Lesson 3:
Assessing Data

SEARCH

RESOURCES

CONCEPTS

Mentor Help
Ask a mentor on our Q&A platform

Peer Chat 3
Chat with peers and alumni

How Data Gets Dirty and Messy

Sources of Dirty Data

Dirty data = low quality data = content issues

There are lots of sources of dirty data. Basically, anytime humans are involved, data. There are lots of ways in which we touch data we work with.

- We're going to have user entry errors.
- In some situations, we won't have any data coding standards, or where w they'll be poorly applied, causing problems in the resulting data
- We might have to integrate data where different schemas have been usec item.
- We'll have legacy data systems, where data wasn't coded when disc and r much more restrictive than they are now. Over time systems evolve. Neec changes.
- Some of our data won't have the unique identifiers it should.
- Other data will be lost in transformation from one format to another.
- And then, of course, there's always programmer error.
- And finally, data might have been corrupted in transmission or storage by physical phenomenon. So hey, one that's not our fault.

Sources of Messy Data

Messy data = untidy data = structural issues

Messy data is usually the result of poor data planning. Or a lack of awareness c data. Fortunately, messy data is usually much more easily addressable than m dirty data mentioned above.