

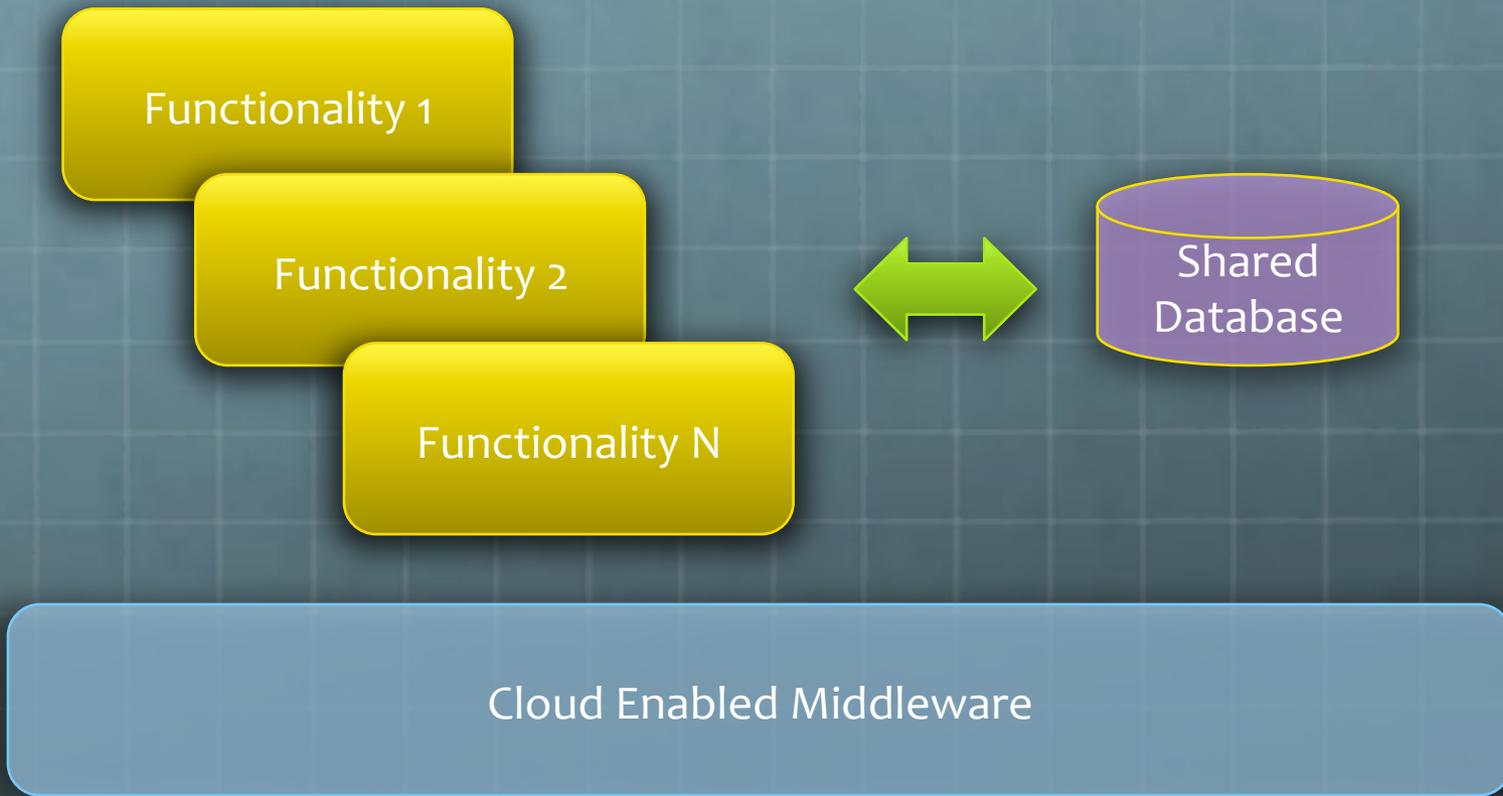
Modernizing HealthCare Applications with Micro Services

Edison Ting
Pivotal Solutions Architect

The screenshot displays the OpenEMR interface for a patient named Sal Goodman. At the top, the patient's name and DOB (1991-07-09) are shown, along with the date Monday, July 21, 2014. A navigation menu on the left includes options like 'Patient/Client', 'Patients', 'New/Search', 'Summary', 'Visits', 'Create Visit', 'Current', 'Visit History', 'Records', 'Visit Forms', 'Import', 'Fees', 'Procedures', and 'Administration'. The main content area shows a calendar for July, a list of providers, and a list of encounters. One encounter is highlighted for 2014-07-21, titled '2014-07-21 Encounter for Sal Goodman'. Below this, there are sections for 'Notes', 'Patient Reminders', 'Disclosures', and 'Vitals'. A 'Billing Manager' window is overlaid on the right, showing criteria for generating bills (X12, CMS 1500 PDF, etc.) and a 'Fee Sheet' section at the bottom.

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**

OpenEMR

Patient Name:

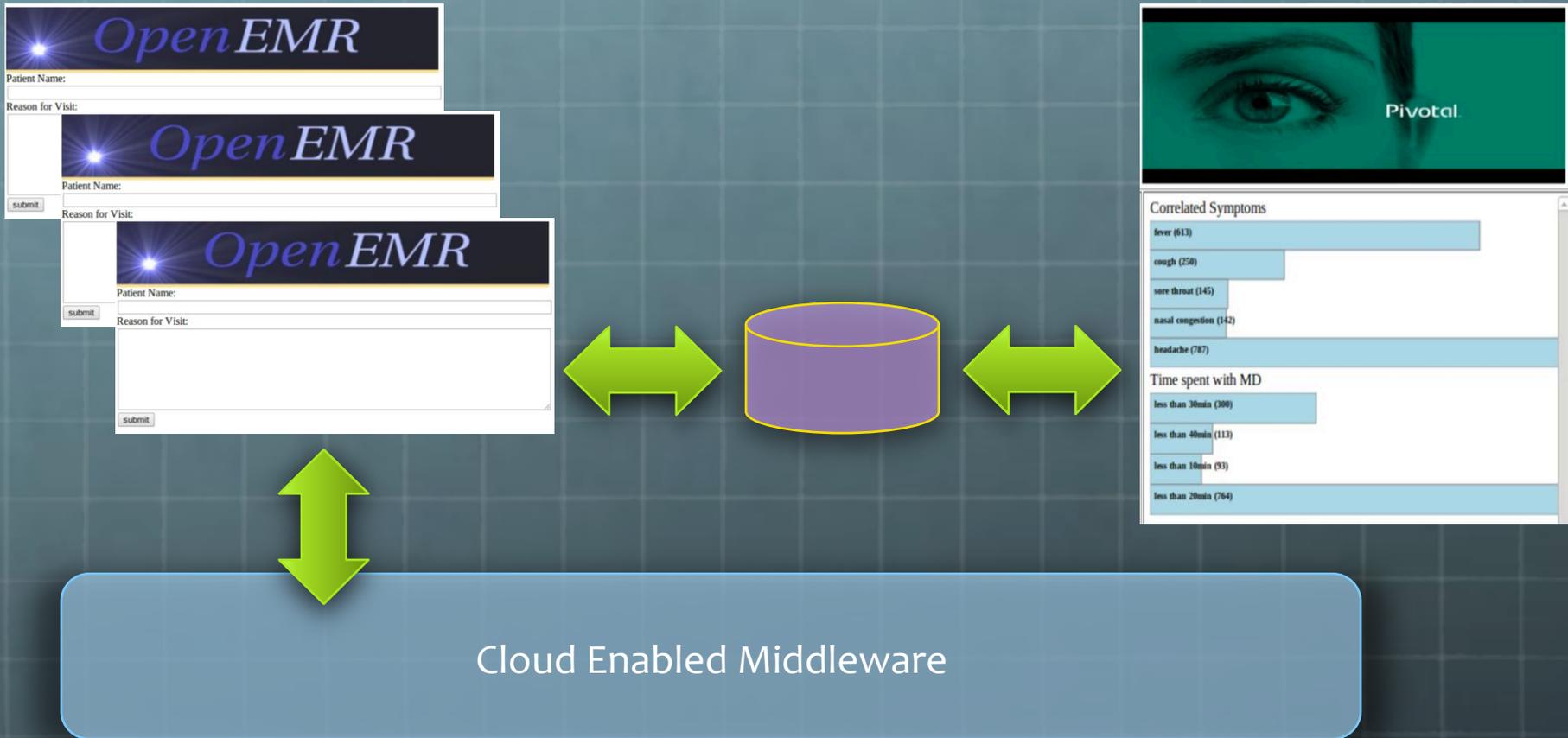
Reason for Visit:

submit



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



- 🌐 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.

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.

EMC², EMC, EMC Proven, the EMC logo, and where information lives are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2012 EMC Corporation. All rights reserved. Published in the USA. 01/12



Ecosystem Challenges Around Data Use

Leonid Zhukov



Ancestry.com

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



Our mission is to help everyone
discover, preserve and share
their family history.



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

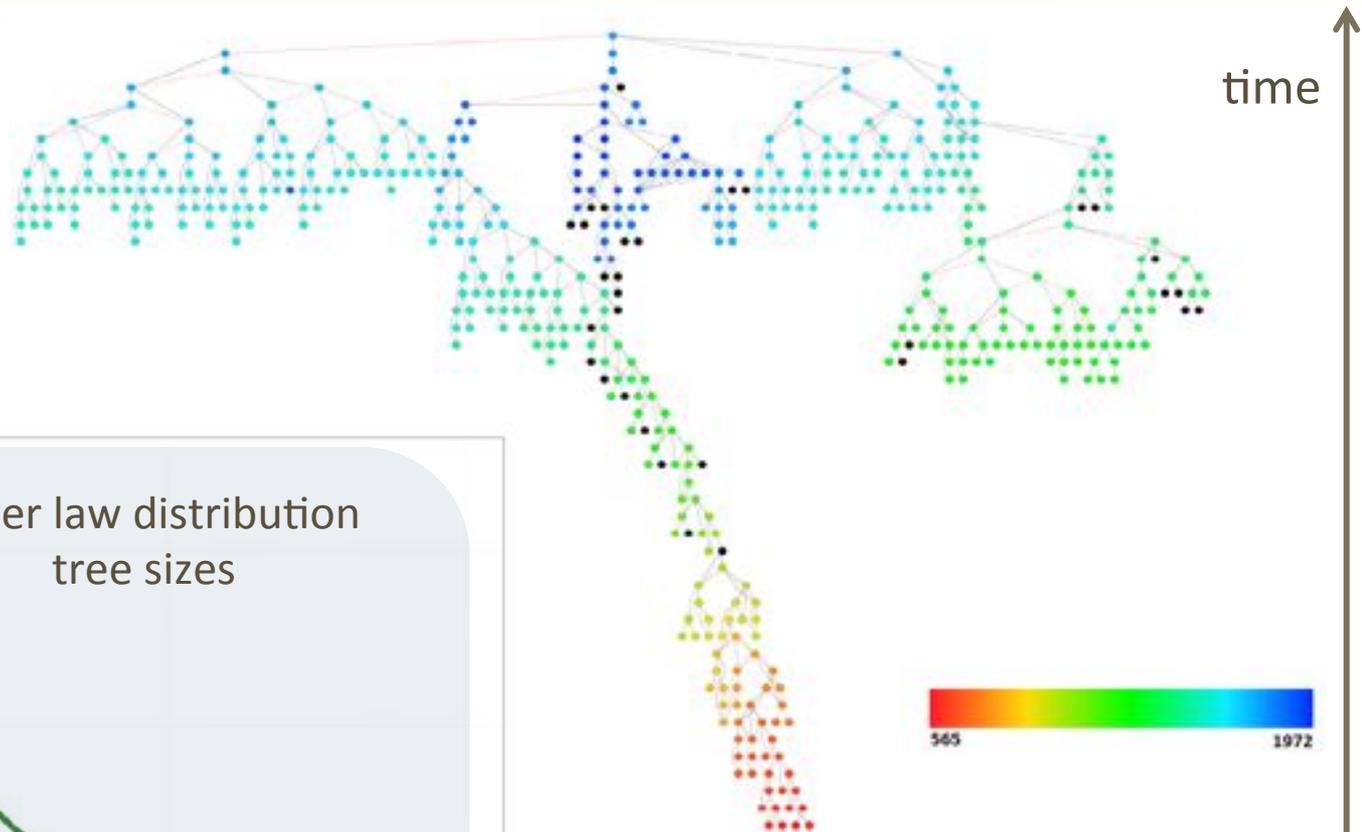
A large, detailed financial ledger or account book. It features a header with the title "ALPHABETICAL LIST of Names in Union No. ..." and a table with columns for names, amounts, and other financial data. The entries are handwritten and include names like "John Smith", "Jane Doe", and "Robert Brown".A handwritten ledger or account book. It has columns for names, amounts, and other financial data. The entries are handwritten and include names like "John Smith", "Jane Doe", and "Robert Brown".

User family trees

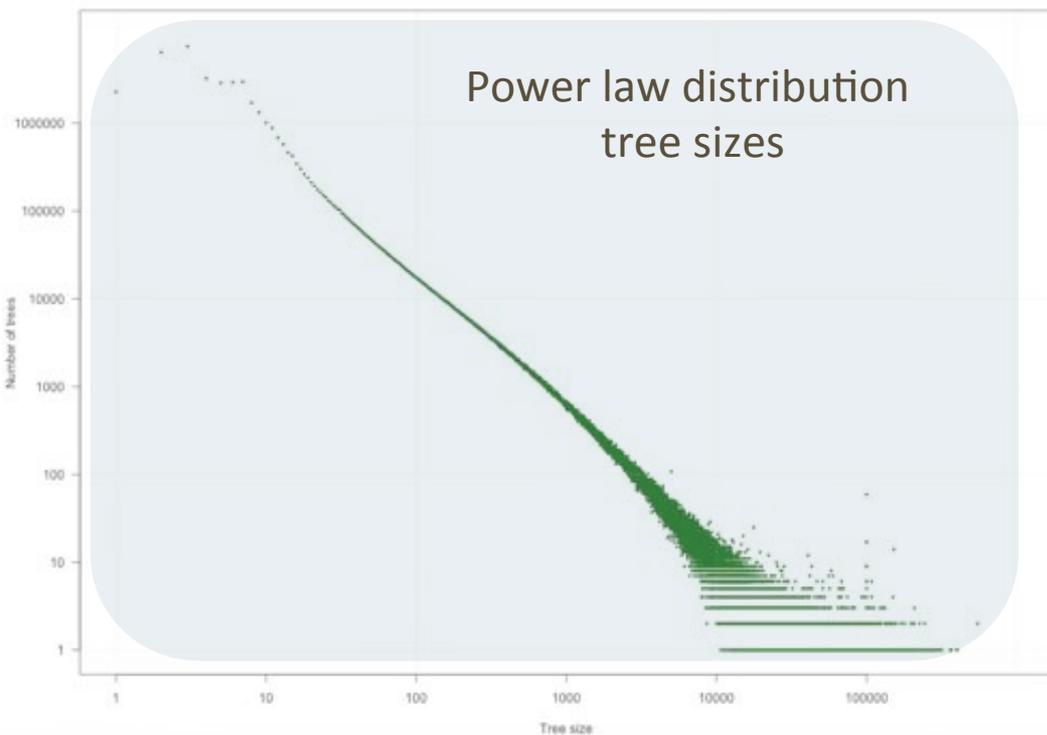
- Family trees:
 - 60 million family trees
 - 6 billion profiles



Family trees



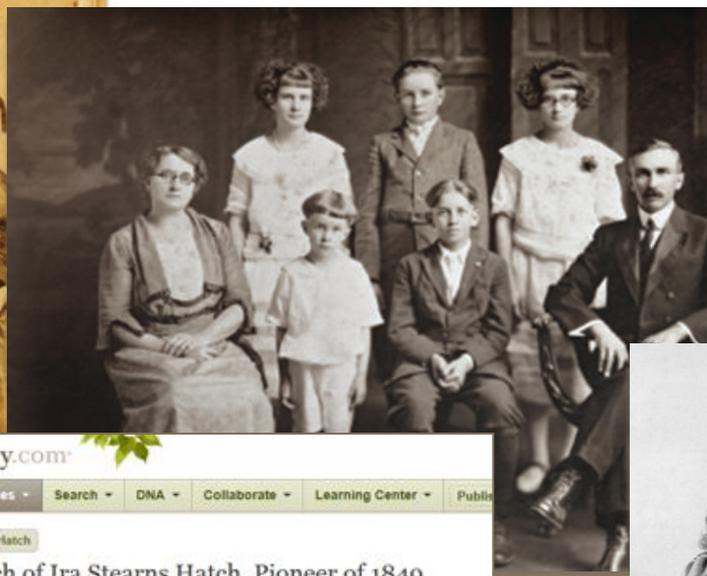
Power law distribution
tree sizes



500 nodes
700 edges
55 generations

User contributed content

- 200 million uploaded family photos and stories



ancestry.com

Home Family Trees Search DNA Collaborate Learning Center Public

Return to Melzar Hatch

Life Sketch of Ira Stearns Hatch, Pioneer of 1849

Life Sketch of Ira Stearns Hatch Pioneer of 1849

When the News of the successful venture of the Pilgrim Fathers' reached the homeland, other honest, sincere people were seized with a desire to also seek a haven of religious freedom in the new land. The Hatches were mostly middle class, neither rich nor poor, mostly small landowners and farmers, pious industrious people. In fact good citizens. One of the descendants of the above mentioned Hatches was Ira Hatch, the son of Jeremiah and Mary Stearns Hatch, who was born at



Person and record search

- Search query

First & Middle Name(s) Robert
Last Name Johnson
Name a place your ancestor might have lived New York, USA
Estimated birth year 1869 [Calculate it](#)

Year: 1949 Location: New York, USA [Remove](#)
[+ Add life events](#) (birth, marriage, death, and more)

Searching for...

Name: "Robert " Johnson
Birth: 1869, New York, USA
Death: 1949, New York, USA

[Edit Search](#)
or Start a new search

Narrow by Category

- All Categories

Census & Voter Lists	5,000+
Birth, Marriage & Death	5,000+
Military	5,000+
Immigration & Travel	2,328
Schools, Directories & Church Histories	5,000+
Tax, Criminal, Land & Wills	3,631
Reference, Dictionaries & Almanacs	1,905

Matches 1-20 of 247,347 Sorted By Relevance [View](#) [Sorted by relevance](#)

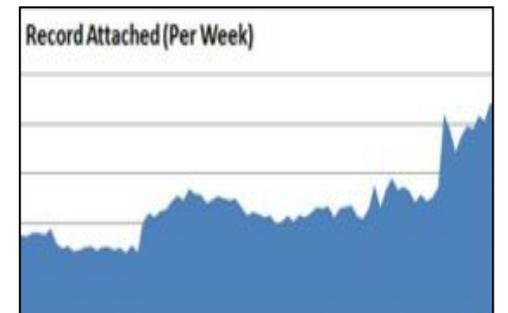
Menands, New York, Albany Rural Cemetery Burial Cards, 1791-2011 Birth, Marriage & Death ★★★★★ View Image	NAME: Sylvester Robert Johnson SPOUSE: Rosa Anna Ebel Johnson MOTHER: Gertrude Lansing Johnson FATHER: Jay Johnson MORE: See all information...
1940 United States Federal Census Census & Voter Lists ★★★★★ View Image	NAME: Robert J Johnson SPOUSE: Anna Johnson BIRTH: abt 1868 - New York RESIDENCE: 1935 - North Salem, Westchester, New York RESIDENCE: North Salem, Westchester, New York
1940 United States Federal Census Census & Voter Lists ★★★★★ View Image	NAME: Robert Johnson SPOUSE: Lena Johnson BIRTH: abt 1870 - New York RESIDENCE: 1935 - Hague, Warren, New York RESIDENCE: Hague, Warren, New York
1940 United States Federal Census Census & Voter Lists ★★★★★ View Image	NAME: Robert Johnson Junior BIRTH: abt 1870 - New York RESIDENCE: 1935 - White Plains, Westchester, New York RESIDENCE: White Plains, Westchester, New York
1930 United States Federal Census Census & Voter Lists ★★★★★ View Image	NAME: Robert M Johnson SPOUSE: Louise W Johnson BIRTH: abt 1863 - New York RESIDENCE: 1930 - Oakland, Alameda, California

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

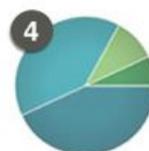
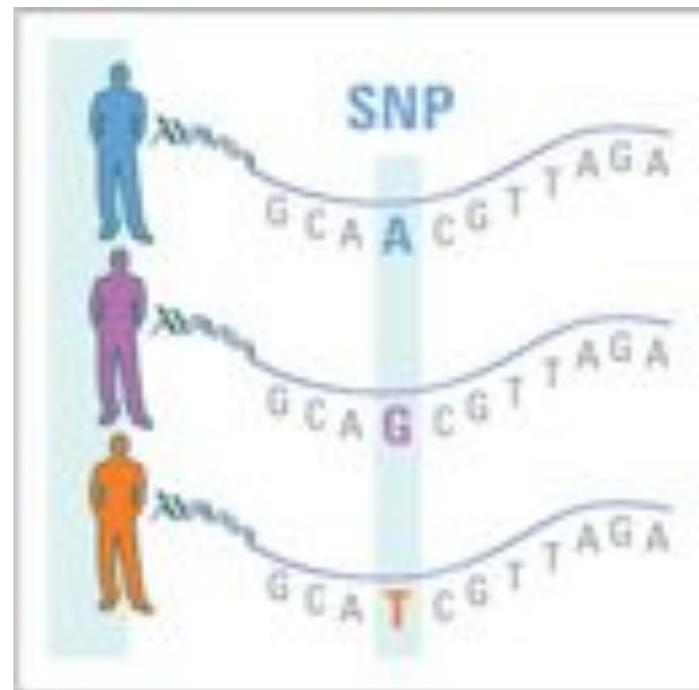
- User behavior data:
 - 75 mln searches daily
 - 10 mln profiles added daily
 - 3.5 mln records attached daily



A screenshot of the Ancestry.com website interface. The main area features a world map with a red location marker. To the left, there is a list of 'People Added to Trees' with names and dates. To the right, there is a 'Latest Searches' section with a search result for 'John Warston born 1850 died 1875'. Below the map, there are several smaller panels: a document image, a family photograph, and two pie charts. The top pie chart is titled 'Top 10 Subscriber Locations' and the bottom one is '10 Most Popular Collections'. The interface is dark-themed with white text.

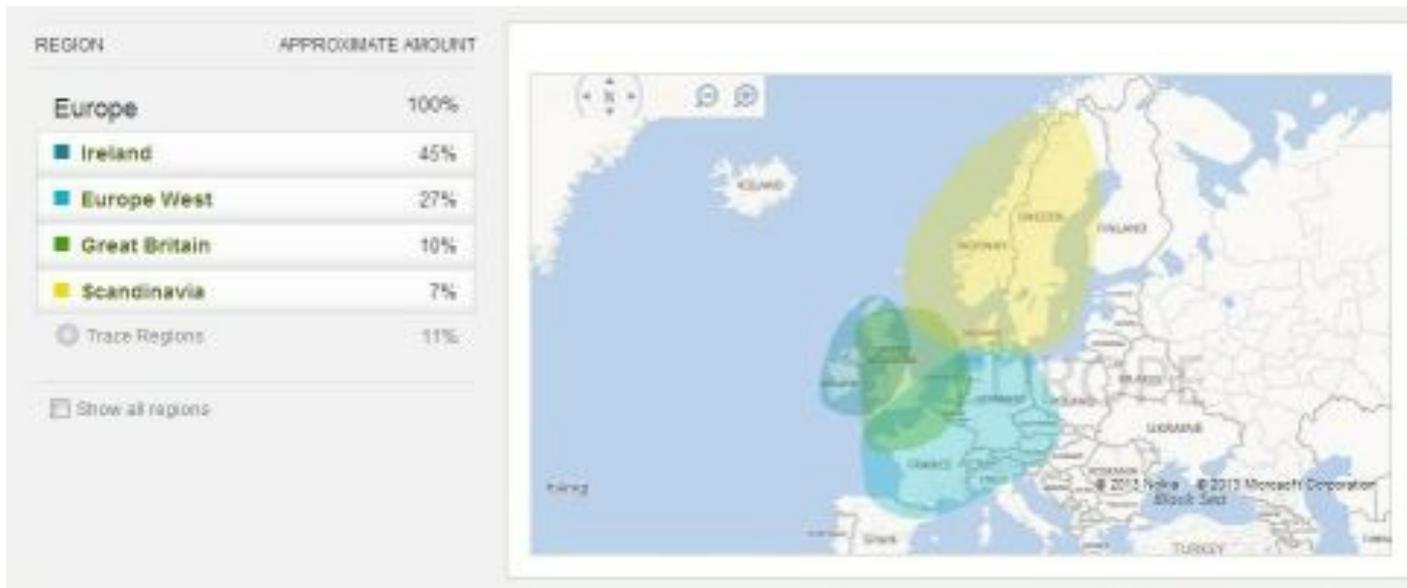
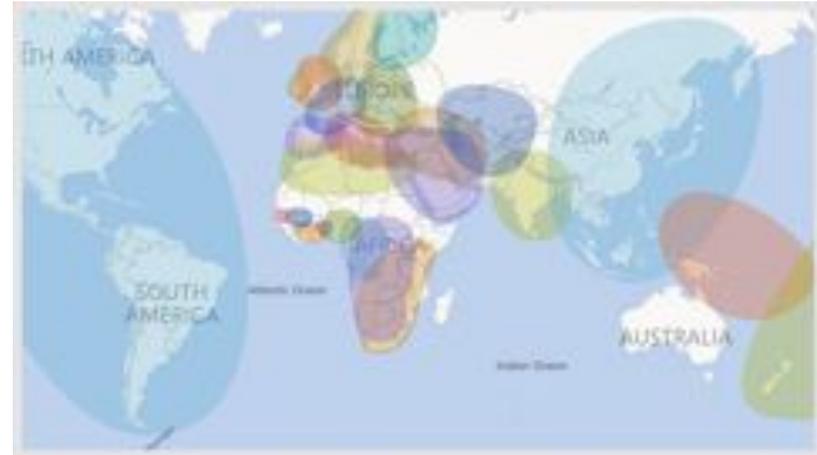
DNA Data

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



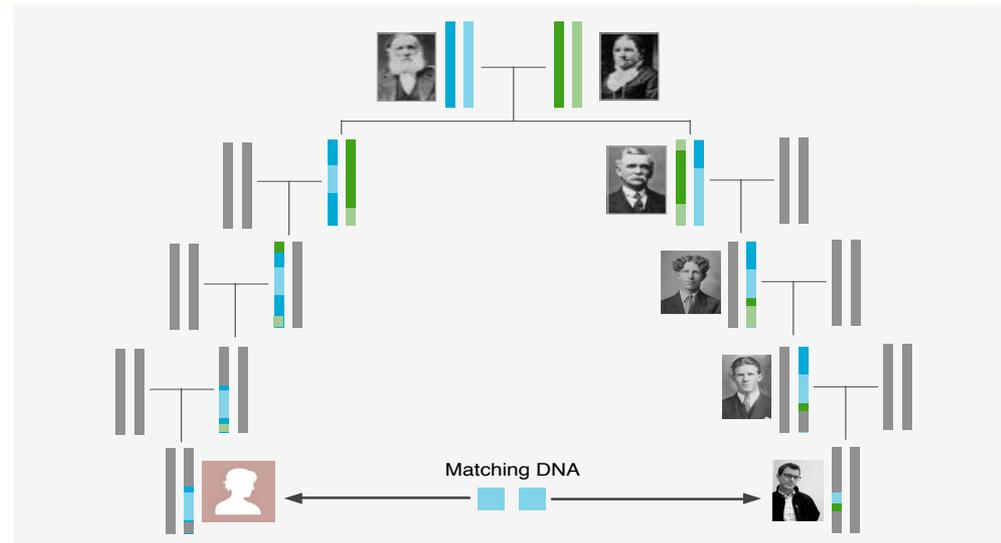
Ancestry DNA

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



Ancestry DNA

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



4TH COUSIN

 **jomacb** 3889 people [Review Match](#)
Possible range: 4th - 5th cousins ⓘ
Last logged in Apr 12, 2012
96% confidence

DISTANT COUSIN

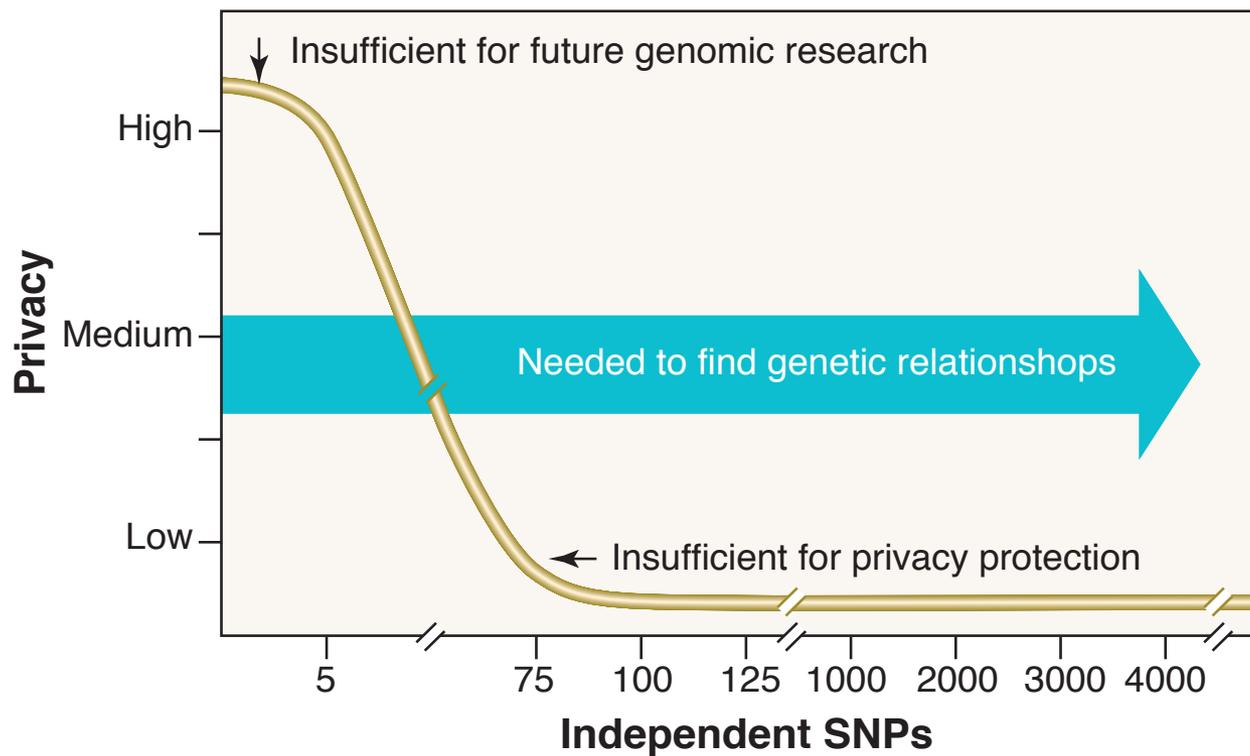
Note: Distant cousin matches (5th cousins or greater) are less sure than 3rd and 4th cousins. These matches are still good leads, but there is a low degree of certainty (50% or less) that you are related. ✕

 **Florence Minar** 3446 people [Review Match](#)
Possible range: 5th - 6th cousins ⓘ
Last logged in Apr 18, 2012
50% confidence

 **admsacollins** 1462 people [Review Match](#)
Possible range: 5th - 6th cousins ⓘ
Last logged in Apr 15, 2012
50% confidence

 **ajam3541956** 228 people [Review Match](#)
Possible range: 5th - 6th cousins ⓘ
Last logged in Apr 16, 2012
50% confidence

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