Towards Personalized Medicine Health: Building Predictive Models for “Segments of One”

Joydeep Ghosh
Schlumberger Centennial Chaired Professor,
Dept of Electrical and Computer Engineering,
University of Texas at Austin
<table>
<thead>
<tr>
<th>Customer Segment</th>
<th>NPV Per Customer</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convenience Seekers</td>
<td>$</td>
<td>• Value convenience in delivery/ordering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• High income</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Long relationship, large referrals</td>
</tr>
<tr>
<td>Brand Buyers</td>
<td>$</td>
<td>• Brand buyers, not price sensitive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Highest income, more often male</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Expensive to acquire, but buy most initially and refer more</td>
</tr>
<tr>
<td>Casual Buyers</td>
<td>$</td>
<td>• Not concerned with perishables or delivery time windows</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Small spending growth</td>
</tr>
<tr>
<td>Relationship Seekers</td>
<td>$</td>
<td>• Influenced by retailer brand, suggestions, and promotions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Low income</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Small spending growth/referral</td>
</tr>
<tr>
<td>Bargain Hunters</td>
<td>$</td>
<td>• Price is primary and perishables are not important</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Low income</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Small purchases</td>
</tr>
</tbody>
</table>

*Source: Bain/Mainspring Online Retailing Survey*
Mass Customization: Segments of One

- Rich population data + high-end analytics
- Pandora
- Yahoo/Facebook/Google
- Siri/Cortona/
- ...........
- So, where is the “doctor”? 

Joydeep Ghosh  UT Austin
Collaborative Email-Spam Filtering with the Hashing-Trick

Joshua Attenberg
Polytechnic Institute of NYU
Five MetroTech Center
Brooklyn NY, 11201
josh@cis.poly.edu

Kilian Weinberger, Anirban Dasgupta
Alex Smola, Martin Zinkevich
Yahoo! Research
4401 Great America Pkwy.
Santa Clara, CA 95054 USA
{kilian, anirban, smola, maz}@yahoo-inc.com

ABSTRACT
This paper delves into a recently proposed technique for collaborative spam filtering [7] that facilitates personalization with finite-sized memory guarantees. In large scale open membership email systems most users do not label enough messages for an individual local classifier to be effective, while the data is too noisy to be used for a global filter across all users. Our hybrid global/individual classifier is particularly effective at absorbing the influence of users who label emails very differently from the general public – because of strange taste or malicious intent. We can accomplish this while still providing sufficient classifier quality to users with a dictionary that can use up a significant portion of a servers’ memory. Furthermore, the dynamic nature of language in both spam and not-spam requires that the dictionary adapt to new words, essentially growing over time. The hashing-trick overcomes this obstacle by rendering the dictionary unnecessary; words are hashed directly to indices. While hash collisions may result in several words being mapped into the same index, and therefore being falsely considered identical, such collisions rarely affect classification results. The decision whether an email is spam and not-spam is rarely based on a single word but on a combination of many slightly indicative words. In fact, without a dictionary much more
Guidance for Personalized Health

• Patient as Customer/Consumer
  – Patient User interface (smartphone)
  – Active user participation

• Data Driven:
  – Diverse Data Sources
  – Data sharing & Privacy issues
    • Attitudinal changes
  – New ML algorithms
    • Collaborative learning to Rank (LeTOR)
    • Semi-supervised, online learning

• Would these algos be admissible under Evidence Based Medicine guidelines? HIPAA?

Joydeep Ghosh  UT Austin
Towards Cancer Care Segments of One

May 1998: Slamon announces results of Herceptin trails for Her-2 positive breast cancer patients.

Joydeep Ghosh  UT Austin
200+ oncogenes: promote tumor formation, can be targeted; 
+ changes to tumor suppressor genes e.g. P53 
+ cancer sniffer 
+....
Healthcare ≠ Cancer Care

• Chronic conditions
• Care coordination and engagement
• Cost-aware personalization of treatments
• ...
• How about 140/90 BP levels
Outline (The Stars Align)

• Ingredients of a Perfect Storm
  – Cost squeeze
  – Demand for better care
  – Regulatory changes
  – Technology (smartphones, cloud, sensors,...)

• Some early successes
  – Research
  – organizations
  – Startups

• The Future
Metrics for Healthcare Quality

Cost
Quality
Coverage
Choice
Towards Digital Health: Regulations, Attitudes, Technology

- **Learning Healthcare system** reports from IOM (2007-)
  - Push toward electronic records + HIEs + meaningful use

- HiTech Act of 2009
- PPACA ("Obamacare")

- Grass-roots, from large employers (Intel...) to empowered patients
- Cloud, smart wearables,...
MegaTrends

• doctor centric \(\rightarrow\) patient centric
  – Medicare: average patient: 6-7 physicians,
    • (14 for those with 4+ chronic conditions)

• Billing Centric Medical Records \(\rightarrow\) records for proactive care

• Passive Patient \(\rightarrow\) activist consumer
The Coming Revolution..

• “In the next 10 years, data science and software will do more for medicine than all of the biological sciences together”
  —Vinod Khosla, 2013
Getting to “Segments of One”

- Data/Knowledge
- Platforms for sharing
- Predictive Models
- Personal Technology
- Market Drivers
- ..... 

- Evidence Based Medicine is so 20th century
Data and Knowledge

- Sources
- Sharing: Privacy vs. Utility
- Actionable Insights
Joydeep Ghosh  UT Austin
## BioBanks

<table>
<thead>
<tr>
<th>Biobank</th>
<th>HPO system size</th>
<th>Current Biobank Size</th>
<th>Recruitment method</th>
<th>Time to achieve size (during active enrollment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Million Veteran Project</td>
<td>6 million</td>
<td>400,000</td>
<td>Mailed veterans info about MVP and enrolled at visit</td>
<td>4 years in 54 sites</td>
</tr>
<tr>
<td>Kaiser Permanente</td>
<td>10.1 million</td>
<td>245,000 (goal of 500,000)</td>
<td>Mailed consent and mailed saliva sample (N = 189,500); electronic or in-person consent and blood samples (N=50,000)</td>
<td>3.5 years using direct mail to 2 million</td>
</tr>
<tr>
<td>Partners Healthcare Biobank</td>
<td>6 million</td>
<td>&gt;30,000</td>
<td>In-person at outpatient visits and inpatient floors; Electronic consent via emails using patient portal</td>
<td>5 years since launch: 2 year pilot study; 3 years via in person recruitment; eConsent for past 1 year; current rate is 1100/month</td>
</tr>
<tr>
<td>Geisinger MyCode</td>
<td>1.3 million with an EHR</td>
<td>&gt;86,000</td>
<td>In-person during routine outpatient</td>
<td>10 years; however, current rate is 1000/</td>
</tr>
</tbody>
</table>
MIMIC-II: ICU data for critical care

- (courtesy Mornin’ Feng, MIT)
Harvard Open Data Effort

Sharing Personal Genomes

The Personal Genome Project was founded in 2005 and is dedicated to creating public genome, health, and trait data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices—our approach is to invite willing participants to publicly share their personal data for the greater good.

Learn more

Participation

Donating your genome and health data to science is a great way to enable advances in understanding human genetics, biology, and health. We seek volunteers willing to donate diverse personal information to become a public resource.

Learn about participating

Open Data

Open data is a critical component of the scientific method, but genomes are both identifiable and predictive. As a result, many studies choose to withhold data from participants and restrict access to researchers. The PGP’s public data is a common ground to collaborate and improve our understanding of genomes.

Use PGP data

Global Network

We are a member of the Global Network of Personal Genome Projects. Since the Personal Genome Project was launched at Harvard Medical School in 2005, the network has grown to include researchers at many leading institutions around the globe.

Find out about the network

Joydeep Ghosh  UT Austin
PatientsLikeMe; Smart Patients

Specially effective for rare diseases, getting new insights

https://crohnology.com/

What if we could learn from the collective experience of patients everywhere?

<table>
<thead>
<tr>
<th>Top Medications</th>
<th>Top Diets</th>
<th>Top Supplements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remicade</td>
<td>No Beer</td>
<td>Vitamin B12</td>
</tr>
<tr>
<td>Prednisone</td>
<td>No Spicy Food</td>
<td>Probiotics</td>
</tr>
<tr>
<td>Flagyl</td>
<td>No Dairy</td>
<td>Vitamin D</td>
</tr>
</tbody>
</table>

1998 people
3546 people
1776 people
1681 people
1342 people
1384 people
1741 people
2170 people
2204 people

Michael Weiss
West Orange, NJ
Multivitamin
Sharing Data/Methods at Scale

• Common Data Models
• Common Tools
• Assessment
OHDSI: Data Mappers + Analyzers (from Patrick Ryan)

OHDSI tools built to help:
- WhiteRabbit: profile your source data
- RabbitInAHat: map your source structure to CDM tables and fields
- ATHENA: standardized vocabularies for all CDM domains
- Usagi: map your source codes to CDM vocabulary
- CDM: DDL, index, constraints for Oracle, SQL Server, PostgreSQL; Vocabulary tables with loading scripts
- ACHILLES: profile your CDM data; review data quality assessment; explore population-level summaries

OHDSI Forums:
Public discussions for OMOP CDM Implementers/developers

Joydeep Ghosh - UT Austin
http://github.com/OHDSI
HIPAA

When HIPAA became law, there were few EHRs, no patient portals, no smartphones or “ubiquitous connectivity”, Facebook was 8 years away.....

“While consumers vehemently agree they should be in control of health data access, the majority are willing to share data for personal and public health, along with discounts.” – Rock Health, Oct 15, n=4017

My Prediction: HIPAA will go away soon!
PCAST REPORT OF 2010

REPORT TO THE PRESIDENT
REALIZING THE FULL POTENTIAL OF
HEALTH INFORMATION TECHNOLOGY
TO IMPROVE HEALTHCARE
FOR AMERICANS:
THE PATH FORWARD

Executive Office of the President
President’s Council of Advisors
on Science and Technology

December 2010

~100 pages and readable!
SOME KEY PCAST RECONS

• Robust exchange of health information
  (using universal exchange infrastructure/language)

• Data Centric
  – Records at both patient (meta-tagged) and aggregated levels but not centralized
    (beyond HIPAA)

• 3rd party mgt: cloud services, health management/analytics
  – (Near) real-time data acquisition
  – analysis for comparative effectiveness, personalized medicine

• Save $1 TRILLION in 10 years.
Perturbed Gibbs Sampler (Park & Ghosh, ‘14): Generating “Realistic But Not Real” Data

- **Disintegrate**: Use a hash function for getting compressed conditional distributions
- **Inject Noise**: further smoothes data to achieve desired privacy level
  - Differential privacy
  - L-diversity
- **Synthesize**: Transform a (random) seed to a synthetic sample by iteratively sampling each feature from the statistical building blocks.
PeGS
SUMMARY

• PeGS synthesizes data that
  – are realistic but not real; similar statistical properties with the original data
  – guarantee the privacy of data with respect to the choice of privacy metrics
• PeGS requires almost no parameter tuning
• Domain knowledge is minimally needed
• Also, PeGS is parallelizable and scalable
• Thus, easier and safer data sharing
EXPERIMENTAL RESULTS ON L-DIVERSITY
(USING CMS Inpatient Claims Public Use Files)

\[ y = \{\text{Payment}\}, \quad X = \{\text{Age, Gender, Drg, \ldots}\} \]

\[ \beta^* = \arg \min_{\beta} \| y_{\text{synth}} - X_{\text{synth}} \beta \|_2^2 + \lambda \| \beta \|_2^2 \]

\[ \text{MSE} = \| y_{\text{orig}} - X_{\text{orig}} \beta \|_2^2 \]
Towards Personalized Health

• Personalized =
  – Predictive
  – Prescriptive & Proactive
  – Precise
  – Private
  – Pervasive (and continually learning)
Predictive Modeling

• **High-throughput Phenotype Generation from Heterogeneous Health Records**

• **Individualized Predictions from Longitudinal Data**
**Non-standardization**

Alternative definitions for Acute Liver Injury

- **D**: Occurrence of at least one diagnosis code

- **D+P**: Occurrence of at least one diagnosis code AND (diagnostic procedure \( \leq 30 \text{d before} \) OR treatment procedure \( \geq 60 \text{d after} \))

- **D+P+L**: Occurrence of at least one diagnosis code AND (diagnostic procedure \( \leq 30 \text{d before} \) OR treatment procedure \( \geq 60 \text{d after} \)) AND laboratory results indicative of Hy's law: (ALT \( \geq 3 \times \text{ULN} \) OR AST \( \geq 3 \times \text{ULN} \)) AND Bilirubin \( \geq 2 \times \text{ULN} \) within 7 days

- **L**: Laboratory results indicative of Hy's law: (ALT \( \geq 3 \times \text{ULN} \) OR AST \( \geq 3 \times \text{ULN} \)) AND Bilirubin \( \geq 2 \times \text{ULN} \) within 7 days
PHENOTYPING VIA TENSOR FACTORIZATION

Nonnegative tensor factorization model used to discover phenotypes (reveal patient clusters)

Elements represent number of times patient is prescribed medication to treat a disease

LIMESTONE: CANDIDATE PHENOTYPE

- Clinical characteristics are nonzero elements
- Each element represents conditional probability given the phenotype and mode
Personalization via Soft Phenotyping

• Each patient is a point in “soft phenotype” space
  – Clinically relevant semantics
• Longitudinal data: trajectories in this space
• Metric Learning: customize goal-specific patient similarity
Monitoring of ICU Patients

Using latent variable/state switching models with stochastic volatility

- e.g. Predict which patients who will experience cardiac arrest in the next 4 hours using stochastic volatility models
Model learns subtype trajectories while estimating sources of variability across individuals at multiple scales.

Saria, Goldenberg, IEEE Intelligent Systems (Forthcoming)
Digital Health Startups!

**FOLLOWING THE INCENTIVES**

*Top six trends of digital health (Q1-Q3 2014)*

- **$381M**
  - **ANALYTICS AND BIG DATA**
  - Data aggregation and analysis to support a wide range of healthcare use cases

- **$280M**
  - **DIGITAL MEDICAL DEVICES**
  - Software/hardware designed to treat a specific disease or condition

- **$238M**
  - **HEALTHCARE CONSUMER ENGAGEMENT**
  - Consumer tools for the purchasing of healthcare services or health insurance (B2B and B2C)

- **$223M**
  - **PAYER ADMINISTRATION**
  - Management and administration tools for payers

- **$195M**
  - **POPULATION HEALTH MANAGEMENT**
  - Comprehensive platforms for managing the health of populations under the shift to risk-based payment models

- **$172M**
  - **TELEMEDICINE**
  - Delivery of healthcare services through non-physical means (e.g., telephone, digital imaging, videoconferencing)

*Note: Flatiron Health was recategorized from Personalized Medicine to Analytics and Big Data*
Opportunities include..


What opportunities or needs can digital health companies help payers address?

We are looking at those that focus on intervention, quality of care, and communication in the following areas:

• **1) Intervention:** Population Health Management: Tools that help identify the highest need patients, who can be reached by a clinician and who are amenable to changing behaviors with specific actions. Resulting in higher quality and lower cost.

• **2) Quality of Care:** Point of Care Solutions: Tools that help patients and doctors make the best decisions possible at the point of care and are a) clinically validated, b) account for patient preference, and c) result in better care (higher quality or lower cost).

• **3) Communication:** Patient/Provider Engagement: Tools that help patients and providers bridge a disconnected and fractured system to either a) communicate better across the system or b) increase self-care/self-management capabilities and success.
Results

User Interface: Cost and Outcome Prediction

- **Skilled Nursing Facility (64%)**
  - This cohort costs on average $30,132 over 90 days.
  - All-Cause Readmission Rates:
    - 9.6% over 30 days
    - 22.31% over 90 days

- **Home Health Care (24%)**
  - This cohort costs on average $30,842 over 90 days.
  - All-Cause Readmission Rates:
    - 21.68% over 30 days
    - 38.46% over 90 days

- **Inpatient Rehab Care (13%)**
  - This cohort costs on average $32,466 over 90 days.
  - All-Cause Readmission Rates:
    - 1.81% over 30 days
    - 1.81% over 90 days

---

Joydeep Ghosh, UT Austin
Results

User Interface: Risk Trajectories
Wearables and Embedded Sensors

- Sensors for DNA assays, micro-fluidics, breath, saliva, blood,..
- iPhone 6s keynote: Airstrip
- AliveCor Mobile ECG; Scanadu
Tyto

Joydeep Ghosh  UT Austin
What can personal monitoring anticipate?

(from Eric Topol, “The patient will see you now”)

<table>
<thead>
<tr>
<th>Condition</th>
<th>S</th>
<th>Metrics</th>
<th>Labs</th>
<th>Imaging</th>
<th>Actionability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Failure</td>
<td>W</td>
<td>Cardiac output, stroke volume, fluid status, vital signs, weight, sleep</td>
<td>BNP, Kidney tests</td>
<td>US</td>
<td>Fluid offload, heart unload, medication adherence</td>
</tr>
<tr>
<td>Depression</td>
<td>W</td>
<td>Voice, vital signs, breathing, communication, activity, facial expression, sleep, HRV, HRR, GSR</td>
<td>Neuro-hormones</td>
<td>EEG</td>
<td>Counseling, anti-depressant, medication adherence</td>
</tr>
<tr>
<td>Asthma</td>
<td>W</td>
<td>FEV₁, air quality, inhaler, GPS, allergens, vital signs</td>
<td>NO Microbiome</td>
<td>--</td>
<td>Preventive Rx of medications, avoidance of trigger</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>W</td>
<td>HRV, GSR, sleep, activity, vital signs</td>
<td>--</td>
<td>EEG</td>
<td>Medication, avoidance of vulnerability</td>
</tr>
<tr>
<td>Autoimmune Diseases</td>
<td>E</td>
<td>Sequence of B and T cell repertoires in blood cells</td>
<td>Microbiome</td>
<td>--</td>
<td>Immunomodulation</td>
</tr>
<tr>
<td>Cancer</td>
<td>E</td>
<td>Presence of ctDNA and CTCs</td>
<td>--</td>
<td>--</td>
<td>Sequencing to determine whether/what Rx is necessary</td>
</tr>
<tr>
<td>Heart Attack</td>
<td>E</td>
<td>Presence of ceDNA or RNA</td>
<td>CT angiogram</td>
<td>--</td>
<td>Anti-clotting medication</td>
</tr>
</tbody>
</table>
A New Initiative on Precision Medicine

Francis S. Collins, M.D., Ph.D., and Harold Varmus, M.D.

"Tonight, I’m launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier."

— President Barack Obama, State of the Union Address, January 20, 2015

The proposed initiative has two main components: a near-term focus on cancers and a longer-term aim to generate knowledge applicable to the whole range of health and disease. Both components are now within our reach because of advances in basic research, including molecular biology, genomics, and bioinformatics. Furthermore, the initiative
• Initial focus on Cancer and in 1 mil. Cohort construction
  – $70 m/215m for NCI: use genomics to identify and target molecular vulnerabilities of individual cancers
    http://deainfo.nci.nih.gov/advisory/ncab/0615/05%20Doroshow.pdf

• Next 4 slides selected from a presentation by Dr. Francis Collins, Director, NIH
  Feb 2015

• http://www.nih.gov/precisionmedicine/presentations.htm
Precision Medicine Initiative

- **National Research Cohort**
  - >1 million U.S. volunteers
  - Numerous existing cohorts (many funded by NIH)
  - New volunteers

- Participants will be centrally involved in design and implementation of the cohort
- They will be able to share genomic data, lifestyle information, biological samples – all linked to their electronic health records
Precision Medicine Initiative

The National Research Cohort will:

- Provide scientists with a ready platform for:
  - Observational studies of drugs and devices
  - Tests of wearable sensors for monitoring health
  - More rigorous interventional studies
- Forge new model for scientific research that emphasizes engaged participants and open, responsible data sharing with privacy protections
Other Diseases: What Success Might Look Like

50-year-old woman with type 2 diabetes visits her doctor

- **Future: + 5 years**
  - Receives word from her doctor about a new drug based upon improved molecular understanding of type 2 diabetes
  - When she enters drug’s name into her smartphone’s Rx app, her genomic data show she’ll metabolize the drug slowly
    - Her doctor alters the dose accordingly
Vision: even BP, Cholesterol, HbA1c levels .... will get personalized.
The Role of Physicians in the Era of Predictive Analytics

Every day, more information becomes available about factors that affect the risk of a clinical event. Based on the conventional risk factors, participants were first classified into 3 levels of cardiovascular risk:

From JAMA, Jul 7, 2015

Conclusions

Predictive analytics can improve clinical care by providing general recommendations for populations that can be incorporated into clinical guidelines. Predictive algorithms are an essential component of guideline recommendations. However, because predictive models imperfectly explain clinical outcomes, they do not estimate individual risk very well even when they accurately explain group risk. Consequently, these models cannot replace the physician in the process of care. Physicians are responsible for assessing an individual’s constellation of problems and risk factors and then selecting with the patient the appropriate treatment for that individual. Physicians must understand the limitations as well as the strengths of the evidence as it applies to individual patients and must recognize additional factors not included in prediction models may alter risks or benefits.

Joydeep Ghosh  UT Austin
An Exciting Time

• Evidence Based Medicine
  – Needs turbocharging

• Learning Healthcare System
  – Right directions but too timid

• 20% doctor included...
  (Khosla, the “technology optimist)
  – Achievable!

  We can do better!

Health Informatics front and center
## Top Match

<table>
<thead>
<tr>
<th>PheWAS code</th>
<th>Description</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>V₇</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>411.4</td>
<td>Coronary atherosclerosis</td>
<td>0.31</td>
</tr>
<tr>
<td>411.2</td>
<td>Myocardial infarction</td>
<td>0.22</td>
</tr>
<tr>
<td>496</td>
<td>Chronic airway obstruction</td>
<td>0.10</td>
</tr>
<tr>
<td>415</td>
<td>Pulmonary heart disease</td>
<td>0.08</td>
</tr>
<tr>
<td>418</td>
<td>Nonspecific chest pain</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>N₁₃</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>411.4</td>
<td>Coronary atherosclerosis</td>
<td>0.56</td>
</tr>
<tr>
<td>411.2</td>
<td>Myocardial infarction</td>
<td>0.15</td>
</tr>
<tr>
<td>272.1</td>
<td>Hyperlipidemia</td>
<td>0.11</td>
</tr>
<tr>
<td>443.9</td>
<td>Peripheral arterial disease</td>
<td>0.02</td>
</tr>
<tr>
<td>418</td>
<td>Nonspecific chest pain</td>
<td>0.02</td>
</tr>
</tbody>
</table>