A few thoughts on health care data categories and acceptable use ?

Bruno Schröder

CTO, Microsoft BeLux

brunosch@microsoft.com

Pourquoi?

Une nouvelle approche de recherche scientifique basée sur les données et des méthodes statistiques avancées (souvent appelées AI ou ML)

Très prometteuse en prévention, médecine personnalisée et épidemiologie

Health care data are everywhere

Wearable sensors can tell when you are getting sick

New research from Stanford shows that fitness monitors and other wearable biosensors can tell when an individual's heart rate, skin temperature and other measures are abnormal, suggesting possible illness.

Geneticist Michael Snyder was wearing seven biosensors collecting data about his health when he noticed changes in his heart rate and oxygen level during a flight. When he later developed a fever, he suspected he had been infected with Lyme disease. Subsequent tests confirmed his suspicion.

https://med.stanford.edu/news/all-news/2017/01/wearable-sensors-can-tell-when-you-are-getting-sick.html

DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection

Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S. Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, Alex D. Leow

(Submitted on 23 Mar 2018)

The increasing use of electronic forms of communication presents new opportunities in the study of mental health, including the ability to investigate the manifestations of psychiatric diseases unobtrusively and in the setting of patients' daily lives. A pilot study to explore the possible connections between bipolar affective disorder and mobile phone usage was conducted. In this study, participants were provided a mobile phone to use as their primary phone. This phone was loaded with a custom keyboard that collected metadata consisting of keypress entry time and accelerometer movement. Individual character data with the exceptions of the backspace key and space bar were not collected due to privacy concerns. We propose an end-to-end deep architecture based on late fusion, named DeepMood, to model the multi-view metadata for the prediction of mood scores. Experimental results show that 90.31% prediction accuracy on the depression score can be achieved based on session-level mobile phone typing dynamics which is typically less than one minute. It demonstrates the feasibility of using mobile phone metadata to infer mood disturbance and severity.

Comments: KDD 2017
Subjects: **Human-Computer Interaction (cs.HC)**; Artificial Intelligence (cs.AI)
DOI: 10.1145/3097983.3098086
Cite as: **arXiv:1803.08986 [cs.HC]**
(or **arXiv:1803.08986v1 [cs.HC]** for this version)

How web search data might help diagnose serious illness earlier

Ryen White, chief technology officer for Microsoft Health and an information retrieval expert, says search queries may be used to predict pancreatic cancer.

Posted June 7, 2016 By **Mike Brunker**

The potential of using engagement with search engines to predict an eventual diagnosis – and possibly buy critical time for a medical response — is demonstrated in a new study by Microsoft researchers Eric Horvitz and Ryen White, along with former Microsoft intern and Columbia University doctoral candidate John Paparrizos.

“We find that signals about patterns of queries in search logs can predict the future appearance of queries that are highly suggestive of a diagnosis of pancreatic adenocarcinoma,” – the medical term for pancreatic cancer, the authors wrote. “We show specifically that we can identify 5 to 15 percent of cases while preserving extremely low false positive rates” of as low as 1 in 100,000.

The researchers used large-scale anonymized data and complied with best practices in ethics and privacy for the study.

Read more at http://blogs.microsoft.com/next/2016/06/07/how-web-search-data-might-help-diagnose-serious-illness-earlier/#hs2xBTWpJR85AWLH.99

Project

Premonition

Problem

When trying to predict the spread of a disease, every second counts. About 60%–75% of all emerging infectious diseases originate in animals, but it's difficult to pinpoint how, when, and where.

Solution

Microsoft Researchers are using mosquitos to collect blood samples from animals in the wild and identify the diseases they're carrying.

Project Premonition uses drones to find mosquito breeding grounds, robotic traps to gather specimens, and cloud-scale genomics powered by machine learning to search the specimens' DNA for pathogens.

"Using the Microsoft Cloud, we can analyze more than 100 million pieces of DNA in every sample," says Microsoft Researcher Ethan Jackson.

The Microsoft Cloud is fighting disease by turning mosquitos into data-gathering devices, and analyzing pathogen data … **so we may one day stop outbreaks before they begin.**

For more information, please see the Premonition video case study.

Source control is not possible

Correlation matters more than the data

Most often

Personal use and Public use

Care and Prevention use cases

Accessing data from the whole population:

- Managing better care through burn out, diabete and cancer prevention
- Reducing social security cost

An impact approach

What is the impact of a data confidentiality breach?

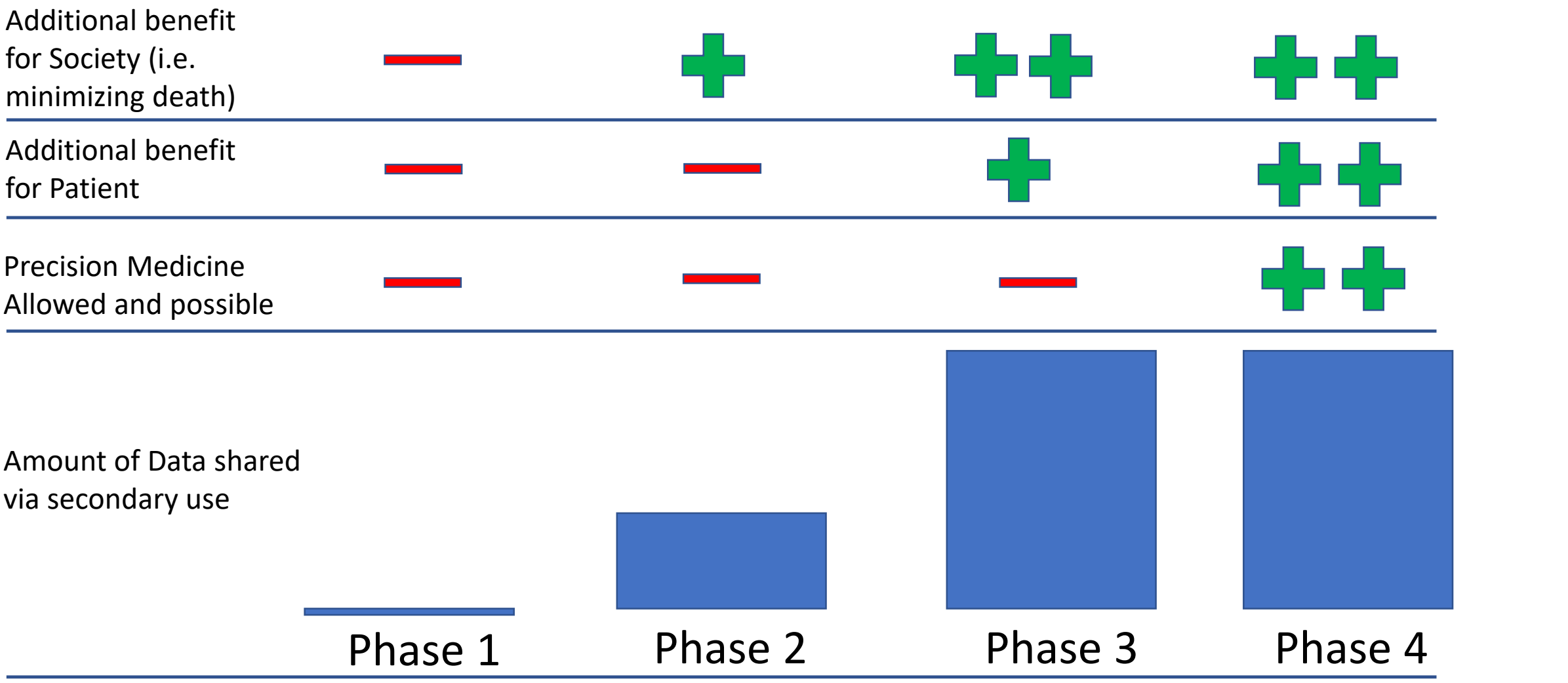
- Genome leak
- Condition leak
- Risk leak
- Wearable data leak
- Keywords search leak

An impact approach

What is the impact of a reuse of the data?

- Patient twin
- Early detection

Data Use Maturity Model



A possible classification

Some data are private, owned by the patient or the hospital

Some data are public

Some data should be public

Control should mostly be at use case level

Impact on individuals

Minor: history of consultation, visit to the doctor, blood test taken

Social: HIV or cancer testing

Work or revenue: HIV or cancer diagnosis

Descendants: genome, hereditary diseases

Impact on society

Data for prevention: pseudo anonymized data used to train detection algorithm (INAH)

Consent probably required

Data for cure: pseudo anonymized patient records used to improve cure

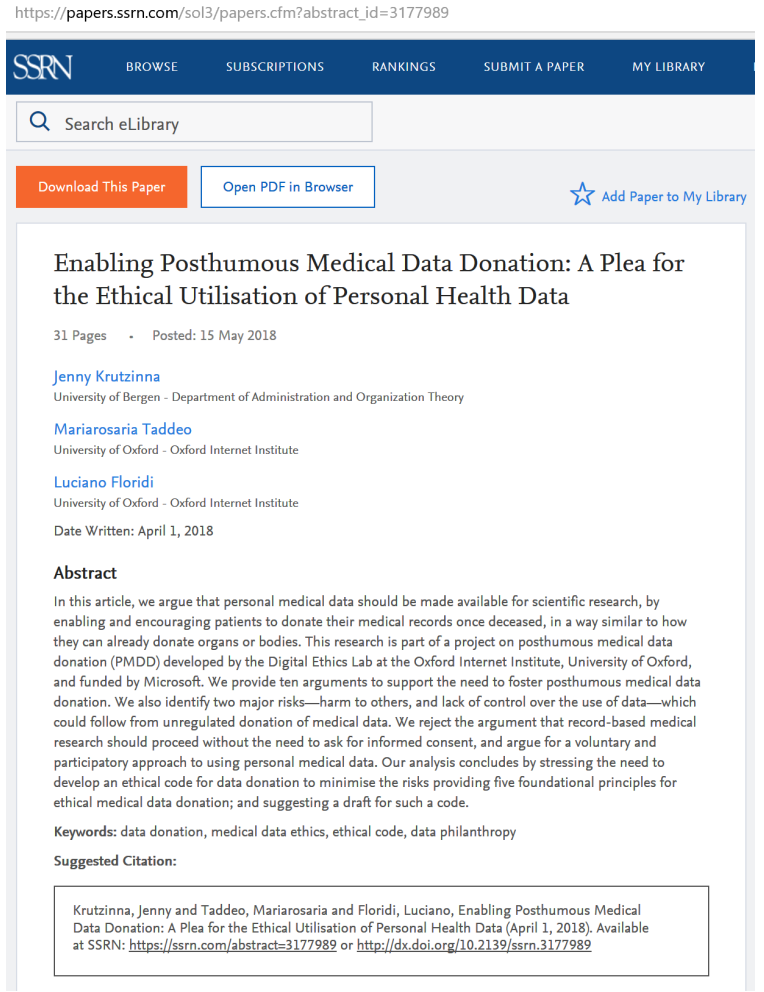
Data for cure: personal records used to cure the patient twin

Creating a data pool

Posthumous Medical Data Donation

Abstract

In this article, we argue that personal medical data should be made available for scientific research, by enabling and encouraging patients to donate their medical records once deceased, in a way similar to how they can already donate organs or bodies. This research is part of a project on posthumous medical data donation (PMDD) developed by the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford, and funded by Microsoft. We provide ten arguments to support the need to foster posthumous medical data donation. We also identify two major risks—harm to others, and lack of control over the use of data—which could follow from unregulated donation of medical data. We reject the argument that record-based medical research should proceed without the need to ask for informed consent, and argue for a voluntary and participatory approach to using personal medical data. Our analysis concludes by stressing the need to develop an ethical code for data donation to minimise the risks providing five foundational principles for ethical medical data donation; and suggesting a draft for such a code.



Common Definition

ISO/IEC FDIS 19944

Information technology — Cloud computing

Cloud services and devices : data flow, data categories and data use

This document provides a description of the ecosystem of devices and cloud services and the related flows of data between cloud services, cloud service customers, cloud service users and their devices. These are necessary to provide guidance about how data is used on the devices in the context of the cloud computing ecosystem, and the associated location and identity issues that emerge from such use.

This document proposes a scheme for the structure of data use statements that can be used by cloud service providers to help cloud service customers understand and protect the privacy and confidentiality of their data and their users' data through increased transparency of policies and practices.

ISO/IEC FDIS 19944: data categories

- Identified data
  - Data that can unambiguously be associated with a specific person because PII is observable in the information. Guidance on what can be considered as identifiers can be found in 4.4.1 of ISO/IEC 29100:2011[04].
- Pseudonymized data
  - Data for which all identifiers are substituted by aliases for which the alias assignment is such that it cannot be reversed by reasonable efforts of anyone other than the party that performed them.
    This corresponds to data defined as “pseudonymization” in 2.24 and described as “pseudonymous data” in 4.4.4; both in ISO/IEC 29100:2011.
- Unlinked pseudonymized data
  - Data for which all identifiers are erased or substituted by aliases for which the assignment function is erased or irreversible, such that the linkage cannot be re-established by reasonable efforts of anyone including the party that performed them.
- Anonymized data
  - Data that is unlinked and which attributes are altered (e.g., attributes’ values are randomized or generalized) in such a way that there is a reasonable level of confidence that a person cannot be identified, directly or indirectly, by the data alone or in combination with other data.
    This corresponds to data defined as “anonymized data” in 2.3 and process defined as “anonymization” in 2.2; both in ISO/IEC 29100:2011.
- Aggregated data
  - Statistical data that does not contain individual-level entries and is combined from information about enough different persons that individual-level attributes are not identifiable.

Data types: ISO/IEC DIS 17788 definitions

- **Cloud service customer data**: class of data objects under the control of the cloud service customer that were input to the cloud service or resulted from exercising the capabilities of the cloud service by or on behalf of the cloud service customer on those data objects
- **Cloud service provider data**: class of data objects, specific to the operation of the cloud service, under the control of the cloud service provider
- **Cloud service derived data**: class of data objects under cloud service provider control that are derived as a result of interaction with the cloud service by the cloud service customer

A PSD-2 in Health?

Once we have a common framework

Work in progress!

Thank you.