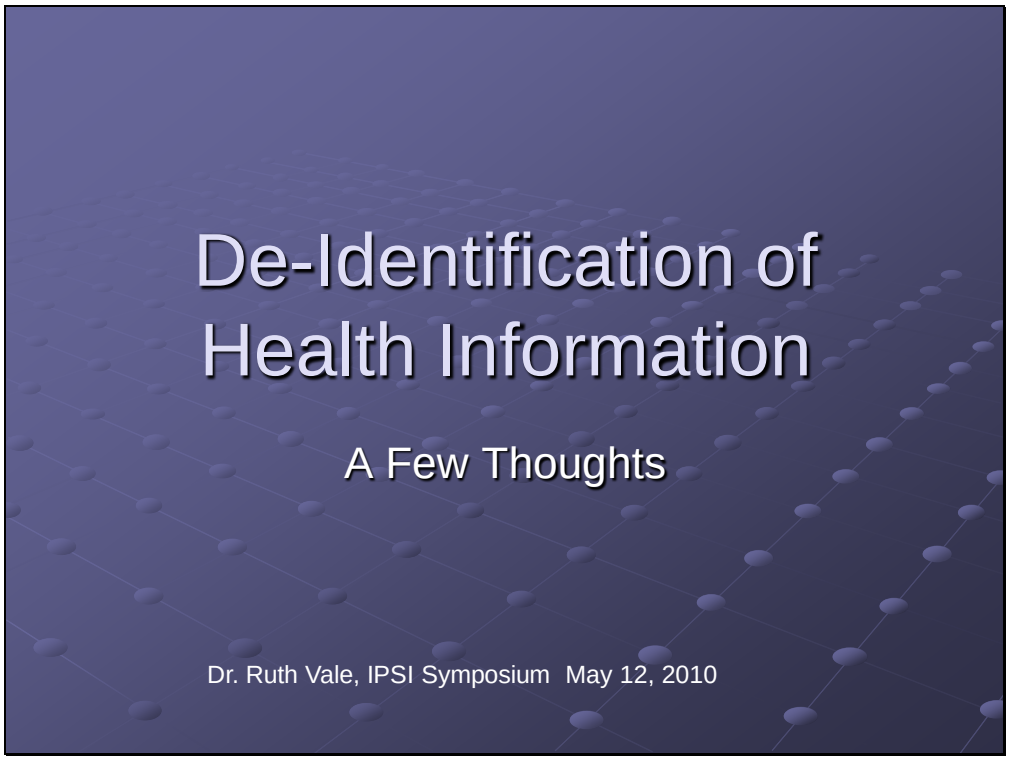
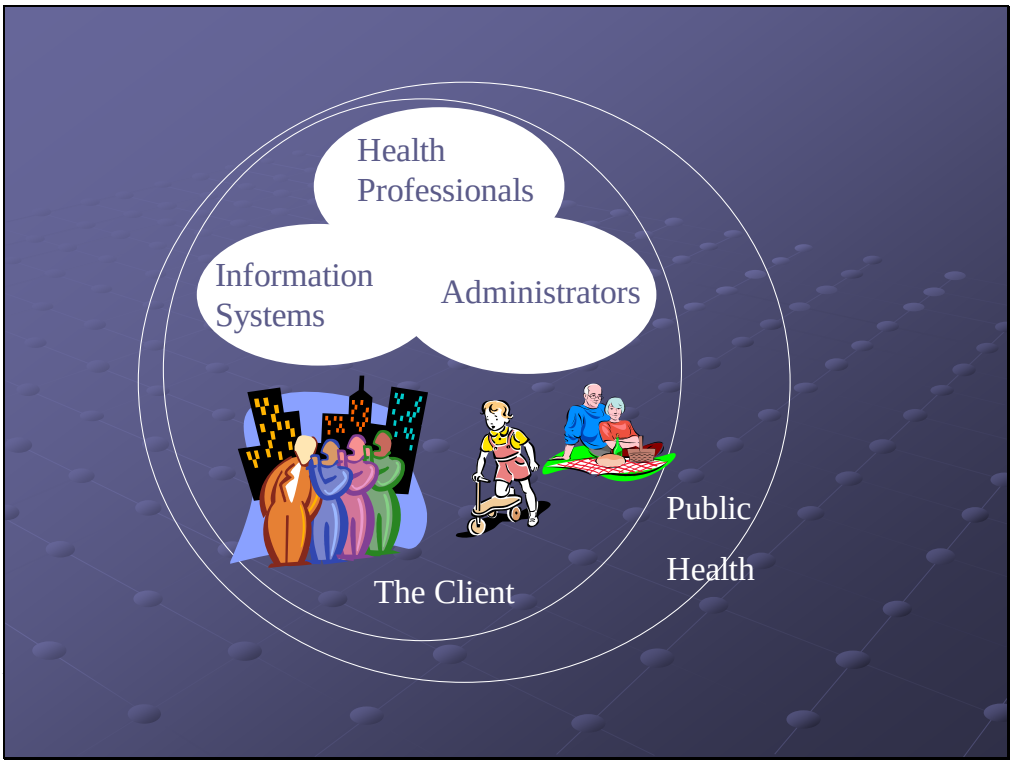
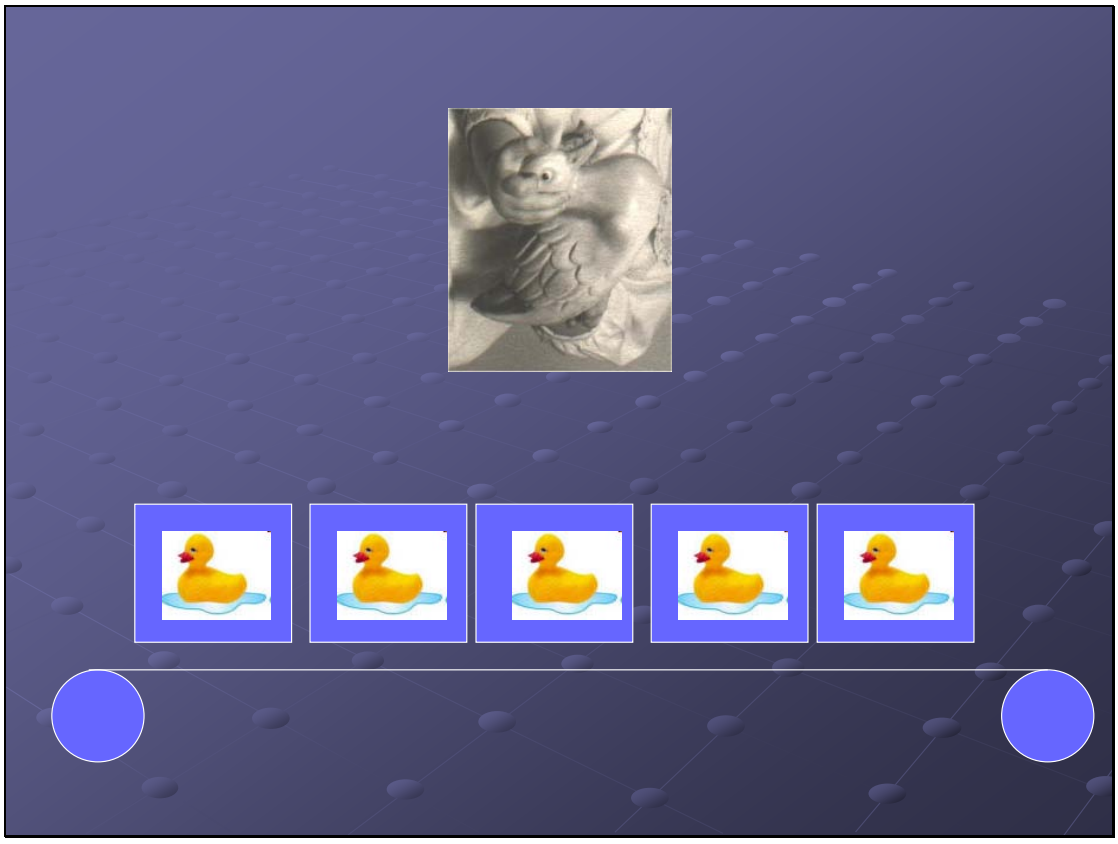


There is a notice on the door of the walk in clinic in my community which says “You may present only ONE problem to the doctor per visit.” The notice then explains why this is a normal and acceptable practice. The practice is intended to manage the volume and the complexity of health care, and the brevity of time, in a service context where a health problem is identified, examined and treated in a single transaction.

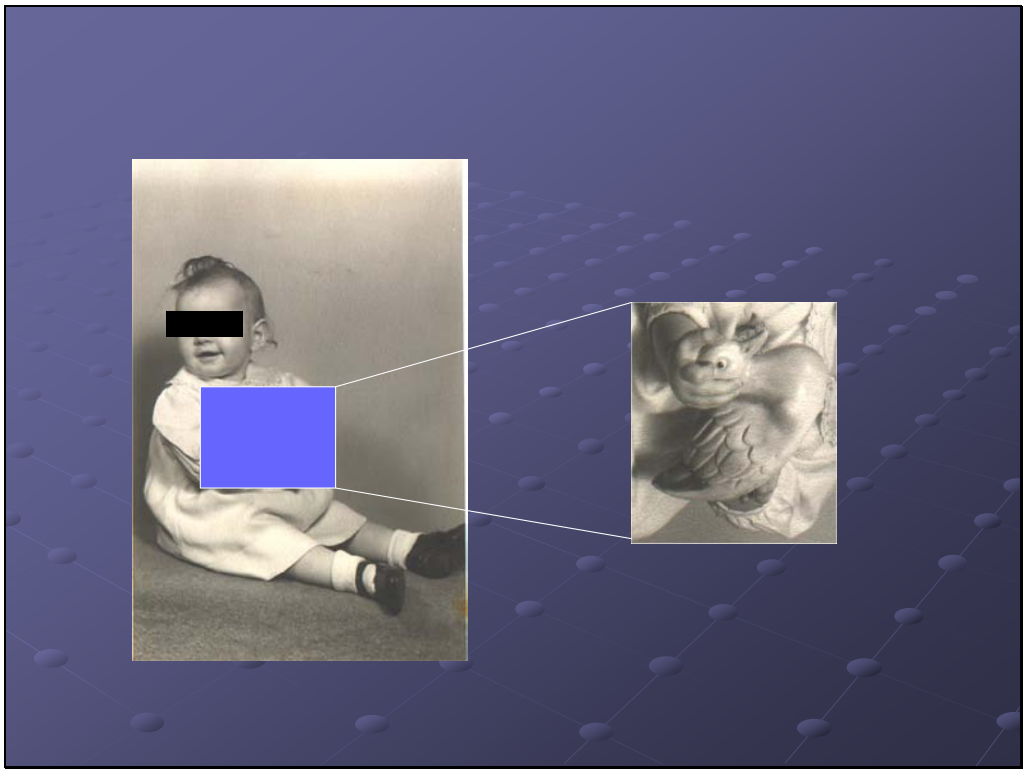
In brief, the individual presents him or herself for observation. The doctor may take samples and then draw conclusions about the particular remedies which might address the client's complaints. The information collected in this transaction is passed to an increasingly complex array of specialists, examiners and administrators for purposes increasingly distant from the single encounter between the individual client and their health care provider.



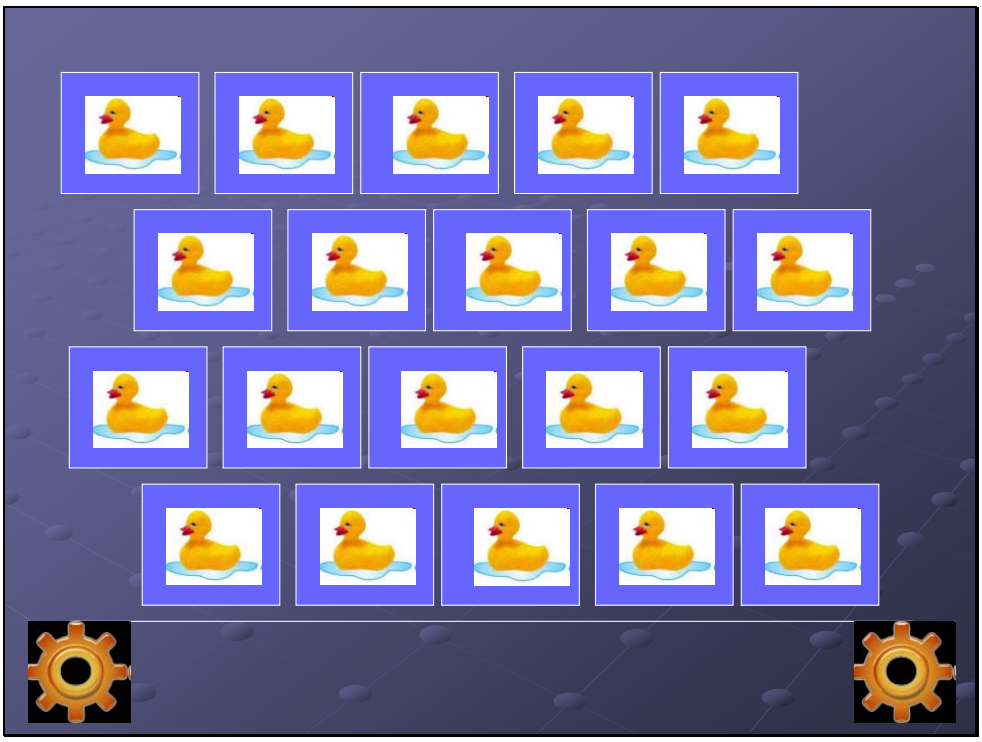
The information is gathered into systems, spawning more systems to collect, use, store and disclose. We could say that the value of the initial transaction is returned to the individual when these systems have completed their work. Studies have been concluded. Remedies are implemented and on the basis of this work, the doctor issues a prescription. Information which pertains to the patient or client as an individual is collected and in exchange the individual receives medical services, on completion of which the patient walks out the door.



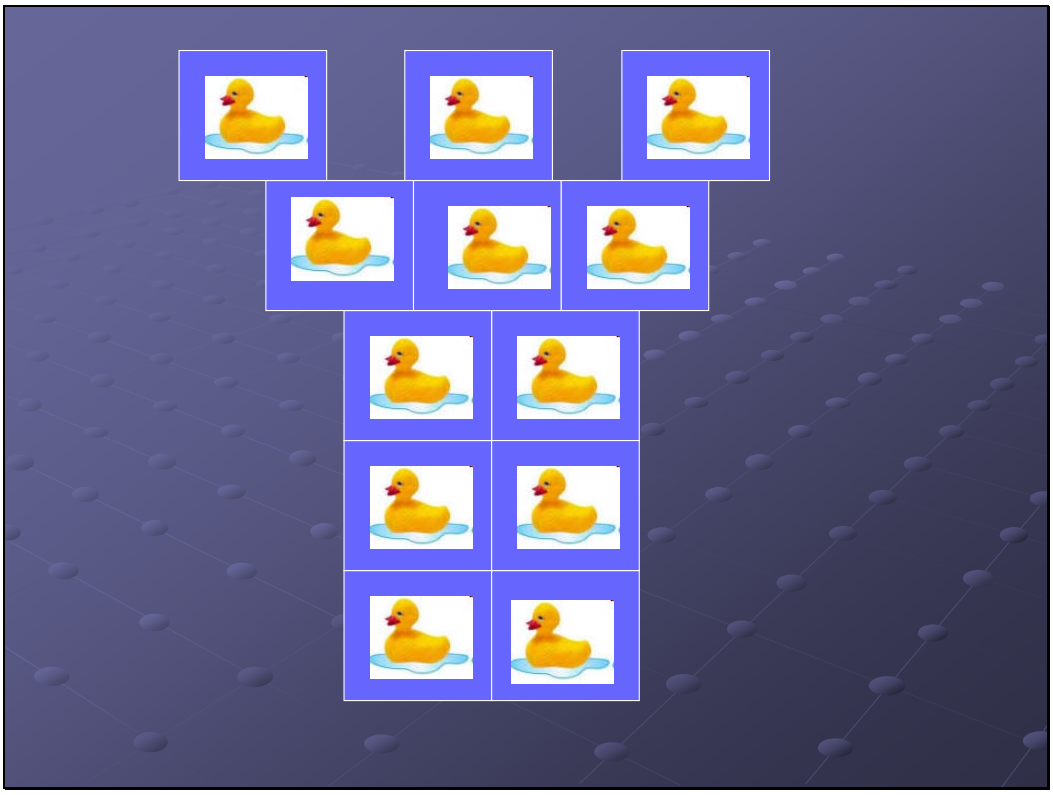
Let me get to the heart of the matter by talking about precisely what I mean by “de-identification” as it relates to the notice on the clinic door. The issue, from a privacy point of view, is that personal health information is collected into databases of such volume and complexity that it is beyond the ability of the individual to exercise any meaningful or practical control over their miniscule part.



The resolution to this dilemma is that we acknowledge this gap and address it by removing those elements which mark ownership, which make it “about me” and “mine” and therefore something that “ought to be within my control”. The perception, and perhaps misperception, is that if we can de-identify in sufficient degree, then society can benefit from the use of such collections and still give individuals a meaningful sense of privacy – that is – control over the information that originates from them, and in a sense, remains theirs as long as it is identified.



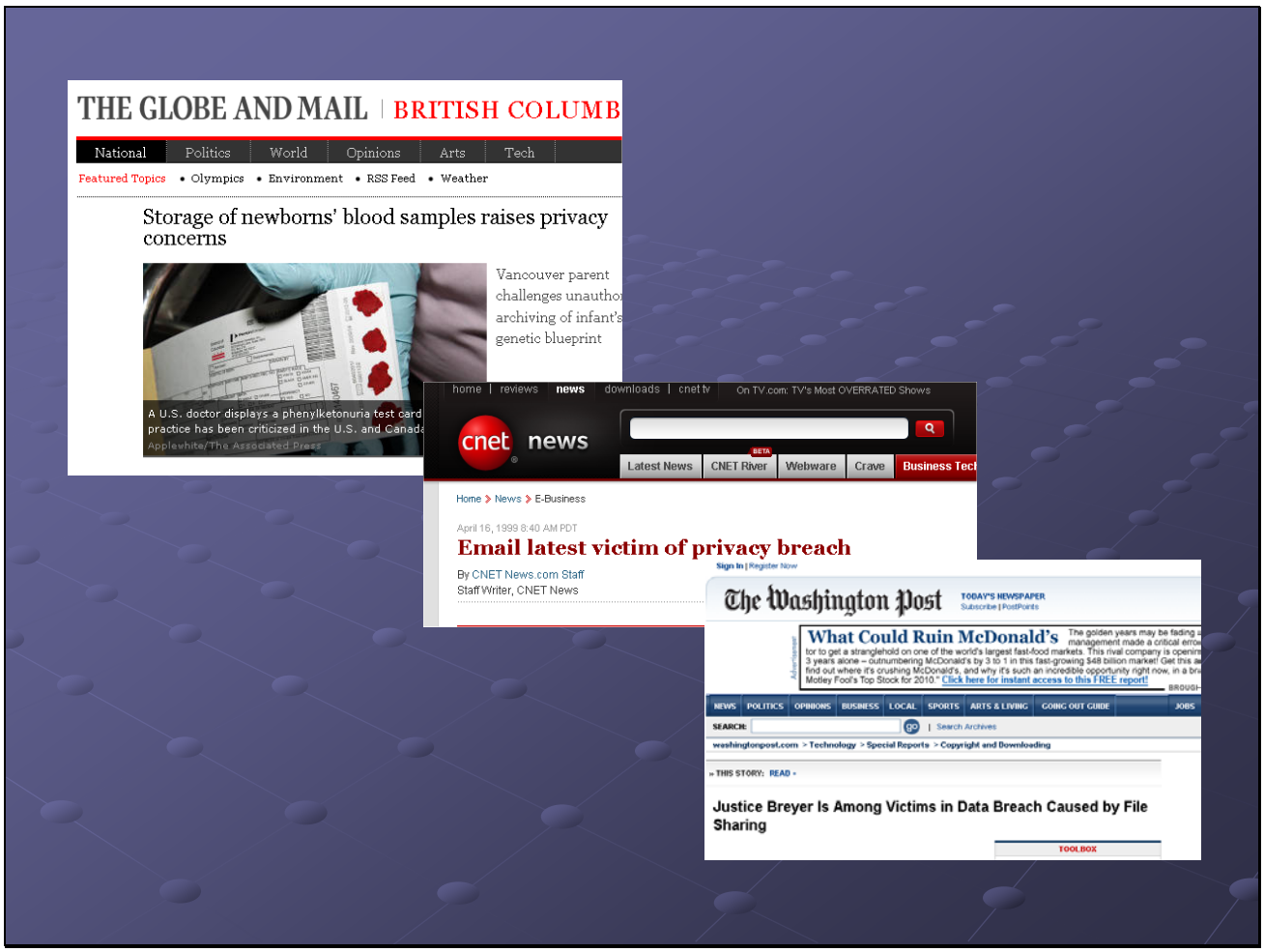
The individual’s information is added to the data sets which support the medical community in its search for knowledge about our species. To accomplish that end, it may be detached from the individual who provided the information. As the volume and complexity of these contributions accumulate into collections, and such collections are examined, parsed and their parts redistributed…



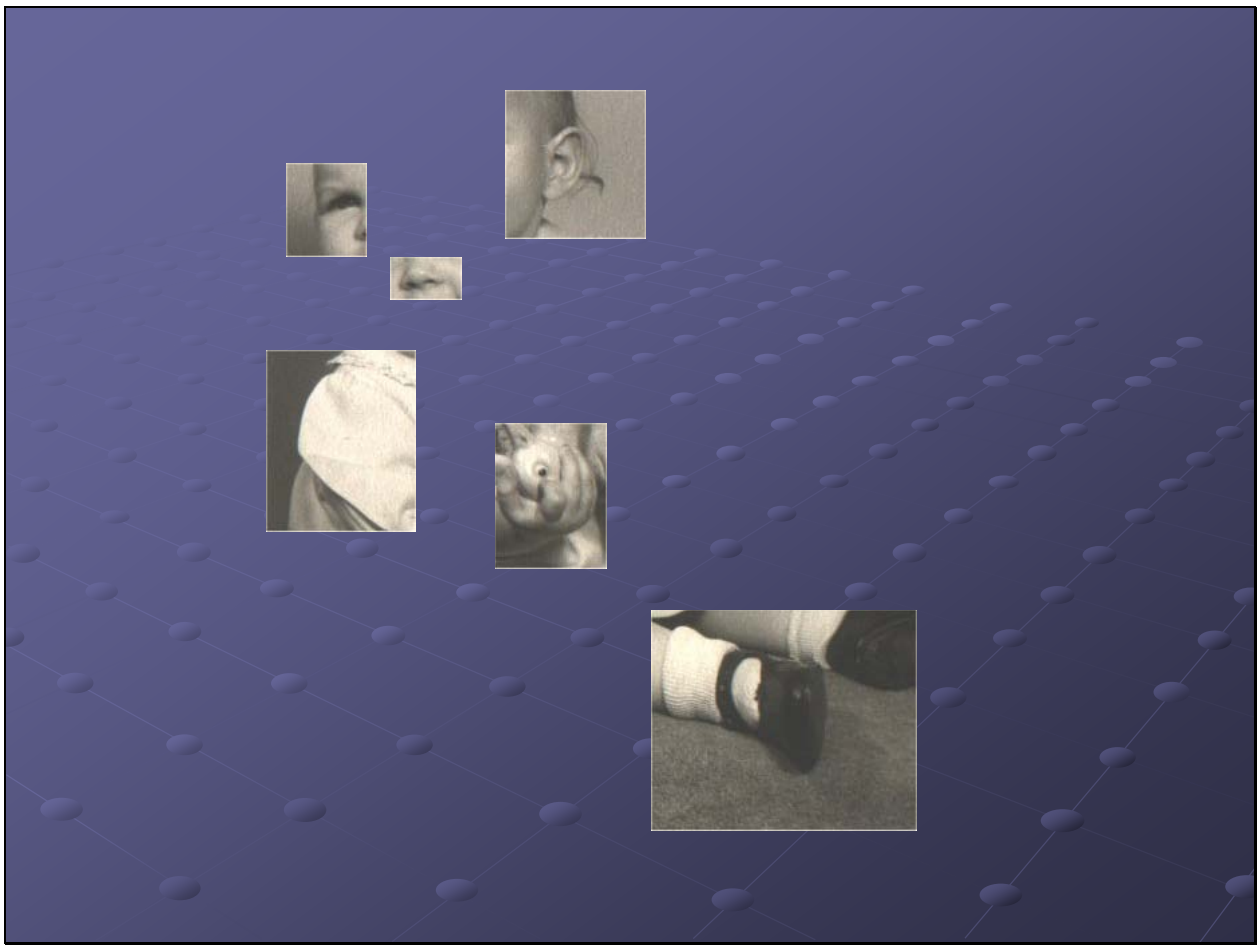
…the byproduct of the basic transaction becomes a commodity with a life and a value all its own. Harvested from individuals, the value of the individual's contribution is returned as informed care. Detached from the individual, and added to information from others, it forms a valuable database from which modern medicine grows a body of knowledge about the human species. This, then, is the context and the argument for de-identification.

This panel will be talking about protecting individual identity and privacy in an atmosphere of ubiquitous surveillance. As beneficial as they might be from the viewpoint of the patient, **universal health records** provide an opportunity for socially-accepted surveillance, growing from an increased and pervasive variety of collection points in a society where observation is ambient and surveillance is ubiquitous.

Individual information is collected at a detail level and handled in large scale and complex information systems which run on automated networks. Personal health information initially presented in the context of a single problem at a walk in clinic is a commodity in health information systems.

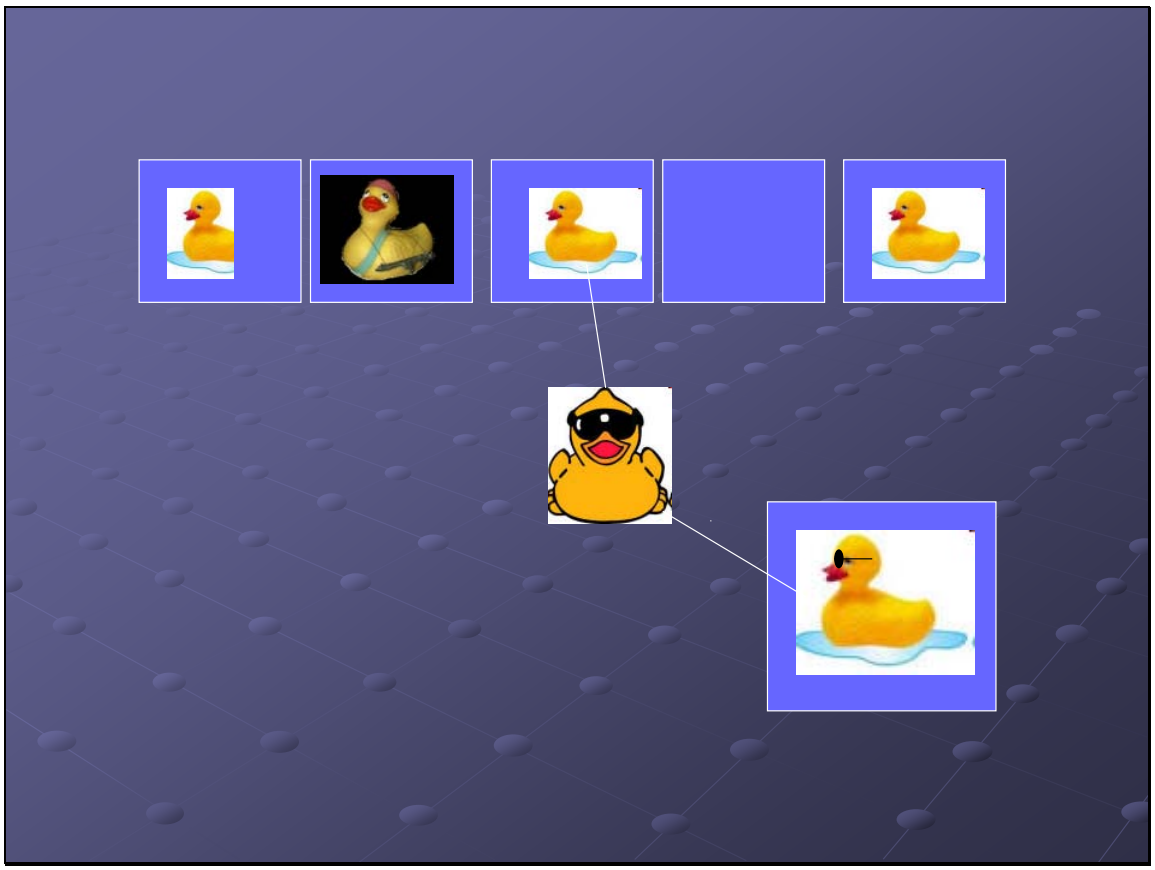


Too frequently, we trip over the privacy risks, amounting to: “Oops, so sorry!” uttered in the rush to harvest information collected by means of our many tools, sensors and databases.

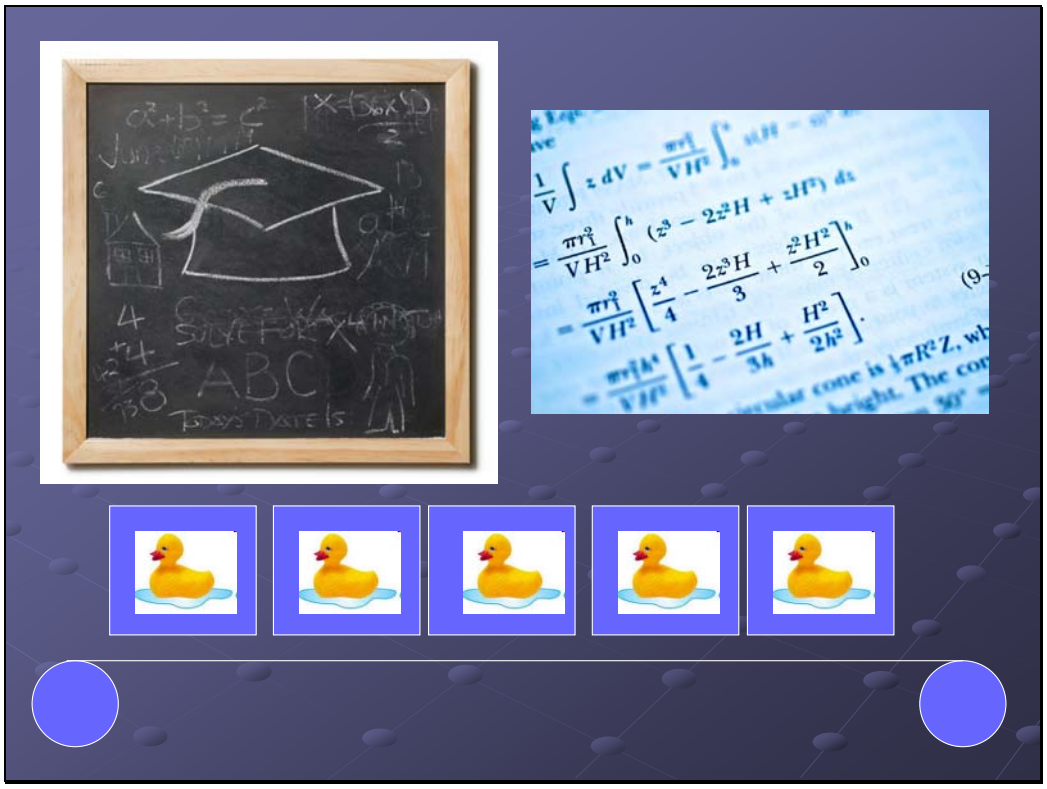


In terms of method, de-identification is the process of altering associations so that the data subject and the object are detached from each other. The degree of de-identification is not binary, as an either/ or proposition, but it moves along a continuum from fully identified records to completely anonymous data. The difficulty is that the less you alter the data the more useful it is in the context of the research questions and more likely that output will expose the individual who is the source of the information. If we alter a record such that it is unknown and unrecognized as belonging to an identified individual, then we may well lose the value it has by virtue of its attachment to the individual while we gain the use of the rest of the data set to generalize to the species as a whole.

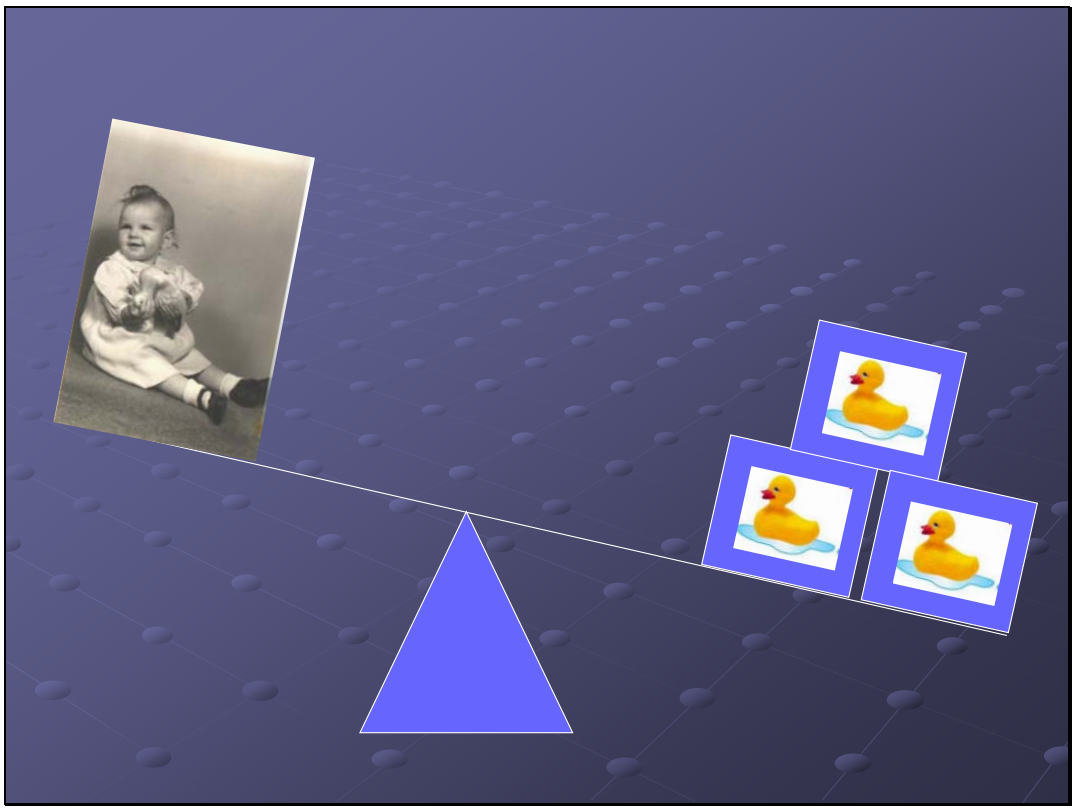
We have generally taken the simplest approach to de-identification by isolating specific data fields and treating them as offenders. For example, HIPPAA identifies eighteen data fields which can be removed or substituted, after the data set is considered “de-identified” and anonymous. We have discovered, however that it is possible to remove the obvious identifiers and from there find it relatively easy to fill in the blanks on the basis of context and inference.



In brief, other methods include the following: **Data detail** may be rounded, truncated or sampled. Alternatively, the data may be altered with random addition of noise to the data, randomization of the data values and data swapping within similar values. **Data Suppression** is accomplished by leaving out fields or content. **Pseudonymity** is accomplished when a series of records are created which appear anonymous to the user but the records can be related by the system to an individual from one case to another without revealing identity. The pseudonym is used to link otherwise de-identified data to the same individual across multiple data records or information system without revealing the identity of the individual, which remains hidden when the key is in the possession of a trusted third party.



Heuristic Methods create generalizations using mathematical and statistical techniques to exclude or obscure variables from a data set when it is disclosed. Statistical studies in re-identification have shown that some fields render a higher re-identification risk than others: especially gender, date of birth, date of service and location co-ordinates of service and address codes, the latter being particularly sensitive to the population size attached to that code.



There has been some progress in developing tools to measure re-identification risk, using a statistical approach, which enables organizations to make risk-based decisions about releasing part or all of their holdings. These studies have identified thresholds beyond which the re-identification risk for some data sets, especially small ones, is significant. The methods remain in debate, but it is clear that de-identified result is never absolute. The risk is never zero. There is ongoing risk and challenge to the exercise of detaching an individual from their information profile.

The CREATE proposal is exciting because where it proposes to address de-identification (and here I read from the proposal itself), the program addresses an “inherently interdisciplinary challenge, which needs highly skilled personnel who not only have deep knowledge in the various technical areas, but who can also appreciate alternative perspectives and collaborate with others who have complementary skill sets.” I look forward to hearing more about the contribution this practice can make to secure and privacy sensitive infrastructure.