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

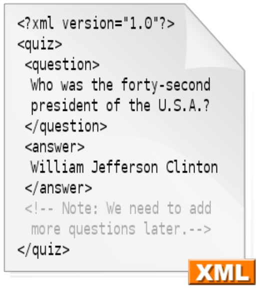
Application specification

Access optimization

Data storage and index creation

Ward No.	Ward name	Type	No. of Beds
3	Carey	Medical	8
6	Bracken	Medical	16
7	Brent	Surgical	12
8	Meavy	Surgical	10

relational tables
HTM/XML backend

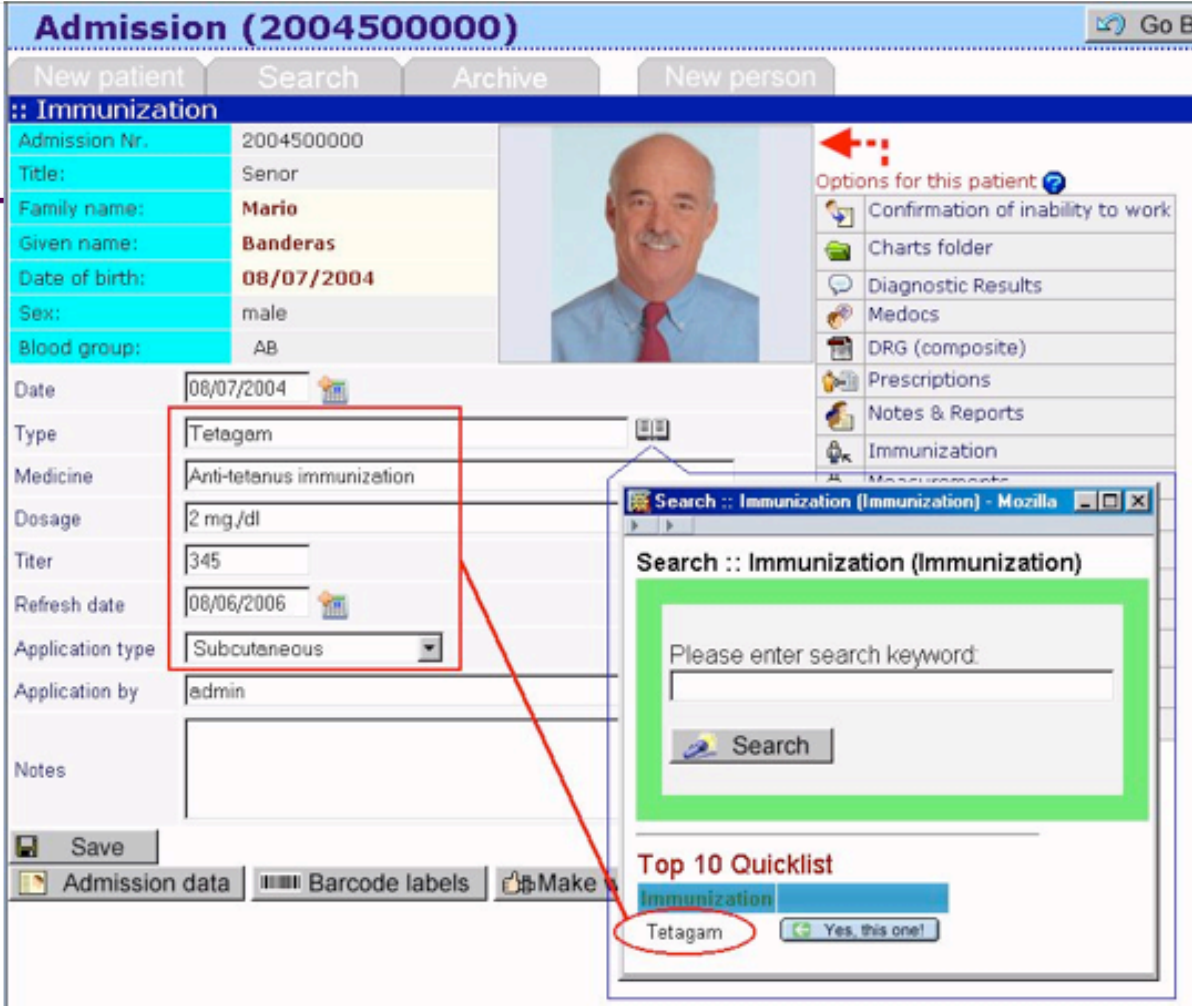


User Health Data

- 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


Admission (2004500000)
Go Ba
New patient
Search
Archive
New person
:: Immunization
Admission Nr.
2004500000
Title:
Senor
Family name:
Mario
Given name:
Banderas
Date of birth:
08/07/2004
Sex:
male
Blood group:
AB
Options for this patient
Confirmation of inability to work
Charts folder
Diagnostic Results
Medocs
DRG (composite)
Prescriptions
Notes & Reports
Immunization
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
Save
Admission data
Barcode labels
Search :: Immunization (Immunization) - Mozilla
Search :: Immunization (Immunization)
Please enter search keyword:
Search
Top 10 Quicklist
Immunization
Tetagam
Yes, this one!

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

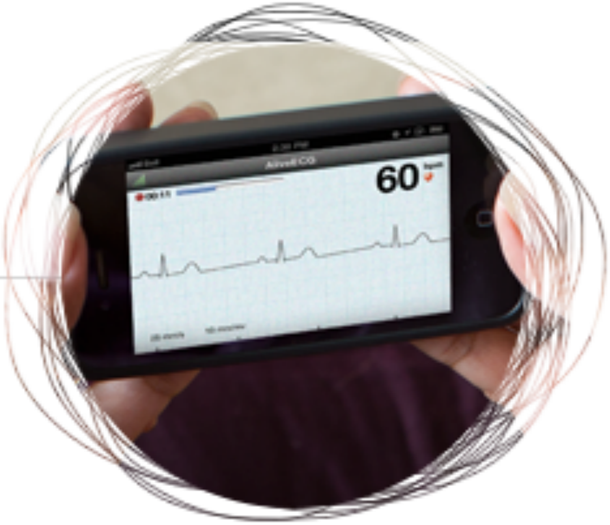
Mobile Medical Apps (MMAs)

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.**

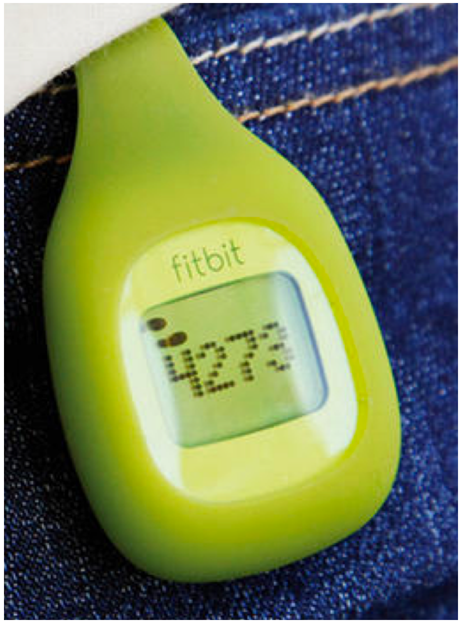
Devices and tools

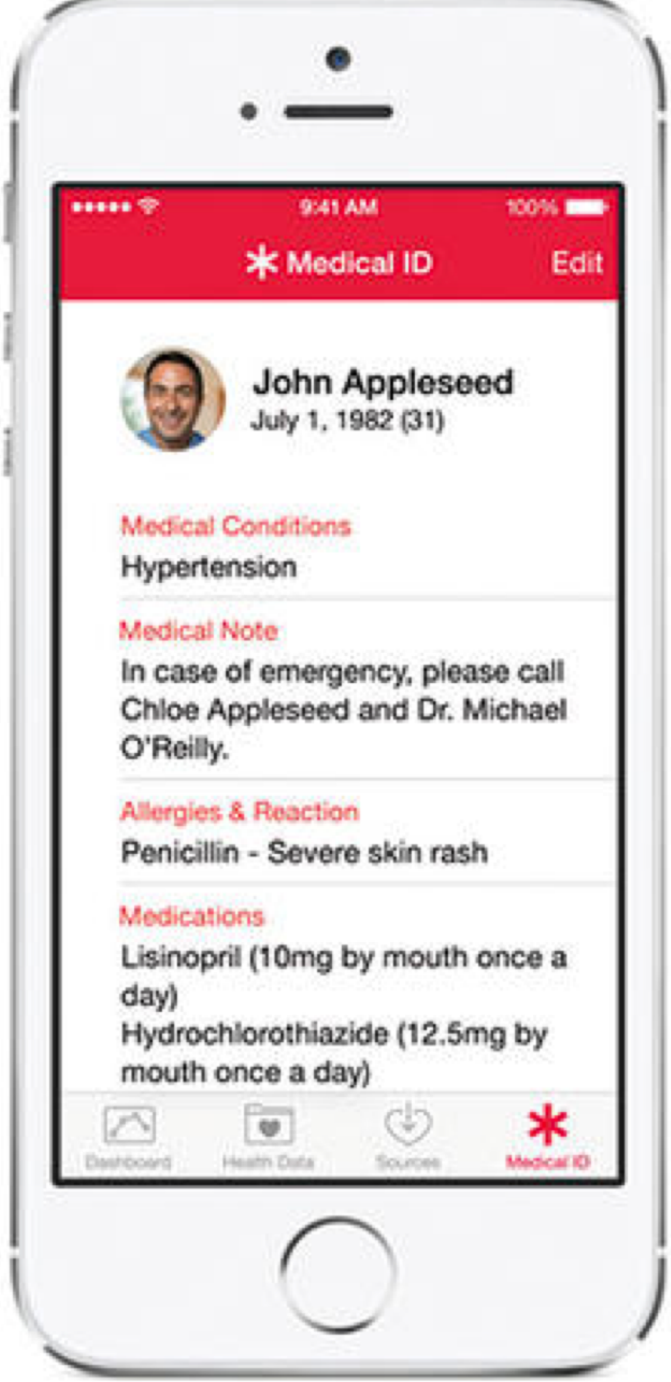
- **Medical devices as wearables or as MMAs**
  - thermometers, scales, blood pressure cuffs, electrocardiogram, sleep patterns
  - **MyFitnessPal** (> 40M users): tracks nutritional intake and monitors weight goals, connects to friends, API to connect to apps (e.g., **Withings Scale** and **RunKeeper**)
  - 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)





Quantified self with Withings

MMAs categories

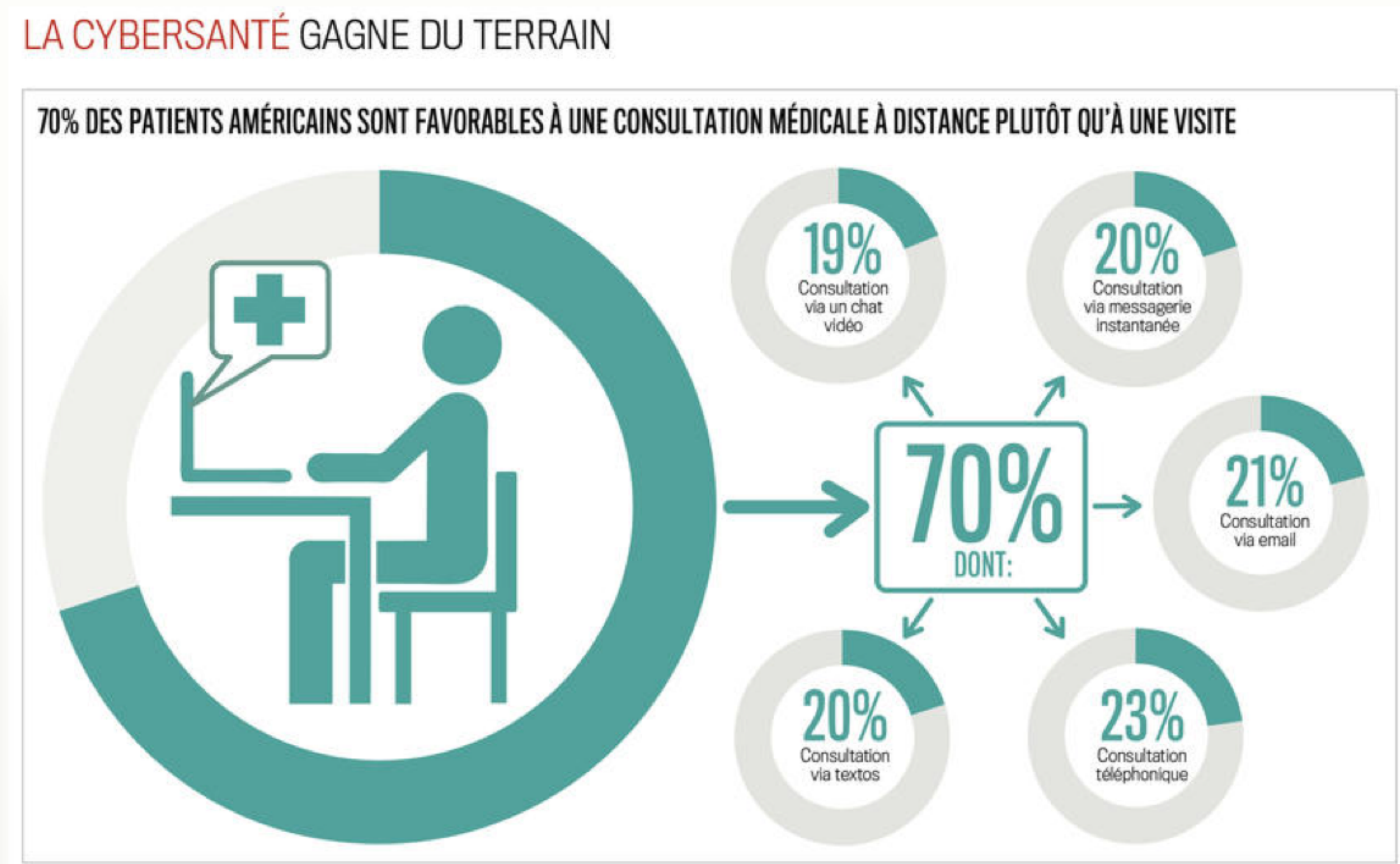
- **General healthcare and fitness**
  - Fitness & nutrition
  - Health tracking tools
  - Managing medical conditions
  - Medical compliance
  - Wellness (traditional and corporate)
- **Medical information**
  - Diagnostic Tools including predispositions
  - Continuing Medical Education (CME)
  - Alerts 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*;

MMAs' adoption (USA)

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



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


patientslikeme
Home
Patients
Treatments
Conditions & Symptoms
Research
Search
Help
Crisis
Conditions & Symptoms > Bipolar Type II
Bipolar Type II
About this Condition: Bipolar II disorder is a bipolar spectrum disorder characterized by at least one hypo-manic 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 psyc.. Read more
Synonyms: Bipolar 2, Manic-Depressive Disorder
Do you have this? Join PatientsLikeMe and start filling out your profile today!
Who has this Condition?
11,879 patients have this condition
405 New patients this month
For 6,789 this is their primary condition
Age
<20 20-29 30-39 40-49 50-59 60-69 70+
Gender
80% Women
20% Men
Top Treatments #Patients #Evaluations
Ibuprofen 113 32
Acetaminophen 97 20
Topiramate 67 0
Excedrin Migraine 42 16
Butalbital-acetaminophen-caffeine 31 0
Chiropractic 24 0
See all
Efficacy Major Moderate Slight None Can't tell
Top Symptoms #Patients
Mania 113
Depression 97
Headache 67
Migraine 42
Fatigue 31
Insomnia 24
See all
Severity Severe Moderate Mild None
Demographics for this Condition
Find Patients Just Like You »
Join Now! (It's free!)
Patient Spotlight
MS: 6 yrs
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**

    < *timestamp, lattitude/longitude, hashtags, text, user* >

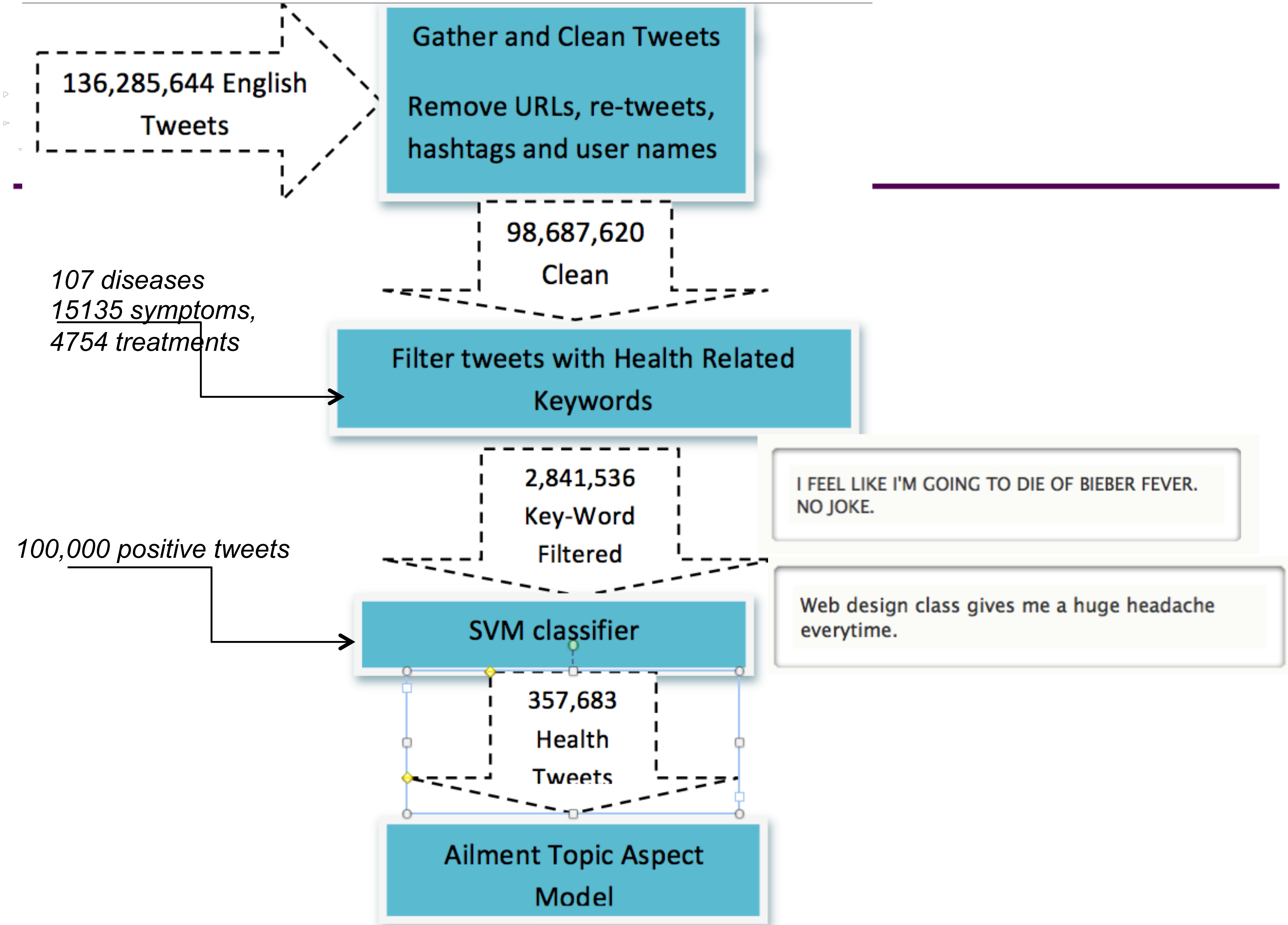
- **The collection process:**
  - Collector (in Python) connects to the Twitter API
  - Collector asks for geo-tagged tweets
  - Twitter API provides a stream of tweets
  - Collector stores obtained tweets in a PostgreSQL

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

Oct/Nov tweets

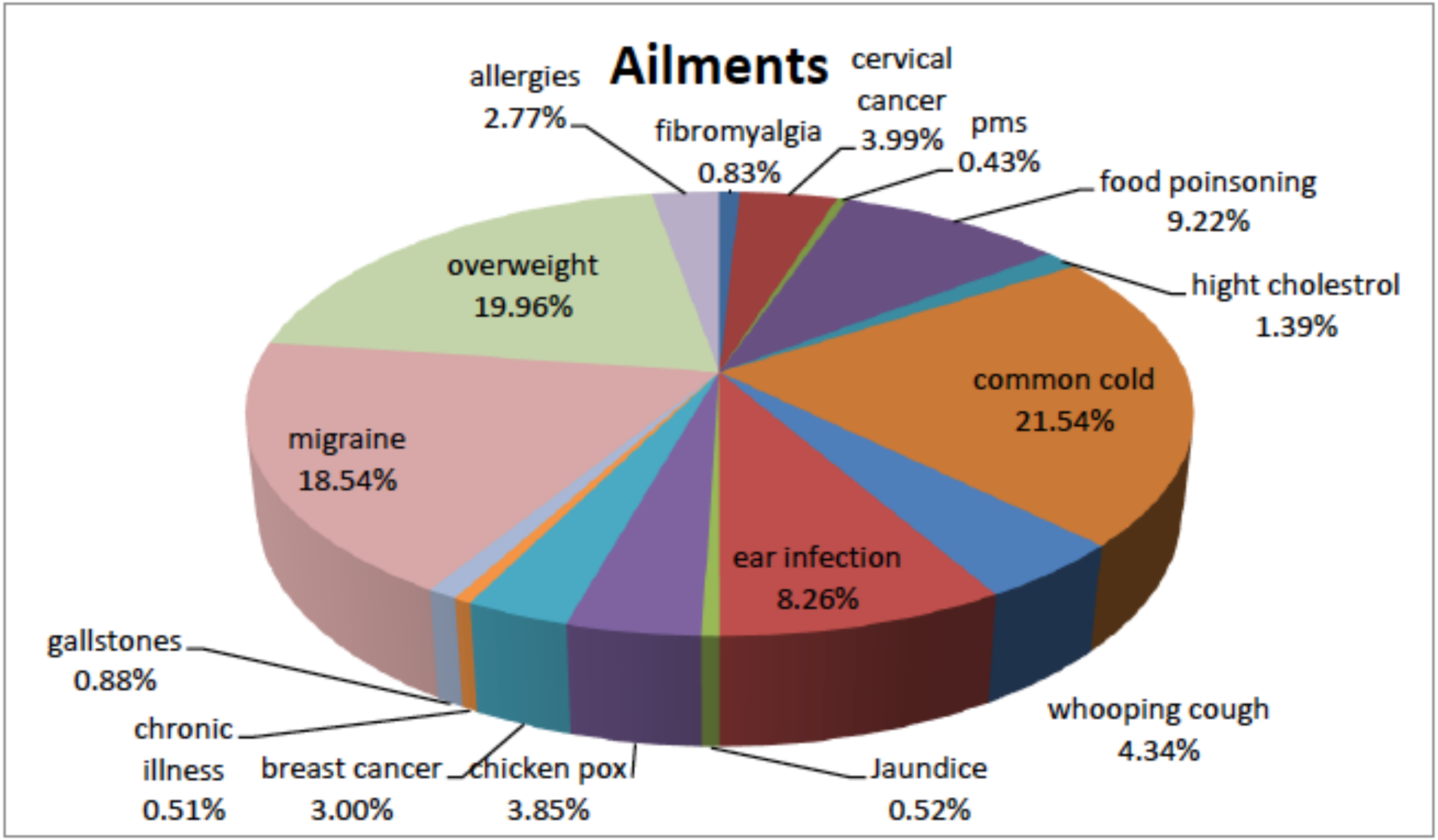
Provenance
north-america 27.4%
south-america 23.6%
asia 22.8%
europe 16.5%
oceania 5.6%
africa 4.1%

langage	pourcentage
en	44%
es	19%
pt (portugais)	15%
tr (turc)	4.7%
ja	3.3%
fr	3.0%
id (indonésien)	2.7%
ru	2.4%
ar (arabe)	1.4%
it	0.9%



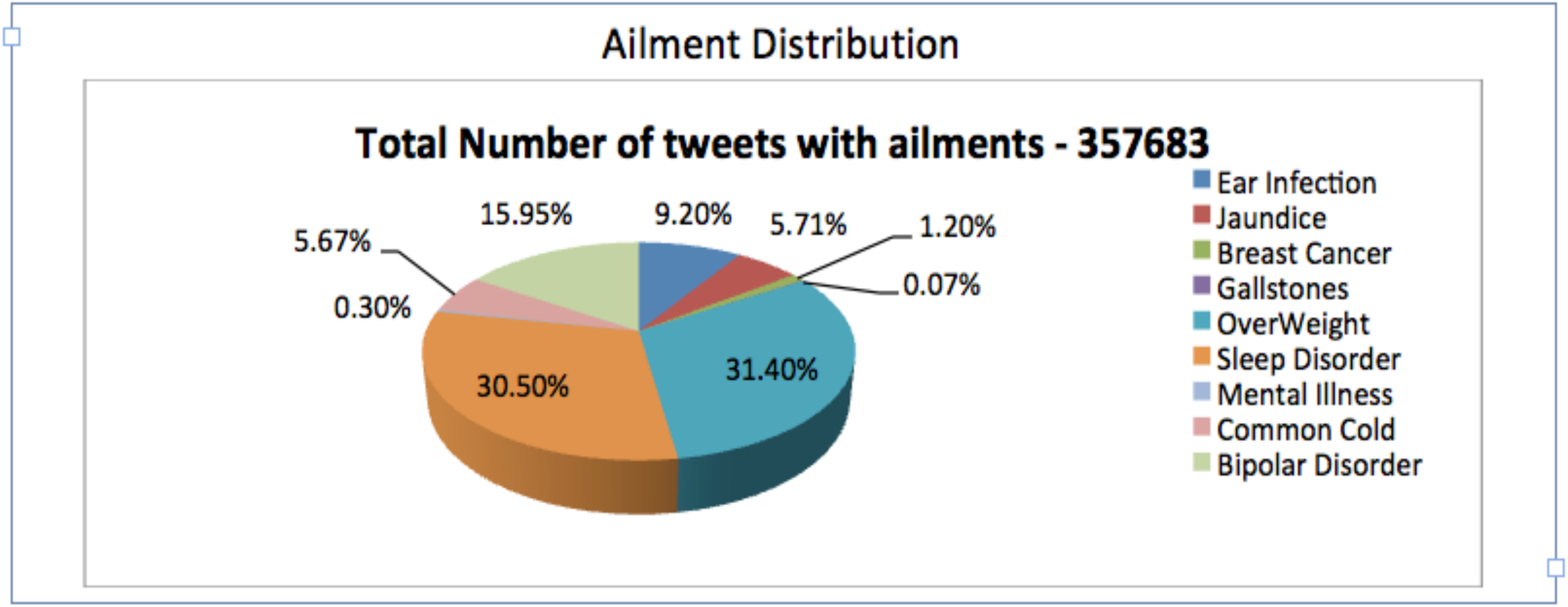
Prospectom 2014

Input: 100,000 positive tweets



Prospectom 2014

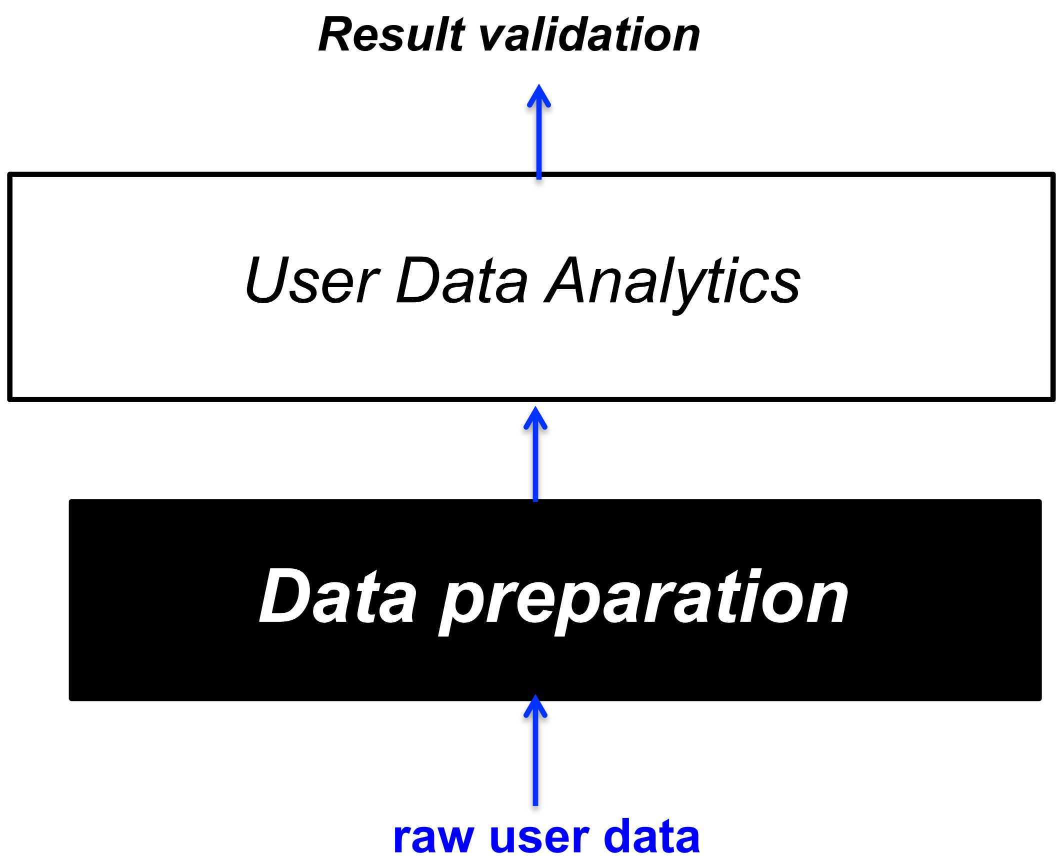
Output: 357,683 tweets
9 (oct/nov) ailments discovered



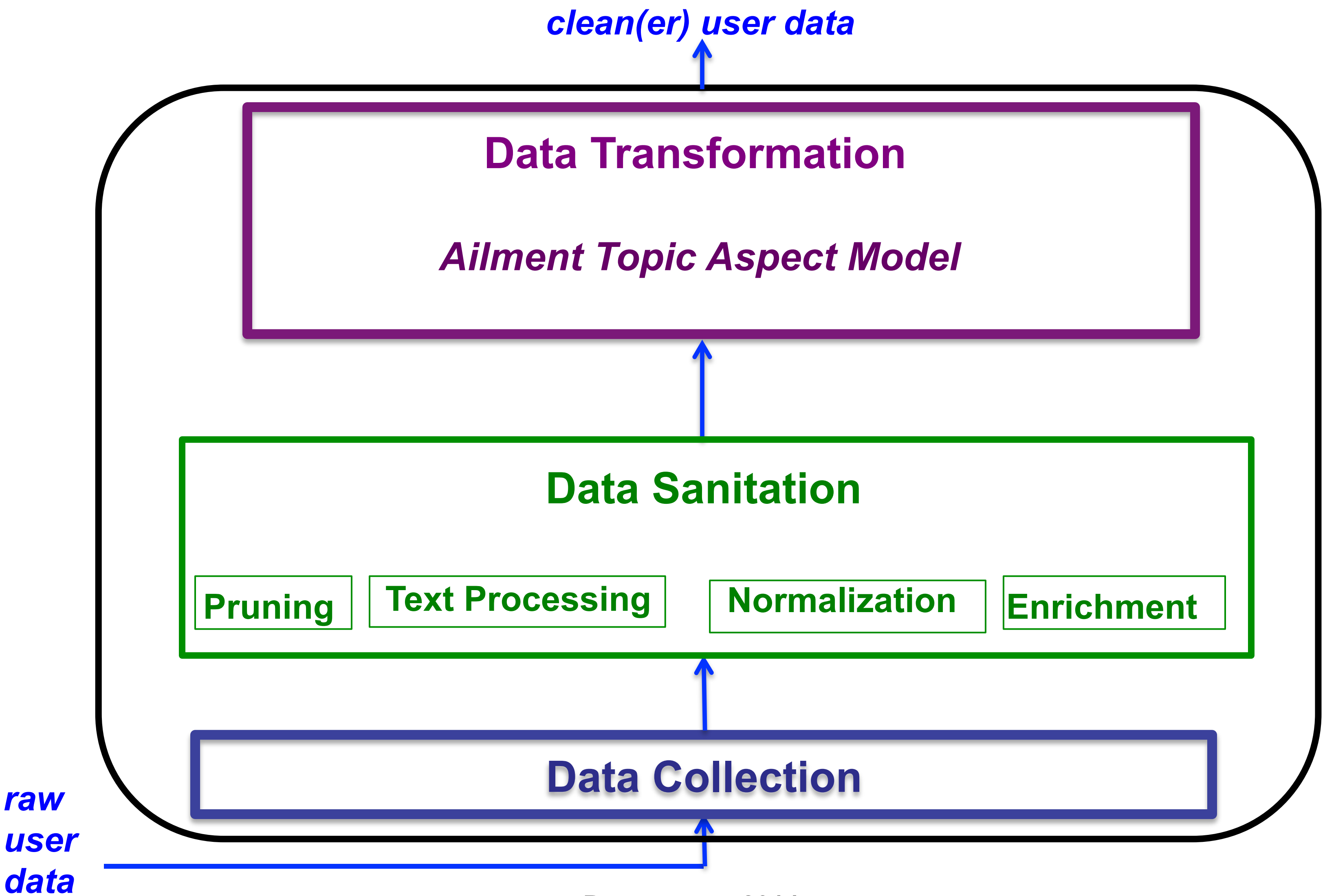
Each tweet is a probability distribution over 9 ailments

Ailment	Allergies	Aches/Pains	Dental
General Words	allergies stop eyes allergic	body head need hurts	meds killers dentist teeth
Symptoms	sneezing cold coughing	pain aches stomach	pain toothache sore
Treatments	medicine benadryl claritin	massage “hot bath” ibuprofen	braces surgery antibiotics

User data management stack



User data preparation



Positive/negative examples

- *Insanity does wonders for me. I lost 30 lbs the first month I started.*

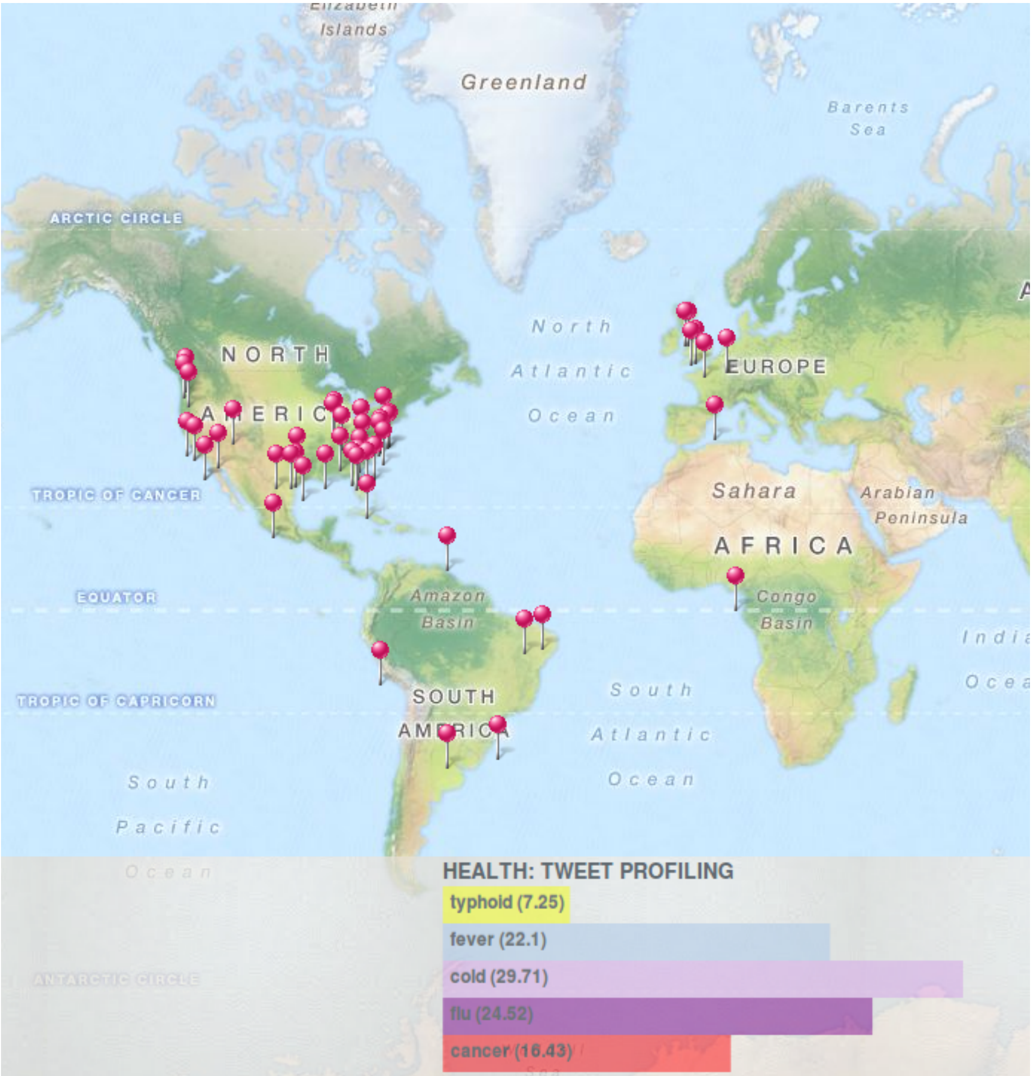
  Ailment – Overweight
- *Wasn't sure how much Benadryl to take for my allergies...here's to hoping I'm not passed out at my desk in an hour*

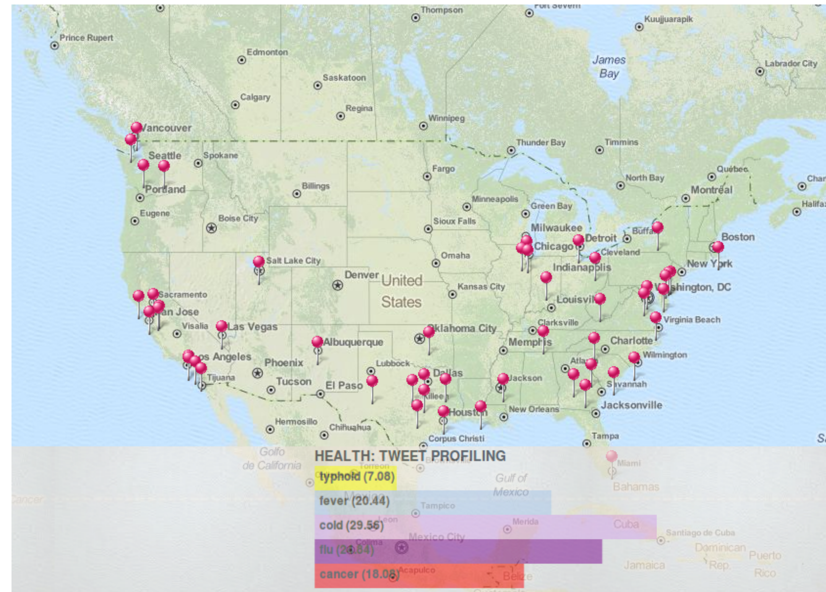
  Ailment – Common cold
- *I always forget I have a nose ring until I accidentally rip it out and start crying.*

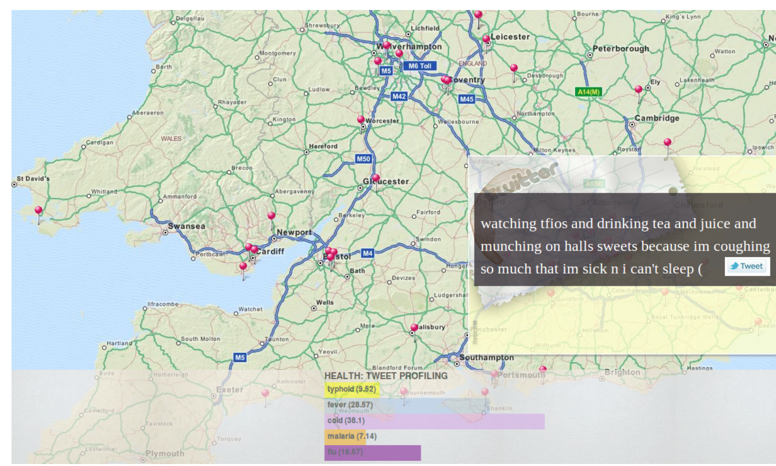
  Ailment –Ear Infection
- *Ending my night with an unused luke Bryan ticket, an allergic reaction to a dog, and 10 other misfortunes at least I have my health !!!!!!*

  Ailment – Overweight

An exploration dashboard

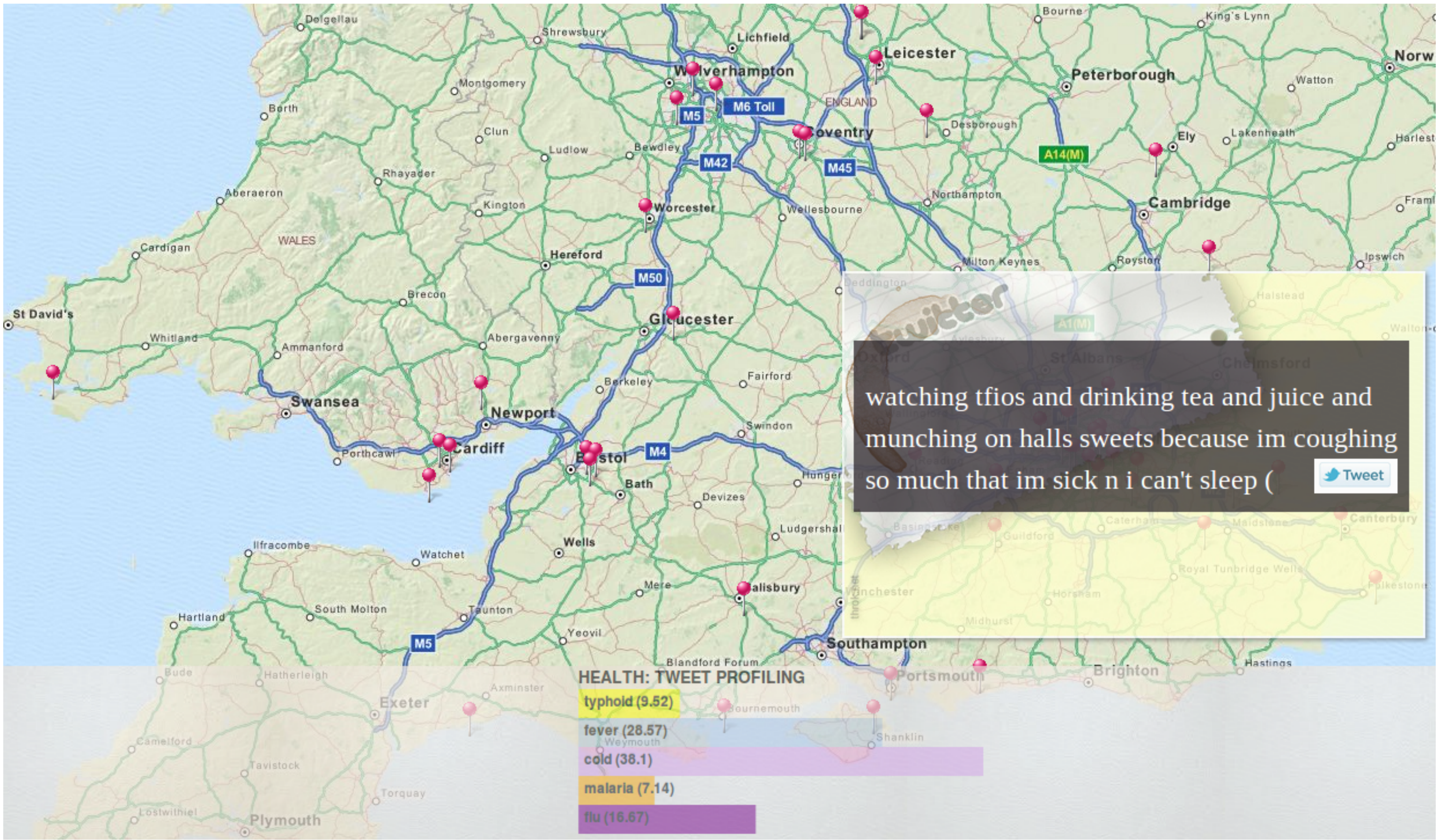






Prospectom 2014

From London



Prospectom 2014

Analytics: drug use

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

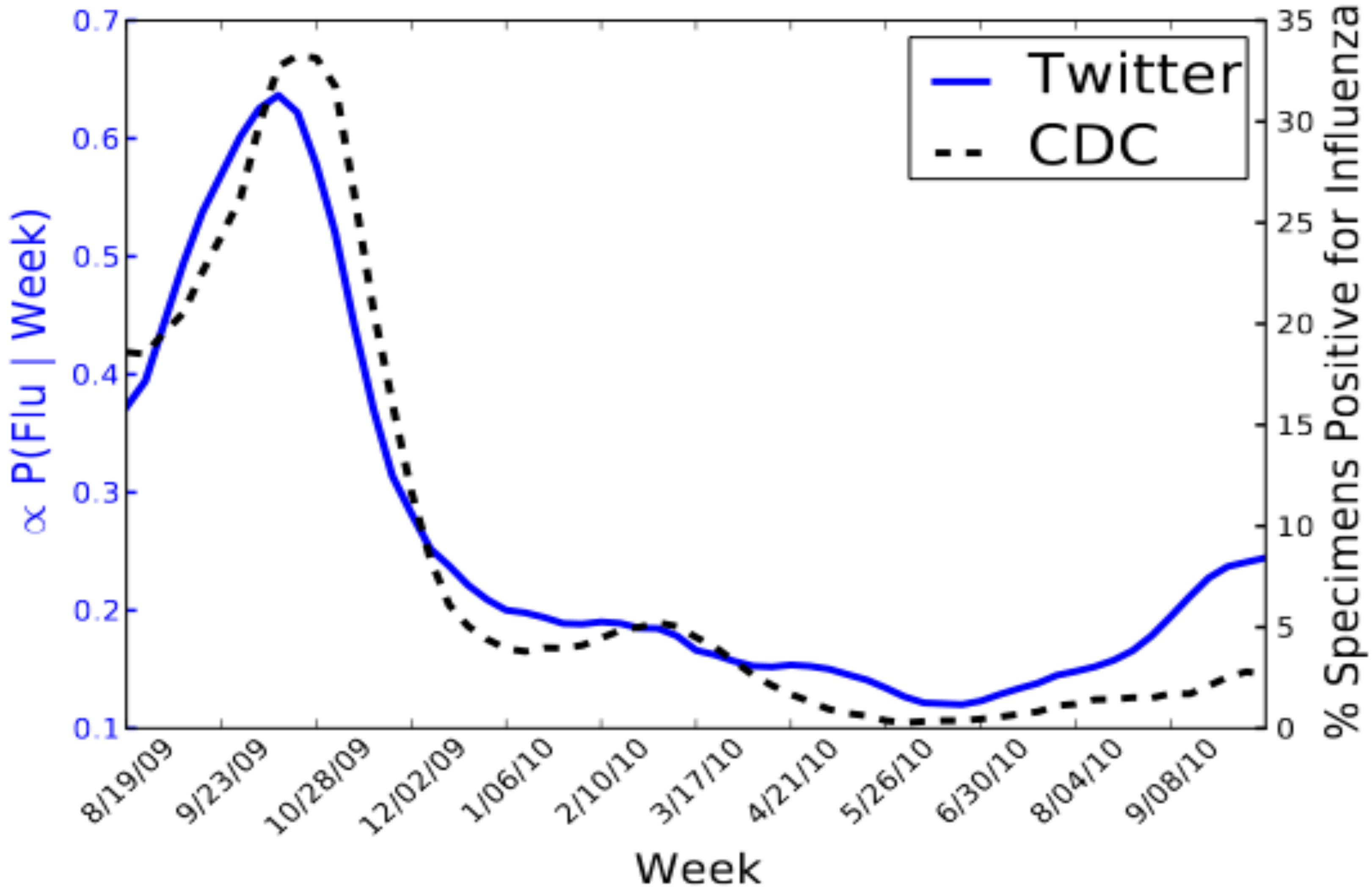
Word	#	Ent.	Most Common Ailments
Pain Relief Medication			
tylenol	1807	1.57	HA (39%), IN (30%), Cold (9%)
ibuprofen	1125	1.54	HA (37%), DN (21%), Aches (17%)
advil	1093	1.08	HA (61%), Cold (6%), DN (5%)
aspirin	885	1.04	HA (69%), IN (10%), Aches (10%)
vicodin	505	1.33	DN (61%), Injuries (11%), HA (10%)
codeine	406	1.94	Cold (25%), DN (19%), HA (17%)
morphine	206	1.17	DN (59%), Infection (22%), Aches (9%)
aleve	183	1.10	HA (62%), IN (15%), DN (14%)
Allergy Medication			
benadryl	871	1.24	Allergies (64%), Skin (13%), IN (12%)
claritin	417	0.54	Allergies (88%), HA (5%)
zyrtec	386	0.49	Allergies (90%)
sudafed	298	1.61	Allergies (39%), Cold (21%), HA (20%)

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*



Crowd4U

Your Web browser

What is a good cafe in Tokyo?

Tasks and Data

Contributors in Academia

Crowd4U Terminals

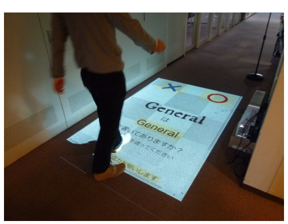
Which are good cafes in Grenoble?

☑ A ☐ B ☑ C

Task Pool

Crowd4U

Requesters

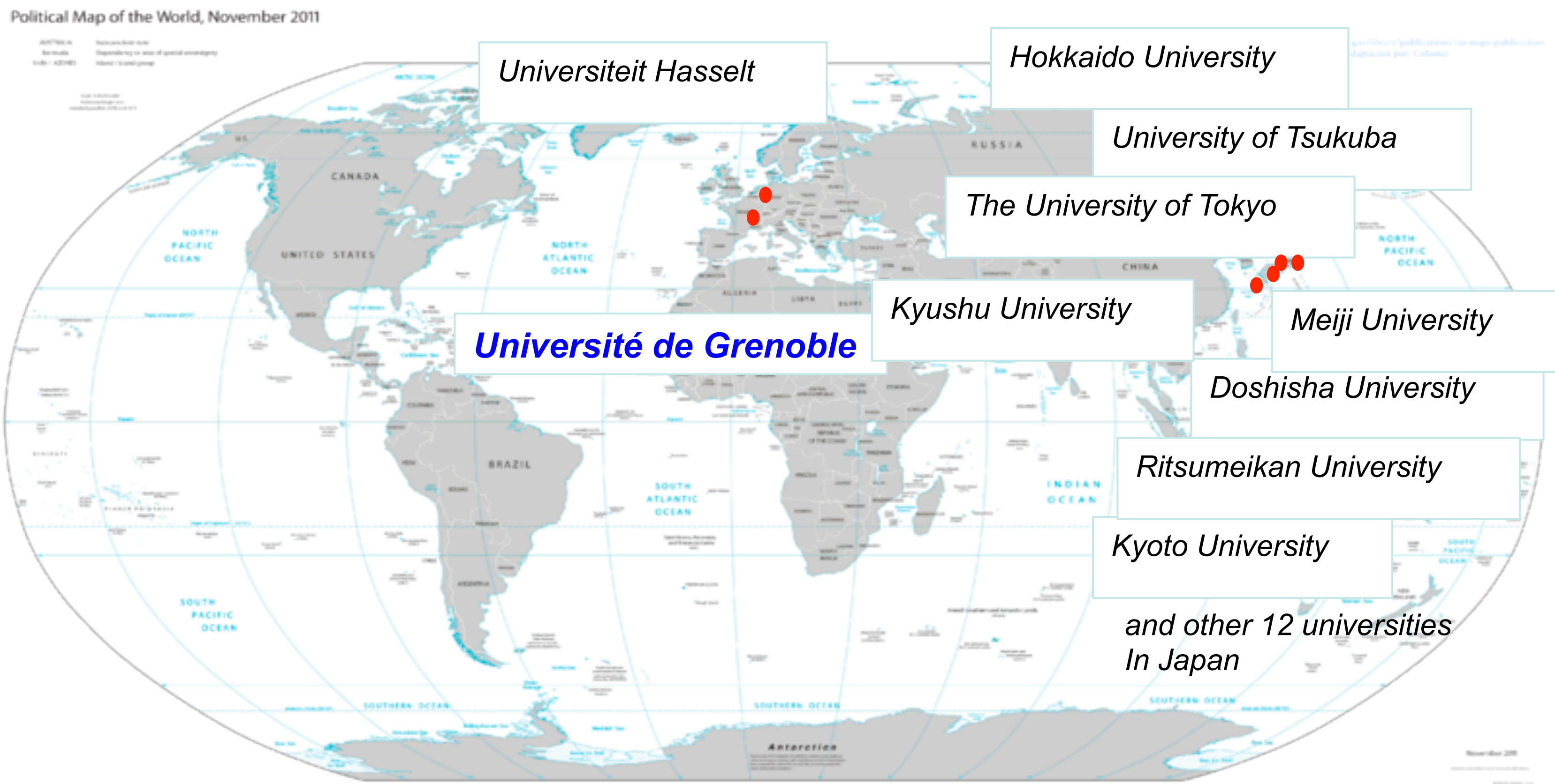


Floor

Is A is better than B?

Yes No

Crowd4U deployment



>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