory, CIRRO, alteryx, visual.ly, AYATA

Data As A Service
factual, kaggle, knoema, GNIP, DATASIFT, INRIX, LexisNexis, LOQATE

Analytics Infrastructure
Hortonworks, VERTICA, MAPR, cloudera, INFOBRIGHT, PARACCEL, EMC², GREENPLUM, NETEZZA, kognitio, DATASTAX, EXASOL

Operational Infrastructure
COUCHBASE, 10gen, TERADATA, HADAPT, TERRACOTTA, VoltDB, MarkLogic, INFORMATICA

Infrastructure As A Service
amazon web services, Windows Azure, infochimps, Google BigQuery

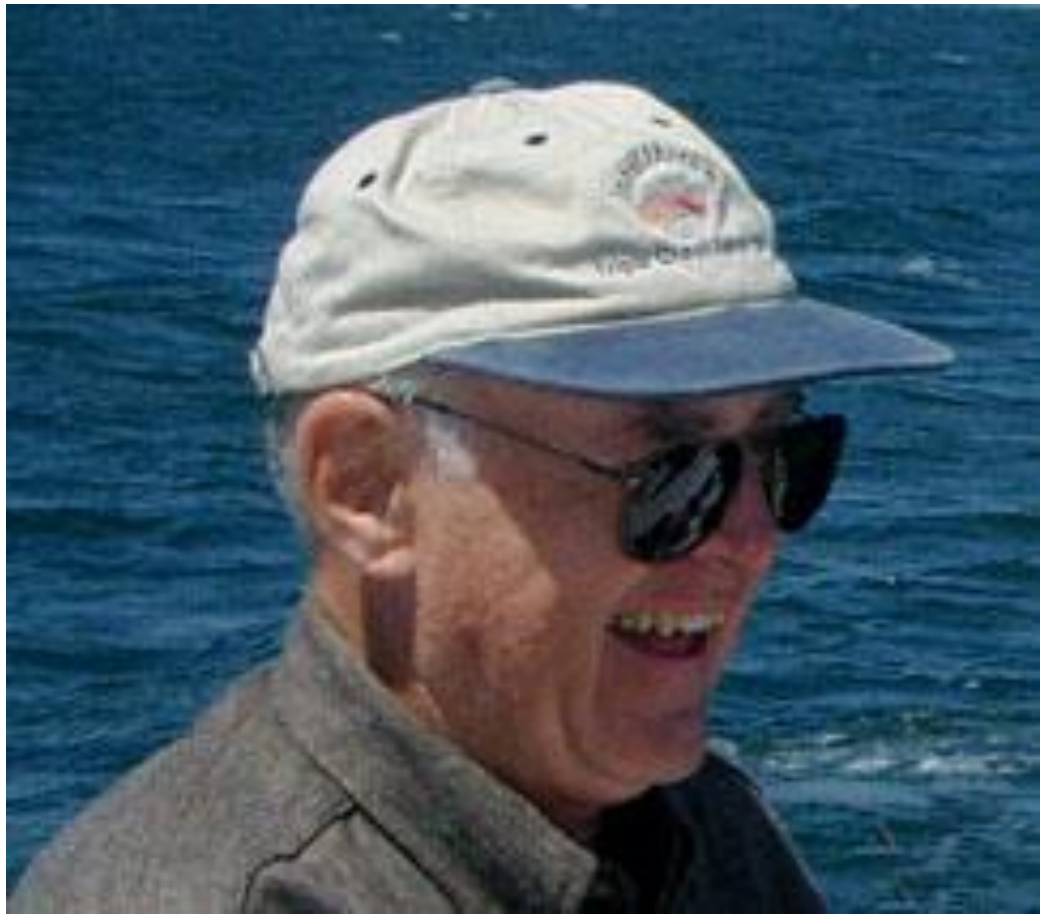
Structured Databases
ORACLE, MySQL, SQL Server, PostgreSQL, IBM DB2, SYBASE, memsql

Technologies
hadoop, hadoop MapReduce, mahout, APACHE HBASE, Cassandra

Big Data
Cloud Computing

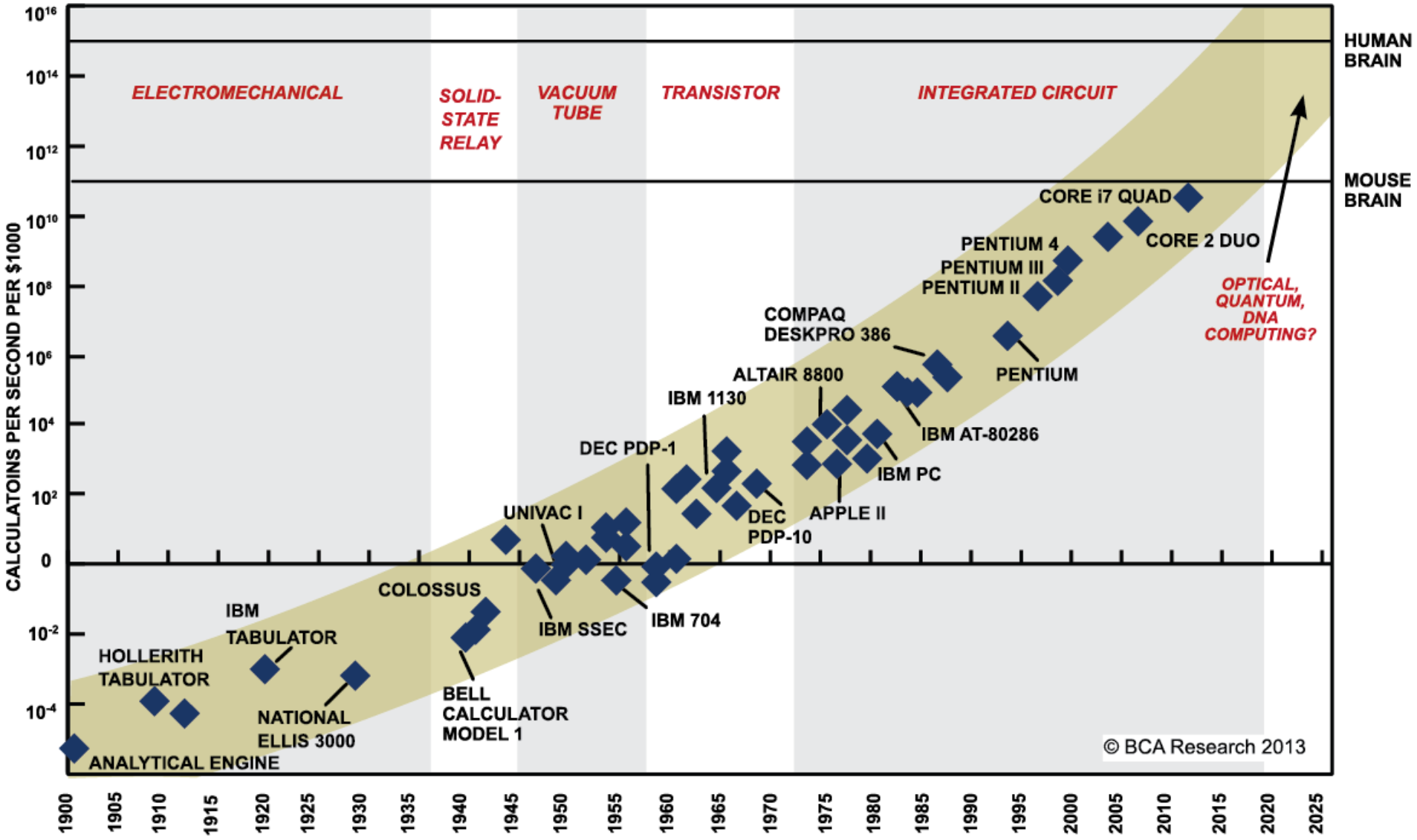
Moore’s Law

- Moore's law is actually about computing power, not data, but data growth seems to following the same trend.
- So far, computing power/capacity seems to double every 1.5-2 years...
- That means it grows exponentially...
  Exponential growth is hard for humans to fathom...
- Yet we need to plan for it.



Gordon Earle Moore (born January 3, 1929) is an American businessman, engineer, and the co-founder and chairman emeritus of Intel Corporation. He is also the author of Moore's law.

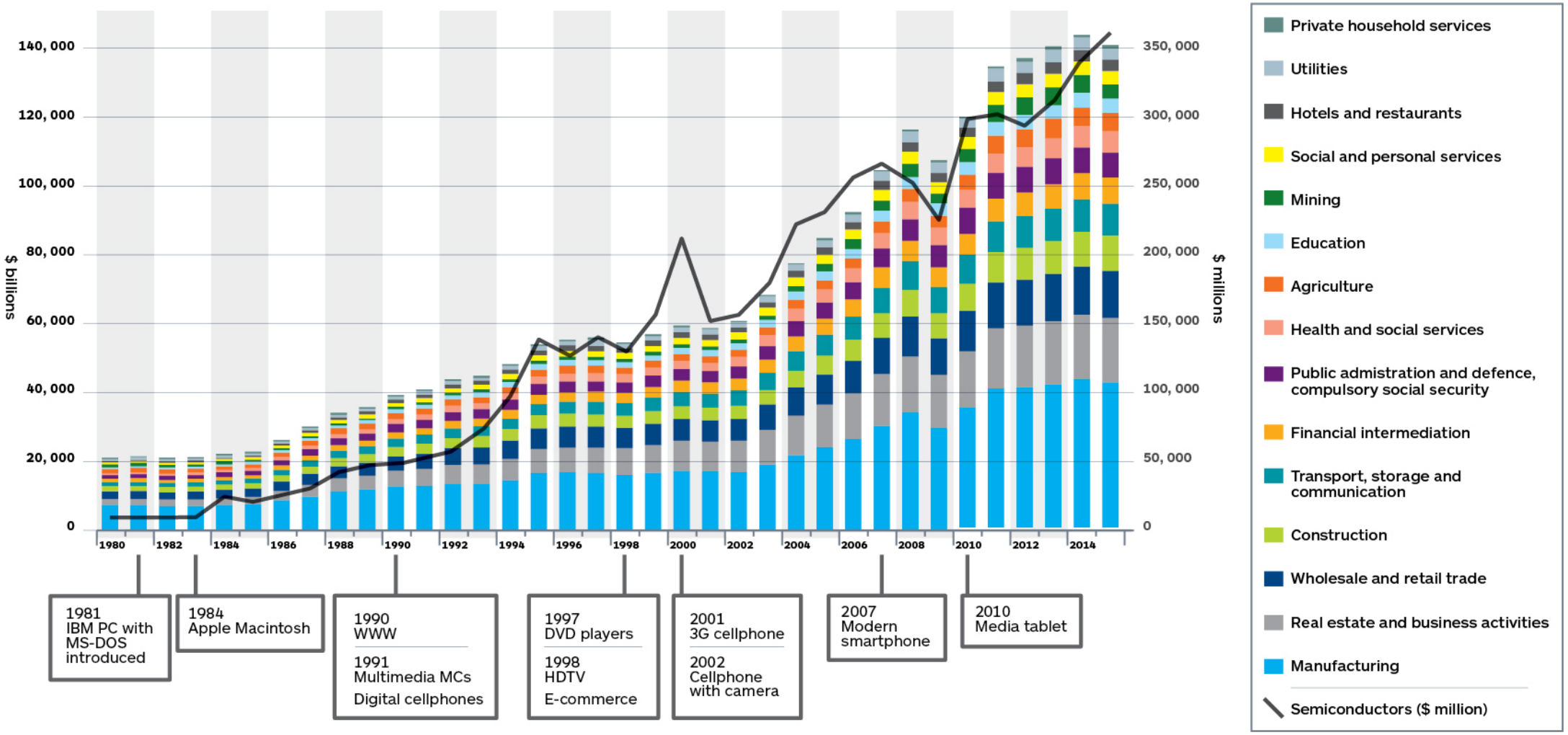
As of April 2019, Moore's net worth is reported to be $10.4 billion.



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Semiconductor growth tracks global output

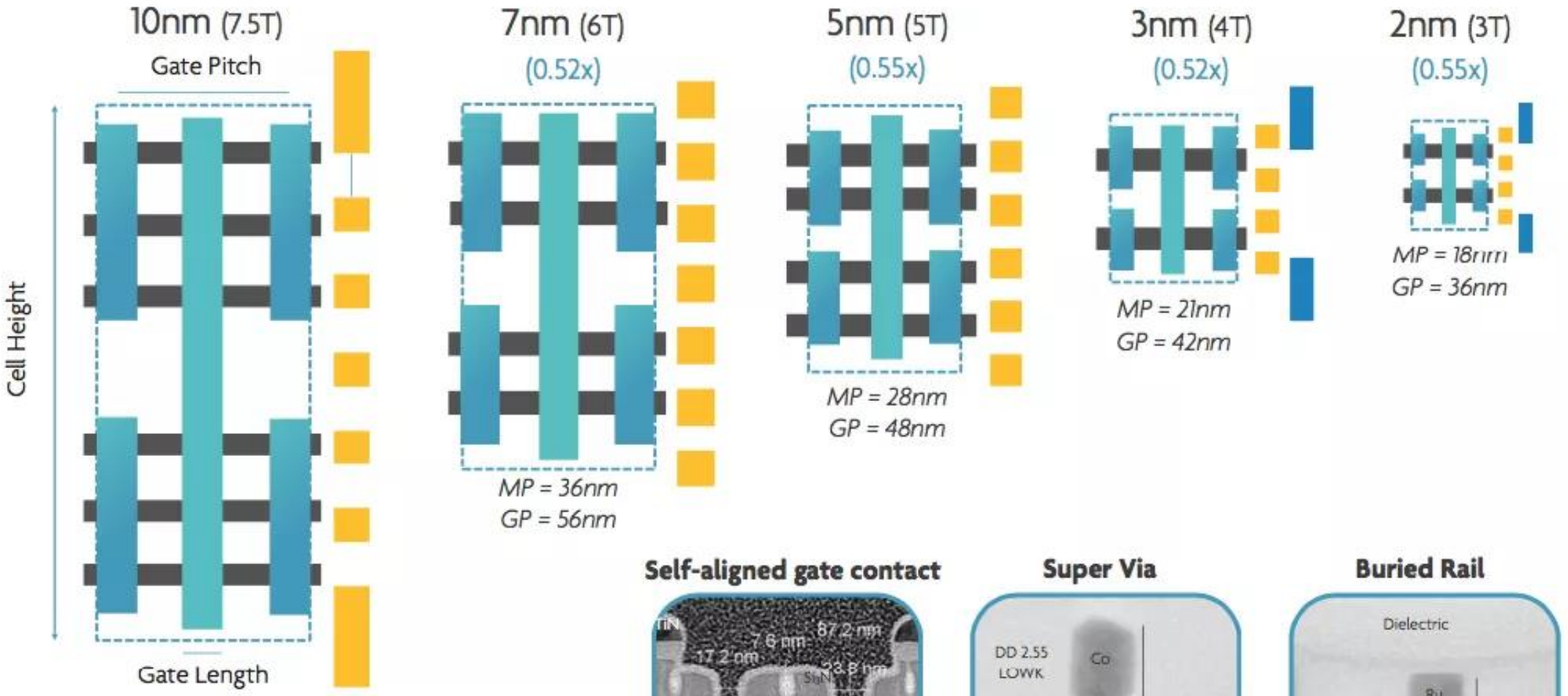
Worldwide gross output by industry ($ in billions, left axis) and semiconductor sales ($ in millions, right axis)



CELEBRATING THE 30TH ANNIVERSARY OF MOORE'S LAW

Source: IHS

eC Learning Channel



Cell library scaling enabled by scaling boosters

IMEC VIEW OF TRANSISTOR ROADMAP

Nanowire/sheet

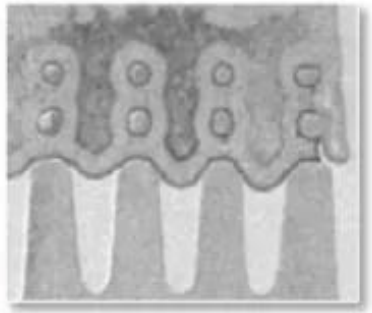
CFET (n-on-p)

VFET

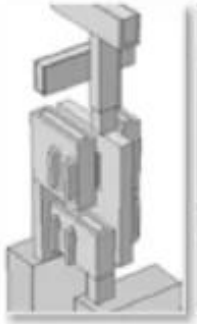
6nm	5nm	4nm	3nm	2nm
2020	2021	2022	2023	2024
48nm 36nm 6.5T	42nm 32nm 6.5T	42nm 32nm 6.5T	36-42nm 21-24nm 5.5T-4.5T	36-42nm 21-24nm 4.5T-3T

Gate Pitch
Metal Pitch
Tracks

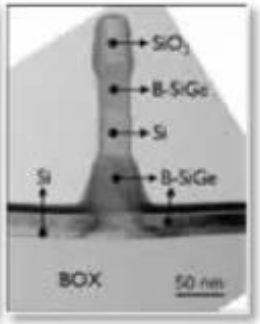
21nm finfet pitch



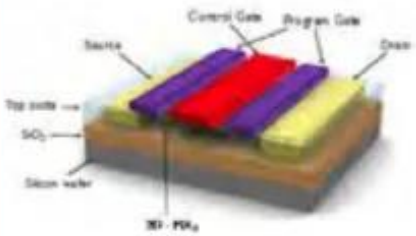
Nanowire FET



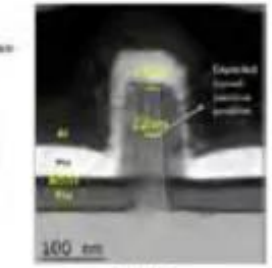
CFET



VFET

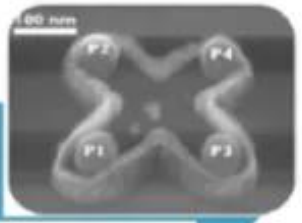


Ambipolar



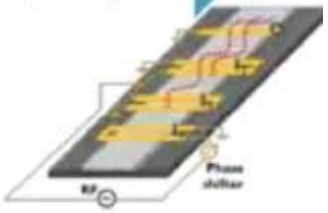
TFET

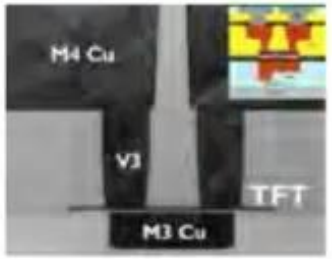
2D MX2 FET



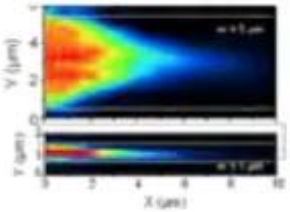
Spin-Torque Majority Gate

CNTFET





IGZO



Spin Wave Majority Gate

imec

eC Learning Channel

Activity

Distribute: Activity Guide - Big Data Sleuth Card - Activity Guide.

Directions:

- With a partner, select one of the tools in the list to the right.
- Determine what the tool is showing.
- Find the source of the data it allows you to explore.
- Complete the table below.

Websites

1 Web archive
http://www.archive.org

2 Measure of America
http://www.measureofamerica.org/maps/

3 Wind Sensor network
http://earth.nullschool.net/

4 Twitter sentiment
https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

5 Alternative Fuel Locator
http://www.afdc.energy.gov/locator/stations/

Website Name
What is this website potentially useful for? What kinds of problems could the provided information be used to solve?
Is the provided visualization useful? Does it provide insight into the data? How does it help you look at a lot of information at once? How could it improve?
Where is the data coming from? Check for "About", "Download", or "API". You may also need to do a web search. • Is the data from one source or many? • Is it static or live? • Is the source reputable? Why or why not? • Add a link to the raw data if you can find one.
Do you consider this "big" data? Explain your reasoning.

Student tasks are to follow the resource and answer prompts related to:

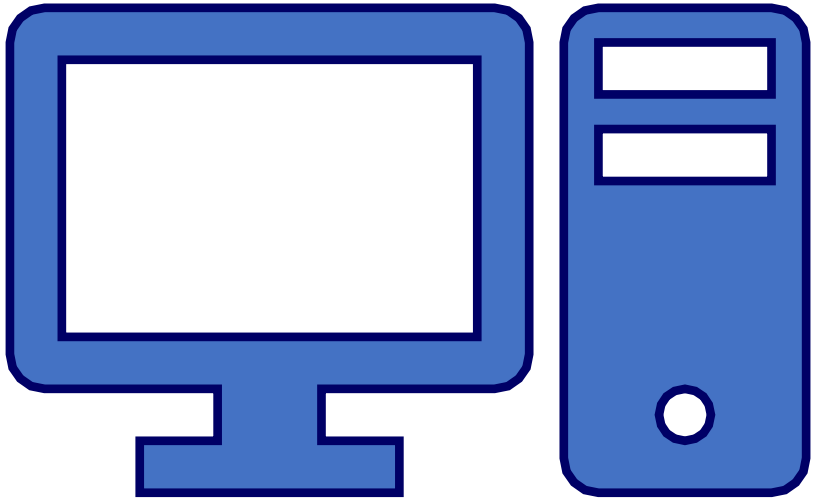
- the visualization tool provided
- the original source of the data
- evaluating the usefulness of both the data and the visualization.

Definition of Big Data

Here is a general-purpose definition of Big Data (taken from Wikipedia: Big Data):

> **“Big data is a broad term for datasets so large or complex that traditional data processing applications are inadequate."**

The fact that big data is increasingly important across industries reflects rapid changes in how much data we're collecting, and the ways we're using it.

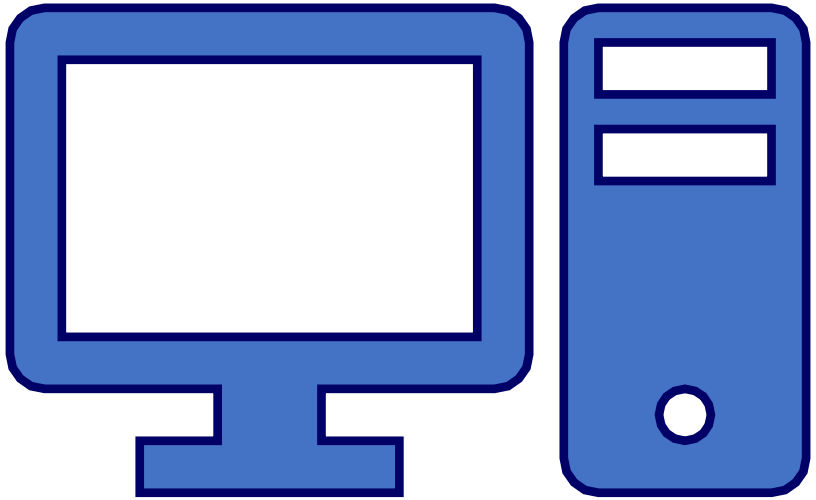


One Column

APP LAB

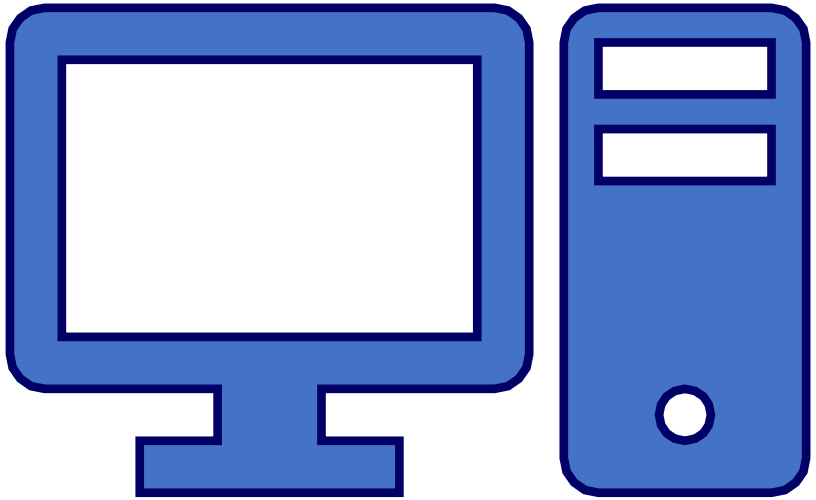
App Lab

- Reading a Bar Chart
- Making Bar Charts
- Reading a Histogram



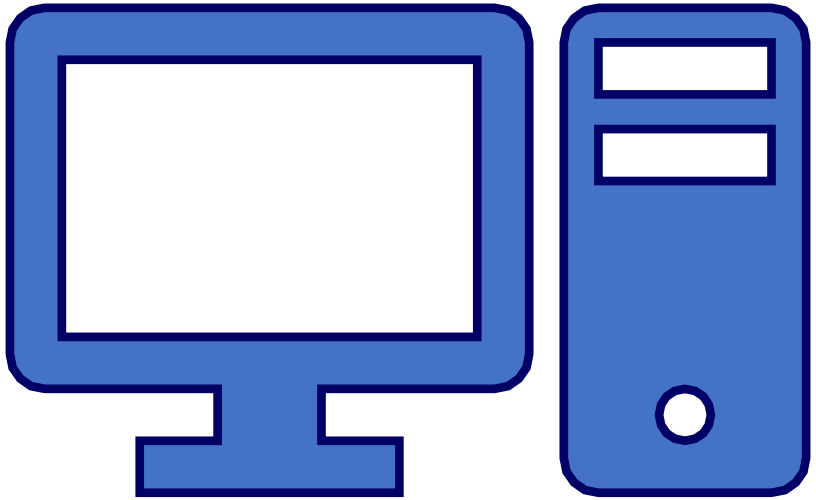
Filtering and Cleaning Data

APP LAB



Two Column

APP LAB



Big, Open, and Crowdsourced Data

APP LAB

Finding Trends with Visualizations

LECTURE 4

Vocabulary

- **Google Trends** - a tool which visualizes data taken from Google search histories all around the world from the past several years.
- **Digital Divide:** an economic and social inequality with regard to access to, use of, or impact of information and communication technologies

Objectives

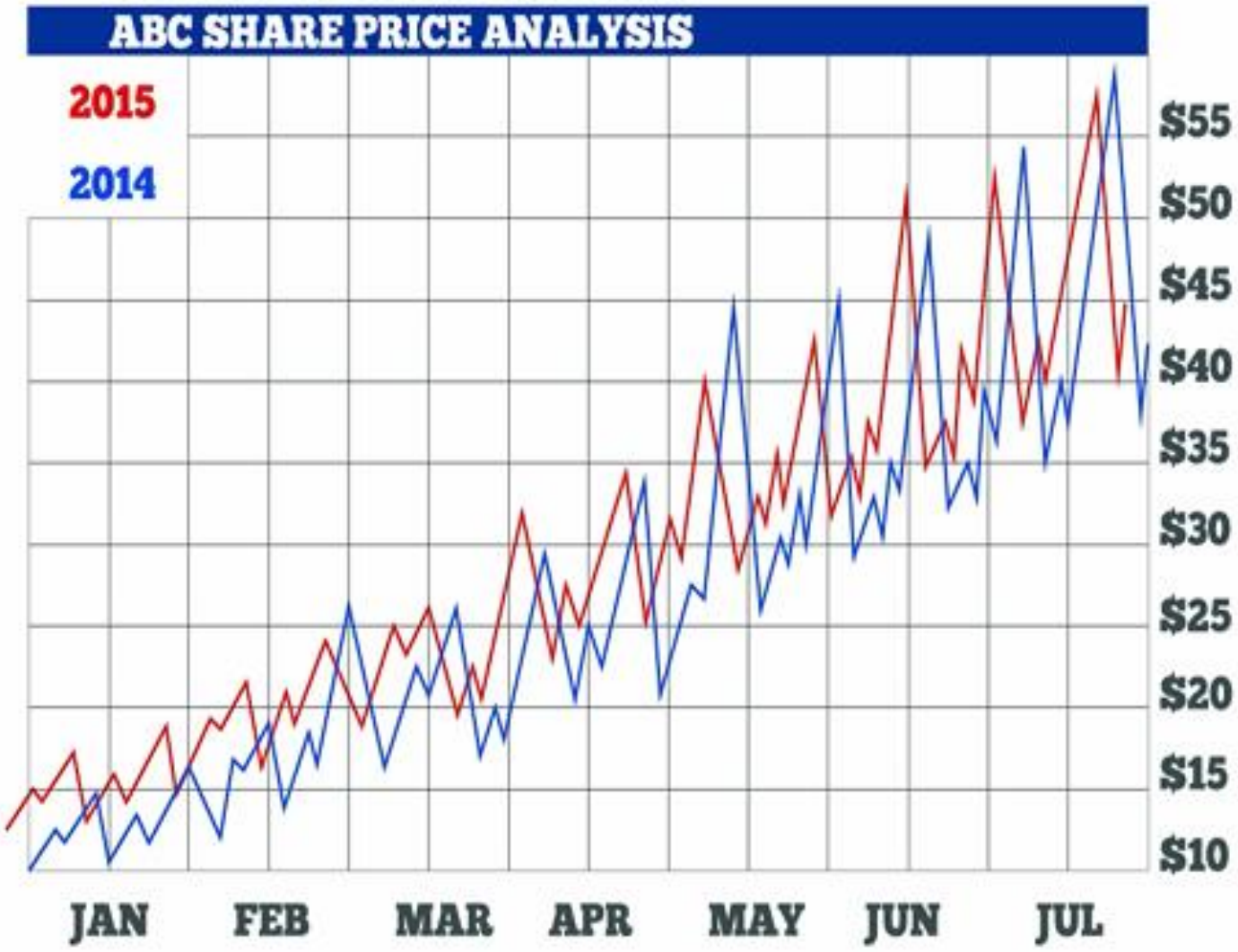
- Use Google Trends to identify and explore connections and patterns within a data visualization.
- Accurately describe what a data visualization of a trend is showing.
- Provide plausible explanations of trends and patterns observed within a **data visualization**.

Activity Guide - Exploring Trends

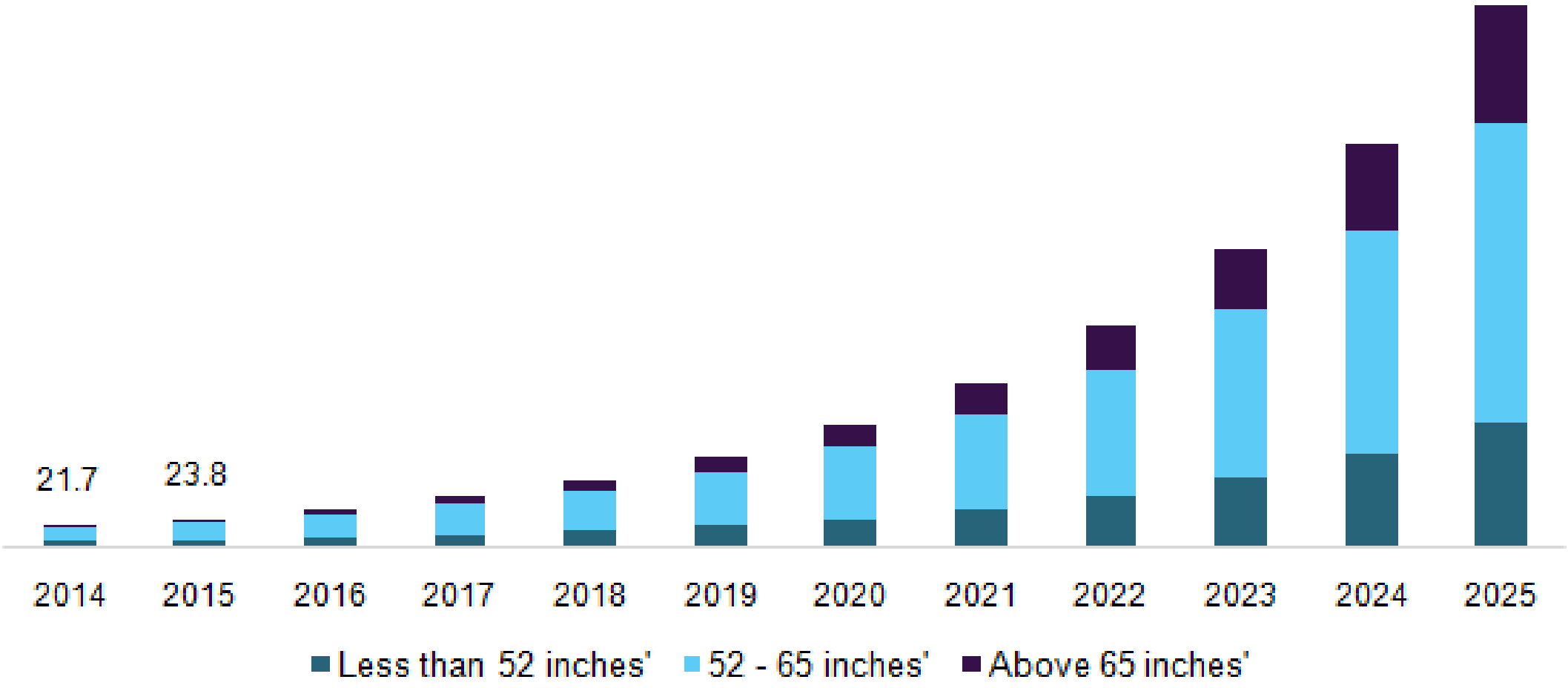
What's a Trend?

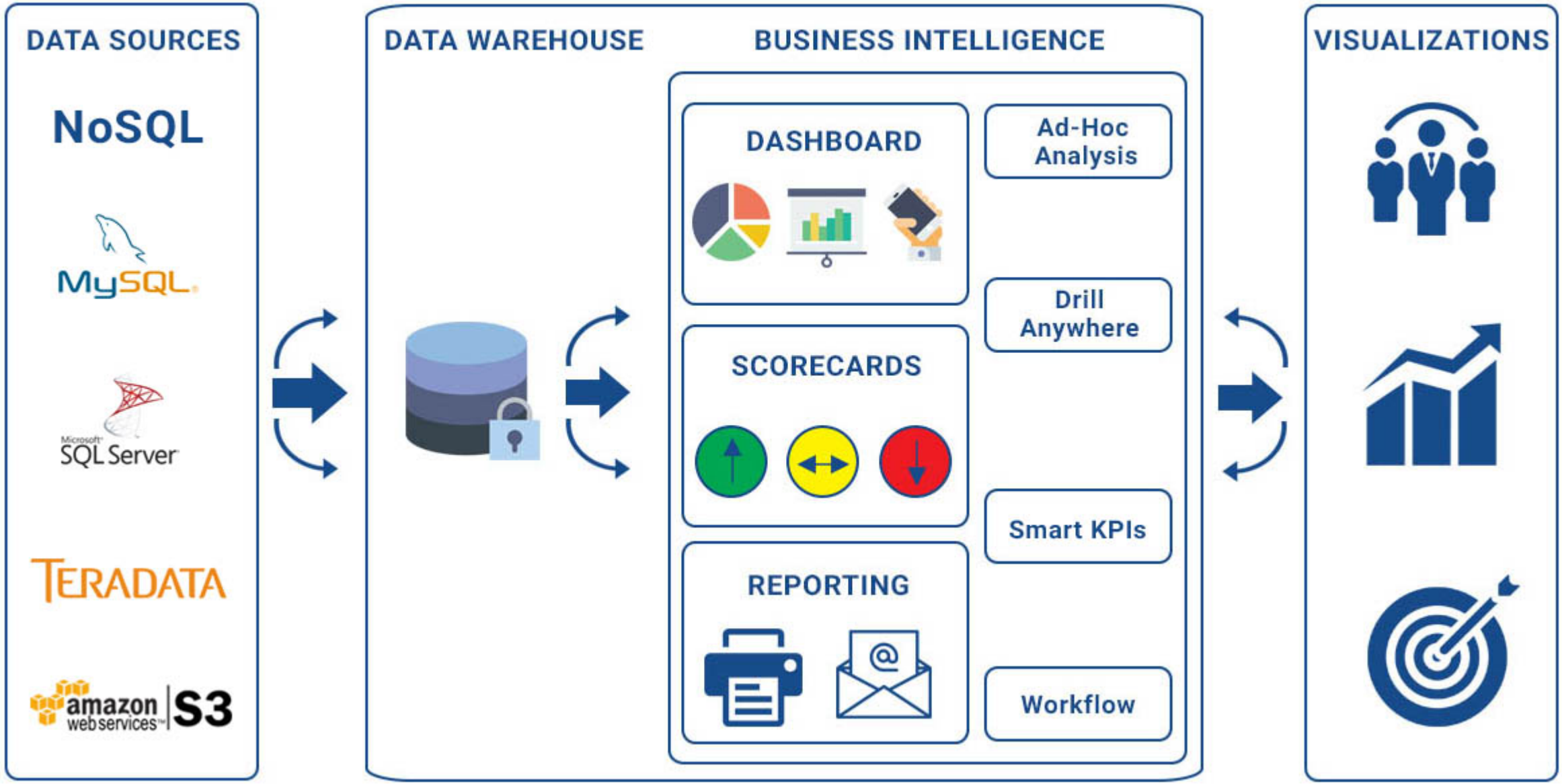
- When you post information to a **social network**, **watch a video online**, or simply **search for information** on a search engine, some of that data is collected, and you reveal what topics are currently on your mind.
- When a topic is quickly growing in popularity, it is often said to be **trending**, but there are many different trends or patterns we might find in this data, including historical trends. These patterns might help us to identify, understand, and predict how our world is changing.

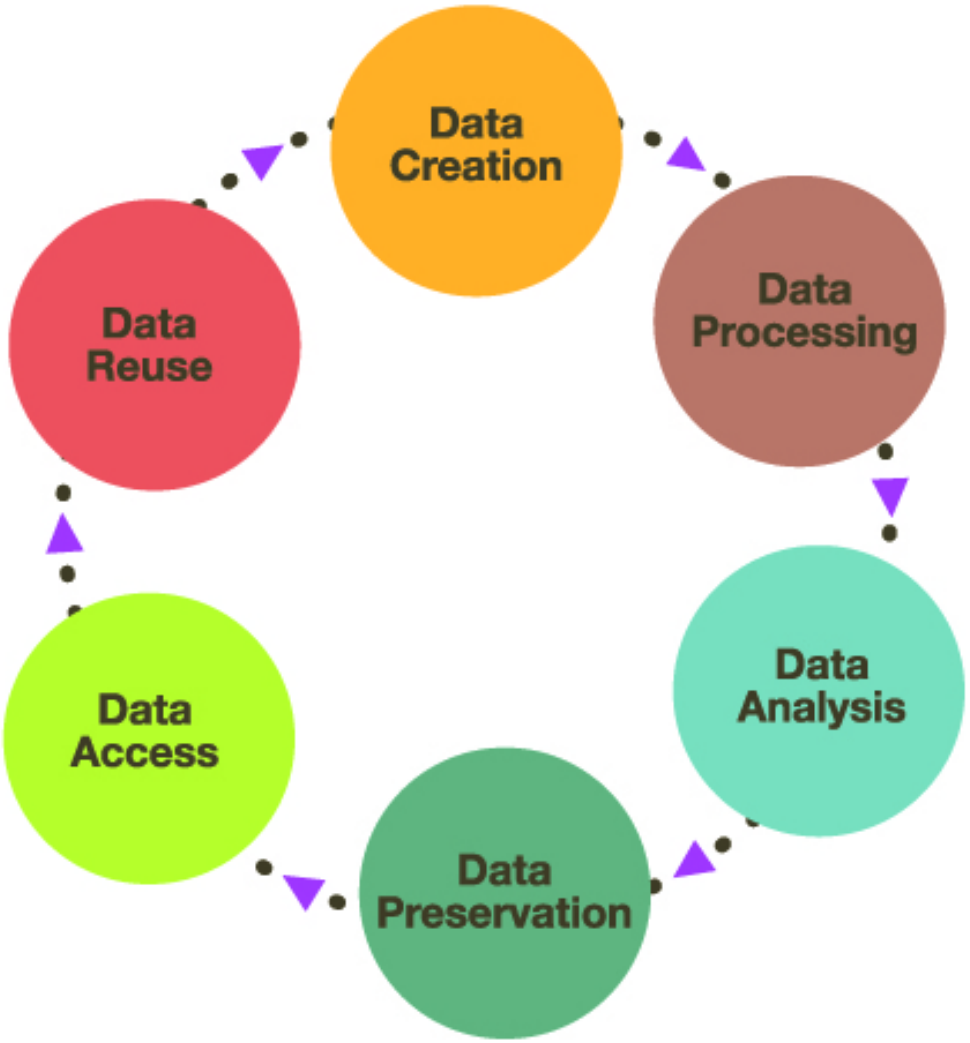
100,000
80,000
60,000
40,000
20,000
0
563
563
83,385
83,386
549
548
83,169
83,170
99,470
99,470
521
521
23,730
23,730
Trend Reporting
1.0M
0.5M
0.0M
June, 2016
July, 2016
August, 2016
September, 2016
October, 2016
November, 2016
December, 2016
January, 2017
February, 2017
March, 2017
April, 2017
May, 2017


ABC SHARE PRICE ANALYSIS
2015
2014
$55
$50
$45
$40
$35
$30
$25
$20
$15
$10
JAN
FEB
MAR
APR
MAY
JUN
JUL

HDTV Trends


21.7
23.8
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Less than 52 inches'
52 - 65 inches'
Above 65 inches'


DATA SOURCES
NoSQL
MySQL
Microsoft SQL Server
TERADATA
amazon web services S3
DATA WAREHOUSE
BUSINESS INTELLIGENCE
DASHBOARD
SCORECARDS
REPORTING
Ad-Hoc Analysis
Drill Anywhere
Smart KPIs
Workflow
VISUALIZATIONS


Data Creation
Data Processing
Data Analysis
Data Preservation
Data Access
Data Reuse

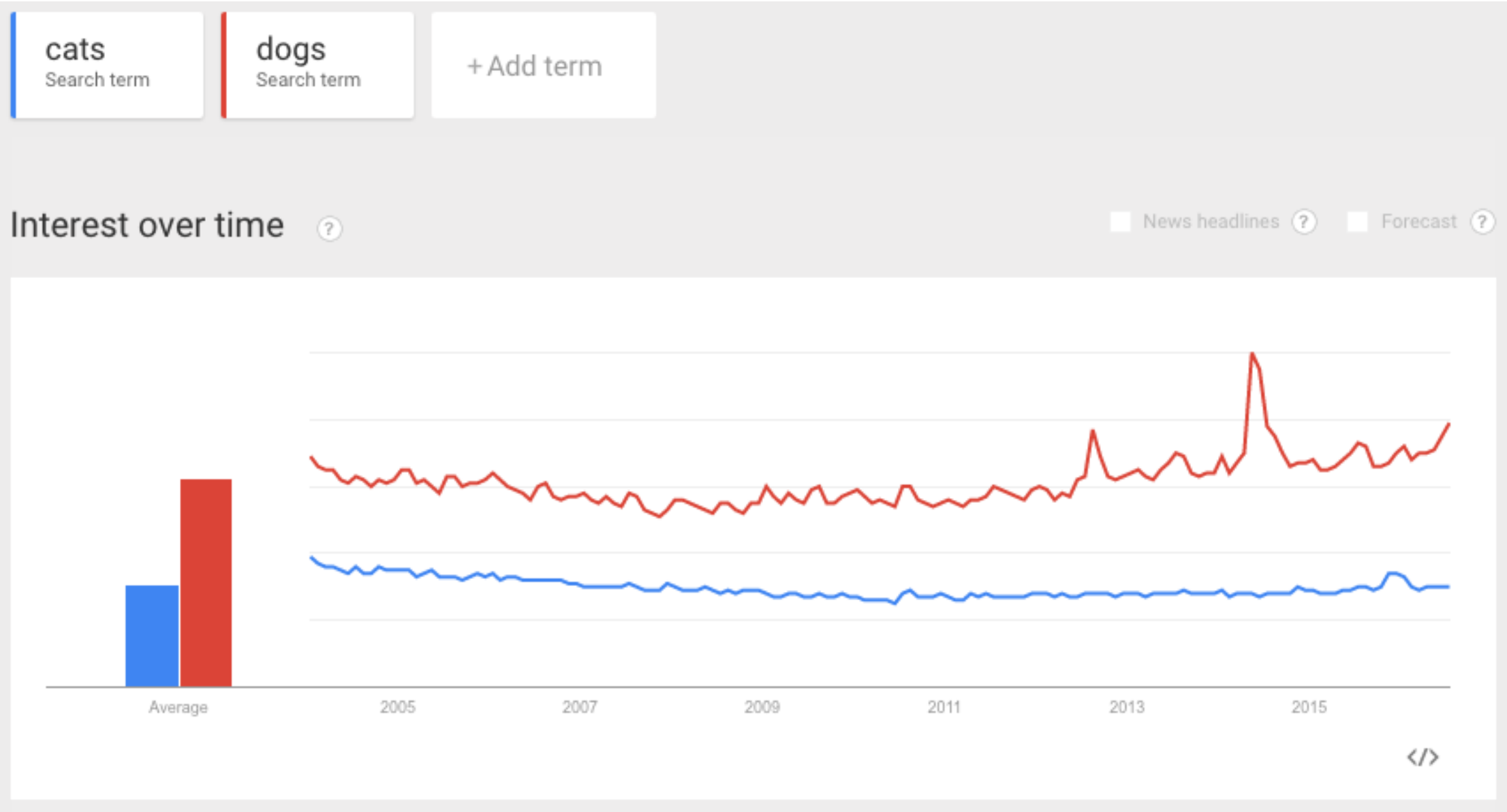
Create
Store
Use
Share
Archive
Destroy

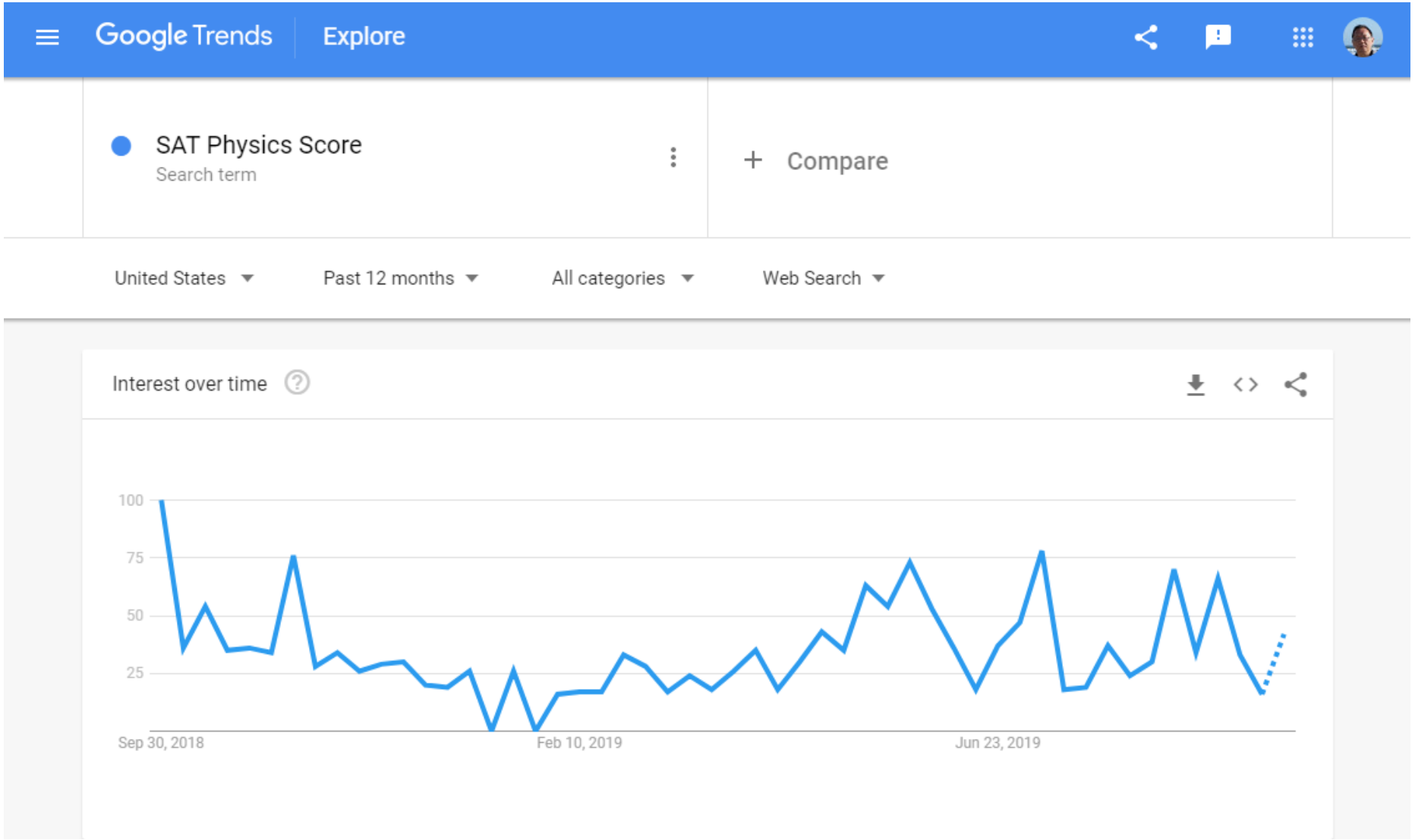
Funding
Motivation
Project Design
Data Collection & Sourcing
Analysis
Interpretation
Comunication & Distribution

Using Google Trends

You will be using Google Trends, which is a tool that allows you to **visualize data** about search history across different times and locations**.** You will be **looking for interesting patterns, trends, or relationships between multiple trends** and **try to tell the story that pattern is showing**.

- Access it here: https://www.google.com/trends/
- To get started, you want to “Explore” a trend of your own.
- Find the Explore text box or go to the Explore section of the site https://www.google.com/trends/explore


cats
Search term
dogs
Search term
+ Add term
Interest over time
News headlines
Forecast
Average
2005
2007
2009
2011
2013
2015


Google Trends
Explore
SAT Physics Score
Search term
Compare
United States
Past 12 months
All categories
Web Search
Interest over time
100
75
50
25
Sep 30, 2018
Feb 10, 2019
Jun 23, 2019

Do your first comparison of trends

- You can enter any two or three things you like. This example compares **streaming**, **mp3**, and **cd**.
- **Understand what is being represented:** Take a moment to actually understand what the Google Trends data is showing. Respond to the questions below. You can find answers on the Google Trends Help Page.
  - Where does Google Trends data come from?
  - How is Google Trends data adjusted? What does a value of 100 mean?

Exploring Trends on your own

Continue exploring comparisons of search trends that are interesting to you. You might start by looking up:

- **a current event / social movement / hashtag / meme**
- **your favorite hobby / movie / song / book / celebrity**
- **popular apps / businesses / products / websites**

Exploring Trends on your own

As you try different terms, here are **some things to look for:**

- **Long-term trends:** Is your topic becoming more popular over time? Less?
- **Short-term trends:** Does your topic suddenly spike or dip in popularity?
- **Patterns:** Does your topic follow some predictable repeated pattern?
- **Relationships:** Does one topic increase or decrease in popularity when another one does?
- **Zoom-in:** You can narrow your search to particular regions, times, and categories

Tell a Story

As you explore, you should eventually **settle on one topic** you find particularly interesting. For your chosen topic, **be prepared to present or share** what you found, responding to the following:

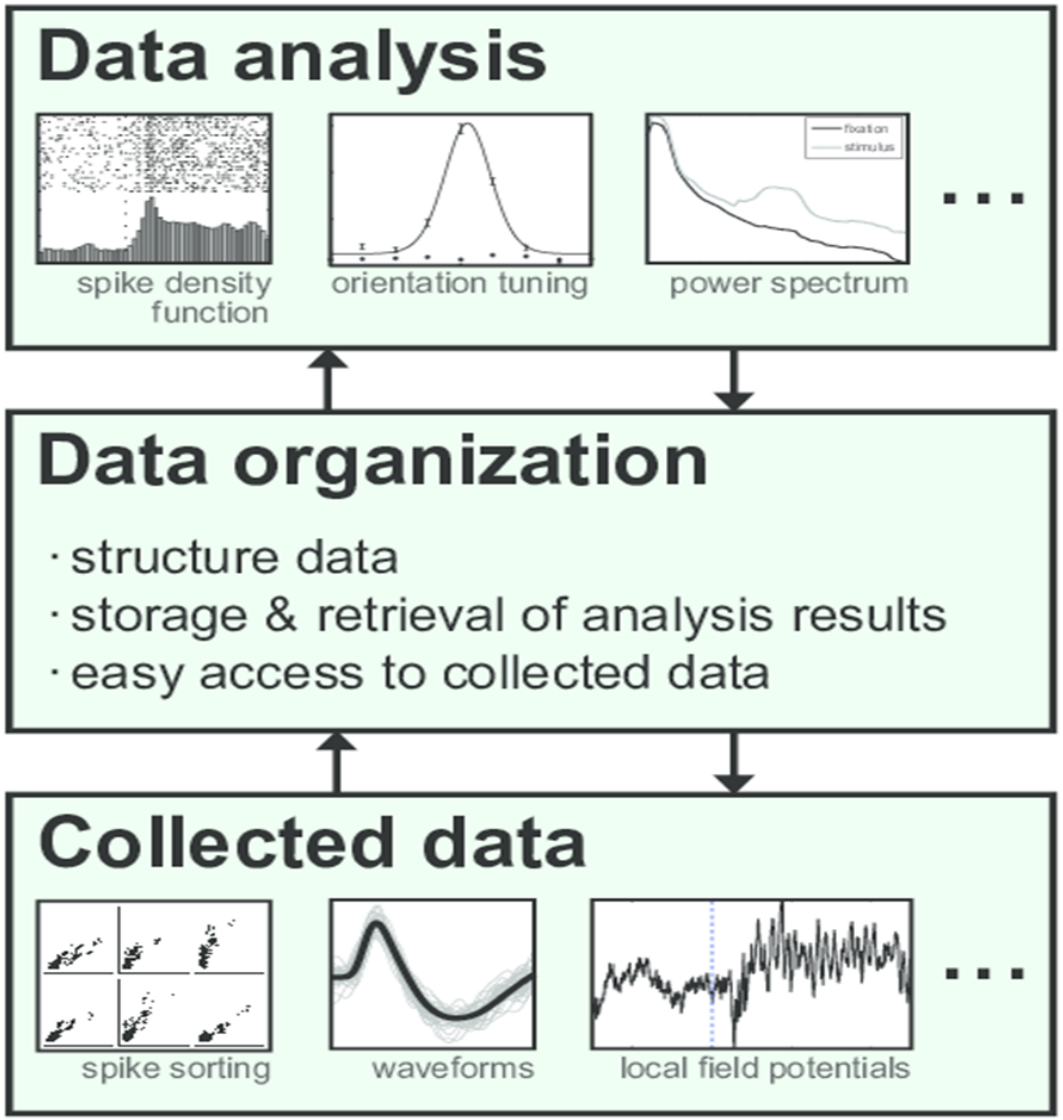
- **Describe** what terms you compared and whether you narrowed your search by using filters.
- **Accurately describe** what the charts or other visualizations are showing.
- Come up with a **possible story or explanation of why** the trend you described might have happened

Visualization of Data

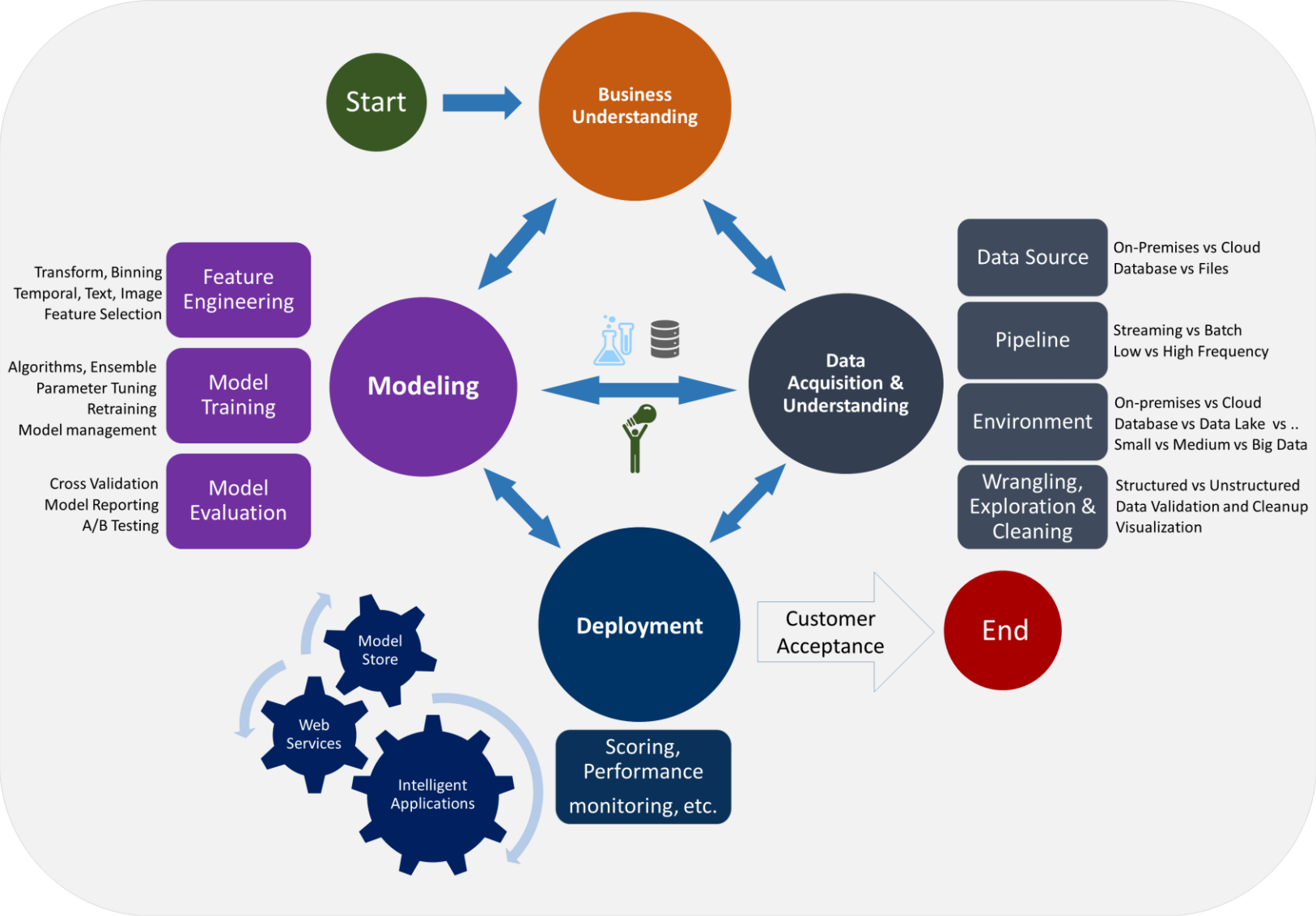
A graph of the planes in the air in real time (got into a discussion about safety with so much going on!)
https://www.flightradar24.com/-5.06,-4.1/2

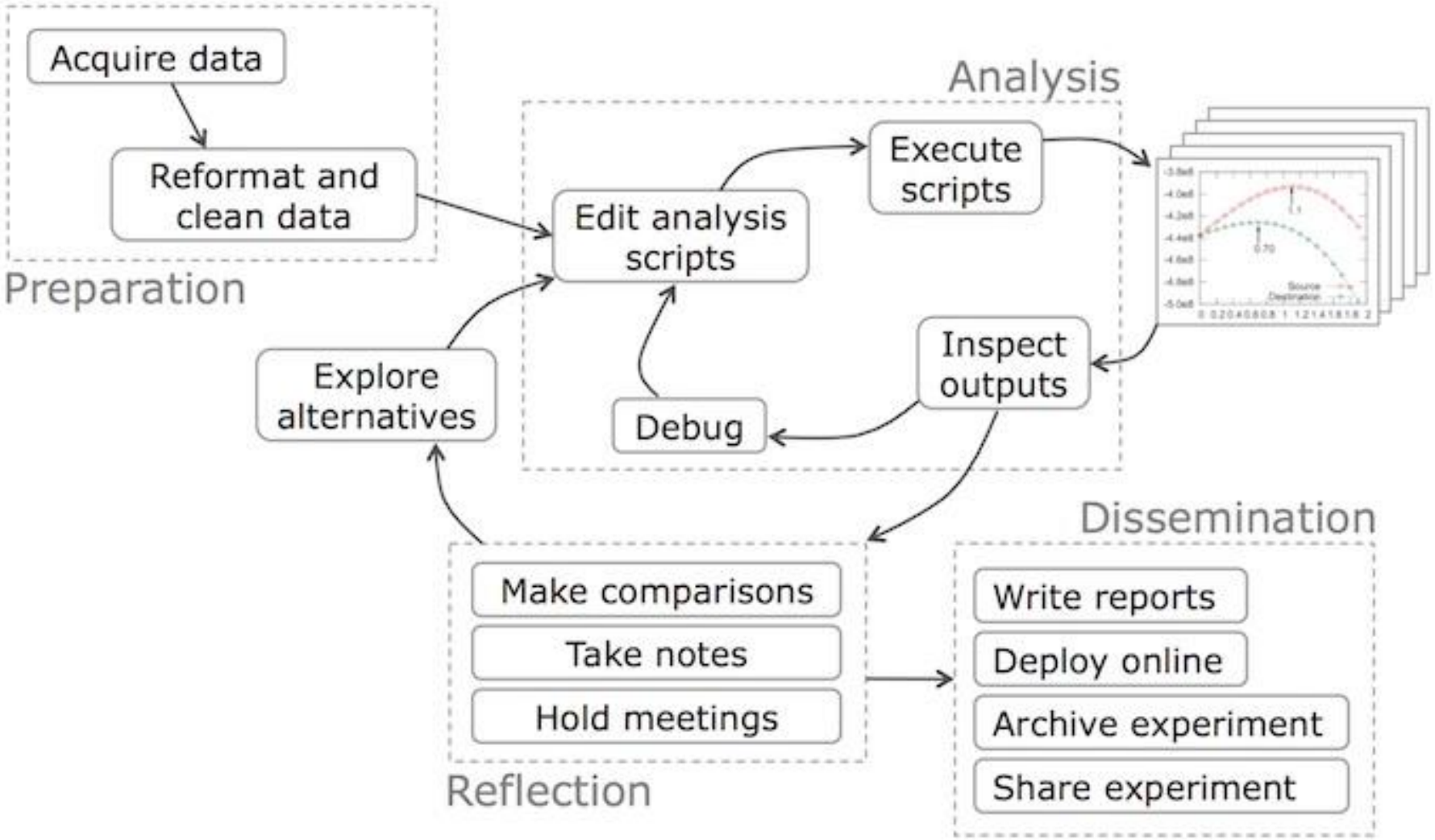
A graph of the planes in the air on Wednesday, Nov 21, 2018 (busiest travel day of the year in the US)
http://mentalfloss.com/article/550073/aircraft-tracker-shows-one-busiest-days-ever-air-travel

A graph (actually a video of graphs) of flights in the air on September 11, 2001 as the events of 9/11 progressed
https://www.youtube.com/watch?v=xh4V9PZT2VY


Data analysis
spike density function
orientation tuning
fixation
stimulus
power spectrum
...
Data organization
· structure data
· storage & retrieval of analysis results
· easy access to collected data
Collected data
spike sorting
waveforms
local field potentials
...

Data Science Lifecycle




Acquire data
Reformat and clean data
Preparation
Analysis
Execute scripts
Edit analysis scripts
Inspect outputs
Debug
Explore alternatives
Dissemination
Make comparisons
Take notes
Hold meetings
Reflection
Write reports
Deploy online
Archive experiment
Share experiment

Data Analysis

Check Your Assumptions

LECTURE 5

Overview

- This lesson asks students to consider carefully the assumptions they make when **interpreting data** and **data visualizations**. The class begins by examining how the Google Flu Trends project tried and failed to use search trends to predict flu outbreaks.
- They will then read a report on the **Digital Divide** which highlights how access to technology differs widely by personal characteristics like race and income. This report challenges a widespread assumption that data collected online is representative of the population at large.

Overview

- To practice **identifying assumptions** in data analysis, students are provided a series of scenarios in which data-driven decisions are made based on flawed assumptions.
- They will need to identify the assumptions being made (most notably those related to the digital divide) and explain why these assumptions lead to incorrect conclusions.

Objectives

- Define the digital divide as the variation in access or use of technology by various demographic characteristics.
- Identify assumptions made when drawing conclusions from data and data visualizations

Video

- Data-Driven Decision
- Bridging the Digital Divide

Activity - The “Digital Divide”

Perhaps one of the easiest assumptions to make, when looking at data collected online, is that it actually is a good representation of what the average person is thinking, doing, or cares about. Look through this report from Pew Research, which shows the large differences in access to technology, even in the modern day. Digital Divides 2018 (http://www.pewinternet.org/fact-sheet/internet-broadband/)

1. What is the “digital divide”?
2. What groups are overrepresented or underrepresented online as a result of the digital divide?
3. What was the most surprising piece of information or visualization you found in this report?

Identifying Assumptions in Data Analysis

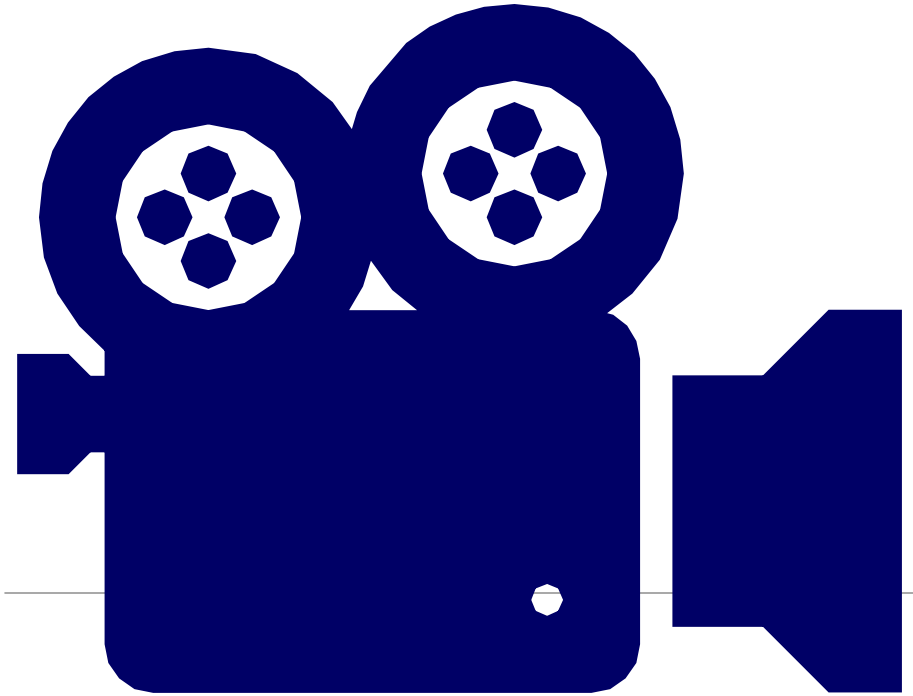
- When you use data to make decisions, you need to be careful to **identify your assumptions** and **reflect on how those assumptions impact your analysis.**
- **Pick one of the scenarios below.** With a partner, respond to the questions you find there about the assumptions made to conduct that analysis.

Scenario	Data	Decision
A city would like to more efficiently locate potholes that need to be filled.	The city builds an app that allows residents to report potholes from their smartphone.	The city will use this app as the primary method of identifying potholes.
A news agency would like to predict the outcome of a coming election.	A social media company (e.g., Twitter) keeps track of how many times each candidate has been mentioned on the platform.	The news agency will use this data to predict the outcome of the election.
A state government is trying to determine which issues are most important for the upcoming year.	The government creates an online survey where citizens can vote for the issue they care most about.	The government will use the results of this survey to help prioritize issues in the coming year.
A chef is deciding where to open a new restaurant in the city.	A restaurant reviews website keeps track of the areas of the city that receive the most restaurant reviews.	The chef will use the data to choose the location for opening a new restaurant.

Respond

Answer each question below for the scenario you chose.

1. What kind of trends or patterns might the decision makers in your scenario be observing in the data when they make their decision? How might they use these patterns to help make their decision?
2. The people making decisions in each instance have not taken into account the "digital divide." How has the assumption that all people have equal access to technology **possibly biased their analysis?**
3. Are there any other assumptions being made in your scenario? How might they affect the final decision made? What additional data might you want to collect before making the decision for your scenario?



Data Analysis and Interpretation

VIDEO - HTTPS://YOUTU.BE/DSLWZFLDAWM

Data Regression

ACTIVITY

Overview

- Learn about the different regression types in machine learning, including linear and logistic regression
- Each regression technique has its own regression equation and regression coefficients
- We cover 7 different regression types in this article

Introduction

- Linear and Logistic regressions are usually the first algorithms people learn in data science. Due to their popularity, a lot of analysts even end up thinking that they are the only form of regressions. The ones who are slightly more involved think that they are the most important among all forms of regression analysis.
- The truth is that there are innumerable forms of regressions, which can be performed. Each form has its own importance and a specific condition where they are best suited to apply. In this article, I have explained the most commonly used 7 types of regression in data science in a simple manner.
- Through this article, I also hope that people develop an idea of the breadth of regressions, instead of just applying linear/logistic regression to every machine learning problem they come across and hoping that they would just fit!

What is Regression Analysis?

- Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

- Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. I'll explain this in more details in coming sections.

Why do we use Regression Analysis?

- As mentioned above, regression analysis estimates the relationship between two or more variables. Let's understand this with an easy example:
- Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

Why do we use Regression Analysis?

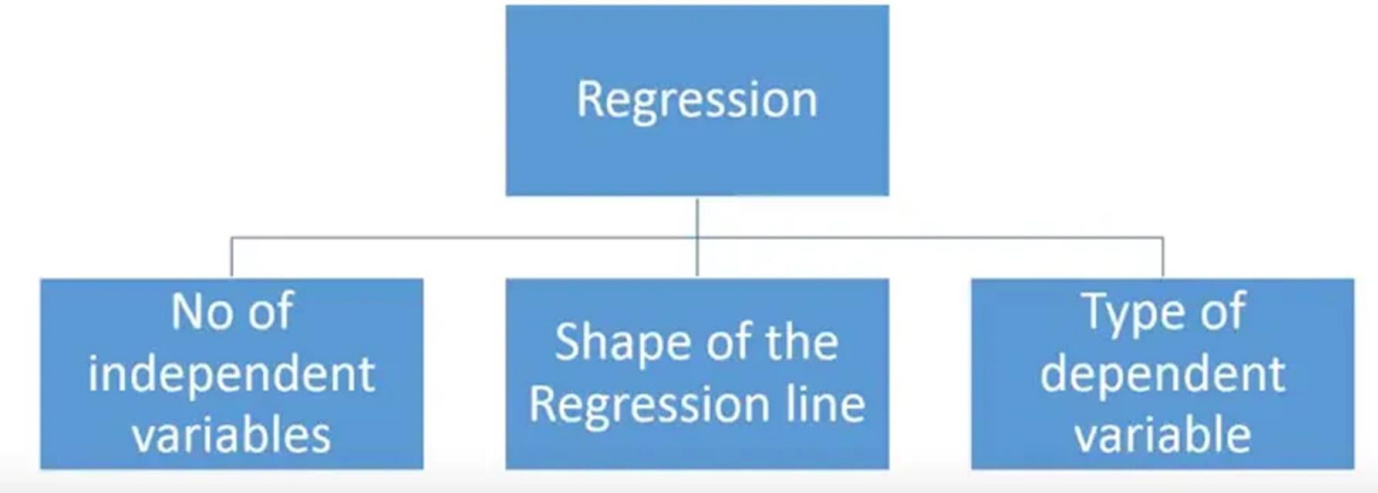
There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

How many types of regression techniques do we have?

- There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).

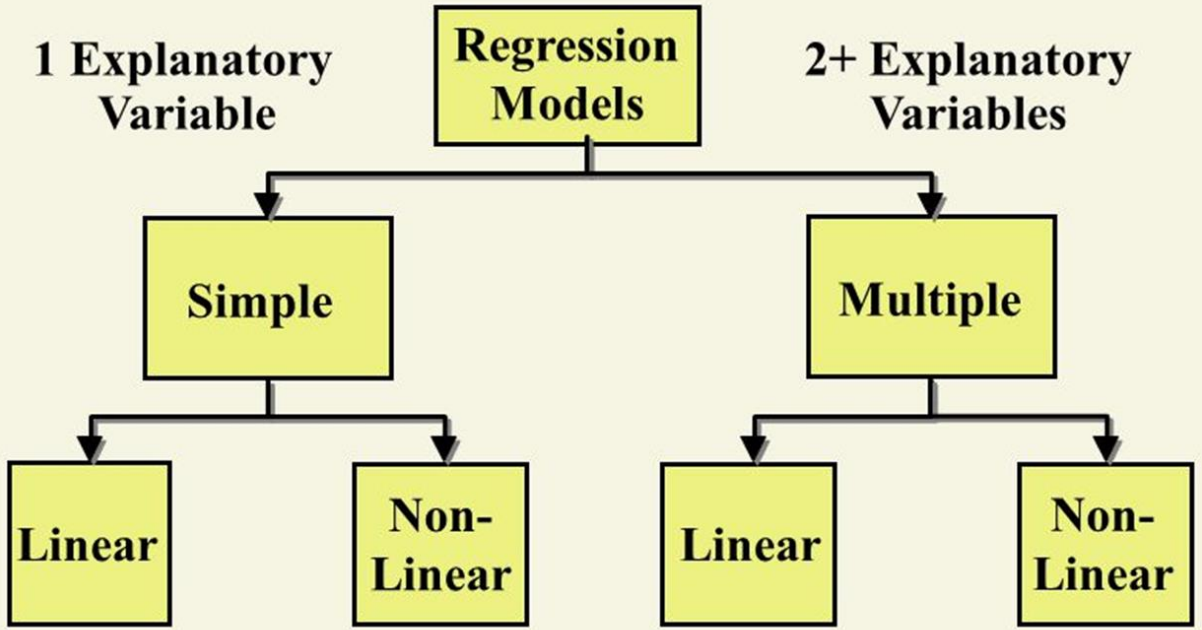


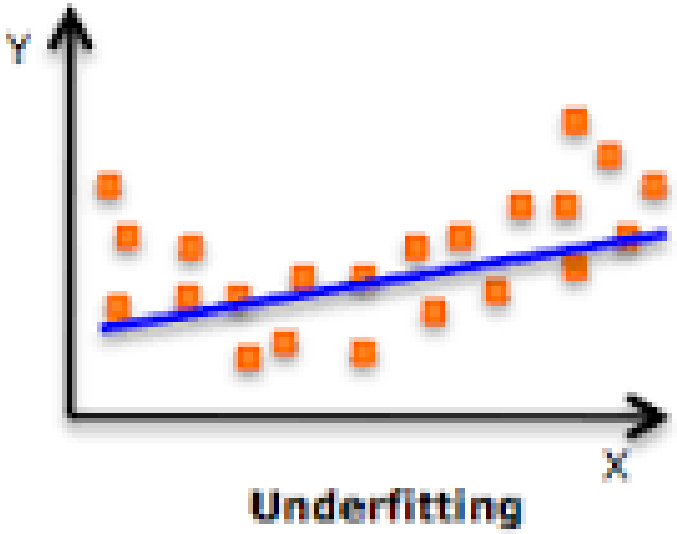
How many types of regression techniques do we have?

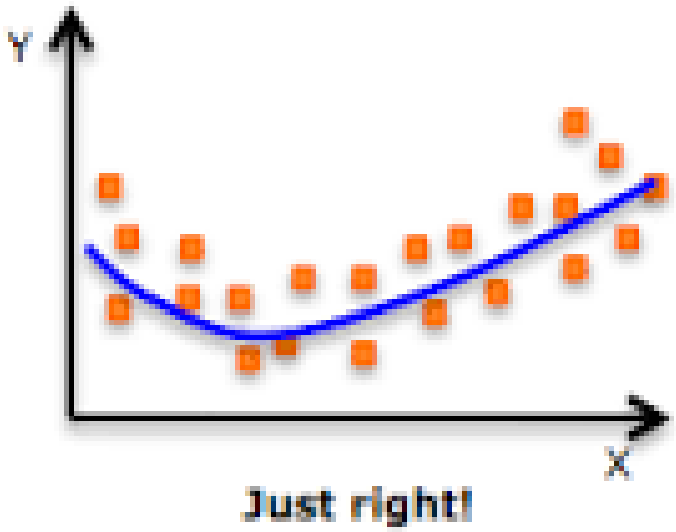
For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven't used before. But before you start that, let us understand the most commonly used regressions:

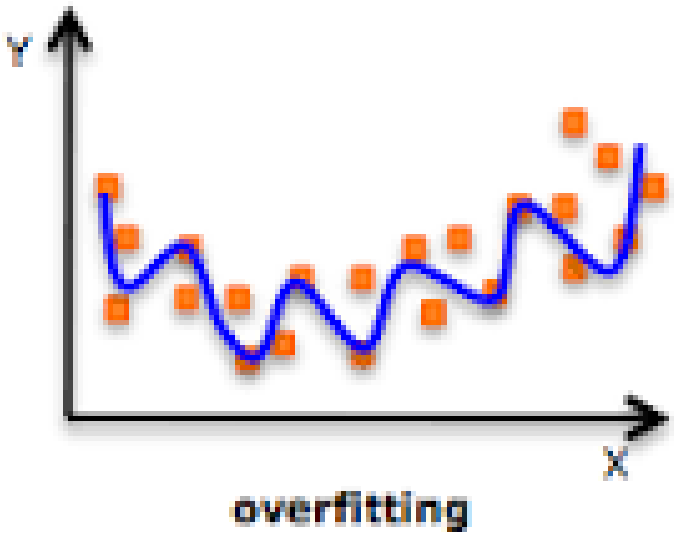
1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. ElasticNet Regression

Types of Regression Models



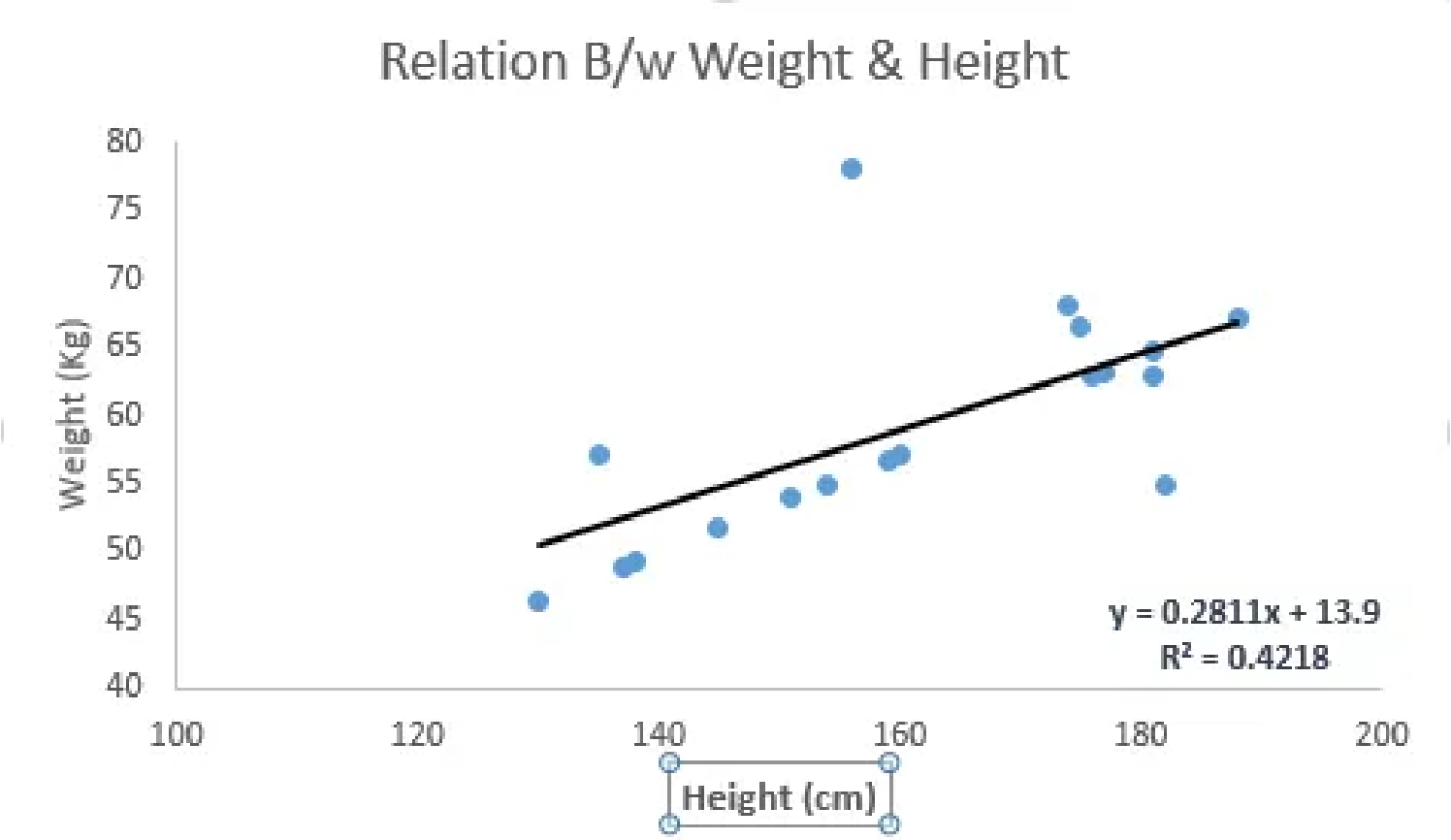

Y
X
Underfitting


Y
X
Just right!


Y
X
overfitting

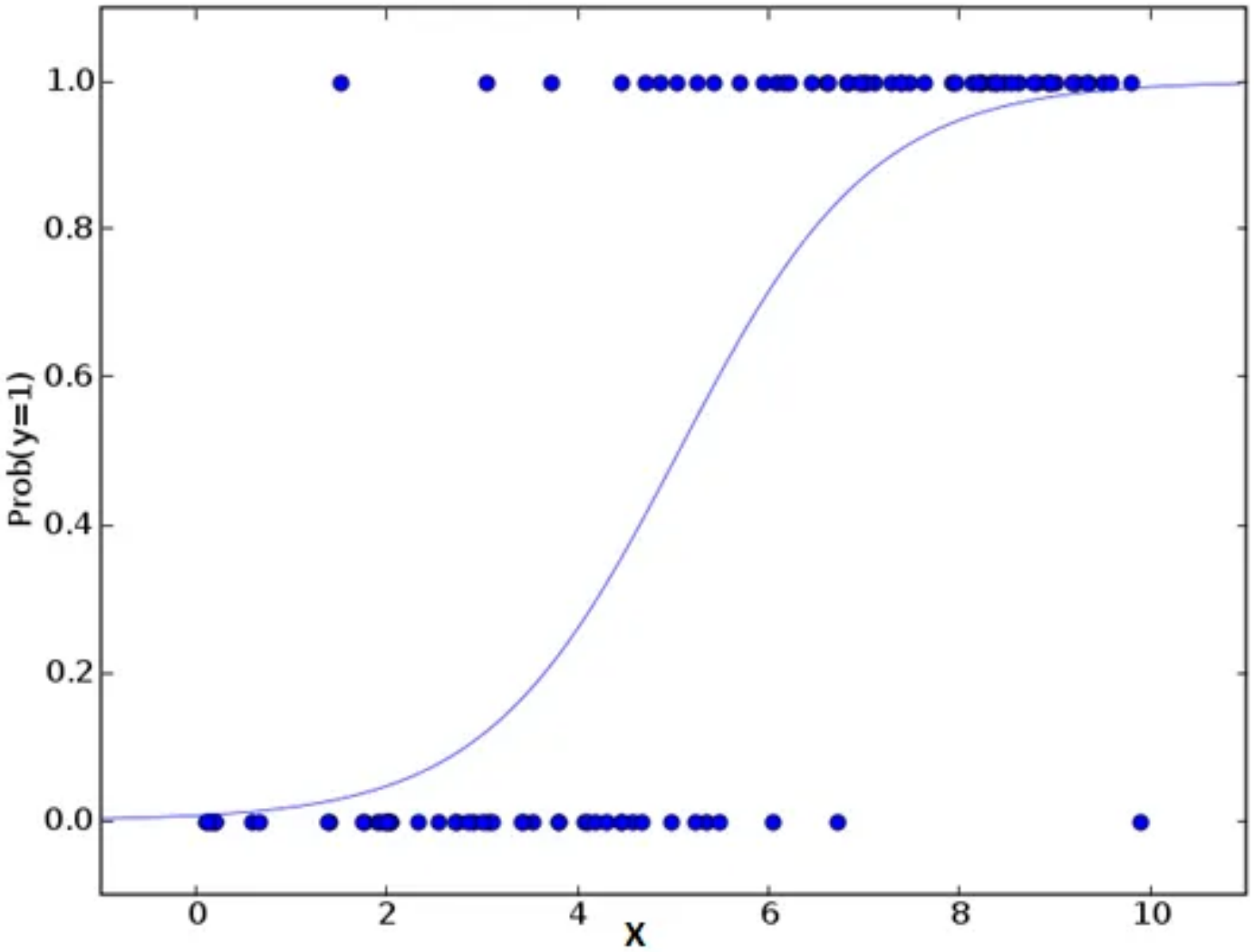
Linear Regression

*Y=a+b*X + e*



Logistic Regression

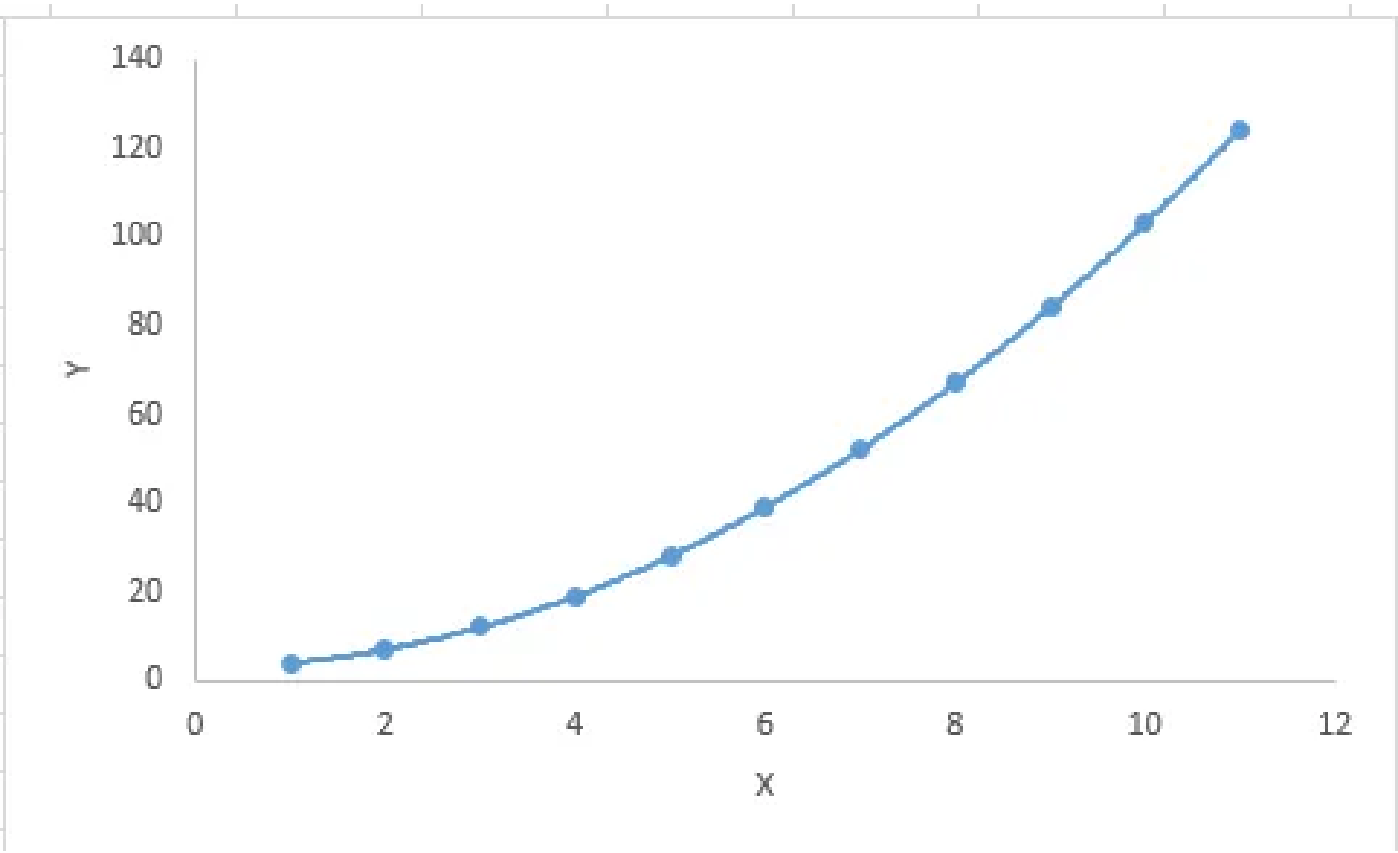
- **Logistic regression** is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.
- odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence ln(odds) = ln(p/(1-p)) logit(p) = ln(p/(1-p)) = $\mathbf{b_0+b_1X^1+b_2X^2+b_3X^3....+b_kX^k}$
- Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is "why have we used log in the equation?".
- Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is **logit** function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).



Logistic Regression

Polynomial Regression

$y = a x^2 + b x + c$



Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

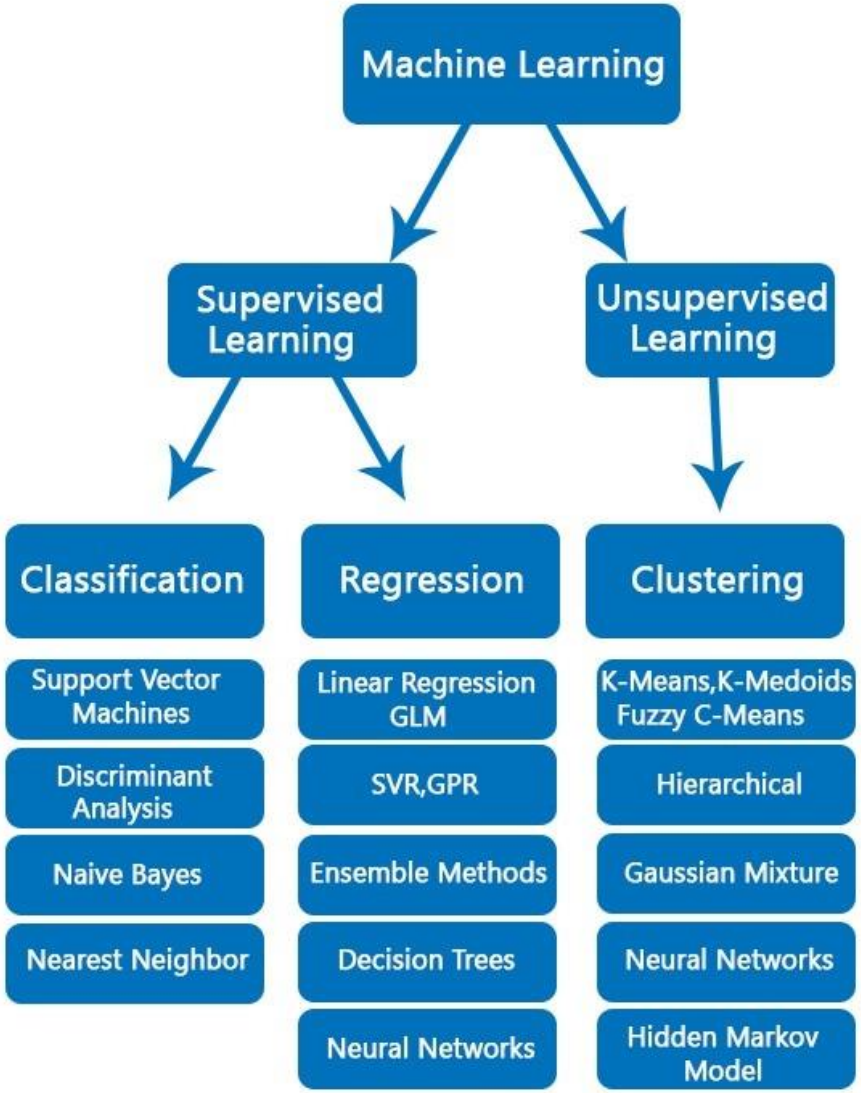
Stepwise Regression

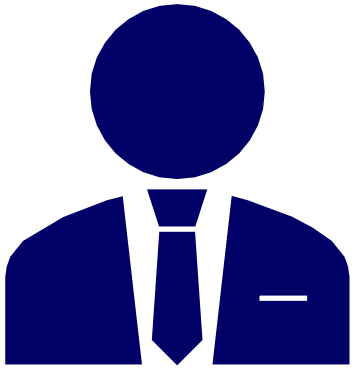
- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

Other Regression Models

1. Ridge Regression
2. Lasso Regression
3. ElasticNet Regression


Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Linear Regression GLM
SVR,GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means,K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Neural Networks
Hidden Markov Model



Qualitative Research

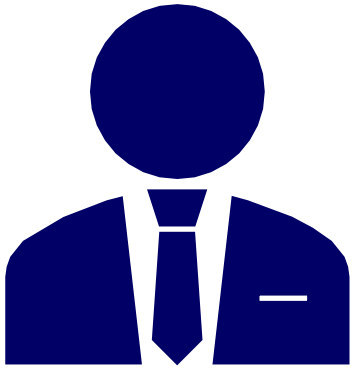
ACTIVITY

Qualitative Research

- The word qualitative implies an emphasis on the qualities of entities and on processes and meanings that are not experimentally examined or measured [if measured at all] in terms of quantity, amount, intensity, or frequency.
- Qualitative researchers stress the socially constructed nature of reality, the intimate relationship between the researcher and what is studied, and the situational constraints that shape inquiry. Such researchers emphasize the value-laden nature of inquiry.

Qualitative Research

- They seek answers to questions that stress how social experience is created and given meaning. In contrast, quantitative studies emphasize the measurement and analysis of causal relationships between variables, not processes.
- Qualitative forms of inquiry are considered by many social and behavioral scientists to be as much a perspective on how to approach investigating a research problem as it is a method.



Quantitative Research

ACTIVITY

Quantitative Research

- Quantitative methods emphasize objective measurements and the **statistical, mathematical**, or **numerical analysis** of data collected through polls, questionnaires, and surveys, or by manipulating pre-existing statistical data using computational techniques.
- Quantitative research focuses on gathering numerical data and generalizing it across groups of people or to explain a particular phenomenon.

Machine Learning and Bias

LECTURE 6

Machine Learning

- Machine learning refers to a computer that can recognize patterns and make decisions without being explicitly programmed. In this activity you're going to supply the data to train your own machine learning model. Imagine an ocean that contains creatures like fish, but also contains trash dumped by humans. What if we could train a computer to tell the difference and then use that technology to help clean the ocean?

Video: What is Machine Learning?

https://youtu.be/KHbwOetbmbs

Material

- All material available on GitHub
  - this presentation
  - conda environments
  - Jupyter notebooks

https://github.com/gjbex/PRACE_ML

or

https://bit.ly/prace2019_ml

Introduction

Machine learning is making great strides

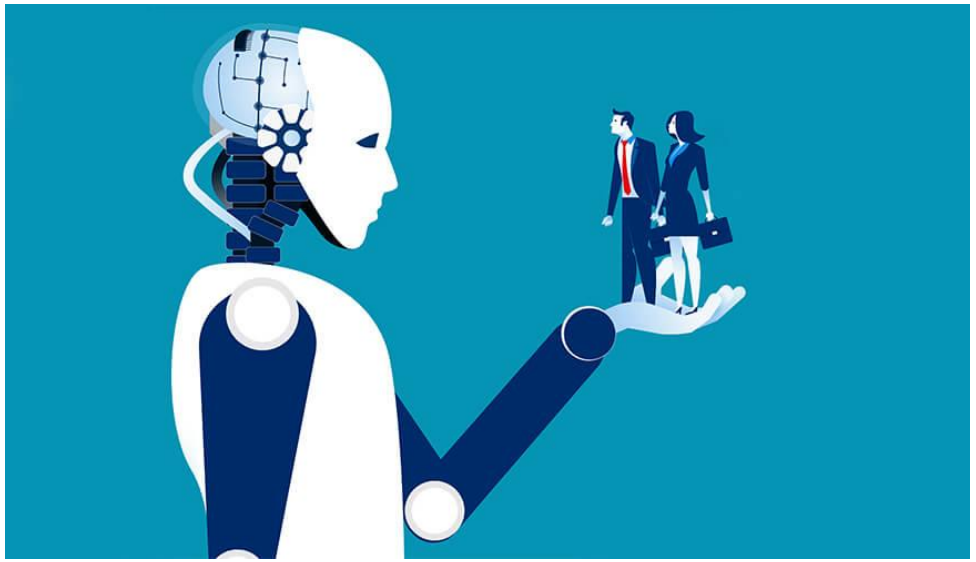
- Large, good data sets
- Compute power
- Progress in algorithms

Many interesting applications

- commercial
- scientific

Links with artificial intelligence

- However, AI $\neq$ machine learning



Machine learning tasks

Supervised learning

- regression: predict numerical values
- classification: predict categorical values, i.e., labels

Unsupervised learning

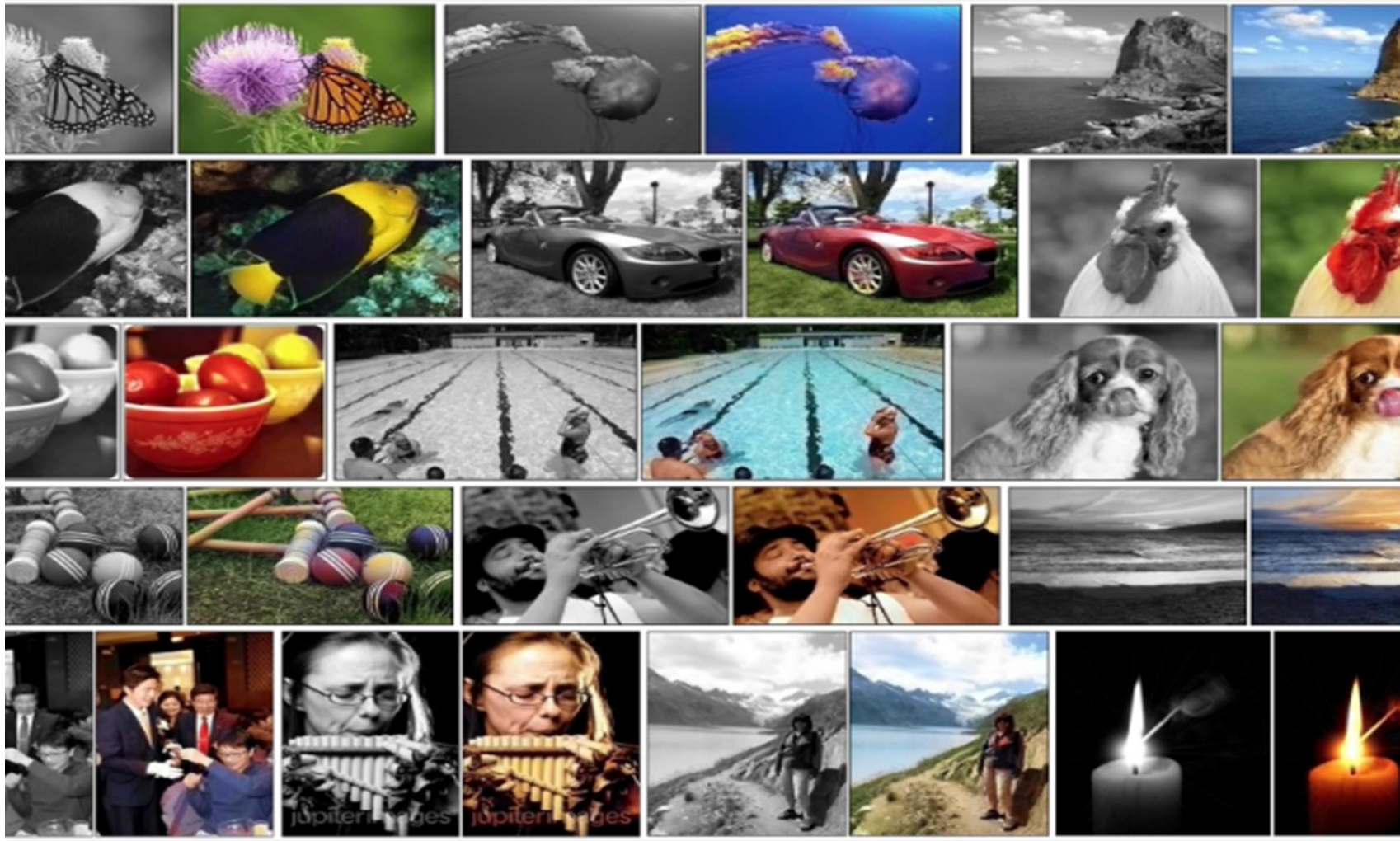
- clustering: group data according to "distance"
- association: find frequent co-occurrences
- link prediction: discover relationships in data
- data reduction: project features to fewer features

Reinforcement learning

Regression

Colorize B&W images automatically

https://tinyclouds.org/colorize/

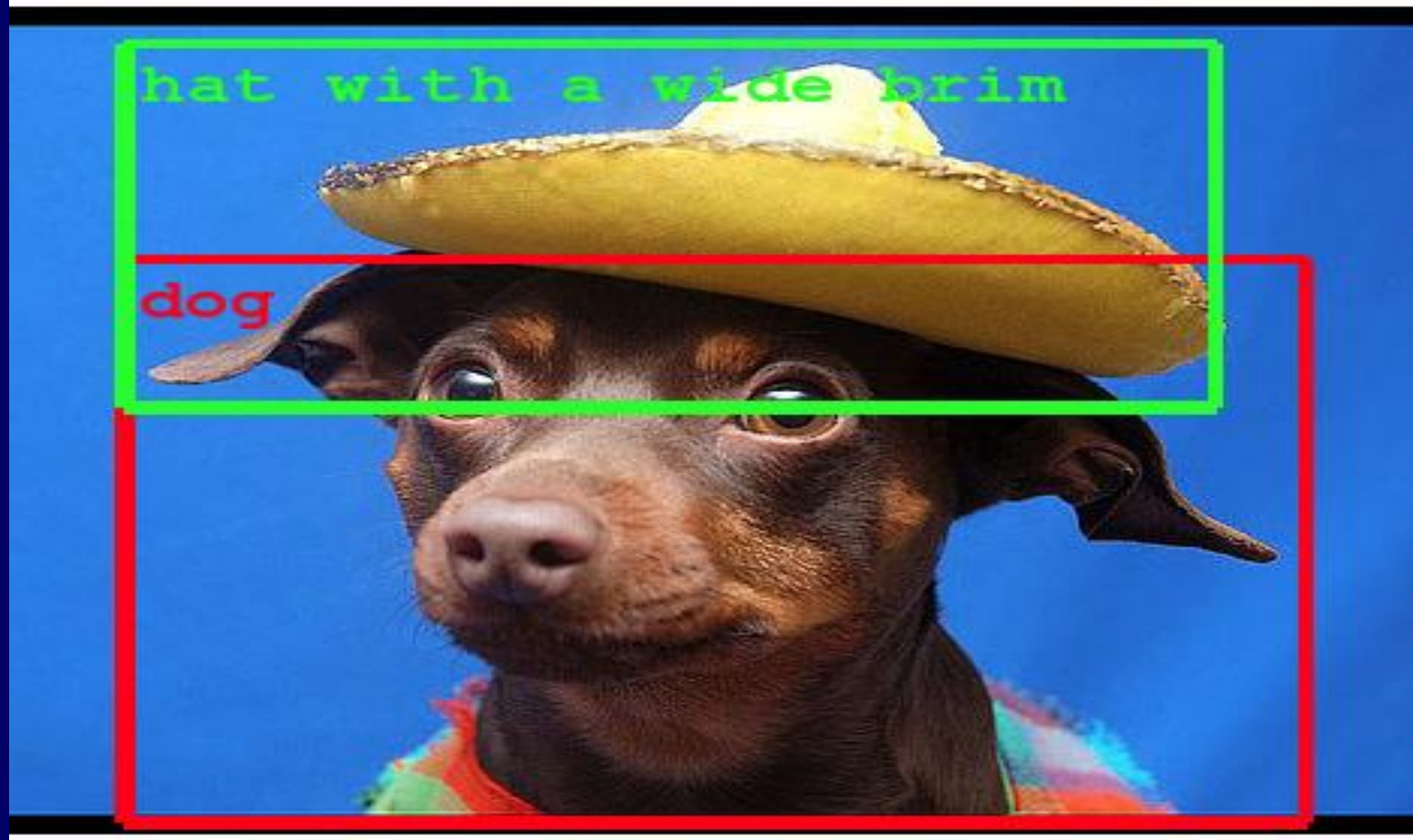


eC Learning Channel

Classification

Object recognition

https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html

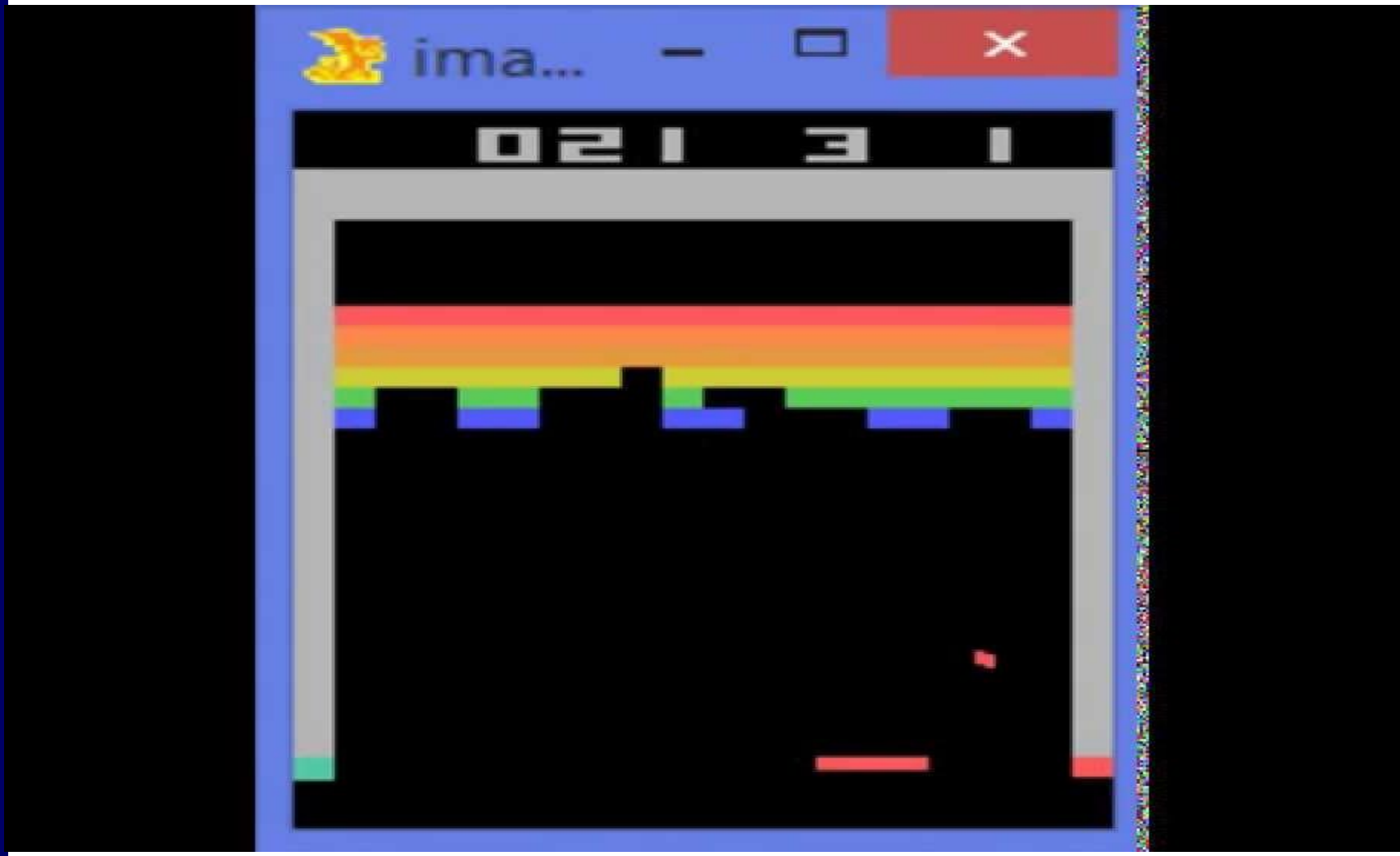


eC Learning Channel

Reinforcement learning

Learning to play Break Out

https://www.youtube.com/watch?v=V1eYniJ0Rnk

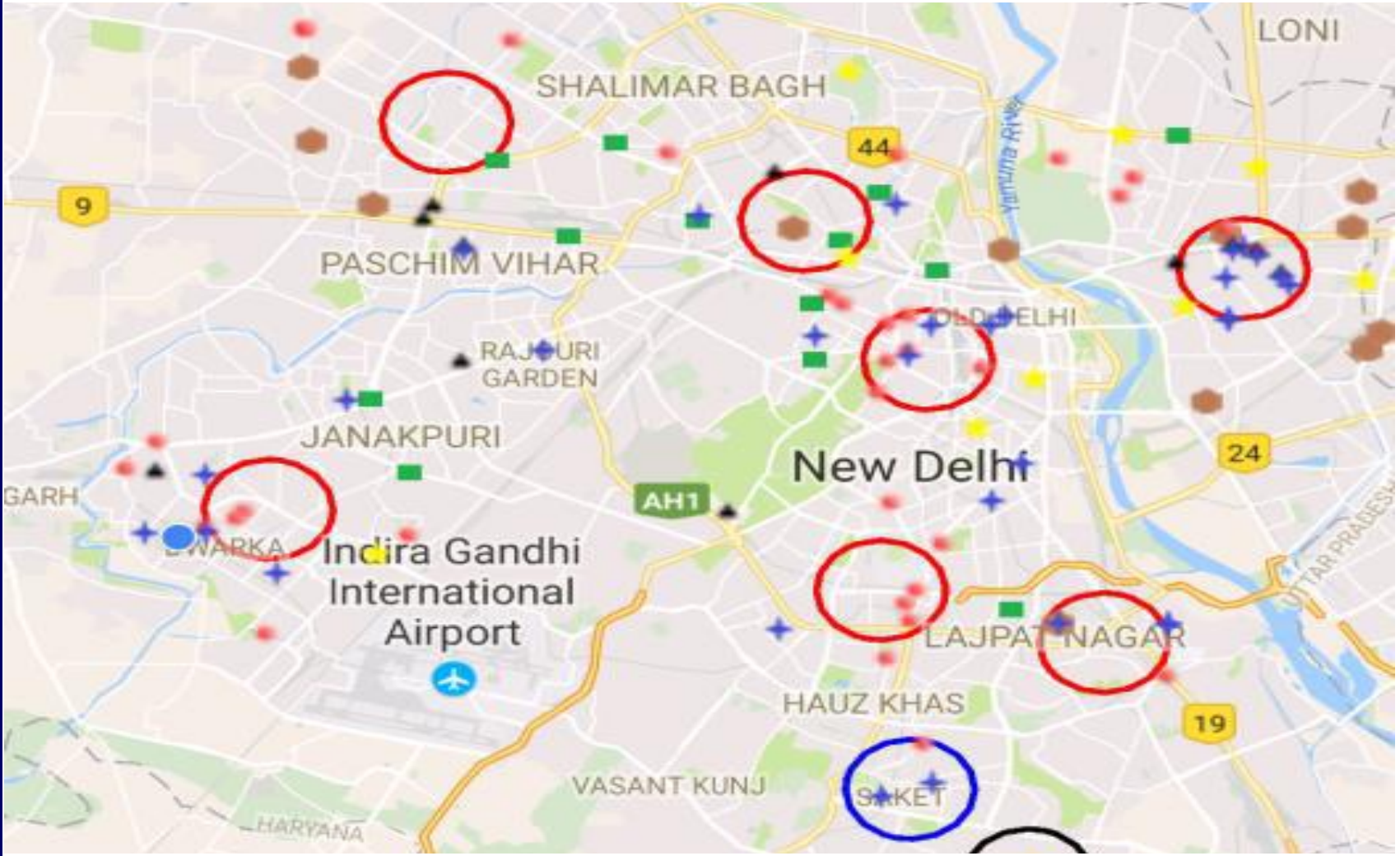


eC Learning Channel

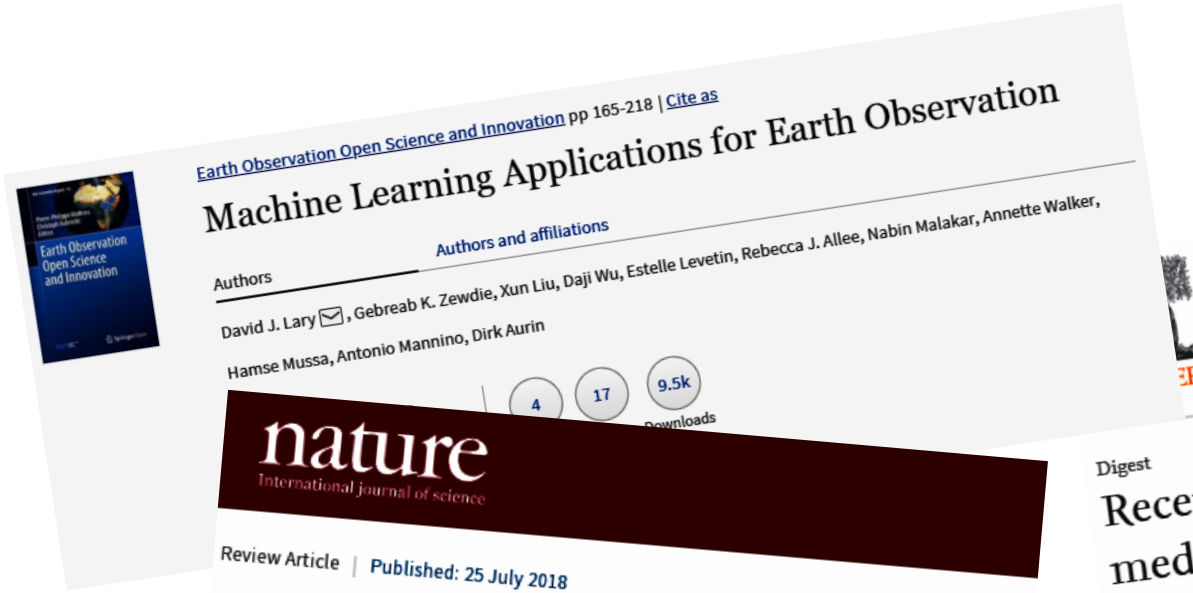
Clustering

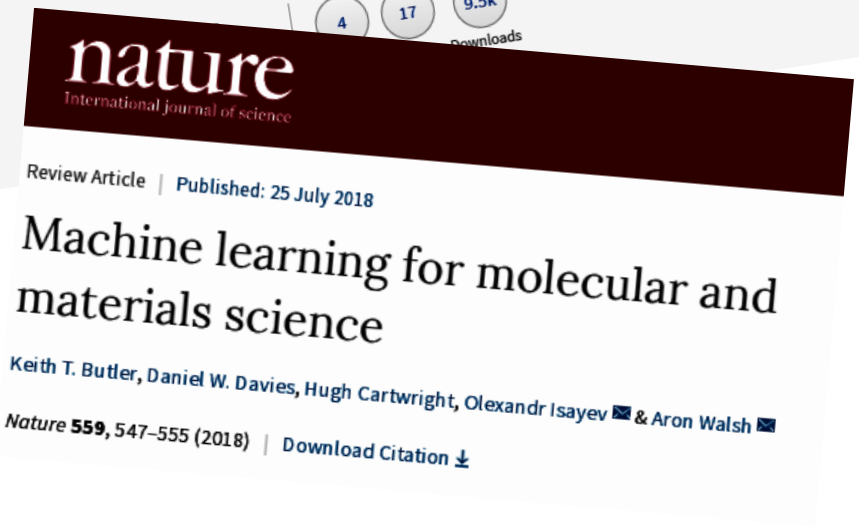
Crime prediction using k-means clustering

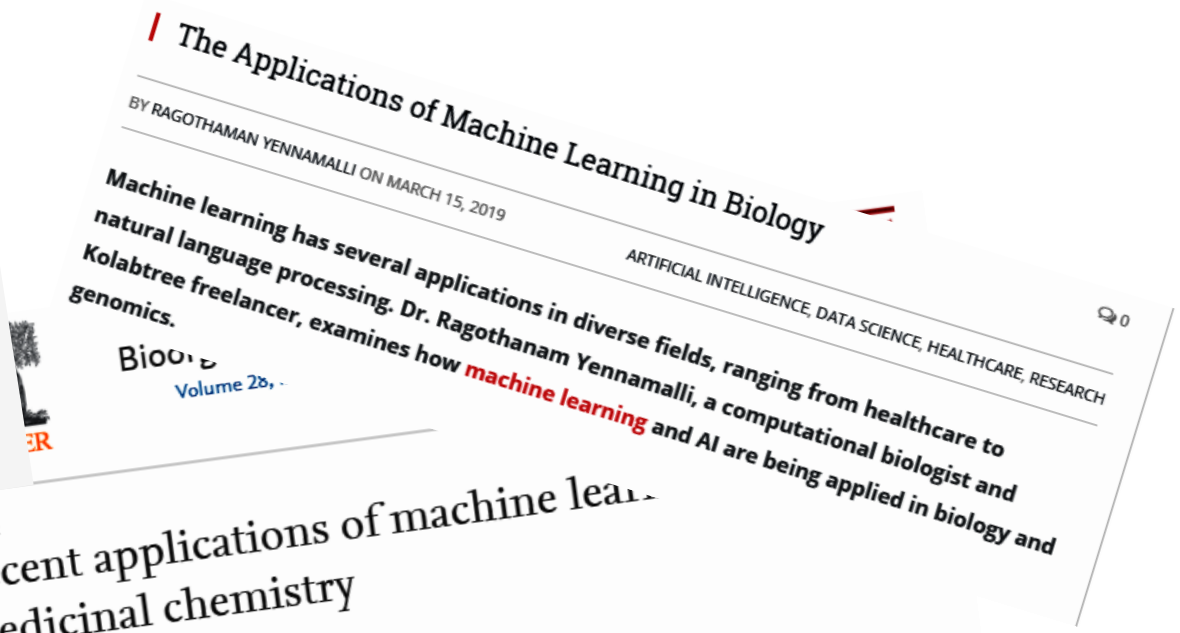
http://www.grdjournals.com/uploads/article/GRDJE/V02/I05/0176/GRDJEV02I050176.pdf

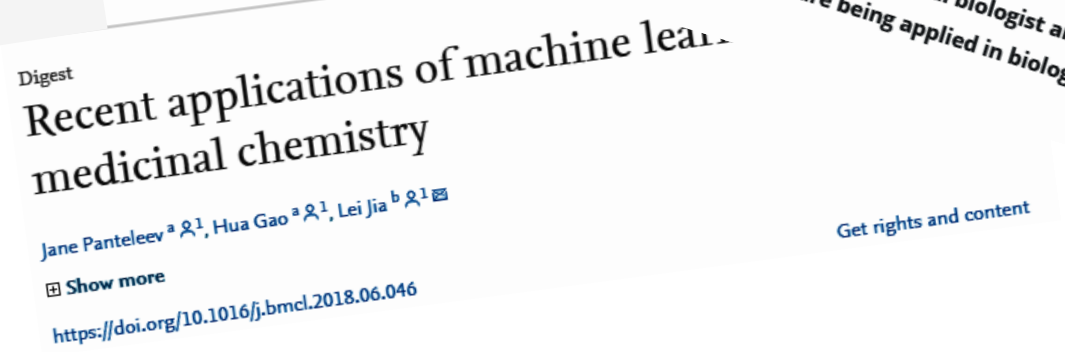


eC Learning Channel









Applications in Science

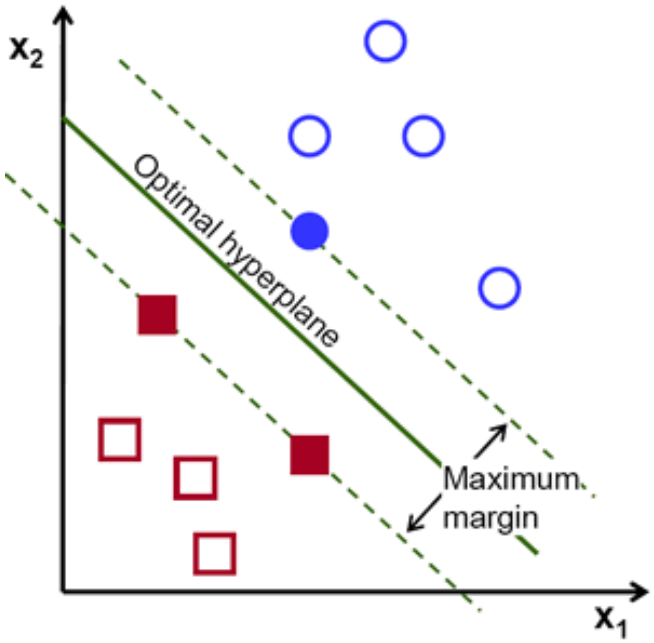
eC Learning Channel

Machine learning algorithms

Regression:
Ridge regression, Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

Classification:
Naive Base, , Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

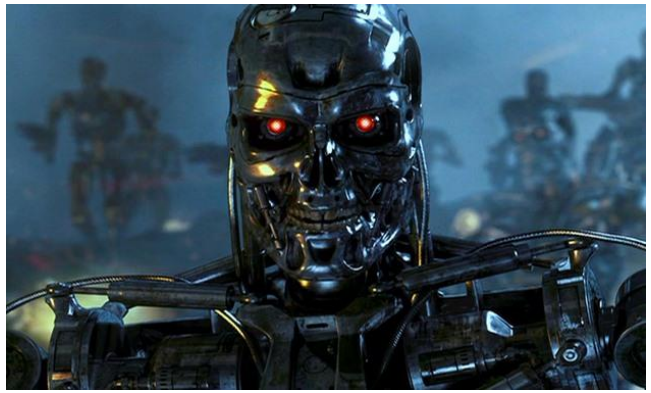
Clustering:
k-Means, Hierarchical Clustering, ...



Issues

Many machine learning/AI projects fail (Gartner claims 85 %)

Ethics, e.g., Amazon has/had sub-par employees fired by an AI automatically



Reasons for failure

Asking the wrong question

Trying to solve the wrong problem

Not having enough data

Not having the right data

Having too much data

Hiring the wrong people

Using the wrong tools

Not having the right model

Not having the right yardstick

Frameworks

Programming languages

- Python
- R
- C++
- …

Many libraries

- scikit-learn
- PyTorch
- TensorFlow
- Keras
- …

Fast-evolving ecosystem!

classic machine learning

deep learning frameworks

scikit-learn

Nice end-to-end framework

- data exploration (+ pandas + holoviews)
- data preprocessing (+ pandas)
  - cleaning/missing values
  - normalization
- training
- testing
- application

"Classic" machine learning only

https://scikit-learn.org/stable/

Keras

High-level framework for deep learning

TensorFlow backend

Layer types
- dense
- convolutional
- pooling
- embedding
- recurrent
- activation
- …

https://keras.io/

Data pipelines

Data ingestion

- CSV/JSON/XML/H5 files, RDBMS, NoSQL, HTTP,...

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

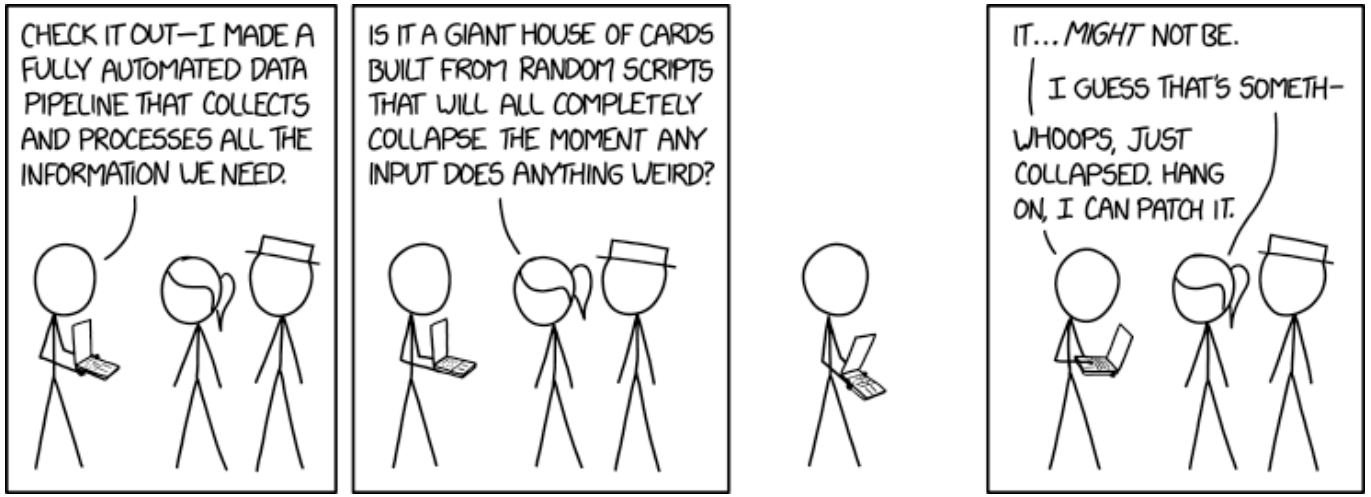
Data cleaning

- outliers/invalid values? $\rightarrow$ filter
- missing values? $\rightarrow$ impute

Must be done systematically

Data transformation

- scaling/normalization



Supervised learning: methodology

Select model, e.g., random forest, (deep) neural network, ...

Train model, i.e., determine parameters

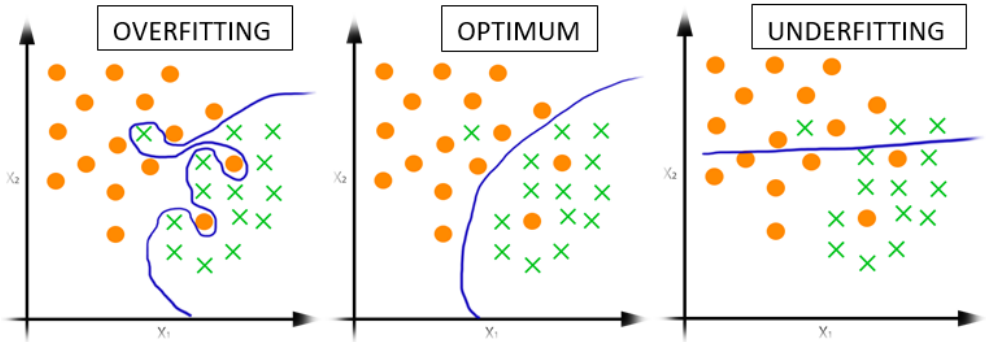
- Data: input + output
  - training data → determine model parameters
  - validation data → yardstick to avoid overfitting

Test model

- Data: input + output
  - testing data → final scoring of the model

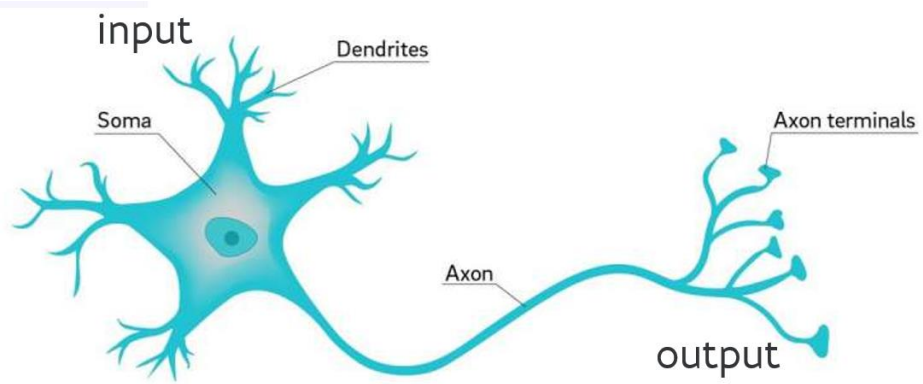
Production

- Data: input → predict output

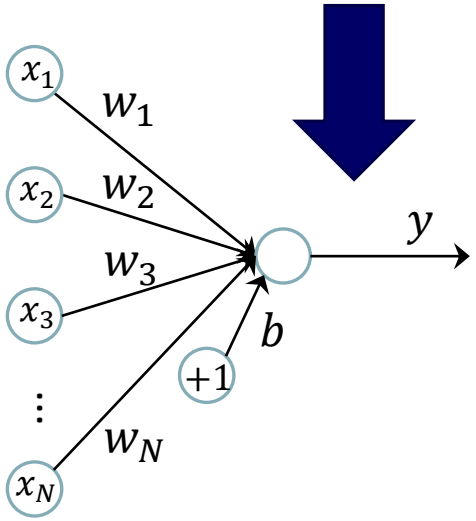


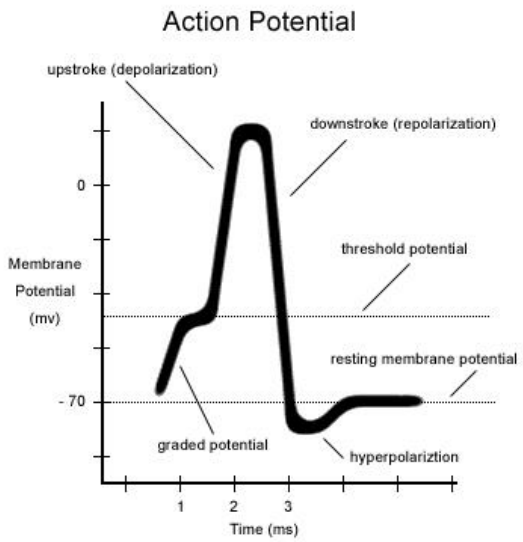
Experiment with underfitting and overfitting:
010_underfitting_overfitting.ipynb

From neurons to ANNs

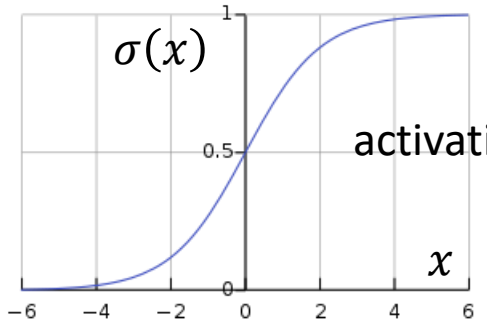


inspiration



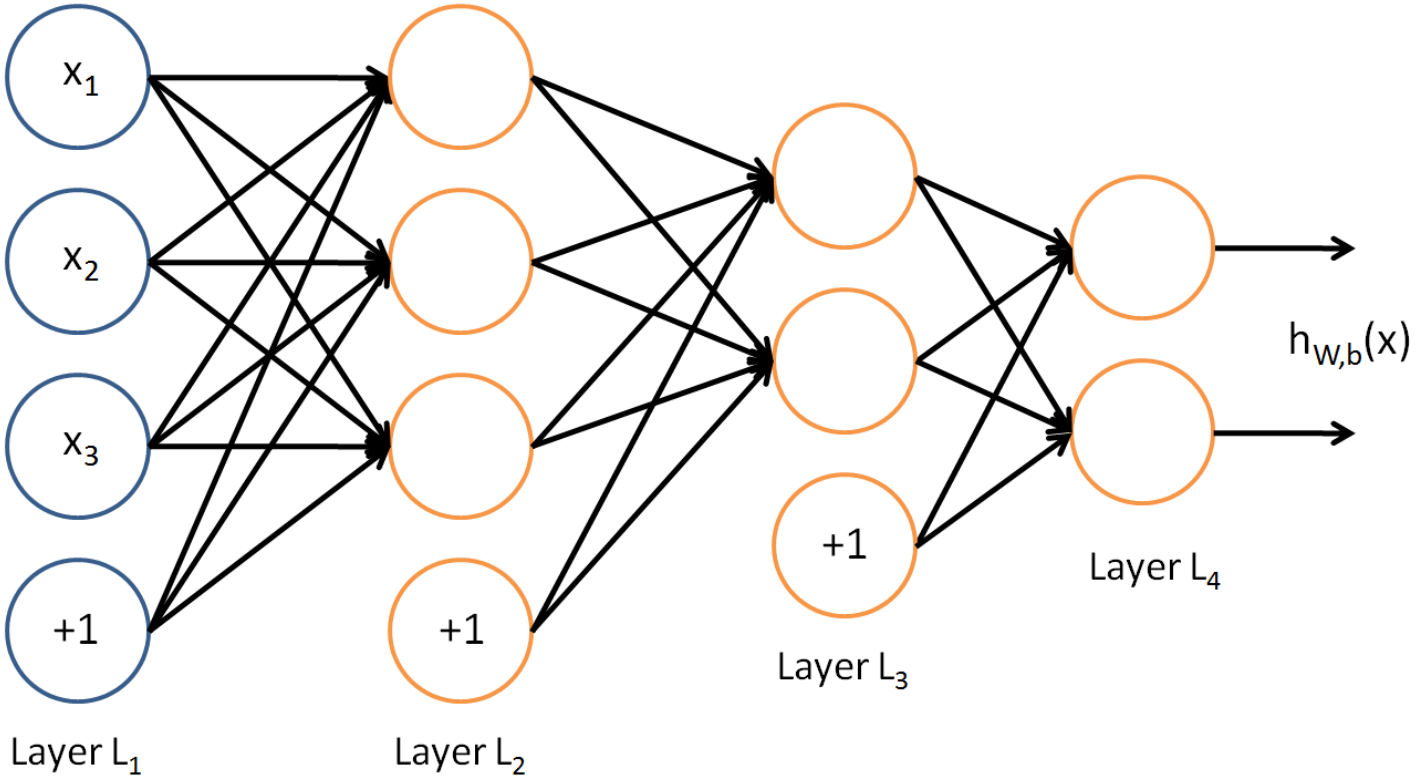


$$y = \sigma\left(\sum_{i=1}^{N} w_i x_i + b\right)$$



activation function

Multilayer network



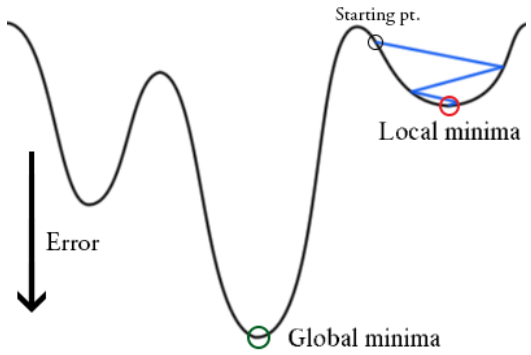
How to determine weights?

Training: backpropagation

Initialize weights "randomly"

For all training epochs

- for all input-output in training set
  - using input, compute output (forward)
  - compare computed output with training output
  - adapt weights (backward) to improve output
- if accuracy is good enough, stop



Task: handwritten digit recognition

Input data

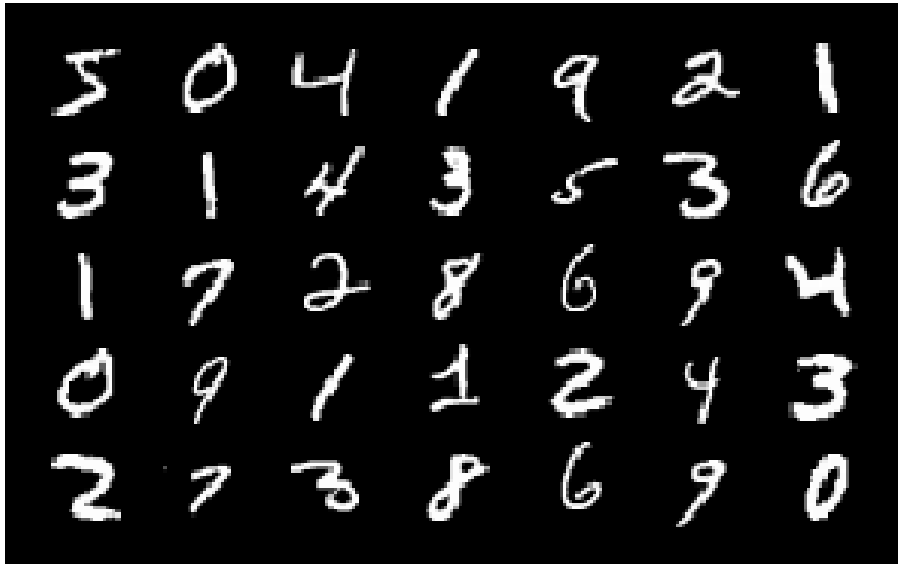
- grayscale image

Output data

- digit 0, 1, ..., 9

Training examples

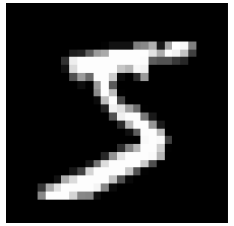
Test examples



Explore the data: 020_mnist_data_exploration.ipynb

First approach

- Data preprocessing
  - Input data as 1D array
  - output data as array with one-hot encoding
- Model: multilayer perceptron
  - 758 inputs
  - dense hidden layer with 512 units
  - ReLU activation function
  - dense layer with 512 units
  - ReLU activation function
  - dense layer with 10 units
  - SoftMax activation function



array([0.0, 0.0,..., 0.951, 0.533,..., 0.0, 0.0], dtype=float32)

5

array([0, 0, 0, 0, 0, 1, 0, 0, 0, 0], dtype=uint8)

Activation functions: 030_activation_functions.ipynb

Multilayer perceptron: 040_mnist_mlp.ipynb

Deep neural networks

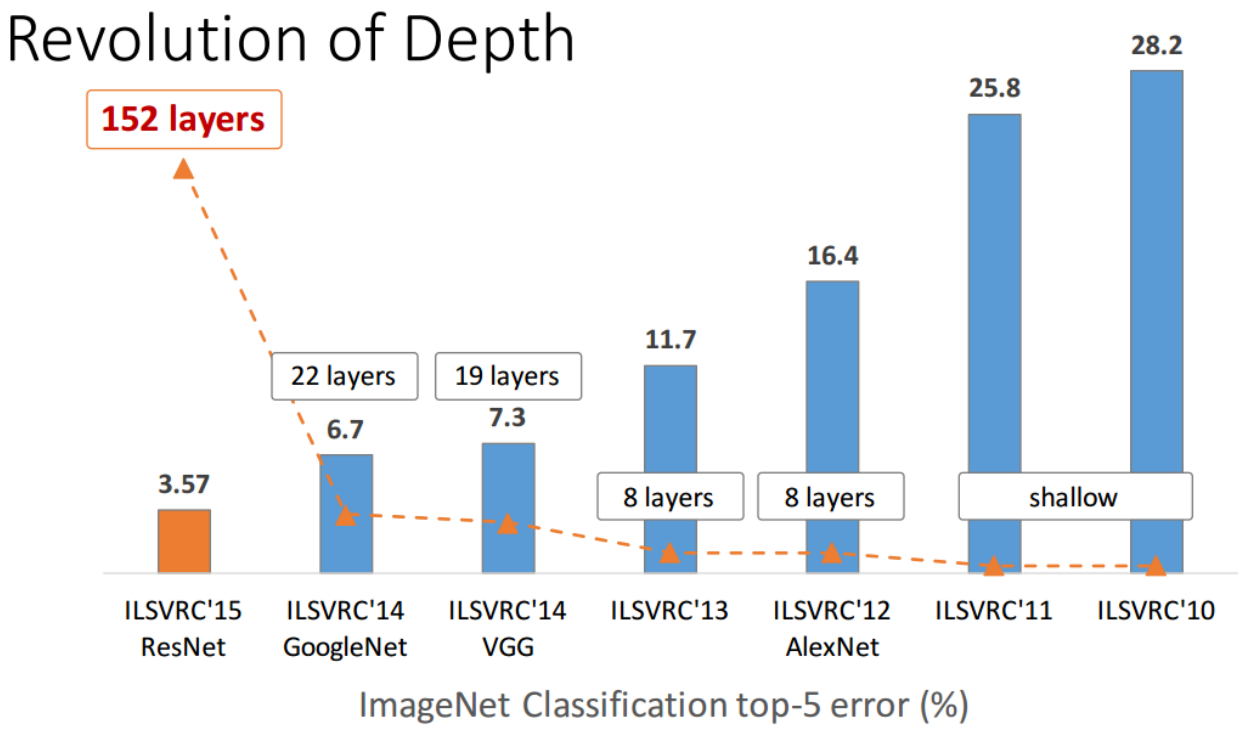
Many layers

Features are *learned*, not given

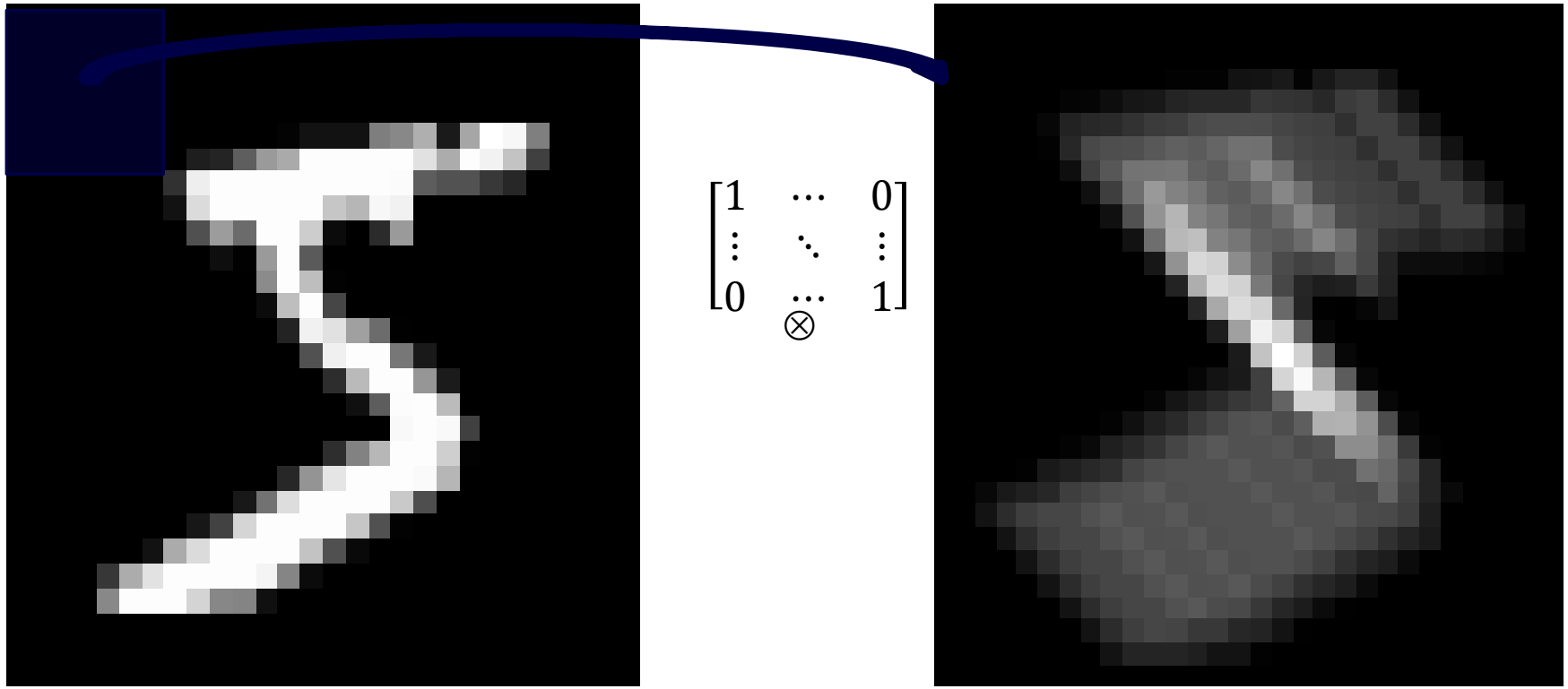
Low-level features combined into high-level features

Special types of layers

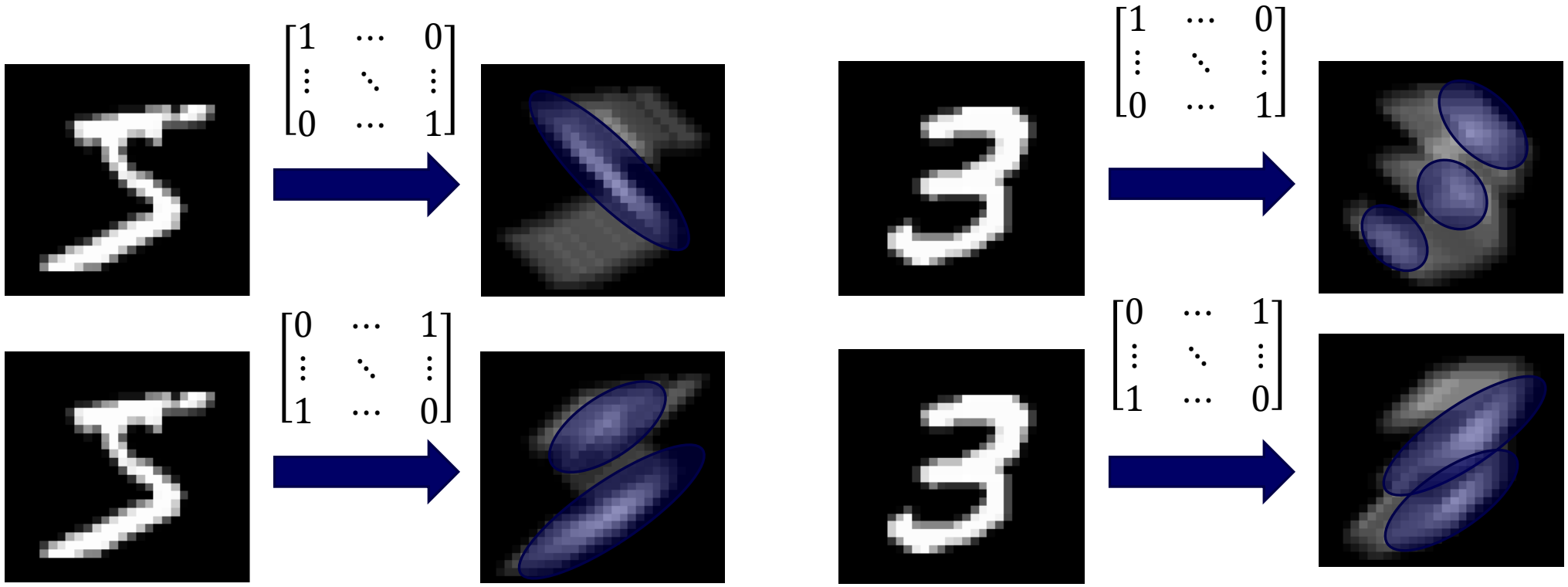
- convolutional
- drop-out
- recurrent
- ...



Convolutional neural networks



Convolution examples

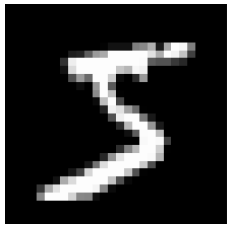


Convolution: 050_convolution.ipynb

Second approach

- Data preprocessing
  - Input data as 2D array
  - output data as array with one-hot encoding

- Model: convolutional neural network (CNN)
  - 28×28 inputs
  - CNN layer with 32 filters 3×3
  - ReLU activation function
  - flatten layer
  - dense layer with 10 units
  - SoftMax activation function



array([[0.0, 0.0,..., 0.951, 0.533,..., 0.0, 0.0]], dtype=float32)

5

array([0, 0, 0, 0, 0, 1, 0, 0, 0, 0], dtype=uint8)

Convolutional neural network: 060_mnist_cnn.ipynb

Task: sentiment classification

Input data

- movie review (English)

Output data

Training examples

Test examples

<start> this film was just brilliant casting location scenery story direction everyone's really suited the part they played and you could just imagine being there Robert redford's is an amazing actor and now the same being director norman's father came from the same scottish island as myself so i loved the fact there was a real connection with this film the witty remarks throughout the film were great it was just brilliant so much that i bought the film as soon as it …

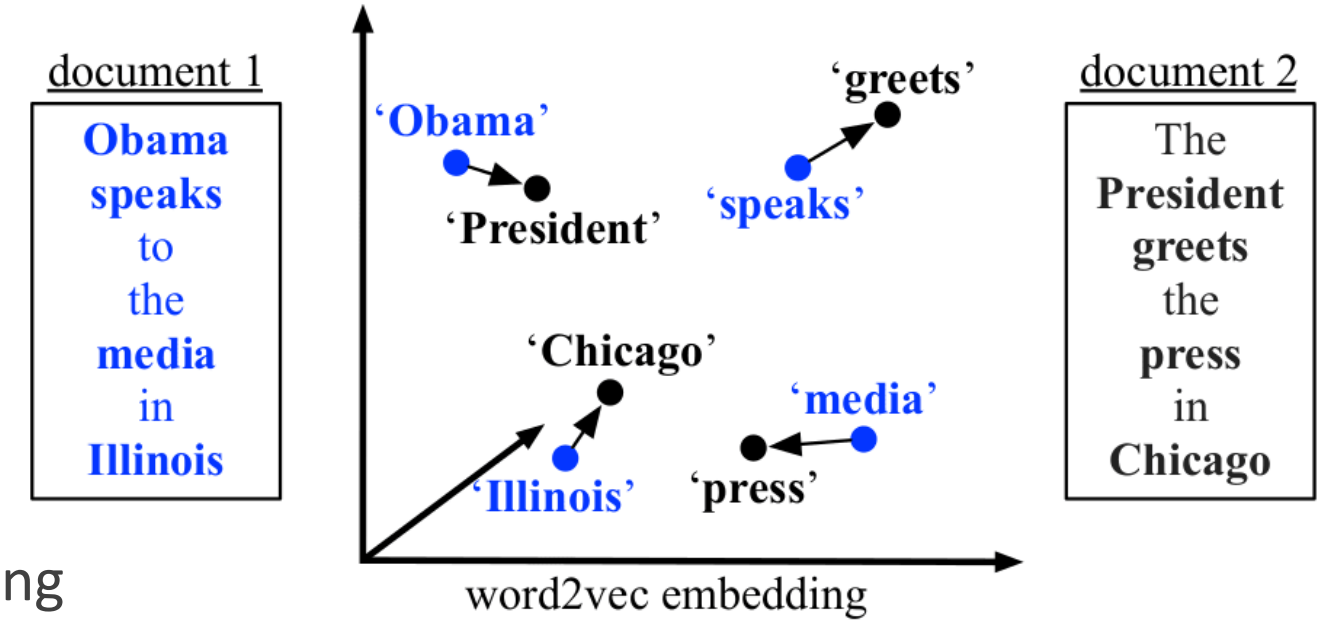
Explore the data: 070_imdb_data_exploration.ipynb

Word embedding

Issues:

- unwieldy
- no semantics

- Represent words as one-hot vectors
  length = vocabulary size
- Word embeddings
  - dense vector
  - vector distance ≈ semantic distance
- Training
  - use context
  - discover relations with surrounding words



How to remember?

Manage history, network learns

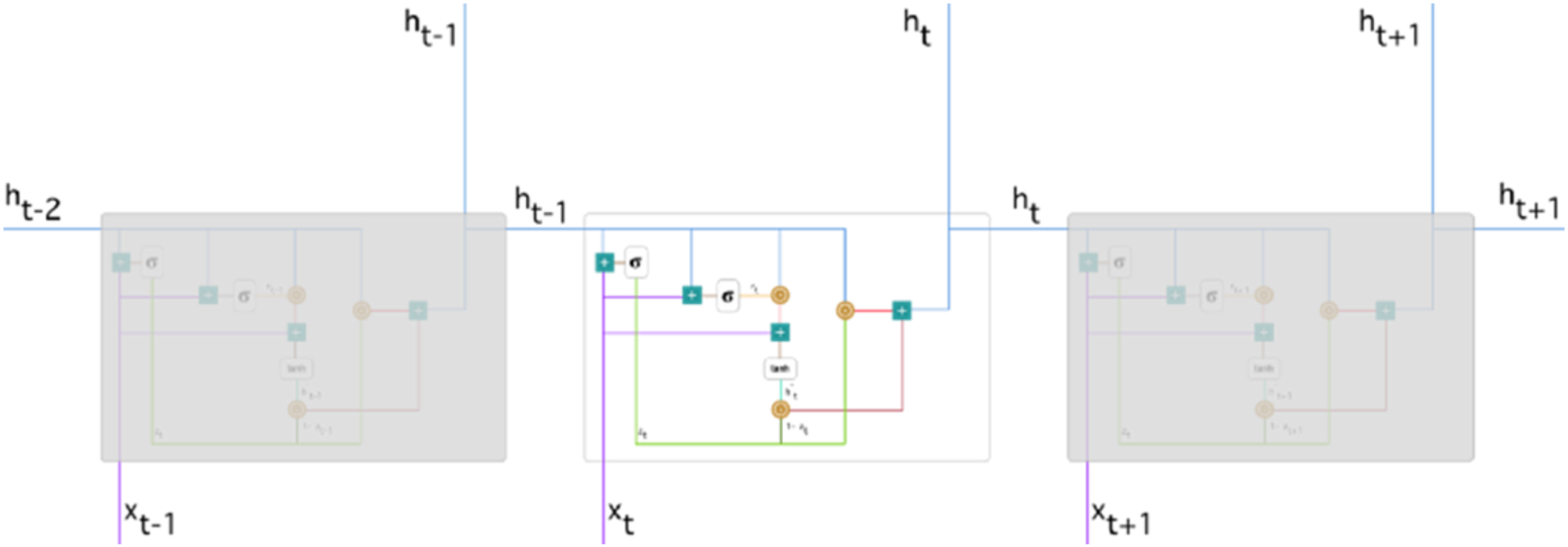
- what to remember
- what to forget

Long-term correlations!

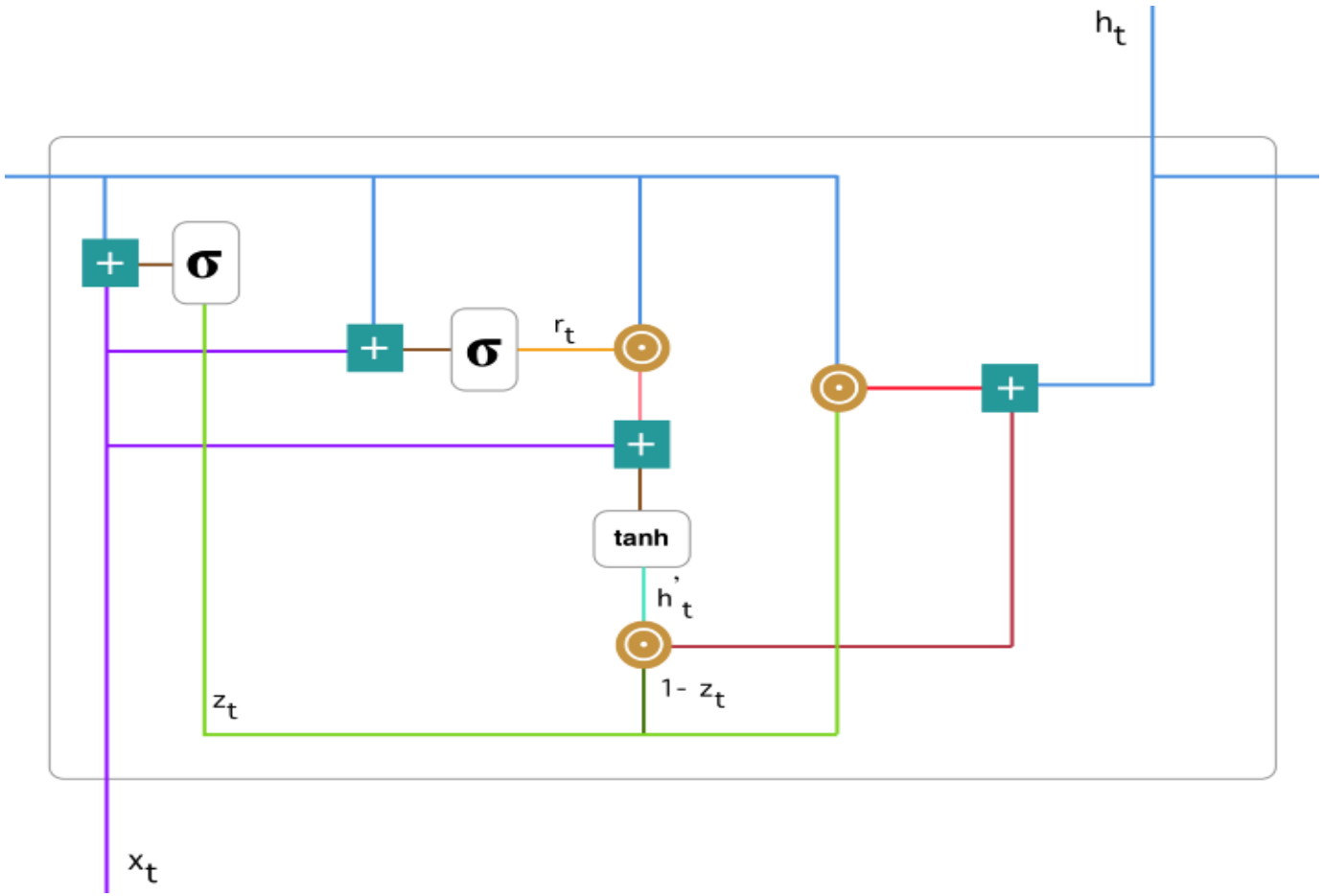
Use, e.g.,

- LSTM (Long Short-Term Memory
- GRU (Gated Recurrent Unit)

Deal with variable length input and/or output


h_{t-1}
h_t
h_{t+1}
h_{t-2}
h_{t-1}
h_t
h_{t+1}
x_{t-1}
x_t
x_{t+1}

Gated Recurrent Unit (GRU)



$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$h'_t = \tanh(W x_t + r_t \odot U h_{t-1})$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t$$

- Update gate
- Reset gate
- Current memory content
- Final memory/output

Approach

- Data preprocessing
  - Input data as padded array
  - output data as 0 or 1
- Model: recurrent neural network (GRU)
  - 100 inputs
  - embedding layer, 5,000 words, 64 element representation length
  - GRU layer, 64 units
  - dropout layer, rate = 0.5
  - dense layer, 1 output
  - sigmoid activation function

Recurrent neural network: 080_imdb_rnn.pynb

Tell a Story

LECTURE 7

Overview [Code.org U9L7/L8]

- You will create an effective visualization that you will use to tell a data story.