Big (User) (Health) Data Management

Sihem Amer-Yahia
DR CNRS @ LIG
Sihem.Amer-Yahia@imag.fr

Prospectom workshop
21 Nov. 2014
Questions to be addressed

- What is (and what is not) data management?
- Where is user data in the health domain?
- Why is Big User (Health) Data management different from traditional data management?
  - Big (User) (Health) Data preparation
  - Big (User) (Health) Data mining
  - Validation
Data Management

• **What it is.** It is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise.” DAMA International}

• **What it isn’t.** It is not about deciding which data to gather and which applications to build for this data (need for a domain expert)

• **What makes it a science.** It develops principled and reusable methods for gathering, organizing and exploiting data

  • **Gathering:** dumps, crawlers, Application Programming Interfaces (APIs)
  • **Organizing:** pre-processing and storing (because data is made persistent on disk) structured and semi-structured content
  • **Exploiting:** via exploration (search and querying languages and algorithms) and recommendation (functions and algorithms)
(Traditional) data management

Separation between physical and logical layers

Application specification

Access optimization

Data storage and index creation

relational tables

HTM/XML backend
User Health Data

- Electronic Health Records (EHRs)
- User-Generated Content (UGC)
EHRs

- Physiological monitoring
- Genomics
- Anatomical imaging
- Also includes demographics, medical history, medication and allergies, immunization status, laboratory test results, personal statistics like weight, and billing information
<table>
<thead>
<tr>
<th><strong>Admission Nr.</strong></th>
<th>20045000000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Senor</td>
</tr>
<tr>
<td><strong>Family name</strong></td>
<td>Mario</td>
</tr>
<tr>
<td><strong>Given name</strong></td>
<td>Banderas</td>
</tr>
<tr>
<td><strong>Date of birth</strong></td>
<td>08/07/2004</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td>male</td>
</tr>
<tr>
<td><strong>Blood group</strong></td>
<td>AB</td>
</tr>
</tbody>
</table>

**Date:** 08/07/2004

**Type:** Tetagam

**Medicine:** Anti-tetanus immunization

**Dosage:** 2 mg/dl

**Titer:** 345

**Refresh date:** 08/06/2006

**Application type:** Subcutaneous

**Application by:** admin

**Notes**

---

**Search:**

Please enter search keyword:

**Top 10 Quicklist**

- Tetagam

---

**Options for this patient**

- Confirmation of inability to work
- Charts folder
- Diagnostic Results
- Medocs
- DRG (composite)
- Prescriptions
- Notes & Reports
- Immunization
- Measurements
Cost benefits of EHRs

http://www.bilan.ch/economie-plus-de-redaction/lheure-de-cybersante-sonne/page/0/5

- McKinsey claims that cyberhealth technology can save between $300 and $450 billions in yearly health costs in the USA.
- PricewaterhouseCoopers estimates it at $99 billion euros in Europe.
EHRs’ adoption
http://www.bilan.ch/economie-plus-de-redaction/lheure-de-cybersante-sonne/page/0/5

• Slow adoption in Switzerland and in France
  • Dossier Medical Personnel (DMP) in France (since August 2004) with a lot of controversy for adoption: www.dmp.gouv.fr
  • 503 751 DMPs today (against 381,015 last year)
  • The word « cybersanté » does not mean anything to 81% of the Swiss population: Institut de recherche GFS Berne for InfoSocietyDays (Feb 2014)

• In the USA
  – $24.4 billions to 4600 hospitals and 400,000 professionals
  – since the adoption of the HITECH Act in 2009 by the Obama administration
    • number of private doctors using an e-health system (21.8% in 2009 against 48.1% in 2013)
    • 44% in 2013 hospitals against 12.2% in 2009

• But no clear evidence of cost reduction
User data

• EHRs
• UGC blurs the boundaries between user-generated, expert-generated/approved, clean/noisy data and is growing at a fast rate
  • Mobile Medical Apps (MMAs)
  • Online forums
Mobile Medical Apps (MMAs)

http://fda.gov

- Mobile apps are software programs that run on smartphones and other mobile communication devices. They can also be accessories that attach to a smartphone or other devices.
- MMAs are medical devices that are mobile apps, meet the definition of a medical device and are an accessory to a regulated medical device or transform a mobile platform into a regulated medical device.
Devices and tools

• Medical devices as wearables or as MMAs
  • thermometers, scales, blood pressure cuffs, electrocardiogram, sleep patterns
  • MyFitnessPal (> 40M users): tracks nutritional intake and monitors weight goals, connects to friends, API to connect to apps (e.g., Withings Scale and RunKeeper)
  • The 20 most popular MMAs on exercise and wellness account for 231 million downloads worldwide

• 5 years ago, it was difficult to find sensors that could fit into a handheld device, not drain power and provide good signal.
The AliveCor handheld heart monitor and a sample ECG that can be emailed to doctors and patients.

Mobisante's smartphone-based ultrasound imaging system (costs 1/10th of a regular ultrasound)

Prospectom 2014
Quantified self with Withings

Prospectom 2014
MMAs categories

• General healthcare and fitness
  • Fitness & nutrition
  • Health tracking tools
  • Managing medical conditions
  • Medical compliance
  • Wellness (traditional and corporate)

• Medical information
  • Diagnostic Tools including predispositions
  • Continuing Medical Education (CME)
  • Alerts and Awareness

• Remote monitoring, collaboration and consultation

• Healthcare management
  • Logistical & payment support
  • Patient health records
MMAs’ availability

• About 100,000 applications are available via iTunes and Google Play (European Commission - last Spring)

• For MMAs that pose minimal risk to patients, the FDA will not expect manufacturers to register:
  – E.g., help patients self-manage their disease or condition without providing specific treatment suggestions;
  – E.g., provide patients with simple tools to organize and track their health information;
LA CYBERSANTÉ GAGNE DU TERRAIN

70% DES PATIENTS AMÉRICAINS SONT FAVORABLES À UNE CONSULTATION MÉDICALE À DISTANCE PLUTÔT QU’À UNE VISITE

- 19% Consultation via un chat vidéo
- 20% Consultation via messagerie instantanée
- 21% Consultation via email
- 20% Consultation via textos
- 23% Consultation téléphonique

70% DONT:
Online forums

- **General:** WebMD.com (> 80M/month), health.nih.gov, healthfinder.gov, intelihealth.com, mayoclinic.org

- **Personalized:** HealthTap, “triage” system, where consumers ask doctors for the most effective way to get specific care

- **Collaborative:** YellowCard (public), PatientsLikeMe (private)
  - help patients with a chronic condition answer: “Given my status, what is the best outcome I can hope to achieve, and how do I get there?”
  - experience sharing via patient, reported outcomes, finding similar patients matched on demographic and clinical characteristics, and aggregated stats
  - CureTogether, Diabetic Connect
Bipolar Type II

About this Condition: Bipolar II disorder is a bipolar spectrum disorder characterized by at least one hypomanic episode and at least one major depressive episode; with this disorder, depressive episodes are more frequent and more intense than manic episodes. It is believed to be under-diagnosed because hypomanic behavior often presents as incredibly high-functioning behavior. Indeed, to a physician or psych.. Read more

Synonyms: Bipolar 2, Manic-Depressive Disorder

Who has this Condition?

11,879 patients have this condition
405 New patients this month
For 6,789 this is their primary condition

Top Treatments

<table>
<thead>
<tr>
<th>Treatment</th>
<th>#Patients</th>
<th>#Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen</td>
<td>113</td>
<td>32</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>97</td>
<td>20</td>
</tr>
<tr>
<td>Topiramate</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>Excedrin Migraine</td>
<td>42</td>
<td>16</td>
</tr>
<tr>
<td>Butalbital-acetaminophen-caffeine</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>Chiropractic</td>
<td>24</td>
<td>0</td>
</tr>
</tbody>
</table>

See all

Efficacy: Major, Moderate, Slight, None, Can't tell

Top Symptoms

<table>
<thead>
<tr>
<th>Symptom</th>
<th>#Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>113</td>
</tr>
<tr>
<td>Depression</td>
<td>97</td>
</tr>
<tr>
<td>Headache</td>
<td>67</td>
</tr>
<tr>
<td>Migraine</td>
<td>42</td>
</tr>
<tr>
<td>Fatigue</td>
<td>31</td>
</tr>
<tr>
<td>Insomnia</td>
<td>24</td>
</tr>
</tbody>
</table>

See all

Severity: Severe, Moderate, Mild, None

Demographics for this Condition

Age:
- <20: 29
- 20-29: 30
- 30-39: 40
- 40-49: 50
- 50-59: 60
- 60-69: 70+
- 70+: 0

Gender:
- 80% Women
- 20% Men

Find Patients Just Like You

Join Now! (It’s free!)

Patient Spotlight

Rachel44
Female, 57 years, Lancaster, PA
Hi, I am Rachel and I am married to my wonderful husband of 20 years, and we have 2 wonderful sons and a daughter in law. I was diagnosed back in '84, wh... See Profile

Links

PatientsLikeMe Research Updates
Take a minute to read recently published reports from the PatientsLikeMe R&D team

What are others saying about PatientsLikeMe?
Tune in to Twitter
Find us In the News
Read patient Testimonials

PatientsLikeMe Videos
And also, Twitter

- **Web search:** users express *need for information* flu medicine
- **Social media:** users express self information sick with the flu
Mining user data

- UGC contains loads of valuable information for business intelligence, public health, ...
- But is also noisy, unreliable, incomplete, uncertain and subjective!
  - how to extract valuable information from raw user data?
CrowdHealth (CNRS MASTODONS) with Noha Ibrahim, Etienne Dublé, Sumit Sidana, Ankita Atray

- **Goal**
  - enable *health and nutrition-related hypothesis testing* in physical and virtual spaces
  - extract observations from Twitter
  - a collaboration between 2 CNRS institutes (INS2I and SHS) and Paris Descartes and UREN (Univ. de Villetaneuse)

- **Partners**
  - researchers on *scalable algorithms for mining personal data and times series*, nutritionists, and researchers studying *individuals in geographical spaces*

- **Budget**: 30K euros since May 2014
Data collection

- Started 10/2014 for a total of 366 million tweets
- This past week: 68 million tweets collected
- At this rate: 3 billion tweets/year
  
  \[ <\text{timestamp},\text{latitude/longitude},\text{hashtags},\text{text},\text{user}> \]

- The collection process:
  - Collector (in Python) connects to the Twitter API
  - Collector asks for geo-tagged tweets
  - Twitter API provides a stream of tweets
  - Collector stores obtained tweets in a PostgreSQL

You Are What You Tweet: Analyzing Twitter for Public Health
Michael J. Paul, Mark Drezde (Johns Hopkins U.) ICWSM 2011
Oct/Nov tweets

**Provenance**
north-america 27.4%
south-america 23.6%
asia 22.8%
europe 16.5%
oceania 5.6%
africa 4.1%

<table>
<thead>
<tr>
<th>langage</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>44%</td>
</tr>
<tr>
<td>es</td>
<td>19%</td>
</tr>
<tr>
<td>pt (portugais)</td>
<td>15%</td>
</tr>
<tr>
<td>tr (turc)</td>
<td>4.7%</td>
</tr>
<tr>
<td>ja</td>
<td>3.3%</td>
</tr>
<tr>
<td>fr</td>
<td>3.0%</td>
</tr>
<tr>
<td>id (indonésien)</td>
<td>2.7%</td>
</tr>
<tr>
<td>ru</td>
<td>2.4%</td>
</tr>
<tr>
<td>ar (arabe)</td>
<td>1.4%</td>
</tr>
<tr>
<td>it</td>
<td>0.9%</td>
</tr>
</tbody>
</table>
136,285,644 English Tweets

Gather and Clean Tweets
Remove URLs, re-tweets, hashtags and user names

98,687,620 Clean

Filter tweets with Health Related Keywords

2,841,536 Key-Word Filtered

100,000 positive tweets

SVM classifier

357,683 Health Tweets

Ailment Topic Aspect Model

I FEEL LIKE I'M GOING TO DIE OF BIEBER FEVER. NO JOKE.

Web design class gives me a huge headache everytime.
Input: 100,000 positive tweets
Output: 357,683 tweets
9 (oct/nov) ailments discovered
Each tweet is a probability distribution over 9 ailments

<table>
<thead>
<tr>
<th>Ailment</th>
<th>Allergies</th>
<th>Aches/Pains</th>
<th>Dental</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Words</td>
<td>allergies stop eyes</td>
<td>body head need</td>
<td>meds killers</td>
</tr>
<tr>
<td></td>
<td>allergic</td>
<td>hurts</td>
<td>dentist teeth</td>
</tr>
<tr>
<td>Symptoms</td>
<td>sneezing cold</td>
<td>pain aches stomach</td>
<td>pain toothache</td>
</tr>
<tr>
<td></td>
<td>coughing</td>
<td></td>
<td>sore</td>
</tr>
<tr>
<td>Treatments</td>
<td>medicine benadryl</td>
<td>massage “hot bath”</td>
<td>braces surgery</td>
</tr>
<tr>
<td></td>
<td>claritin</td>
<td>ibuprofen</td>
<td>antibiotics</td>
</tr>
</tbody>
</table>
User data management stack

Result validation

User Data Analytics

Data preparation

raw user data
User data preparation

Data Collection

Data Sanitation

Data Transformation

Ailment Topic Aspect Model

clean(er) user data

Pruning
Text Processing
Normalization
Enrichment

raw user data
Positive/negative examples

• Insanity does wonders for me. I lost 30 lbs the first month I started.
  **Ailment – Overweight**

• Wasn't sure how much Benadryl to take for my allergies...here's to hoping I'm not passed out at my desk in an hour
  **Ailment – Common cold**

• I always forget I have a nose ring until I accidentally rip it out and start crying.
  **Ailment – Ear Infection**

• Ending my night with an unused luke Bryan ticket, an allergic reaction to a dog, and 10 other misfortunes at least I have my health !!!!!!!
  **Ailment – Overweight**
An exploration dashboard
From London

watching tfios and drinking tea and juice and munching on halls sweets because i'm coughing so much that i'm sick n i can't sleep 😐
# You Are What You Tweet: Analyzing Twitter for Public Health

Michael J. Paul, Mark Drezde (Johns Hopkins U.) ICWSM 2011

## Drug Use Analytics

<table>
<thead>
<tr>
<th>Word</th>
<th>#</th>
<th>Ent.</th>
<th>Most Common Ailments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tylenol</td>
<td>1807</td>
<td>1.57</td>
<td>HA (39%), IN (30%), Cold (9%)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>1125</td>
<td>1.54</td>
<td>HA (37%), DN (21%), Aches (17%)</td>
</tr>
<tr>
<td>Advil</td>
<td>1093</td>
<td>1.08</td>
<td>HA (61%), Cold (6%), DN (5%)</td>
</tr>
<tr>
<td>Aspirin</td>
<td>885</td>
<td>1.04</td>
<td>HA (69%), IN (10%), Aches (10%)</td>
</tr>
<tr>
<td>Vicodin</td>
<td>505</td>
<td>1.33</td>
<td>DN (61%), Injuries (11%), HA (10%)</td>
</tr>
<tr>
<td>Codeine</td>
<td>406</td>
<td>1.94</td>
<td>Cold (25%), DN (19%), HA (17%)</td>
</tr>
<tr>
<td>Morphine</td>
<td>206</td>
<td>1.17</td>
<td>DN (59%), Infection (22%), Aches (9%)</td>
</tr>
<tr>
<td>Aleve</td>
<td>183</td>
<td>1.10</td>
<td>HA (62%), IN (15%), DN (14%)</td>
</tr>
</tbody>
</table>

## Allergy Medication

<table>
<thead>
<tr>
<th>Word</th>
<th>#</th>
<th>Ent.</th>
<th>Most Common Ailments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benadryl</td>
<td>871</td>
<td>1.24</td>
<td>Allergies (64%), Skin (13%), IN (12%)</td>
</tr>
<tr>
<td>Claritin</td>
<td>417</td>
<td>0.54</td>
<td>Allergies (88%), HA (5%)</td>
</tr>
<tr>
<td>Zyrtec</td>
<td>386</td>
<td>0.49</td>
<td>Allergies (90%)</td>
</tr>
<tr>
<td>Sudafed</td>
<td>298</td>
<td>1.61</td>
<td>Allergies (39%), Cold (21%), HA (20%)</td>
</tr>
</tbody>
</table>
Prepare, mine then validate

• User data contains loads of valuable information for business intelligence, public health, ...
• But is noisy, unreliable, incomplete, uncertain and subjective!
  – how to extract valuable information from raw user data?
  – how to prepare raw user data then extract valuable information?
  – how to validate findings?
    • automatically
    • with humans
Flu trends on Twitter

- You Are What You Tweet: Analyzing Twitter for Public Health
- Michael J. Paul, Mark Drezde (Johns Hopkins U.) ICWSM 2011
Tasks and Data

What is a good cafe in Tokyo?

Which are good cafes in Grenoble?

Is A is better than B?

Your Web browser

Crowd4U Terminals

Contributors in Academia

Floor

Requesters
Crowd4U deployment

Universiteit Hasselt
Hokkaido University
University of Tsukuba
The University of Tokyo
Kyushu University
Meiji University
Doshisha University
Ritsumeikan University
Kyoto University
and other 12 universities In Japan

>900 tasks/day (Dec., 2013)
Next steps

• Apply classification to nutrition
  – Easier?
    • If someone tweets about pizza, he/she is likely to have had it
  – Count one serving and compute carbohydrates, fat, calories
  – Aggregate

• Compute observations that relate health and nutrition
Summary

• Managing UGC is a Big Data problem
• New opportunities for domain experts
  • Large-scale analytics on user-generated content
  • Can researchers help public health officials reduce current more expensive and time consuming method?
• New opportunities for researchers in data management, data mining and related areas
  • Data preparation
  • Crowdsourcing for targeted human involvement