Does “Big Data” spell “Big Trouble” for Lawyers?

Paul Karlzen  
Director  HR Information & Analytics  
April 1, 2015

Presented to Colorado Bar Association, Communications & Technology Law Section  
Denver, Colorado
Reflection

There are three kinds of lies — lies, damned lies and statistics.

(Mark Twain)
**Mission:**
The Mission of Catholic Health Initiatives is to nurture the healing ministry of the Church, supported by research and education. Fidelity to the Gospel urges us to emphasize human dignity and social justice as we create healthier communities.

---

**Vision:**
Our **vision** is to live up to our name as One CHI:

**Catholic:** Living our Mission and Core Values  
**Health:** Improving the health of the people and communities we serve  
**Initiatives:** Pioneering models and systems of care to enhance care delivery
Our ministry calls for exceptional people.

About 46 million people – or nearly 17% of the US population – live within a 60-mile radius of a CHI hospital.

Operations in 19 states:
- 105 hospitals, including:
  - 4 academic health centers and teaching hospitals
  - 30 critical access hospitals
  - 13 clinically integrated networks
  - 10 insurance plans

AT A GLANCE
CHI Today
AT A GLANCE
Fiscal Year 2014 Statistics

FINANCIAL HIGHLIGHTS

$21.8 BILLION IN ASSETS

$13.9 BILLION IN TOTAL ANNUAL OPERATING REVENUES

$910 MILLION IN TOTAL ANNUAL CONTRIBUTIONS TO CHARITY CARE AND COMMUNITY BENEFIT

MISSION & MINISTRY FUND
MORE THAN $55 MILLION IN GRANTS FOR BUILDING HEALTHY COMMUNITIES

CHI: “Our ministry calls for exceptional people.”
What is “Big Data”? 

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.

Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy.

The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

What is “Big Data”? 

Every day, we create **2.5 quintillion bytes of data** — so much that 90% of the data in the world today has been created in the **last two years** alone.

This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data**.

Source: IBM
What is “Big Data”?

“Big data does not exist”

“Big Data” is grammatically incorrect. It’s like saying “big water”. Rather, it should be “a lot of data” or “plenty of data”.

Data doubles every three years. This will bring an estimated 8 zettabytes by 2015.

(1 zettabyte = 250 billion DVDs or 36 million years of HD Video).

Source: “Predictive Analytics” by Eric Siegel, page 78
My definition of “Big Data”

“I shall not today attempt to further define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so.

But I know it when I see it...”
Case Maker© could not find the term “big data” in any Colorado Statute or regulation.

Case Maker© could not find the term “big data” in any Colorado case law.
Text Analytics
(aka Text Mining, Content Analysis)

An exploration and analysis of **textual (natural-language) data** by automatic and semi automatic means to discover new knowledge.

**Text mining** can help an organization derive potentially valuable business insights from text-based content such as:
- Word documents
- Email
- Social media stream posts like Facebook, Twitter and LinkedIn

**Methods** used include:
- word clustering/lexical analysis (to study word frequency distributions)
- taxonomies
- pattern recognition
- tagging/annotation

Source: [http://searchbusinessanalytics.techtarget.com/definition/text-mining](http://searchbusinessanalytics.techtarget.com/definition/text-mining)
In the 1990s, Fortune magazine had named Enron "America's Most Innovative Company" six years in a row. In 2001, Enron filed for bankruptcy due to financial fraud. The email dataset played a role in showing the communications among executives in the Enron corporation.

Source: http://flowingdata.com/2008/01/24/6-influential-datasets-that-changed-the-way-we-think/
Mentions by Detail L4
2015 YTD

Decisions (74.3%)

Dress Code (2.2%)
Healthcare Related Benefits (2.7%)
Coffee (5.3%)
Direct Supervisors (15.5%)
Legal Implications of Textual Analysis

Much more than E-discovery

If computers can read through thousands, if not hundreds of thousands of documents and find specific documents related to your litigation or legal work –

why should companies pay law firms to have associates do this work?
Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.
Dr. Eric Siegel, discussing his book *Predictive Analytics*:

The shortest definition is my book's subtitle, the power to predict who will click, buy, lie, or die. Predictive analytics is the technology that learns from data to make predictions about what each individual will do -- from thriving and donating to stealing and crashing your car.

By doing so, organizations boost the success of marketing, auditing, law-enforcing, medically treating, educating, and even running a political campaign for president.
Predictive Analytic Examples

Netflix
Amazon
Credit Score
Weather Forecast
Advanced predictive modeling and analytics technologies to provide a comprehensive and accurate view across all providers, recipients, and geographic locations within the Medicaid program in order to:

I. Identify and analyze those billing and utilization patterns that are present a high risk of fraudulent activity;

II. Be easily integrated into the existing Medicaid program claims operations;

III. Undertake and automate such analysis before payment is made to minimize disruptions to state department operations and speed claim resolution;

IV. Prioritize the identified transactions for additional review before payment is made based upon the likelihood of potential waste, fraud, or abuse;

V. Obtain outcome information from adjudicated claims to allow for refinement and enhancement of the predictive analytics technologies based on historical data and algorithms with the system; and

Colorado Statutes
Title 25.5 HEALTH CARE POLICY AND FINANCING ADMINISTRATION
Article 1. Department of Health Care Policy and Financing
Part 1. GENERAL PROVISIONS
Current through Chapter 2 of the 2015 Legislative Session
§ 25.5-1-114.5. Medicaid fraud detection – request for information
Colorado Cases – Predictive Analytics

CHI: “Our ministry calls for exceptional people.”
What is Correlation?

In statistics and probability theory, correlation means how closely related two sets of data are related.

Correlation usually has one of two directions. These are positive or negative. If it is positive, then the two sets go up together. If it is negative, then one goes up while the other goes down.

(Source: Wikipedia)

Correlation does not always mean that one causes the other. It is very possible that there is a third factor involved.
Causation

The act or process of causing something to happen or exist

or

The relationship between an event or situation and a possible reason or cause

(Causeation) The causing or producing an effect <the plaintiff must prove causation>.

Benjamin Cardozo, The Paradoxes of Legal Science 83 (1928)

(Source: Black’s Law Dictionary 10th ed. 2014)
Divorce rate in Maine correlates with Per capita consumption of margarine (US)

Correlation: 99%
Sources: US Census & USDA tylervigen.com

<table>
<thead>
<tr>
<th>Divorce rate in Maine</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divorces per 1000 people (US Census)</td>
<td>5</td>
<td>4.7</td>
<td>4.6</td>
<td>4.4</td>
<td>4.3</td>
<td>4.1</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Per capita consumption of margarine (US)</td>
<td>8.2</td>
<td>7</td>
<td>6.5</td>
<td>5.3</td>
<td>5.2</td>
<td>4</td>
<td>4.6</td>
<td>4.5</td>
<td>4.2</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Correlation: 0.992558

CHI: “Our ministry calls for exceptional people.”

Source: Spurious Correlations - www.tylervigen.com
Online revenue on Black Friday correlates with People killed by foreign body entering into or through eye or natural orifice

<table>
<thead>
<tr>
<th>Year</th>
<th>Online revenue on Black Friday (Millions of dollars)</th>
<th>People killed by foreign body entering into or through eye or natural orifice (Deaths (US) (CDC))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>534</td>
<td>22</td>
</tr>
<tr>
<td>2009</td>
<td>595</td>
<td>25</td>
</tr>
<tr>
<td>2010</td>
<td>648</td>
<td>29</td>
</tr>
</tbody>
</table>

Correlation: 0.992471

CHI: “Our ministry calls for exceptional people.”

Source: Spurious Correlations - www.tylervigen.com
Lawyers in Colorado correlates with Visitors to Disney World's Animal Kingdom

<table>
<thead>
<tr>
<th>Year</th>
<th>Lawyers in Colorado (ABA)</th>
<th>Visitors to Disney World's Animal Kingdom (TEA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>18,376</td>
<td>9.49</td>
</tr>
<tr>
<td>2008</td>
<td>18,894</td>
<td>9.54</td>
</tr>
<tr>
<td>2009</td>
<td>19,340</td>
<td>9.59</td>
</tr>
<tr>
<td>2010</td>
<td>19,737</td>
<td>9.686</td>
</tr>
<tr>
<td>2011</td>
<td>20,211</td>
<td>9.783</td>
</tr>
<tr>
<td>2012</td>
<td>20,768</td>
<td>9.998</td>
</tr>
</tbody>
</table>

Correlation: 0.966559

Source: Spurious Correlations - www.tylervigen.com

CHI: “Our ministry calls for exceptional people.”
CHI: “Our ministry calls for exceptional people.”

Source: Spurious Correlations - www.tylervigen.com
Legal Implications of Predictive Analytics

What if computers could predict the outcomes of cases given textual analysis and predictive analytics?
Legal Implications of Predictive Analytics
Could the US government extend their current programs and use big data, textual analysis and predictive analytics to arrest and imprison terror suspects before they commit a crime?

What if your client uses big data, textual analysis and predictive analytics to determine if a potential employee is a “health risk” and doesn’t extend them an offer based on this prediction?
Questions?