C²METADATA
Continuous Capture of Metadata

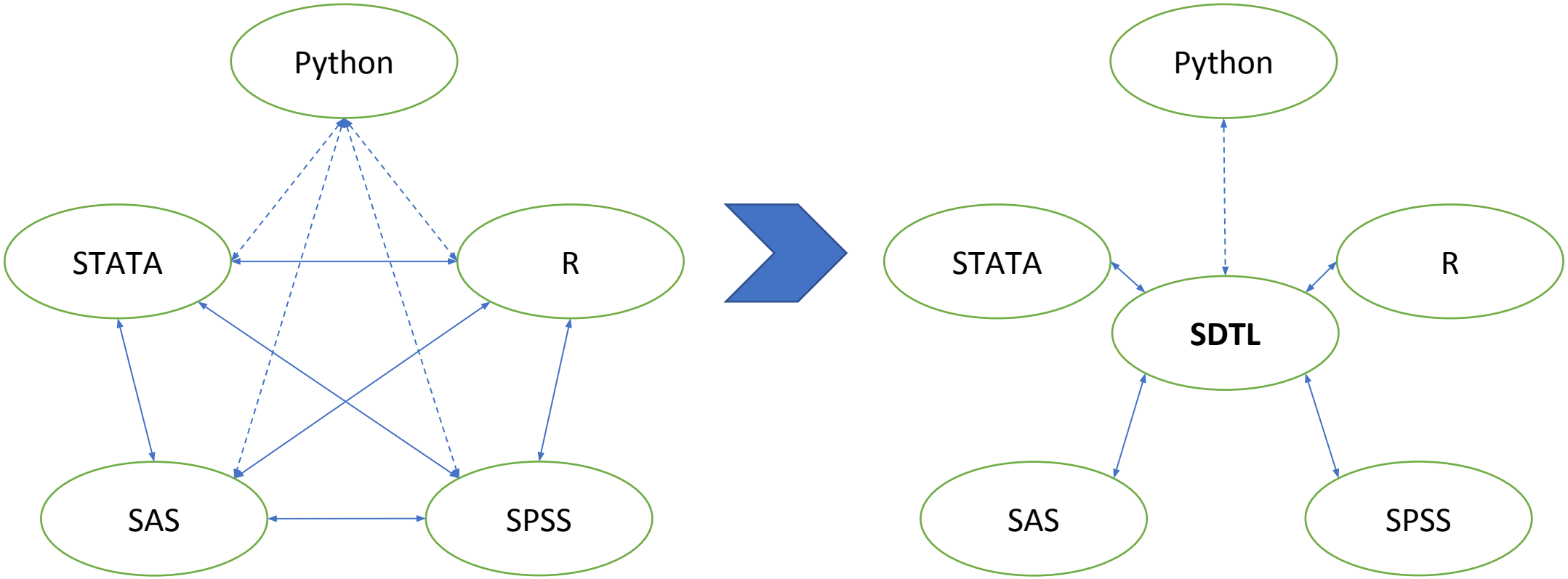
Design Motivation and Impact on the Database Community

Jie Song
Computer Science Engineering
University of Michigan

Supported by the Data Infrastructure Building Blocks (DIBBs) program of the National Science Foundation through grant NSF ACI-1640575

From Language Conversion To Data Transformation Conversion

- Inter-conversion between each of the languages are expensive, O(n^2)
- Bridge by an intermediate language, O(n)



From Language Conversion To Data Transformation Conversion

- Traditional general programming language conversion
  - A simple syntax replacing system
  - More recent works begin to consider different levels of conversion
    - E.g., convert code of a procedural language into a purely object oriented language
- Statistical programming languages
  - Simpler programming languages that are mostly procedural
  - R, Python are more general purpose while SPSS, STATA and SAS are limited
  - Focus on data transformation and analysis

From Language Conversion To Data Transformation Conversion

D1

V1	V2
2	3
1	4

V3=V2+1

D2

V1	V2	V3
2	3	4
1	4	5

Sort by V1

D3

V1	V2	V3
1	4	5
2	3	4

Rename V1 as Index

D4

Index	V2	V3
1	4	5
2	3	4

V4=Count(V1) groupby V1

D5

V1	V4
1	1
2	1

R: D1$V3=D1$V2+1 **Or**

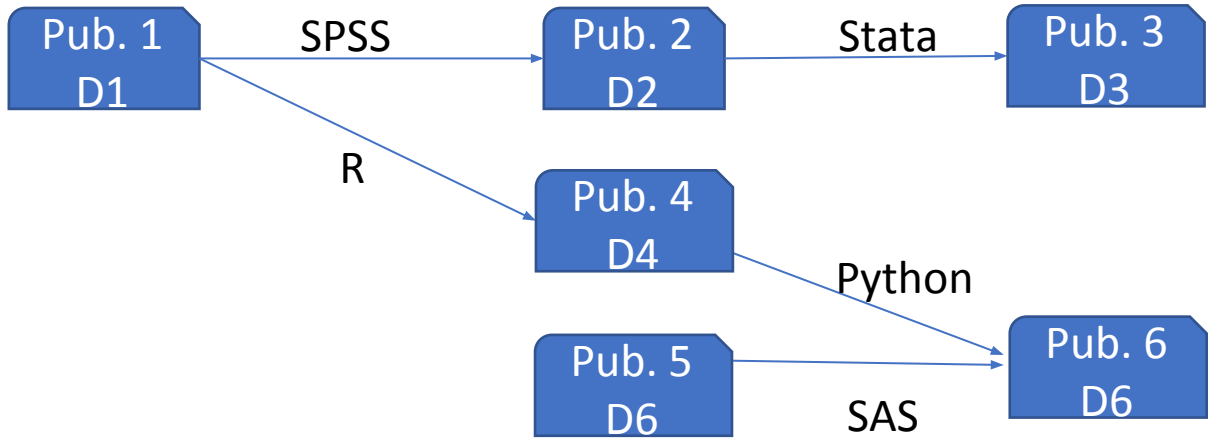
```
for (i in 1:nrow(D1))
  D1$V3[i] = D1$V2[i]+1
```

Stata: generate V3=V2+1

- Data transformation as a directed acyclic graph
  - Data as the node and transformation as the edge
- Semantic level conversion
  - Regardless of the underlying syntax used by hiding implementation details
  - From higher level, e.g. data set level, to lower level, e.g., column/row level, metadata level
- Graphical Visualization for better understanding
  - Intermediate result when processed with the raw data
  - Data lineage or provenance
  - Code association with transformations

The Future of C^2Metadata

- A systematic guideline for building converter that maps from a new statistical language to SDTL
  - Phrase-based statistical translation
- Promote research in
  - Data services, e.g. data sharing and data publication linage construction
  - Transformation code transfer and reuse
  - Data preparation such as
    - data exploration
    - data quality evaluation
    - data integration



References

- George, D., Girase, P., Gupta, M., Gupta, P., & Sharma, A. (2010). Programming Language Inter-conversion. International Journal of Computer Applications, 1(20), 68–74. http://doi.org/10.5120/419-619
- Karaivanov, S., Raychev, V., & Vechev, M. (2014). Phrase-Based Statistical Translation of Programming Languages. dl.acm.org (pp. 173–184). ACM. http://doi.org/10.1145/2661136.2661148
- Kontogiannis, K., Martin, J., Wong, K., Gregory, R., Müller, H., & Mylopoulos, J. (2010). Code migration through transformations (pp. 201–213). Presented at the CASCON First Decade High Impact Papers, New York, New York, USA: ACM Press. Http://doi.org/10.1145/1925805.1925817