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) Cost of Lost Sales, etc.

3

(2) Cost of data cleanup

4

1-10-100 Rule

"the cost of quality"

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

$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

prevention is less costly than correction is less costly than *failure*

5

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

Demographic data: 2.3-5.2%

Treatment data: 10-26.9%

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: _____

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.

6

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 Rosvally
@DRosvally
Follow

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

10:22 AM - 21 Sep 2018

insecure, unavailable, unstructured (often unnecessarily) and unreliable. 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). The large scale of

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 *Information systems development: A Beginner's experience.* Unpublished MSc thesis, University of Warwick, UK.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

7

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....

8

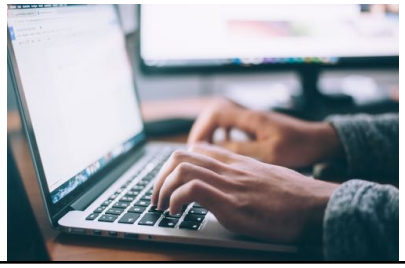
data acquisition

QA

Before Data Collection

During Data Collection

During Data Entry

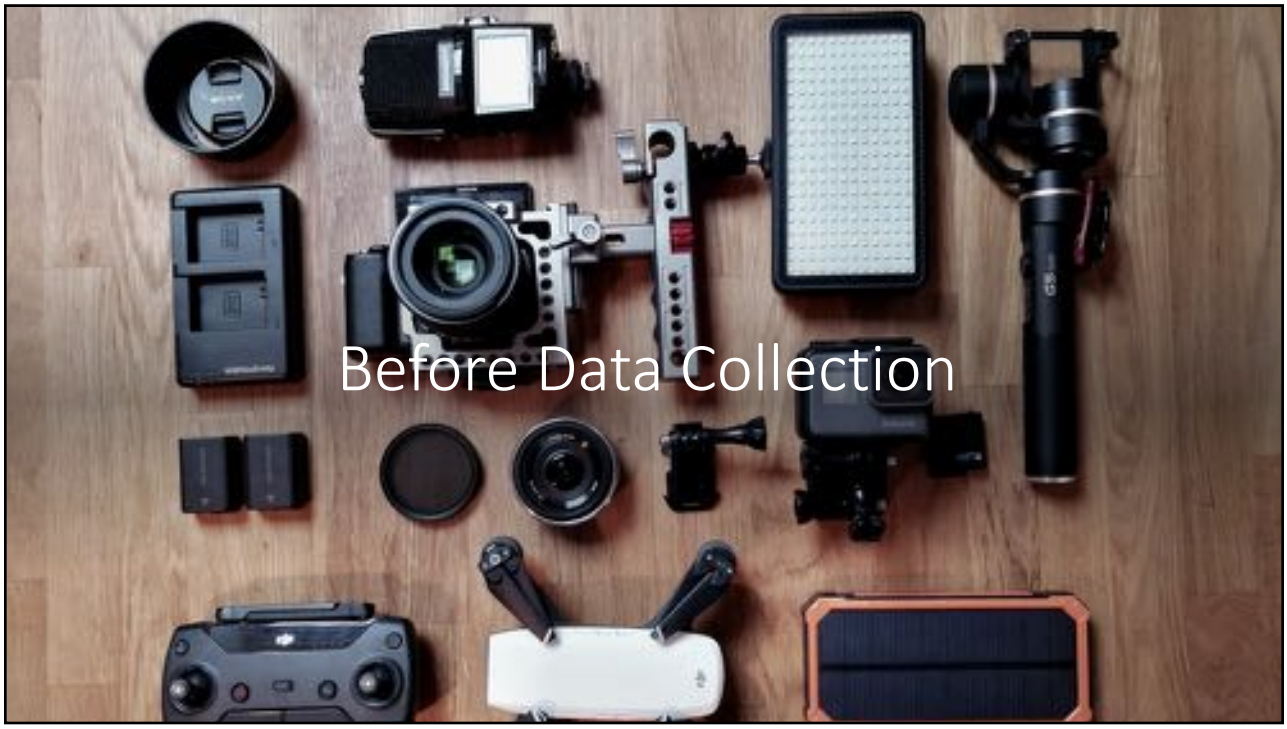


After Data Entry

data processing

QC

9



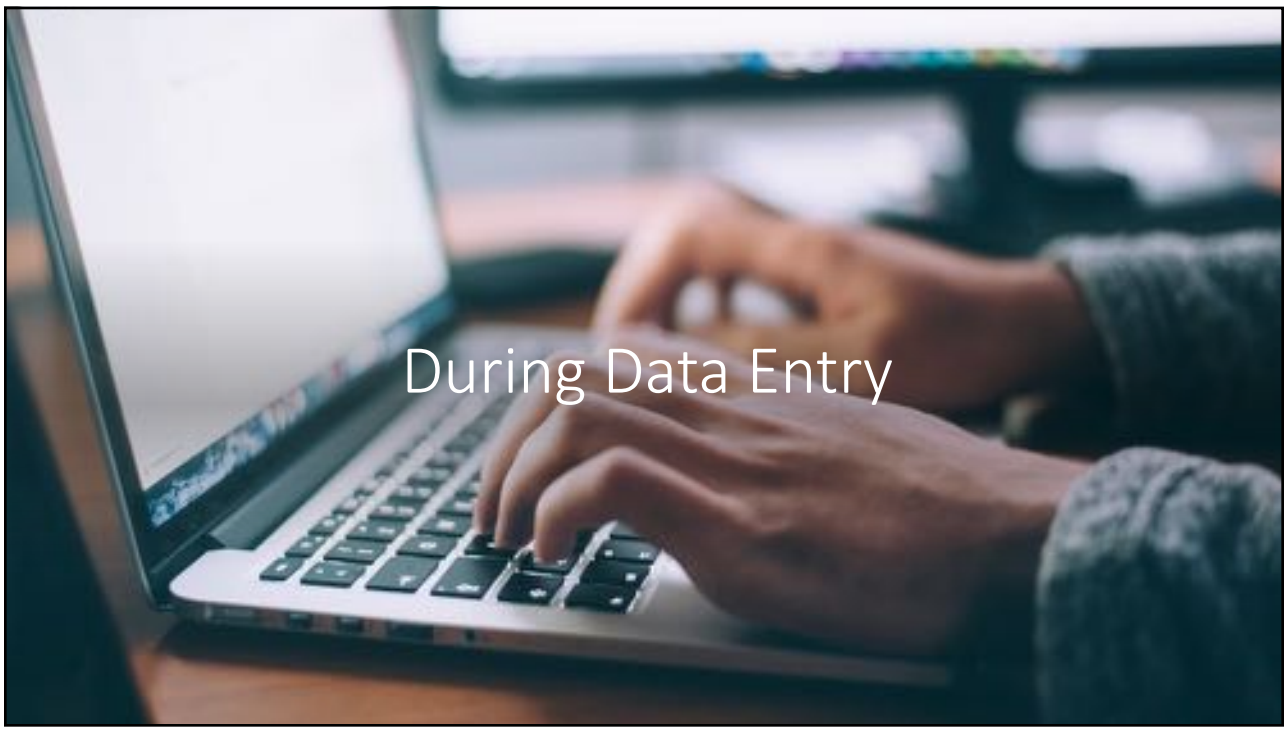
10

11

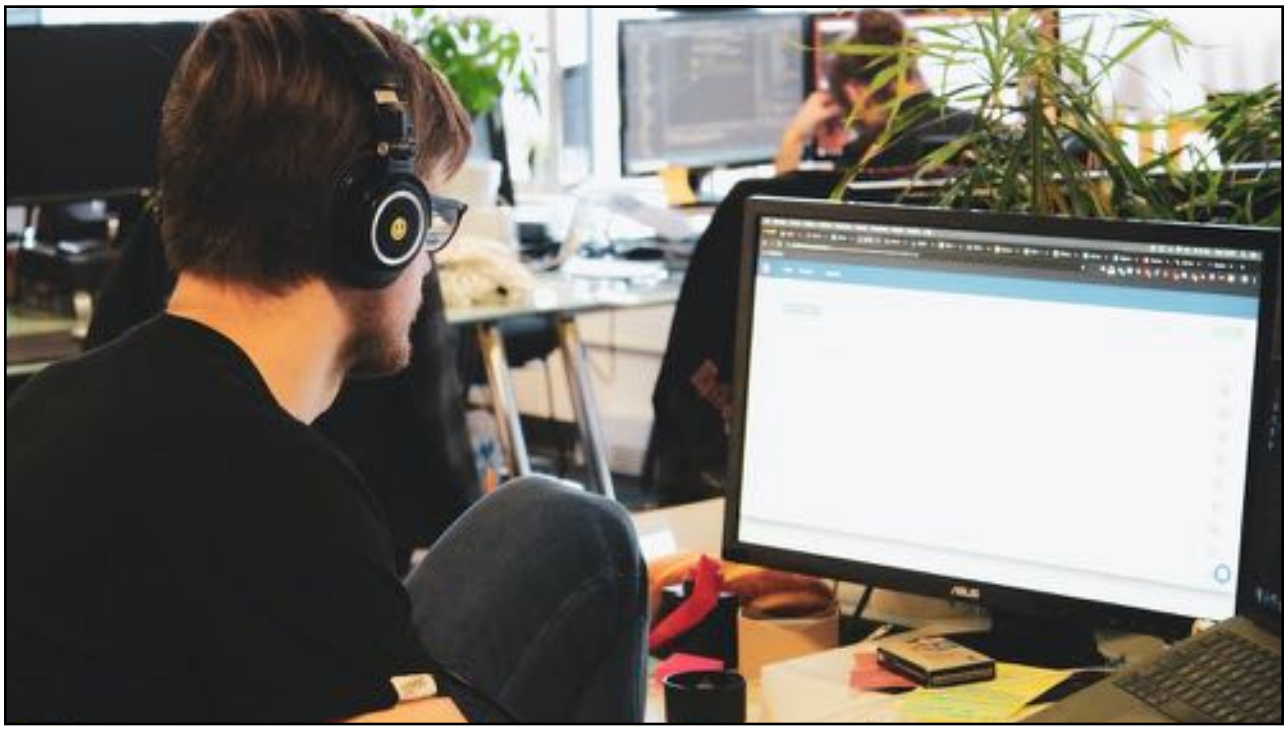
QA/QC

Quality Assurance / Quality Control

12



13



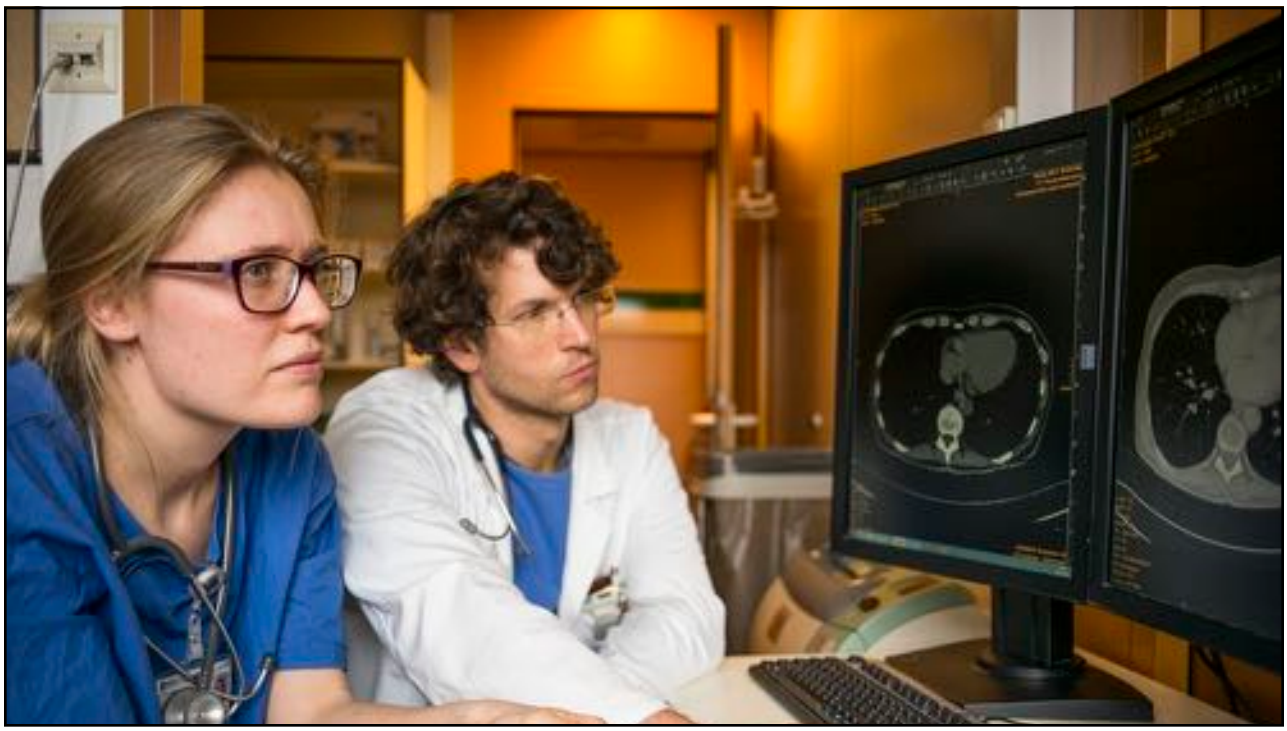
14

15

16

17

18



19

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
3 data entry methods (double entry, visual checking, single entry).
Each entered 30 data sheets, each w/ 6 types of data.

Results
Visual checking: 2958% more errors than double entry
Visual entry = not significantly better than single entry.

Perfect accuracy:
Double entry: 77.4% of participants
Visual checking: 17.1%
Single entry: 5.5%

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

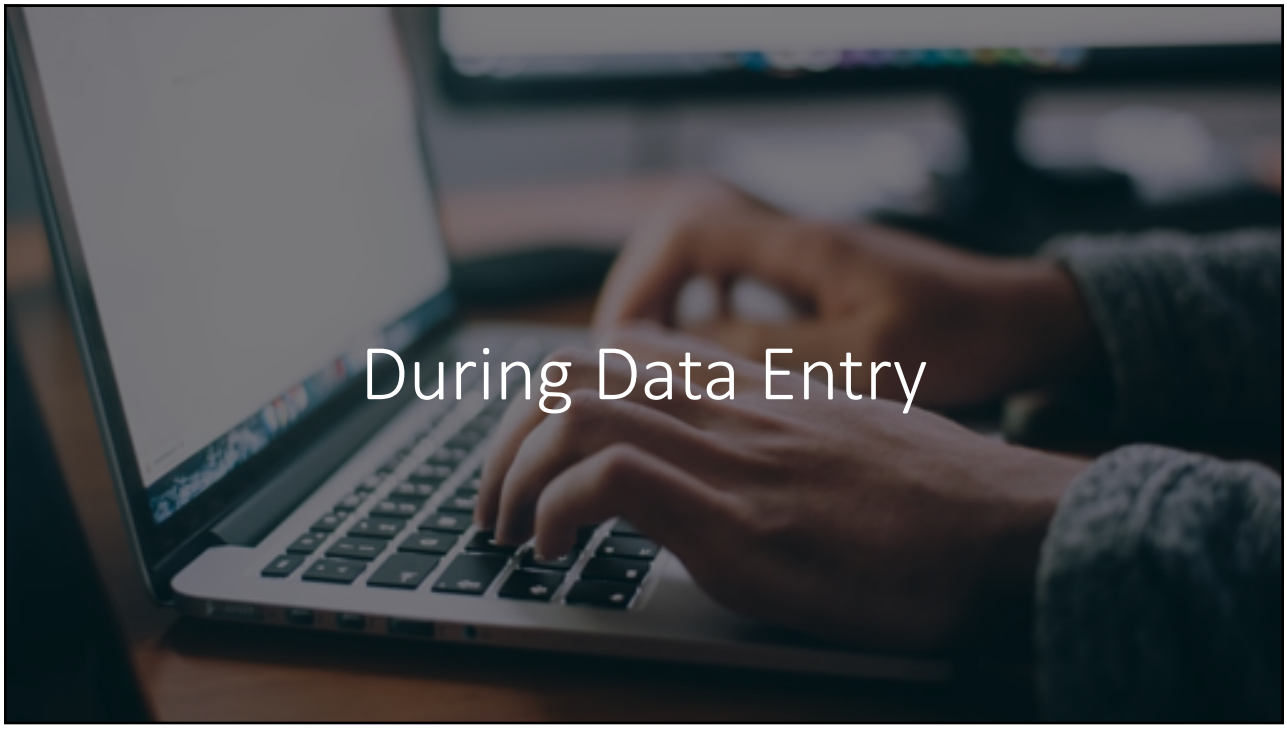
1.	1	2	3	4	(5)
2.	1	2	3	4	(5)
3.	1	2	(3)	4	5
4.	1	(2)	3	4	5
5.	1	(2)	3	4	5
6.	1	2	3	4	(5)
7.	1	2	(3)	4	5
8.	1	2	(3)	4	5
9.	1	(2)	3	4	5
10.	1	(2)	3	4	5

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

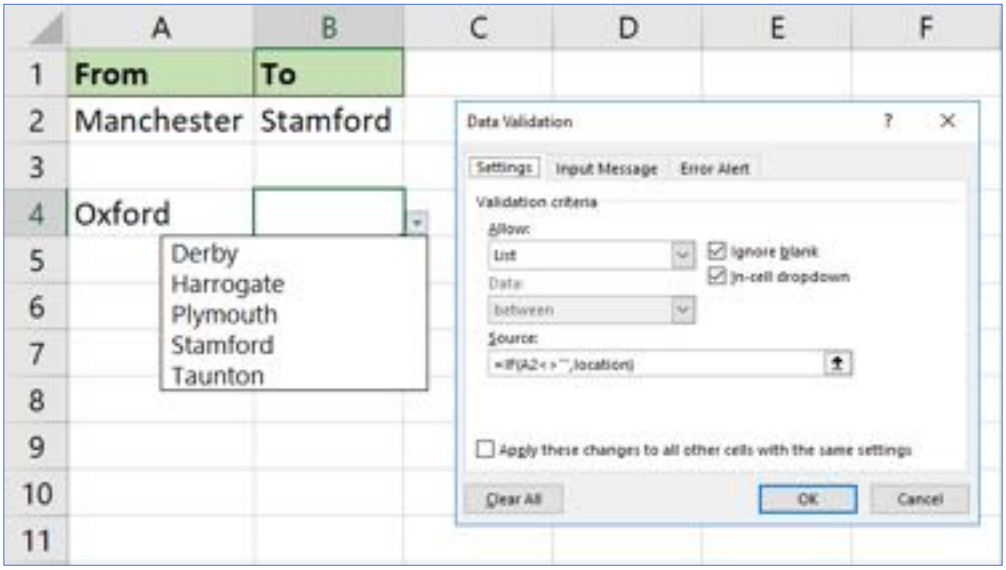
20

21



22

Data Validation



23

Atomized Data Entry

(one data type per cell)

genus	species	family	count
heliconia	acuminata	Heliconiaceae	3
heliconia	latispatha	Heliconiaceae	12
heliconia	tarumaensis	Heliconiaceae	5
heliconia	acuminata	Heliconiaceae	7
heliconia	sylvestris	Heliconiaceae	22
heliconia	bihai	Heliconiaceae	16

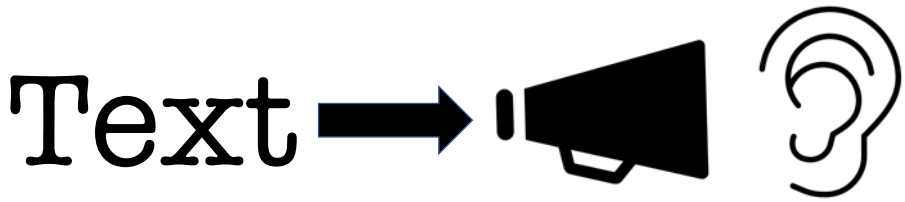
Simplify Entry using Codes, Eliminate duplication

species	count
heac	3
hela	12
heta	5
heac	7
hesy	22
hebi	16

code	genus	species	family
heac	heliconia	acuminata	Heliconiaceae
hela	heliconia	latispatha	Heliconiaceae
heta	heliconia	tarumaensis	Heliconiaceae
hesy	heliconia	sylvestris	Heliconiaceae
hebi	heliconia	bihai	Heliconiaceae

24

Single-user tool #1



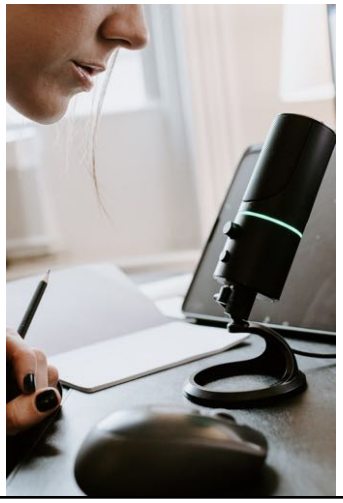
25

How does it work?

26

Single-user tool #2

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

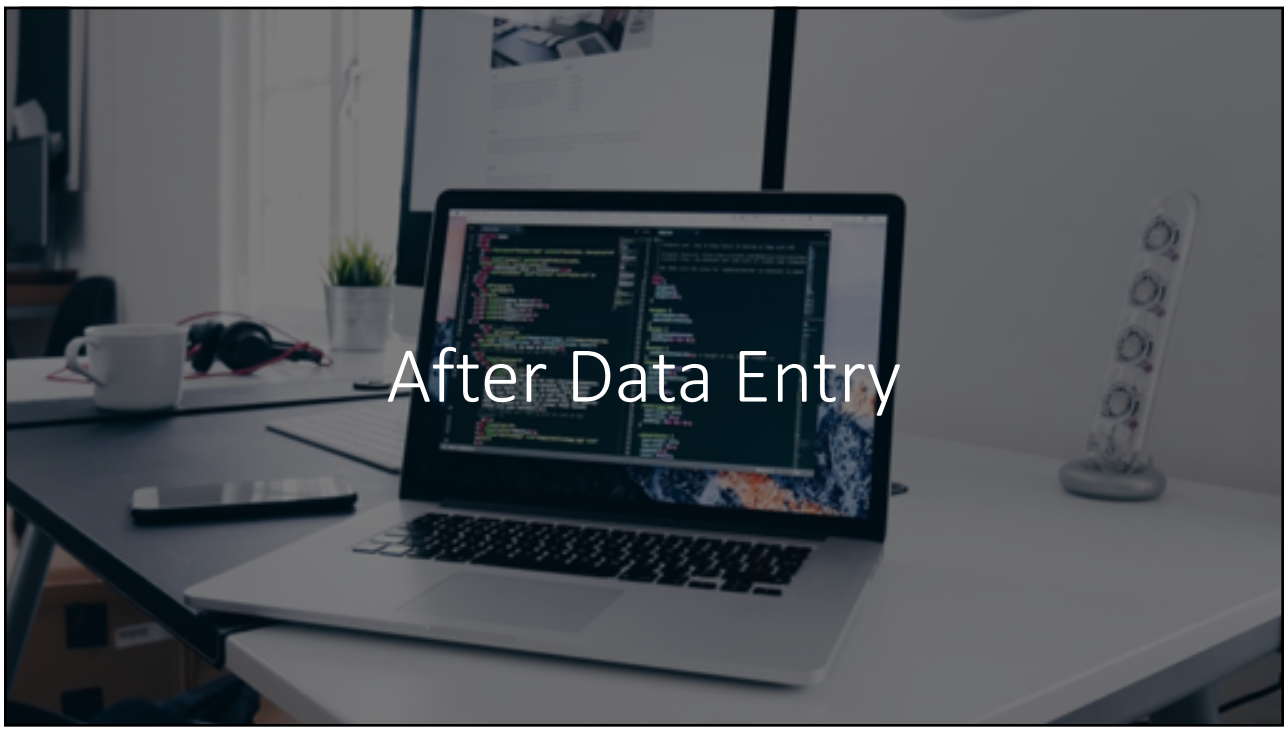


27

QA/QC

Quality Assurance / Quality Control

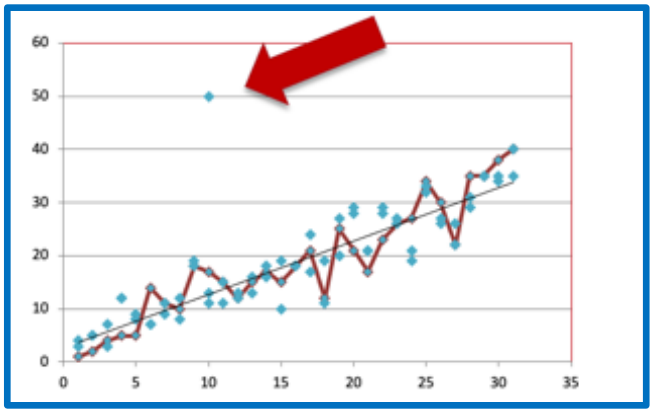
28

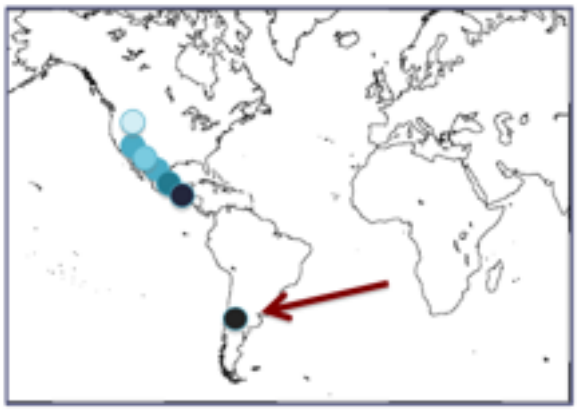


29

1) Visualize the Data

Great way to look for outliers





(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

30

2) Summarize the Data

Do the values look reasonable?

31

3) Annotate the Data

Mark data with quality control flags

- Verified
- Needs review
- Needs correction
- Data interpolated

32

96% → 99%

33