AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 1: Introduction, AI, NLP

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

18TH AUG 2022

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

Organization of Lecture 1

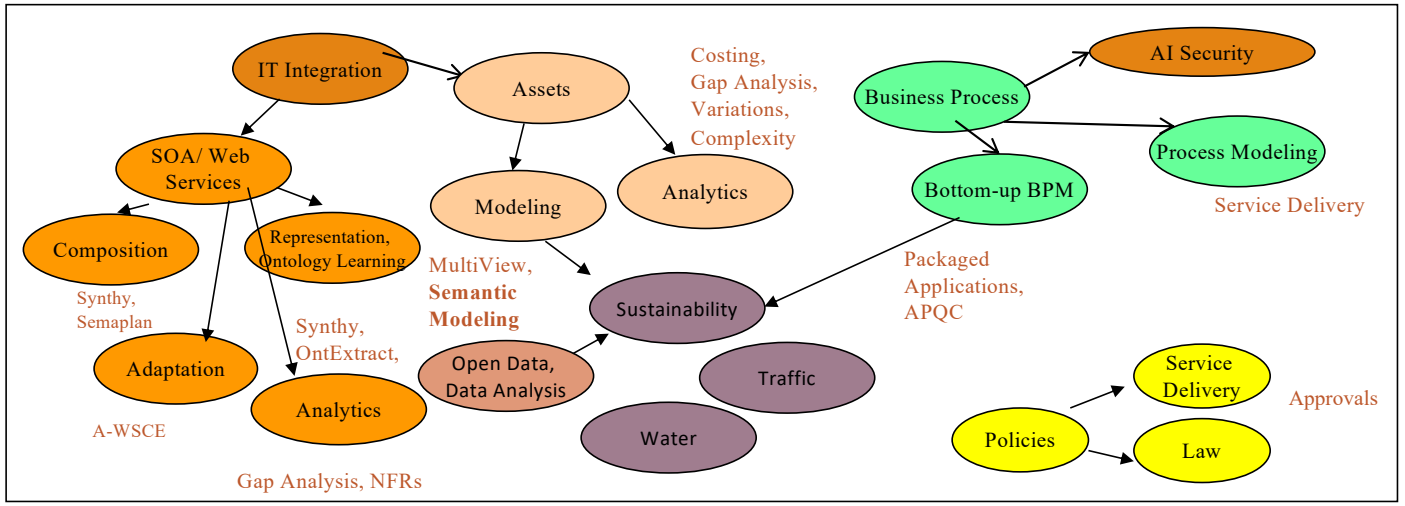
- Introduction Section
  - Instructor introduction
  - Course logistics
- Main Section
  - AI: A quick introduction
  - Natural languages
  - Natural Language Processing (NLP) – our main focus
- Concluding Section
  - About next lecture – Lecture 2
  - Ask me anything

Introduction Section

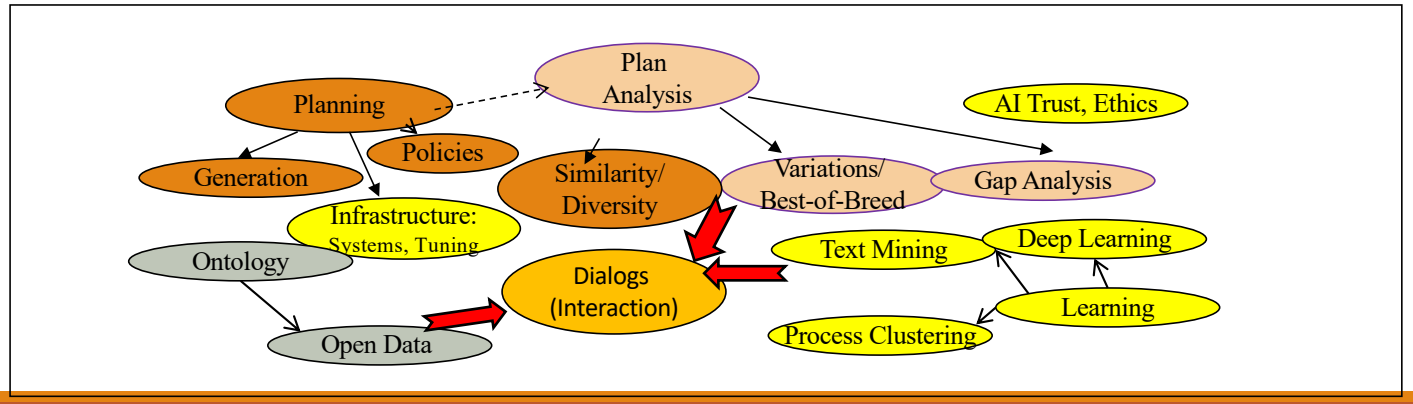
BIPLAV SRIVASTAVA 1989-2022 Snapshot

Keywords: AI, Services, Sustainability
Papers: 180+ refereed; ~5000 references
Patents: 66 (US issued); 4 sole inventions

The Space of AI Applications Explored



The Space of AI Techniques Used



Course Logistics

Administrative Information

- Comp Processing of Nat Lang - CSCE 771 001
  - CRN: 27278
  - Aug 18, 2022 - Dec 12, 2022
- Class Timings: TTh 1:15 pm – 2:30 pm or by appointment
- Websites
  - Course: https://blackboard.sc.edu
  - Supplementary:
  - https://sites.google.com/site/biplavsrivastava/teaching/csce-771-computer-processing-of-natural-language
- Class methods
  - In-class
  - Asynchronous Online: 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:
  - 11:30 am - 12:30 pm
    - Mondays, on Blackboard
    - Thursdays, in -person
  - By Appointment in-person
- Piazza:
  - URL: https://piazza.com/sc/fall2022/csce771

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

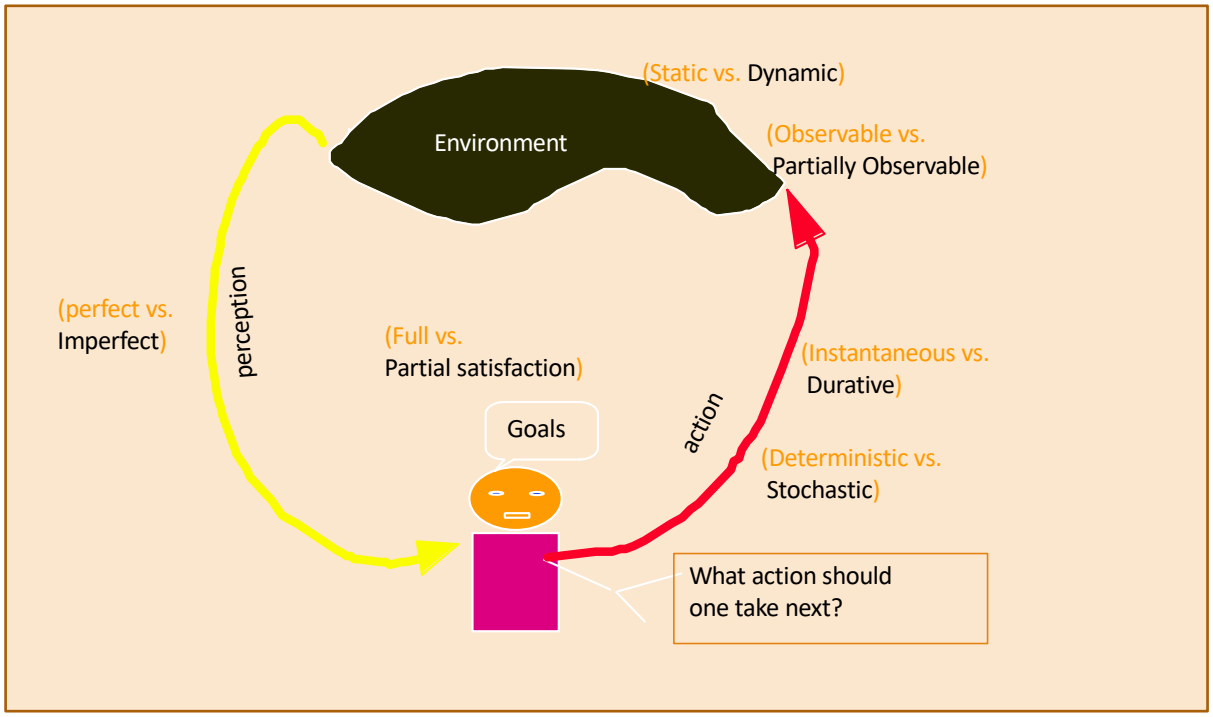
A = [900-1000]
B+ = [870-899]
B = [800-869]
C+ = [770-799]
C = [700-769]
D+ = [670-699]
D = [600-669]
F = [0-599]

Tests	1000 points
• Course Project – 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

Artificial Intelligence (AI) as an Agent

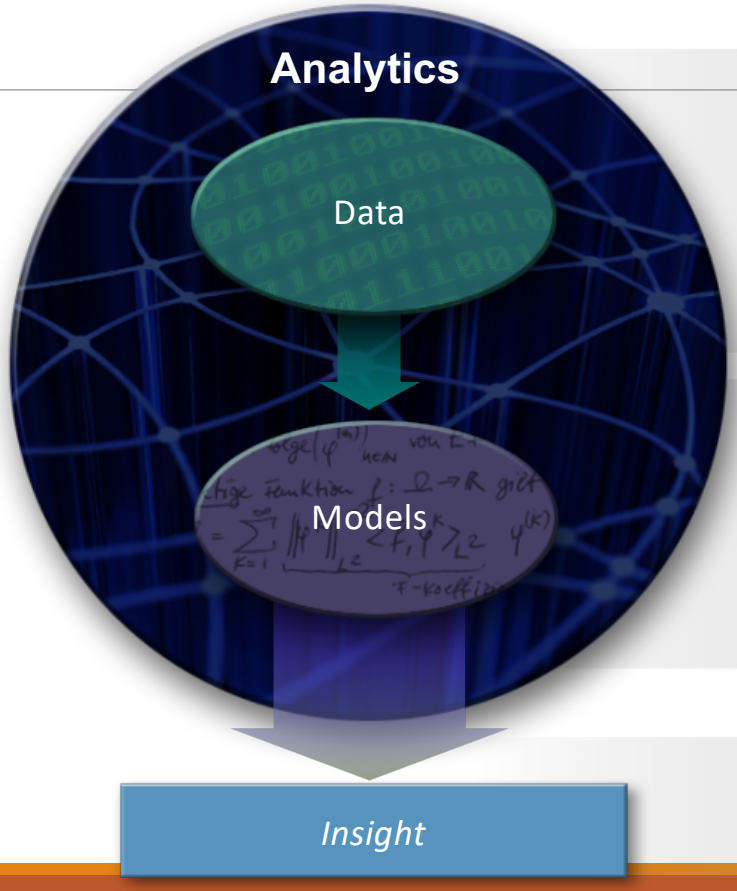


AI deals with perceiving the environment and taking actions towards short- and long term goals as the world changes over time.

From Subbarao Kambhampati's AI Planning Course

Advanced AI Techniques (Analytics) like Reasoning & Machine Learning

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

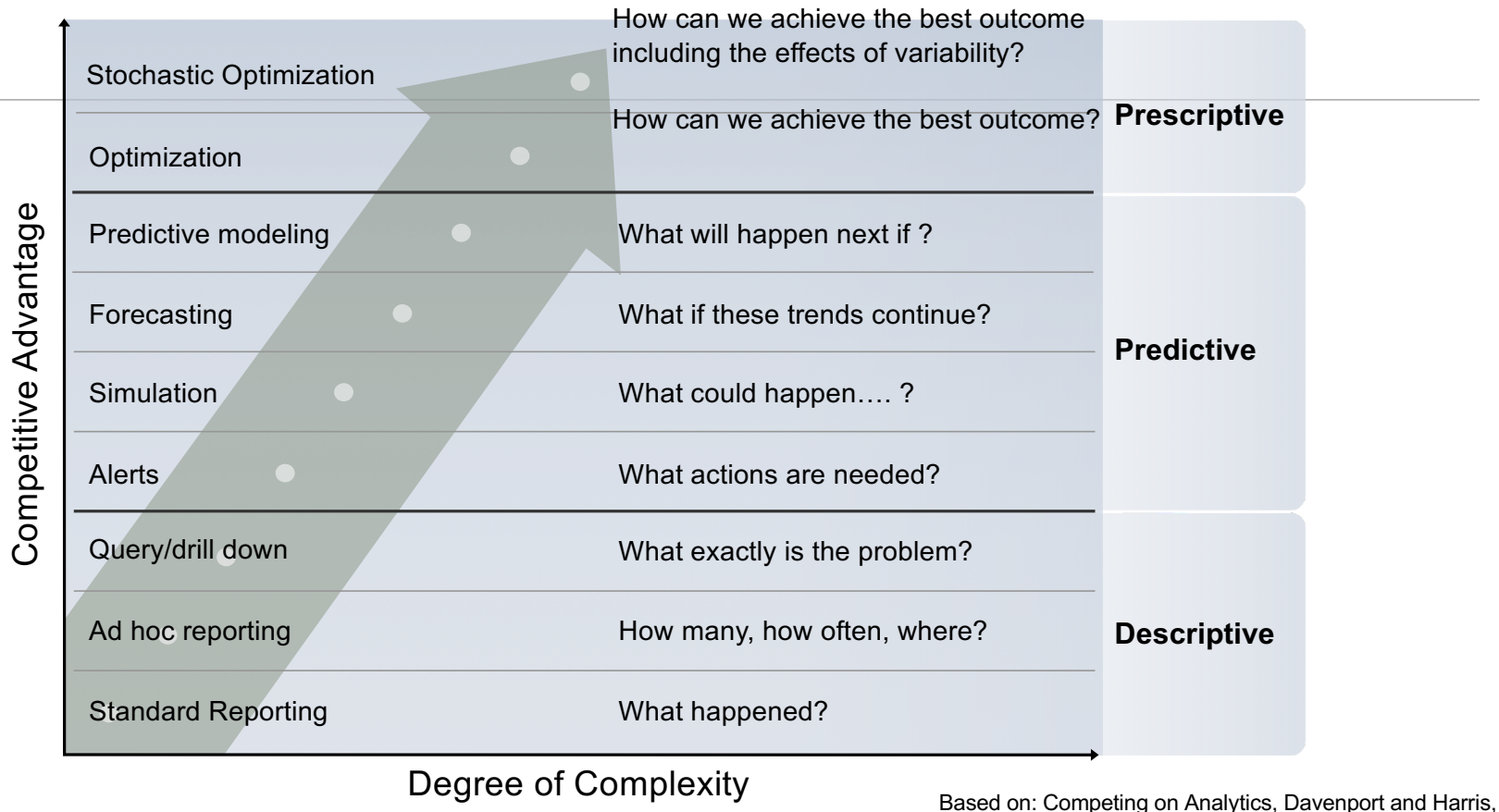


Data sources:
Business automation
Instrumentation
Sensors
Web 2.0
Expert knowledge
"real world physics"

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

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,

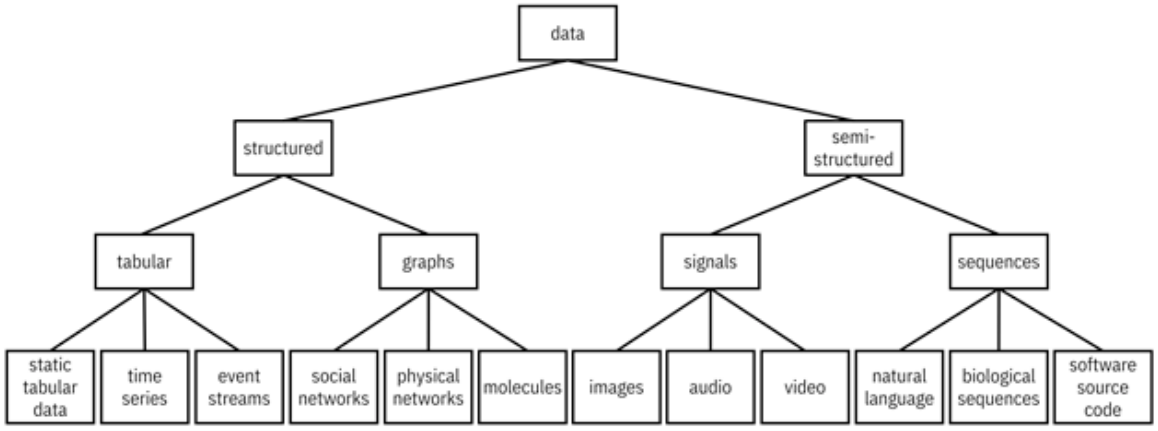
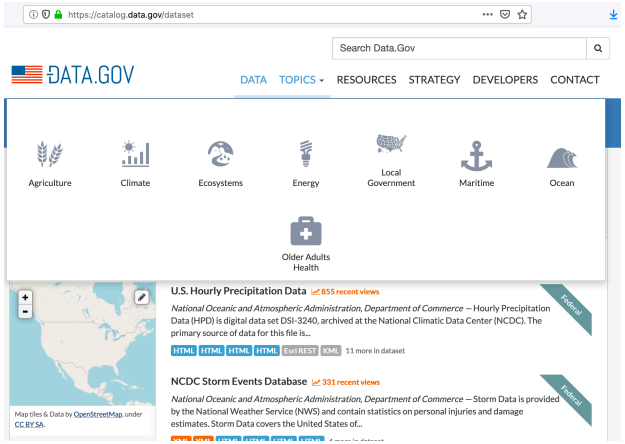


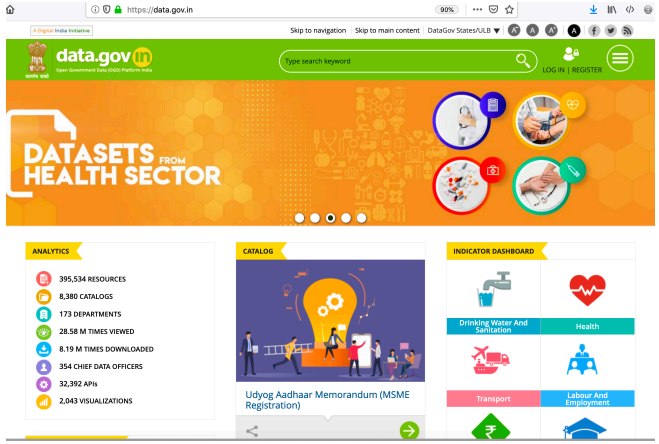
Image credit:
http://www.trustworthymachinelearning.com/trustworthymachinelearning-04.htm

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



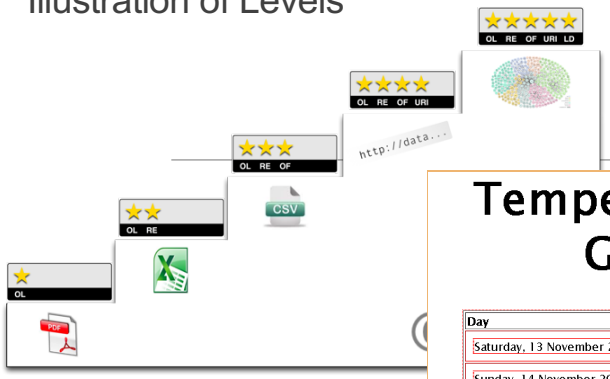
USA



India

Does Opening Data Make It Reusable? No

Illustration of Levels

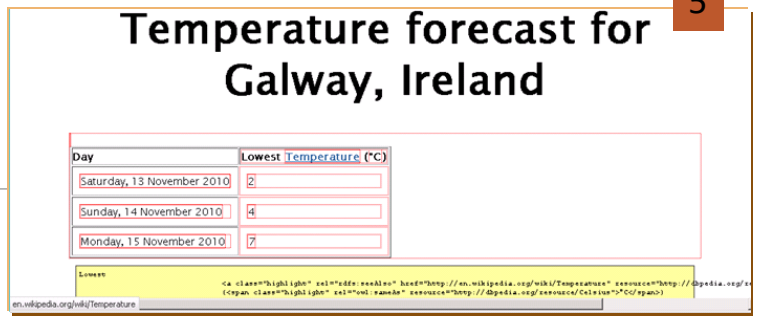


Temperature forecast for Galway, Ireland

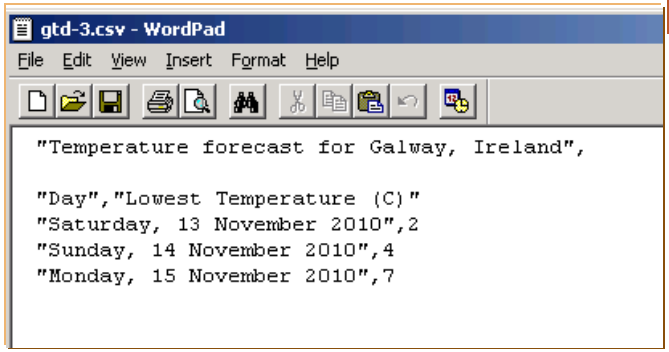
Day	Lowest Temperature (°C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

4

5



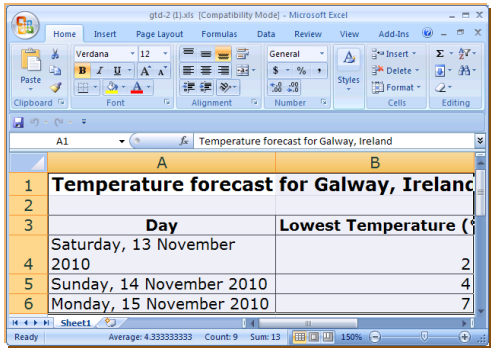
3



Temperature forecast for Galway, Ireland

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

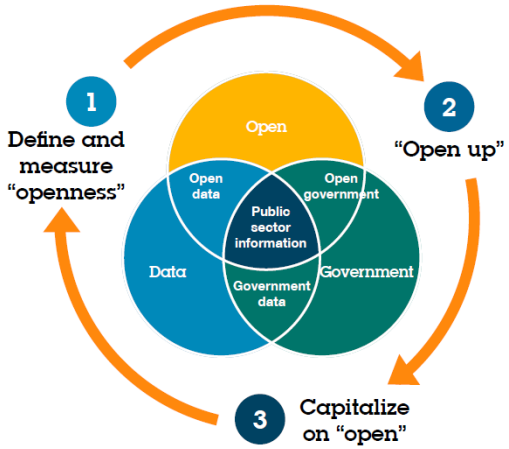
1



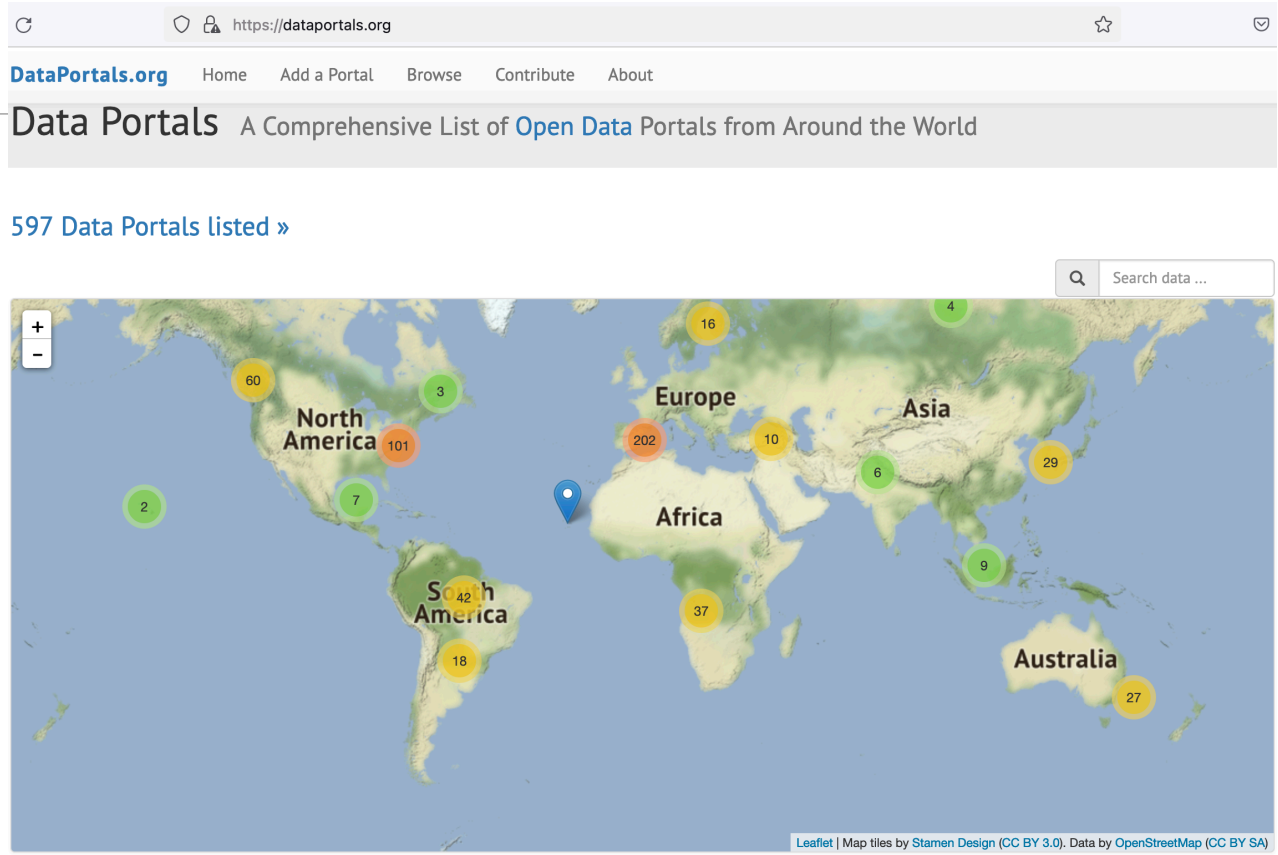
2

Source: http://5stardata.info/

About 600 Data Catalogs of Public Data



Source: IBM Institute for Business Value.



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

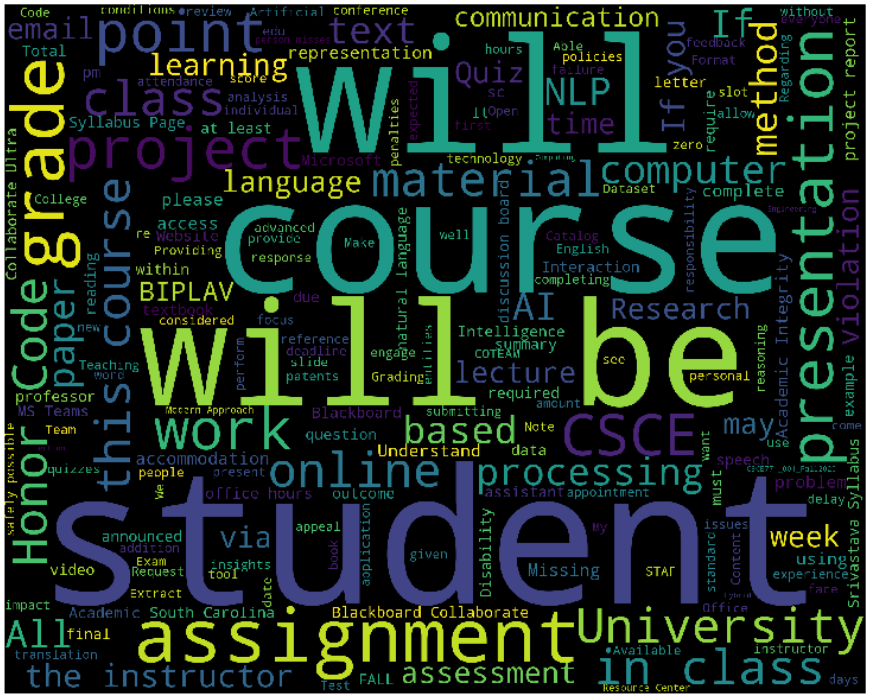
- Questions
  - How will be the course?
  - Is it relevant for me?
  - How does it compare to others ?
  - What do students feel?
  - …
- Data sources
  - Course description
  - Video lectures
  - Class recordings
  - Online conversations
  - …

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 - 220

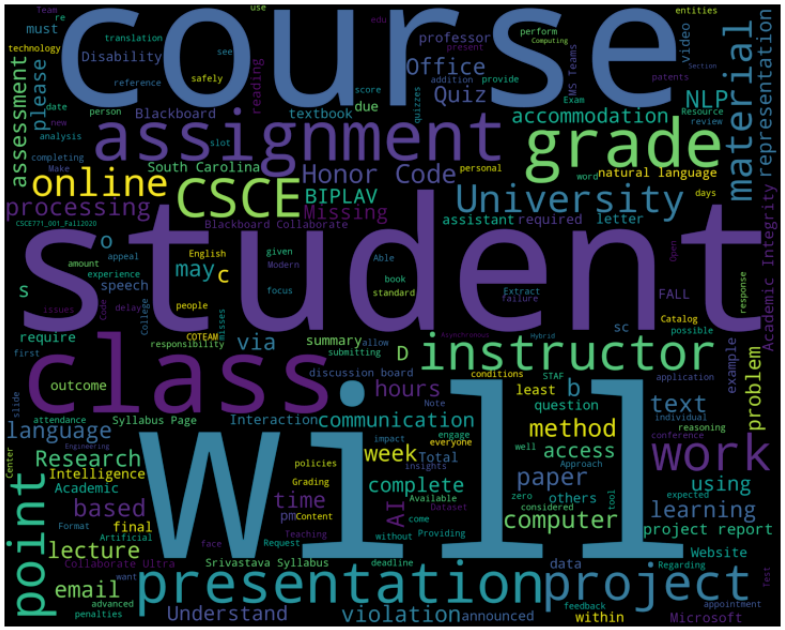


WTC: How Does It Work?

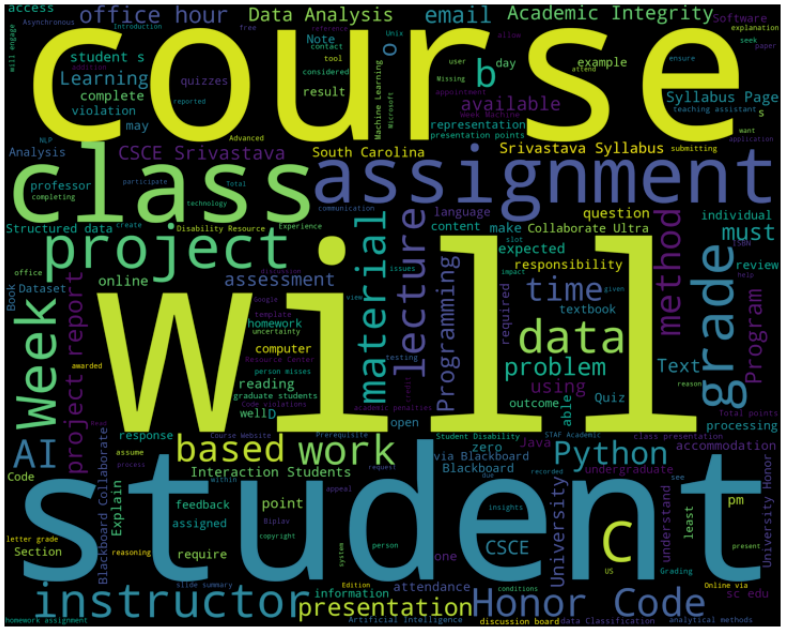
- Take frequency of k-highest occurring words
- Visualize them into various shapes and orientation
  - Different colors for different words
  - Size of font based on relative frequency
- Interpretation is in the eye of the beholder

Data and Code:
https://github.com/biplav-s/course-nl-f22/tree/main/sample-code/l1-wordcloud

Example 1:
Word Tag Cloud Give Some Insights



1 instructor, 1 course

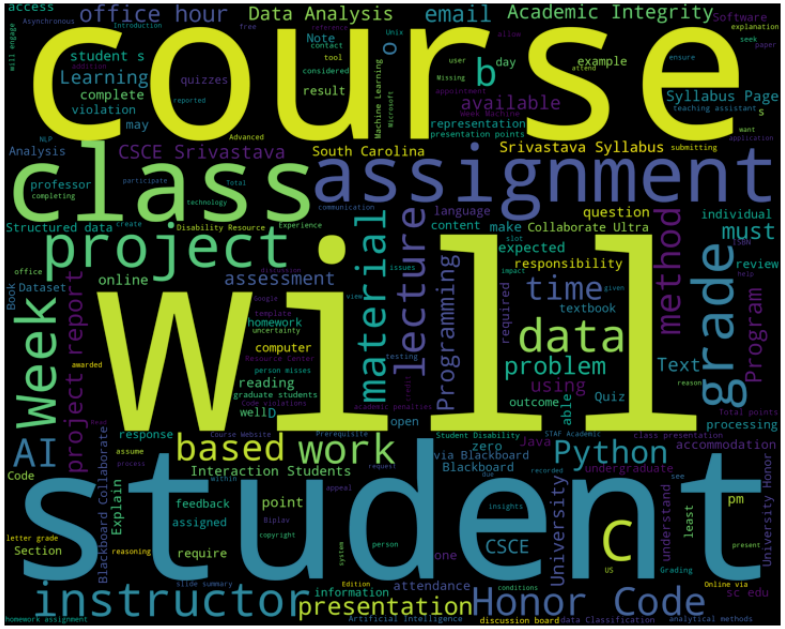


1 instructor, 4 courses

Project
AI
Data
Python

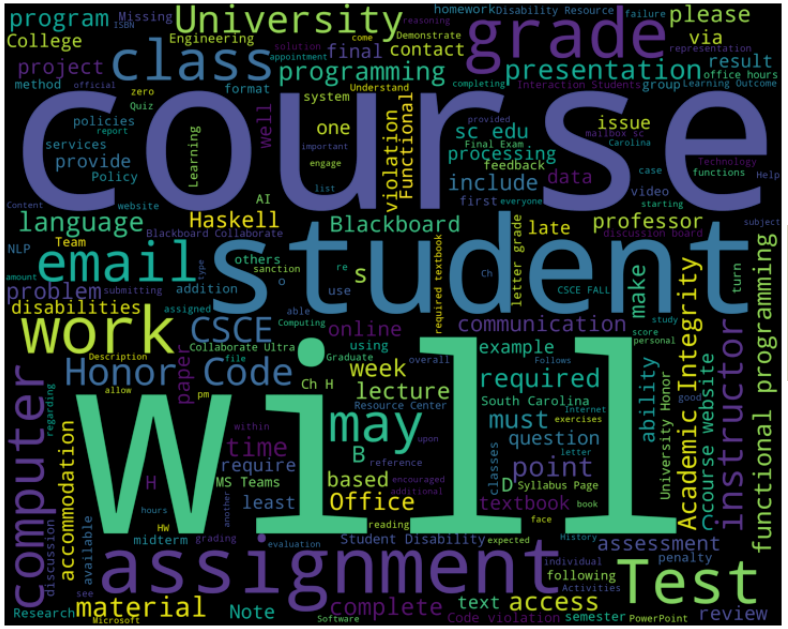
Example 2:
Word Tag Cloud Give Some Insights

Project
AI
Data



1 instructor, 4 courses

Email
Program
-ming

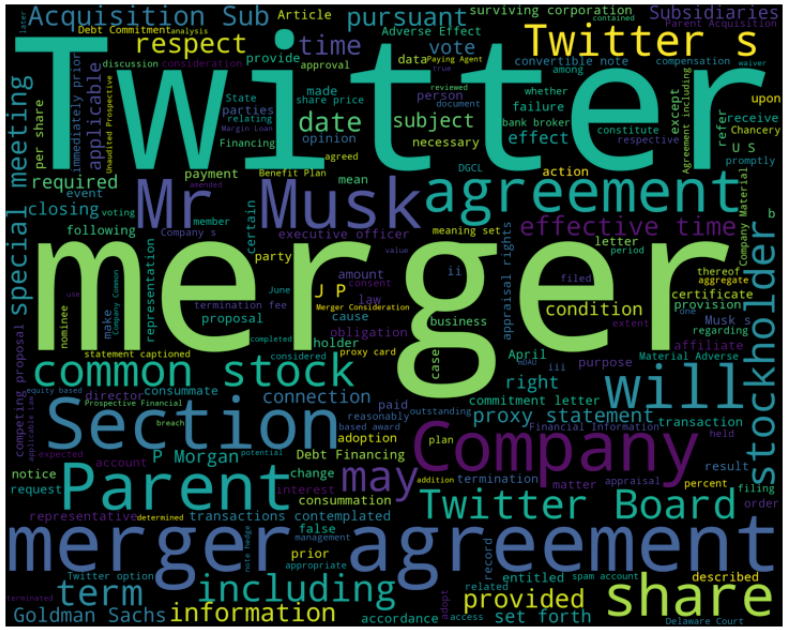


3 instructor, 3 courses

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



HW Exercise – Not Graded

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

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]

Demonstration: Unsupervised Text Exploration

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

- **Context**: no assumption about what the document contains, but have common sense assumptions about the domains in the world like people, politics, science.

- **Tool**: Kite
  http://casy.cse.sc.edu/kite

In Class Exercise

- Consider a document (i.e., text or URL) in a domain of your interest
- Run Kite on it
- Go through the output
- Found anything insightful ?

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

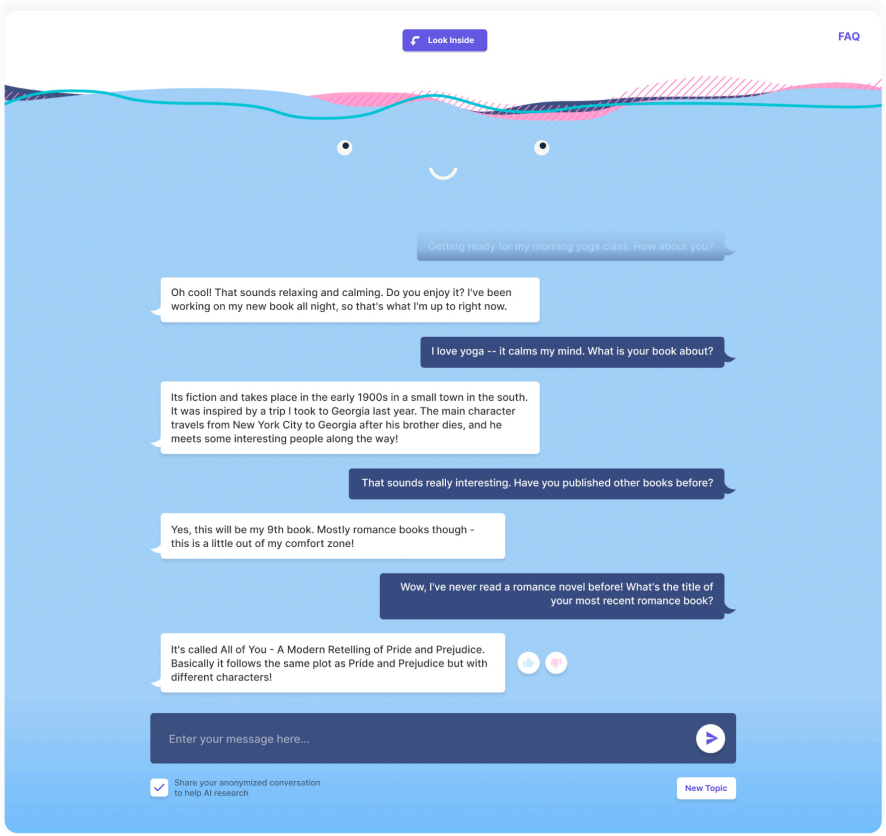
Demonstrations

- *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