Universität Augsburg
Institut für Informatik


# GitHub Klassifizierung

Andreas Grafberger, Martin Keßler, Michael Leimstädtner, Stefan Grafberger

# Team Universität Augsburg

Andreas Grafberger

Martin Keßler

Michael Leimstädtner

Stefan Grafberger

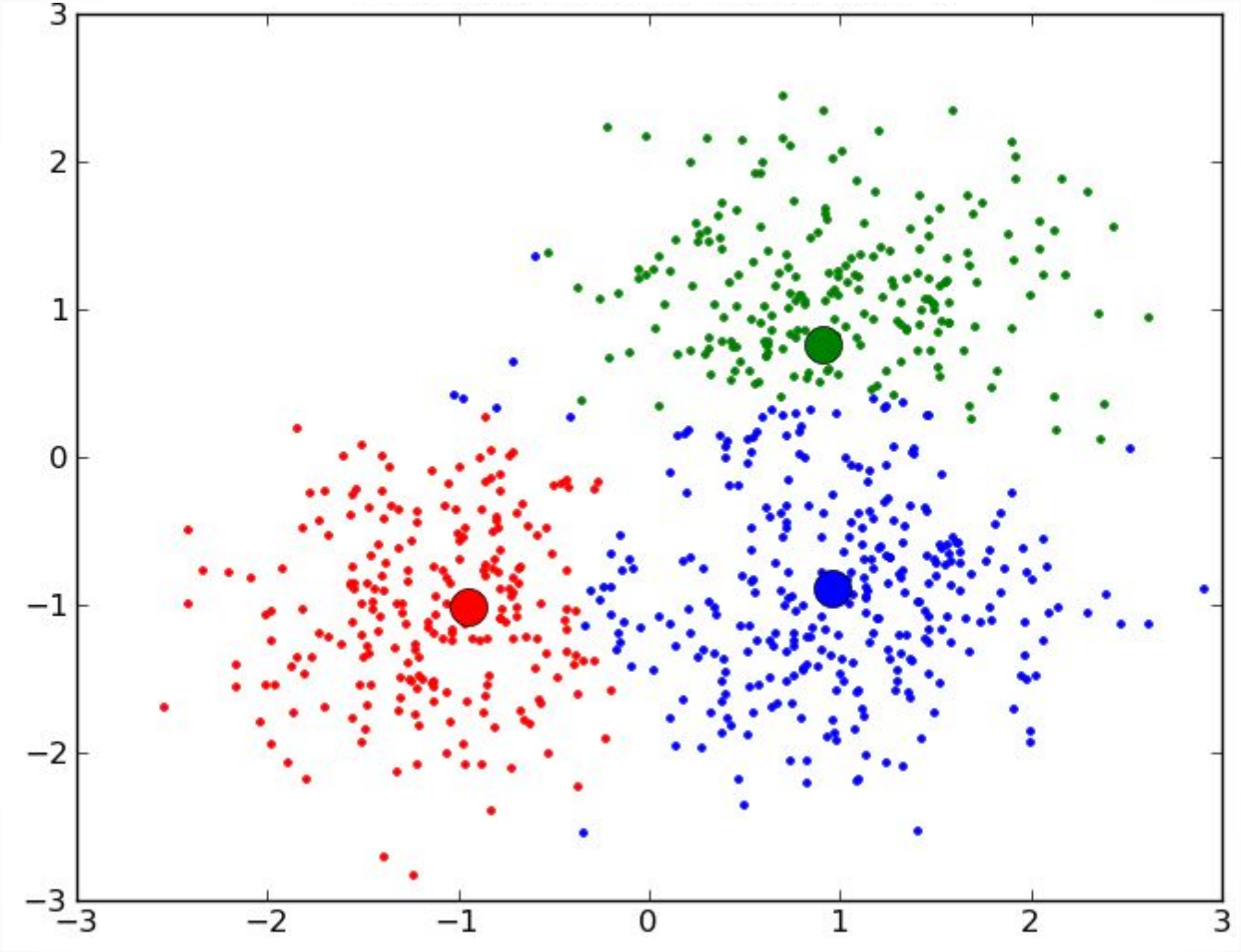
# Gliederung

- Herausforderungen und unser **Lösungsansatz**
- **Demonstration** der Anwendung
- **Features** und **Vorhersagemodelle**
- **Fragen**

# Besonderheiten und Herausforderungen

## Probleme

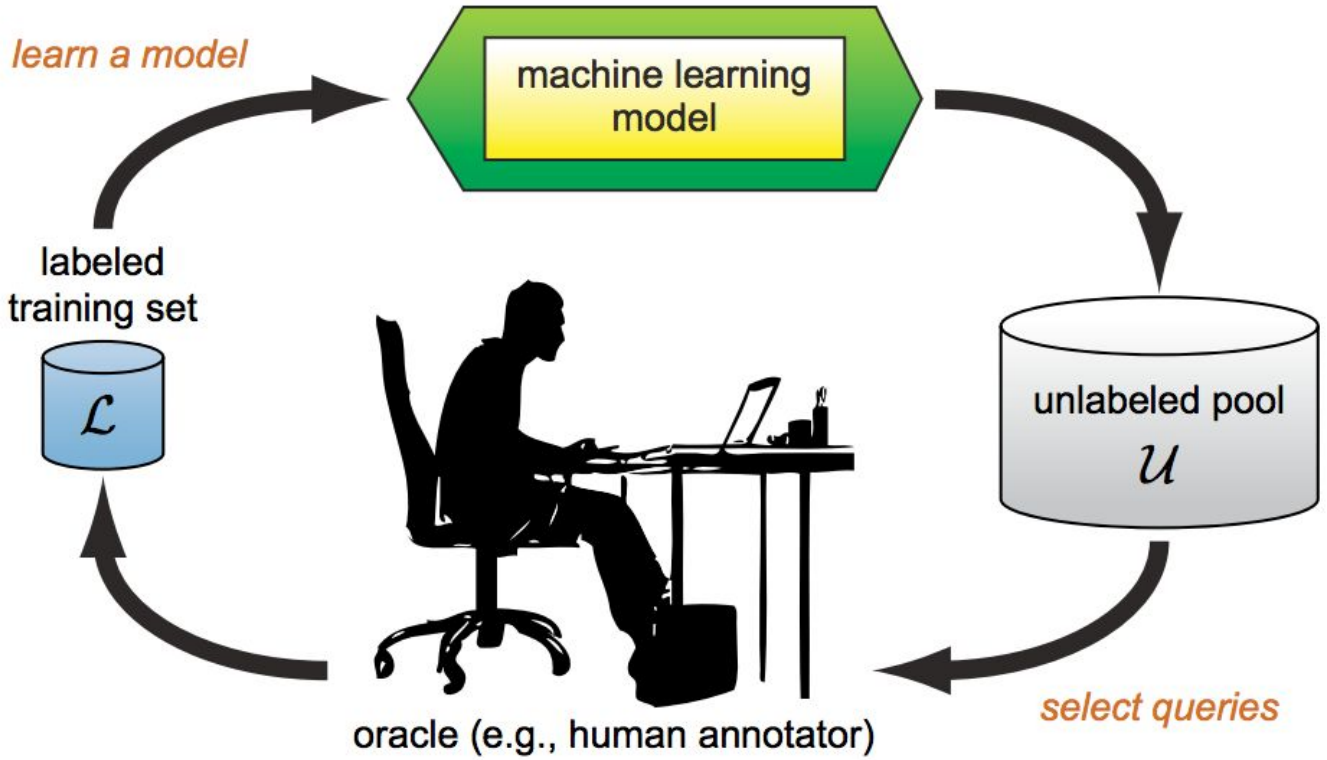
- Unklare **Klassenbeschreibungen**
- **Ungleichgewicht** der Klassenverteilung
- **Hochdimensionaler** unbekannter Feature-Raum




*Quelle: http://scikit-learn.sourceforge.net/0.5/_images/plot_mean_shift.png*

# Generierung der Trainingsdaten - Probleme

- Welche Repositories interessant?
- Keine scharfen **Grenzen** zwischen den Klassen
- **Majority** class
- Großer Feature-Vektor

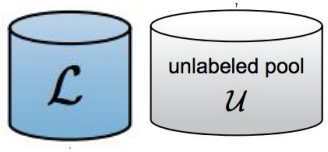


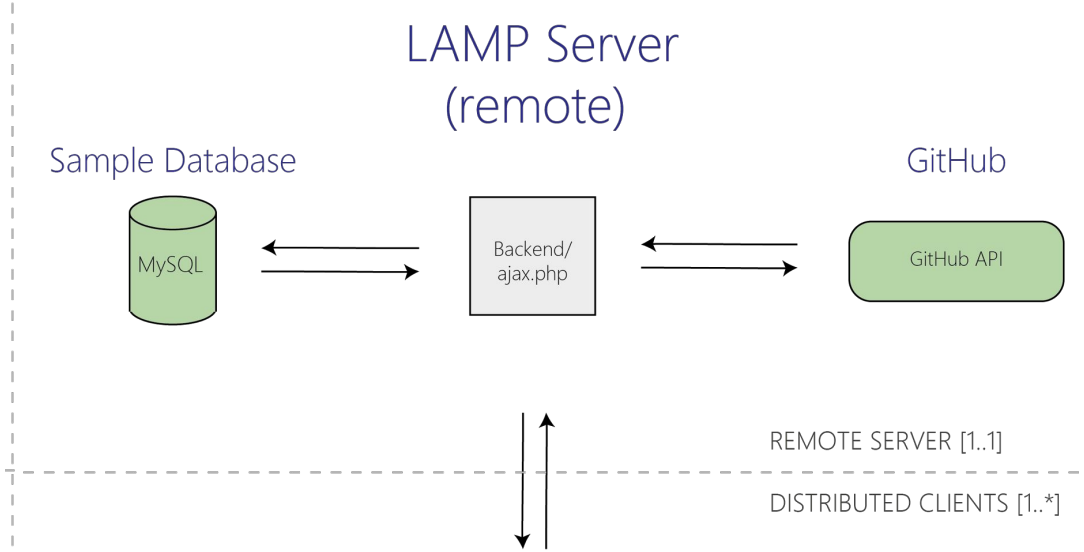

# Generierung der Trainingsdaten - Active Learning

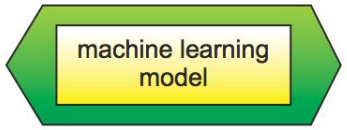
- Stream- und Pool based
- Unsicherheit auf Basis der Klassen-Wahrscheinlichkeiten

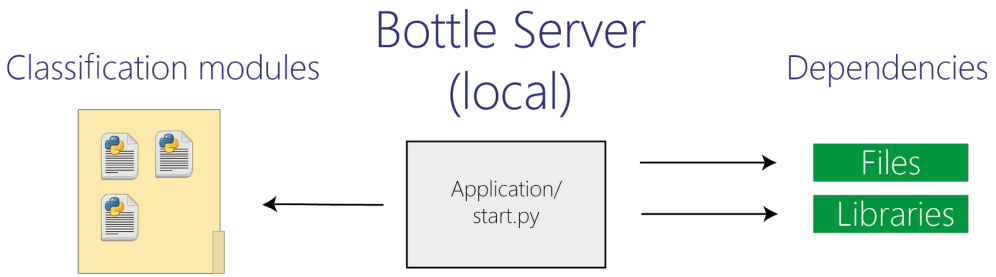
⇒ geringe Redundanz, 2.000 Samples

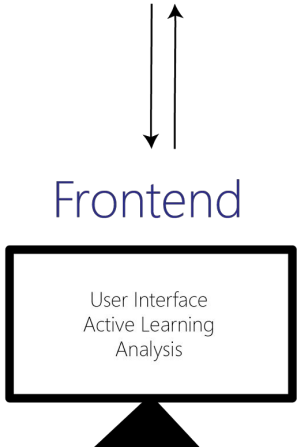
*Quelle: Settles, Burr (2010), "Active Learning Literature Survey" Computer Sciences Technical Report 1648.*



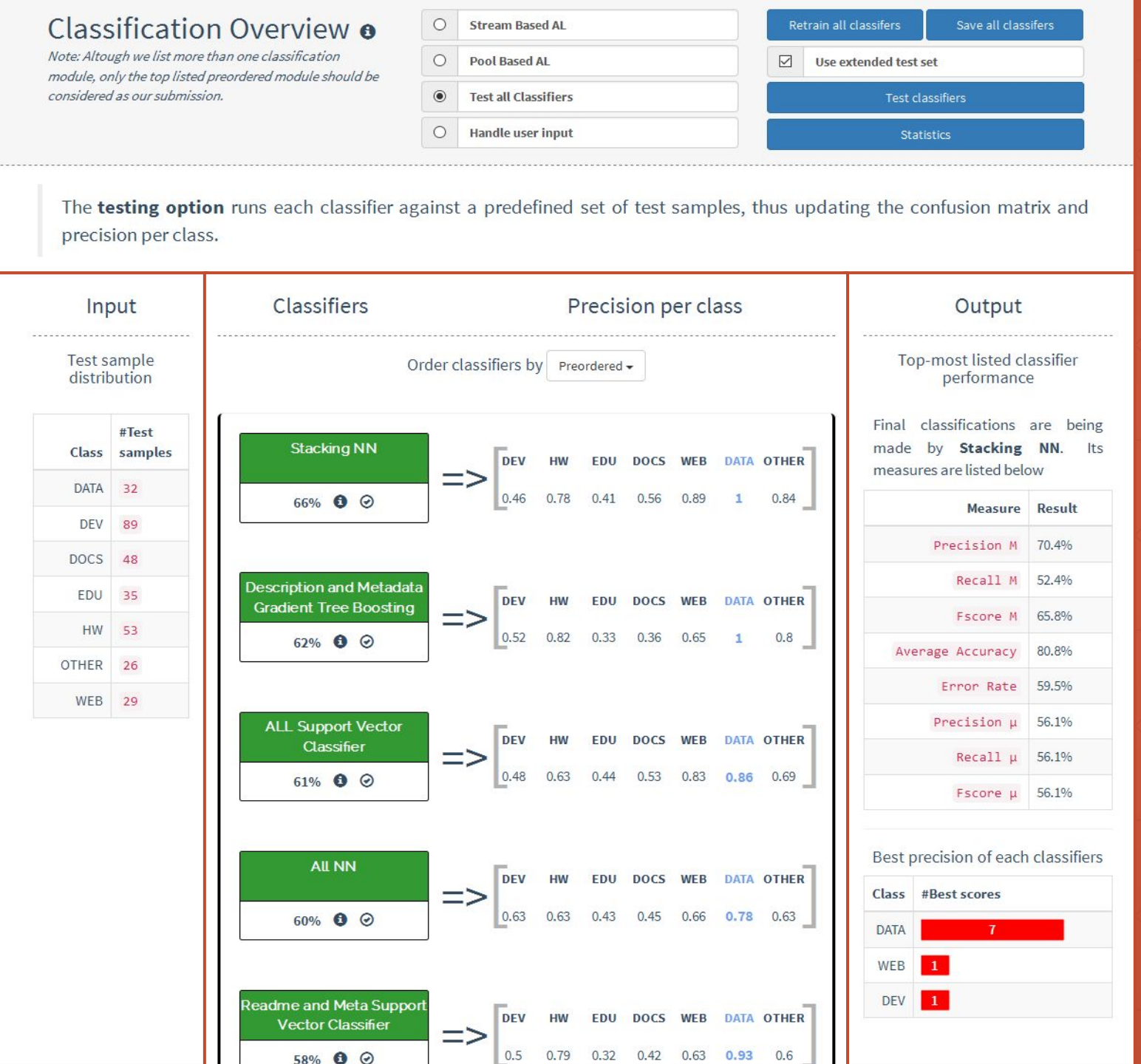




# Softwaredesign

- lokal: **Python - Server**
- remote: **LAMP - Server**
- DB: **MySQL**
- GUI: **HTML, CSS, Vue.JS**



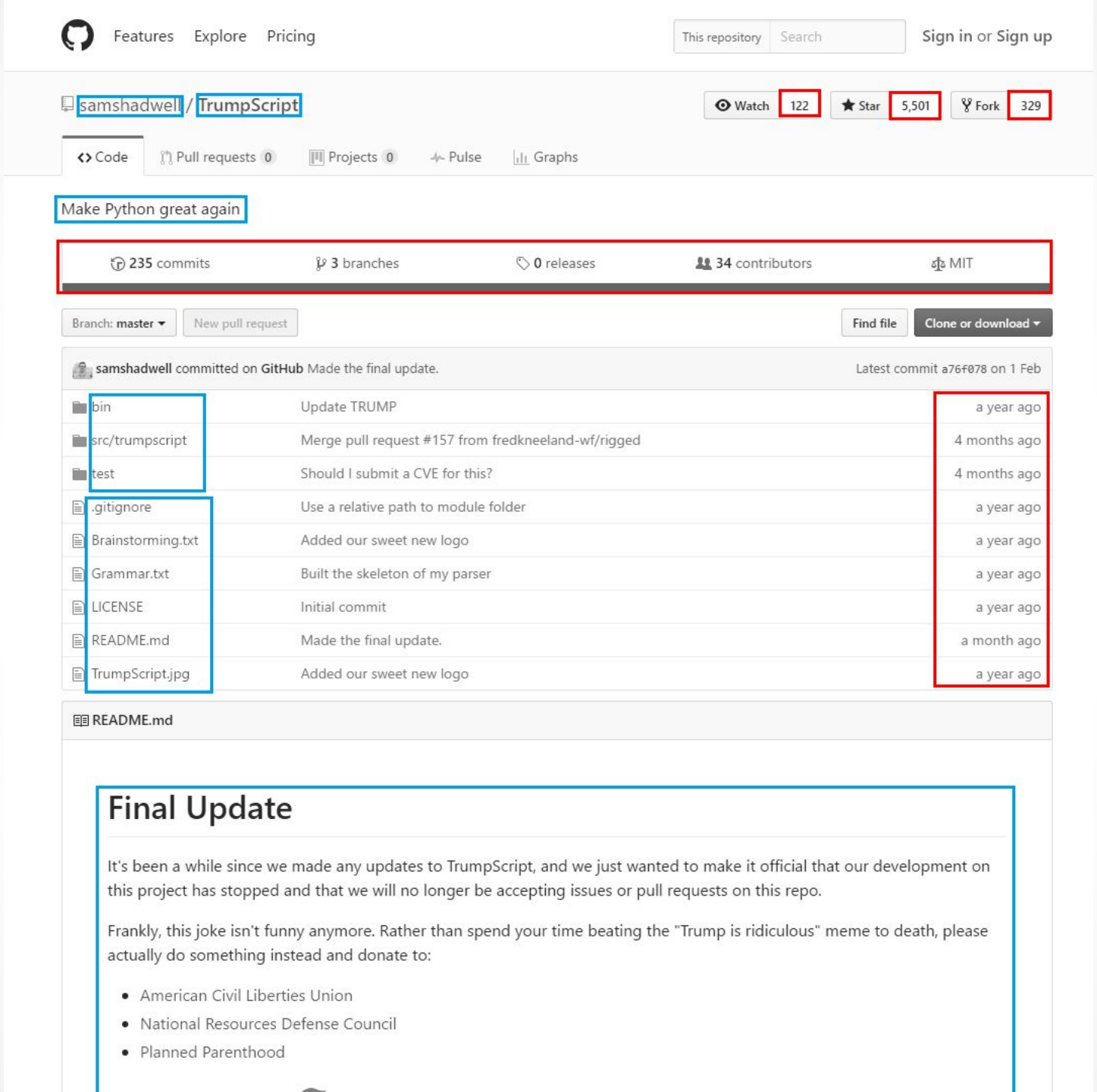

Universität Augsburg
Institut für Informatik

# Live Demo

- Struktur
- Use Cases

Universität Augsburg
Institut für Informatik

informatiCup




# Features

- Welche gibt es?
- Welche brauchen wir Menschen?
- Welche eignen sich?
- Codierung / Pre-Processing

# Text Features - Codierung

1. Bag-of-words model
    a. Term frequency (**Tf**) + inverse document frequency (**Tf-idf**)
    b. Stemming
2. N-grams
3. Character-sequence (**RNN**, **LSTM**, …)
4. Word2Vec

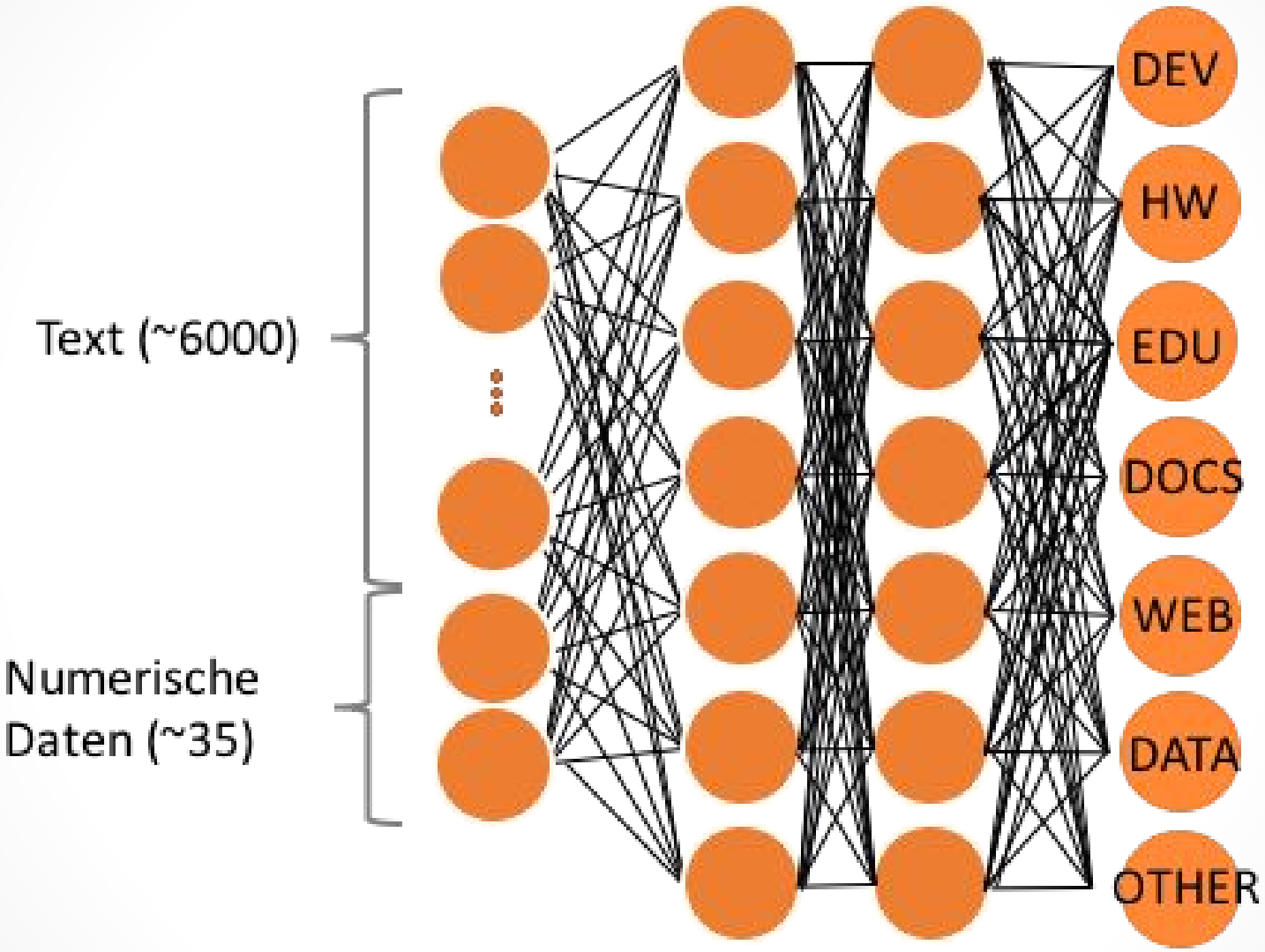
# Vorhersagemodelle

## Base Classifier

- Support Vector Machine
- Linear Classifiers
- Naive Bayes
- Decision Trees
- **Neuronale Netze**

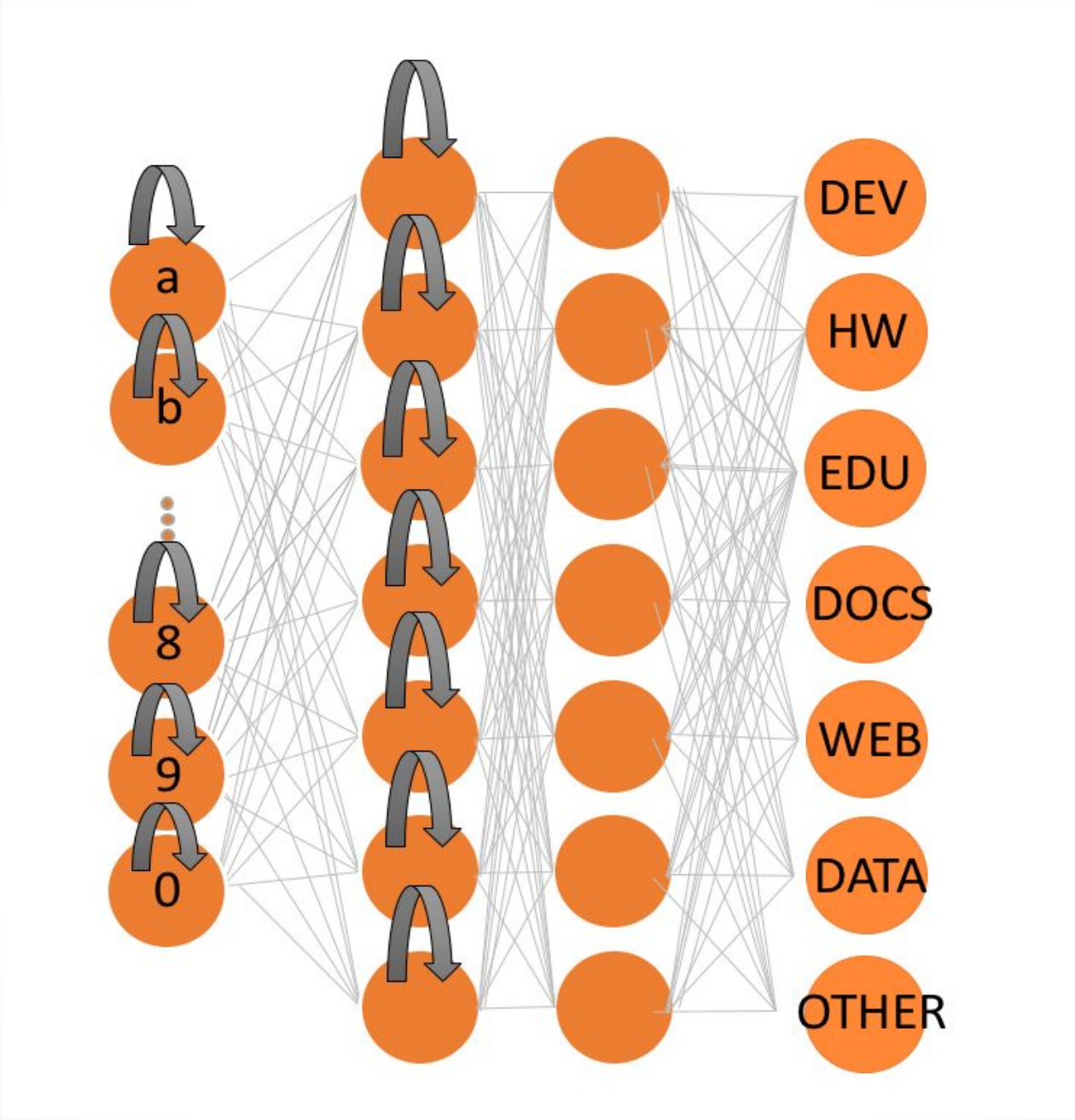
## Ensemble Classifier

- Boosting
  - Gradient Tree Boosting
  - AdaBoost
- Bagging
  - Random Forest
- **Stacking**



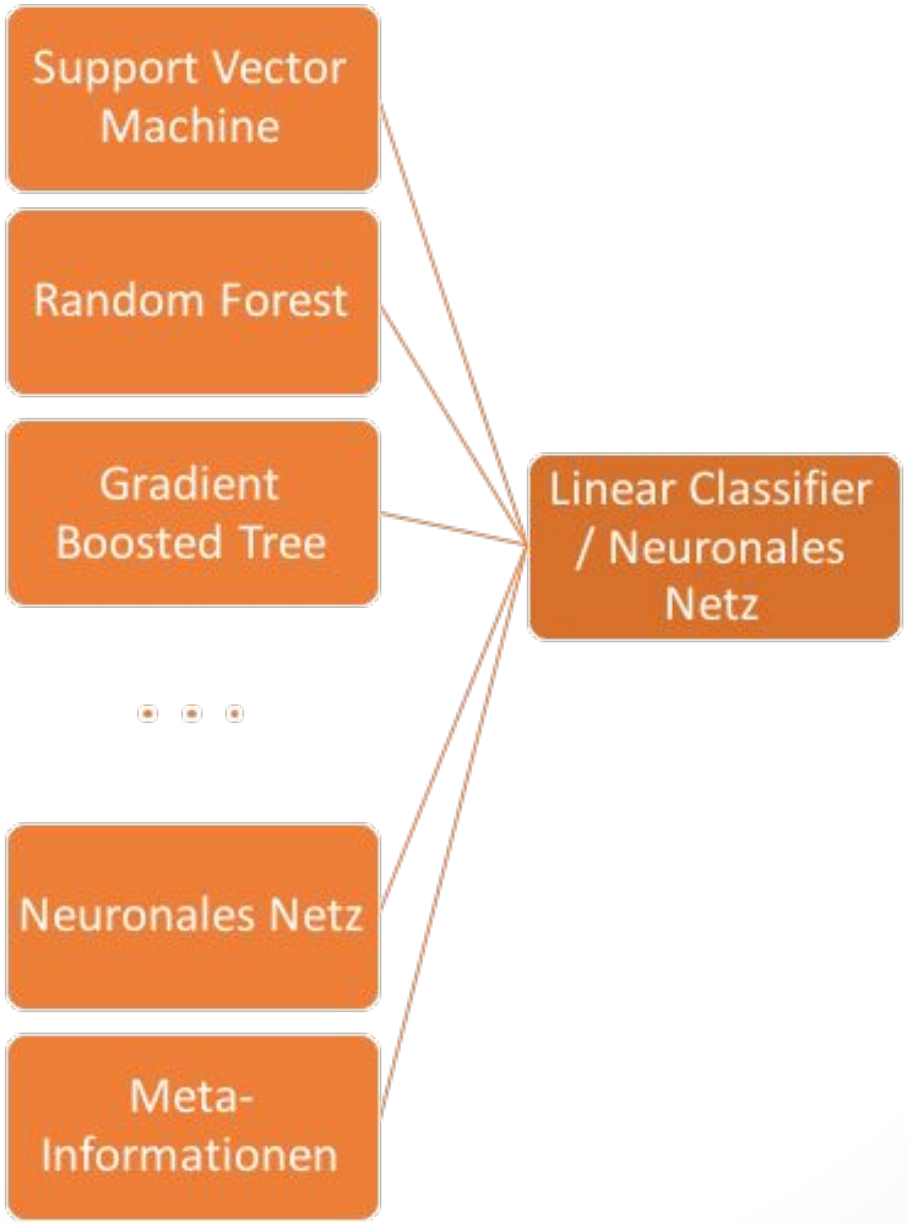

# Vorhersagemodelle - Neuronale Netze

- Feed Forward Network
- 2-3 hidden layer




# Vorhersagemodelle - Neuronale Netze

- Recurrent Neural Network
  ➔ LSTM
- Benutzt Repository-Name




# Vorhersagemodelle - Problemlösungen

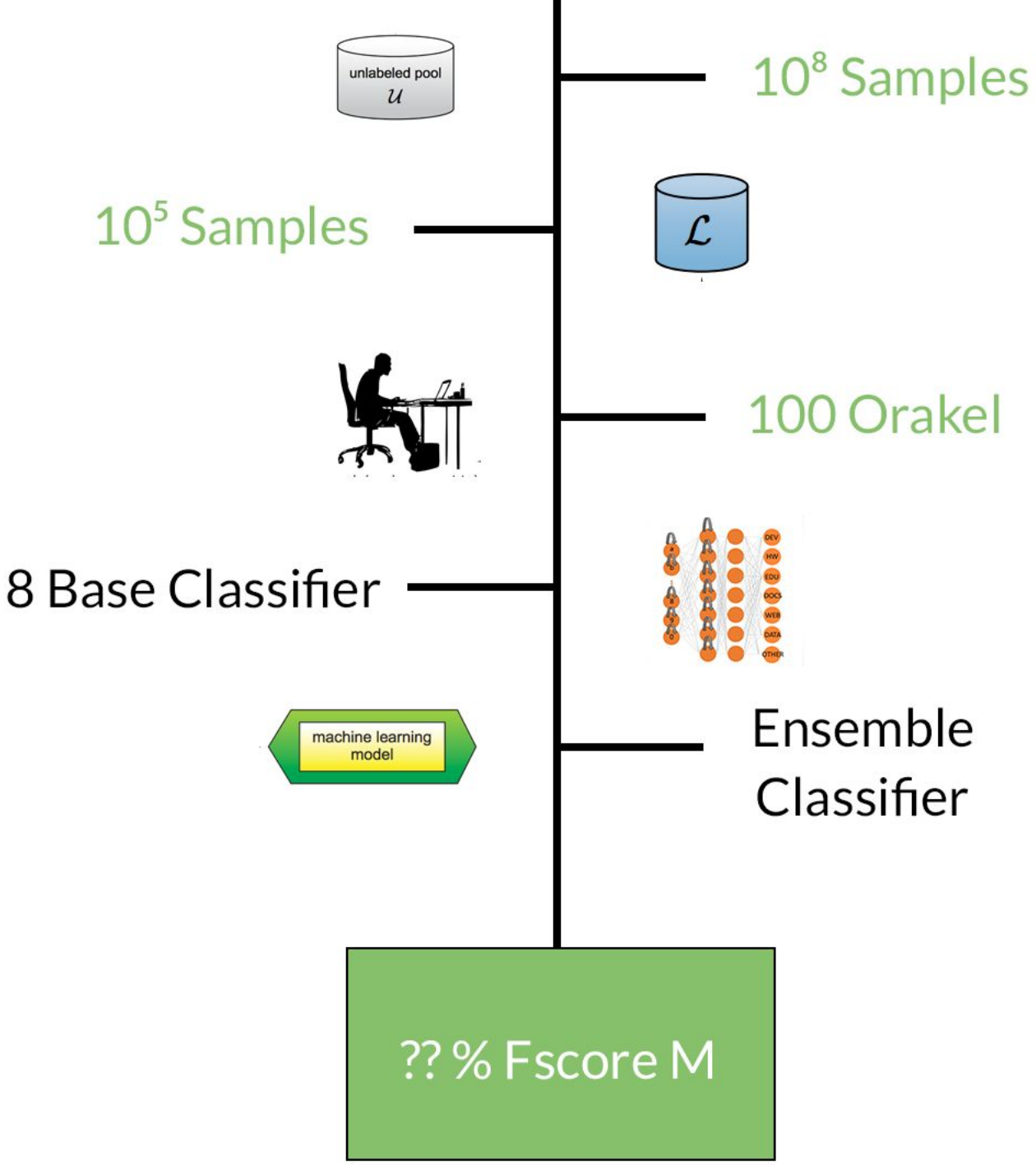
- 60% - Grenze

➔ Kombination aller Vorhersagen

- Class imbalance

➔ Class-Weights

- Overfitting / Vanishing Gradients




# Zusammenfassung

# Vielen Dank für Ihre Aufmerksamkeit



# Index der Zusatzfolien

# Unsere Anwendung - Ziele

- Sandbox für verschiedene Klassifikatoren
- Active Learning
- Features, Trainings- und Testdaten
- Eine GUI für verschiedene Zwecke?
- Plattformunabhängig
- Leichte Evaluation verschiedener Modelle

# Beispiel Decision Tree / Random Forest

# Features

## Text

- Repository - Name
- Autor
- Short - Description
- Readme
- Dateinamen
- Ordnernamen
- Dateitypen

## Numerisch

- # Commits, Forks, Folders / Files, ...
- ø Länge der Commit-Nachricht
- ø Levenshtein - Distanz der Ordner-/Dateinamen
- Verwendete Sprachen

## Boolean (1 oder 0):

- Besitzt Wiki
- Ist ein Fork, ...

**Measure table**

| Measure | Result |
|---|---|
| Precision M | 67.73 % |
| Recall M | 53.93 % |
| Fscore M | 64.43 % |
| Average Accuracy | 79.6 % |
| Error Rate | 62.24 % |
| Precision µ | 58.97 % |
| Recall µ | 58.97 % |
| Fscore µ | 58.97 % |

# Evaluation - Metriken

- Precision vs. Recall
- Interessanter: **Precision**
- Noch besser: **FScore M**

Zurück zum Index

Confusion matrix

| ▼ Class. \ Reference ► | DEV | HW | EDU | DOCS | WEB | DATA | OTHER | Total | Precision |
|---|---|---|---|---|---|---|---|---|---|
| DEV | 9 | 4 | 1 | 1 | 0 | 1 | 0 | 16 | 0.56 |
| HW | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| EDU | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 5 | 0.4 |
| DOCS | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 3 | 0.67 |
| WEB | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 4 | 0.75 |
| DATA | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| OTHER | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 10 | 7 | 6 | 3 | 3 | 2 | 0 | 31 | 0 |
| Recall | 0.9 | 0.29 | 0.33 | 0.67 | 1 | 0 | 0 | 0 | 0.58 |

# Evaluation - Konfusionsmatrix

- Ermöglicht detaillierte Analyse der Klassifikatoren
- Schön ersichtlich, wo der Klassifikator noch Probleme hat

Zurück zum Index

# Evaluation - Unsere Ergebnisse in Zahlen

- **66 %** FScore M mittels **Ensemble** Klassifizierer
- **60 %** FScore M mittels **einzelner** Klassifizierer

- Problematisch bleibt die Unterscheidung von:
  - HW von DEV
  - DEV und nicht DEV

- Die besten Ergebnisse liefern:
  - DEV
  - EDU
  - DATA

Zurück zum Index

# Evaluation - Interpretation

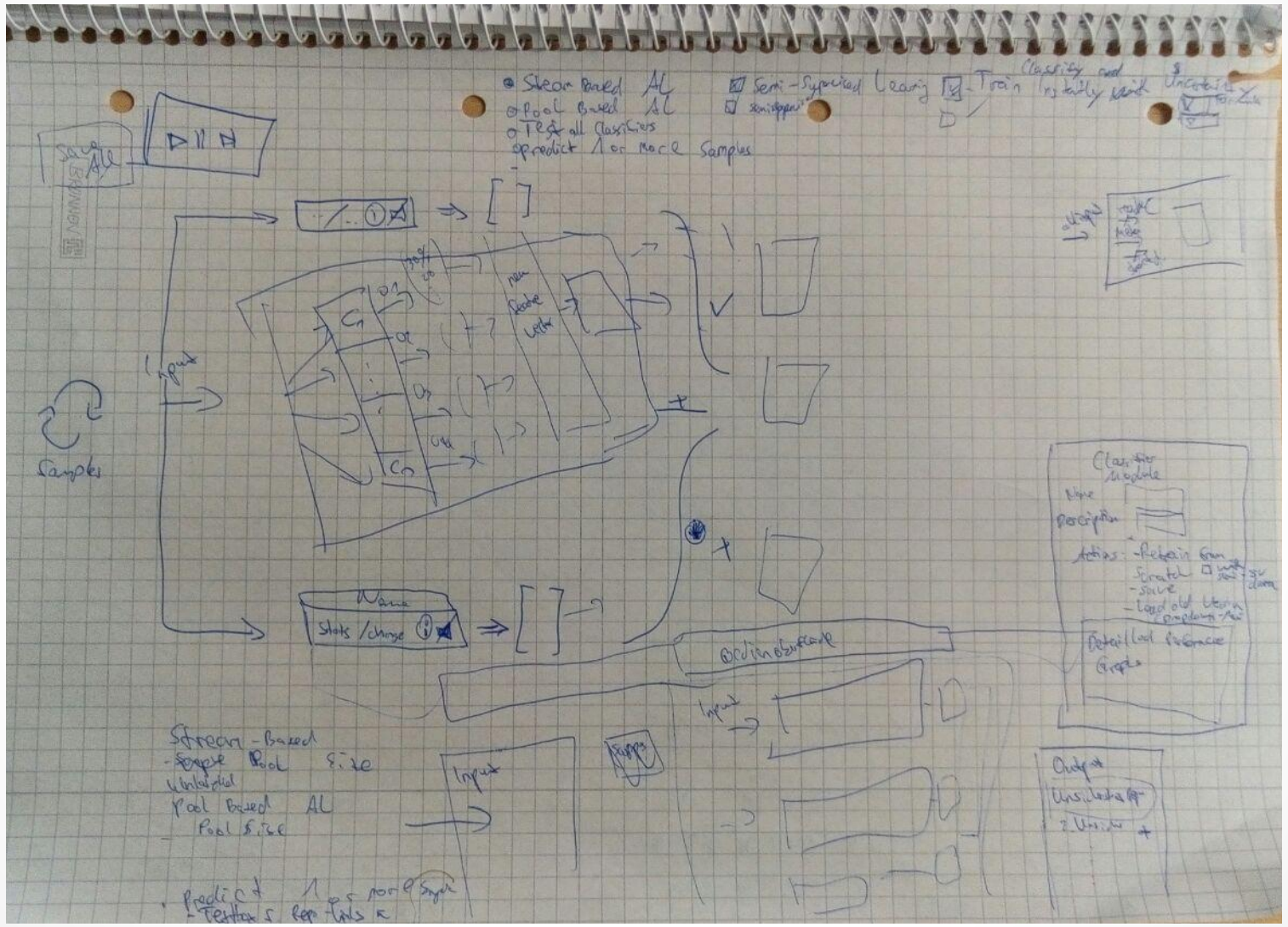
- Ergebnisse der Teams schwer zu vergleichen, da unterschiedliche Trainingsdaten und **unterschiedliche Interpretation** der Klassen
- auch unterschiedlichste Variationen und Klassifikatoren brachten keine Verbesserung => Möglichkeiten bei **Klassifikatoren ausgereizt**
- erwarten bessere Resultate mit höherer Anzahl **Trainingsdaten**

  + zudem weitere **Konsistenzprüfung** der bisherigen Trainingsdaten

  + bessere **Feature-Voranalyse**

- Für 10.000 - dimensionalen Inputvektor erstaunlich gute Ergebnisse, ggf. durch das **Active Learning**

Zurück zum Index

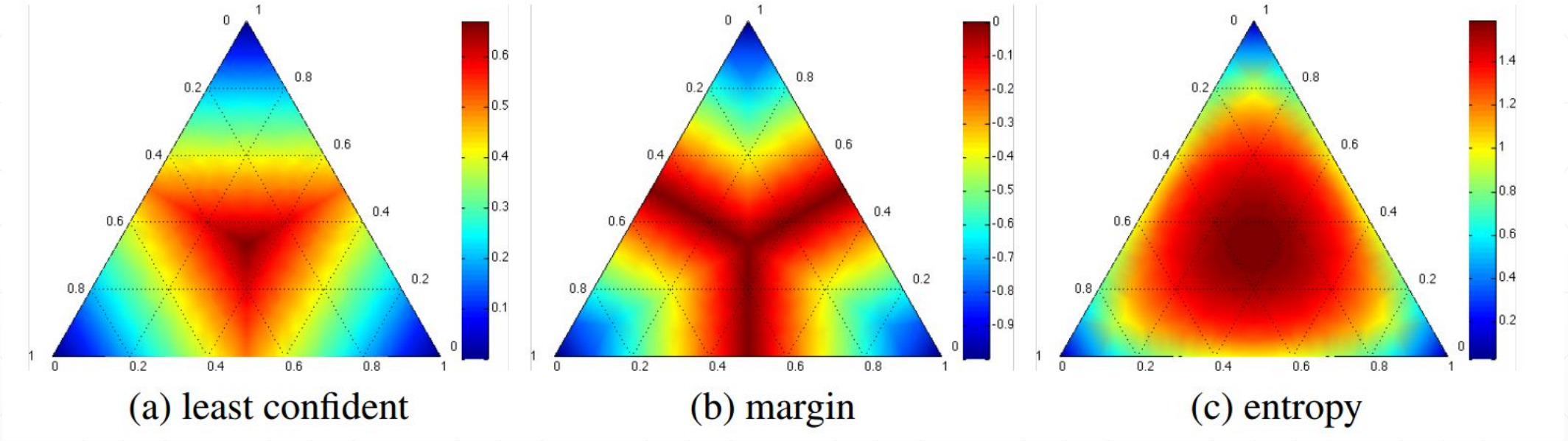



# Softwareentwurf

- Module als “Black Boxes”
- Enge Verknüpfung zwischen GUI und Logik - Schicht
- Model - View - Controller

Zurück zum Index

# Unsicherheitsformeln




*Quelle: Settles, Burr (2010), "Active Learning Literature Survey" Computer Sciences Technical Report 1648.*

Zurück zum Index

# Libraries und Technologien

- Libraries:
  - → numpy
  - → scipy
  - → cherrypy / paste
  - → bottle
  - → keras
  - → sklearn
  - → nltk
  - → gensim
  - → pattern
  - → theano
  - → demjson
  - → Vue.JS
- Technologien:
  - → PHP
  - → Python
  - → MySQL
  - → JavaScript
  - → HTML
  - → CSS

Zurück zum Index

# Besonderheiten und Herausforderungen

| Problem | Unser Ansatz |
| --- | --- |
| ▪ Unklare **Klassenbeschreibungen** | ➜ **“Saubere”** Trainingsdaten |
| ▪ **Ungleichgewicht** der Klassen | ➜ **Active Learning** |
| ▪ Vergleich verschiedener **Algorithmen** | ➜ **Modularisierung, Ensemble Learning** |
| ▪ **Hochdimensionaler** unbekannter Feature-Raum | ➜ **Feature engineering, Active Learning** |