The Shortest Path to Ethics in AI

An Integrated Assignment Where Human Concerns Guide Technical Decisions

Noelle Brown

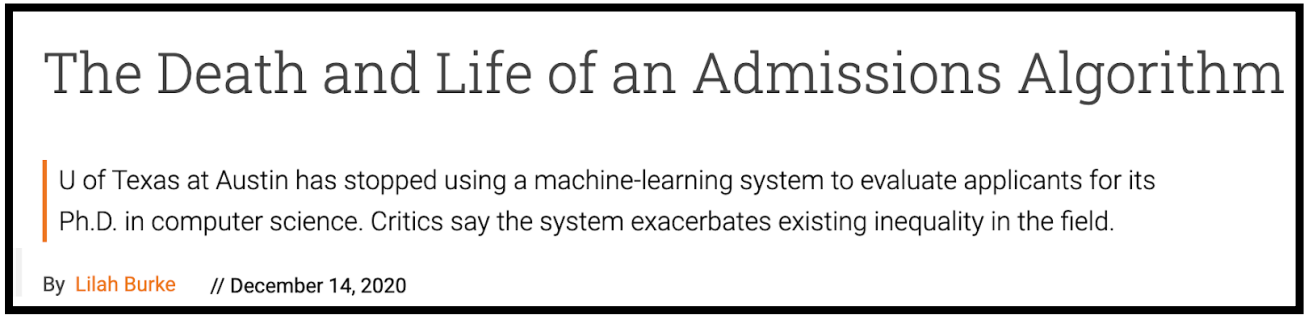
Koriann South

Eliane Wiese

University of Utah

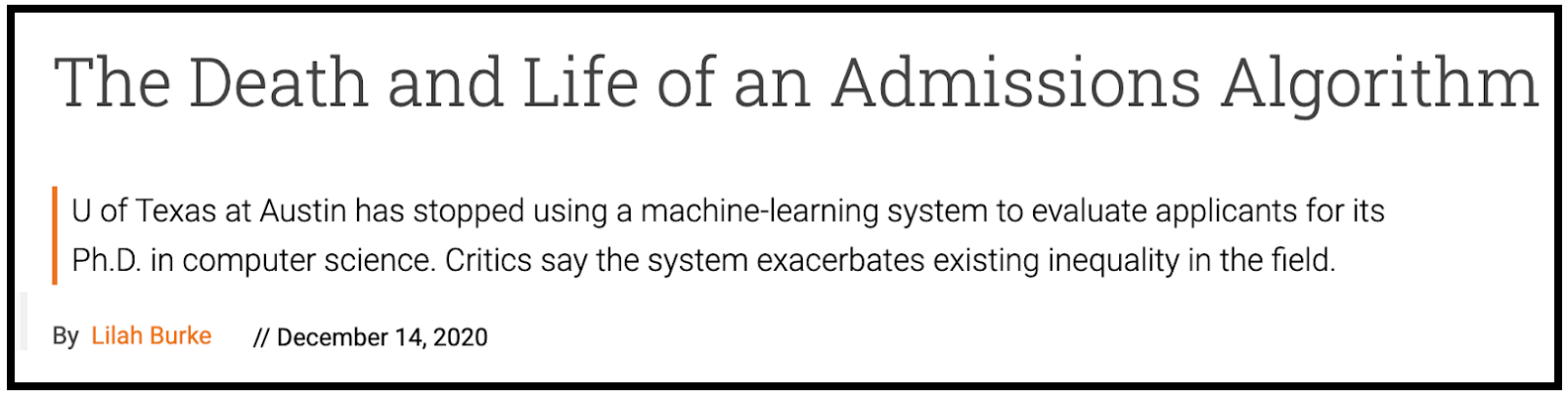
noelle.brown@utah.edu

noelleb.com



AI can be ethically problematic

Common methods for integrating ethics within technical courses



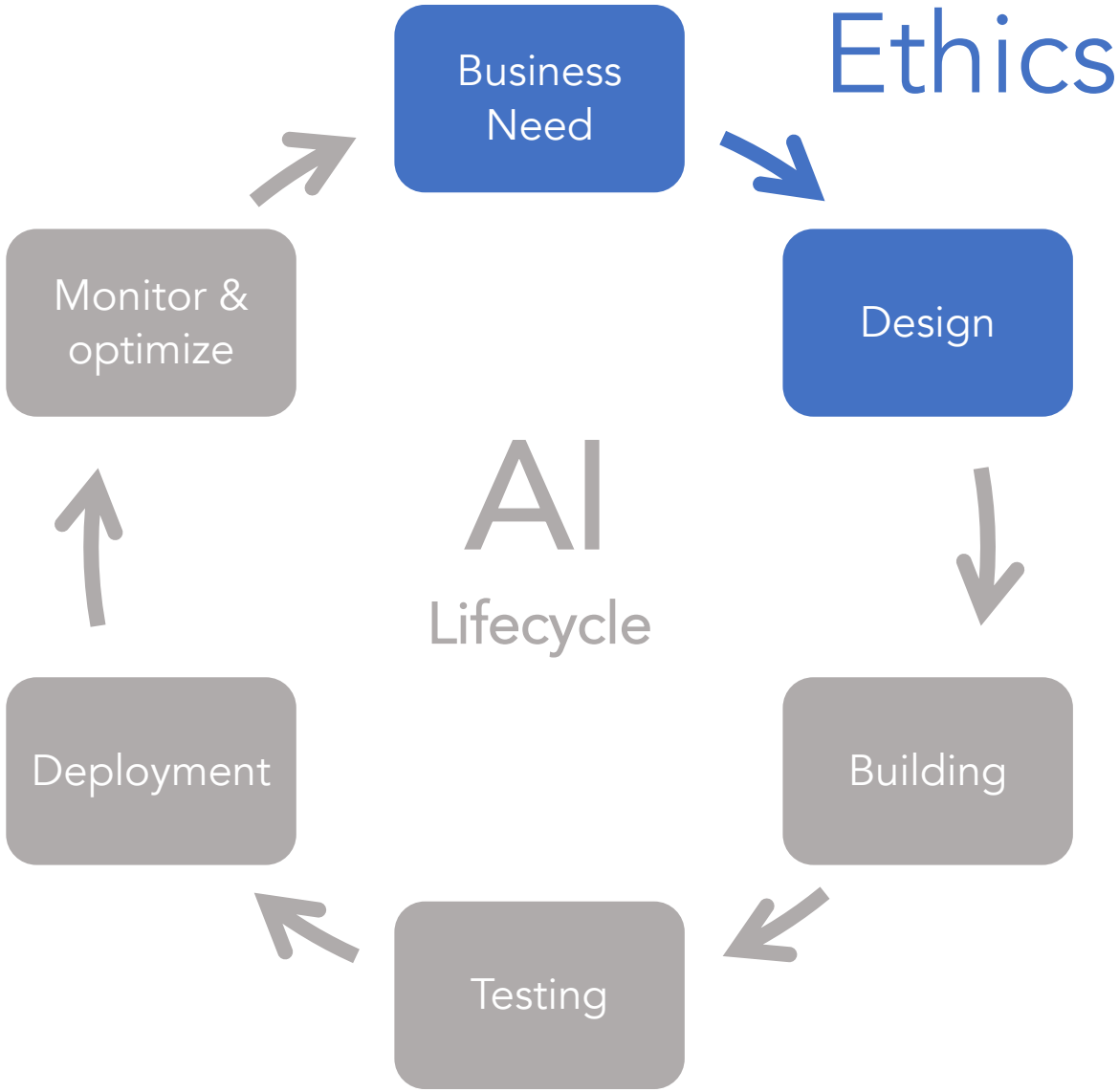
The Death and Life of an Admissions Algorithm

U of Texas at Austin has stopped using a machine-learning system to evaluate applicants for its Ph.D. in computer science. Critics say the system exacerbates existing inequality in the field.

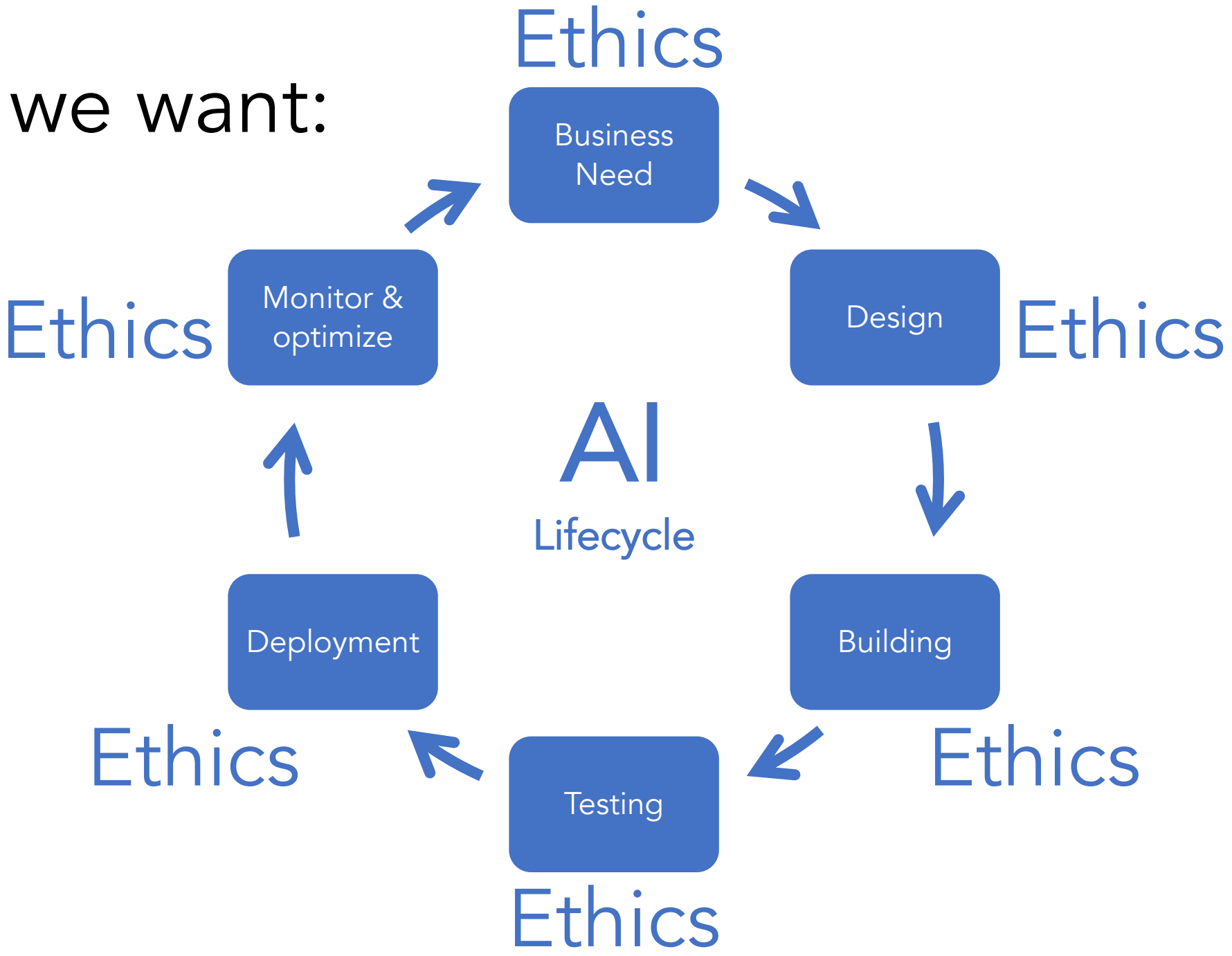
By Lilah Burke // December 14, 2020

- Discuss the benefits and potential harms of using this algorithm
- Write an essay arguing whether this should be used

Currently:


Business Need
Ethics
Design
Building
Testing
Deployment
Monitor & optimize
AI
Lifecycle

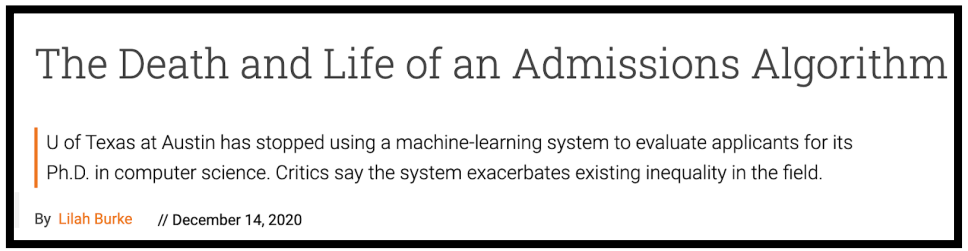
What we want:


Ethics
Business Need
Design
Ethics
Ethics
Monitor & optimize
AI
Lifecycle
Deployment
Building
Ethics
Ethics
Testing
Ethics

AI course learning outcomes

- Graph search algorithms
- Constraint satisfaction problems
- Markov models
- Reinforcement learning
- Bayes' nets
- ML/NLP algorithms
- Computer vision

How can we get students to see the connection?

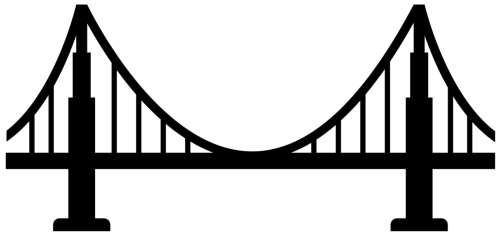


The Death and Life of an Admissions Algorithm

U of Texas at Austin has stopped using a machine-learning system to evaluate applicants for its Ph.D. in computer science. Critics say the system exacerbates existing inequality in the field.

By Lilah Burke // December 14, 2020

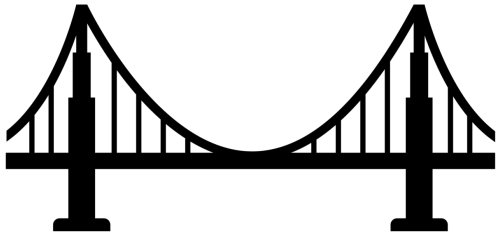
- Discuss the benefits and potential harms of using this algorithm
- Write an essay arguing whether this should be used



- Graph search algorithms
- Constraint satisfaction problems
- Markov models
- Reinforcement learning
- Bayes' nets
- ML/NLP algorithms
- Computer vision

How can we get students to see the connection?

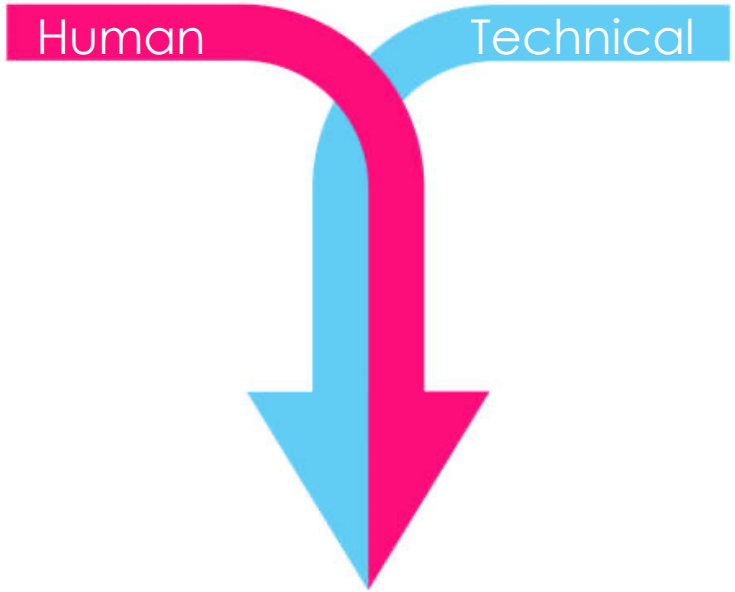
Considering potential human harm



Graph search algorithms

How can we get students to see the connection?

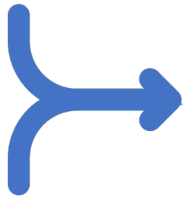
Considering potential human harm



Graph search algorithms

Integrated Assignment

Successes for the Integrated Assignment

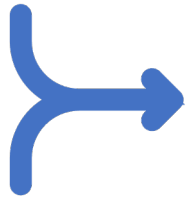


Elicited technical reasoning that responded to human concerns.

Deepened the technical rigor, helping the instructor better understand his students' misconceptions.

Followed a general rubric that offers a template for future integrated assignments.

Successes for the Integrated Assignment

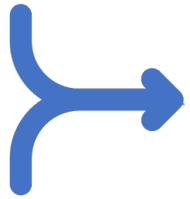


Elicited technical reasoning that responded to human concerns.

Deepened the technical rigor, helping the instructor better understand his students' misconceptions.

Followed a general rubric that offers a template for future integrated assignments.

Successes for the Integrated Assignment



Elicited technical reasoning that responded to human concerns.

Deepened the technical rigor, helping the instructor better understand his students' misconceptions.

Followed a general rubric that offers a template for future integrated assignments.

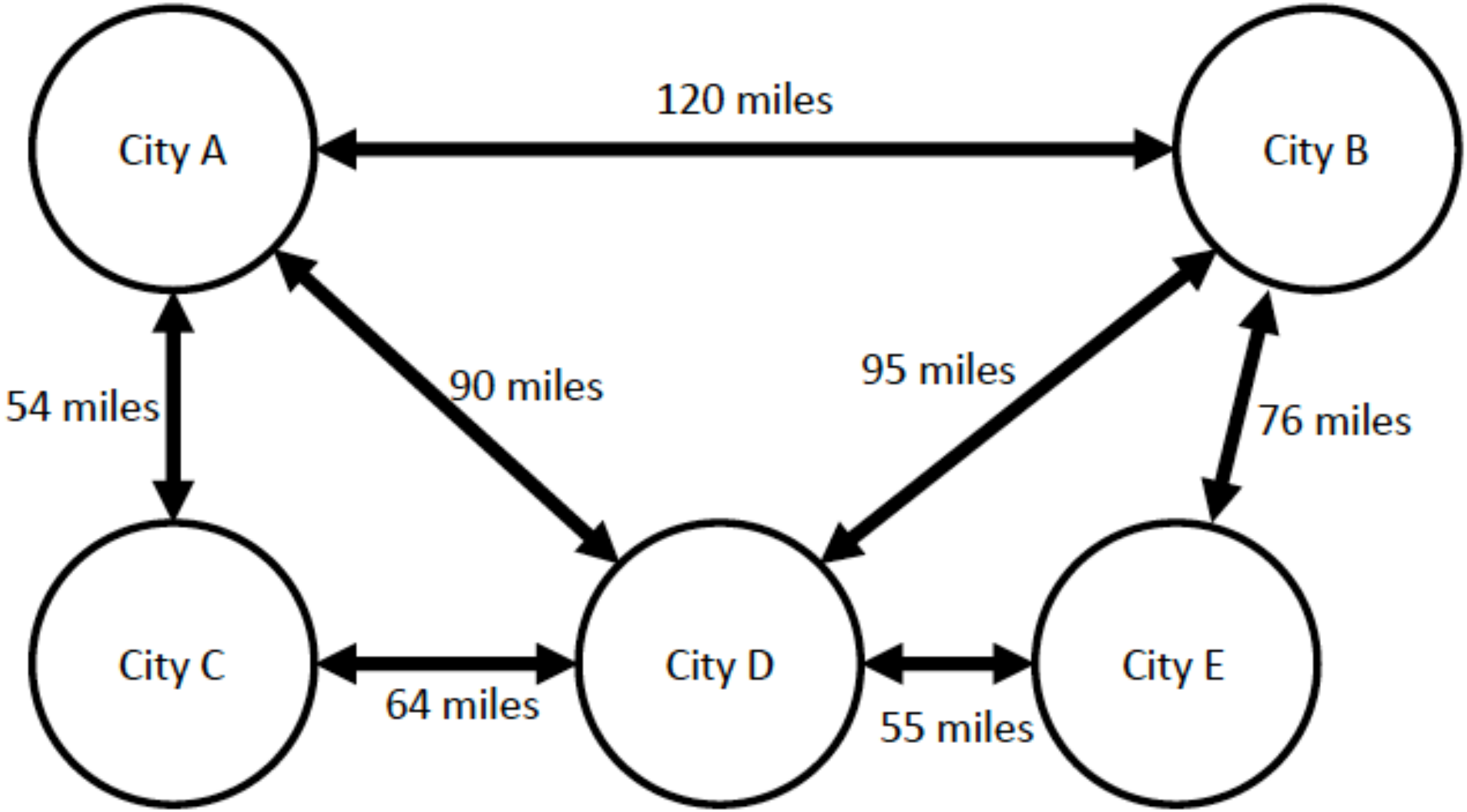
RQ: How can an assignment simultaneously assess technical knowledge and judgements in response to human contexts?

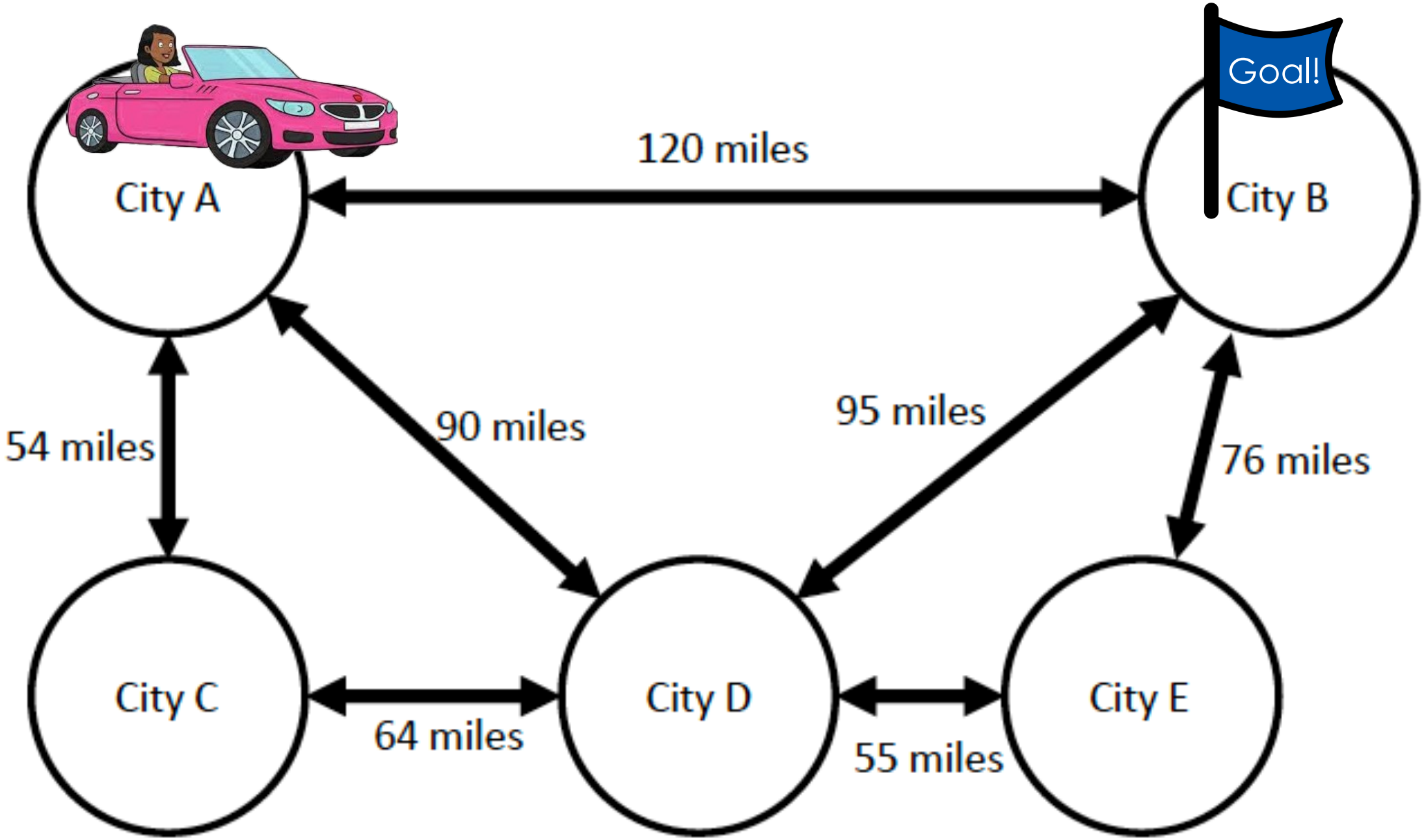
Overview

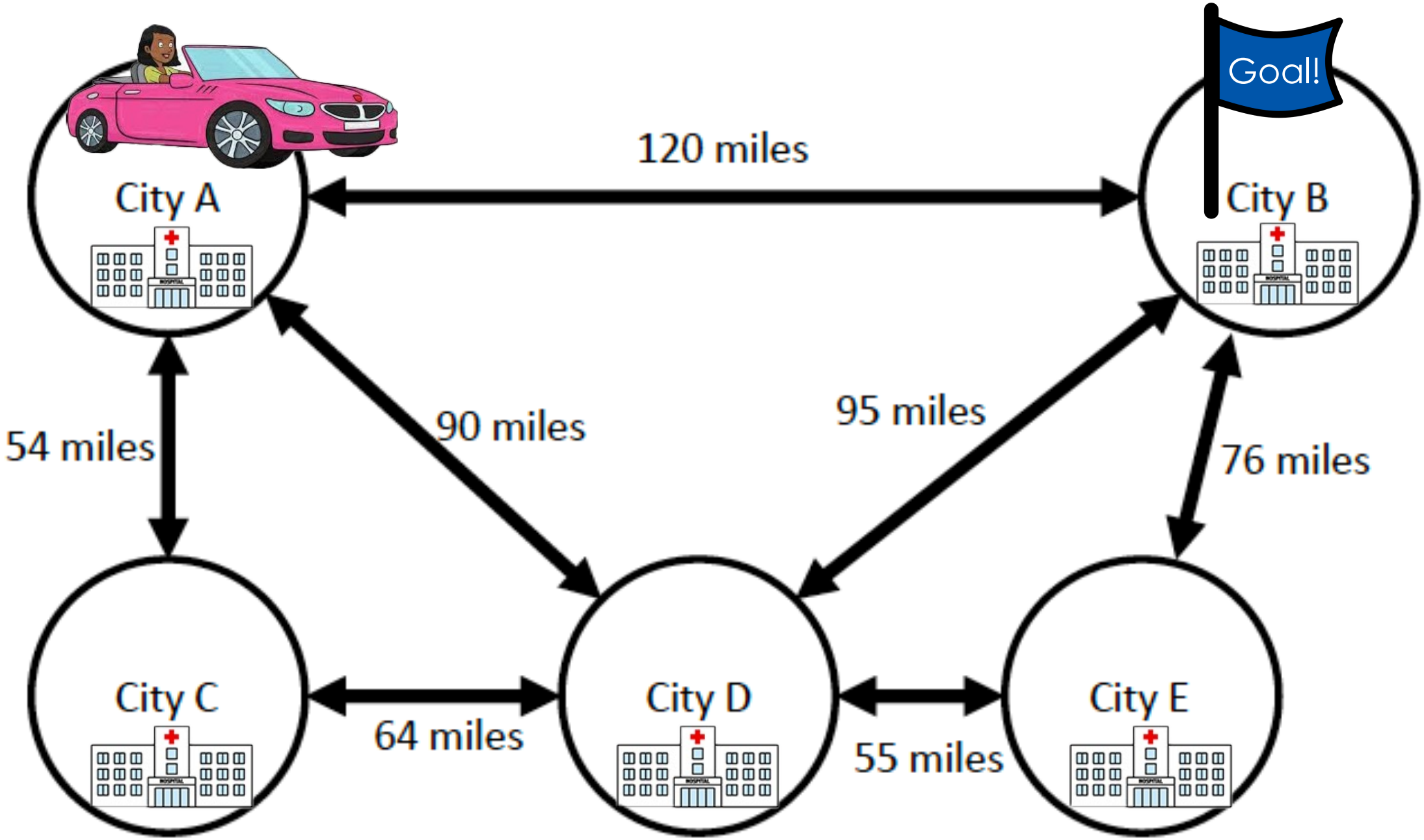
1. Course assignment
2. Rubric & scoring
   - Technical
   - Human
   - Integrated
3. Benefit: illuminates student misconceptions
4. Challenge: grading

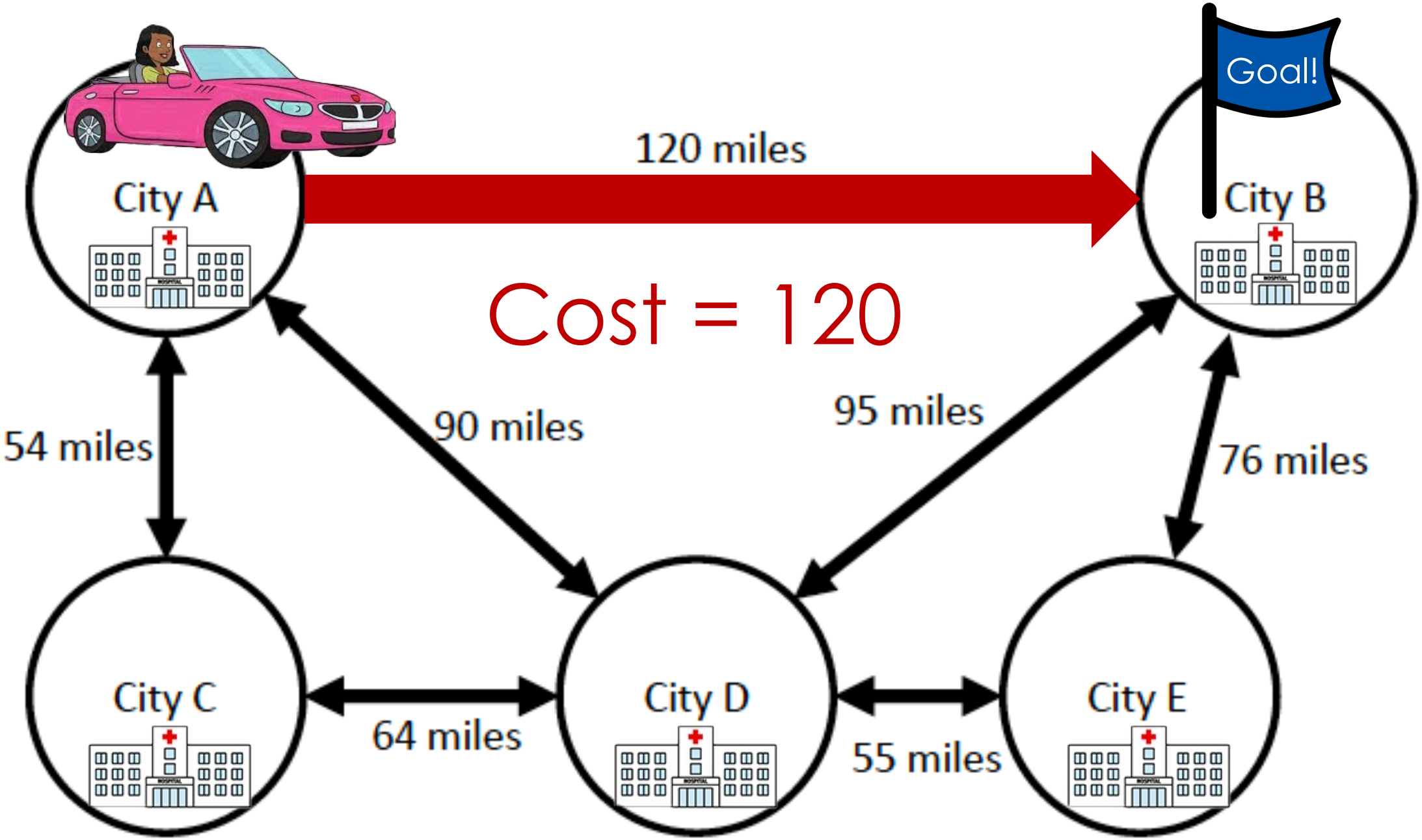
Overview

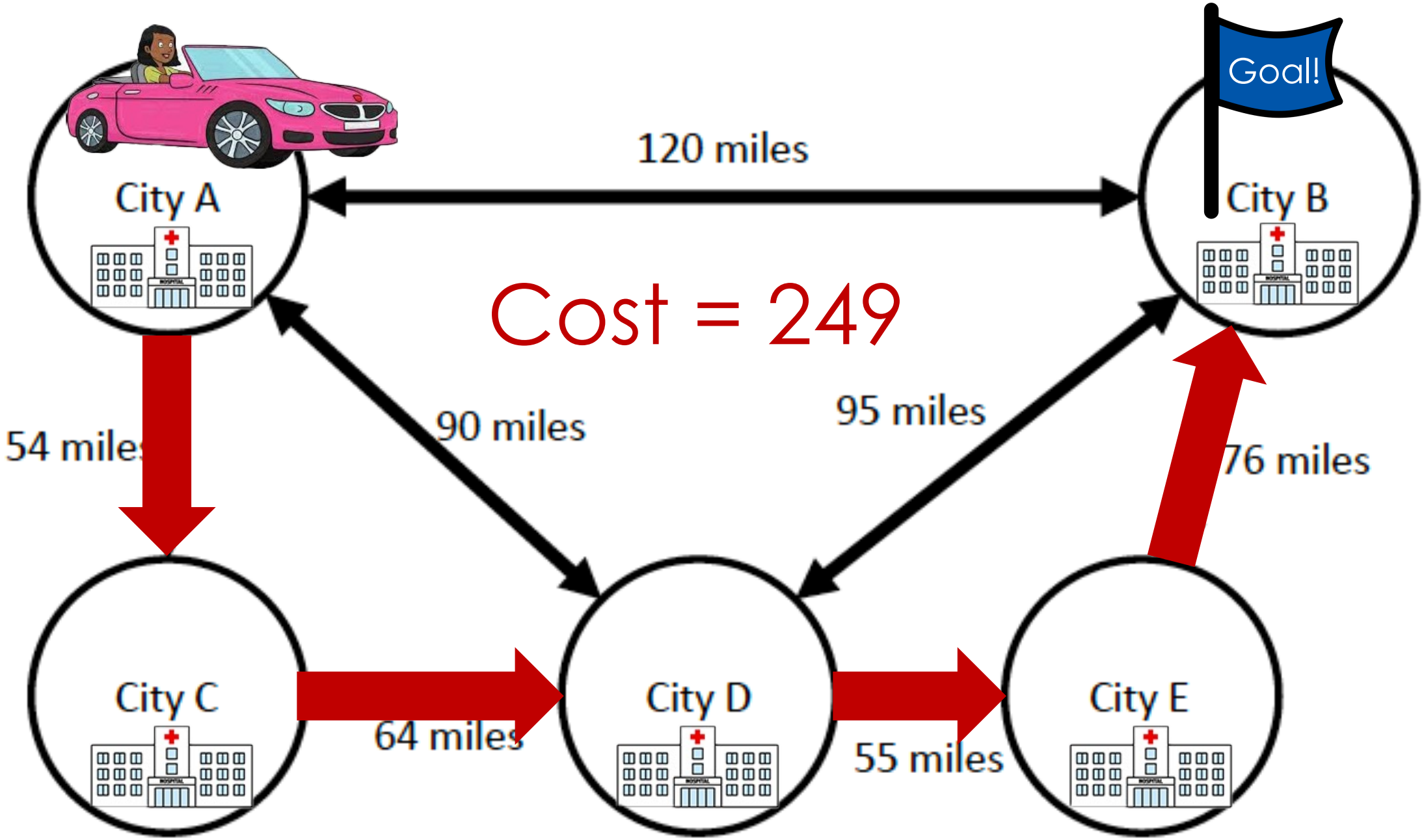
1. Course assignment
   - Given to 120 students
2. Rubric & scoring
   - Technical
   - Human
   - Integrated
3. Benefit: illuminates student misconceptions
4. Challenge: grading

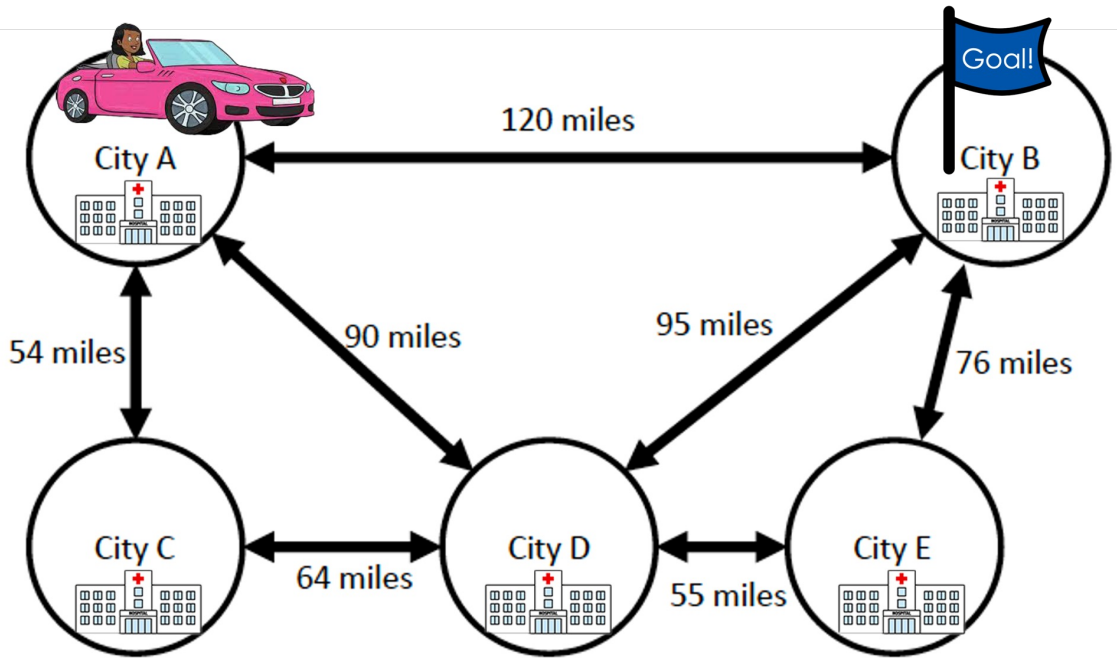

City A
City B
City C
City D
City E
120 miles
54 miles
90 miles
95 miles
76 miles
64 miles
55 miles


Goal!
City A
120 miles
City B
54 miles
90 miles
95 miles
76 miles
City C
64 miles
City D
55 miles
City E


Goal!
120 miles
City A
City B
54 miles
90 miles
95 miles
76 miles
City C
City D
City E
64 miles
55 miles


Goal!
120 miles
City A
City B
Cost = 120
54 miles
90 miles
95 miles
76 miles
City C
City D
City E
64 miles
55 miles


Goal!
120 miles
City A
City B
Cost = 249
54 miles
90 miles
95 miles
76 miles
City C
64 miles
City D
55 miles
City E



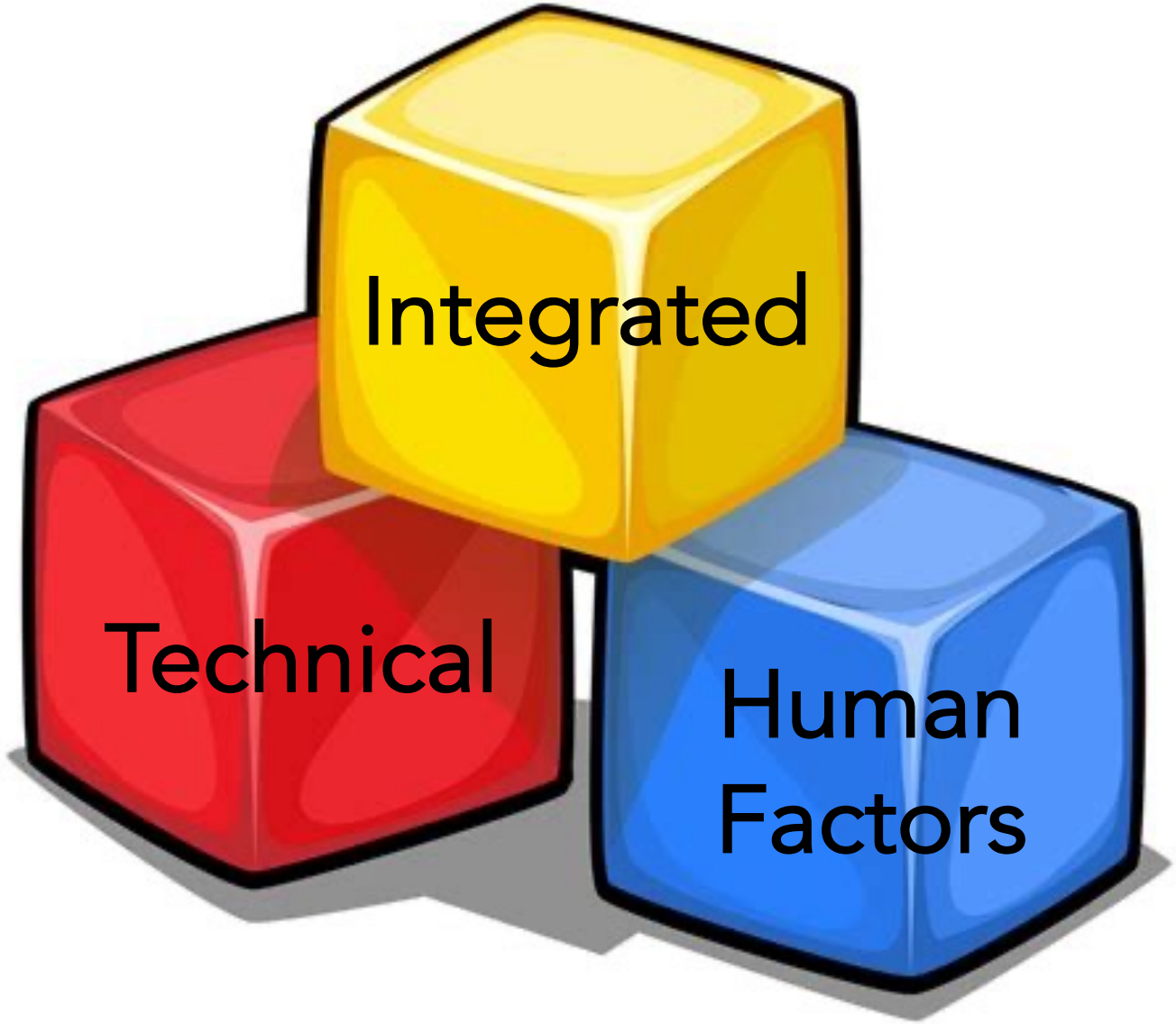
1. Describe an alternative cost function to use for this scenario.
2. Calculate the optimal path from A to B under your chosen cost function.

To make an ethical decision, there must be a context where there are concepts of right and wrong.

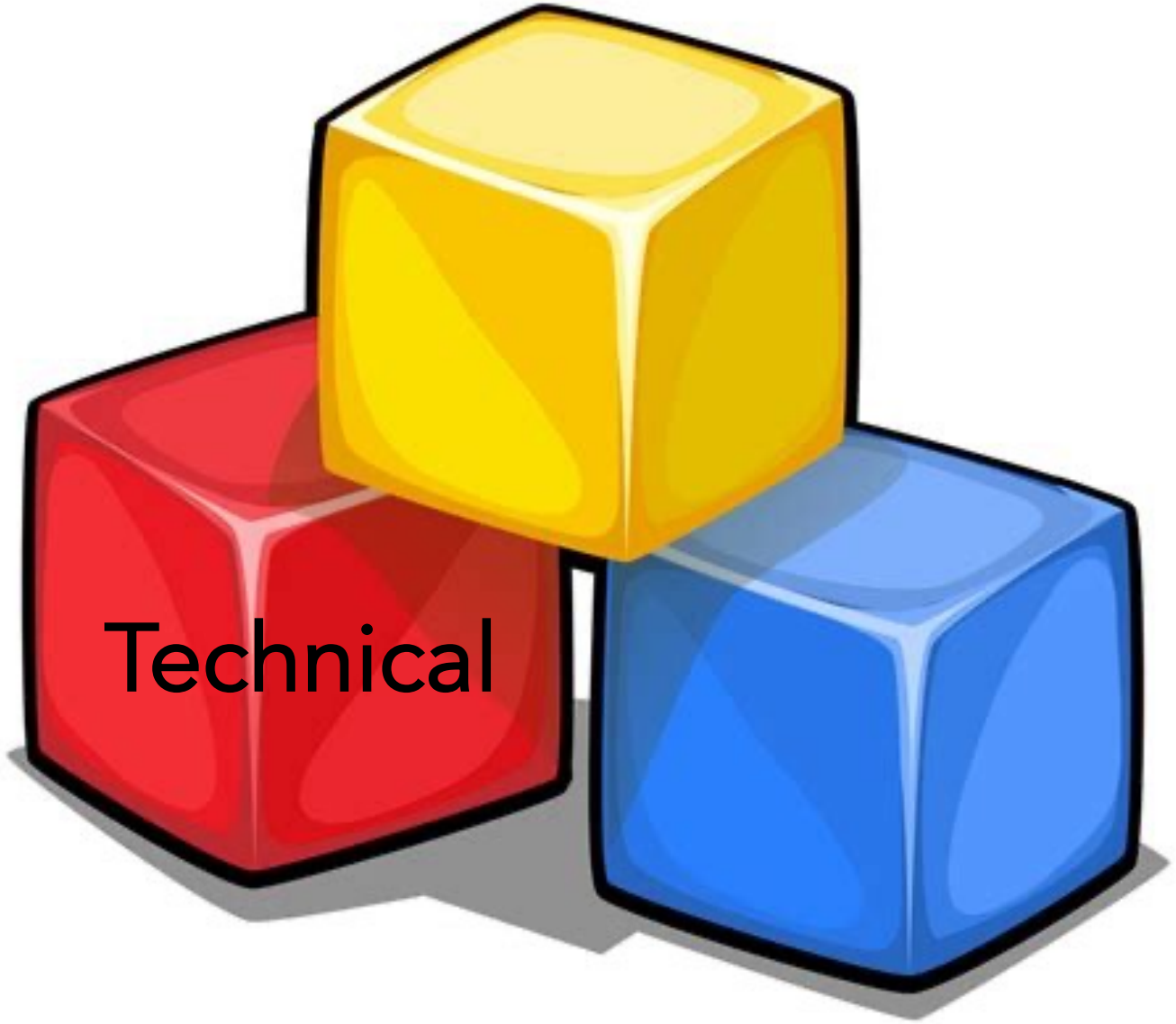
Overview

1. Course assignment
2. Rubric & scoring
   - Technical
   - Human
   - Integrated
3. Benefit: illuminates student misconceptions
4. Challenge: grading

Rubric Development



Demonstration of AI knowledge targeted by the course.



How well can a student define a technical solution?

Level 1
32%

No well-defined technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
24%

Partial technical implementation.

"A cost function could be to take the shortest path between cities: at each step, always take the shortest road."

Level 3
44%

Complete technical solution.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost* = *Distance*^{2}."

How well can a student define a technical solution?

Level 1
32%

No well-defined technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
24%

Partial technical implementation.

"A cost function could be to take the shortest path between cities: at each step, always take the shortest road."

Level 3
44%

Complete technical solution.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost* = *Distance*^{2}."

How well can a student define a technical solution?

Level 1
32%

No well-defined technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
24%

Partial technical implementation.

"A cost function could be to take the shortest path between cities: at each step, always take the shortest road."

Level 3
44%

Complete technical solution.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost = Distance*^{2}."

How well can a student define a technical solution?

Level 1
32%

No well-defined technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
24%

Partial technical implementation.

"A cost function could be to take the shortest path between cities: at each step, always take the shortest road."

Level 3
44%

Complete technical solution.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost = Distance*^{2}."

How well can a student define a technical solution?

Level 1
32%

No well-defined technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
24%

Partial technical implementation.

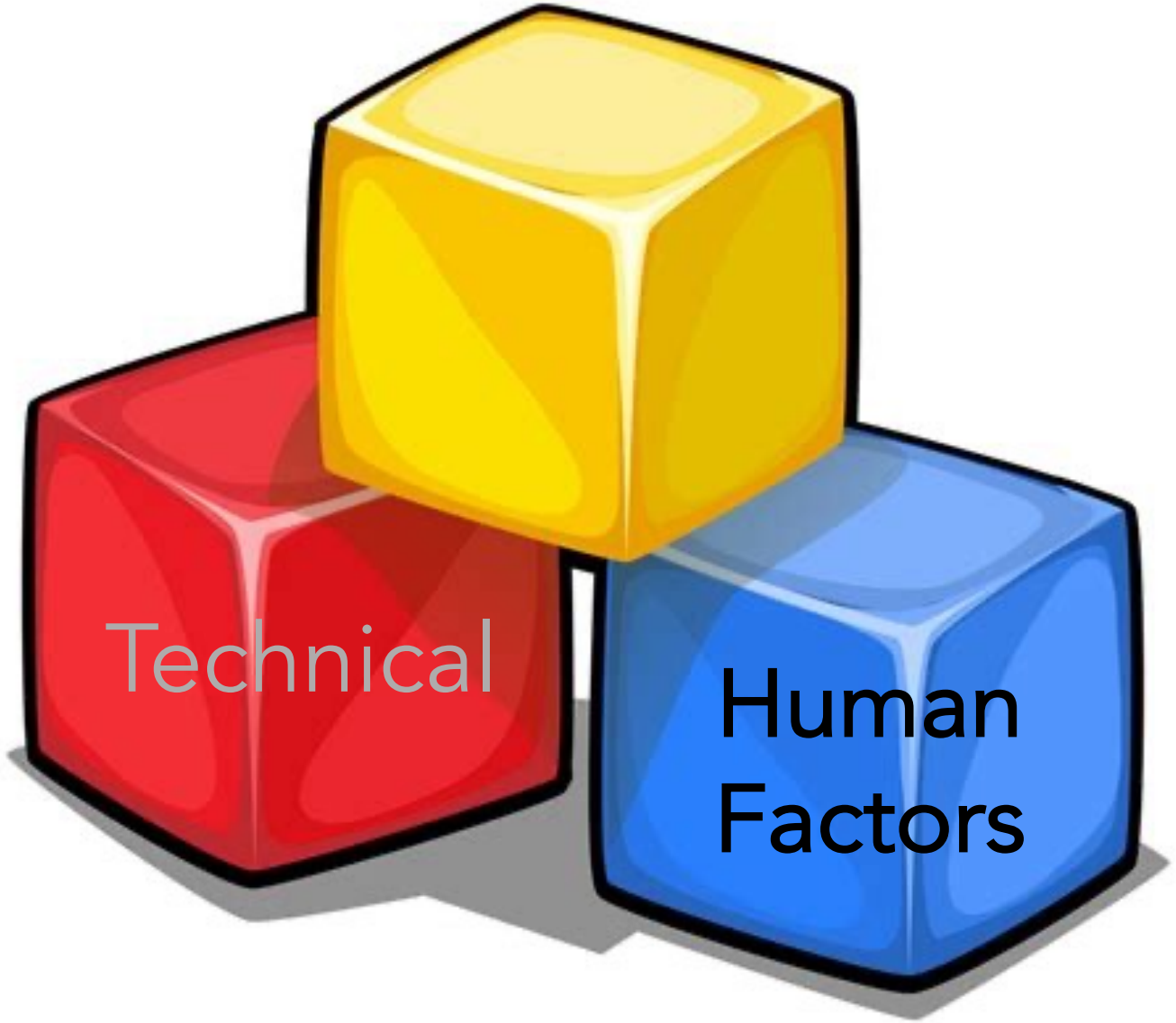
"A cost function could be to take the shortest path between cities: at each step, always take the shortest road."

Level 3
44%

Complete technical solution.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost = Distance*^{2}."

Identification of human concerns based on the problem context.



To what extent does a student consider the human context?

Level 1
0%

No consideration of human factors or context.

Level 2
84%

Considers human needs as they are explicitly presented.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 3
16%

Considers nuances of human needs.

"My answer depends on Jessie's priorities. If they want to travel quickly, then the shortest path is the best. But, if they want to travel safely, then this isn't the best option since it is far from medical care."

To what extent does a student consider the human context?

Level 1
0%

No consideration of human factors or context.

Level 2
84%

Considers human needs as they are explicitly presented.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 3
16%

Considers nuances of human needs.

"My answer depends on Jessie's priorities. If they want to travel quickly, then the shortest path is the best. But, if they want to travel safely, then this isn't the best option since it is far from medical care."

To what extent does a student consider the human context?

Level 1
0%

No consideration of human factors or context.

Level 2
84%

Considers human needs as they are explicitly presented.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 3
16%

Considers nuances of human needs.

"My answer depends on Jessie's priorities. If they want to travel quickly, then the shortest path is the best. But, if they want to travel safely, then this isn't the best option since it is far from medical care."

To what extent does a student consider the human context?

Level 1
0%

No consideration of human factors or context.

Level 2
84%

Considers human needs as they are explicitly presented.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 3
16%

Considers nuances of human needs.

"My answer depends on Jessie's priorities. If they want to travel quickly, then the shortest path is the best. But, if they want to travel safely, then this isn't the best option since it is far from medical care."

To what extent does a student consider the human context?

Level 1
0%

No consideration of human factors or context.

Level 2
84%

Considers human needs as they are explicitly presented.

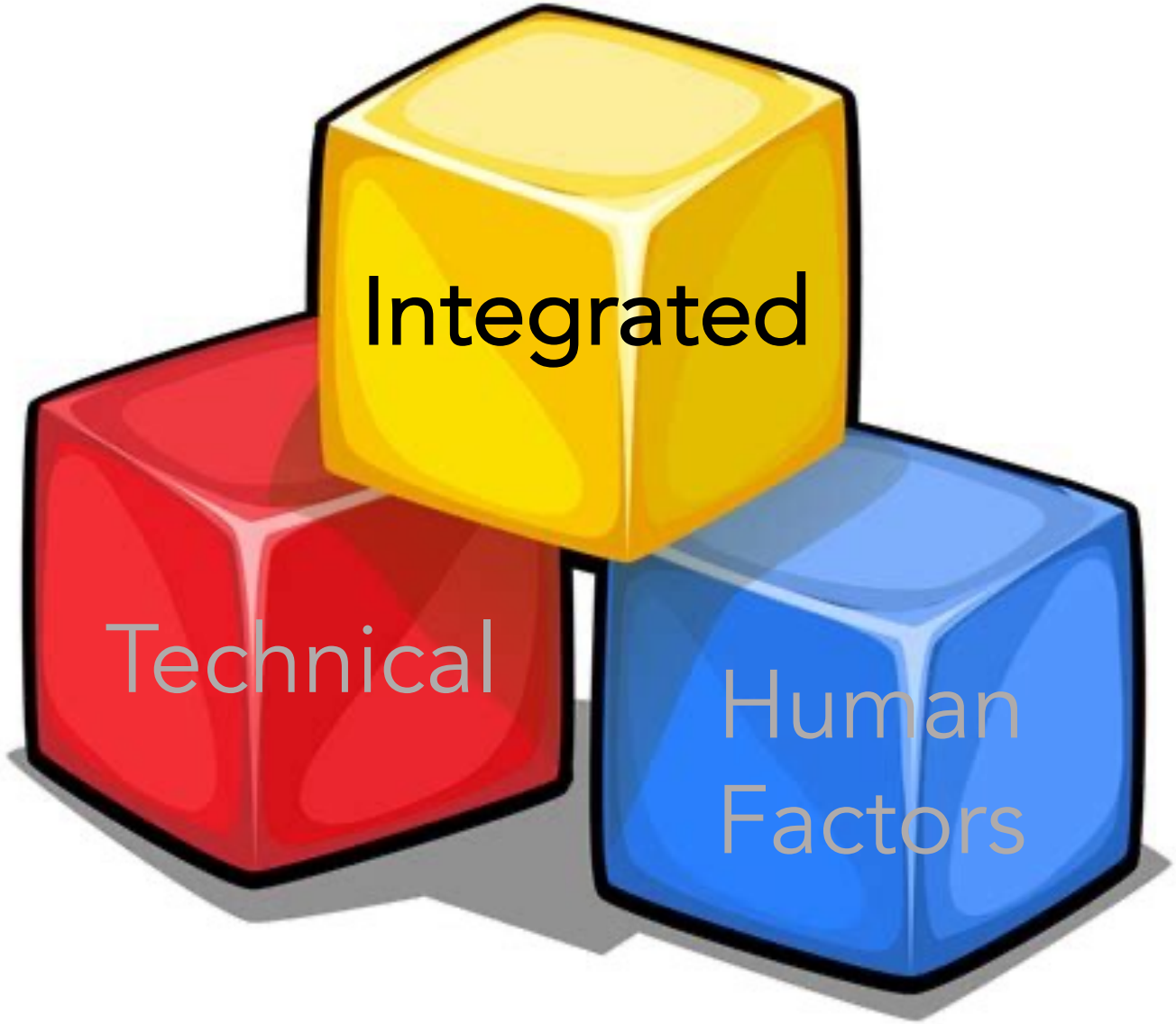
"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 3
16%

Considers nuances of human needs.

"My answer depends on Jessie's priorities. If they want to travel quickly, then the shortest path is the best. But, if they want to travel safely, then this isn't the best option since it is far from medical care."

Alignment between the proposed algorithm and the identified human concerns.



Does a student's technical solution accomplish their human goal?

Level 1A
15%

Well-defined technical solution does not match human needs.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost* = *Distance*^{2}."

Level 1B
56%

Considers human factors, but no technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
29%

Well-defined technical solution aligned with human needs.

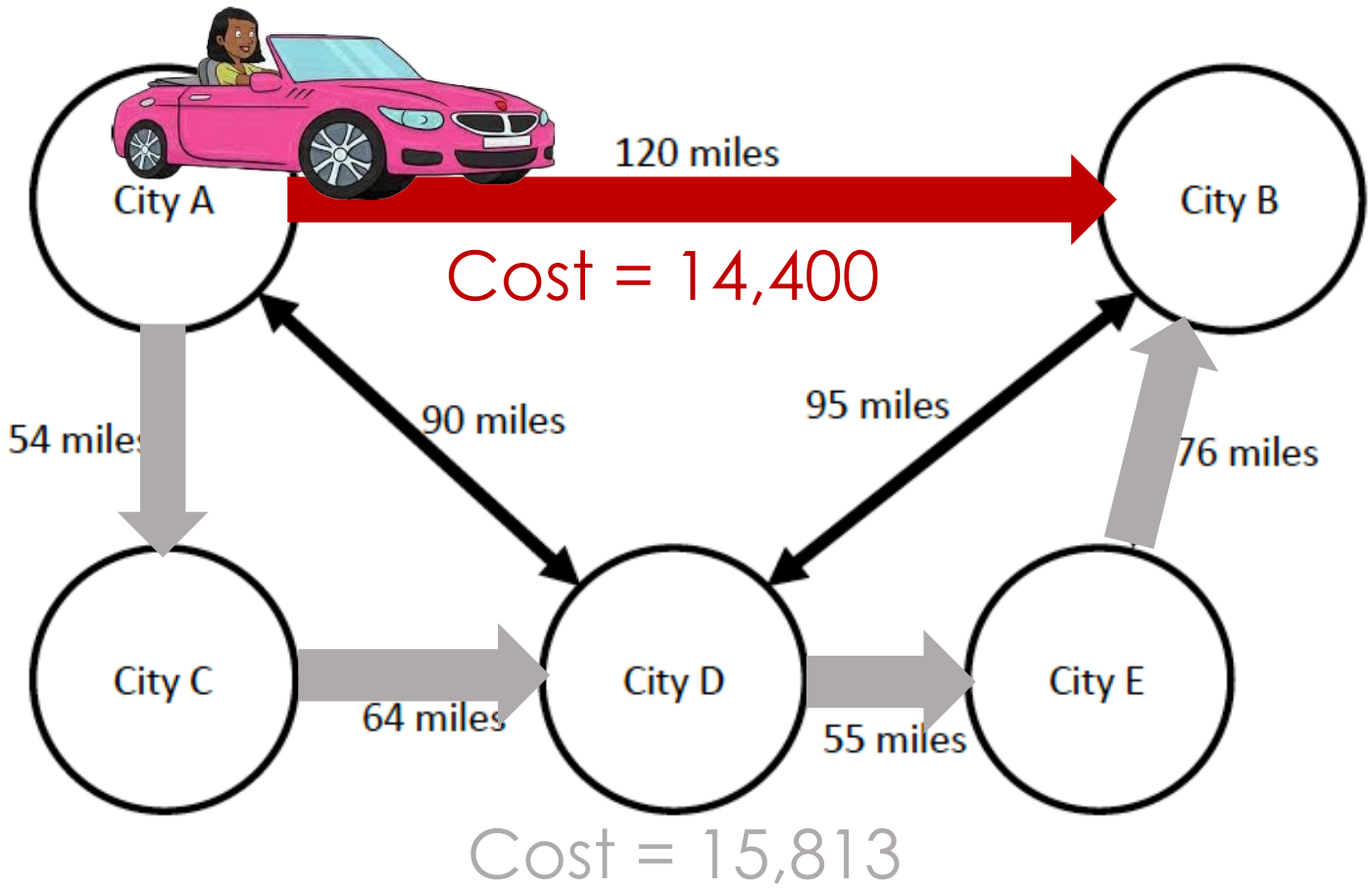
"Sum the distances raised to the power of 3 to keep Jessie on short roads."

Student response

"To keep Jessie off the long road, exponentially punish long edges:

Edge Cost = Distance^2."

Integrated Level 1A



Does a student's technical solution accomplish their human goal?

Level 1A
15%

Well-defined technical solution does not match human needs.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost = Distance*^{2}."

Level 1B
56%

Considers human factors, but no technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
29%

Well-defined technical solution aligned with human needs.

"Sum the distances raised to the power of 3 to keep Jessie on short roads."

Does a student's technical solution accomplish their human goal?

Level 1A
15%

Well-defined technical solution does not match human needs.

"To keep Jessie off the long road, exponentially punish long edges: *Edge Cost = Distance*^{2}."

Level 1B
56%

Considers human factors, but no technical solution.

"Use a cost function that keeps Jessie the closest to medical help as possible."

Level 2
29%

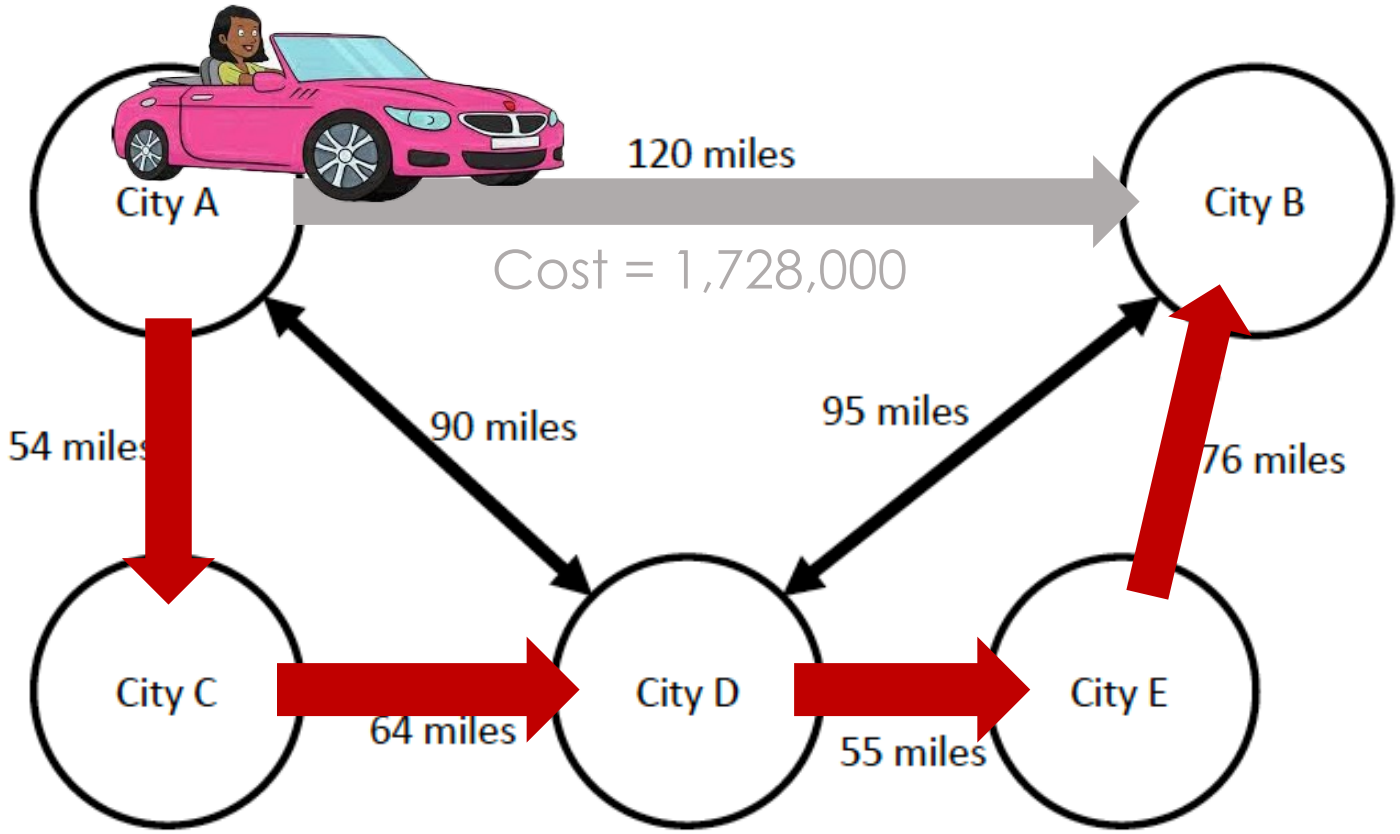
Well-defined technical solution aligned with human needs.

"Sum the distances raised to the power of 3 to keep Jessie on short roads."

Student response

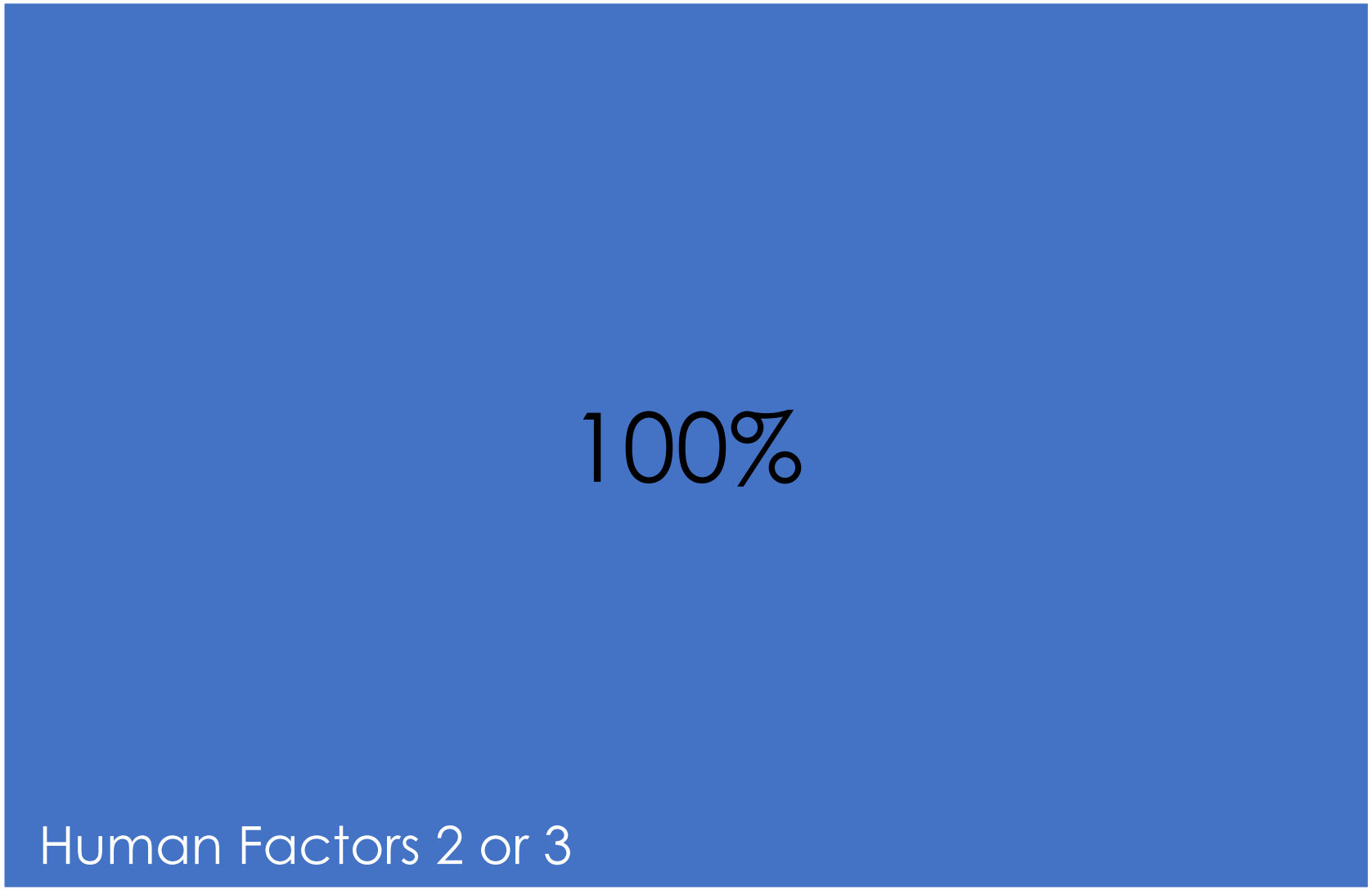
"Sum the distances raised to the power of 3 to keep Jessie on short roads."

Integrated Level 2

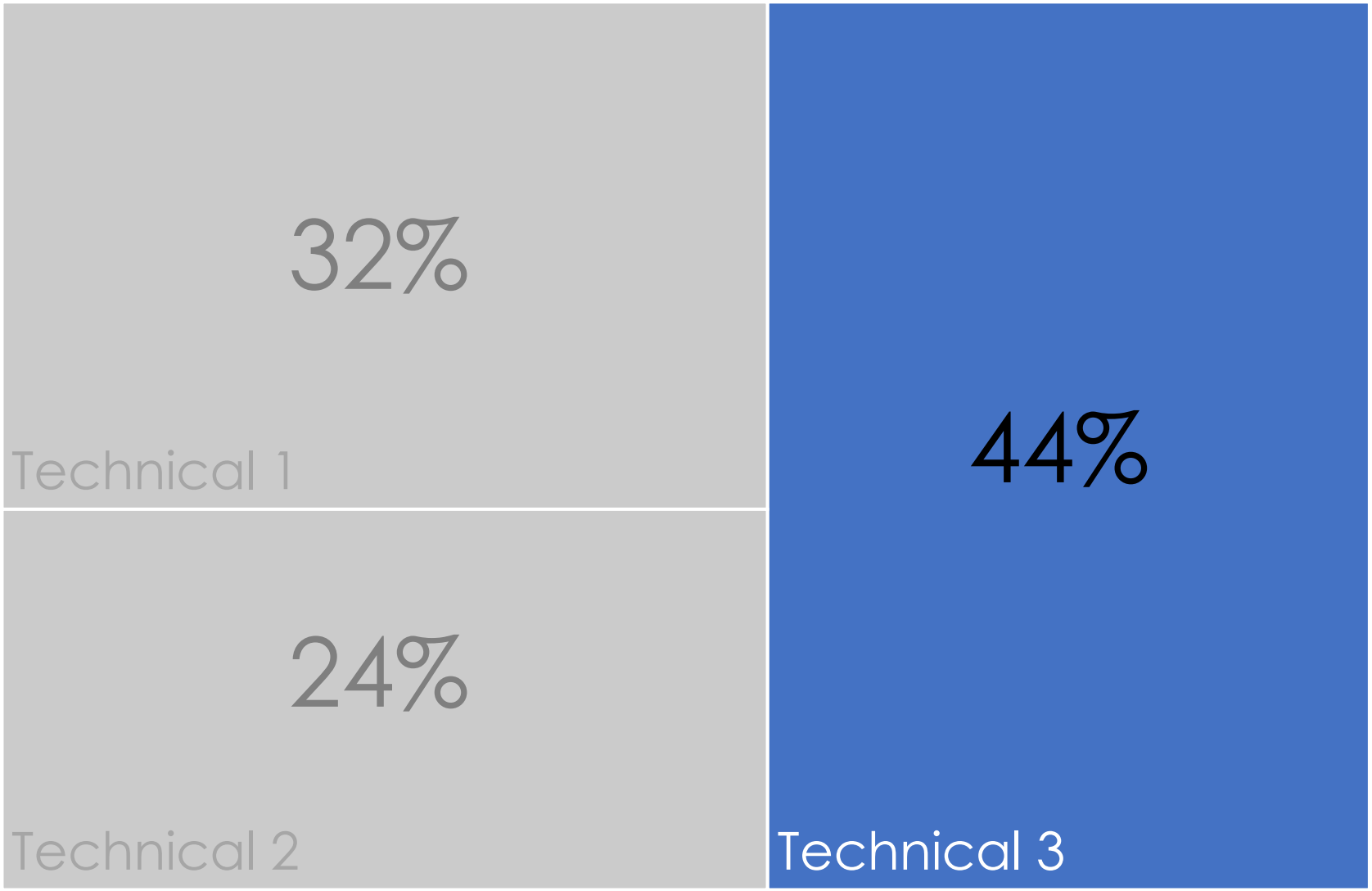


The integrated rubric is more than just the sum of its parts.

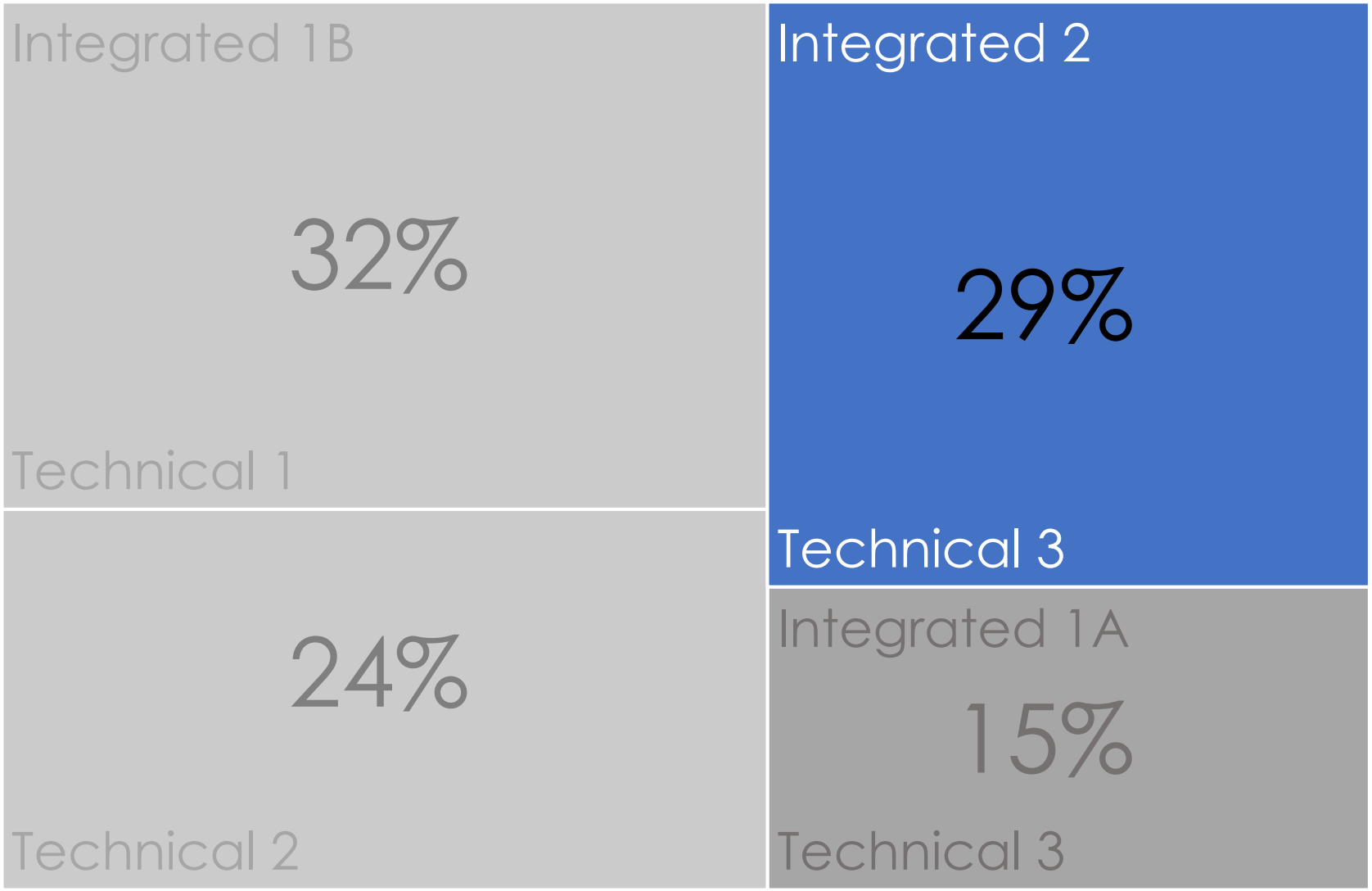
All students were eligible for highest integrated level based on human factors.



44% were eligible for highest integrated level: ≥ 2 on human and 3 on technical.



Only 29% of students scored in the highest level of the integrated rubric.



Overview

1. Course assignment
2. Rubric & scoring
   - Technical
   - Human
   - Integrated
3. **Benefit: illuminates student misconceptions**
4. Challenge: grading

Improved instructor pedagogical content knowledge by revealing misconceptions:

- What makes a cost function well-defined
- What makes a cost function well-suited for graph search
- Cost function ≠ cost of segment
- What are the assumptions of each algorithm

Overview

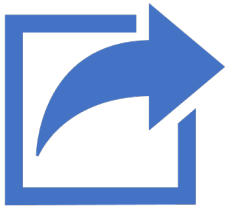
1. Course assignment
2. Rubric & scoring
   - Technical
   - Human
   - Integrated
3. Benefit: illuminates student misconceptions
4. Challenge: grading

It’s hard to reason through a student’s technical (or non-technical) implementation to determine if it is consistent with their defined human goals.

Limitations & Future Work

How can we combine this approach with other ethics instruction?

What are the benefits and challenges of this approach in other CS courses?

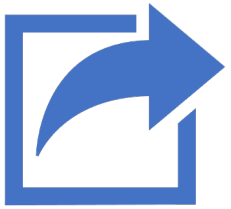


Will this approach foster the attitude that ethical decisions are embedded throughout development?

Limitations & Future Work

How can we combine this approach with other ethics instruction?

What are the benefits and challenges of this approach in other CS courses?

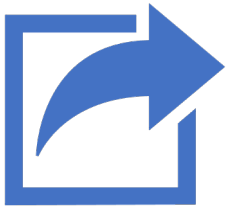


Will this approach foster the attitude that ethical decisions are embedded throughout development?

Limitations & Future Work

How can we combine this approach with other ethics instruction?

What are the benefits and challenges of this approach in other CS courses?



Will this approach foster the attitude that ethical decisions are embedded throughout development?

What makes an assignment integrated?

	Technical	Human	Integrated
1	No solution	No consideration	Missing either human or technical
2	Partial solution	Basic consideration	Human and technical are misaligned
3	Complete solution	Careful consideration	Human and technical are aligned

What makes an assignment integrated?

	Technical	Human	Integrated
1	No solution	No consideration	Missing either human or technical
2	Partial solution	Basic consideration	Human and technical are misaligned
3	Complete solution	Careful consideration	Human and technical are aligned

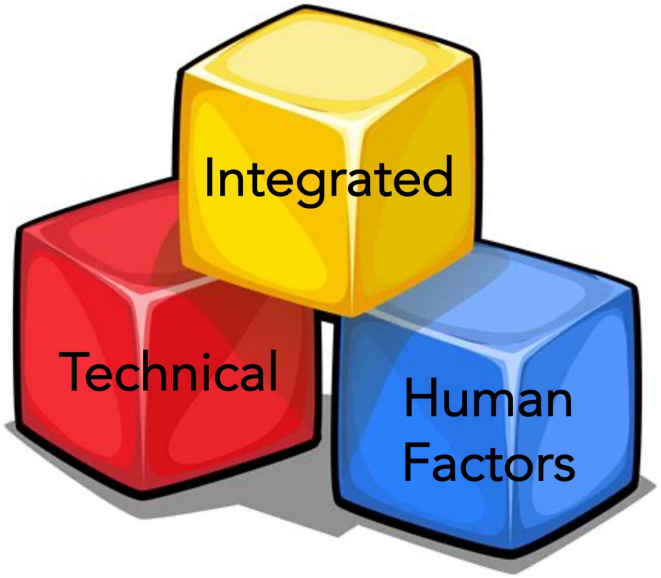
Overall, we showed that ethics can be integrated into technical assignments such that the technical decisions rely on concerns for humans.

The Shortest Path to Ethics in AI:
An Integrated Assignment Where Human Concerns Guide Technical Decisions

Noelle Brown | noelle.brown@utah.edu | noelleb.com

Co-authors: Koriann South & Eliane S. Wiese

Partner instructor: Alan Kuntz



Takeaways:

- We analyzed students' responses on three dimensions of reasoning: technical, human factors, and integrated.
- Successful consideration of technical and human factors individually does not guarantee successful integration.
- This process can improve instructor PCK by revealing technical misconceptions.
- Our rubrics can help with evaluation and design.