References and Related Work for **NLP** Components

**I)**  Pasupat, Panupong, and Percy Liang. “Compositional Semantic Parsing on Semi-

Structured Tables.” In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1470–1480. Beijing, China: Association for Computational Linguistics, 2015.<https://doi.org/10.3115/v1/P15-1142>.

*Pasupat and Liang of Stanford discuss methodology for algorithmic processing of natural*

*language in dataset analysis. They suggest “knowledge graphs” that bind and process logical forms of queries. The authors use and provide a “Wiki Table Questions” for model training and outline floating parsing algorithms and pruning choices for a set of problems very similar to ours.*

**II)** Vakulenko, Svitlana, and Vadim Savenkov. “TableQA: Question Answering on

Tabular Data.” *ArXiv:1705.06504 [Cs]*, August 30, 2017.<http://arxiv.org/abs/1705.06504>. *Vienna University of Economics and Business.* Accessed November 26, 2019.<http://ceur-ws.org/Vol-2044/paper1/paper1.html>.

*Vakulenko and Savenkov examine a similar set of problems – still related to processing natural language queries in dataset analysis situations. They write about deep learning and neural network approaches, using end-to-end memory networks architecture.*

**III)** Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software.* August 2014,

Volume 59, Issue 10. [jstatsoft.org](https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf)

*Wickham writes about the uses and conventions of “tidy data,” a standardized way of formatting data that aims to lessen the “huge amount of effort [that] is spent cleaning data to get it ready for analysis.” Data is tidy if 1) every variable forms a column; 2) every observation forms a row; 3) every type of observational unit forms a table. We plan to ask our users to put their data in tidy form so there is less plumbing on our end to deal with the hairy problem of variable identification.*

References and Related Work for **Econometric** Components

**IV)** Phillips, Peter C. B. “Automated Discovery in Econometrics.” Econometric Theory 21, no. 1 (February 2005): 3–20. <https://doi.org/10.1017/S0266466605050024>.

*Phillips comments on state and future of automated discovery in economic statistical querying. He writes that his “experience with automated discovery algorithms in econometrics leads me to believe that these methods will play an important role in the future use of applied econometrics.” Pointing to a growing practical research agenda – as policy decisions become increasingly driven by empirical relationships – he sees a great degree of under-realized potential in this sort of automation.*

**V)** Glymour, Clark. “The Automation of Discovery.” Daedalus – MIT Press 133, no. 1 (January 2004): 69–77. <https://doi.org/10.1162/001152604772746710>.

*Like Phillips, Clark writes excitedly about the future of automated model selection in science. He points to a few examples from the late twentieth century in which the power of computing helped define important variable relationships. And he sees great promise in machines’ potential to grow their influence and further assist in empirical research. He goes so far as to compare traditional research methods to searching for needles in a haystack and automated model selection methods to “[running] a magnet through the haystack.”*

**VI)** Bollen, Kenneth A., and Daniel J. Bauer. “Automating the Selection of Model-

Implied Instrumental Variables.” Sociological Methods & Research 32, no. 4 (May 2004): 425–52. <https://doi.org/10.1177/0049124103260341>.

*Bollen appears to be the eminent scholar on automated instrument selection. He suggests a maximum likelihood technique for selecting variables for 2SLS analysis.*