Nuclear Forensics Attribution as a Digital Library Search Problem

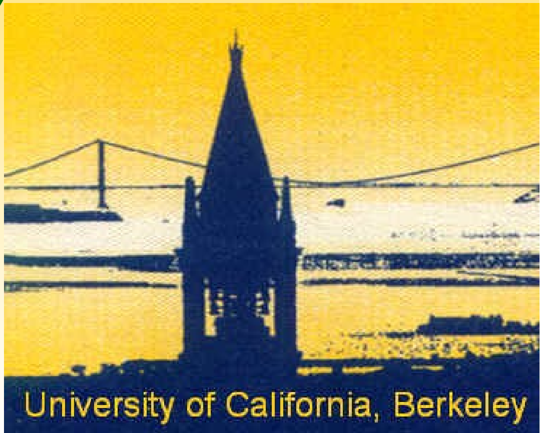
ARI Grant Oral Presentation, Washington DC, July 23, 2012

- Fredric Gey (PI), Ray R Larson (co-PI), Electra Sutton (scientist) (students) Chloe Reynolds, David Weisz, Matthew Proveaux

  - Institute for the Study of Societal Issues, The Information School and
  - Nuclear Engineering Department
  - University of California, Berkeley
  - http://metadata.berkeley.edu/nuclear-forensics

- First year funding source
- National Science Foundation Grant ***#1140073:***
  "ARI-MA Recasting Nuclear Forensics as a Digital Library Search Problem"

- Thanks to Bethany Goldblum UCB Nuclear Engineering for helpful comments



Nuclear Forensics Attribution as a Digital Library Search Problem

- **Reframes the problem** of nuclear forensics discovery (identifying the source of smuggled nuclear material) as a digital library search problem against large libraries of analyzed nuclear materials, i.e.
  - Spent fuel from a nuclear reactor after fission
  - Enriched uranium or plutonium in the nuclear fuel
  - Refined uranium ore (yellow cake) from mines
- **Develops multiple models of the nuclear forensics search process** similar to how traditional forensics (fingerprint and DNA matching) benefited from specialized data representations and efficient search algorithms

Nuclear Forensics Search Models

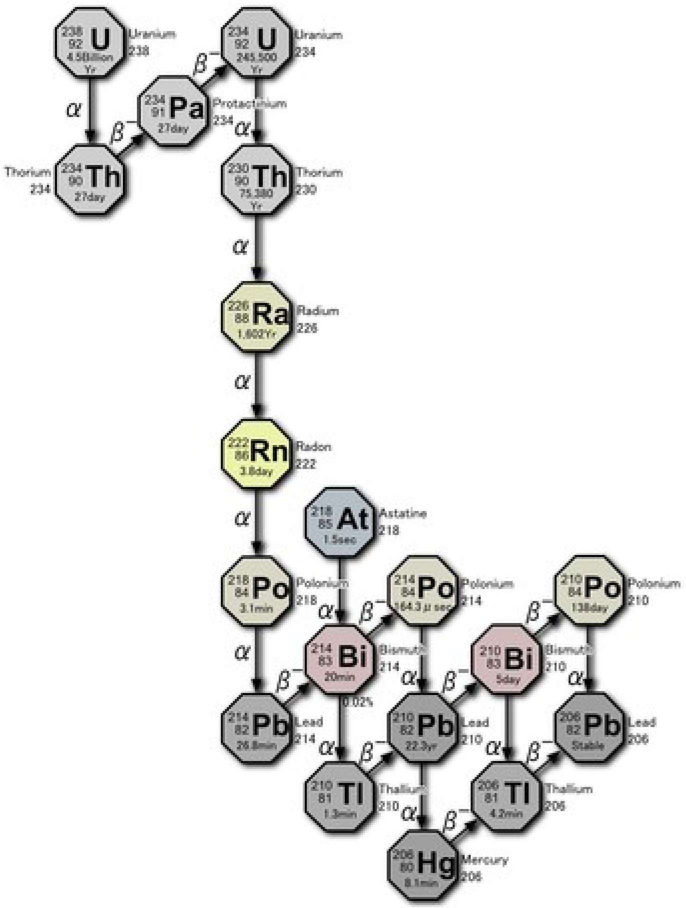
Nuclear forensics search can be framed as a:

1. **Directed graph matching problem** (in particular a weighted, labeled directed graph matching problem)
2. **Automatic classification problem** where machine learning is applied to classify a seized sample
3. **Process logic problem**, whereby the forensic investigation capture the steps and logic which a human nuclear forensics expert would approach

Search Model: Directed Graph Matching

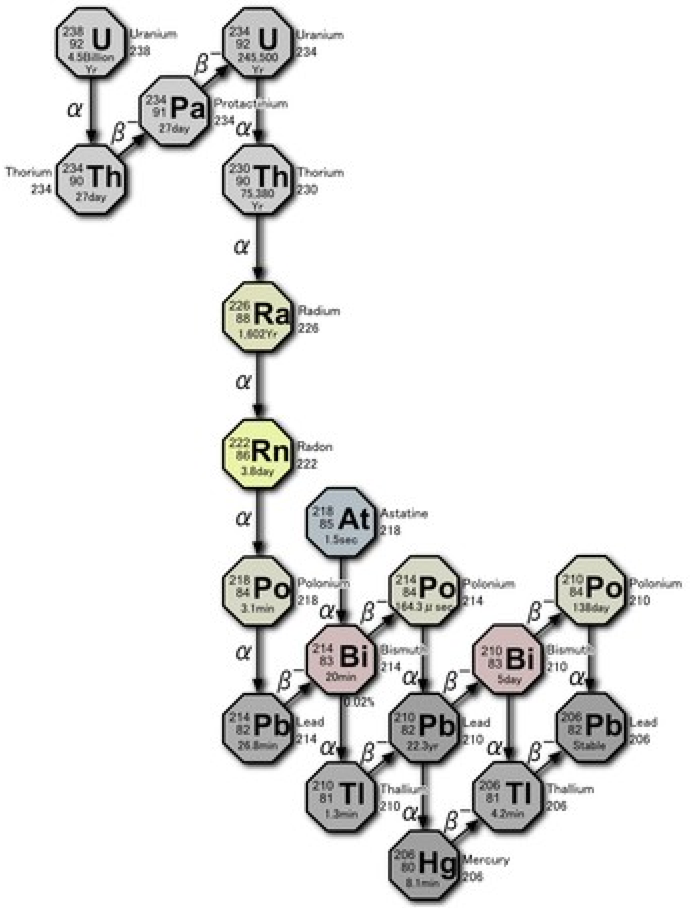
Represented as a Graph G = (V,E), a nuclear sample consists of a finite number of vertices (sometimes referred to as nodes) v_1 ... v_n representing elements in a decay chain.

For Uranium 238, n=19, v_1 = ^{238}U v_2 = ^{234}Th and v_{19} = ^{206}Pb the terminal stable element of lead. Associated with each vertex at time t_m, is an amount $m(t_m)$, the measured mass of the element at the time of measurement. The edges (or arcs) between elements represent the decay direction: thus $e_{7,8}$ = (^{226}Ra,^{222}Rn), represents the decay path from Radium to Radon.



Search Model: Directed Graph Matching

A seized material sample at time t_m, is referred to as $G_s(t_m,)$. Let us further say that there exist a digital library of k samples each measured at different times LIB=$\{G_1(t_1) \ldots G_k(t_k)\}$. We wish to match the seized sample to appropriate library samples. But there are differences in times of measurement – to do the match we have to forwardly compute each of the library samples from $t_{k,}$ to time t_m (or backwardly compute the seized sample from time time t_m to time $t_{k,}$). Thus we seek a similarity function:



$$SIM\,(G_s(t_m,),G_i(t_i)\ \varepsilon\ LIB) = SIM(G_s(t_i)=f_b(G_s(t_m,),G_i(t_i))\ \varepsilon\ LIB)$$

for the ith sample in the library and where f_b is the backward computation function.

This is the simplest model – in reality, all samples may have additional geolocation clues L (manufacturing, irradiation period, operation history, etc) which may or may not have a time dependency. Thus G = (V,E,L) for a more complex model.

Nuclear Reactor Database (Unifying Multiple Datasets)

University of California, Berkeley

We wanted a comprehensive detailed database about worldwide nuclear reactors including geographic coordinates

Searches for “nuclear dataset” and similar terms

- 200+ datasets found on web
- 80+ datasets downloaded (arbitrary subset)
  - Sorted into useful (65) / not useful (15) categories
  - Not useful example: Nuclear capacity by country
- Consolidation, done by choosing 5 reputable datasets (e.g. IAEA) and creating a unified database
- Unified dataset into a Google Earth viewer

Nuclear material could come from any of about 500 nuclear power plants worldwide

(Worldwide Nuclear Power Plants using Google Earth)

Original data source: http://maptd.com/worldwide-map-of-nuclear-power-stations-and-earthquake-zones

Supplemented with additional nuclear plant data from IAEA

Other Data Sets Assembled or Being Assembled in Support of the Project

The Nuclear Wallet Cards, J.K. Tuli, National Nuclear Data Center, Brookhaven National Laboratory.

Plutonium Metal Standards Exchange Program, Los Alamos National Laboratory (to benchmark code)

Reactor Isotopic composition data from Spent Fuel Isotopic Composition Database (SFCOMPO), OECD Nuclear Energy Agency (NEA)

Atomic Mass Data Center, CSNSM Orsay, France and hosted by National Nuclear Data Center (BNL, USA)

International Atomic Energy Agency (IAEA) nuclear material processing practices and telltale isotopic

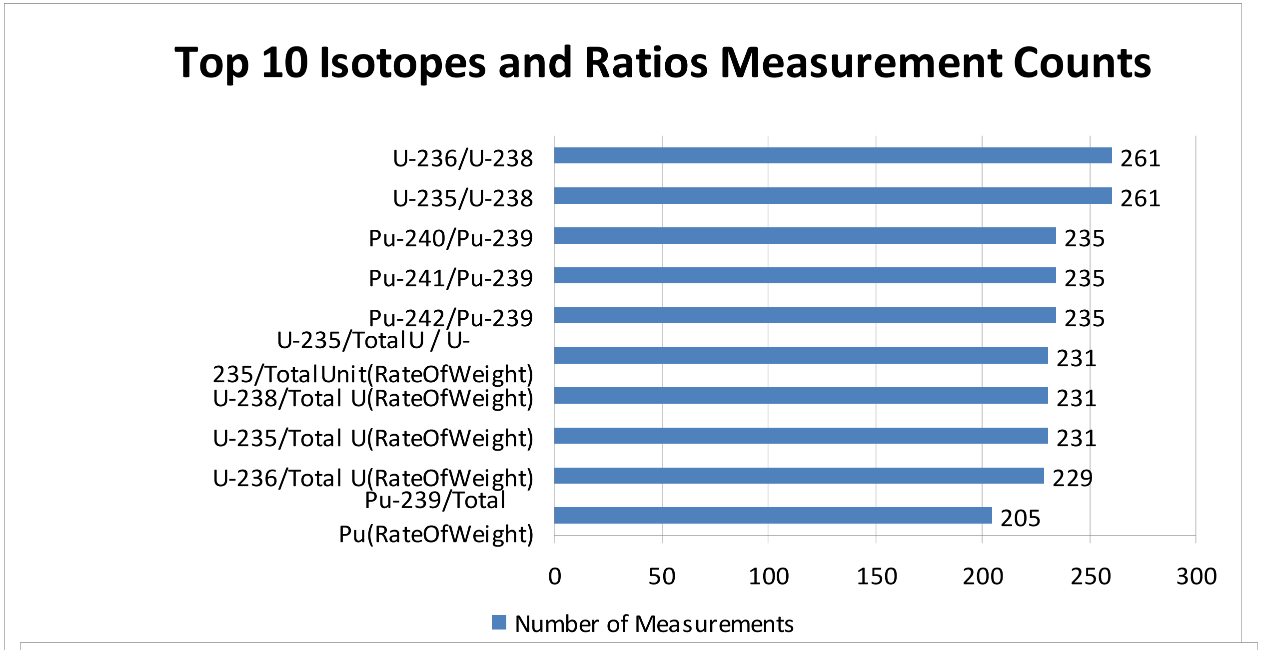
Nuclear Fuel Cycle and Weapon Development Cycle, Prepared for DOE by the Pacific Northwest National Laboratory.

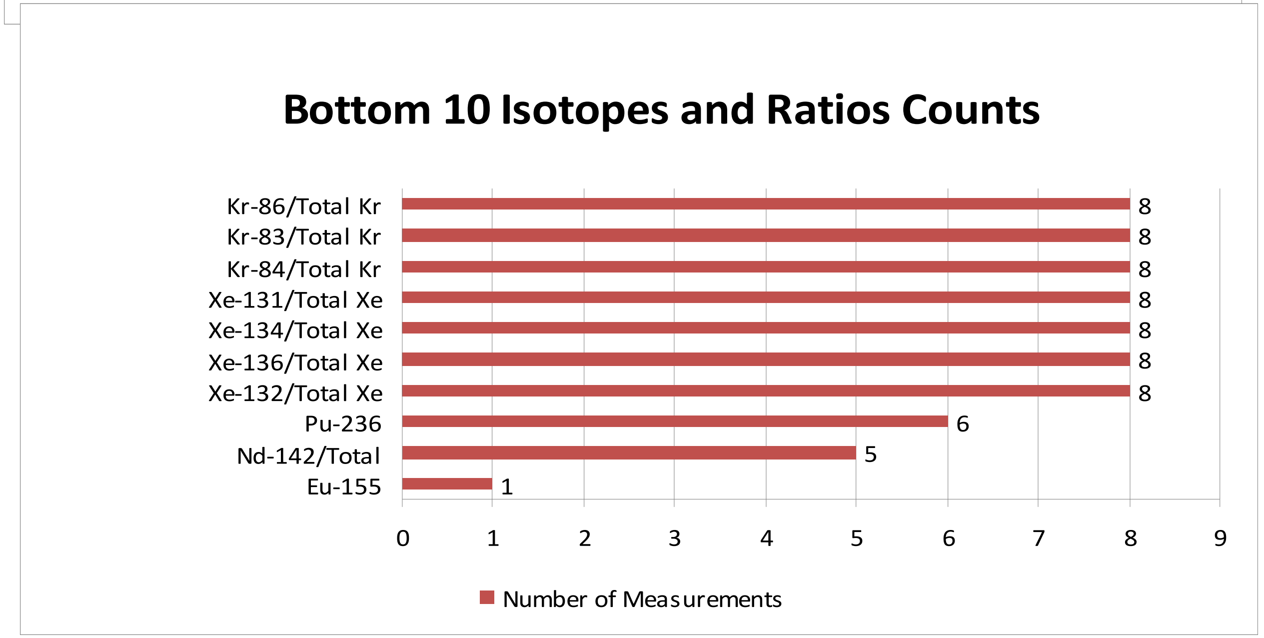
Spent Nuclear Fuel Database SFCOMPO (source: OECD Nuclear Energy Agency)

To experiment, we downloaded this spent fuel measurement database (html tables) from the web :

- **14 reactors from 4 countries (light water. BWR,PWR) Germany, Italy, Japan, USA**
- **261 Samples (variable number per reactor)**
  - **Maximum samples (Trino Vercellese, IT): 39**
  - **Minimum samples (Genkai-1, JA): 2**
- **10,340 Measurements of Isotopes, Isotope Ratios and Burnup, (variable number for each sample)**

SFCOMPO Spent Nuclear Fuel Variable Measurement Characteristics





Nuclear Murder and Attribution

- On November 1, 2006, Alexander Litvinenko, former Russian Federal Security officer was poisoned by Polonium-210 isotope while having lunch at a London sushi restaurant. He died of radiation poisoning three weeks later.
- According to doctors, "Litvinenko's murder represents an ominous landmark: the beginning of an era of nuclear terrorism."
- Polonium-210 (^{210}Po) is an isotope of Polonium with a significant half-life (138 days). It decays by emitting alpha particles which can be easily shielded by even pieces of paper or the human skin
- UK authorities traced the material to a specific nuclear reactor in Russia **HOW DID THEY DO THIS?**

SFCOMPO Spent Nuclear Fuel Data
A Naive Search Experiment: Structure

1. Assume the temporal effects are negligible on measurements and measurement ratios
2. A single sample is removed from the set of samples in the database. This sample becomes the "query sample" (the seized sample of unknown origin) and all other 260 samples are the "document samples" (to invoke search terminology).
3. A similarity matching algorithm is developed which matches each measurement in the query sample with its equivalent measurement in each document sample. This match results in a number between zero and 1 called a Retrieval Status Value (RSV) (ideally it is a estimate of a matching probability).
4. Document samples are ranked by this RSV similarity value.
5. Relevance of the document sample to the query sample is assessed as follows:
   1. If a document sample comes from the same reactor as the query sample, then the document sample is judged relevant.
   2. Otherwise it is Irrelevant
6. Standard web retrieval performance measure (precision at rank 10) is used

Search Experiment Performance Measure

1. The standard measure of performance for web retrieval is the computation of precision at rank ten.
2. Precision for each ranked document (web page) is the fraction of relevant documents divided by the rank number, i.e.
    1. If the first document is relevant, precision at 1 is 1.0
    2. If the second document is irrelevant, precision at 2 is 0.5
    3. If the third document is relevant, precision at 3 is .667
    4. If the fourth document is irrelevant, precision at 4 is again 0.5
3. Only the first ten ranked web pages are judged for relevance or irrelevance

SFCOMPO Search Experiment: Overall and Performance by Reactor

University of California, Berkeley

Precision-at- Rank-10, by Reactor					
Average Precision@10 over 261 query samples				0.34	
Reactor	Country	Number of Samples	Maximum Possible Precision	Precision (per Reactor)	Actual / Maximum Precision
JPDR	Japan	30	1.00	1.00	100%
Monticello	USA	30	1.00	0.85	85%
Tsuruga-1	Japan	10	0.90	0.53	59%
Trino_Vercellese	Italy	39	1.00	0.24	24%
Fukushima-Daini-2	Japan	18	1.00	0.21	21%
Takahama-3	Japan	16	1.00	0.16	16%
Fukushima-Daiichi-3	Japan	36	1.00	0.16	16%
Obrigheim	Germany	23	1.00	0.15	15%
Genkai-1	Japan	2	0.10	0.10	100%
H.B.Robinson-2	USA	7	0.60	0.09	14%
Cooper	USA	6	0.50	0.07	13%
Gundremmingen	Germany	12	1.00	0.06	6%
Mihama-3	Japan	9	0.80	0.06	7%
Calvert_Cliffs-1	USA	9	0.80	0.06	7%

Search Experiment Implications and Next Steps

1. Performance seems promising considering the crudeness of the assumptions (however we may only be correlating burn-up -- needs further investigation)
2. What might happen if the following improvements were made?
   1. All measurements are available instead of selected ones
   2. All measurements are normalized to a uniform precise time
3. Our collaborators at PNNL (funded by DNDO/NTNFC) are doing just that, by computationally:
   1. Filling in (imputing) the missing values
   2. Normalizing the actual and imputed measurements to a precise time
4. The target date for PNNL to complete this task is August 15.
5. We will then re-run our search experiment on the “improved” database
6. PNNL is expanding the database to other reactor types (e.g. graphite moderated)

Future Directions and Activities

1. Expand collaboration to forensics and data groups at LLNL and ORNL
2. Attend the SFCOMPO meetings at OECD/NEA September 17-18 in Paris
3. Seek to assemble data for Uranium mines/ores for equivalent search experiments
4. Initiate contacts with IAEA data groups and the Institute for Trans-Uranium Elements in Karlsruhe, Germany. Seek out equivalent groups in Japan.
5. Begin to create nuclear forensics educational materials in collaboration with the UCB Nuclear Engineering Department

Collaborators/Subject Matter Experts

Department of Nuclear Engineering, University of California, Berkeley (Bethany Goldblum, Prof. Jasmina Vujic)

Nuclear Systems Design, Engineering and Analysis, Pacific Northwest National Laboratory, Richland, WA (Michaele (Mikey) Brady Raap, Jon Schwantes)

Nuclear Science Division Isotopes Project, Lawrence Berkeley National Laboratory, (Richard Firestone)

Chemistry & Materials Science Division, Los Alamos National Laboratory, Los Alamos, NM (Lav Tandon. Kevin Kuhn)

Students

University of California, Berkeley

Chloe Reynolds, Masters of Information Management and Systems, School of Information, June 2012

Matthew Proveaux, incoming Masters student, Nuclear Engineering (MS 2014) (BS Physics, UC Davis)

David Weisz, incoming PhD student, Nuclear Engineering (MS Health Physics, nuclear non-proliferation track, Georgetown University), summer only.

Charles Wang, incoming Masters student, School of Information (MIMS 2014) (B.S. computer science)

Actively recruiting for fall 2012

Planning a steady state of 2+ graduate students until the FY 2013 budget situation is clarified.

Seeking NSF REU Undergraduate funding for summer 2013

Publications and Presentations

University of California, Berkeley

"Database Heterogeneity in a Scientific Application," poster presentation at the IASSIST 2012 conference, June 6, Washington DC

"Applying Digital Library Technologies to Nuclear Forensics" to be published at the *International Conference on Theory and Practice of Digital Libraries (TPDL),* Cypress September 23-27, 2012

"Nuclear Forensics: A Scientific Search Problem" submitted to LWA 2102: *Lernen, Wissen, Adaption* Dortmund, Germany, September 12-14, 2012

Nuclear Forensics Search

Grant home page

http://metadata.berkeley.edu/nuclear-forensics

- Contacts
  - Fred Gey (gey at berkeley dot edu)
  - Ray Larson (ray at ischool.berkeley dot edu)
  - Electra Sutton (electra at berkeley dot edu)