# Learning a Sampling Strategy for Named Entity Recognition

Christopher Kümmel (845567)

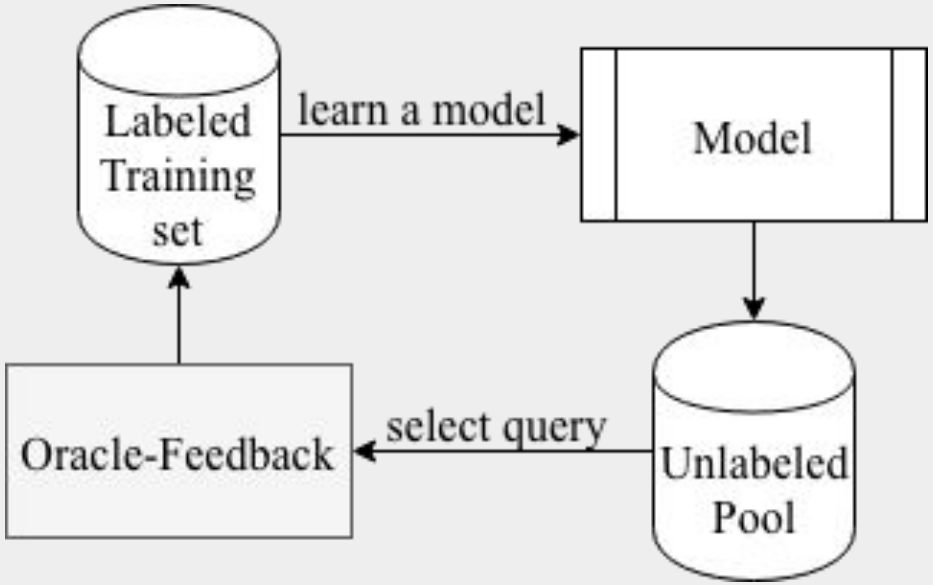
# Design Challenges in Named Entity Recognition

[Pers Pace], a junior, helped [Org Ohio State] to a 10-1 record and a berth in the [Misc Rose Bowl] against [Org Arizona State].

- Entity classification is not trivial
- Information of non-local features are needed
- Noisy data complicates the task
- Requires a well labeled dataset

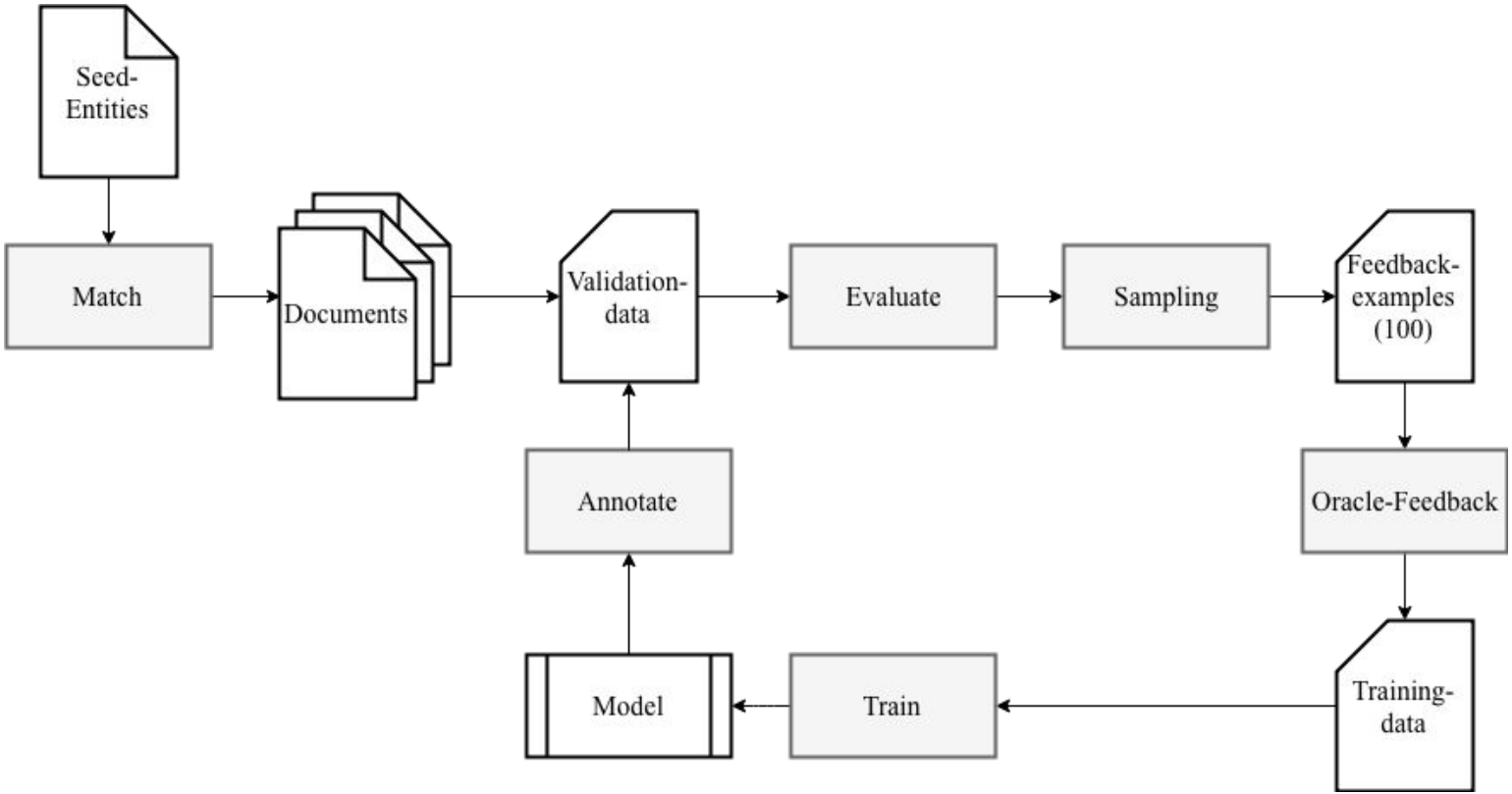
*[Ratinov and Roth - Design Challenges and Misconceptions in Named Entity Recognition, 2009]*
*[CoNLL2003 Shared-Task, 2003]*

# Active Learning




*[Burr Settles - Active Learning, 2012]*

# Active Learning Workflow



# Reduce Effort for Active Learning

- Improve performance by asking questions to an oracle
  - Loosen the strict requirements of well-defined datasets
  - Minimized initial costs turn into manual actions
- Generate data based on the oracles feedback

# Which data should be provided to the oracle?

# Classification Example

… Pace, a junior, helped [Loc Ohio] State to a 10-1 record
and a berth in the [Misc Rose Bowl] against [Org Arizona State]. ...

… [Pers Pace], a junior, helped [Org Ohio State] to a 10-1 record
and a berth in the [Misc Rose Bowl] against [Org Arizona State]. ...

*[CoNLL2003 Shared-Task, 2003]*

# **Baseline Strategies**
Pre-defined Strategies to Sample Sentences

# Random Sampling

Randomly selecting sentences out of the dataset

# Uncertainty Sampling

$$t_{i_{LC}} = argmax_{t_i} 1 - P_\theta(\hat{y}|t_i)$$

Where ŷ is defined as:

$$\hat{y} = argmax_y P_\theta(y|t_i)$$

For all tokens in a sentence x:

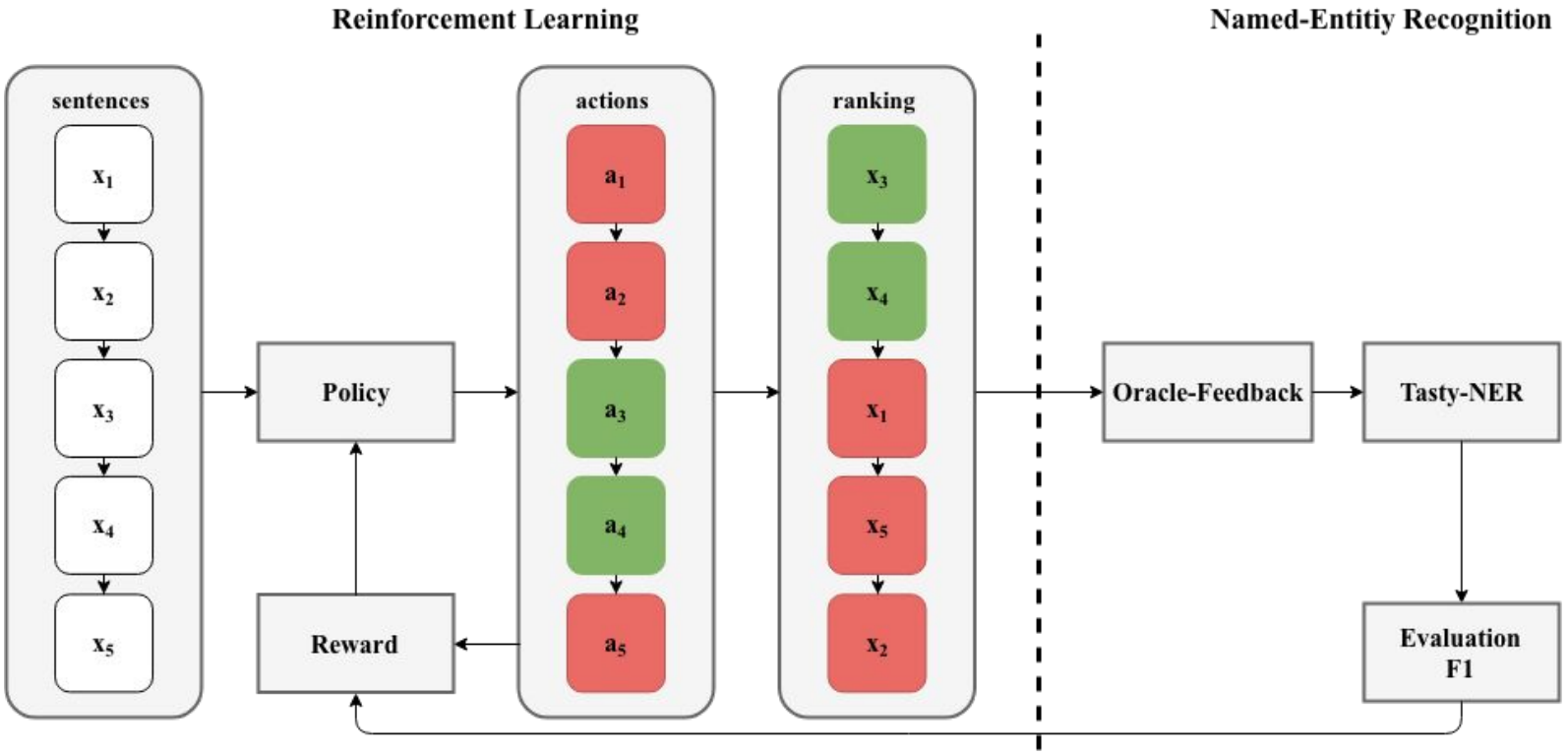
$$x_{i_{LC}} = \frac{\sum_{i=0}^{|x|} t_{i_{LC}}}{|x|}$$

*[Burr Settles - Active Learning, 2012]*

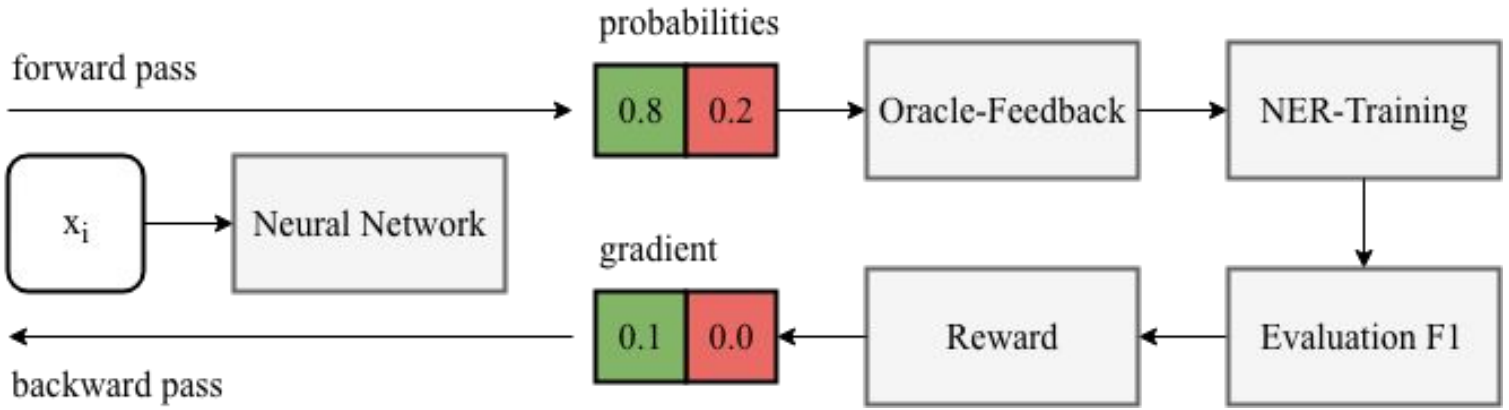
# **Reinforcement Learning Approach**

Sampling data with a Neural Network

# Markov Decision Process



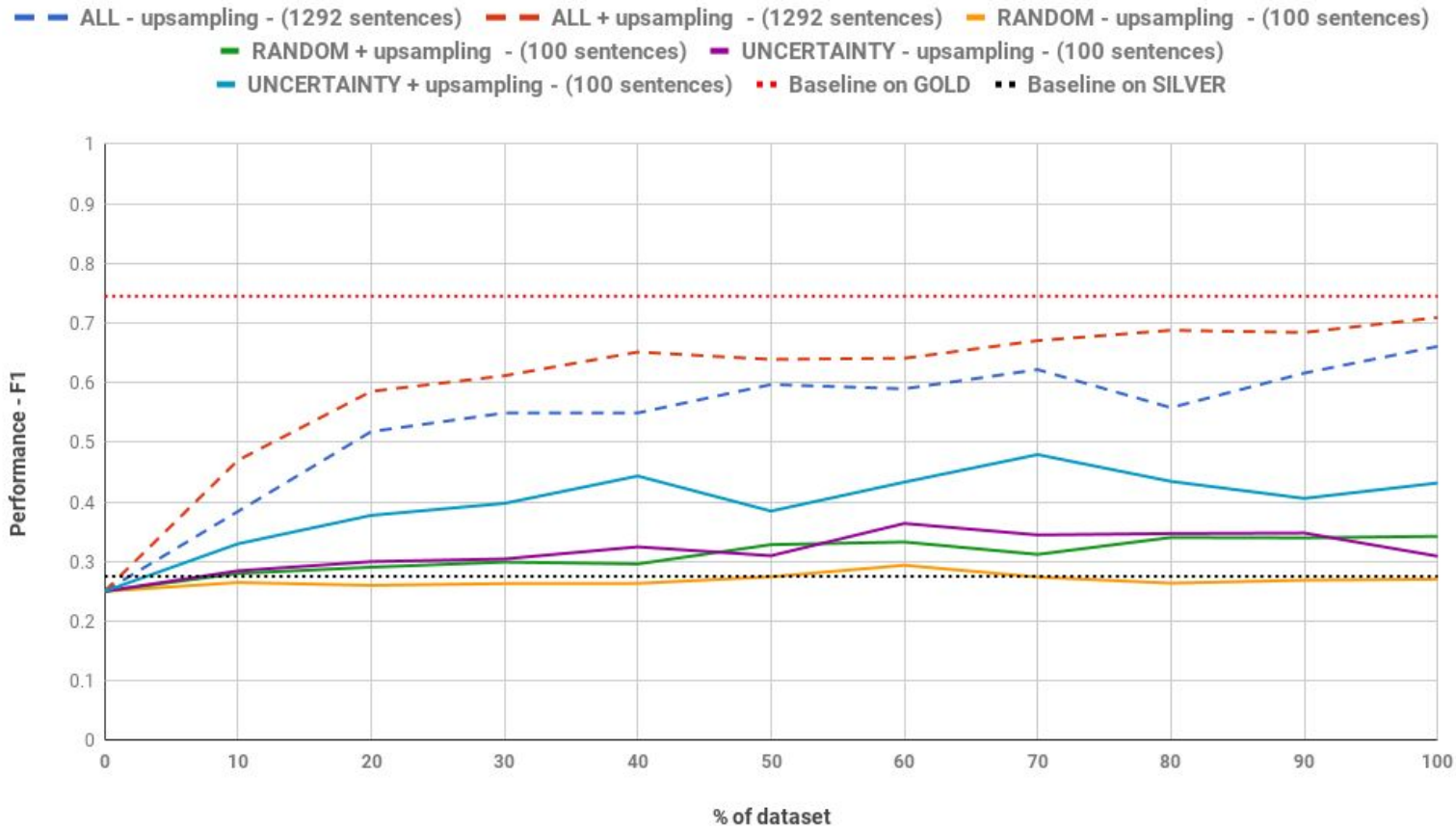
# Policy Gradient



# **Results**
How did the Strategies perform?

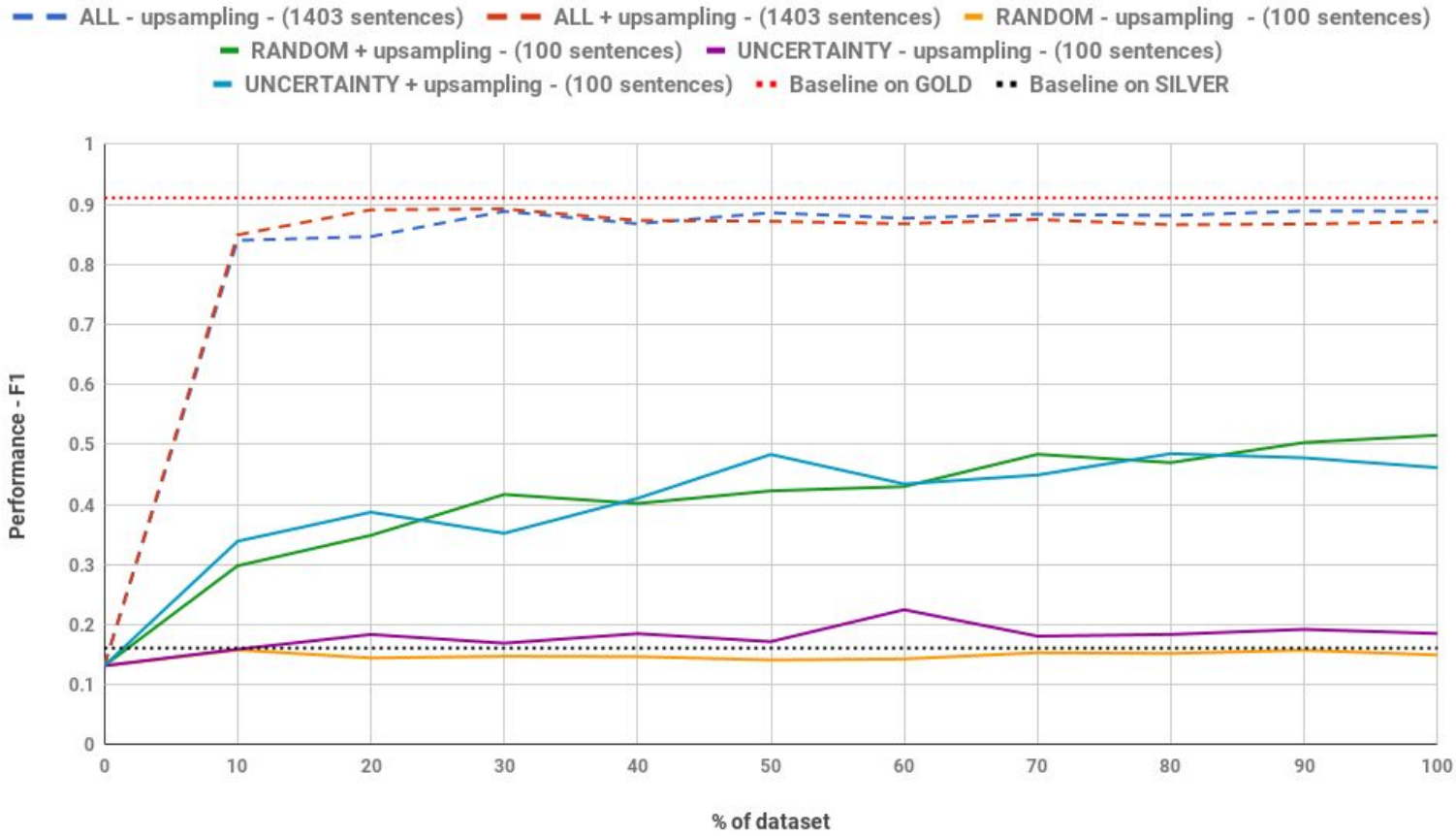
## Baseline Strategies on i2b2 - F1 Score

PRED vs. GOLD - Train on Silver & User



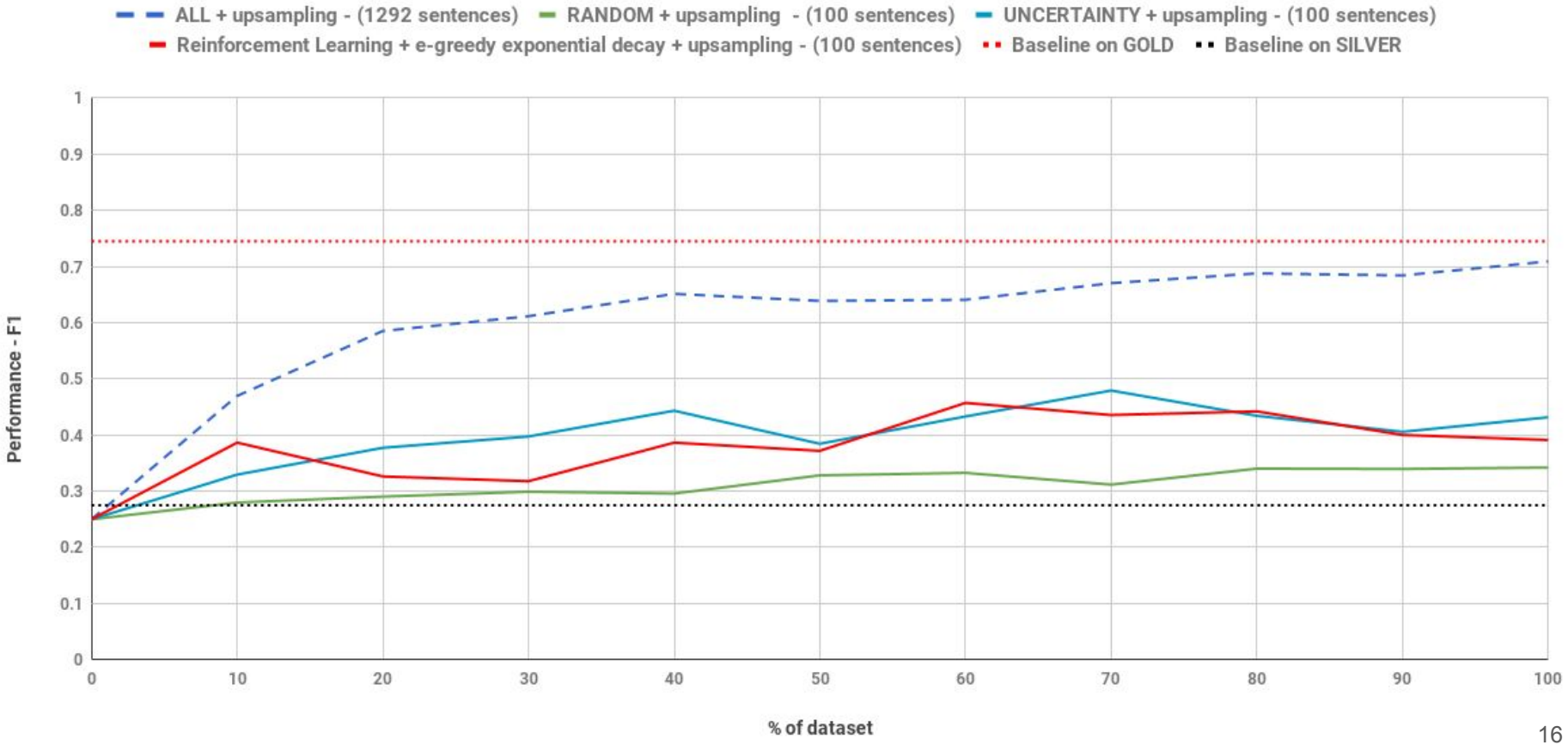
## Baseline Strategies on CoNLL2003 - F1 Score

PRED vs. GOLD - Train on Silver & User



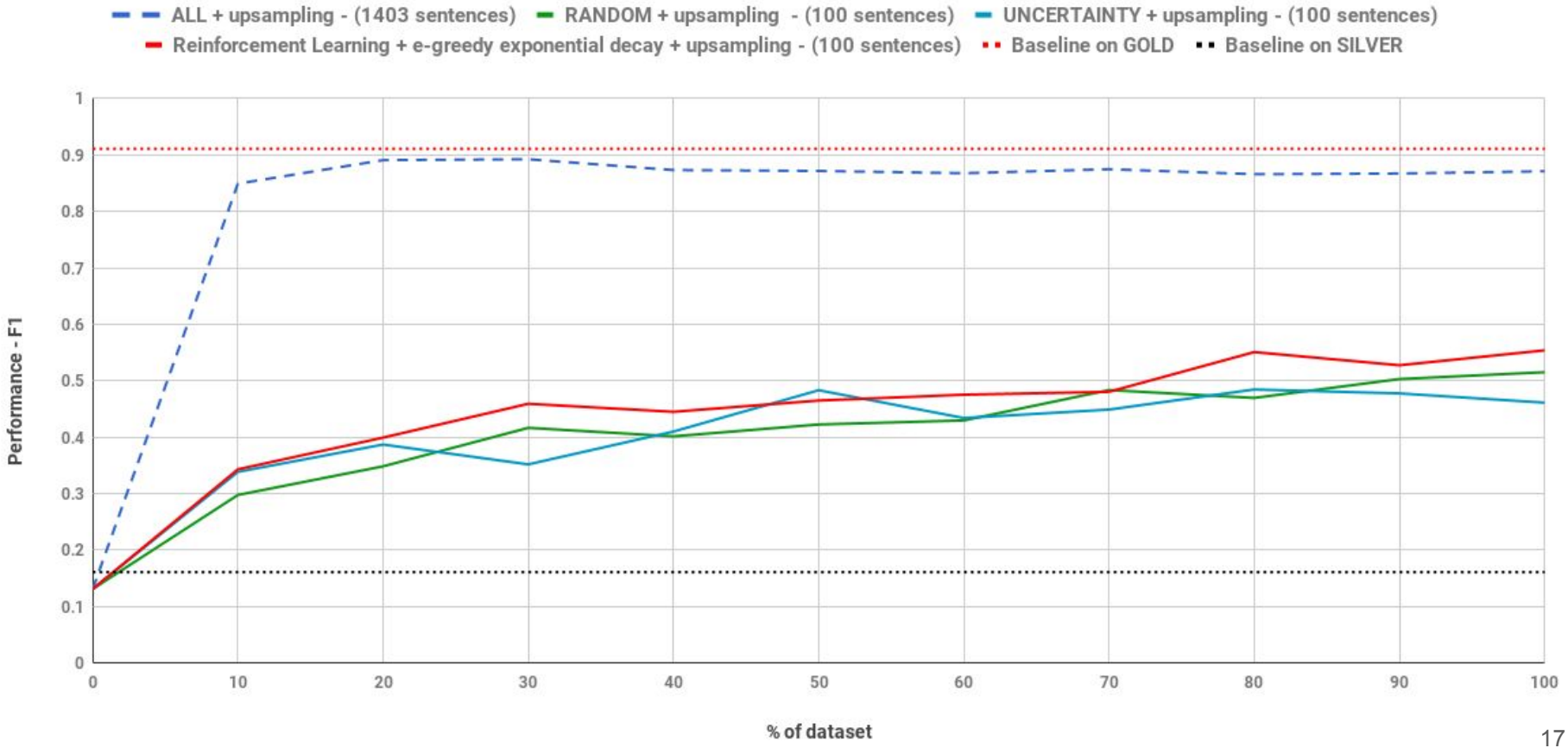
## Sampling Strategies on i2b2 - F1 Score

PRED vs. GOLD - Train on Silver & User



## Sampling Strategies on CoNLL2003 - F1 Score

PRED vs. GOLD - Train on Silver & User



# Conclusion

- RL-sampling does not improve performance on the same dataset
- Cross-domain usage can boost performance
- Does not justify the extra effort

# Outlook

- Execute tests with more episodes
- Include word and context features
- Integration with Active Learning GUI
- Deep Q Approaches

# References


1. Sebastian Arnold, Robert Dziuba, and Alexander Löser. **TASTY: Interactive Entity Linking As-You-Type.** In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 111–115, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.
2. Sebastian Arnold, Felix A. Gers, Torsten Kilias, and Alexander Löser. **Robust Named Entity Recognition in Idiosyncratic Domains.** arXiv:1608.06757 [cs], August 2016.
3. Meng Fang, Yuan Li, and Trevor Cohn. **Learning how to Active Learn: A Deep Reinforcement Learning Approach.** In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 595–605, 2017.
4. Christopher D. Manning and Hinrich Schütze. **Foundations of Statistical Natural Language Processing.** MIT Press, Cambridge, Mass., 2005.
5. Lev Ratinov and Dan Roth. **Design Challenges and Misconceptions in Named Entity Recognition.** In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL ’09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
6. Burr Settles. **Active Learning**. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.
7. Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. **Policy Gradient Methods for Reinforcement Learning with Function Approximation.** In Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999], pages 1057–1063, 1999.
8. Erik F. Tjong Kim Sang and Fien De Meulder. **Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition.** In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL ’03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
9. Özlem Uzuner,Brett R. South, Shuying Shen, and Scott L Du Vall. **2010i2b2/VA challenge on concepts, assertions, and relations in clinical text.** Journal of the American Medical Informatics Association, 18(5):552–556, September 2011.

# Learning a Sampling Strategy for Named Entity Recognition

Christopher Kümmel

- Generate training data based on the oracle’s feedback
- Minimize costs of training NER models
- Sampling strategy performance evaluation