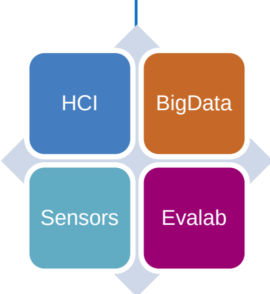
CERN openlab Healthcare Workshop

Big Data in Healthcare

November 2014

Frederic Ehrler, PhD
Christian Lovis, MD MPH

Hôpitaux Universitaires de Genève



Axes

Advanced Human-Computer interactions

- Interfaces design, ergonomic, dimensionality, NLP, 3D, contactless, GG, …

Massive Data

- Representation, semantic, interopérability, interactions, dimensionnality, analysis, …

Self-monitoring

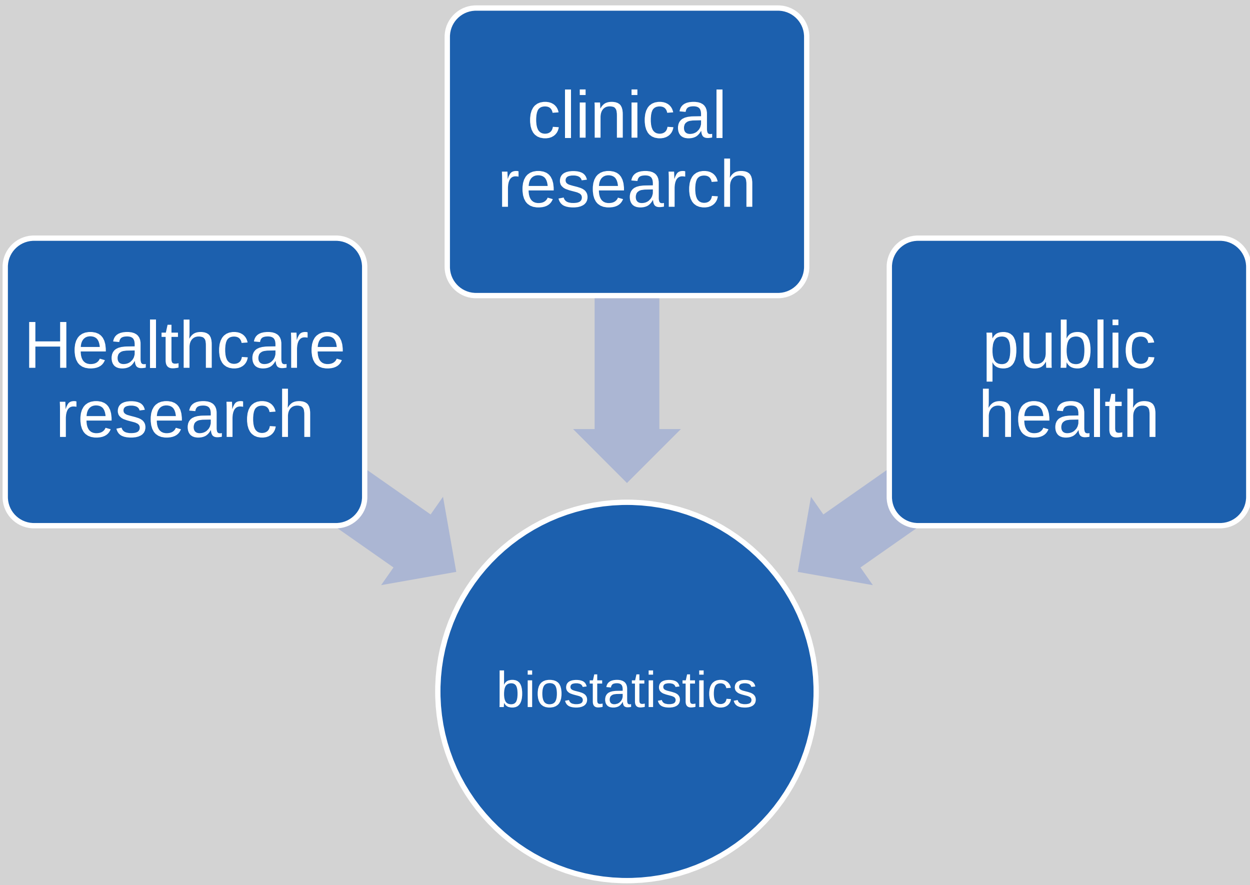
- Bio-sensors, sensors networks, interoperability, …

Evaluation - Evalab

Agenda

- Big data compliant bio-statistics project
- Time persistent interpretable Big data
  - Quality
  - Semantic

Need for biostatistics



Analyzing distributed BigData

- Life sciences requires specific and robust analytics

Frequently used statistical tests (modified from [3])

Statistical Test	Description
Fisher's exact test	Suitable for binary data in unpaired samples: the 2 x 2 table is used to compare treatment effects or the frequencies of side effects in two treatment groups
Chi-square test	Similar to Fisher's exact test (albeit less precise). Can also compare more than two groups or more than two categories of the outcome variable. Preconditions: sample size >ca. 60. Expected number in each field ≥5.
McNemar test	Preconditions similar to those for Fisher's exact test, but for paired samples
Student's t-test	Test for continuous data. Investigates whether the expected values for two groups are the same, assuming that the data are normally distributed. The test can be used for paired or unpaired groups.
Analysis of variance	Test preconditions as for the unpaired t-test, for comparison of more than two groups. The methods of analysis of variance are also used to compare more than two paired groups.
Wilcoxon's rank sum test (also known as the unpaired Wilcoxon rank sum test or the Mann-Whitney U test)	Test for ordinal or continuous data. In contrast to Student's t-test, does not require the data to be normally distributed. This test too can be used for paired or unpaired data.
Kruskal-Wallis test	Test preconditions as for the unpaired Wilcoxon rank sum test for comparing more than two groups
Friedman test	Comparison of more than two paired samples, at least ordinally scaled data
Log rank test	Test of survival time analysis to compare two or more independent groups
Pearson correlation test	Tests whether two continuous normally distributed variables exhibit linear correlation
Spearman correlation test	Tests whether there is a monotonous relationship between two continuous, or at least ordinal, variables

Choosing Statistical Tests
Part 12 of a Series on Evaluation of Scientific Publications
Jean-Baptist du Prel, Bernd Röhrig, Gerhard Hommel, Maria Blettner

1) Bigdata & statistics

- The problem:
  - Very large data volumes
  - Limited bandwidth
  - Heterogeneous legal framework

→**Distributed storage**

→**Very limited ways of aggregating data**

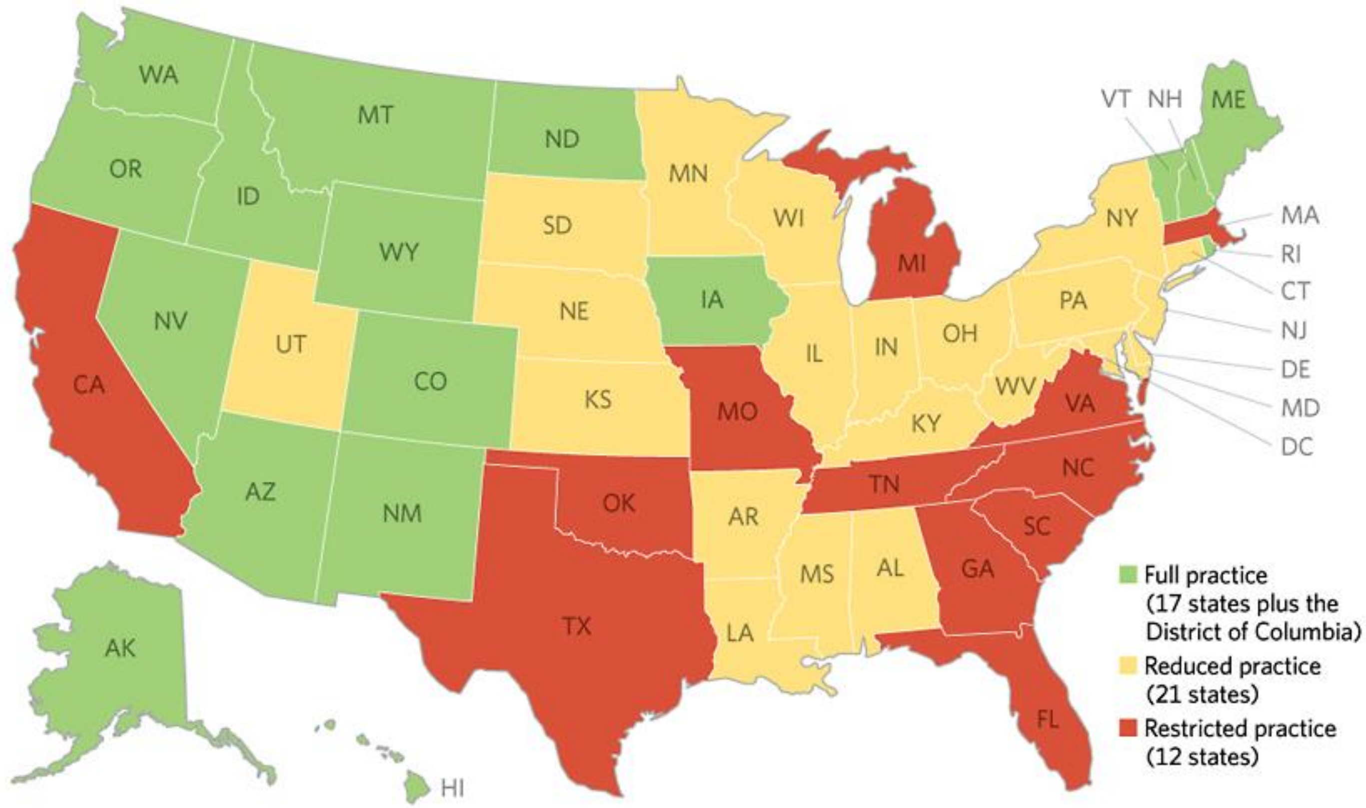
Heterogeneous sources

- **Phenotype**
  - Electronic patient records
  - Self monitoring, health records
- **Genomics**
  - Epigenetics
  - Proteomics, Metobolomics, *omiocs…
- **Environment**
  - Air, water, etc… quality
  - Pollutants, pollens, …
  - Sanitation
  - Microbiomes, …
- **Lifestyle**
  - Social media
  - Diet, nutrition, sport, etc…

Distributed sources

MAP 1

Advanced Practice Registered Nurses—Scope of Practice Variability



Source: American Association of Nurse Practitioners, "State Practice Environment," http://www.aanp.org/legislation-regulation/state-practice-environment (accessed December 2, 2013).

B2887 heritage.org

Limitations to the fast transport of very large amount of data through existing networks

HUG
Hôpitaux Universitaires de Genève

Which can be distributed ?

$$\sqrt{\sum_{i=0}^{n} a_i} <> \sum_{i=0}^{n} \sqrt{a_i}$$

- Mean
  - Number and sum of each source
- Median ?
- Correlation ?
- Etc …

Hôpitaux Universitaires de Genève

Need to characterize each test

Can be computed in distributed datasources

- Min, max, etc …

Can be indirectly…

- Define robust methodology, such as for mean

Cannot be distributed…

- R&D to find solution or proxys

HUG
Hôpitaux Universitaires de Genève

What we want to achieve

- The final product of this project is a clear, defined and robust biostatistics framework that can be used in analytics for truly distributed environments.

Quality of time persistent data

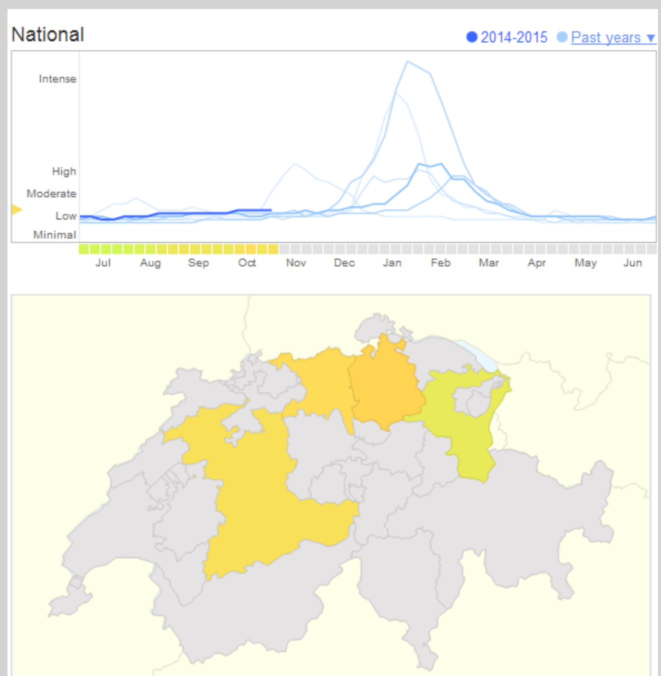
question in the present based on a known dataset the past

A large flow of data in the present and past that can be used to predict events in the present or near future

in the future, to answer questions that were not known when the database was built.

Identification of current trend

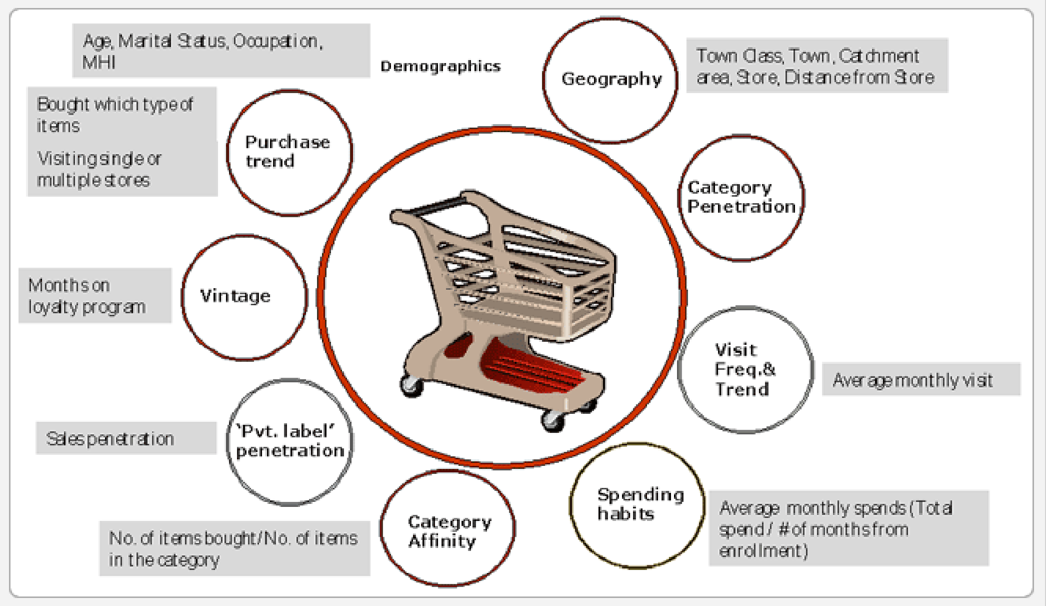
- An existing database that covers the field of the questions, to evaluate a known question in the present based on a known dataset the past. This is exemplified by Google Flu trends, or internet fingerprinting if the data are individualized.



HUG
Hôpitaux Universitaires de Genève

Prediction of future event

- A large flow of data in the present and past that can be used to predict events in the present or near future. This is the approach of risk evaluation in insurers for example



Building database for future research

- In that third situation, the database has to be built in the present in such a way that it will be usable, in the future, to answer questions that were not known when the database was built.

HUG Hôpitaux Universitaires de Genève

GiGo – Garbage In – Garbage Out

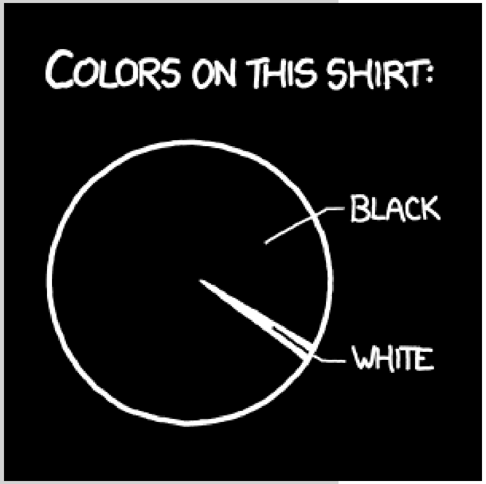
- Everybody is concerned by data quality
- Everybody is trying to improve data quality

GiGo – Garbage In – Gold Out

- Improving data quality will always be behind data production
- Improving data quality is very expensive
- Keeping data quality is not sustainable
- **So, don’t care about data quality, but describe it !**

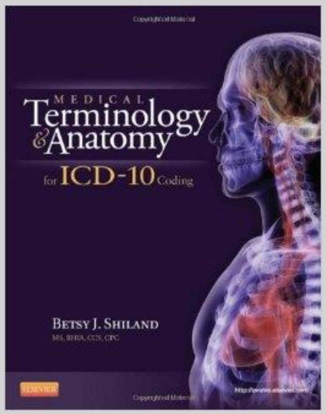
What we want to achieve

- This project aims at developing a metadata framework that
  - allows to develop a model able to support a structured description of the quality of the data, the reliability of the sources, with the temporal variation of these elements, in order to support further interpretation of the data.
  - The objective of the project is to lower the dependence on data quality by having a formal description of it in the data space.



HUG
Hôpitaux Universitaires de Genève

Semantic of time persistent data



SNOMED® INTERNATIONAL

LOINC®
Logical Observation Identifiers Names and Codes

GO the Gene Ontology

HUG
Hôpitaux Universitaires de Genève

What we want to achieve

- This project aims at exploring new and innovative ways of representing semantics in databases characterized by a constant growth in depth and width.
- Such type of databases handle data that need:
  - a strong evolutionary representation of meaning,
  - to keep the meaning of each element at each temporal incidence of data,
  - to be able to map to existing terminologies, classifications and ontologies in the field; to handle the evolution of them;
  - to adapt to semantic extension and sludge;
  - to provide a multi-dimensional framework of meanings
  - to build a strong language-dependent mapping of meanings.

HUG
Hôpitaux Universitaires de Genève