Life After Record Linkage:

Tackling the Downstream Task with Error Propagation

Andee Kaplan

Duke University
andrea.kaplan@duke.edu

October 10, 2018
http://bit.ly/kaplan-ncsu-2018

Outline

1. Record linkage and the downstream task
   What is record linkage, and how is it useful?
2. Representative records
   How to represent a linked dataset via prototyping?
3. Results
   How well does prototyping work for different data situations and downstream tasks?
4. Prototyping in practice
   Some guidelines for using prototyping and directions for future work
5. Other work
   What else have I been up to?

Record Linkage and the downstream task

What is record linkage?

Record linkage is the process of merging noisy databases to remove duplicate entities without the use of a unique, identifying attribute

Records to link

name	bill
Lula Monahan	7193.77
Shane Nikolaus	56.60
Darcy Orn	314.86
Merritt Littel	0.00
Tera Greenfelder-Smitham	0.00

Name	Glucose	BP	Insulin	Age
Luila Molnahan	117	92	0	38
Shane Nikolaus	131	68	166	28
Dgarcy Orn	95	72	0	27
Merritt Littel	116	78	180	25
Tera Grewenfelder-Smitham	134	72	0	60

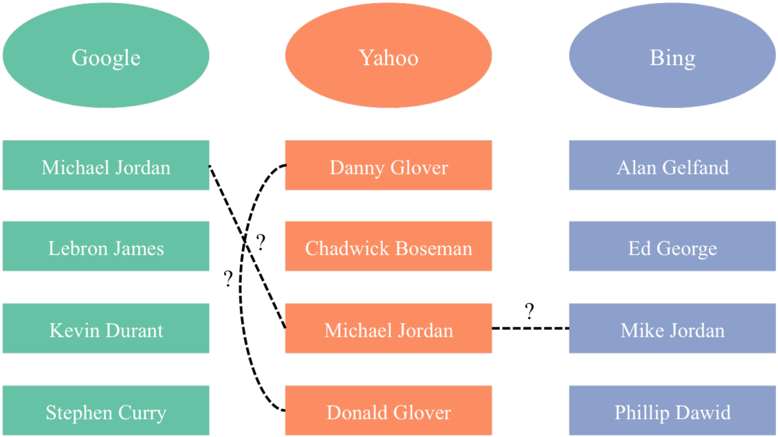
Link on names?

Client name	Patient name
Lula Monahan	Luila MoInahan
Shane Nikolaus	Shane Nikolaus
Darcy Orn	Dgarcy Orn
Merritt Littel	Merritt Littel
Tera Greenfelder-Smitham	Tera Grewenfelder-Smitham

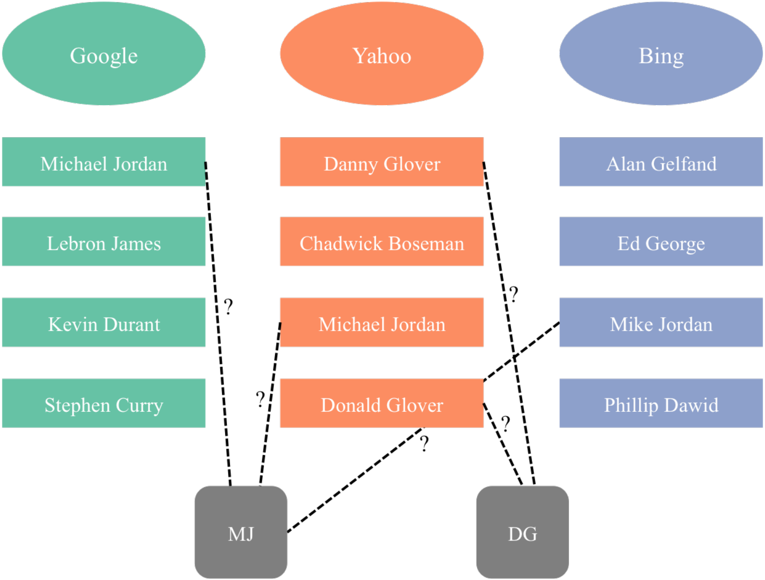
Probabilistic record linkage

Put lit review here

Probabilistic record linkage

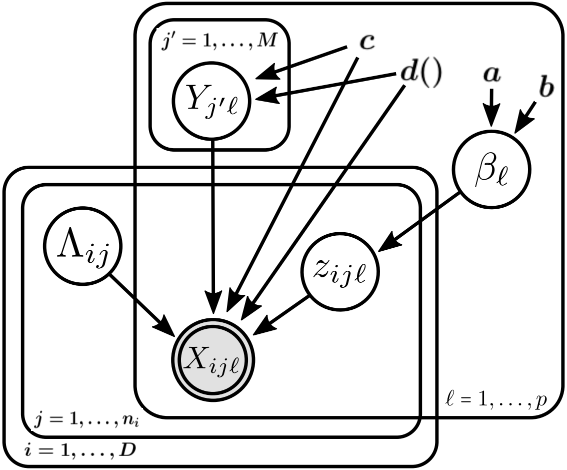


Latent clustering approach



Bayesian hierarchical model

Proposed by Steorts (2015) with a package on CRAN (`blink`) (Steorts 2017)



Downstream task

Representative records

Results

Prototyping in practice

Other work

Thank you

- Slides – http://bit.ly/kaplan-ncsu-2018
- Contact
  - Email – andrea.kaplan@duke.edu
  - Web – http://andeekaplan.com
  - GitHub – http://github.com/andeek

Appendix

Appendix Slide

References I

Steorts, Rebecca. 2017. *Blink: Record Linkage for Empirically Motivated Priors*. https://CRAN.R-project.org/package=blink.

Steorts, Rebecca C. 2015. "Entity Resolution with Empirically Motivated Priors." *Bayesian Analysis* 10 (4). International Society for Bayesian Analysis: 849–75.