ORIGAMI: Oak Ridge Graph Analytics for Medical Innovation

An AI Workflow for Discovering Novel Associations in Massive Medical Knowledge Graphs

"Rangan" Sukumar , PhD

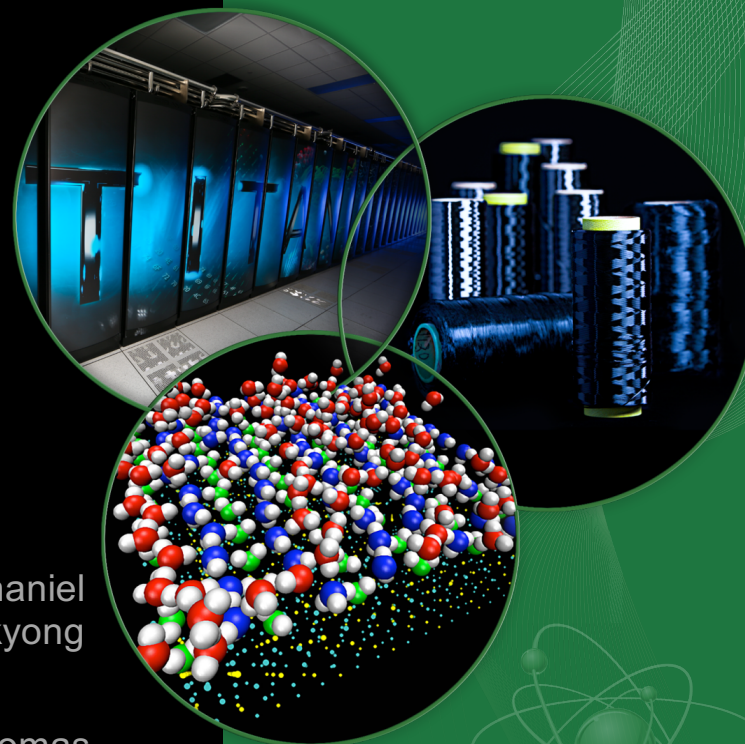## Data Scientist and Group Leader @ NCCS/OLCF

ORNL Team:

Larry Roberts, Matt Lee, Seung-Hwan Lim, Keela Ainsworth, Nathaniel Bond, Tyler Brown, Katie Senter, Alexandra Zakrzewska, Seokyong Hong, Yifu Zhao, Derek Kistler

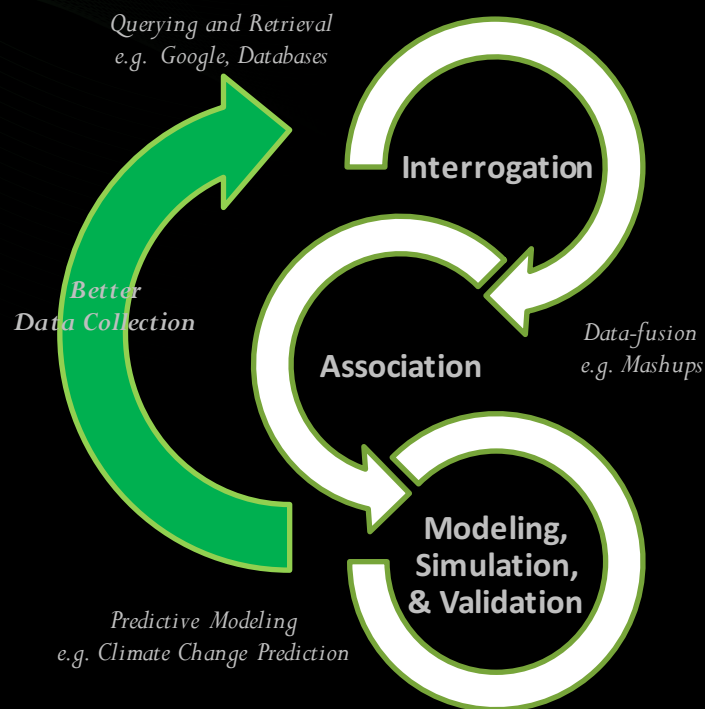**Collaborators at the National Library of Medicine**: Thomas Rindflesch, Marcelo Fiszman, Mike Cairelli

Academic Collaborators: Drs. Elliot Siegel (Maryland and Veterans Administration), Edward Chaum (UTHSC), Mark Wallace (Vanderbilt)

OAK RIDGE
National Laboratory

Group Vision: On-demand Data, Analytics and Workflows

The Lifecycle of Data-Driven Discovery

Querying and Retrieval
e.g. Google, Databases

Interrogation

Better
Data Collection

Association

Data-fusion
e.g. Mashups

Modeling, Simulation, & Validation

Predictive Modeling
e.g. Climate Change Prediction

Domain Scientist's View

The Process of Data-Driven Discovery

Science of scalable predictive functions

Pattern Discovery   **Pattern Recognition**

e.g. Hypothesis generation   *e.g. Classification, Clustering*

Science of data (Data-aware)

e.g. Deep learning, Feature extraction, Meta-tagging

Data science (Infrastructure-aware)

Shared-storage, shared-memory, shared-nothing
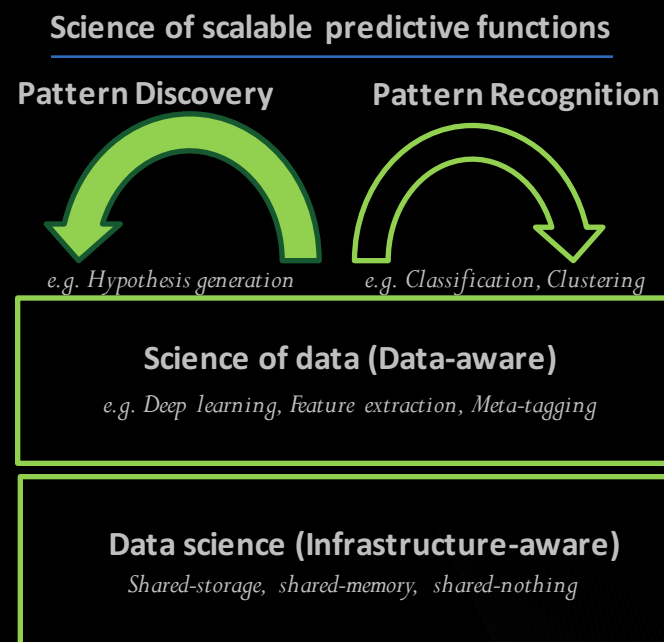
Data Scientist's View
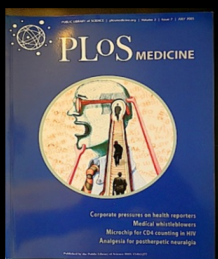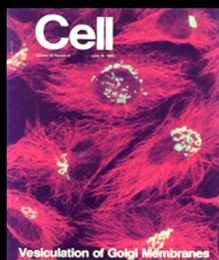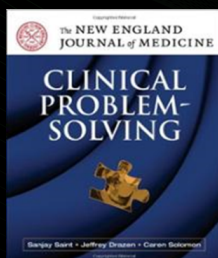
S.R. Sukumar, "Open Challenges in the Big Data Era – A Data Scientist's Perspective, IEEE Big Data , 2015.

OAK RIDGE
National Laboratory

What is ORiGAMI ?

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

The Genesis of ORiGAMI: On-demand Data

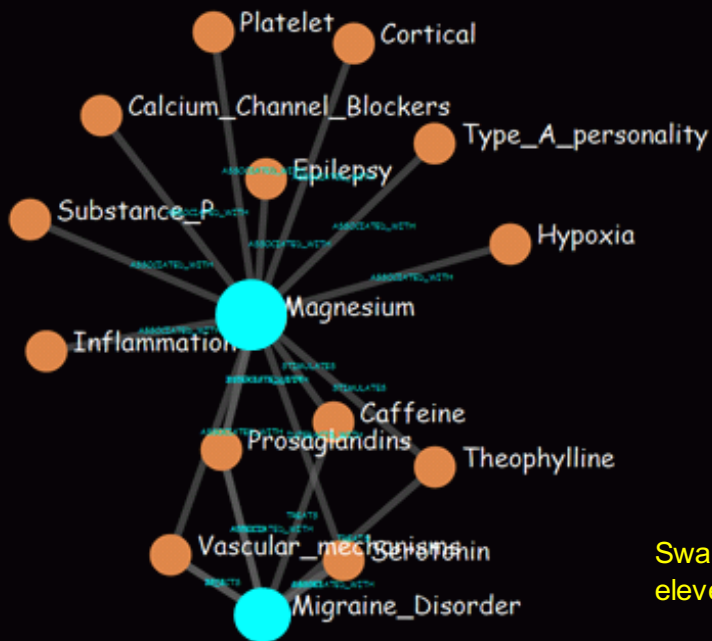## The National Library of Medicine

Number of papers processed: ~23.5M
Number of predications   : ~70 M
Number of distinct terms   : ~ 2 M
Number of "node-types"   :   133
Number of "relationships"   :   69

Dr. Thomas Rindflesch @ Health DataPalooza 2014

"We have a massive "heterogeneous" semantic graph that *researchers* need to query interactively…"
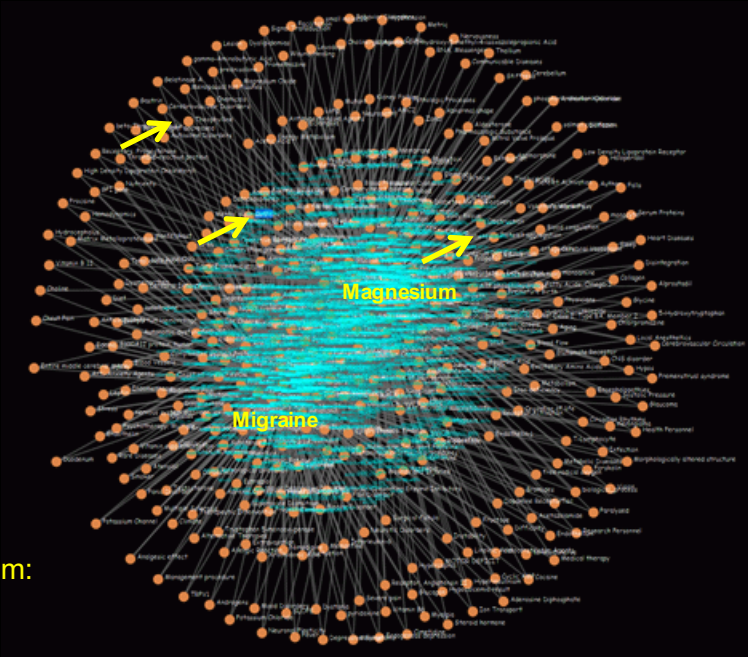
Where is the "Big Data" problem?

1987

2014

Swanson, Don R. "Migraine and magnesium: eleven neglected connections." (1987).

Today: There are 133,193 connections between migraine and magnesium.

OAK RIDGE
National Laboratory

We need a system to....

1. **Big Data:** Store, process, retrieve and reason with massive-scale datasets for newer discoveries.

2. **Signal-to-Noise Ratio:** Separate signal from noise when noise > signal

3. **Hypothesis Generation:** Given knowledgebase + new data, formulate 'new interesting questions'.

4. **Significance Ranking:** Given different knowledge nuggets, find significant associations or predict future connections.

OAK RIDGE
National Laboratory

ORiGAMI Today: An Open Eco-System for Data-Driven Discovery

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

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

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

Open Algorithms

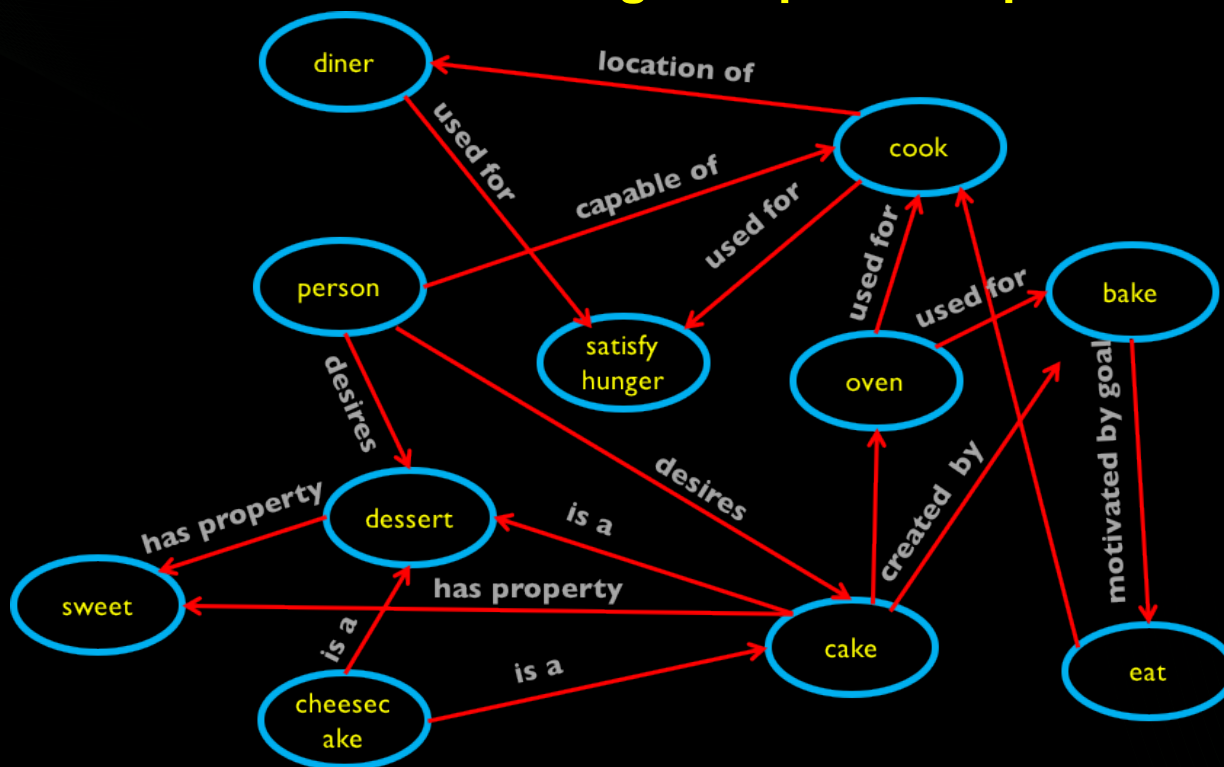http://github.com/ssrangan

- Data-driven reasoning
  - Semantic
  - Graph-theoretic
  - Statistical

- Model-driven reasoning
  - Term-based
  - Path-based
  - Meta-pattern
  - Context-based
  - Analogy

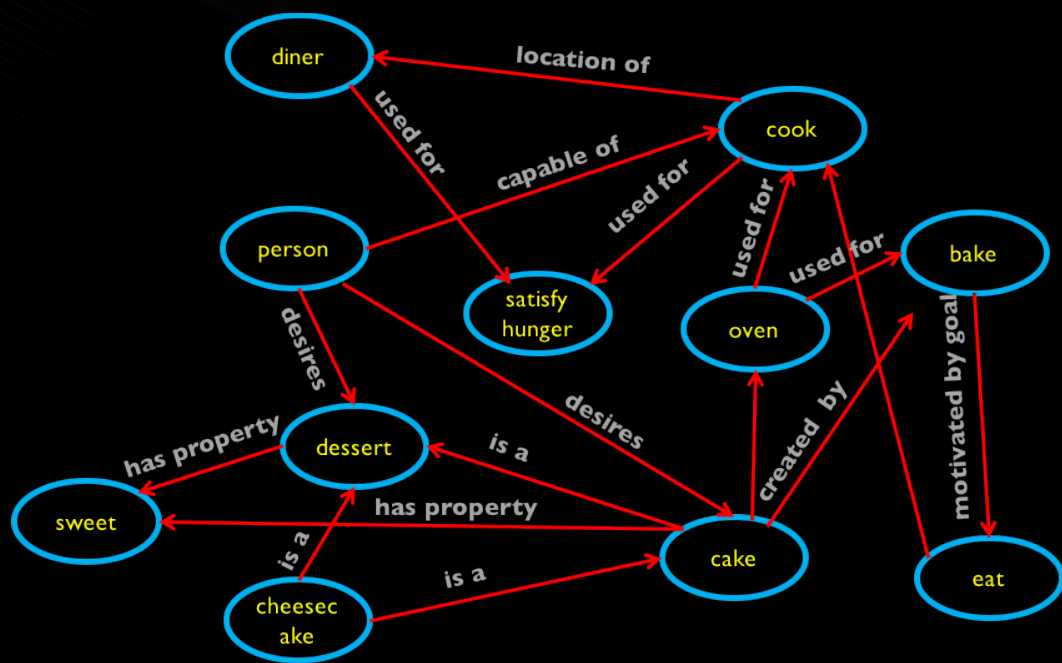## Open API: http://hypothesis.ornl.gov

How does ORIGAMI work?

Takes Knowledge Graphs as input…

~3.4 M words, ~57 K types of relationships, ~10 million assertions

What do these "Graph-operations" enable ?

Semantic, Statistical and Logical Reasoning at Scale

Curious Question:

Are giraffes vegetarian ?

~3.4 M words, ~57 K types of relationships, ~10 million assertions

What do these "Graph-operations" enable ?

Semantic Reasoning : "Connect the dots"

Answer #1

Giraffes only eat leaves.

Leaves are parts of trees, which are plants.

Plants and parts of plants are disjoint from animals and parts of animals.

Vegetarians only eat things which are not animals or parts of animals.

Answer #2

Giraffes is a mammal.

Human is a mammal.

People are human.

People who eat only vegetables are vegetarians.

Answer #3

Giraffes are desired by people.

Horses are desired by people.

Horses do not eat meat.

Herbivores do not eat meat.

Herbivores are vegetarians.

What do these "Graph-Operations" enable ?

Statistical Reasoning : "Search for the evidence"

Potential Answer #1

Deer is an animal.	Giraffe is an animal.
Deer is a mammal.	Giraffe is a mammal.
Deer has four legs.	Giraffe has four legs
Deer is a herbivore.	?
Herbivores are vegetarians.	?

Potential Answer #2

Horse is a large animal.	Giraffe is a large animal.
Horse sleeps standing.	Giraffe sleeps standing.
Horse has four legs.	Giraffe has four legs.
Horse is a mammal.	Giraffe is a mammal.
Horse is an animal.	Giraffe is an animal.
Horse only eat plants.	Giraffe eat leaves.
Horse is a herbivore.	?

OAK RIDGE
National Laboratory

What do these "Graph-Operations" enable ?

Logical Reasoning : "Filter noise from signal"

Extract Rules and Exceptions and Resolve:

A "is like" B. B "do not" C. D "only if" C. Therefore A "is not" D ?

A "is not like" B. B "do not" C. D "only if" C. Therefore A "may be" D ?

Giraffes are similar to deer, elephants, horses, cows, goats, sheep, toads, and humans.

Elephants, horses, cow and goat are herbivores, while humans and toads are not.

Result: Giraffes 'may be' vegetarian than not.

OAK RIDGE
National Laboratory

At scale: "How does Nexium treat Heartburn?"

This is an example of how our search 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.
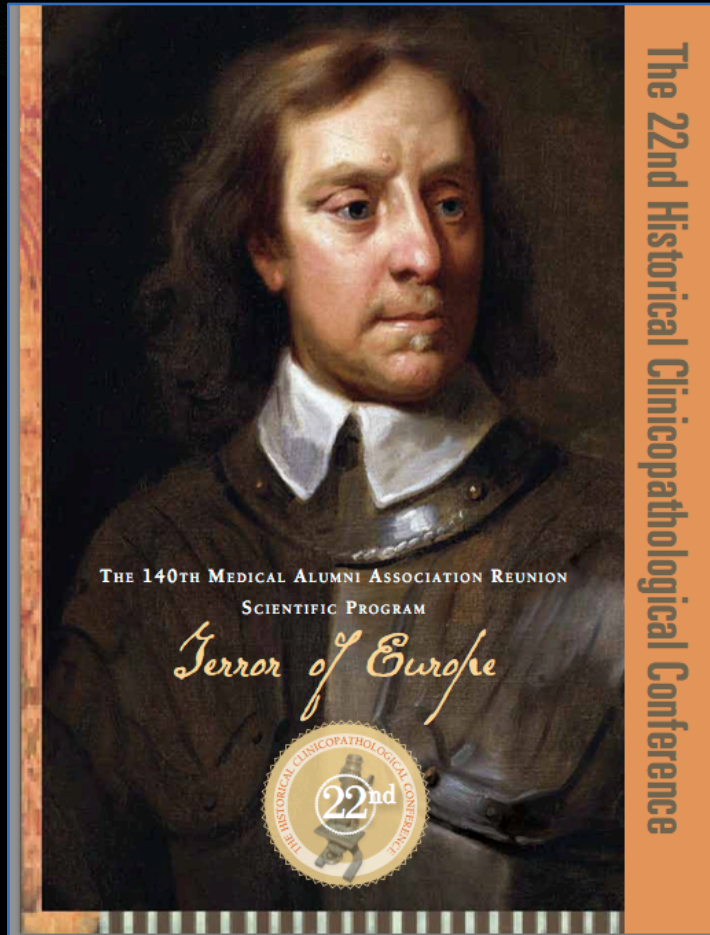
Abstract
There were authentic distinctions between the groups of healthy volunteers and patients with a peptic ulcer disease in Cmax, Tmax, AUC(0-t), AUC (0-infinity), Clt, Vd of omeprazole and Cmax of esomeprazole (Nexium, AstraZeneca). When the pharmacokinetics of omeprazole and ezomeprazole were compared 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.

OAK RIDGE National Laboratory

13

ORIGAMI @ Work...

The 22nd Historical Clinicopathological Conference

THE 140TH MEDICAL ALUMNI ASSOCIATION REUNION
SCIENTIFIC PROGRAM

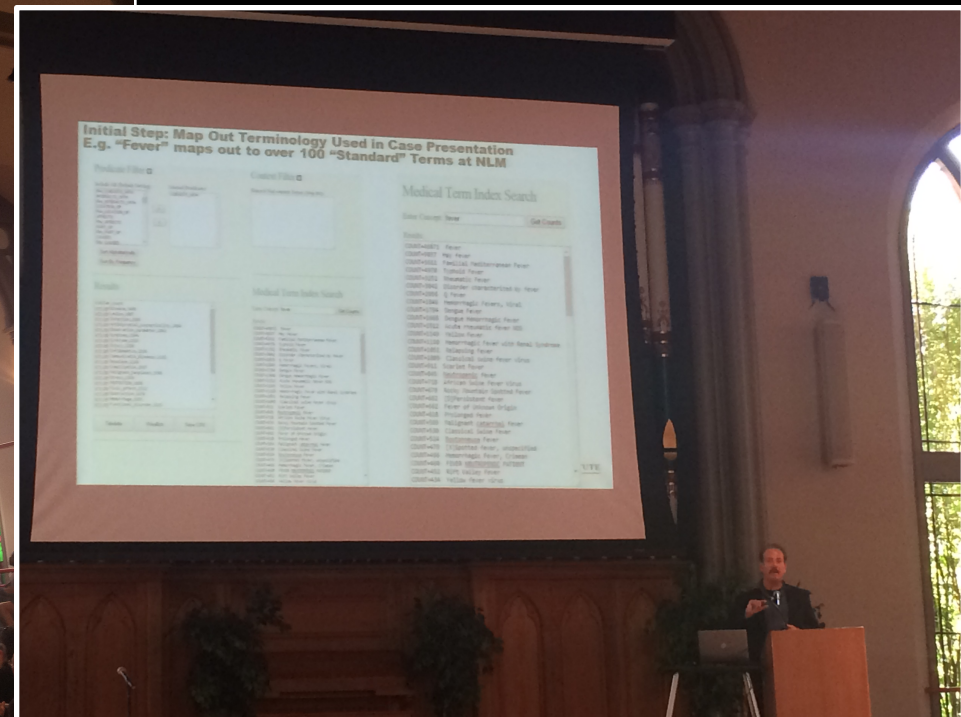*Terror of Europe*

22nd

Success Stories : ORiGAMI @ Work

1. Historical Clinicopathological Conference, October 2015

2. Hamilton Eye Institute, UTHSC, Summer 2014

…..and many more in the pipeline.

OAK RIDGE
National Laboratory

ORiGAMI @ Work: Historical CPC 2015, Baltimore

OAK RIDGE
National Laboratory

The Historical CPC Question

Other ailments : Arthritis, Dysentery, Herpes Simplex Infections, Vesicular eruption, Nephrolithiasis, Primary Insomnia, Soldier

Tertian fever
Cardiac arrhythmia
Intermittent fever
Intermittent pain
Sore Throat
Sweating

Welsh population
English Population
Longevity Hereditary

Peptic Ulcer
Irritable Bowel Syndrome
Pimples
Pustular acne

Multiple boils

Neck Injuries

Breast Cyst
Rash erythematous

Age

Birth
(1599)

29

45   48

57

59

Death
(1658)

Our Method

Using Semantic and Logical Heuristics

Symptoms

Intermittent fever

Tertian fever

Arrhythmia

Intermittent pain

Sore Throat

Sweating

Multiple boils

Specific context heuristic

Still_s_Disease__Adult_Onset
Brachyspira_pilosicoli
Malaria
Reactive_Hemophagocytic_Syndrome
Brachyspira_aalborgi
Chronic_Childhood_Arthritis
Endemic_Diseases

Pattern-similarity heuristic

HIV_Infections
Tuberculosis
Rheumatoid_Arthritis
Anemia
Functional_disorder
Acquired_Immunodeficiency_Syndrome

Semantic-similarity heuristic

Malaria
Malaria__Cerebral
Malaria__Vivax
Plasmodium_falciparum_infection
Disseminated_Intravascular_Coagulation
Myocarditis
Quartan_malaria
Abdominal_mass
Airway_disease
Aspergillosis__invasive
Ataxia__cerebellar__acute
Auricular_chondritis

Soldier

Aggressive_reaction
Albers_Schonberg_disease
Hallucinogen_Persisting_Perception_Disorder
Stress_Disorders_Post_Traumatic
Leishmaniasis_Cutaneous
Oppenheim's_Disease
Achilles_bursitis

Welsh

iodothyronine_deiodinase_type_I
Iodide_Peroxidase
EMD_21388
thyronamine
type_II_5__deiodinase
2_mercaptoimidazole
Acrocephaly
Acute_appendicitis_NOS
Adrenocortical_carcinoma
Anxiety_neurosis__finding
Bronchiectasis
CDK8_gene_CDK8
Central_Nervous_System_Agents
Complement_component_C4_C4A
Congenital_anomaly_of_cartilage
DIO1_DIDO1
DIO2_gene_DIO2
DUOX1
DUOX2

Multiple boils

Empyema
Gangrenous_pneumonia
MRSA_infection
Meniere_s_Disease
Nosocomial_pneumonia
Pericarditis
Septicemia
Stenosis_unspecified

Our Method

Step 2: Conceptual Reasoning and Validation

Probabilistic filtering using context proximity

Intermittent fever

Tertian fever

Plasmodium_ovale
Malarial_hepatitis
Anopheles_darlingi
Mixed_infectious_disease
Plasmodium_species
Quartan_malaria
Plasmodium_falciparum_infection
Congenital_malaria

Sore Throat

Sweating

Multiple Boils

Gomphosis_structure
Adductor_spastic_dysphonia
Goblet_Cells
Cold_symptoms
Acute_radiation_proctitis
Common_Cold
Genu_varum
Amyloidosis__Familial
GLI3_gene_GLI3
Diverticulitus

18
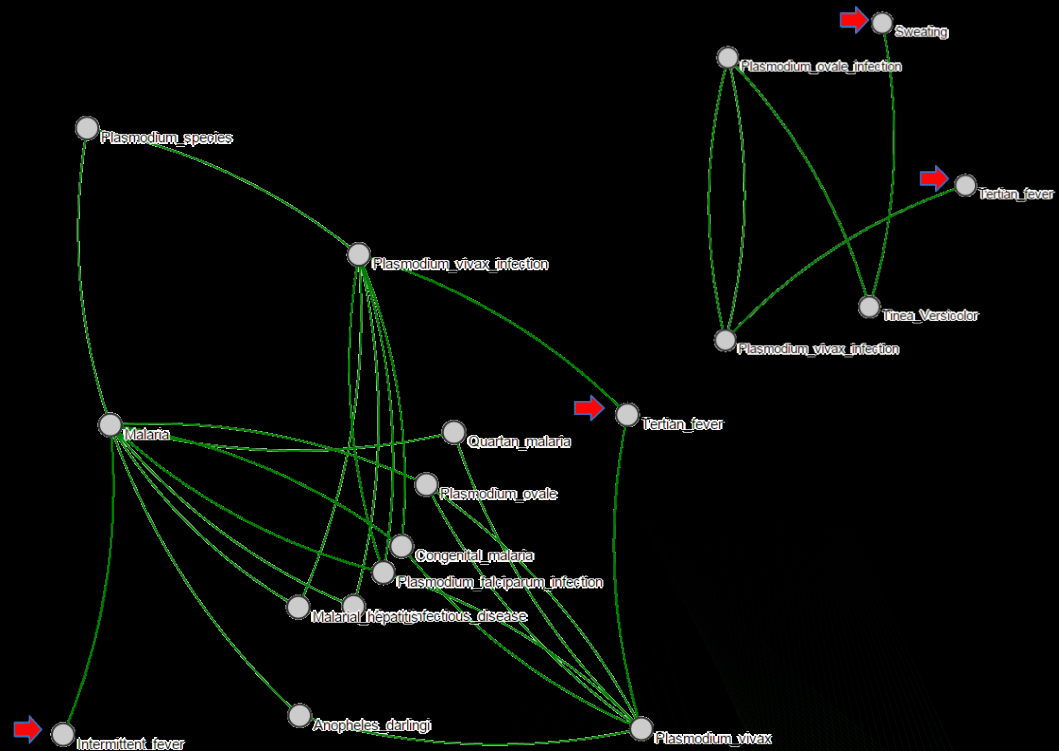
Our Method

Step 3: Personalized Hypothesis Generation

Using Random walks with restarts
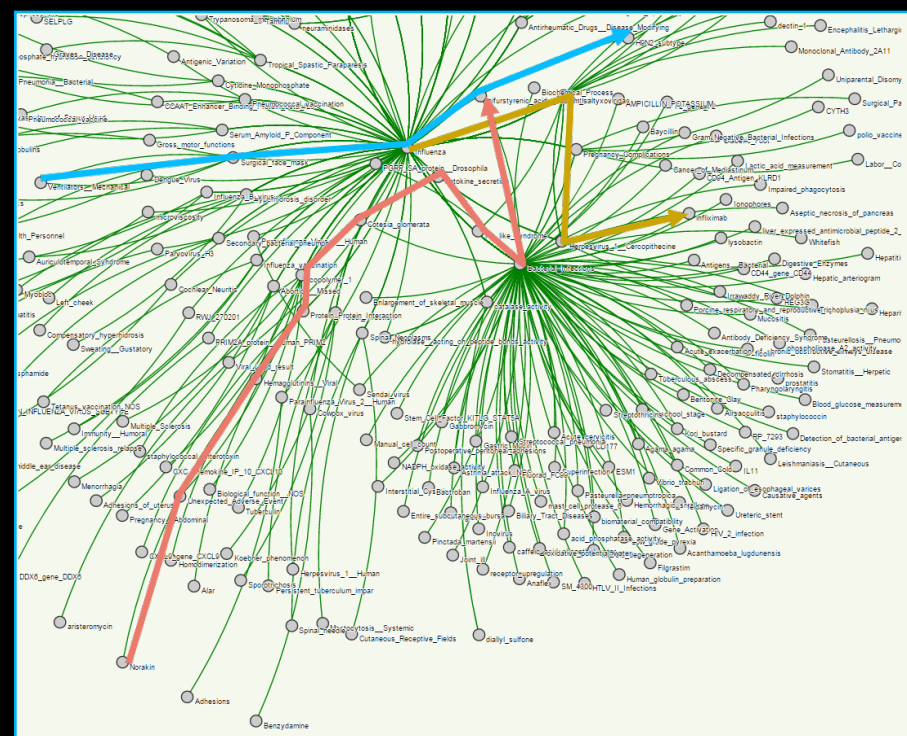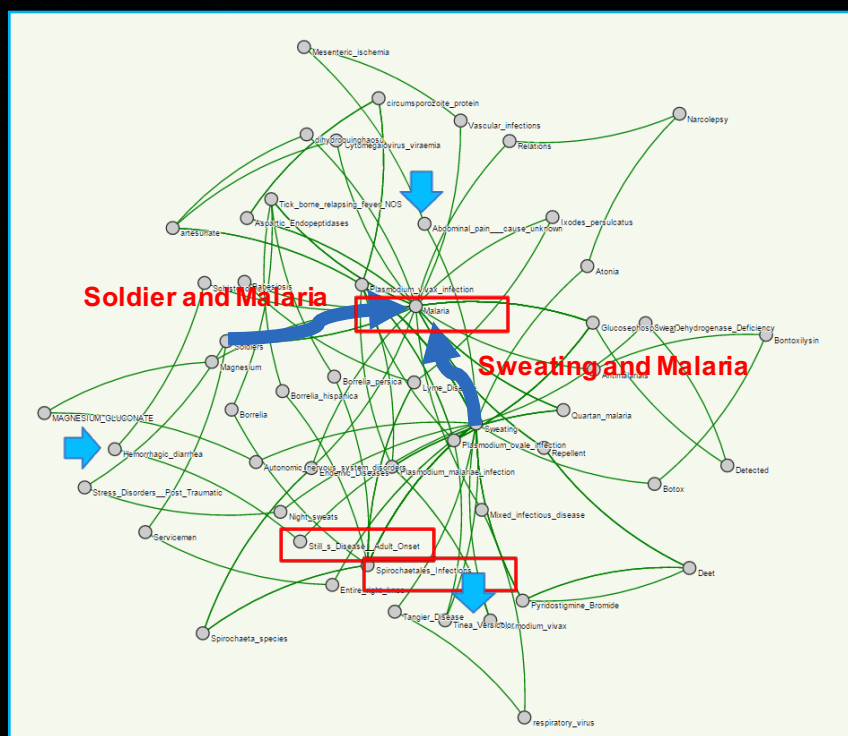
ORiGAMI 's Answers

Top 10 Hypothesis generated using ORiGAMI

Malaria
Lichen_disease
Urinary_tract_infection
Coccidiosis
Bacteremia
Encephalomyelitis_Western_Equine
Poisoning_syndrome
Adult_Still's Disease
MRSA (Staph Infection)
Septecemia

OAK RIDGE
National Laboratory

ORiGAMI @ Work: Historical CPC 2015, Baltimore

Crowdsourcing the experts

ORiGAMI @ LA Times and Washington Post

What killed Cromwell? Or Mozart? Sleuthing doctors take on the 'ultimate whodunit'
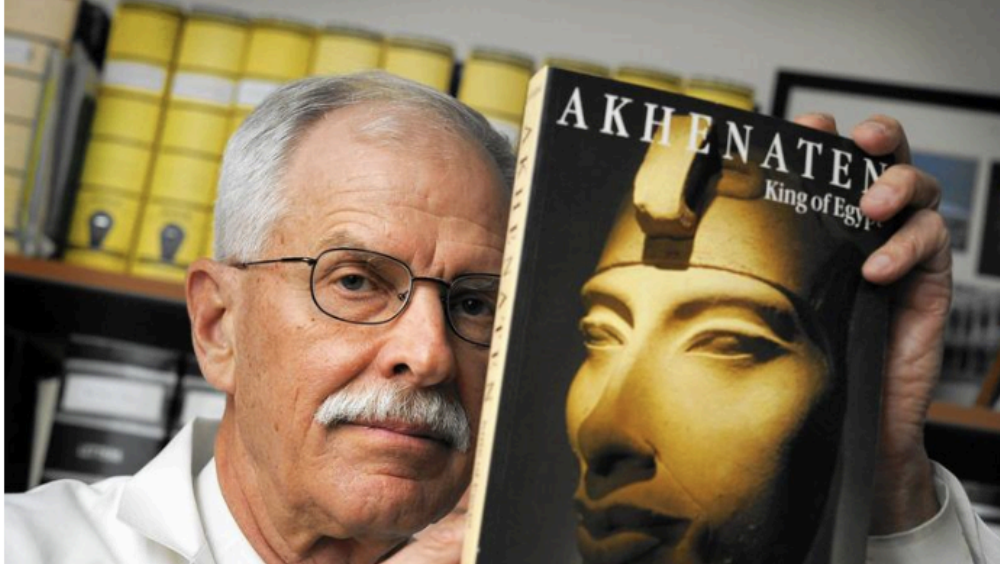
The Historical Clinicopathological Conference, launched by Dr. Philip A. Mackowiak, has examined the deaths of Herod the Great, Pericles and Akhenaten. (Lloyd Fox / Baltimore Sun)
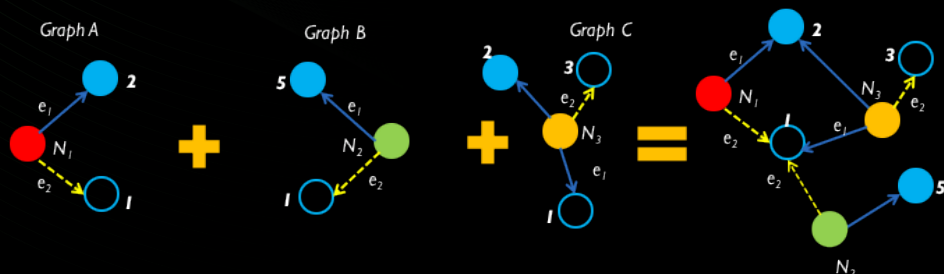
In 4.5 seconds, ORIGAMI — short for Oak Ridge Graph Analytics for Medical Innovation — converged on virtually the same conclusion drawn after weeks of research and deliberation by Saint: Cromwell was done in by malaria.

"Who would I rather have making a diagnosis? It would be hands-down Dr. Saint," Siegel said. On the other hand, "if you told me there was a mystery disease and no one had any ideas about it and we needed some new insight ... I like the computer."
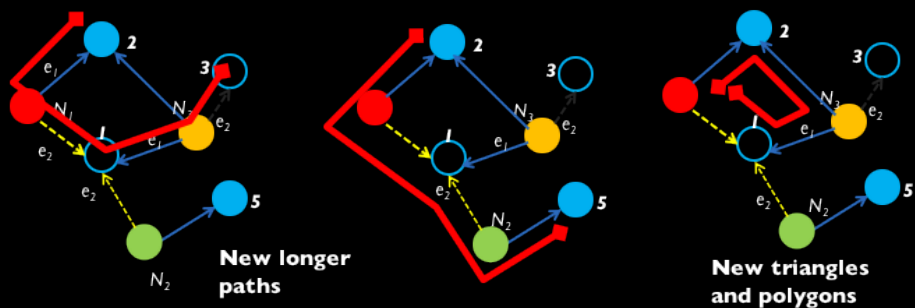
OAK RIDGE
National Laboratory

Questions?

Extracting Novel and Useful Associations

Predictive Inferencing using Recommender Algorithms

Where are the new connections?
What are the "important" new connections?

New longer paths

New triangles and polygons

Knowledge Graph

Streaming Data
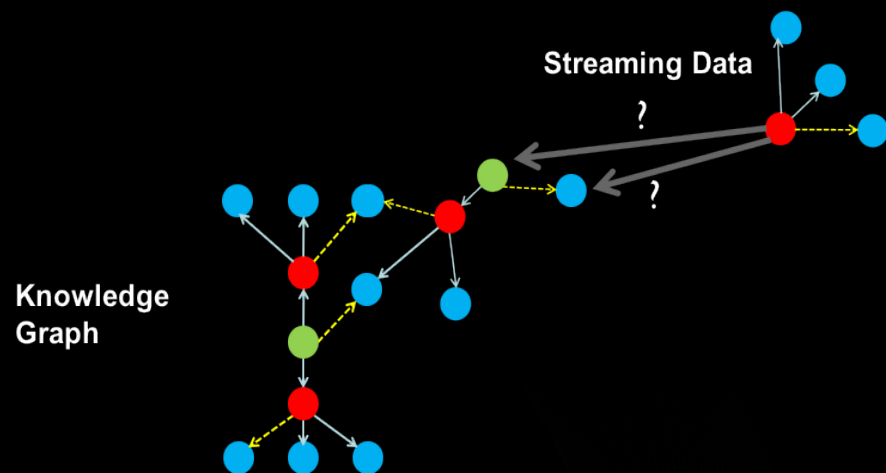
What is the probability of a particular edge to occur ?

OAK RIDGE
National Laboratory

23