Course Overview

Anupam Datta

John Mitchell

Stanford CS 329T, Spring 2021

Course Logistics

- Lectures/"Fireside chats": Tue 2:30pm-3:50pm PST
- Labs (attendance required in your assigned session):
  - Session 1/2: Thu 2:30pm-3:30pm PST
  - Session 3/4: TBD (based on your input)

THIS WEEK: Thurs 2:30 for all

- Web page: http://web.stanford.edu/class/cs329t/

- Gradescope (assignment submissions)
- Canvas (grades)
- Piazza (announcements, for all other communication)

- Stanford Honor Code

Course staff

- Instructor: Anupam Datta
  - Email: danupam@stanford.edu
  - Office hours: TBD

- Instructor: John C. Mitchell
  - Email: jcm@cs.stanford.edu
  - Office hours: Wed 11-12 noon (starting April 7)

- Office hours available remotely:
  - Zoom meeting links on website

Course staff

- TA: Soham Gadgil
  - Email: sgadgil@stanford.edu
  - Office hours: TBD

- TA: Shreya Singh
  - Email: ssingh16@stanford.edu
  - Office hours: TBD

- Office hours available remotely:
  - Zoom meeting links on website

Course staff

- Contributing instructor: Divya Gopinath
  - Email:
  - Office hours:

- Contributing instructor: Piotr (Peter) Mardziel
  - Email: piotrm@gmail.com
  - Office hours:

- Office hours available remotely:
  - Zoom meeting links on website

Today

- Goals
- Modules
  - Background
  - Explainability
  - Fairness
  - Privacy
  - Robustness
- Classwork
  - One homework for each module

- Logistics covered further in lab Thursday 2:30
  - Course Format
  - Prerequisites
  - Grading
  - Policies

Machine Learning Systems are Ubiquitous

Google

Big Data in Government, Defense and Homeland Security 2015 - 2020

April 3, 2013, Vol 309, No. 13 >

< Previous Article | Next Article >

Viewpoint | April 3, 2013

The Inevitable Application of Big Data to Health Care

Travis B. Murdoch, MD, MSc; Allan S. Detsky, MD, PhD

[+] Author Affiliations

NEW YORK, May 12, 2015 /PRNewsv

How Big Data Could Replace Your Credit Score

Credit scores are useful in determining who gets loans, but they're far from perfect. AvantCredit determines loan-worthiness based on all sorts of factors, including your use of social media and prepaid cell phones.

Big Data in Education

Learn how and when to use key methods for educational data mining and learning analytics on large-scale educational data.

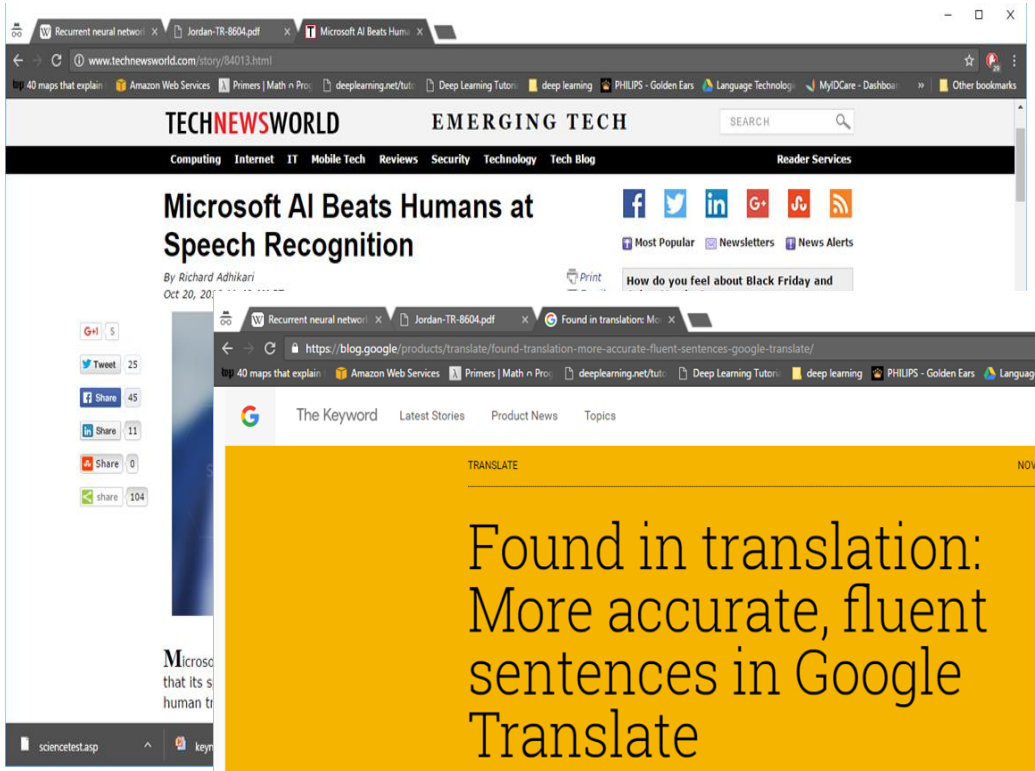
TEACHERS COLLEGE
COLUMBIA UNIVERSITY

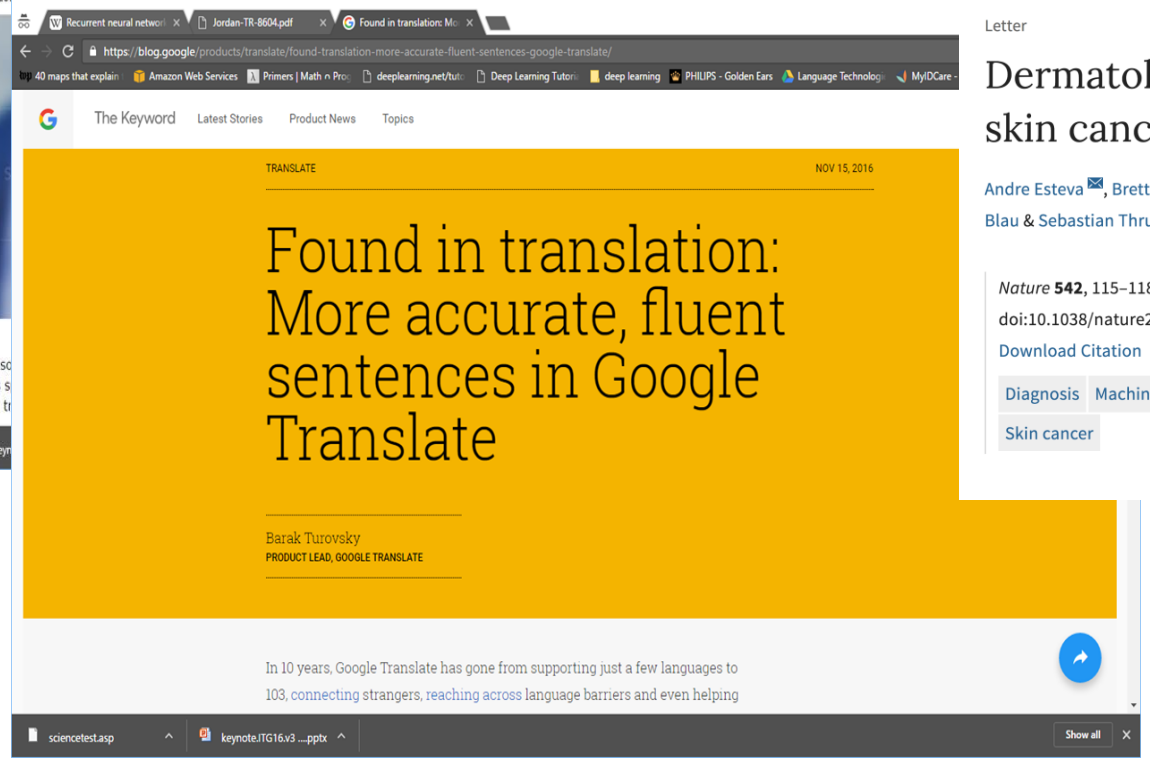
amazon

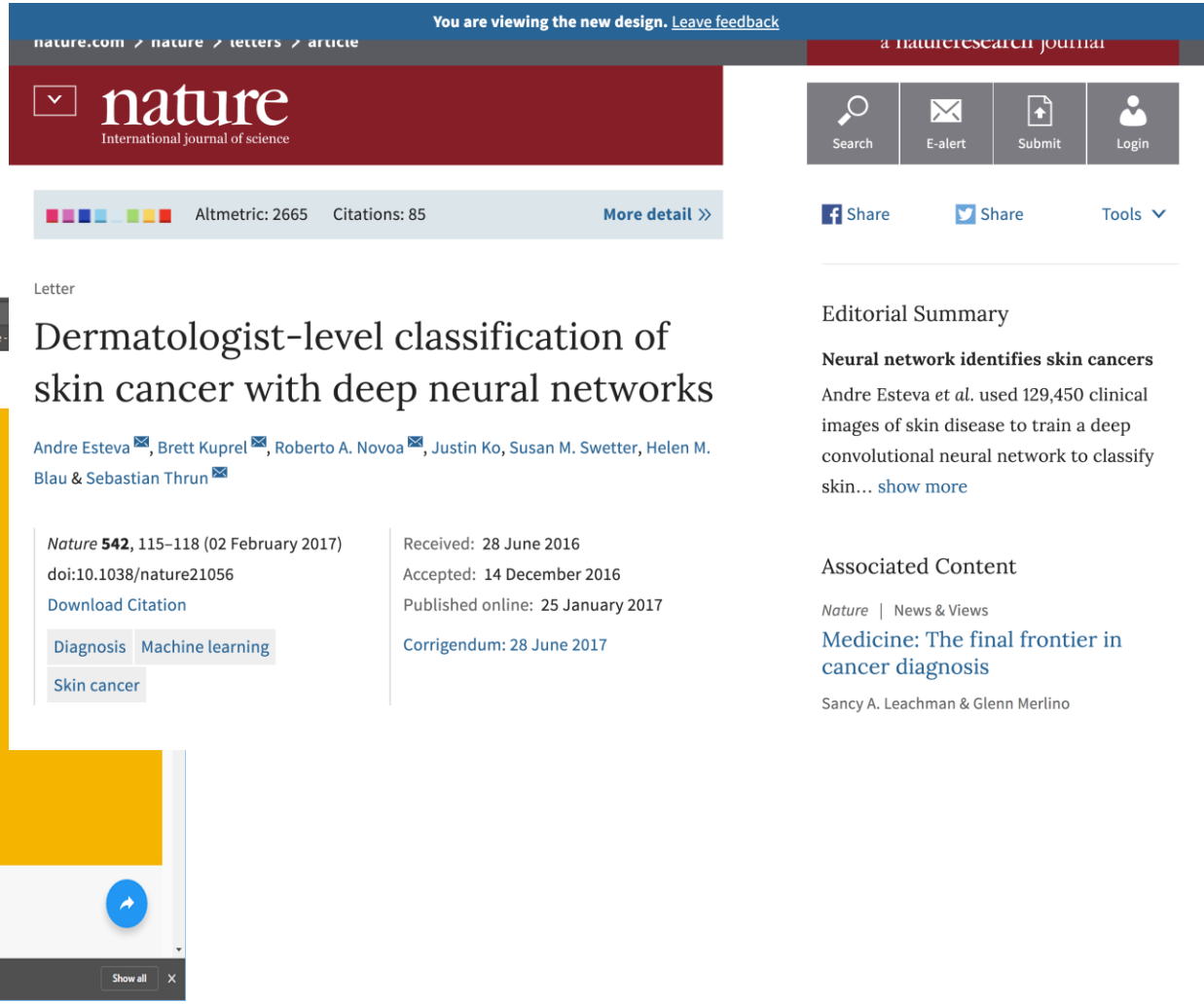
facebook

bing

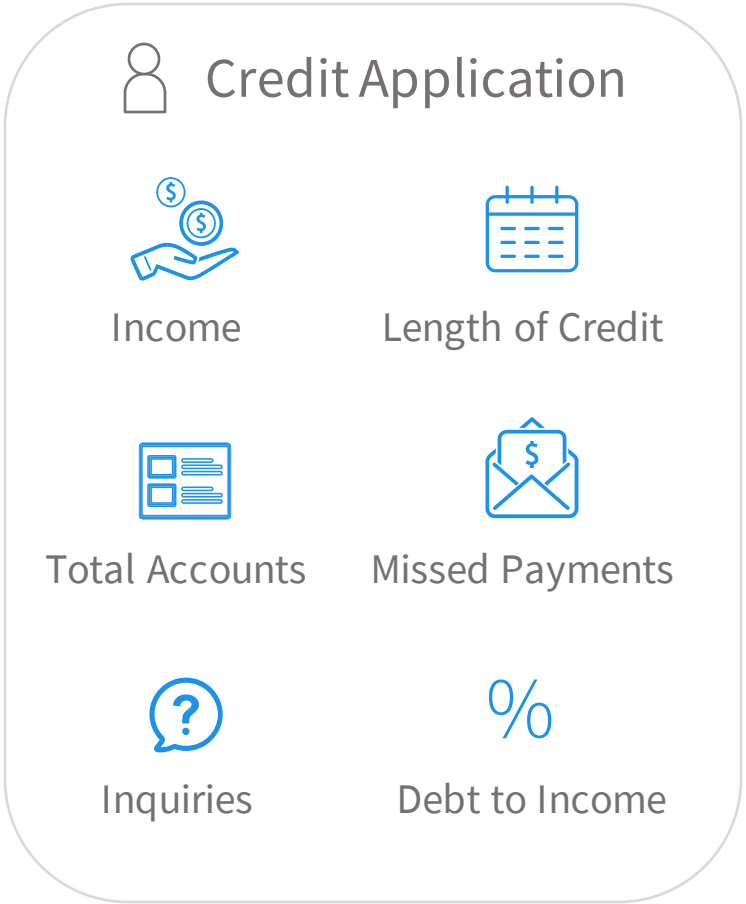
Continuing successes of deep learning

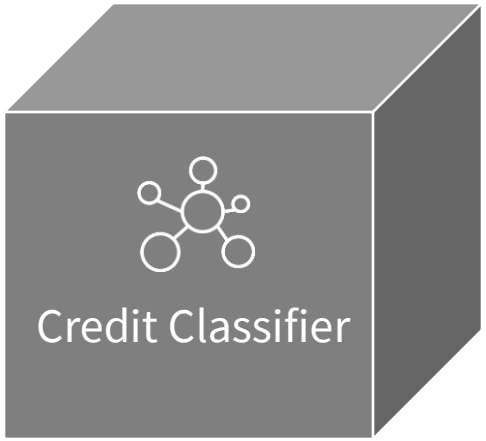




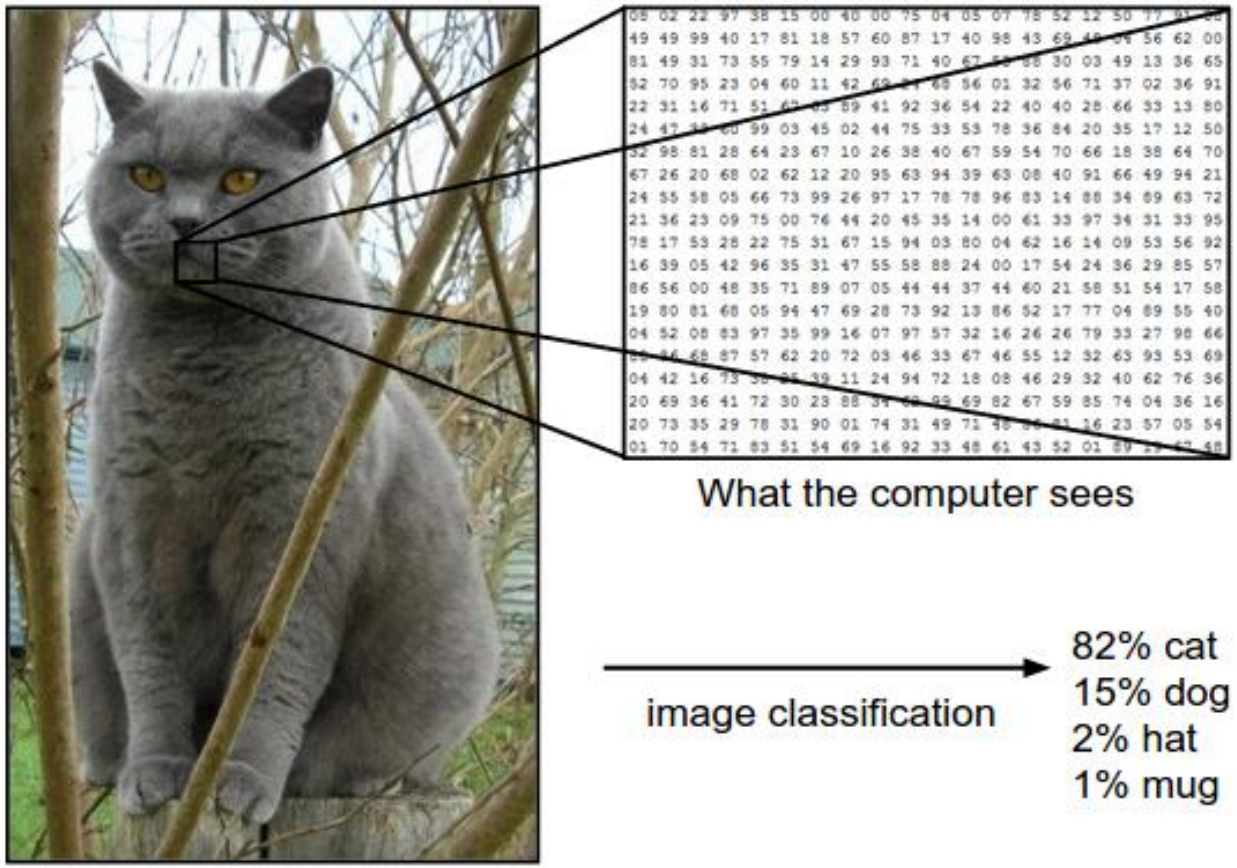


ML Models over Structured Data

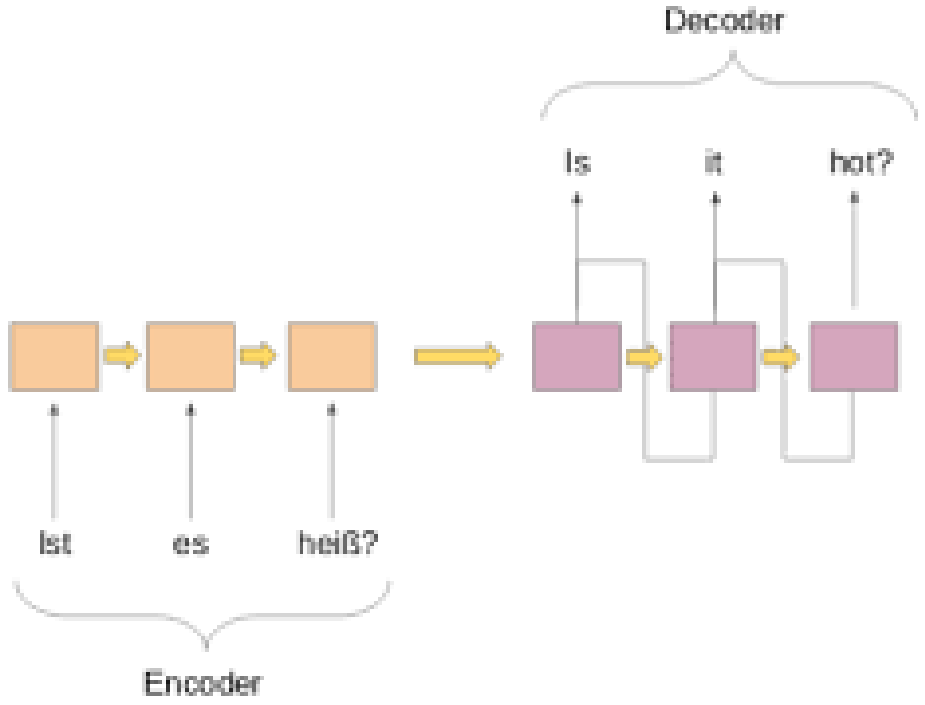


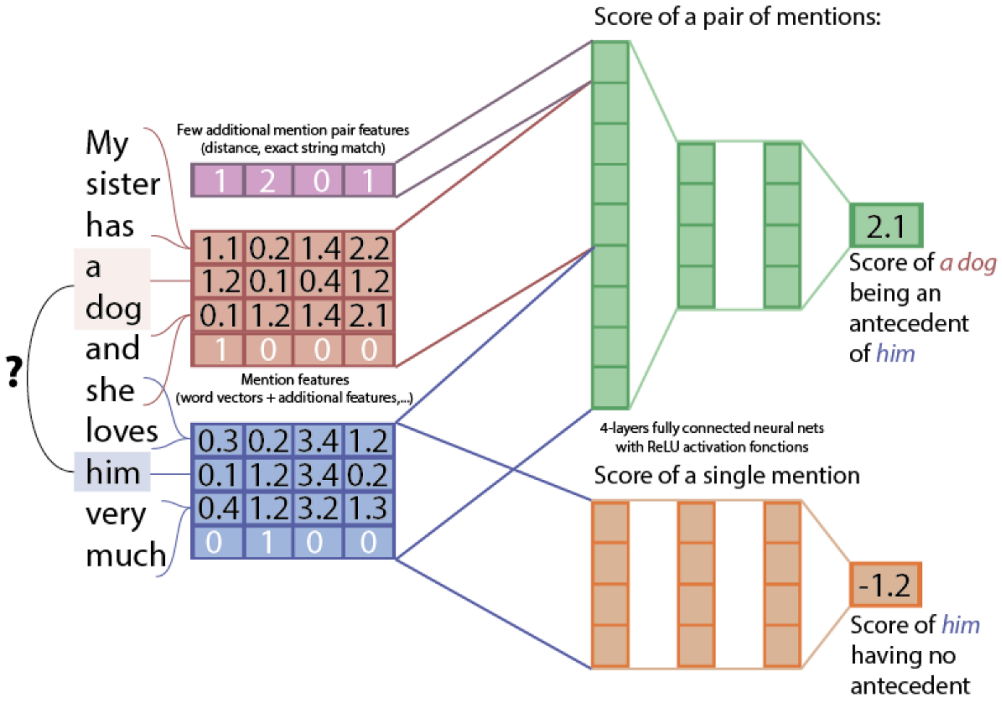


ML Models over Image Data



ML Models for Natural Language Processing





Significant Barrier to Adoption of ML

Assessing and improving trustworthiness of ML systems

- National Academies Workshop (March 2021)
- Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (December 2020)
- EU Ethics Guidelines for Trustworthy Artificial Intelligence (December 2018)
- FEAT Principles (November 2018)
- ...

Attributes of Trustworthiness

- Explainability
- Fairness
- Privacy
- Robustness
- Accuracy
- Auditability
- Reproducibility
- …

Course objective

Understand how to assess and improve trustworthiness of ML Models

Course modules

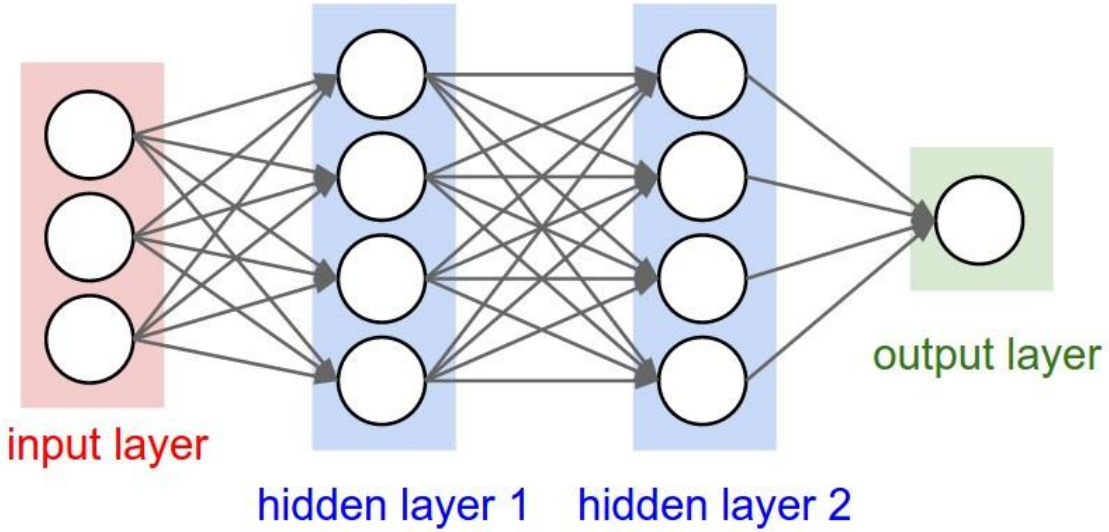
Pre-req/review: Fundamentals of machine learning & deep learning

Elements of Trustworthy Machine Learning

1. Explainability
2. Fairness
3. Privacy
4. Robustness

Spanning traditional Statistical Machine Learning and Deep Learning

Course module 0: Machine Learning Review



python

NumPy

Keras

Homework 1

Practice using basic tools:

- Scikit-learn
- Numpy
- Keras

By implementing digit classifiers using:

- Logistic regression
- Deep neural network models

Sample application exercises with logistic regression:

- Explanations
- Adversarial examples
- Model stealing

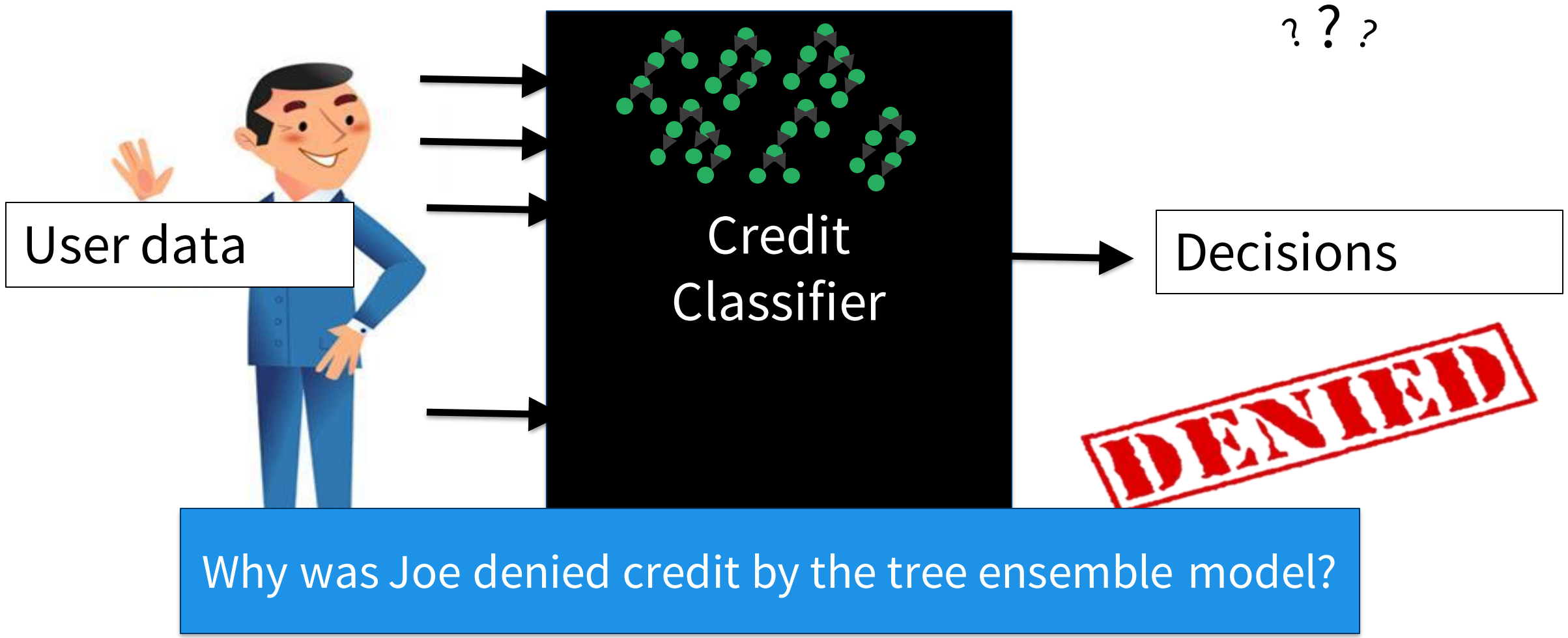
python

Course module 1: Explainability

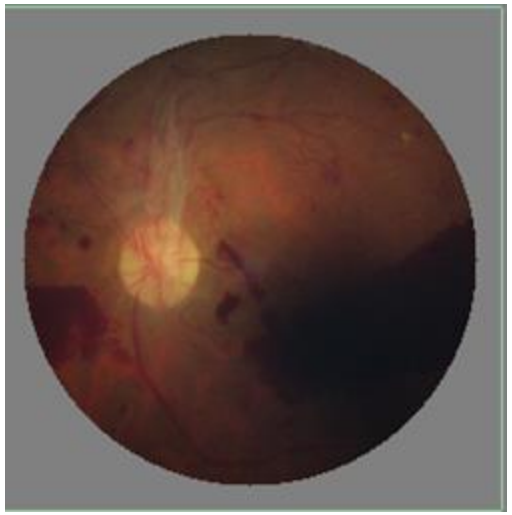
Explanations enable understanding and debugging of ML models

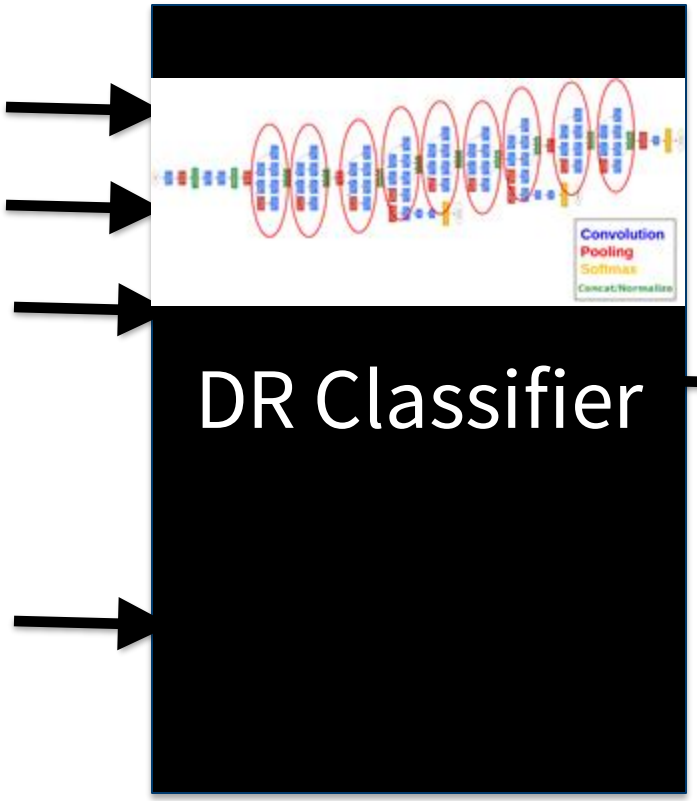
- How can we explain individual predictions from ML models?
- How can we explain an ML model's behavior at a global level?
- How do we evaluate explanation methods in a principled manner?
- How can explanations enable testing, debugging, and improvement of ML models?

Explanations: Structured Data, Statistical ML Models



Explanations: Unstructured Data, Deep Neural Networks





? ? ?

Diabetic retinopathy Stage 5

Why this diagnosis from the GoogleNet neural network?

Homework 2

Explanations for traditional models

- Model-agnostic: Shapley Values, local approximations (QII, LIME, SHAP)
- Model-specific optimizations: TreeSHAP

Explanations for deep models

- Saliency Maps
- Integrated Gradients
- Influence-directed Explanations

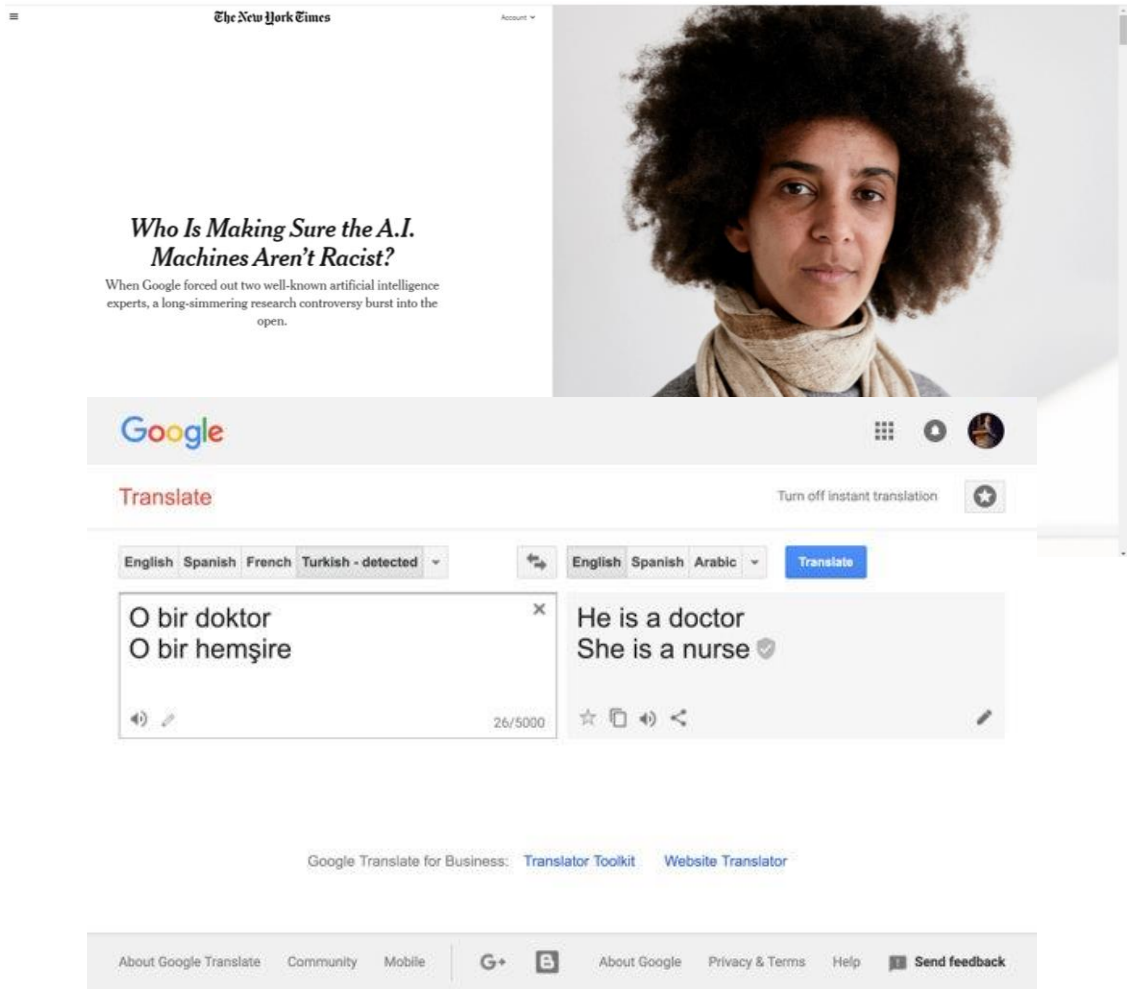
What will I be doing?

- Understand, reason about, and extend the theoretical principles backing explanation frameworks like QII, LIME, SHAP, Integrated Gradients, etc.
- Implement a subset of these methods yourself (LIME, some deep explanations)
- Play around with open-source packages that enable explanations (SHAP, Trulens)

Course module 2: Fairness

Unfair bias measurement and mitigation

Fairness Matters



- Facial Recognition
- Online Advertising
- Natural Language Processing

Homework 3

Fairness metrics

- Demographic fairness in classification models (structured data, deep neural networks)
- Biased word embeddings (text data, word2vec)

Debiasing

- Adversarial training (deep neural networks)
- Subspace projection (text data, word2vec)

Course module 3: Privacy

ML Models may leak information about their training data subjects

- How do we formalize privacy risks from ML?
- How do we mitigate these privacy risks?

Homework 4

Privacy risks

- Membership inference: models that remember too much about training data (deep networks, image data).

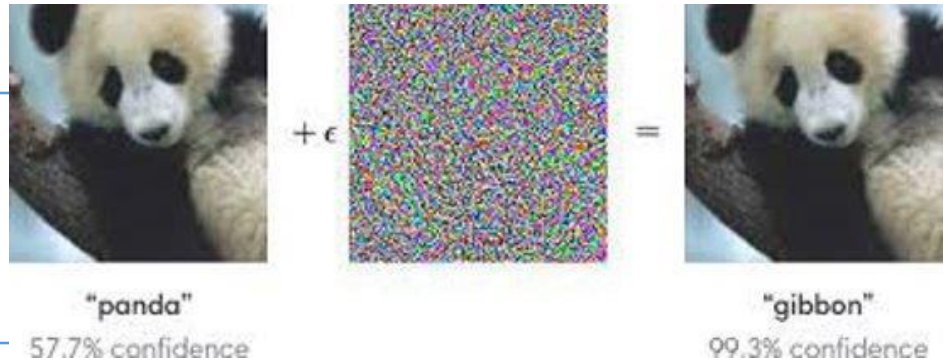
Course module 4: Robustness

Attacks on classifiers and defenses

Homework 5

Adversarial attacks.

- CNN's over image datasets.



Defenses.

- Defensive distillation, adversarial training

Summary

- Goal
  - Understand how to assess and improve trustworthiness of ML Models
- Modules
  - Background
  - Explainability
  - Fairness
  - Privacy
  - Robustness
- Classwork
  - One homework for each module
  - Final project (10% of grade)

Course Structure

- Activities: Offline lectures, "fireside chats", labs
- Weekly:
  - Offline/prerecorded lectures
    - Prerecorded video
  - Fireside chats
    - Occasionally guest lectures
  - Labs
    - Background, software, homework intros, homework help

Prerequisites

Grading

- Homework: 80%
  - 5 x 16%
- Final report: 10%
- Class participation: 10%
  - Be present and engaged in class and piazza
  - Informed questions for guest lecturers

Collaboration policy on homework

Acknowledgment

- Builds on material from
  - CMU Spring 2018-2020 18739: Security and Fairness of Deep Learning