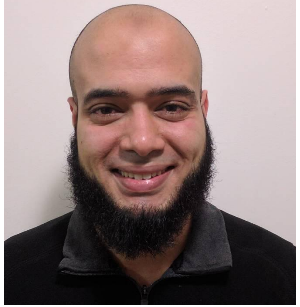
Machine Learning
Data in Practice

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Note

- This is a **high-level** lecture
- Most of the topics can be either a lengthy lecture or experimentations
- In the next of your ML journey, you can pick and explore!

Quotes!

- *"Data is the new oil." – Clive Humby*
- *"Without data, you're just another person with an opinion." – W. Edwards Deming*
- *"Garbage in, garbage out."*
- *"Data are just summaries of thousands of stories – tell a few of those stories to help make the data meaningful." – Chip & Dan Heath*
- *"Data matures like wine, applications like fish." – James Governor*
- *"Not everything that can be counted counts, and not everything that counts can be counted." – Albert Einstein*
- *"Data really powers everything that we do." – Jeff Weiner*

Data is the new oil

- Most of the recent machine learning success is about:
  - Huge amount of data
  - **Supervised learning**
  - Deep learning
  - Smart tricks in Modeling and Representation Learning
- For the data part!
  - We need to **collect** data
  - We need **store** data
  - We need to **annotate** data: during and after collection
  - We need to **process** data efficiently
  - We need to **track** our experimentations on different datasets

Data Types

- Structured data
  - numbers (discrete, continuous) and labels
  - We process into **tabular data** (rows examples and cols features)
    - Example: predict house price from its features
  - Very common in real life and kaggle competitions
  - **Boosting** techniques are common and sometimes deep learning
- Semi/Unstructured data
  - Text (e.g. emails, tweets, articles)
  - Images and Videos
  - Audio
  - Area where **deep learning** spiked in performance

Labeling Types

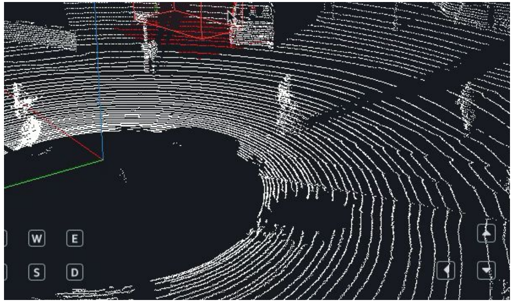
- **Natural Labeling**: use natural feedback (estimated trip time / predicted stock)
  - Short feedback (tweets reactions) vs long feedback (detect fraud later)
- **Hand** Labeling: Human annotators manually label the data
  - **Crowdsourced** Labeling: Platforms like **Amazon's Mechanical Turk** allow for distributing the labeling task among a large number of human annotators online.
  - Cost-effective but quality challenges to mitigate
- **Automated** Labeling: Use available models to label the data (common tasks
  - Aka **data** distillation. Results might be inaccurate / Domain gaps in the data
- **Weak Supervision** Labeling: Find/build cheap noisy/coarse labels
  - Rule-based from text, meta data, etc
- **Semi-supervised Labeling**: Train on subset to annotate others
- **Active Learning Labeling**: Select next subset to label

Data & Labels

- Think about data **size**: small, medium to large: The larger the better
- Think about data **quality**
  - Covering diverse cases or just redundant data!
    - Redundant data is misleading as it doesn’t add value!
    - You may think adding data doesn’t help. But you have the wrong data
- Think about labels **depth**
  - Fine-grained vs coarse grained. Diverse or narrow
- Think about labels **quality**
  - Noise or clean - Can we spot mistakes or cheating early?
  - Can we create **objective/consistent definitions** for the annotation process?
- Think about labels **nature**: fully annotated, semi-annotated, weakly-annotated
- Think about labelers: In-house, crowdsourcing or outsourced

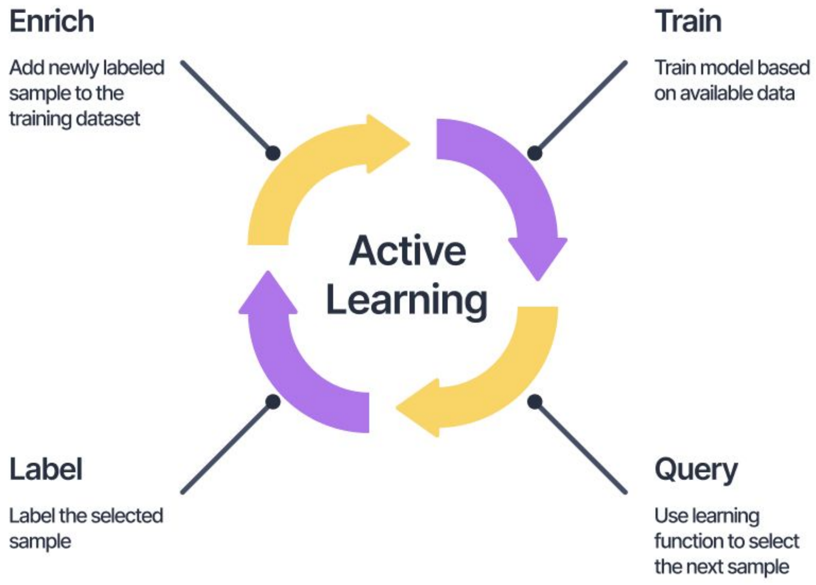
Hand Labeling

- It can be time consuming and expensive! Examples:
  - Diagnosing diseases from X-rays, MRI scans, or other medical images
    - You need an expert! Not just crowdsourcing
  - Semantic Segmentation in Autonomous Vehicles
    - Every pixel need to be labeled
    - Boundaries are challenging!
  - 3D Point Cloud Labeling (Lidar)
- You need
  - To review to fix mistakes
  - To handle people cheating (multiple annotators?)
  - To repeat with many for subjectivity?
    - Is this a spam email? Ask 10 persons!
    - Describe this image
      - How to handle disagreements?!
        - Your instructions also can be vague!



Active Learning Labeling

- **Goal**: Minimize the number of labeled examples ⇒ less cost
- Active learning involves **iteratively** labeling the data instances that the model finds **most confusing**
- Annotators only annotate the most **informative** subset
  - Hence overall **avoid annotation** examples that won't help the model
- Participated in this [paper](): Active learning for structured prediction from partially labeled data



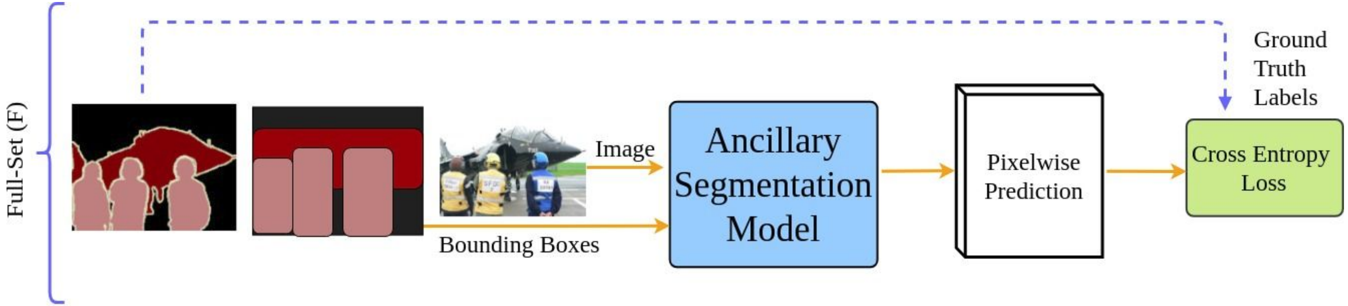
[Img src]()

Human-in-the-Loop (HITL)

- HITL also involve **human interaction** but broader than data labeling
- Humans may provide guidance, correct mistakes, or interpret complex data
- Humans may interact with the system at **various stages**, including data preprocessing, **model training**, evaluation, or even at inference time
    - Or gather feedback from the users and build insights on mistakes
- Humans here are experts: domain experts / ML experts
- So while both active learning and HITL involves human, they are different
- Surveys
    - Human-in-the-loop machine learning: a state of the art
    - A SURVEY OF HUMAN-IN-THE-LOOP FOR MACHINE LEARNING

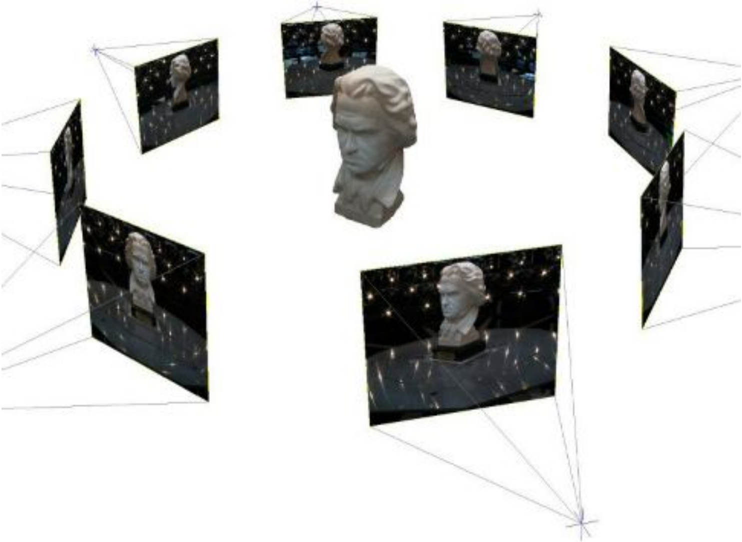
Semi-supervised Labeling (Learning)

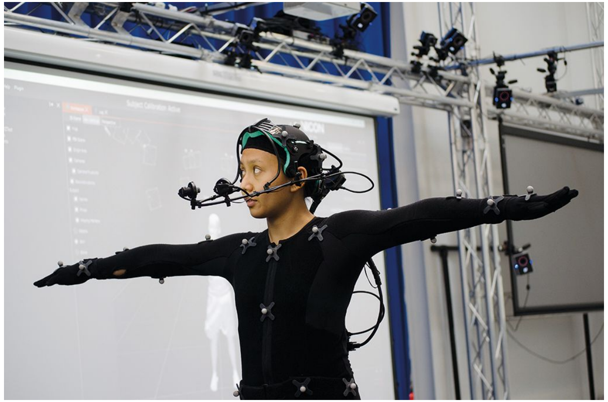
- Start with a **Small** Labeled Dataset ⇒ Build initial model
- Collect more data (same distribution)
  - Label them with the previous model
    - You may filter low-confidence results
  - Create **Pseudo**-Labels from the unlabeled data (a bit noisy)
- Train a new model with labeled and pseudo labeled data
- You may repeat the process



Labeling in 3D Computer Vision

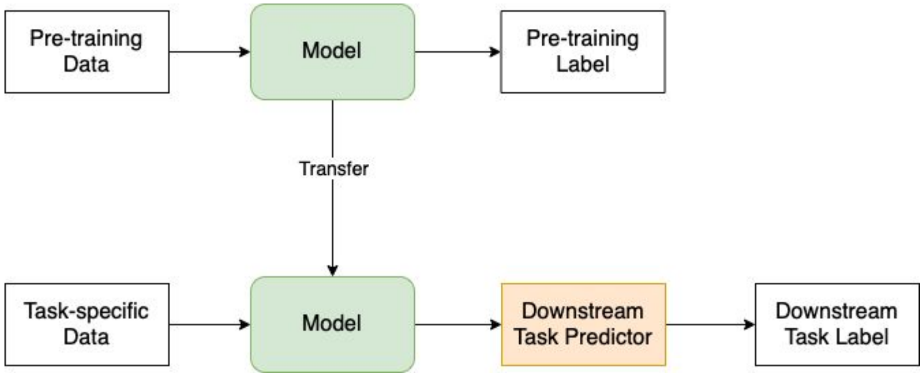
- Motion Capture Suits
- Multi-view camera: 2D - 3D View





Transfer Learning

- Popular technique that helped many small-datasets (in vision, nlp) to perform strongly. Main successful application is deep learning
- First train a deep learning on a close task with a lot of labeled data
  - General image classifier trained on 1 million images (cars, animals, etc)
- Now, resume the training (**fine-tuning**) on a new task with less data
  - Your work dataset has 5000 examples for 7 categories
  - Called **downstream** task
- Another path: **Feature Extraction**
  - Extract representations
  - Train another model!
- Popular models: BERT / ResNet



Img src

Data Collection

- There are many **public datasets** that can be utilized for POC
  - Be careful from licence / Don't use for production models
- **Web scraping**
  - Crawl web pages to get data (images, text, videos) and their metadata
  - Issue: websites legal limitations
- User-generated data
  - This is the most common format. Your business may collect data (social media, banks, mobile apps, health care, customer reviews, discussion boards, surveys)
  - We may aggregate information from multiple sources
- Sensor-data
  - Temperature (weather forecast), humidity, motion (fitness in gem), camera, etc
- Synthetic data (e.g. body / hand-pose data)
- *Launch a product without ML and collect data*

Data Collection on Large Scale

- Sometimes, we have to collect terabyte/petabyte of data (e.g. AV)
- However, there are many challenges:
  - In a fine grained data collection, we get more and more cases and scenarios to define!
    - Data variance covering all needed
  - Data Quality and Consistency (across various **sources**)
  - Annotation Quality (how much noise/mistakes?
  - Scalability: to collect, store, process and annotate
    - Time and Cost
  - Legal and Regulatory Compliance
  - **Big troubles:**
    - **Recollect** missing scenarios
    - **Relabel** for new approaches or we decided to move from N classes to M classes
    - **Data Drift**: Does data nature changes over time?

Data Collection: Stages

- Assume we want to solve a computer vision task for automotive industry
    - For example, hands-on-wheel task
- Discuss how we reach close-to-perfect model

Stage #1

- We must start on a **small** scale
- Annotate the data by ourselves for hand labeling
    - This helps us really know **what we need**
- Use the data to build your model.
- Get insights on what you need
    - Model changes
    - Data diversity you need
- Tip: you may develop as a POC (proof-of-concept)
    - Get cycle fast. Decide if it worth continue or not
    - If yes, make your code more production-ready
- Meta-data: it is important to save all relevant information
    - Time, Location, Car Info, Cameras Info/Calibration, angle, user IDs, etc

Stage #2

- Time to **scale**
- Hire specialists to collect data for you
  - Clear requirements. Gradual to early verify and fix
- Explore the different ways to **annotate** the data (natural, automatic, etc)
  - You can add extra sensors to help annotation (e.g. multi-camera), but we won't use in production

Stage #3

- Time for deployment
- For example like a tesla car that someone buys
  - It can’t have extra devices
- We get user agreement to collect data
  - Privacy issues
- This is the hardest in annotation and largest in scale

Data Collection - Behind the scenes

- ML team design a **script**: describing the flow of the data collection
  - User enter car, Open window, Put hands on the wheel for 10 seconds in this position
  - You must be clear on requirements
- Some team may develop DC (data collection) tools or setup
  - ML team guides the DC team based on their needs to build the required tool
- Common variables
  - List all variations we need (e.g. 50% men, 20% age 20-35, lighting conditions, etc)
  - List all behaviour based on the task (e.g. instructions for hands on wheel, for gaze estimation, etc)
- A lot of early, regular and last reviews to fix issues as early as possible

Data Verification

- A process to review the **accuracy** (represents what we expect), **consistency** (no conflicts between data sources / integrity), and **quality** of data
- **Format** Verification: such as date formats, phone numbers, email addresses
- **Completeness** Check: amount of missing attributes
- **Uniqueness** Check: any duplicates?
- **Range** Check: check ranges of specific features
- **Compliance** Check, regulatory or policy requirements (e.g. in healthcare, finance)

Data Lineage

- In scenarios with multiple sources of data, we may need **data lineage**
- Data lineage refers to the **tracking** of the flow and transformation of data from a storage/source to another
- Why?
  - Helps identifying potential **quality issues** or errors
  - Data **governance** (is everything you do to ensure data is secure, private, accurate, available, and usable)
  - **Compliance** and Auditing

From Data to Modeling

- There are several factors that affect our model choice
  - Nature of the problem itself (e.g. classification, recommendation, temporal data, interpretable model, domain experts, calibration-sensitive apps, safety critical, real time, etc)
  - Data: quantity, labelts, etc
  - Constraints: computational resources
- In the future, you should develop better sense on how data affects modeling
  - Structured vs unstructured data
  - Semi-supervised, weakly surprised model and fully supervised model, Unsupervised Learning
  - Only a small amount of labeled data ⇒ Few-shot learning
  - High-Dimensional Data / Imbalanced Data
  - Online Learning / Transfer learning
  - Multimodal Data (text, image, audio)

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."