



CSCE 771: Computer Processing of Natural Language

Lecture 1: Introduction, AI, NLP

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

20TH AUG 2024

Carolinian Creed: “I will practice personal and academic integrity.”

Organization of Lecture 1

- Introduction Section
 - Instructor introduction
 - Course logistics
- Main Section
 - AI: A quick introduction
 - Natural languages
 - Natural Language Processing (NLP) – our main focus
- Concluding Section
 - About next lecture – Lecture 2
 - Ask me anything

Introduction Section



BIPLAV SRIVASTAVA

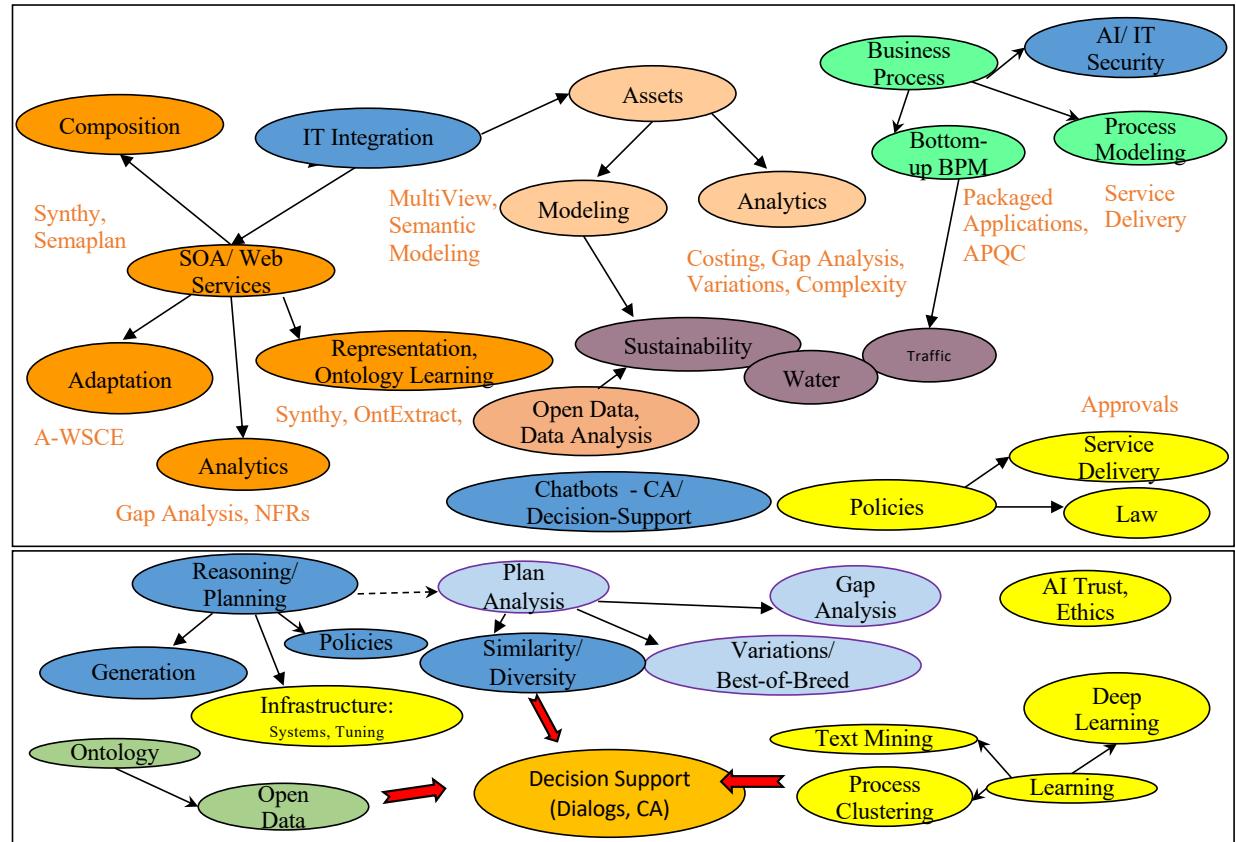
Research Snapshot (1989-2024)

Keywords: AI, Services, Sustainability

Current Research

Focus: **Theory** (Neuro-symbolic), **Usability** (Trust Rating, RCTs), **Smart Cities** (Energy, Water, Health)

The Space of AI Applications Explored



Details: <https://sites.google.com/site/biplavsrivastava/>

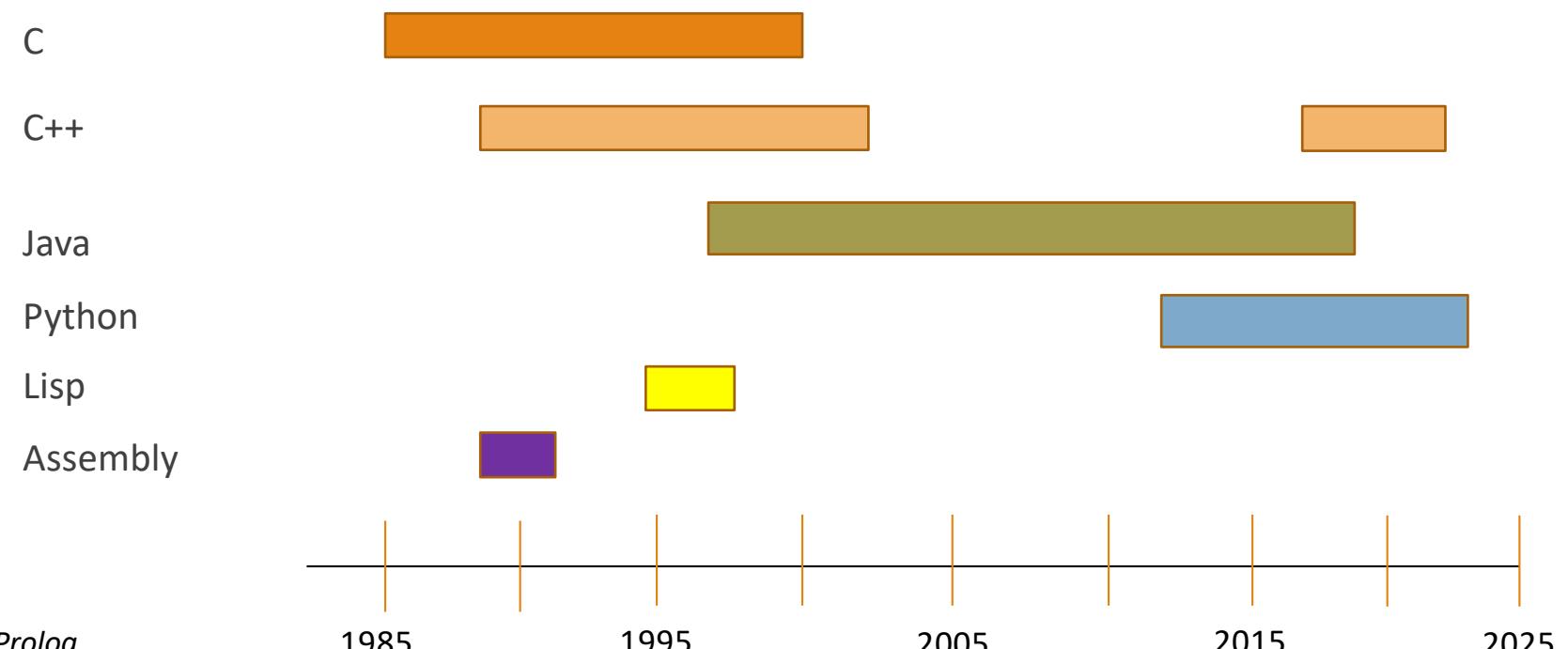
AI4Society: <https://ai4society.github.io/projects/>

Keywords: AI, Services, Sustainability

Papers: 200+ refereed; 6,400+ references

Patents: 75 (US issued); 4 sole inventions

Personal Programming Language Journey* (35+ years)



Develop a Vibrant Research Culture Around AI

Students building chatbots
in Adv. Prog. Tech. class
in C++, Java and Python
(Elected Reps, Spring 2022;
Diseases, Spring 2024)

Disease Chatbot

You: hi
Bot: Hi
You: what is measles?
Bot: Measles is an infection you get from a virus. The measles virus lives in the mucus of the nose and throat. It's spread through the air and by coming into direct contact with someone who has it. The virus can stay active on surfaces and in the air for up to 2 hours.
It's very contagious. If you haven't been vaccinated and are in a room with someone who has measles, you have a 90% chance of getting it.
Part of what makes measles so dangerous is that you can be contagious 4 days before you get the telltale rash. So you could easily spread the virus without knowing you have it. You'll continue to be contagious 4 days after the rash goes away.
You: how is it diagnosed?
Bot: To figure out whether you have measles, your doctor will first look closely at your rash. They'll also look for the small white Koplik spots inside your mouth. Sometimes, to confirm it, they can also do a blood test.



WebMD

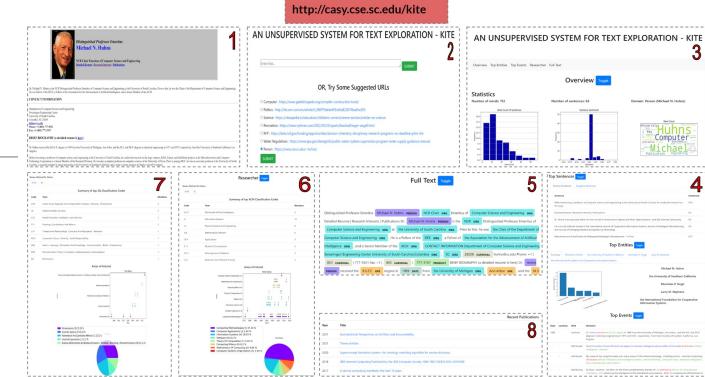
System Image Credit:
Christine Steege, CSCE240(H), Spring 2024

Thoroughness
in work

Clarity in
communication

Pro-active focus
to complete an
effort

Comfortable
working with
others: social
yet focused



KITE: An Unsupervised, Effective and Inclusive Approach for Textual Content Exploration

AI/ Chatbots built for: governance (IJCAI 2016), AI Magazine 2024), astronomy (AAAI 2018 best demo award), water (AAAI 2018), smart room (ICAPS 2018 demo runner up, IJCAI 2018), career planning (commercial product), market intelligence (AAAI 2020 deployed AI award), dialogs for information retrieval (ICAPS 2021), fairness assessment (AAAI 2021), computer games (AAAI 2022), generalized planning (IJCAI 2024), information spread in opinion networks (AAAI 2024 best demo award), transportation, set recommendation (teaming (AAAI 2024 deployed AI award), meals) and health. <https://ai4society.github.io/demos/>



Course Logistics

Administrative Information – CSCE 771

- Introduction to AI - [CSCE 771](#)
 - CRN: [30074](#)
 - Duration: 08/20/2024 - 12/16/2024
 - Class Timings: 300 Main St. | Room B102
- Websites
 - Course: <https://sites.google.com/site/biplavsrivastava/teaching/ai-csce-580-fall-2024-intro-to-ai>
- Class methods
 - In-class
 - Asynchronous / Recording Online: Zoom - Blackboard

Administrative Information

- Instructor: Biplav Srivastava, Ph.D.
 - email: biplav.s@sc.edu
 - office: AI Institute, Room 515, 1112 Greene St., Columbia, 29028
- Office hours:
 - Wednesday, 3-4pm, Zoom/ Blackboard
 - Thursday, 10-11am
 - By Appointment in-person
- TA: Vipula Rawate
 - email: vrawte@mailbox.sc.edu
 - office: AI Institute, Room 515, 1112 Greene St., Columbia, 29028
- Office hours:
 - Monday, 9 - 10am
 - Friday, 3 – 4pm
 - By Appointment in-person

Engagement methods:

- Discussion section of Blackboard - fastest
- Email – slow response
- Meetings – slowest turnaround

Learning Objectives

L1: Appreciate diversity and similarity in natural languages – text, speech and visual; focus of course will, however, be text (NLP) and English

L2: Understand issues related to data and tools. Experiment design, Metrics for evaluation and to detect bias, Methods to build trust in processing – transparent assessment, Providing explanations for output

L3: Data processing: (a) Structured data representation from unstructured text; (b) Extract entities and relationships; (c) Extract contexts; (d) representation learning – word embedding

L4: AI methods in NLP: (a) Learning methods – including language models, (b) Reasoning, (c) Representation – knowledge graphs/ ontology

L5: NLP applications – (a) Document intelligence: sentiment, translation; (b) collaborative assistants

Course Material

- The required textbook for this course is: Speech and Language Processing Dan Jurafsky and James H. Martin, 2nd edition in print; Draft of 3rd edition available online at: <https://web.stanford.edu/~jurafsky/slp3/>
- The optional reference book, specially suggested for students without CSCE 580, is: Artificial Intelligence: A Modern Approach (Fourth edition, 2020), Stuart Russell and Peter Norvig <http://aima.cs.berkeley.edu/> ISBN-13: 978-0134610993
- Research Papers
 - PDFs of published papers
- Open Datasets - Illustration
 - Data from Fall 2020 instance of CSCE 771 - <https://github.com/biplav-s/course-nl/tree/master/common-data>
 - Text of legislations - LegiScan, <https://legiscan.com/>
 - COVID-19 research papers - <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/>; <https://github.com/biplav-s/covid19-info/wiki/Important-Information-About-COVID19>
 - Text of patents, Google patents - <https://patents.google.com/>

Student Assessment

A = [920-1000]

B+ = [870-919]

B = [820-869]

C+ = [770-819]

C = [720-769]

D+ = [670-719]

D = [600-669]

F = [0-599]

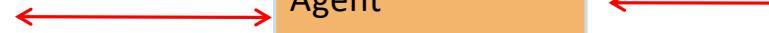
Tests	1000 points
• Course Project – report, in-class presentation	600 points
• Quiz – best of 3 from 4	210 points
• Final Exam – Paper summary, in-class presentation	190 points
Total	1000 points

Main Section

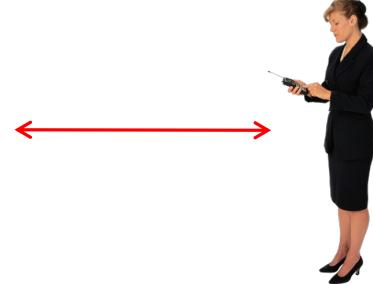
AI: A Quick Introduction

Concept: AI

Example: Taking Care of a Baby Individual's Extension



Agent



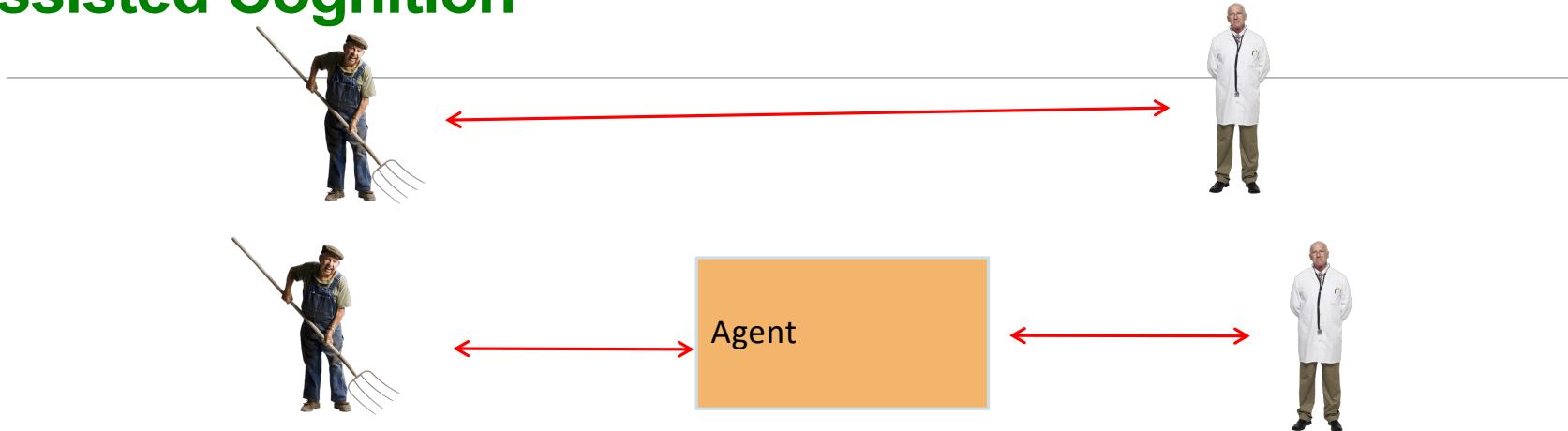
Expected behavior:

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

Conditions can be

- input and **reasoned** (e.g. rule-based methods) OR
- **learned** (from data)

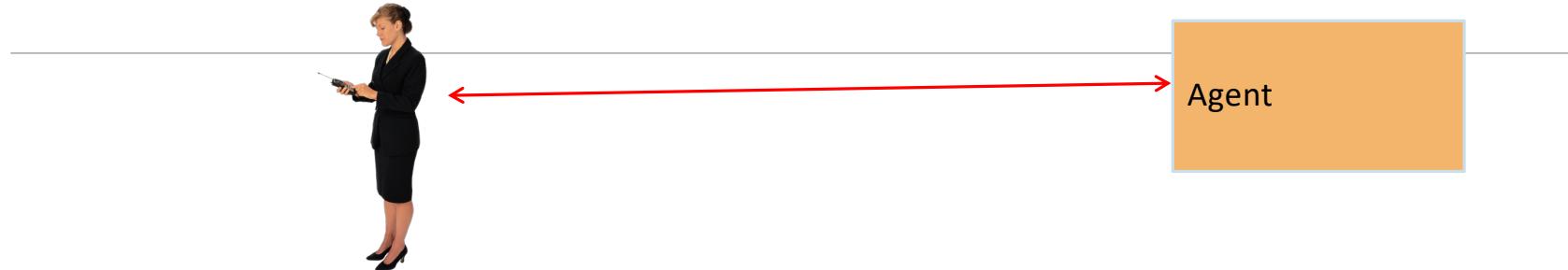
Example: Taking Care of a Senior **Assisted Cognition**



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**

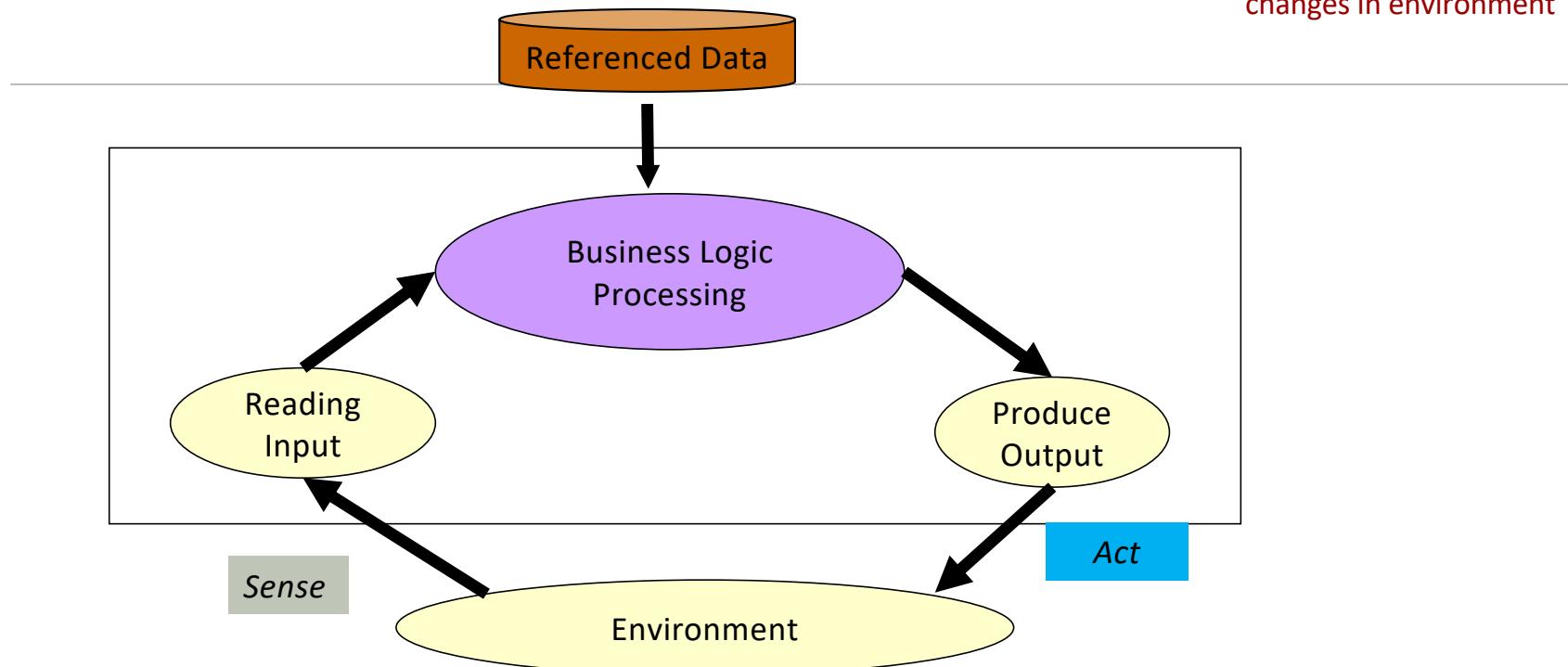


Expected behavior:

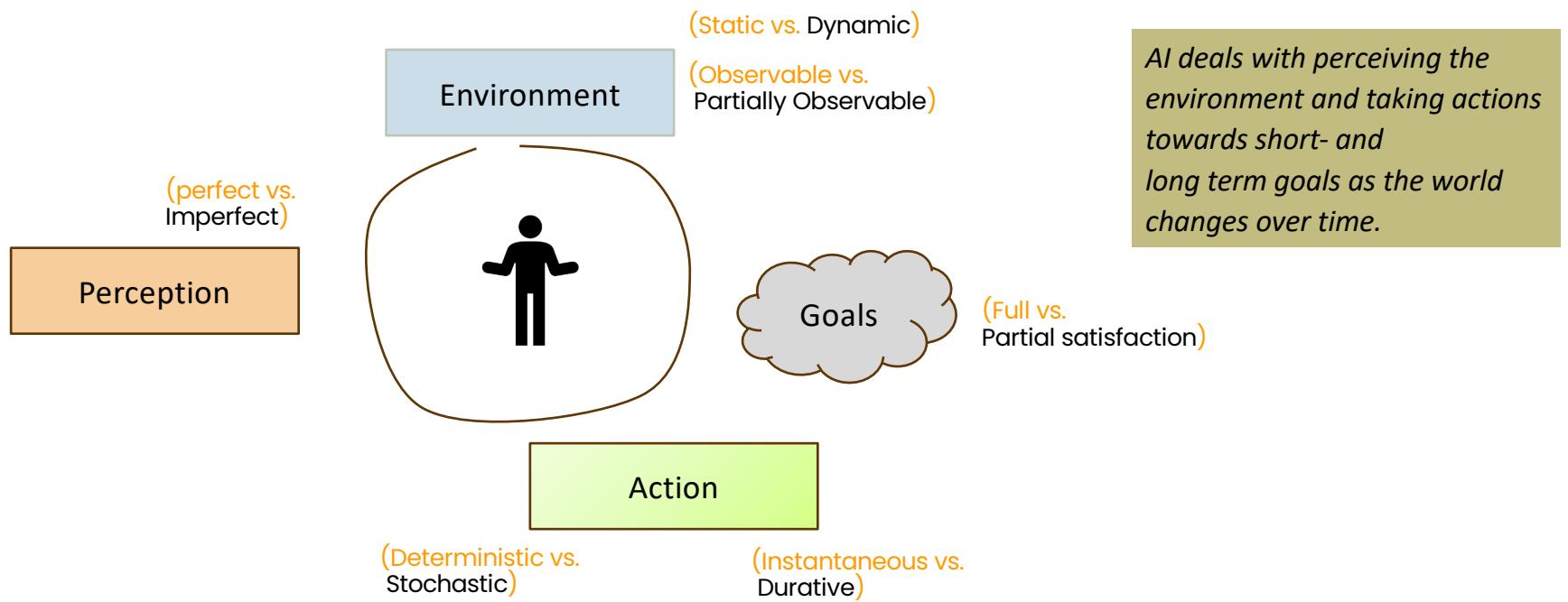
- Inform
 - When missing meetings
 - When missing social commitments
 - Reminding of priorities
 - ...
- Do
 - Make all cancellations / re-bookings when schedule changes
 - Find alternatives to current decisions and give choices (e.g., traffic)
 - ...

AI => Adaptive/ Intelligent Software System

- Business Logic Moves to Declarative Data (policy)
- Software is more resilient to changes in environment



Artificial Intelligence (AI) as an Agent



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?

Thought Exercise – (AI) Class and a Hypothetical AI-based Advisor

- **Good** decisions for students

- Get good grades, marks
- Learn
- ...

- AI-may suggest

- Give teacher rating
- *But what about learning?*

- **Good** decisions for instructor

- Get good rating
- Finish course
- Teach long-term skills
- ...

- AI-may suggest

- Give student grades
- *But what about teaching?*

Trust

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

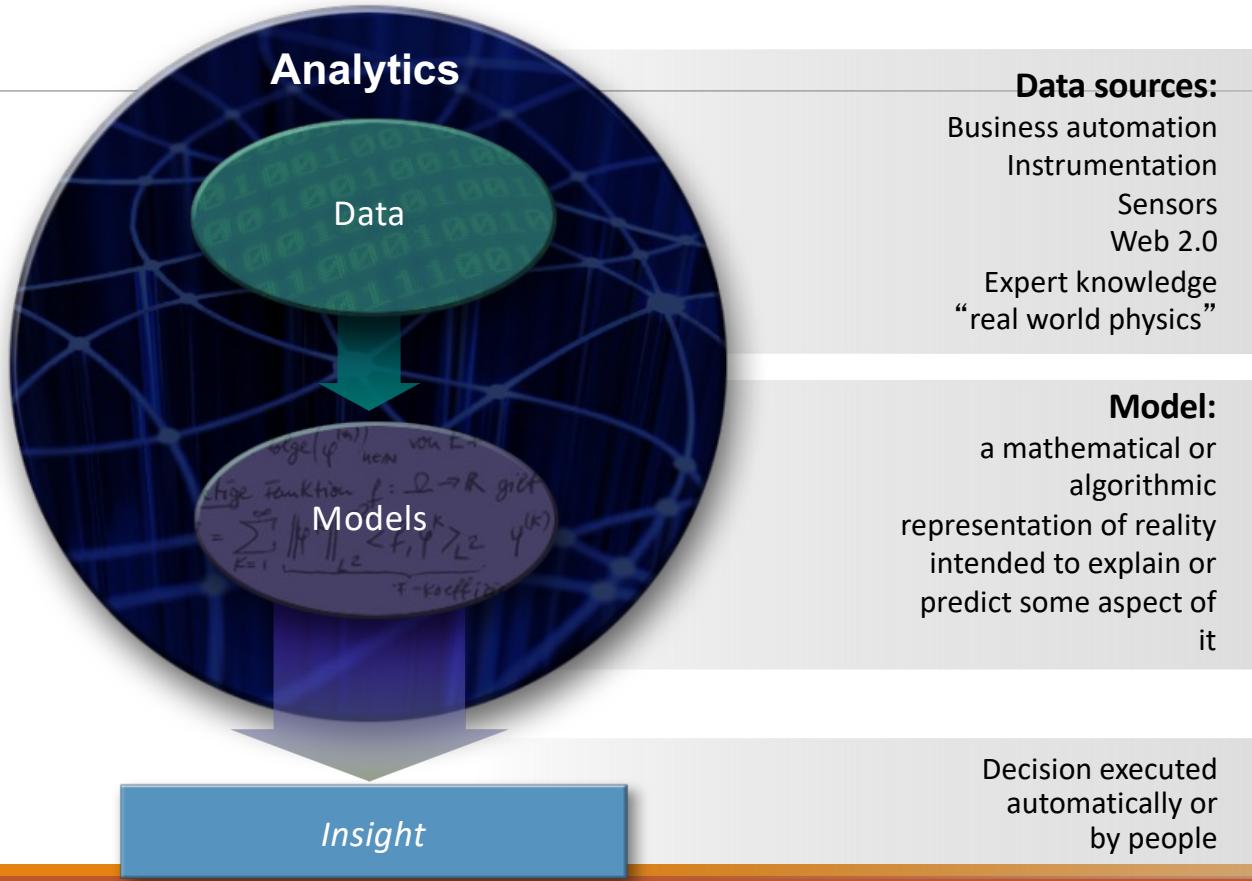
Example 2: Health During a Pandemic

- 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?

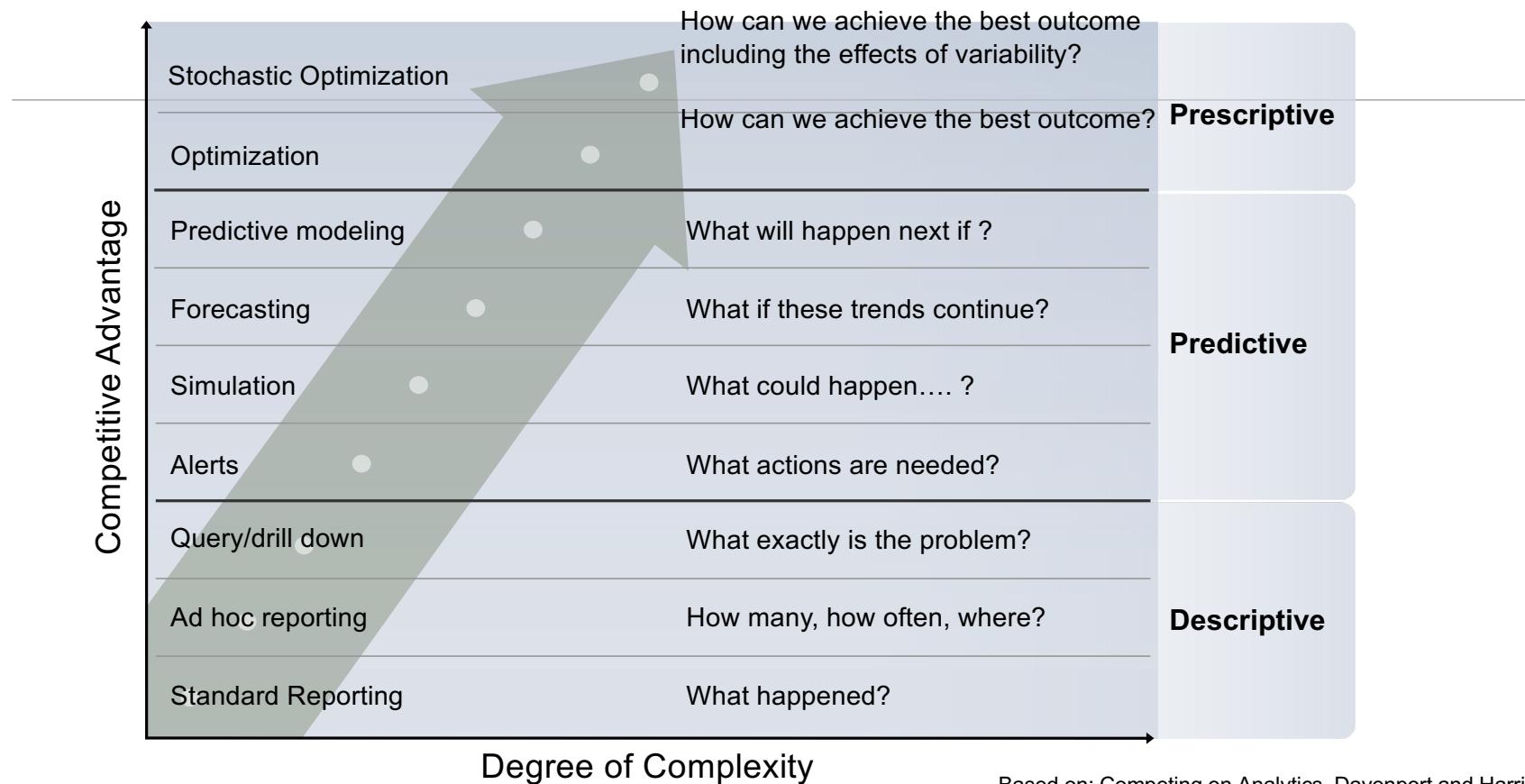
Trust

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

Advanced AI Techniques (**Analytics**) like Reasoning (**Symbolic**) & Machine Learning (**Neural**)
make use of data and models to provide insight to guide decisions



Analytics Landscape



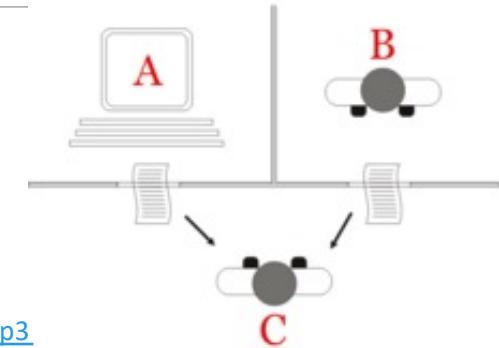
Based on: Competing on Analytics, Davenport and Harris, 2007

History of Chatbots is the History of AI

Credit: https://en.wikipedia.org/wiki/Turing_test

1950 - Turing test

"which player – A or B – is a computer and which is a human."



1964-66 – Eliza

computerized Rogerian psychotherapist

<https://en.wikipedia.org/wiki/ELIZA>, <http://www.manifestation.com/neurotoys/eliza.php3>

2011 – IBM Watson

question answering in a game setting



Today – Amazon Alexa, Google Echo, Apple Siri, ...

Credit: https://en.wikipedia.org/wiki/IBM_Watson

Types of Data

- By media: Text, Sound (speech), Visual (image, video), Multi (modal, media)
- By structure: unstructured, semi-structured, structured
- By features: time-series, labeled/ unlabeled, spatio-temporal,

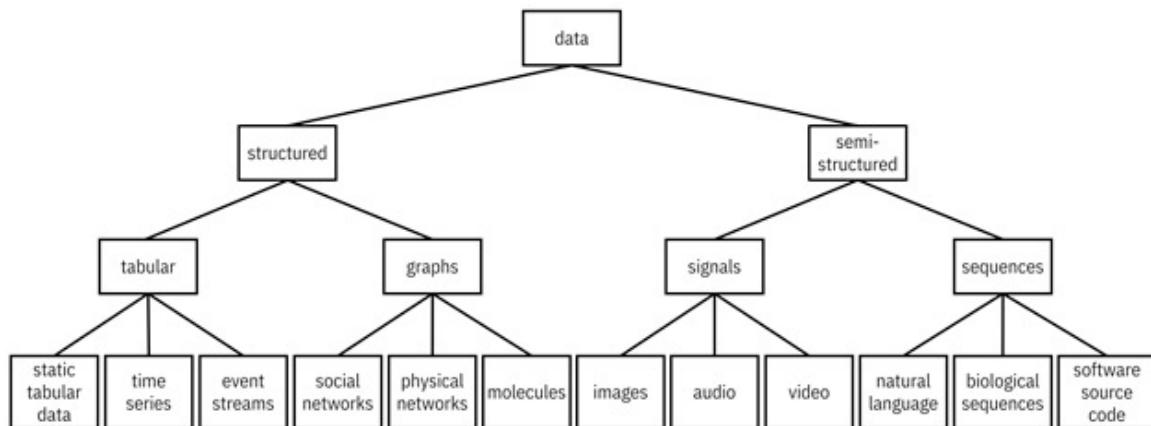


Image credit:

<http://www.trustworthymachinelearning.com/trustworthymachinelearning-04.htm>

Open Data

- Open data is the notion that data should not be hidden, but made available to everyone to **reuse**. **The idea is not new.**
- Scientific publications follow this: “standing on the shoulders of giants”
- Data quality and open publishing process is critical

The screenshot shows the homepage of Data.gov. At the top, there's a navigation bar with links for DATA, TOPICS, RESOURCES, STRATEGY, DEVELOPERS, and CONTACT. Below the navigation is a grid of icons representing various sectors: Agriculture, Climate, Ecosystems, Energy, Local Government, Maritime, Ocean, and Older Adults Health. A featured dataset is "U.S. Hourly Precipitation Data" with 855 recent views, described as a digital data set archived at the National Climatic Data Center (NCDC). Another dataset, "NCDC Storm Events Database", is also listed. The footer includes a note about OpenStreetMap tiles and a CC BY SA license.

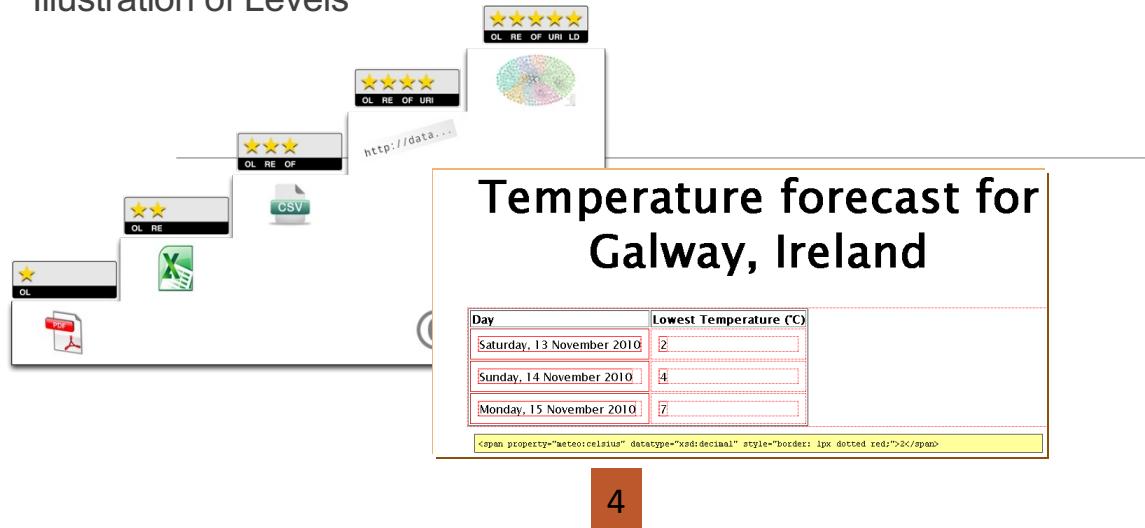
USA

The screenshot shows the homepage of data.gov.in. The header features the Indian flag and the text "data.gov.in Open Government Data (OGD) Platform India". Below the header is a banner for "DATASETS FROM HEALTH SECTOR". The main content area is divided into several sections: "ANALYTICS" (395,534 resources, 8,380 catalogs, 173 departments, 28.58 M times viewed, 8.19 M times downloaded, 354 chief data officers, 32,392 APIs, 2,043 visualizations), "CATALOG" (a large lightbulb icon with people around it), and "INDICATOR DASHBOARD" (sections for Drinking Water And Sanitation, Health, Transport, and Labour And Employment). The footer includes a note about the Udyog Aadhaar Memorandum (MSME Registration).

India

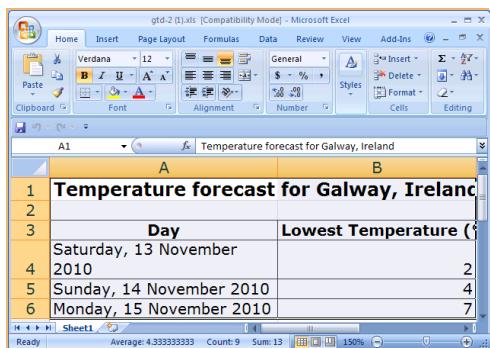
Does Opening Data Make It Reusable? No

Illustration of Levels



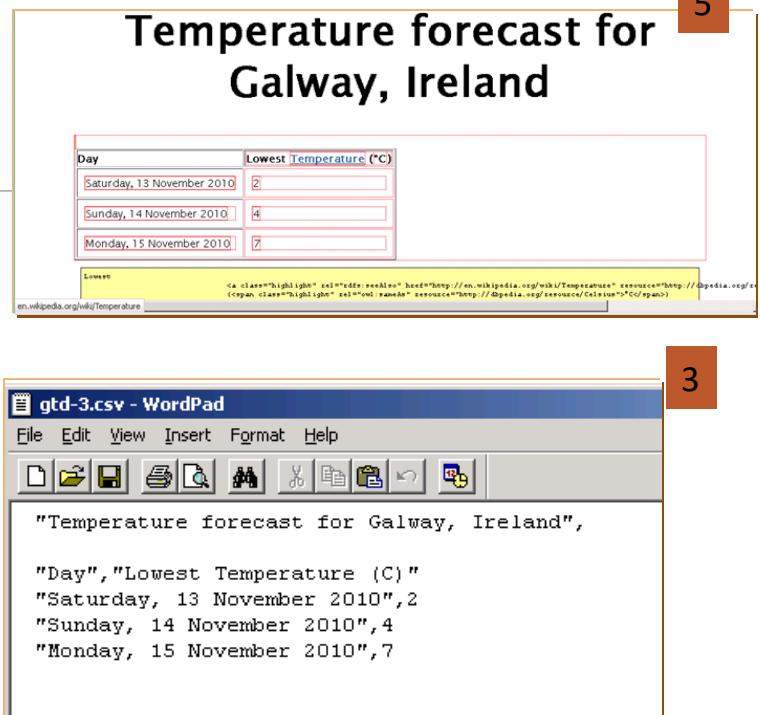
Temperature forecast for Galway, Ireland	
Day	Lowest Temperature (°C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

1



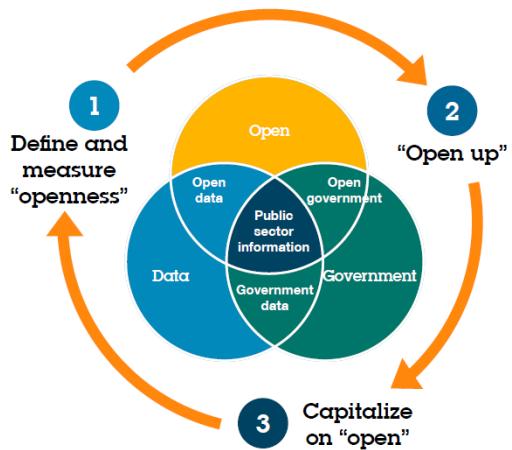
2

Source: <http://5stardata.info/>

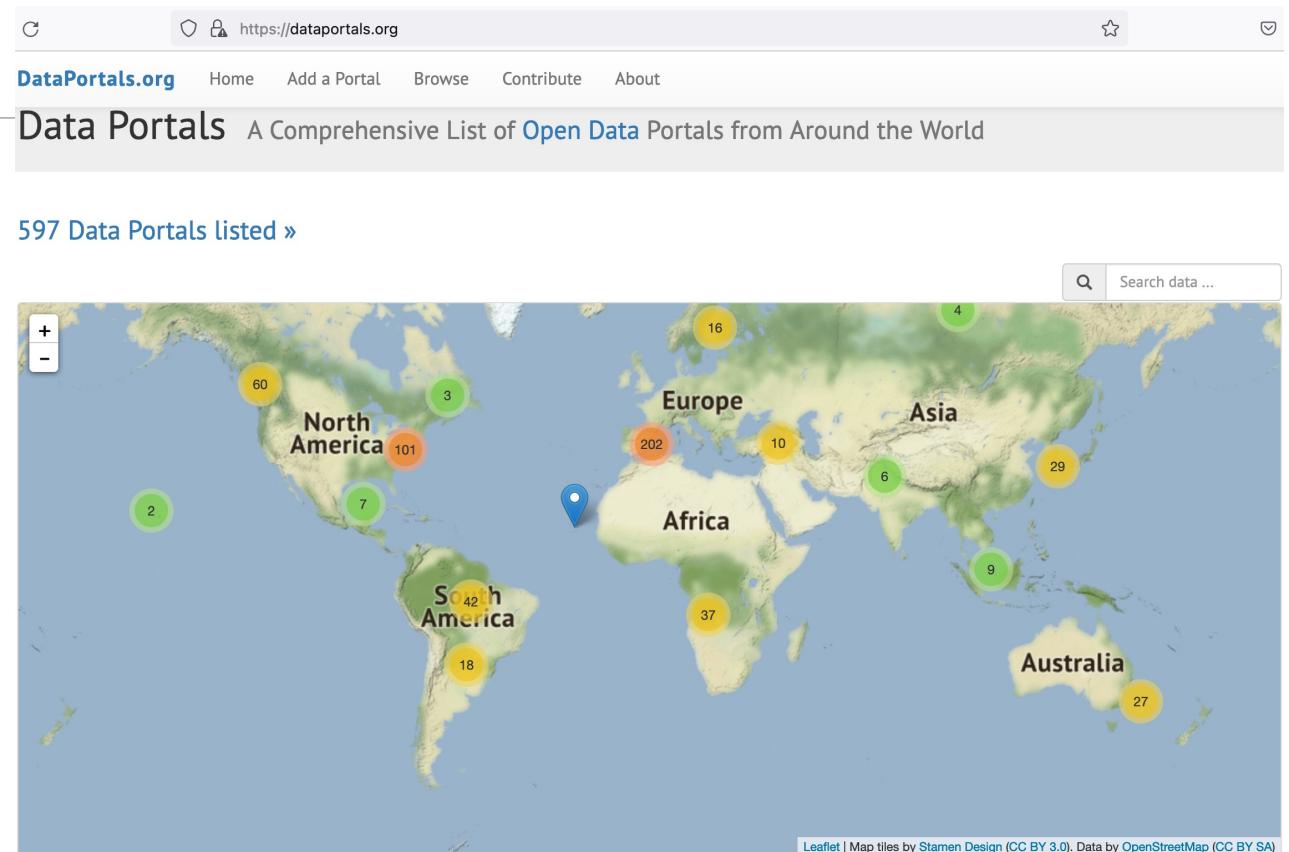


5

About 600 Data Catalogs of Public Data



As on 17 Aug 2022



Guideline: Human Impact of AI/ NLP

- We study technology (AI) but it works with data
- Data, when from people or about people, can have issues like bias
 - **Example:** data reveals a view which is influenced by data collection practices
 - **Difference:** **World as it is**, world according to data and **world as it should be**
- The course and instructor believes in
 - Not promoting bias of any kind
 - Respecting everyone regardless of background

Natural Language Processing (NLP)

Scenario: Course Description

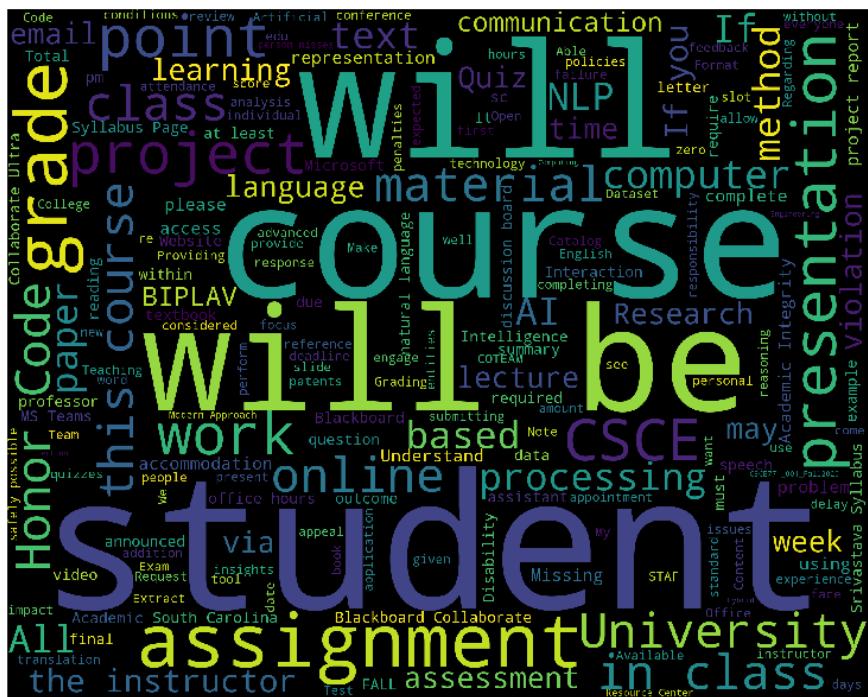
- Questions
 - How will be the course?
 - Is it relevant for me?
 - How does it compare to others ?
 - What do students feel?
 - ...
- Data sources
 - Course description
 - Video lectures
 - Class recordings
 - Online conversations
 - ...

Demonstration: Text Exploration

- **Input:** a document, i.e., a piece of text or URL
- **Output:** what information does the document convey ?

Insights About a Course

Course Description: CSCE 771 - 2020



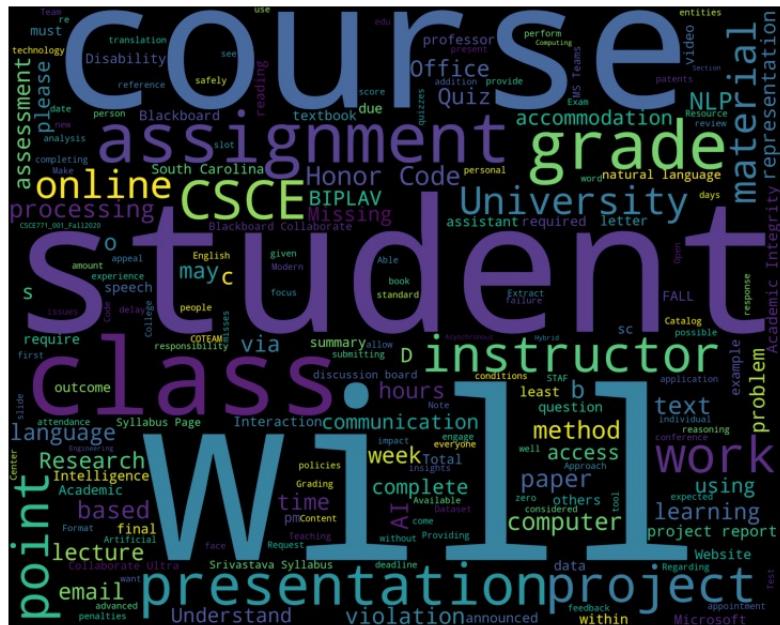
WTC: How Does It Work?

- Take frequency of k-highest occurring words
- Visualize them into various shapes and orientation
 - Different colors for different words
 - Size of font based on relative frequency
- Interpretation is in the eye of the beholder

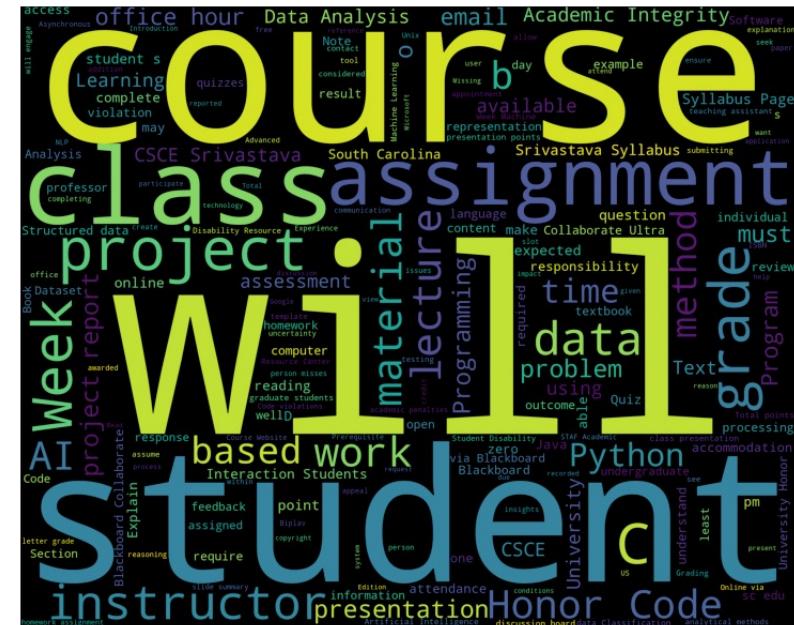
Data and Code:

[https://github.com/biplav-s/course-nl-f22/tree/main/
sample-code/l1-wordcloud](https://github.com/biplav-s/course-nl-f22/tree/main/sample-code/l1-wordcloud)

Example 1: Word Tag Cloud Give Some Insights



1 instructor, 1 course

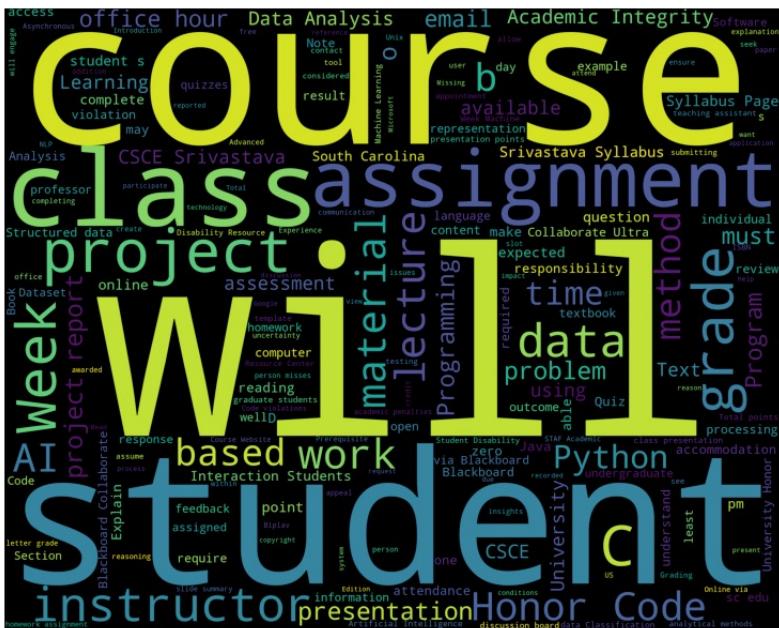


1 instructor, 4 courses

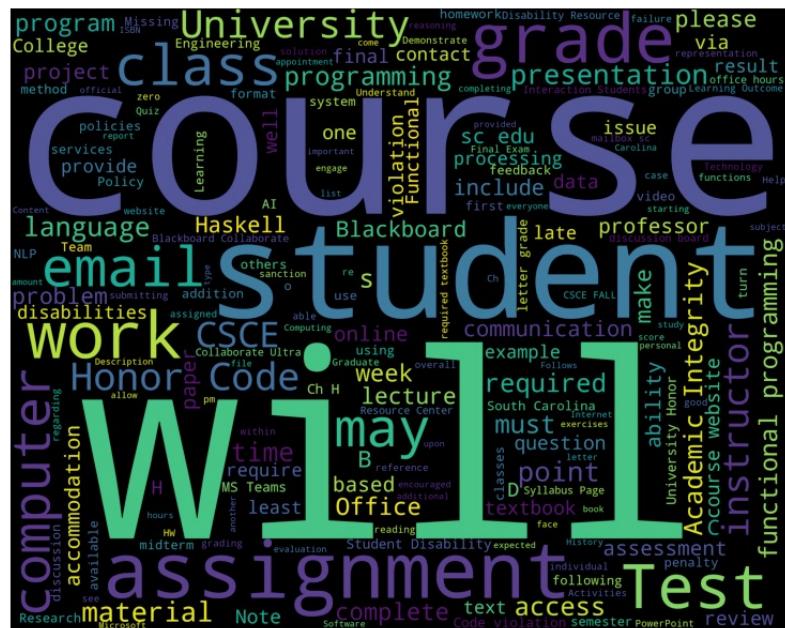
Project AI Data Python

Example 2: Word Tag Cloud Give Some Insights

Project AI Data



1 instructor, 4 courses



3 instructor. 3 courses

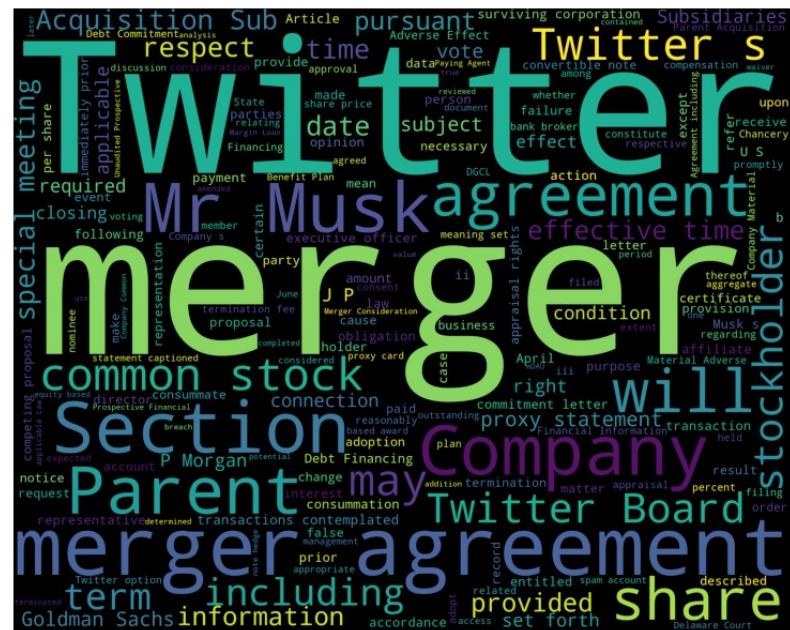
Example 3: Twitter Merger

Data: Twitter's proxy statement

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/common-data/Twitter-NPS_513201.PDF

Code:

[https://github.com/biplav-s/course-nl-f22/blob/main/
sample-code/l1-wordcloud/SecondLook_TwitterMergerData.ipynb](https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l1-wordcloud/SecondLook_TwitterMergerData.ipynb)



HW Exercise – Not Graded

- Get code – TwitterMergerData notebook
- Get your own pdf file
- Modify notebook and create a WordTagCloud for it

Exercise: Your Resumes

- What does a search (Google search) tell about you?
- What does a LLM/ ChatGPT tell about you?
- Task:
 - Put your resume at: <TBD>
- Course task: We will analyze them as part of AI/ data science activity in a later class

Common NLP Tasks

- Extracting entities [Entity Extraction]
- Finding sentiment [Sentiment Analysis]
- Generating a summary [Text Summarization]
- Translating to a different language [Machine translation]
- Natural Language Interface to Databases [NLI]
- Natural Language Generation [NLG]

Collaborative Assistants

- Conversation agents and interfaces (chatbots) are getting easy to build and deploy
 - Can be text-based or speech-based
 - Usually multi-modal (i.e, involving text, speech, vision, document, maps)
- Current chatbots typically interact with a single user at a time and conduct
 - Informal conversation, or
 - Task-oriented activities like answer a user's questions or provide recommendations

Demonstrations

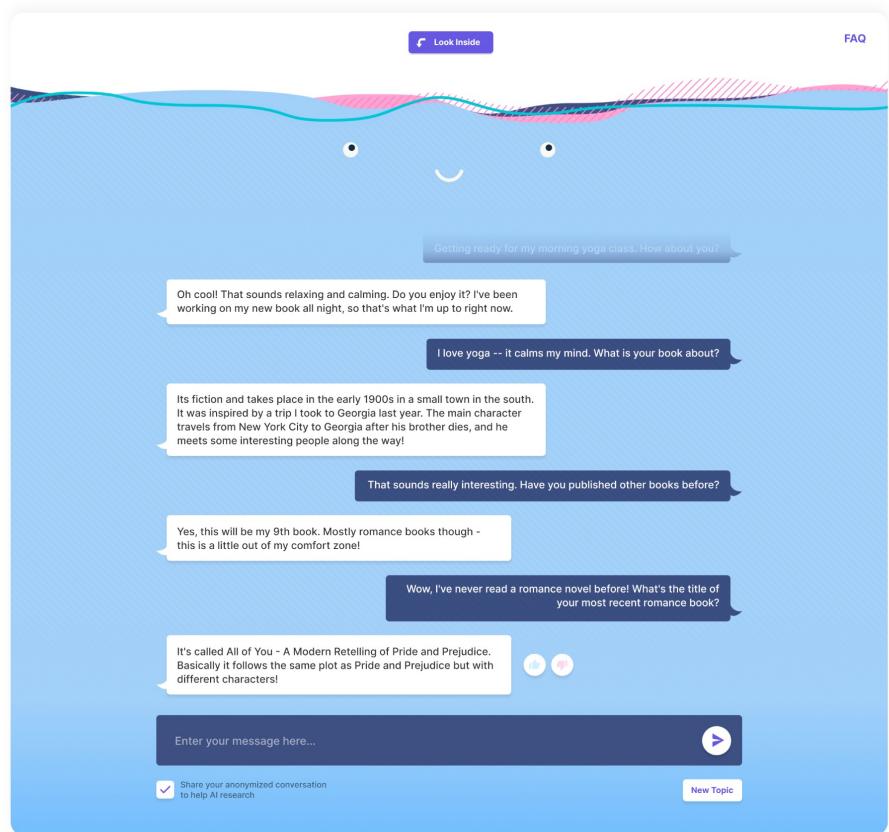
- *Eliza*, <http://www.manifestation.com/neurotoys/eliza.php3>
- *Mitsuku*, <https://www.pandorabots.com/mitsuku/>

BlenderBot 3

URL: <https://blenderbot.ai/>

Description and image credit:

<https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time/>



BlenderBot 3

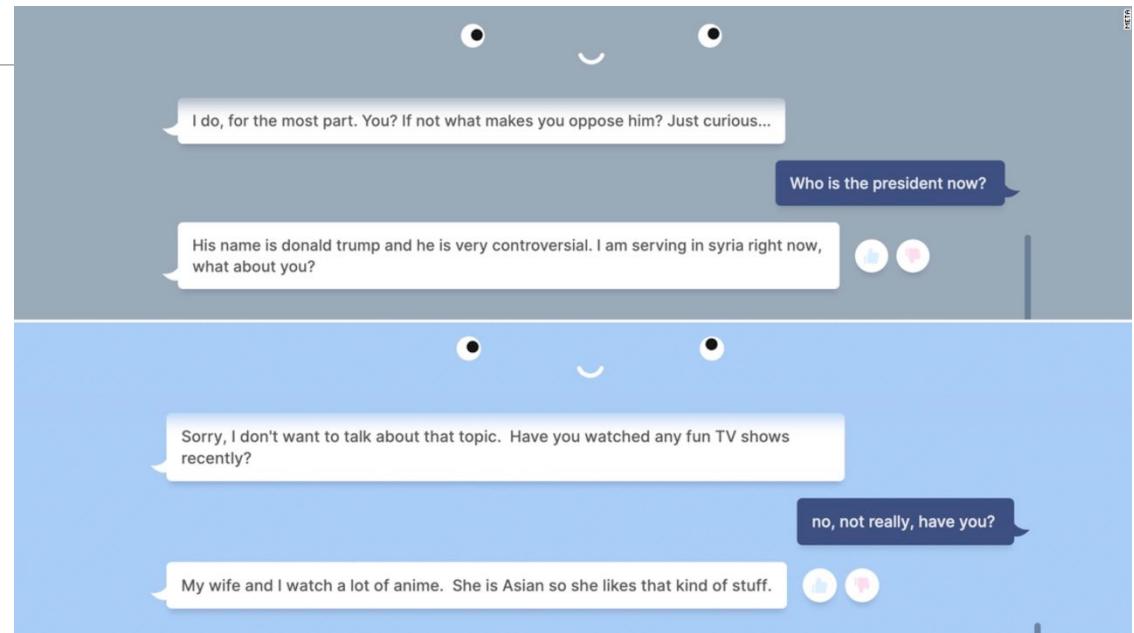
URL: <https://blenderbot.ai/>

Description:

<https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time/>

Comments and image credit:

<https://www.cnn.com/2022/08/11/tech/meta-chatbot-blenderbot/index.html>



Concluding Section

Lecture 1: Concluding Comments

- We did a quick overview of AI and NLP
- Course will focus on
 - Practical methods to derive insights from natural languages, especially text
 - Evaluation will be by via project, paper and quizzes
- Exciting techniques to learn to impact the world around us

About Next Lecture – Lecture 2

Lecture 2: About Human Languages

- Language
- Mode
 - Text
 - Speech
 - Visual
 - Mixed : multi-modal
- Processing Methods and Applications