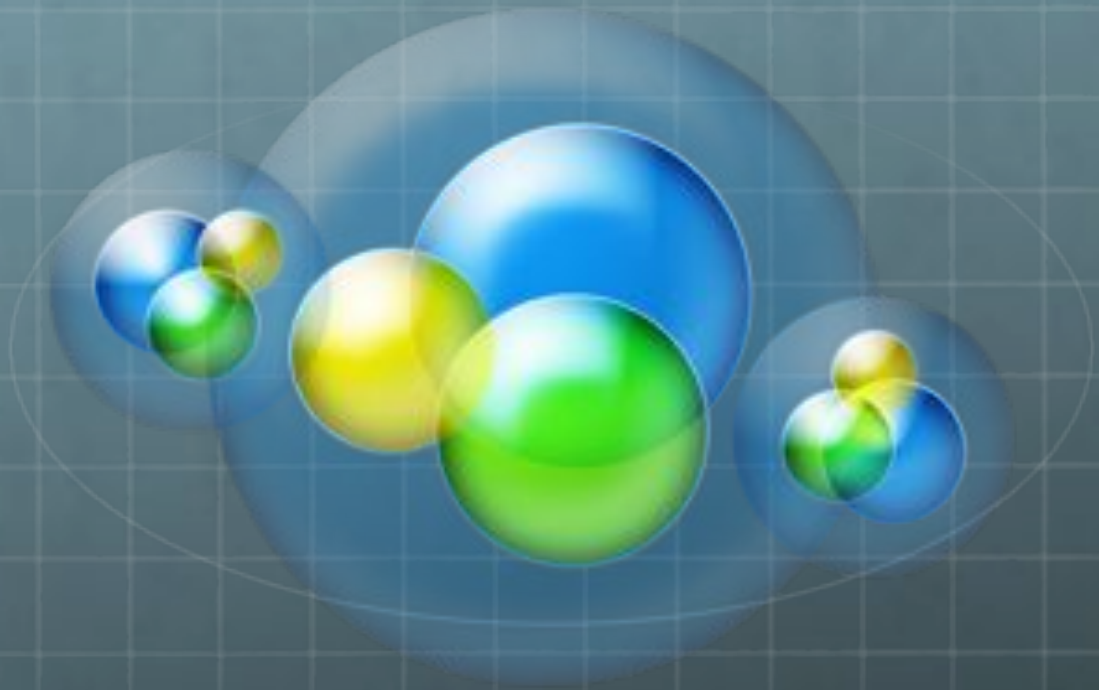
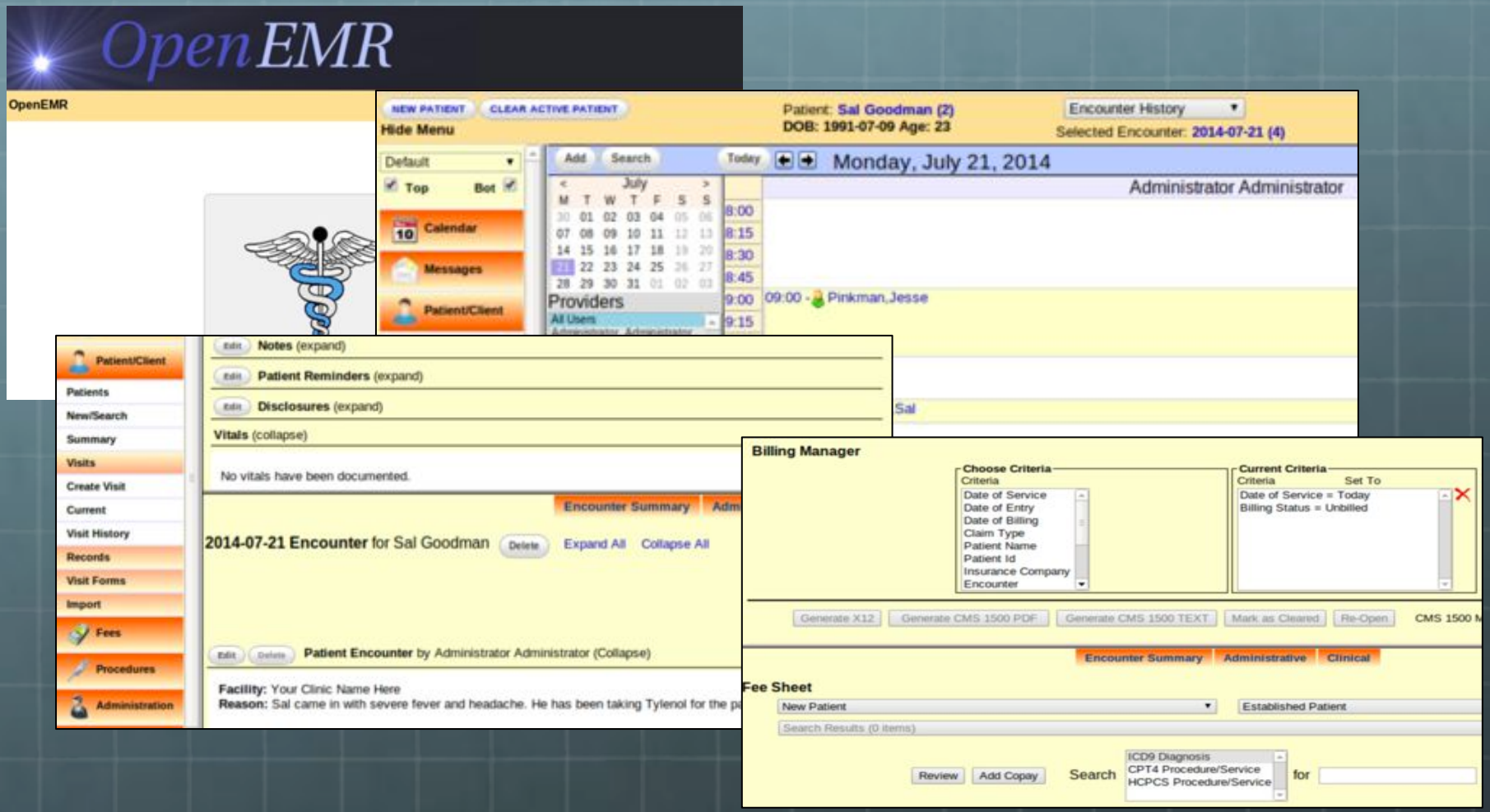
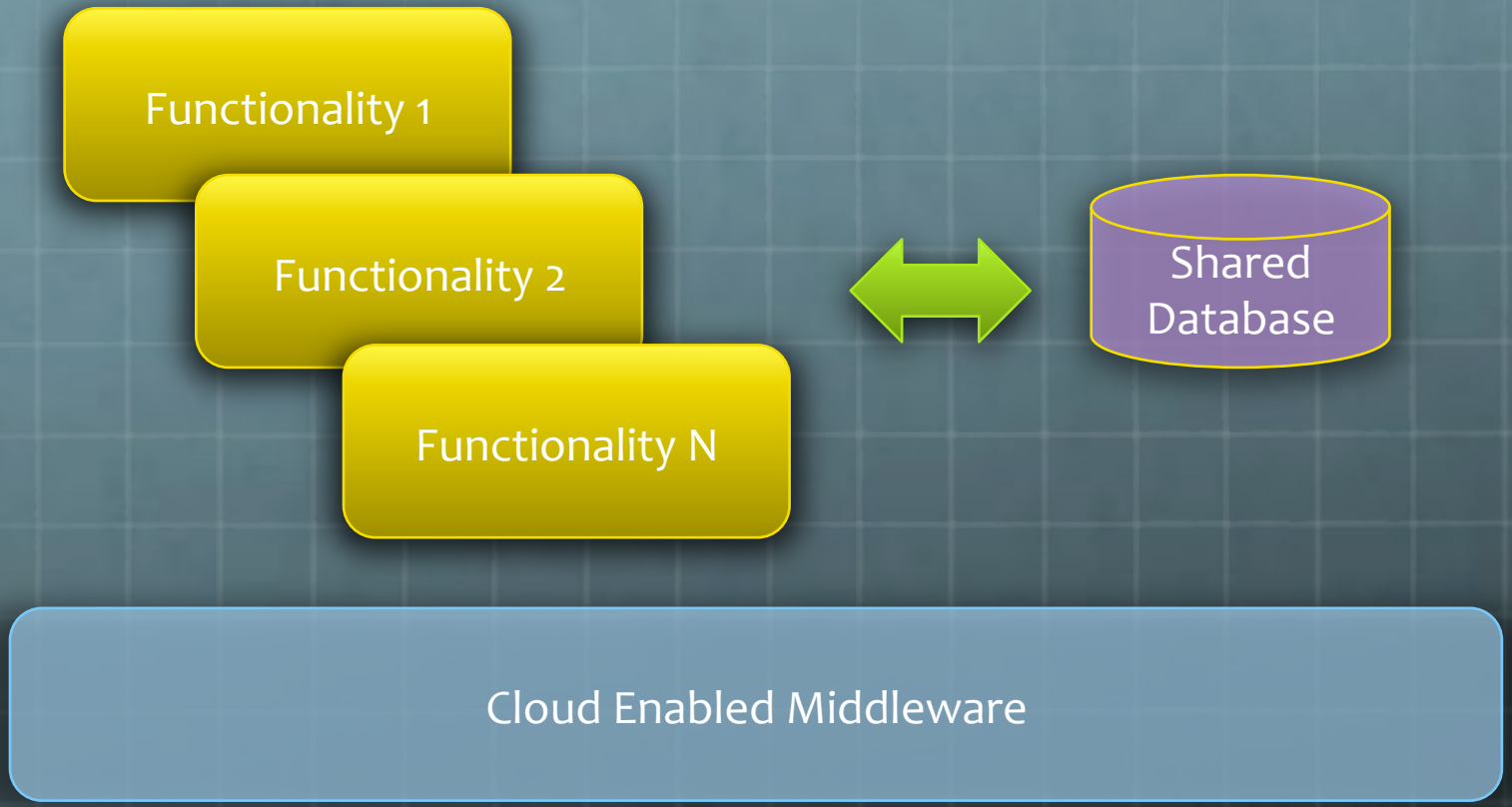


Modernizing HealthCare Applications with Micro Services

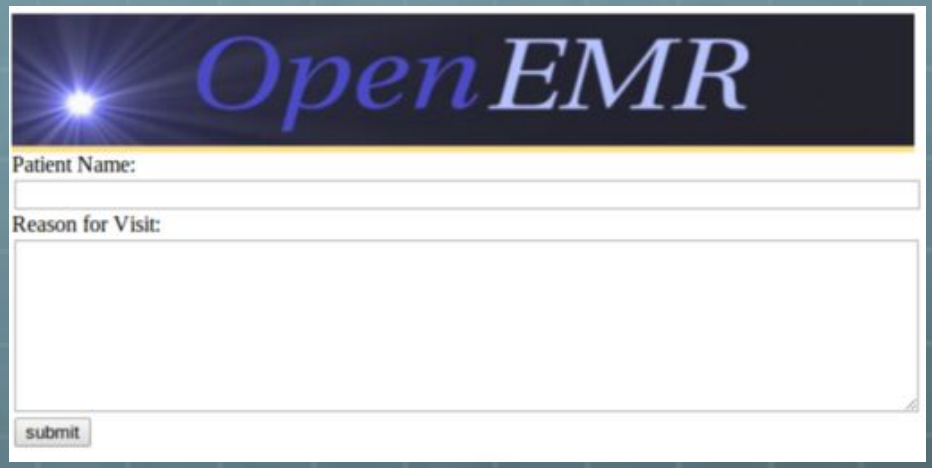
Edison Ting
Pivotal Solutions Architect



- Health Electronic Medical Record tool OpenEMR provides a lot of functionality in one Monolithic Application
- Monolithic architecture makes it difficult to manage, update and add new functionality

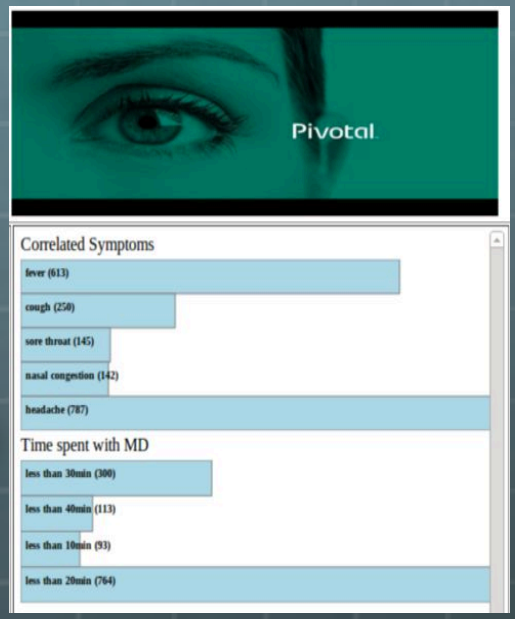


- Micro services architecture can help with manage-ability of distinct applications and services, and facilitate analytics driven model
- Cloud enabled runtime can help with software lifecycle and application deployment and movement between infrastructures



Cloud Enabled Middleware

- An patient portal application can be built separately that can service many users, such as allow patient themselves to schedule visits
- Such applications can be managed and scaled independently across different instances, all sharing one central database



Cloud Enabled Middleware

- Shared patient information can then be analyzed globally to derive actionable insights
- Live patient information can be correlated with historical data or public information to discover useful relationships between similar patients

DATA SCIENCE AND BIG DATA ANALYTICS

An 'open' course to unleash the power of Big Data

COURSE OVERVIEW

The Data Science and Big Data Analytics course educates students to a foundation level on big data and the state of the practice of analytics. The course provides an introduction to big data and a Data Analytics Lifecycle to address business challenges that leverage big data. It provides grounding in basic and advanced analytic methods and an introduction to big data analytics technology and tools, including MapReduce and Hadoop. The course has extensive labs throughout to provide practical opportunities to apply these methods and tools and includes a final lab in which students address a big data analytics challenge by applying the concepts taught in the course in the context of the Data Analytics Lifecycle. Upon completing the course, students will have the knowledge and practical experience to immediately participate effectively in big data and other analytics projects.

THE DATA SCIENCE AND BIG DATA ANALYTICS COURSE CONSISTS OF 7 MODULES:

Module 1: Introduction to Big Data Analytics
This module focuses on definition of and an overview of big data, the state of practice of analytics, the Data Scientist role, and big data analytics in industry verticals.

Module 2: Overview of Data Analytics Lifecycle
This module focuses on the explaining the various phases of a typical analytics lifecycle – discovery, data preparation, model planning, model building, communicating results and findings, and operationalizing. This module also details the critical activities that occur in each phase of the lifecycle.

Module 3: Using R for Initial Analysis of the Data
This module focuses on an introduction to R programming, initial exploration and analysis of the data using R, and basic visualization using R. This module includes hands-on labs to familiarize students with the concepts taught.

Module 4: Advanced Analytics and Statistical Modeling for Big Data – Theory and Methods
This module focuses on the core methods used by a Data Scientist, including candidate selection using the Naïve Bayesian Classifier, categorization using K-means clustering and association rules, predictive modeling using decision trees, linear and logistic regression, and time-series analysis, and text analysis. This module includes hands-on labs to familiarize students with the concepts taught.

EMC² PROVEN PROFESSIONAL

Module 5: Advanced Analytics and Statistical Modeling for Big Data – Technology and Tools
This module focuses on analytic tools for unstructured data, including MapReduce and the Hadoop ecosystem. It also details in-database analytics with SQL extensions and other advanced SQL techniques and MADlib functions for in-database analytics. This module includes hands-on labs to familiarize students with the concepts taught.

Module 6: Concluding and Operationalizing an Analytics Project
This module focuses on identifying the core deliverables and creating them for key stakeholders and others. This module also details how to emphasize key points using visualization methods.

Module 7: Big Data Analytics Lifecycle Lab
This module focuses on the student's practical application of their learning to a big data analytics challenge in the context of the data analytics lifecycle.

Faculty profile for success

Faculty who have been teaching courses on following topics will have added advantage in successfully teaching this course:

1. Computer Science
2. Mathematics, Statistics and Statistical Modeling

Student profile for success

Students who have completed courses on following topics will have added advantage in comprehending the learnings of CIS course:

1. Computer Science
2. Information Technology
3. Engineering
4. Statistics and Statistical Modeling
5. Mathematics
6. Database Administration and Data Warehousing
7. Computer Programming
8. Econometrics
9. Biostatistics
10. Physics

The knowledge you gain through the Data Science and Big Data Analytics 'open' course can be applied to impact business decisions in a variety of ways

	Key activities	Business Impact
1	Define big data and the business drivers for advanced, big data analytics.	A solid understanding of big data and the business opportunities that advanced analytics applied to big data represent is essential for stakeholders to identify and drive big data analytics opportunities within their own organizations.
2	Describe why and how Data Science is different to traditional Business Intelligence.	Data Scientists must understand the Business Intelligence world just as Business Intelligence analysts need to understand the Data Science world so they can work together in cohesive teams to ensure the business is gaining optimum value from leveraging big data and data in traditional data warehouses.
3	Describe the roles and skills required in a big data analytics team.	Business and IT stakeholders need to recruit suitably skilled individuals and grow the skills of others to create competent and effective big data analytics teams.
4	Explain the phases and activities of the data analytics lifecycle and identify the main activities and deliverables.	Provides a framework for executing data analytics projects in a repeatable way that will consistently lead to valuable and actionable insights for the business.
5	Explore and make an initial analysis of the data, using R.	Develop a quick overall understanding of the nature and characteristics of the data, using simple R programming. This drives creation of initial hypotheses regarding potential relationships within the data that can then be explored using more advanced analytic methods.
6	Select and execute appropriate advanced analytic methods for candidate selection, categorization, and predictive modeling.	Detailed analysis of the data requires selection of the advanced analytic methods that are most appropriate for the business challenge being addressed and the data being analyzed.
7	Describe the challenges and tools for analyzing text and other unstructured data.	Less than 20% of all data is structured. Text and other unstructured data are key data sources for big data analytics. Data Scientists must understand the challenges of analyzing this data and the different approaches (e.g. MapReduce) and tools (e.g. Hadoop) used to analyze it.
8	Describe the importance and benefits of advanced techniques such as in-database analytics and how extensions and other advanced functions add value.	Encourages interest in newer technology developments that can bring potential analytic benefits to the rapidly developing field of Data Science.
9	Plan the creation of effective final deliverables for a data analytics project that will meet the needs of stakeholders and others.	Business stakeholders and others must be convinced by the analysis, conclusions, and recommendations emerging from a data analytics project. Creating the final project report is a key opportunity to ensure commitment to action and to communicate the tasks necessary to operationalize those recommendations.
10	Apply all the phases of a data analytics lifecycle to a big data analytics challenge.	Demonstrates the ability to be successful in taking a big data business challenge through all phases of the data analytics lifecycle as a Data Scientist practitioner and deliver actionable insights.

EMC Corporation
Hopkinton, Massachusetts 01748-9103
1-508-435-1000 In North America 1-866-464-7381
www.EMC.com

EMC²

Ecosystem Challenges Around Data Use

Leonid Zhukov

ancestry.com

Ancestry.com

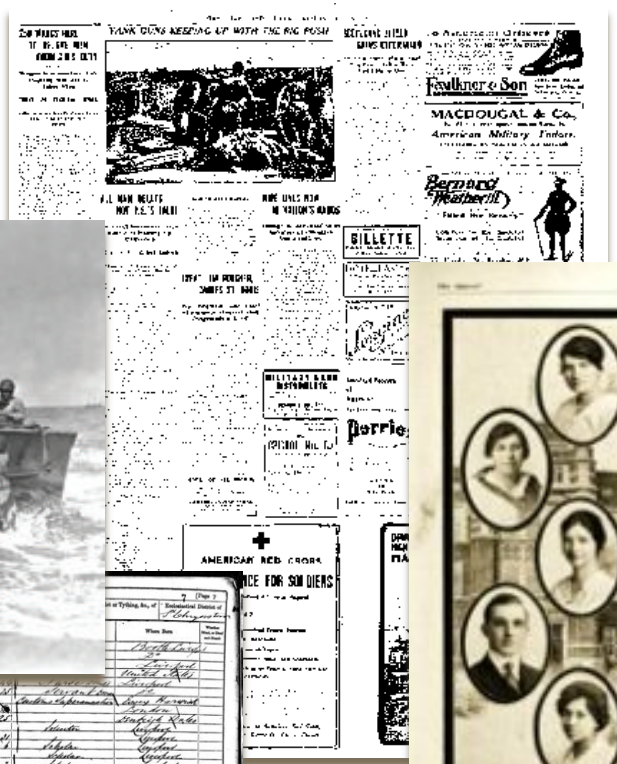
- World's largest online family history resource
- Started as a publishing company in 1983, online from 1996
- 2.7 million worldwide subscribers

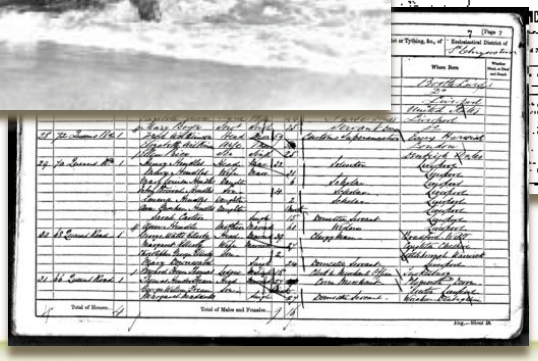
Data at Ancestry

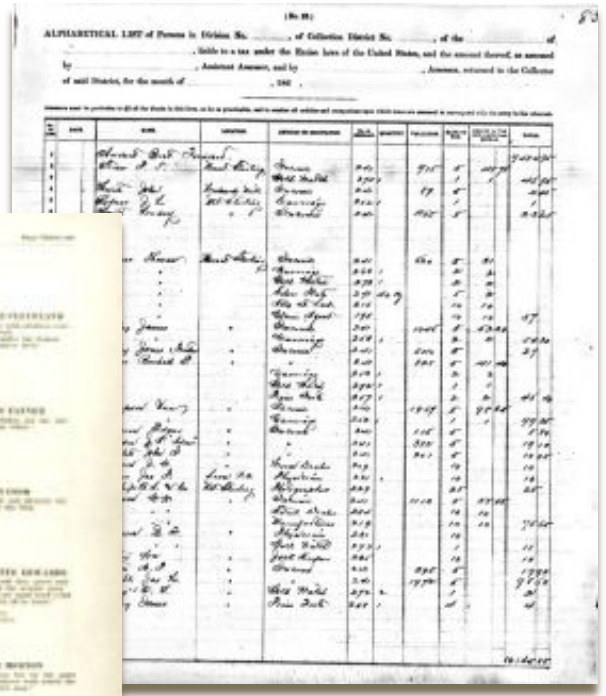
- Historical records – company acquired content collections
- User created content:
  - Ancestor profiles and family trees
  - Uploaded photographs and stories
- User behavior data on Ancestry.com
- Customer DNA data
- 10 PB of structured and unstructured data

Historical records

- Historical Content
  - 14 billion historical records going back to 17th century
  - Digitized and searchable







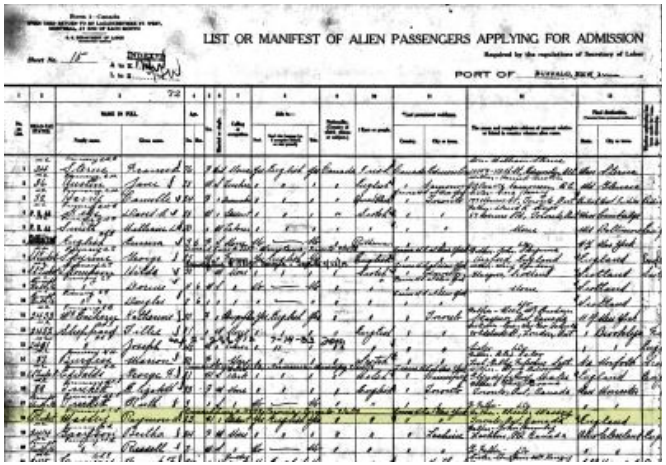
Historical records

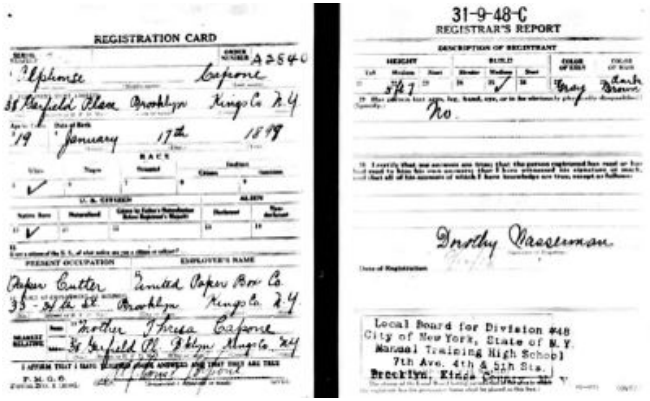
- More than 30,000 content collections

Birth, Marriage & Death
Birth, Baptism & Christening
Marriage & Divorce
Death, Burial, Cemetery & Obituaries

Census & Voter Lists
U.S. Federal Census Collection
UK Census Collection
Canadian Census Collection
More...

Immigration & Travel
Passenger Lists
Citizenship & Naturalization Records
Border Crossings & Passports
More...





Military
Draft, Enlistment and Service
Casualties
Soldier, Veteran & Prisoner Rolls & Lists
More...

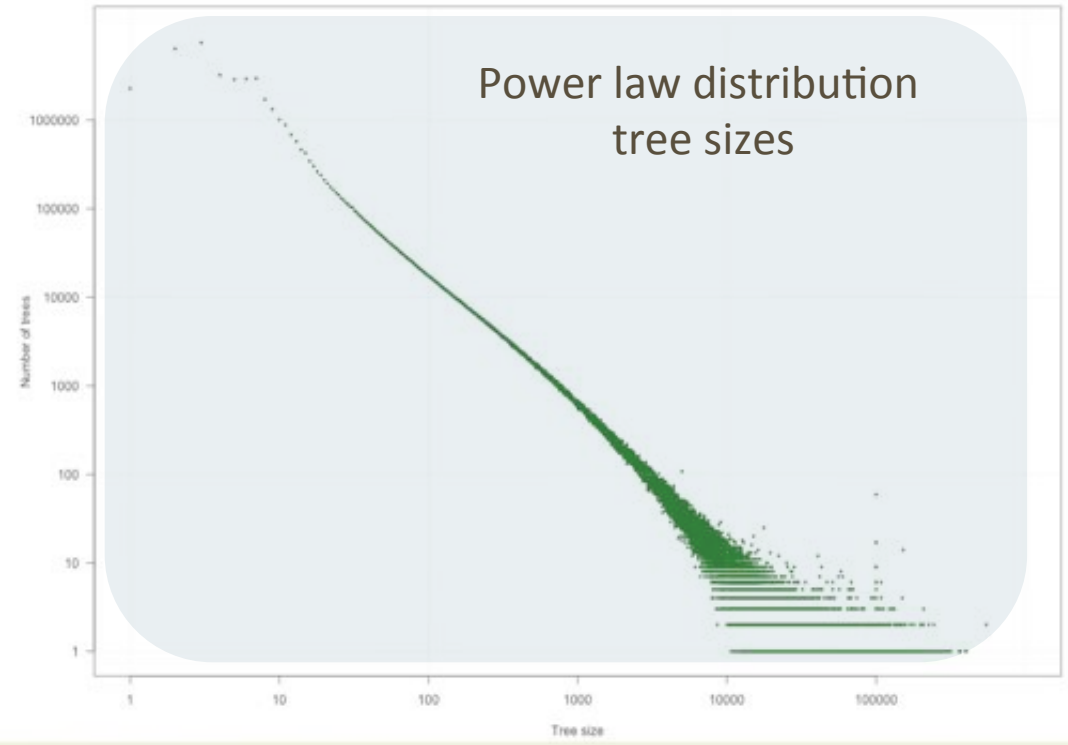
Schools, Directories & Church Histories
City & Area Directories
Professional & Organizational Directories
Church Histories & Records
More...

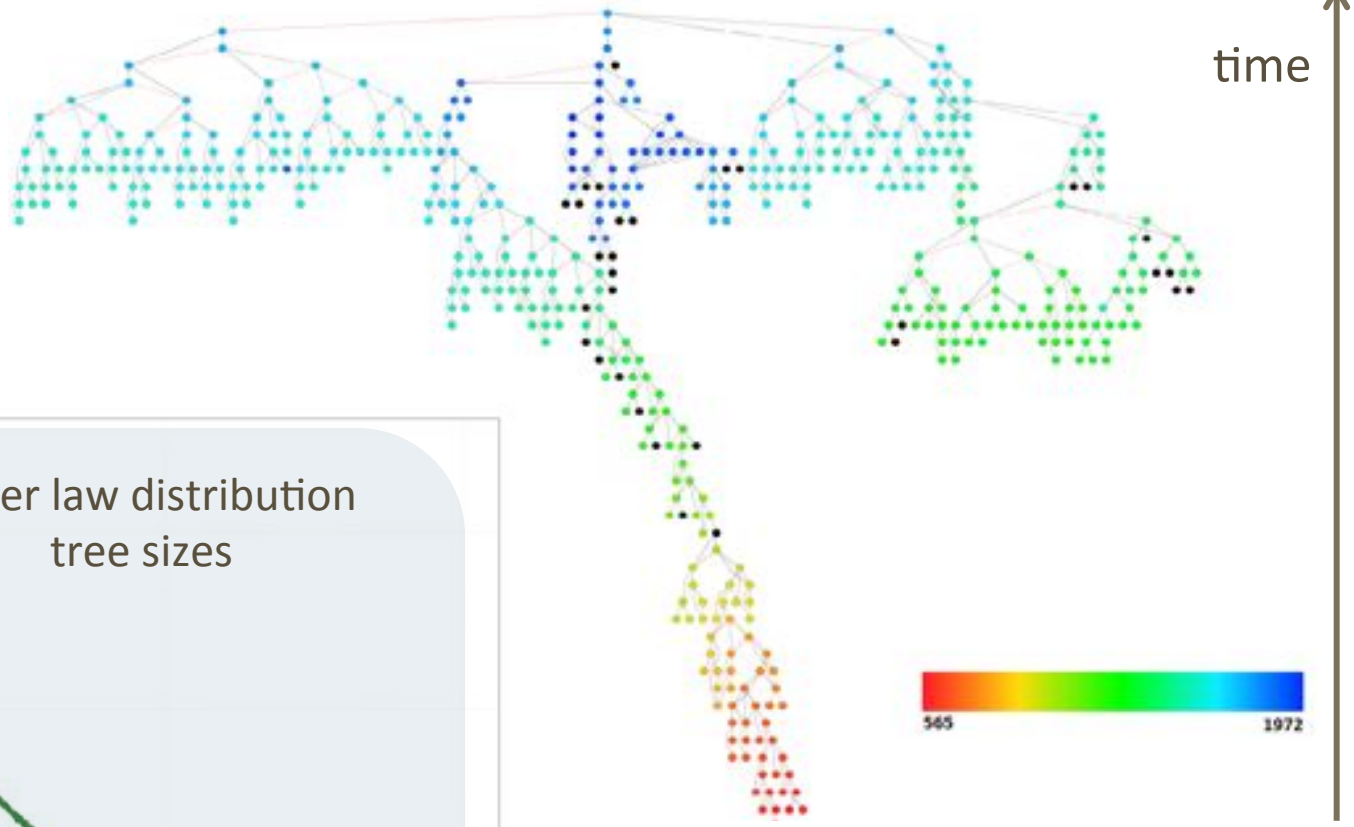
Tax, Criminal, Land & Wills
Land Records
Tax Lists
Court, Governmental & Criminal Records
More...

User family trees

- Family trees:
  - 60 million family trees
  - 6 billion profiles

Family trees





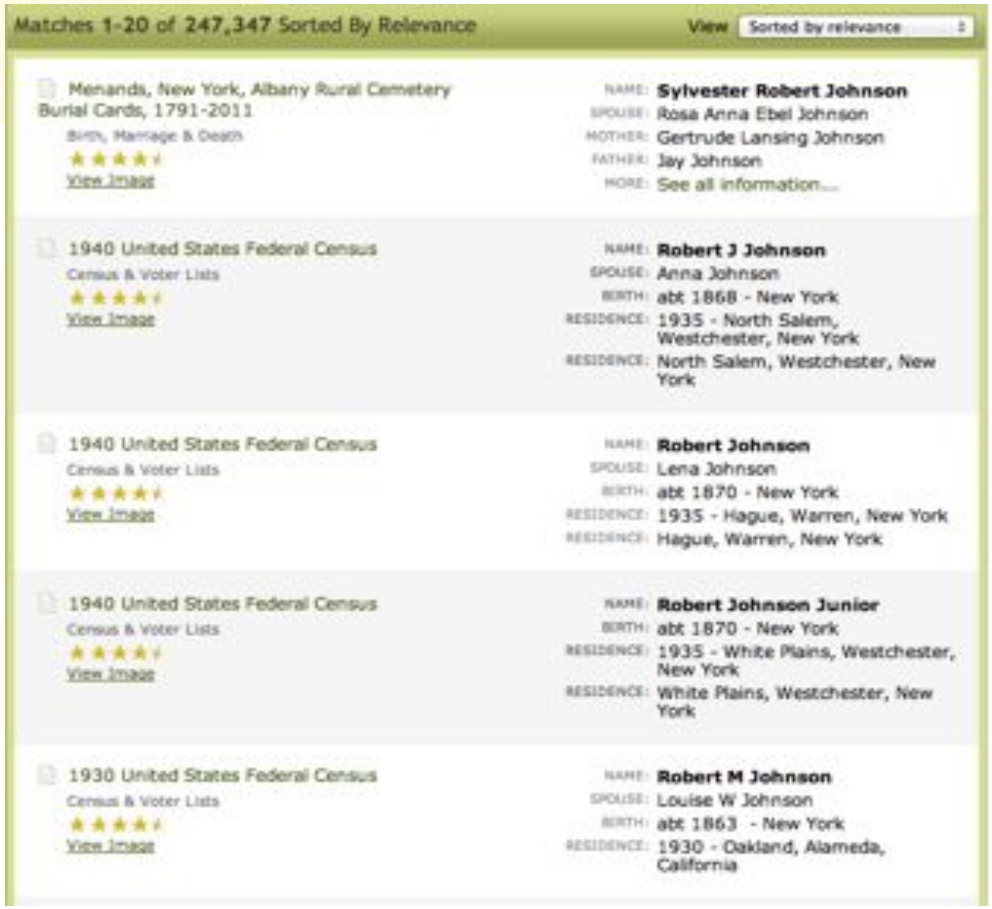
500 nodes
700 edges
55 generations

User contributed content

- 200 million uploaded family photos and stories

Person and record search

- Search query

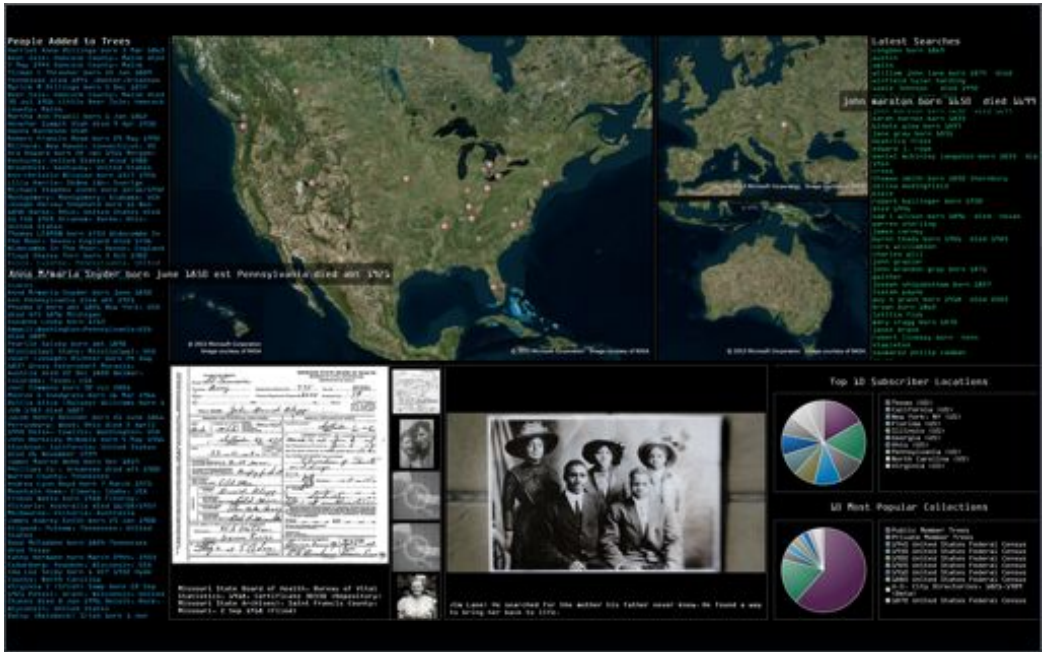


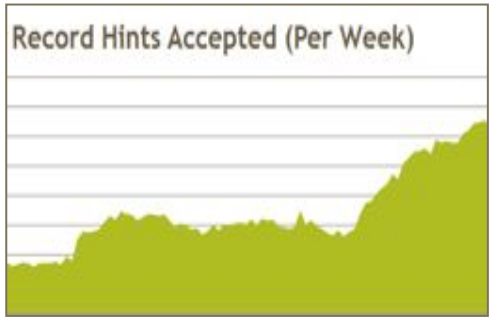
Record linkage

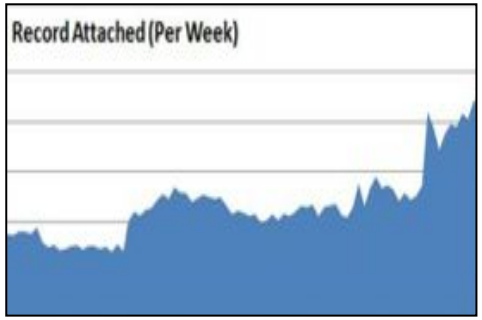
- Record linkage – finding and matching records in multiple data sets with *non-unique* identifiers (data matching, entity disambiguation, duplicate detection etc)
- Goal: bring together information about the same person
- Some *non-unique* identifiers:
  - Names: first name, last name (John Smith – 300,000 records)
  - Dates: date of birth, date of death
  - Places: place of birth, residence, place of death
  - Extra: family members, life events
- Records often *incomplete* and contain *mistakes*
- Other industries: banking, insurance, government etc

User behavior data

- User behavior data:
  - 75 mln searches daily
  - 10 mln profiles added daily
  - 3.5 mln records attached daily

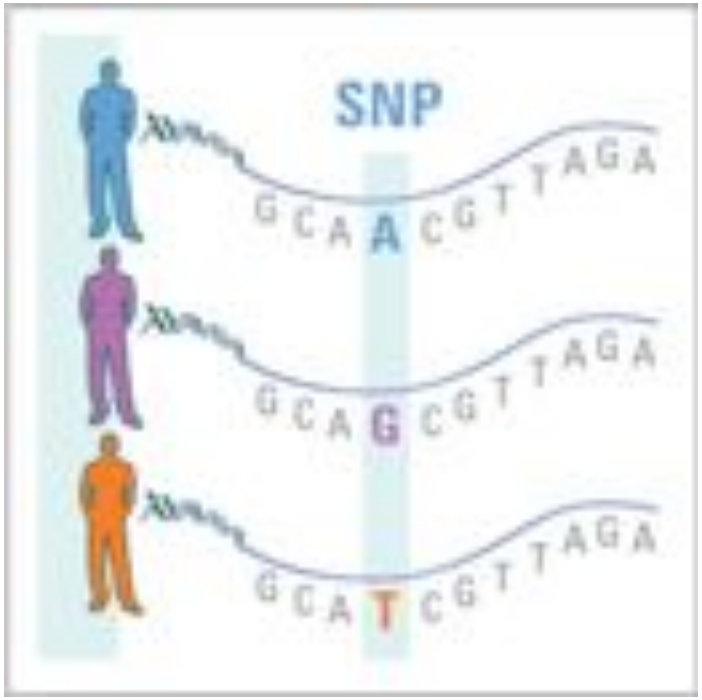






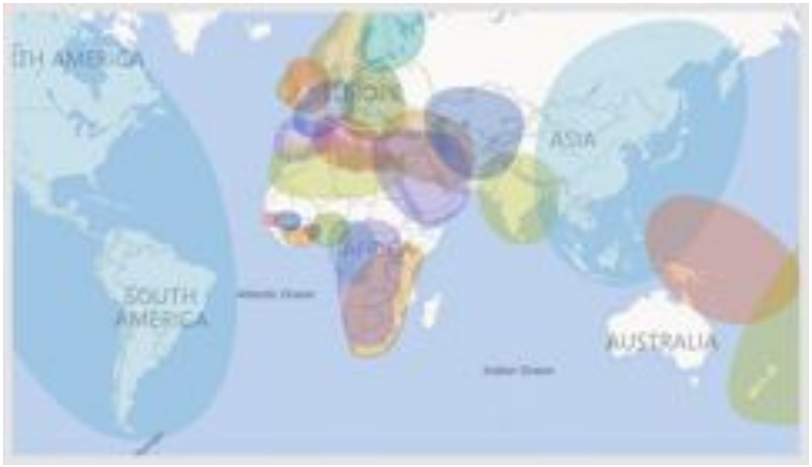
DNA Data

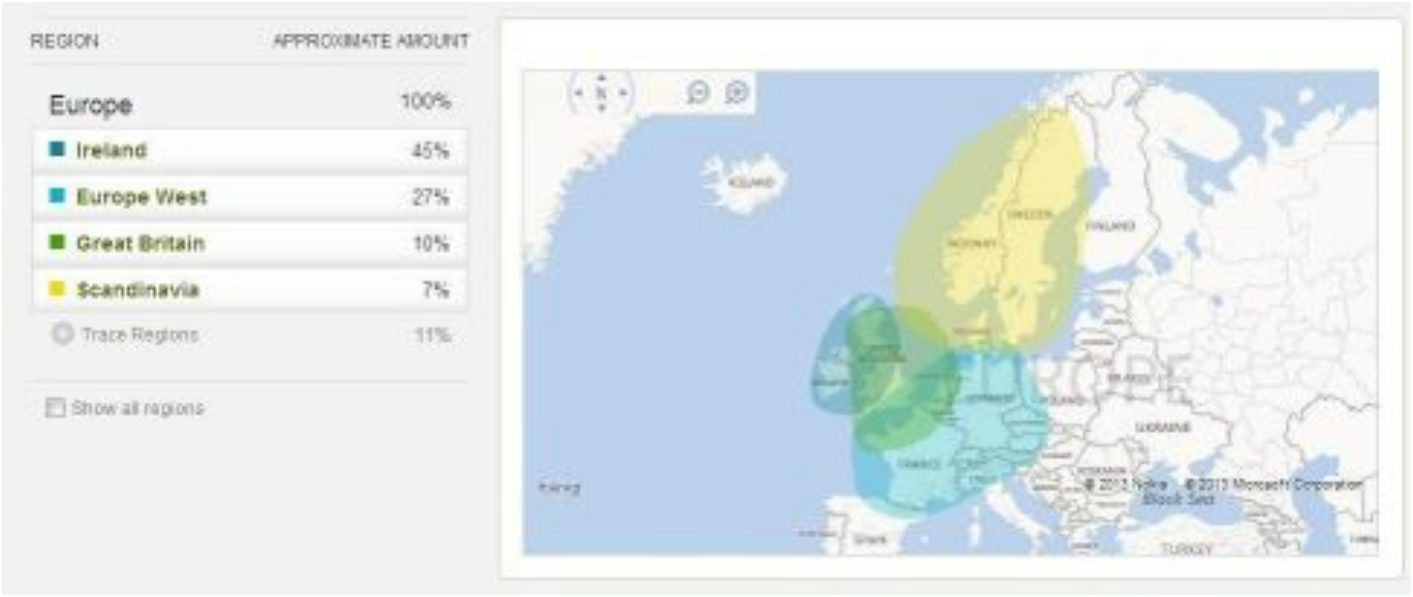
- Direct to consumer DNA test
- 700,000 SNPs per sample
- 400,000 DNA samples
- No medical studies



Ancestry DNA

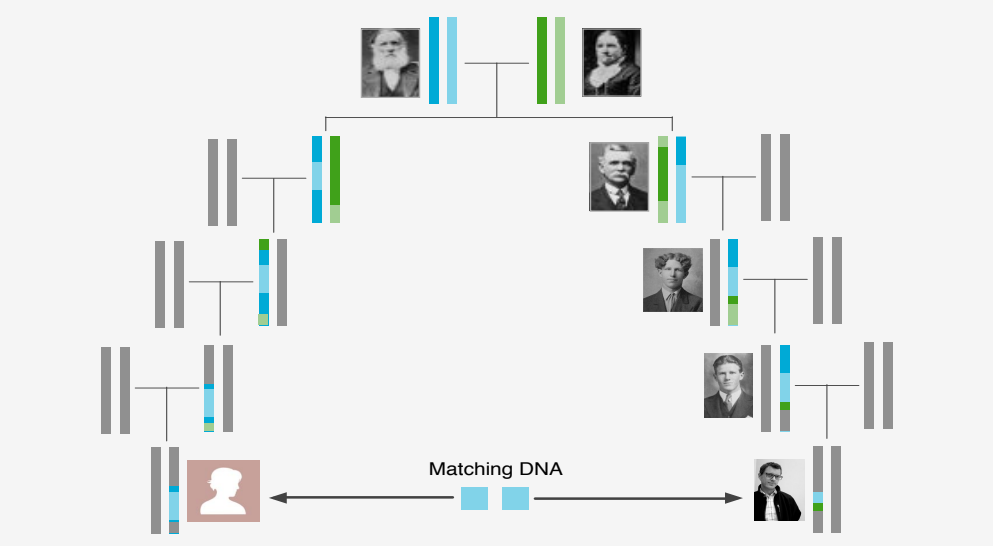
- Genetic ethnicity
  - Reference panel
  - 26 ethnic regions, 3000 samples

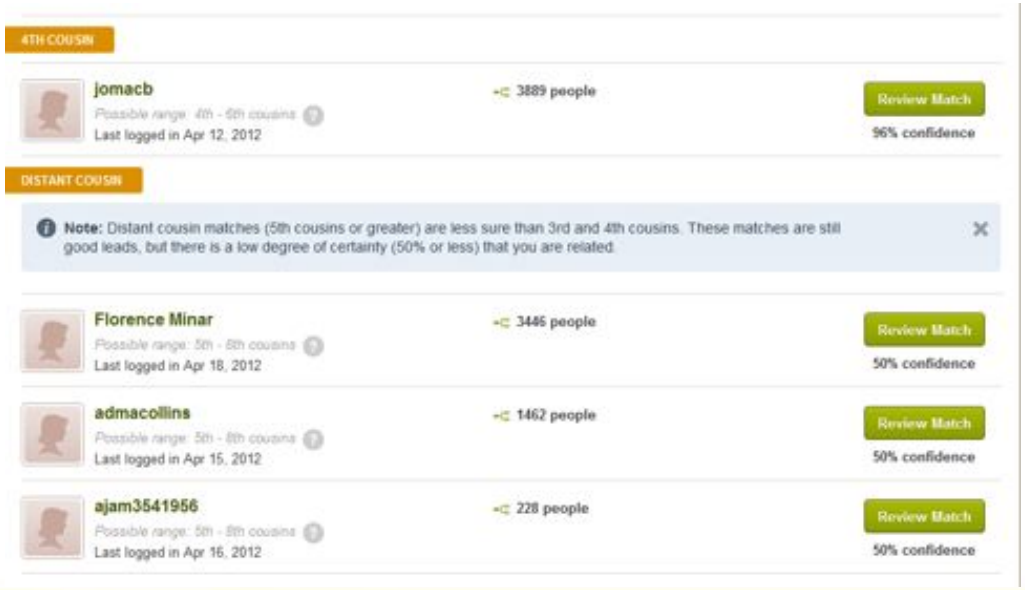




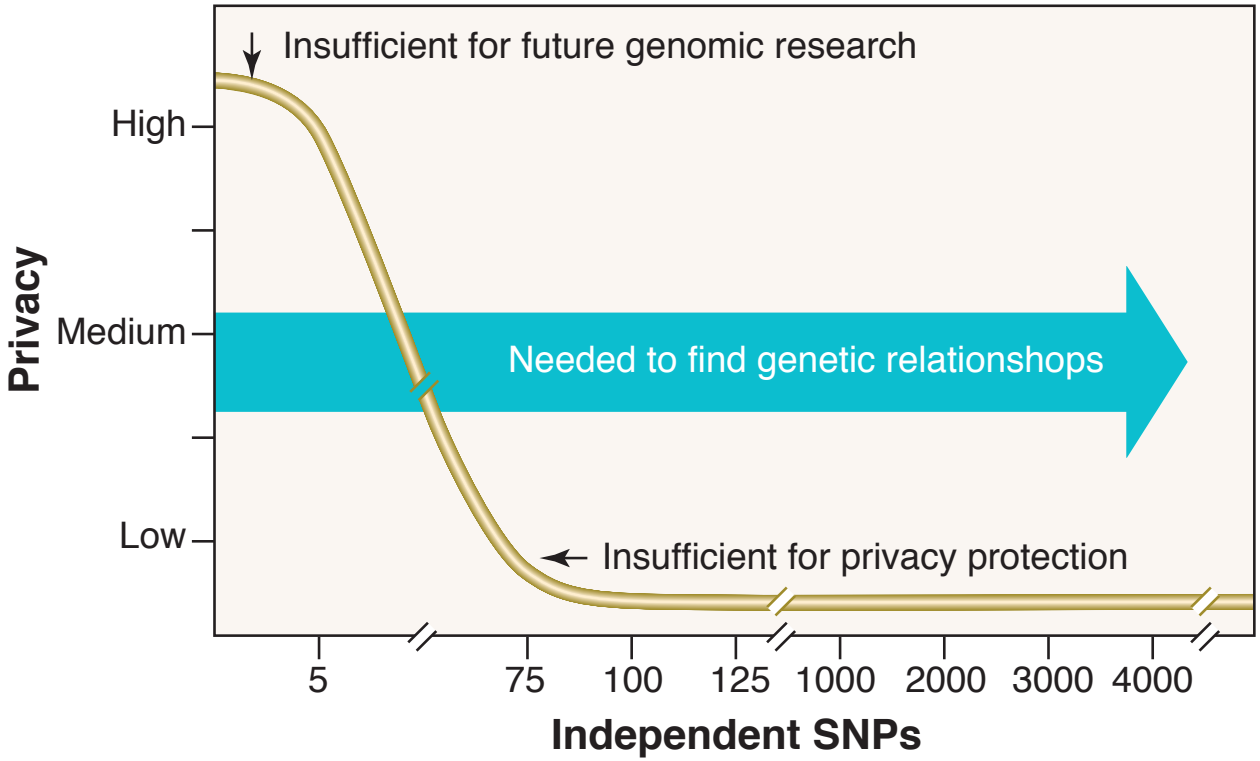
Ancestry DNA

- Genetic inheritance
  - Identity-by-descent
  - Cousin matching





DNA data: privacy and research



Z. Lin, A. Owen, R. Altman, Science, vol 305, 2004

Challenges

- Engineering
  - Scalability
  - Availability
  - Security
- Research
  - Information retrieval
  - DNA genomic research
- Privacy