MIS 0855 Fall 2016 – Data Science

Week 8 – Dirty Data

Min-Seok Pang

Management Information Systems
Fox School of Business, Temple University
minspang@temple.edu
Oct. 20th, 2016

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?

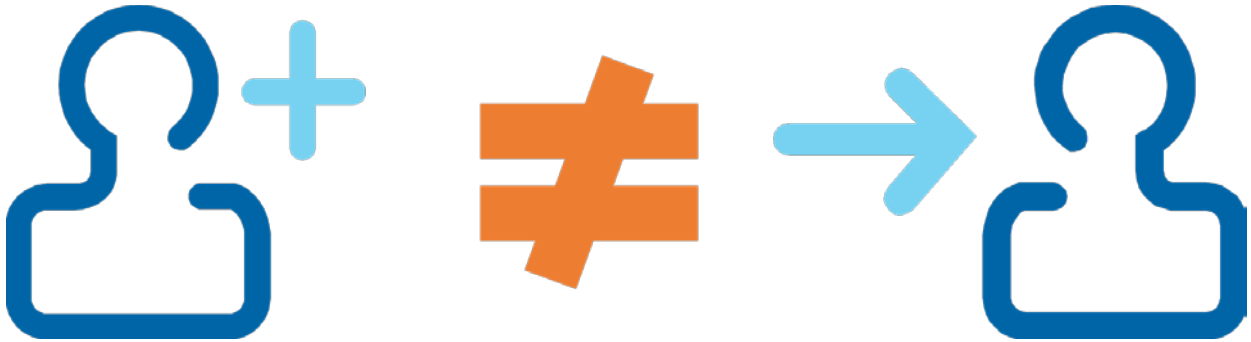
- Think of Ms. Pamela Smith O'Brien
  - How many different names can she have?
- How about an address?
  - 1303 North Taylor Street, Apt. #102, Philadelphia, Pennsylvania 19123, USA
  - How many different addresses can be *valid*?

Origin of Dirty Data

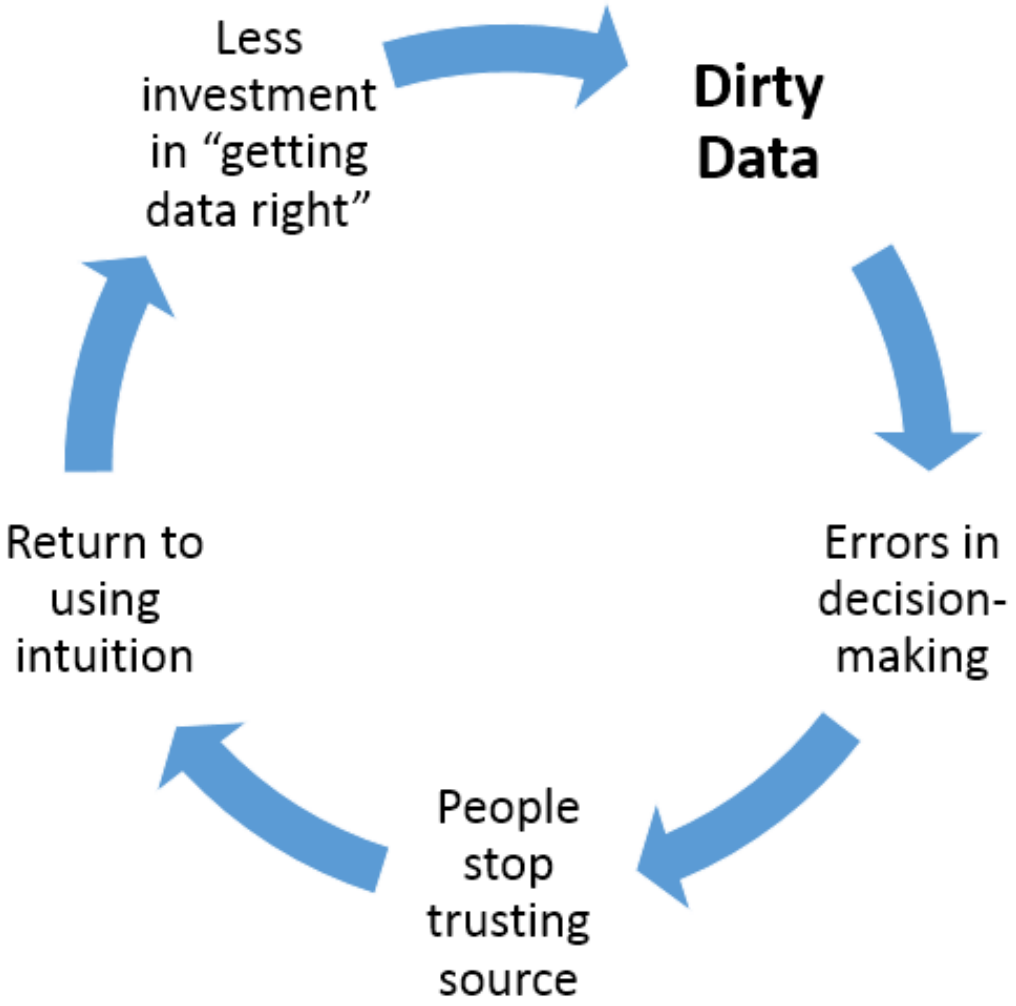
- Measurement can be inaccurate
  - Name – a person's name or a company's name?
- Instrument : the question may be wrong or ambiguous
  - Phone number – home, work, or cell?
- Consistency : the question can be answered inconsistently

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

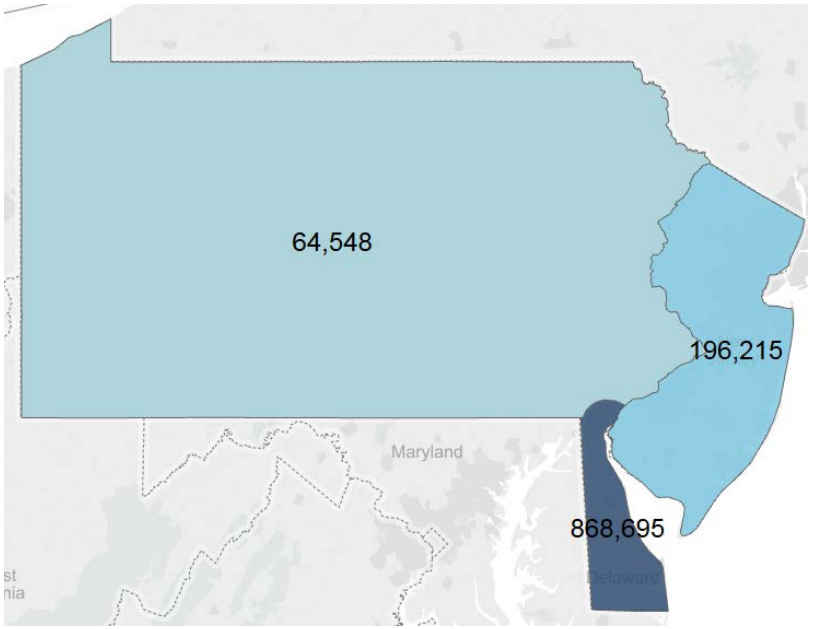
The good news is that a little communication goes a very long way. Time and time again, in meetings with data creators and data users, I've heard "We didn't know that anyone used that data set, so we didn't spend much time on it. Now that we know it's important, we'll work hard to get you exactly what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

Cleaning Data

- What are the problems in this dataset?
- What should you do before analysis?

	A	B	C
1	Year	State	Sales
2	2009	NJ	25690
3	2009	Penna.	17685
4	2009	DE	79034
5	2010	NJ	31240
6	2010	Penna.	27583
7	2010	DE	549460
8	2011	NJ	94690
9	2011	PA	39336
10	2011	DE	71149
11	2012	NJ	10852
12	2012	Pennsylvania	69108
13	2012	DE	89395
14	2013	NJ	33743
15	2013	PA	25212
16	2013	DE	79657

Without Data Cleansing

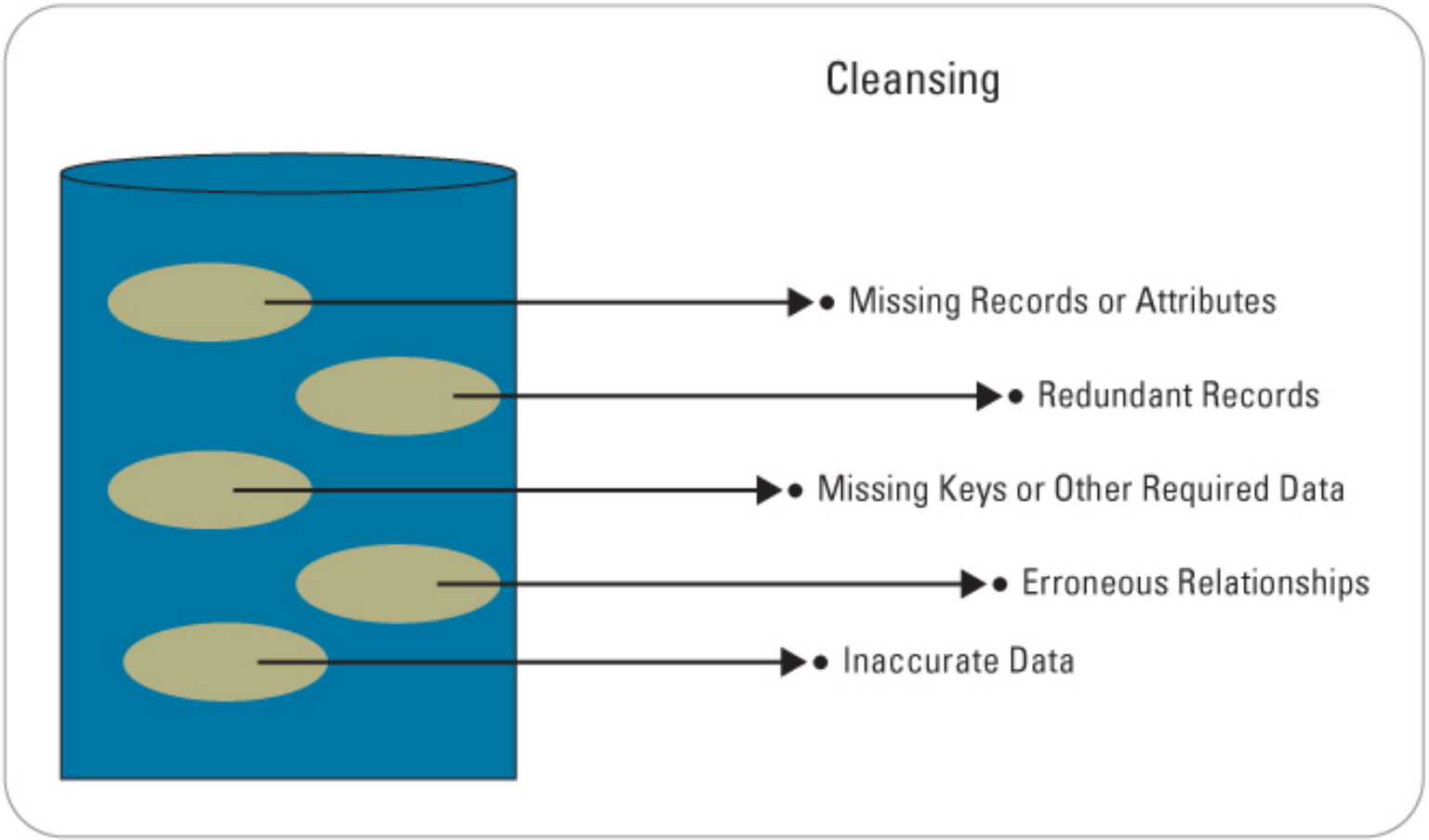


State	
DE	868,695
NJ	196,215
PA	64,548
Penna.	45,268
Pennsylvania	69,108

Row Labels	Sum of Sales
DE	868695
2009	79034
2010	549460
2011	71149
2012	89395
2013	79657
NJ	196215
2009	25690
2010	31240
2011	94690
2012	10852
2013	33743
PA	64548
2011	39336
2013	25212
Penna.	45268
2009	17685
2010	27583
Pennsylvania	69108
2012	69108
Grand Total	1243834

- How would you fix this?
- if you have millions of sales records?

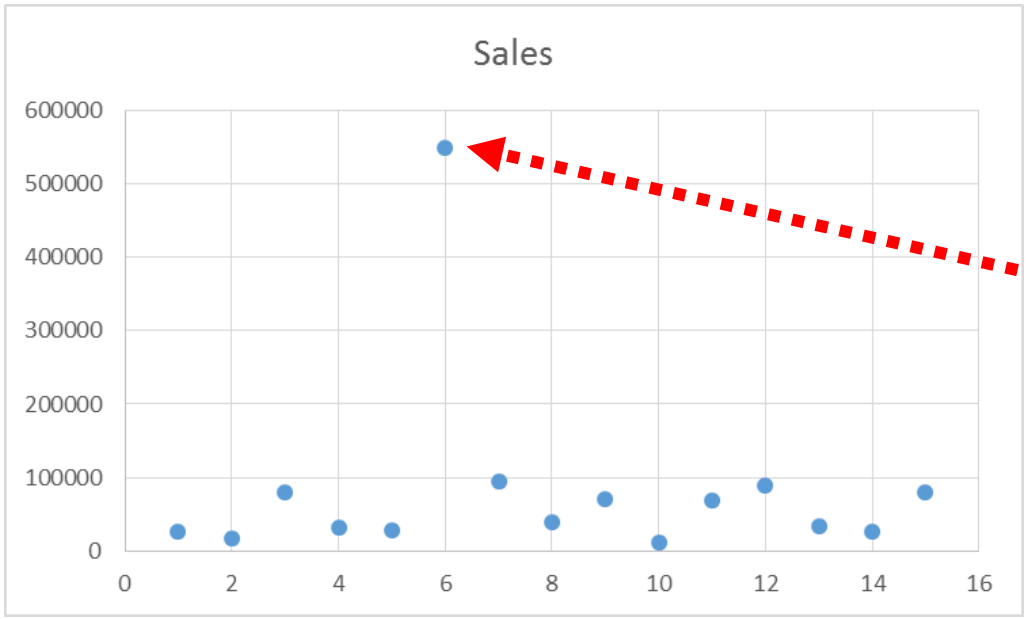
Activities in Data Cleansing



Characteristics of High-Quality Data

Accuracy	Are all the values correct? For example, is the name spelled correctly? Is the dollar amount recorded properly?
Completeness	Are any of the values missing? For example, is the address complete including street, city, state, and zip code?
Consistency	Is aggregate or summary information in agreement with detailed information? For example, do all total fields equal the true total of the individual fields?
Uniqueness	Is each transaction, entity, and event represented only once in the information? For example, are there any duplicate customers?
Timeliness	Is the information current with respect to the business requirements? For example, is information updated weekly, daily, or hourly?

Be Careful in Cleaning Data

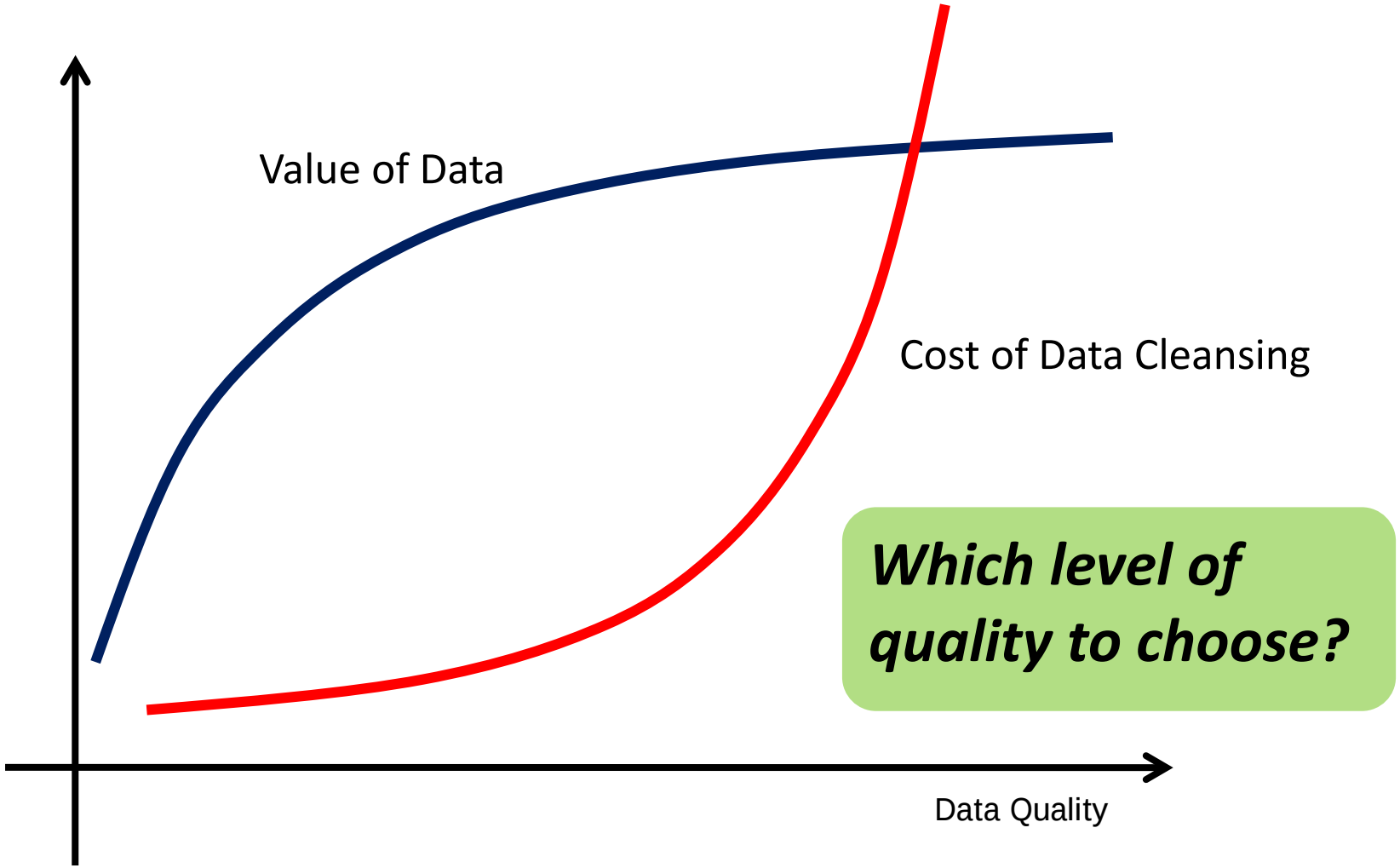


	A	B	C
1	Year	State	Sales
2	2009	NJ	25690
3	2009	Penna.	17685
4	2009	DE	79034
5	2010	NJ	31240
6	2010	Penna.	27583
7	2010	DE	549460
8	2011	NJ	94690
9	2011	PA	39336
10	2011	DE	71149
11	2012	NJ	10852
12	2012	Pennsylvania	69108
13	2012	DE	89395
14	2013	NJ	33743
15	2013	PA	25212
16	2013	DE	79657

- Do we have to remove this?
- If yes, then what?
- Is this really an error or a simply unusual but correct data?

Trade-Off in Data Cleansing (1/2)

Trade-Off in Data Cleansing (2/2)



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