MIS 0855 Spring 2015 – Data Science

Day 18 – Dirty Data

Min-Seok Pang

Management Information Systems
Fox School of Business, Temple University
minspang@temple.edu
Feb. 23rd, 2015

Fox School of Business
TEMPLE UNIVERSITY

Data Users Spend 50% of Their Time In

- searching for data
- correcting errors
- verifying correctness

Find Dirty Stains in This Data!

Customer ID	Customer First Name	Customer Last Name	Address	City	State	Zip	Phone
1771	Larry	Shimk	143 S.	Denver	NY	178908	911
1771	Caroline	Shimk	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Shimk	Caroline	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Heather	Schwiter	55 N. W. S. Miss	LaGrange	GA	14321	716-333-4567
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1773	Justin	Justin	34 Kerry Rd.	Littleton	CO	98987	716-67-9087
1774	Pam		66 S. Carlton	North Glen	CO	98765	343-456-6857
1775	D.	Fernandez	3514 S. Main	Denver	CO	80252	303-333-8965
1776	PepsiCo		15365 K St. NW	Washington	DC	20035	202-353-1535
1777	Sam	Esteban	4413 Madison Rd	Ann Arbor	MI	48109	734-140-2531 ext 354
1778	Caroline	Smith	143 N. West St.	Buffalo	NY	14321	716-333-4567

Why Does Data Get Dirty?

- Think of Ms. Pamela Smith O'Brien
  - How many different names can she have?
- How about an address?
  - 1303 North Taylor Street, Apt. #102, Philadelphia, Pennsylvania 19123, USA
  - How many different addresses can be *valid*?

Origin of Dirty Data

- Measurement can be inaccurate
  - Name – a person's name or a company's name?
- Instrument : the question may be wrong or ambiguous
  - Phone number – home, work, or cell?
- Consistency : the question can be answered inconsistently
  - Address – South Taylor St., S. Taylor Street, So Taylor St., S. Taylor St.

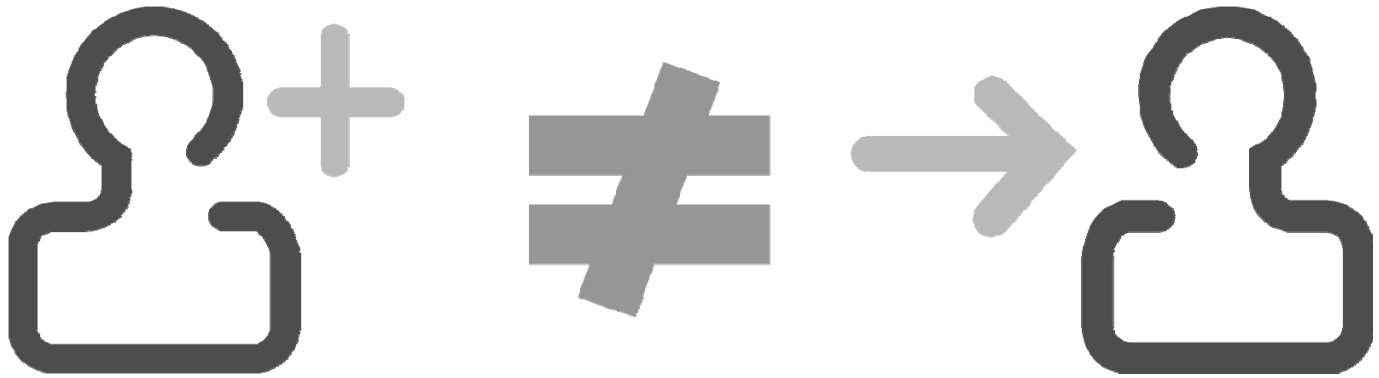
Five Characteristics of High-Quality Data

- Accuracy, Completeness, Consistency, Uniqueness, Timeliness

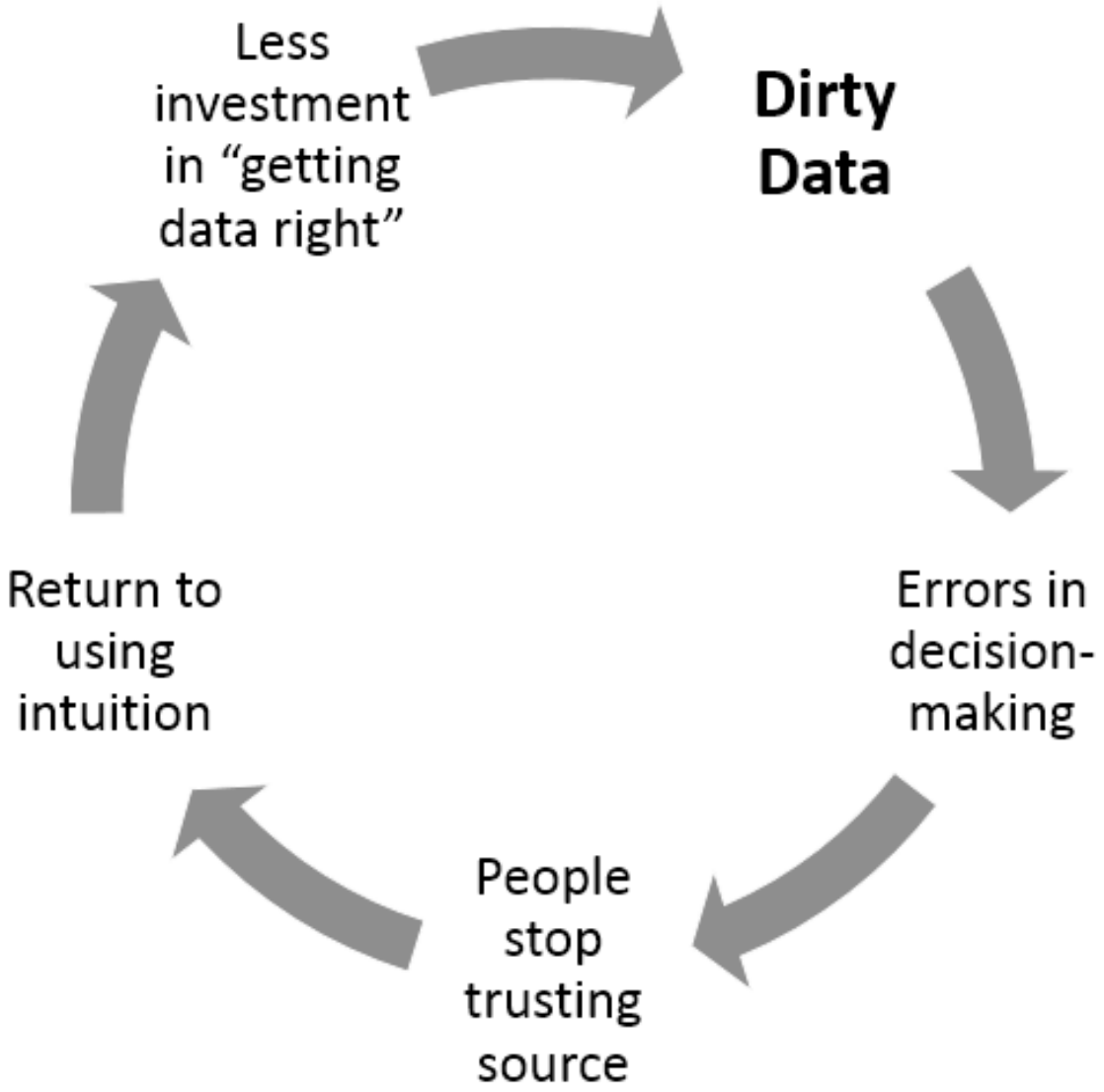
Customer ID	Customer First Name	Customer Last Name	Address	City	State	Zip	Phone
1771	Larry	Shimk	143 S.	Denver	NY	178908	911
1771	Caroline	Shimk	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Shimk	Caroline	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Heather	Schwiter	55 N. W. S. Miss	LaGrange	GA	14321	716-333-4567
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1773	Justin	Justin	34 Kerry Rd.	Littleton	CO	98987	716-67-9087
1774	Pam		66 S. Carlton	North Glen	CO	98765	343-456-6857
1775	D.	Fernandez	3514 S. Main	Denver	CO	80252	303-333-8965
1776	PepsiCo		15365 K St. NW	Washington	DC	20035	202-353-1535
1777	Sam	Esteban	4413 Madison Rd	Ann Arbor	MI	48109	734-140-2531 ext 354
1778	Caroline	Smith	143 N. West St.	Buffalo	NY	14321	716-333-4567

Why Is This Happening?

- "The Agency Problem"
- The data creator is usually not the data consumer.
  - Data creator – sales, customer service
  - Data consumer – marketing dept.
- When the creator doesn't care much about how the data would be used, data is likely to get dirty.



Vicious Cycle from Dirty Data



One Solution

Data's Credibility Problem

Management—not technology—is the solution.
by Thomas C. Redman

The good news is that a little communication goes a very long way. Time and time again, in meetings with data creators and data users, I've heard "We didn't know that anyone used that data set, so we didn't spend much time on it. Now that we know it's important, we'll work hard to get you exactly what you need." Making sure that creators know how data will be used is one of the easiest and most effective ways of improving quality.