Discussion of “Entity Resolution: Measuring and Reporting Quality” by Rebecca Steorts

Goal: Combine sets of files to create larger, cleaner sets of data for policy analyses.

Economics- Companies

```
 Agency A            Agency B

fuel          ------>  outputs
feedstocks    ------>  produced
```

Health- Individuals

```
 Receiving             Agencies
  Social Benefits       B1, B2, B3

 Incomes               Agency I

 Use of Health         Agencies
  Services              H1, H2
```

File A	Common	File B
A_{11} , ... A_{1n}	Name1,Addr1,DOB1	B_{11},...B_{1m}
A_{21} , ... A_{2n}	Name2,Addr2,DOB2	B_{21},...B_{2m}
.		.
.		.
.		.
A_{N1} , ... A_{Nn}	NameN,AddrN,DOBN	B_{N1},...B_{Nm}

Issues:

1. Clean-up original source files (**A** and **B**)
   a. Modeling/edit/imputation
   b. Data linkage (duplication)
2. Create merged file (data linkage)
3. Adjust statistical analysis for linkage error
   (research problem, easiest 5-20% solved)
   a. Enhancements to current elementary models
   b. Extensions using modeling/edit/imputation and statistical matching

For 100s of millions of records, computational algorithms need to be 2-6 orders faster than those used previously.

Cautions

5% error in each of files **A** and **B**

5% matching error

Errors are additive

15% error in $(A_{j1}, \ldots A_{jn}, B_{j1}, \ldots B_{jm})$ data

Are there any analyses that are possible?

If the error is reduced to 5% overall, what analyses are possible?

How will we even know how much error is in the files?