**PROVISIONING THE 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...”*

```bash
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**

- Accepts arbitrary inductive specifications: I/O examples (+ or ×), constraints
- Fast: < 2 sec per iteration
- Effective: 1–6 examples until convergence
- Interactive: asks disambiguating questions: “Score” is currently ambiguous. Which highlighting is correct?

**Typical performance**

- 531 Flash Fill scenarios
- 6464 Flash Extract scenarios

**Program Synthesis Framework**

- Program ranking function $h(P)$
- Domain-specific language $\mathcal{L}$
- Disambiguation
- Refined intent
- Ranked set of valid programs $s$
- Best program

**User (Debugging)**

- Interactive questions
- Test inputs
- Deployable code in Python/R/C#/...

**Power**

- Deployed in numerous Microsoft products:
  - Convert-String, ConvertFrom-String
  - 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

**References**


https://microsoft.github.io/prose
https://microsoft.github.io/prose/playground
https://microsoft.sharepoint.com/teams/msprose