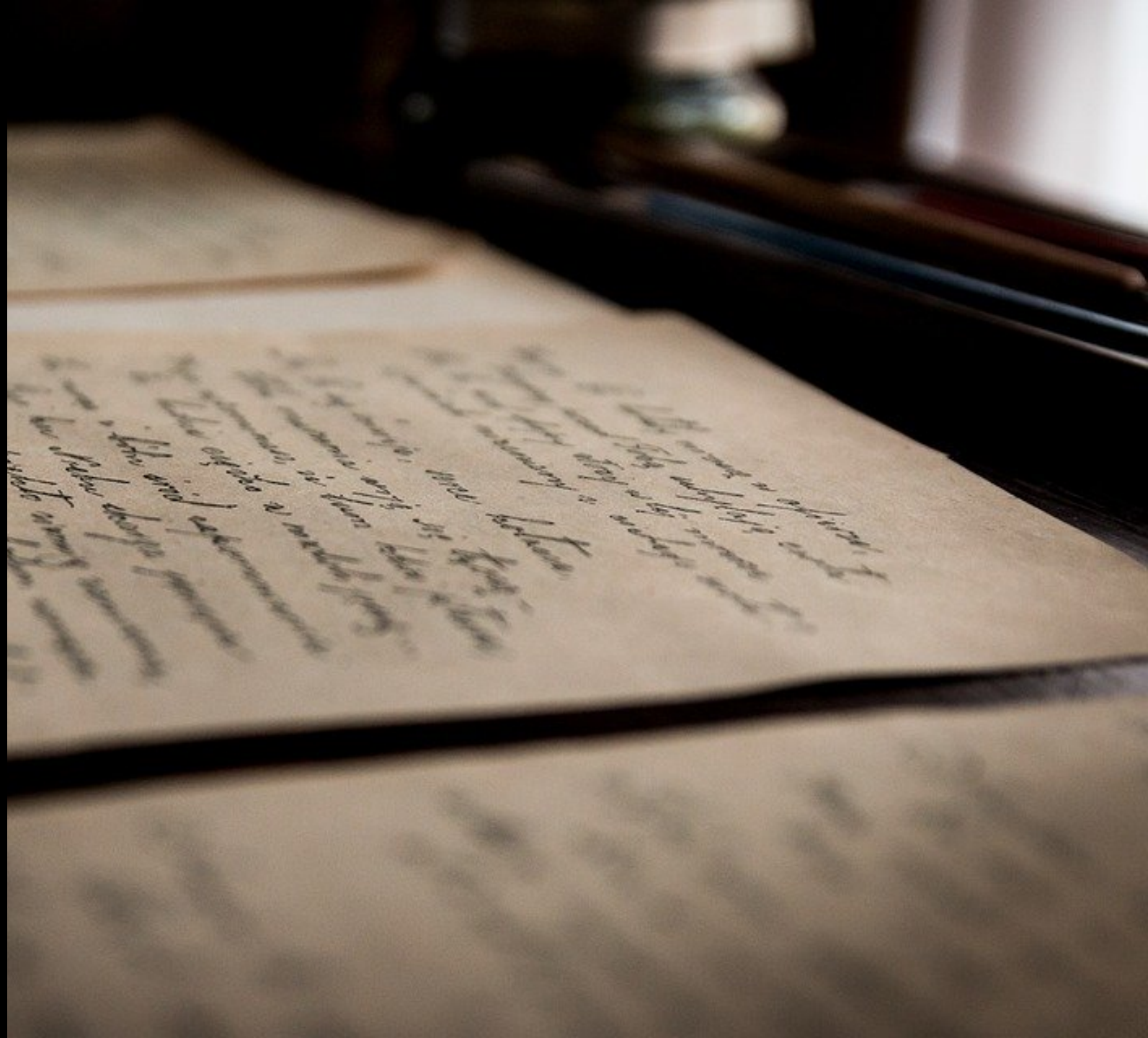
Cursive & Recursive:

Generating Transcriptions of Archival Documents Using Machine Learning



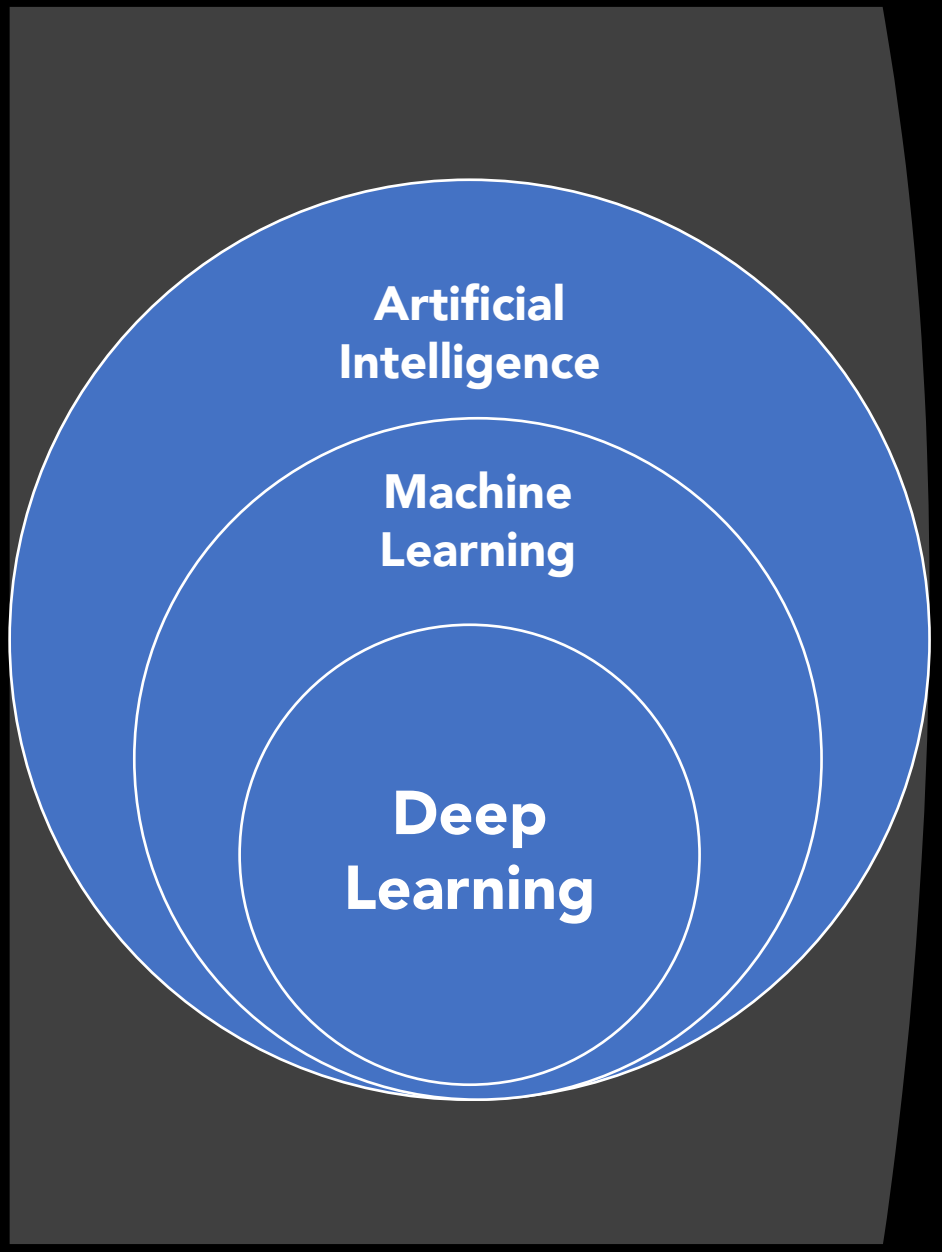
Week 5

Buchanan Fellowship, Spring 2020

Vanderbilt University Library

Machine Learning

What do we really mean?



Artificial Intelligence: any technique that enables computers to mimic human intelligence – to sense, reason, act and adapt

Machine Learning: AI technique whose performance improves as exposed to more data

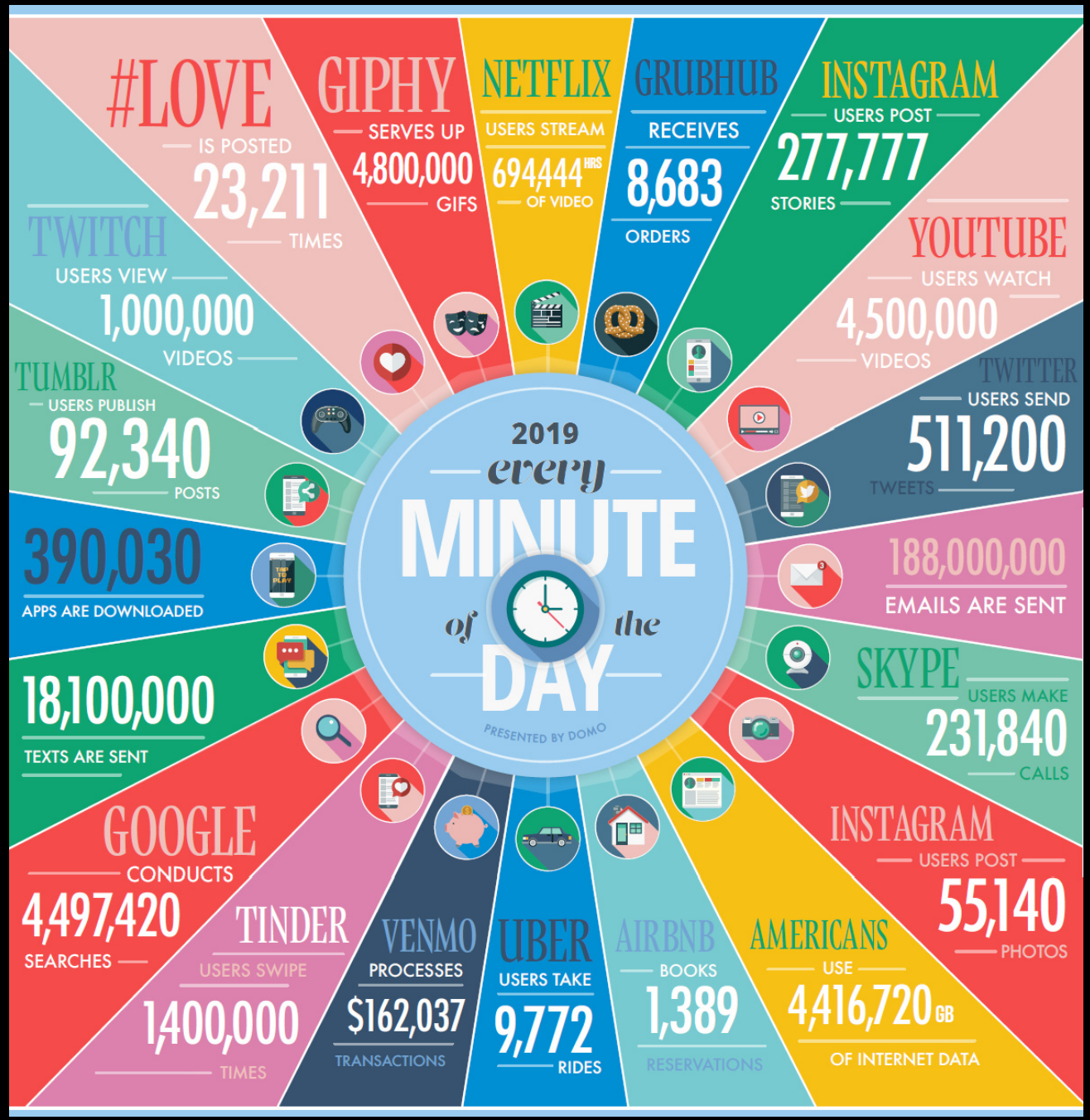
Deep Learning: ML technique based on multilayered neural networks reliant on vast amounts of data

Weak and Strong AI

- Weak AI – built for a specific purpose
  - Spam Detection
  - Text extraction for depositing checks
  - Reading handwriting of a specific person
- Strong AI – generalizable
  - Reading any handwriting
  - Still mostly science fiction

Recent Explosion

- Improved parallel processing makes computing faster and cheaper
- Minimal storage costs makes it easy to store vast amounts of data
- Big Data accumulation – text, transactions, location data, images
- Internet of Things (IoT) and sensors everywhere sometimes mean data accumulated just for the sake of accumulating
- 2012 breakthrough with neural networks



https://www.visualcapitalist.com/big-data-keeps-getting-bigger/

Machine Learning Basics

- Machine learning needs data, deep learning needs even more

- Give the computer lots of data, identify patterns, then predict outcomes based on previously learned patterns
- It’s about statistics not rules

Supervised vs. Unsupervised Learning

Supervised

- Classification algorithm
- Needs a training set (human labelers)
- Classifies new data according to the predetermined labels by finding patterns

Unsupervised

- Clustering algorithm
- No training set, machine ‘learns’ on its own
- Clusters data according to features it finds
- May uncover hidden relationships

Different Machine Learning Models

- K-Nearest Neighbors
- Linear Models
- Naïve Bayes Classifier
- Principle Components Analysis
- Decision Trees
- Random Forests
- Gradient Boosting Decision Trees
- Support Vector Machines
- Neural Networks

Which model?

- What size is your data?
- What type of data do you have?
- What do you want to do with it?
- How long do you have to train?
- How explainable / trustworthy does the model need to be?

Neural Networks Explained

- https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

- Don’t worry about the match, focus on the big picture

Alternative Intro to ML

- Visual intro to modeling housing prices

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

- If you want more math:

http://neuralnetworksanddeeplearning.com/index.html