Persistent and Hackathons

Smart India Hackathon
Our country needs audacious digital transformation to reach its potential.

- 25% of India between the age of 16-25
- Over 300 Million young people
- The largest exporter of information technology services and skilled manpower among developing countries

...but a vast majority of its population still lacks the skills to meaningfully participate in the digital economy.
Smart India Hackathon

World’s biggest open innovation model !!

Smart India Hackathon is a non-stop digital product development competition, where problems are posed to technology students for innovative solutions. It...

- Harnesses creativity & expertise of students
- Sparks institute-level hackathons
- Builds funnel for ‘Startup India’ campaign
- Crowdsources solutions for improving governance and quality of life
- Provides opportunity to citizens to provide innovative solutions to India’s daunting problems

How it works

- Problem statements released by ministries
- Students submit proposals
- Smart India Hackathon finals
- Evaluation + selection of finalists
Ingredients for a successful hackathon

• Executive sponsorship
• Theme identification
  • Broad areas of problem domains, solution requirements identified
• Technology available:
  • Communication and collaboration
  • Platforms, hardware
  • Internal and external resources
  • Data
• Enablement
  • Coaches
  • Demos
  • Licenses
• Internalization
  • Innovation as a habit
  • Traction
• Contact: sinaihackathon@persistent.com  shriram@persistent.com
VEGA: Healthcare data platform

Integration Engine
Create interfaces and channels easily using standard templates. Integrate with EMRs and other healthcare back-end systems using HL7, X12, web services.

Experiences
Create experiences at speed that focus on patient engagement and care-coordination.

Algorithms
Repository of solutions for various condition types and specialities. Use a existing catalogue or create new additions.

Data Pipeline
Organize the data layer activities into a set of repeatable pipelines. A pipeline consists of work-flow steps consisting of actions. Actions supported include shell scripts, python code, map-reduce patterns.

Data Sets
A sample set of real world (de-identified) data that covers patients, claims, outcomes and other example data required to simulate a real world experience. New data sets are added continuously to the catalogue.

Visualizations
Interactive visualization tailored for healthcare experiences. A rich catalogue of ready to use examples and ability add new ones.
Marketing List Segmentation Analytics

Advanced analytics provided by ShareInsights helps a global tech product conglomerate to generate meaningful, actionable insights and minimize risks for future campaigns.

Audience Engagement Analytics

Know how ShareInsights enabled a media organization to get an accurate representation of its viewership numbers by analyzing multi-channel data.
SanGeniX: NGS data analysis solution

Introduction: SanGeniX is a comprehensive and user friendly NGS data analysis solution. It’s rich graphical user interface (GUI) which enables user to perform primary, secondary and tertiary analysis of NGS data in a seamless fashion.

Business Problem: NGS data analysis is very complicated and requires specialized bioinformatics as well as computation resources.

Domain: Genomics

Areas of application: Personalized medicine, Clinical diagnostics, Animal genomics, Plant genomics, and Microbial genomics

Impact: Simplifies analysis and interpretation of NGS data for biologists.
SanGeniX: NGS data analysis with workflows and visualization

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Business Value</th>
<th>Important features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Integration of industry standard algorithms ensures data reliability</td>
<td>• Increased productivity</td>
<td>• Predefined workflows for all the common NGS workflows</td>
</tr>
<tr>
<td>• Automation - less hands-on time.</td>
<td>• Cost effective due to cloud hosted pay per use solution</td>
<td>• Integration of industry standard algorithms</td>
</tr>
<tr>
<td>• Point-and-click - Easy to use</td>
<td>• Highly customizable to include customer specific workflows</td>
<td>• Flexibility to include custom algorithms</td>
</tr>
<tr>
<td>• Intuitive and innovative data visualization</td>
<td>• Allows focus on biological questions and analysis instead of pondering over data</td>
<td></td>
</tr>
<tr>
<td>• Role based access to ensure security and compliance</td>
<td>• HIPAA Compliance</td>
<td></td>
</tr>
<tr>
<td>• Scalable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graphically enriched visualization
A comprehensive literature mining solution
for biomedical literature

www.biogyan.com
Introduction

User’s Query

- Gene names
- Biological processes
- Cell-type/Tissue

Collate & Process PubMed articles

- Compute article relevance
- Highlight searched terms & sentences
- Generate holistic graphs & tables

User Action

- Study results
- Mark validity, add comments
- Export results to Spreadsheet, EndNote, BibTeX

Key features:

- Allows combinatorial search of genes, cell-types and cellular processes
- Ranks scientific articles based on relevance to the search terms
- Displays biological pathways and 3D structures of queried genes
- Allows offline reading of articles
- Available at: http://biogyan.com/
BioGyan - Screenshots

Result Screen highlighting ranked articles

Biological pathway & 3D structure

Queried genes & their participation

Distribution of articles for a queried gene
Market Space

• Life Sciences Research Space:
  • Academic research labs
  • Research Hospitals
  • Pharma/Biotech research labs

• Key market players:
  • All first and second tier Universities and Pharma/Biotech companies

• Industry Issues:
  • Literature mining is an integral part of R&D, drug discovery and regulatory filings.

• Key Services:
  • Licensed product
  • Literature mining service
  • Development of customized solution
Multiomics Project
Need of the hour: Omics data integration and analysis for precision medicine

Multi-omics Platform

- Development of a comprehensive multi-omics data analysis platform
- Implementation of diverse analysis pipelines for addressing use-cases such as biomarker discovery and disease sub-typing
- Data analytics and advanced visualization for deriving actionable outcomes
- Open for collaboration

Inter-relationships between omic data (e.g. genomics, transcriptomics, proteomics, metabolomics)
Leverage bioinformatics, data analytics and machine learning
Address diverse use-cases (e.g. diagnostic, drug discovery, precision medicine)
Multi-omics approach for biomarker prediction from large scale data
Biomarker discovery pipelines: Multiple Approaches

**Input layer:**
Multiomics datasets (multiple samples)

**Analysis layer:**
Multiple approach

**Prediction layer:**
Putative biomarker prediction

**Statistics/Bioinformatics analysis**

**Pattern recognition: Machine Learning**

**Pattern recognition: Data integration methods/tools**

**Features (genes) selection:**
Statistically significant genes

**Features (genes) selection:**
Contribution to model building

**Biomarker Prediction: Tertiary/Downstream analysis**

- **Transcriptomics data:** mRNA, miRNA
- **Genomics data:** CNV, SNP, etc
- **Proteomics data**
- **Metabolomics data**
- **Epigenetics data:** Methylation, ChIPSeq, etc
- **Clinical data**

**Individual dataset**

**Individual/combined dataset**

**Combined dataset**
Statistical Approach for biomarker prediction

**Input Layer**
- **Genomics (CNV, SNP)**
- **Transcriptomics (MRNA, MiRNA)**
- **Proteomics (RPPA, Ms /MS)**
- **Metabolomics (LC/GC-MS)**
- **Epigenomics (Methylation, ChiP-Seq)**

**Individual Omic analysis layer**
- **Normalization & Analysis** (Gistic2, CNTOOLS, MAFTOOLS)
- **Significant genes/ Genomic regions**
- **Normalization & Analysis** (Limma, EdgeR, DESeq2)
- **Significant Genes/miRNAs**
- **Normalization & Analysis** (Galaxy-P, PGA)
- **Significant proteins**
- **Normalization & Analysis** (SIMCA-P, R)
- **Significant metabolites**
- **Normalization & Analysis** (Limma, MACS)
- **Significant genomic regions** (Methylation, Enriched sites)

**Integrative layer**
- **Correlation Analysis / Clustering / Cluster of clusters / Multivariate Analysis / Pathway analysis / Network Analysis**

**Biomarker Prediction / Tertiary analysis**
- **Literature Annotation**
  - Outcome: Literature based gene process associations
  - Tool: BioGyan
- **Network based analysis**
  - Outcome: List of risk pathways
  - Tool: iSubPathway Miner
  - Outcome: List of cross talking pathway pairs and network
  - Tool: Cross-Talk module
  - Outcome: PPI network, pathway enrichment, functional annotations
  - Tool: REACTOME FI
- **Annotation from Databases**
  - Outcome: Annotations like molecular function, diseases, drugs, pathways, TFBS etc.
  - Tool: ToppGene

**Prediction layer**
Machine Learning for biomarker prediction: Approach 1

**Input Layer**

- Genomics (CNV, SNP)
- Transcriptomics (mRNA, miRNA)
- Proteomics (RPPA, Ms /MS)
- Metabolomics (LC/GC-MS)
- Epigenomics (Methylation, ChIP-Seq)

**Normalization & Analysis**

- Genomics (Gistic2, CNTOOLS, MAFTOOLS)
- Transcriptomics (Limma, edgeR, DESeq2)
- Proteomics (Galaxy-P, PGA)
- Metabolomics (SIMCA-P, R)
- Epigenomics (Limma, MACS)

**Significant genes/miRNAs**

- Significant genes/miRNAs
- Significant proteins
- Significant metabolites
- Significant genomic regions (Methylation, Enriched sites)

**Feature Selection**

- Model1
- Model2
- Model3
- Model4
- Model5

**Integrated Analysis**

- Correlation Analysis | Clustering / Cluster of clusters | Multivariate Analysis | Pathway analysis / Network Analysis

**Biomarker Prediction / Tertiary analysis**

- Literature Annotation
  - Outcome: Literature based gene process associations
  - Tool: BioGyan
- Network based analysis
  - Outcome: List of risk pathways
  - Tool: iSubPathwayMiner
- Annotation from Databases
  - Outcome: Annotations like molecular function, diseases, drugs, pathways, TFBS etc.
  - Tool: ToppGene

**Prediction layer**

- Nearest Neighbors
- Linear SVM
- RBF SVM
- Gaussian Process
- Decision Tree
- Random Forest
- Neural Networks
- AdaBoost
- Naive Bayes
- QDA
- Logistic Regression
- Decision Tree
- Random Forest
- Neural Networks
- AdaBoost
- Naive Bayes
- QDA
- Logistic Regression
Machine Learning for biomarker prediction: Approach 2

Integrated Input Layer

- Genomics (CNV, SNP)
  - Normalization & Analysis (Gistic2, CNTOOLS, MAFTOOLS)
  - Significant genes/Genomic regions
- Transcriptomics (MRNA, MiRNA)
  - Normalization & Analysis (Limma, EdgeR, DESeq2)
  - Significant Genes/miRNAs
- Proteomics (RPPA, Ms /MS)
  - Normalization & Analysis (Galaxy-P, PGA)
  - Significant proteins
- Metabolomics (LC/GC-MS)
  - Normalization & Analysis (SIMCA-P, R)
  - Significant metabolites
- Epigenomics (Methylation, ChiP-Seq)
  - Normalization & Analysis (Limma, MACS)
  - Significant genomic regions (Methylation, Enriched sites)

Machine Learning layer

- Machine learning
  - Nearest Neighbors
  - Linear SVM
  - RBF SVM
  - Gaussian Process
  - Decision Tree
  - Random Forest
  - Neural Networks
  - Adaboost
  - Naive Bayes
  - QDA
  - Logistic Regression
- Combined Model
- Feature Selection
- Two or more data set combined
  - Correlation Analysis / Clustering / Cluster of clusters / Multivariate Analysis

Prediction layer

- Literature Annotation
  - Outcome: Literature based gene process associations
  - Tool: BioGyan
- Network based analysis
  - Outcome: List of risk pathways
  - Tool: iSubPathwayMiner
  - Outcome: List of cross talking pathway pairs and network
  - Tool: Cross-Talk module
  - Outcome: PPI network, pathway enrichment, functional annotations
  - Tool: REACTOME FI
- Annotation from Databases
  - Outcome: Annotations like molecular function, diseases, drugs, pathways, TFBS etc.
  - Tool: ToppGene
Data Integration methods for biomarker prediction

**Input Layer**
- Genomics (CNV, SNP)
  - Normalization & Analysis (Gistic2, CNOTOOLS, MAFTOOLS)
  - Significant genes/Genomic regions
- Transcriptomics (MRNA, MiRNA)
  - Normalization & Analysis (Limma, EdgeR, DESeq2)
  - Significant Genes/miRNAs
- Proteomics (RPPA, Ms/MS)
  - Normalization & Analysis (Galaxy-P, PGA)
  - Significant proteins
- Metabolomics (LC/GC-MS)
  - Normalization & Analysis (SIMCA-P, R)
  - Significant metabolites
- Epigenomics (Methylation, ChIP-Seq)
  - Normalization & Analysis (Limma, MACS)
  - Significant genomic regions (Methylation, Enriched sites)

**Integrated layer**
- OncoImpact
- Driver Genes
  - mixOmics
  - MOcluster
  - iClusterplus
  - Feature Ranking
  - Cluster membership of samples
  - Visualization: Heatmaps, plots, etc

**Prediction layer**
- Literature Annotation
  - Outcome: Literature based gene process associations
  - Tool: BioGyan
- Network based analysis
  - Outcome: List of cross talking pathway pairs and network
  - Tool: Cross-Talk module
- Annotation from Databases
  - Outcome: PPI network, pathway enrichment, functional annotations
  - Tool: REACTOME FI
  - Outcome: Annotations like molecular function, diseases, drugs, pathways, TFBS etc.
  - Tool: ToppGene

**Biomarker Prediction / Tertiary analysis**
Current Use Case:

TCGA Breast cancer data analysis for biomarker prediction

**Data filtering: Quality control**

- **Data filtering: Quality control**
- **DESeq2**
- **edgeR**
- **Limma**
- **Machine learning**
- **Nearest Neighbors**
- **Linear SVM**
- **RBF SVM**
- **Gaussian Process**
- **Decision Tree**
- **Random Forest**
- **Automated feature selection**
- **Statistically significant genes**

**Statistical analysis**

- **DESeq2**
- **edgeR**
- **Limma**
- **Machine learning**
- **Nearest Neighbors**
- **Linear SVM**
- **RBF SVM**
- **Gaussian Process**
- **Decision Tree**
- **Random Forest**
- **Automated feature selection**
- **Statistically significant genes**

**Gene expression data (read counts per gene)**

- **MOcluster**
- **mixOmics**
- **iClusterPlus**
- **Machine learning**
- **High scoring genes**
- **Automated feature selection**

**TCGA (Breast cancer datasets)**

- **Normal TNBC Non-TNBC**

**Copy number variation data (Segmentation mean data)**

- **GISTIC**
- **CNTools**
- **Machine learning**
- **Nearest Neighbors**
- **Linear SVM**
- **RBF SVM**
- **Gaussian Process**
- **Decision Tree**
- **Random Forest**
- **Automated feature selection**
- **Statistically significant genes**

**Biomarker prediction: Tertiary analysis**

**766 samples**

**831 samples**

**825 samples**