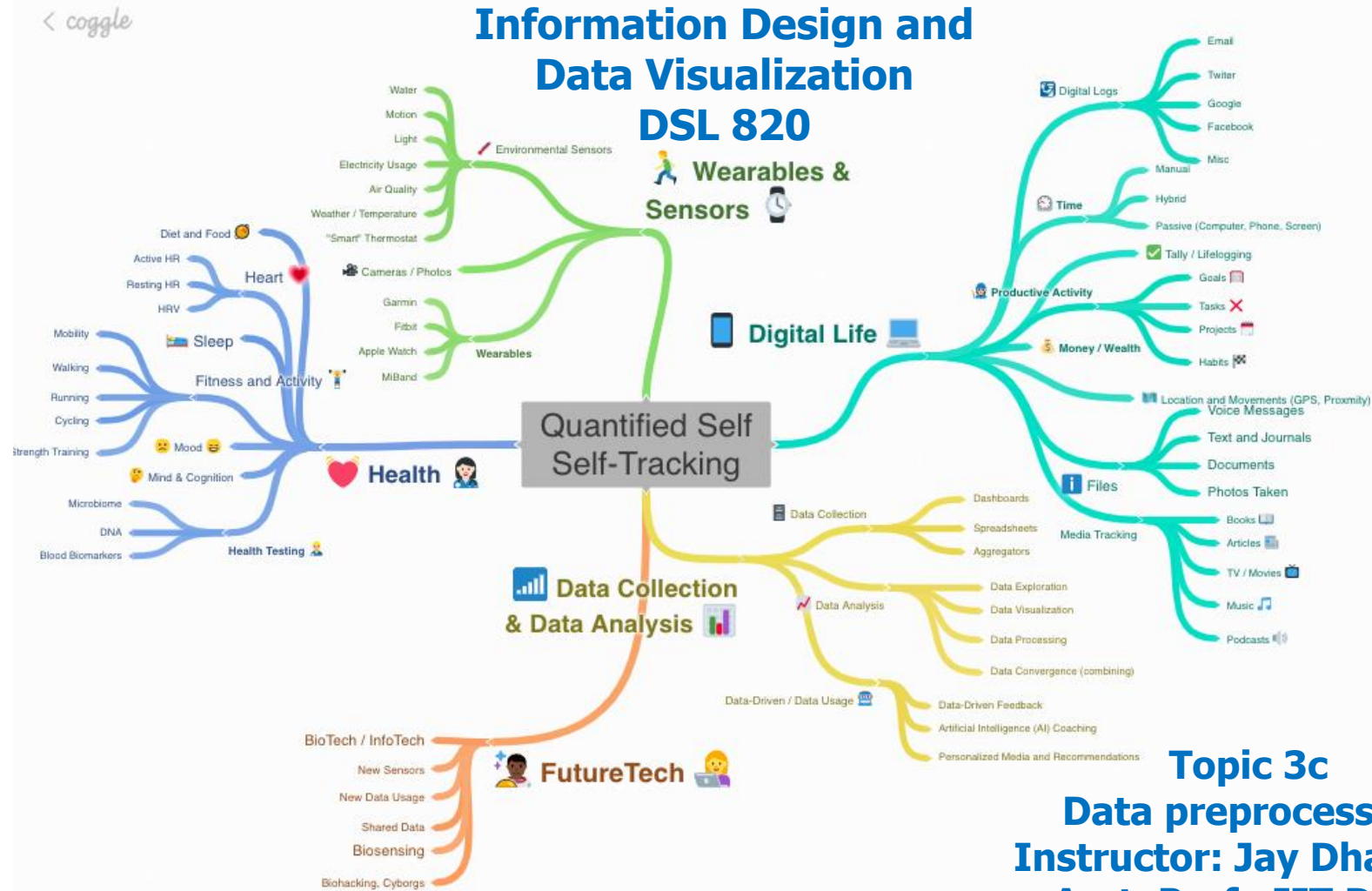


Materials and Prototyping DSL 732 &

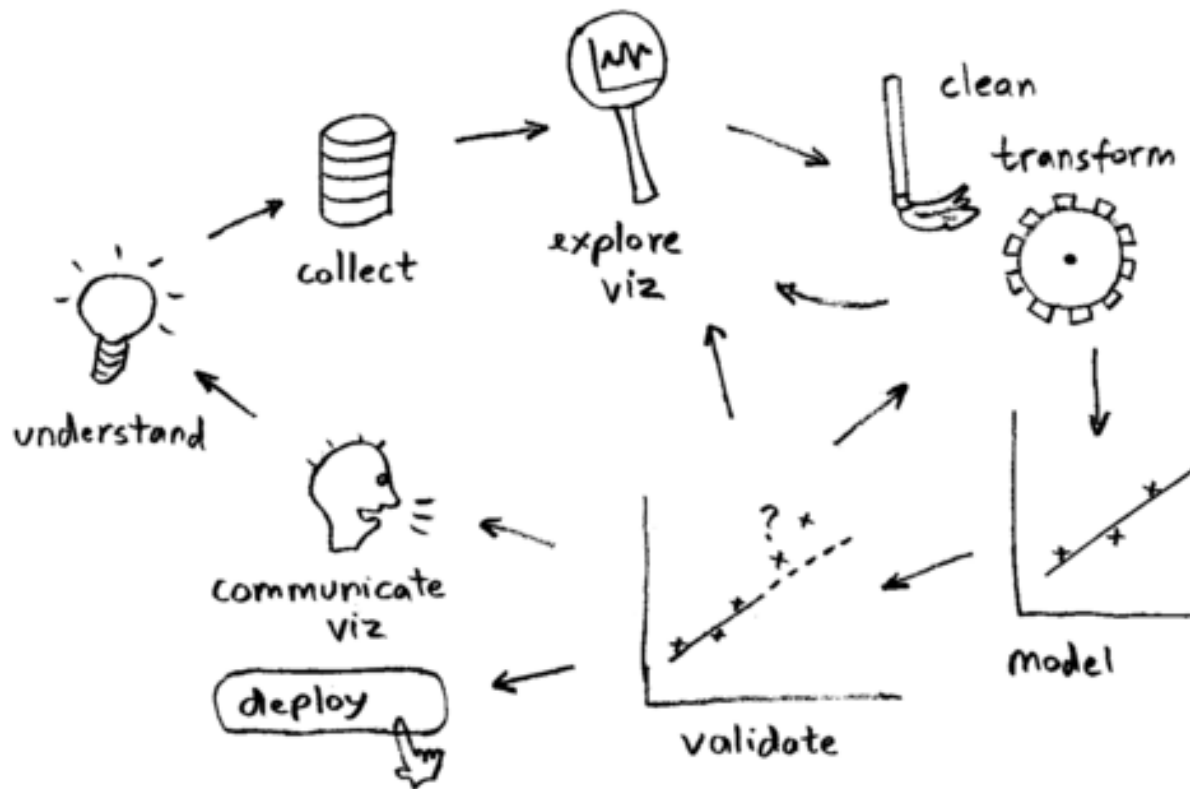
Information Design and Data Visualization DSL 820



Topic 3c

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

Dated: 20th January, 2020

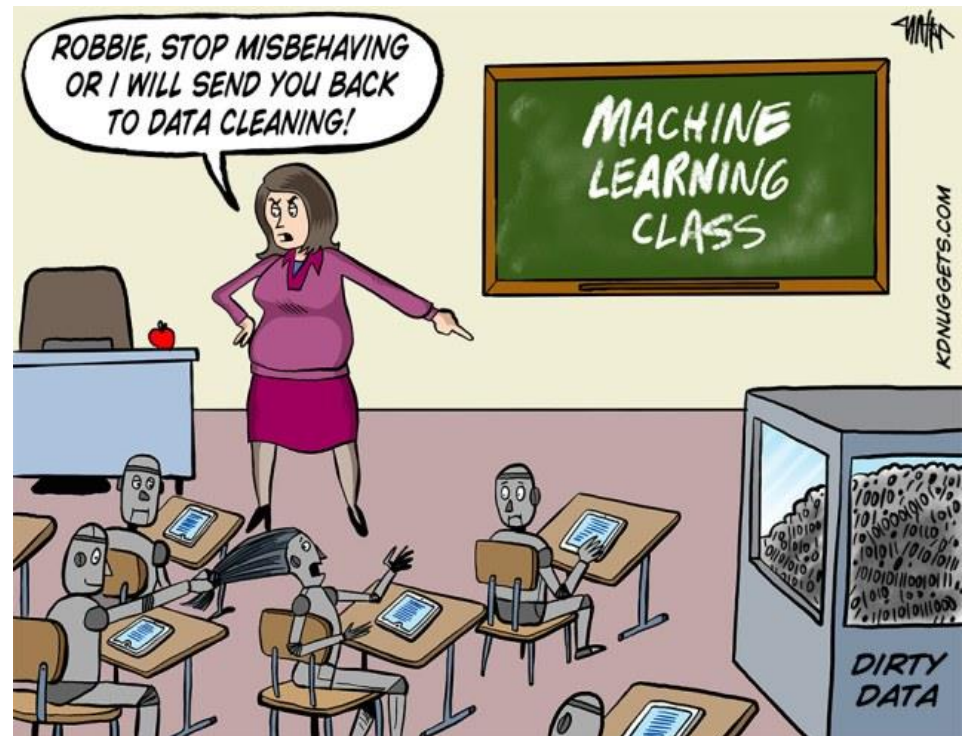


Data science workflow

Source: <http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/>

Data preprocessing

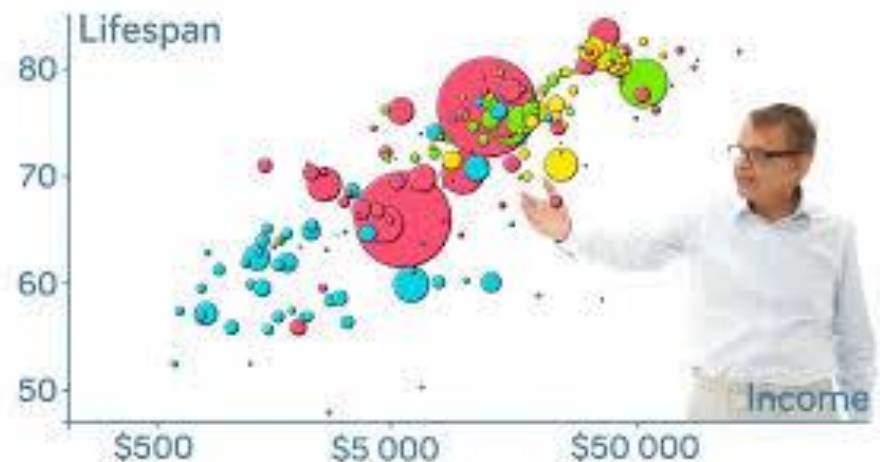
- Data cleaning (duplicates, inconsistencies, missing data, removing outliers)
- Combining data, data aggregation
- Data transformation (dimensionality reduction, normalization)



Source: <https://dunyaoguz.github.io/my-blog/dataframemapper.html>

Data Visualization

- Information visualization is a set of technologies that use visual computing to amplify human cognition with abstract information.[Card, 2008, p. 542]
- Visuals are worth a thousand words
- Detect patterns, trends and outliers in the data
- [Out of sight, out of mind](#)



Source: [Data sings with Hans Rosling](#)
[Gapminder slides](#)

Example 1: Covid-19 and Environmental Sustainability

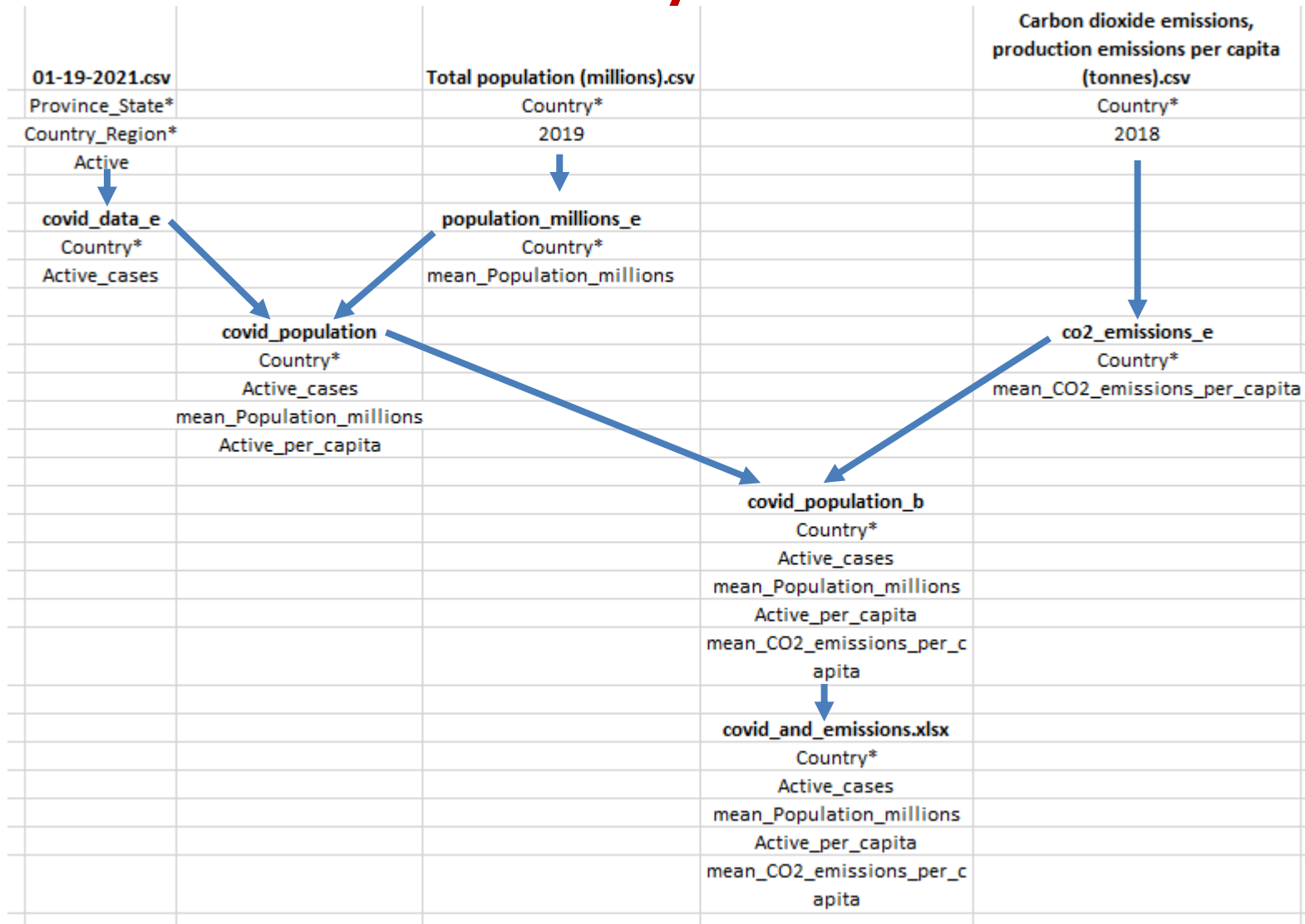
- Does Covid-19 affect those countries more which have higher CO2 emissions or higher aging population or lower HDI?
- Source of Covid-19 data:
<https://github.com/CSSEGISandData/COVID-19>
- Source of human development data:
<http://hdr.undp.org/en/data>
- Download [covid hdi.zip](#)



shutterstock.com • 1827313010

Schema for covid19_hdi data analysis

*key



Working with MATLAB online

- [MATLAB online account](#)
- Files at
https://web.iitd.ac.in/~jay/dsl732/dsl732_iot.html
- How would you compare Covid-19 active cases in those countries which have higher aging population or lower HDI?

Example 2: need for adequate ventilation (high CO₂ levels)

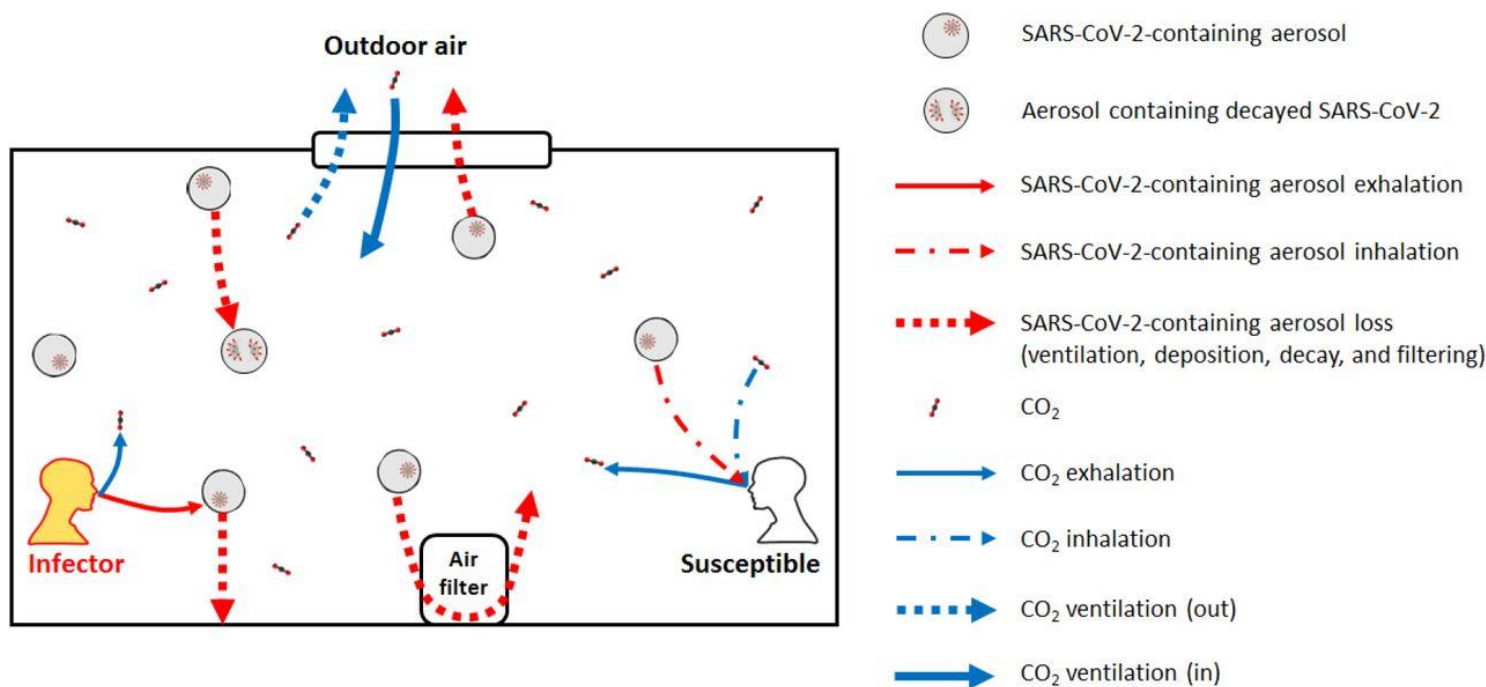
Table 1 | Overview of potential health effects

	CO ₂ concentration (ppm)	Duration	Selected key references
Adverse health outcomes associated with acute CO ₂ exposure			
CO ₂ retention	1,000–5,000	<4 h	Zhang et al. ⁷⁵ ; Zhang et al. ⁷³ ; Vehvilainen et al. ⁷⁷ ; Shiraram et al. ⁷⁶
Inflammation	2,000–4,000	2 h	Thom et al. ^{80,81} ; Schneberger et al. ⁸²
Cognitive effects	1,000–2,700	1–6 h	Kajtar and Herczeg ⁸⁵ ; Satish et al. ⁸⁶ ; Allen et al. ^{87,88} ; Zhang et al. ⁷⁵ ; Zhang et al. ^{73,74} ; Rodeheffer et al. ⁹¹ ; Snow et al. ⁹⁰
Adverse health outcomes associated with chronic CO ₂ exposure			
Chronic, low-grade systemic inflammation	~3,000	13 d	Zappulla ^{2,69} ; Beheshti et al. ¹⁰¹
Bone demineralization and kidney calcification	~2,000–3,000	60–90 d	Schaefer et al. ^{102,103}
Chronic, low-grade (sub-clinical) respiratory acidosis	Unknown	Decades	Carnauba et al. ¹⁰⁹ ; Robertson ^{61,106}
Behavioural changes and physiological stress	700–3,000	13–15 d	Beheshti et al. ¹⁰¹ ; Wade et al. ¹⁰⁴ ; Martrette et al. ¹¹¹ ; Kiray et al. ¹¹²
Hedonic feeding behaviours	Unknown	Ecological	Hersoug et al. ¹¹³ ; Zheutlin et al. ¹
Oxidative stress and endothelial dysfunction	3,000–5,000	13 d to 6 months	Beheshti et al. ¹⁰¹ ; Thom et al. ^{80,81} ; Zwart et al. ¹¹⁹

Exposure levels, including magnitude and duration, for which health effects may manifest. The selected key references are most relevant to the health end-point.

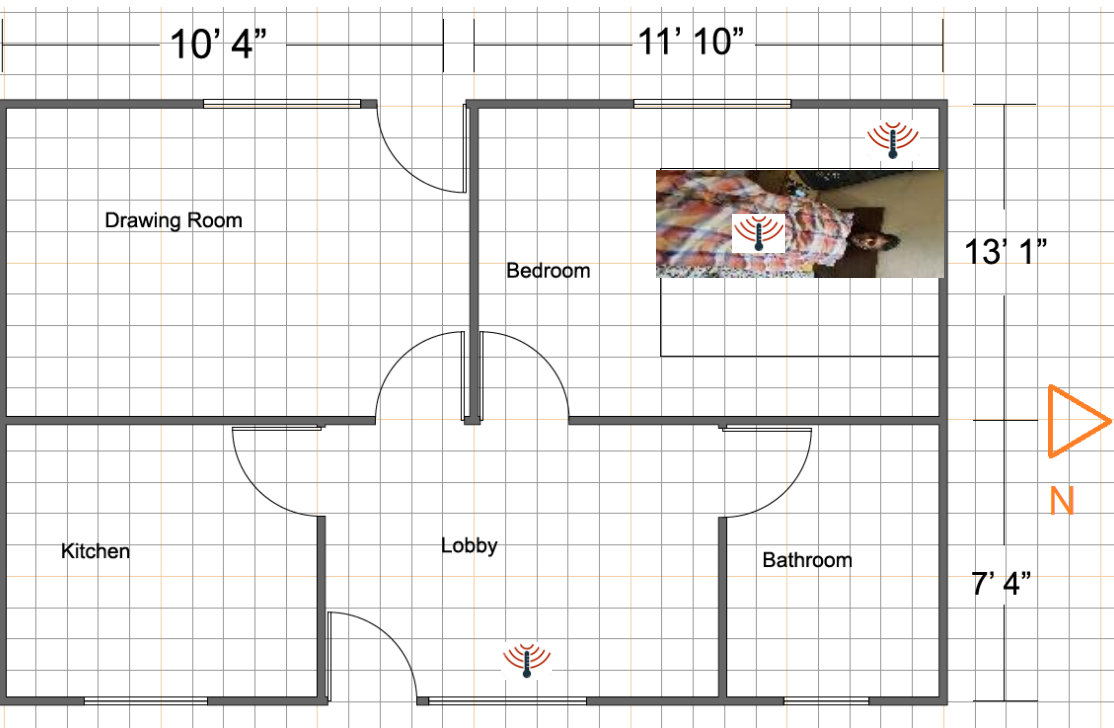
Jacobson, Tyler A, Jasdeep S Kler, Michael T Hernke, Rudolf K Braun, Keith C Meyer, and William E Funk. 2019. “Direct Human Health Risks of Increased Atmospheric Carbon Dioxide.” *Nature Sustainability* 2 (8): 691–701. <https://doi.org/10.1038/s41893-019-0323-1>.

CO2 levels as a proxy for Covid-19 transmission



Peng, Zhe, and Jose L. Jimenez. 2021. "Exhaled CO₂ as a COVID-19 Infection Risk Proxy for Different Indoor Environments and Activities." *Environmental Science and Technology Letters* 8 (5): 392–97. <https://doi.org/10.1021/acs.estlett.1c00183>.

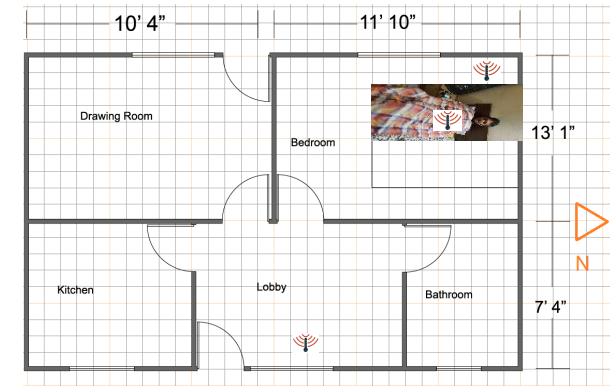
Indoor air quality experiment setup



Experiment: CO₂ levels in a bedroom (no heater)

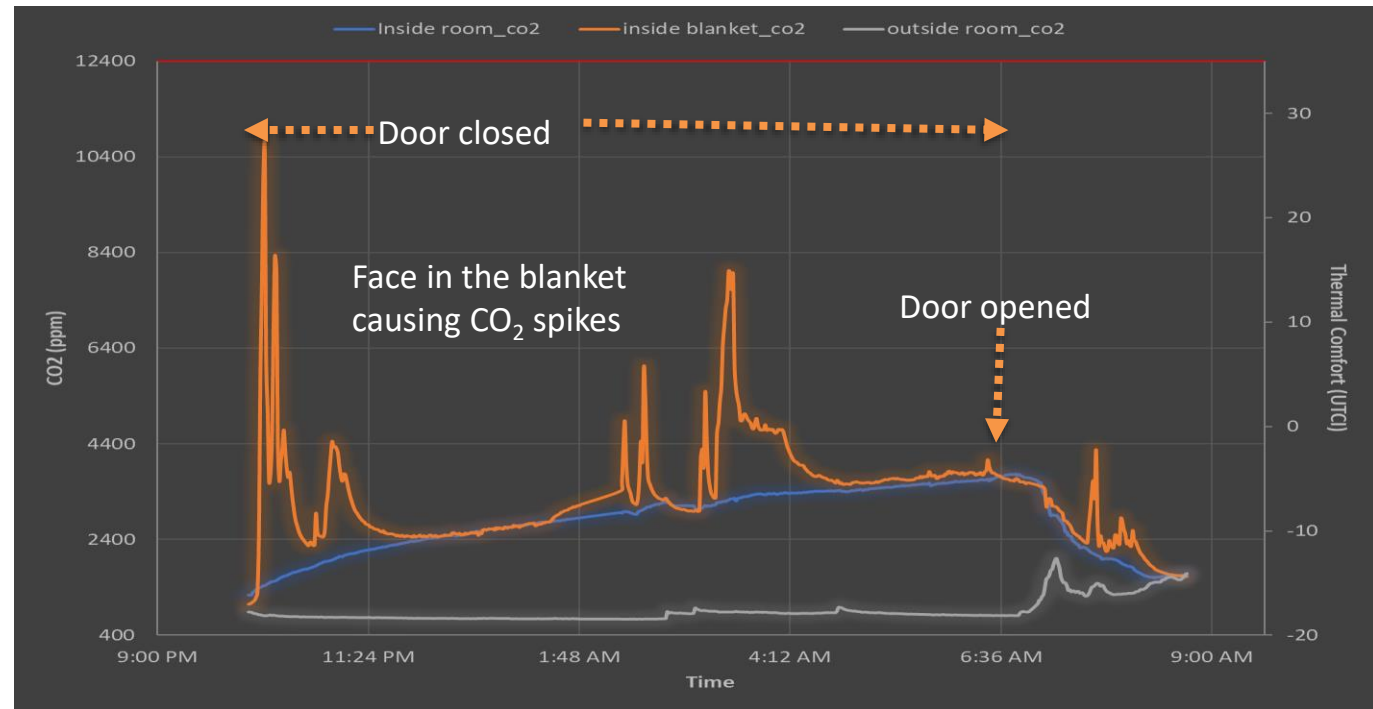
Date: 25th December, Location: New Delhi

Download [co2_expt.zip](#)



Insights:

1. Outside room CO₂ levels were around 800 ppm
2. Indoor room CO₂ levels reached 3500+ for two occupants. It's a fact as pointed out by the Nature paper.
3. CO₂ levels inside the blanket spiked to 10000+ ppm also. Its better to keep the face outside the blanket as we spend 1/3rd of our life sleeping.
4. Door opening led to mixing of CO₂ levels for inside and outside of room.

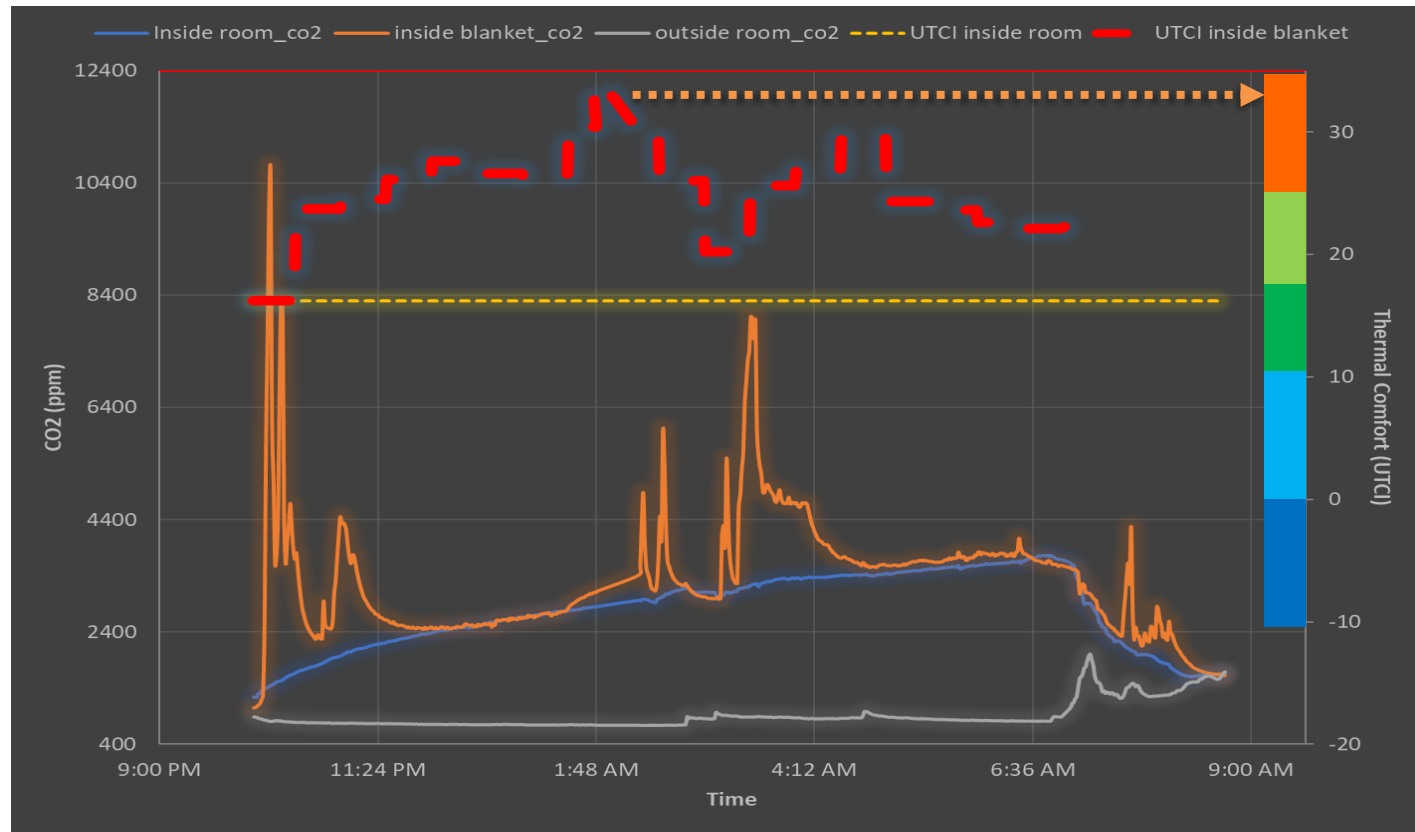


Experiment: CO₂ levels vs. thermal comfort in a bedroom (no heater)

Date: 25th December, Location: New Delhi

Insights:

1. Outside and inside room UTCI was around 16 C
2. Inside blanket UTCI reached even 34 C
3. With no heater, it is better if the door can be kept open to have low CO₂ levels.

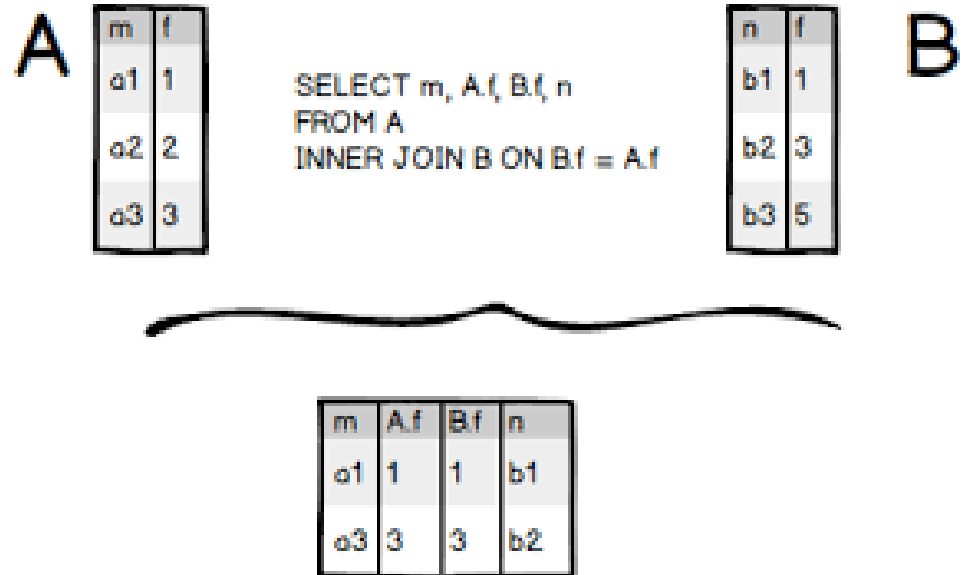


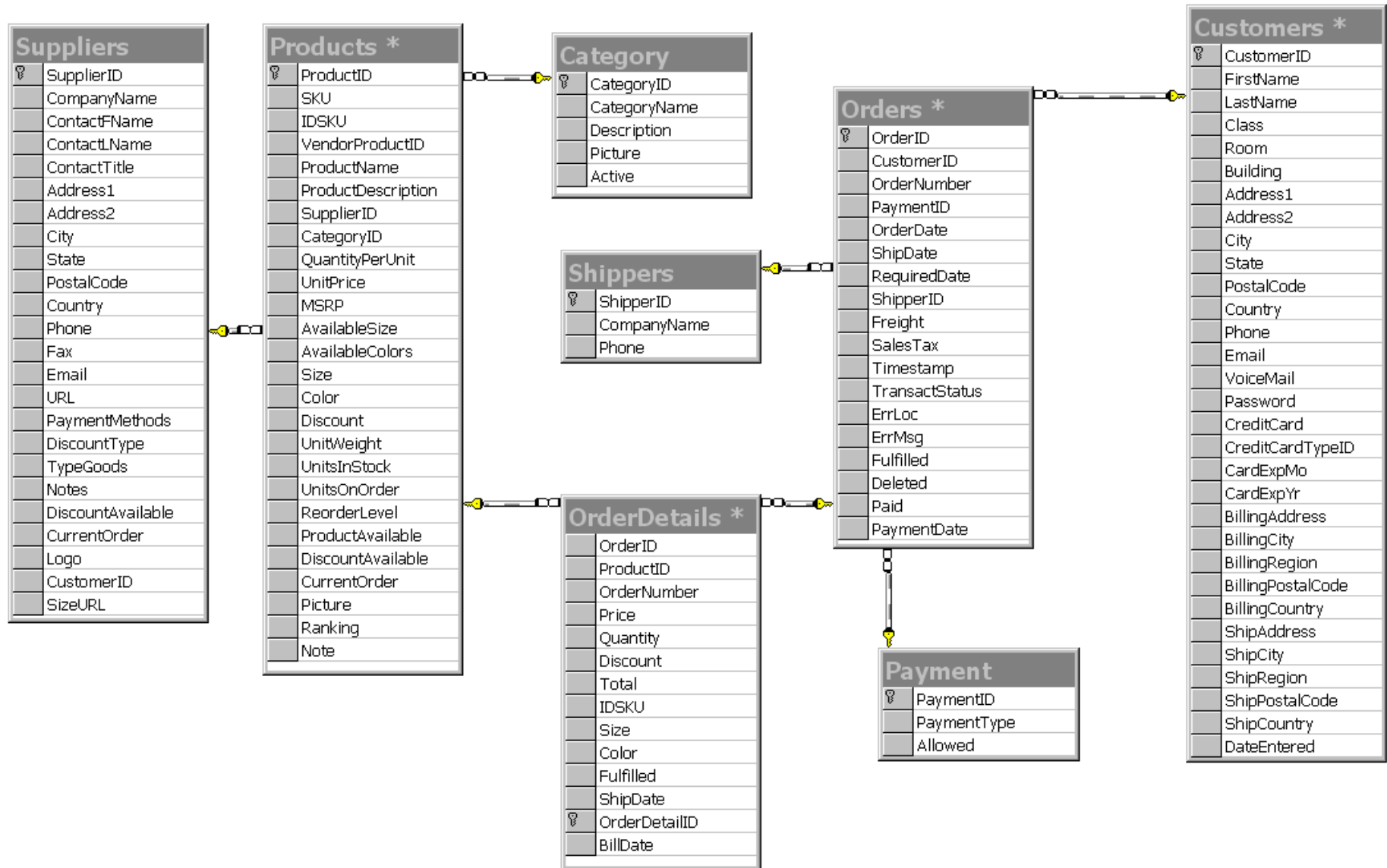
Concepts learnt

- Excel, MATLAB Cloud vs. PC
- Importing csv data into MATLAB
- Importing many tables of data through scripting
- MATLAB Live Script Editor, grouping data, keys in a table, get rid of duplicates (example: `duplicates_covid_hdi.mlx`), missing values
- MATLAB Fundamentals, MATLAB onramp, MATLAB for Data Processing and Visualization
- [Data Driven Design \(Data science with Python, R\)](#)

Combining data

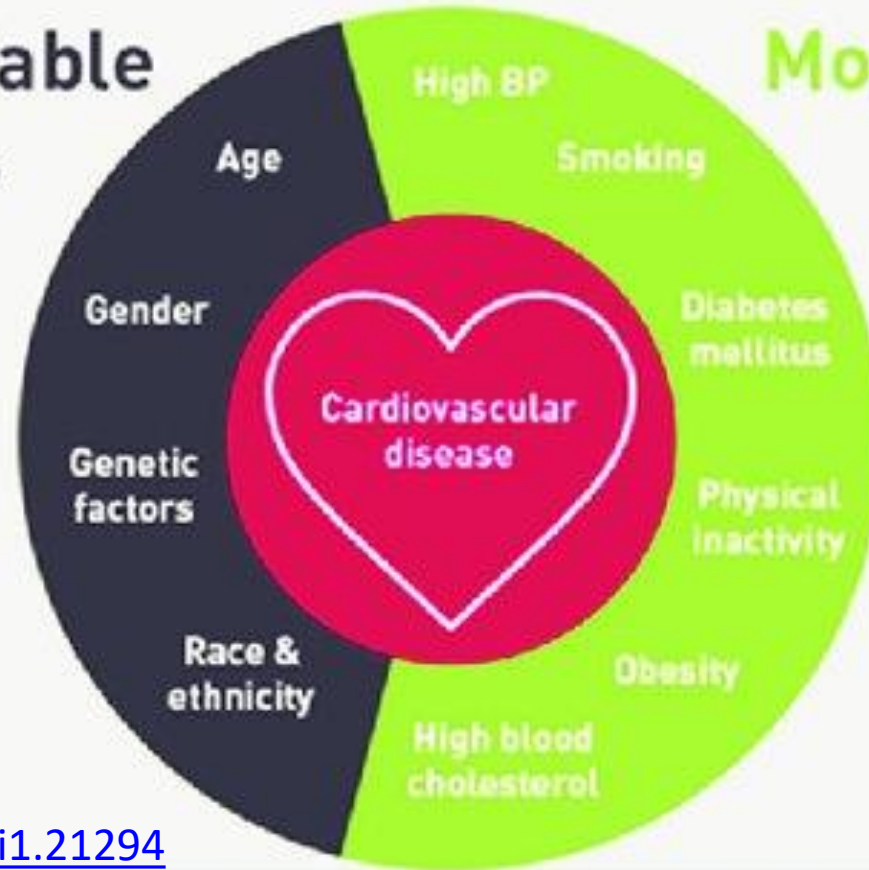
- DBMS – many tables, have different kinds of data
- Inner join (between tables with fields having unique values - keys)
- Queries to filter data
- Avoid duplicates
- Outer join





Relational database design: <https://www.princeton.edu/~rcurtis/ultradev/ecommdatabase.html>

Non-modifiable risk factors



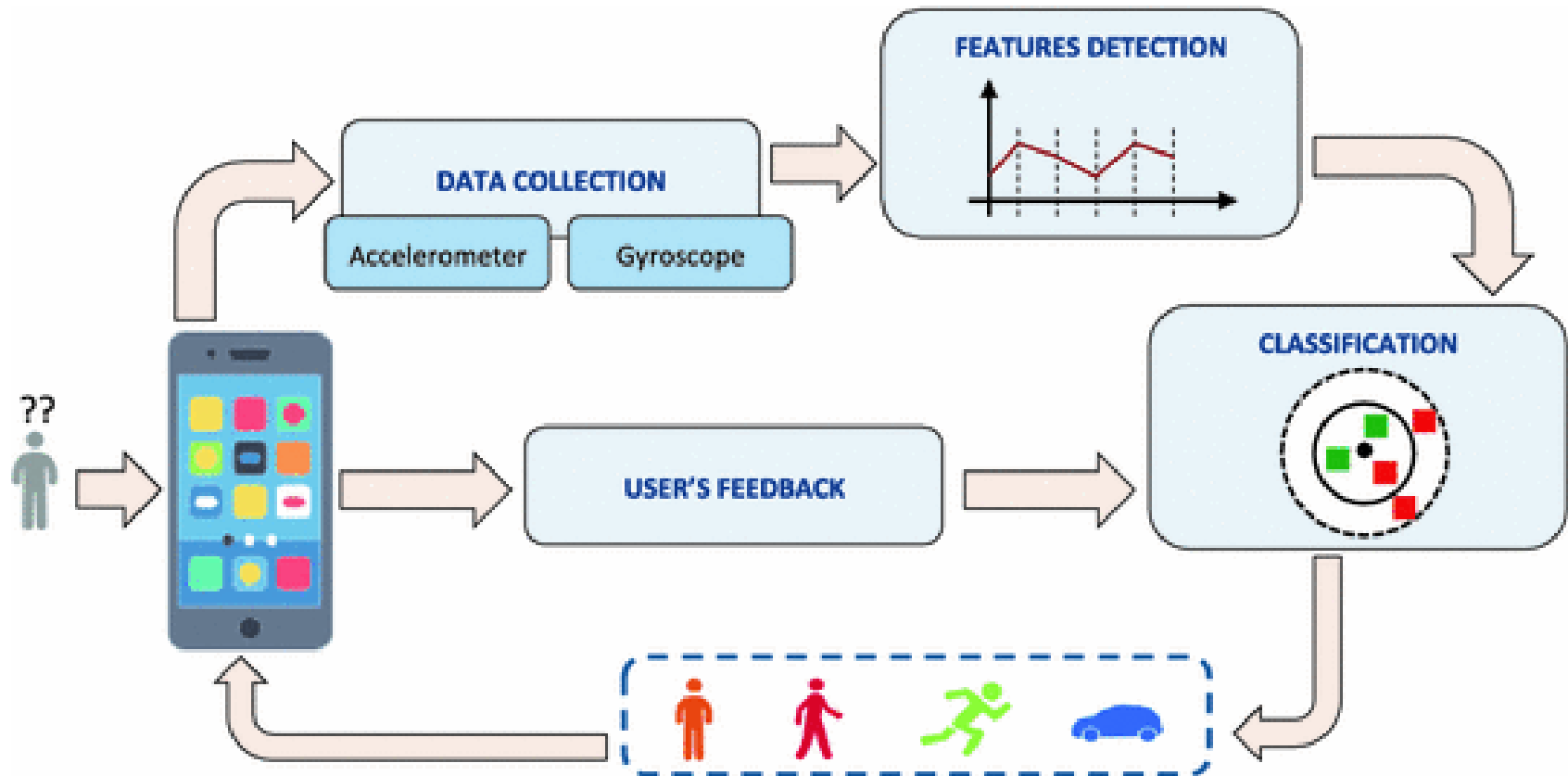
Modifiable risk factors

DOI: [10.3126/ajms.v10i1.21294](https://doi.org/10.3126/ajms.v10i1.21294)

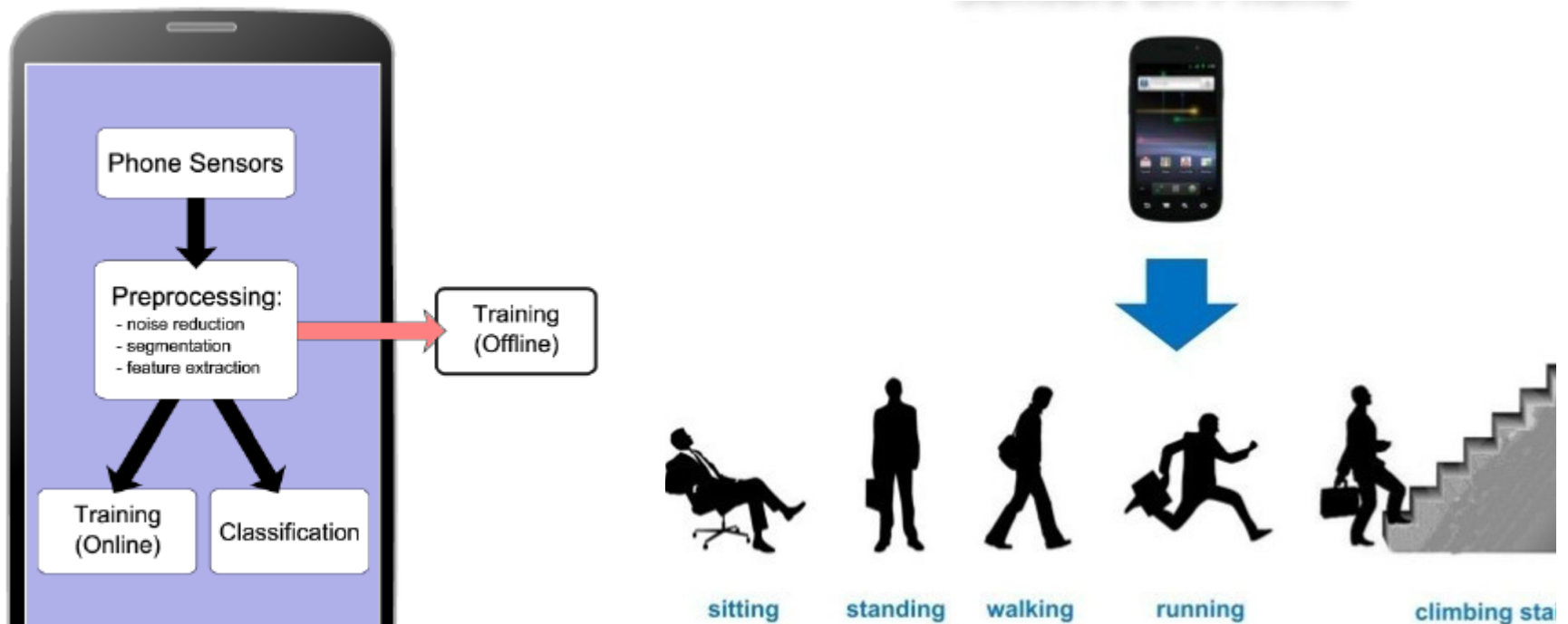
Heart disease prediction

[Heart disease data set source](#)

Smartphone data for Human Activity Recognition



Human activity recognition



doi:10.3390/s150102059

[MATLAB video tutorial](#)
[Data source](#)

Classify acceleration data for posture

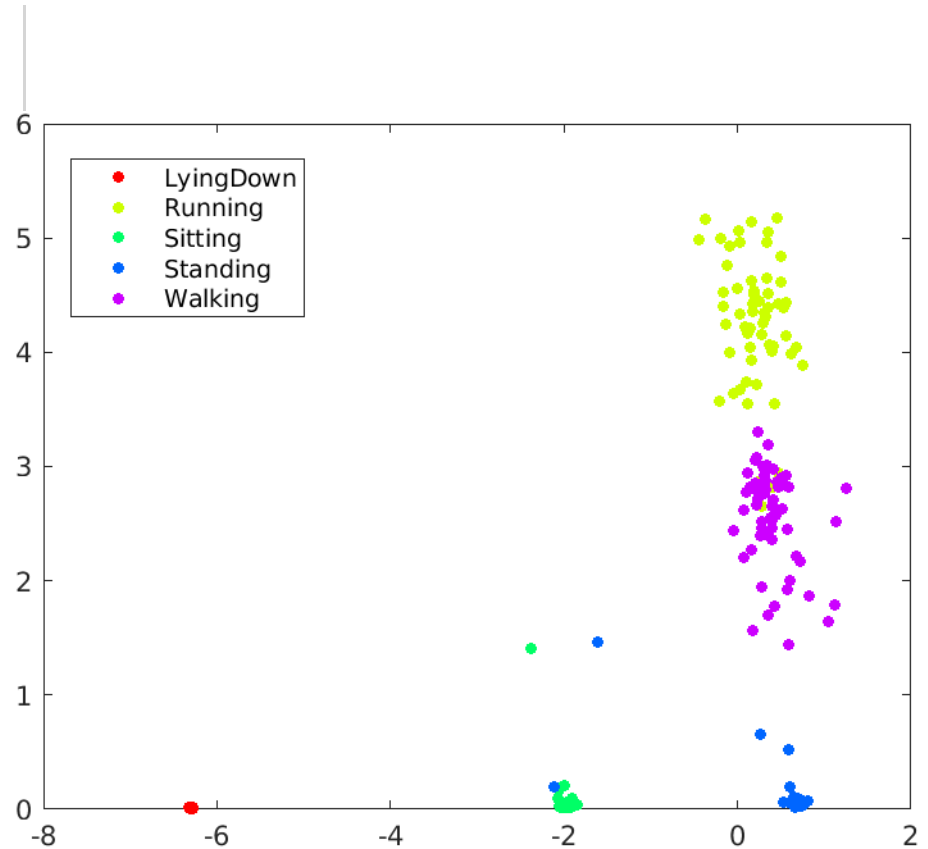
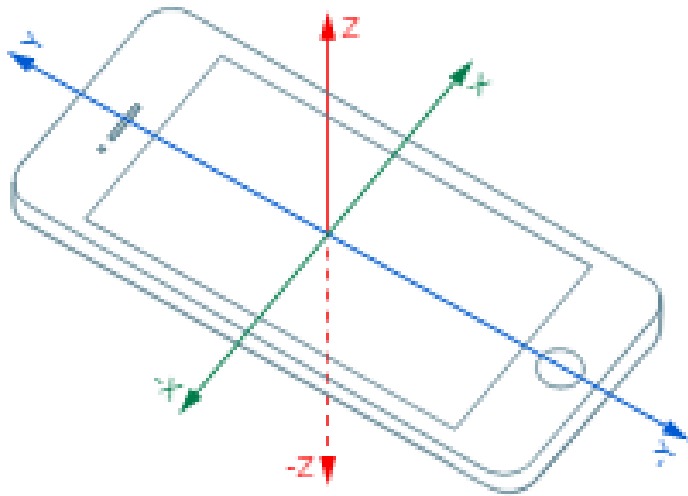
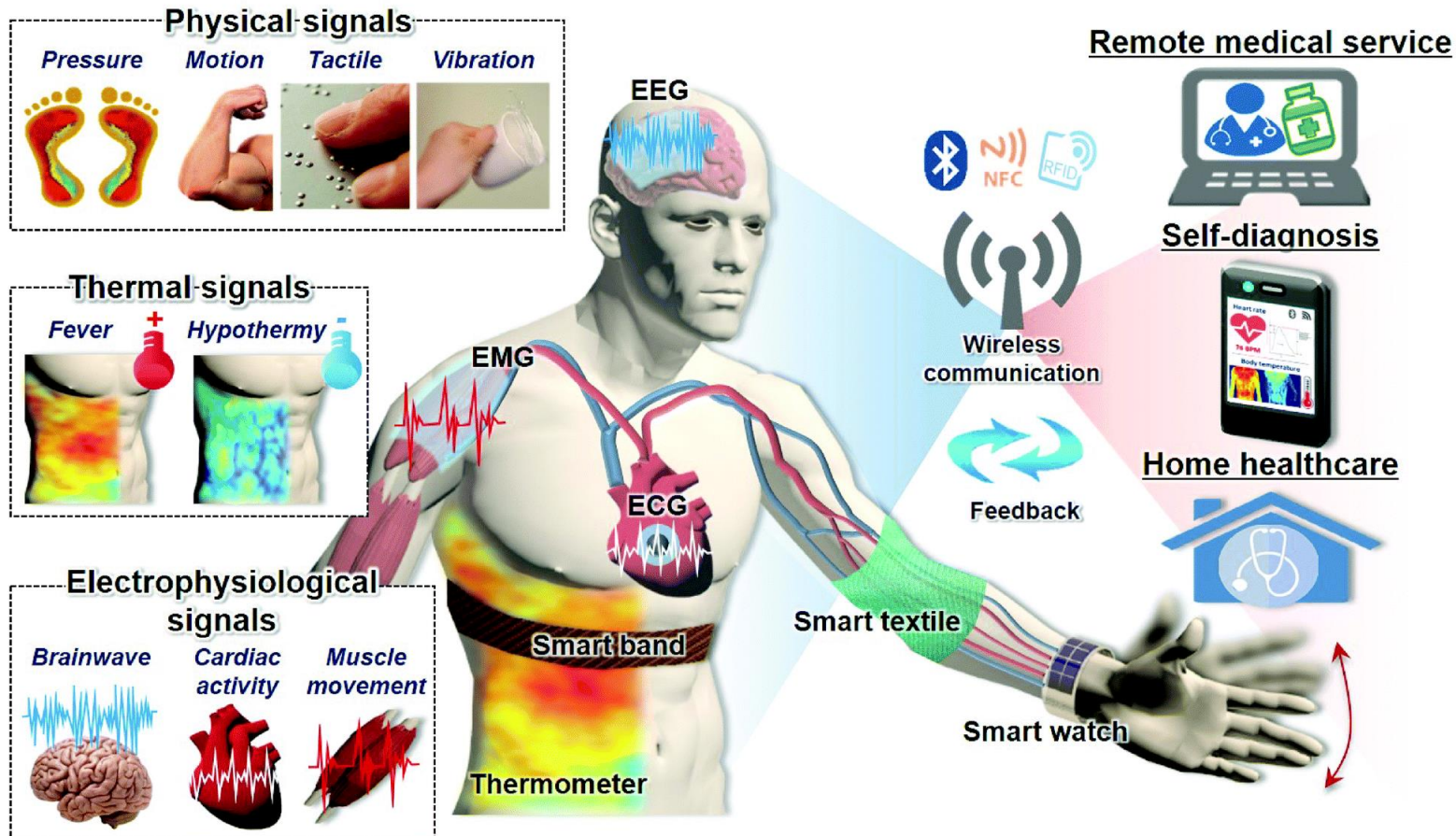


Image source: <https://www.w3.org/TR/accelerometer/>

Wearables for health monitoring

Physiological bio-signals and sensors

User-interactive system



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>.

Assignment 3c

- Take any data science problem and briefly describe the problem you are looking at. What are the variable(s) you would like to study? What are the sources of data? Please share the rough schema of your project data. Please perform the first cut analysis using data collection, data pre-processing and data visualization and share insights for your project. Please also upload your scripts properly commented and the links to the data sources with your assignment.