Industry Insights 4: Susan Mcgregor

Steve Crawshaw

Watched 2022-07-20

Susan E McGregor is an Associate Research Scholar at Columbia University's Data Science Institute. Former data journalist at WSJ. She has published books on information security essentials and practical data wrangling and data quality.

Presentation

Integrity, Fit & Authenticity: Prerequisites for Effective Data

How to use journalistic rigour in the data analysis field and apply in a range of industries?

Data science = stats + computing + social science etc. Generate insight by combining disciplines. Models predict data - we are trying to generate insights about the world (!= data). Social science brings the ethical dimension and how implementation of models impacts society and individuals.

Data Integrity

- Timely (recent and available)
- Complete (redaction, truncation, sparse)
- Well annotated (meta data, data dictionary, units etc.)
- High Volume (possible to derive trends, principles for selecting a range)
- Historical (enough data to make an inference or prediction)
- Consistent (ppb vs ugm-3, SI vs imperial units, different q's wordings)
- Multivariate (how many features available and how generated)
- Atomic (granularity, averaging period)
- Clear (annotation, augmentation required to infer meaning)
- Dimensionally Structured (can values be connected to widely accepted standards - SI units, lat long etc)

Data Fit

- Validity (can the data provide the insight required? arrest rate == crime rate?)
- Reliability (does repetition of the analysis provide similar results)
- Representativeness (does the data generalise to a population [twitter is not real population])

Data Authenticity

The malleability of digital data can undermine confidence, which can be critical in applying the insights from analysis or data.

Can a digital journalist cite sources in a transparent and persistent way? (difficult). This is problematic.

Project Starling (photographs). Stanford. Twitter bot to track changes (diffs) in MSM (NYT). Not transparent to readers and therefore changes can create a different story from the original.

Offer transparency in an unchallengable way to increase trust.

Proposal to use blockchain to track changes in news by hashing diffs and updating blockchain - accessible with browser extension. Useful for archiving news and sources for legal cases based on stories.

Questions

Q - How to report on data quality in journalism? A - Happening a bit with covid tracker in US. Data viz can be a problem for unsophisticated readers. Need to be conversational to explain meanings.

Q - where is the blockchain project now? A - receiving requirements, pitching.

Q - how to DS ensure they don't introduce bias? A - Need to establish objectives *a priori* and not go fishing.

Q - for non stationary data, a sample cannot generalise to the population. How do we do that? A - recognise that you can only talk about the sample.

Q - What to do about deep fakes etc? A - Not a new phenomenon. Need to re - establish transparency e.g. local news outlets. Approach stories with an open mind. Disinformation is profitable, incentives are perverse.

Q - How will the increase of data shape journalism? A - There has been a shift to collaboration to analyse large data sets. E.g. wikileaks, Panama papers. This will continue, though it is

anithetical to competition for stories between news companies. Co - operation may drive complementary angles to large stories.