Background for Machine Learning

AnHai Doan

Why?

- **To solve problems in data exploration, cleaning, integration**
  - we often have to use multiple techniques
    - database, ML, big data scaling, effective user interaction, cloud
  - ML will play a big part
- **So we need to cover ML, but only how to use it, not how it works**
  - this is not a ML course

Important Concepts from ML Courses

- **Three main branches**
  - supervised learning (classification): traditional + deep learning (DL)
  - unsupervised learning (clustering)
  - reinforcement learning
  - we will mostly focus on classification, will also occasionally use clustering
- **Classification**
  - a training set U of pairs <example,label>
  - each example is a feature vector
  - train a model M on U
  - apply M to a test set V of new examples
- **Key concepts to know in classification**
  - example, label, features, feature vectors
  - training, testing, ML model, trained model
  - search for the model that best fits the training data
  - popular models: decision tree, random forest

Using Classification in Practice

- **Given a problem X, first decide if X can be solved using classification**
- **If yes, then create training data**
  - generate a set of examples E
  - for each example in E, convert it into a feature vector
    - to do this, must define a set of features F
  - for each example in E, assign a correct label
    - must define a set of labels
- **Train a classifier M on the training data**
- **Apply M**
  - generate new examples, apply M to them
- **The aboves form a ML pipeline**
  - draw the picture here
- **Often this pipeline need to be augmented with preprocessing and postprocessing steps**
  - which often use rules
- **The final pipeline often has both rules and ML models**

How to Develop a Good Pipeline?

- **Show with two examples**
  - person name extraction from text documents
  - classify a column in a table as “age” or not
- **Key concepts**
  - development stage: create a good pipeline, often using data samples
  - production stage: deploy the pipeline on the entirety of data

Clustering

- Motivation for why we need to use this
- Many clustering techniques exist

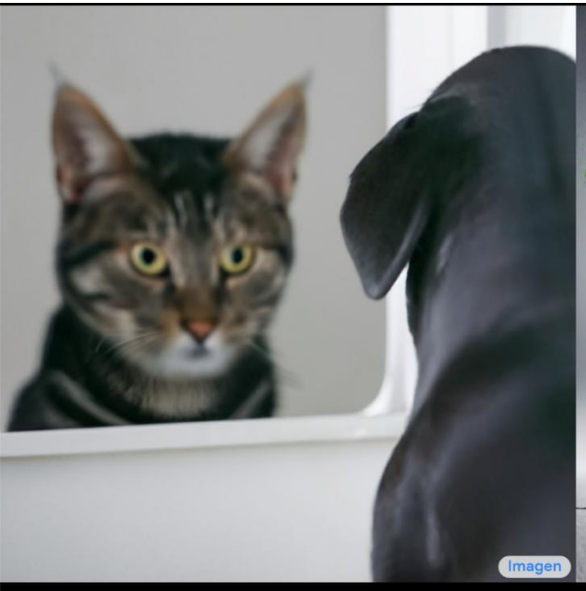
Deep Learning

- **Latest sexiest model: deep learning (DL) models**
- **Important ideas to know**
  - different ML models have different levels of expressive power (can be very simple or very complex)
    - e.g., suppose we classify each people as “likely to repay a loan” (yes/no), if a model can only have rules that involve a single feature, the model has very little expressive power
    - if a model can have rules that involve up to two features, it has more expressive power,
    - and so on
  - models with more expressive power can be more accurate
    - surely if we can write rules with two features, we can classify people more accurately
  - but models with more expressive powers need more training data
    - or else it will just “remember” the training data instances, and can fit training data very well, but has poor generalization power ➔ overfitting the training data
  - DL models are the ones with the most expressive power that we know today
  - but they need tons of training data to work well

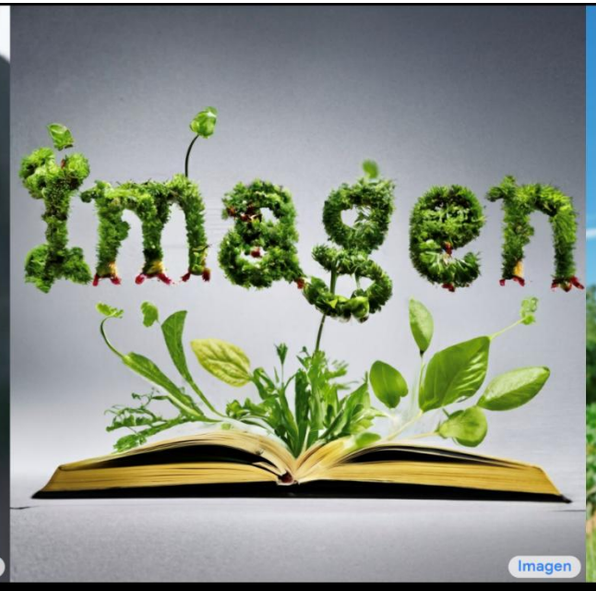
An Extremely Brief History of DL

- **Neural network models were invented long time ago, in the 1940s**
- **For a long time, they were ignored**
  - not deep (that is, big) enough to have enough expressive power to do interesting things
  - if model is big, don't know how to train fast (no known algorithm)
  - no hardware that can help train fast (no GPU for a long time)
  - not enough labeled data to demonstrate its power
- **Around 2012, the above problems were solved, and a clear demonstration in computer vision showed the dramatic power of DL models**
  - things exploded from there
- **A few years later, generative DL became popular**
  - traditional classification: new example ➔ label
  - in generative DL, label can be more general: a sentence, a textual answer, an image, etc.
    - given an input, generate an output
- **And now, foundation models are the latest craze**

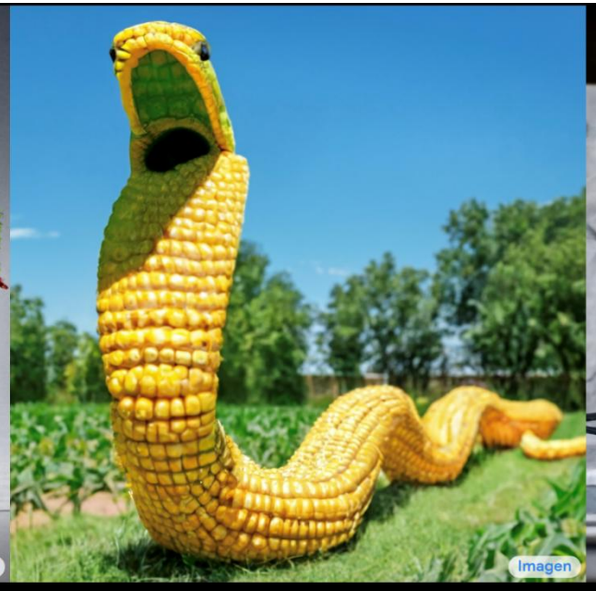
Example of Generative DL



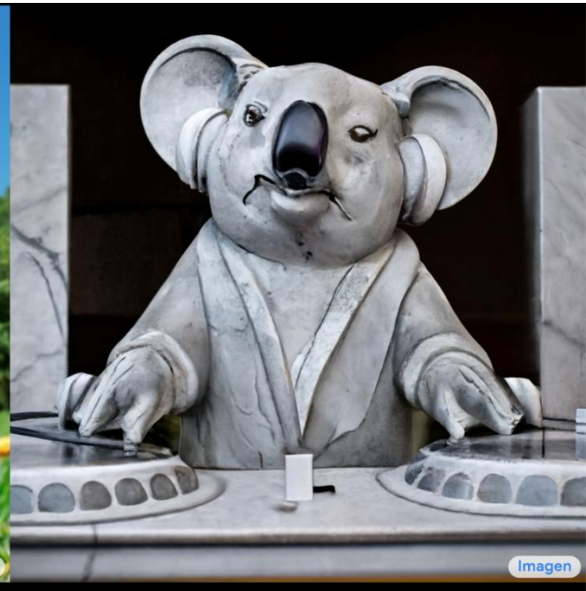
A dog looking curiously in the mirror, seeing a cat.



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.



A giant cobra snake on a farm. The snake is made out of corn.



A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala has wearing large marble headphones.

- **And ChatGPT of course**

Who is AnHai Doan?

AnHai Doan is a computer science researcher and professor at the University of Wisconsin-Madison. He is a leading expert in the field of data management and has published numerous papers and articles on topics such as data integration, data cleaning, data quality, and data provenance. He is a well-known researcher in the database community and has received several awards for his contributions to the field. His research focuses on developing techniques and systems to make data integration and cleaning more effective and efficient. His team develop system like DataWrangler, Data Cleaner, and Data Profiler etc. He is also known for his contributions to the field of data management in general, including work on data integration, data cleaning, data quality, and data provenance.

Foundation Models

- **Big complex DL models**
  - hundreds of millions or billions of weights
- **Trained on tens of billions of documents**
- **Can be trained further (fine tuning) for a wide variety of tasks**
- **We will touch upon them later in the course**

Others

- **Python**
  - pandas, Jupyter notebook
  - why Python? Lot of data science stuff
- **Big data scaling, eg Spark**
  - you don’t have to know this, we will introduce the minimum if necessary
- **Other topics will be introduced as they arise**
  - e.g., keyword search