ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

FINDING SIMILAR ITEMS FROM THE PAST

SANJOY DASGUPTA, PROFESSOR

UC San Diego

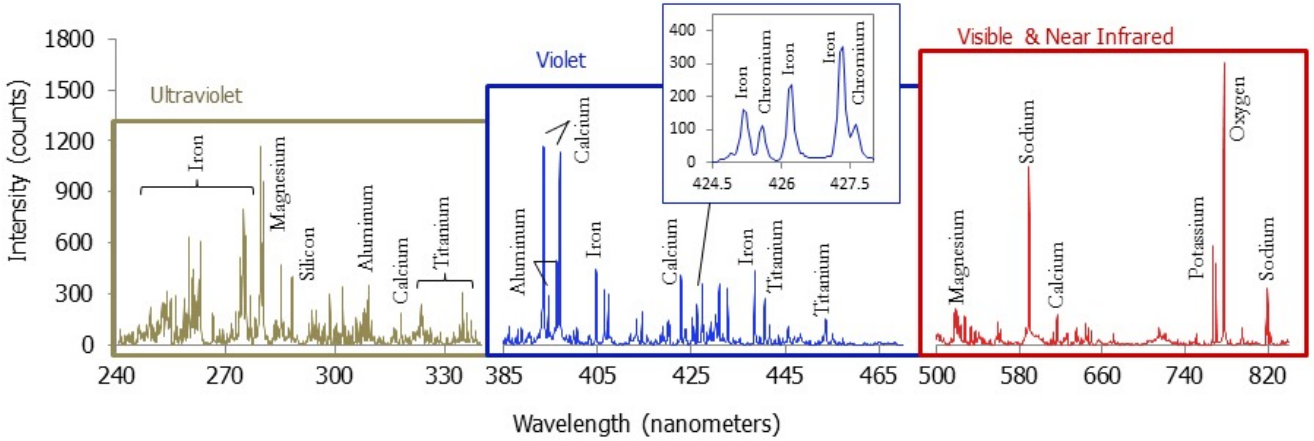
COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Example: an Instrument on the Mars Rover

Mars Curiosity Rover: the ChemCam instrument

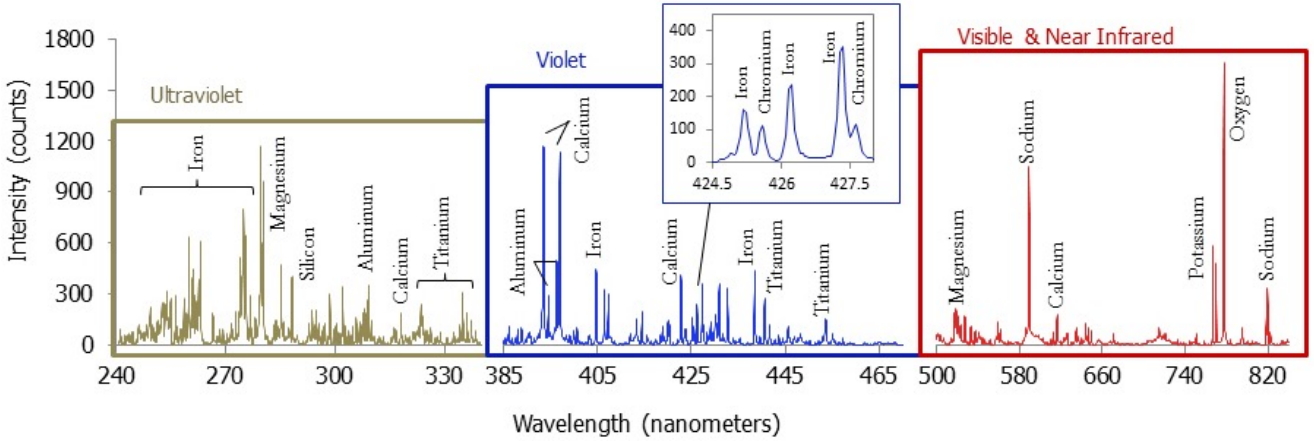
- Laser-induced breakdown spectroscopy (LIBS)
- Gives detailed information about the chemical composition of rock



Example: an Instrument on the Mars Rover

Mars Curiosity Rover: the ChemCam instrument

- Laser-induced breakdown spectroscopy (LIBS)
- Gives detailed information about the chemical composition of rock



Given an observation: Have we seen something like this before?

Novelty Detection

- Past observations: $x_1, x_2, \ldots, x_n$ from some space $\mathcal{X}$
- Now you see x
- Is it something familiar, or something new that warrants attention?

Novelty Detection

- Past observations: $x_1, x_2, \ldots, x_n$ from some space $\mathcal{X}$
- Now you see x
- Is it something familiar, or something new that warrants attention?

Nearest neighbor approach:

- Fix a distance function d on $\mathcal{X}$
- Find $\min_i d(x_i, x)$
- If this distance is large: x is something new

Novelty Detection

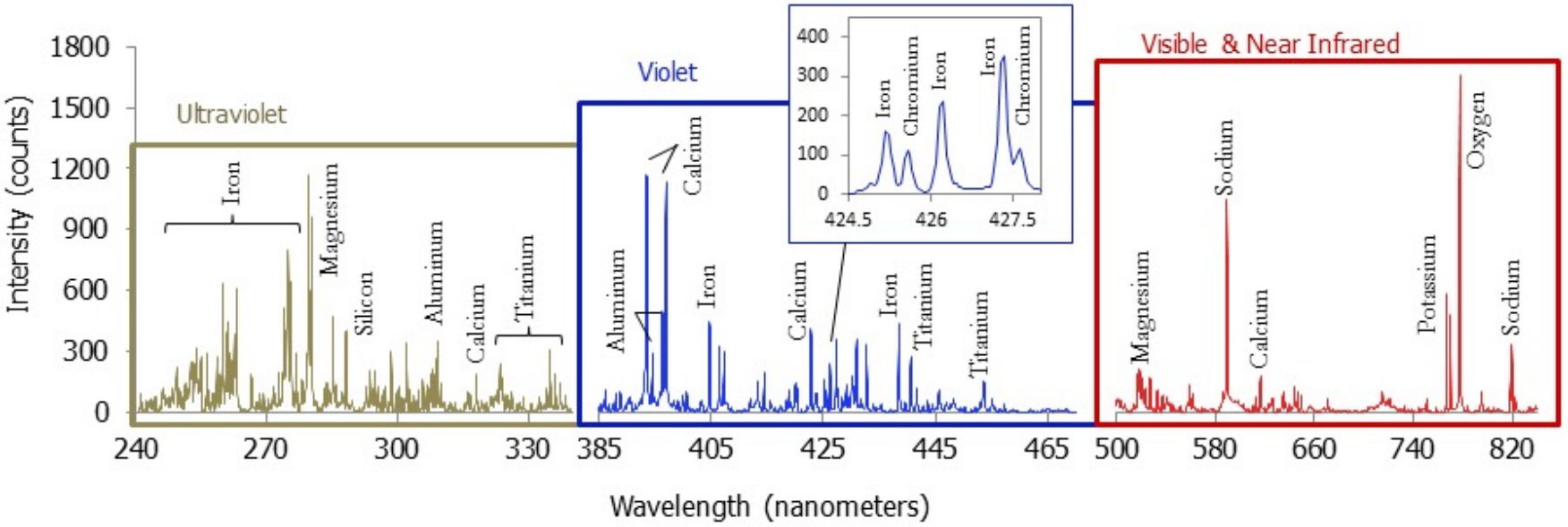
- Past observations: $x_1, x_2, \ldots, x_n$ from some space $\mathcal{X}$
- Now you see x
- Is it something familiar, or something new that warrants attention?

Nearest neighbor approach:

- Fix a distance function d on $\mathcal{X}$
- Find $\min_i d(x_i, x)$
- If this distance is large: x is something new

ChemCam example: What is $\mathcal{X}$, and what is the distance function?

A ChemCam Observation



A ChemCam Observation

# wave	shot1	shot2	shot3	shot4	shot5	shot6
240.811	2.97E+11	2.61E+11	3.45E+11	2.99E+11	2.93E+11	3.07E+11
240.86501	1.50E+11	1.32E+11	1.22E+11	1.17E+11	6.16E+10	9.10E+10
240.918	1.06E+11	1.31E+11	8.70E+10	7.35E+10	1.04E+11	7.50E+10
240.972	1.09E+11	1.09E+11	1.67E+11	1.92E+11	1.43E+11	1.75E+11
241.02699	3.59E+11	4.78E+11	5.33E+11	4.23E+11	4.35E+11	5.27E+11
241.07899	8.83E+11	9.92E+11	1.13E+12	1.01E+12	1.04E+12	1.08E+12
241.133	1.06E+12	1.18E+12	1.42E+12	1.26E+12	1.28E+12	1.38E+12
241.188	7.63E+11	8.49E+11	1.06E+12	9.59E+11	9.22E+11	1.02E+12
241.24001	2.88E+11	3.21E+11	4.30E+11	4.09E+11	3.71E+11	4.04E+11
241.29401	1.88E+11	1.79E+11	2.78E+11	2.30E+11	1.85E+11	2.15E+11
241.34801	3.14E+11	4.13E+11	4.12E+11	4.25E+11	4.04E+11	3.66E+11
241.401	4.71E+11	5.03E+11	5.99E+11	4.97E+11	5.12E+11	5.65E+11
241.45599	3.62E+11	3.52E+11	3.61E+11	3.50E+11	3.69E+11	4.13E+11
241.508	1.10E+11	1.65E+11	1.89E+11	1.72E+11	1.27E+11	1.92E+11
241.562	5.23E+10	6.94E+10	1.30E+11	3.23E+10	6.19E+10	9.61E+10

A single observation can be represented by a 6144-dimensional vector.

Picking a Distance Function

The most familiar option: **Euclidean, or ℓ_2, distance.**