AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 1: Introductions: AI, Trust, Ethics

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

19TH AUG 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 1

- Introduction Segment
  - What is this course about? Some examples.
  - Course logistics
- Main Segment
  - AI: A quick introduction
  - Trust (in Technology)
  - Ethics
  - What to expect in the course?
- Concluding Segment
  - About next lecture – Lecture 2
  - Ask me anything

Example 1: Courses for a Student

- Decision: Student deciding which courses to take for their program
- Data
  - **Public**: About courses
  - **Public**: About faculties
  - **Public**: About job opportunities
  - **Public**: About research opportunities and industry trends
  - Private: what the student wants to do
- Analysis
  - Courses offered in different semesters
  - Teachers offering courses – background, hardness of classes, …

Trust
- Are the insights reliable?
- Do they cause short- or long-term harm?
- Will users adopt the insights?

Example 2: Health During a Pandemic

- Decision: Individual staying healthy during a pandemic like COVID19
- Data
  - **Public**: About disease, cases, deaths, variants
  - **Public**: About mitigation steps: e.g., mask wearing restrictions and practices, lockdowns, hospital conditions
  - Private: pre-existing health conditions
- Analysis
  - Regions with high and low cases
  - Whether to eat inside a restaurant?
  - How to make an urgent road trip ?
  - How to hold classes at a University?

Trust
- Are the insights reliable?
- Do they cause short- or long-term harm?
- Will users adopt the insights?

Course Logistics

Administrative Information

- Timings
  - Class: Tu, Th - 2:50 pm – 4:05 pm
  - Office Hours: 1130-1230 Mon, Wed; other times by appointment. Can be in-person or virtual.
- Websites
  - Course: https://blackboard.sc.edu
  - Supplementary: https://sites.google.com/site/biplavsrivastava/teaching/csce-590-trusted-ai
  - GitHub: https://github.com/biplav-s/course-tai (code samples)
- Class methods
  - Primary: In-class
  - Secondary: Synchronous online for recording and slides via Blackboard

Learning Objectives

Undergraduate students will be able to:

L1: **Explain, execute and create AI-based analytical methods to process data**: (a) unstructured data, (b) semi-structured data, (c) structured data

L2: **Explain AI methods in data analysis**: (a) Learning methods, (b) Reasoning, (c) Representation and standardization – knowledge graphs/ ontology, (d) Preferences, (e) Handling Uncertainty

L3: **Identify trust issues in AI methods**: (a) fairness and bias, (b) harmful language, (c) data privacy

L4: **Methods and tools to promote trust**: (a) Data sampling and synthetic data, (b) Testing and rating for communication, (c) Algorithmic innovations like differential privacy and explanations

Graduate students will be able to do all of the above, and:

L5: **Evaluate gaps in Trusted AI tools and create new datasets to handle them**

L6: **Explain emerging standards, frameworks and laws**.

L7: **Explain research findings in open areas and critique their contributions**

Time Allocation

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data) - Classification
- Week 9: Machine Learning (Text data) - Classification – Trust Issues
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

Course Material

- AI Fairness
  - Trisha Mahoney, Kush R. Varshney, and Michael Hind, Available at: https://krvarshney.github.io/pubs/MahoneyVH2020.pdf
  - In AI We Trust: Ethics, Artificial Intelligence, and Reliability, Mark Ryan. Available at: https://link.springer.com/article/10.1007/s11948-020-00228-y
- Python for Data Analysis
  - Latest: Python for Data Analysis Book, by Wes McKinney, 2nd Edition. On Amazon at: https://www.amazon.com/gp/product/1491957662/, ISBN-13: 978-1491957660, ISBN-10: 1491957662
  - Book Data and Code Notebooks: https://github.com/wesm/pydata-book
  - 1st edition (free download): https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Analysis.pdf
- Artificial Intelligence: A Modern Approach (Fourth edition, 2020), Stuart Russell and Peter Norvig,
  - http://aima.cs.berkeley.edu/, ISBN-13: 978-0134610993

Open Datasets

- US: https://www.data.gov/ or any US state
- Text of legislations - LegiScan, https://legiscan.com/
- Kaggle datasets: https://www.kaggle.com/datasets
- Google datasets search: https://datasetsearch.research.google.com/

Undergraduate Student Assessment

Tests	1000 points
• Course Project – report, in-class presentation	600 points
• Quiz – best of 3 from 4	200 points
• Final Exam	200 points
Total	1000 points

- Project: 50% + 10%:
  project report (50%) and code, for elevator presentation to class (10%)
  - Data analysis project
    - Dataset must be from given catalog
    - Use analytical methods to present new insights
- Quiz: 20%
  - 4 based on preceding lectures
- Exam: 20%
- Total 100%

Graduate Student Assessment

Tests	1000 points
• Course Project – report, in-class presentation	500 points
• Quiz – best of 3 from 4	200 points
• Papers: summary, in-class presentation	200 points
• Final Exam	100 points
Total	1000 points

- Project: 40% + 10%:
  project report (40%) and code, for elevator presentation to class (10%)
  - Data analysis project OR
    - Dataset must be from given catalog
    - Use analytical methods to present new insights
  - Create or explore new methods (preferred for graduate students) project
    - Problem to be discussed with instructor
    - Example: Analyze sound signals to estimate crowd
- Quiz: 20%
  - 4 based on preceding lectures
- Paper presentation: 10% + 10%
  - Research paper reading (10%) and presentation to class (10)% - Total 20%
    - Read a paper accepted at a top Data / AI conference: AAAI 2019-2021, IJCAI 2019-2021, NeurIPS 2019-2021, KDD 2019-2021, SIGMOD 2019-2021. Make a 1-page summary highlighting the key points, what you liked and what you did not. Try any code given in the paper
    - Present a 1-slide summary to class (10%)
- Exam: 10%
- Total 100%

Student Assessment

A = [900-1000]
B+ = [870-899]
B = [800-869]
C+ = [770-799]
C = [700-769]
D+ = [670-699]
D = [600-669]
F = [0-599]

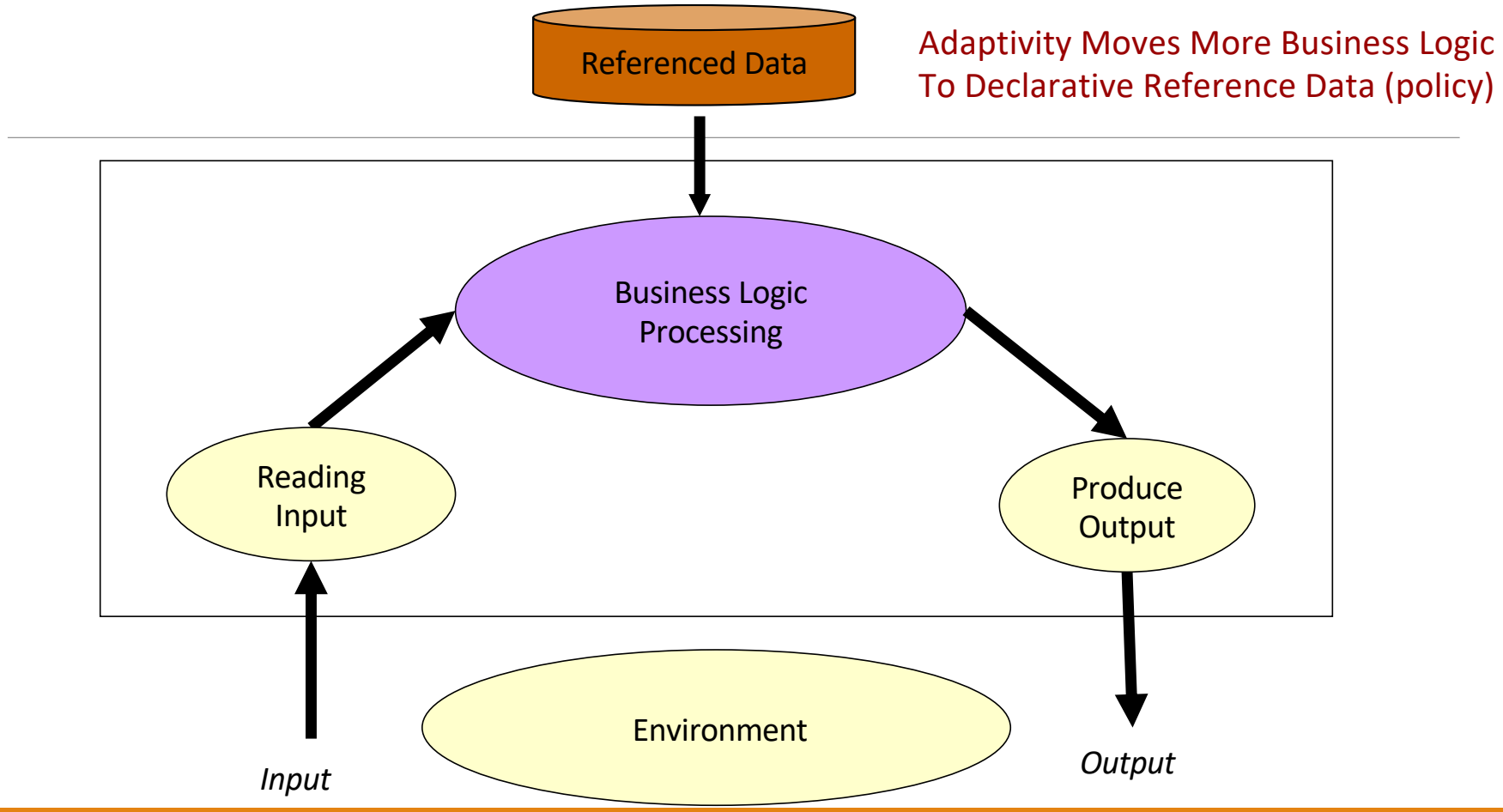
Tests	1000 points
• Course Project	600 points
• Quiz – best of 3 from 4	200 points
• Final Exam	200 points
Total	1000 points

Tests	1000 points
• Course Project	500 points
• Quiz – best of 3 from 4	200 points
• Papers	200 points
• Final Exam	100 points
Total	1000 points

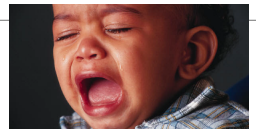
Main Segment

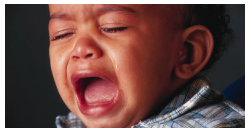
AI: A Quick Introduction

(Adaptive) Software System



Example: Taking Care of a Baby
Individual's Extension





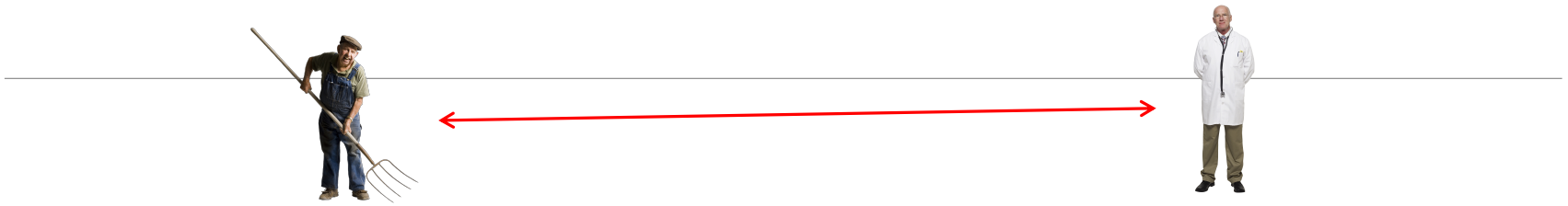
Agent

Expected behavior:

- Inform
  - Alert when crying
  - Alert when awake
  - Alert when idle
- Do
  - Raise temperature of room
  - Play music
  - …

Can be specified and reasoned (**reasoning/ rule-based methods**) OR learned **(learnt from data)**

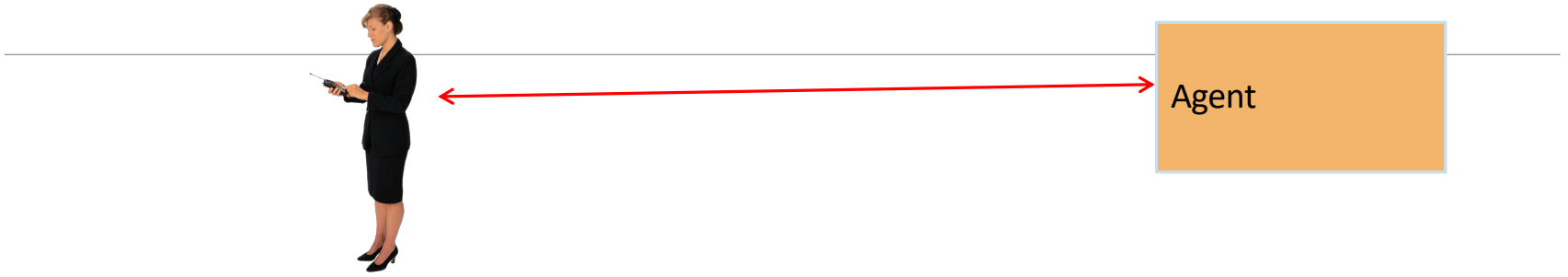
Example: Taking Care of a Senior
Assisted Cognition



Expected behavior:

- Inform
  - Alert when idle
  - Alert when away from known locations
  - Alert when checkup/ medicines due
- Do
  - Send body parameters periodically
  - …

Example: Taking Care of Oneself
Personal Digital Assistants



Expected behavior:

- Inform
  - When missing meetings
  - When missing social commitments
  - Reminding of priorities
  - …
- Do
  - Make all cancellations / re-bookings when schedule changes
  - Find alternatives to current decisions and give choices (e.g., traffic)
  - …

What Are Intelligent Agents?

Agents are active, persistent software components that perceive, reason, act, and communicate. (Huhns and Singh)

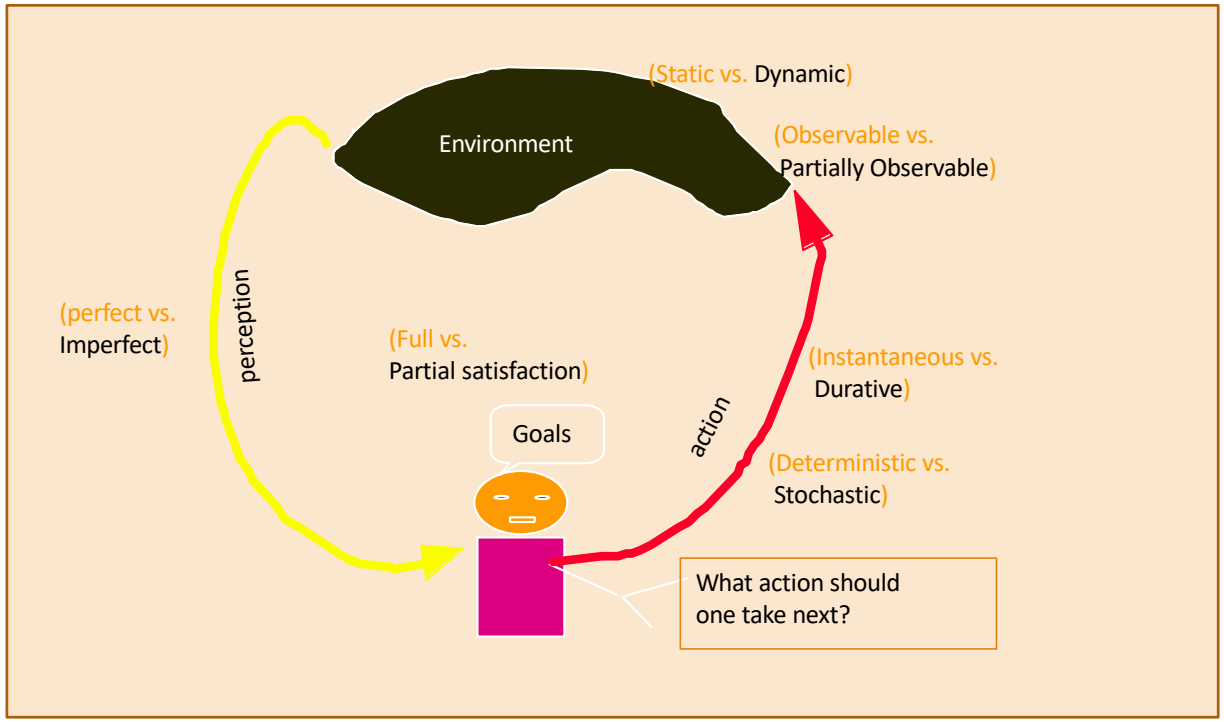
- Software that **assists** people and acts on their behalf
- Agents can help ***people*** and ***processes***
- Agents are used for automation and control

- finding and filtering information
- personalizing your environment
- negotiating for services
- automating tedious tasks
- taking actions you delegate
- learning about you over time
- collaborating with other agents
- capturing individual and organizational knowledge
- sharing knowledge

- finding and fixing problems
- automating complex procedures
- finding "best fit" procedures
- pattern recognition and classification
- predictions and recommendations
- negotiate and cooperate with other organizations' agents

Credits:
Joe Bigus, Intelligent Agents

Artificial Intelligence (AI) as an Agent

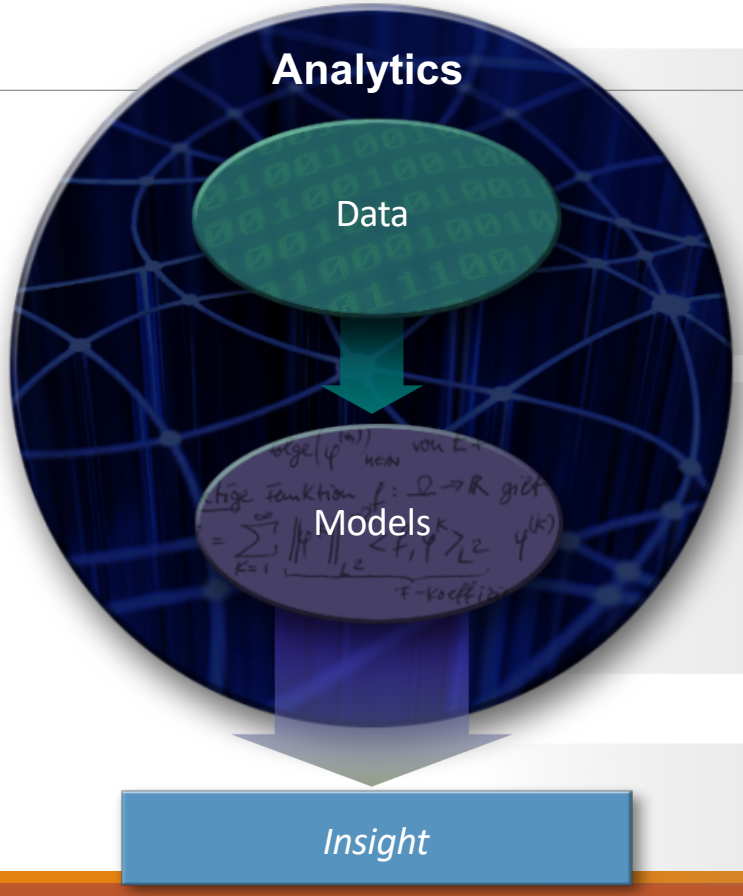


AI deals with perceiving the environment and taking actions towards short- and long term goals as the world changes over time.

Credits:
From Subbarao Kambhampati's AI Planning Course
Russell & Norvig, AI – A Modern Approach

Advanced AI Techniques (Analytics) like Reasoning & Machine Learning

make use of data and models to provide insight to guide decisions

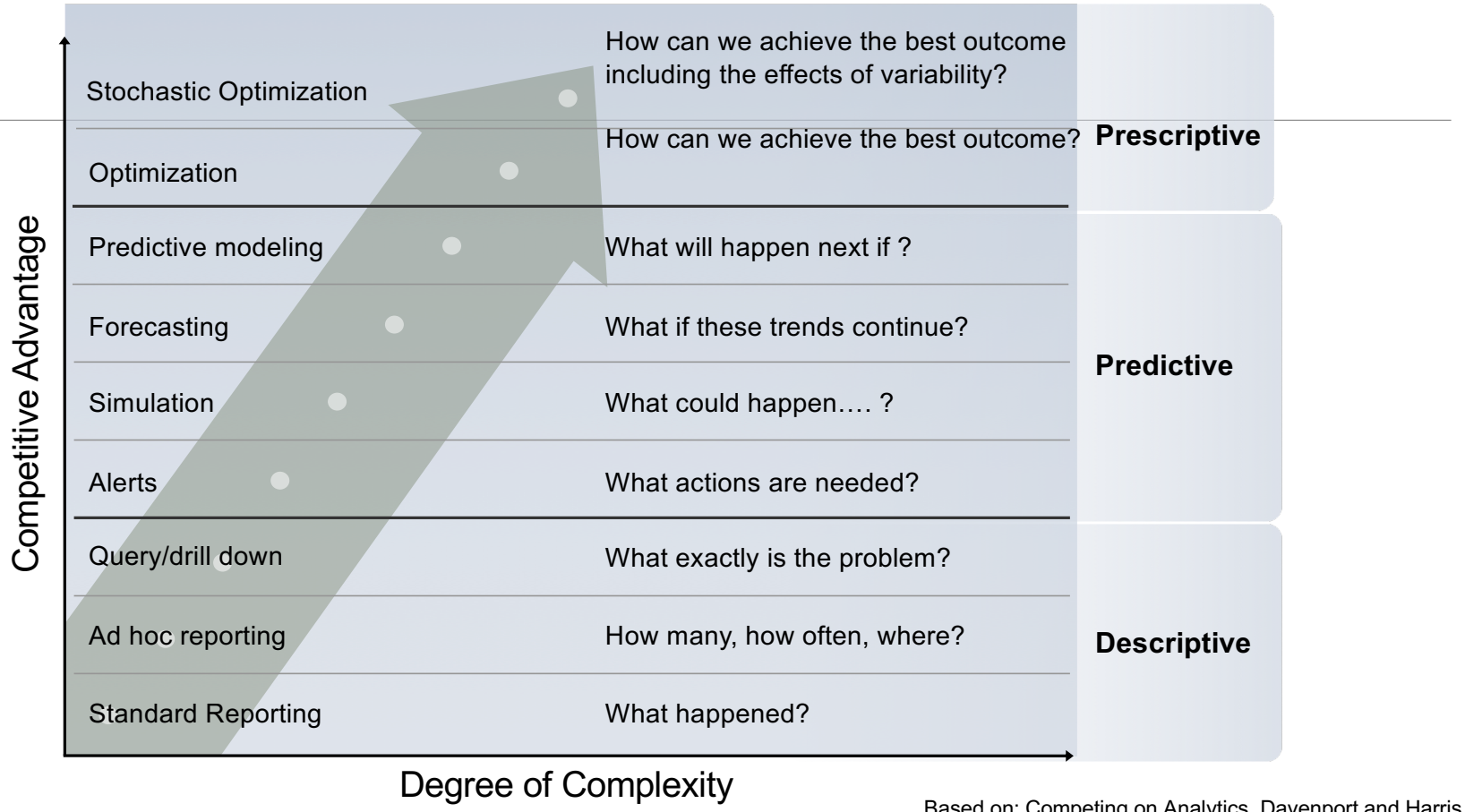


Data sources:
Business automation
Instrumentation
Sensors
Web 2.0
Expert knowledge
"real world physics"

Model:
a mathematical or algorithmic representation of reality intended to explain or predict some aspect of it

Decision executed automatically or by people

Analytics Landscape



Based on: Competing on Analytics, Davenport and Harris, 2007

Example: Courses for a Student

- Decision: Student deciding which courses to take for their program
- Data
  - **Public**: About courses
  - **Public**: About faculties
  - **Public**: About job opportunities
  - **Public**: About research opportunities and industry trends
  - Private: what the student wants to do
- Analysis
  - **Descriptive**: Courses offered in different semesters; Teachers offering courses
  - **Predictive**: How fill will be a particular class next semester?
  - **Prescriptive**: Should a student a particular course?

Example: Health During a Pandemic

- Decision: Individual staying healthy during a pandemic like COVID19
- Data
  - **Public**: About disease, cases, deaths, variants
  - **Public**: About mitigation steps: e.g., mask wearing restrictions and practices, lockdowns, hospital conditions
  - Private: pre-existing health conditions
- Analysis
  - **Descriptive**: Regions with high and low cases
  - **Predictive**: Does wearing mask help reduce cases?
  - **Prescriptive**:
    - Whether to eat inside a restaurant?
    - How to make an urgent road trip ?
    - How to hold classes at a University?

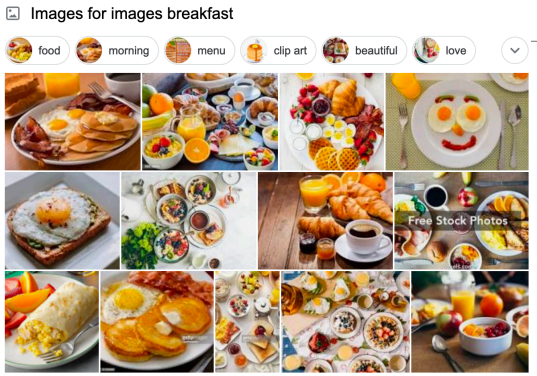
Resources: https://github.com/biplav-s/covid19-info/wiki/Important-Information-About-COVID19

Sub-areas of AI

- **Representation**: formal representation of knowledge.
  - Illustration: entities and their relationships, like last Russian Czar's family tree
  - Methods: Ontology, knowledge graph, word embedding, "Model"
- **Reasoning**: deriving conclusions from formally represented knowledge.
  - Illustration: Modus ponen – P implies Q. P is True. Hence Q must be true.
  - Methods: Deduction, Induction, Abduction, Proposition logic, First-order logic, Fuzzy logic
- **Learning**: drawing insights from data
  - Illustration: predict COVID cases in USA by end of the month
  - Methods: Machine Learning – Classification, Clustering, Association; Deep Neural Network
- *Additionally*, ***human interaction considerations***:
  - Collaborative assistants
  - Explanations

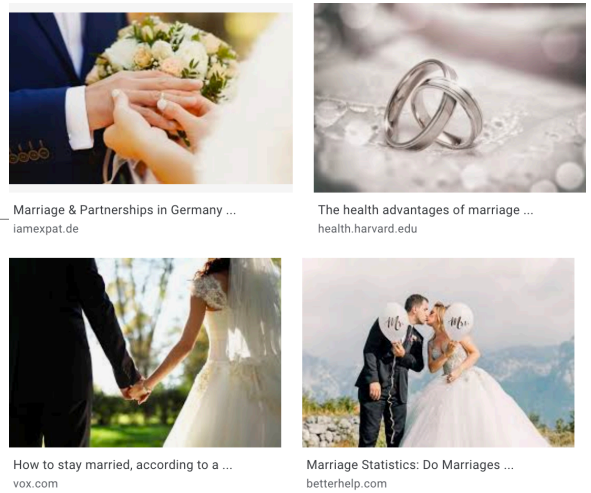
Trust – In Technology

AI Results – Are They Representative ?



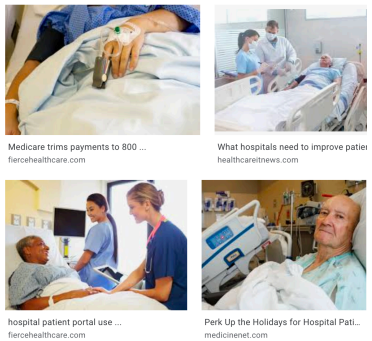
Breakfast items searched on Google

Employees



Marriage

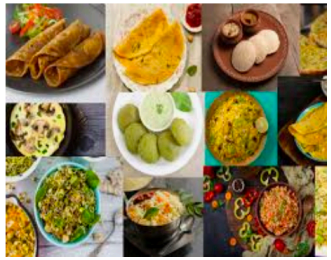
Waiting Customers



Hospital Patients

Many perspectives missing by default, and the system does not even inform about it

Example: Regional Perspectives Matters for Trust



Breakfast items searched on Google

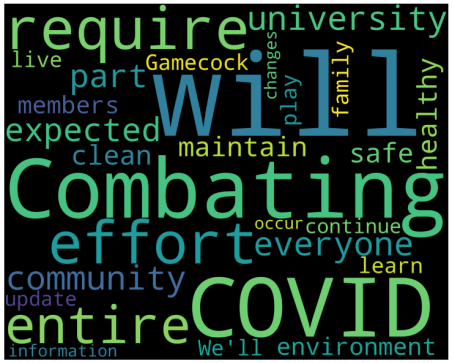
Waiting Customers

How does this impact ?

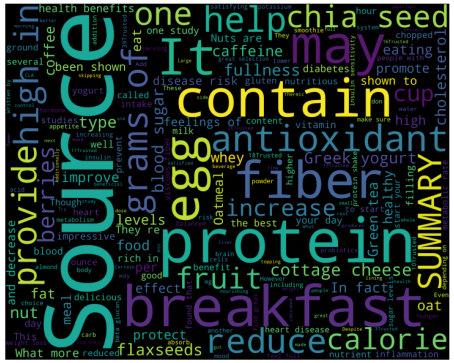
- Training data
- Preferences and constraints
- Inductive bias (i.e., implicit defaults in the algorithms)
- Integration of solution to wider ecosystem

Code Time: Insights from Text

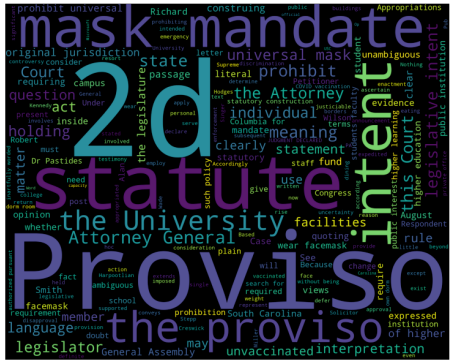
https://github.com/biplav-s/course-tai/blob/main/sample-code/l1-intro/Introduction%20Text%20Issues.ipynb



Covid from UoSC



12 Healthy breakfast dishes



Mask ruling by SC Supreme Court

Components of Trust

1. Competent – including, well tested
2. Explainable – how does the system give its result
3. Holds human values
    1. Fairly and ethically used
    2. Adequate data management & preserves privacy
4. Allows human-technology interaction

Components of Trust - Examples

1. Competent – including, well tested
2. Explainable – how did it give the result
3. Holds human values
    1. Fairly and ethically used
    2. Adequate data management & preserves privacy
4. Allows Human-technology interaction

	Car – cruise control	Doctor	Nuclear Energy
Competent	X	X	X
Explainable	X	?	X
Holds human values	?	X	?
Allows human interaction	X	X	

Components of Trust for AI

1. Competent – including, well tested
2. Explainable – how did it give the result
3. Holds human values
    1. Fairly and ethically used
    2. Adequate data management & preserves privacy
4. Allows Human-technology interaction

	AI – Self-driving Car	AI-powered Chatbot: Medical Guide	AI – Image Captioner
Competent	?	X	X
Explainable	?	?	?
Holds human values: Fair and ethical	?	?	?
Holds human values: data and privacy	?	?	X
Allows human interaction	X	X	X

AI Ethics

Why is Ethics Even an Issue?

- When a technology works with humans and relates to inter-personal issues, the question of ethics comes into picture
- Examples: medicine (opiods), food (genetically modified)

AI Ethics

Multidisciplinary field of study

How to optimize AI's beneficial impact while reducing risks and adverse outcomes

How to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios

To identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

What is Specific to AI?

- AI needs data
  - Data privacy and governance
- AI is often a black box
  - Explainability and transparency
- AI can make decisions/recommendations
  - Fairness and value alignment
- AI is based on statistics and has always a small percentage of error
  - Who is accountable if mistakes happen?
- AI can infer our preferences and manipulate them
  - Human and moral agency
- AI is very pervasive and dynamic
  - Larger negative impacts for tech misuse
  - Fast transformation of jobs and society

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

Main AI Ethics Issues

DATA GOVERNANCE AND PRIVACY

FAIRNESS AND INCLUSION

HUMAN AND MORAL AGENCY

VALUE ALIGNMENT

ACCOUNTABILITY

TRANSPARENCY AND EXPLAINABILITY

TECHNOLOGY MISUSE

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

Lecture 1: Concluding Comments

- We did a quick overview of the course
- Looked at AI, Trust and Ethics
- Course will focus on
  - Practical methods to derive insights from open data
  - Awareness of trust issues and how to address them
  - Student evaluation will be by via project, quizzes, paper presentation (graduate) and final exam
  - **Bring your ideas to your project**
- Exciting techniques to learn to impact the world around us

Concluding Segment

About Next Lecture – Lecture 2

Lecture 2: Trusted Decisions

- The quality of our decisions
- Trusted decisions
- Importance of data