Building Blocks for AI

Marco Antonio Gutierrez
Informatics Division &
Biomedical Informatics Laboratory
Heart Institute, University of Sao Paulo Medical School
São Paulo, BRAZIL
Heart Institute (InCor)
University of Sao Paulo Medical School, Brazil

- 500 beds
- 2,000,000 lab tests/yr
- 260,000 outpatients/yr
- 200,000 Image diagnosis/yr
- 12,000 inpatients/yr
- 12,000 CATH procedures/yr
- 4,500 surgeries/yr
- Graphic methods 140,000/yr
- 500 beds
- 260,000 outpatients/yr
- 12,000 inpatients/yr
- 4,500 surgeries/yr
Clinics Hospital
University of Sao Paulo Medical School, Brazil

- Largest hospital in Latin America
- 7 Institutes
- 3,000 beds

- 3,500 outpatients/day
- 1,000,000 lab tests/month
Big Data Analytics in Health and Life Sciences

Today: Many disparate data types, streams...

Meds & labs
Clinical
Genomics

Personal data

Claims & transactions

Patient experience

Future: Integrated computing and data

Payer
Provider
Life Sciences
Patient

Accountability

Fraud Detection
Population risk management
Imaging
Population risk management

Revenue realization
Genomics
Bioinformatics
Clinical Trials

Personal Health

Evidence-based clinical decision support
Wellness
Health monitoring data
Personalization
Decision Transparency

Treatment Value
Today Data Sources

- Legacy Systems
  - Print to Text or Delimited String
- All EMR Initiated Data (Stored Procedures)
- Medical Imaging Devices (DICOM/HL7)

- Physiological Monitors (HL7)
- Ventilators (HL7)
- Lab/Analysis Equipment (HL7)
- Smart Pumps (HL7)
- Home Monitoring
- Hospital Sensors
Newer Data Sources

• External Streaming Device Data
• Wearables
• Home Devices
• Geographic Information System (GIS) Data
• Omic Data

• Open Data
  – http://dados.gov.br
  – http://data.gov.uk
  – http://www.data.gov

• Adverse Drug Event
  – http://www.researchae.com

• Internet of Things (IoT)
  – Telematics
  – 5G

EMR: Future data sources

- **Patient-Generated Data**
  - Home Devices (Scale, Vital Signs, Glucose)
  - Exercise & Diet (Fit Bit, Jawbone, Nike)

- Combining Phenotype Data with Genotype Data

- Patient Threat Analysis
  - Patient caregivers and outcomes

---

BigData & InCor ~0.5 PetaByte (0.5 \(10^{15}\)) and counting

- 1.4 million patients
- 40 million LA exams
- 2 million Image exams
- > 1 million EKG

### Modality

<table>
<thead>
<tr>
<th>Modality</th>
<th># Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest X-Ray</td>
<td>CR 840.159</td>
</tr>
<tr>
<td>ECO/Ultrasound</td>
<td>US 528.916</td>
</tr>
<tr>
<td>EKG</td>
<td>ECG 317.614</td>
</tr>
<tr>
<td>Tomography</td>
<td>CT 235.355</td>
</tr>
<tr>
<td>X-Ray Angiograph</td>
<td>XA 174.470</td>
</tr>
<tr>
<td>Nuclear Medicine</td>
<td>NM 101.403</td>
</tr>
<tr>
<td>Magnetic Resonance</td>
<td>MR 73.107</td>
</tr>
<tr>
<td>Endoscopy</td>
<td>ES 14.696</td>
</tr>
<tr>
<td>PET</td>
<td>PT 3.674</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,289,394</strong></td>
</tr>
</tbody>
</table>
Our initiatives in AI

I. Biomedical signal processing (e.g. EKG analysis);

II. Medical Image processing (e.g. Cardiac CineMRI & Chest X-Ray Analysis);

III. NLP to extract non-structured information from ePR (e.g. medical report or Corona Virus outbreak)

IV. Integration of biomarkers & clinical data to produce predictive risk algorithms to assist the management of CVD.
EKG Analysis

Database w/ 200K EKG + report

Mapped to

Priorization levels

Critical
Urgent
Non-Urgent
Normal

Initial Dataset: 20K EKGs
Extended Dataset: 200K EKGs

<table>
<thead>
<tr>
<th>TPR</th>
<th>TNR</th>
<th>FPR</th>
<th>FNR</th>
<th>PPV</th>
<th>F1 Score</th>
<th>Accuracy</th>
<th>AUC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.94</td>
<td>0.71</td>
<td>0.29</td>
<td>0.06</td>
<td>0.90</td>
<td>0.92</td>
<td>0.88</td>
<td>0.82</td>
</tr>
</tbody>
</table>
CineMRI Automated Analysis

<table>
<thead>
<tr>
<th></th>
<th>DICE index</th>
<th>APD (mm)</th>
<th>Good Contours (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epicardium</td>
<td>0.93 ± 0.06</td>
<td>0.72 ± 0.62</td>
<td>98.59 ± 4.28</td>
</tr>
<tr>
<td>Endocardium</td>
<td>0.86 ± 0.13</td>
<td>1.19 ± 1.29</td>
<td>94.98 ± 14.04</td>
</tr>
</tbody>
</table>
Chest X-Ray Automated Classification

Dataset of annotated Chest X-Ray images & reports w/ phenotypes

Mapped to

Critical priorization levels

Critical
Urgent
Non-Urgent
Normal
Any AI initiative needs GPU power

5 h (NV 1050ti) \rightarrow 5 \textbf{min} (M100)

1 h (NV 1050ti) \rightarrow 1 \textbf{min} (M100)
InCor Data Science Enterprise Resource

- EHR
- Images & Signals
- Omics

- De-ID API

- 1.4 million patients
- 40 million LA exams
- 2 million Image exams
- > 1 million EKG

Projects:
- OHDSI
- R
- TensorFlow
- KNIME

Targeted Groups:
- Bench Scientists
- Clinical Researchers
- Students & Trainees
Area 200 m²
166,535 BTU/h (dual)
120,000 BTU/hr (backup)
Redundants power source
Nobreak 50 KVA
Moto-generator
Challenges in Life Sciences

**Big Data in Life Sciences**
- Sequencer advances – 4x data in 50% less time: 0.5TB/device/day
- 4D molecular imaging produces 2TB/device/day
- Fragmented software ecosystem, lots of open source

**Burdens of Data Management**
- Store, manage, share, ingest and move PBs of research & clinical data
- Need to reliably ‘snapshot’ pipelines with archive to tiered storage

**Innovation Drives Change**
- Rapid iteration of algorithms far outpace IT, requiring flexibility, agility
- Most applications do not fully leverage available infrastructure

**Converged Infrastructure**
- Workloads converging between local and cloud-based HPC/BiD Data
- Advanced orchestration required to maximize throughput & efficiency

Enabling extreme-scale computing on massive data sets
How to codify and represent our knowledge

Experiments & Instruments
- Facts

Simulations
- Facts

Literature
- Facts

Other Archives
- Facts

Questions

Answers

Generic Problems

- Managing peta/exabytes
- How to organize it
- How to reorganize it
- How to share it?

- Tools for visualization & presentation
- Building and executing models
- Documenting experiments
- Curation and long-term preservation

Adapted from: The Fourth Paradigm
Data-intensive Scientific Discovery
Edited by Tony Hey, Stewart Tansley, and Krist in Tolle, Microsoft Corp, 2009
"Not everything that counts can be counted, not everything that can be counted, counts."
Albert Einstein
Thank you!

Email: marco.gutierrez@incor.usp.br