VECTOR INSTITUTE

Bias in AI:
Week #5: Course Review & Capstone Project

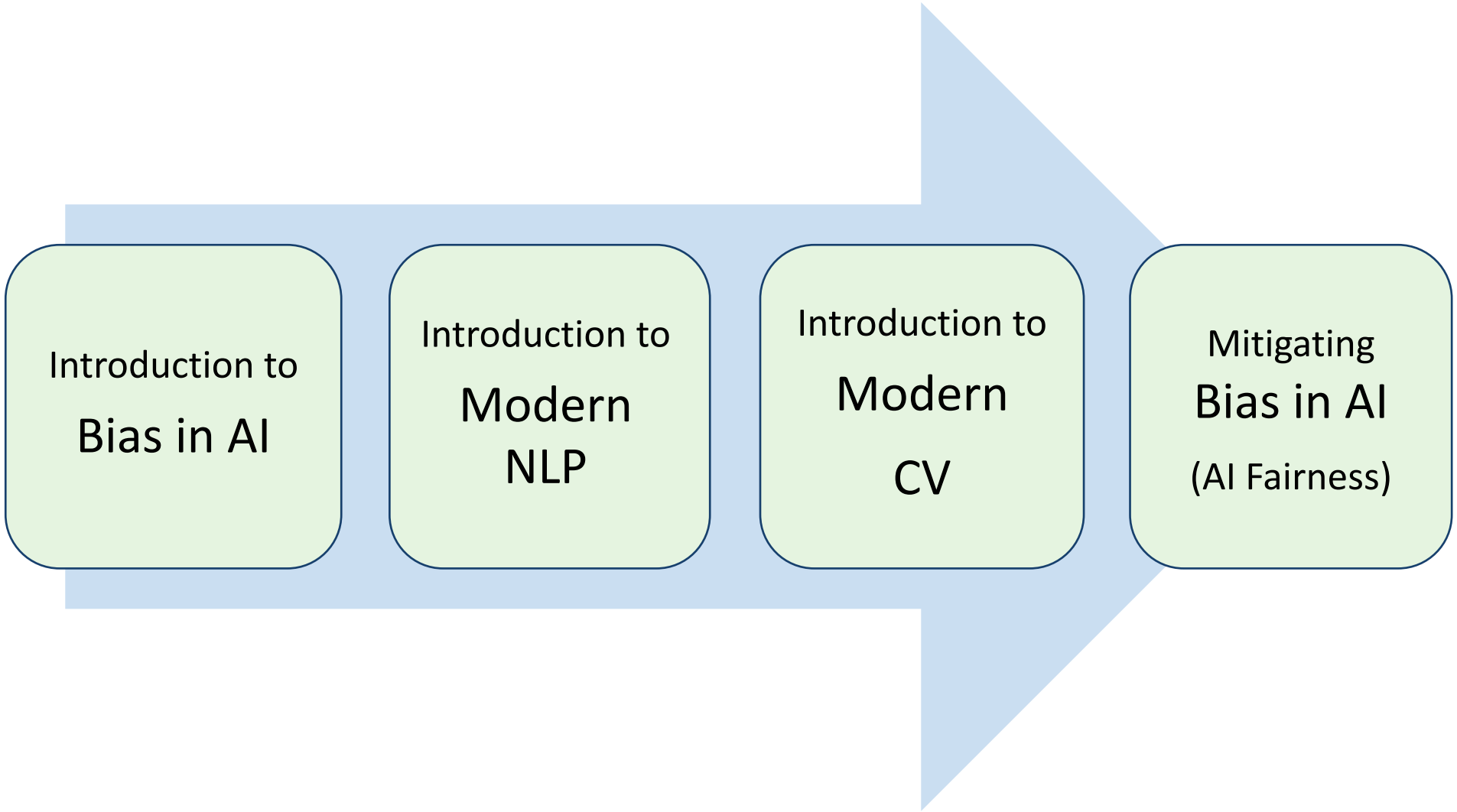
Instructor:
Sayyed Nezhadi

Winter 2023

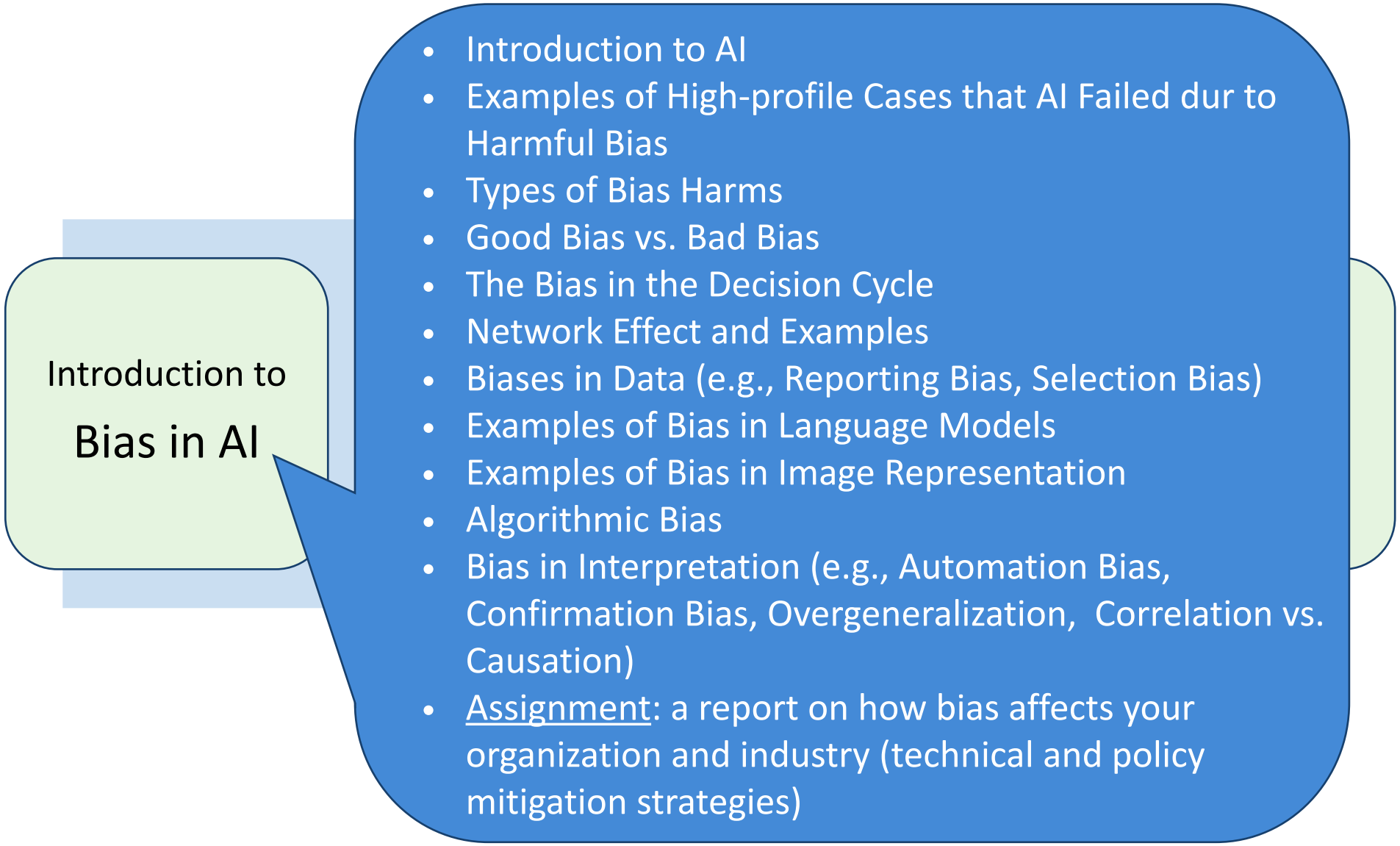
What was the Goal?

- Create **awareness** about Bias in AI, sources and types, and potential harmful impacts
- Introduce potential approaches to identify and mitigate bias in AI systems
- Practical Examples in NLP & Computer Vision
- **Apply** learnings to your own businesses, to innovate, improve processes, solve problems

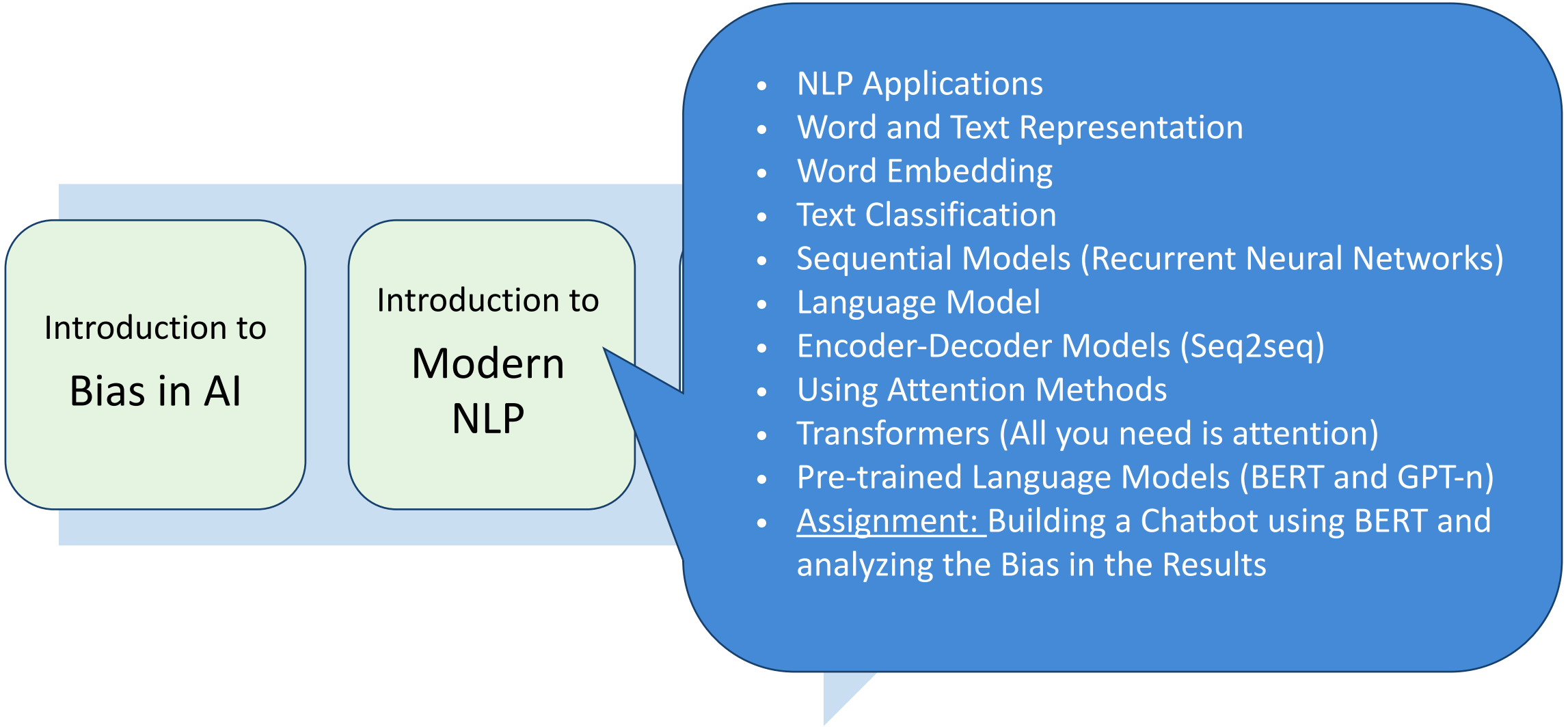
What We Discussed (Review)



What We Discussed (Review)



What We Discussed (Review)



What We Discussed (Review)

- CV Applications
- Image Representation
- Image Classification
- Convolutional Neural Networks
- ImageNet Models
- Transfer Learning and Fine-tuning
- Object Detection Methods
- Image to Text (Image Captioning)
- Encoder-decoder Models
- Attention Methods
- Generative Models
- Generative Adversarial Networks

Introduction to

Modern CV

- <u>Assignment:</u>
  - Training a computer vision model for gender classification
  - Analyzing the bias in the results

What We Discussed (Review)

Final Thoughts

Class Discussion

Capstone Project (one project by company*)

- Subject: Identify and reduce bias in your AI models.
- Objectives:
  - Explore and identify a possible bias issue in your company's AI models, products, or processes
  - Assess the type and the source of the bias, the potential harmful effects, the business and societal impacts, and associated regulations
  - Explore the potential technical and policy mitigation strategies
  - Choose the best feasible strategies to make the model/product/process more fair
  - Apply the selected technical strategies
  - Analyze the results and their impacts and/or trade-offs
  - Define future direction

Project Report

- Your project report should be written like a research-based business paper and should include the following sections :
  - Executive Summary (~250 words)
  - Introduction
  - Problem Definition
  - Model, Product, or Process
  - Explaining Data (e.g., Training Data)
  - Identified Bias
  - Impact Analysis
  - Mitigation Strategies
  - Experiments
  - Results
  - Conclusion & Future Work
  - References

Project Report

- Due by the end of Monday Feb 20th, 2023
- 8-10 pages long (excluding references)
- No need to include super sensitive content/ trade secrets
- The report should be created by LaTex (if you don't have LaTex on your computer, use https://www.overleaf.com/)
- Follow the suggested format by NeurIPS conference
- You can find the style files for NeurIPS 2020 here:
  https://nips.cc/Conferences/2020/PaperInformation/StyleFiles

Final Presentations

- Presentation Date: Monday Feb 27th 2023, 10 AM-12 PM
- Presentation Length: 15 minutes followed by 5 minutes Q&A
- Capstone Panel: Vector Institute & CRA Stakeholders
- Submission Deadline: Presentation documents can be submitted before the presentations by the end of Sunday Feb 26th, 2023

Evaluation (Rubric)

Item	L4: Excellent (100%)	L3: Good (75%)	L2: Satisfactory (50%)	L1: Needs Improvement (25%)
Identifying Bias (15%)	Clearly defined problem statement, evaluated multiple candidates, selected an impactful case	Good problem statement, evaluated a couple of candidates, selected a moderately impactful case	Loosely defined problem statement, evaluated only one use-case, selected a case with low impact	No problem statement, No exploration / evaluation, selected a case with low impact
Impact Analysis (15%)	Clearly identified the type and source of the bias, analyzed the harms, impact analysis using an impact assessment framework	Identified the type and source of the bias, good impact analysis but without a standard framework	Either type or source of the bias is identified, loosely impact analysis	Type and source are not defined, no impact analysis is done, impacts ae only named
Mitigation Strategy (15%)	Explored the potential technical and policy mitigation strategies, reasoned the strategy selection clearly with limitations and strengths	Explored the potential technical and policy mitigation strategies, reasoned the strategy selection	Explored either the potential technical and policy mitigation strategies, loosely reasoned the strategy selection	Selected the strategy with no exploration and without reasoning

Evaluation (Rubric) - continued

Item	L4: Excellent (100%)	L3: Good (75%)	L2: Satisfactory (50%)	L1: Needs Improvement (25%)
Analyzing the Results (15%)	Applied the strategy completely, Analyzed the results clearly, evaluated impacts and trade-offs, drew concrete conclusions	Applied the strategy, Analyzed the results, reviewed some impacts and trade-offs, drew conclusions	Applied the strategy loosely, Analyzed the results, drew conclusions	Incomplete application, provided the results without analysis and without impact analysis, conclusions without strong evidence
Project Report (20%)	Well formatted, included executive summary and all other listed sections (including references) withing the page limit, clearly articulated, used proper visualization to convey the message, submitted on time	Well formatted, included executive summary and most other listed sections withing the page limit, used some visualization to convey the message, submitted on time	Included executive summary and most other listed sections, used some visualization to convey the message, submitted on time	Formatted incorrectly, no executive summary or conclusion, no visualization to convey the message, submitted late
Presentation (20%)	Well formatted, used effective visualizations, conveyed the message clearly, presented within the time limit	Well formatted, used some visualizations, conveyed the message, presented within the time limit	Moderately formatted, used some visualizations, conveyed the message, presented close to the time limit	Not formatted properly, Mostly text, difficult to understand the message, too short or long