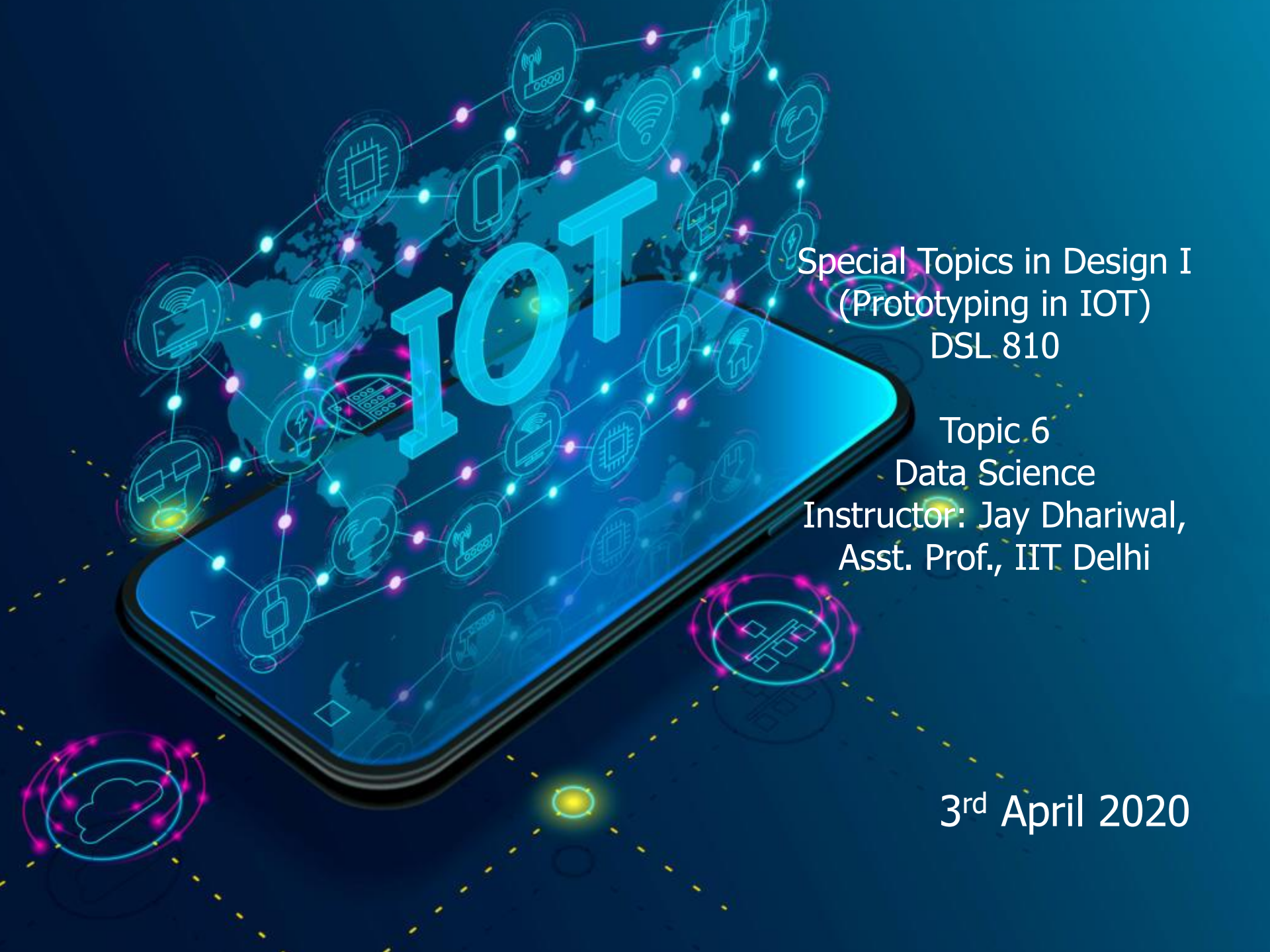


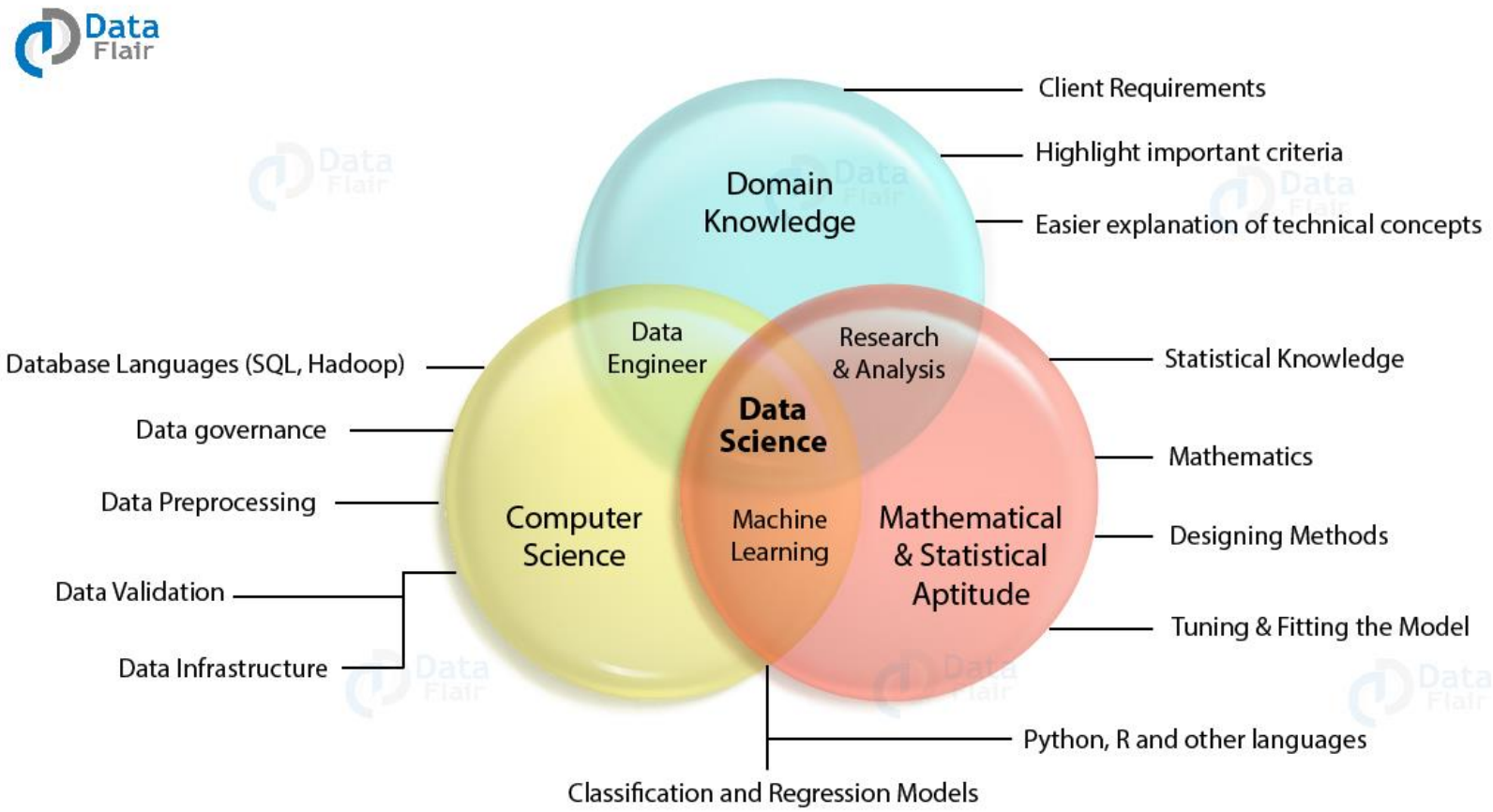
Special Topics in Design I (Prototyping in IOT) DSL 810

Topic 6
Data Science

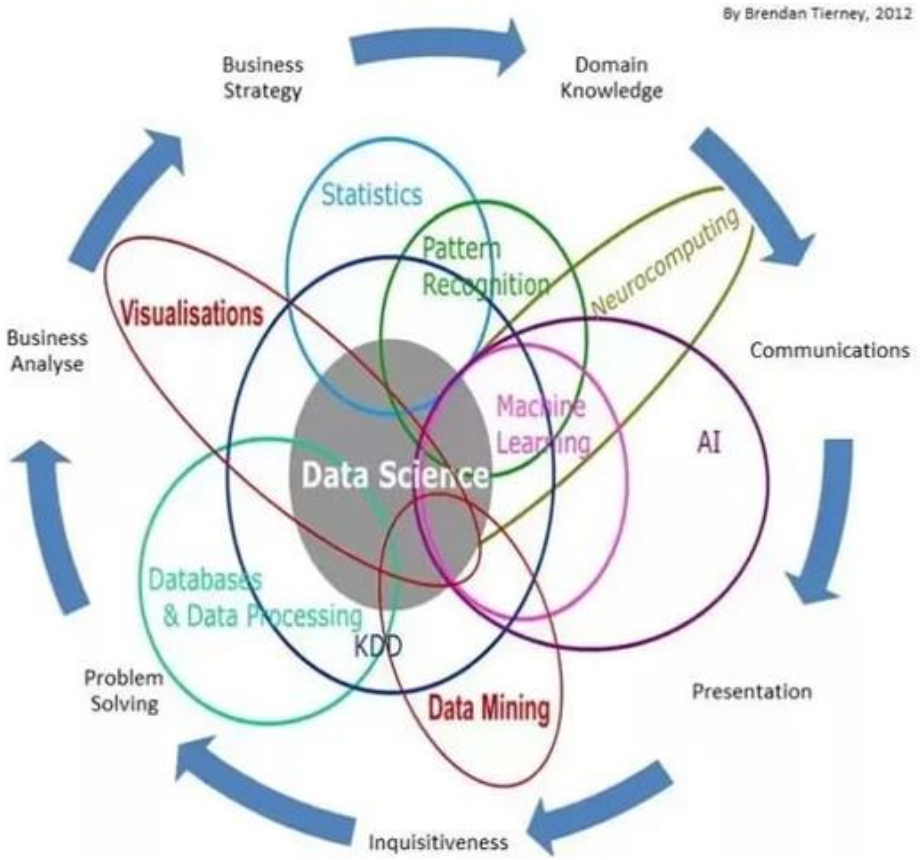
Instructor: Jay Dhariwal,
Asst. Prof., IIT Delhi

3rd April 2020

Data Science: understand and analyze actual phenomena with data



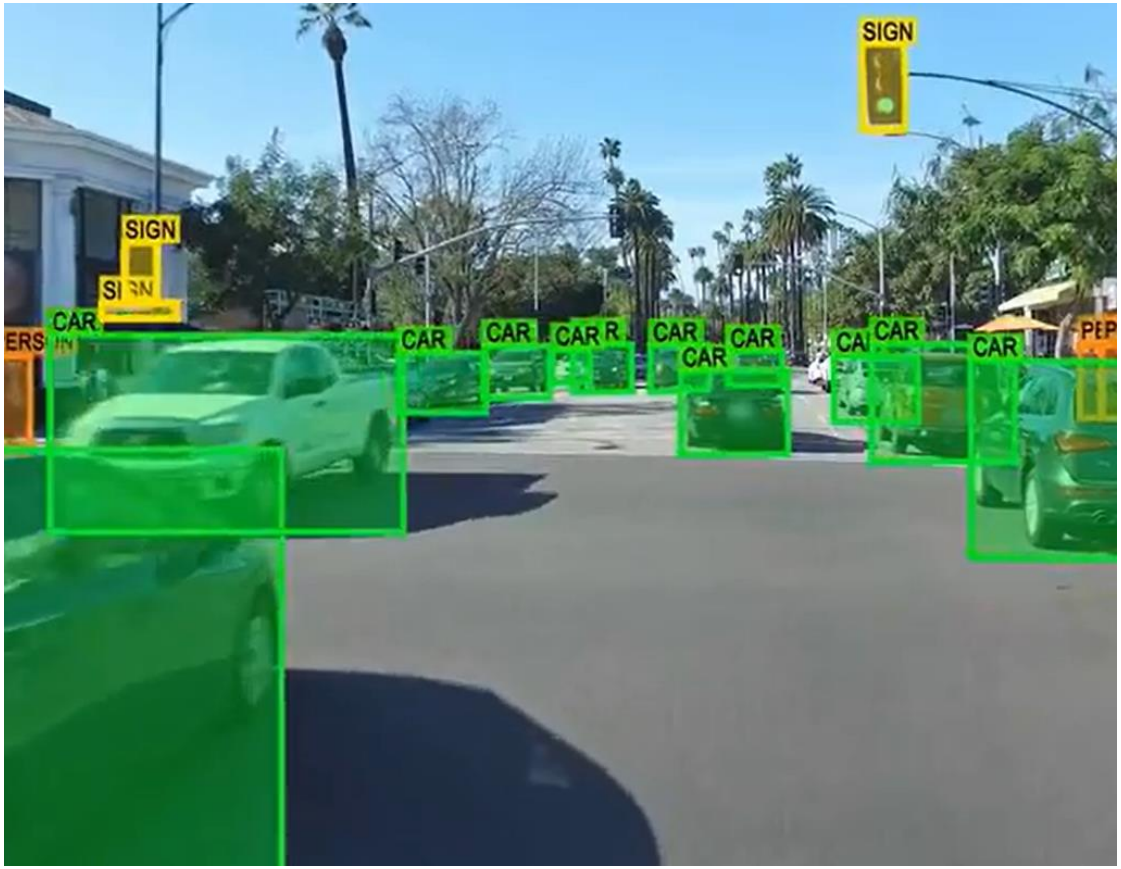
Data Science is multi-disciplinary



[Source]

Have you used AI/ML before?

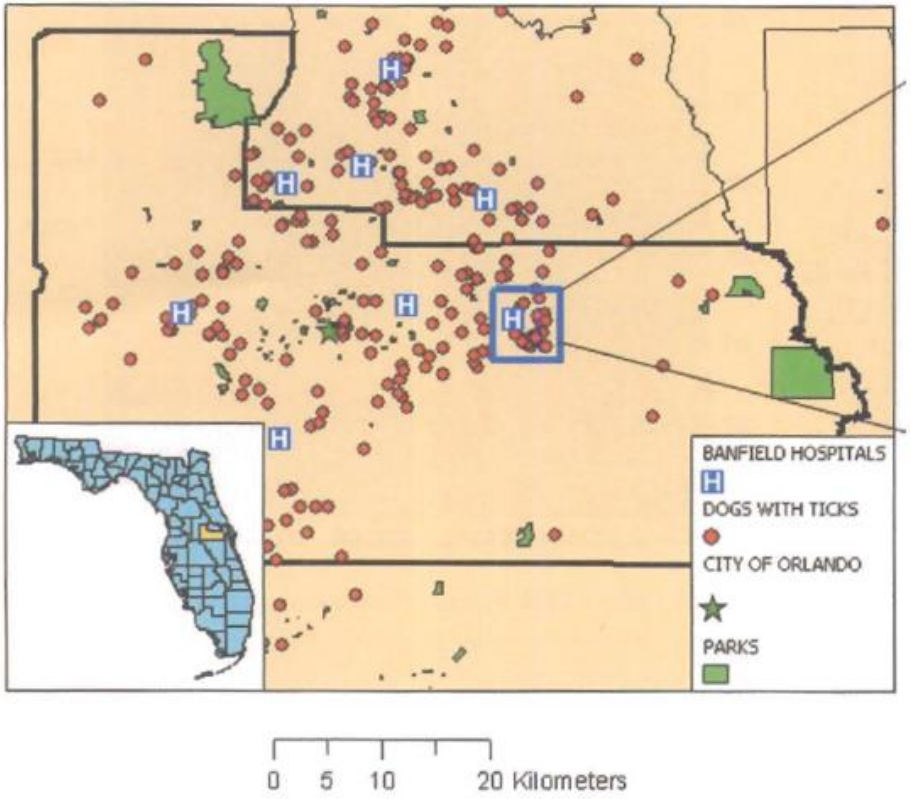
- Siri, Google Assistant, Alexa.
- Amazon suggestions to buy stuff, Film suggestions on Netflix



Why should you care about AI/ML?

- AI vs. ML
- Google's AI AlphaGo Is Beating Humanity At Its Own Games
- Elon Musk's concerns about Artificial Intelligence
- Eric Schimdt: Former Chairman, Alphabet (parent company of Google): Self Driving Cars are the future, AI assisted health care.
- Vinod Khosla about Generative Design (CAD+AI). Bike example

Case studies in Data Science

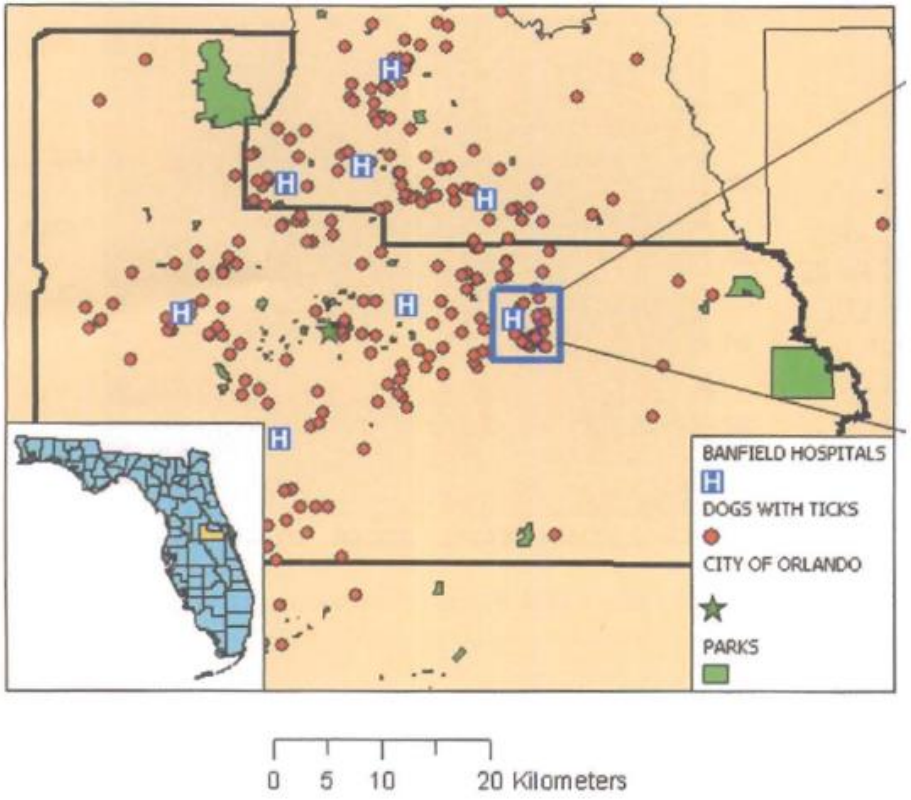


- Spatio-temporal clusters for early epidemic detection
- Data Science for COVID-19
- NodeMCU + sensors + ThingSpeak
- Big Data analysisn from MIT North Court study
- Machine Learning for Building simulation
- Marta González - Mobile Data for Urban Transformation
- Sensor data from smart phones in MATLAB

Moore G.E., Ward M.P., **Dhariwal J.**, Wu C.C., Glickman N.W., Lewis H.B., Glickman L.T., 'Development of a national companion animal syndromic surveillance system for bioterrorism', *2nd International Conference on the Applications of GIS and Spatial Analysis to Veterinary Science (GISVET 04), Univ. Guelph, Ontario, Canada*, Durr, P. A. and Martin, S. W., Jun 2004.

Understanding congested travel in urban areas Serdar Çolak, Antonio Lima & Marta C. González *Nature Communications* **volume 7**, Article number: 10793 (2016)

Spatio-temporal clusters for early epidemic detection



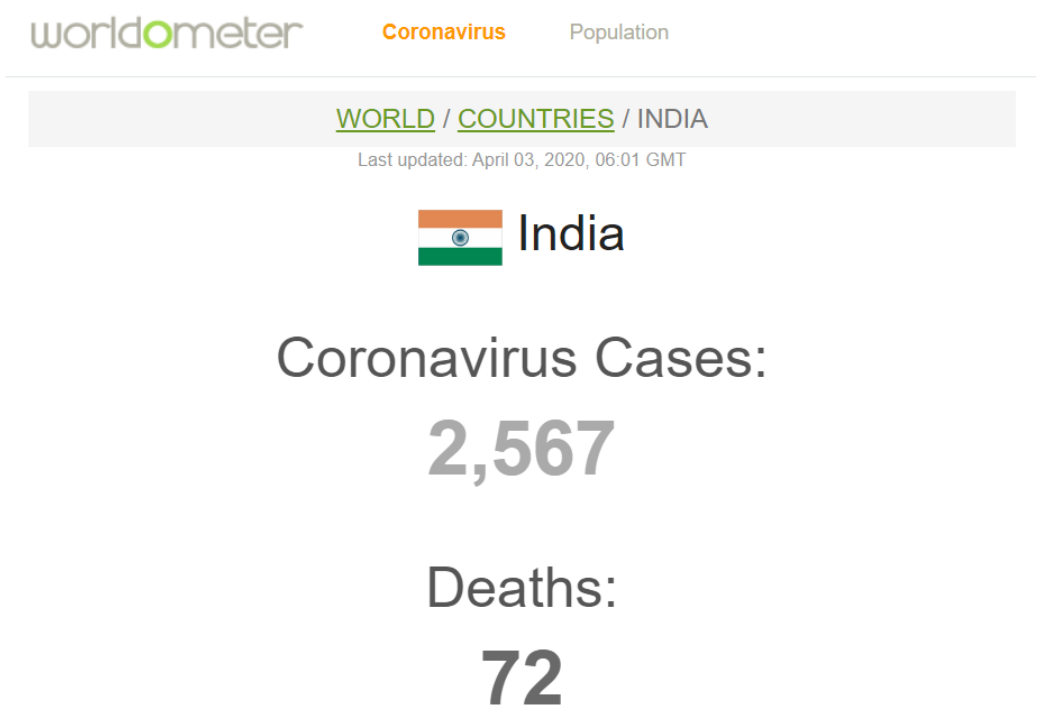
Data for Orlando, Florida

- Banfield Chain of Pet Hospitals with practices all over USA
- Electronic real-time database for dogs and cats
- Disease events statistically random or unusually high
- Surveillance system
- ArcGIS for spatial data
- SAS for temporal data
- SATSCAN for Spatio-temporal cluster analysis
- Vets would have a local picture vs. Surveillance system have the bigger picture

Moore G.E., Ward M.P., **Dhariwal J.**, Wu C.C., Glickman N.W., Lewis H.B., Glickman L.T., 'Development of a national companion animal syndromic surveillance system for bioterrorism', *2nd International Conference on the Applications of GIS and Spatial Analysis to Veterinary Science (GISVET 04), Univ. Guelph, Ontario, Canada*, Durr, P. A. and Martin, S. W., Jun 2004.

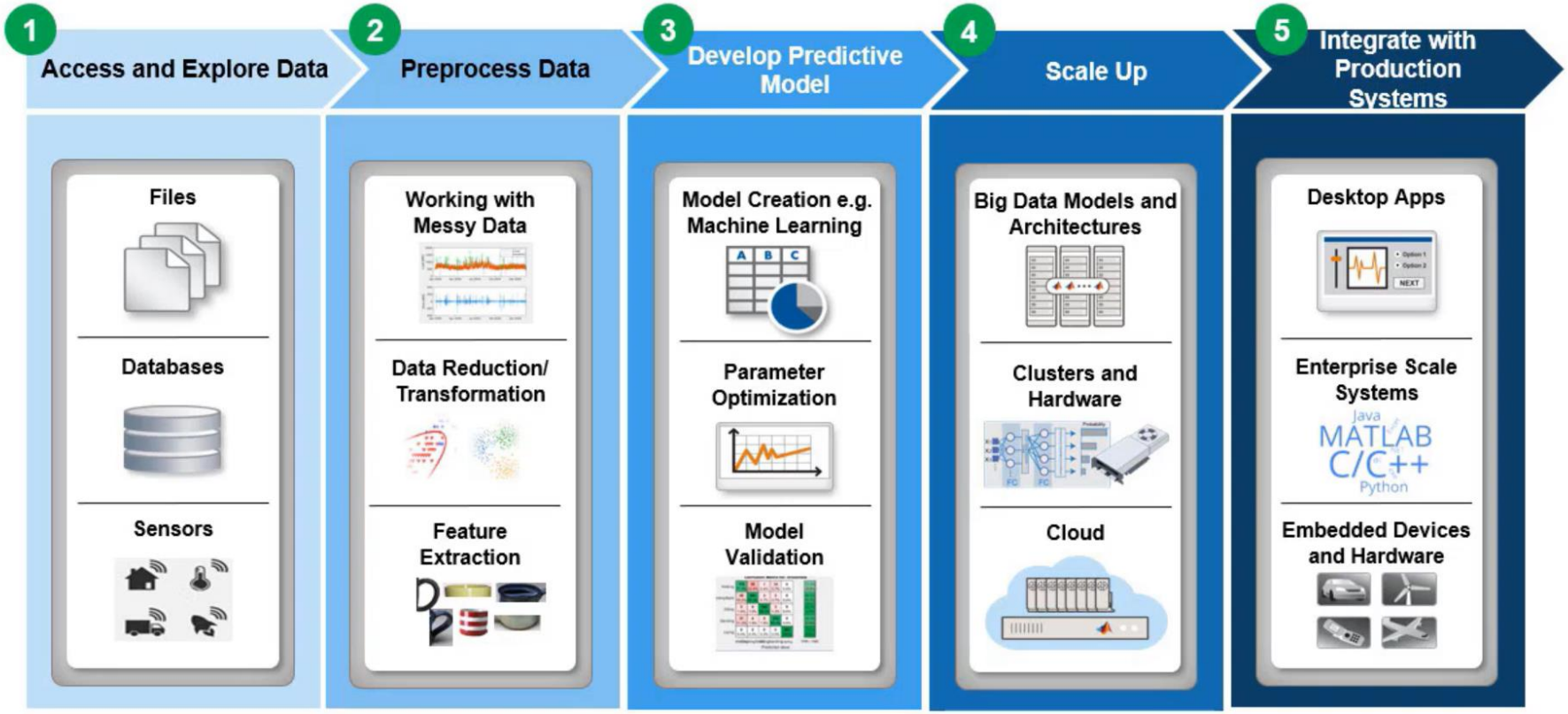
Data Science for COVID-19 Trend analysis

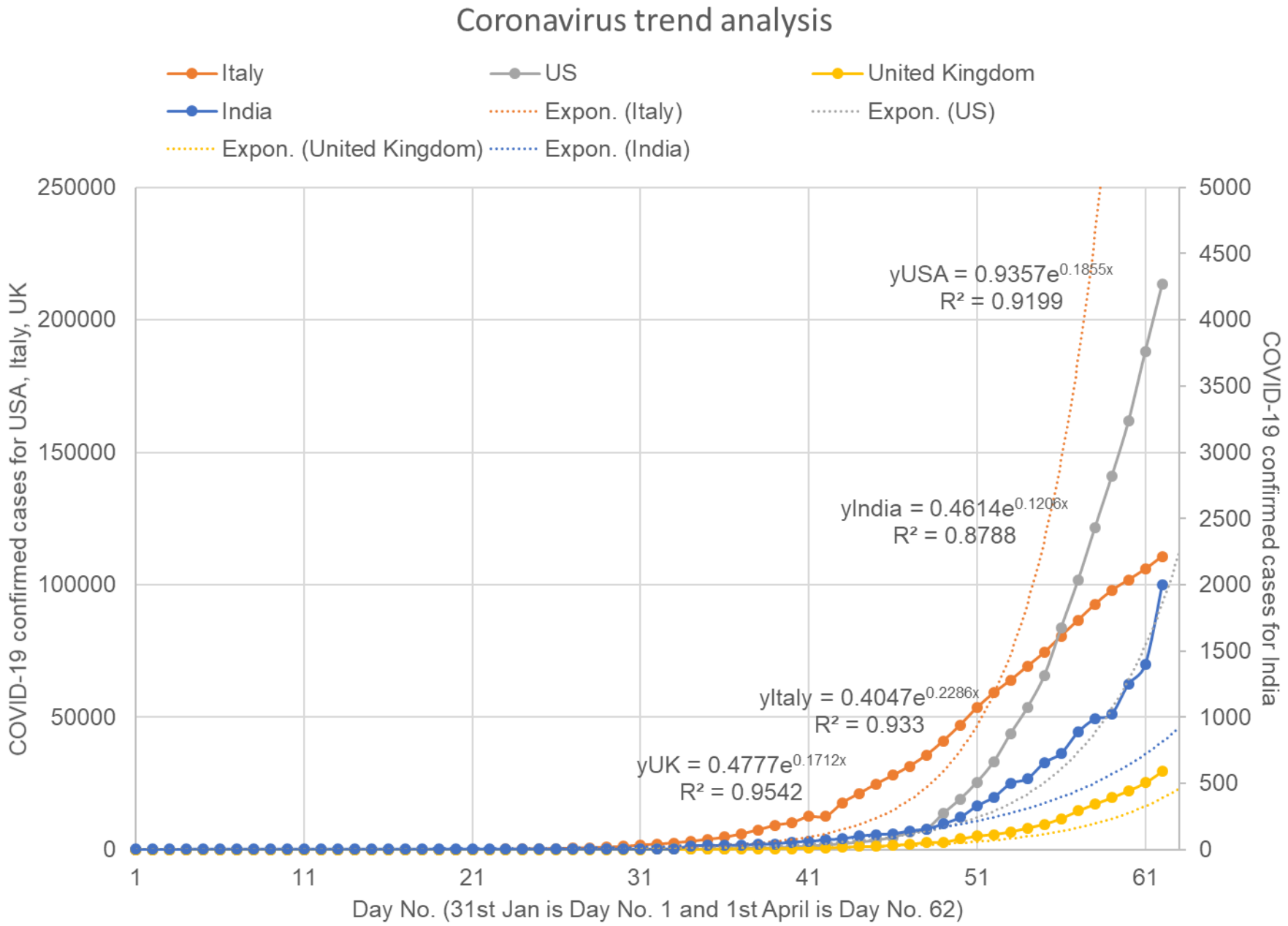
- Forecast COVID-19 cases for India, UK, USA in the next two weeks
- Help with lockdown policies, building up capacity to deal with this crisis



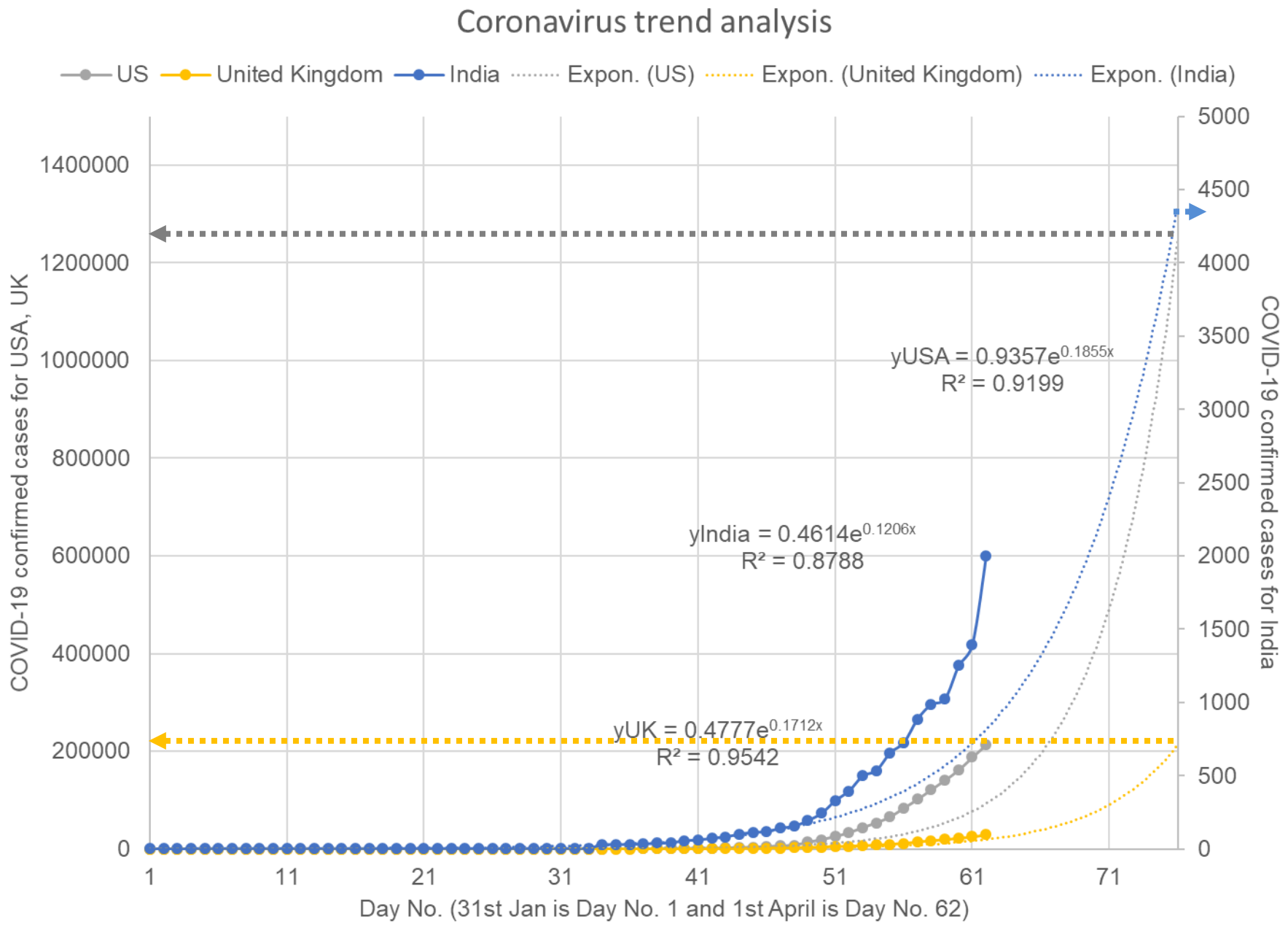
Data science analysis workflow

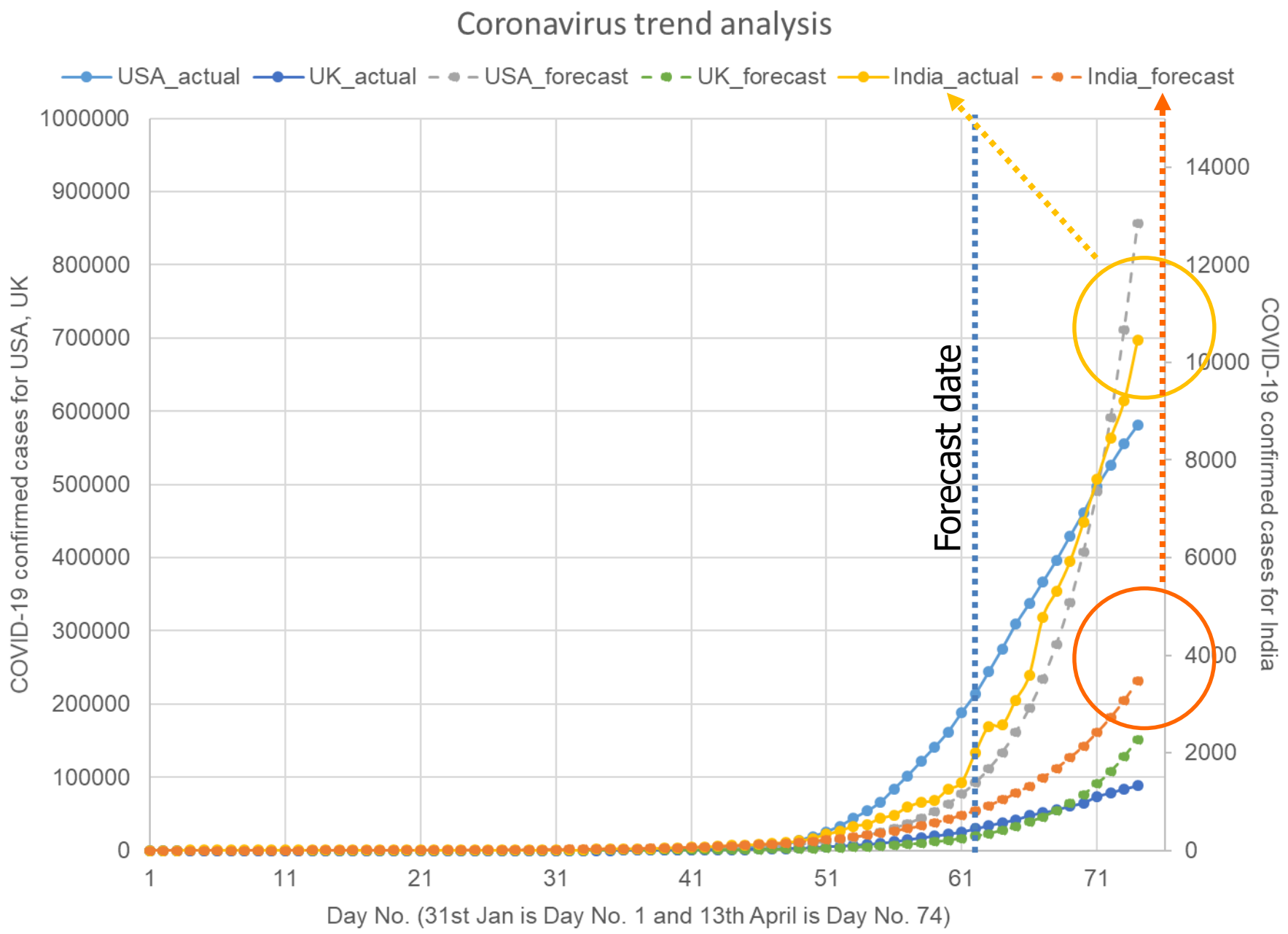
Source: Data Science with MATLAB





Data source: Johns Hopkins CSSE, Visual dashboard


Coronavirus trend analysis
US
United Kingdom
India
Expon. (US)
Expon. (United Kingdom)
Expon. (India)
COVID-19 confirmed cases for USA, UK
COVID-19 confirmed cases for India
Day No. (31st Jan is Day No. 1 and 1st April is Day No. 62)
yUSA = 0.9357e^0.1855x
R² = 0.9199
yIndia = 0.4614e^0.1206x
R² = 0.8788
yUK = 0.4777e^0.1712x
R² = 0.9542
1400000
1200000
1000000
800000
600000
400000
200000
0
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
11
21
31
41
51
61
71


Coronavirus trend analysis
USA_actual
UK_actual
USA_forecast
UK_forecast
India_actual
India_forecast
Forecast date
COVID-19 confirmed cases for USA, UK
COVID-19 confirmed cases for India
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
14000
12000
10000
8000
6000
4000
2000
0
1
11
21
31
41
51
61
71
Day No. (31st Jan is Day No. 1 and 13th April is Day No. 74)

Why is COVID-19 spreading the way it is?

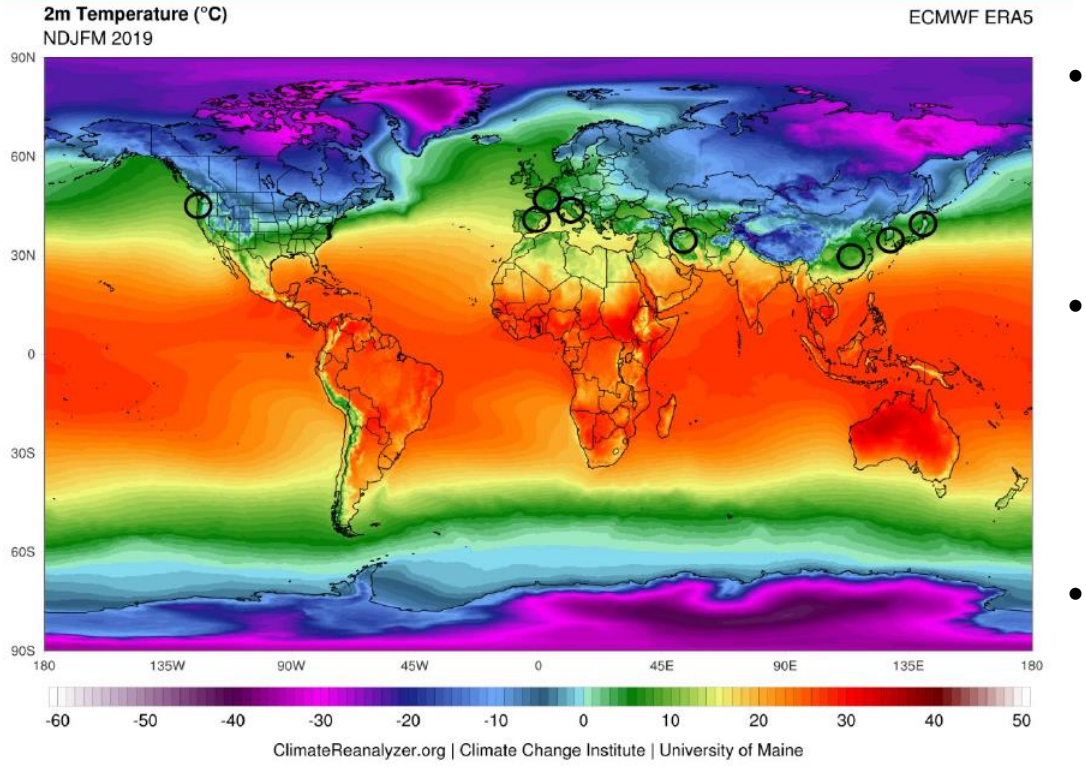
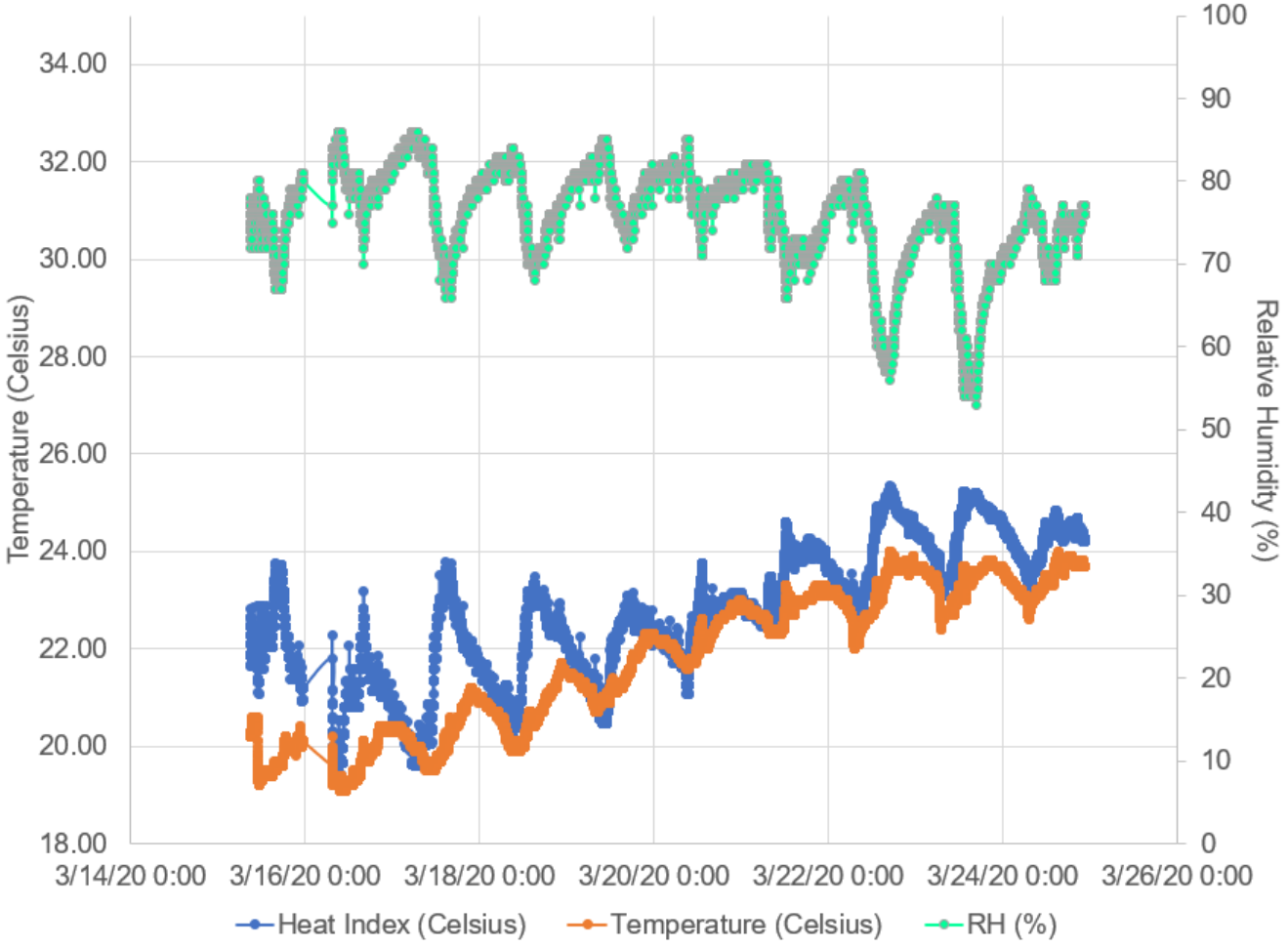


Figure 1. World temperature map November 2018-March 2019. Color gradient indicates 2-meter temperatures in degrees Celsius. Black circles represent countries with significant community transmission (≥ 10 deaths as of March 10, 2020). Image from Climate Reanalyzer (https://ClimateReanalyzer.org), Climate Change Institute, University of Maine, USA.

- Regions between 30 ° to 50 ° Latitude had the Temperature (5-11 °C), Rh (44-84%) in Jan, Feb 2020 for virus to spread
- Predictions for March and April from this paper – The virus spread would potentially move northwards to UK, North-Eastern USA, Manchuria belt
- Can we use this domain knowledge to have predictions for India using machine learning?

Source: Sajadi, Mohammad M. and Habibzadeh, Parham and Vintzileos, Augustin and Shokouhi, Shervin and Miralles-Wilhelm, Fernando and Amoroso, Anthony, Temperature, Humidity and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19 (March 5, 2020)

Heat Index profile for my room
NodeMCU + DHT11 + ThingSpeak

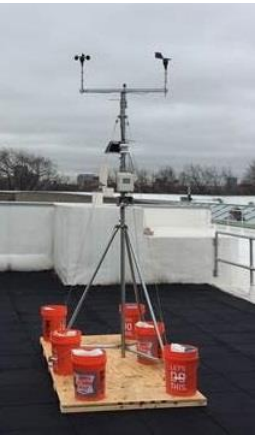


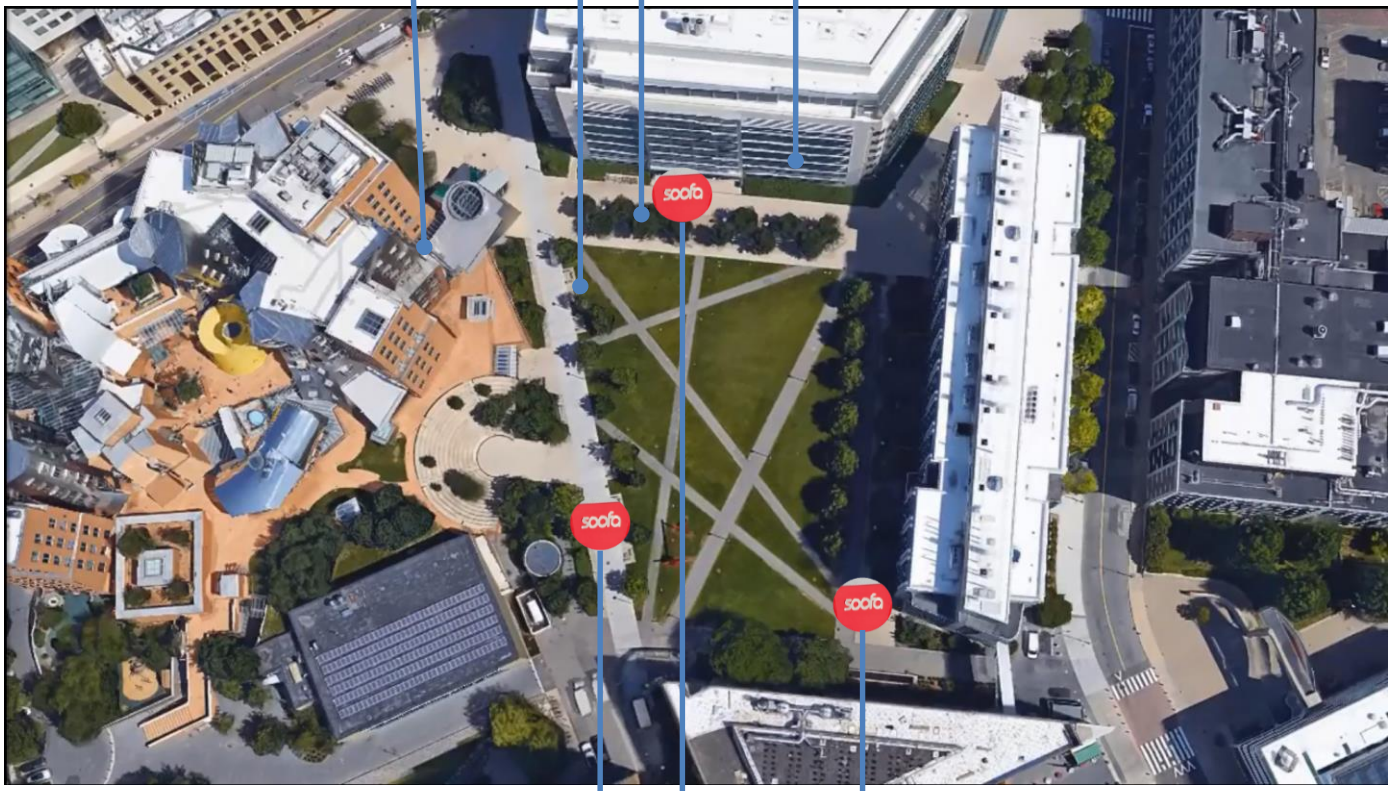
Observations

1. 48500 data points
2. T, Rh every 18 sec for 10 days
3. Cyclical pattern
4. Daily Temperature Increase
5. In the context of COVID-19, this analysis helps to know what Temp, Rh to avoid which is conducive for the virus
6. Assignment 5 on thermal comfort

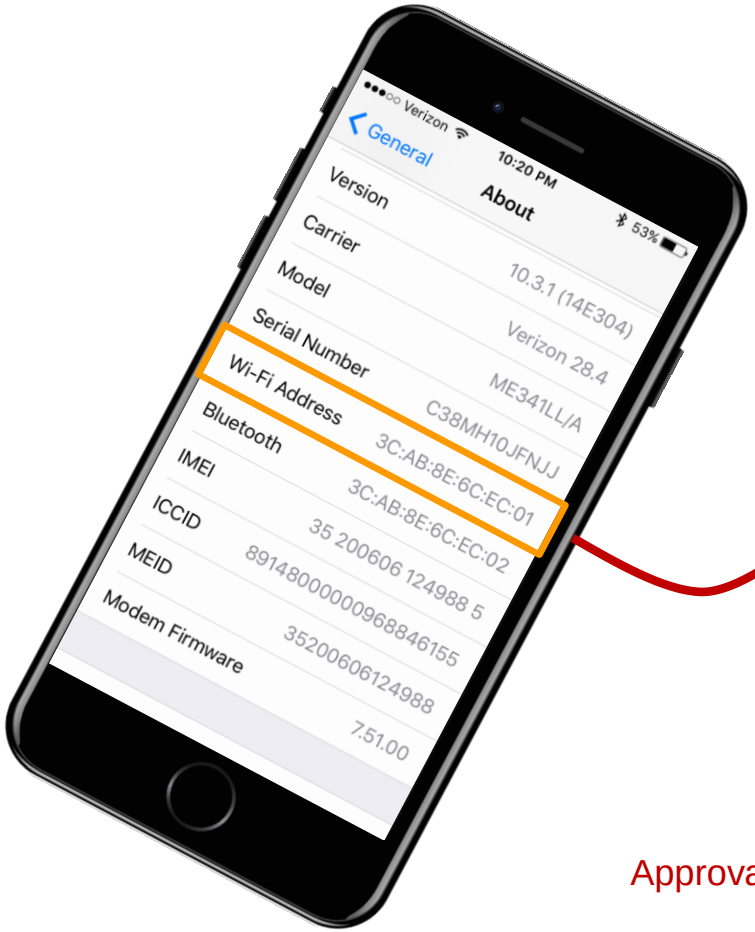
Big Data Analytics: Design of Outdoor Public Spaces

Reinhart C., Dhariwal J. and Gero K., 'Biometeorological indices explain outside dwelling patterns based on Wi-Fi data in support of sustainable urban planning', *Building and Environment*, 126, 2017, 422–430.




Outdoor seating
Koch cafeteria
Stata Center cafeteria
soofa
soofa
soofa
Soofa benches
Weather station at roof of nearby building

Privacy in the Modern Age

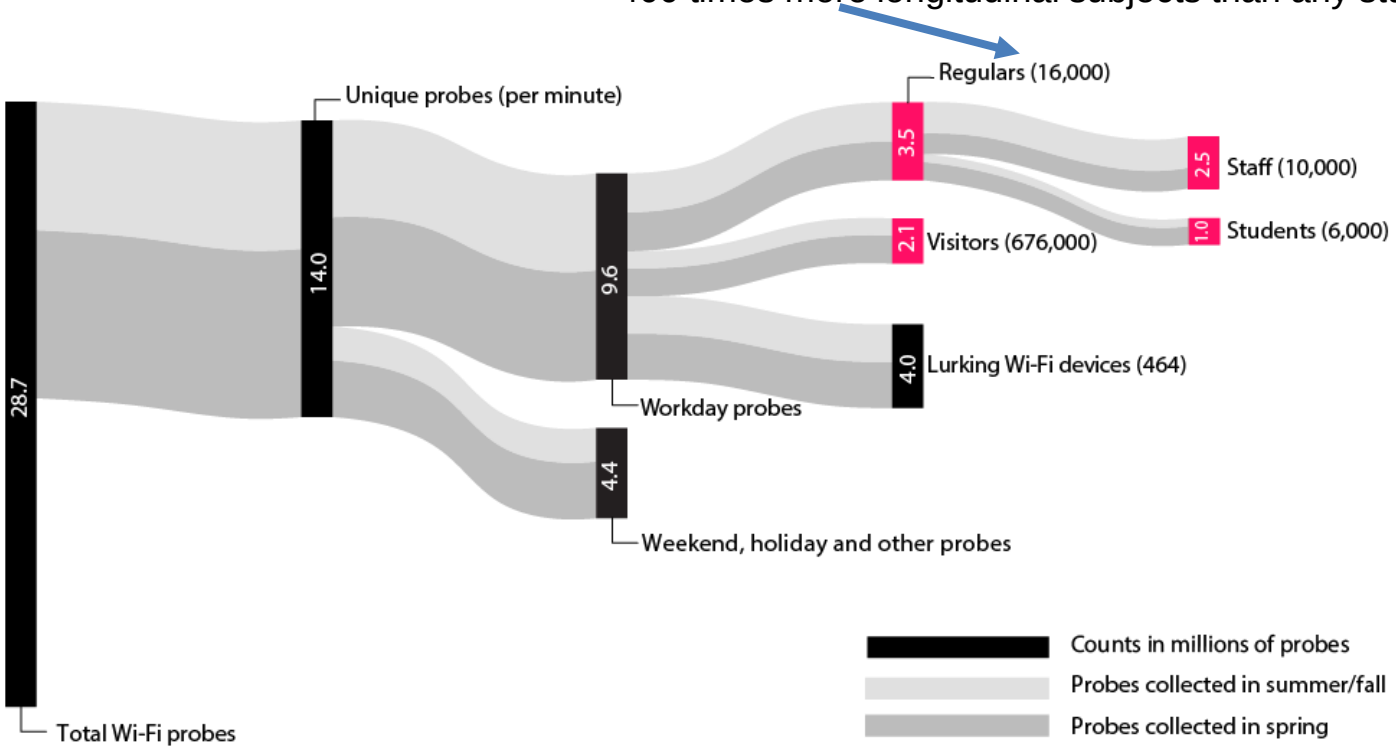


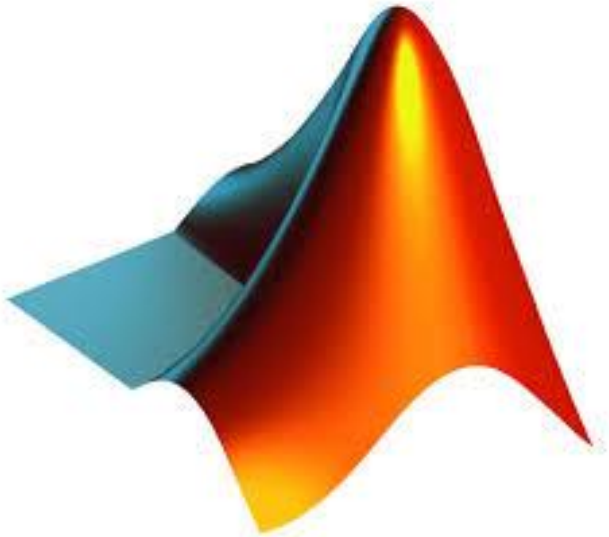
Encrypted device ID

Approvals from COUHES at MIT

Results from July 2016 – May 2017

400 times more longitudinal subjects than any study in the past

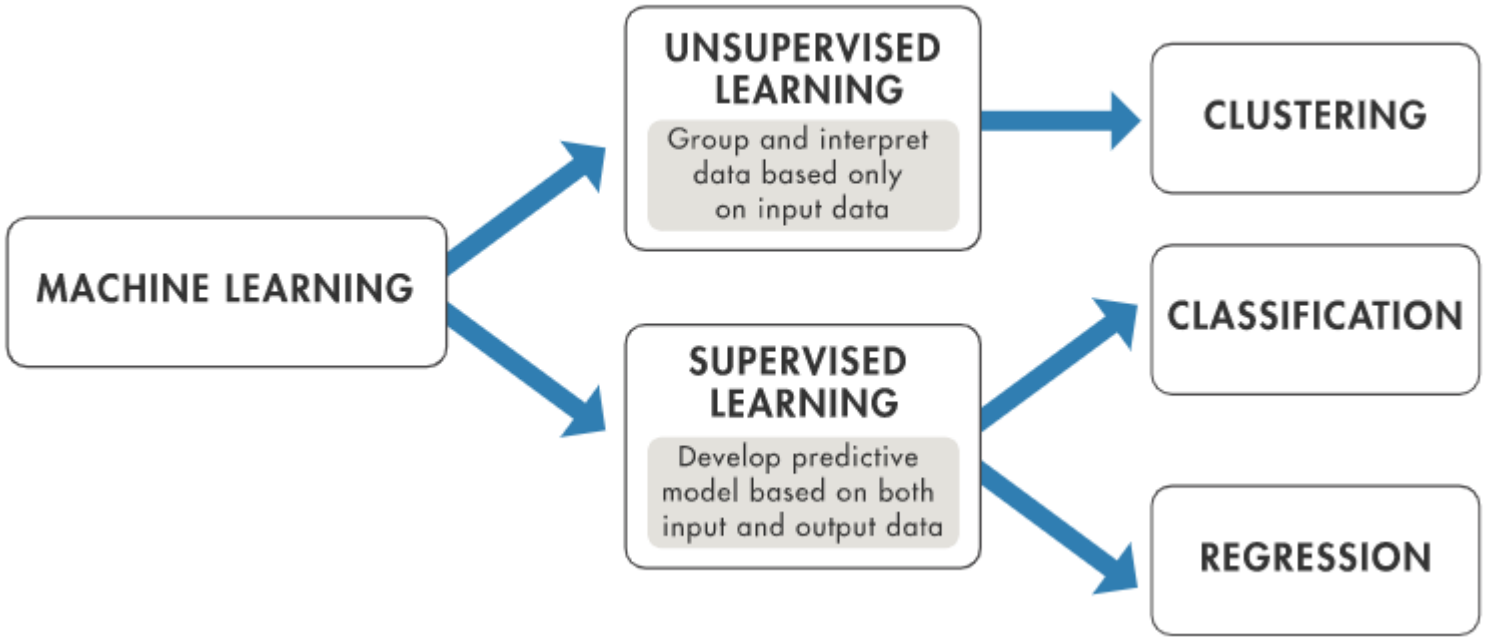




Data science using MATLAB

- Download MATLAB. IIT Delhi student license
- Get started with MATLAB
- Data science for COVID-19: trend analysis, effect of Temp/ Humidity and other variables to predict COVID-19 spread
- Machine learning using sensor data from smart phones in MATLAB

Machine Learning Techniques



Source: MATLAB

Regression Learner Toolbox in MATLAB

Workflow

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV) — Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Residuals, Error, ϵ: $N(0, \sigma^2)$

Steps in MATLAB

Regression Learner Toolbox in MATLAB

Regression Analysis on Life Expectancy

Dataset for Linear Regression

Life expectancy = f(Birth Rate, Cancer Rate, Dengue Cases,
Environmental Performance Index (EPI), Gross Domestic Product (GDP),
Health Expenditure, Heart Disease Rate, Population, Area, Pop Density, Stroke Rate)

Model:

Life expectancy = 75.4657 – 0.4074*Birth Rate + 0.1642*EPI – 0.0581*Stroke Rate

Improvement Focus based on the Model Predictions

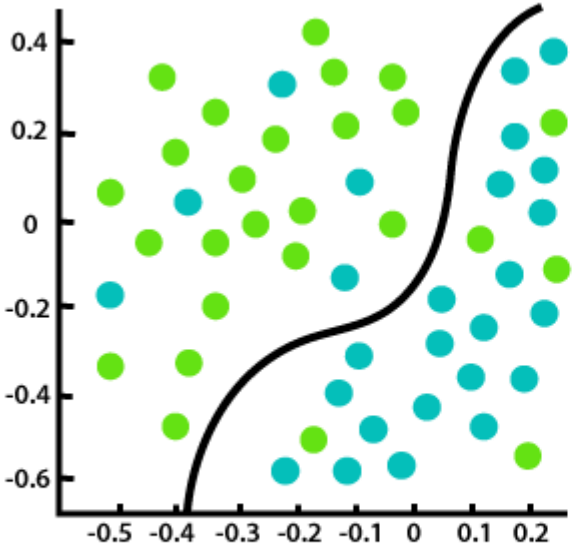
Country	Birth Rate	EPI	Stroke Rate	Life expectancy
India	21.8	30.57	71.48	67.45
India	10	40	50	75.05

Data Science with R

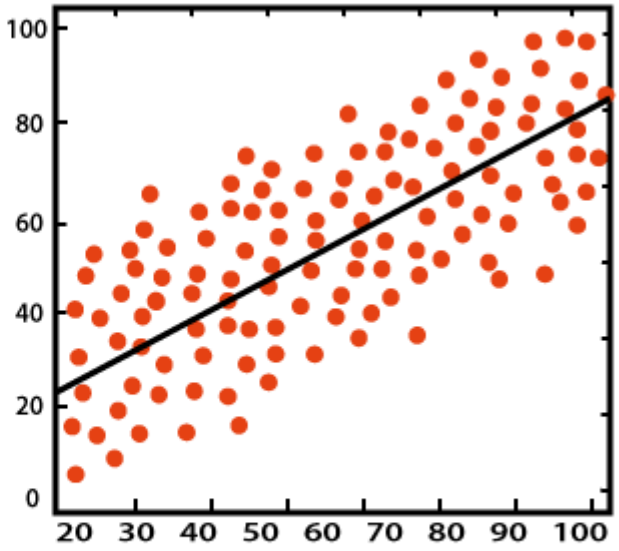
- Free and open source
- Download R, RStudio
- edX course on Data Science: R Basics
- Another course on R Basics
- Machine Learning in R step-by-step

Classification

- Getting excited about it – real world examples
- Applying classification models to datasets



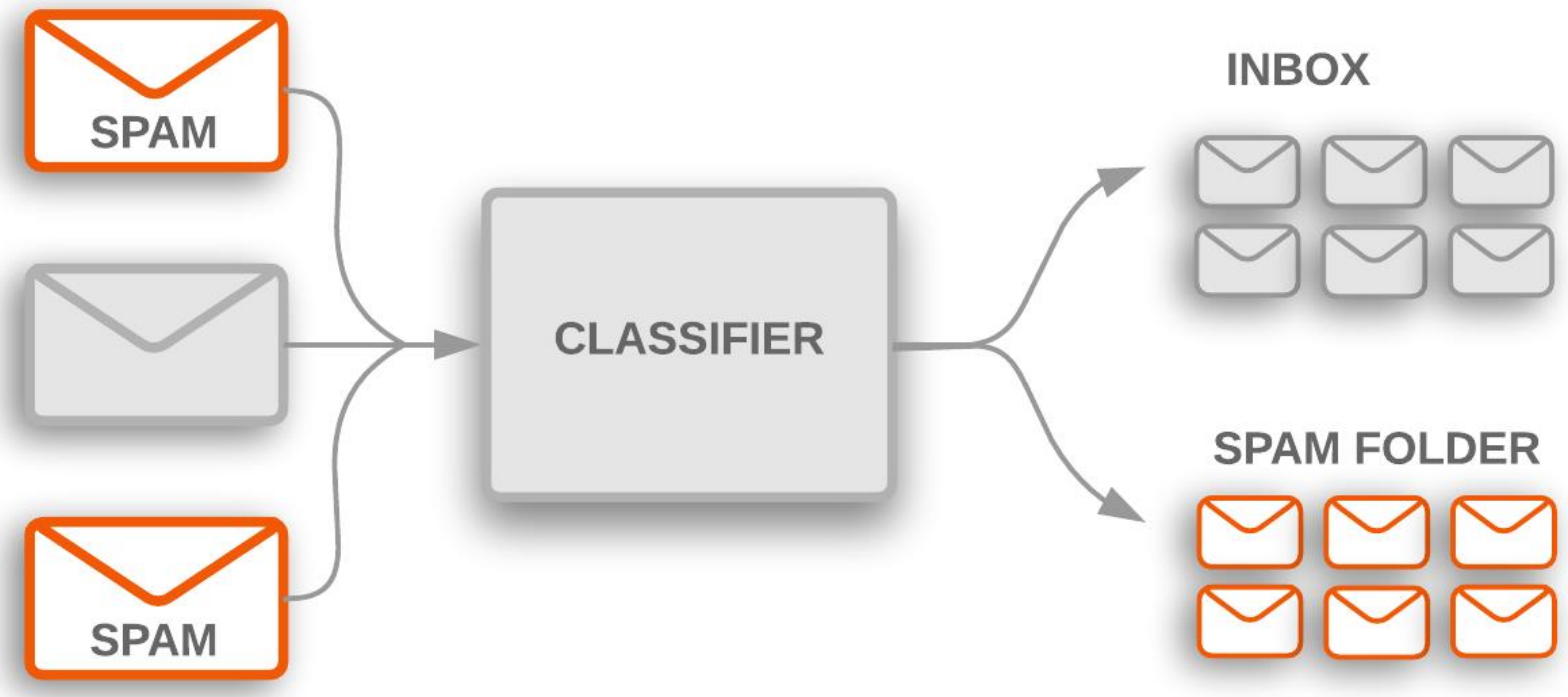
Classification



Regression

[Classification vs. Regression]

Text classification



Source

Image Classification

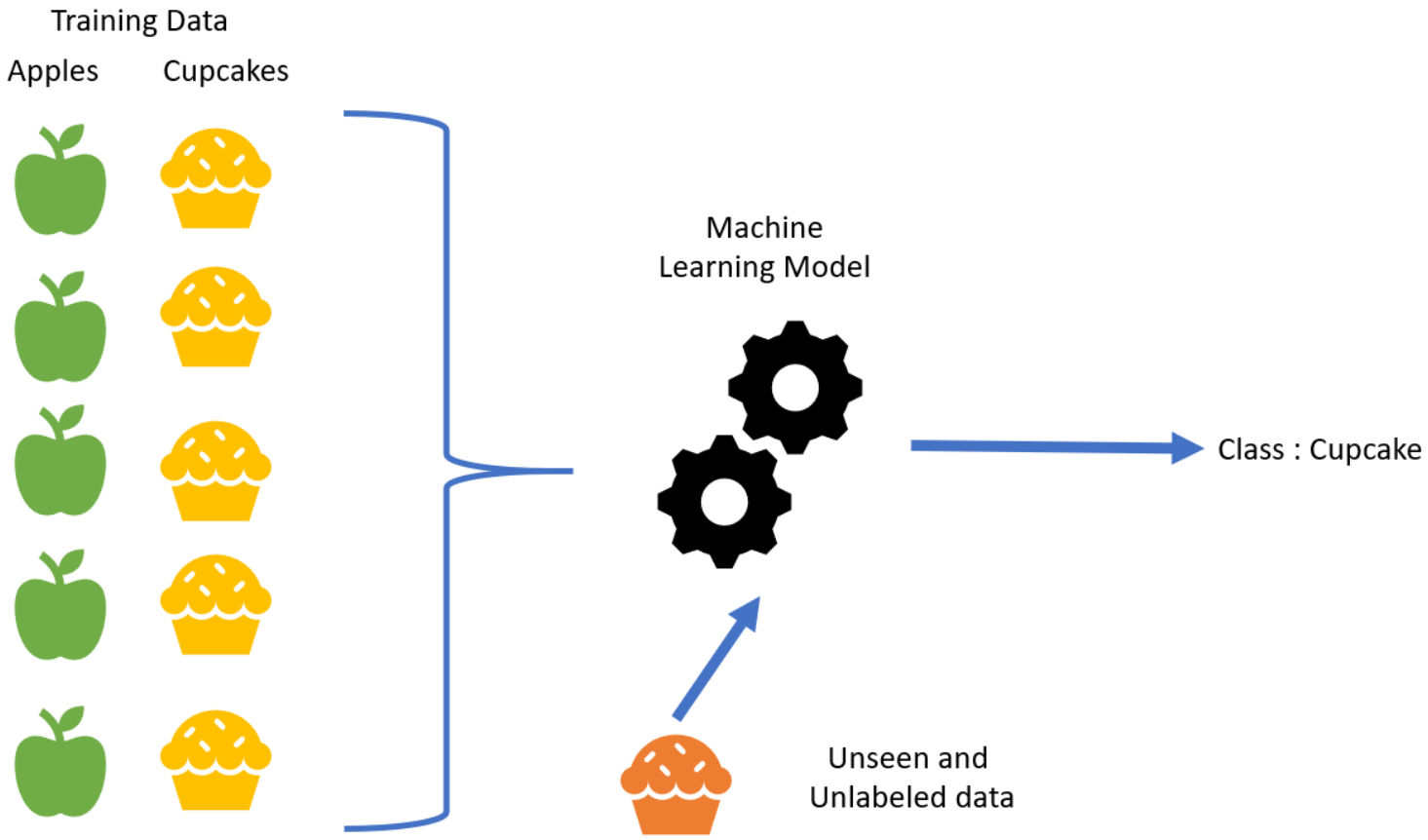
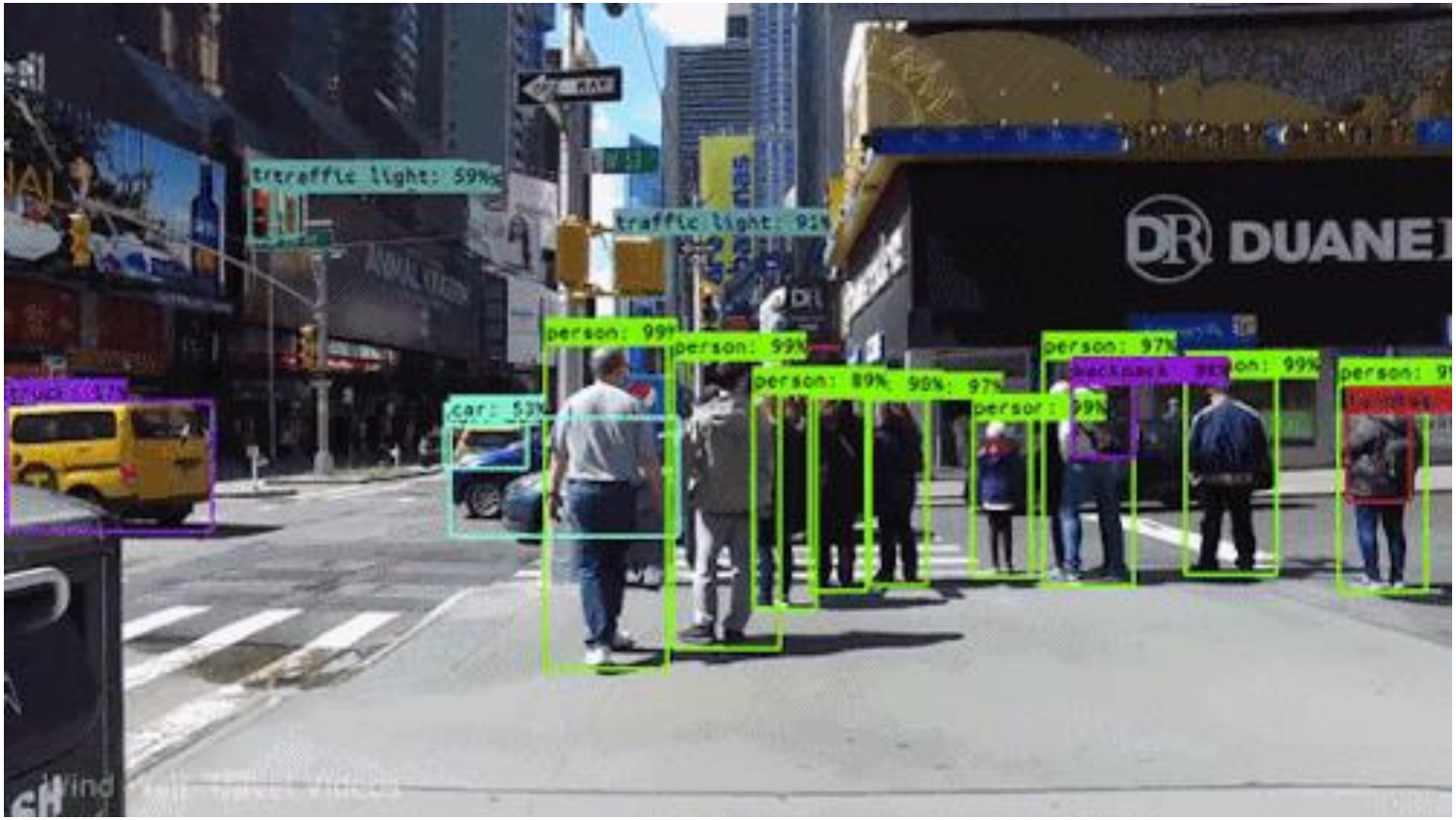
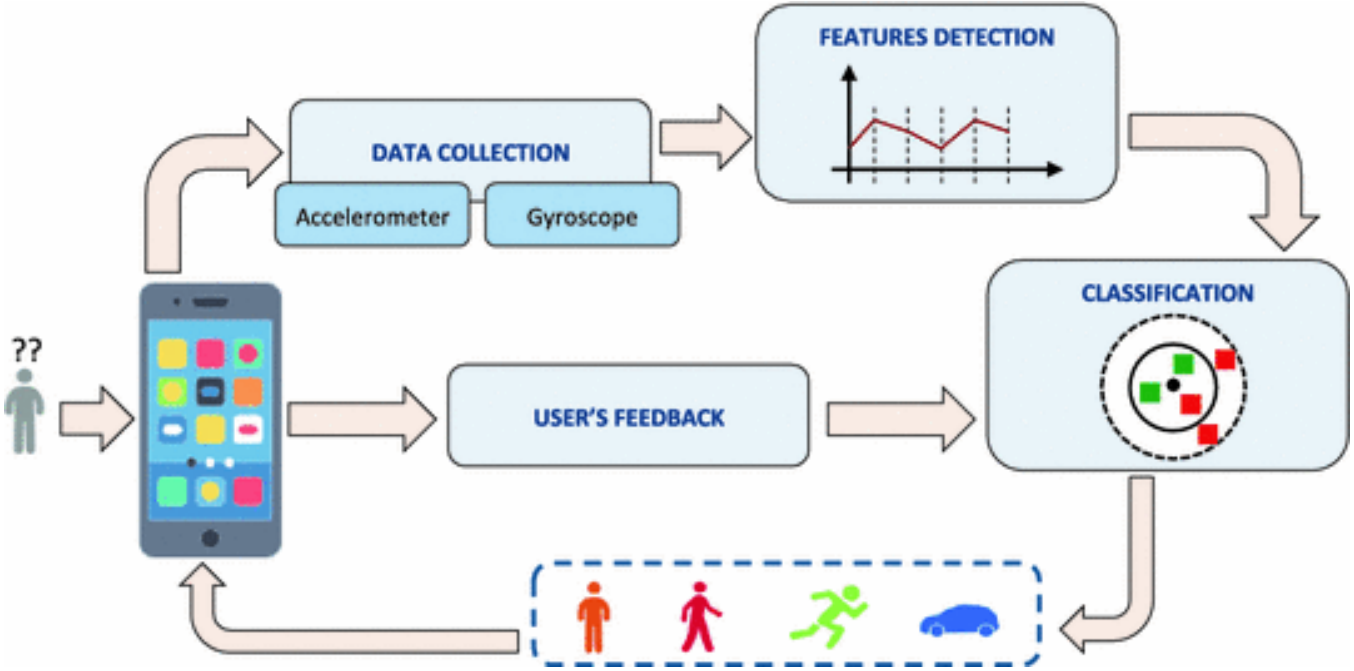


Image classification: Object Detection



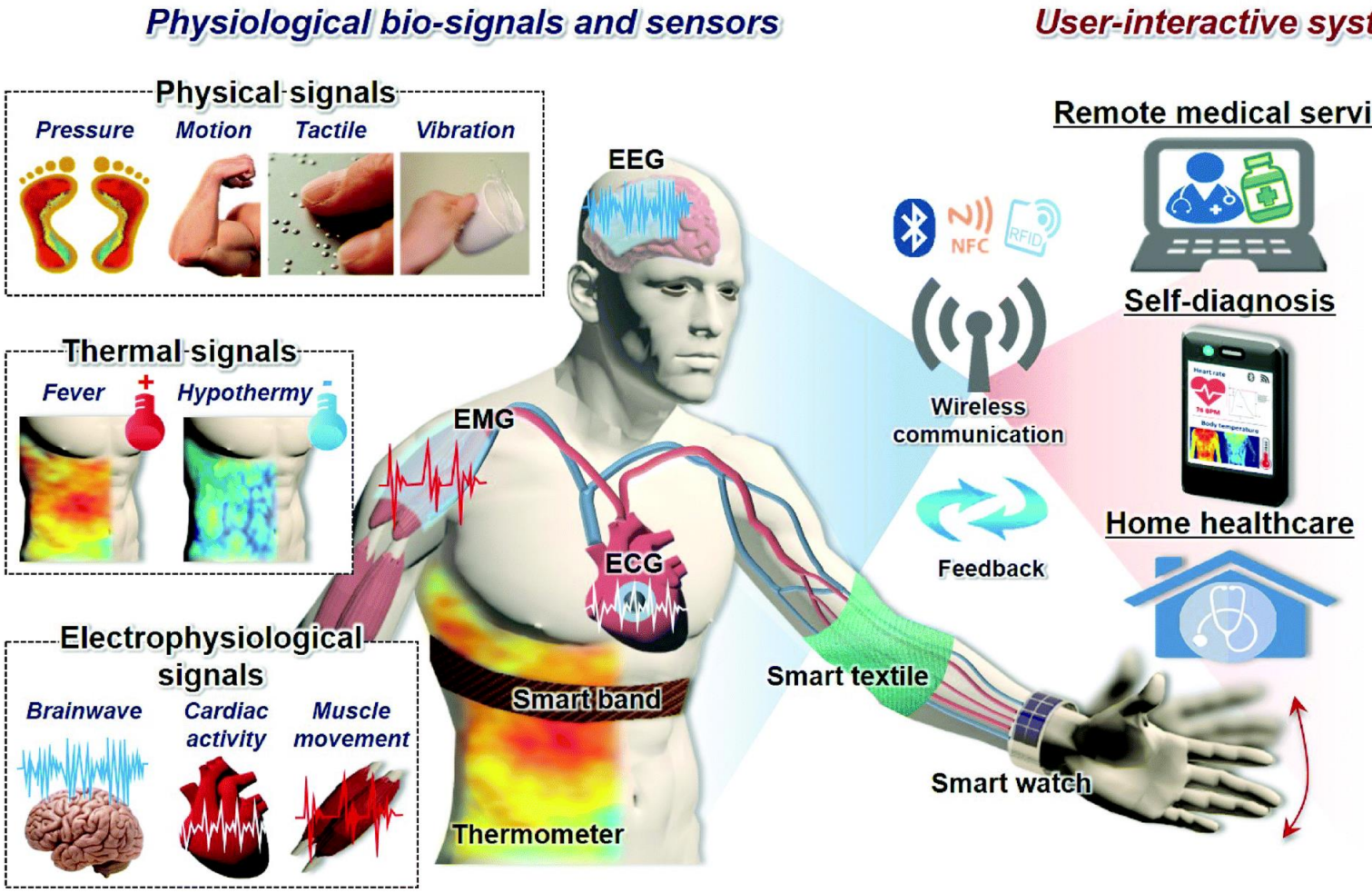
[How do self-driving cars see?]

Smartphone data for Human Activity Recognition



Source

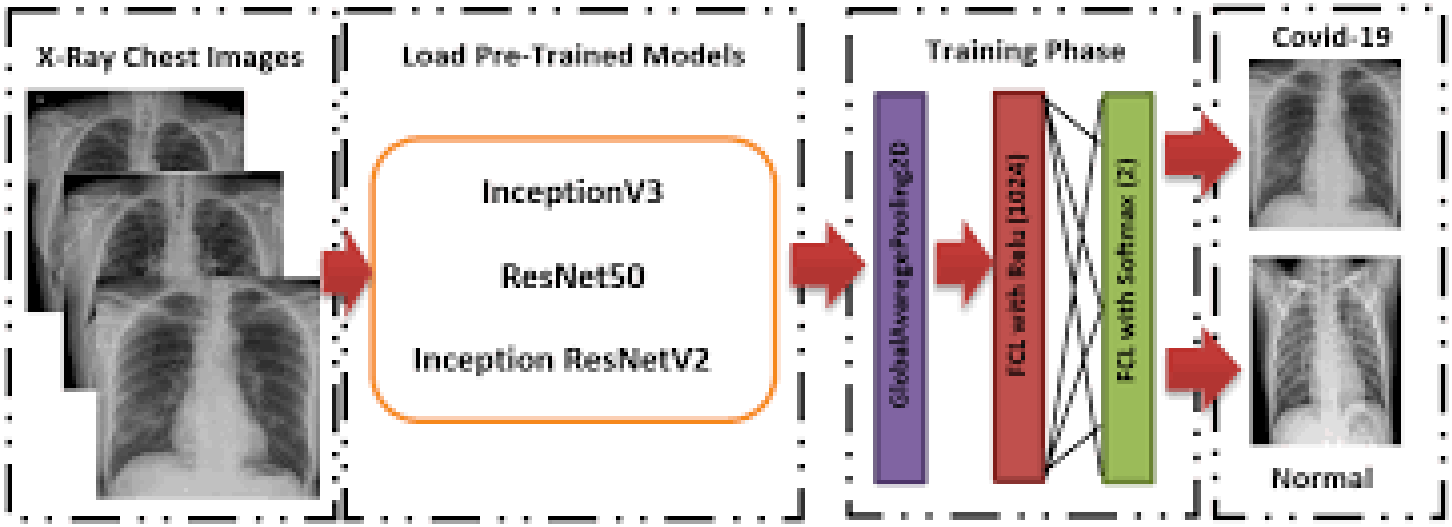
Wearables for health monitoring



Ha, Minjeong, Seongdong Lim, and Hyunhyub Ko. 2018. "Wearable and Flexible Sensors for User-Interactive Health-Monitoring Devices." Journal of Materials Chemistry B 6 (24): 4043–64. https://doi.org/10.1039/c8tb01063c.

Applications for COVID-19

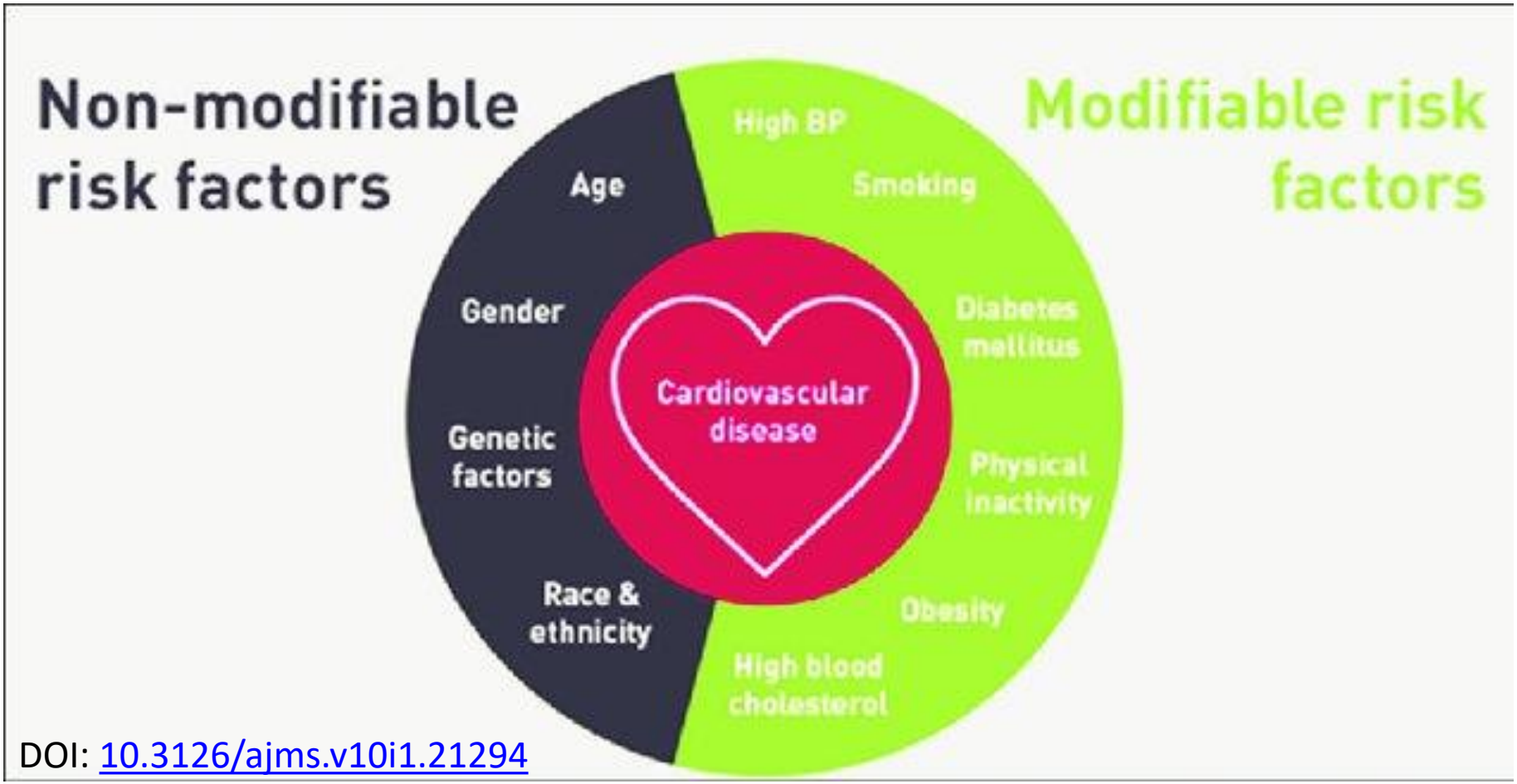
- Chest X-ray data as diagnostic to detect COVID-19
- Cough sounds to detect COVID-19
- Symptoms tracking
- Data for training an AI/ML model, validate and use for predictions
- Using all together for better prediction model?



Source

Classification Learner App in MATLAB Workflow

- Heart disease
- Human activity recognition
- Text classification (SMS spam)
- Image classification (AlexNet)

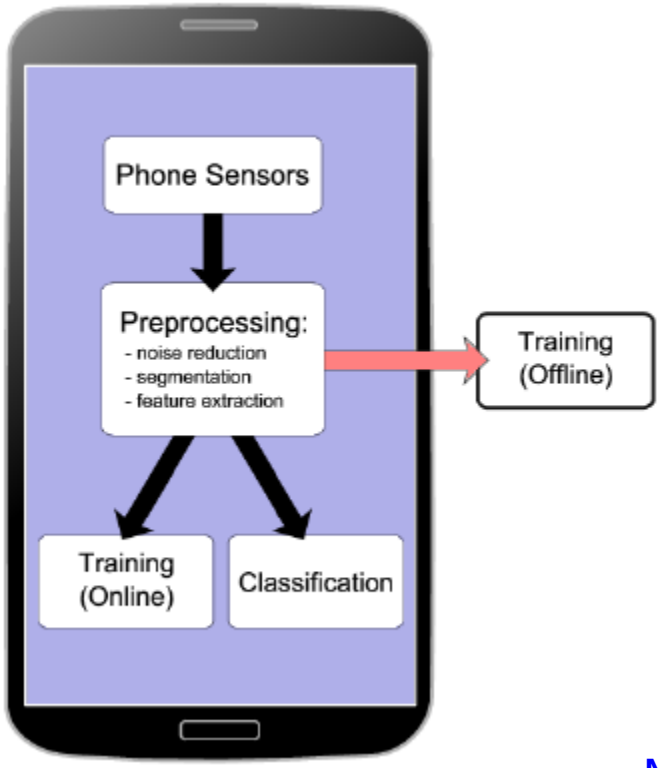


DOI: 10.3126/ajms.v10i1.21294

Heart disease prediction

Heart disease data set source

Human activity recognition



doi:10.3390/s150102059

[MATLAB video tutorial]
[Data source]

Text classification

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

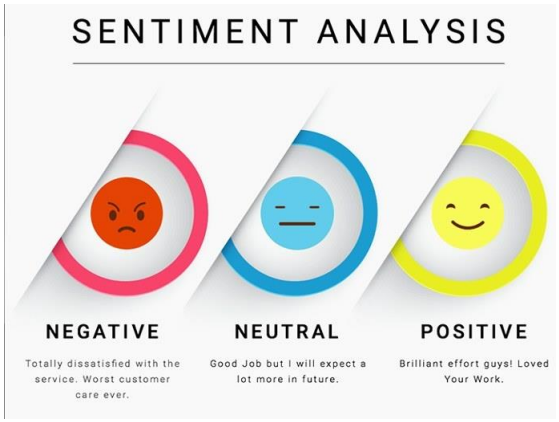
Now is the time for all good men to come to the aid of their party.

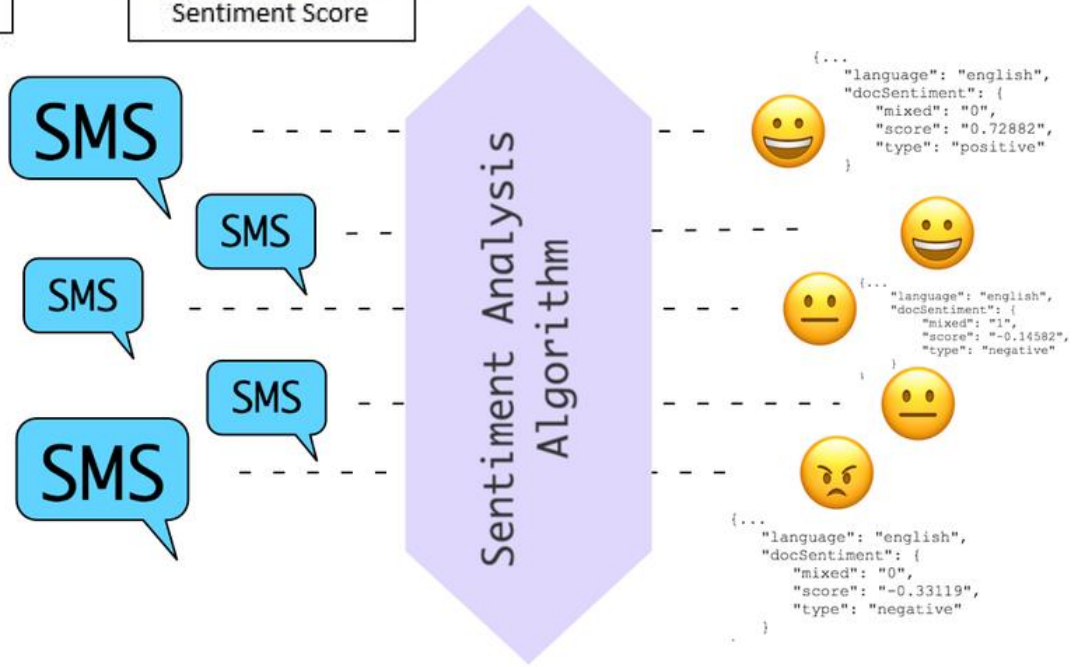
Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

- SMS spam or not dataset source
- Do not disturb messages
- MATLAB script Dataset source
- Bag of words
- Governments taking suggestions from citizens with lakhs of responses



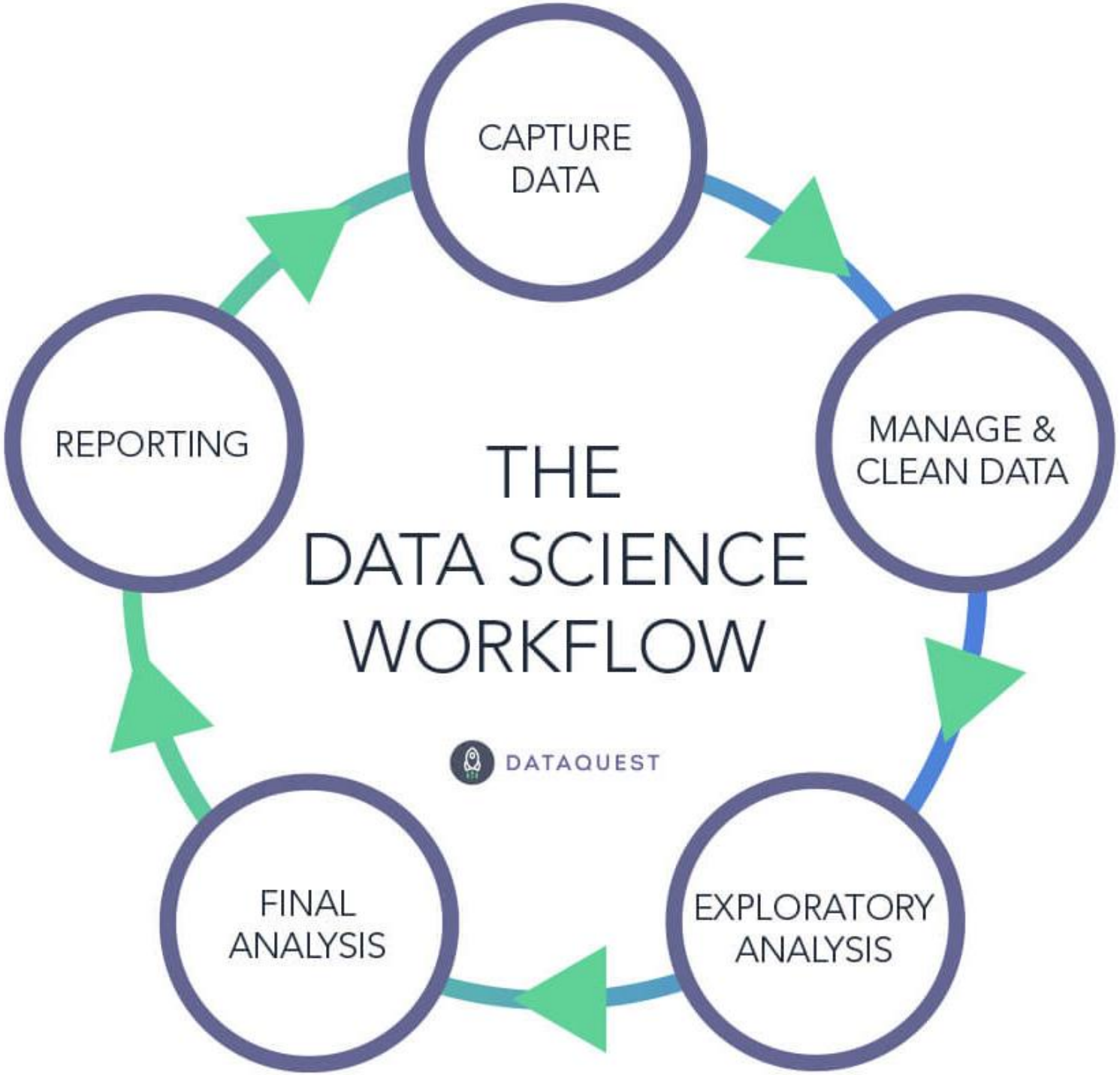


Sentiment Classification

- Sentiment classifier in MATLAB
- Social Media text mining
- US elections analysis through Twitter data – Deb Roy, MIT Paper

Image classification

- AlexNet – CNN (1000 object categories, 227X227 pixels). Download package for MATLAB.
- MATLAB script for image classification with AlexNet


CAPTURE DATA
MANAGE & CLEAN DATA
EXPLORATORY ANALYSIS
FINAL ANALYSIS
REPORTING
THE DATA SCIENCE WORKFLOW
DATAQUEST

Effect of meteorological factors on COVID-19 spread

- This project intends to build a data science based model to study the spread of COVID-19 in India based on meteorological and other factors
- COVID-19 spread = f(average daily T_{air}, diurnal temperature range, humidity, sunlight, wind speed, precipitation, day number, lockdown strategy – mobility, population density, fraction of population > 65 years of age, quality of health care, GDP, GDP per capita, HDI, viral factors) = predictor variables
- COVID-19 response variables = cases, deaths, R-number, cases/population, deaths/population, cases/tests, deaths/tests
- This would help in predicting the areas at greater risk of community spread in the coming months, allowing to focus public health efforts accordingly.
- A long term view on this project can be taken to build a decision support system which would be helpful with future pandemics as well

Literature review

- Araujo, Miguel B., and Babak Naimi. 2020. "Spread of SARS-CoV-2 Coronavirus Likely to Be Constrained by Climate." MedRxiv, https://doi.org/10.1101/2020.03.12.2003472
- Ma, Yueling, Yadong Zhao, Jiangtao Liu, Xiaotao He, Bo Wang, Shihua Fu, Jun Yan, Jingping Niu, Ji Zhou, and Bin Luo. 2020. "Effects of Temperature Variation and Humidity on the Death of COVID-19 in Wuhan, China." Science of the Total Environment 724. https://doi.org/10.1016/j.scitotenv.2020.138226.
- Wang, Jingyuan, Ke Tang, Kai Feng, and Weifeng Lv. 2020. "High Temperature and High Humidity Reduce the Transmission of COVID-19." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3551767.
- Breton, Theodore R. 2020. "The Effect of Temperature on the Spread of the Coronavirus in the U.S." SSRN Electronic Journal, no. March. https://doi.org/10.2139/ssrn.3567840.
- Yan, Ning Ning, Tian Yu Zhang, and Xiao Jun Li. 2020. "Synthesis of 1, 2, 2, 6, 6-Pentamethylpiperidinol." Xiandai Huagong/Modern Chemical Industry 40 (5): 186–89. https://doi.org/10.16606/j.cnki.issn0253-4320.2020.05.040.
- Xu, Bo, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L. Cohn, et al. 2020. "Epidemiological Data from the COVID-19 Outbreak, Real-Time Case Information." Scientific Data 7 (1): 1–6. https://doi.org/10.1038/s41597-020-0448-0.
- Sajadi, Mohammad M, Parham Habibzadeh, Augustin Vintzileos, Fernando Miralles-wilhelm, and Anthony Amoroso. 1992. "This Preprint Research Paper Has Not Been Peer Reviewed. Electronic Copy Available at: Https://Ssrn.Com/Abstract=3550308," no. 410: 6–7.
- Luo, Wei, Maimuna S Majumder, Dianbo Liu, Canelle Poirier, Kenneth D Mandl, Marc Lipsitch, and Mauricio Santillana. 2020. "The Role of Absolute Humidity on Transmission Rates of the COVID-19 Outbreak." MedRxiv, 7. https://doi.org/10.1101/2020.02.12.20022467.
- https://home.iitd.ac.in/research-pracriti.php

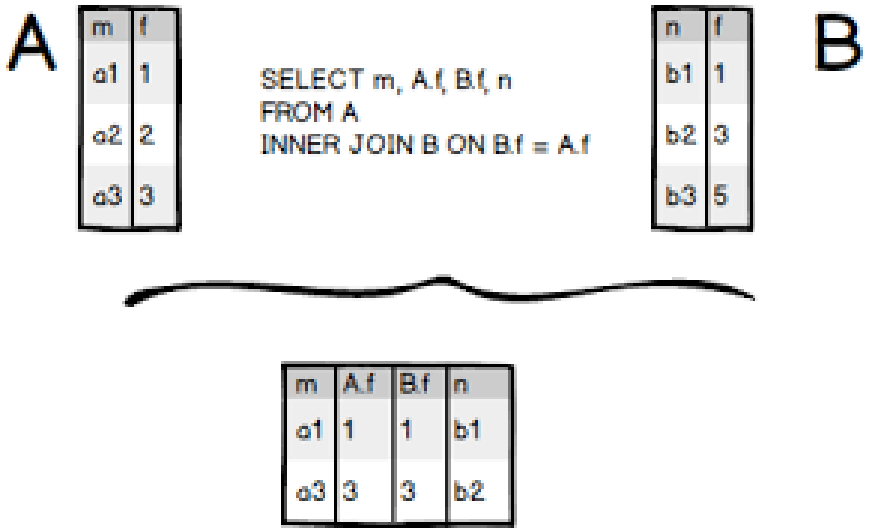
Papers suggest important predictors, data sources

Data sources

- COVID-19 cases by location – global
- COVID-19 cases for India
- Demographic and socio-economic indicators
- Weather data

Data preprocessing

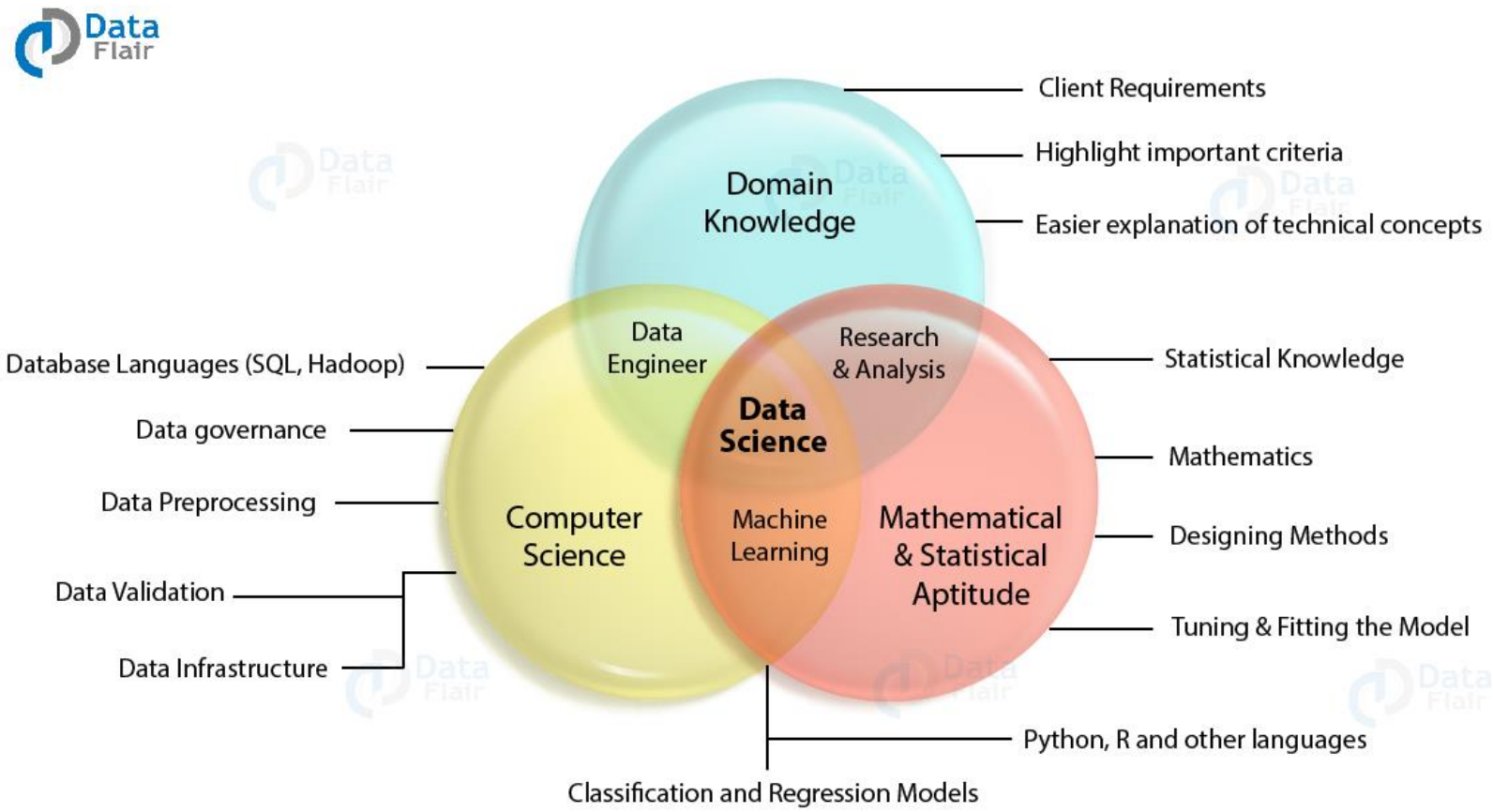
- SQL inner join (between tables with fields having unique values - keys)
- Avoid duplicates



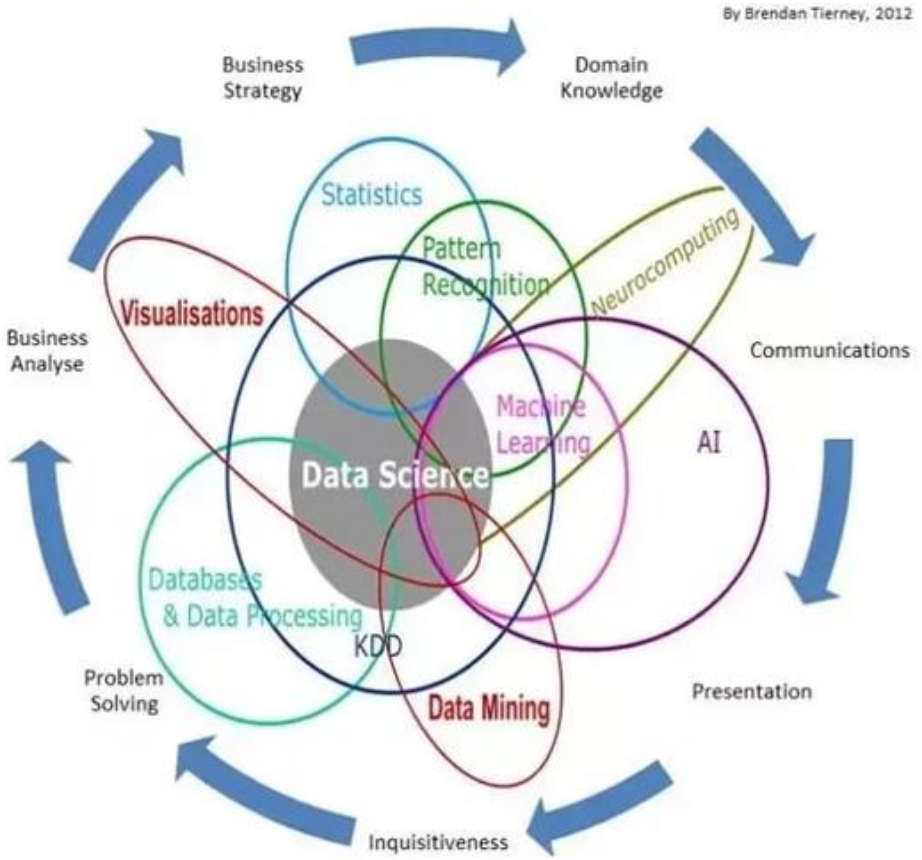
Future work for this project

- Greater spatial resolution of the data
- Use weather related factors, socio-economic factors, covid-19 factors
- Use other response variables
- Learn from epidemiological models from previous pandemics

Data Science: understand and analyze actual phenomena with data



Data Science is multi-disciplinary



Source

Note: All the images in these slides without a source have been taken from google images.