

QA/QC

Quality Assurance / Quality Control

1

The Annual Cost of ‘Bad Data’
to US Businesses

\$14.2 million - \$3.1 trillion

Gartner Research

IBM

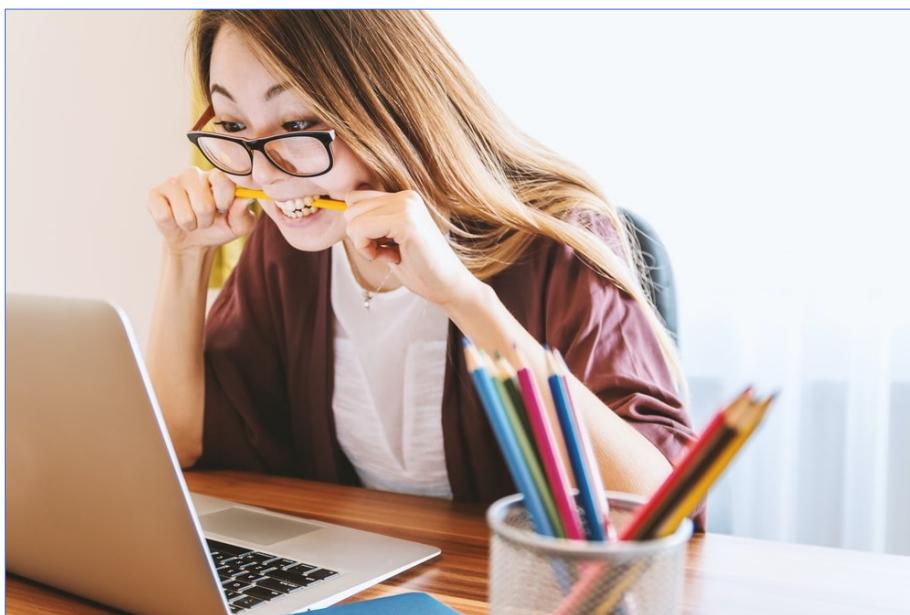
2

(1) Focus efforts in the wrong direction
marketing to the wrong audience, lost sales, inefficient hiring / capital investments



3

(2) Cost of data cleanup
Staff, Software, Opportunity Costs



4

The 1-10-100 Rule

a.k.a. "the cost of quality"

Data entry errors multiply costs exponentially according to the stage at which they are identified and corrected.

5

\$1: Price to check the data at first point of entry



\$10: Price to find and correct error when it is part of a batch



\$100: Cost of fixing the mistake when it reaches customers



6

IBM InfoSphere QualityStage

Investigate, cleanse and manage data to gain more value from your information assets

See why IBM is a leader in the 2020 Gartner Magic Quadrant for Data Quality Solutions →

Offers rich capabilities to create and monitor data quality

IBM InfoSphere® QualityStage™ is designed to support your data quality and information governance initiatives. It enables you to investigate, cleanse and manage your data, helping you maintain consistent views of key entities including customers, vendors, locations and products. The solution helps you deliver quality data for your big data, business intelligence, data warehousing, application migration and master data management projects. Also available for IBM System z®.

IBM InfoSphere Revenue 2021: \$79.1 Billion

SAP Data Services

Maximize the value of all your organization's structured and unstructured data with exceptional functionalities for data integration, quality, and cleansing.

Request a demo Request a quote

SAP Data Services Revenue 2021: \$27.5 Billion

LexisNexis Risk Solutions

Maintain an accurate, consistent customer profile.

Open an overlooked address update. Don't let compromised data impact your performance.

LN Risk Solutions sales 2021: \$466.64 Million

7

Prevention
is less expensive than
Correction
which is much less expensive than
Failure

8

Goldberg et al. (2008): error rates* in two clinical research databases

Demographic data:
2.3-5.2%

HEALTH UNIT AND PHYSICIAN INFORMATION:					
Health Unit ID: _____	Date: ____/____/____	Physician name: _____	Physician ID: _____		
PATIENT DEMOGRAPHY INFORMATION:					
<i>Patient information (Woman):</i>					
Name: _____	Date of birth: ____/____/____	Home address: _____	City: _____		
Occupation: _____	Education: _____	Landline phone: _____	Cell phone: _____		
Marital status: _____	Language: _____	Race: _____	Ethnicity: _____	Religion: _____	
<i>Basic health information and Vital signs of patient:</i>					
Blood type: _____	Rh factor: _____	Pulse: _____	Body temp: _____	Blood pressure: _____	Height: _____ Weight: _____
<i>Emergency contact:</i>					
Name: _____	Phone number: _____	Relationship: _____			
<i>Husband information:</i>					
Name: _____	Phone number: _____	Date of birth: ____/____/____	Occupation: _____	Education: _____	

*Includes errors made during data entry and misinterpretation of data contained on the original forms.

Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. *AMIA Annual Symposium proceedings. AMIA Symposium, 2008*, 242-246.

9

Treatment data:
10-26.9%

Mr. XXXXXX, a 50-year-old male, was involved in a head-on motor vehicle collision on December 23, YYYY. He was attended by medical personnel from City of XYZ Municipal Ambulance Service for the first aid of his collision injuries. Upon their arrival, he was ambulatory as he had extricated himself. He had sinus tachycardia and complained of pain in his left flank and low back. There were contusions on his thorax and minor abrasions in his upper and lower extremities observed. An intravenous access was established through which Normal Saline was infused. His neck and back was secured with complete spinal precautions. He was transported to XXXX West XYZ Memorial Hospital for the further treatment of his injuries (**PDF REF: 78-79**).

Includes errors made during data entry and misinterpretation of data contained on the original forms.

Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. *AMIA Annual Symposium proceedings. AMIA Symposium, 2008*, 242-246.

10

 **Darryl Pieroni**
@DarrylPieroni

Follow ▾

Most **#datascientists** spend only 20 percent of their time on actual **#data** analysis. The remaining 80 percent of data scientists' time is spent locating, unifying and cleaning data before they can begin their real work.

Source: IDG

3:22 PM - 23 Jan 2019

 **Dr. Danielle Rosvaly**
@DRosvaly

Follow ▾

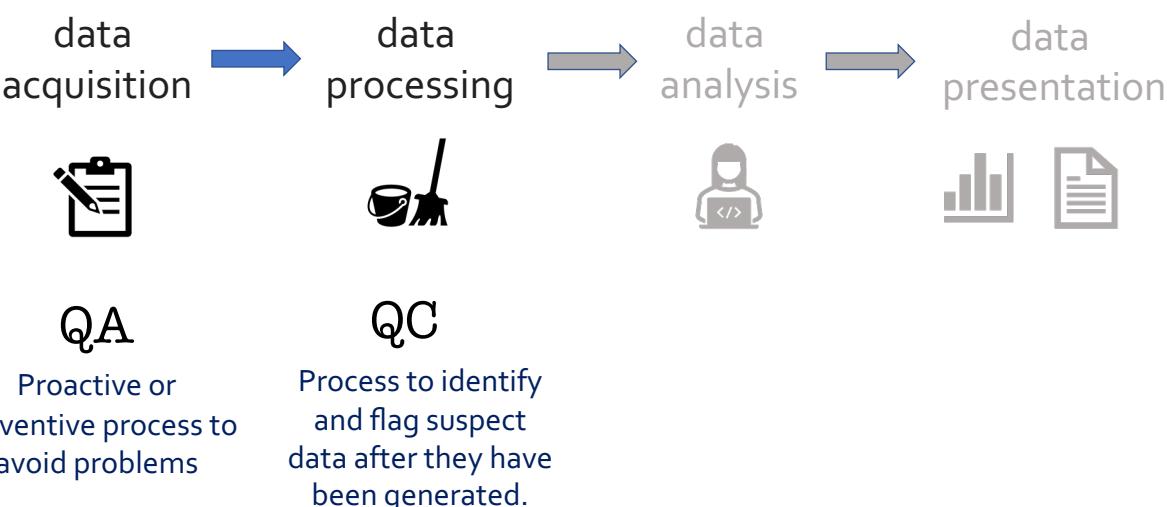
eighty percent of our time as Digital humanists is spent data cleaning. That sounds about right. **#folgermiranda**

10:22 AM - 21 Sep 2018

In a recent student consultancy project setup to create a master pricing database, approximately 35% of the time was spent in simply cleaning up 'dirty data' ([Cokayne-Naylor, 2009](#)).

Cokayne-Naylor, 2009 Cokayne-Naylor, K. (2009). *Information systems development: A Beginner's experience*. Unpublished MSc thesis, University of Warwick, UK.
[Google Scholar](#)

11



```

graph LR
    A[data acquisition] --> B[data processing]
    B --> C[data analysis]
    C --> D[data presentation]
  
```

data acquisition → **data processing** → **data analysis** → **data presentation**

QA
Proactive or preventive process to avoid problems

QC
Process to identify and flag suspect data after they have been generated.

Put another way....

12



13

QA/QC

Quality Assurance / Quality Control

14



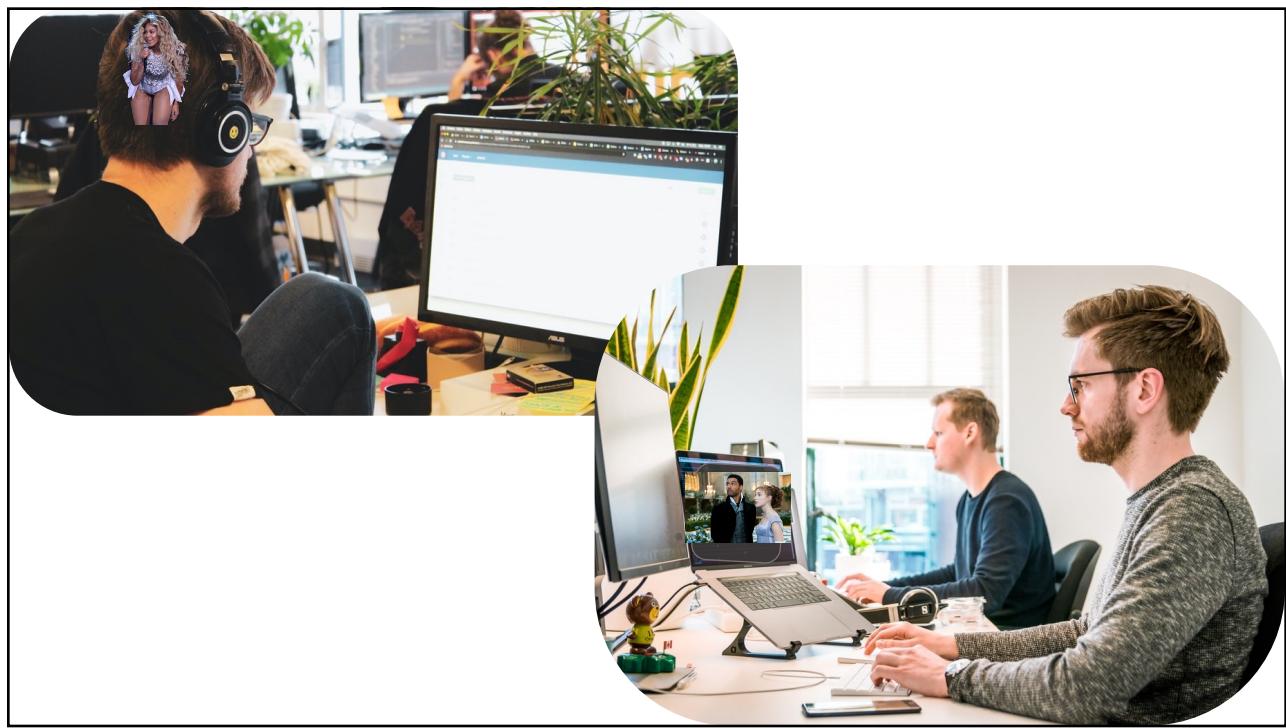
15



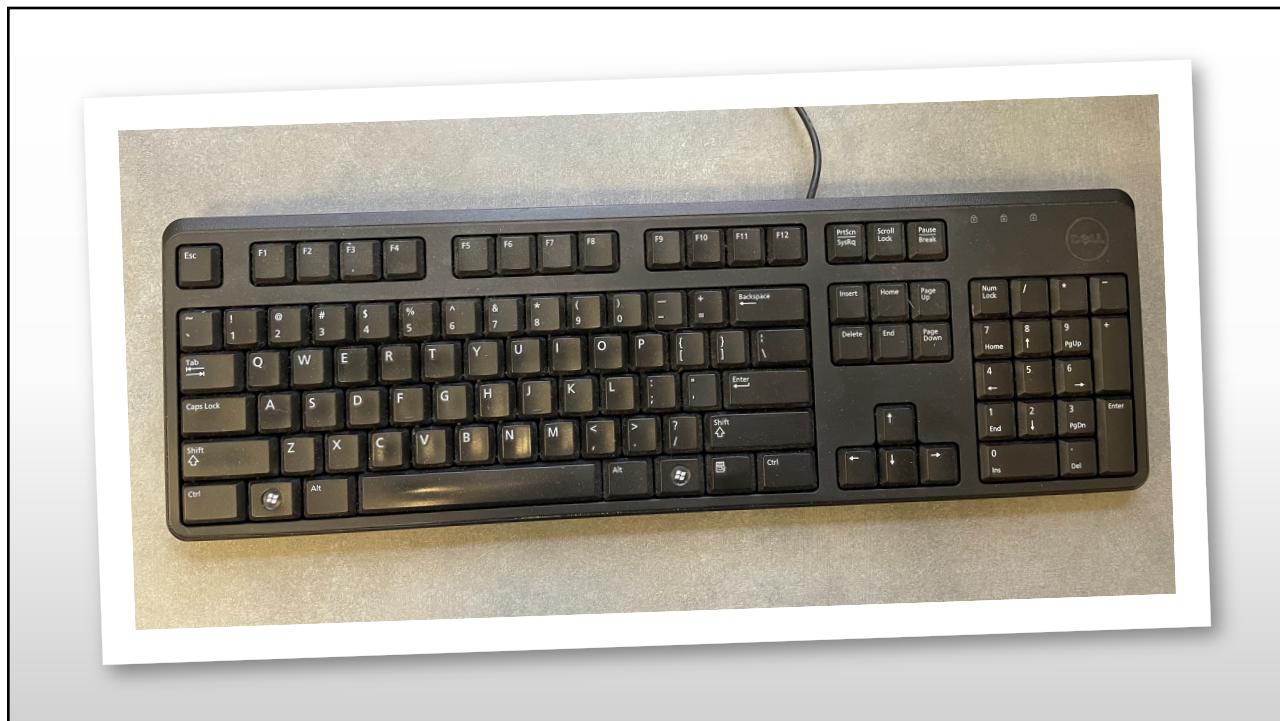
16



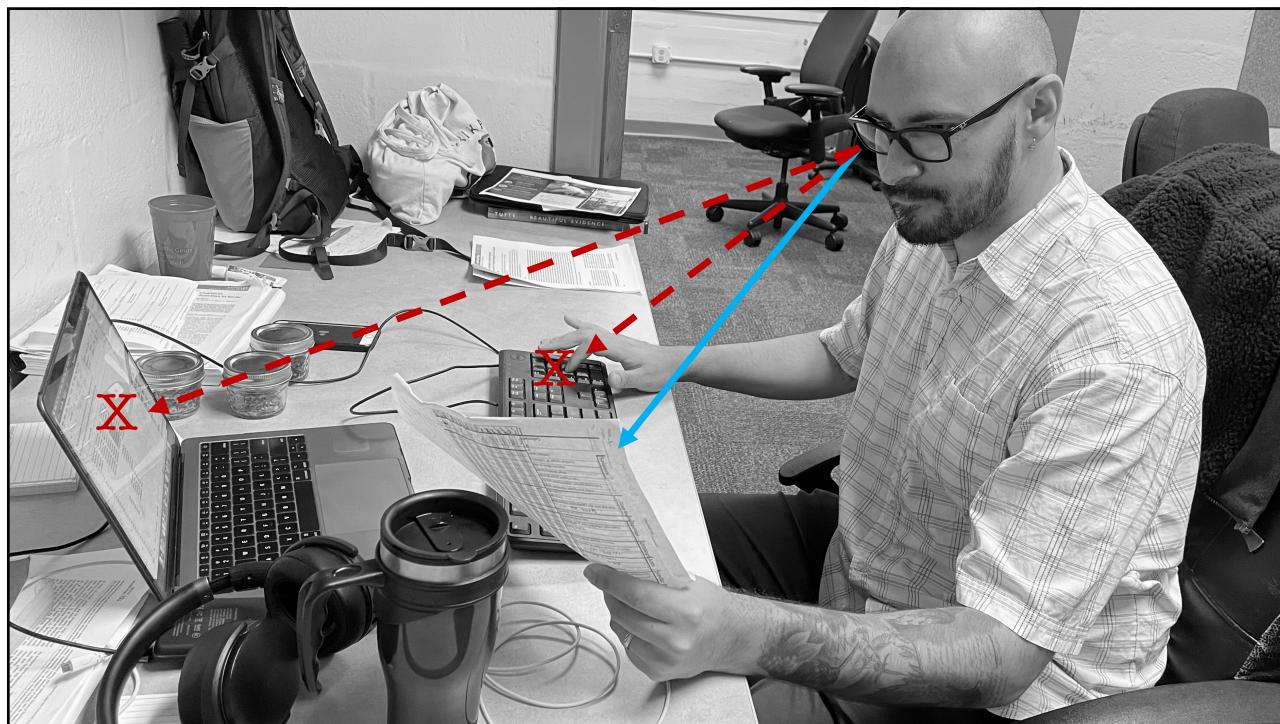
17



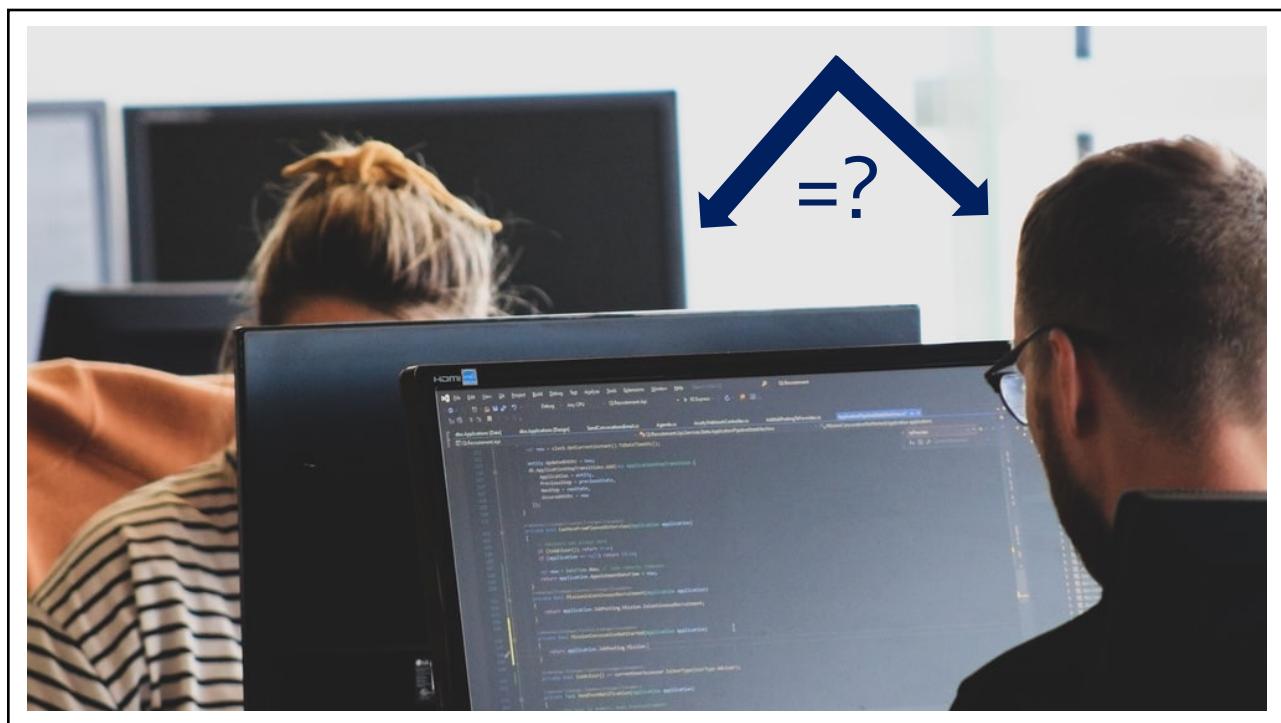
18



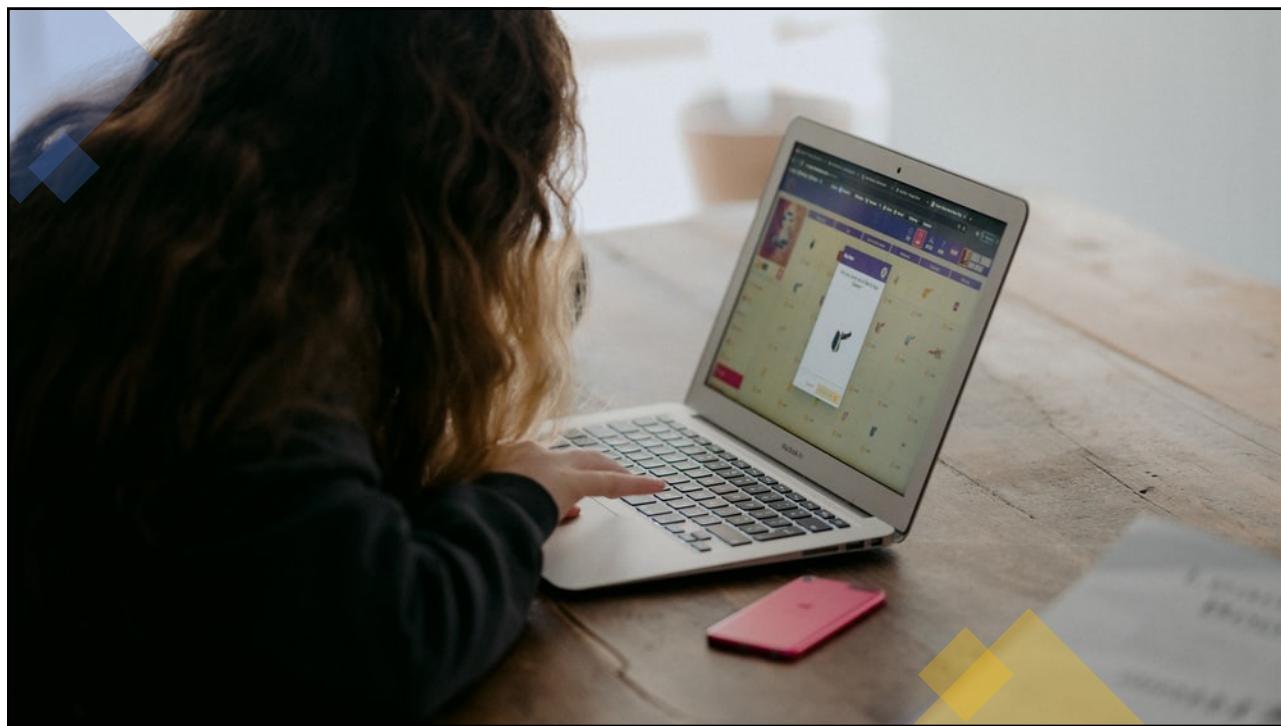
19



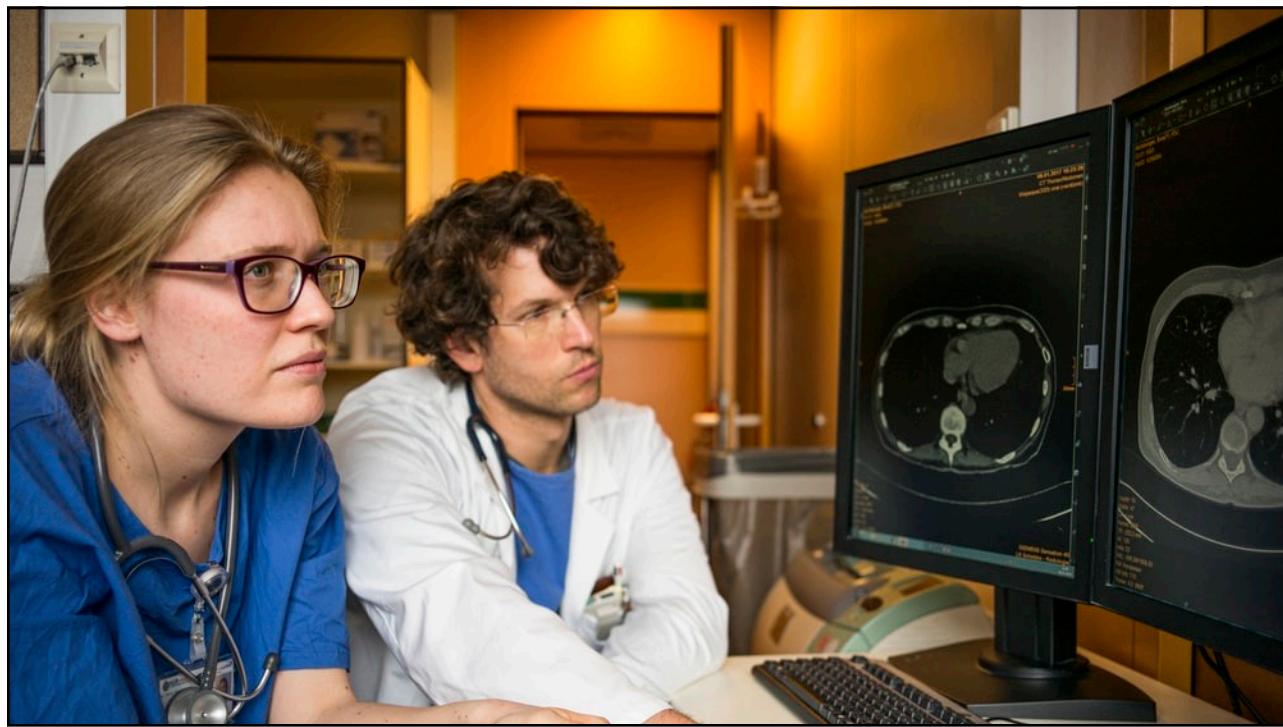
20



21



22



23

Computers in Human Behavior
Volume 27, Issue 5, September 2011, Pages 1834-1839

Preventing human error: The impact of data entry methods on data accuracy and statistical results

Kimberly A. Barchard ^{a, b}, Larry A. Pace ^{b, 1}

Methods
195 undergraduates
Each entered 30 data sheets, each w/ 6 types of data.
3 data entry methods

Family Background					School Experiences				
1.	(SD)	D	N	A	SA	1.	D	N	A
2.	(SD)	D	N	A	SA	2.	D	(N)	A
3.	SD	D	N	A	(SA)	3.	(D)	N	A
4.	SD	D	N	(A)	SA	4.	D	N	A
5.	SD	D	(N)	A	SA	5.	D	(N)	A
6.	SD	(D)	N	A	SA	6.	D	N	A
7.	SD	D	(N)	A	SA	7.	(D)	N	A
8.	SD	D	(N)	A	SA	8.	D	N	(A)
9.	(SD)	D	N	A	SA	9.	D	(N)	A
10.	(SD)	D	N	A	SA	10.	(D)	N	A

Extraversion					Social Skills Test				
1.	1	2	3	4	(5)	1.	1	2	(3)
2.	1	2	3	4	(5)	2.	(1)	2	3
3.	1	2	(3)	4	5	3.	1	(2)	3
4.	1	(2)	3	4	5	4.	1	(2)	3
5.	1	(2)	3	4	5	5.	(1)	2	3
6.	1	2	3	4	(5)	6.	1	(2)	3
7.	1	2	(3)	4	5	7.	(1)	2	3
8.	1	2	(3)	4	5	8.	(1)	2	3
9.	1	(2)	3	4	5	9.	1	(2)	3
10.	1	(2)	3	4	5	10.	(1)	2	3

double entry visual checking single entry

24

ERRORS

single entry



visual checking



visual checking



2958% more
than

double entry



25

PERFECT ACCURACY

double entry



77.4%
of participants

visual checking



17.1%
of participants

single entry

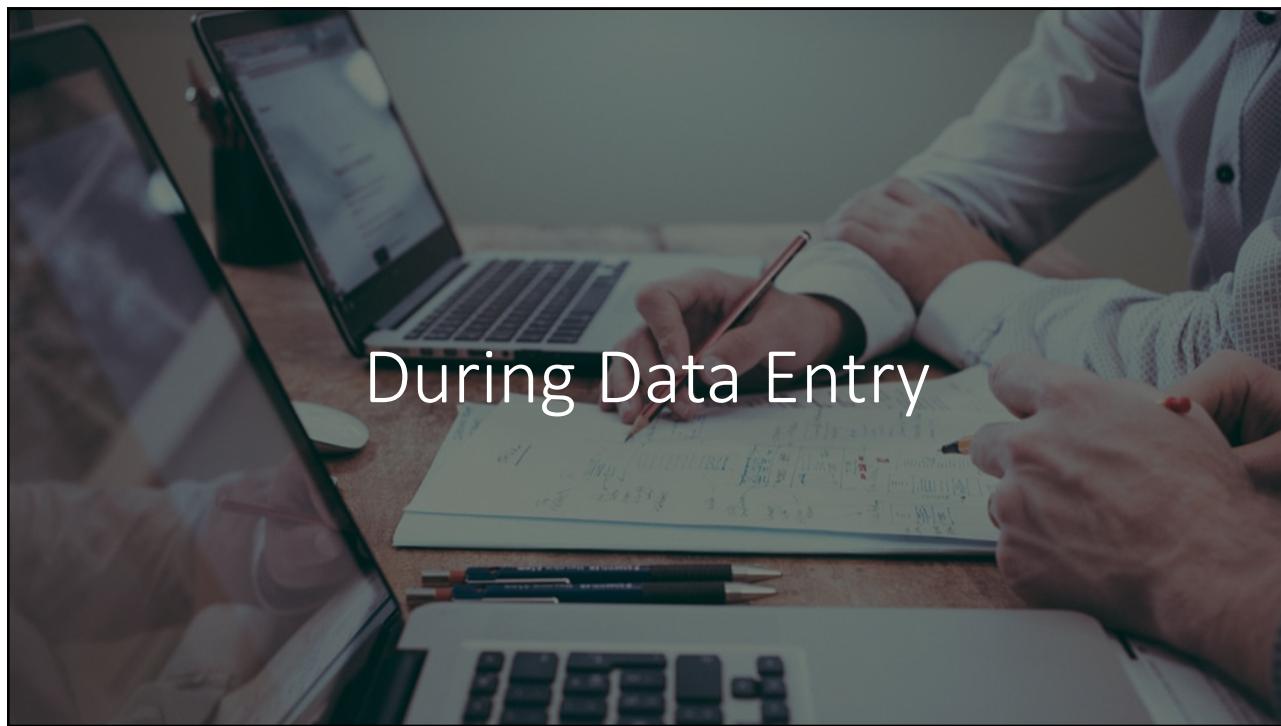


5.5%
of participants

26



27

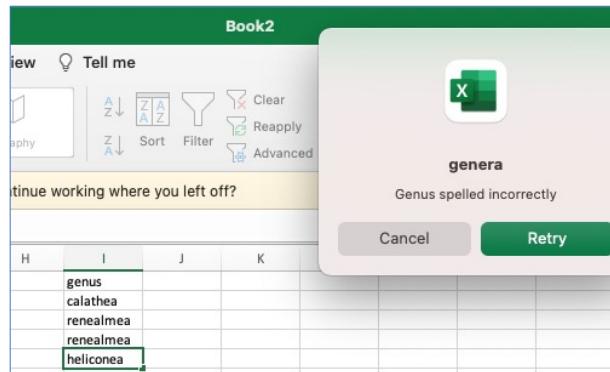


During Data Entry

28

Tool #1: Restricted Data Entry aka Data Validation

limit the values that can be entered to those that are actually possible



29

Tool #2: Atomized Data Entry

one type of data per cell

Instead of this...

	A	B
1	species	count
2	heliconia acuminata (heliconiaceae)	3
3	heliconia latispatha (heliconiaceae)	12
4	heliconia tarumaensis (heliconiaceae)	5
5	heliconia sylvestris (heliconiaceae)	7
6	heliconia bihai (heliconiaceae)	22
7	renealmia floribunda (zingiberaceae)	16

...do this.

	A	B	C	D
1	genus	species	family	count
2	heliconia	acuminata	heliconiaceae	3
3	heliconia	latispatha	heliconiaceae	12
4	heliconia	tarumaensis	heliconiaceae	5
5	heliconia	sylvestris	heliconiaceae	7
6	heliconia	bihai	heliconiaceae	22
7	renealmia	floribunda	zingiberaceae	16

30

Tool #3: Codes

fewer keystrokes = fewer errors, eliminates duplication, faster

Instead of this...

	A	B	C	D
1	genus	species	family	count
2	heliconia	acuminata	heliconiaceae	3
3	heliconia	latispatha	heliconiaceae	12
4	heliconia	tarumaensis	heliconiaceae	5
5	heliconia	sylvestris	heliconiaceae	7
6	heliconia	bihai	heliconiaceae	22
7	renealmia	floribunda	zingiberaceae	16

...do this.

	A	B	C	D
1	code	genus	species	family
2	ha	heliconia	acuminata	heliconiaceae
3	hl	heliconia	latispatha	heliconiaceae
4	ht	heliconia	tarumaensis	heliconiaceae
5	hs	heliconia	sylvestris	heliconiaceae
6	hb	heliconia	bihai	heliconiaceae
7	rf	renealmia	floribunda	zingiberaceae

Set up in advance
Table 1:
species_codes

	A	B
1	code	count
2	ha	3
3	hl	12
4	ht	5
5	hs	7
6	hb	22
7	rf	16

Enter data like this
Table 2:
survey_data

31

89% fewer chances to make a mistake.

196 keystrokes

(not counting return/tab)

	A	B	C	D
1	genus	species	family	count
2	heliconia	acuminata	heliconiaceae	3
3	heliconia	latispatha	heliconiaceae	12
4	heliconia	tarumaensis	heliconiaceae	5
5	heliconia	sylvestris	heliconiaceae	7
6	heliconia	bihai	heliconiaceae	22
7	renealmia	floribunda	zingiberaceae	16

21 keystrokes

(not counting return/tab)

	A	B
1	code	count
2	ha	3
3	hl	12
4	ht	5
5	hs	7
6	hb	22
7	rf	16

32

Tool #4: Validate data entry with “Text-to-Speech”.

“Three,
Four,
Nine,
Apple...”



33

Tool #5: Enter data with ‘Speech-to-Text’, then validate.

Record a reading of the data and then transcribe from the recording

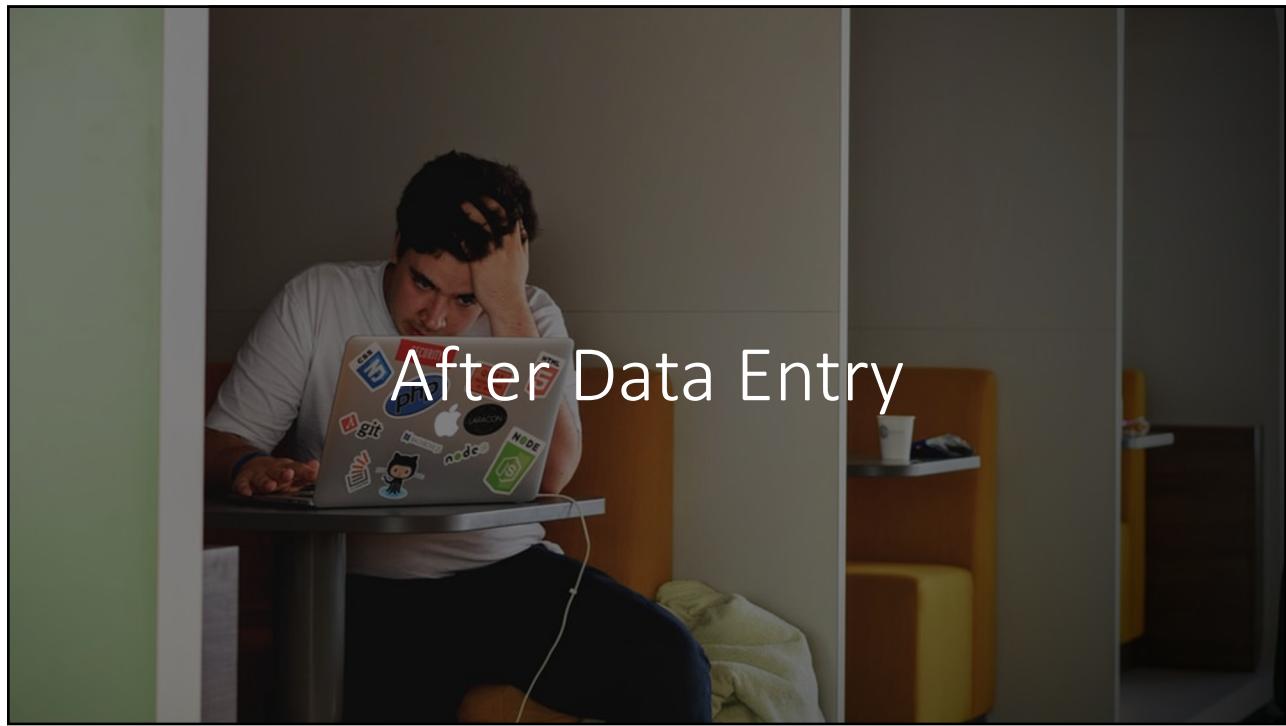


34

QA/QC

Quality Assurance / Quality Control

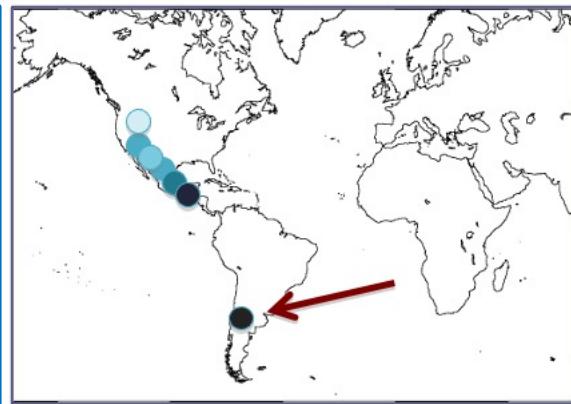
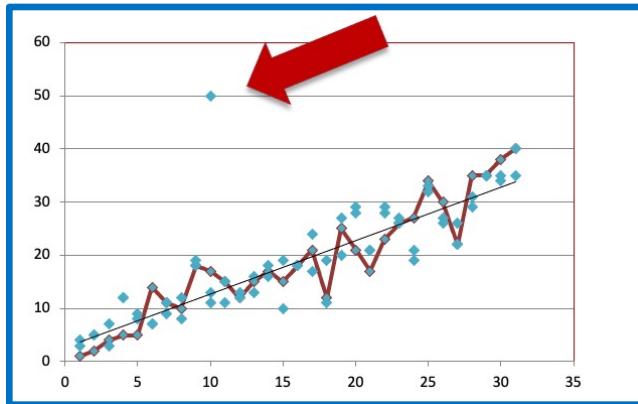
35



36

(1) Visualize the Data

Do any seem values seem unusual?



- (i) Normal probability plots
- (ii) Regression
- (iii) Scatter plots
- (iv) maps
- (v) Subtract values from mean
- (vi) change from last year's measurement
- (vii) statistical tests for outliers

37

(2) Summarize the Data

Do the values look reasonable?



... Oh, do let us go in a caravan!" *Caravan and other stories* by Mrs. Russell
"I know it sounds lovely, darling! I've got a car, and we could have a holiday if they're still here. I would get a nice big pension to buy one if we had one, Daddy couldn't afford this now. No, we'll have to do without a holiday this year; but I'll tell you all we'll all go to Southend for the day, as we did last year, and have a nice tea with Mummy and a special present for her. There are lots of nice gardens there," "Oh, oh! I do wish I could go and ride," he added unexpectedly. "You don't know how I like my," he continued sighing deeply as he remembered the blissful days of his friend and the share his little pony had given him. "Southend is nothing but gardens and people," cried Phyllis; "it's in this place; and oh! Mummy, I do so long for fields and flowers," she added pensively; and she shook her long brown hair like the sand in her open palm. "Never mind, darling, you shall have them one day," answered her mother with easy vagueness.
This really was a very disappointing, and it was the most fortunate of all the evenings a car stopped at the door.
"Uncle Edward!" shouted Bob, rushing from the room. Phyllis tears so hastily from her eyes that she arrived at the front door as he did, and both flung themselves on the tall, kindly-looking man.
"Uncle Edward! Uncle Edward!" they cried. "You've come! We've been longing to see you. Oh, how glad we are you're here!" Now the delightful thing was that their uncle was just as pleased to see them as they were, and returned their happy and most cordiality. They were just on the point of dragging him in, hand in hand, when he said: "Stop, no so fast. There's things to do in front of the car, when he said: "Stop, no so fast. So said he, he began diving into the back of it and bringing out, packages, but various parcels, which he handed out one by one,
There's this case of chickens I've brought for your mother to cook for

38

(3) Annotate the Data

Mark data with quality control flags



- Verified
- Needs review
- Needs correction
- Data interpolated

39

96%  99%

40