Rich Information Hides in Missing Data

Maria Wolters
Reader in Design Informatics
University of Edinburgh
maria.wolters@ed.ac.uk
@mariawolters

12th International Digital Curation Conference, Edinburgh

http://www.slideshare.net/mariawolters
Key Points

- Missing data can be informative, if they occur in a pattern
- Missing data can improve the overall data quality
- But only if we understand why data are missing
How figure was found


Source of the figure
Why this figure?

❖ Does the figure really show two triangles, or is it a design of an 80’s PacMan game?
❖ Do we want to know how the figure was found?
❖ Do we want to attribute the figure?
❖ Do we want to learn more about how graphics are discussed in fan forums?
What Is Missing Data?
Missing Data

- informally: observations that we would like to be there, or that should be there, but that are not

- Statistical treatment differs according to whether missing data are
  - completely random
  - predictable from existing data
  - not predictable from existing data
Research Goal

- Tell the hidden stories behind missing data by understanding and describing data generation processes
  - qualitatively for deeper understanding
  - quantitatively to feed into data analysis and visualisation
Missing Data is Informative: Appropriating Help4Mood
Depression is a change relative to an individual baseline.
Help4Mood: Supporting People with Depression

- daily monitoring
  - of **activity** using actigraph
  - of **mood, thought patterns & psycho-motor** symptoms using talking head GUI
- weekly one-page reports to clinicians

Pilot Randomised Controlled Trial

- Participants with Major Depressive Disorder (SCID diagnosed)
- Use Help4Mood for 4 weeks every day
- Background measures include demographics and attitudes to computers
- (Pre/Post measures to establish change)
- Qualitative interviews at intake and debriefing for those randomized to Help4Mood
Usage Patterns during Pilot RCT

- 18 in Romania, 7 in Scotland, 2 in Spain (EU Project)
- 14 treatment as usual (age 42 years +/- 10), 13 Help4Mood (age 35 +/- 12)
- None formally tracked or measured their mood before, but some used introspection
Even For Regular Users, Half the Data Were Missing!

- Half did not use it regularly, and half used it regularly
- Regular use was not daily; instead, it was 2-3 times per week. Why?
  - **Lack of mobility**: Platform was installed on a laptop, difficult to take on trips
  - **Self-Reporting is Work**: boring, tedious; or demanding
  - **Appropriation**: Users tweak technology to fit their needs, departing from initial design

---

If at all possible,
it would be good not to have the same questions every day;
or even if the questions are the same, the phrasing should be
different.
At some point it gets boring—I think this could be changed.
(RO15, female, 30–39)
“This wasn’t very pleasant. Because you don’t go to therapy every day. You wouldn’t go every day; you would go maybe once a week or two or three times maybe, but not every day. It’s a bit too much to use it every day.”
(P01, Case Studies)
The monitoring part helped me understand some things ... sometimes I did not realize how I felt that day, how happy I was or how active I was. The system helped me observe these things and also control them.

(RO14, female, 20–29)
Missing Data Is Informative

- People used Help4Mood in idiosyncratic ways
- Use versus non-use means different things for different people:
  - some may be bored by the questions
  - others may feel unable to confront them
Missing Data Can Improve Overall Data Quality: The ForgetIT Project
Humans Forget

% of syllables remembered vs. Elapsed Time

https://en.wikibooks.org/wiki/Cognitive_Psychology_and_Cognitive_Neuroscience/Memory
But That’s a Good Thing!

- We forget (irrelevant) details
- We forget what is not practiced
- We tend to forget distressing events
- We remember relevant information and patterns
Input:
Pigment on parchment

Processing
at various
levels of granularity;
feature detection

Output:
A woman.
The Mona Lisa
ForgetIT

- EU FP7 IP project, 2013-2016, led by L3S, Hanover
- Components:
  - Managed Forgetting
  - Contextualised Remembering
  - Synergetic Preservation
What is Managed Forgetting?

- Inspired by human forgetting

- Relies on two value judgements:
  - Is it important to preserve? (*preservation value*)
    e.g., papers given to be published in the International Journal of Digital Curation
  - Is it relevant at the moment? (*memory buoyancy*)
    e.g., conference hashtag #idcc17; presentation slides

Case Study: Personal Digital Photos

❖ **Question:**
   How do we implement managed forgetting for the personal use case of digital photos?

❖ **Method:**
   Large online survey (n=1221 valid answers) distributed via project web site, crowdsourcing, universities

❖ Sample culturally diverse with range of age groups, but recruitment doesn’t allow systematic cross cultural comparisons
Clustering of Commonly Used Strategies

- Strategies for
  - Photo Management
  - Photo Preservation
- Items taken from archivist recommended practices, Human Computer Interaction research on digital photo and digital memento management
- Clustered using agglomerative hierarchical clustering
Dimensions of Data Management

- Photo Management Practices: File versus Curate
  - Curators are more likely to annotate and label their photos and file them carefully
- Photo Preservation Practices: Safe vs Basic
  - Safe more likely to keep multiple copies in multiple places
  - Basic less likely to use any photo preservation practice
4 Personas

- Combinations of Preservation and Management Strategies
  - File and Forget (31.1%)
  - Safe Curator (28.9%)
  - Safe Filer (22.6%)
  - Basic Curator (17.4%)

- File and Forget less likely to worry about losing digital photos and less likely to find preservation important
Pilot ForgetIT Solution

- Developed by DFKI using Semantic Desktop infrastructure
- Managed forgetting:
  - duplicate detection
  - quality of photo
- Contextualised Remembering:
  - Automatic labelling of photos with content and spatiotemporal context
- Preservation value is set manually
Managed Forgetting Requires

- analysis of photos to detect content, common themes and events
- support for curating the photos that are kept
- educating users about preservation strategies and their cost
Can We Learn What to Forget?

- Key uses of digital photos are documentation and reminiscence.
- In reminiscence, photos serve as cues for autobiographical memory.
- Cues can be highly individual, which limits the degree to which they can be learned automatically.

Forgetting Data Improves Data Quality

- By abstracting from irrelevant (to be forgotten) detail, it is easier to process and analyse data.
- Judicious forgetting can make the remaining data easier to find and manage.
We Benefit Most From Missing Data If We Know Why It is Missing
Help4Mood

- Modelling individual tendencies using priors

- Examples:
  - For P01 („hard to cope with questions“):
    \[ p(\text{non-use} \mid \text{unwell}) > p(\text{use} \mid \text{unwell}) \]
  - For RO15 („boring!“):
    \[ p(\text{non-use} \mid \text{unwell}) = p(\text{use} \mid \text{unwell}) \]
  - For RO14 („helps make sense of feelings“):
    \[ p(\text{non-use} \mid \text{unwell}) < p(\text{use} \mid \text{unwell}) \]
Digital Photos

- Data from field evaluation of ForgetIT Pilot Prototype using photos of the Edinburgh Fringe Festival taken using cheap phone camera
- Amateur photographer deleted all their photos, because they weren’t good enough
- Foreign student kept photos that particularly reminded them of their time in Edinburgh
Summary

Missing data can be informative and inform data quality, but this requires triangulation with other data sources to understand reasons for missingness.

maria.wolters@ed.ac.uk * mariawolters.net
@mariawolters (also on LinkedIn)
Always open for collaboration!
It Doesn’t (Quite) Work This Way

Peer Support

Constant Unobtrusive Data Stream

Self-Help Internet-Based Therapy

http://www.acog.org/About-ACOG/ACOG-Departments/Long-Acting-Reversible-Contraception


http://imgarcade.com/1/depressed-stick-figure.png
It’s a complex adaptive system

Individualised monitoring based on what person has & does

Productive reflection and self-experimentation

Coping and getting better:
- Twitter, exercise, kindness
- Friends
- Medications
- GP

http://www.thebolditalic.com/articles/3609-the-stick-figure-guide-to-kicking-depression
Trackers also act as constant reminders, be they positive (desire to be healthy) or negative (reminder of illness).
The Story of Hidden Missing Data
Data go dark because of quality issues in data entry and management (e.g., Chan et al., 2014, Medical Care Research and Review)

People go to the doctor when worried about something, which increases likelihood of detection of other problems.

So does diabetes really increase your cancer risk, or is your cancer more likely to be spotted in regular checkups? (e.g., Badrick and Renehan, 2014, Eur J Cancer)
We can only detect what is being observed. Does that mean regular screening and monitoring?

What is the human cost of a false positive?

What is the human cost of a false negative?

What are the numbers needed to treat?