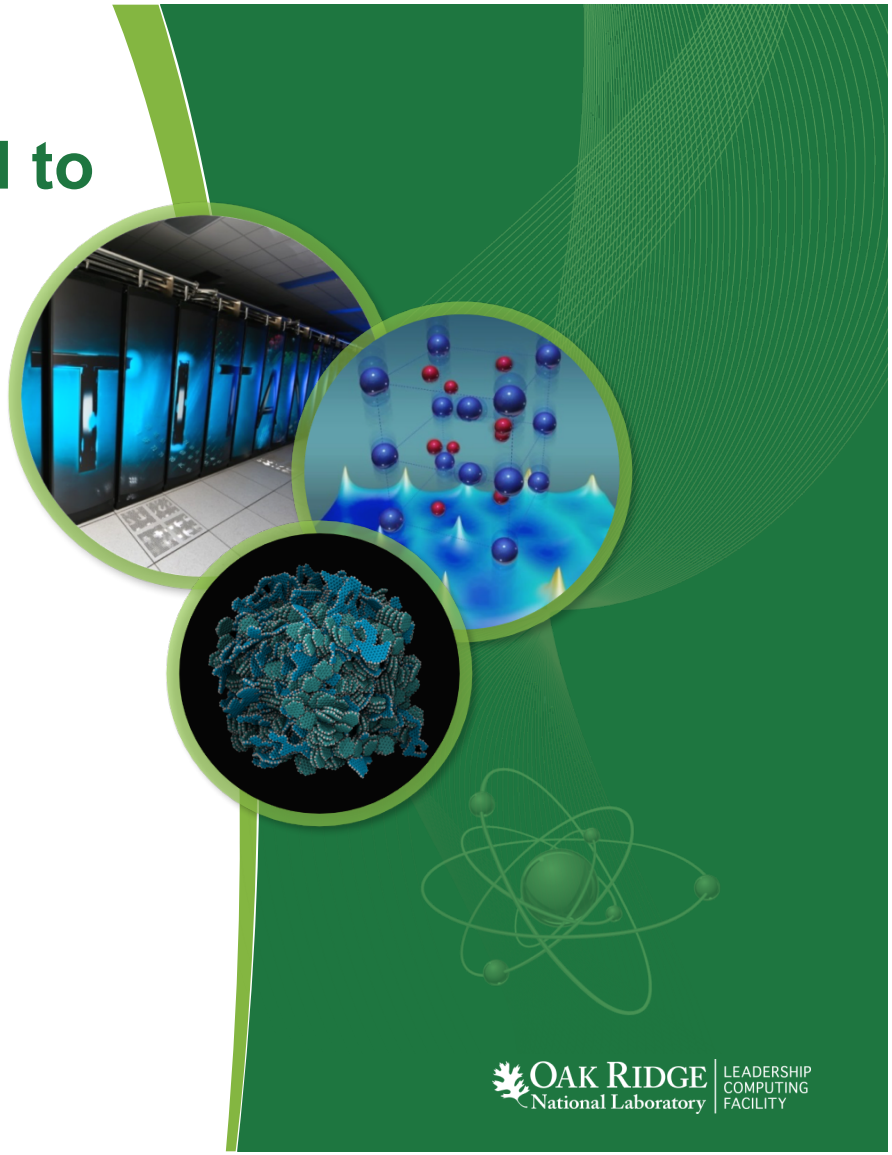
Data Science Stories Using Open Health Data : From Fighting Fraud to Solving Medical Mysteries



Rangan Sukumar, PhD
Senior Analytics Architect, Office of the CTO, Cray Inc.

Previously,
Chief Data Scientist[1] and Group Leader[2]

[1]Health Data Sciences Institute
Computational Sciences and Engineering Division, ORNL

[2]Advanced Data and Workflows
Oak Ridge Leadership Computing Facility, ORNL

Email: ranganutk@gmail.com

ORNL is managed by UT-Battelle
for the US Department of Energy

Health Data Sciences Institute @ Oak Ridge National Lab

Health Data + **ORNL Compute** = **Smarter Healthcare**

- **Claims**
- **Images**
- **Omics**
- **EHRs**
- **Sensors**

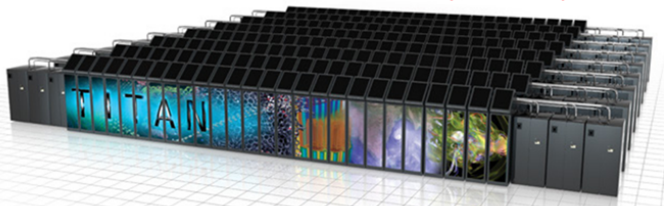
- **Hardware**
  - Infrastructure
- **Software**
  - Architectures
  - Algorithms
  - Simulations
- **Data Scientists**

- **Better Policy**
- **Better Quality**
- **Better Integrity**
- **Better Science**

Urika - Extreme Analytics

Urika – Graph Discovery

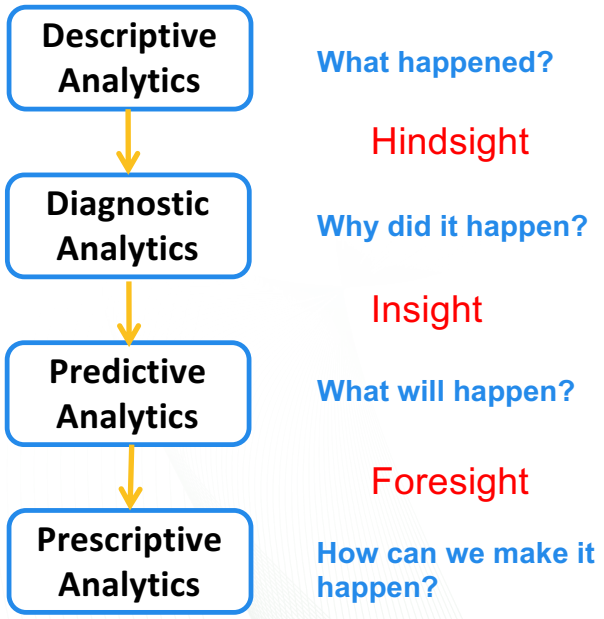
TITAN (CPU + GPU)



Began as an inter-agency agreement between Department of Energy and the Human Health Services/ Centers for Medicare and Medicaid Services in 2011.

Since 2011...

Open Data	Source
National Plan and Provider Enumeration System (NPPES)	Centers for Medicare and Medicaid Services (CMS)
Doctor Referral Graph (DocGraph)	DocGraph (Fred Trotter FOIA)
List of Excluded Individuals and Entities (LEIE)	Office of Inspector General – Health and Human Services (OIG)
Semantic MEDLINE	National Library of Medicine (NLM)
Clinicaltrials.gov	National Institute of Health (NIH)
Surveillance, Epidemiology and End Results (SEER)	National Cancer Institute (NCI)
Provider Utilization and Synthetic Public Use File	Centers for Medicare and Medicaid Services (CMS)
OpenFDA	Federal Drug Administration (FDA)
The Cancer Genome Atlas	National Cancer Institute (NCI)
Exposure to Environmental Chemicals	Centers for Disease Control (CDC)



Concept adapted from Gartner's Webinar on Big Data

Story #1: Fighting Fraud, Waste and Abuse

Partnership Fund for Program Integrity Innovation

Fifth Semiannual Report to Congress

January 2013

the Provider Enrollment Process for Risk Assessment and Comparative Analysis
Health and Human Services, Centers for Medicare & Medicaid Services (CMS)
$2.9 million (*$2.7 million obligated*[9])
Texas, California, Arizona, Arkansas, Washington, Tennessee, Oregon, Minnesota, New Mexico, North Carolina, and New Jersey have submitted provider files for the analysis.

The pilot tests an automated tool to screen Medicaid providers for potential fraud by cross-checking their credentials, background, and history among states and with Federal Medicare data. Currently, CMS and states lack standardized Medicaid provider data, which hampers effective analysis to assess providers for risk of fraud. The National Health Care Anti-Fraud Association estimates that three percent of all health care spending is lost to health care fraud.[10] In Medicaid, where fraud is difficult both to measure and to prevent, that would equate to approximately $12 billion in Federal and state funds in FY 2011.

***Implementation*:** The pilot will leverage and advance existing CMS fraud detection efforts through the following steps:

1. Capture Medicaid Statistical Information System (MSIS) data from pilot states to help form a complete, cross-program data set for a specific provider type.
2. Validate Medicaid provider data using CMS Center for Program Integrity Analytics Lab Tools (now used in Medicare).
3. Use this validated data to test an automated cross-check that identifies providers enrolled in both Medicare and Medicaid and to determine an effective provider risk-assessment model.
4. Apply the model to other pilot states to identify high-risk providers for follow-up.

Based on its analysis of the provider files submitted by the participating states, Oak Ridge National Labs (ORNL) has identified probable risk factors and high-risk vendors. The findings from the pilot have been validated and risk-scored by CMS in partnership with Rainmakers, a contractor.

How a computing powerhouse delivers health care

Health datasets come in many orders of magnitude, but few are as large as the public health big data being gathered and analyzed by computers at the Energy Department's Oak Ridge National Lab.

About four years ago, the ORNL decided to amass as much public health care data as it could and subject it to the analytics engines of its most powerful computers.

"We were in a unique position with our leadership computing resources and data science expertise, and we saw an opportunity to use health data to discover data-driven insights for better health care quality, integrity and policy," said Sreenivas Sukumar, a researcher in ORNL's computational sciences division.

To analyze the datasets, researchers used the Lab's multicore Titan, the second-most powerful computer in the world, Apollo, an in-memory Urika graph-computer built by Yarcdata, and distributed cloud computing-based machines.

The lab also tapped some of the biggest producers of health related data, including the Cancer Genome Atlas, clinicaltrials.gov, Semantic MEDLINE, openFDA, DocGraph and the National Plan and Provider Enumeration System.

In working with the data, the researchers initially encountered computing silos created by existing information architectures that did not scale to the analytics requirements of the large datasets. Consequently, the lab turned to an approach using graph computing, a scalable computing solution capable of uncovering relationships hidden in the data.

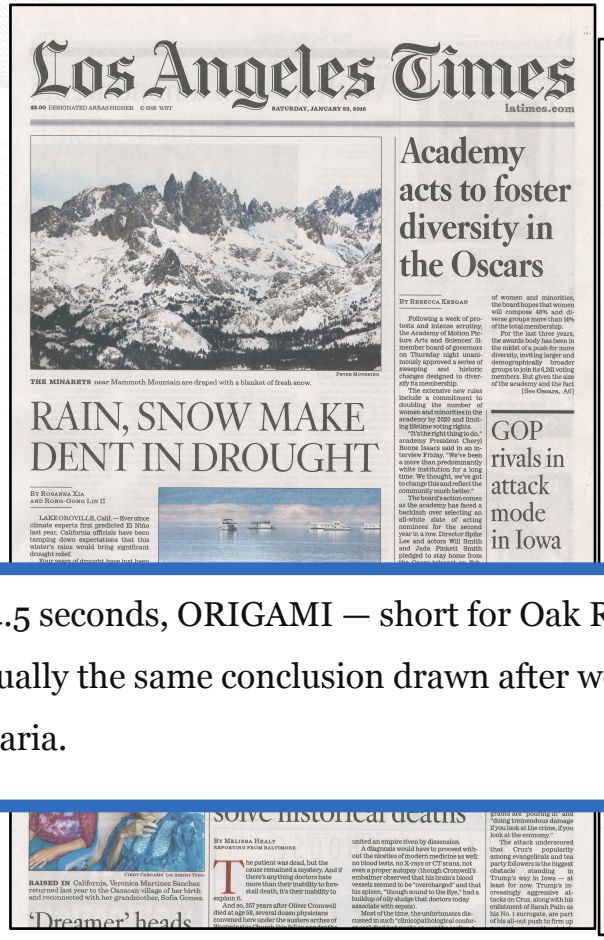
The graph computing almost immediately provided insights into some of the datasets, including feedback on understanding fraud, waste and abuse within the federal health care system, according to ORNL researchers.

In one case, the lab was able to identify a health care provider using multiple identities to bill patients. Another case showed guilt-by-association patterns that highlighted the potential for fraud before the provider began billing.

Georgia Tourassi, director of

Story #2: Solving Medical Mysteries

LA Times, January 23, 2016



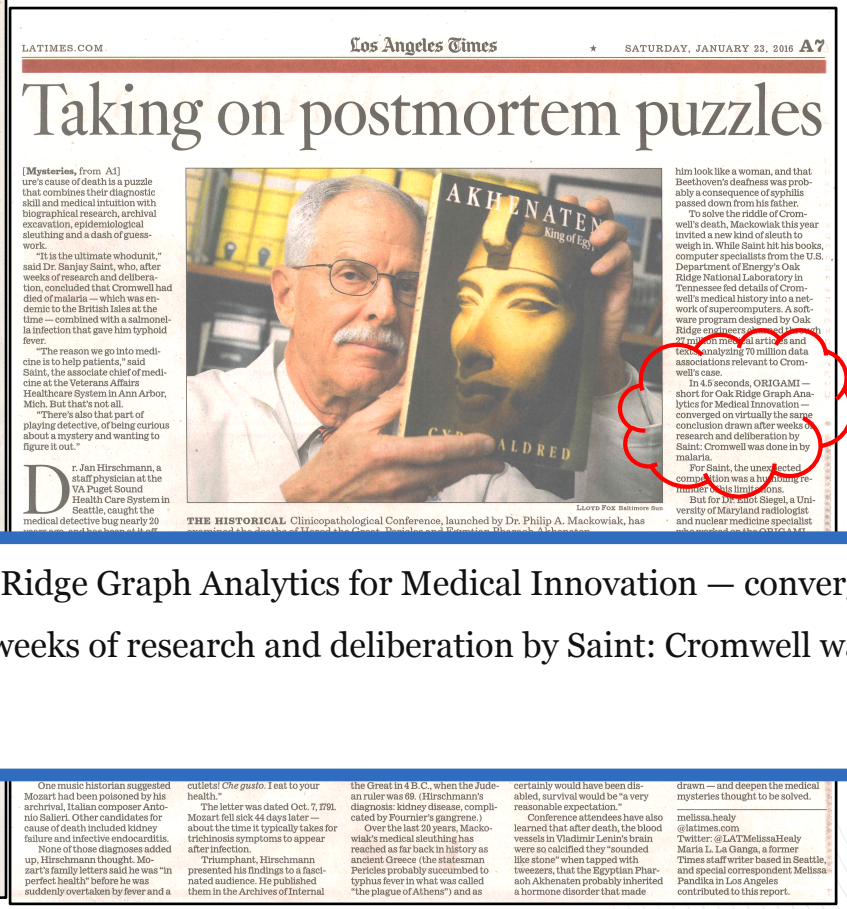
Los Angeles Times

latimes.com

Academy acts to foster diversity in the Oscars

RAIN, SNOW MAKE DENT IN DROUGHT

GOP rivals in attack mode in Iowa



LATIMES.COM Los Angeles Times ★ SATURDAY, JANUARY 23, 2016 A7

Taking on postmortem puzzles

LLOYD FOX Baltimore Sun

THE HISTORICAL Clinicopathological Conference, launched by Dr. Philip A. Mackowiak, has

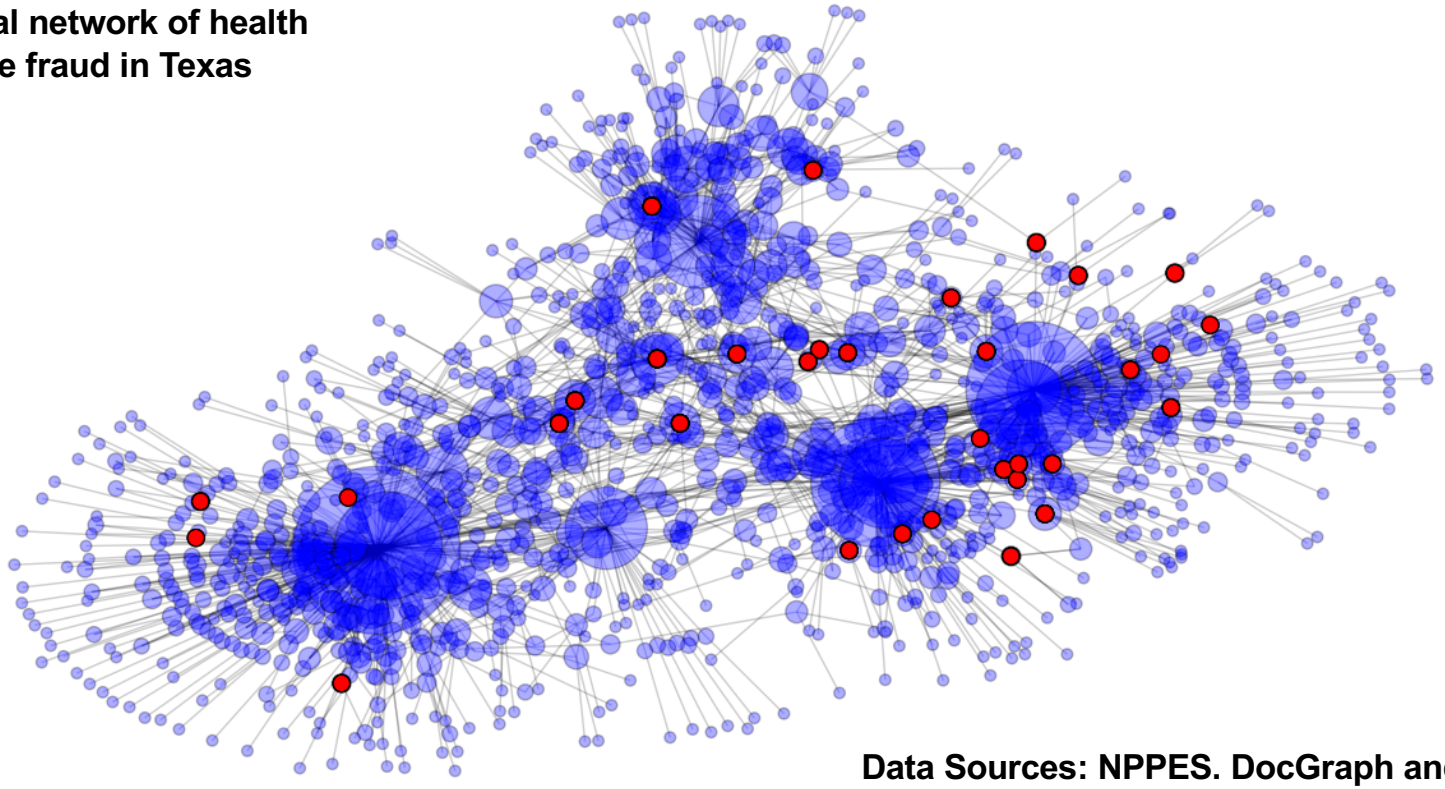
2016 R&D 100 WINNER

In 4.5 seconds, ORIGAMI — short for Oak Ridge Graph Analytics for Medical Innovation — converged on virtually the same conclusion drawn after weeks of research and deliberation by Saint: Cromwell was done in by malaria.

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Story #1: Fraud Detection and Prevention

A social network of health care fraud in Texas



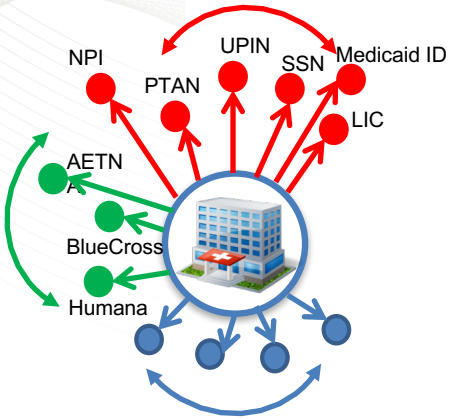
Data Sources: NPPES. DocGraph and LEIE

Chandola, Varun, Sreenivas R. Sukumar, and Jack C. Schryver. "Knowledge discovery from massive healthcare claims data." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

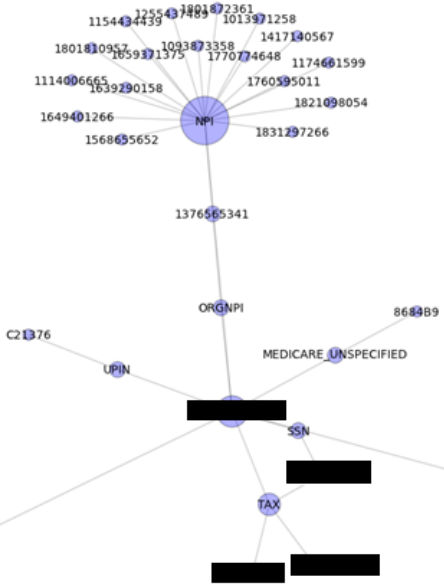
Story #1: Fraud Detection and Prevention

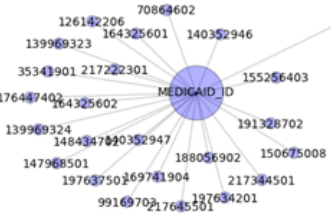
Multiple identities

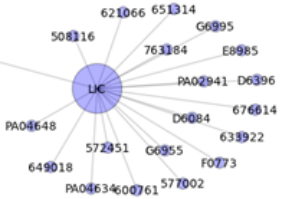
Government issued IDs



Attributes : Name, Address, Telephone, etc.

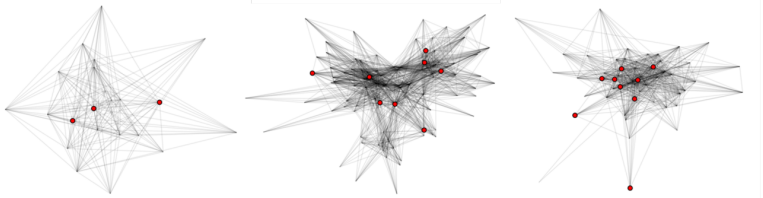




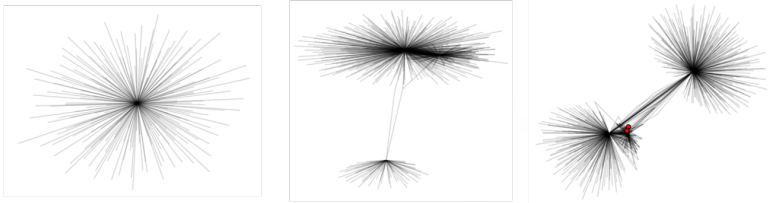


19 distinct people using the same address and phone in TX.

High-risk patterns (Y-networks and triangles, polygons)



Low-risk patterns (* networks)

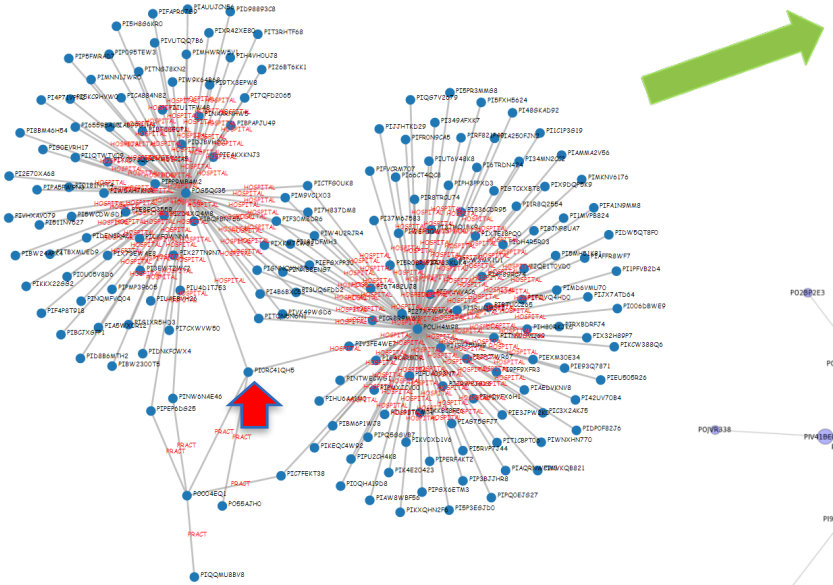


Different names associated with the ID in TX-provider file

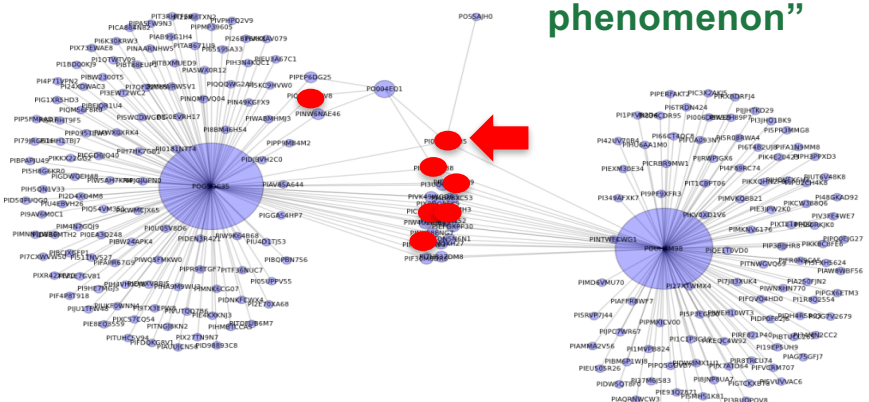
HEALTHARE MEDICAL ASSOCIATES
HEALTHCARE MEDICAL
HEALTHCARE MEDICAL ASSOCIATES
HEATLHCARE MEDICAL ASSOCIATES
MEDISTAT GROUP ASSOCIATES

Story #1: Fraud Detection and Prevention

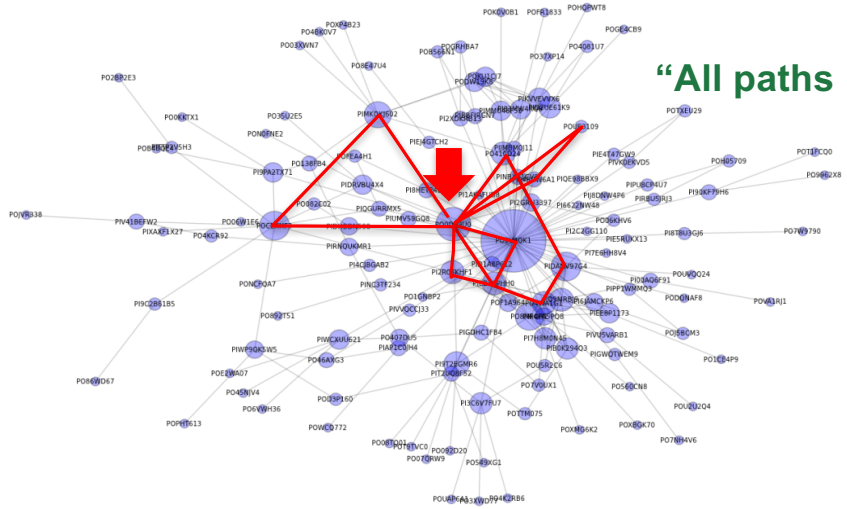
“Affiliations to multiple hospitals while also owning private and group practice are strong indicators of potential suspicious activity.”



“The country club phenomenon”

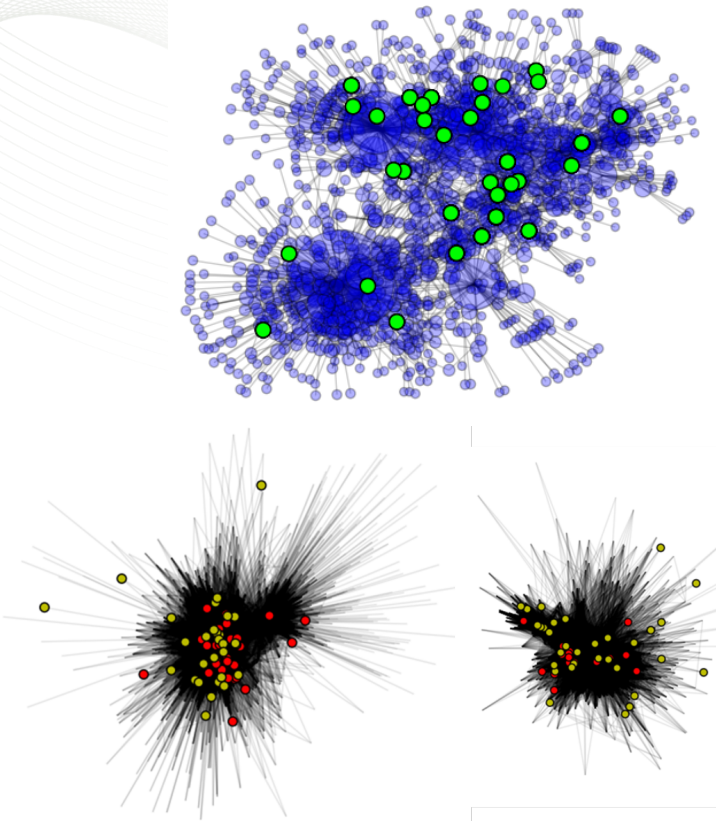


“All paths lead to Rome”

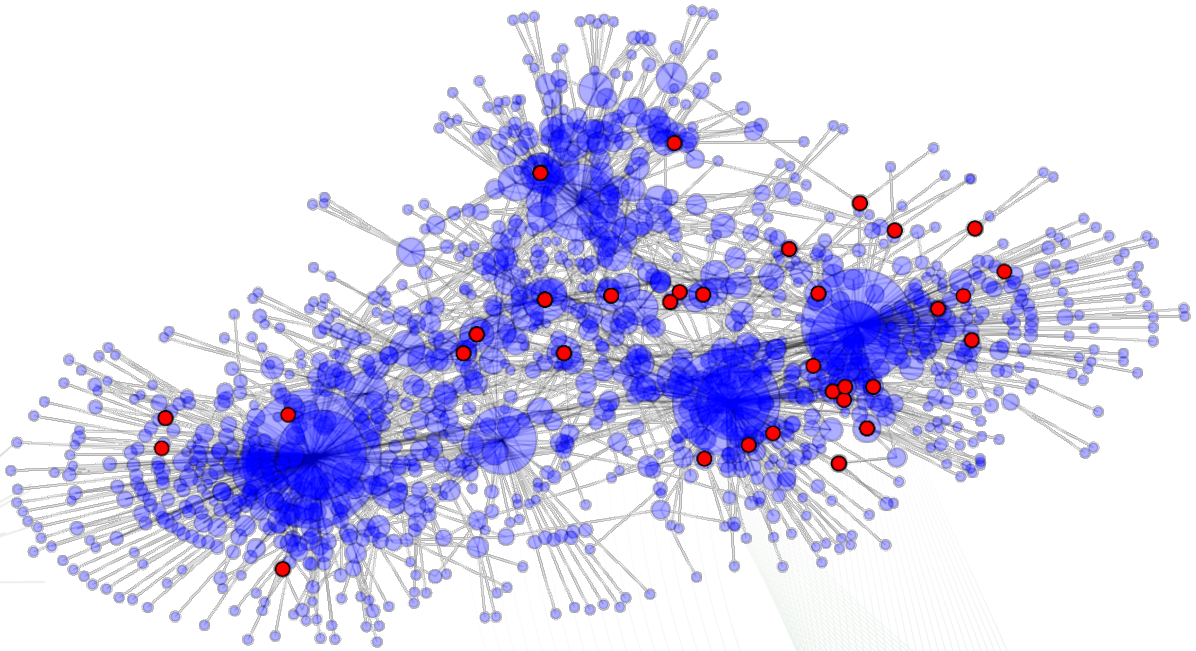


Story #1: Fraud Detection and Prevention

Extrapolating to unseen data



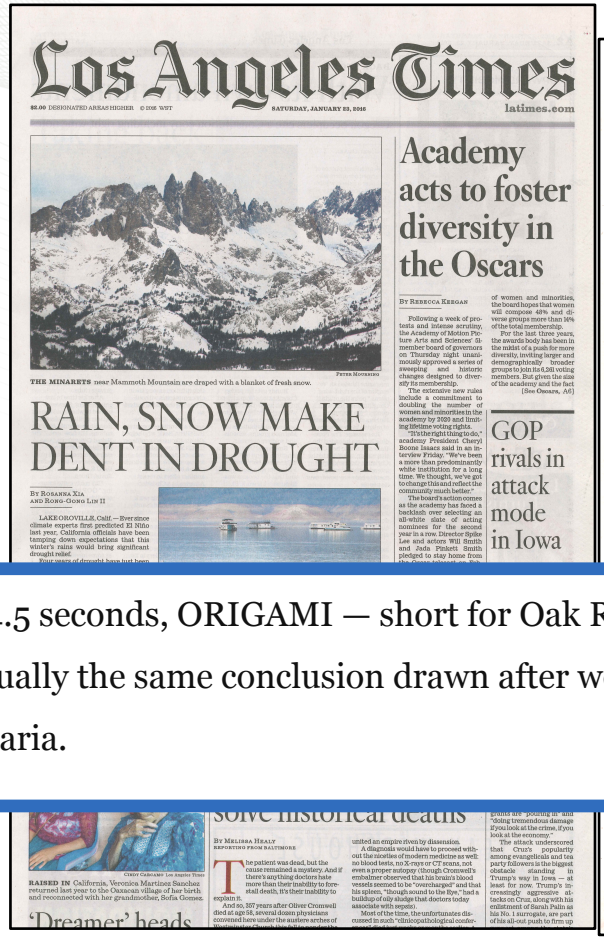
Now that we now there is a “network-science” behavior to healthcare fraud, any ideas on how we “act” on it ?



Hint: What is Facebook and LinkedIn doing to you?

Story #2: Solving Medical Mysteries

LA Times, January 23, 2016



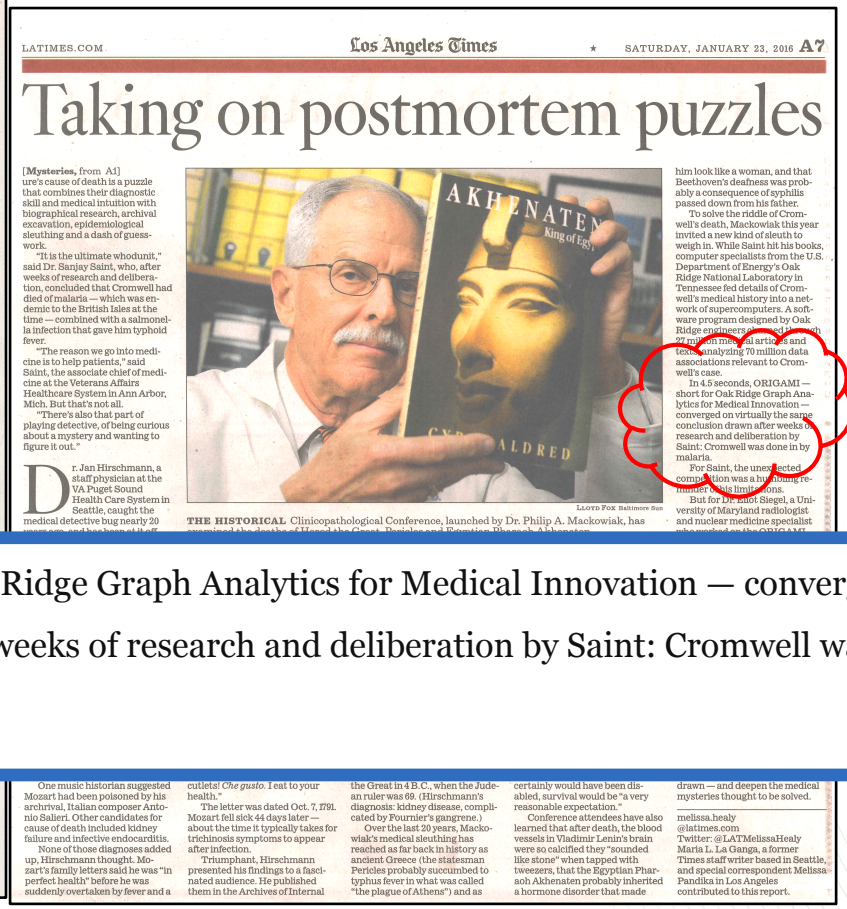
Los Angeles Times

latimes.com

Academy acts to foster diversity in the Oscars

RAIN, SNOW MAKE DENT IN DROUGHT

GOP rivals in attack mode in Iowa



LATIMES.COM Los Angeles Times ★ SATURDAY, JANUARY 23, 2016 A7

Taking on postmortem puzzles

In 4.5 seconds, ORIGAMI — short for Oak Ridge Graph Analytics for Medical Innovation — converged on virtually the same conclusion drawn after weeks of research and deliberation by Saint: Cromwell was done in by malaria.

2016 R&D 100 WINNER

Story #2: ORIGAMI is an eco-system

Open Data

National Library of Medicine's *Semantic Medline*

Subject	Predicate	Object
Influenza	ISA	RNA_Virus_Infections
Influenza	ISA	Viral_upper_respiratory_tract_infection
Influenza	INTERACTS_WITH	Influenza_A_Virus__H1N1_Subtype
Influenza	ISA	Acute_viral_disease
Influenza	ISA	Influenza_with_pneumonia_NOS
Influenza	AFFECTS	Maori_Population
Influenza	COEXISTS_WITH	Influenza_A_Virus__H3N2_Subtype
Influenza	COEXISTS_WITH	Mental_alertness
Influenza	INTERACTS_WITH	Dengue_Virus
Influenza	AFFECTS	Influenza_with_encephalopathy
Influenza	AFFECTS	Swine_influenza
.	.	.
.	.	.
.	.	.
Influenza	CAUSES	UPPER_RESPIRATORY_SYMPTOM
Influenza	CAUSES	Wheezing_symptom

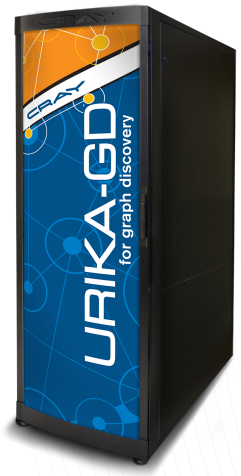
70 million predications from 23.5 million PubMed articles

Compute

ORNL's Compute and Data Environment for Science

504 compute cores, 5.4 TBs of distributed memory, and 576 TBs of local storage

64 Threadstorm processors, 2 TBs of shared memory connected to 125 TB of storage



Open Algorithms

http://github.com/ssrangan

- Data-driven reasoning
  - Semantic
  - Graph-theoretic
  - Statistical
- Model-driven reasoning
  - Term-based
  - Path-based
  - Meta-pattern
  - Context-based
  - Analogy

Open API: http://hypothesis.ornl.gov

Story #2: ORIGAMI is intelligent

Here is an example of how ORIGAMI works….

Nexium						Heartburn

Filling in the blanks….and then ranking it for significance…

Nexium	'Is a'	Esomeprazole	*'Reverse(Is a)'*	Proton Pump Inhibitors	*'Disrupts'*	Heartburn

Eksp Klin Gastroenterol. 2009;(4):86-92.

[Omeprazol and ezomeprazol pharmacokinetics, duration of antisecretory effect, and reasons for their probable changes in duodenal ulcer].

[Article in Russian]

Serebrova SIu, Starodubtsev AK, Pisarev VV, Kondratenko SN, Vasilenko GF, Dobrovol'skiĭ OV.

Abstract

There were authentic distinctions between the groups of healthy volunteers and patients with a peptic ulcer disease in Cmax, Tmax, AUC(0-t), AUC (0-infinity), Clt, Vd of omeprazole and Cmax of esomeprazole (Nexium, AstraZeneca). When the pharmacokinetics of omeprazole and ezomeprazole were compared in both groups, there were authentic distinctions in Cmax, AU(0-t), AUC(0-infinity), Clt, T1/2. The patients who had taken omeprazole the time of hypoacide condition was much shorter than in other groups. Disintegration test modeling pHmax for pH oscillation with large amplitude, that is typical for ulcer disease, demonstrated a possibility of early partial release of omeprazole, its acid-depended degradation and reduction of its bioavailability.

Aliment Pharmacol Ther. 2006 Sep 1;24(5):743-50.

Systematic review: proton pump inhibitors (PPIs) for the healing of reflux oesophagitis - a comparison of esomeprazole with other PPIs.

Edwards SJ[1], Lind T, Lundell L.

Author information

Abstract

BACKGROUND: No randomized controlled trial has compared all the licensed standard dose proton pump inhibitors in the healing of reflux oesophagitis.

AIM: To compare the effectiveness of esomeprazole with licensed standard dose proton pump inhibitors for healing of reflux oesophagitis (i.e. lansoprazole 30 mg, omeprazole 20 mg, pantoprazole 40 mg and rabeprazole 20 mg).

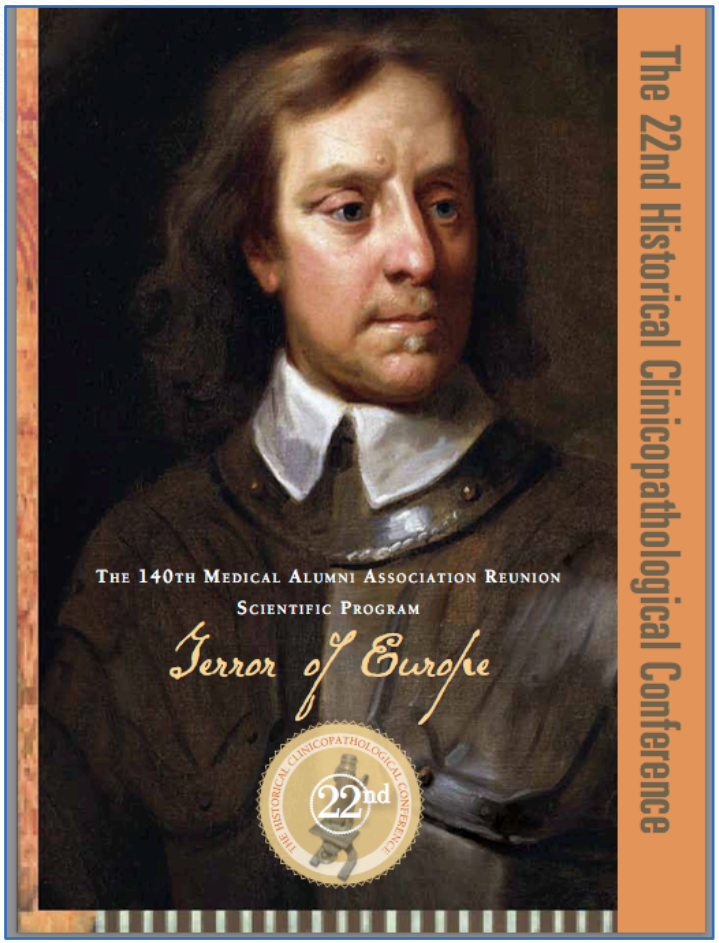
METHODS: Systematic review of CENTRAL, BIOSIS, EMBASE and MEDLINE for randomized controlled trials in patients with reflux oesophagitis. Searching was completed in February 2005. Data on endoscopic healing rates at 4 and 8 weeks were extracted and re-analysed if not analysed by intention-to-treat. Meta-analysis was conducted using a fixed effects model.

RESULTS: Of 133 papers identified in the literature search, six were of sufficient quality to be included in the analysis. No studies were identified comparing rabeprazole with esomeprazole. A meta-analysis of healing rates of esomeprazole 40 mg compared with standard dose proton pump inhibitors gave the following results: at 4 weeks [relative risk (RR) 0.92; 95% CI: 0.90, 0.94; P < 0.00001], and 8 weeks (RR 0.95; 95% CI: 0.94, 0.97; P < 0.00001). Publication bias did not have a significant impact on the results. The results were robust to changes in the inclusion/exclusion criteria and using a random effects model.

CONCLUSION: Esomeprazole consistently demonstrates higher healing rates when compared with standard dose proton pump inhibitors.

Story #2: ORIGAMI at the Historical Clinicopathological Conference

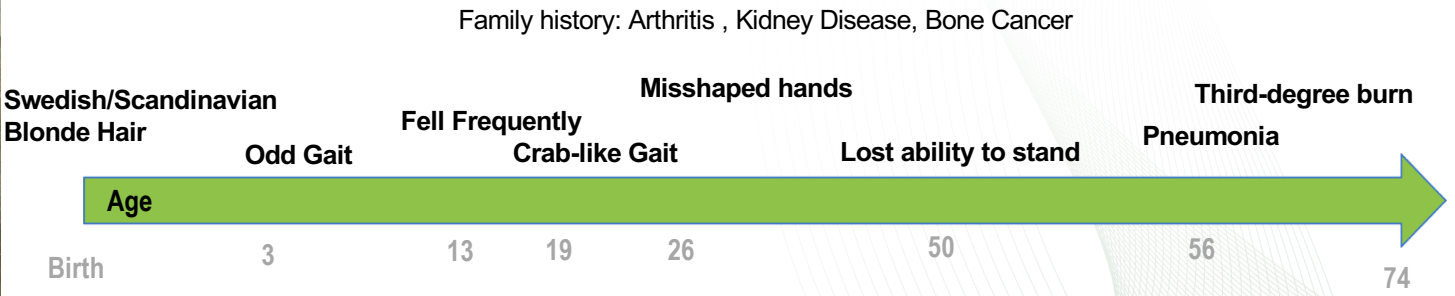
2015



2016

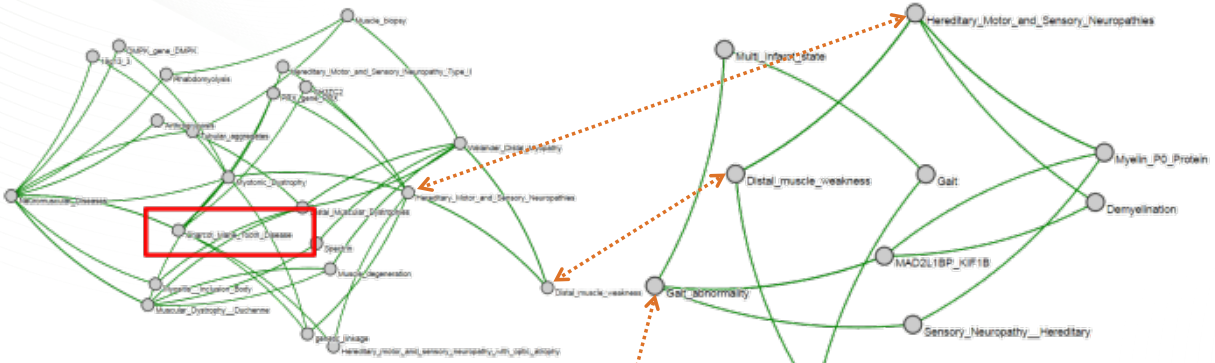
Story #2: ORIGAMI at Historical Clinicopathological Conference

This patient, like the artist who made her famous, was a "cripple and an outsider," though she was not always so. She began life as a small, blond-haired girl with a "silver giggle," who seemed no different from other children. However, by the time she reached the age of three, she was walking on the outside of her feet with an odd gait. Even so, she was a bright, fiercely determined child, who stomped around ignoring her disability as it gradually increased in severity. By the time she was 13, she stumbled and fell frequently, though with a mind so "bright, curious and hungry," her teacher had hopes that she too would one day be a teacher. The patient, if not beautiful at age 19, was slim and handsome enough to attract a suitor, who claimed that during their brief courtship, she could do anything–"row a boat, climb a tree, harness a horse, and drive a carriage." Her letters at that time, however, told a different story, one involving a series of "bad falls." Her one and only suitor vanished from her life as suddenly as he had appeared. The patient's balance soon worsened to the point that it was unsafe for her to look up without having a firm grip on something for steadiness. Although she was still able to walk, her crablike gait forced her to use the entire width of the road when ambulating. Her mother made her kneepads to wear under her skirt as protection against her many falls. Her hands, as yet unaffected, were capable of the intricate work of a talented seamstress. By the time she reached 26, the patient could walk only three or four steps without assistance, and her hands had become so misshaped and unsteady she had to use her wrists, elbows, and knees to do those things formerly done with her hands. Offers of help were gently but firmly refused. By the end of her fifth decade, she had lost the ability to stand and resorted to crawling to get where she wanted to go. Her mind continued to be as sharp as ever. No neurological disorders are known to have affected other members of the patient's family. Her father was a Swedish sailor with a disabling arthritis, who died at age 72 of unknown cause. Her mother developed kidney disease in her 40s and died edematous at age 68 of either renal failure or congestive heart failure. There were three brothers, one who died in his 80s of metastatic bone cancer. The medical histories of the other two are unknown. The patient was evaluated medically just once, when she was 26, at the Boston City Hospital. After a week of observation and tests failed to produce a diagnosis, she was told "to just go on living as [she] had always done. When the patient was 56, she developed a severe illness thought to have been pneumonia. One evening, while recuperating, she sat with one leg stretched out beneath a stove and fell asleep. When she awoke, the heat from the fire had seared the flesh from her withered leg. The third-degree burn healed slowly in response to repeated application of cod liver oil. At age 74, the patient finally consented to the use of a wheelchair and died shortly thereafter.



Story #2: ORIGAMI at work

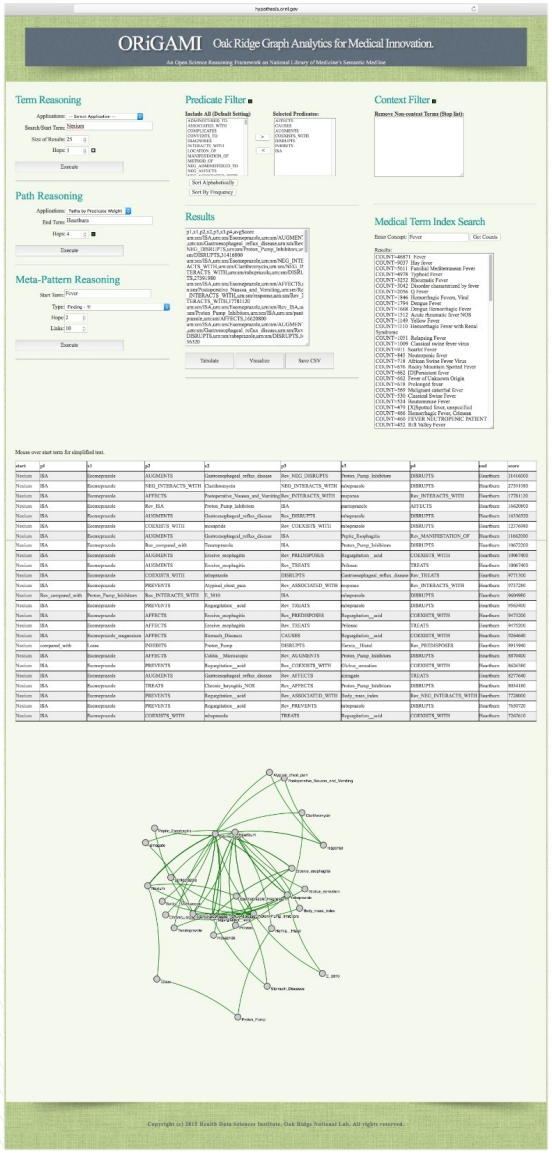
An Artificial Intelligence Workflow for Discovering Novel Associations in Massive Medical Knowledge Graphs



Neuromuscular disease - Distal muscle weakness

Gait - Distal muscle weakness

Gait abnormality - Falls frequently



Story #2: ORiGAMI at Historical Clinicopathological Conference

2015

2016

Story #2: ORIGAMI at Historical Clinicopathological Conference

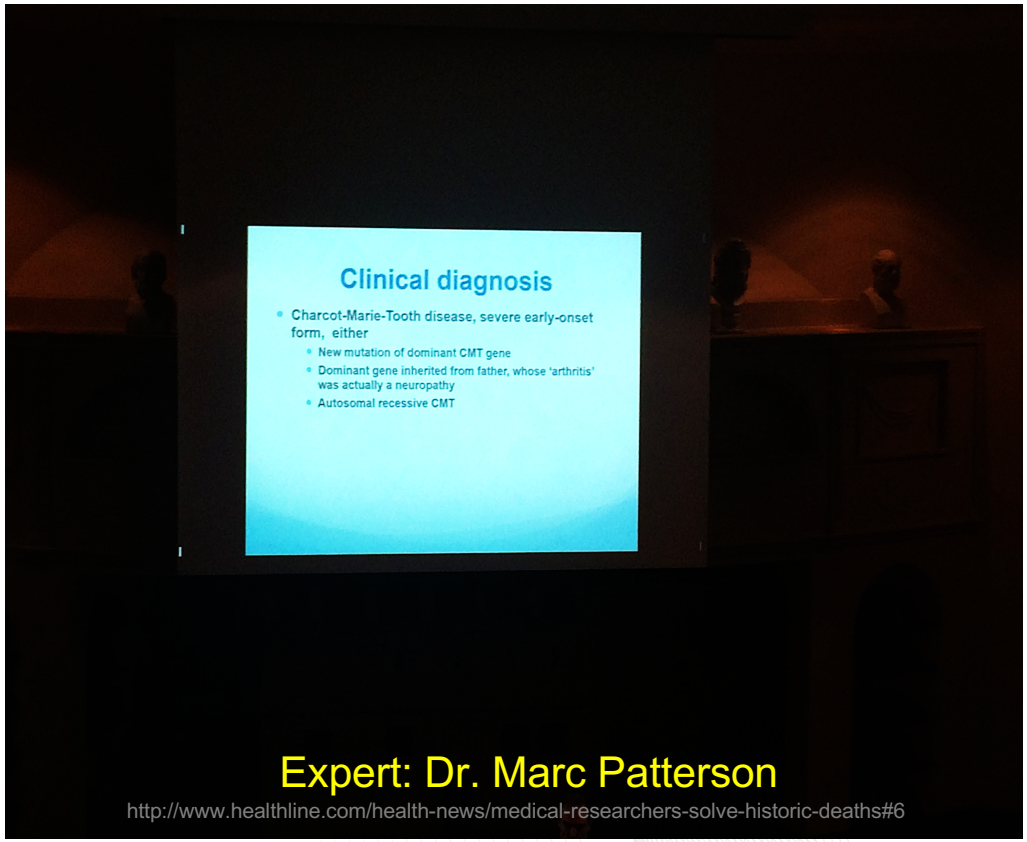
2015	2016
Malaria	**Charcot Marie Tooth Disease**
Lichen_disease	**Welander Distal Myopathy**
Urinary_tract_infection	Fasciitis Plantar
Coccidiosis	Talocalcaneal coalition
Bacteremia	**Cerebellar atrophy**
Encephalomyelitis_Western_Equine	Friedreich Ataxia
Poisoning_syndrome	Hypolipoproteinemia
Adult_Still's Disease	Multi infarct state
MRSA (Staph Infection)	Neuroleptic Induced Parkinson
Septecemia	Quadriplegic spastic cerebral palsy

Story #2: ORIGAMI at Historical Clinicopathological Conference

2015

Expert: Dr. Sanjay Saint + House staff

2016



Expert: Dr. Marc Patterson

http://www.healthline.com/health-news/medical-researchers-solve-historic-deaths#6

ORIGAMI now has answers for 23 previous mysteries….

Story #3: Since then…

- ORIGAMI has hypothesized the following “novel” associations
  - Beta-blockers accelerate diabetic retinopathy (Dr. Ed Chaum, UTHSC-Memphis)
  - Xylene is a potential “environmental” carcinogen (Dr. Gina Tourassi, ORNL)

- And is working on the following “potential” leads
  - Dysferlinopathy (Dr. Plavi Jain, Jain Foundation)
  - Sesamin as a nutraceutical (Dr. Ryan Yates, urxdna.com)
  - Identifying potential compounds for viruses like Ebola and Zika (?)
  - Extending beyond medicine to material science (Dr. Khalinin)

Summary and Conclusions

- Open Data + Open Software = Productive Reproducible Research
- Open Science is collaborative, empowering and rewarding
  - Careers : (Data journalists, citizen scientists, etc.)
  - Business models : (e.g. Careset, Accordian Health, etc.)
- Open Data requires "*quality patrols*" and Open Science requires "*courage, generosity and a sustainable incentive model*"
- The alternative to not supporting "Open Science" is scary - curbs creativity.