**Specific for AI monitoring / problems in a clinical field**

*AutoDC: Automated data-centric processing*

* Introduced an automated data-centric tool to speed up data improvement processes
* Utilized image embeddings and outlier detection by Isolation Forest
* Workflow includes label correction (human-in-the-loop), edge case selection (human / Isolation Forest), data augmentation (mainly for the edge cases)
* Users saved 80% of the time compared to fully manual process, model accuracy boosted to 10-15% in three test image datasets

*Combining Data-driven Supervision with Human-in-the-loop Feedback for Entity Resolution*

* Discussed a human-in-the-loop and data-centric solution to closing the training-production performance divergence. Tested on entity resolution system (language model)
* Includes system development with several versions due to feedback and improvement
* Human evaluation and feedback: group the errors into different types
* Data augmentation: Built with specifying some error patterns
* Data-derived rules: Useful for solving some error patterns, and could even be prioritized over decision made by database learning

*Highly Efficient Representation and Active Learning Framework and Its Application to Imbalanced Medical Image Classification*

* Proposed a framework combining an unsupervised representation learning of a CNN and a Gaussian Process (GP) method which produces prediction and uncertainty in one shot
* Representation learning: maximizes similarity of augmented (from the same image) image pairs and leverages a contrastive loss
* GP classifier: non-parametric Bayesian method, uses contrastive learning as input to a GP with RBF kernel, prediction and uncertainty in one shot since it’s Bayesian
* ResNet-50 for CNN and Sparse Variational GP from GPflow package for GP classifier
* Only <10% labeled data is needed to reach accuracy from training all available data. More samples from rare classes are automatically selected in early active learning cycles

Self-supervised Semi-supervised Learning for Data Labeling and Quality Evaluation

nferX: a case study on data-centric NLP in biomedicine

**Generic Data-centric**

*A Data-Centric Behavioral Machine Learning Platform to Reduce Health Inequalities*

* Developed a data-centric ML platform featuring a software development kit (SDK)
* The SDK unifies behavioral data labeling and ensures data format consistency

*CogALex 2.0: Impact of Data Quality on Lexical-Semantic Relation Prediction*

* Proposed a manually improved multilingual dataset for lexical-semantic relation prediction (CogALex 2.0) to solve difference in data quality across languages and linguistic relations

*Small Data in NLU: Proposals towards a Data-Centric Approach*

* Sketched some proposals to integrate existing methods and principles from Computational Linguistics / NLP into data-centric ML practices

*LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*

* Proposed LAION-400M dataset with CLIP-filtered 400 million image-text pairs, their CLIP embeddings and kNN indices that allow efficient similarity search

***Simultaneous Improvement of ML Model Fairness and Performance by Identifying Bias in Data***

* Proposed a methodology to remove potential bias-creating samples with respect to a protected attribute. Cosine similarity is used to flag similar samples
* Testing on UCI Adult dataset and German Credit dataset shows that there is an increase in accuracy and a decrease in biasness (using fairness metrics AOD, SPD)

*Augment & Valuate : A Data Enhancement Pipeline for data-centric AI*

* Introduced a domain-agnostic data enhancement pipeline. Includes a data cleansing via data valuation, data cleansing by contrastive learning, data augmentation (Faster AutoAugment)
* With a fixed ResNet-50 model, the accuracy has increased from 0.64 (baseline) to 0.85

*Can machines learn to see without visual databases?*

* Sustained the position that the time has come for thinking of learning machines that conquer visual skills in a truly human-like context

*Homogenization of Existing Inertial-Based Datasets to Support Human Activity Recognition*

* Proposed a platform (CLP) enabling the integration and the distribution of data coming from heterogeneous sources. It can be used for human activity recognition tasks

*Sim2Real Docs: Domain Randomization for Documents in Natural Scenes using Ray-traced Rendering*

* Created a framework for synthesizing datasets and performing domain randomization of documents in natural scenes, using Blender for 3D rendering

*Data Expressiveness and Its Use in Data-centric AI*

* Introduced a data-centric version of EM algorithm to measure the value / expressiveness of the data by using expressive data structure like a binary search tree
* Training time is reduced when compared to the previous EM approaches