Application of Data Science

CMP020L014 S/A

What to expect

- This module will review different applications where data science has had a tremendous impact in the last decade.
- Data Science has dominated almost all the industries of the world today. Various sectors like marketing, agriculture, journalism, security, healthcare, etc., use data science.
- The aim of this module is to develop students' awareness of the application areas of data science techniques.
- Learning Outcomes
  - LO1: Demonstrate a comprehensive understanding of current developments in data science.
  - LO2: Systematically and critically analyse and evaluate diverse sources of data to solve a problem.
  - LO3: Propose and develop a data science solution for a complex dataset.

Topics and Assessment

Topics	Week(s)
Introduction to explainable artificial intelligence	1
Mathematical morphology and noise filters	2
Image data padding and thresholding	2
Video analytics	2
Visual feature extraction: Histogram of oriented gradients	1
Image data enhancement and augmentation	1
Time-dependent data	2
Sampling theorem	1
Advanced parametric and non-parametric statistical tests	If time allows
Structural equation modelling	

Assessment	Weight	Submission method	Due date
Presentation (Coursework Milestone 1)	50%	In-class	**Week-9 or later**
Final Coursework	50%	Moodle Submission	**Assessment week**

Make a group of 3

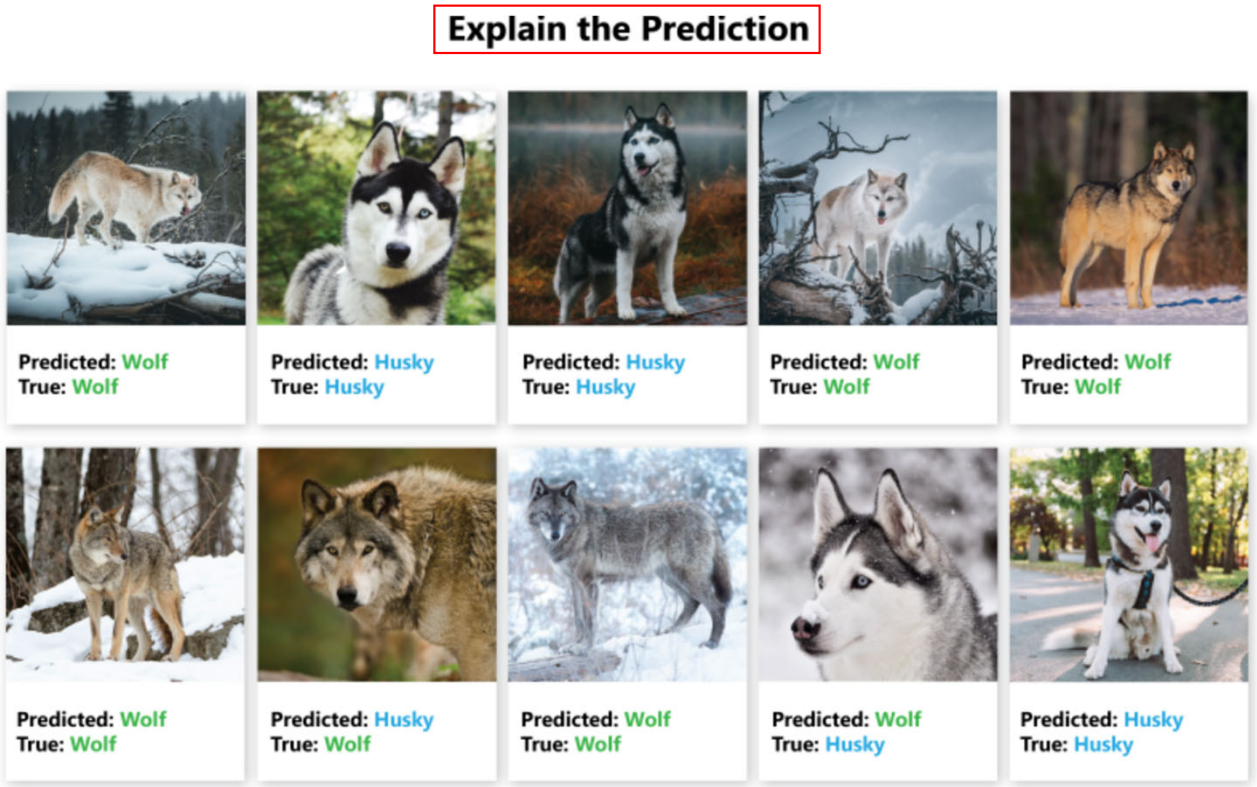
Plagiarism during coursework submission

- Plagiarism while submitting the coursework is seen as presenting someone else's work or ideas as your own, with or without their consent, by incorporating it into your work without full acknowledgment.
  - All published and unpublished material, whether in manuscript, printed or electronic form, is covered under this definition.
- Plagiarism during the assessment is unethical and can have serious consequences.
  - The University regards plagiarism in examinations/assessments as a serious matter.
  - Cases will be investigated, and penalties may range from deduction of marks to expulsion from the University, depending on the seriousness of the occurrence.
- You have come to university to learn to know and speak your own mind, not merely to reproduce the opinions of others.
- While submitting, if any coursework shows more than 20% similarity will likely be rejected.

Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI)

- XAI is a field that focuses on making AI systems and their decisions understandable to humans.
- In simpler terms, it's about adding transparency to AI, so we can understand why it makes certain choices and the reasoning behind its actions.



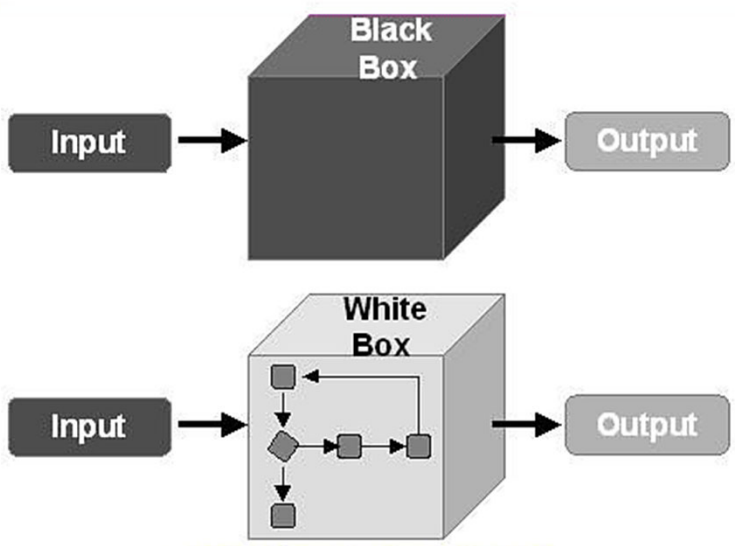
What is bias in AI?

Why XAI is Important?

- Trust and acceptance: When we understand how AI works, we're more likely to trust its decisions and accept its recommendations.
- Bias detection and mitigation: AI models can be susceptible to biases, leading to unfair or discriminatory outcomes. XAI techniques can help us identify these biases and mitigate them.
- Debugging and improvement: Understanding how AI models work makes it easier to debug them when they make mistakes. We can also use this knowledge to improve their accuracy and performance.

Approaches to XAI

- Depending on the complexity of the AI model and the desired level of explanation. Some common techniques include:
  - Feature importance: This identifies which features of the input data have the most influence on the AI's decision.
  - Counterfactual explanations: These explain how the AI's decision would have changed if the input data had been different.

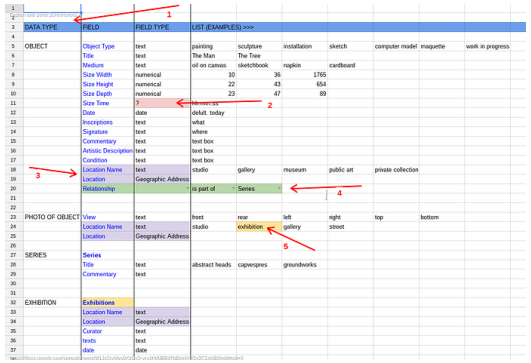


Do we need to glow the whole box?

Data & Processing

Types of Data

- Structured data: This type of data is organized in a fixed format, like rows and columns in a spreadsheet or database.



- Unstructured data: This type of data has no predefined format and can be messy and complex. Examples include text documents, images, videos, audio, etc.

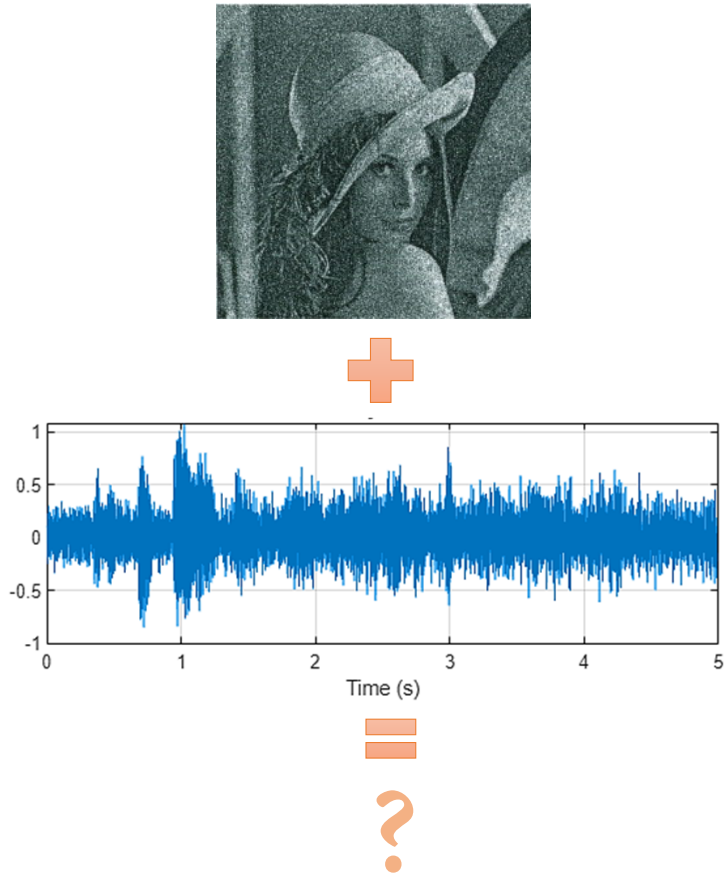
Which one is more relevant to real-world?

Good Data vs Big Data

- The notion of "good data" being more important than "big data“.
  - It's more about prioritising quality over quantity.
  - Big data boasts impressive volume, while good data emphasises meaningful information and accurate insights.
- Quality over Quantity:
  - Garbage in, garbage out
  - Focus on relevance: Not all data is created equally.
- Ethical Considerations:
  - Bias and privacy: Big data often raises concerns about privacy and algorithmic bias.
  - Transparency and trust: Trust in data-driven decisions requires transparency in data collection and analysis. Good data fosters trust by adhering to ethical principles and responsible practices.

Advantages of Data Pre-processing

- Improved explainability and interpretability:
  - Cleaning and filtering data: Removing noise, outliers, and inconsistencies in the data.
  - Feature scaling and normalisation: Ensuring features are on the same scale can prevent dominant features from overshadowing the contributions of others.
  - Feature engineering: Creating new features or combining existing ones to better capture relevant information can make explanations more concise and understandable.

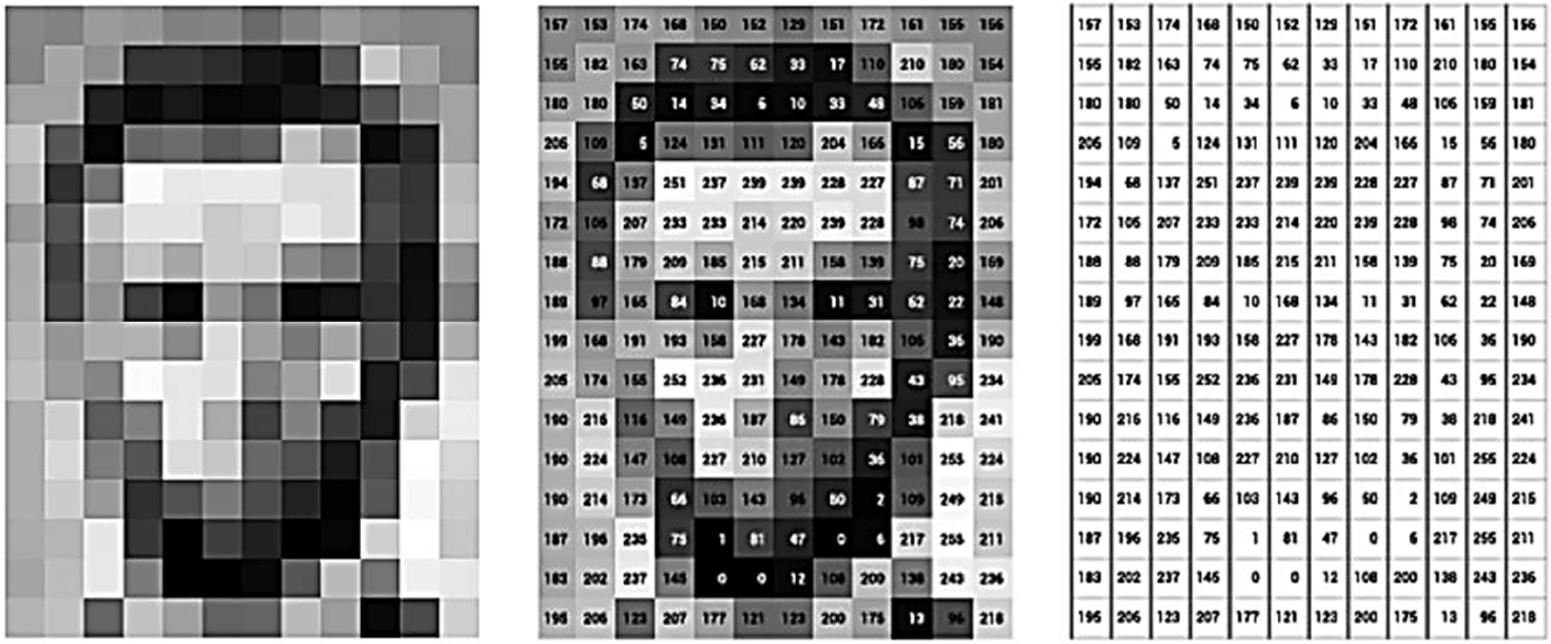


Advantages of Data Pre-processing

- Efficient model training and explanation generation:
  - Reduced complexity: Pre-processed data is often smaller and easier to work with, leading to faster training times and more efficient explanation generation, especially for complex models.
  - Resource-saving: By ensuring data quality and consistency, pre-processing can optimize resource utilization during model training and explanation, reducing computational costs.

Image Data

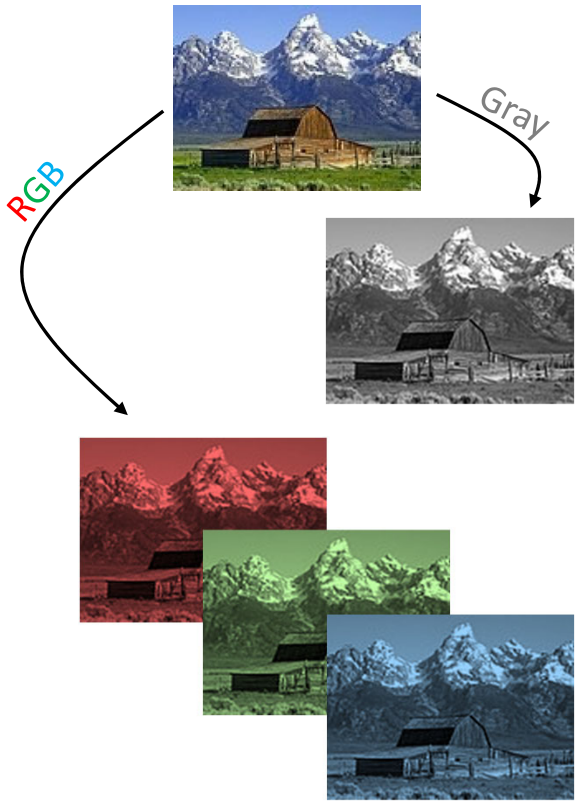
Abraham Lincoln (Matrix)



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Image of Abraham Lincoln as a matrix of pixel values

Image

- An image is a visual representation of something.
- It can be 2D/3D, which can be fed into the visual system to convey information.
- In the context of data processing, an image is a distributed amplitude of color(s).
- The smallest element of image is called pixel.
  - Pixel is a point on the image that takes on a specific shade or color. In data science, 2D/3D images usually represented in the following way:
    - Grayscale - A pixel is an integer with a value between 0 to 255 (0 is completely black and 255 is completely white).
    - RGB - A pixel is made up of 3 integers between 0 to 255 (the integers represent the intensity of red, green, and blue).