Messy spreadsheet example

NAME	EMAIL	FIRST CONTACT	CLASS	PACKET SENT	FOLLOW UP	PACKET RECEIVED
Jill	jill@gmail.com	9/8/2015	A	9/9/2015	9/12/2015	
Judy	judy@gmail.com	9/1/2015	B	9/2/2015	9/4/2015	
John	john@gmail.com	8/15/2015	C	8/16/2015	8/18/2015	
Mark	mark@gmail.com	7/2/2015	A	7/3/2015	7/11/2015	
Steve	stevee@gmail.com	6/9/2015	B	6/10/2015	6/22/2015	
Amy	amy@gmail.com	5/12/2015	C	5/13/2015	5/17/2015	

Can we automatically improve it?

E.g. The values of the first column occur in the second column as substrings, which means they are related and can be grouped out

Ideas

1. column clusteing – closeness metrics, substrings
2. pattern as new tables or values?
3. google refine – refinment language
4. canonical forms extracted for particular data instances

Emerging from

1. Databases – extract schema and functional dependencies from the data
2. Predicate invention – create new table, columns or values for common values or patterns, such as substrings e.g. domain in the email such as gmail
3. Data cleaning – generalize tricks from excel and google docs cleaning methods
4. Data compression – we can refer to a range of columns see first

   figurem, like:

A:	Steve	Ann
B:	Alex	John
C:	Peter	Amy

5. Interactive mining – we can give explanation for transformation adn allow user to cancel them out

What is really needed

1. Objective: can we define better?
2. What methods we gonna use, some relational learning?
3. Can we find good convincing examples?
4. In general, what kind of experimental evidence can we provide?