Data Science Applications

Prof. Dr. Ziawasch Abedjan / Mohammad Mahdavi/ Felix Neutatz / Mahdi Esmailoghli

Big Data Management

Big Data Management Group (BigDaMa)

Group leader

Ziawasch Abedjan

Mohammad Mahdavi

PhD Students

Larysa Visengeriyeva

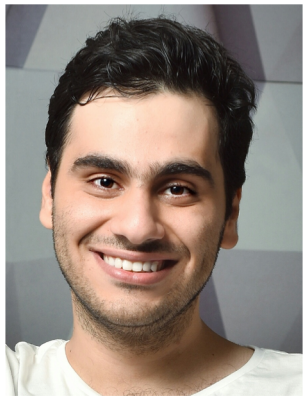
Data Cleaning
Machine Learning

Felix Neutatz

Feature Engineering
Machine Learning

Maximilian Dohlus

Data Lake Integration
Data Profiling



Mahdi Esmailoghli

Stream Data Analytics

Goal of the Seminar

- Learn to find and use scientific papers
- Build a system that benefits from integrating different sources
- Gaining experience in building components of an analytical application
- First-hand experience of applied machine learning

Prerequisites

- Reading and Presenting in English
- Being literate in at least one programming language
  - Preferably python or java
- Having enthusiasm for formally solving problems
- Having successfully attended the data integration lecture

Seminar Organization 1

1. Pick a group, select top 3 topics, propose solution – Deadline Sunday 12:00 am
2. Assignment of topics 24.10.
3. Find relevant literature and datasets for your application 31.10.
4. Discuss the paper with your advisor – Individual meetings (10% of the grade)
5. Give a 20+5 minutes presentation on ongoing progress– 28/29.11 (30% of the grade)
6. Give a 20+10 minutes final presentation – End of semester (30% of the grade)
7. Write an 8 Page summary on your system– Deadline February 20th (30% of the grade)

First Presentation

- How do you manage the data at hand?
- What is your application going to solve?
- Which resources are you going to use?
- Preliminary results and demo

Second Presentation

- Discuss your system and sell its advantages with experimental results
- Your written manuscript will have a similar structure.

Topics

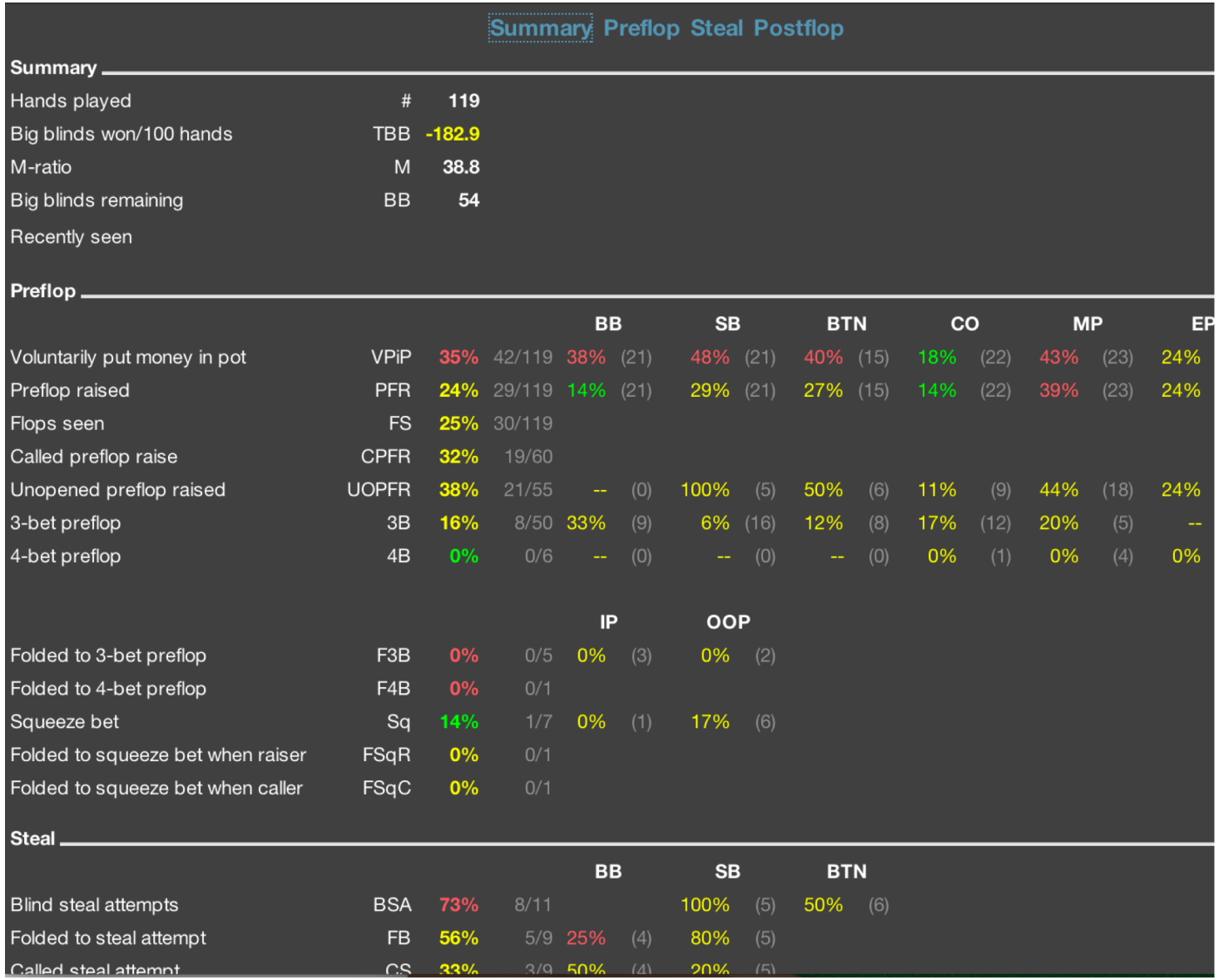
- Each topic has a data handling aspect and a machine learning task
  - Poker assistant
  - Predicting World Cup matches
  - Summarizing Daily news
  - Data Cleaning for Machine Learning
  - Outlier Explanation

Use Case: Texas Holdem Poker

Challenges:

- Assessing how strong your hand is in the current pot?
- How strong could your opponent be?
- What is the most profitable play in the current situation?

Statistics that you can obtain:



Summary Preflop Steal Postflop

Summary

Hands played	#	119
Big blinds won/100 hands	TBB	-182.9
M-ratio	M	38.8
Big blinds remaining	BB	54
Recently seen		

Preflop

				BB		SB		BTN		CO		MP		EP
Voluntarily put money in pot	VPiP	35%	42/119	38%	(21)	48%	(21)	40%	(15)	18%	(22)	43%	(23)	24%
Preflop raised	PFR	24%	29/119	14%	(21)	29%	(21)	27%	(15)	14%	(22)	39%	(23)	24%
Flops seen	FS	25%	30/119											
Called preflop raise	CPFR	32%	19/60											
Unopened preflop raised	UOPFR	38%	21/55	--	(0)	100%	(5)	50%	(6)	11%	(9)	44%	(18)	24%
3-bet preflop	3B	16%	8/50	33%	(9)	6%	(16)	12%	(8)	17%	(12)	20%	(5)	--
4-bet preflop	4B	0%	0/6	--	(0)	--	(0)	--	(0)	0%	(1)	0%	(4)	0%

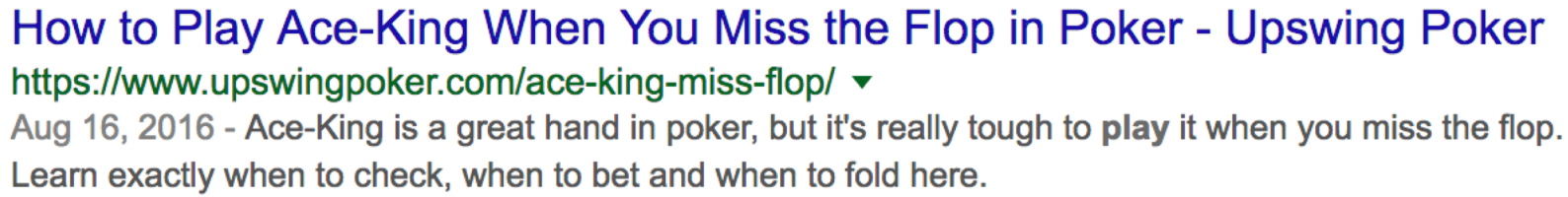
				IP		OOP	
Folded to 3-bet preflop	F3B	0%	0/5	0%	(3)	0%	(2)
Folded to 4-bet preflop	F4B	0%	0/1				
Squeeze bet	Sq	14%	1/7	0%	(1)	17%	(6)
Folded to squeeze bet when raiser	FSqR	0%	0/1				
Folded to squeeze bet when caller	FSqC	0%	0/1				

Steal

				BB		SB		BTN	
Blind steal attempts	BSA	73%	8/11			100%	(5)	50%	(6)
Folded to steal attempt	FB	56%	5/9	25%	(4)	80%	(5)		
Called steal attempt	CS	33%	3/9	50%	(4)	20%	(5)		

Further helpful data

- Identify player patterns
- Connect your dataset to external datasets:
  - Other hand histories
- Web sources:



Data: Hand history

PokerStars Hand #175821392060: Hold'em No Limit ($0.02/$0.05 USD) - 2017/09/18 16:35:37 ET
Table 'Dawson' 9-max Seat #4 is the button
Seat 1: ShadarNyr ($10.80 in chips)
Seat 2: Natch_Flush ($9.63 in chips)
Seat 3: totojkee ($5.15 in chips)
Seat 4: Luigi_198909 ($5 in chips)
Seat 5: karakasss ($4.95 in chips)
Seat 6: BigDaMaHero ($5 in chips)
Seat 7: SwimmingPool ($5 in chips)
Seat 8: (PSY)De_Ange ($5 in chips)
Seat 9: Langolier86 ($5.37 in chips)
karakasss: posts small blind $0.02
BigDaMaHero: posts big blind $0.05
*** HOLE CARDS ***
Dealt to BigDaMaHero [Tc 6d]
SwimmingPool: raises $0.10 to $0.15
(PSY)De_Ange: folds
Langolier86: folds
ShadarNyr: folds
Natch_Flush: folds
totojkee: folds
Luigi_198909: folds
karakasss: folds
BigDaMaHero: folds
Uncalled bet ($0.10) returned to SwimmingPool
SwimmingPool collected $0.12 from pot
*** SUMMARY ***
Total pot $0.12 | Rake $0
Seat 1: ShadarNyr folded before Flop (didn't bet)
Seat 2: Natch_Flush folded before Flop (didn't bet)
Seat 3: totojkee folded before Flop (didn't bet)
Seat 4: Luigi_198909 (button) folded before Flop (didn't bet)
Seat 5: karakasss (small blind) folded before Flop
Seat 6: BigDaMaHero (big blind) folded before Flop
Seat 7: SwimmingPool collected ($0.12)
Seat 8: (PSY)De_Ange folded before Flop (didn't bet)
Seat 9: Langolier86 folded before Flop (didn't bet)

Tasks

- Information extraction
  - Crawling data sources
  - Parsing data
  - Making data structured
- Storage and indexing
  - Designing data model
  - Providing efficient load/save operations
- Application
  - Modeling states and actions
  - Designing game play strategies

Information Extraction

- Some papers
  - Etzioni, Oren, et al. "Web-scale information extraction in knowitall:(preliminary results)." *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004.
  - Chang, Chia-Hui, et al. "A survey of web information extraction systems." IEEE transactions on knowledge and data engineering 18.10 (2006): 1411-1428.
  - Cowie, Jim, and Wendy Lehnert. "Information extraction." *Communications of the ACM* 39.1 (1996): 80-91.
  - Banko, Michele, et al. "Open information extraction from the web." *IJCAI*. Vol. 7. 2007.
  - Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In Proceedings of the Association of Computational Linguistics (ACL), 2015.

Information Extraction

- Some libraries and tools
  - For requesting web pages
    - E.g., Requests: HTTP for Humans
      - http://docs.python-requests.org/en/master/
  - For parsing texts
    - E.g., Beautiful Soup
      - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
    - E.g., Scrapy
      - https://scrapy.org/
  - For natural language processing
    - E.g., NLTK
      - https://www.nltk.org/
  - For information extraction
    - E.g., Stanford OpenIE
      - https://nlp.stanford.edu/software/openie.html
    - E.g., MITIE: MIT Information Extraction
      - https://github.com/mit-nlp/MITIE

Storage and Indexing

- Simple DBMS
  - Postgres
- Key-value-store
  - MongoDB
- Document indexer
  - Elastic search

Application

- Some papers
  - Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1. No. 1. Cambridge: MIT press, 1998.
  - Dahl, Fredrik A. "A reinforcement learning algorithm applied to simplified two-player Texas Hold'em poker." *European Conference on Machine Learning*. Springer, Berlin, Heidelberg, 2001.
  - Bowling, Michael, et al. "Heads-up limit hold'em poker is solved." *Science* 347.6218 (2015): 145-149.
  - Shi, Jiefu, and Michael L. Littman. "Abstraction methods for game theoretic poker." *International Conference on Computers and Games*. Springer, Berlin, Heidelberg, 2000.
  - Billings, Darse, et al. "Approximating game-theoretic optimal strategies for full-scale poker." *IJCAI*. 2003.
  - Baker, Roderick JS, and Peter I. Cowling. "Bayesian opponent modeling in a simple poker environment." *Computational Intelligence and Games, 2007. CIG 2007. IEEE Symposium on*. IEEE, 2007.

Application

- Some libraries and tools
  - For machine learning
    - E.g., scikit-learn
      - http://scikit-learn.org/stable/
    - E.g., Tensorflow
      - https://www.tensorflow.org/
  - For reinforcement learning
    - E.g., TensorForce
      - https://github.com/reinforceio/tensorforce

Predicting World Cup 2018 Matches

- The main focus of this project is on collecting various data from various sources and integrating them
- You should collect and integrate various data such as
  - Population of the countries
  - Their annual budget
  - Price of their players
  - Rank of their national football team
- Given a football match such as "France vs Croatia (Final)", your model should predict the wining probability for each of the teams

Predicting World Cup 2018 Matches

- A paper
  - [1] Suzuki, Adriano Kamimura, et al. "A Bayesian approach for predicting match outcomes: the 2006 (Association) Football World Cup." Journal of the Operational Research Society 61.10 (2010): 1530-1539.
- Some data sources
  - [2] https://www.scoreboard.com/en/soccer/world/world-cup/results/
  - [3] Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." The semantic web. Springer, Berlin, Heidelberg, 2007. 722-735.

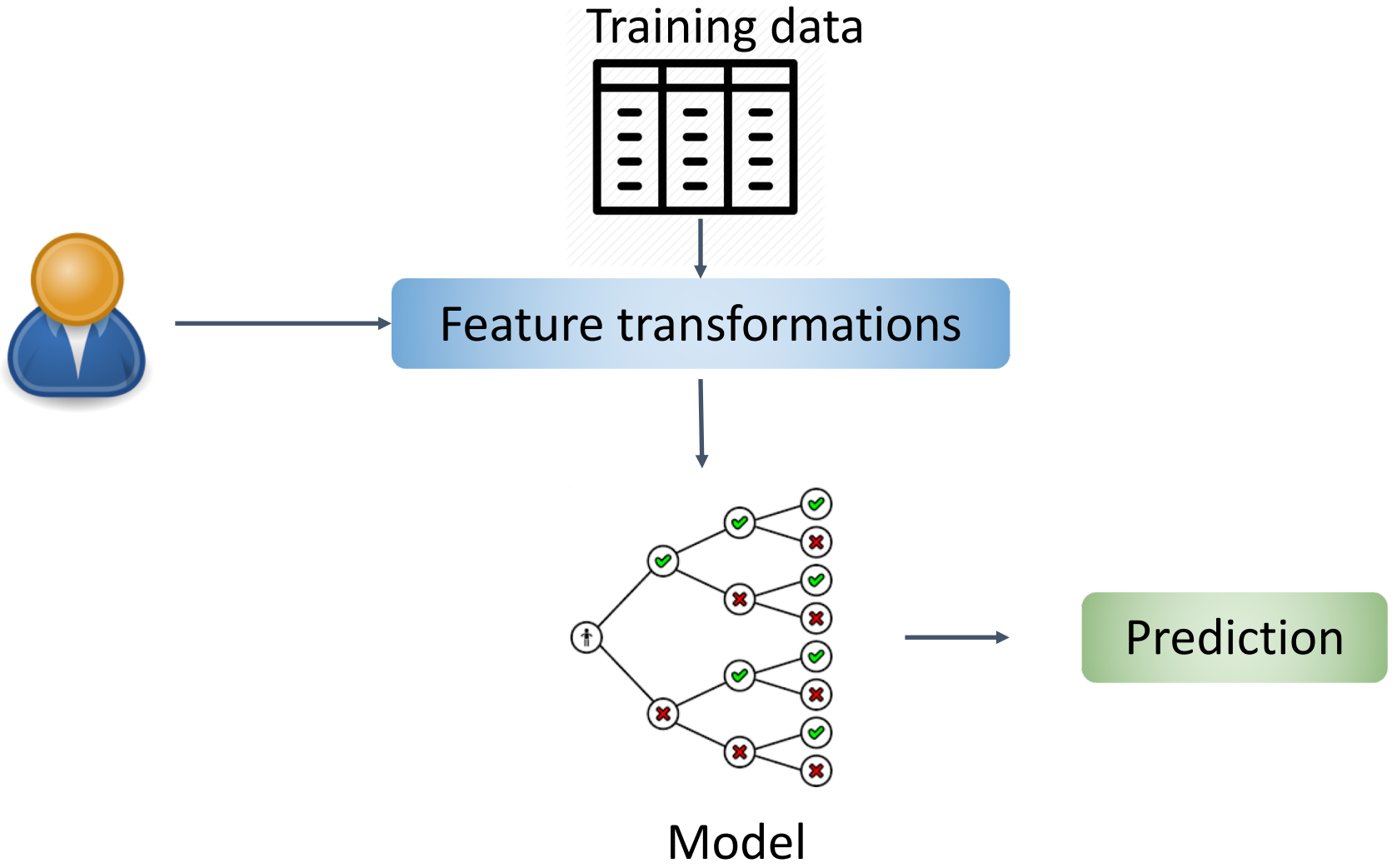
Summarizing Daily News

- The goal of this project is to build an online approach collecting and clustering the data
- You should build a system to collect news from various sources such as
  - News agency web sites
  - RSSes
  - Twitter
- Your system should cluster the collected news to present the summary of the hot daily news based on the clusters

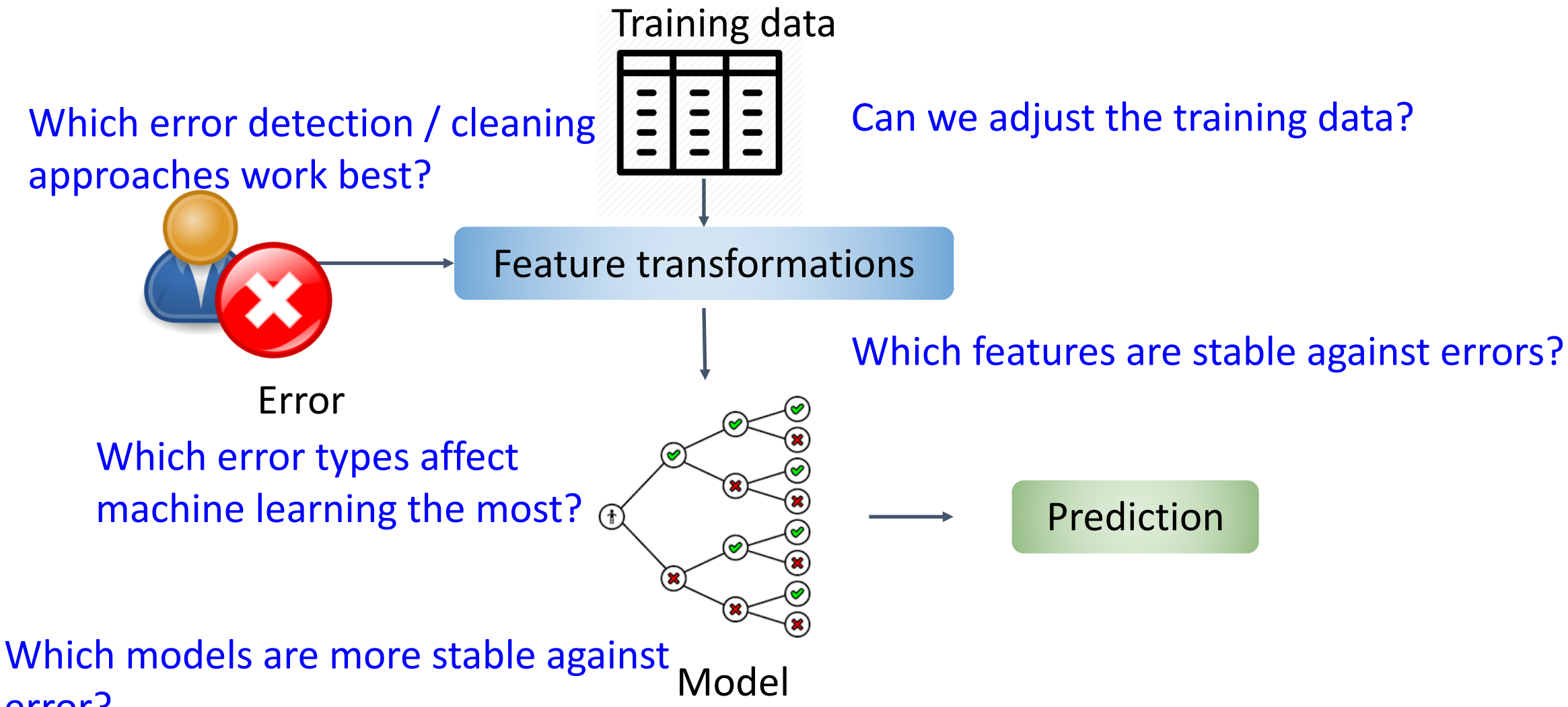
Summarizing Daily News

- A paper
  - [1] Huang, Anna. "Similarity measures for text document clustering." Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008.
- A data source
  - [2] http://feeds.bbci.co.uk/news/world/rss.xml

Data Cleaning for Machine Learning



Data Cleaning for Machine Learning



Tasks

Paper: Read about data cleaning in general:

Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M. and Tang, N., 2016. Detecting data errors: Where are we and what needs to be done?. *Proceedings of the VLDB Endowment*, *9*(12), pp.993-1004.

Datasets: Kaggle datasets, such as House Price Prediction, Wine Rating Prediction

Feature Engineering: Be creative! Start with simple features, such as one hot encoding.

Models: Try different models, such as XGBoost and Linear Regression.

Error Types: Use our error generator framework to create different test scenarios.

Cleaning Methods: Start with traditional cleaning methods using our abstraction layer framework. Then, use novel approaches, such as ActiveClean and ED2.

Extra: Come up with ways to augment the training to improve the model performance in case of errors.

Comparison of SVM and Random forest regarding to outlier Explainability

- Read “Anomaly detection: a survey”:
  - https://dl.acm.org/citation.cfm?id=1541882.
  - First 6 sections of the paper should be read.
  - Students should just know the techniques, challenges and limitations of techniques (No need to go into the details of techniques).
- Read an outlier explanation paper.
  - The paper will be announced.
- Read section 2.2 of EXstream paper.
  - https://par.nsf.gov/servlets/purl/10033440
  - Why this paper is of the opinion that Decision Tree and Logistic Regression are not suitable for explanation?

Comparison of SVM and Random forest regarding to outlier Explainability

- Implement an SVM and a Random forest outlier detection algorithm.
- Apply them on a labelled dataset such as KDD Cup 1999 dataset:
  - http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
- Explain detected outliers manually.
  - Using responsible features and values.
- Compare learnt models' ability in explaining detected outliers.
  - Learnt hyperplane in SVM
  - Learnt rules in Decision trees
- Compare the explanations of SVM and Random Forest with manual explanation (as ground truth) and report TP, FP, TN, and FN.
- Compare the explanation results with Macrobase as a state-of-the-art solution.

Next step:

- Submit until Sunday 12:00 am
  - Create groups of two students
  - Make a top 3 list of topics
  - Write an application/ implementation proposal (300 words)
  - CC in your email your group member!
- Meeting next week: Topic assignment