L. Sweeney, Information Explosion. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, L. Zayatz, P. Doyle, J. Theeuwes and J. Lane (eds), Urban Institute, Washington, DC, 2001.

# Information Explosion

Latanya Sweeney, Ph.D.
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA USA
E-mail: latanya@cs.cmu.edu

## 1 Abstract

In this chapter, I examine the tremendous growth in information being collected on individuals. From the examples provided in this chapter, it is clear that many details in the lives of most people are being documented in databases somewhere. I provide examples that exemplify recent behavioral tendencies in the collection of person-specific data. These tendencies are: (1) given an existing person-specific data collection, expand the number of fields being collected; I term this the "*collect more*" trend; (2) replace an existing aggregate data collection with a person-specific one; I term this the "*collect specifically*" trend; and, (3) given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection related to the question, problem or opportunity; I term this the "*collect it if you can*" trend. No matter how you look at it, all three tendencies result in more and more information being collected on individuals. Having so much sensitive information available makes it even more difficult for other organizations to release information that are effectively anonymous.

## 2 Introduction

Society is experiencing unprecedented growth in the number and variety of data collections as computer technology, network connectivity and disk storage space becomes increasingly affordable. Data holders operating autonomously and with limited knowledge are left with the difficulty of releasing information that does not compromise privacy, confidentiality or national interests. In many cases the survival of the database itself depends on the data holder's ability to produce anonymous data because not releasing such information at all may obstruct the goals for which the data were collected, while on the other hand, failing to provide proper protection within a release may create circumstances that harm the public or others. Ironically, the broad availability of public information makes it increasingly difficult to provide data that are effectively anonymous.

| Examples | 1983 | 1996 |
|---|---|---|
| Each birth | 280 | 1864 |
| Each hospital visit | 0 | 663 |
| Each grocery visit | 32 | 1272 |

**Figure 1 Estimated growth in data collections (per encounter) in Illinois (in bytes)**

Figure 1 demonstrates how some data collections expanded from 1983 to 1996 for some person-specific encounters in the State of Illinois. The values are the number of *bytes* (letters, digits and other printable characters) that were stored for each person per encounter in the collection shown. Figure **2** describes how the estimates used in Figure 1 were computed. The values shown in Figure 1 and Figure **2** are the number of printable characters that are reserved for providing the information.

In the next subsections, I further explain these figures and show that they are representative of many experiences in most states.

| | |
|---|---|
| 1983 Birth certificate (280) based on field sizes:<br>40 Name<br>80 Address<br>80 Parent's names<br>80 Hospital name & address | 1996 Birth certificate (1864) based on field sizes:<br>40 Name<br>80 Address<br>80 Parent's names<br>80 Hospital name & address<br>1584 Birth characteristics |
| 1983 Health care cost data (0) based on field sizes:<br>*no such collection existed* | 1996 Health care cost data (663) based on field sizes with noted codes expanded to include textual description:<br>263 Primary fields (see Figure 2-6)<br>80 Expand hospital name, patient name, location<br>160 Diagnosis codes described (8 x 20)<br>160 Procedure codes described (8 x 20) |
| 1983 Grocery purchases (32) based on field sizes:<br>18 subtotal, tax, total<br>14 date and time | 1996 Grocery purchases (1272) based on field sizes and average number of items expanded to include textual notation:<br>240 amount (6)<br>x number of items (40)<br>18 subtotal, tax, total<br>14 date and time<br>120 Name, address<br>80 Payment information<br>800 Item description (20)<br>x number of items (40) |

**Figure 2 Estimations of the sizes of some person-specific data collections (per encounter)**

## 3 Growth in birth certificate information

As shown in Figure 3, birth certificate information historically had only 7 to 15 fields of information but today in Illinois (as well as in most states) more than 100 fields of information are collected about each child's birth even though the parents may only receive the traditional few fields.

| **Field name** |
|---|
| Child's first name |
| Child's middle name (sometimes or initial) |
| Child's last name |
| Day, month and year of birth |
| City and/or County of birth (sometimes hospital) |
| Father's name |
| Mother's name (including maiden name) |

**Figure 3 Minimal Set of Birth Certificate Fields**

Figure 3 contains the minimal list of fields available on a birth certificate from almost any state or county in the United States, post 1906 [1].

| **Field name** |
|---|
| Child's first name |
| Child's middle name (sometimes or initial) |
| Child's last name |
| Day, month and year of birth |
| City and/or County of birth (sometimes hospital) |
| Father's name |
| Mother's name (including maiden name) |
| Place of birth (address and town/city) |
| Mother's age and address |
| Mother's birthplace (town/city, state, county) |
| Mother's occupation |
| Mother, number of previous children |
| Father's age and address |
| Father's birthplace (town/city, state, county) |
| Father's occupation |

**Figure 4 Typical Set of Birth Certificate Fields, post 1925**

Figure 4 contains the typical list of fields available on a birth certificate from most states or counties in the United States, post 1925. This list was provided from the Commonwealth of Massachusetts [2].

| Field# | Size | Field name | Field# | Size | Field name |
|---|---|---|---|---|---|
| 1 | 1 | File Status | 31 | 3 | Mother's State of Birth |
| 2 | 50 | Baby's First Name | 32 | 7 | Mother's Residence Address |
| 3 | 50 | Baby's Middle Name | 33 | 2 | Mother's Residence Direction |
| 4 | 50 | Baby's Last Name | 34 | 20 | Residence Street Address |
| 5 | 1 | Baby's Suffix Code | 35 | 10 | Residence Type |
| 6 | 3 | Baby's Suffix Text | 36 | 2 | Residence Extension |
| 7 | 8 | Baby's Date of Birth | 37 | 10 | Residence Apartment # |
| 8 | 5 | Baby's Time of Birth | 38 | 20 | Mother's Town of Residence |
| 9 | 1 | AM/PM Indicator | 39 | 1 | Mother's Residence in City Limits |
| 10 | 1 | Baby's Sex | 40 | 14 | Mother's County of Residence |
| 11 | 3 | Blood Type | 41 | 3 | Mother's State of Residence |
| 12 | 1 | Born Here? | 42 | 10 | Mother's Residence Zip Code |
| 13 | 40 | Place of Birth | 43 | 38 | Mother's Mailing Address |
| 14 | 1 | Facility Type | 44 | 19 | Mother's Mailing City |
| 15 | 20 | City of Birth | 45 | 2 | Mother's Mailing State |
| 16 | 20 | County of Birth | 46 | 10 | Mother's Mailing Zip Code |
| 17 | 6 | Certifier's Code | 47 | 1 | Mother Married? |
| 18 | 30 | Certifier's Name | 48 | 50 | Father's First Name |
| 19 | 1 | Certifier's Title | 49 | 50 | Father's Middle Name |
| 20 | 30 | Attendant's Name | 50 | 50 | Father's Last Name |
| 21 | 1 | Attendant's Title | 51 | 1 | Father's Suffix Code |
| 22 | 23 | Attendant's Address | 52 | 9 | Father's Suffix Text |
| 23 | 19 | Attendant's City | 53 | 9 | Father's Social Security Number |
| 24 | 2 | Attendant's State | 54 | 8 | Father's Date of Birth |
| 25 | 10 | Attendant's Zip Code | 55 | 3 | Father's State of Birth |
| 26 | 50 | Mother's First Name | 56 | 14 | Mother's Origin |
| 27 | 50 | Mother's Middle Name | 57 | 14 | Mother's Race |
| 28 | 50 | Mother's Last Name | 58 | 2 | Mother's Elementary Education |
| 29 | 9 | Mother's Social Security Number | 59 | 2 | Mother's College Education |
| 30 | 8 | Mother's Date of Birth | 60 | 11 | Mother's Occupation |

**Figure 5 Typical Set of Electronic Birth Certificate Fields in 1999 -*starting fields 1-60***

| Field# | Size | Field name |
|---|---|---|
| 61 | 11 | Mother's Industry |
| 62 | 14 | Father's Origin |
| 63 | 14 | Father's Race |
| 64 | 2 | Father's Elementary Education |
| 65 | 2 | Father's College Education |
| 66 | 11 | Father's Occupation |
| 67 | 11 | Father's Industry |
| 68 | 1 | Plurality |
| 69 | 1 | Birth Order |
| 70 | 2 | Live Births Still Living |
| 71 | 2 | Live Births Now Dead |
| 72 | 4 | Month/Year Last Live Birth |
| 73 | 2 | Number of Terminations |
| 74 | 4 | Month/Year Last Termination |
| 75 | 1 | Baby's Weight Unit |
| 76 | 5 | Baby's Weight |
| 77 | 6 | Date of Last Normal Menses |
| 78 | 1 | Month Prenatal Care Began |
| 79 | 2 | Total Number of Visits |
| 80 | 2 | Apgar Score – 1 Minute |
| 81 | 2 | Apgar Score – 5 Minute |
| 82 | 2 | Estimate of Gestation |
| 83 | 6 | Date of Blood Test |
| 84 | 22 | Laboratory |
| 85 | 1 | Mother Transferred In |
| 86 | 30 | Facility Mother Transferred From |
| 87 | 1 | Baby Transferred Out |
| 88 | 30 | Facility Baby Transferred To |
| 89 | 1 | Tobacco Use During Pregnancy |
| 90 | 3 | Number of Cigarettes/Day |
| 91 | 1 | Alcohol Use During Pregnancy |
| 92 | 3 | Number of Drinks/Week |
| 93 | 3 | Mother's Weight Gain |
| 94 | 1 | Release Info For SSN |
| 95 | 6 | Operator Code |
| 96 | 12 | Hospital ID |
| 97 | 1 | Sent to Romans |
| 98 | 1 | Sent to APORS |
| 99 | 16 | Other Certifier Specify |
| 100 | 12 | Temporary Audit Number |
| 101 | 16 | Other Facility Specify |
| 102 | 16 | Other Attendant Specify |
| 103 | 1 | Mother's Race |
| 104 | 1 | Father's Race |
| 105 | 2 | Mother's Origin |
| 106 | 2 | Father's Origin |
| 107 | 1 | Attendant Same YN |
| 108 | 1 | Mailing Address Same YN |
| 109 | 1 | Capture Father's Info YN |
| 110 | 2 | Mother's Age |
| 111 | 2 | Father's Age |
| 112 | 12 | Baby's Hospital Med. Rec. |
| 113 | 1 | High Risk Pregnancy YN |
| 114 | 1 | Care Giver (For Chicago) |
| 115 | 1 | Record Selected For Download |
| 116 | 1 | Downloaded |
| 117 | 1 | Printed |
| 118 | 12 | Form Number |
| | | **MEDICAL RISK FACTORS** |
| 119 | 1 | Anemia |
| 120 | 1 | Cardiac Disease |

**Figure 6 Typical Set of Electronic Birth Certificate Fields in 1999 - *continued fields 61-120***

| Field# | Size | Field name |
|---|---|---|
| 121 | 1 | Acute/Chronic Lung Disease |
| 122 | 1 | Diabetes |
| 123 | 1 | Genital Herpes |
| 124 | 1 | Hydramnios/Oligohydramnios |
| 125 | 1 | Hemoglobinopathy |
| 126 | 1 | Hypertension, Chronic |
| 127 | 1 | Hypertension, Preg. Assoc. |
| 128 | 1 | Eclampsia |
| 129 | 1 | Incompetent Cervix |
| 130 | 1 | Previous Infant 4000+ Grams |
| 131 | 1 | Previous Preterm or SGA Infant |
| 132 | 1 | Renal Disease |
| 133 | 1 | Rh Sensitization |
| 134 | 1 | Uterine Bleeding |
| 135 | 1 | No Medical Risk Factors |
| 136 | 40 | Other Medical Risk Factors |
| | | **OBSTETRIC PROCEDURES** |
| 137 | 1 | Amniocentesis |
| 138 | 1 | Electronic Fetal Monitoring |
| 139 | 1 | Induction of Labor |
| 140 | 1 | Stimulation of Labor |
| 141 | 1 | Tocolysis |
| 142 | 1 | Ultrasound |
| 143 | 1 | No Obstetric Procedures |
| 144 | 40 | Other Obstetric Procedures |
| | | **COMPLICATIONS OF LABOR & DELIVERY** |
| 145 | 1 | Febrile (>100 or 38C) |
| 146 | 1 | Meconium Moderate, Heavy |
| 147 | 1 | Premature Rupture (>12 Hrs) |
| 148 | 1 | Abruptio Placenta |
| 149 | 1 | Placenta Previa |
| 150 | 1 | Other Excessive Bleeding |
| 151 | 1 | Seizures During Labor |
| 152 | 1 | Precipitous Labor (<3 Hrs) |
| 153 | 1 | Prolonged Labor (>20 Hrs) |
| 154 | 1 | Dysfunctional Labor |
| 155 | 1 | Breech/Malpresentation |
| 156 | 1 | Cephalopelvic Disproportion |
| 157 | 1 | Cord Prolapse |
| 158 | 1 | Anesthetic Complications |
| 159 | 1 | Fetal Distress |
| 160 | 1 | No Complications of L&D |
| 161 | 40 | Other Complications of L&D |
| | | **METHOD OF DELIVERY** |
| 162 | 1 | Vaginal |
| 163 | 1 | Vaginal After Previous C-Section |
| 164 | 1 | Primary C-Section |
| 165 | 1 | Repeat C-Section |
| 166 | 1 | Forceps |
| 167 | 1 | Vacuum |
| | | **ABNORMAL CONDITIONS OF NEWBORN** |
| 168 | 1 | Anemia |
| 169 | 1 | Birth Injury |
| 170 | 1 | Fetal Alcohol Syndrome |
| 171 | 1 | Hyaline Membrane Disease/RDS |
| 172 | 1 | Meconium Aspiration Syndrome |
| 173 | 1 | Assisted Ventilation <30 |
| 174 | 1 | Assisted Ventilation >30 |
| 175 | 1 | Seizures |
| 176 | 1 | No Abnormal Conditions of Newborn |
| 177 | 40 | Other Abnormal Condition of Newborn |
| | | **CONGENITAL ANOMALIES OF CHILD** |
| 178 | 1 | Anencephalus |
| 179 | 1 | Spina Bifida/Meningocele |
| 180 | 1 | Hydrocephalus |

**Figure 7 Typical Set of Electronic Birth Certificate Fields in 1999 -*continued fields 121-180***

| Field# | Size | Field name |
|---|---|---|
| 181 | 1 | Microcephalus |
| 182 | 40 | Other CNS Anomalies |
| 183 | 1 | Heart Malformations |
| 184 | 40 | Other Circ./Resp. Anomalies |
| 185 | 1 | Rectal Atresia/Stenosis |
| 186 | 1 | Tracheo-Esophageal Fistula/Esophageal Atresia |
| 187 | 1 | Omphalocele/Gastroschisis |
| 188 | 40 | Other Gastrointestinal Ano. |
| 189 | 1 | Malformed Genitalia |
| 190 | 1 | Renal Agenesis |
| 191 | 40 | Other Urogenital Anomalies |
| 192 | 1 | Cleft Lip/Palate |
| 193 | 1 | Polydactyly/Syndactyly/Adactyly |
| 194 | 1 | Club Foot |
| 195 | 1 | Diaphragmatic Hernia |
| 196 | 40 | Other Musculoskeletal/Integumental Anomalies |
| 197 | 1 | Down's Syndrome |
| 198 | 40 | Other Chromosomal Anomalies |
| 199 | 1 | No Congenital Anomalies |
| 200 | 40 | Other Congenital Anomalies |
| | | **CODE STRIP** |
| 201 | 1 | Record Complete YN |
| 202 | 1 | Record Type |
| 203 | 4 | Facility ID |
| 204 | 4 | City of Birth |
| 205 | 3 | County of Birth |
| 206 | 2 | Mother's State of Birth |
| 207 | 2 | Mother's State of Residence |
| 208 | 4 | Mother's Town of Residence |
| 209 | 3 | Mother's County of Residence |
| 210 | 2 | Father's State of Birth |

| Field# | Size | Field name |
|---|---|---|
| 211 | 14 | Certifier's License Number |
| 212 | 6 | Laboratory ID Number |
| 213 | 4 | Mother Xfer Code |
| 214 | 3 | Mother Xfer County Code |
| 215 | 4 | Baby Xfer Code |
| 216 | 3 | Baby Xfer County Code |
| 217 | 4 | Year of Birth |
| 218 | 7 | Certificate # |
| 219 | 1 | Unique Code |
| 220 | 8 | File Date |
| 221 | 2 | Community Area |
| 222 | 4 | Census Tract |
| 223 | 2 | Century of Last Live Birth |
| 224 | 2 | Century of Last Termination |
| 225 | 2 | Century of Last Menses |
| 226 | 2 | Century of Blood Test |

**Figure 8 Typical Set of Electronic Birth Certificate Fields in 1999 -*continued fields 181-226***

## 3.1 Electronic birth certificate systems

Figure 5, Figure 6, Figure 7 and Figure 8 show the dramatic increase in fields being collected today in reporting a live birth. This particular schema was from the State of Illinois, but is typical of most states even though exact specifications differ from state to state. In the State of Washington, as well as in most states, the filing of live births and fetal deaths are completed by the hospital or birth attendant and then forwarded to the state department of public health [3]. Electronic birth certificate systems, which are computer programs that facilitate the filing, have enabled this dramatic increase in data collection. They began around 1984 and are currently used by more than 32 state departments of public health and 2000 hospital and birthing centers and account for more than 50% of all births in the United States [4]. The information is not entered by typing the values in from scratch. Instead, many entries are entered by selecting a value from a limited list of options and other values may be directly transferred from the hospital's information system; thereby, reducing data entry time and typing mistakes. An electronic birth certificate system usually creates a "hard copy" birth certificate as well as a computer file of live births. The hard copy provides an original birth certificate. Until recently, the computer file was usually sent to the state's central repository by diskette or by direct modem transmission. In more recent years, however, the World Wide Web has been used.

The additional fields of information included in the reporting of live births have contributed to new reports on birth characteristics, infant and maternal mortality, birth weight and gestational age, and adequacy of prenatal care [5]. Without discounting the usefulness of such information, it is important to note that in terms of birth certificate information, the growth in the volume of information being collected has been tremendous: from about 7 fields of information in 1906, to around 15 fields from 1925 to 1980, to more than 200 fields of information by 1999. Clearly, technology has provided the means to make these collections practical.

## 4 Growth in health care cost data

The Illinois Health Care Cost Containment Council (IHCCCC) did not collect health care cost data in 1983, but today a record of each person's hospital visit is recorded. IHCCCC reports more than 97% compliance by Illinois hospitals in providing the information [6]. Figure 9 contains a sample of the kinds of fields of information that are not only collected, but also disseminated. The fields shown in Figure 9 are provided to researchers needing detailed patient specific data to further knowledge; the data are considered useful for measuring access, quality and outcomes.

| # | Field description | Size |
|---|---|---|
| 1 | HOSPITAL ID NUMBER | 12 |
| 2 | PATIENT DATE OF BIRTH (MMDDYYYY) | 8 |
| 3 | SEX | 1 |
| 4 | ADMIT DATE (MMDYYYY) | 8 |
| 5 | DISCHARGE DATE (MMDDYYYY) | 8 |
| 6 | ADMIT SOURCE | 1 |
| 7 | ADMIT TYPE | 1 |
| 8 | LENGTH OF STAY (DAYS) | 4 |
| 9 | PATIENT STATUS | 2 |
| 10 | PRINCIPAL DIAGNOSIS CODE | 6 |
| 11 | SECONDARY DIAGNOSIS CODE - 1 | 6 |
| 12 | SECONDARY DIAGNOSIS CODE - 2 | 6 |
| 13 | SECONDARY DIAGNOSIS CODE - 3 | 6 |
| 14 | SECONDARY DIAGNOSIS CODE - 4 | 6 |
| 15 | SECONDARY DIAGNOSIS CODE - 5 | 6 |
| 16 | SECONDARY DIAGNOSIS CODE - 6 | 6 |
| 17 | SECONDARY DIAGNOSIS CODE - 7 | 6 |
| 18 | SECONDARY DIAGNOSIS CODE - 8 | 6 |
| 19 | PRINCIPAL PROCEDURE CODE | 7 |
| 20 | SECONDARY PROCEDURE CODE - 1 | 7 |
| 21 | SECONDARY PROCEDURE CODE - 2 | 7 |
| 22 | SECONDARY PROCEDURE CODE - 3 | 7 |
| 23 | SECONDARY PROCEDURE CODE - 4 | 7 |
| 24 | SECONDARY PROCEDURE CODE - 5 | 7 |
| 25 | DRG CODE | 3 |
| 26 | MDC CODE | 2 |
| 27 | TOTAL CHARGES | 9 |
| 28 | ROOM AND BOARD CHARGES | 9 |
| 29 | ANCILLARY CHARGES | 9 |
| 30 | ANESTHESIOLOGY CHARGES | 9 |
| 31 | PHARMACY CHARGES | 9 |
| 32 | RADIOLOGY CHARGES | 9 |
| 33 | CLINICAL LAB CHARGES | 9 |
| 34 | LABOR-DELIVERY CHARGES | 9 |
| 35 | OPERATING ROOM CHARGES | 9 |
| 36 | ONCOLOGY CHARGES | 9 |
| 37 | OTHER CHARGES | 9 |
| 38 | NEWBORN INDICATOR | 1 |
| 39 | PAYER ID 1 | 9 |
| 40 | TYPE CODE 1 | 1 |
| 41 | PAYER ID 2 | 9 |
| 42 | TYPE CODE 2 | 1 |
| 43 | PAYER ID 3 | 9 |
| 44 | TYPE CODE 3 | 1 |
| 45 | PATIENT ZIP CODE | 5 |
| 46 | Patient Origin COUNTY | 3 |
| 47 | Patient Origin PLANNING AREA | 3 |
| 48 | Patient Origin HSA | 2 |
| 49 | PATIENT CONTROL NUMBER | |
| 50 | HOSPITAL HSA | 2 |

**Figure 9 IHCCCC Research Health Data**

The National Association of Health Data Organizations (NAHDO) reported that 37 of the 50 states (or 74%) had legislative mandates to gather hospital-level data [7] similar to that described in Figure 9. In addition, by 1996, 17 states reported they had started collecting ambulatory care (outpatient)

data from hospitals, physician offices, clinics, and so on. These data collections began in an effort to help reduce healthcare costs. Many states have subsequently distributed copies to researchers, sold copies to industry and made versions publicly available. While there are many possible sources of patient-specific data, these represent a class of data collections that were not available prior to 1983.

## 5 Growth in supermarket transaction data

Private sector information about individuals has expanded also. For example, supermarket transactions consisted only of summary price information in 1983 and were not identified to individuals. In many supermarkets today in Illinois, the complete list of purchased items is often stored along with the identity of the consumer. This increase in the volume of data collected about individuals from supermarket purchases is the topic of discussion in this subsection.

As shown in Figure 2, a consumer in 1983 could purchase items from a supermarket and the only recorded evidence left behind were roughly an inventory debit and a record of the total amount purchased and the amount of tax paid. There was no knowledge necessarily of the identity of the consumer or of the consumer's personal habits and behaviors in terms of goods typically purchased and the times and days of the consumer's shopping experiences. Analyses of consumer behaviors were based on aggregated sales, and so it was nearly impossible to identify multiple purchases from the same customer. Inferred patterns drawn from the data therefore included uncertainty.

Today's computer technology has changed all this and makes it possible to now dramatically reduce the uncertainty. Nowadays consumer transactions can be stored and analyzed, and by doing so, information about each consumer's lifestyle, behavior, beliefs and habits can usually be revealed. After all, watching an individual consumer's purchases week after week provides clues about demands on the consumer's time, economic status and life experiences. In the remainder of this subsection, I will discuss the evolution of this transformation.

### 5.1 Catalina Marketing

In March 1996, Catalina Marketing, Inc. (Catalina) began a data collection enterprise that currently stores the shopping patterns of an estimated 143 million shoppers each week from more than 11,000 supermarkets nationwide [8]. By July 1998, the 2-terabyte database had reportedly 18 billion rows of data. Catalina's objective was to use technology to improve its ability to measure consumer behavior.

Retailers obtained chain-specific information from Catalina over the World Wide Web. After they entered a username and password, retailers queried the data from a browser and the results were displayed in HTML or e-mailed back in a comma-delimited format for easy use in spreadsheet programs. Such reports were useful in developing time of day specials, designing direct mail campaigns and assessing traffic flow through checkout lanes. [9]. However, in these early years, results were typically compiled from transactions in which the consumer was not only anonymous but also independent across multiple transactions. That is, multiple purchases attributable to the same consumer were not recognizable as such in the data.

## 5.2 Supermarket loyalty cards

As Catalina Marketing began its data collection, awareness arose among retailers that with the help of personally identifying cards and state-of-the-art database technology, retailers, such as supermarkets, could analyze millions of transactions quickly to identify their best customers and build loyalty through special rewards such as discounted prices. These are called *loyalty programs* and the accompanying card is termed a *loyalty card*.

Here's how a loyalty program usually works. A consumer applies for a loyalty card; Figure 10 provides examples of the kinds of fields of information requested on applications for loyalty cards from some major supermarket chains stores in the Washington, DC, California and Massachusetts areas in 1998. All consumers are normally eligible (with no restrictions on teenagers, for example, who may purchase condoms and other sensitive items) and acceptance is almost always guaranteed by merely completing the application. The information provided on the application is not customarily reviewed for accuracy. There is expected to be only one loyalty card per household in some programs, but other programs seek to have each person within each household have his/her own card. Upon reaching the check-out counter, the cashier asks each consumer for a loyalty card. If the consumer has a loyalty card, the card is scanned or the identifying number found on the card manually entered into the computerized register. Purchased items are then scanned as normal, deducting savings automatically. In most loyalty card programs, the final receipt includes an itemized list of the savings that resulted from using the card. Consumers that have no loyalty card are charged the higher, non-discounted prices.

| Field name | Food Lion | Fresh Fields | Safeway | Star Market |
|---|---|---|---|---|
| Name | yes | yes | yes | yes |
| Home street address | yes | yes | yes | yes |
| Homy city | yes | yes | yes | yes |
| Home state | yes | yes | yes | yes |
| Home ZIP | yes | yes | yes | yes |
| Home phone number | yes | yes | yes | yes |
| Social Security Number | | | | yes |
| | | | | |
| **Additional data sometimes requested** | | | | |
| Birth date | | | yes | yes |
| ZIP code of work place | | yes | | |
| Other stores where you shop | yes | yes | | |
| Number of people in household | yes | yes | | |
| Age each person in household | yes | yes | | |
| How much do you spend each week | yes | yes | | |
| | | | | |
| **Additional data for accepting checks** | | | | |
| Bank | | | yes | yes |
| Bank account number | | | yes | yes |

**Figure 10 Typical set of fields on loyalty card applications**

Here are examples of price differences based on the use of a loyalty card in the Georgetown Safeway in Washington, DC on or about December, 1998 [10]. A bag of bagels were $1.59 with no loyalty card and $.99 with a loyalty card. A frozen gourmet dinner was $4.19 with no loyalty card and $2.99 with a loyalty card. While these savings are dramatic, overall reported savings tend to be 20-40%.

On the other hand, Idaho-based Albertson's, who operates in 37 states, did not have a loyalty program in 1997. In a press release, company officials implied that discounted prices based on loyalty cards at other supermarkets may be comparable to their everyday-low-prices or promotional prices that didn't require a card. If this is so, the comparative result would imply an inflated price (or penalty) to consumers who do not use a loyalty card in supermarkets that offer them rather than a discount to consumers who do.

Ideally, with a loyalty program, retailers could become better at serving individual consumers because discounts and product selections could become specific to each consumer. But this new approach tugs at two historical traditions. The first is the issue of privacy with respect to collecting so much consumer-specific data; and, the second concerns moving away from charging the same price for the same product, to charging a price based on the consumer's perceived value to the retailer. In this subsection however, I limit discussion to the data collection itself.

When supermarket chains combine the kind of detailed transactions data kept by Catalina Marketing with the demographic information gathered from their own loyalty-card programs, the idea is that highly targeted marketing campaigns can be created and better relationships cultivated with

customers. In fact, targeted, scanner-based campaigns seem to have a redemption rate double that of direct mail [11].

As of December 1998, Food Marketing Institute reported that 6 of 10 supermarket companies electronically collected customer transaction data or planned to do so soon, as opposed to 3 out of 10 such supermarket companies in 1993 [12]. If correct, more than half of the supermarkets in the United States could soon require consumers to be cardholders in order to receive discounted prices.

## 6 Opting out

So far in this chapter, I have provided examples of data collection activities that have expanded as the supporting technology became readily available. Of course, the particular data collections mentioned are representative of many other data collections currently underway, as I will discuss in the next subsection. But before I move on, let me say that most privacy discussions consider whether the individual has the ability to decide not to participate (or to "*opt out"*) of the data collection. The collections of birth and health data that were described in this subsection provide no such possibility, being dictated by regulation and law and intended to capture the entire population.

On the other hand, a consumer can decide to pay a higher price and not use a loyalty card at the supermarket; or alternatively, a consumer can provide false information when applying for such a card. But these actions are not wholly satisfactory. The first option requires consumers, who want to opt out, to pay (by means of higher prices) for privacy that they historically enjoyed for free; and, the later option encourages deliberate deceptive action by innocent, irreproachable consumers. The individual seems to risk unknown future uses of their purchasing patterns that could become personally damaging. Yet, the primary beneficiaries for loyalty programs appear to be manufacturers and retailers who can better avoid waste in promoting products. In other works, I propose that newer technology for privacy protection integrated with effective policy can offer better solutions in these kinds of situations so that opting out, if available at all, is not the only possible choice for individuals who want privacy protection.

## 7 Behaviors in data collecting today

The examples provided in the previous subsections exemplify recent behavioral tendencies in the collection of person-specific data. These informally observed "trends" are enumerated below and further discussed in this chapter.

**Behavior 1.** Given an existing person-specific data collection, expand the number of fields being collected. I casually refer to this as the "*collect more*" trend.

**Behavior 2**. Replace an existing aggregate data collection with a person-specific one. I casually refer to this as the "*collect specifically*" trend.

**Behavior 3.** Given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection related to the question, problem or opportunity. I causally refer to this as the "*collect it if you can*" trend.

No matter how you look at it, all three tendencies result in more and more information being collected on individuals. Below is a discussion of each behavior.

## 7.1 Behavior 1. Collect more

The earlier presentation on the increase in the number of fields collected as part of the birth certificate is an example of this behavior. In fact, I found increases in many other older established person-specific collections. Figure 11 contains an overview in which 13 of 21 (or 62%) person-specific data collections that have historically collected person-specific data expanded the number of fields being collected from 1983 to 1996.

| Old Collections | 1983 | 1996 |
|---|---|---|
| bank account | • | • |
| birth certificate | • | 💻 |
| census survey | • | 💻 |
| credit card | • | 💻 |
| credit history | • | 💻 |
| driver license | • | 💻 |
| legal actions | • | 💻 |
| medical record | • | 💻 |
| marriage license | • | 💻 |
| military service | • | • |
| motor vehicle registration | • | • |
| phone calls | • | • |
| professional license | • | 💻 |
| property (& tax) records | • | • |
| public assistance | • | 💻 |
| real estate | • | • |
| recreational license | • | 💻 |
| selective service | • | • |
| tax filings | • | 💻 |
| voting list | • | • |
| worker's compensation | • | 💻 |
| | | |
| Percentage that increased | | 62% |

**Figure 11 Expansion in some historic person-specific collections**

## 7.2 Behavior 2. Collect specifically

In places where tabular statistics were once the form of reporting or sampling the method of collection, person-specific data collection is becoming the new standard. The earlier discussion on

supermarket transactions illustrates how decisions that once relied on aggregate statistical data now use highly detailed, person-specific data.

Another example is student-specific educational data. By 1996 in some states, recordings of student-specific data for kindergarten through 12th grade began. These collections typically include days absent, number of school lunches consumed, immunizations, allergies, and so forth for each student. In 1983, this information was provided in aggregate student-body summaries at the school level. Under the new collection practice, this information is specific to the student and shared in that form. These examples demonstrate the growing number of new entity-specific data collections.

## 7.3 Behavior 3. Collect it if you can

The earlier discussion on healthcare cost data collected by states was representative of new person-specific data collections that recently begun or are being initiated, but that did not exist in 1983. There are many other examples as well. Here are a few, including Immunization Registries and The National Directory of New Hires.

### 7.3.1 National Directory of New Hires

In a report prepared for the Chairman of the Subcommittee on Federal Services in 1994, nonpayment of child support was shown to contribute to childhood poverty as well as to increases in the number of families receiving welfare [13]. This report stated that in 1994, more than one-fifth of America's children lived in poverty, and it recited an estimate that half of those would live in single-parent families at some point in their lives. To help obtain the financial support that parents owe their children and to reduce welfare costs, Congress passed the Personal Responsibility and Work Opportunity Reconciliation Act of 1996. This mandated the establishment of new resources at the Federal and State levels to assist state child support enforcement agencies and included provisions for the establishment of a National Directory of New Hires as well as State Directories of New Hires.

The goal of these newly created worker-specific data collections, at their inception in 1996, was and remains to monitor all individuals with jobs in order to better track down parents who owe child support. That is, these collections do not merely contain information about Americans found to be delinquent parents, but include information on almost all working Americans, the vast majority of whom have been accused of nothing.

Employers must file timely reports on every person they hire and, quarterly, the wages of every worker [14]. Figure 12 contains a list of the fields employers are required to report. In addition, states must regularly report all people seeking unemployment benefits and all child-support cases, even if the parents and children involved do not receive public assistance or ask for help in collecting support.

| Field name | Reported when newly hired | Updated quarterly on all employees |
|---|---|---|
| Employee name | yes | yes |
| Employee SSN | yes | yes |
| Employee address: street | yes | |
| Employee address: city | yes | |
| Employee address: state | yes | |
| Employee address: ZIP | yes | |
| Employer name | yes | yes |
| Employer address: street | yes | yes |
| Employer address: city | yes | yes |
| Employer address: state | yes | yes |
| Employer address: ZIP | yes | yes |
| Federal employer identification number (FEIN) | yes | yes |
| Employee wage amount | yes | yes |
| Reporting period | yes | yes |
| | | |
| **Additional Fields States Can Require Be Reported** | | |
| Employee date of birth | may be required | |
| Employee date of hire | may be required | |
| Employee state of hire | may be required | |

**Figure 12 Fields collected in National Directory of New Hires**

As of 1997, more than 7.4 million delinquents owed more than $43 billion in past child support. These registries are credited for increasing payments from $12 billion to $14.4 billion in 1998 and helped locate more than 1.2 million delinquents [15]. As of June 1999, the registry had information on almost 12 million families involved in child-support cases.

### 7.3.2 Immunization Registries

Children can be immunized against most serious childhood diseases with little or no cost to parents. The problem is that many of these diseases appear so rarely in society nowadays that parents conclude immunizations are not necessary and are concerned about possible side-effects and the increased number of immunizations required. Absence of immunizations leads to vulnerabilities in society of disease outbreaks which can present serious health risks.

For example, a measles epidemic broke out among Rutgers University students in the spring of 1994. Measles is a deadly viral disease, highly contagious and airborne. Students entering a university should have had two inoculations, but many have had only one [16]. According to the Centers for Disease Control and Prevention (CDC), 18,000 cases were reported in 1989 and 27,000 in 1990.

In order to encourage parents to have their children immunized at a proper age, state and national registries were created as a solution to this problem. The objective is to maintain a record of the immunization history of each child so that when a child appears at a clinic or physician office, the registry can be consulted to ensure proper immunizations are administered. Figure 13 has a copy of fields

legislated to be collected in the State of Texas and serves as an example of the kinds of fields maintained in immunization registries.

| Field name |
|---|
| |
| **CHILD INFORMATION** |
| Child's name (first, middle, last) |
| Child's address: street |
| Child's address: city |
| Child's address: state |
| Child's address: ZIP |
| Child's Social Security Number (if available) |
| Child's gender |
| Child's date of birth |
| Mother's maiden name |
| |
| **HEALTHCARE PROVIDER'S INFORMATION** |
| Health care provider's name (first, middle, last) |
| Health care provider's business address: street |
| Health care provider's business address: city |
| Health care provider's business address: state |
| Health care provider's telephone |
| Health care provider's business address: ZIP |
| |
| **VACCINE INFORMATION** |
| Date vaccine was administered |
| Vaccine lot number (if known) |
| Dose or series number (if known) |
| Name of vaccine manufacturer (if known) |

**Figure 13 Fields collected in Immunization Registry in the State of Texas**

Child-specific immunization registries began around 1997 and as of March 2000, 21 of 50 states (or 42%) have a law authorizing the creation of an immunization registry. The specifics vary from state to state. But records from the electronic birth certificate database often seed immunization registries with new records. Copies of the information are forwarded from the state to the national database maintained by CDC and in some cases such as in Texas, copies are made available to the local public health department, the child's physician, the school in which the child is enrolled, and the childcare facility in which the child is enrolled [17].

### 7.3.3 Others

Not all new collections result from legislative mandate. There are many in the private sector as well, and of course, technology makes more and more collections possible.

New sources of person-specific collections result from using the World Wide Web. Online companies can track what participating customers look at, where customers go and how long customers linger on a particular page. One of the leaders is Amazon.com, the online bookseller. It tracks the purchases of more than 4.5 million customers and then offers them suggestions about what they might enjoy, based on personal reading patterns correlated by the computer. Another company, DoubleClick, Inc., links together traces of locations where customers and potential customers have been to construct the browsing profiles of customers and potential customers.

New technology keeps emerging that fuels demand for more collections. For example, Visionics, Inc., has a product that given a collection of identified faces, such as those available from driver's license photos or earlier surveillance photos, can automatically locate faces in complex scenes, and then track and identify individuals by matching faces to those stored in the collection. The technology is currently being used at casinos in Las Vegas where gaming investigators have used the system to identify in real-time, known casino cheaters, card counters and their associates. Similar technology is deployed on streets in England.

Some new trial uses of global positioning systems installed in vehicles are being piloted in which each vehicle reports to a central source location its geographic coordinates with date and time. That is, the vehicle reports where it's been, and when and how long it was there. A version of the system is being piloted by an insurance company in Texas to help set car insurance premiums based on actual travel patterns.

There is no doubt that a tremendous amount of information is already being collected on individuals, collections are expanding, and new ones being created at an alarming rate. Technology is the catalyst, so the behaviors are expected to continue.

## 8 Disk storage per person

In an attempt to characterize the growth in person-specific information, I introduce a new metric termed *disk storage per person (DSP)*, which is measured in megabytes/person (where megabytes is $10^6$ bytes and is abbreviated MB). *Global DSP* (*GDSP*) is the total rigid disk drive space (in MB) of new units sold in a year divided by the world population in that year. I do not consider removable and other storage mediums. I do not include used and refurbished drives if they were first purchased in previous years even though they may be operational during the evaluation year. And I recognize disk storage is used for more than person-specific data. So the GDSP metric is a crude measure of how much disk storage could possibly be used to collect person-specific data on the world population. This value appears to be dramatically increasing with time.

In 1983, one million fixed drives with an estimated total of 90 terabytes ($10^{12}$ bytes written TB) were sold worldwide; 30MB drives had the largest market segment [18]. In 1996, 105 million drives, totaling 160,623 terabytes were sold with 1 and 2 gigabyte ($10^9$ bytes written GB) drives leading the industry [19]. By the year 2000, with 20GB drive leading the industry, rigid drives sold for the year are projected to total 2,829,288 terabytes [20]. A summary of these storage values appears in the top row of the chart in Figure 14.

The world population in 1985 was 4.5 billion ($10^9$), in 1996 was roughly 5.8 billion and is expected to be 6 billion by the year 2000 [21]. Therefore, in 1985, there were 20,000 bytes per person ($GDSP_{1985} = 0.02$); in 1996, there was about 28 MB/person ($GDSP_{1996} = 28$); and, by the year 2000, there may be 472 MB/person ($GDSP_{2000} = 472$). These values are summarized in Figure 14.

| | 1983 | 1996 | 2000 |
|---|---|---|---|
| **Storage (TB)** | 90 | 160,623 | 2,829,288 |
| **Population(mil)** | 4,500 | 5,767 | 6,000 |
| **GDSP (MB/person)** | 0.02 | 28 | 472 |
| **Person-time/page** | 2 mon | 1 hr | 3.5 min |

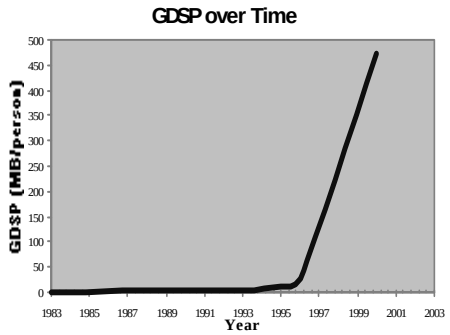


**Figure 14 Characterizing computer storage available for recording person-specific information**

GDSP values signal the amount of storage possibly available to record all the events for each person throughout the year. Here's an analogy to better understand the storage space implied by these values. In 1985, $GDSP_{1985} = 0.02$ is similar to reserving a small file that could reside on a diskette as the amount of storage available to hold the union of all information collected on a single person in 1985. In 1996, which is the knee of the curve in Figure 14, $GDSP_{1996} = 28$ makes roughly a 30MB hard drive (or 20 diskettes) available to store information about each person during that year. By 2000 with $GDSP_{2000} = 472$, most of a CD (or 338 diskettes) could be used to record the events of each person in that year. Clearly, the amount of storage space possibly available to store information on each person is growing rapidly.

I attempt to estimate how much of an adult's life could be documented on a single piece of letter-size paper (8.5in x 11in) and then stored in computers as the technology progressed from 1983 to the year 2000. Assume a printed page of text contains 54 lines by 60 characters; this can be stored in 3,240 bytes with no compression. I can use these GDSP figures to compute the amount of a person's time that can be documented on such a page. In 1983 a page could be used to document two months of a person's life. Actual recordings in 1983 did include itemized long distance phone calls, credit card purchases, the volume of electricity used, and so forth. In 1996, a page could be used to document each hour of a person's life. Recordings in 1996 did expand in both size and number. Examples of new collections included items purchased at the grocery store, web sites visited, and the date and time in some locations a

car proceeded through a tollbooth. By the year 2000, with 20 gigabyte drives leading the industry, it is projected that a page could be used to document every 3.5 minutes of a person's life. Collections are expanding to include visual data, surveillance information, genetic and biometric information such as, heart rate, pulse and temperature. These values are included in the last row of the chart on Figure 14. So, GDSP provides a way of characterizing how much space could be used to record daily events for each person in the world population using only new rigid disk drives sold that year.

## 9 Discussion

In summary, there is no doubt that society is moving towards an environment in which society could have almost all the data on all the people. As a result, it is becoming increasingly difficult to produce anonymous and declassified information in today's globally networked society. Most data holders do not even realize the jeopardy at which they place financial, medical, or national security information when they erroneously rely on security practices of the past. Technology has eroded previous protections leaving the information vulnerable. In the past, a person seeking to reconstruct private information was limited to visiting disparate file rooms and engaging in labor-intensive review of printed material in geographically distributed locations. Today, one can access voluminous worldwide public information using a standard handheld computer and ubiquitous network resources. Thus from seemingly innocuous anonymous data, and available public and semi-public information, one can often draw damaging inferences about sensitive information. However, one cannot seriously propose that all information with any links to sensitive information be suppressed. Society has developed an insatiable appetite for all kinds of detailed information for many worthy purposes, but unfortunately current practices tend to distribute information widely.

### 9.1 Past practices may no longer be applicable

In 1997, the New England Journal of Medicine published an article by Melton [22] in which he described an environment at the Mayo Clinic that had enjoyed a long tradition of sharing patient records with researchers in an open manner with little or no privacy problems. Among other things, he questiond why established and old data sharing practices that seem to have proven themselves to work sufficiently in the past were no longer considered acceptable. An answer is that until recently there existed natural limits that protected patient privacy which technology now erodes at an alarming rate. It was not our old practices that protected our privacy. Instead, it was our old practices in the absence of current technology that provided the protection.

For example, in an earlier time, if I wanted to receive research information from Mayo Clinic's records, I would have to take time off from work, take a plane to Rochester, Minnesota, and then have access to their files only during the times in which their records room was open. I could only leave with that information I could write down during that time (assuming the absence of copiers). The physical labor involved in manually reviewing records as well as the physical restrictions on entering the records room provided economic boundaries that restricted the dissemination of person-specific data. Now consider what is involved today if all of the records of the Mayo Clinic were available electronically. I could access all of their information from anywhere in the world using a standard handheld computer and ubiquitous network resources. I could have an exact copy in a matter of seconds and could further

distribute it widely to others, around the world, in a matter of minutes. Today's technology does pose unparalleled threats to patient privacy.

Today's technology also makes access to the information easier within Mayo Clinic itself. As a result, more data tend to be shared internally than ever before.

## 9.2 Data sharing and risk

Many details about our lives are documented on computers and when this information is linked together, the resulting profiles can be as identifying as fingerprints even when the information contains no explicit identifiers such as name and address. The increase in the availability of detailed data as well as inexpensive technology to process it, is having a dramatic impact on research. Having more information available will probably lead to even more and more studies because additional data can often help ensure the validity and generalizability of study results. These will likely result in continued increase in data collected and shared. Therefore the time is right to seriously examine data collection and sharing practices. Most person-specific data are autonomously controlled and much of the information is replicated across collections. So, coherent and comprehensive approaches are needed. The time to make policy changes is now in order to prevent data holders and governments from succumbing to the financial incentives that encourage sales of data.

## 9.3 Risk and liability

Citizens in the United States are largely unaware of the loss of privacy and the resulting ramifications that stem from having so much person-specific information available. Clearly a loss of dignity and financial income can result when personal medical information is widely and publicly distributed. Yet, data holders make data sharing decisions to benefit themselves and minimize their own risk. Doing so, does not always provide desirable protections for the persons whose information is contained within the data.

The idea of “risk” concerns the likelihood of experiencing loss or damage. As a liability in the context of this work, “risk” refers to an obligation the data holder has to the subjects whose information is contained within the data and to society. So, both the data holder and the subjects of the data want no harm to result from the sharing of data, but from the data holder’s perspective such harm appears as legal liability. As a result, actions the data holder may take to “protect” data are not the same as actions that would “protect” the identities of the subjects. Instead, such actions are aimed at limiting the data holder’s liability regardless of their inefficiency in protecting subjects. Examples of such self-serving actions include, but are not limited to: (1) making it difficult to identify the data holder as the source of shared information; (2) making it difficult for society to know what is collected and to whom copies are given; and, (3) making it legally difficult for a recipient of the data to publicly admit to being able to identify subjects in data that the data holder asserts are anonymous. These kinds of actions help protect the data holder, but do not protect the identities of the subjects. On the other hand, protections for subjects are limited almost entirely to the protections that can be made available through policies, regulations and laws. Therefore, it is essential that measurements of risk and characterizations of access policies be based on society’s perspective.

## 9.4 Tension between privacy and secondary uses of data

Below is an empirically proven claim.

Informal claim . Data collected for one reason tends to get used for another.

This claim gives rise to additional concerns over privacy, because decisions that led to the inclusion of information in the primary data collection typically did not also consider secondary uses of the data. This happens because the demand from secondary uses typically appears after the data are collected. Even in cases where there appears to have been meaningful discussion of secondary uses beforehand, such as the employment of a consent form, care is not always taken to ensure that the resulting decision was not coerced or made with little or no understanding of the ramifications. In the next few paragraphs, I look at different ways in which society has made decisions about secondary uses of data, and I provide a way to reason about these findings.

### 9.4.1 Quality versus anonymity

There is a natural tension between the quality of data and the techniques that provide anonymity protection. Consider a continuum that characterizes possible data releases. At one end of the continuum are person-specific data that are fully identified. At the other end are anonymous data that are derived from the original person-specific data, but in which no person can be identified. Between these two endpoints is a finite partial ordering of data releases, where each release is derived from the original data but for which privacy protection is less than fully anonymous. See Figure 15.

The first realization is that any attempt to provide some anonymity protection, no matter how minimal, involves modifying the data and thereby distorting its contents. So, as shown in Figure 15, movement along the continuum from the fully-identified data towards the anonymous data adds more privacy protection, but renders the resulting data less useful. That is, there exists some tasks for which the original data could be used, but those tasks are not possible with the released data because the data have been distorted.

So, the original fully identified data and the derived anonymous data are diametrically opposed. The entire continuum describes the domain of possible releases. Framed in this way, a goal of this work is to produce an optimal release of data so that for a given task, the data remain practically useful yet render minimally invasive to privacy.

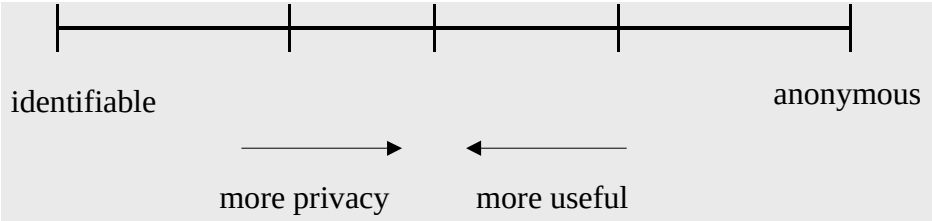


**Figure 15 Optimal releases of data**

### 9.4.2 Tug-of-war between data holders and recipients

The second realization that emerges from Figure 15 is that the usefulness of data is determined by the task to which the recipient puts the data. That is, given a particular task, there exists a point on the continuum in Figure 15 that is as close to anonymous as possible, yet the data remain useful for the task. A release of data associated with that point on the continuum is considered optimal. In the next paragraphs, I provide a skeletal depiction of current practices that determine who gets access to what data. I show that the result can be characterized as a tug-of-war between data holders and data recipients.

In general, the practices of data holders and related policies do not examine tasks in a vacuum. Instead, the combination of task and recipient together are weighed against privacy concerns. This can be modeled as a tug-of-war between the data holder and societal expectations for privacy on one side, and the recipient and the recipient's use for the data on the other. In some cases such as public health legislation, the recipient's need for the data may overshadow privacy protections, allowing the recipient (a public health agent) to get the original, fully-identified health data. See Figure 16 in which a tug-of-war is modeled. The privacy constraints on the data holder versus the recipient's demand for the data are graphically depicted by the sizes of the images shown. In the case illustrated, the recipient receives the original, fully identified data.

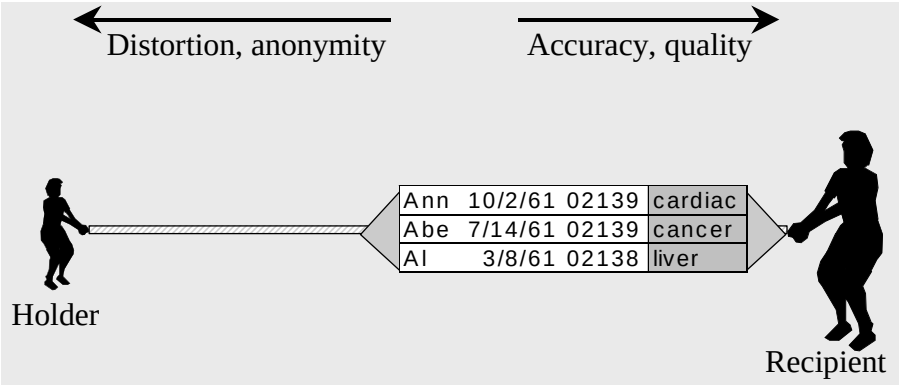


**Figure 16. Recipient's needs overpower privacy concerns**

Figure 17 demonstrates the opposite extreme outcome to that of Figure 16. In Figure 17, the data holder and the need to protect the confidentiality or privacy of the information overshadows the recipient and the recipient's use for the data and so the data is completely suppressed and not released at all. Data collected and associated with national security concerns provides an example. The recipient may be a

news-reporting agent. Over time the data may eventually be declassified and a release that is deemed sufficiently anonymous provided to the press, but the original result is as shown in Figure 17, in which no data is released at all.

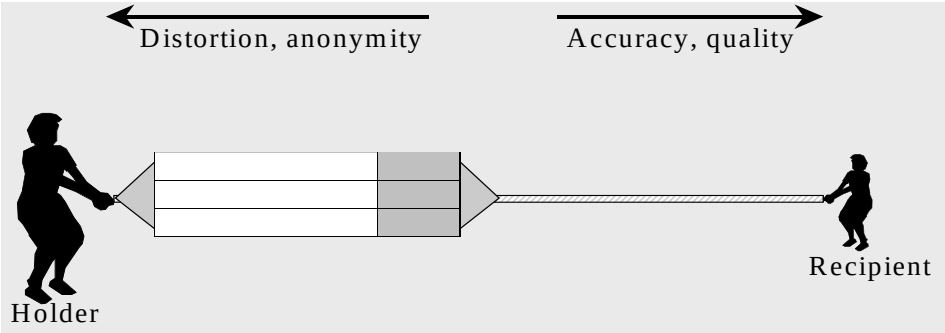


**Figure 17 Data holder and privacy concerns overpower outside uses of the data**

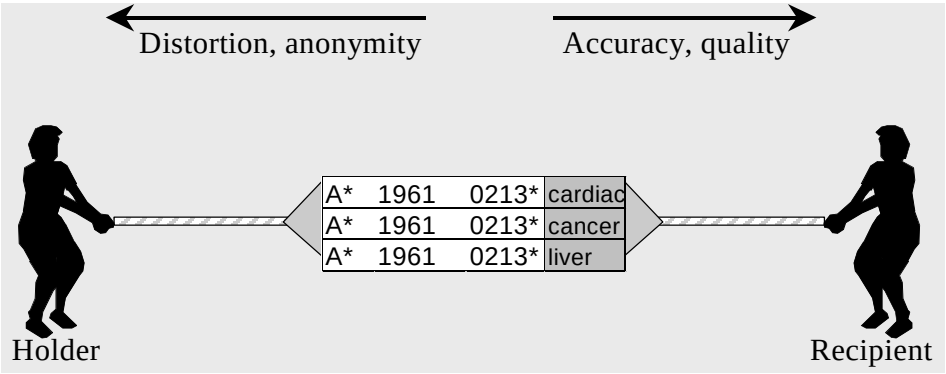


**Figure 18. An optimal balance is needed between privacy concerns and uses of the data**

Figure 16 and Figure 17 depict situations in which society has made explicit decisions based on the needs of society as a whole. But secondary uses of medical data, for example, by marketing firms, pharmaceutical companies, epidemiological researchers and others do not in general lend themselves to such an explicit itemization. Figure 18 demonstrates situations in which the needs for privacy are weighed equally against the demand for the data itself. In such situations, a balance should be found in which the data are rendered sufficiently anonymous, yet remain practically useful. As an example, this situation often occurs with requests by researchers for patient-specific medical records in order for them to undertake clinical outcomes, or administrative research that could possibly provide benefits to society. At present, decisions are primarily based on the recipient receiving the original patient data or no data at all. Attempts to provide something in-between typically results in data with poor anonymity protection or data that is overly distorted. This work seeks to find ways for the recipient to get data that has adequate privacy protection, therefore striking an optimal balance between privacy protection and the data's fitness for a particular task.

At present, data holders often make decisions arbitrarily or by ad hoc means. Figure 19 portrays the situation some state and federal agencies find themselves when they seek to produce public-use files

for general use. Over the past few years, there has been a tremendous effort to make more data that is collected by government agencies available over the World Wide Web. In these situations, protecting the reputation of the agency, and the guarantees for privacy protection for which some agencies are legally bound, outweighs the demands of the recipient. In these cases, a strongly distorted version of the data is often released; the released data are typically produced with little or no consideration to the tasks required. Conversely, some state and federal agencies release poorly protected data. In these cases, the individuals contained in the data can be easily re-identified.

Neither way of releasing data yields optimal results. When strongly distorted data are released, many researchers cannot use the data, or have to seek special permission to get far more sensitive data than what are needed. This unnecessarily increases the volume of sensitive data available outside the agency. On the other hand, data that do not provide adequate anonymity may harm individuals.

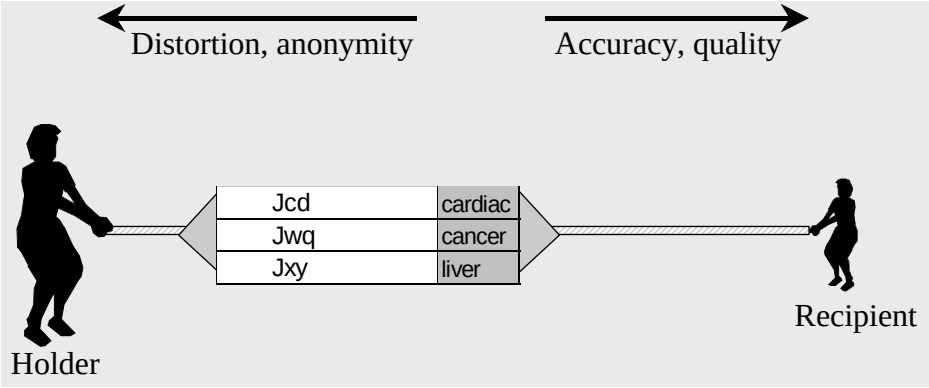


**Figure 19. Data holder and privacy concerns limit uses of the data**

In examining the different struggles between privacy and secondary uses of person-specific data, I make the following claims:

Informal claim . Current policies and practices support crude decisions. A recipient today tends to receive the sensitive data itself, no data at all, overly distorted data that is of little or no use, or poorly protected data in which individuals can be re-identified.

Informal claim . Ultimately, the data holder must be held responsible for enforcing privacy protection because the data holder typically reaps a benefit and controls both data collection and dissemination.

While the claims above are independent of the content of data, the study of secondary uses of medical data in particular provides a natural incentive to find optimal solutions between researchers and data holders. After all, there are no legislative guidelines to empower one party so that it can overwhelm the other as was shown in Figure 16 and Figure 17. Also, state and federal agencies tend to be small in number and highly visible in comparison to the dramatic number of holders of medical data. Because there are so many data holders, it is hard to scrutinize their actions, and the resulting damage to individuals can be devastating yet hard to prove. And there exists strong financial incentives not to provide adequate protection in medical data. On the other hand, research from data may lower health

costs or save lives. These reasons provide motivation for finding optimal releases of data and for integrating technology with policy for maximal benefit.

A Harris-Equifax poll [23] implies that the public would be willing to share information for research provided researchers and others could not identify any person included in the released data. Rendering data sufficiently anonymous would be a way in which data could be more freely shared.

## 10 Future Work

Below are proposed projects of varying degrees of difficulties and skill requirements that extend this work.

1. Track the growth and expansion of common person-specific data collections. An example provided in this chapter was birth certificate data, but similar expansions in collection and sharing exist in other collections. Document the growth as well as the rationale behind the growth and make predictions about future collection and sharing of the information.

2. Historically, many person-specific pieces of data were limited to few collections in isolated locations. An example would be birth certificate information. Today however, there are many possible sources of such information independent of the original collecting organization. An example would be health data. Inferences about the health of individuals can be found in prescription data, a log of visitors to web sites containing information about particular diseases, mailing lists, warranty data, and so forth. Similarly, information about an individual's birth information or criminal record can be inferred from different kinds of data, each having different quality issues. Select a piece of information and document how many different ways such information could be inferred through public and semi-public sources. Take care to comment on the quality of the information provided from different sources.

L. Sweeney, Information Explosion. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, L. Zayatz, P. Doyle, J. Theeuwes and J. Lane (eds), Urban Institute, Washington, DC, 2001.

# References

1 Massachusetts Department of Public Health, Registry of Vital Records & Statistics, *Massachusetts Vital Record Content* (1999) (available at http://www.mass-doc.com/massachusetts_vital_records_content.htm).

2 See note 1 Massachusetts Department of Public Health.

3 Washington State Department of Public Health, Washington State Vital Statistics, *Information Sources for Vital Statistics* (1999) available at http://www.doh.wa.gov/EHSPHL/chs/sub1.htm

4 Genesis Systems, Inc., *History of Genesis Systems and Electronic Birth Certificate* (1999) available at http://www.genesisinfo.com/website-v1/prodebc.html and http://www.genesisinfo.com/website-v1/history.html

5 Massachusetts Department of Public Health, Bureau of Health Statistics, Research and Evaluation, *Advance Data: Births 1997* (1999) available at http://www.state.ma.us/dph/birth97/ab97xsum.htm

6 State of Illinois Health Care Cost Containment Council, *Data release overview.* (Springfield: State of Illinois Health Care Cost Containment Council, March 1998).

7 National Association of Health Data Organizations, *A Guide to State-Level Ambulatory Care Data Collection Activities* (Falls Church: National Association of Health Data Organizations, Oct. 1996).

8 "Get to Know Your Customer -- Catalina Marketing lets retailers find out who's really buying," *Information Week*, 691, July 13 (1998).

9 See note 8 Information Week.

10 "Bargains at a Price: Shopper's Privacy," *Washington Post*, December 31, (1998): A01.

11 See note 8 Information Week.

12 See note 10 Washington Post.

13 General Accounting Office, Income Security Issues, *Child Support Enforcement: Families Could Benefit From Stronger Enforcement Program.* Washington: General Accounting Office (HEHS-95-24) December 1994.

14 United States Department of Health and Human Services, Office of Child Support Enforcement, *National Directory of New Hires: NDNH Guide for Data Submission*. Washington: United States Department of Health and Human Services (HHS-OCSE-1996-1012) August 1997.

15 "Uncle Sam Has All Your Numbers," *Washington Post*, June 27, (1999): A01.

16 "Students Need Two Measles Shots," *Oregon Daily Emerald*, November 10, (1997).

17 "Chapter 100 Immunization Registry," *Texas Register*, (23) 50 December 11, (1998).

18 "Disk/Trend report 1983," *Computer Week.* Mountain View, CA. (46) 11/11/83.

19 "Rigid disk drive sales to top $34 billion in 1997," *Disk/Trend News.* Mountain View, CA: Disk/Trend, Inc., 1997.

20 See note 19 Disk/Trend News.

21 Data obtained using FAOSTAT, a database produced by the Food and Agriculture Organization (FAO) of the United Nations. Rome, Italy: FAO, 1997.

22 Melton, LJ. The threat to medical-records research. *New England Journal of Medicine* 1997;337:1466-1469.

23 Louis Harris and Associates, The Equifax-Harris Consumer Privacy Survey. (Atlanta: Equifax, 1994).