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

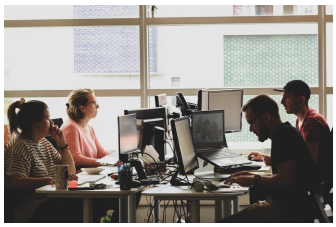
a.k.a. "the cost of quality"

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

5

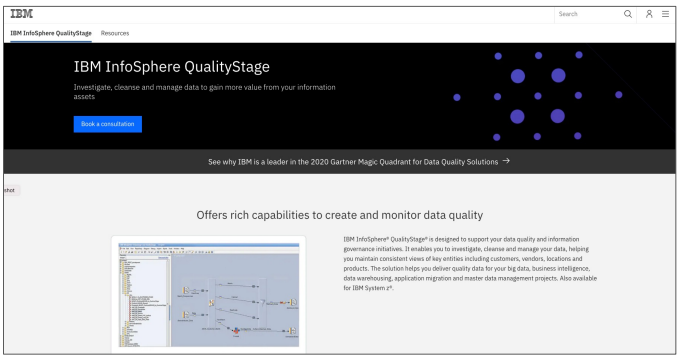
$1: Price to check the data at first point of entry

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

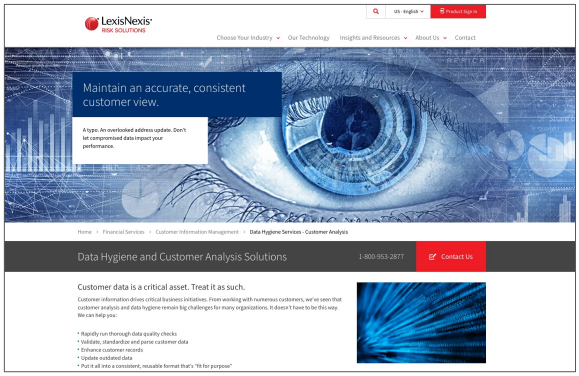


$100: Cost of fixing the mistake when it reaches customers

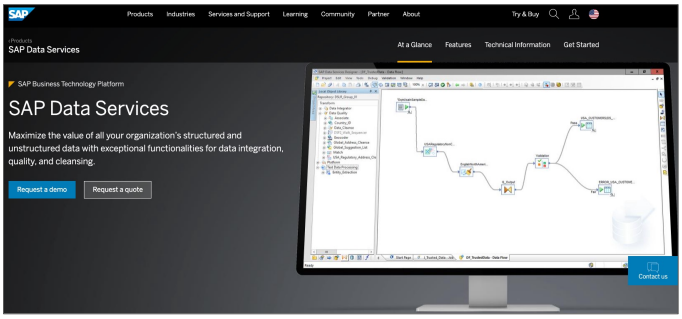
6



IBM InfoSphere Revenue 2021: $79.1 Billion



LN Risk Solutions sales 2021: $466.64 Million



SAP Data Services Revenue 2021: $27.5 Billion

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:

Patient information (Woman):

Name: _____ Date of birth: __/__/_ Home address: _____ City: _____

Occupation: _____ Education: _____ Landline phone: _____ Cell phone: _____

Marital status: _____ Language: _____ Race: _____ Ethnicity: _____ Religion: _____

Basic health information and Vital signs of patient:

Blood type: __ Rh factor: ___ Pulse: ____ Body temp: __ Blood pressure: ___ Height: __ Weight: __

Emergency contact:

Name: _____ Phone number: _____ Relationship: _____

Husband information:

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

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

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

data acquisition

QA

Before Collection

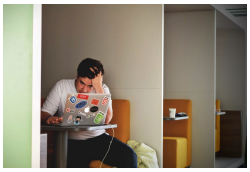
Data Collection

Data Entry

data processing

QC

After Data Entry



13

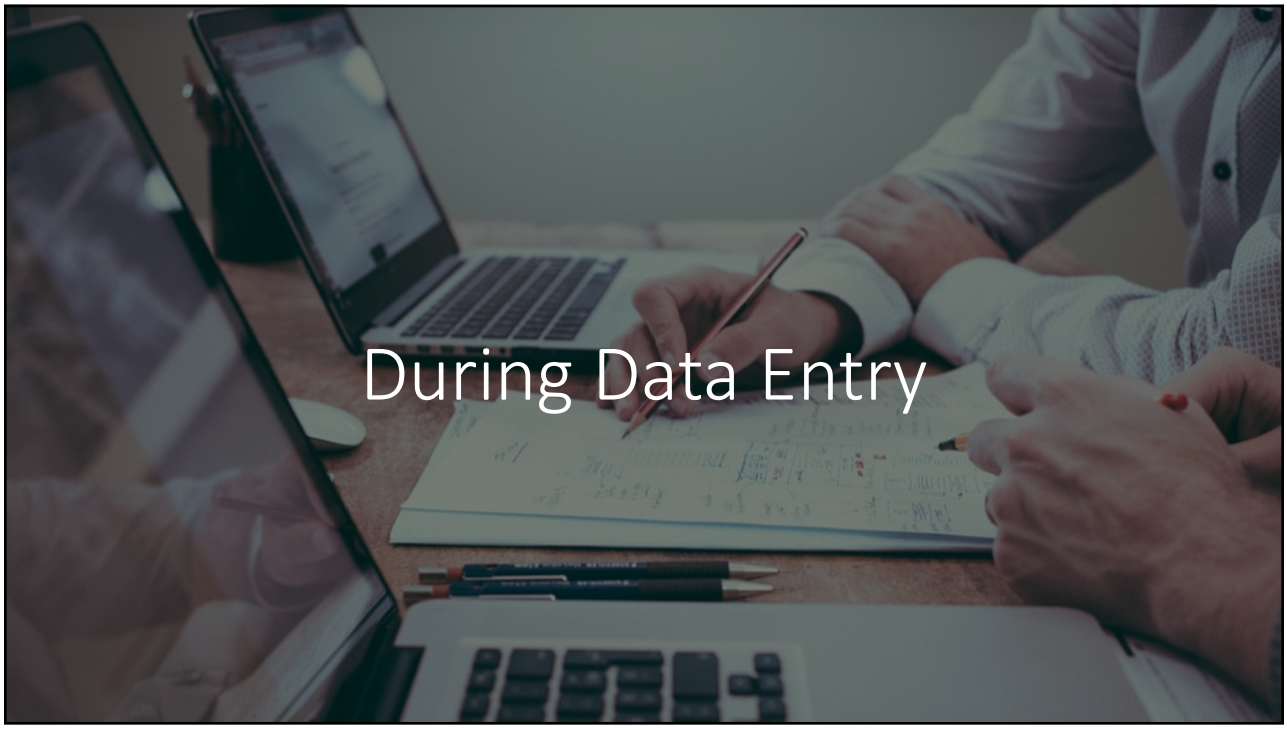
QA/QC

Quality Assurance / Quality Control

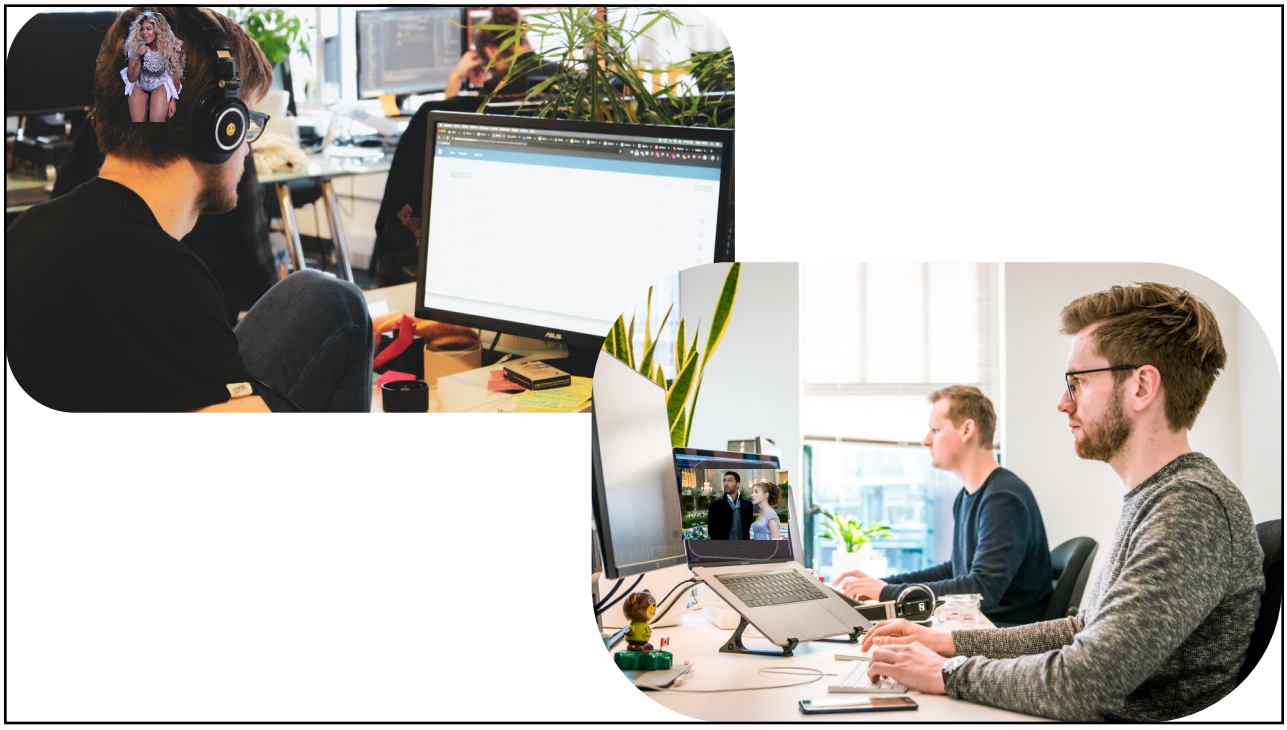
14

15

16

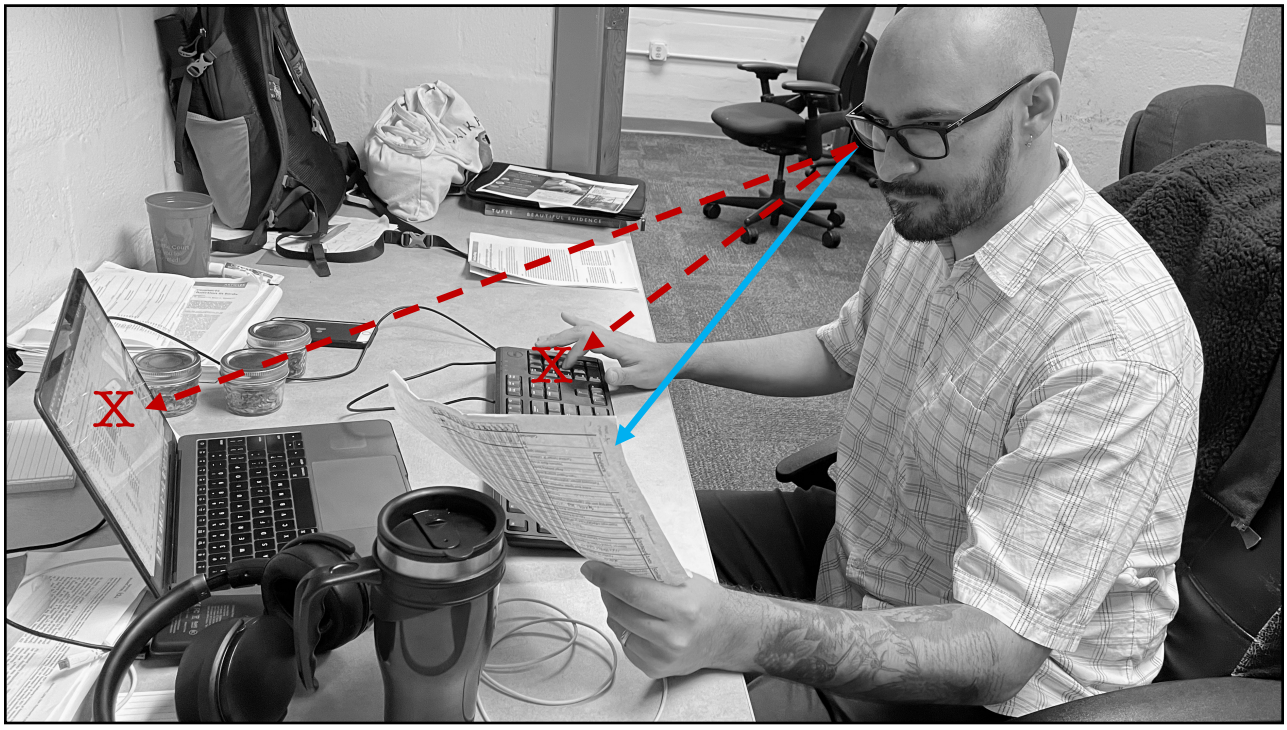


17



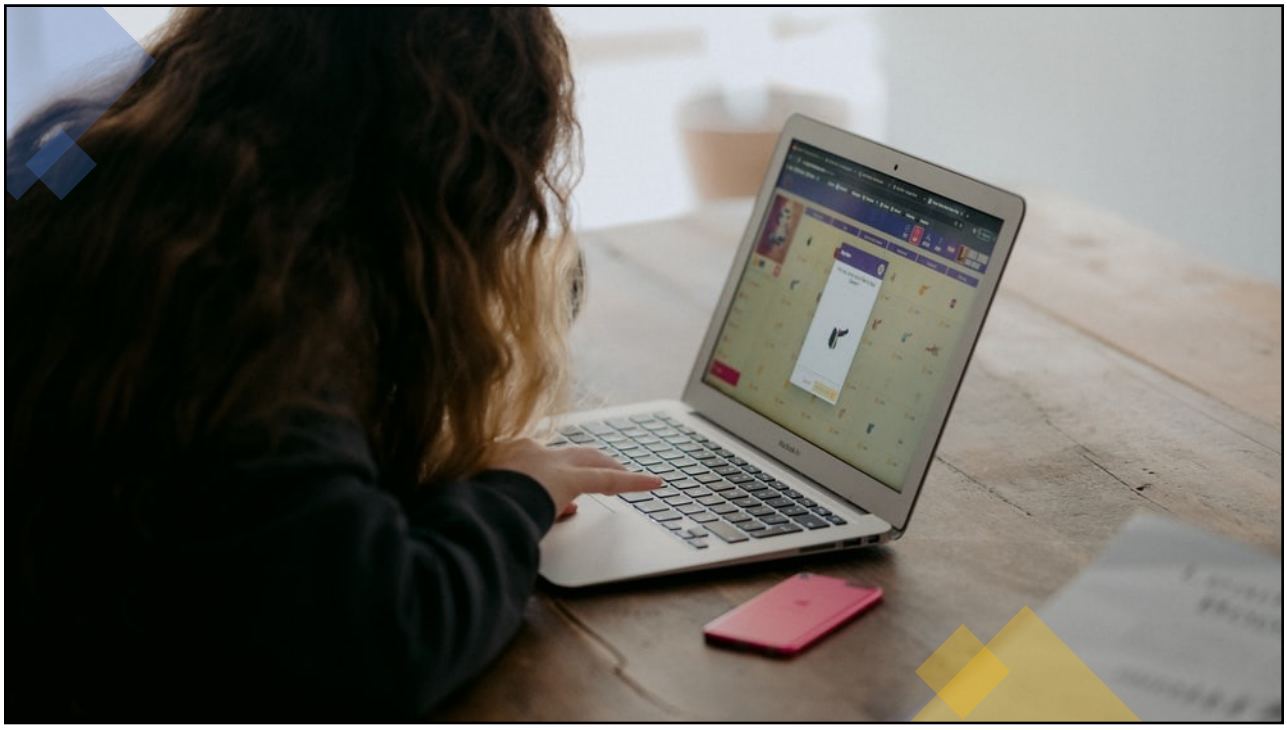
18

19

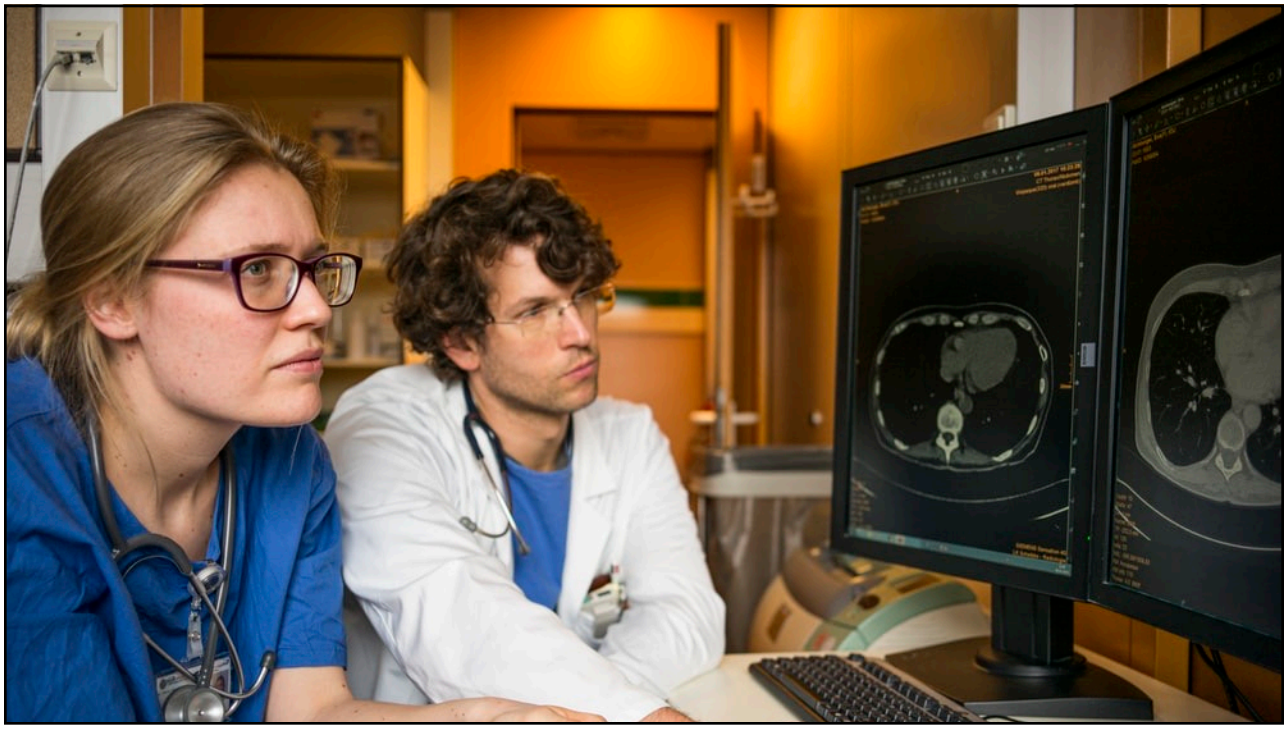


20

21



22



23

ELSEVIER
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], Larry A. Pace [b, 1]

Methods

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

double entry | visual checking | single entry

Family Background

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

School Experiences

1.	D	N	(A)
2.	D	(N)	A
3.	(D)	N	A
4.	D	N	(A)
5.	D	(N)	A
6.	D	N	(A)
7.	(D)	N	A
8.	D	N	(A)
9.	D	(N)	A
10.	(D)	N	A

Extraversion

1.	1	2	3	4	(5)
2.	1	2	3	4	(5)
3.	1	2	(3)	4	5
4.	1	(2)	3	4	5
5.	1	(2)	3	4	5
6.	1	2	3	4	(5)
7.	1	2	(3)	4	5
8.	1	2	(3)	4	5
9.	1	(2)	3	4	5
10.	1	(2)	3	4	5

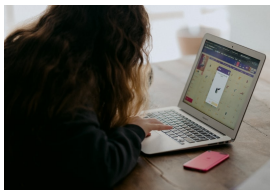
Social Skills Test

1.	1	2	(3)
2.	(1)	2	3
3.	1	(2)	3
4.	1	(2)	3
5.	(1)	2	3
6.	1	(2)	3
7.	(1)	2	3
8.	(1)	2	3
9.	1	(2)	3
10.	(1)	2	3

24

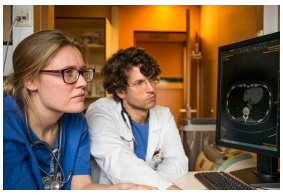
ERRORS

single entry

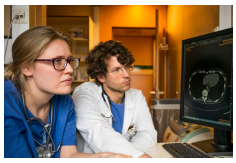


≈

visual checking

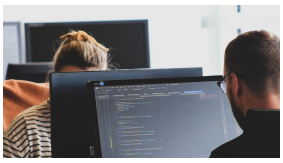


visual checking



2958% more than

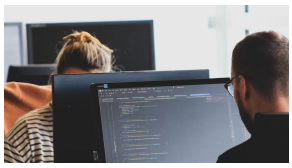
double entry



25

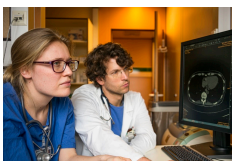
PERFECT ACCURACY

double entry



77.4% of participants

visual checking



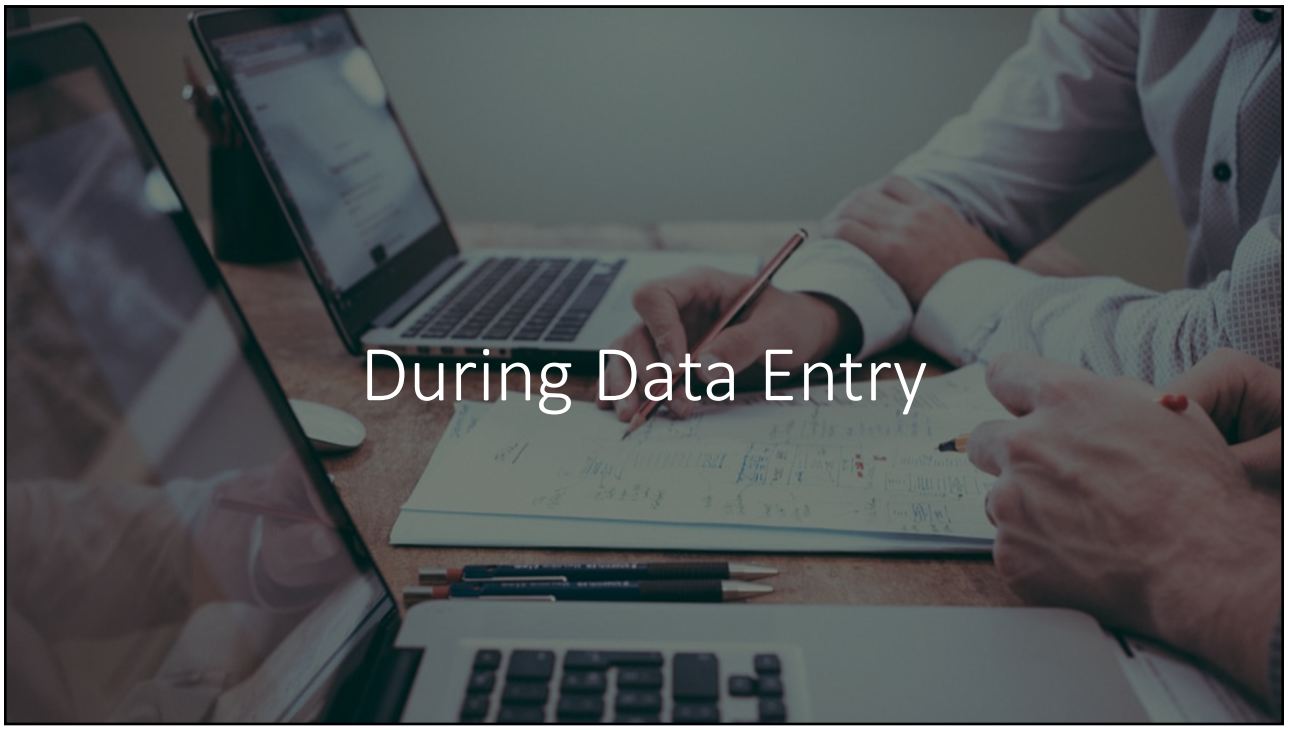
17.1% of participants

single entry

5.5% of participants

26

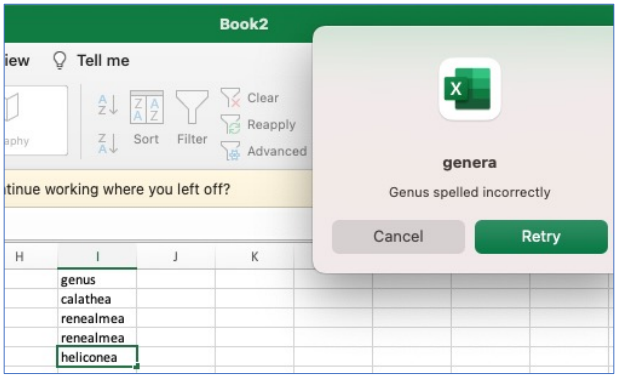
27



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.

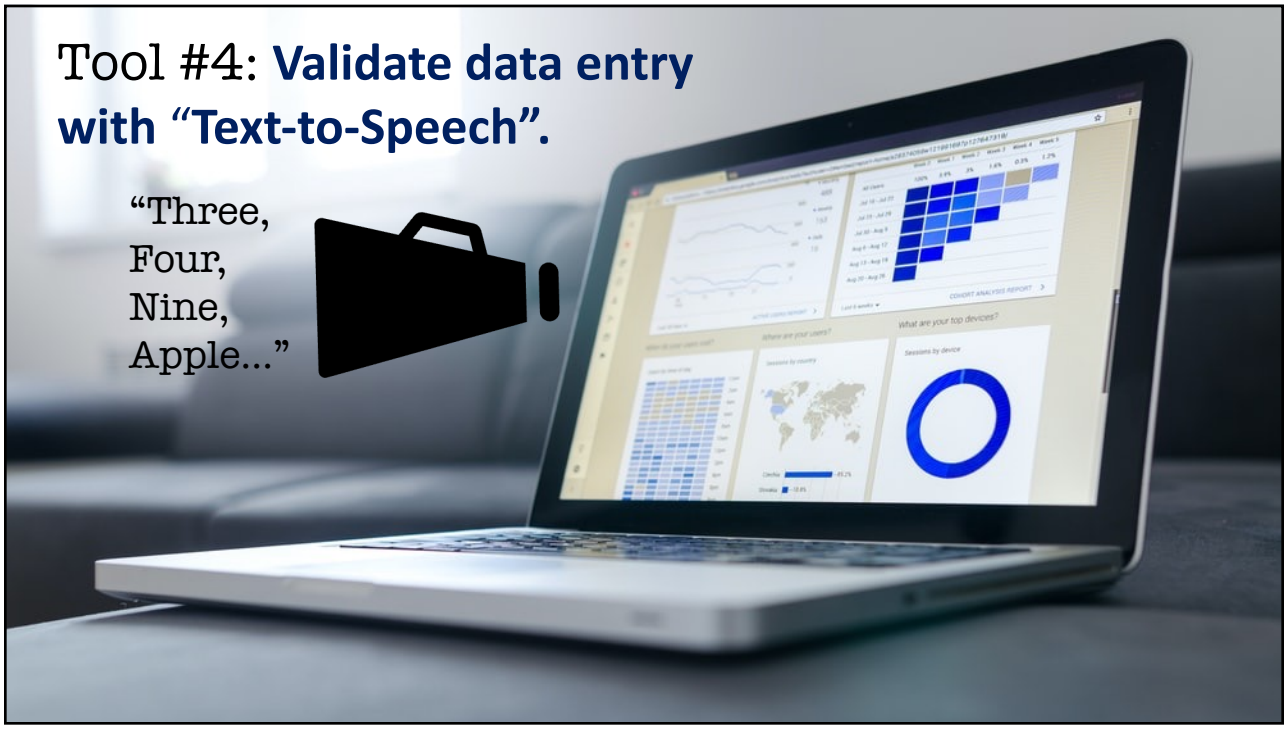
196 keystrokes
(not counting return/tab)

	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



33

Tool #5: Enter data with 'Speech-to-Text`, then validate.

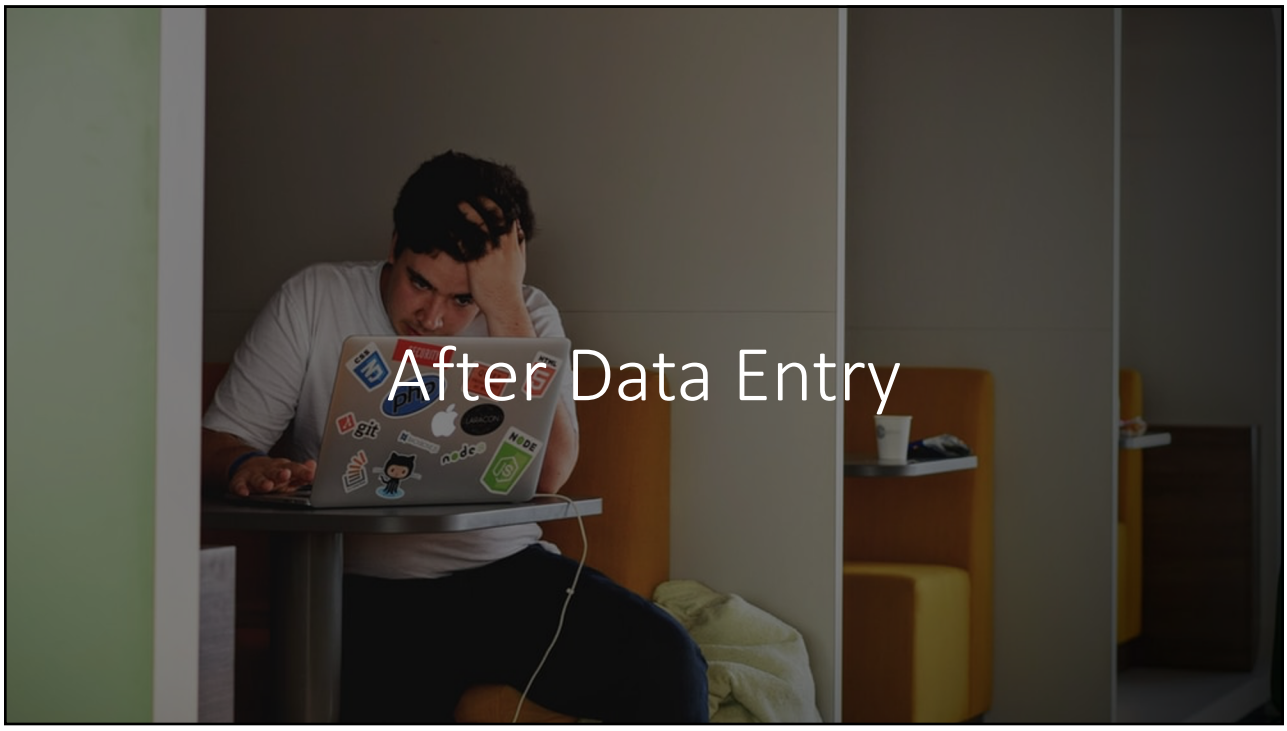
Record a reading of the data and then transcribe from the recording

34

QA/QC

Quality Assurance / Quality Control

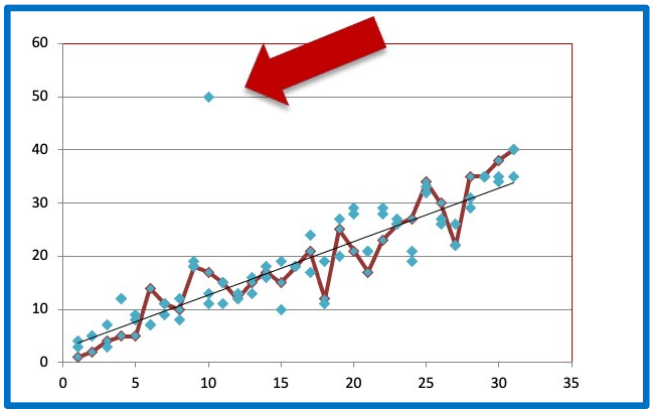
35

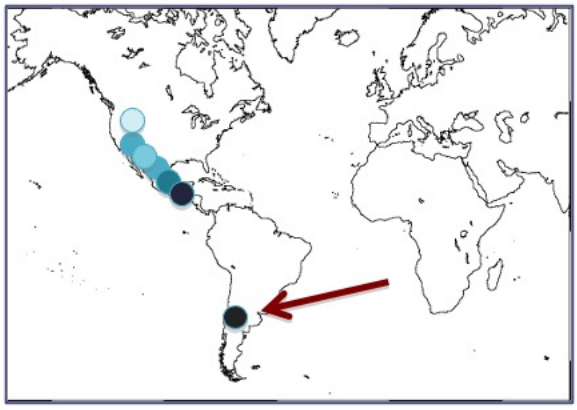


36

(1) Visualize the Data

Do any seem values seem unusual?



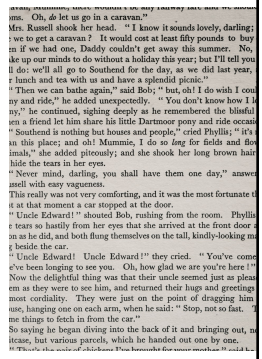


(i) Normal probability plots (ii) Regression (iii) Scatter plots (iv) maps
(v) Subtract values from mean (vi) change from last year's measurement (vii) statistical tests for outliers

37

(2) Summarize the Data

Do the values look reasonable?



38

(3) Annotate the Data

Mark data with quality control flags

- Verified
- Needs review
- Needs correction
- Data interpolated

39

96% → 99%

40