The Value of Variety

*Methods, strategies and architectures to deal with the most intriguing “V” of biomedical big data*

Riccardo Bellazzi, PhD

Centre for Health Technologies and Dept. Electrical, Computer and Biomedical Engineering
University of Pavia

ICS Maugeri Hospital, Pavia
Big Data pillars

Volume
- Data at Rest
  - Terabytes to exabytes of existing data to process

Velocity
- Data in Motion
  - Streaming data, milliseconds to seconds to respond

Variety
- Data in Many Forms
  - Structured, unstructured, text, multimedia

Veracity
- Data in Doubt
  - Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Variability

Value

From: N. Peek, UManchester
Variety – the spice of life

"Variety's the very spice of life, that gives it all it's flavour."
William Cowper, The task, 1785
Variety, Not Volume, Is Driving Big Data Initiatives

For large corporations, data variety trumps volume when looking for insights.
Big data analytics to improve cardiovascular care: promise and challenges

John S. Rumsfeld, Karen E. Joynt & Thomas M. Maddox
Variety is the spice of biomedical big data

Exposome

Genome

Phenome

By F. Martin Sanchez – Medinfo 2015
Are we ready for the variety challenge?
Care informs research, research informs care

An informatics research agenda to support precision medicine: seven key areas


DOI: http://dx.doi.org/10.1093/jamia/ocv213  ocv213 First published online: 23 April 2016
From care to research

Data

HIS

Anonymized data

Match IDs

Researcher

Anonymized samples

Biobank

Clinical patient management

Samples

Laboratory

Patient
Supporting biomedical research

EHR4CR
Electronic Health Records for Clinical Research

EMIF — European Medical Information Framework
One Platform for data discovery, assessment and (re)use

i~HD

Enriching knowledge and enhancing care through health data

The European Institute For Innovation Through Health Data
An open-source software program developed at the NIH research center "i2b2"

i2b2 Academic Users' Group: a wide community that contributes to i2b2 development
I2b2-based Pavia projects

- A number of initiatives to support clinical research informatics based on data integration (> 100,000 patients)
- All Pavia research hospitals involved
  - Arrhythmogenic diseases
  - Breast cancer
  - Onco-Hematology
  - Diabetes
  - Neurology
- Hospitals in Bergamo and Milan
4th i2b2 EU AUG Meeting

Sept 15th 14:30 - Sept 16th 12:30

Pavia – ITALY

Invited Talks by
Paul Avillach
Shawn Murphy

further info http://cht.unipv.it

Sponsored by: i2b2 TriNetX BIOMERIS
A Network-Based Data Integration Approach to Support Drug Repurposing and Multi-Target Therapies in Triple Negative Breast Cancer

Francesca Vitali, Laurie D. Cohen, Andrea Demartini, Angela Amato, Vincenzo Eterno, Alberto Zambelli, Riccardo Bellazzi
Temporal clusters of breast cancer patients

A Dagliati, L Sacchi, A Zambelli, V Tibollo, L Pavesi, JH Holmes, R Bellazzi
Temporal Electronic phenotyping by mining Careflows of Breast Cancer Patients 2016, Journal of Biomedical Informatics
Benefits of big data technologies

- Flexible data storage
  - SQL and NoSQL
- Easier parallel computing
  - Hadoop/Sparc, Map-Reduce
- More efficient virtualization
  - Docker
- Improved analytics platforms
  - R, Python
- Improved analytics and data processing methods
  - NLP, embedding, projections, factorization

Source: Ventana Research The Challenge of Big Data Benchmark Research
Open and commercial Ecosystems
NIH Big Data to Knowledge (BD2K) Program

- a massively scalable toolkit to enable large, multi-center Patient-centered Information Commons (PIC) at local, regional, and national scale
- alignment of all available biomedical data per individual.

1. Count Everything: Integrating Clinical, Genomics and mHealth APIs across the BD2K Program
2. Global Rare Diseases Registry
3. Integrating Data and Toolkits Across Institutional Boundaries
4. Sync for Science (S4S)
Big data architectures and omics

Journal of Biomedical Informatics
Volume 64, December 2016, Pages 288–295

Special Communication

Evaluation of relational and NoSQL database architectures to manage genomic annotations

Wade L. Schulz, MD PhD, Brent G. Nelson, MD, Donn K. Felker, MD, Thomas J.S. Durant, MD, Richard Torres, MD MS
Big data architectures to extend i2b2 - NoSQL-NGS

Gabetta et al, BMC Bioinformatics, 2016, Murphy et al PLOS One 2017
**i2b2 – BigQ NGS 1.0**

**Filters:** specific gene, exomic function, Polyphen2 score, Frequency, …

**Performance:** 55 exomes – 20k variants = 1M ➔ specific gene = 1 sec
Marco Ramoni (TBI) Distinguished Paper Award
Joel Saltz, Stony Brook University "Towards Generation, Management, and Exploration of Combined Radiomics and Pathomics Datasets for Cancer Research"
Human Brain Project

From Data to models of the Brain

Neuroscience
  Unify

Medicine
  Classify

Computing
  Produce
Multiscale disease signatures

1. Data Federation
   - Behavior
   - Neuro-psychology
   - Brain imaging

2. Data Integration
   - Genetics
   - Proteomics
   - Knowledge
   - Simulation

3. Data Mining

4. Causal Modeling

5. Biological Signature of Disease

- Disease Definition
- Pharmacology
- Clinical Trial

Disease Signature Rule Space
- Clinical features
- Causal features
- Physical features

Disease Space
- Phenomenology
- Memory
- Learning
- Behavior

Human Brain Project

Centre for Health Technologies
University of Pavia
Based on Federated data

MIP PLATFORM
(GUI + iPython Notebook)
mip.humanbrainproject.eu

MIP CENTRAL SERVICES
(Computational workflow, MIP Ontology, API to connect HPC – GRID – Hospital Resources)

Federated Queries/Computation
(EXAREME/Woken)

HBP Platforms
- Reference Neuroimaging
- Genetic data
- Disease Models
- Disease Atlases

HPC Platforms
- Machine Learning
- Feature Engineering
- Disease Models

Hospital node 1
- DATA FACTORY
- Pre-Processing
- Normalization
- Segmentation
- Fiber tracking
- Etc...

ALGORITHM FACTORY
- SVM
- KNN
- CART
- RF
- Bayesian
- TSNE

Hospital node 2
- DATA FACTORY
- Pre-Processing
- Normalization
- Segmentation
- Fiber tracking
- Etc...

ALGORITHM FACTORY
- SVM
- KNN
- CART
- RF
- Bayesian
- TSNE

LDSM

Brown SA¹.
Types data integration

Early integration

Intermediate integration

Late integration

Input dataset

Processing

D1

D2

D3

Processing

Category 1

Category 2

Category 3

Category 4

Category 5

Series 1

Series 2

Series 3
Flat, deep or knowledge-based?
Word embeddings

- A neural word embedding represents a word with numbers.

\[ W(“cat”) = (0.2, -0.4, 0.7, ...) \]
\[ W(“mat”) = (0.0, 0.6, -0.1, ...) \]

https://deeplearning4j.org/word2vec#intro
Image embedding

NSN - Not Competent for development    SN - Competent for development

Oocytes
Image embedding on the shelves


105 images of cell development
Clustering and RF classification

<table>
<thead>
<tr>
<th></th>
<th>NSN</th>
<th>SN</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>32</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>48</td>
<td>50</td>
</tr>
<tr>
<td>Σ</td>
<td>34</td>
<td>51</td>
<td>85</td>
</tr>
</tbody>
</table>
Multimodal Deep Belief Networks

- Multimodal Representation
  - Visible Layer
    - miRNA
  - Hidden Layer
  - Visible Layer
    - GE
  - Visible Layer
    - MU

Unimodal models

Top-level DBN

G. Gerard - University of Pavia - Cancer Data Analysis with a Multimodal Deep Belief Network
Multimodal Representation Layer

G. Gerard - University of Pavia - Cancer Data Analysis with a Multimodal Deep Belief Network
Classification Results

- Integrative analysis of 'multi-dimensional genomic data' obtained TCGA project.
- Unsupervised clustering of patients in groups based on their multi-platform genomic profiles.
- **Data**: 200 AML miRNA expression, Gene expression and DNA mutations from TCGA Project.
Flat, deep or knowledge-based?
Recommender Systems

Working on sparse data

Winners on Netflix price

Adapted from Zitnik and Zupan, AIME 2015
### Movies

<table>
<thead>
<tr>
<th>Seven Samurai</th>
<th>Memento</th>
<th>The Pianist</th>
<th>Big Fish</th>
<th>Pulp Fiction</th>
<th>A Beautiful Mind</th>
<th>The Hobbit</th>
<th>V for Vendetta</th>
<th>La Haine</th>
<th>Rain Man</th>
<th>Into the Wild</th>
<th>Fight Club</th>
<th>Star Wars</th>
</tr>
</thead>
</table>

### Users

- Alenka
- Teja
- Miha
- Luka
- Jasna
- Gregor
- Tina
- Urban
- Ana
- Uroš
- Andreja
- Andraž
- Sanja
- Klemen
- Petra
- Primož

*Adapted from Zitnik and Zupan, AIME 2015*
Matrix Factorization


Adapted from Zitnik and Zupan, AIME 2015
Prediction

<table>
<thead>
<tr>
<th>Seven Samurai</th>
<th>Memento</th>
<th>The Pianist</th>
<th>Big Fish</th>
<th>The Godfather</th>
<th>Pulp Fiction</th>
<th>A Beautiful Mind</th>
<th>The Hobbit</th>
<th>V for Vendetta</th>
<th>La Haine</th>
<th>Rain Man</th>
<th>Into the Wild</th>
<th>Fight Club</th>
<th>Star Wars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alenka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teja</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jasna</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uroš</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andreja</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andraž</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanja</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klemen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petra</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primož</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mustafa & Morris, 2012
Moreau & Tranchevent, 2012

Adapted from Zitnik and Zupan, AIME 2015
Tensor Factorization for Precision Medicine in Heart Failure with Preserved Ejection Fraction

Authors

Yuan Luo, Faraz S. Ahmad, Sanjiv J. Shah
Tensor factorization
Extension of matrix trifactorization

Adapted from Zitnik and Zupan, AIME 2015
Tri-factorization algorithm

\[ \text{obj1} \quad \text{obj2} \quad \text{obj3} \]

\[ R + \theta \quad \text{sparse} \]

\[ G_1 \quad G_2 \quad G_3 \]

\[ S \]
Tri-factorization algorithm

\[ R + \theta \]

dense

\[ obj1 \quad obj2 \quad obj3 \]

\[ obj1 \]

\[ obj2 \]

\[ obj3 \]
Measuring Similarity

- TCGA data – AML (200 patients)
  - Clinical data, gene expression, mutations from TCGA AML;
  - Gene interactions, diseases, pathways from online databases.
Care informs research, research informs care

An informatics research agenda to support precision medicine: seven key areas


DOI: http://dx.doi.org/10.1093/jamia/ocv213  ocv213  First published online: 23 April 2016
The “sidecar” approach

SMART-on-FHIR implemented over i2b2

Kavishwar B Wagholkar,1,*, Joshua C Mandel,2,3 Jeffery G Klann,1,4 Nich Wattanasin,5 Michael Mendis,6 Christopher G Chute,6 Kenneth D Mandl,2,3 Shawn N Murphy7
Predicting the development of Diabetes complications and assessing the evolution of the disease
The MOSAIC project (2013 - 2016)
Data integration

I2b2 Data Warehouse

Air quality maps from satellite data

1000 patients (+2000) for ten years
The architecture

- EHR data
- Administrative data
- Environmental data

[Diagram showing various components of the architecture including Hospital Adapter, DB Query Engine, Broadcast/Aggregator, CORE ONTOLOGY, DB Cache, ORCHESTRATOR, Security, Track, Bayesian Network, Cox Regression, Data Mining Module, Temporal Abstraction, Process Mining, Process Modelling.]
Mosaic Dashboard

Traffic Lights – Predictive Models of complications and other monitoring parameters

<table>
<thead>
<tr>
<th>Traffic Lights</th>
<th>Clinical data</th>
<th>Therapies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PATIENT:</strong></td>
<td><strong>ONSET YEAR:</strong> 1981</td>
<td><strong>LAST VISIT:</strong> 30/05/2014</td>
</tr>
</tbody>
</table>

**DIASTOLIC BLOOD PRESSURE**
- Value: 70 mm Hg - Date: 24/10/2013
- Value: 80 mm Hg - Date: 18/01/2013

**SYSTOLIC BLOOD PRESSURE**
- Value: 120 mm Hg - Date: 24/10/2013
- Value: 140 mm Hg - Date: 18/01/2013

**DEI**
- Value: Good - Date: 24/10/2013
- Value: Good - Date: 20/08/2013

**PHYSICAL ACTIVITY**
- Value: No - Date: 24/10/2013
- Value: No - Date: 20/08/2013

**BMI**
- Value: 35.88 - Date: 24/10/2013
- Value: 35.88 - Date: 20/08/2013

**Hba1c**
- Value: 50.82 mmol/mol - Date: 17/10/2013
- Value: 50.82 mmol/mol - Date: 16/08/2013

**CARDIO VASCULAR RISK**
- Value: II - Date: 24/10/2013
- Value: II - Date: 17/10/2013

**LEVEL OF COMPLEXITY**
- Value: 3rd Level - Date: 24/02/2009
- Value: 2nd Level - Date: 31/07/2001
MOSAIC system validation in clinical practice

- Validation at ICSM – September 2015 to March 2016
  - DSS in use during clinical practice, BI module evaluated during periodic meetings
- Without MOSAIC system
  - 352 PATIENTS: Start: 15 September 2015, End: 1 December 2015
- With MOSAIC system
  - 353 PATIENTS: Start: 1 December 2015, End: 30 March 2016

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male = 3; Female = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>40.6 ± 15</td>
</tr>
<tr>
<td>Years of Professional Experience</td>
<td>13 ± 12</td>
</tr>
<tr>
<td>Role</td>
<td>7 diabetologists, 1 epidemiologist, 1 researcher</td>
</tr>
<tr>
<td>IT Literacy (Self-evaluation)</td>
<td>High = 2; Medium = 5; Low = 2</td>
</tr>
<tr>
<td>Affiliation</td>
<td>ICSM = 7; ASL = 2</td>
</tr>
</tbody>
</table>
DSS Prospective Validation – Results

Visit duration significantly lower WITH the MOSAIC system ($p<<0.01$)

More screening exams prescribed WITH the MOSAIC system ($p<0.01$)

More interventions on physical activity ($p = 0.05$)
Challenges....
Requirements to fully exploit variety

- Data models, sharing and standardization
- Trustworthy, Security and Privacy
- Knowledge maintenance and distribution
- Learning organizations, organizational learning
THI - Translational Health Informatics

EHR - Electronic Health Records
The i~HD Quality Seal for Research Platforms (QS4RP)


on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
VERTIcal Grid IOgistic regression (VERTIGO)


WebDISCO: a web service for distributed cox model learning without patient-level data sharing


- Codify and preserve useful knowledge,
- Learn how best to share and disseminate findings
- Critically assess the quality of the evidence in decision-making
- Provide a rationale or explanation for the recommendation
- Assess confidence in the recommendation
- Describe the data and knowledge sources and the reasoning model
- Learn from experience
Thanks to...

BMI LABS “Mario Stefanelli”
Dark times lie ahead of us and there will be a time when we must choose between what is easy and what is right
- Albus Dumbledore (J.K. Rowling)
Thank you for the invitation

Lelio Menozzi, “l’area VUL” (Pavia, seen from near my house)