The AI Data Revolution
Doing More with Less Data Labeling

David C. Anastasiu
Assistant Professor
San Jose State University
(Starting Fall 2019, Santa Clara University)
According to Miriam-Webster:

1. A branch of computer science dealing with the simulation of intelligent behavior in computers.

2. The capability of a machine to imitate intelligent human behavior.

What is AI?
Types of AI

Narrow
- Pattern Recognition
- Prediction

General
- Can think like a human

Augmented
- Can help a human think
Why Care About AI?

• There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies

• New mantra
  • Gather whatever data you can whenever and wherever possible.

• Machine learning is becoming ubiquitous in society
  • Here’s a movie you might like (recommender systems)
  • You should try this product (advertising)
  • Found a shortcut that will save you 10 minutes (maps)
  • Jessica is a friend suggestion for you (social media)
  • Caution! Vehicle approaching in right lane (drive assist / self-driving cars)
Machine Learning Tasks

Data

<table>
<thead>
<tr>
<th>#</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>80K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Supervised Learning

Learn a function

\[ Y = f(X) + \epsilon \]

Fundamental Assumption of Learning:

- The distribution of training examples is identical to the distribution of test examples (including future unseen examples).
- Training examples must be sufficiently representative of (future) test data.
General Supervision

- Time series analysis problem
- Data collected during a sensory challenge protocol (SCP) in which the reactions to eight stimuli were observed.
- Based on electrocardiogram (ECG) and skin conductance.
- Multivariate time series /w 2M+ samples for each subject.

Example:
Autism Spectrum Prediction

- /w Dr. Megan C. Chang
- Student: Manika Kapoor
Deep Learning

• A subfield of machine learning.
• Uses artificial neural networks (ANNs) with many layers for pattern discovery.
• Building block: Perceptron
  • $\mathbf{w} \cdot \mathbf{x} + b > 0$
• DNNs can approximate infinite functions
• Needs sufficient labeled input
  • Avoid overfitting

Credit: Figure adapted from Yu et al., Artificial Intelligence in healthcare, Nature Biomedical Engineering, VOL 2, OCTOBER 2018, 719–731, https://www.nature.com/articles/s41551-018-0305-z.pdf
What Made Deep Learning Possible

**Many Core Hardware**

- GPUs, TPUs
- Thousands of cores, really good at dense matrix operations.

**Distributed Computing**

- Supercomputing
- Shared-nothing Computing
- Split the work among many systems.

**Lots of Labelled Data**

- Mechanical Turk
- Label Generation
- Split the work among many humans, or be clever about creating labels.

David C. Anastasiu The AI Data Revolution
Deep Learning

- Convolutional Neural Network (CNN)
- SegNet Architecture
- Compute the probability of each pixel belonging to a polyp

Example: **Real-time detection of polyps during colonoscopy using machine learning.**

Example:
Kidney Health Monitoring

/w Dr. Alessandro Bellofiore
Students: Rathna Ramesh, Ragwa Elsayed

Borrowed Supervision
(Transfer Learning)

Localization
Transfer Learning
Deep Learning – YOLO
Alternatives

Feature Extraction
Color-based features:
RGB, Color Histogram, Gradient Histogram

Prediction
Creatinine level (regression)
Kidney health level (classification)
General/Borrowed Supervision

2017 AI City Challenge
- Collaborative annotation
- Over 150,000 annotations from 80 h video
- Localization and classification

2018 AI City Challenge
- GPS-based annotation
- 27 videos, 3 locations
- Speed estimation, anomaly detection, re-identification and tracking

Semi/Aided Supervision

David C. Anastasiu

Charles MacKay

Thomas Tang & team - UW
2019 AI City Challenge
Multi-car multi-video tracking
- Applications in corridor-level traffic flow optimization
- Videos from 40 cameras within 6 km² of a city (3 hours)
- Hybrid models (single-camera tracking, deep-learning based Re-ID, camera linking)

Vehicle Re-ID
- Applications in vehicle counting, traffic volume estimation
- 56K vehicle images, 1K queries
- Visual features from CNNs + semantic features from travelling direction and vehicle type classification
2019 AI City Challenge

Multi-car multi-video tracking
○ Applications in corridor-level traffic flow optimization
○ Videos from 40 cameras within 6 km² of a city (3 hours)
○ Hybrid models (single-camera tracking, deep-learning based Re-ID, camera linking)

Vehicle Re-ID
○ Applications in vehicle counting, traffic volume estimation
○ 56K vehicle images, 1K queries
○ Visual features from CNNs + semantic features from travelling direction and vehicle type classification
2019 AI City Challenge
Traffic Anomaly Detection

- Speed up emergency response
- 50 h of highway videos from Iowa
- Foreground segmentation + spatio-temporal anomaly detection

Challenge Stats

- 334 teams (1,004 researchers) signed up
- Competitive evaluation
- 129 track submissions (1,337 solution submissions) from 96 teams (355 researchers)

CVPR 2019 Workshop

Check out AI City Challenge 2020

David C. Anastasiu
2019 AI City Challenge
Traffic Anomaly Detection

- Speed up emergency response
- 50 h of highway videos from Iowa
- Foreground segmentation + spatio-temporal anomaly detection

Challenge Stats

- 334 teams (1,004 researchers) signed up
- Competitive evaluation
- 129 track submissions (1,337 solution submissions) from 96 teams (355 researchers)
- CVPR 2019 Workshop

Check out AI City Challenge 2020
**Generated Supervision**

**CGI-based Labelling**
- Blender-generated objects w/ motion
- Bounding box automatically generated
- GAN-based smoothing to improve blend

**Anomaly Detection in Expense Reports**
- Receipt localization dataset generation
- Small representative set of receipts
- GAN-based model for background
- Generate millions of receipts & combinations
Guided Supervision

- Example Weak Supervision Sources
- Technical Challenge: Integrating & Modeling Diverse Sources
- Use Weak Supervision to Train End Model

○ Weak Supervision Work by Alex Ratner & Chris Ré at Stanford
○ Provide inexact rules covering many (but not all) cases
○ Learn how to combine weak predictors into a strong one

\[ \hat{\theta} = \text{argmin}_\theta \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, f_\theta(x))] \]
Taught Supervision

https://youtu.be/jwSbzNHGfIM?t=28
**Few Labels?**

No worries...

<table>
<thead>
<tr>
<th>Label</th>
<th>Borrow</th>
<th>Prioritize</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd-Source Label Generation</td>
<td>Transfer Learning</td>
<td>Semi-Supervised Learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Generate</th>
<th>Guide</th>
<th>Teach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data + Label Generation</td>
<td>Weak Supervision</td>
<td>Reinforcement Learning</td>
</tr>
</tbody>
</table>
Challenges of Deep Learning

Data Driven Behaviour
- GPUs, TPUs
- What is the path that will be taken at inference time?

Input Bias
- (Un)intended consequences
- Who’s being left out?

XAI
- Explainable AI
- Understand why the model made the decision
Explainable AI

- Critical requirement for AI in many contexts beyond healthcare
  - DARPA/DoD priority
  - Already required in the General Data Protection Right (GDPR)

- RISE: randomized input sampling for explanation of black-box models

https://www.darpa.mil/program/explainable-artificial-intelligence
Deep Learning

• Convolutional Neural Network (CNN)

• Improved prediction of cardiovascular risk factors

• Predicted side-factors:
  • Age, gender, smoking, diabetic, BMI

• “Soft attention” used to point out the salient pixels

Example: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Credit: Figures adapted from Poplin1 et al., Nature Biomedical Engineering, volume 2, pages 158–164 (2018) https://www.nature.com/articles/s41551-018-0195-0.pdf
Conclusions

- AI is here to stay (this time)
- Will permeate CS/Engineering (and many other fields)
- Learn machine learning
- Or partner with a data scientist