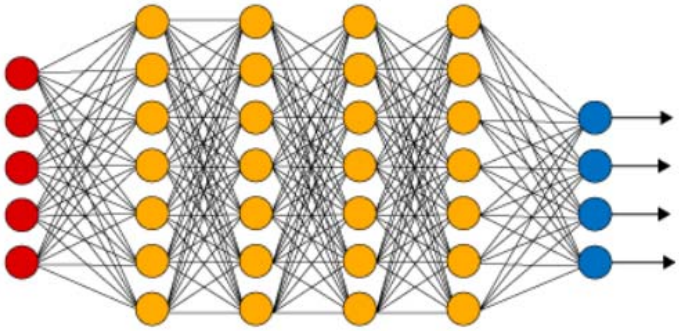
New Trends in Machine Learning and Data Analytics

Prof. Dr. E. Rahm
und Mitarbeiter

Seminar, WS 2020/21



SEMINARZIELE

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
  - vorgegebene Literaturempfehlungen können ergänzt werden (zB Recherche in Google Scholar)
- Diskussion
- schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilte(n) Betreuer/in

MODULZUORDNUNG

- Seminar ist beschränkt auf Master Data Science
- Teil des Pflichtmoduls Skalierbare Datenbanktechnologien 1
  - daneben noch 2 Vorlesungen aus IDBS1, Cloud and Big Data Management, Data Mining
  - inoffizielle Belegungsalternative in Ausnahmefällen: 3 VL
- Erhöhung des Teilnehmerlimits von 20 auf 30

BEWERTUNG/ MODULTEILPRÜFUNG

- selbständiger Vortrag mit Diskussion (ca. 22+8=30 Minuten)
  - Abnahme der Folien durch Betreuer/in
  - Folien: Englisch, Vortrag: Deutsch oder Englisch
- schriftliche Ausarbeitung (15-20 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer/in
  - Abgabe-Deadline **31.3.2021**
- Voraussetzungen für erfolgreiche Seminarleistung
  - aktive Teilnahme an allen Vortragsterminen
  - beide Teilleistungen werden erbracht (Vortrag/Ausarbeitung)
  - Bewertung: 50% Vortrag/Diskussion, 50% Ausarbeitung
- Seminar-Workload (gemäß Modulbeschreibung) 150 h:
  - 30h Präsenzzeit
  - 120 h Selbststudium (Vorbereitung Vortrag, Ausarbeitung)

ScaDS DRESDEN LEIPZIG

TERMINE

UNIVERSITÄT LEIPZIG

- Themenzuordnung
  - Präferenzen von 1 bis 5 bis **bis 01.11.2020 23:59 Uhr**
  - Themenzuteilung ab 3.11. im Moodlekurs einsehbar
- Absprache mit Betreuer/in bezüglich Stoffauswahl / Vortragsgestaltung
  - möglichst frühzeitig
- Vortragstermine
  - 5 Freitag Nachmittage in Moodle (**8.1.,15.1., 22.1., 29.1, 5.2**.)
  - 2 Sitzungen mit je 3 Vorträgen ab 13:15 Uhr

44 Themen

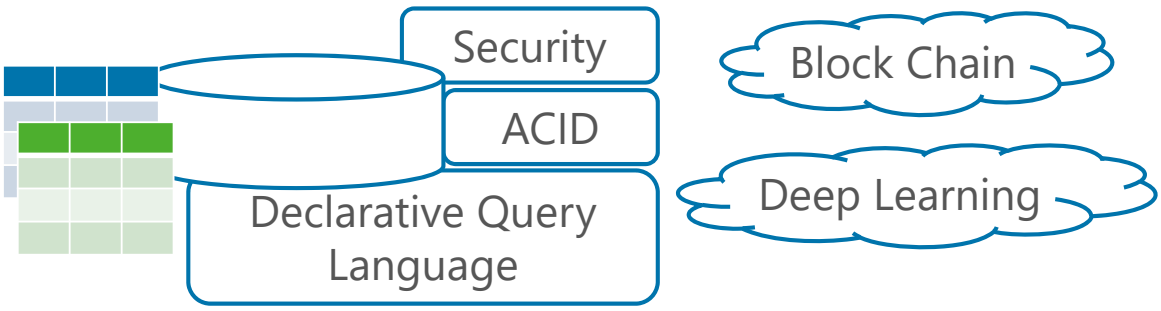
Nr	Topic	Supervisor
Machine Learning in Databases		
DB1	Data Management in Machine Learning:Challenges, Techniques, and Systems	Christen
DB2	DBMS Tuning with ML-Techniques	Christen
DB3	Security and Privacy on Blockchain	Franke
Privacy & Security		
P1	Membership Inference Attacks Against Machine Learning Models	Schneider
P2	Preventing Membership Inference Attacks with PATE	Schneider
P3	Generating Differential Private Datasets Using GANs	Schneider
P4	Clustered federated Learning: Model-Agnostic Distributed Multitask Optimization under Privacy Constraints	Sehili
P5	Practical Secure Aggregation for Privacy-Preserving Machine Learning	Sehili
P6	ABY3: A Mixed Protocol Framework for Machine Learning	Sehili
P7	Privacy-Preserving Classification on Deep Neural Network	Sehili
P8	Crime Data Analysis	Franke
Techniques for limited labeled data		
LD1	Human in the Loop for Entity Resolution	Köpcke
LD2	Cross-Modal Entity Resolution Based on Co-Attentional Generative Adversarial Network	Köpcke
LD3	Transfer Learning for Entity Resolution	Wilke
LD4	Effective and Efficient Data Cleaning for ER	Köpcke
LD5	Semi-automated Labelling for ML	Wilke
LD6	Machine Learning for Entity Resolution	Saeedi

Nr	Topic	Supervisor
Time Series Analysis		
TS1	Time-series forcasting	Täschner
TS2	Time Series Classification with ML: HIVE-COTE and InceptionTime	Burghardt
Graphs		
G1	Programming Abstractions for Distributed Graph Processing	Rost
G2	Graph Stream Summarization Techniques	Rost
G3	Dynamic/Stream Graph Neural Network	Alkamel
G4	Graph Analytics on GPUs	Gomez
G5	The Message Passing Framework for Graph Neural Networks	Petit
G6	Graph Neural Networks from a Spectral Perspective	Petit
G7	Attention Models in Graphs	Petit
G8	Large-Scale Machine Learning on Graphs	Schuchart
G9	Bootstrapping Entity Alignment with Knowledge Graph Embeddings	Obraczka
G10	Multi-view Knowledge Graph Embedding for Entity Alignment	Obraczka
Signal processing		
SP1	Location Tracking using Mobile Device Sensors	Rohde
SP2	Automated Reverse Engineering and Privacy Analysis of Modern Cars	Grimmer
SP3	Advances in pedestrian detection systems	Täschner
SP4	Person Detection With a Fisheye Camera	Burghardt
SP5	Bird Voice Recognition	Franke
SP6	Marine Bioacoustics I : ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning	Lin
SP7	Marine Bioacoustics II : Marine Mammal Species Classification using Convolutional Neural Networks and a Novel Acoustic Representation	Lin
Deep Learning in Physics		
PH1	Physics Informed Neural Networks	Uhrich
PH2	Deep Neural Networks Motivated by Partial Differential Equations	Uhrich
Bio-Medical Applications		
BM1	Deep Learning for Prediction of Survival of Brain Tumors	Martin
BM2	Machine Learning for Genomics Data	Christen
BM3	Construction of biomedical knowledge graphs	Christen
BM4	Electronic Health Record Data Quality	Rohde
BM5	Human Behavioural Analysis For Ambient Assisted Living	Burghardt
BM6	Active survival learning in precision medicine	Pogany

ML & DATABASES

Machine Learning in Databases		
DB1	Data Management in Machine Learning:Challenges, Techniques, and Systems	Christen
DB2	DBMS Tuning with ML-Techniques	Christen
DB3	Security and Privacy on Blockchain	Franke

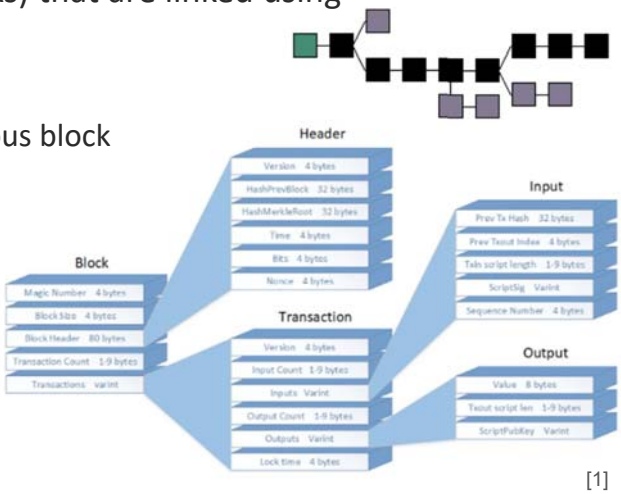
DBMS AND ML (CHRISTEN, DB1/DB2)



- bridge between relational DBMS and Machine Learning
- utilize ML techniques for DBMS tuning (**index selection**, query reformulation, etc.)
- DB-inspired ML Systems
  - declarative ML-language

SECURITY AND PRIVACY ON BLOCKCHAIN (FRANKE, DB3)

- blockchain: list of records (blocks) that are linked using cryptography
- each block contains a
  - cryptographic hash of the previous block
  - timestamp
  - transaction data
- vulnerabilities in blockchain?
  - hashing operations
  - identity attacks
  - routing attacks
  - …



[1]

[1] Dasgupta, D. et al.: A survey of blockchain from security perspective. *J BANK FINANC TECHNOL* **3,** 1–17, 2019

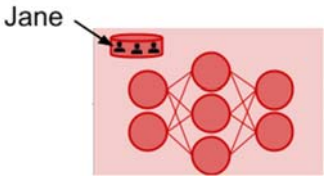
PRIVACY & SECURITY

P1	Membership Inference Attacks Against Machine Learning Models	Schneider
P2	Preventing Membership Inference Attacks with PATE	Schneider
P3	Generating Differential Private Datasets Using GANs	Schneider
P4	Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization under Privacy Constraints	Sehili
P5	Practical Secure Aggregation for Privacy-Preserving Machine Learning	Sehili
P6	ABY3: A Mixed Protocol Framework for Machine Learning	Sehili
P7	Privacy-Preserving Classification on Deep Neural Network	Sehili
P8	Crime Data Analysis	Franke

MEMBERSHIP INFERENCE (SCHNEIDER, P1)

- avoid membership inference attacks (MIA)
  - find out whether a specific record/person is in training data used for ML model

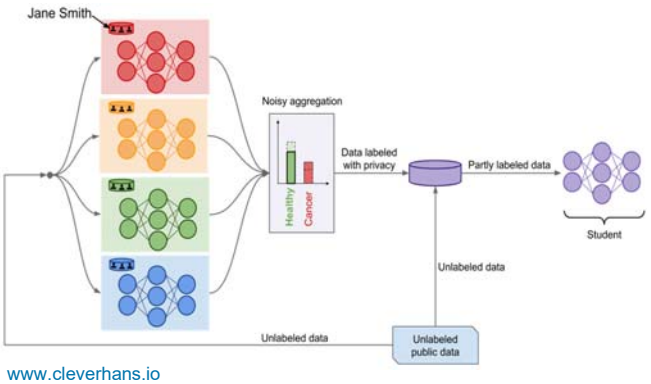
Train ML model
recommend treatment for HIV patient



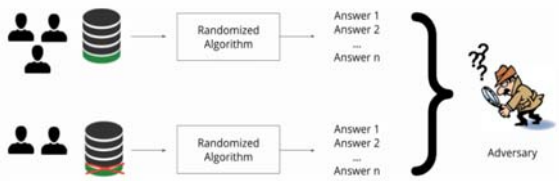
MIA
Was Jane part of training data?
Exposes Jane's HIV status!

PATE FRAMEWORK (SCHNEIDER, P2)

- **Preventing Membership Inference Attacks with PATE**



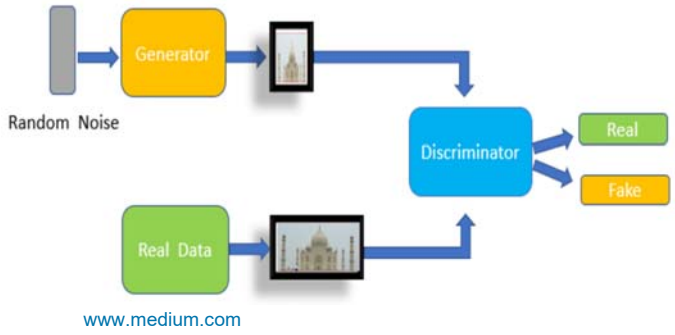
www.cleverhans.io



PATE
applies Differential Privacy and Ensemble Learning

Differential Privacy
no adversary can detect if Jane contributed to a query result

GENERATING SYNTHETIC DATA WITH GANS (SCHNEIDER, P3)



www.medium.com

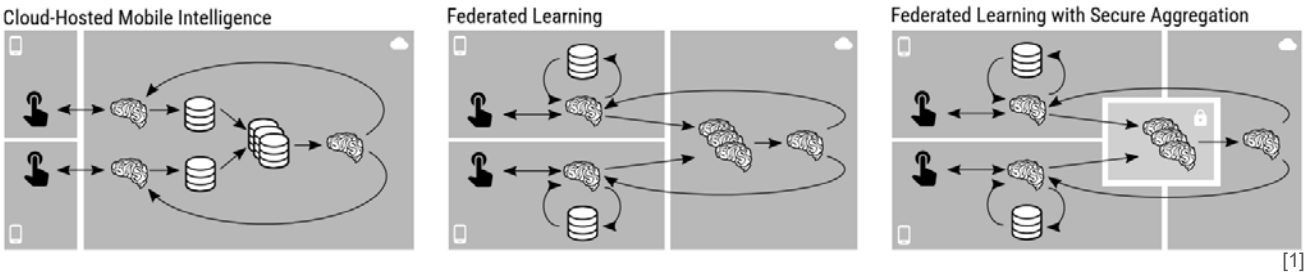
Generator
generate realistic fake data and fool the discriminator

Discriminator
detect fakes

- Generative Adversarial Networks (GAN): adversarial training of two neural nets
- protection of sensitive training data
- synthetization of training data using GANs based on real data

PRIVACY-PRESERVING FEDERATED LEARNING (SEHILI, P4-P5)

- **federated learning** models are trained and evaluated locally.
- summaries of local models are shipped to a server to be **aggregated in a new global model**.
- utilize **secure multiparty aggregation** to aggregate local models.

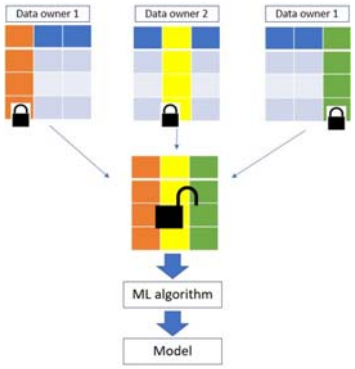


[1]

[1] Bonawitz, K. et al.: Practical Secure Aggregation for Privacy-Preserving Machine Learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191, 2017

PRIVACY PRESERVING MACHINE LEARNING (SEHILI, P6)

- ML algorithms often use person-related data e.g. in medicine, finance …
- problems: data sharing not possible due to competitive or regulatory reasons.
- solution: **encrypt data at owner –> (decrypt) data at use (MPC)**
- **ABY3** presents several protocols for multi-party computation (MPC)
- uses encrypted data for:
  - linear/logistic regression
  - neural networks
  - extendable to other models



https://dl.acm.org/doi/10.1145/3243734.3243760

PRIVACY PRESERVING CLASSIFICATION ON DNN (SEHILI, P7)

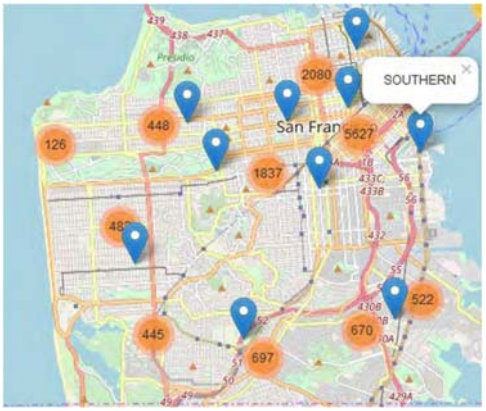
- use of **homomorphic encryption** for privacy-preserving classification:
  - user encrypts its data and sends it to server that holds a trained model.
  - server makes prediction over encrypted data and generates an encrypted result.
  - the server sends encrypted result to the user who can decrypt it.
  - **the server knows nothing about user's data and result.**
  - **the user know nothing about the model hold by the server**
- important requirements: **efficiency and accuracy for deep neural networks**

[1] Pengtao Xie et al.: Crypto-nets: Neural networks over encrypted data. arXiv:1412.6181, 2014

[1]

SECURITY – CRIME DATA ANALYSIS (FRANKE, P8)

- pattern recognition for violent crimes and accidents
- goal: assistance for crime prevention
  - tracking, prediction and prevention
- challenges
  - large amounts of heterogeneous multi-sourced data
  - real-time processing and forecasting



[1]

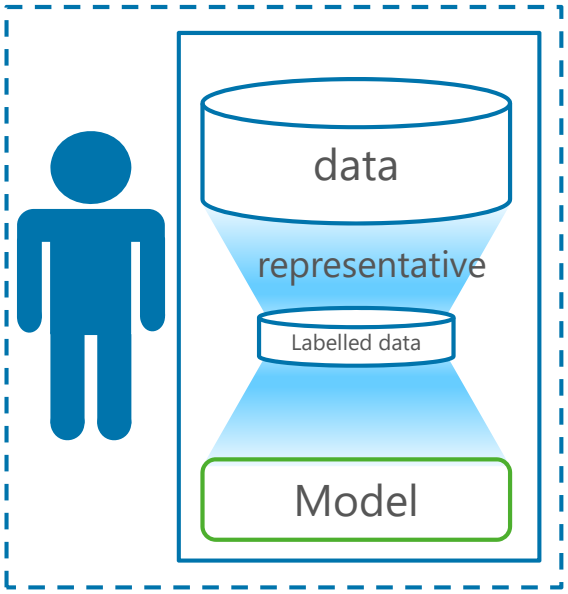
[1] M. Feng *et al.*, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in *IEEE Access*, vol. 7, pp. 106111-106123, 2019

TECHNIQUES FOR LIMITED LABELLED DATA

LD1	Human in the Loop for Entity Resolution	Köpcke
LD2	Cross-Modal Entity Resolution Based on Co-Attentional Generative Adversarial Network	Köpcke
LD3	Transfer Learning for Entity Resolution	Wilke
LD4	Effective and Efficient Data Cleaning for ER	Köpcke
LD5	Semi-automated Labelling for ML	Wilke
LD6	Machine Learning for Entity Resolution	Saeedi

LIMITED LABELLED DATA

- huge amount of data
- effective models require representative labelled data
  - mostly small amount of labelled data available
  - data labelling is expensive and time consuming



HUMAN IN THE LOOP FOR ENTITY RESOLUTION (KÖPCKE, LD1)

- entity resolution
  - identification of representations for the same real world object

CID	Name	Street	City
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503
24	Christian Smith	Hurley St 2	S Fork MN

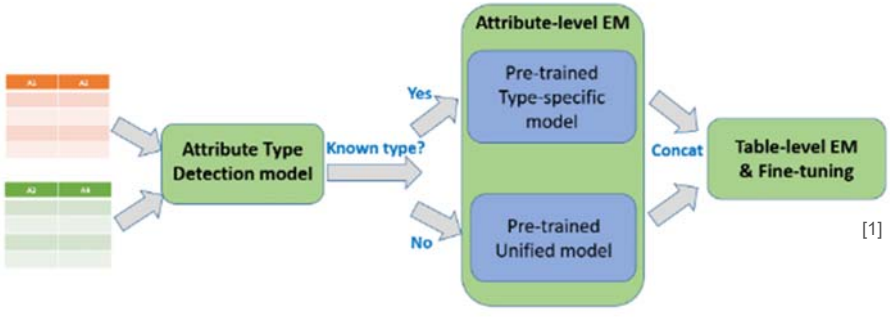
- learning-based approaches have become state of the art
- challenges
  - sufficient amount of labelled examples for learning high quality models
  - explainability: Black Box models are difficult for humans to interpret
- approach
  - explainable active learning

CROSS-MODAL ENTITY RESOLUTION (KÖPCKE, LD2)

- cross-modal entity resolution aims to find semantically similar items from objects of different modalities (e.g. image and text).
- key challenge: How to bridge the modality gap
- approach:
  - co-attentional Generative Adversarial Network (CAGAN) : Generative adversial network with co-attention mechanism
  - co-attention: eliminate the imbalance of information between modalities and generate more consistent representations

TRANSFER LEARNING FOR ENTITY RESOLUTION (WILKE, LD3)

- **transfer learning**: reduce the amount of training data by using models that are pretrained on very large and 'generic' data
- can we combine **ER** with **transfer learning**?

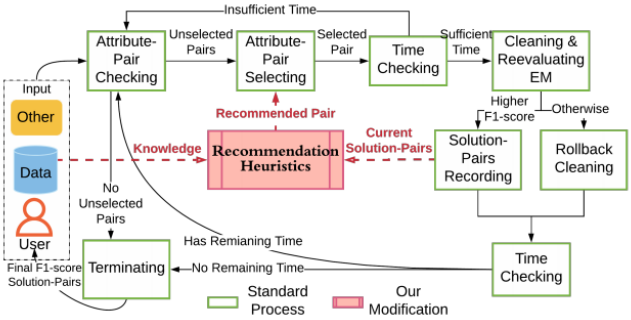


[1]

[1] Zao, He. Auto-EM: End-to-end Fuzzy Entity-Matching using Pre-trained Deep Models and Transfer Learning. WWW, 2019.

EFFECTIVE AND EFFICIENT DATA CLEANING (KÖPCKE, LD4)

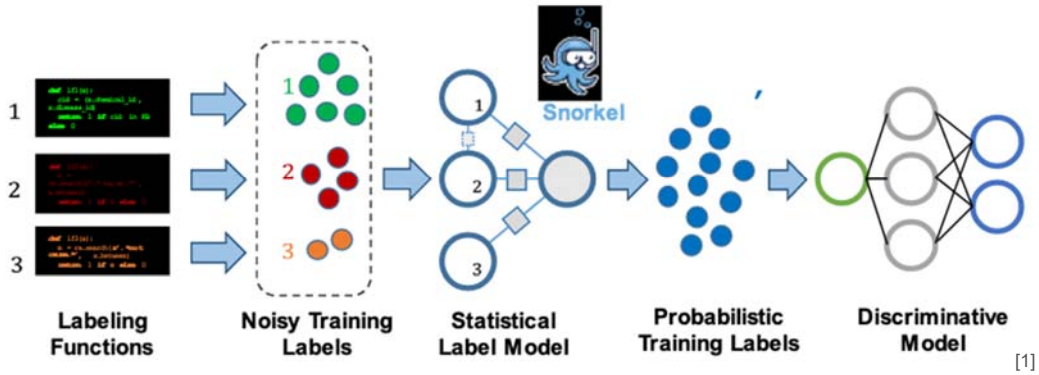
- ER quality can be improved by data cleaning
- but: time cost of data cleaning by human experts could be prohibitive
- approach:
  - maximize ER quality by data cleaning under a time constraint on the cleaning efforts by users
  - recommend to human experts a time-efficient order in which values of attributes could be cleaned in the given data



[1] Ao, J. et al.: Effective and Efficient Data Cleaning for Entity Matching. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA'19) Article 2, 1–7, 2019

SEMI-AUTOMATED LABELLING FOR ML (WILKE, LD5)

- "Labeling training data is increasingly the largest bottleneck in deploying machine learning systems"
- rule-based systems are much faster to train/configure but lack expressivity
- idea: combine many "weak" sources to create a generative model of the training data

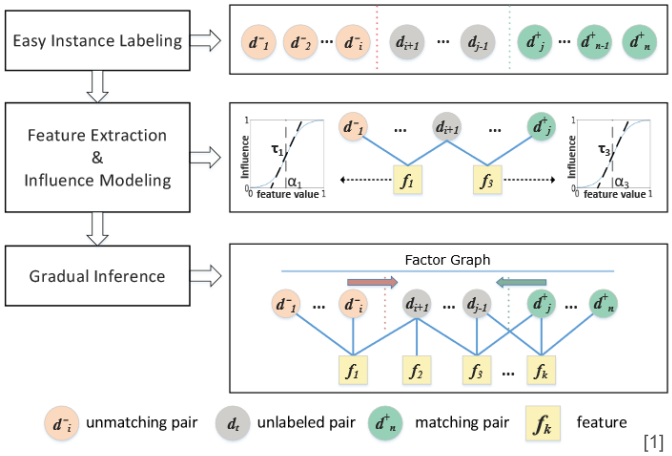


[1]

[1] Ré. Software 2.0 and Snorkel: Beyond Hand-Labeled Data. KDD 2018.

ML FOR ENTITY RESOLUTION (SAEEDI, LD6)

- Entity Resolution (ER): the task of disambiguating records that correspond to real world entities
- recent approaches are focused on Machine Learning (ML)
- challenge: little amount of labelled training data
  - One solution: Gradual ML
    - Steps:



[1]

[1] B. Hou et al.: "Gradual Machine Learning for Entity Resolution," in *IEEE Transactions on Knowledge and Data Engineering, 2020*

TIME SERIES ANALYSIS

TS1	Time-series forcasting	Täschner
TS2	Time Series Classification with ML: HIVE-COTE and InceptionTime	Burghardt

TIME-SERIES FORECASTING (TÄSCHNER, TS1)

- key area in academic research with applications in climate modeling, biological science, decision making in retail and finance, …
- extraction of meaningful characteristics from historical data
- **time-series forecasting**
  - prediction of future values based on **previous observations** + **uncertainty estimation**
- objectives of recommended literature:
  - overview over traditional (domain expertise) and modern data-driven approaches
  - comparison of different methods (one-step-ahead, multi-horizon, …)
  - evaluation of recent approaches using ML / Deep Learning

TIME SERIES CLASSIFICATION WITH ML: HIVE-COTE AND INCEPTIONTIME (BURGHARDT, TS2)

- TSC = ML area for categorization (or labelling) of time series
- significant progress in accuracy of classifiers in last decades

- Flat Collective of Transformation-based Ensembles (Flat-COTE)
  - combines 35 classifiers on 4 representations
- HIVE-COTE: Hierarchical Vote Collective of Transformation-based Ensembles

Questions:

- What is TSC?
- What are the application areas for TSC?
- Which methods are there for TSC (e.g. Time Series Forest)?
- How does these methods work and differ among each other?
- What special properties must the data have for successful application of TSC?
- What use cases can be solved with TSC methods?
- How does FLAT-COTE differ from the classic methods?
- What makes HIVE-COTE different from FLAT-COTE?
- Strengths and weaknesses of HIVE-COTE?
- How does InceptionTime differ from HIVE-COTE and the other methods?
- Which method is suitable for a special application from production (band saw, classification of condition band saw).

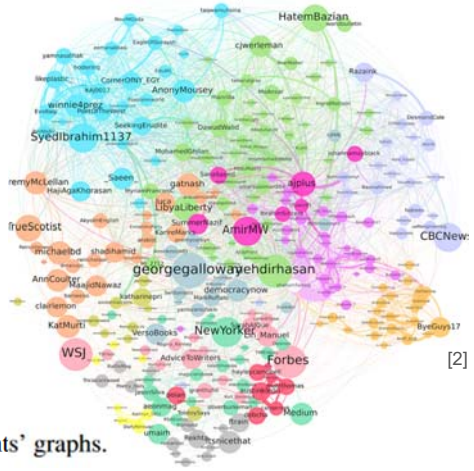
GRAPH MACHINE LEARNING

G1	Programming Abstractions for Distributed Graph Processing	Rost
G2	Graph Stream Summarization Techniques	Rost
G3	Dynamic/Stream Graph Neural Network	Alkamel
G4	Graph Analytics on GPUs	Gomez
G5	The Message Passing Framework for Graph Neural Networks	Petit
G6	Graph Neural Networks from a Spectral Perspective	Petit
G7	Attention Models in Graphs	Petit
G8	Large-Scale Machine Learning on Graphs	Schuchart
G9	Bootstrapping Entity Alignment with Knowledge Graph Embeddings	Obraczka
G10	Multi-view Knowledge Graph Embedding for Entity Alignment	Obraczka

GRAPH-ML – CHALLENGES

UNIVERSITÄT LEIPZIG

- analysis of large graphs
- analysis/prediction of graph evolution
  - new nodes/edges, deletions …
- interpretable graph representations for ML
  - graph node embeddings



[2]

Table 6: The sizes of the participants' graphs.

(a) Number of vertices.

Vertices	Total	R	P
< 10K	22	11	11
10K–100K	22	9	13
100K–1M	19	7	12
1M–10M	17	6	11
10M–100M	20	10	10
> 100M	27	10	17

[1]

(b) Number of edges.

Edges	Total	R	P
< 10K	23	11	12
10K–100K	22	9	13
100K–1M	13	3	10
1M–10M	9	5	4
10M–100M	21	8	13
100M–1B	21	8	13
> 1B	20	8	12

(c) Total uncompressed bytes.

Size	Total	R	P
< 100MB	23	12	11
100MB–1GB	19	9	10
1GB–10GB	25	9	16
10GB–100GB	17	5	12
100GB–1TB	20	8	12
> 1TB	17	5	12

[1] Sahu, S., Mhedhbi, A., Salihoglu, S. et al. The ubiquity of large graphs and surprising challenges of graph processing: extended survey. The VLDB Journal 29, 595–618, 2020

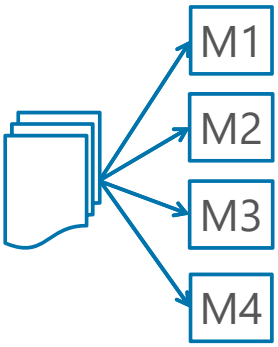
[2] https://towardsdatascience.com/information-flow-within-twitter-community-def9e939bb99

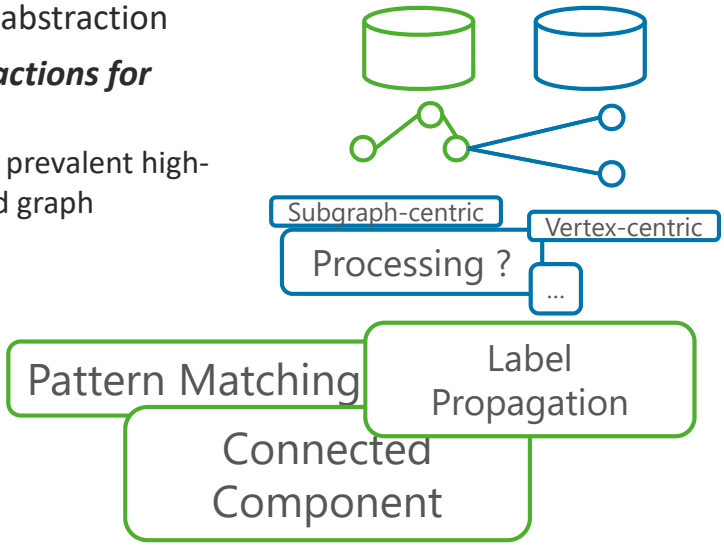
ScaDS DRESDEN LEIPZIG

EFFICIENT ANALYSIS OF LARGE AND EVOLVING GRAPHS (ROST, G1)

UNIVERSITÄT LEIPZIG

- using distributed graph processing systems
- algorithms require a high-level abstraction
- ***high-level programming abstractions for distributed graph processing***
  - goal : present overview of most prevalent high-level abstractions for distributed graph processing



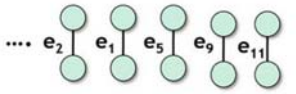


ScaDS DRESDEN LEIPZIG

GRAPH STREAM SUMMARIZATION TECHNIQUES (ROST, G2)

UNIVERSITÄT LEIPZIG

- Graph streams
  - continuous sequence of edges, including its two endpoints and attributes
  - streaming graphs are very large and change fast
- **Graph stream summarization**
  - G = (V, E) → Gh = (Vh, Eh)
    where |Vh| ⩽ |V| and |Eh| ⩽ |E|
  - structural, attribute-based or hybrid
  - requirements of summarization:
    - (1) the linear space cost
    - (2) the constant update time
- ***Graph Stream Summarization Techniques***



Edge Stream [1]

[1] Khan, A., Aggarwal, C. Toward query-friendly compression of rapid graph streams. Soc. Netw. Anal. Min. 7, 23, 2017

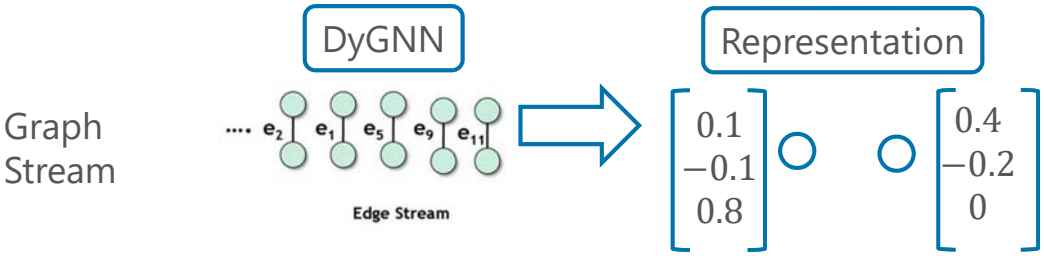
ScaDS DRESDEN LEIPZIG

DYNAMIC STREAMS & GRAPH NEURAL NETWORK (ALKAMEL, G3)

UNIVERSITÄT LEIPZIG

- Graph Neural Network
  - generation of representations consisting of structural and local features
  - only for static graphs

- **adaptation of existing GNNs for graph streams**



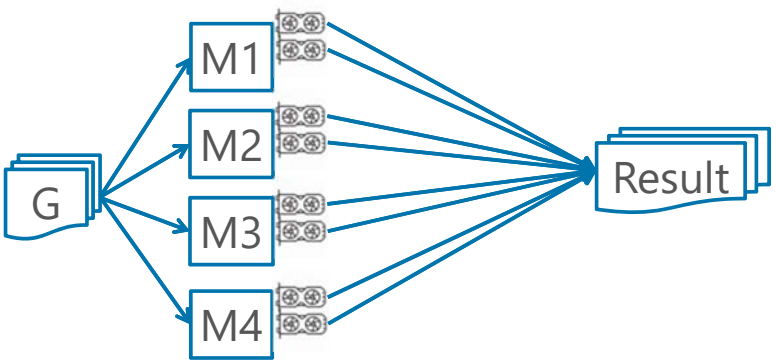
ScaDS DRESDEN LEIPZIG

GRAPH ANALYTICS ON GPU (GOMEZ, G4)

UNIVERSITÄT LEIPZIG

- challenges on GPU graph processing
  - partition large graphs among GPUs
  - implement efficient communication among GPUs
  - ensuring efficient computation on each GPU
  - for multi-host: efficient communication

- **graph analytics for massive datasets on distributed GPUs**

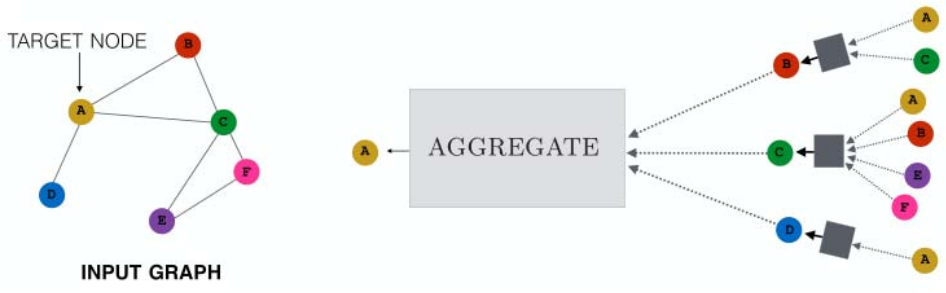


GRAPH EMBEDDINGS (PETIT, SCHUCHART)

- embedding (knowledge) graph entities into a low-dimensional space
  - encode entity properties and relationships to neighbors in graph
- generation and application of embeddings
  - ***the Message Passing Framework for Graph Neural Networks***
  - ***Graph Neural Networks from a spectral perspective***
  - ***Graph Attention Networks***

THE MESSAGE PASSING FRAMEWORK (PETIT, G5)

- general formulation for Graph Neural Networks (GNNs)
- different approaches can be formulated in MPNN (message passing neural networks) framework
- aggregate from local neighborhood and then update

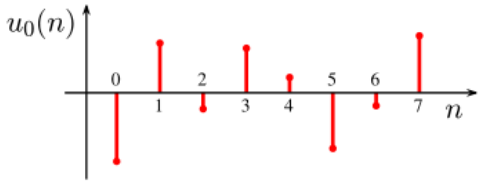


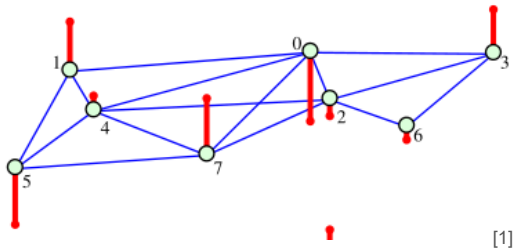
[1]

[1] Hamilton, William L.:Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* 14.3 (2020): 1-159.

GNNS FROM A SPECTRAL PERSPECTIVE (PETIT, G6)

- GNNs = Graph Neural Networks
- connection between graph signal processing and GNNs
- GNNs as filters on graph signals
- advantageous: background knowledge in signal processing



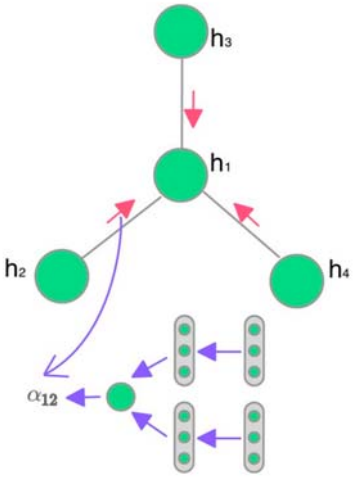


[1]

[1] Stanković L., Daković M., Sejdić E.:Introduction to Graph Signal Processing. In: Stanković L., Sejdić E. (eds) Vertex-Frequency Analysis of Graph Signals. Signals and Communication Technology. Springer, Cham., 2019

ATTENTION MODELS FOR GRAPHS (PETIT, G7)

- attention mechanism gained a lot of traction in the past years: de-facto standard in many sequence-based tasks
- transfer of attention mechanism to graph domain: attending over the neighborhood of a graph

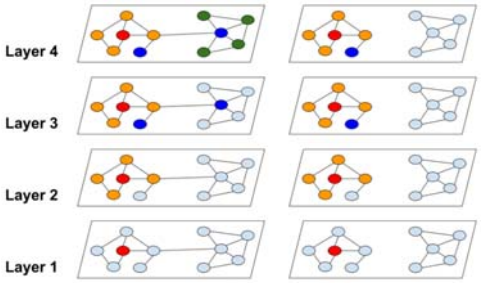


[1]

[1] https://dsgiitr.com/blogs/gat/

LARGE-SCALE ML ON GRAPHS (SCHUCHART, G8)

- Graph Convolutional Networks highly effective on small to medium sized graphs, but challenging for large graphs
  - high memory consumption
  - hard to compute
- ClusterGCN:
  - cluster small sub-graphs
  - learn on subset of sub-graphs



[1]

[1] Chiang, Wei-Lin et al.: Cluster-GCN. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019

GRAPH EMBEDDINGS FOR ENTITY RESOLUTION (OBRACZKA, G9/G10)

- ER to identify the same entities in two knowledge graphs (KG)
- challenges
  - supervised ER approaches require already matched entity pairs
  - features for generating embeddings (attributes, relationships, etc.)
- solutions
  - sampling strategies to extend the training data
  - → ***bootstrapping ER with KG embeddings (G9)***
  - ***multi-view KG embeddings for ER (G10)***
    - train different models per feature and combine them

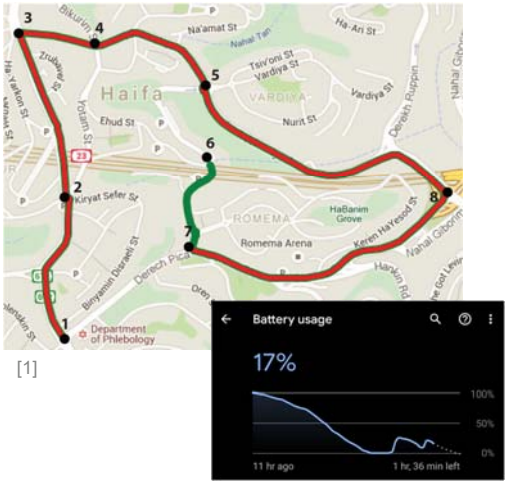
SIGNAL & IMAGE PROCESSING

SP1	Location Tracking using Mobile Device Sensors	Rohde
SP2	Automated Reverse Engineering and Privacy Analysis of Modern Cars	Grimmer
SP3	Advances in pedestrian detection systems	Täschner
SP4	Person Detection With a Fisheye Camera	Burghardt
SP5	Bird Voice Recognition	Franke
SP6	Marine Bioacoustics I : ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning	Lin
SP7	Marine Bioacoustics II : Marine Mammal Species Classification using Convolutional Neural Networks and a Novel Acoustic Representation	Lin

LOCATION TRACKING USING MOBILE DEVICE SENSORS (ROHDE, SP1)

UNIVERSITÄT LEIPZIG

- tracking mobile phones essentially means tracking people
- even without permission to access GPS or WIFI (SSID) attackers might infer the location e.g. from
  - the varying power consumption depending on the distance to the base station and obstacles between them (arXiv:1502.03182)
  - gyroscope, accelerometer, and magnetometer information (DOI 10.1109/SP.2016.31)

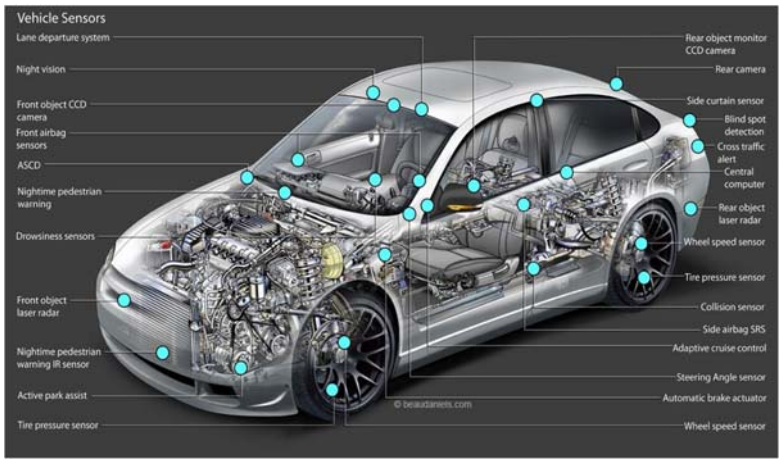


[1]

[1] Michalevsky, Y. et al.:PowerSpy: Location Tracking using Mobile Device Power Analysis., *CoRR* abs/1502.03182, 2015
[2] Google

AUTOMATED REVERSE ENGINEERING AND PRIVACY ANALYSIS OF MODERN CARS (GRIMMER, SP2)

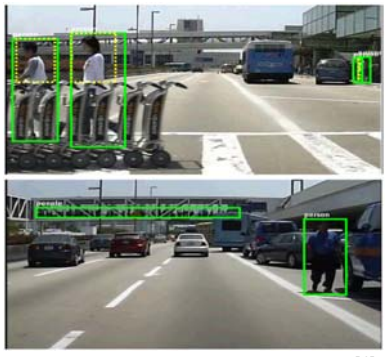
- "I Know Where You Parked Last Summer"
- what "private" information is recorded by your car?
- can we obtain this information?



https://carfromjapan.com/article/car-maintenance/types-of-sensors-used-in-automobile-engine/

PEDESTRIAN DETECTION SYSTEMS (TÄSCHNER, SP3)

- problem area of object detection and tracking
- application in fields of video surveillance, autonomous driving, human-computer interaction, …
- objectives of recommended literature:
  - compilation of different (recent) approaches
  - comparison of approaches and techniques used
  - methods and results of evaluation regarding accuracy, performance, robustness, …



[1]

[1] Brunetti, A. et al.: Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing 300, 17-33,2018

ANIMAL VOICE RECOGNITION (FRANKE, LIN, SP4-SP6)

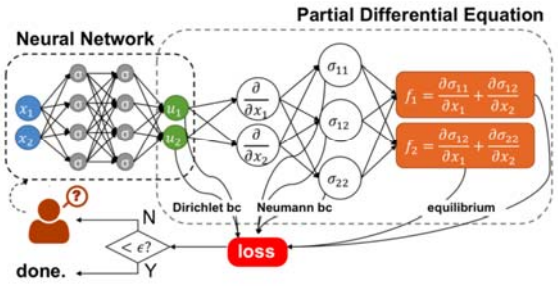
- motivation
  - studies about behaviour and implicitly about climate change or agriculture
  - build bioacoustics archives to identify reappearing communication patterns
- challenges
  - small amount of data
  - noisy recordings (cars, construction areas, etc.)
- ***bird voice recognition (SP4)***
  - concurrent singing, large amount of species
- ***marine mammal recognition using CNNs (SP5, SP6)***

DEEP LEARNING IN PHYSICS

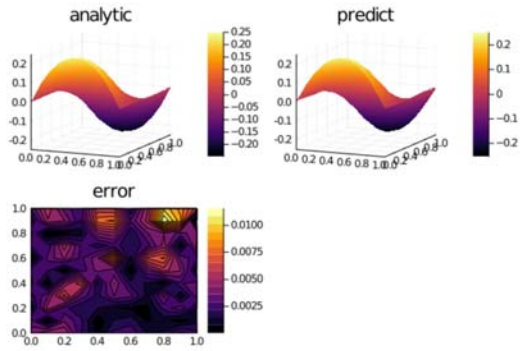
PH1	Physics Informed Deep Learning	Uhrich
PH2	Deep Neural Networks Motivated by Partial Differential Equations	Uhrich

PHYSICS INFORMED NEURAL NETWORKS (UHRICH, PH1)

- modeling and simulation of physical systems
- data-driven solution and discovery of nonlinear partial differential equations
- new class of data-efficient function approximators that naturally encode underlying physical laws as prior information
- opportunity for better predictions with limited labeled data



[1]

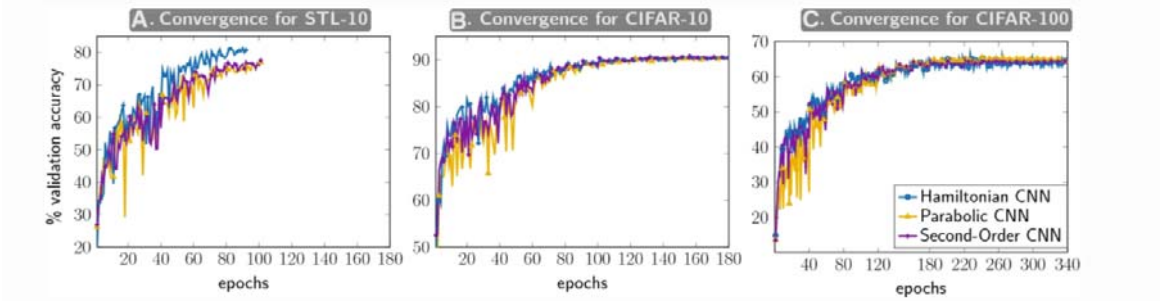


[2]

[1] Peng, G.C.Y *et al.*: Multiscale Modeling Meets Machine Learning: What Can We Learn?. *Arch Computat Methods Eng*, 2020
[2] https://nextjournal.com/kirill_zubov/physics-informed-neural-networks-pinns-solver-on-julia-gsoc-2020-second-evaluation

DEEP NEURAL NETWORKS MOTIVATED BY PARTIAL DIFFERENTIAL EQUATIONS (UHRICH, PH2)

- network design is central task in deep learning
- approach for designing, analyzing and training of more effective models
- Inspired by partial differential equations
- improves training outcomes and generalizes neural network architecture design with less trial and error



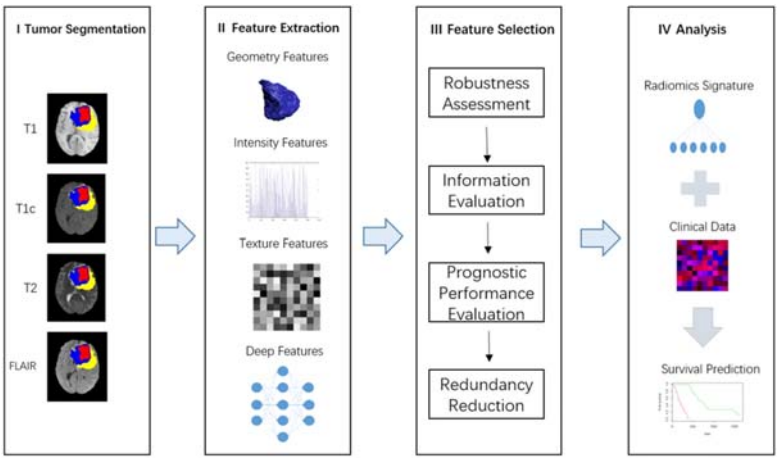
[1]

[1] Ruthotto, L., Haber, E.: Deep Neural Networks Motivated by Partial Differential Equations. *J Math Imaging Vis* **62,** 352–364, 2020

BIO-MEDICAL APPLICATIONS

BM1	Deep Learning for Prediction of Survival of Brain Tumors	Martin
BM2	Machine Learning for Genomics Data	Christen
BM3	Construction of biomedical knowledge graphs	Christen
BM4	Electronic Health Record Data Quality	Rohde
BM5	Human Behavioural Analysis For Ambient Assisted Living	Burghardt
BM6	Active survival learning in precision medicine	Pogany

DEEP LEARNING FOR SURVIVAL PREDICTION FOR BRAIN TUMORS (MARTIN, BM1)



[1]

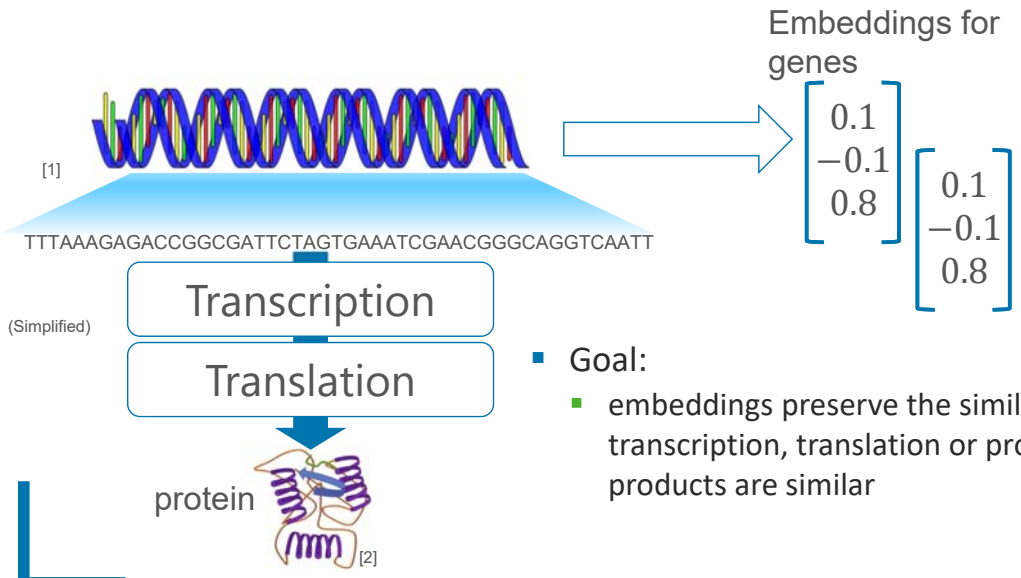
Keywords

- Radiomics
- MR images
- Feature Extraction
- Feature Selection
- Image Processing
- Deep Learning
- CNN
- Transfer Learning
- Statistical Analysis
- Signature Construction

[1] Lao et al.: A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme, Nature Scientific Reports, 7:10353, 2017

MACHINE LEARNING FOR GENOMICS DATA (CHRISTEN, BM2)

- effective representations of genomics data enable the application of ML-techniques
  - clustering and classification

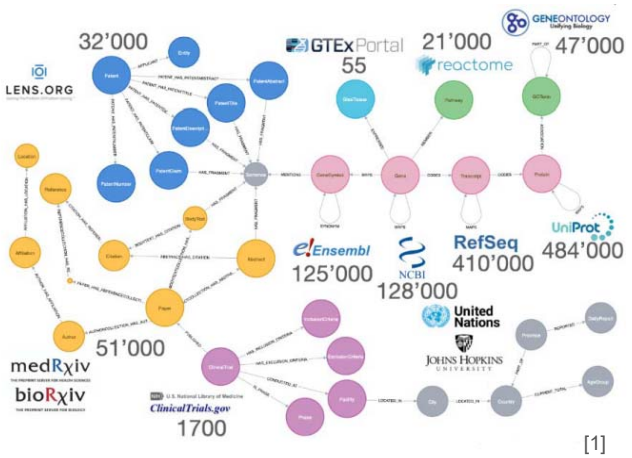


- Goal:
  - embeddings preserve the similarity if transcription, translation or protein products are similar

[1] by Clker-Free-Vector-Images from Pixabay
[2] Shen, Chang-Hui.:Gene Expression: Translation of the Genetic Code, Diagnostic Molecular Biology, 2019

REPRESENTATION OF BIOMEDICAL DATA (CHRISTEN, BM3)

- large amount of heterogeneous data (omics data, medical publications, case report forms)
  - unified representation in knowledge graphs (KG)
- **construction of biomedical KGs**
  - overview of the KG generation process
  - analysis opportunities (disease-gene associations)
  - application of embedding techniques
- **representation of genomics data**
  - methods to generate representations for gene expression data



[1]

https://covidgraph.org/

DATA QUALITY OF ELECTRONIC HEALTH RECORDS (ROHDE, BM4)

- electronic health records (EHR) are valuable sources for secondary use in research, operational analytics etc.
- utility depends on the data quality
- goal
  - review existing data quality assessment terminologies and tools

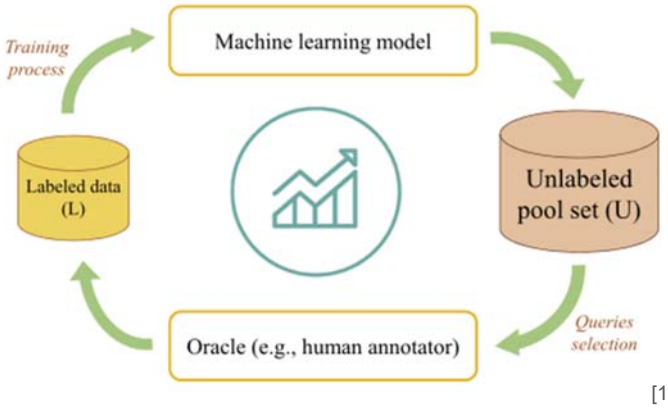
[1]

[1] https://www.aamc.org

HUMAN BEHAVIOURAL ANALYSIS FOR AMBIENT-ASSISTED LIVING (BURGHARDT, BM5)

- elders at home with chronic disease requiring ongoing care
  - so far: medical monitoring, e.g. blood glucose levels, heart rate, weight, blood pressure ....
- new research trends focus on unobtrusive monitoring of behaviour, including activity levels, falls and adherence to health behavior
- questions:
  - what are the challenges in the field of human behavioural analysis for ambient assisted living?
  - what methods can be applied in human behavioural analysis for ambient assisted living?
  - what are the trends in human behavioural analysis?
  - what are further application areas for human behavioural analysis?

ACTIVE SURVIVAL LEARNING IN PRECISION MEDICINE (POGANY, BM6)



[1]

- in medicine, labelled data is scarce and is often high-dimensional
- use deep learning to reduce features in unsupervised way
- use active learning to overcome labels problem

[1] Nezhad, M. Z. et al.: A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer, Expert Systems with Applications 115, 16-26, 2019

44 Themen

Nr	Topic	Supervisor
Machine Learning in Databases		
DB1	Data Management in Machine Learning:Challenges, Techniques, and Systems	Christen
DB2	DBMS Tuning with ML-Techniques	Christen
DB3	Security and Privacy on Blockchain	Franke
Privacy & Security		
P1	Membership Inference Attacks Against Machine Learning Models	Schneider
P2	Preventing Membership Inference Attacks with PATE	Schneider
P3	Generating Differential Private Datasets Using GANs	Schneider
P4	Clustered federated Learning: Model-Agnostic Distributed Multitask Optimization under Privacy Constraints	Sehili
P5	Practical Secure Aggregation for Privacy-Preserving Machine Learning	Sehili
P6	ABY3: A Mixed Protocol Framework for Machine Learning	Sehili
P7	Privacy-Preserving Classification on Deep Neural Network	Sehili
P8	Crime Data Analysis	Franke
Techniques for limited labeled data		
LD1	Human in the Loop for Entity Resolution	Köpcke
LD2	Cross-Modal Entity Resolution Based on Co-Attentional Generative Adversarial Network	Köpcke
LD3	Transfer Learning for Entity Resolution	Wilke
LD4	Effective and Efficient Data Cleaning for ER	Köpcke
LD5	Semi-automated Labelling for ML	Wilke
LD6	Machine Learning for Entity Resolution	Saeedi

Nr	Topic	Supervisor
Time Series Analysis		
TS1	Time-series forcasting	Täschner
TS2	Time Series Classification with ML: HIVE-COTE and InceptionTime	Burghardt
Graphs		
G1	Programming Abstractions for Distributed Graph Processing	Rost
G2	Graph Stream Summarization Techniques	Rost
G3	Dynamic/Stream Graph Neural Network	Alkamel
G4	Graph Analytics on GPUs	Gomez
G5	The Message Passing Framework for Graph Neural Networks	Petit
G6	Graph Neural Networks from a Spectral Perspective	Petit
G7	Attention Models in Graphs	Petit
G8	Large-Scale Machine Learning on Graphs	Schuchart
G9	Bootstrapping Entity Alignment with Knowledge Graph Embeddings	Obraczka
G10	Multi-view Knowledge Graph Embedding for Entity Alignment	Obraczka
Signal processing		
SP1	Location Tracking using Mobile Device Sensors	Rohde
SP2	Automated Reverse Engineering and Privacy Analysis of Modern Cars	Grimmer
SP3	Advances in pedestrian detection systems	Täschner
SP4	Person Detection With a Fisheye Camera	Burghardt
SP5	Bird Voice Recognition	Franke
SP6	Marine Bioacoustics I : ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning	Lin
SP7	Marine Bioacoustics II : Marine Mammal Species Classification using Convolutional Neural Networks and a Novel Acoustic Representation	Lin
Deep Learning in Physics		
PH1	Physics Informed Neural Networks	Uhrich
PH2	Deep Neural Networks Motivated by Partial Differential Equations	Uhrich
Bio-Medical Applications		
BM1	Deep Learning for Prediction of Survival of Brain Tumors	Martin
BM2	Machine Learning for Genomics Data	Christen
BM3	Construction of biomedical knowledge graphs	Christen
BM4	Electronic Health Record Data Quality	Rohde
BM5	Human Behavioural Analysis For Ambient Assisted Living	Burghardt
BM6	Active survival learning in precision medicine	Pogany