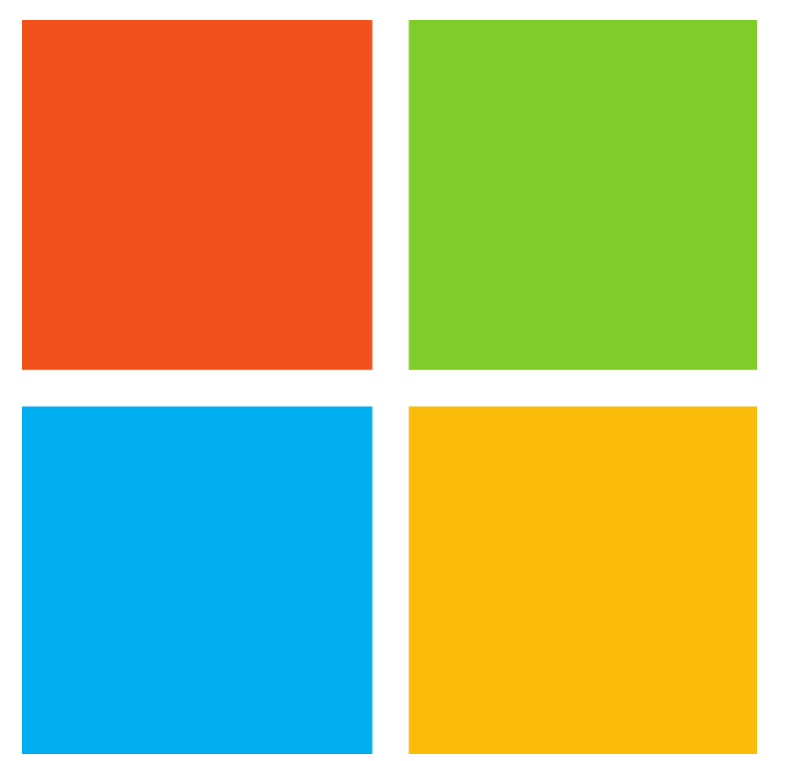


PROGRAMMING BY EXAMPLES FOR INDUSTRIAL DATA WRANGLING



Alex Polozov

polozov@cs.washington.edu

Sumit Gulwani

sumitg@microsoft.com

Microsoft PROSE team

prose-contact@microsoft.com

Wrangling & Cleaning: 80% of your “data science” time

Example: CS1951A “Data Science” online class at Brown University. Assignment #2, “Data Cleaning & Integration”:
“Transform this Super Bowl table in Wiki markup into a more usable (CSV) format. Here’s a template to get you started...”

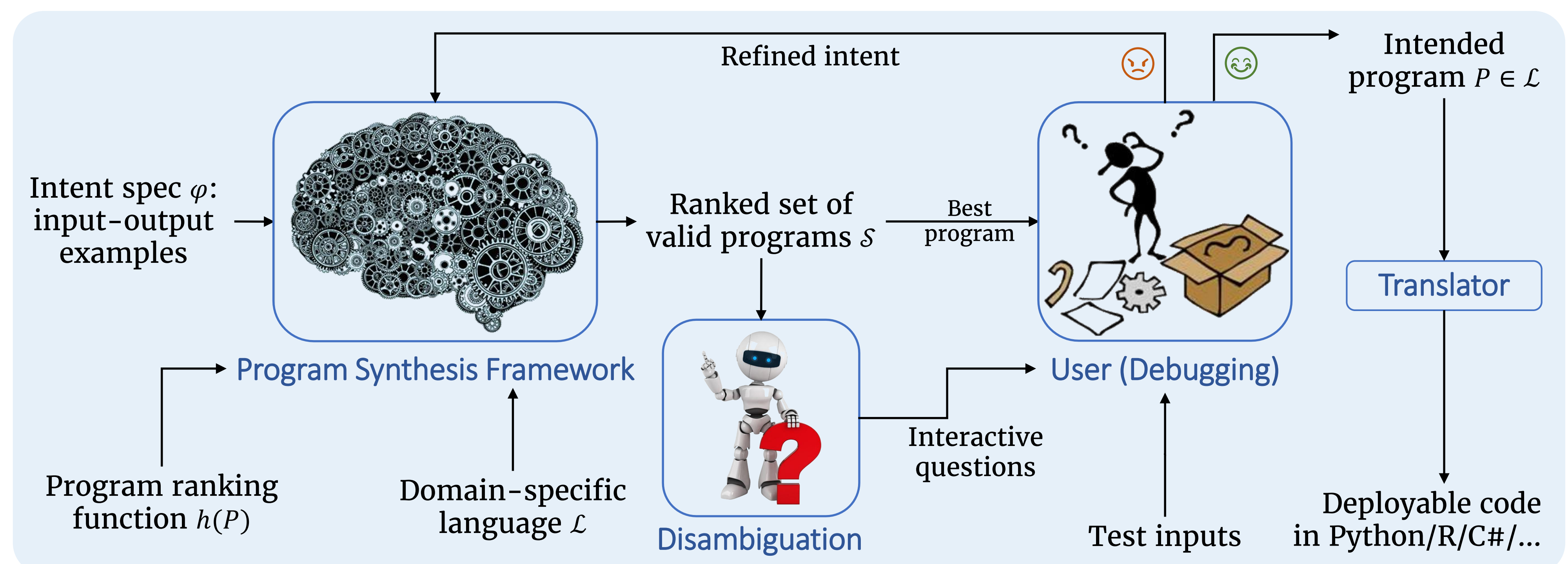
```
cat superbowl.txt | awk '$1=$1' ORS=' ' | sed 's/|-|/\n/g' | grep '^| style=\"text-align: center;\"' | grep -v "Championship"
```

PROSE Playground: Data Extraction by Examples

Excel Flash Fill: Data Transformation by Examples

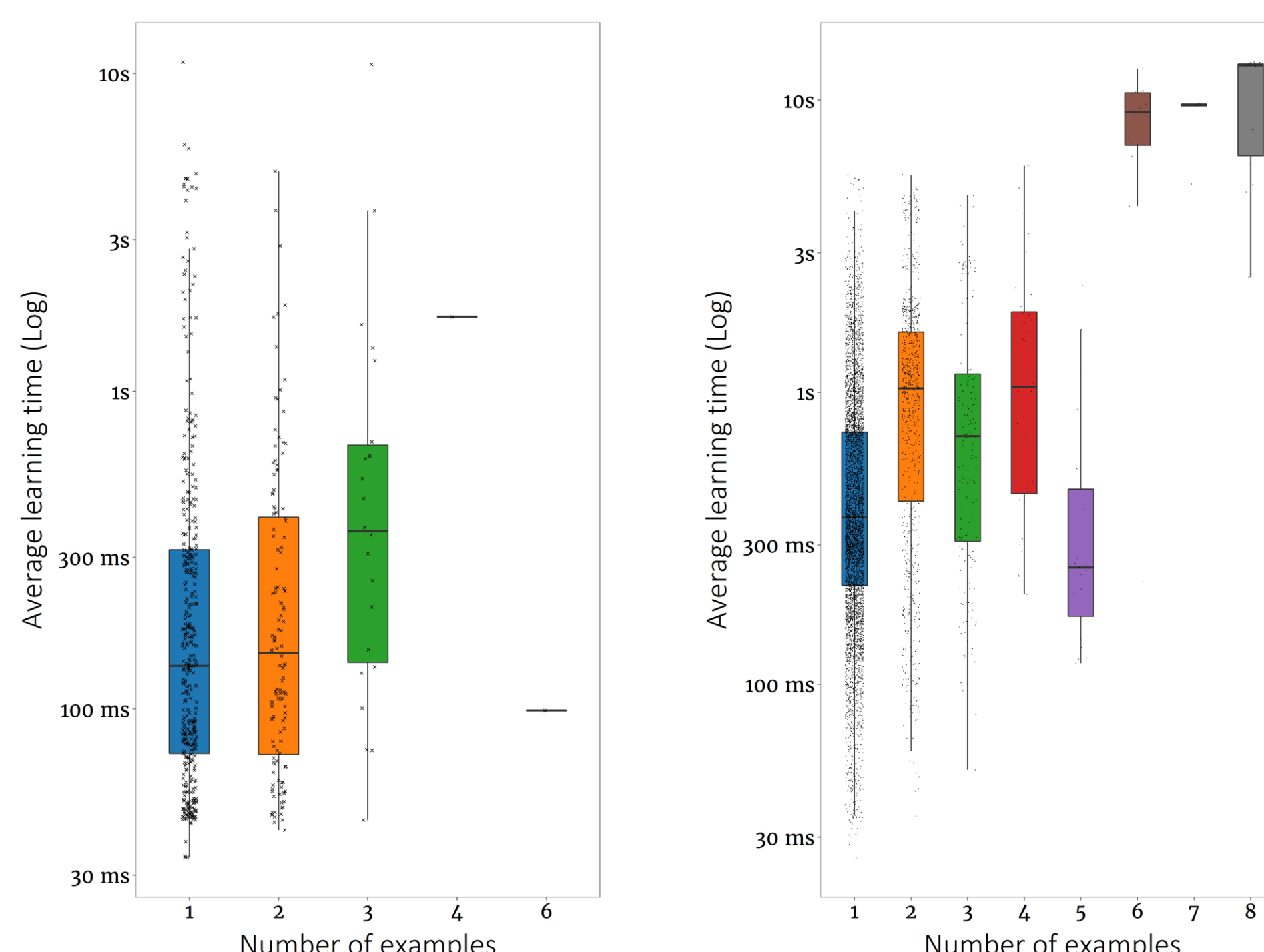
Powered by a universal framework for inductive program synthesis

- Accepts arbitrary **inductive specifications**: I/O examples (+ or X), constraints
- **Fast**: < 2 sec per iteration
- **Effective**: 1-6 examples until convergence
- **Interactive**: asks disambiguating questions:
 "Score" is currently ambiguous. Which highlighting is correct?
- Synthesizes programs in **real-life DSLs**:
 - Regexes, DOM processing, date/time
 - Conditionals, loops, recursion
 - Executable C# semantics



Typical performance

531 Flash Fill scenarios 6464 Flash Extract scenarios



References

1. Gulwani, S., 2011. Automating string processing in spreadsheets using input-output examples. In *ACM SIGPLAN Notices* (Vol. 46, No. 1, pp. 317-330). ACM.
2. Le, V. and Gulwani, S., 2014. FlashExtract: a framework for data extraction by examples. In *ACM SIGPLAN Notices* (Vol. 49, No. 6, pp. 542-553). ACM.
3. Andersen, E., Gulwani, S. and Popović, Z., 2014. Programming by Demonstration Framework Applied to Procedural Math Problems. Technical Report MSRTR-2014-61.
4. Polozov, O. and Gulwani, S., 2015. FlashMeta: A framework for inductive program synthesis. In *ACM SIGPLAN Notices*, 50(10), pp.107-126.
5. Mayer, M., Soares, G., Grechkin, M., Le, V., Marron, M., Polozov, O., Singh, R., Zorn, B. and Gulwani, S., 2015. User interaction models for disambiguation in programming by example. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (pp. 291-301). ACM.
6. Singh, R. and Gulwani, S., 2015. Predicting a correct program in programming by example. In *International Conference on Computer Aided Verification* (pp. 398-414). Springer International Publishing.
7. Rolim, R., Soares, G., D'Antoni, L., Polozov, O., Gulwani, S., Gheyi, R., Suzuki, R. and Hartmann, B., 2016. Learning Syntactic Program Transformations from Examples. *arXiv preprint arXiv:1608.09000*.

<https://microsoft.github.io/prose>

<https://microsoft.github.io/prose/playground>

<https://microsoft.sharepoint.com/teams/msprose>

- Deployed in numerous Microsoft products:



- Domain development:
 - No synthesis expertise necessary
 - 1-2 months until production-ready
- Machine-learned ranking, often one-shot
- Novel synthesis algorithm: **backpropagation**
- Applications beyond data wrangling!
 - Automatic code refactoring
 - Intelligent tutoring systems