Data Centric AI and Software 2.0

Jayanthi Suryanarayana

Bio

Name : Jayanthi Suryanarayana

https://www.linkedin.com/in/jayanthi-suryanarayana-2313841/

Electronics engineer by training, pivoted to software engineering. Understand the nuances of app, data and ML engineering. Expertise in Data and AI strategy, ML Engineering and Synthetic Data. Curious and constantly learning in this fast evolving AI space.

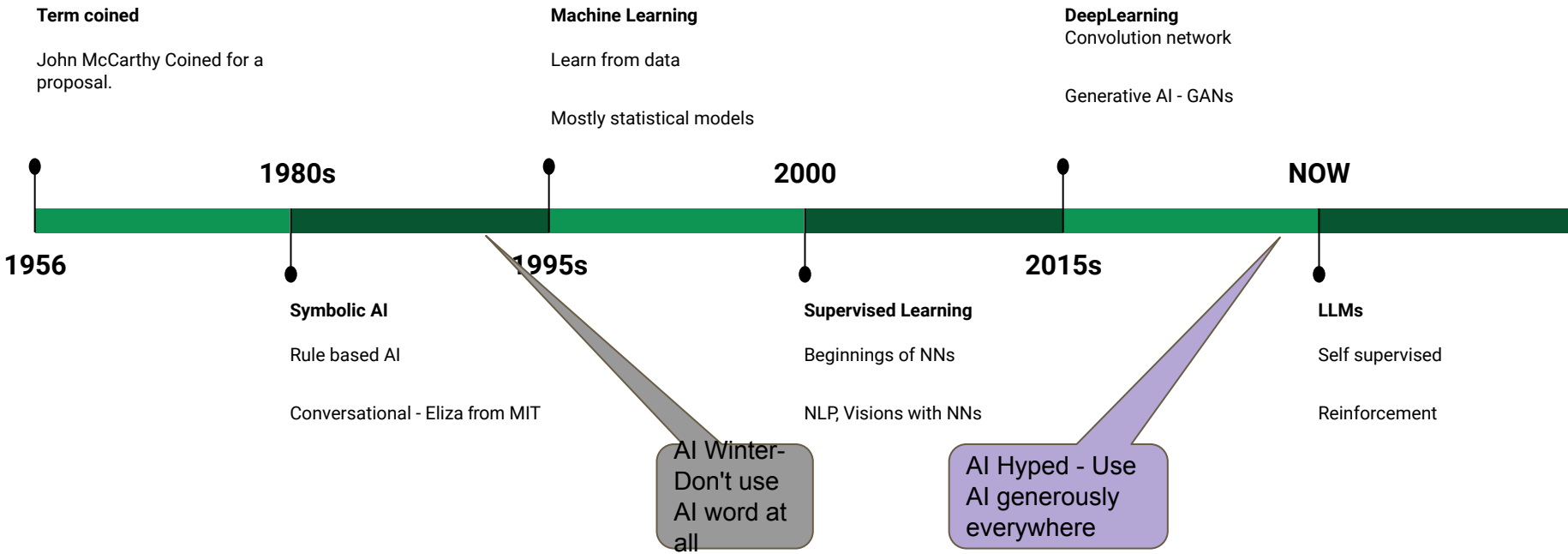
Built data platforms for enterprise, Passionate about the use of synthetic data to enable people who work with data (analysts, engineers, scientists anyone who wants to get value out of data)

Outside of work:
Enjoy time with family, cook and do yoga

Agenda

- What is the context - AI,Software 2.0, Data centric AI
- Motivation for Data Centric AI
- Framework - techniques for data centric AI
- Data centric AI in enterprise data management
- Takeaways
- Resources

AI - and its implied meaning - a timeline view



Software 2.0

Andrej Karpathy:

www.youtube.com/watch?v=y57wwucbXR8

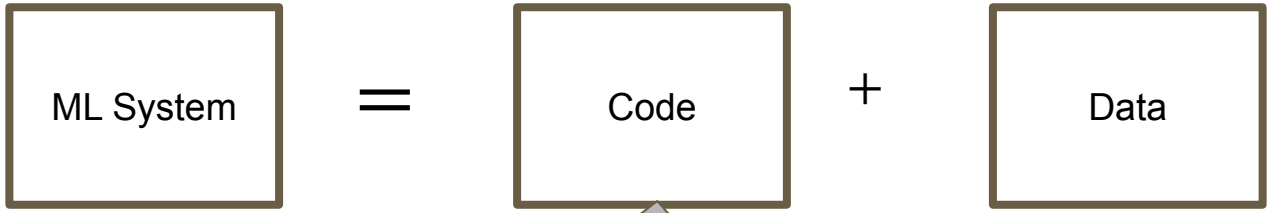
https://karpathy.medium.com/software-2-0-a64152b37c35

Software 1.0 : Explicit - Code only

Software 2.0: Abstract - Code + Data (Machine Learning mostly deep learning)

most of the active "software development" takes the form of curating, growing, massaging and cleaning labeled datasets. This is fundamentally altering the programming paradigm

Machine Learning Systems - Software 2.0



Model Centric AI:
Keep data constant
Iterate on Model Selection and hyperparameters tuning

Data Centric AI

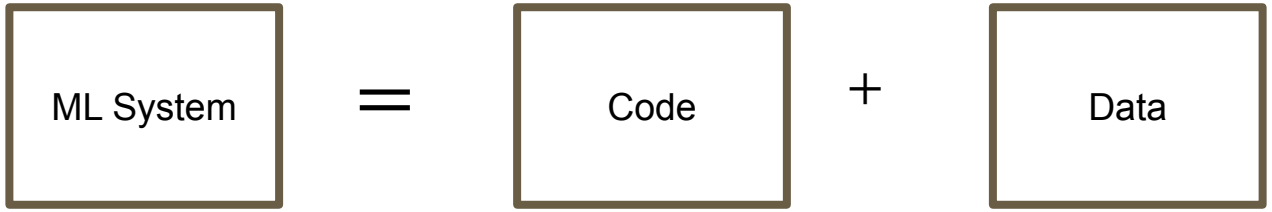
Andrew Ng:

Momentum Since couple of years

Data-Centric AI is the discipline of systematically engineering the data used to build an AI system.

https://landing.ai/data-centric-ai/#:~:text=Data%2DCentric%20AI%20is%20the,on%20data%20instead%20of%20code.

Data Centric AI - Software 2.0



Data Centric AI:

- ❑ **Systematic approach** understand/update data to improve ML System performance
- ❑ Use machine learning approaches, techniques

Data Ingredient quality - How to think about

Label errors:

https://labelerrors.com/

Label errors are prevalent (3.4%) across benchmark ML test sets.

Feature Errors:

Acquisition, transformations , systematic, non systematic etc. etc.

Motivation for Data Centric AI

- You can do only so much with model centric approach to improve performance
- Real world datasets are messy compared to research/academic setting.
- When put in production, model does not perform well as it did during the development.
- Data scientists spend 70-80% of the time in data preparation in adhoc ways
- Poor Data quality - loss in revenue and reputation for enterprises

How are we managing now? - Ad Hoc ways, non systematic

Go buy more data

Fix in pipelines with special and custom transforms

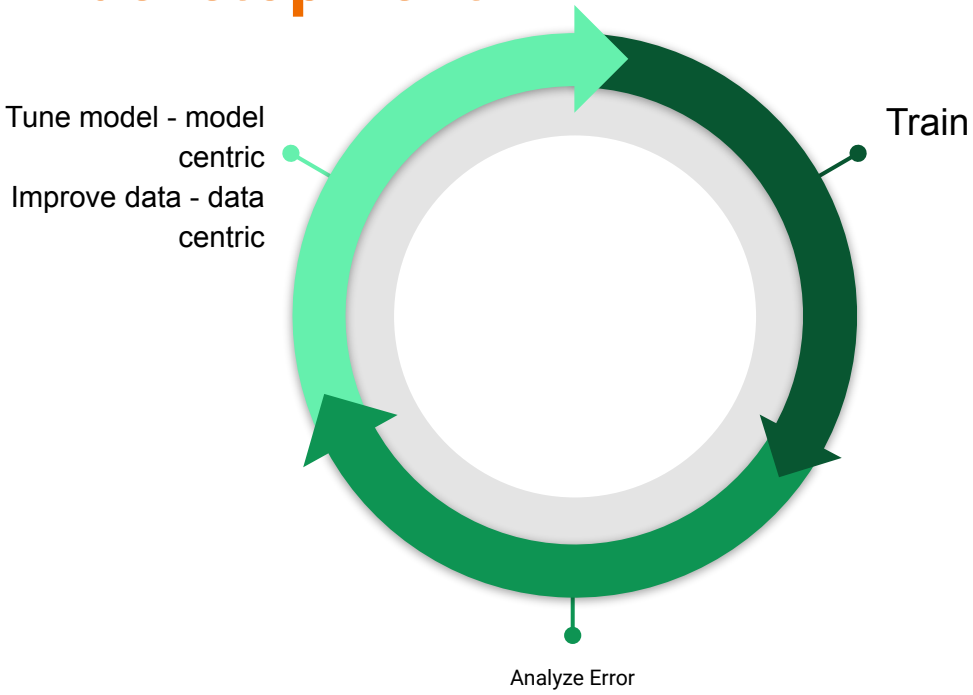
Labelling processes - not integrated , not at platform level, deal with inconsistencies inconsistently

Data set selection - Manual

Team composition and tasks - project based

MLOps in different levels of maturity

Iterative ML development



Model Centric vs Data Centric Techniques

Model Centric

Change models

Tune Hyper parameters

Regularization

Loss Function

Data centric

Data Augmentation

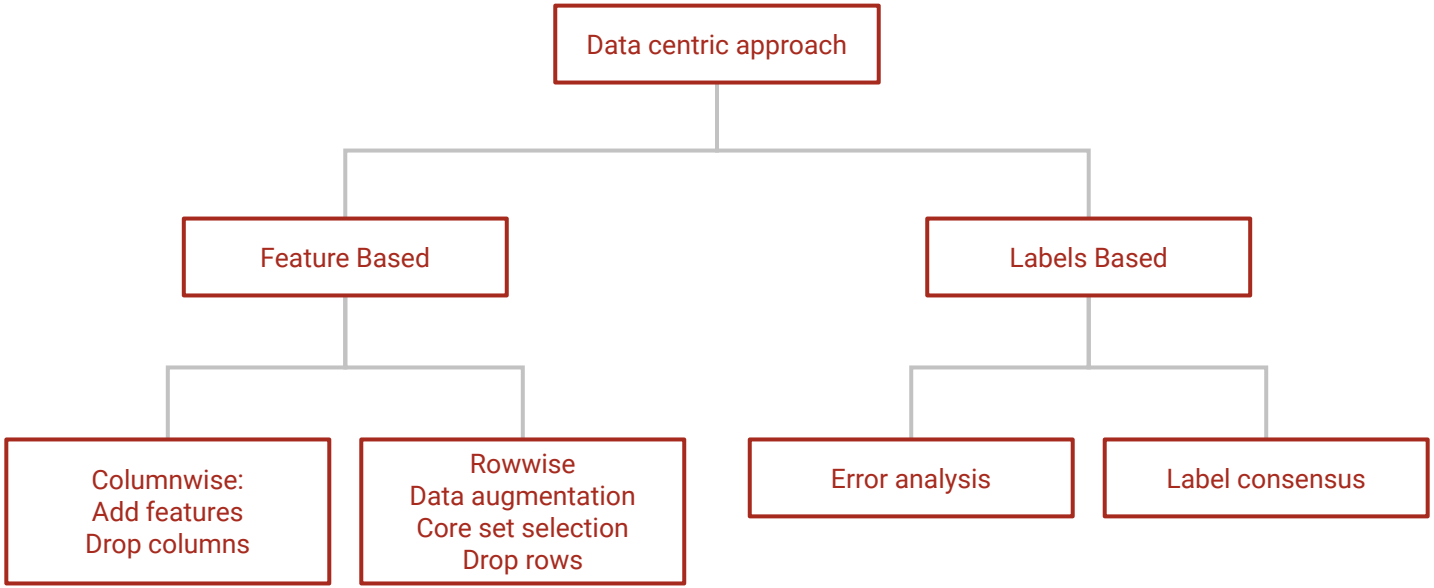
Feature engineering

Label consensus

Active Learning

Traditional : Outlier detection, error detection in data (missing value handling etc.)

Framework for data centric techniques



Non consumer AI - Data limitations

Consumer space - Billions of rows of data

Problem becomes how to pick most valuable data set relevant for the mlapp.

Non consumer space - Orders of 1000s of rows

Problem becomes how to grow data

Enterprise data management

Enterprise data management capabilities

Data Governance

Data Quality Management

Metadata Management

Master Data Management (MDM)

Data Lineage

Data Reconciliation

Data Forensics

Data Certification

Data Discovery

Business Intelligence

Data quality - Enterprise data

Poor Data quality costs not only revenue loss but destroys reputation of companies

Enterprises continue to invest in data quality as part of data management capabilities

Data centric AI can fit right in there and can be integrated as part of data strategy

Take aways

It is a change in mindset - iterative and systematic examination of data as part of the development process

Model centric and data centric approaches co-exists

Tool ecosystem in this space is growing.

Resources

Courses:

https://dcai.csail.mit.edu/

Weblinks:

https://www.youtube.com/watch?v=Yqj7Kyjznh4&t=1966s

https://www.youtube.com/watch?v=06-AZXmwHjo&t=702s

Literature Survey:

https://arxiv.org/abs/2303.10158

Github:

https://github.com/daochenzha/data-centric-AI