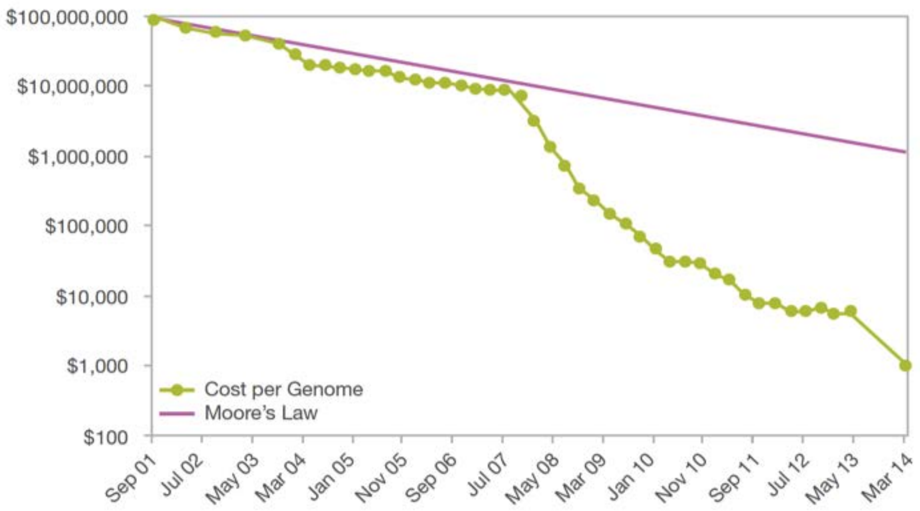
Personalized Medicine and IT

Translating Genomic-Based Research for Health

www.intel.com/healthcare/bigdata

Ketan Paranjape
General Manager, Life Sciences
Intel Corp.

$1000 Genome

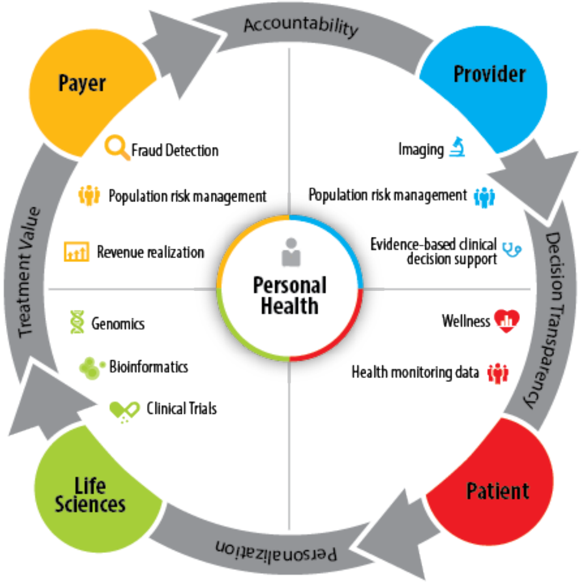

$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Cost per Genome
Moore's Law
Sep 01
Jul 02
May 03
Mar 04
Jan 05
Nov 05
Sep 06
Jul 07
May 08
Mar 09
Jan 10
Nov 10
Sep 11
Jul 12
May 13
Mar 14

- **Agenda**
  - Genomics = “Big Data”
  - Barriers
  - Solutions

- **Takeaways**
  - Genomics, Personalized is a “Big Data” problem; Lots to learn from other industries or may be not ..
  - Barriers – Clinical, Literacy and Societal, Economic and Commercial, Technical, Ethical; Validated in 12 Countries
  - It is all about the ecosystem – payers, providers, pharma, and .. the patient

Personalized Medicine @ Intel

Today: Many disparate data types, streams…



Future: Integrated computing and integrated data

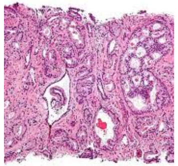
Leading to better decisions

- Improved patient experience
- Healthier population outcomes
- Reduced costs

Promising Areas for Genomic PM

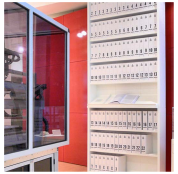
Pathogens

- Epidemics control
- Infection control
- Epidemics resources management
- Pharmacogenomics (tailored response)



Tissue

- Cancer characterisation
- Pharmacogenomics (tailored treatment)
- Symptoms and side effects management



Human

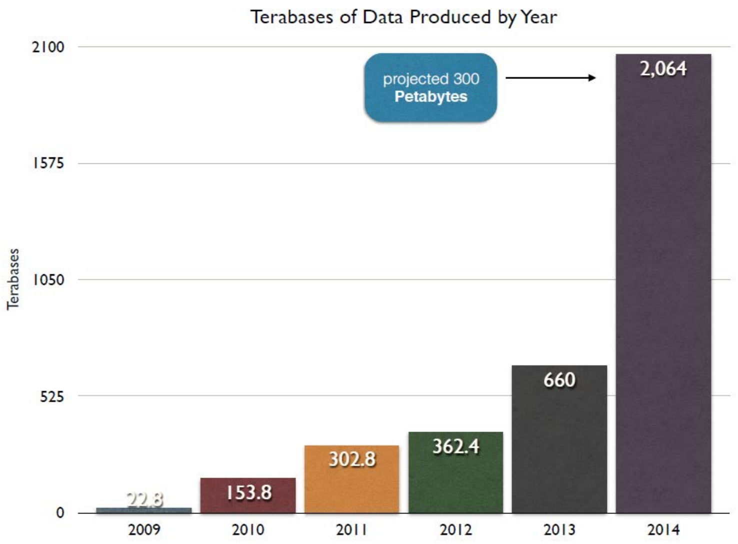
- Rare diseases
- Pre-* and Neo-natal screening*
- Common disease predispositions*
- Pharmacogenomics (tailored prevention and treatment)*

*some societal and/or ethical debates ongoing in these areas

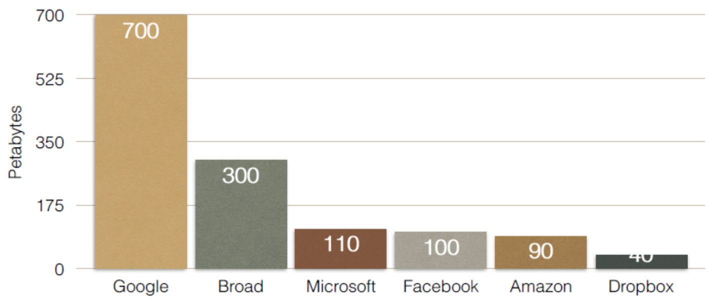
Acknowledgements

It Takes a Village …

Genomics is a Big Data Problem



We produce as much data as the big cloud providers



The Broad Institute will produce more data than Microsoft, Facebook and Amazon combined by 2015 ….

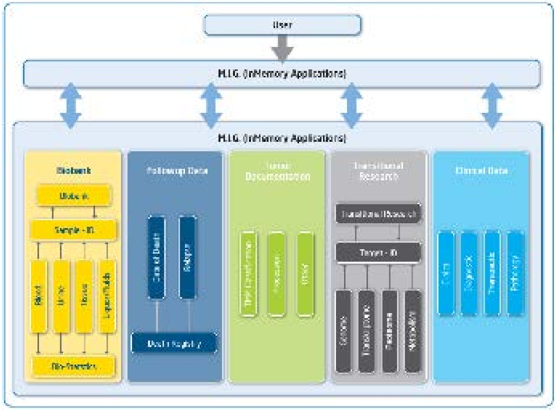
BROAD INSTITUTE

The Challenges of analyzing hundreds of thousands of genomes; Mauricio Carneiro, PhD, Broad Institute

Technical Barriers

Data Generation (science)	Data Management (technical)	Data Interpretation (multidisciplinary)
Reproducibility	Proliferation of open source tools	Creation of large sophisticated linked dataset
Data Quality	Transfer & Sharing	Smart analytics and novel algorithm development
Methods transparency	Scalability	Information Visualisation Tools
Experimental design	Security	Integration with EHR
Creation of common ontologies	Interoperability	Clinical Decision Support Tools
	Compression	
	Storage & Curation (short term and long term)	
	Computing Infrastructures	

Charite **“Real-time” Cancer Analysis** – Matching proper therapies to patients using **in-memory techniques**



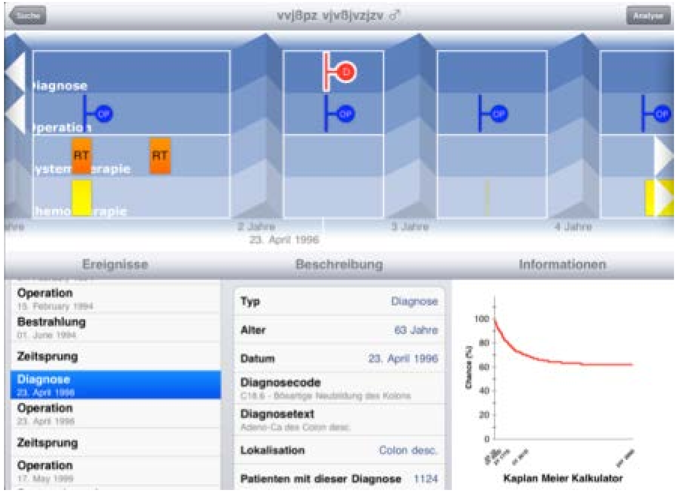
- **Challenge**: Real-time analysis of cancer patients (3.5M Data points per Patient, Up to 20 TB of data/patient)
- **Solution**: Using structured and unstructured data to collect and analyze tables used to take up to **two days -- now takes seconds**
- **Benefits**: Improves medical quality in disruptive way for Patient, Doctor, Hospital, Research

HANA Oncolyzer

- Ad-hoc Analysis of heterogeneous tumor data for cancer research
  - Medical records from decades of tens of thousands of patients
  - Structured and unstructured data (records, time series, free text, etc.)

Solution

  - Integrated into condensed but exhaustive view
  - On-the-fly analyses
    (e.g. Kaplan-Meier estimation, cohort statistics)
  - Includes external data sources
    (e.g. PubMed, pharmaceutical databases)
  - Attributes can be native, views, freetext-extracted, calculated



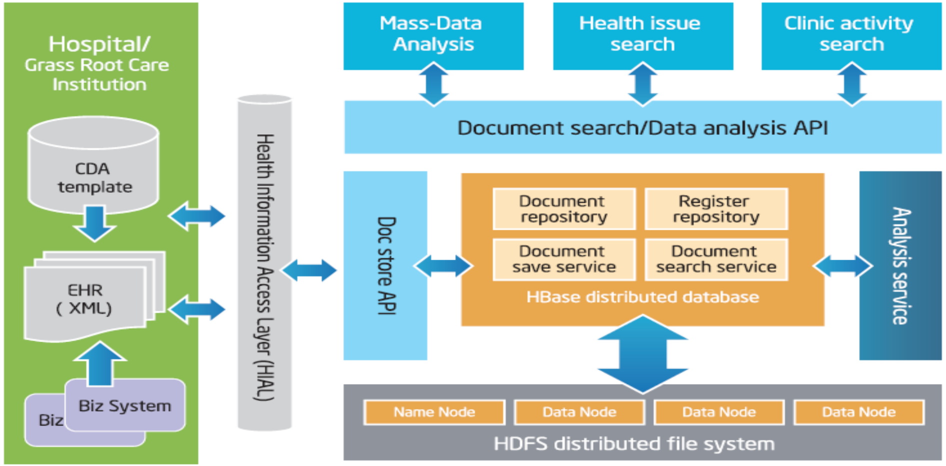
Regional Health Information Network RHIN – China (Jinzhou, Pop 3M)

Challenge: RHIN has challenges with scalability, performance and maintenance. Data storage is expensive

Solution: EMR data and healthcare services running on Apache Hadoop

Benefits: High performance and scalability demonstrated via POC and stress testing. Significantly reduced storage cost

1/5 Reduction in Response Time; 5x Concurrent Users



Data processing flow of RHIN platform

http://hadoop.intel.com/pdfs/IntelChinaHealthyCityAnalyticsCaseStudy.pdf

Collaborative Care Analytics

Business Need

Identify relevant features and patterns behind diseases, their similarities and differences in order to more accurately identify suspect conditions and link ICD9/10 codes with HCCs.

Benefits to Business

- Efficiently and Intelligently Identify disease concepts and patterns
- Identify suspect conditions to identify previously undiagnosed cases
- Predicting a substance abuser
- Predict insurance company rejections/adjustments
- Identifying forgeries altered prescriptions
- Improving inventory control by analyzing patterns of drug usages and thus maintaining optimal inventory levels

Solution

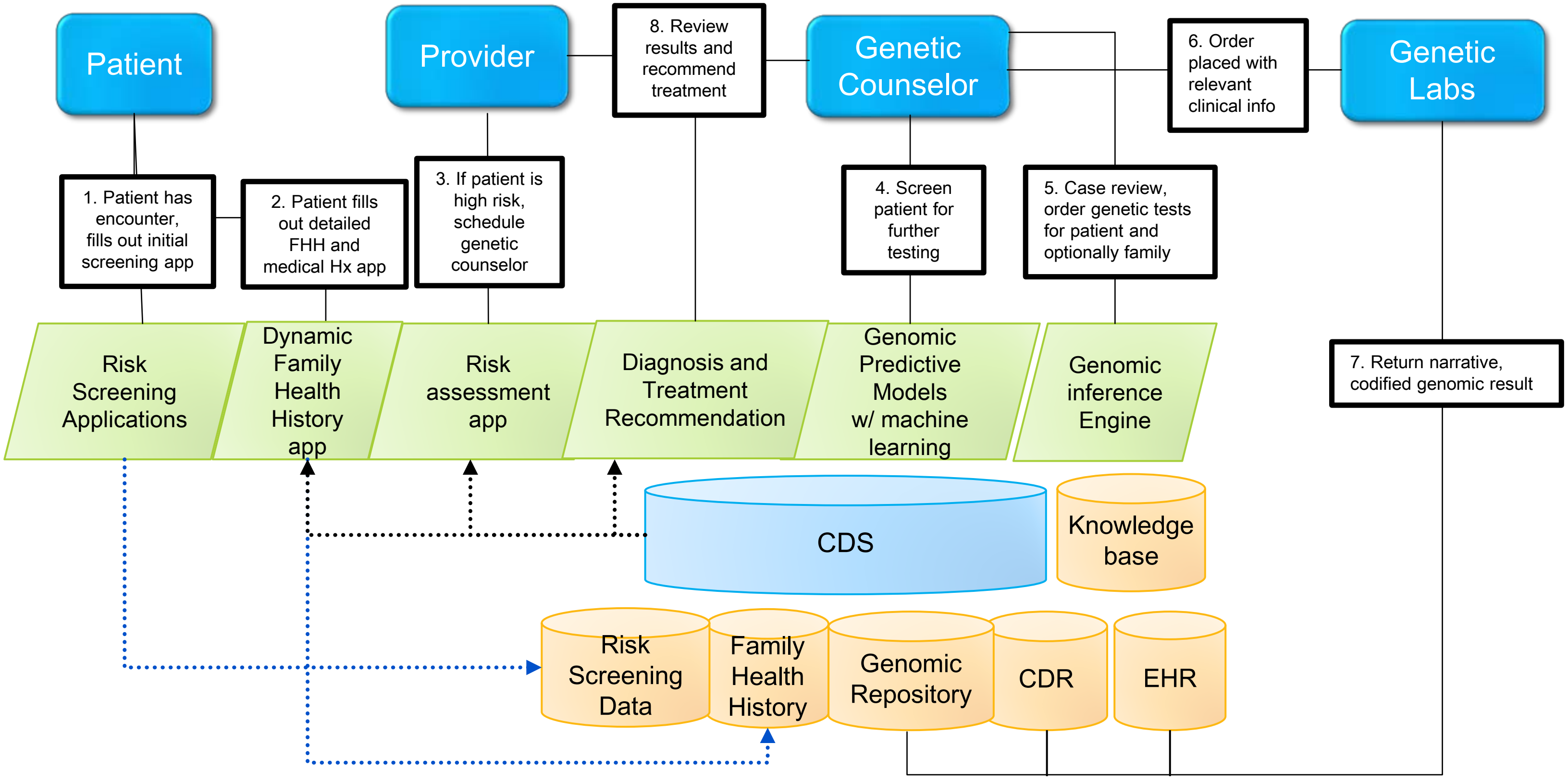
Phase	Description
Data Attribution	Assigning all various pieces of data (records, images) to respective patients. Examples of potential information pieces to align: Aligning patient medical data, heredity data, lifestyle data
Graph Creation	Create vertices for each Patient, containing a set of attributes such as medical, heredity, and lifestyle data associated with patient's current medical condition and RAF score.
Feature Enrichment	Enriching the data with new more meaningful features can be very useful. • Example 1: a composite metrics built around frequencies of treatments and latency in between treatments • Example 2: a metric capturing change in effectiveness of medications based on their specific order of prescription
Enabled Analysis	Graph Queries • Specific sub-graph searches • Element centrality and importance measures Identifying commonalities • Clustering patients into cohorts • Anticipating potential disease or treatment Anomaly detection • Identifying anomalous behavior/cases

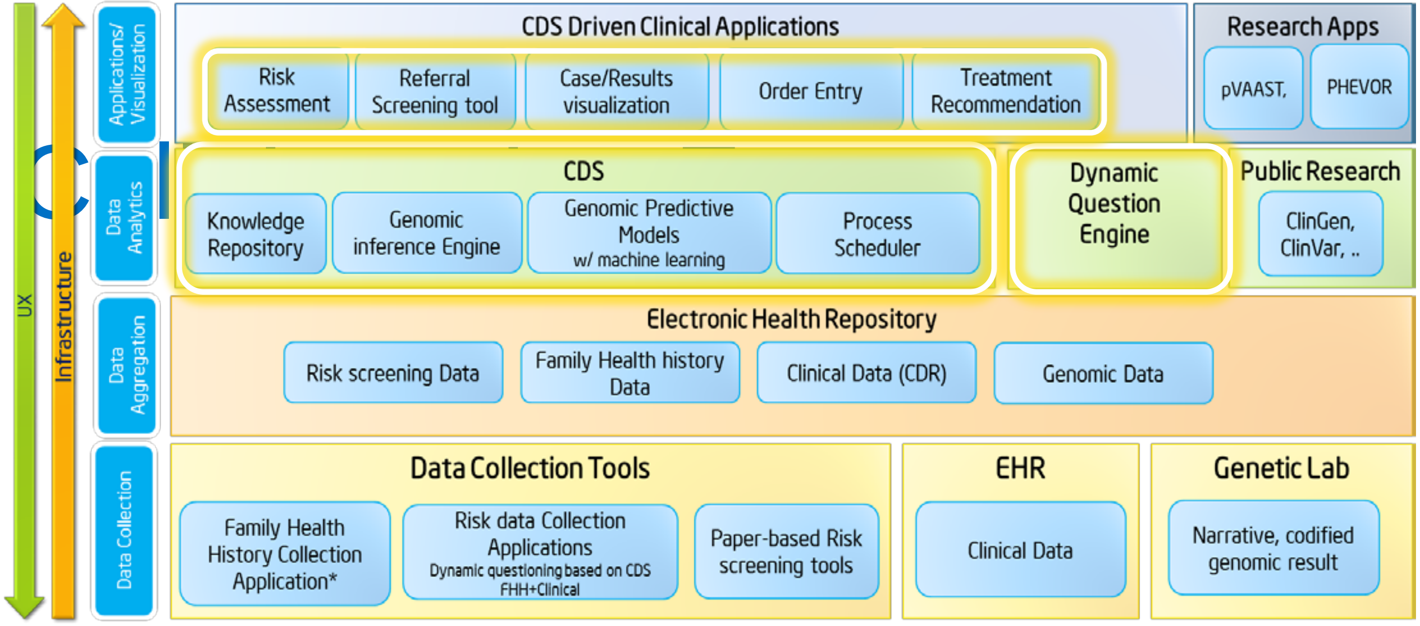
An Idea For New CDS Applications Combining Clinical, Genetic/Genomic, and Family Health History Data

- **Goal** - Promote widespread use of clinical decision support that will help clinicians/counselors in assessing risk and assist genetic counselors in ordering genetic tests.
- Build a scalable CDS that leverages standardized data that includes:
  - Family Health History
  - Clinical and Screening
  - Genomic data

- **The solution will:**
- Be **agnostic to data collection tools**. The solution Be **scaled** to different **clinical domains** (grow beyond Breast Cancer) and other **healthcare institutions**.
- Be **standards based** where they exist
- **Work across all EHRs**, but starting with Cerner.
- Leverage Intel technologies (infrastructure, Intel Data Platform etc.).
- Be flexible to incorporate other data sources (e.g. Imaging data, personal device data)

Sample Clinical Workflow





Baseline infrastructure

- Risk Screening Applications
- Structured and coded Family health History data
- Data Mart that combines FHH, Clinical, and Genomic data

Solution Considerations

Utilize the following where appropriate:

1. Health Services Platform (HSPC), HealtheDecision, Open CDS
2. Intel Data Platform for Machine Learning, Graph Analytics, Mining
3. HL7 standards, FHIR + SMART Apps for clinician facing applications

Demo – Finding Relationships Using Graph Query & Visualization

Scenario: Explore relationship trends over time using Medline Data Set

Challenges: Complexity, query times, too many tools, scalability

Solution: **Intel® Data Platform Analytics Toolkit** end-to-end graph processing capabilities

Demo

Parse and Extract Categories	Transform To Standard Ontology	Engineer Feature Co-occurrence	Build Graph - Link Terms by Co-occurrence	Query/ Visualize Trends	Graph Learning to Detect Clusters
• Medline™ XML (article) • MeSH Ontology XML (terms)	• Left Join article to term • Filter out non-matches • Filter self-matches • Filter malformed data	• Inner Join Article/Terms to Itself • Filter duplicates • Aggregate by year • Compute count by year	• Extract vertices • Assign year & count as edge properties • Assign terms to vertex property	• Gremlin query per visual • Data flow into Python web server & other libraries	• Select optimization parameters • Invoke LDA algorithm

Ingest & Clean → Engineer Features → Structure Graph → Load Data → Query & Analyze

Learn

Visualize

Training Programs

Bioinformatics, Life Sciences,

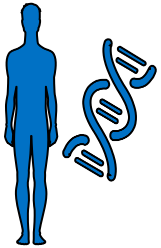
Computer Sciences, Clinicians

Day 1 - Sep 3	NGS data analysis workflows overview, bottlenecks and optimization solutions			
8:30 - 9:00	Registration			
9:00 - 9:30	Overview of workshop	Lecture	Nick Luscombe	Crick
09:30 - 10:15	Workload Characterization & Optimization Tradeoffs	Lecture	Chris Dagdigian	Intel BioTeam
10:15 - 10:45	Coffee break			
10:45 - 11:30	Introduction to Parallelism	Lecture	Clay Breshears	Intel
11:30 - 13:00	Lunch (not provided)			
13:00 - 13:15	Welcome by Jim Smith		Jim Smith	Crick
13:15 - 14:00	Memory, I/O or CPU constraints	Lecture	Clay Breshears	Intel
14:00 - 15:30	Diagnostic Tools	Practical	Clay Breshears	Intel
15:30 - 15:45	Coffee break			
15:45 - 16:45	Mapping strategies overview	Lecture	Ernest Turro	CRUK Cambridge Institute
16:45 - 17:45	System Architecture & Technology Options	Lecture	Chris Dagdigian	Intel BioTeam
17:45-18:15	System Architecture Scavenger Hunt	Practical	Clay Breshears	Intel
18:15 - 18:30	Q&A session		Intel/Crick	
18:30 onwards	Drinks reception			

Day 2 - Sep 4	Mapping			
9:00 - 10:00	Introduction to RNA-seq analysis	Lecture	Vincent Plagnol	UCL
10:00 - 10:15	Coffee break			
10:15 - 11:15	Thread and Process Level Optimizations	Lecture	Clay Breshears	Intel
11:15 - 12:30	Thread and Process Level Optimizations	Practical	Clay Breshears	Intel
12:30 - 14:00	Lunch (not provided)			
14:00 – 15:00	Data Latency, Data Chunking & Placement	Lecture	Clay Breshears	Intel
15:00 – 16:00	Data chunking	Practical	Clay Breshears	Intel
16:00 - 16:15	Coffee Break			
16:15 - 17:30	Data chunking (continued)	Practical	Clay Breshears	Intel
17:30 - 18:00	Q&A session		Intel/Crick	

Day 3 - Sep 5	R optimization			
9:00 - 10:00	Debugging and Profiling in R	Lecture	Robert Sugar	Crick
10:00 - 10:15	Coffee break			
10:15 - 12:30	R-based Optimization	Practical	Robert Sugar/ Kathi Zarnack	Crick
12:30 - 14:00	Lunch (not provided)			
14:00 - 14:45	SPRINT Overview	Lecture	Eilidh Troup	SPRINT Team
14:45 - 15:15	Coffee break			
15:15 - 16:30	R-based Optimization	Demo	Eilidh Troup	SPRINT Team
16:30 - 17:00	SGI UV2 with Xeon Phi	Lecture	Simon Appleby	SGI
17:00 - 17:30	Q&A session		Crick/Intel/SPRINT	

Day 4 - Sep 6	Bench-marking and scaling up			
9:00 - 10:00	On the empirical evaluation of RNA-seq gene profiling pipelines	Lecture	Nuno Fonseca	EBI
10:00 - 10:30	Coffee break			
10:30 - 11:30	Pipelines for large RNA-sequencing projects	Lecture	Steve Searle	Sanger Institute
11:30 - 13:00	Lunch (not provided)			
13:00 - 14:00	Cloud-based Analytics & Map Reduce	Lecture	Ketan Paranjape	Intel
14:00 - 16:00	Cloud-based Analytics & Map Reduce	Practical	Ketan Paranjape	Intel
15:30 - 15:45	Coffee break			
15:45 - 16:45	Scaling up to Production	Lecture	Ketan Paranjape	Intel
16:45 - 17:15	Q&A and wrap up		Crick/Intel	



Genes causing it identified & disease pathways determined

[redacted]on
me[redacted]egime
[redacted]…

Personalized Medicine: All in a day by [redacted] 2020 intel®

Look Inside.™

Clinical Barriers:

- Sample (conservation & validity)
- Labs (validity & repeatable)
- Biomarkers (novelty & interpretation)
- Merging with clinical data (system & EHR)
- Actionability (Pathways & Preventative)
- New Taxonomies & Training

Literacy and societal challenges:

Trust/Trustworthiness

Literacy

- Health Care professionals
  - (lack of) Genomic Medicine Specialists
  - (need for) Statistics and Epidemiology Training
- Citizens
  - (need for increased) General Understanding
  - (unknown) Impact on people's life choices

Transparency/Accountability

- Ownership
  Custodianship
  Access
- Who Benefits?
  - Equality and Justice
  - Data Usage & Revenue

Economic & Commercial barriers:

- Insurance
  - Reimbursement & pricing
  - Insurability
  - [Discrimination]
- Economic Sustainability
  - Health care system (preventative shift)
  - Ownership and regulations for use and profit
  - Pharma / Biotech industry sustainability
    - RCT
    - IP
    - Companion Diagnostic

Ethical barriers:

Clinical Research

- Informed Consent
- Communicating findings of unknown significance (right to know/right not-to-know)

Privacy

- Security & regulations across GEOs
- Anonymisation of WGS
- Confidentiality for *personal* information

Discrimination

- [GINA in US]
- Management of *sensitive* information (potential impact)

Targeted Usages

Areas	Benefits, solutions
Utilization and Treatment Analysis	Combine trends with individual treatment analysis
Treatment Effectiveness	Build large-scale treatment effectiveness monitoring
Diagnosis and Treatment correlation	Discover diagnosis/treatment connections
Managed Care Optimization	Optimize resources for managed care
Diagnosis Treatment and Trends and Predictions	Determine overall trends, but put them at the disposal of individual diagnosticians
Drug Utilization and Expense Prediction	Build dynamic precise drug utilization prediction models
Treatment and Outcomes Analysis and Optimization	Predict treatment prognosis, optimize based on individual's complete picture
Demand Forecasting	Better demand preparedness
Price Analysis and Determination	Optimize quality and revenue through price monitoring
Epidemiology Research	Discover trends by analyzing data from disperse sources
Provider Ratings and Benchmarking	Use all source of data to benchmark and monitor providers
Patient History and Digital Records Archiving and Analysis	Combine patient history records from disparate sources, greatly improve the quality of patient care
Contract Optimization	Optimize contract resource utilization

Actionability and Data driven medicine

- Biomarkers need to be proven to produce better clinical (efficacy) and economic (efficiency) measurements before they are introduced in clinical best practice;

- Browsing EHR as if they were research database to identify possible improvements presents challenges due to lack of consent and of validated pathways for the results (Denny 2012);

Denny 2012: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002823

Actionability and Data driven medicine

- There are difficulties in communicating to patients results that are still being researched (uncertain actionability);
- Finally, there are some extra challenges in curating changing interpretations through time and re-contacting patients once variants have become clinically valid or have changed significance;

Denny 2012: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002823

Change of paradigm:

- New taxonomy of diseases need to be created (these might cut across several traditional disciplines);
- Primary and secondary care are designed to be reactive rather preventative; WGS for screening and susceptibility would push health care towards prevention, something the current system is not necessarily set up to deal with;
- Health care professionals need training to interface with these information (see also literacy section);

Education, trust and society

- There are too few clinical genomic specialists;
- Health care professionals (including pharmacists) are not trained in clinical genomics > shift required in the way they practice;
- New Clinical pathways/guidelines have to be created;
- More training for the general population (and the professionals as well) in genomics, in mathematics and statistics is required to better understand how these data are interpreted;

Education, trust and society

- People are uncertain whether to trust institutions collecting and holding these data about them (access and ownership);
- Systems need to be put in place to increase transparency and accountability of different stakeholders of genomic data usage (trust);
- The impact of people’s life choices is not fully understood;
- Equality and Justice needs to be ensured (trust); protective measure against discrimination need to be in place [e.g. GINA leading the way];

Reimbursement and economic sustainability:

- Reimbursement methods need to consider flexible pricing for tailored therapeutic responses; standardisation and harmonisation are also needed;

- The health care system has to rethink the wider economic implications and sustainability of preventative (less costly and more spread over time) versus reactive health care delivery (more expensive and concentrated in short bursts);

Reimbursement and economic sustainability:

- Reimbursement methods need to consider flexible pricing for tailored therapeutic responses; standardisation and harmonisation are also needed;
- The health care system has to rethink the wider economic implications and sustainability of preventative (less costly and more spread over time) versus reactive health care delivery (more expensive and concentrated in short bursts);

Reimbursement and economic sustainability:

- The insurance industry needs to implement systems for insurability of people with several biomarkers for potential ‘risks’;
- Ownership of data requires increased transparency and strict regulations around access;
- The pharmaceutical industry needs to rethink its own economic model of developing drugs that target only small segments of the population (e.g. development of companion diagnostics) and its economic sustainability (considering patenting genes is probably not viable);

Genomic Data Merging with EHR

- Need to develop a standardised genetic terminology [HL7 has a working group on this; also GA4GH];
- Current EHR do not support browsing annotated WGS (or imaging data from radiography and pathology for that matter); speed is going to be an issue…
- Current EHR would require standardization for *communicating*, *querying*, *storing* and *compressing* large volumes of data while *interfacing* with EHRs identifiable patient information. This requires sophisticated computational tools and skills;
- Processing of structured and unstructured large volume of data requires access to HPC infrastructure; also long term storage costs have to be considered;
- Platforms would need to allow interpretation and re-interpretation of variants through time;

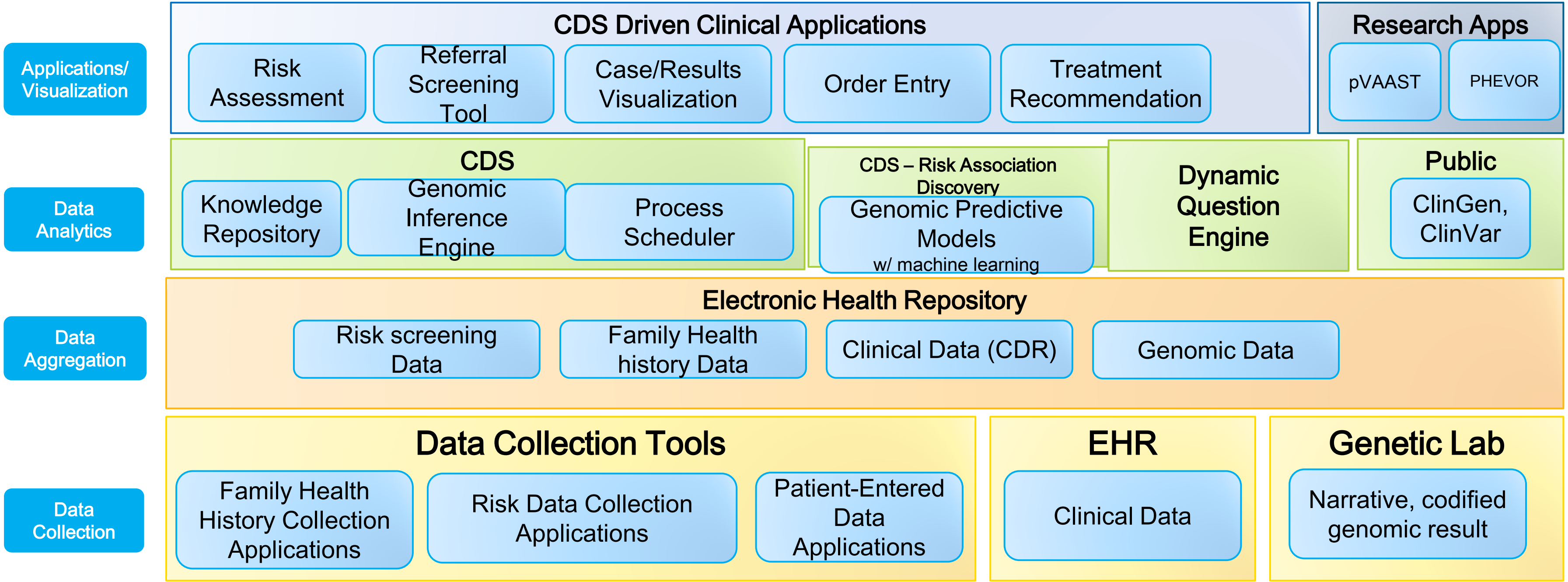
Ethics, Privacy and Discrimination:

- *Confidentiality: Personal* and *Sensitive* information: can potentially affect people *personally* (e.g. anxiety, choices about the future), *socially* (e.g. stigmatization, involvement of family members), *economically* (e.g. increase costs for long term screening), and *professionally* (e.g. discrimination);
- *Consent*: several issues with consent consistency and harmonisation. In addition, informed consent for WGS studies presents some extra challenges (e.g. not knowing in advance the use that will be made of samples, long term consequences, need to share data to get the best value);

Ethics, Privacy and Discrimination:

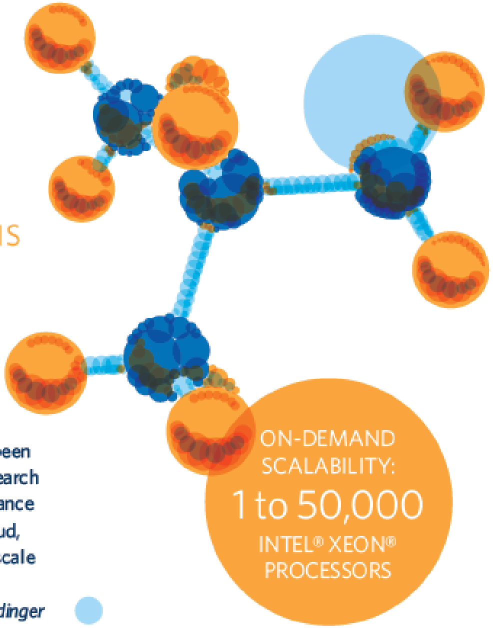
- *Security*: Highest security standards and strict regulations required due to the confidentiality and sensitivity of the data (privacy concerns);
  - the governance landscape across GEOs is fragmented;
  - WGS data cannot be anonymised like it is generally done for similarly sensitive information (sharing for research purposes);
- *Communication*:
  - Post-result counselling by health care worker is recommended (by ethical guidelines) and could encounter a bottleneck (see training);
  - Challenges in communicating findings of unknown or uncertain significance (balance between right-to-know and right-not-to-know);

Component View



High Throughput Science: Embracing Cloud-based Analytics

IN 60 MINUTES,
SCHRÖDINGER
CAN TEST 16 MILLION MOLECULES, SO PHARMACEUTICAL RESEARCHERS CAN MAKE BREAKTHROUGHS SOONER



"For years, pharmaceutical companies have been scaling up staffing resources via contract research organizations. Now, thanks to high-performance computing in the Amazon Web Services cloud, it's incredibly simple and cost-effective to scale their compute resources in a similar way."
—*Scott Becker, VP of Enterprise Products, Schrödinger*

- **Challenge:** Team of cancer researchers had to screen a drug concept with a list of tens of millions of molecules working with a tight deadline, a fixed budget, and strict security and compliance requirements. Schrödinger's* existing in-house servers would be tied up for weeks
- **Solution:** Schrödinger* leveraged software from AWS* partner, Cycle Computing*, to provision a fully secured cluster of 50,000 cores,
- **Enabled the team to run 16M molecular simulations an hour**
  - **Developed target list of 1000 molecules in <8hrs**