This is Sandra. Thanks for joining me today for this presentation about nanoinformatics. This presentation is provided in 3 parts. In part 1, I talk about the big picture background for why I developed N4mics. N4mics is an informatics tool that can be used to explore the data stored in the Nanomaterial-Biological Interactions Knowledgebase. In part 2, I will explain the features of N4mics, focusing on how to interpret the tool output; and in part 3, I will demonstrate the use of the N4mics tool.

A bit about me before we get started; I graduated from Carnegie Mellon University with a PhD in Civil and Environmental Engineering in 2005. Prior to starting graduate school, I worked for several years as a consultant to environmental contractors, helping them organize and format data for inclusion in the Air Force's data repository.

Since graduating, I have completed two postdocs, taught undergraduate courses in water treatment and solid waste management at Geneva College, and taught a graduate level environmental database management course at Carnegie Mellon University.

I am a licensed professional engineer in the state of Pennsylvania.

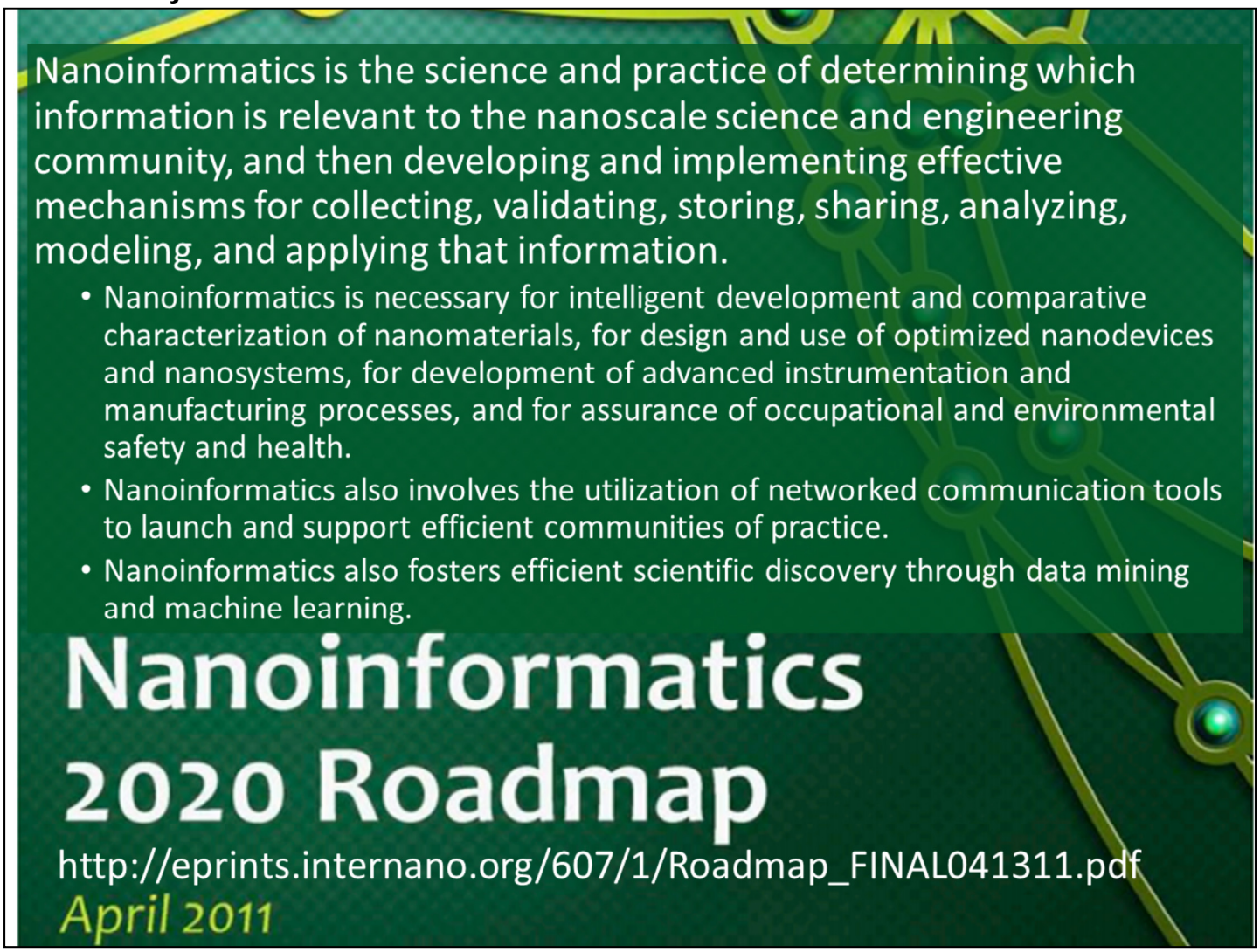
If you need to reach me, the best way to contact me is by sending an email to SandraKarcher44@gmail.com.

In case you are wondering, this photo was taken on I-90 in Washington State, near the Wild Horse Monument. In the background is the Vantage Bridge - crossing the Columbia River.

The Big Picture Background

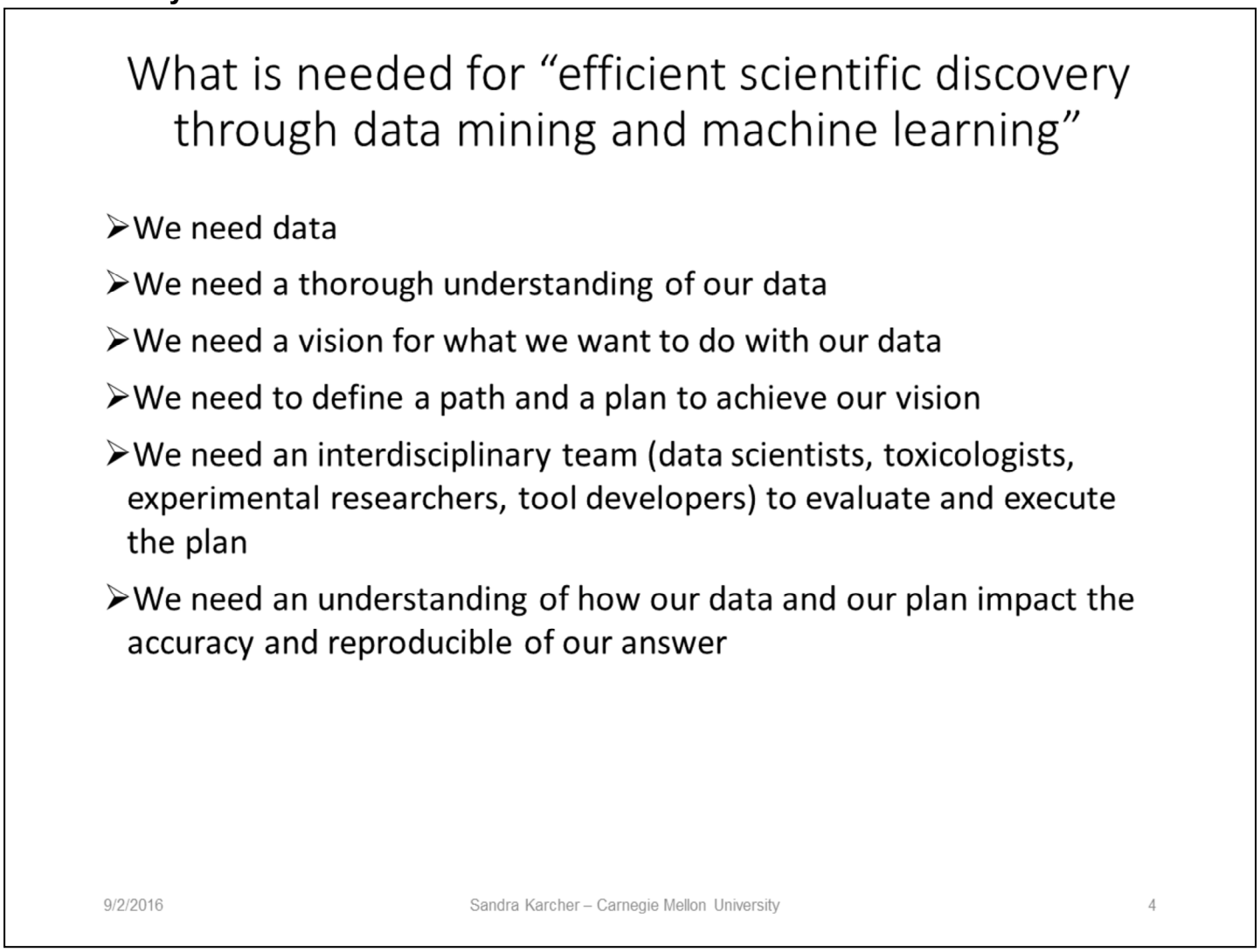
Part 1

Welcome to part 1 of the 3 part series on nanoinformatics by Sandra Karcher.



What is nanoinformatics? A working definition is provided in the Nanoinformatics 2020 Roadmap which states: Nanoinformatics is the science and practice of determining which information is relevant to the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying that information.

This working definition is followed by three bulleted items. I want to focus here on the third of those bullets: “Nanoinformatics also fosters efficient scientific discovery through data mining and machine learning.”

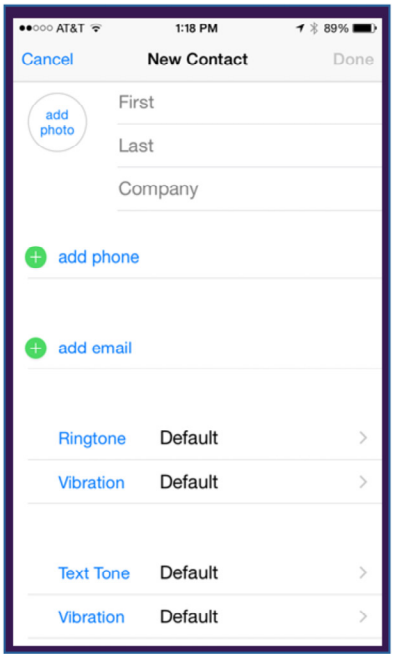


What is needed for "efficient scientific discovery through data mining and machine learning"?

We need data and we need a thorough understanding of our data so that we can use them appropriately. We also need a vision for what we want to do with our data. Do we want to confirm a specific hypothesis, or answer a specific question? Perhaps we just want to explore our data to see what new insights can be gained, or new hypotheses generated, from looking at our data from various perspectives. To use our data efficiently, we must be able to define a path forward and develop a plan for achieving our vision.

Successfully developing a plan can be challenging, as it often requires working in interdisciplinary teams. Evaluating and executing the plan can be challenging for the same reason. While there is much to be gained from working in interdisciplinary teams, problems often arise due to conceptualization and communications gaps between team members. Each field tends to have their own way of thinking about and their own vocabulary for talking about their data. These differences can significantly impede progress, but can be overcome by a commitment on the part of all team members to work together and learn enough about each others disciplines to develop a deep understanding of the appropriate use of the data and the accuracy and reproducibility of any results acquired through mining of those data.

Data Collection



Manual entry

Embedded in another process

Automatic collection via an instrument

Improving Efficiency

9/2/2016 Sandra Karcher – Carnegie Mellon University 5

Where do we get data? There are fields in which data are collected and stored in minable formats automatically via some sort of instrumentation. Notice here some examples, my Garmin vivo keeping track of my steps, and a meter installed in a water treatment plant, automatically recording the temperature, pH, and conductivity of the water. Automatic collection of data into a structured database is the best and most efficient way to collect data to be used for mining.

The structured collection of data through manual entry embedded in another process, shown here as placing an order on Amazon, can also be an efficient way of acquiring data for use in mining.

The least efficient method of collecting data for mining, both in terms of efficiency and standardization, is by manual entry, particularly if the format of entry is not constrained or specified. Shown here is the manual entry of a contact into a cell phone. There is some structure to how the data are to be entered, but if you are like me, you enter only the information you need, in the format that works for you, and that might vary from person to person, and also depend on how much time you have to get the contact into the phone. The lack of standardization associated with manual entry makes mining these data challenging.

Established Method of Sharing Data

MATERIALS AND METHODS

Mesocosms. The mesocosms were rectangular shaped boxes constructed of treated wood and were kept outdoors in an open area of the Duke Forest in Durham, NC. Each mesocosm enclosure was 3.66 m long, 1.22 m wide and 0.8 m deep (Supporting Information (SI) Figure S1a). The bed was sloped at ~13 degrees (SI Figure S1b),

AgNPs. The AgNPs used in this study were purchased from NanoAmor (Houston, TX). They were reported by the manufacturer to be 10 nm in diameter with a PVP coating (10 000 g/mol). The physical properties of these particles have been previously described.[35,41] Briefly, the particles were polydisperse with particle sizes ranging from 30 to 80 nm, and with aggregates of up to 200 nm in diameter based on TEM and DLS measurements (z-average hydrodynamic diameter). The initial Ag(0) content as determined by X-ray absorption spectroscopy was 80–85 wt % with the balance

Analytical Method	Experimental Method	Experimental Measurements	Analytical Measurements

- peer reviewed literature
- pdf narrative (good for human reader)
- mix of methods and measurements
- career advancement and continued funding are often tied to the number of articles published and/or cited

- Researchers fear someone could use their data and reduce the novelty of their work (Reichman et al.; "Challenges and Opportunities of Open Data in Ecology"; Science 2011, 331 (6018), 703–705.)
- Those who had no role in generating data, but who mine it sometimes called "research parasites" (Longo et al.; , D. L.; "EDITORIAL - Data Sharing "; New England Journal of Medicine 2016, 374 (3), 276–277.

In academia, data have historically been collected and shared through peer reviewed publication. The narrative format of a journal article may work well for a human reader, but is awful for facilitating efficient scientific discovery through data mining and machine learning. Curation of data from publications into structures that are more computer friendly can be a huge bottleneck.

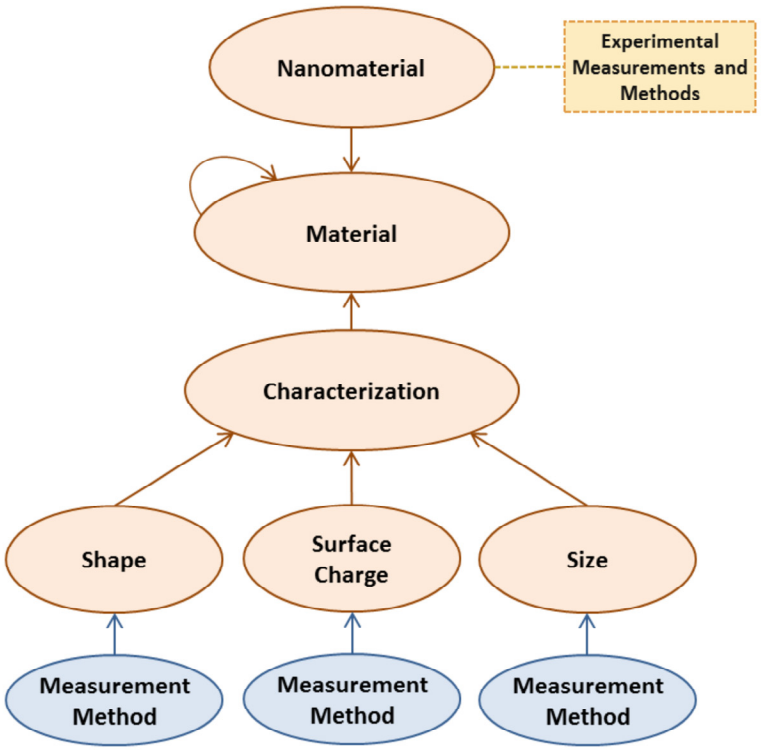
The intermixing of methods and measurements, together, and in with characterization and toxicity information makes it very difficult to mine data formatted in this way. While computer capabilities have increased exponential in the past few decades, the paradigm for sharing data through peer reviewed publication has remain largely unchanged. Career advancement and continued funding opportunities are often tied to the number of articles a researcher has published, and there is little incentive to invest resources in formatting data for broader use. Actually, sharing data can be perceived negatively.

A 2011 article in Science by Reichman et al. indicates that researchers can be reluctant to share their data until they are completely done with it, fearing that someone could take their data and use it in a way that limits or reduces the novelty of their work. In a 2016 article in the New England Journal of Medicine, Longo et al. notes that, those who had no role in generating data, but who mine data generated by others are sometimes referred to as "research parasites".

To move forward with efficient scientific discovery through data mining and machine learning, the engrained culture of publish or parish, which often competes for resources that are needed to prepare data for broader use, will have to transform, recognizing achievements in informatics as indicators of a researcher's value.

Needed to Support Knowledge Discovery

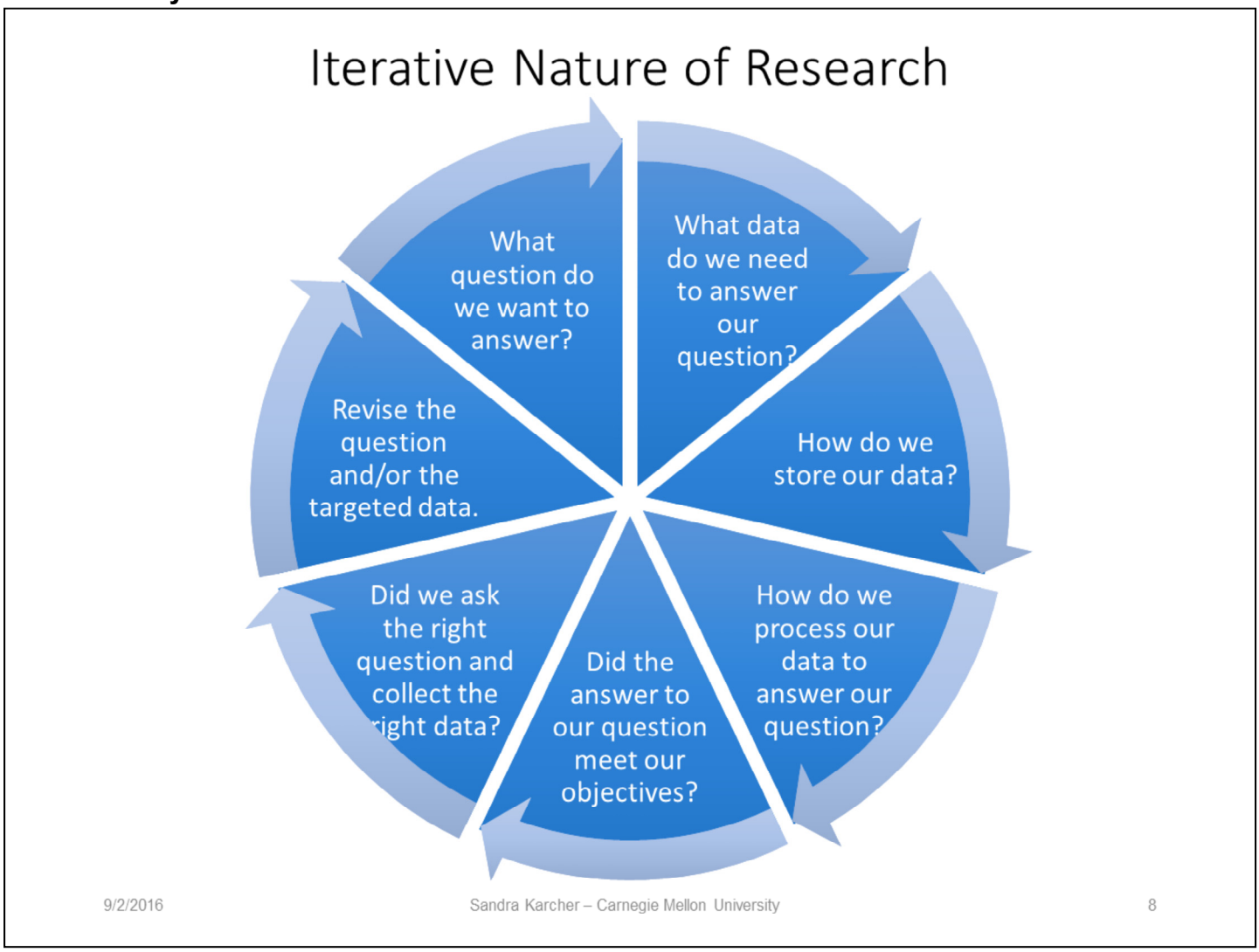
- data repositories
- structured data (good for data mining)
- data of sufficient granularity to support knowledge discovery
- career advancement and continued funding also tied to organizing and formatting data for broader use



Assuming we can change the culture and gain the support needed to embrace standardized data collection and sharing, we need to develop a common model for organizing and storing nanoinformatics data so that data can be shared either through mapping from one repository into another, or through federated integration. The development of a common model for nanoinformatics data is one of the recommendations of the authors of "Integration among Databases and Data Sets to Support Productive Nanotechnology: Challenges and Recommendations", a paper produced by the National Cancer Institute (NCI) Nanotechnology Working Group. This paper is one in a series of papers produced by the Nanomaterials Data Curation Initiative (NDCI).

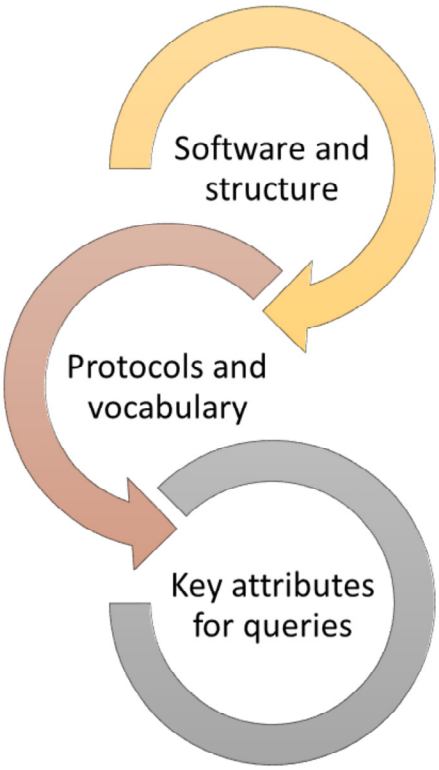
Another paper produced by the group entitled, "How should the completeness and quality of curated nanomaterial data be evaluated?" discusses issues related to data granularity, specifically looking at the definition of quality and completeness.

When considering developing data resources, either warehouses or federated, in addition to taking into consideration the tracking and maintaining of intellectual property, we must also determine who is responsible for appropriate use of those data. Consider the case where a curator extracts information from the published literature and incorporates it into a database. Who is responsible for determining the appropriate use of those data? The data generator, the data curator, or a data miner? Consider that, if a data miner must go back to the publication and read through the mix of methods and measurements to determine the appropriate use of those data, what value was added to the process by the curator? Also consider that, being a data miner requires a different skillset than being a data generator. Unless the data miner has be cross trained, with some depth, in the nuances of the experimental methods, the data miner may lack the appropriate skillset to thoroughly understand the intricacies of the long narrative provided in the publication. There is value in curating information into a database that can be searched as a way of pointing back to a relevant publication, but curating to support identification of relevant literature does not require the same level of metadata collection as does curating in support of mining that will lead to knowledge discovery.



Remember that I told you in the introduction that before I went back to graduate school, I spent several years formatting data for inclusion in the Air Force's environmental data repository. These data were generated as part of site investigations. The methods of sample collection and laboratory analysis were highly standardized and documented. Thus, method metadata could be entered into the database by simply referring to the method (for example, SW6010). Capturing the method metadata to enable data mining of associated results when the methods are evolving adds another layer of complexity. Because research is iterative, and experimentalists are often working to discover the appropriate "standard" method, the design of the data model must be flexible enough to adapt to changing methods while capturing enough metadata to enable mining without requiring a data miner to go back to the peer reviewed literature to work with those data.

Concepts



- Determine the structure for organizing and storing data (selection of software will influence the database structure)
- Develop and/or adopt existing protocols for organizing the data in and populating the database (e.g., architecture, vocabulary)
- Include fields/attributes that can be used to facilitate query development (e.g., sort data, query subsets of data, uniquely identify each measurement and/or observation)

Let's look at three key concepts of storing data. The software selected for storing data will determine the overall structure of the database. For example, software such as MySQL and Microsoft Access store data in columns and rows (a.k.a. fields and records). To enable data mining, the columns and rows need to be designed in a way that allows us to break data down into individual pieces, with each field of a record providing information that, taken all together, fully describes the measured result or observation being stored in that record of the database.

Having established protocols and vocabulary for populating the data structure enables standardization, allowing for consistency in the database to be maintained across time and across users. As a simple example, we could have a protocol that required all states to be entered using their official two letter designation. Without such a protocol, one user might enter "PA", another "penn", and another "Pennsylvania". If we wanted to query all records associated with the state of Pennsylvania, we would need to: (1) know that there were inconsistencies in the database, and (2) develop an algorithm to tell the computer that "PA", "penn", and "Pennsylvania" should be considered the "same".

Key attributes allow us to uniquely identify each piece of data (measurement and/or observation) from all the others in the database and also allow us to select and group data in different ways, depending on what we want to do with those data. Assuming we know the use case for our database (what hypothesis we are testing or what question we are planning to answer), we can include attributes in our structure that allow us to extract relevant data.

At this point, I am sometimes asked how these concepts relate to ISA-TAB-nano. ISA-TAB-nano refers to a set of files in Microsoft Excel, formatted following a specific set of protocols. The content of those files uses some vocabulary from established ontologies. Ontologies contain both vocabulary and relationship information (for example, in the UO ontology - https://bioportal.bioontology.org/ontologies/UO, "hour" is a "time unit", and a "time unit" is a "unit"). Thus, ISA-TAB-nano refers to an Excel based set of files formatted in a specific way using vocabulary and relationships from specific ontologies. ISA-TAB-nano was designed to be a data transfer protocol; while data can be stored in the ISA-TAB-nano format, they cannot be easily queried using SQL.

What is a "Query"?

- "a question or a request for information"
- We use queries to extract data subsets and perform analyses
- When using relational databases, such as MySQL and Microsoft Access, queries can be written using Structured Query Language (SQL)

A query is defined as a question or a request for information.

We use queries to extract subsets of data and to perform calculations.
Queries can be written using Structured Query Language (SQL).
SQL is a programming language designed for managing data in a relational database management system (such as Access, MySQL and ORACLE) – but there are many software options for transforming and querying data.

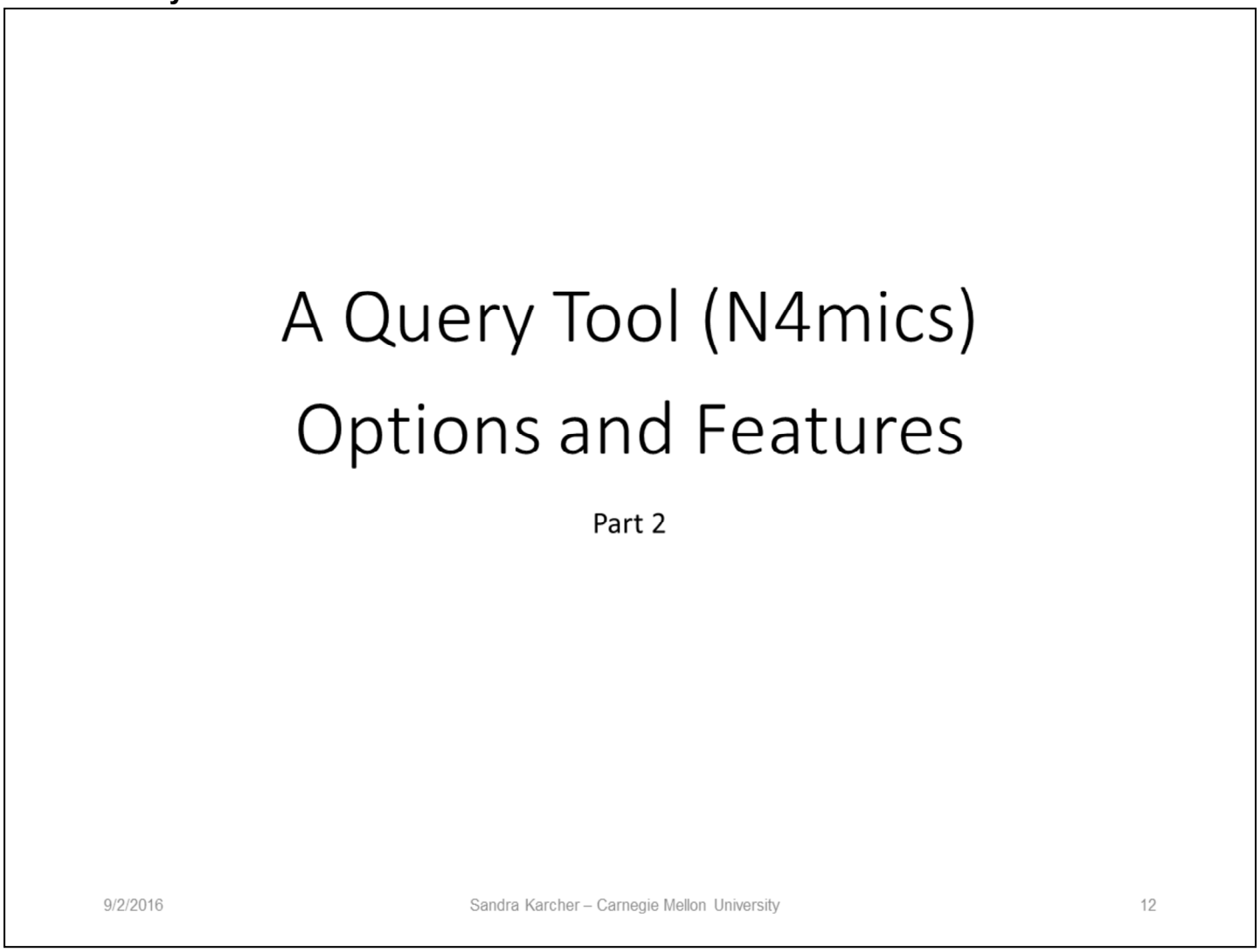
It is important to understand that:
bad data plus good queries leads to bad data analysis and that
good data plus bad queries leads to bad data analysis.

Queries are not a magical fix to problems encountered during data collection. To perform meaningful data analysis, we need high quality data of high enough resolution to enable us to answer our research question, and we need well designed queries to allow us to arrive at the correct answer to our question (and also provide us with some understanding of the error and reproducibility associated with our answer).

End of the Big Picture Background

Part 1

This is the end of part 1 of the nanoinformatics series by Sandra Karcher.



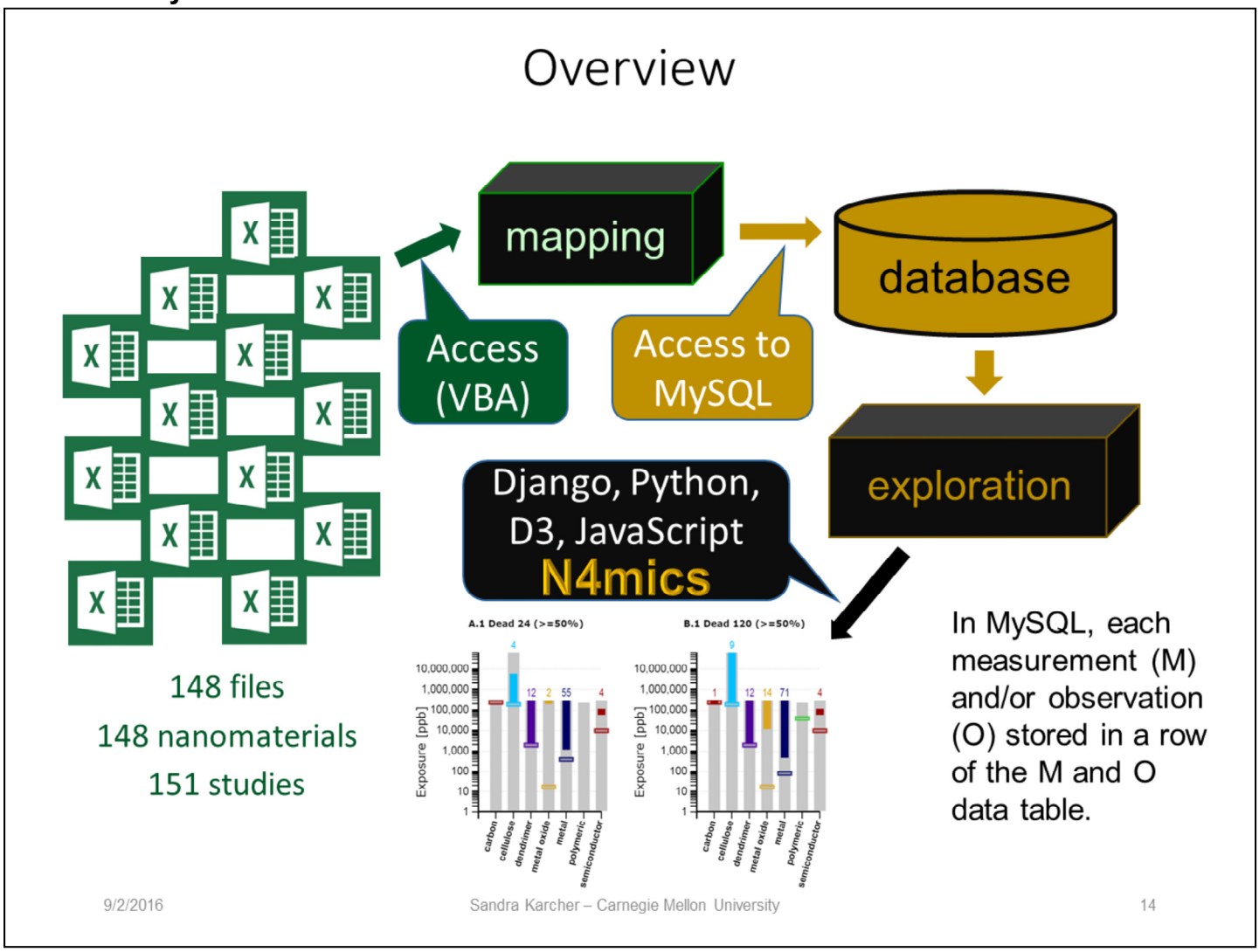
Welcome to part 2 of the 3 part series on nanoinformatics by Sandra Karcher.

When I starting working in nanoinformatics in 2014, I was tasked with designing and developing a nanoinformatics tool that demonstrated value. I approached this task from several angles, and settled in on testing the flexibility of a data structure, designed to hold various kinds of data, by mapping data into the structure from other repositories, and then developing a process for transforming those data for use in a visualization tool.

One of the data sets explored during this testing process was the Nanomaterial-Biological Interactions Knowledgebase - the NBI.

- Goals and Objectives (from http://nbi.oregonstate.edu/goals.php)
  - To serve as a repository for annotated data on the physicochemical properties of nanomaterials and their biological interactions
  - To organize and analyze data and compare results across research platforms in an effort to define robust structure-activity relationships
  - To identify the functional design principles of high performing, environmentally-benign nanomaterials
  - To predict potential biological impacts of unsynthesized nanomaterials
- Experimental Method described by Lisa Truong, Stacey Harper, Robert Tanguay; *Methods in Molecular Biology* (2011) 691:272-279; "Evaluation of embryotoxicity using the zebrafish model"
- NBI Website: http://nbi.oregonstate.edu/

The NBI is a standardized repository of nanomaterial characteristics and associated biological responses from assays performed on zebrafish. Data and information on the NBI is available on their website and the assay methodology is documented in the literature. Data in the NBI is useful for identifying correlations between nanomaterial properties and biological responses. Late in 2015, Stacey Harper and I began collaborating on tool development, and in December, when the first version of the tool was completed, Greg Lowry and Bryan Harper joined in providing feedback and suggesting improvements to the tool. Preliminary findings generated using the tool were presented to the nanomaterial working group in January of 2016.



The tool evolved significantly over time, with new features being added and more components being moved into a web compatible framework.

I have a long history and an existing toolset for mapping Excel spreadsheets into Access, so the mapping portion of the tool was developed using visual basic and run in Access. Once the data were incorporated into my experimental test structure, the database was moved into MySQL.

It may be interesting to note that, in my experimental test structure, each record in the database holds the results and associated metadata for one measurement and/or observation.

The exploration portion of the tool was developed in Django and Python, and uses D3 and JavaScript. Data are prepared for inclusion in the Django framework using MySQL queries. The portion that runs in Django and Python is called N4mics.

Nanomaterial Characteristic Information

NBI Data

NBI Material | NBI-EMZ Exp. Design | NBI-EMZ Data

	A	B
1	**Nanomaterial - Biological Interactions Knowledgebase Data Input/Output Template**	
2		
3		
4		
5		
6	**Field Name**	**Value**
8	NBI Material Identifier	6
9	Particle Descriptor	Gold-TMAT(1.5nm)-pure
11	Investigator / Material Data Contributor: Name	Dr. Stacey Harper
12	Investigator / Material Data Contributor: Affiliation	Oregon State University
13	Investigator / Material Data Contributor: Email	stacey.harper@oregonstate.edu
14	Material Type	metal
15	Manufacture Date	
16	Manufacturer	Jim Hutchison, University of Oregon
17	Synthesis Process	reduction using sodium borohydride in the presence of triphenylphosphine
18	Synthesis Precursors	chloroauric acid (HAuCl4)
19	Purity	pure
20	Types of Impurities	
22	**Primary Particle / Material Core Data:**	
23	Primary Particle Size: Avg. (nm)	1.5
24	Primary Particle Size: Min. (nm)	
25	Primary Particle Size: Max (nm)	
26	Method of Size Measurement	
27	Instrument Used for Size Measurement	
28	Core Shape	spherical
29	Core Structure	solid
30	Crystal Structure	
31	Core Atomic Composition	gold [Au]
32	Number of Core Atoms	101
33	Mass Core Atoms (ng)	19894
35	**Core Shell / Coating (if present):**	
36	Shell Composition	triphenyl phosphine
37	Shell Surface Shape	
38	Shell Linkage	
40	**Surface Linkages / Ligands (if present):**	
41	Outermost Surface Functional Groups	N,N,N-trimethylammoniumethanethiol [TMAT]
42	Surface Chemistry Linkage Group / Type	
43	Density of Surface Covered with Ligands (%)	
44	Minimum Number of Ligands	30
45	Maximum Number of Ligands	35

Characteristics used in N4mics were selected based on the number of spreadsheets in which they were populated.

Let's take a look at the format of the data in the NBI spreadsheets. Information on the nanomaterial characteristics is provided in a material tab.

Characteristics used in N4mics were selected based on the number of spreadsheets in which they were populated.

Characteristics of a Nanomaterial

Characteristic	# populated	Description
Surface Charge	72	Nulls were replaced with "unknown".
Primary Particle Size: Avg. (nm)	117	When the primary particle was null it was left null, except, in the case of four files where a primary size was not provided, but a maximum value was; in those cases, the maximum was used as the primary particle size.
Outermost Surface Functional Groups	89	Nulls were replaced with "none".
Core Atomic Composition	147	Information from the particle descriptor was used to populate the missing value.
Material Type	147	Information from the particle descriptor was used to populate the missing value.
Shell Composition	48	Nulls were replaced with "none".
Purity	107	Nulls were replaced with "unknown".
Core Shape	108	Nulls were replaced with "unknown".
Core Structure	69	Nulls were replaced with "unknown".

This table provides information on the number of spreadsheets in which a specific characteristic was populated and also indicates how missing values were handled in preparing data for use in the N4mics tool. Decisions on how to handle missing data were made in collaboration with representatives from the NBI.

Experimental Design Information

NBI Data

NBI Material | **NBI-EMZ Exp. Design** | NBI-EMZ Data

	A	B
1	**Nanomaterial - Biological Interactions Knowledgebase Data Input/Output Tem**	
2		
3		
4		
5		
6	**Field Name**	**Value**
8	NBI Experiment ID	6
9	NBI Material Identifier	6
10	Exposure Metric / Assay	
11	Primary Exposure Route	dermal
12	Primary Exposure Delivery	waterborne
13	Secondary Exposure Route	oral
14	Secondary Exposure Delivery	waterborne
15	Tertiary Exposure Route	
16	Tertiary Exposure Delivery	
17	Study Factor(s)	
18	Exposure Organism	zebrafish
19	Exposure Organism Life stage	embryo
20	Exposure Length Classification	
21	Duration of Exposure (hours)	112
23	Investigation Title	
24	Investigation Description	
25	Investigator Name	
26	Investigator Affiliation	
27	Investigator Email	
29	Study Start Date	
30	Exposure Organism Gender	mixed
31	Exposure Organism Average Weight (mg)	1 mg
32	Exposure Organism Initial Age (hours post-fertilizati	8
33	Continuity of Exposure	Continuous
34	Exposure Temperature (Celsius)	
35	Exposure Media	fish water
36	Media Composition	
37	Media pH	
38	Material Zeta Potential in Media (mV)	
39	Stable Average Agglomerate Size in Media (nm)	
40	Stable Agglomerate Size in Media Minimum (nm)	
41	Stable Agglomerate Size in Media Maximum (nm)	
42	Nanomaterial Preparation	
44	Experimental Notes	
45	LC50 (ppm)	
46	NOAEL (ppm)	
47	Weighted EZ Metric EC50 (ppm)	
48	Additive EZ Metric EC50 (ppm)	

At the time N4mics was developed, the experimental design of the assays reported in the NBI Knowledgebase were intended to be consistent across all studies.

Information on the experimental design was provided in a design tab.

At the time N4mics was developed, the experimental design of the assays reported in the NBI Knowledgebase were intended to be consistent across all studies. Thus, the design information was not queried for use in the N4mics tool.

Biological Response Data (NBI_6)

M – mortality
J – jaw
E – eye
etc.

NBI Material | NBI-EMZ Exp. Design | **NBI-EMZ Data**

NBI Data

Dosage Concentrations Used (ppm)	Weighted EZ Metric Score	Additive EZ Metric Score	24 hpf evaluation								120 hpf evaluation →											
			M		DP		SM		N		M		Y		A		E		Sn		J	
			yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no
0	0.00	0.00	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24
0.016	0.02	0.17	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24	0	24
0.08	0.26	0.42	3	21	0	21	0	21	0	21	3	18	0	18	0	18	0	18	0	18	1	17
0.4	0.31	0.63	3	21	0	21	0	21	0	21	4	17	0	17	0	17	0	17	0	17	3	14
2	0.51	1.33	5	19	0	19	0	19	0	19	6	13	0	13	0	13	0	13	0	13	8	5
10	0.72	1.54	7	17	0	17	0	17	0	17	9	8	0	8	0	8	0	8	0	8	6	2
50	0.84	1.25	11	13	0	13	0	13	0	13	9	4	0	4	0	4	0	4	0	4	3	1
250	1.00	1.00	24	0																		

Exposure [ppm]	# Fish Added	# Dead Fish at 24 hpf	% Dead to **Total** at 24 hfp	# Living Fish at 120 hpf	# Dead Fish at 120 hpf	% Dead to **Total** at 120 hpf	# Abnormal Jaw at 120 hpf	# Dead plus Abnormal Jaw at 120 hpf	% Harmed Jaw to **Living** at 120 hpf	% Harmed Jaw or Dead to **Total** at 120 hpf
0	24	0	0.0	24	0	0.0	0	0	0.0	0.0
0.016	24	0	0.0	24	0	0.0	0	0	0.0	0.0
0.08	24	3	12.5	18	6	25.0	1	7	5.6	29.2
0.4	24	3	12.5	17	7	29.2	3	10	17.6	41.7
2	24	5	20.8	13	11	45.8	8	19	61.5	79.2
10	24	7	29.2	8	16	66.7	6	22	75.0	91.7
50	24	11	45.8	4	20	83.3	3	23	75.0	95.8
250	24	24	100.0	0	24	100.0	0	24	0.0	100.0
			A			B			C	D

9/2/2016 Sandra Karcher – Carnegie Mellon University 18

The biological response results are stored in a data tab (which is shown in the top half of the slide).
The bottom portion of the slide provides an example of how data provided in the Excel data tab are prepared for use in the exploration tool.

The counts of the number of fish displaying and/or not displaying an abnormal response are extracted from the spreadsheet and normalized into a percent response. Notice the columns labeled at the bottom of the slide as A, B, C, or D. These columns contain normalized responses and it is the information in these columns that is fed into the N4mics tool.

Column A indicates the percentage of fish that were observed to be dead at the 24 hours post fertilization observation. Column B is the cumulative percent dead at the 120 hour post fertilization observation. Notice the number of dead fish were normalized by the number of fish observed in the study at each concentration of exposure. Column C indicates the percentage of fish that were found to have an abnormal jaw when the jaw was observed at 120 hours post fertilization. The count of fish with an abnormal jaw is normalized by the number of living fish at the time of observation. Column D is computed by adding the number of fish dead at 120 hours to the number of fish observed to have an abnormal jaw, then divided by the total number of fish observed and multiplied by 100. This method of normalization provides a way of looking at the sublethal and lethal responses together.

The table shown on the bottom of the slide is prepared in Excel and is just for demonstration purposes. Data used in the N4mics tool are prepared using SQL queries.

Mortality Data - Prepared

Response	Exposure [ppb]	Nano-material	A % Dead to Total @24	B % Dead to Total @120	Fisher's % Dead to Total @24	Fisher's % Dead to Total @120	Minimum [ppb] to reach Fisher's % Dead to Total @24	Minimum [ppb] to reach Fisher's % Dead to Total @120	Value of Characteristic descriptor of nanoparticles (unitless)
mortality	0	NBI_6	0.0	0.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	16	NBI_6	0.0	0.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	80	NBI_6	12.5	25.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	400	NBI_6	12.5	29.2	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	2000	NBI_6	20.8	45.8	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	10000	NBI_6	29.2	66.7	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	50000	NBI_6	45.8	83.3	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	250000	NBI_6	100.0	100.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
mortality	0	NBI_78	8.3	8.3	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	16	NBI_78	25.0	25.0	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	80	NBI_78	12.5	12.5	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	400	NBI_78	20.8	20.8	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	2000	NBI_78	8.3	8.3	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	10000	NBI_78	8.3	8.3	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	50000	NBI_78	12.5	12.5	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	250000	NBI_78	12.5	12.5	33.3	33.3	-	-	Sulfonated Nanocrystaline Cellulose
mortality	0	NBI_159	4.2	8.3	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	48	NBI_159	0.0	0.0	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	240	NBI_159	4.2	8.3	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	1000	NBI_159	8.3	8.3	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	6000	NBI_159	100.0	100.0	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	30000	NBI_159	100.0	100.0	25.0	33.3	6000	6000	Gold Nanorods (10x34nm) #79-6000
mortality	0	NBI_167	8.3	8.3	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	16	NBI_167	2.8	2.8	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	80	NBI_167	8.3	8.3	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	400	NBI_167	8.3	13.9	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	2000	NBI_167	8.3	8.3	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	10000	NBI_167	2.8	5.6	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	50000	NBI_167	0.0	13.9	27.8	27.8	-	250000	Samarium Oxide - Sonicated
mortality	250000	NBI_167	8.3	97.2	27.8	27.8	-	250000	Samarium Oxide - Sonicated

When the data are prepared for use in the exploration tool, they actually look more like this. This table shows a subset of mortality data for four nanomaterial assays.

As shown previously, column A indicates the percentage of fish that were observed to be dead at the 24 hour post fertilization observation and column B shows the cumulative percent dead at the 120 hour post fertilization observation.

Sublethal Response Data (jaw) - Prepared

Response	Exposure [ppb]	Nano-material	C: % Abn. Jaw to Living when observed	D: % (Abn. Jaw or Dead) to Total	Fisher's % Abn. Jaw to Total	Fisher's % (Abn. Jaw or Dead) to Total	Minimum [ppb] to reach Fisher's % Abn. Jaw to Living	Minimum [ppb] to reach Fisher's % (Abn. Jaw or Dead) to Total	Value of Characteristic
jaw	0	NBI_6	0.0	0.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	16	NBI_6	0.0	0.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	80	NBI_6	5.6	29.2	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	400	NBI_6	17.6	41.7	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	2000	NBI_6	61.5	79.2	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	10000	NBI_6	75.0	91.7	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	50000	NBI_6	75.0	95.8	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	250000	NBI_6	0.0	100.0	20.8	20.8	2000	80	Gold-TMAT(1.5nm)-pure
jaw	0	NBI_78	0.0	8.3	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	16	NBI_78	0.0	25.0	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	80	NBI_78	0.0	12.5	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	400	NBI_78	10.5	29.2	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	2000	NBI_78	0.0	8.3	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	10000	NBI_78	0.0	8.3	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	50000	NBI_78	0.0	12.5	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	250000	NBI_78	0.0	12.5	20.8	33.3	-	-	Sulfonated Nanocrystaline Cellulose
jaw	0	NBI_159	0.0	8.3	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	48	NBI_159	0.0	0.0	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	240	NBI_159	0.0	8.3	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	1000	NBI_159	0.0	8.3	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	6000	NBI_159	0.0	100.0	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	30000	NBI_159	0.0	100.0	20.8	33.3	-	6000	Gold Nanorods (10x34nm) #79-6000
jaw	0	NBI_167	0.0	8.3	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	16	NBI_167	0.0	2.8	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	80	NBI_167	0.0	8.3	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	400	NBI_167	0.0	13.9	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	2000	NBI_167	3.0	11.1	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	10000	NBI_167	2.9	8.3	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	50000	NBI_167	9.7	22.2	13.9	27.8	250000	250000	Samarium Oxide - Sonicated
jaw	250000	NBI_167	100.0	100.0	13.9	27.8	250000	250000	Samarium Oxide - Sonicated

9/2/2016 Sandra Karcher – Carnegie Mellon University 20

This table shows the sublethal jaw response for four nanomaterial assays.

As indicated previously, column C indicates the percentage of fish that were found to have an abnormal jaw and column D indicates the percentage of fish that were found to have a jaw abnormality or that were dead at the time of observation.

There are some nuances with regard to these data tables. Not all the fields in the tables are shown here; some have been hidden from view to simplify the presentation of the normalized data. Consider the case where more than one assay was performed using the same nanomaterial. How would we distinguish the results of one assay from the others while still maintaining the ability to associate the biological responses of multiple assays to the same nanomaterial? There are fields in the underlying tables that allow each assay to be distinguished from all the others and that allow multiple assays to be linked to the same nanomaterial, but explaining those fields would require going into details that are beyond the scope of this presentation.

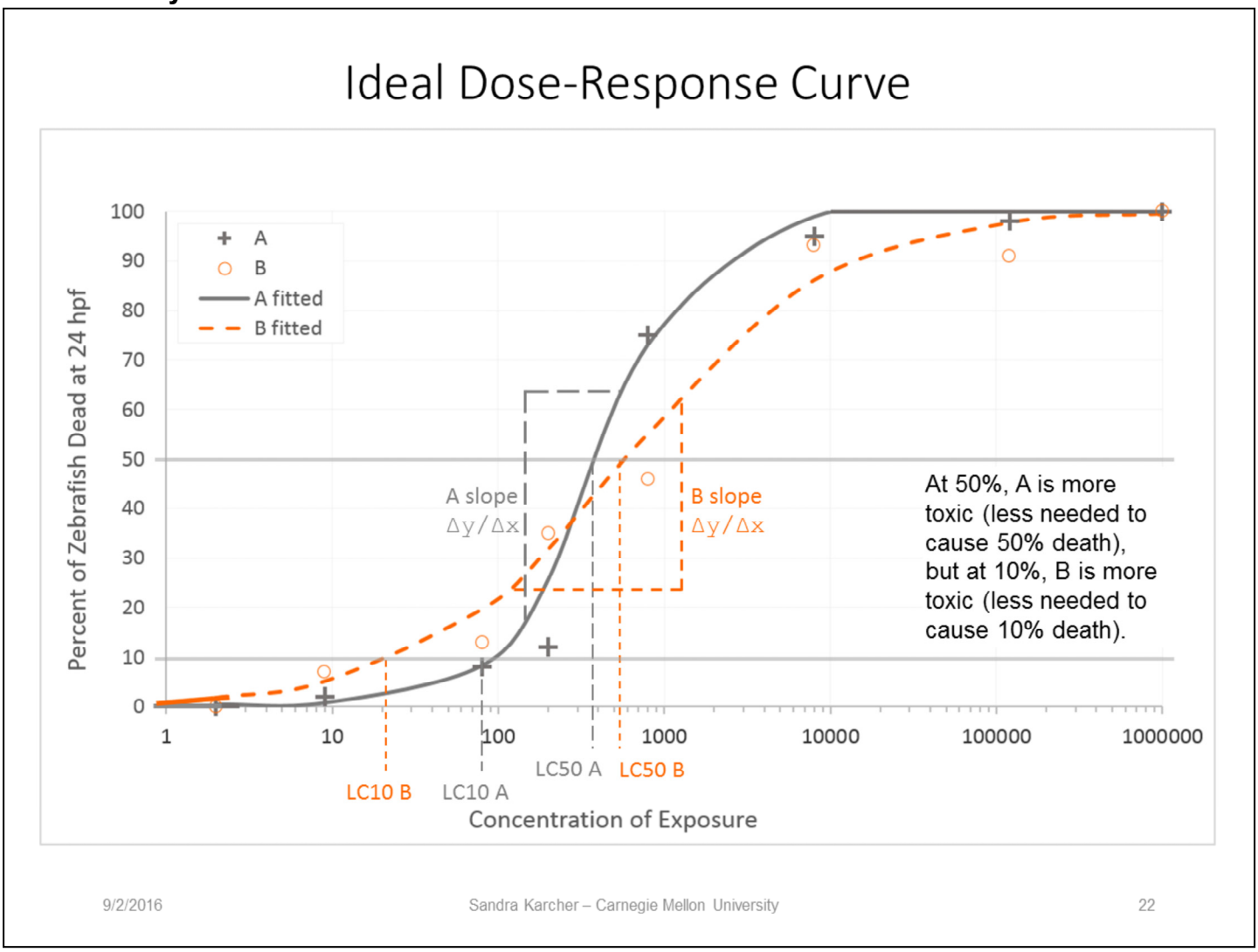
Level of Significance

- Fisher's exact test
  - used to determine the number of dead or abnormal fish that must be observed in a system where the fish have been exposed to a nanomaterial to be considered significantly greater than the number of dead or abnormal fish observed in a similar system where the fish were not exposed to the nanomaterial
  - use a p-value for significance level (p=0.05 with one tail)
  - the number of fish is normalized by the number of fish introduced (at each concentration of exposure)
  - concentration that first reached the Fisher's exact test percentage is identified in data preparation

Fisher's % Abn. Jaw to Total	Fisher's % (Abn. Jaw or Dead) to Total	Minimum [ppb] to reach Fisher's % Abn. Jaw to Living	Minimum [ppb] to reach Fisher's % (Abn. Jaw or Dead) to Total
20.8	20.8	2000	80
20.8	20.8	2000	80
20.8	20.8	2000	80
20.8	20.8	2000	80
20.8	20.8	2000	80

9/2/2016 21

You may have noticed on the previous slides some columns of information relating to the Fisher's exact text. The Fisher's exact test was used to determine the number of dead or abnormal fish that must be observed in a system where the fish have been exposed to a nanomaterial to be considered significantly greater than the number of dead or abnormal fish observed in a similar system where the fish were not exposed to the nanomaterial. In N4mics, a p-value of 0.05 was used to find the number of fish that must be observed in an exposed systems to be considered significant. These numbers were normalized into a percentage, and the concentration where that percent was first observed pulled from the database. This analysis was performed as part of the data preparation process.



The NBI stores the results of dose-response studies. In dose-response studies, we are usually looking for things like, the lethal concentration at a targeted response level, for example the LC50, which is the concentration where 50 percent of the fish are observed to be dead.

While we can learn a lot from looking at the LC50, there is information that could be very important, that we would miss. For example,

At the 50 percent response level, A appears more toxic (meaning less is needed to cause 50 percent of the fish to die) than B, but at the 10 percent response level, B looks more toxic than A.

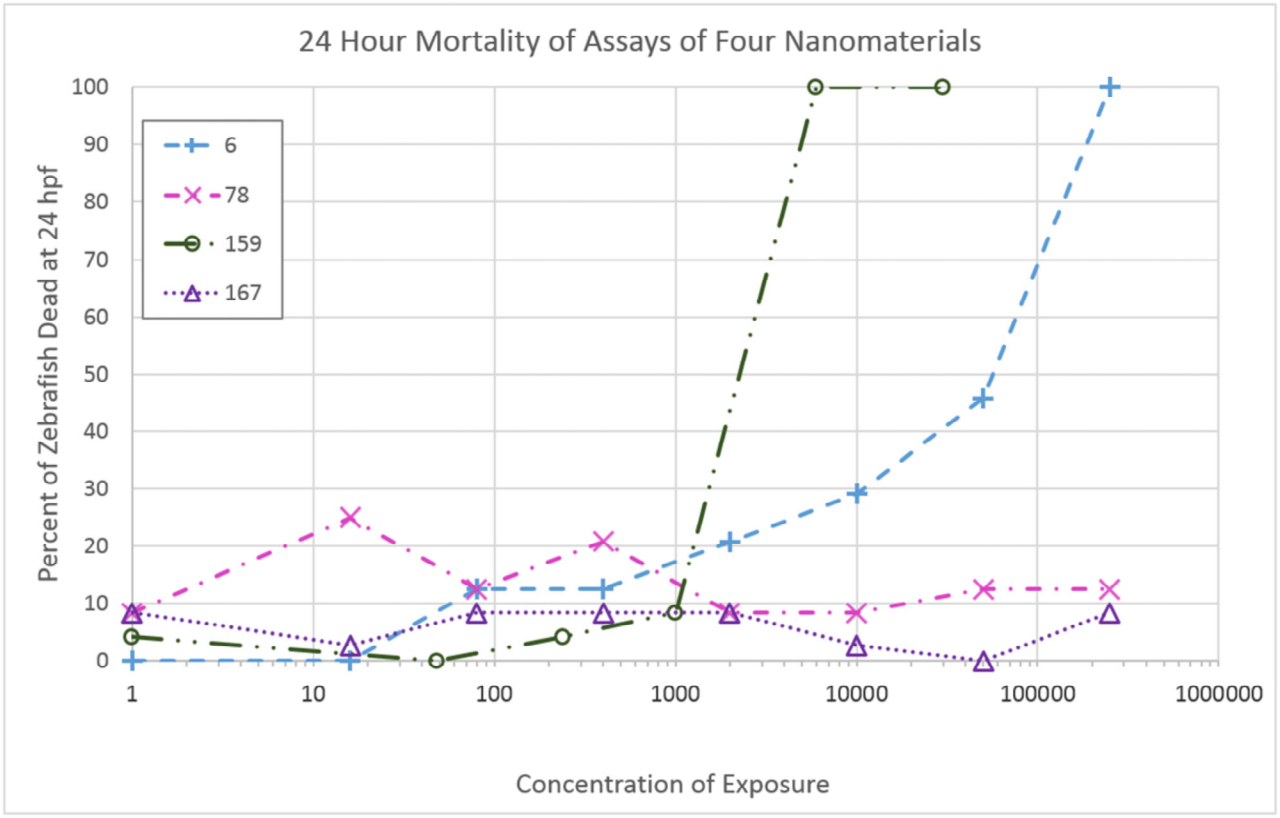
Perhaps we might want to calculate the LC10, which is the concentration where 10 percent of the fish were observed to be dead

and maybe also calculate the slope of the linear portion of the curve.

The great thing about the NBI is that the data are stored at the level of granularity of the counts of zebrafish that were observed to display a specific biological response at each concentration of exposure. So, in theory, we can use these data to determine the lethal concentration at any threshold percentage we want.

But, it is important to understand that the dose-response curves for the NBI data rarely look like the idealized ones displayed here.

Real Data



Here we see real data from the NBI. This graphs shows the normalized mortality responses at the 24 hour post fertilization observation for four nanomaterials. Notice that these do not fit well to a sigmoidal curve. Would an LC50 for these nanomaterials be informative? If so, how should it be calculated?

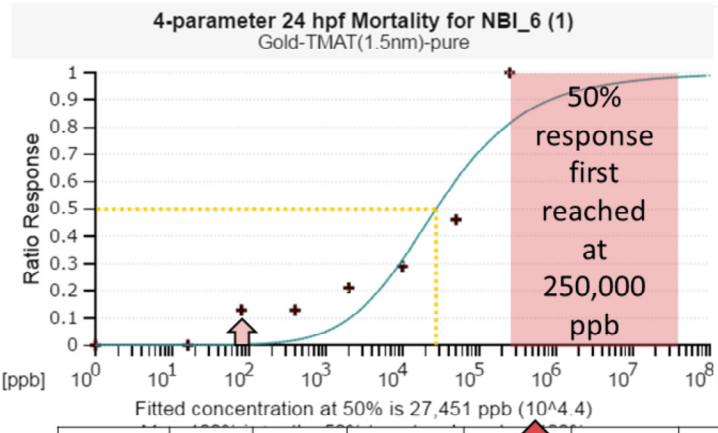
Things to Know about the NBI Data

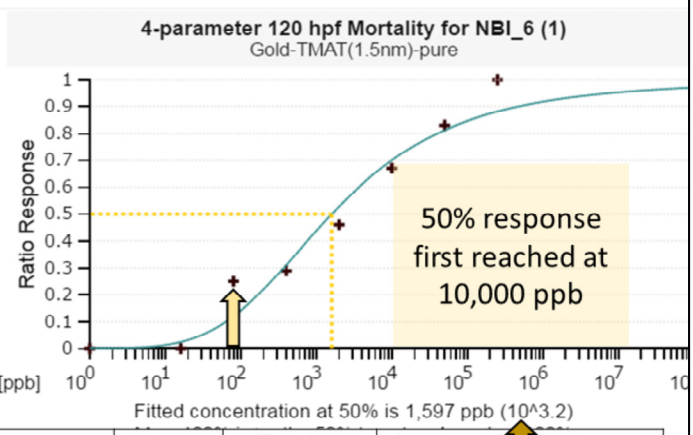
- Of the 148 nanomaterials reported in the NBI as of December 2015:
  - 36 were dosed at high enough concentrations to result in greater than a 99 percent mortality response
  - 54 were dosed at high enough concentrations to result in greater than a 50 percent mortality response
- There are limitations to making some nanomaterials at high enough concentrations to achieve 100 percent mortality.

Before we answer those questions, there are some things that should be considered. Of the nanomaterials in the NBI data set at the time the tool was developed, less then 25 percent were dosed at high enough concentrations to kill all the zebrafish, and 50 percent of the zebrafish were killed in less than half the assays.

There are limitations to making some of the nanomaterials at high enough concentrations to kill all the zebrafish, thus, we need to consider methods of comparing nanomaterial toxicity beyond the LC50.

Real Data





Exposure [ppm]	# Fish Added	# Dead Fish at 24 hpf	% Dead to **Total** at 24 hfp	# Living at 120 hpf	# Dead Fish at 120 hpf	% Dead to **Total** at 120 hpf	# Abnormal Jaw at 120 hpf	# Dead plus Abnormal Jaw at 120 hpf	% Harmed Jaw to **Living** at 120 hpf	% Harmed Jaw or Dead to **Total** at 120 hpf
0	24	0	0.0	24	0	0.0	0	0	0.0	0.0
0.016	24	0	0.0	24	0	0.0	0	0	0.0	0.0
0.08	24	3	12.5	18	6	25.0	1	7	5.6	29.2
0.4	24	3	12.5	17	7	29.2	3	10	17.6	41.7
2	24	5	20.8	13	11	45.8	8	19	61.5	79.2
10	24	7	29.2	8	16	66.7	6	22	75.0	91.7
50	24	11	45.8	4	20	83.3	3	23	75.0	95.8
250	24	24	100.0	0	24	100.0	0	24	0.0	100.0

9/2/2016 Sandra Karcher – Carnegie Mellon University 25

Even though, for many of the assays reported in the NBI, extrapolation is needed to generate a dose-response curve, there is a feature in N4mics that allows the user to perform 4 or 5 parameter logistic regression, using Python's least_squares function, to fit a curve to the points and return a concentration of exposure where a user selected target response percentage would be met. The shape of the curve will depend on the initial, lower, and upper bounds used in fitting the curve.

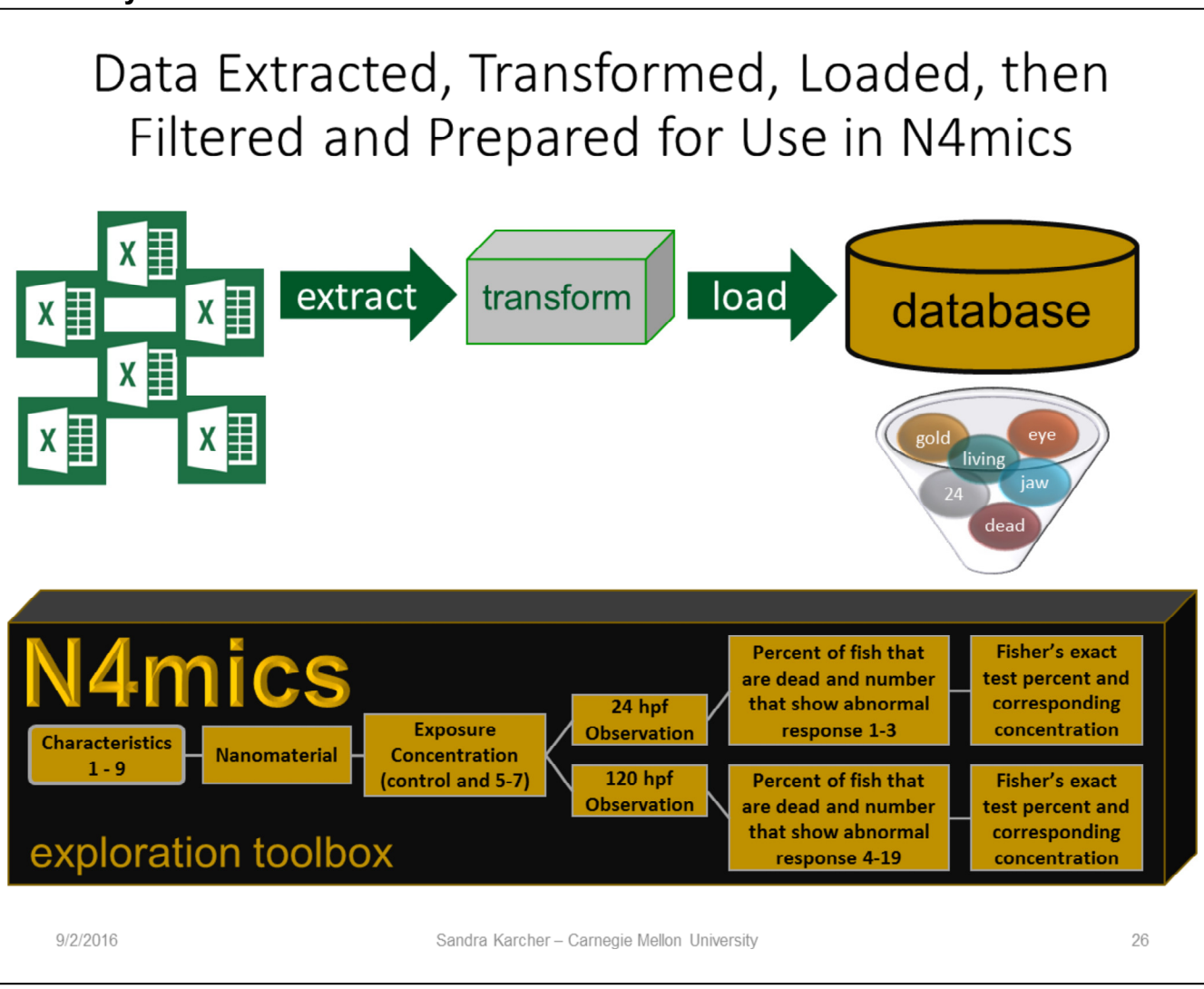
Shown on the slide are the results of curve fitting for the 24 and 120 hour mortality data for nanomaterial NBI_6. In this instance, the bounds were set to try to get a sigmoidal shape to the curve. The tool tells us that the user selected a target percent of 50, and that the fitted concentration where that target threshold was met was at a concentration of 10^4.4 ppb for the 24 hour data, and at a concentration of 10^3.2 ppb for the 120 hour data.

Since the raw data (fish counts) are available in the NBI data set, we have the option of looking at these data in a variety of ways.

Perhaps we might like to know the actual concentration of exposure where the target threshold was first reached. For example, for 24 hour mortality, a 50 percent target response was first reached at 250,000 parts per billion [ppb], and for the 120 hour mortality, at 10,000 ppb.

Or what if we were interested in the concentration where a 10 percent threshold was first reached. We see from the table that the concentration would be the same for the 24 and 120 hour mortality, 80 ppb.

N4mics will allow data exploration in a variety of ways, including returning concentrations where a user selected target threshold was first reached.



The process of preparing data for use in the exploration tool is shown in the figure on the slide. In summary, data are extracted from Excel spreadsheets, transformed and loaded into a MySQL database. Those data are filtered and formatted for use by the exploration tool using SQL queries and the resulting output is appended into the Django exploration toolbox. The portion of the process included in the Django framework is referred to as N4mics.

Features of N4mics

- Allows us to look across studies to find relationships between nanomaterial characteristics and biological responses
- Allows us to look at biological responses collectively, individually, or in user specified groups.
- N4mics has three primary options of visualizing toxicological results:
  1) the minimum concentration and/or range of concentrations where a user specified percent response was met or exceeded
  2) the maximum response by concentration of exposure
  3) the maximum response by individual sublethal response at concentrations at or below a user specified threshold
- When data are aggregated by nanomaterial, two bonus features are provided:
  1) 4 or 5 parameter logistic regression curve fitting
  2) matrix of weighted similarity scores for selected nanomaterials based on characteristics of the nanomaterials

N4mics allows the user to explore the NBI, looking for relationships between nanomaterial characteristics and biological responses. Based on the user's selections, biological responses can be examined individually, or collectively. N4mics offers three options for visualizing toxicological responses. We just looked an example of the first option where N4mics returns the concentration where a user specified percentage was reached. N4mics also provides options for visualizing the maximum response by concentration of exposure and the maximum response by individual sublethal response at concentrations at or below a user specified threshold concentration.

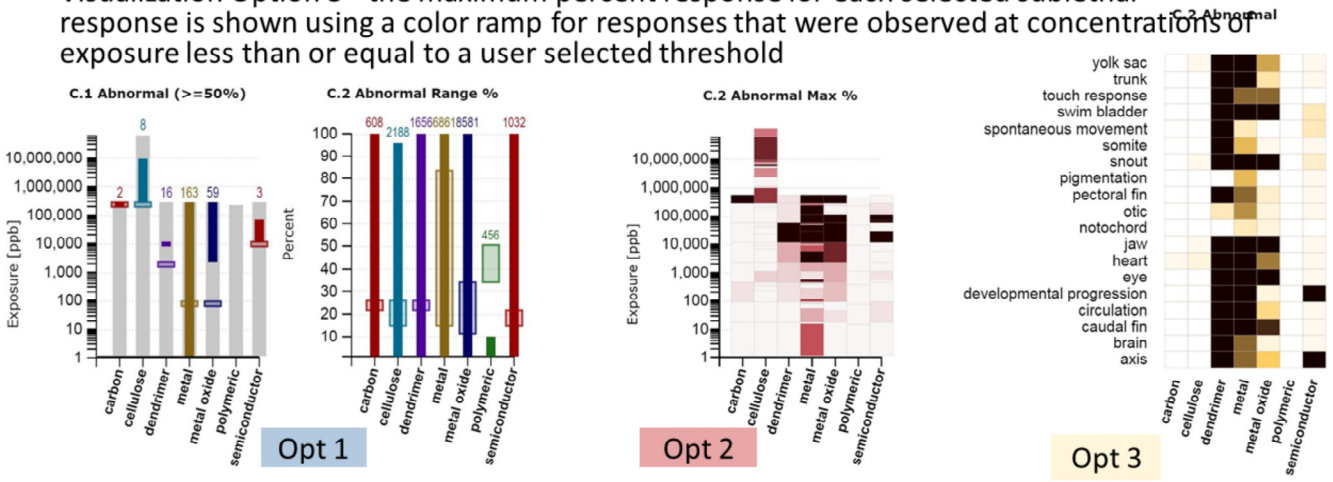
An analysis was conducted using a 50 percent target response rate and a threshold concentration of 100,000 ppb, and the results of that analysis are published in the paper that accompanied the release of the tool (the paper has just been accepted for publication in Environmental Science: Nano at the time this presentation was prepared).

When an N4mics user selects to group data by nanomaterial, N4mics offers two bonus features, the first is an option to fit a curve to the 24 and 120 hour mortality data. The second is a feature that allows the user to enter weights for each characteristic, and then, based on those weights, calculates a similarity matrix for selected nanomaterials.

Let's take a closer look at the graphs generated using the tool.

Overview of Primary Visualization Options

- All the primary visualizations display distinct labels across the ***x-axis*** based on the user selected grouping method (by nanomaterial, response, or characteristic(s))
- The tool aggregates data for display on the ***y-axis*** based on the selected option
  - Visualization Option 1 - two sets of four graphs are provided; a set that shows the concentration of exposure where selected responses reached a target percentage and a corresponding set that shows the full range of responses
  - Visualization Option 2 - the maximum percent response for each concentration of exposure interval is shown using a color ramp
  - Visualization Option 3 - the maximum percent response for each selected sublethal response is shown using a color ramp for responses that were observed at concentrations of exposure less than or equal to a user selected threshold



Sandra Karcher – Carnegie Mellon University

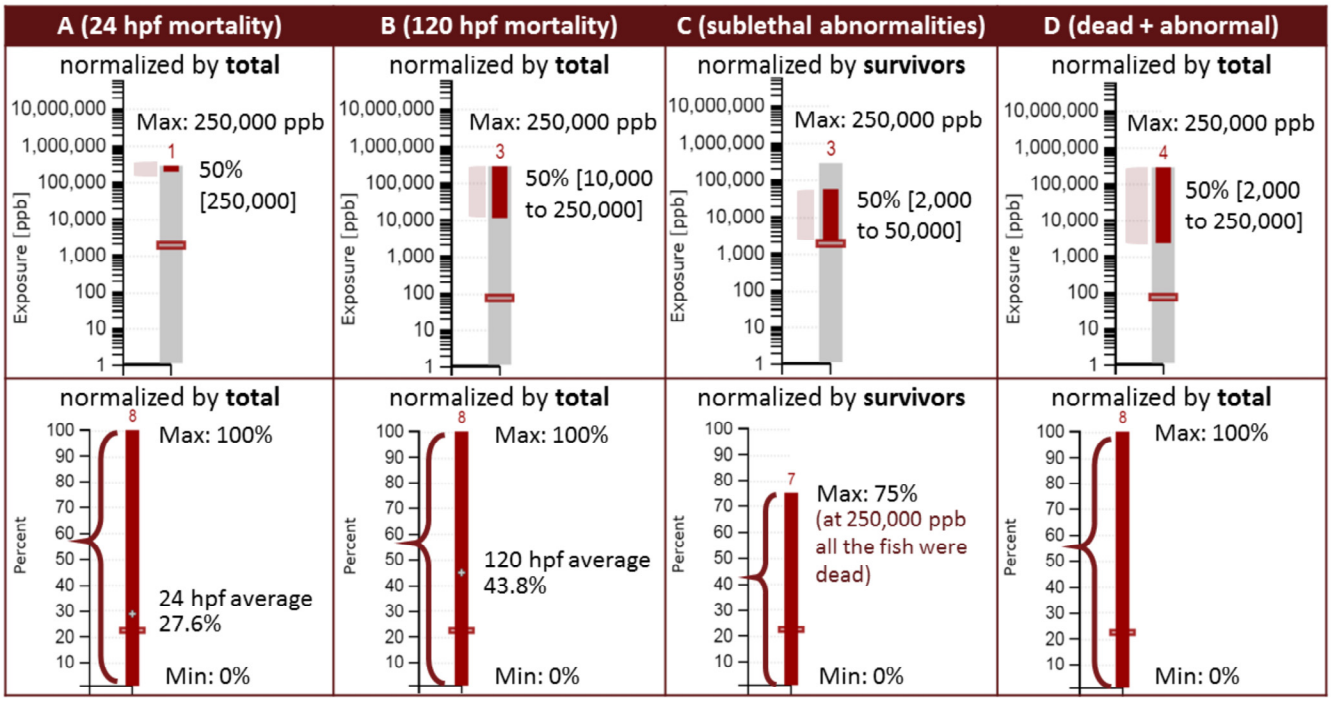
The N4mics user interface guides the user through selecting the method to be used to group data – by nanomaterial, by response, or by characteristic, and also through the process of selecting a target response rate and a threshold concentration. All the graphs generated using the three primary visualization options will display distinct labels across the x-axis based on the selected grouping method. The way data are aggregated on the y-axis depends on the visualization option.

Option 1 generates pairs of graphs for each normalization method (A, B, C, and D). One of the graphs shows the concentration range where the user selected target percentage criterion was met and the other shows the full range of responses.

Option 2 generates a heat map for each normalization method showing the maximum response observed at each concentration interval.

Option 3 also generates a heat map, but this graph shows the maximum of each selected sublethal response that occurred at or below the user selected threshold concentration.

[ppb]	% Dead 24 hfp	% Dead 120 hpf	% Jaw at 120 hpf	% Dead or Jaw at 120 hpf
0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0
80	12.5	25.0	5.6	29.2
400	12.5	29.2	17.6	41.7
2,000	20.8	45.8	61.5	79.2
10,000	29.2	66.7	75.0	91.7
50,000	45.8	83.3	75.0	95.8
250,000	100.0	100.0	0.0	100.0



Let's look at a complete set of graphs generated using primary visualization option 1. As indicated previously, a pair of graphs showing the results for each normalization method (A, B, C, and D) are generated. The top set of graphs on the slide shows the concentration of exposure on the y-axis, and the bottom set shows the percent response on the y-axis.

The exposure graphs (top set) show concentrations where the target response was met or exceeded. For example, a 50 percent response was met or exceeded when observing 24 hour mortality at a concentration of exposure of 250,000 ppb. When looking at 120 hour mortality, a 50 percent response was met or exceeded over a range of concentrations, from 10,000 to 250,000 ppb. Similar information is shown in graphs C and D. The range of the concentration of exposure used in the assay of this nanomaterial is shown on the exposure graphs (top set) with the light grey bar (for this assay, from 0 to 250,000 ppb).

The narrow, hollow band that extends beyond the sides of the solid grey bar, indicates the concentration where the required Fisher's exact test percentage was first met. Recall that the Fisher's exact test was used to determine the number of fish, dead or abnormal, that must be observed in the case where fish were exposed to a nanomaterial to be considered statistically significant.

The percent graphs (bottom set) indicate the percent response over the whole range of exposure.

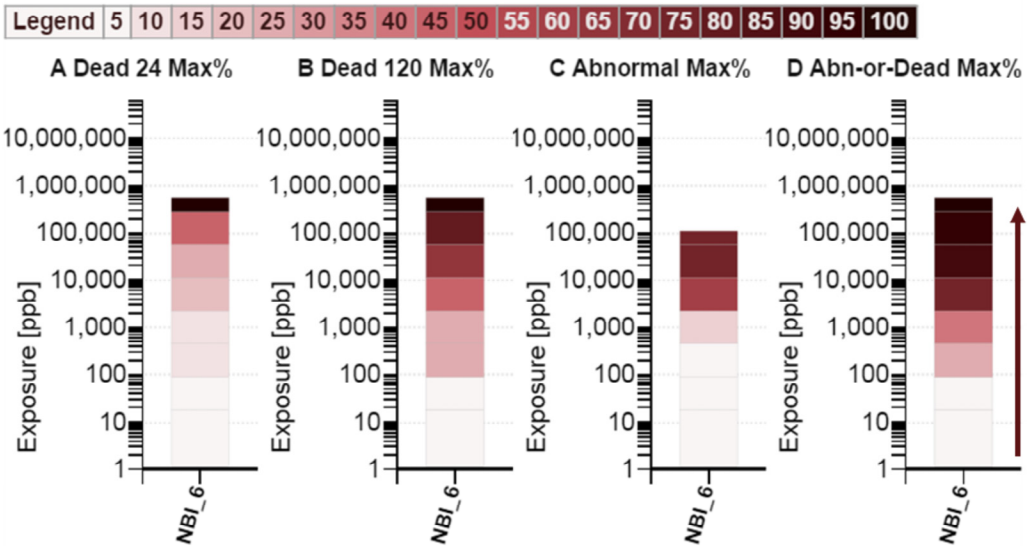
These graphs also show the Fisher's exact test percentage.

The A and B percent graphs also show an average percent response. These values are calculated by taking the average of all the normalized results in column A and column B (as shown in the table at the top of the slide), respectively.

Primary Visualization Option 2

Graphs A - Dead @24, B - Dead @120, C - Abnormal, D - Abnormal or Dead
Graphs A, B, C, and D: Maximum percent response at the indicated exposure concentration interval
Nanomaterials: NBI_6; Responses: jaw, mortality; Characteristic: nanomaterial added; (some nanomaterials were individually selected/excluded)

(in this case, only one nanomaterial is shown in the group, thus, it is a "max" of a group containing one value)



[ppb]	A	B	C	D
250,000	100	100	0	100
50,000	46	83	75	96
10,000	29	67	75	92
2,000	21	46	62	79
400	13	29	18	42
80	13	25	6	29
16	0	0	0	0
0	0	0	0	0

- Nanomaterial NBI_6 showing mortality and jaw abnormalities

9/2/2016 Sandra Karcher – Carnegie Mellon University 30

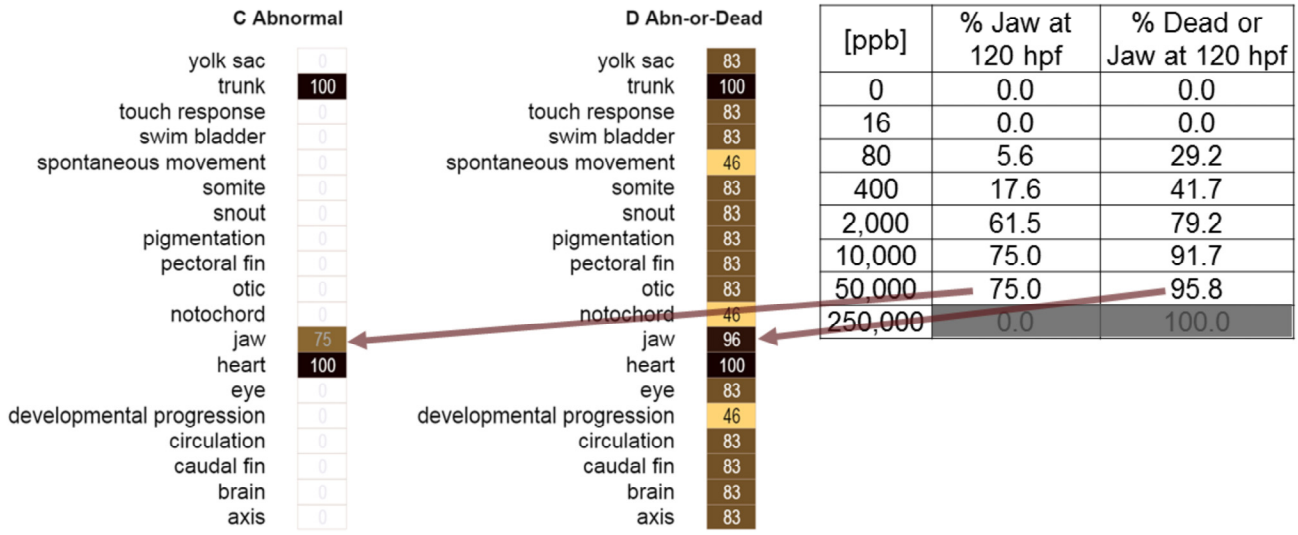
When using visualization option 2, information presented in the table shown is displayed using a heat map. Notice that the concentration of exposure has been displayed in the table in descending order to make it easier to see how the coloration of the heat map is applied.

Primary Visualization Option 3

Graphs C and D: Maximum percent response at exposure concentrations less than or equal 100000 ppb
Nanomaterials: NBI_6; Responses: all; Characteristic: descriptor of nanoparticles (unitless); descriptor: Gold-TMAT(1.5nm)-pure;

C Toggle labels | D Toggle labels

Legend 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 Target concentration: 100000 ppb



[ppb]	% Jaw at 120 hpf	% Dead or Jaw at 120 hpf
0	0.0	0.0
16	0.0	0.0
80	5.6	29.2
400	17.6	41.7
2,000	61.5	79.2
10,000	75.0	91.7
50,000	75.0	95.8
250,000	0.0	100.0

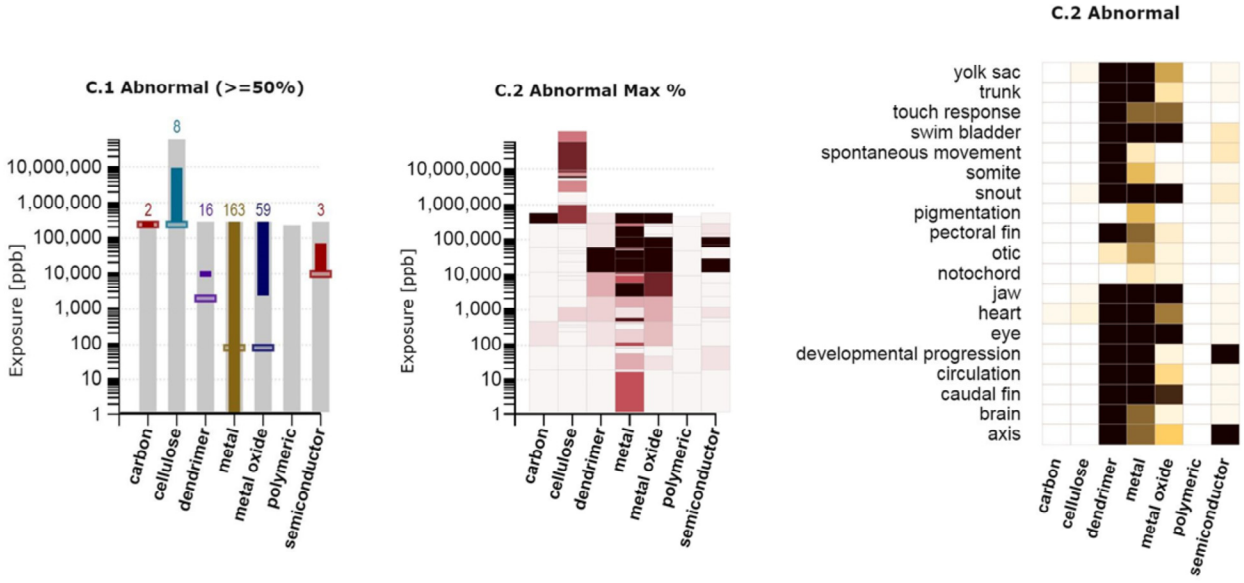
- Nanomaterial NBI_6 showing maximum % response of abnormalities with 100,000 ppb threshold

9/2/2016 Sandra Karcher – Carnegie Mellon University 31

When using visualization option 3, a heat map shows the results of normalization methods C and D, but only responses that were observed at or below a user selected threshold concentration are considered in the application of the color shading. In the example shown on the slide, only responses that were observed at concentrations less than or equal to the selected threshold, in this case, 100,000 ppb, are considered.

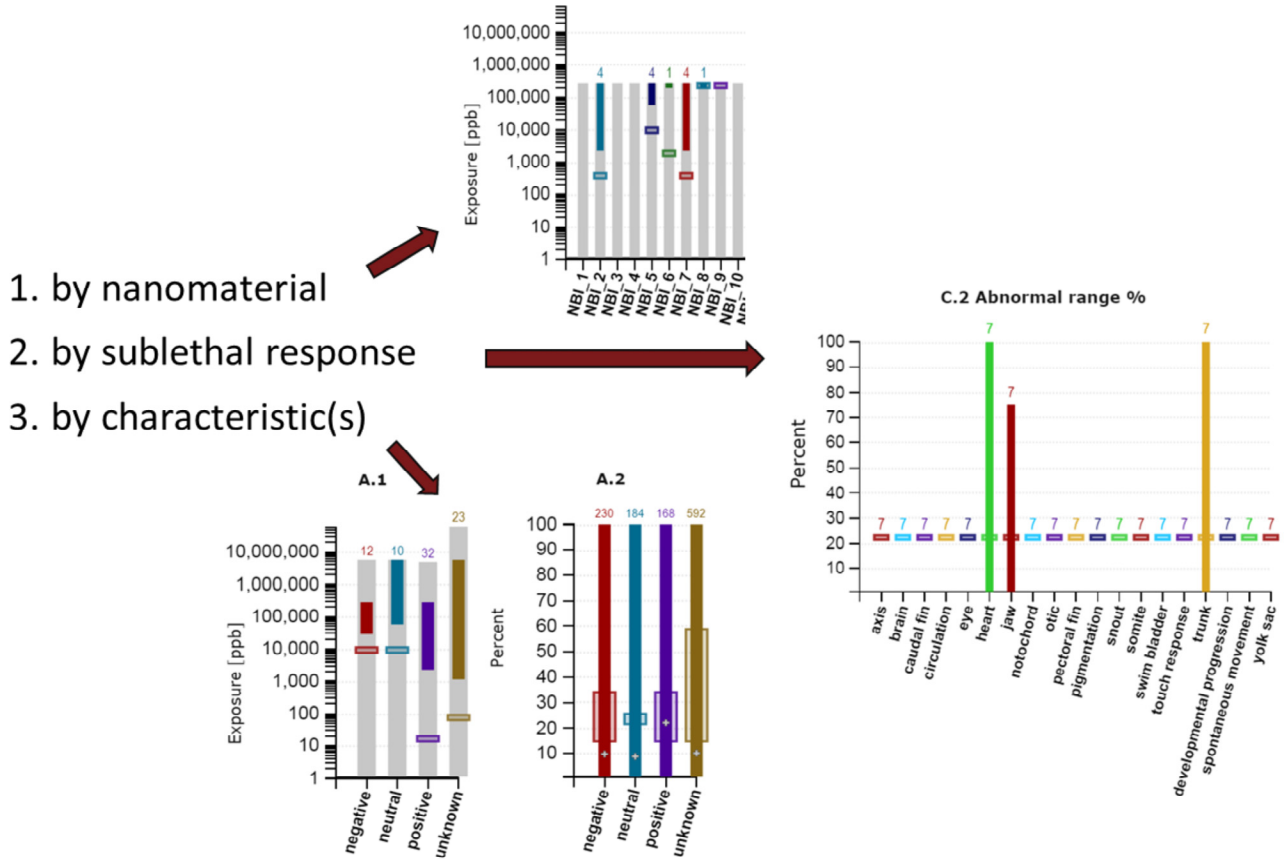
Aggregation

- When more than one response and/or nanomaterial is included in a group, in all cases ***except*** the visualization option 1 percent graphs, the result shown reflects the ***most sensitive response and/or the most toxic nanomaterial*** in the group.



When more than one response and/or nanomaterial is included in a group, in all cases except the visualization option 1 percent graphs, the result shown reflects the most sensitive response and/or the most toxic nanomaterial in the group.

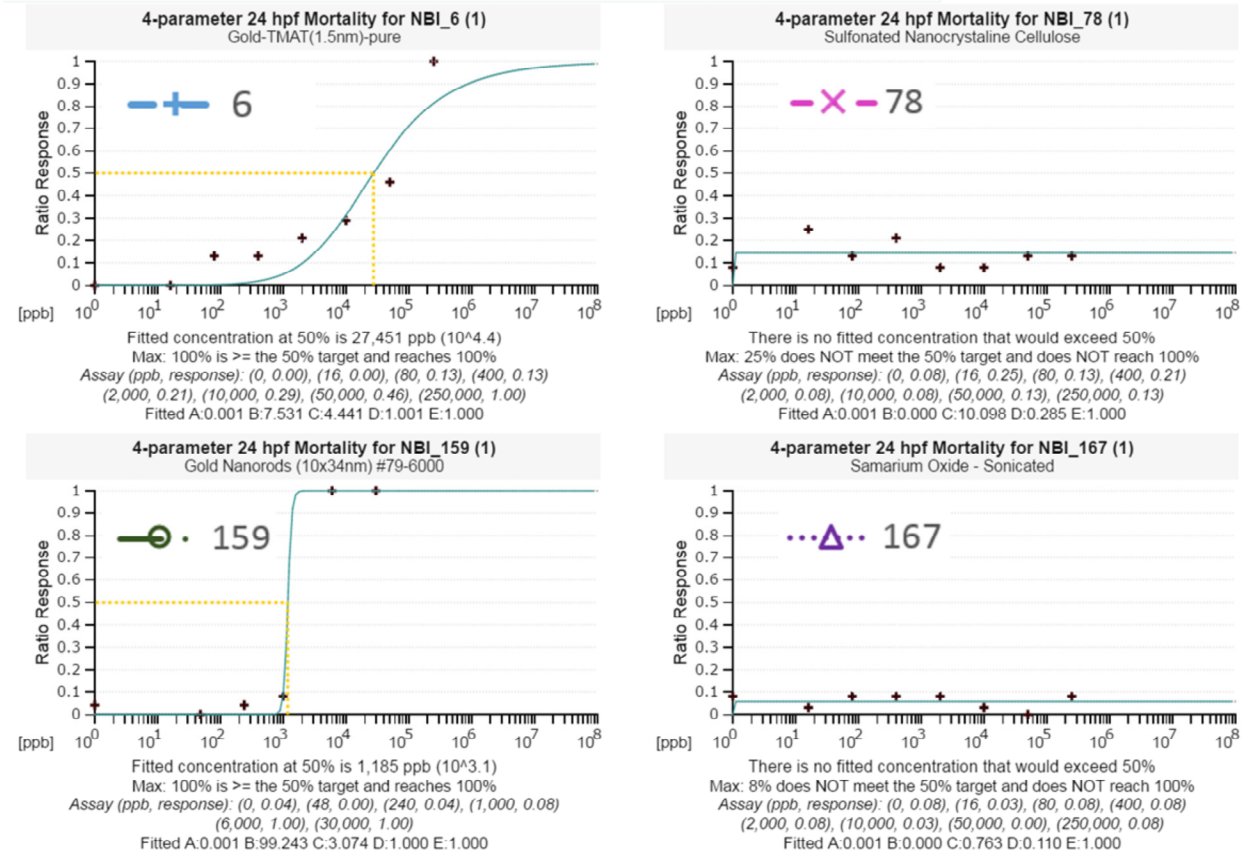
The Three grouping/aggregation methods (x-axis)



9/2/2016 Sandra Karcher – Carnegie Mellon University 33

When using the primary visualization options, there are three methods for grouping data along the x-axis: by nanomaterial, by response, and by characteristics. When data are grouped by nanomaterial, N4mics offers two bonus features, the first is an option to fit a curve to the 24 and 120 hour mortality data. The second is a feature that allows the user to enter weights for each characteristic, and then, based on those weights, calculates a similarity matrix for selected nanomaterials.

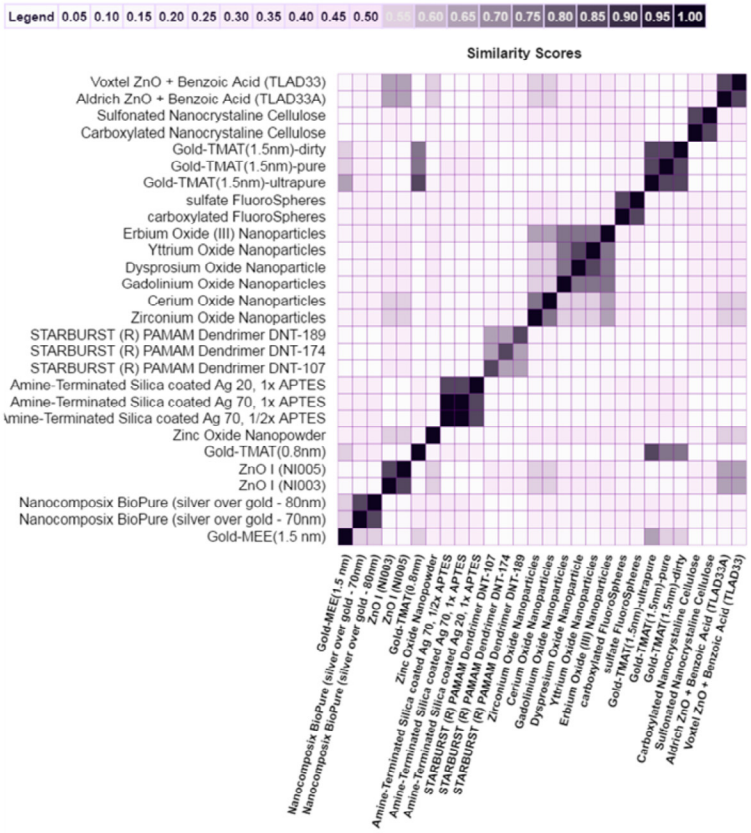
Bonus Feature - Fitted Mortality Curves



As seen earlier in this presentation, N4mics allows the user to perform 4 or 5 parameter logistic regression, using Python's least_squares function, to fit a curve to the 24 and 120 mortality data points and returns the fitted concentration of exposure where the user selected target response percentage is met. The tool allows the user to select the initial values and the lower and upper bounds of the fitting parameters. Curves can be fit to best match the actual data or to achieve more of a sigmoidal look. The tool provides guidance on how to set the parameters. Curve fitting is only active when the user selects to aggregate data by nanomaterial.

The curve fitting feature was developed as a first step in exploring the NBI data using Python's SciPy and NumPy libraries. Unfortunately, available funding for the project did not support continued development in this area. Should more funding become available, development on this path could be further explored.

Bonus Feature – Characteristic Similarity



- darker color indicates nanomaterials that are more similar based on weighted characteristics
- weights assigned by user
- binary comparison for all except size (size is an interval comparison; interval size set by user)

Weights of the Characteristics

Label	Characteristic	Weight
1	descriptor	0.00
2	core material	0.11
3	shell material	0.11
4	surface charge	0.11
5	surface composition	0.11
6	core structure	0.11
7	purity	0.11
8	nanomaterial type	0.11
9	core shape	0.11
10	nanomaterial size	0.11

9/2/2016 Sandra Karcher – Carnegie Mellon University 35

The second bonus feature available in N4mics is also only active when the user selects to aggregate data by nanomaterial. This feature allows the user to select a weight for each nanomaterial characteristic, and then provides a similarity matrix based on those characteristics. The matrix is symmetric about the diagonal that goes from the bottom left to the top right. The diagonal shows the similarity of a nanomaterial to itself (meaning, the diagonal represents the maximum possible similarity).

In the example shown, the characteristics were all weighted equally (note that the descriptor is really not a characteristic, rather, it is a unique identifier for each nanomaterial).

The similarity feature was developed as a first step in defining a path to scoring nanomaterial similarity based on biological response and on patterns of biological response. Unfortunately, available funding for the project did not support continued development in this area. Should more funding become available, development on this path could be further explored.

End of A Query Tool (N4mics) Options and Features

Part 2

This is the end of part 2 of the nanoinformatics series by Sandra Karcher.

Using N4mics

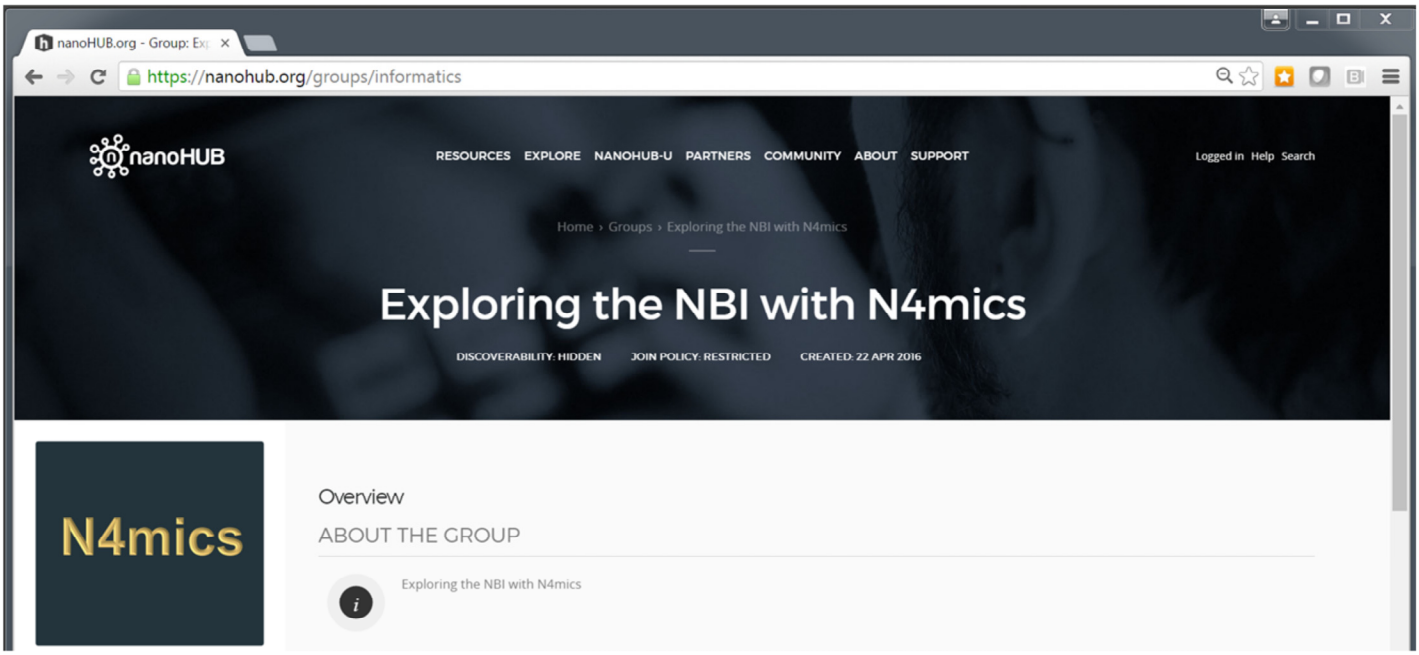
Part 3

Welcome to part 3 of the 3 part series on nanoinformatics by Sandra Karcher.

The N4mics tool is available on nanoHUB.

N4mics on nanoHUB - Group

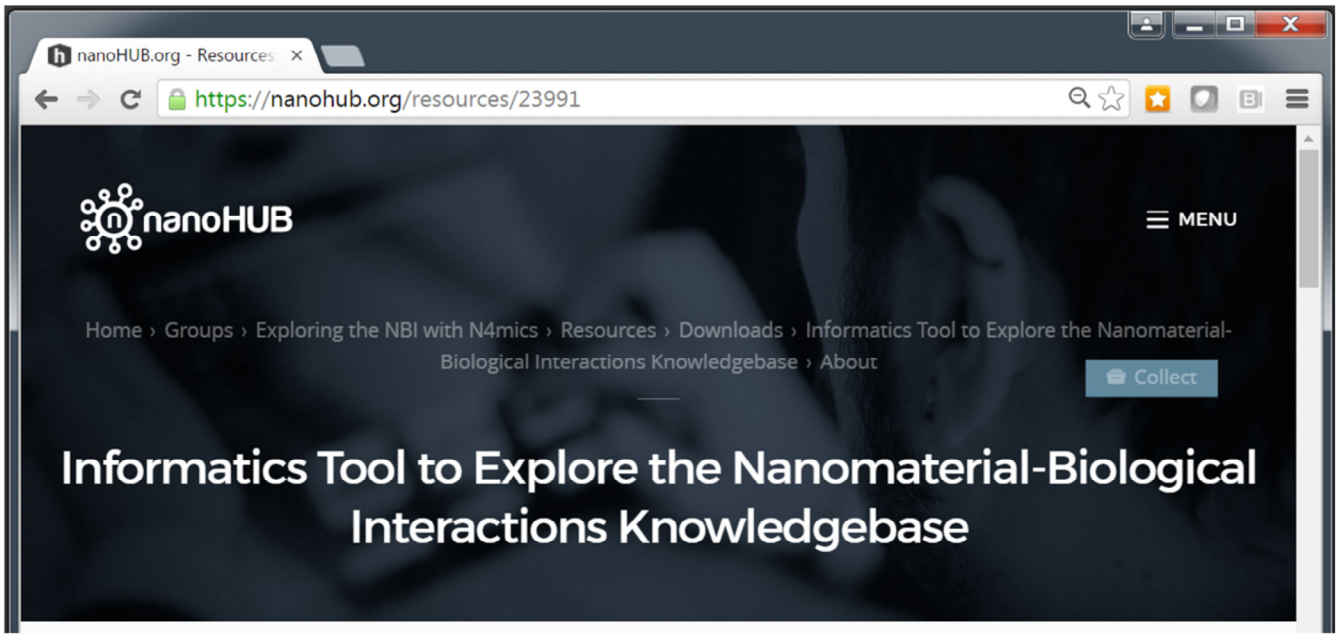
- To access N4mics on nanoHUB, join the group "**Exploring the NBI with N4mics**" located at https://nanohub.org/groups/informatics



To use the tool or view the supporting information available on nanoHUB, you must be a member of the group "Exploring the NBI with N4mics". This group is the place to talk about the tool with colleagues, point out issues with the tool, or make suggestions for tool improvements. Although funding for this project has reached an end, I encourage you to post problems you encounter using the tool. I will try to fix reported bugs as time permits.

N4mics on nanoHUB - Resource

- The N4mics tool is part of the resource "**Informatics Tool to Explore the Nanomaterial-Biological Interactions Knowledgebase**" located at https://nanohub.org/resources/23991



Once you are a member of the "Exploring the NBI with N4mics" group, you should have complete access to the resource, which includes the N4mics tool and a significant amount of supporting information.

When you scroll down the resource page, you will see the N4mics Abstract. The abstract summarizes the information published in the paper that accompanied the release of the tool and introduces the supporting information that is available on the nanoHUB site.

Resource Abstract as of August 31, 2016

About | Citations | Questions | Reviews | Supporting Docs

Category

Downloads

Published on

11 Apr 2016

Abstract

Welcome to the resource: *Informatics Tool to Explore the Nanomaterial-Biological Interactions Knowledgebase*!

Access to this resource is restricted to members of the group Exploring the NBI with N4mics. N4mics was developed to enable cross-study exploration of the NBI Knowledgebase with the goal of identifying relationships between nanomaterial characteristics and biological responses. N4mics allows the user to select nanomaterials and biological responses to investigate and to specify how those data will be aggregated and visualized. N4mics offers three primary options for visualizing data.

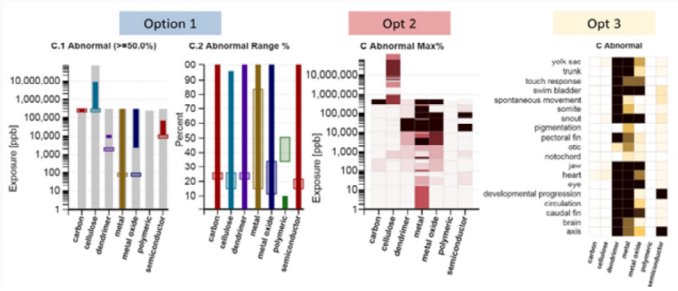
The visualizations generated using the primary options display distinct labels across the x-axis based on the user selected grouping method (by nanomaterial, response, or characteristic(s)). The tool aggregates data for display as follows: (1) two sets of graphs are provided; a set that shows the concentration of exposure where selected responses reached a target percent response, and a corresponding set that shows the full range of responses; (2) the maximum percent response for each concentration of exposure interval is shown using a color ramp; and (3) the maximum percent response for each selected sublethal response is shown using a color ramp for responses that were observed at concentrations of exposure less than or equal to a user selected threshold. Examples of the graphs are shown below.

Exploring the NBI with N4mics

N4mics

Watch resource

When watching a resource, you will be notified of changes made. You may stop watching at any time.



9/2/2016 Sandra Karcher – Carnegie Mellon University 40

The first part of the abstract provides some basic information about the tool and the primary visualization options. This information is also provided in the published paper and in part 2 of this presentation series.

Resource Abstract as of August 31, 2016 (continued)

When the user opts to aggregate data across the x-axis by nanomaterial, N4mics offers two bonus features: 1) logistic regression to fit the mortality curves, and 2) calculation of characteristic similarity scores. More information on using the bonus features is available inside the N4mics tool.

Go to the N4mics tool.

Supporting Information List

Presentation.pdf - This is a three part PowerPoint presentation (Nanoinformatics: Part 1 - The Big Picture Background; Part 2 - N4mics Options and Features; Part 3 – Using N4mics), saved to pdf format with the presentation narrative provided as notes.

Users_Guide.pdf - This document provides instructions on how to navigate within the N4mics tool and explains how to select options for aggregating and visualizing data.

Visualizations_Examples.pdf - Some example visualizations are provided in this document.

Understanding_Aggregation.pdf - This file provides information on how to appropriately interpret the visualizations and on the computational algorithms used to prepare and aggregate data.

NBI_Nano_Characteristics.pdf - This file contains a table of the nanomaterials in the NBI at the time this tool was developed and includes information on the characteristics used by the N4mics tool.

NBI_Excel_File_Format.pdf - The NBI data used in this tool were provided via Microsoft Excel files. This pdf file describes the format of those Excel files.

NBI_SourceData_12082015.zip - The NBI data used in the N4mics tool containing one Excel file for each nanomaterial studied.

Fishers_Exact_Test.pdf - Some of the visualizations generated by the N4mics tool provide information on the statistical significance of observing dead and/or abnormal fish in the assays based on the Fisher's Exact Test. Information on how the Fisher's Exact Test was used is provided in this file.

Why_Zebrafish.pdf - The NBI provides results of experimental studies performed using zebrafish. This file briefly explains the advantages of working with zebrafish.

nbitotxt.py - A python file that can extract the characteristics (used in N4mics) and the associated responses out of the NBI spreadsheets and put them into a pipe delimited text file.

Bio

For more info on Sandra, look here: https://www.linkedin.com/in/sandrakarcher44

When scrolling down, information on the supporting information is provided.

The abstract can be viewed by anyone, but only group members can access the tool and the supporting information.

Summary of How the Tool Works

An "instance" can be conceptualized as a set of user selections and the associated results.

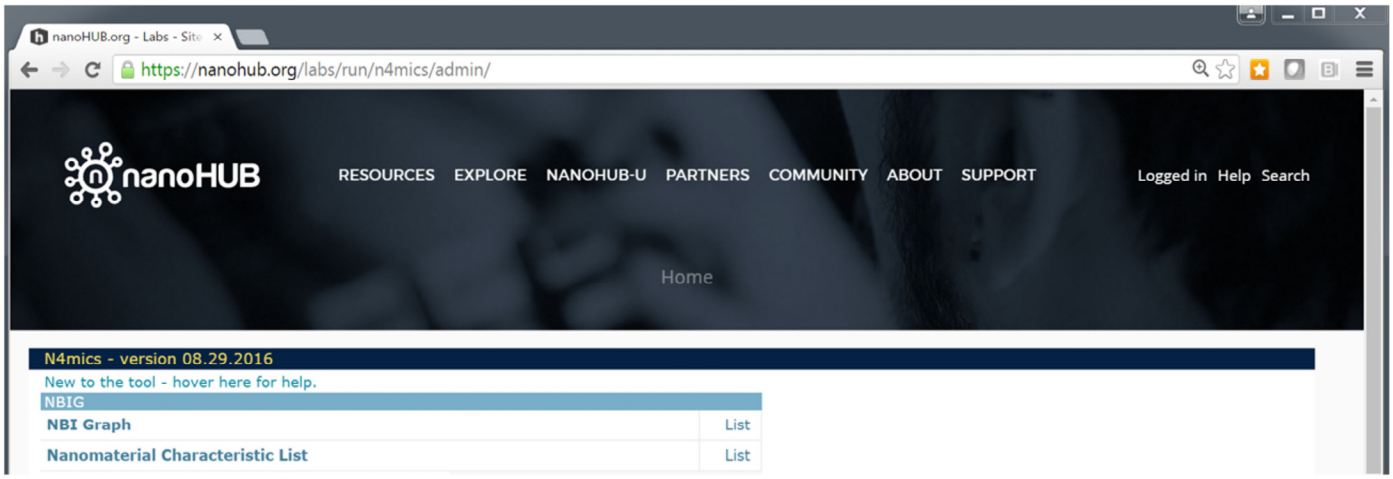
- Typical use of tool
  - add an instance (step 1)
  - customize the instance (step 2)
  - review results
- Can also
  - go back to previously saved (existing) instances and review the results (no need to repeat the step 2 save)
  - edit an existing instance and update the results (must redo the step 2 save)
  - delete existing instances
  - duplicate existing instances
  - download information about existing instances to the screen (where it can be easily copied)

The N4mic tool works by associating data and results through an "instance". An instance can be conceptualized as a set of user selections and the associated results. The user interface guides the user through the process of selecting: data to be used in an analysis, the data grouping method, the target response rate, the threshold concentration, and other parameters needed to process data for visualization, then allows the user access to the generated visualizations.

I envision the tool will typically be used to add an instance, customize that instance, and then review the results of that instance, but the tool also allows the user to store and go back to previously saved instances and review the previously saved results, edit an existing instance and update the results, delete existing instances, duplicate existing instances, and download information about existing instances to the screen (which can be copied and pasted to another file). Some of these features will not be covered in this presentation, but are explained in the user's guide, which is provided in the supporting information available on nanoHUB (https://nanohub.org/resources/23991).

Tool Home Screen

- Two parts of the tool
  - NBI Graph – the part where the visualizations are made
  - Nanomaterial Characteristic List – a table of nanomaterials and associated characteristics

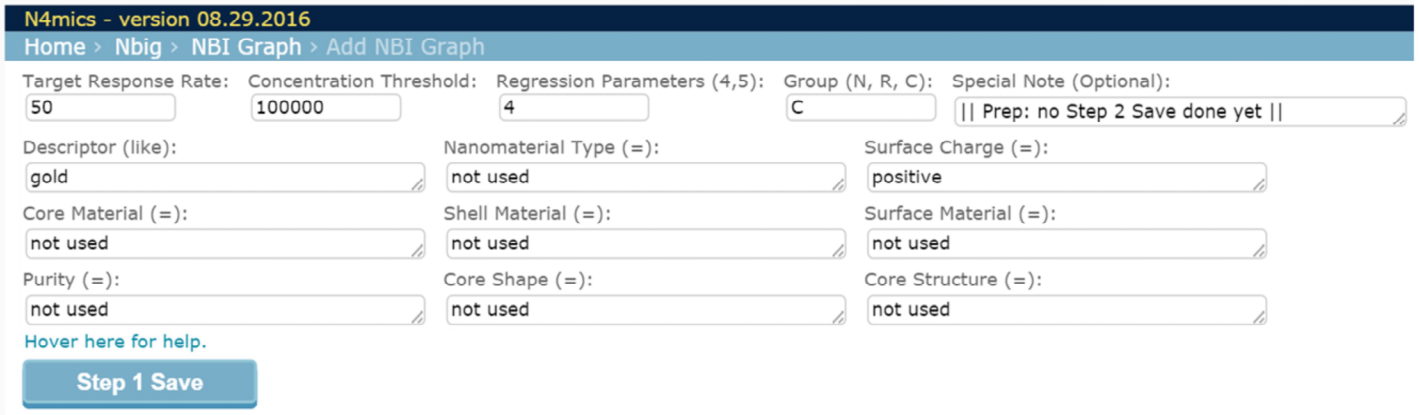


The home screen provides links to the two parts of the tool, the NBI Graph part and the Nanomaterial Characteristic List part. The NBI Graph part of the tool is the subsection that aggregates data and returns visualizations as output. The Nanomaterial Characteristic List part of the tool provides a list of all the nanomaterials explored using N4mics. A list similar to this, along with the maximum concentration of exposure used in the assays of each nanomaterial, is available in pdf format in the supporting information available on nanoHUB (https://nanohub.org/resources/23991).

The rest of this demonstration will focus on the NBI Graph part of the tool - the part where the visualizations are made. To access the graph feature, click on the NBI Graph active link, or on the List button next to the link. Then, click on the ADD NBI GRAPH button.

Adding an Instance – Step 1

- Instances are added using the "ADD NBI GRAPH"
- Fields are prepopulated but can be edited by the user

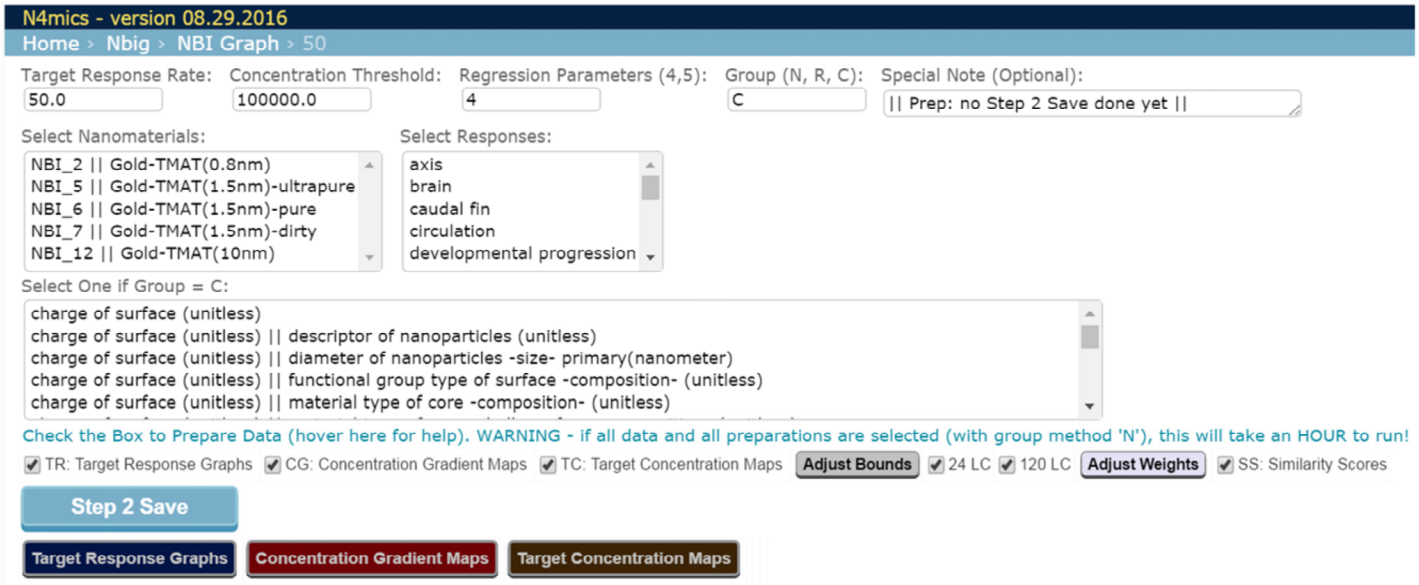


The add new form displays prepopulated boxes of information, all of which can be changed by the user. The visualization option 1 target response rate is prepopulated with 50. The visualization option 3 threshold concentration is prepopulated with 100,000 ppb. The default for the logistic regression bonus feature is 4-parameter regression. The grouping method is prepopulated with "C" (but the user can change it to "N" or "R"). The beginning of the Special Note box is controlled by the tool and is used to indicate how data have been prepared. The end of the Special Note box is free format text and can be populated by the user in a way that will assist them in recalling the appropriate use of the instance.

Moving down the form, the next set of boxes allows the user to limit the nanomaterials that will be explored in the current instance. The Descriptor box is a "like" condition box, the remaining characteristic boxes are "equal" condition boxes. For this demonstration, nanomaterials with a positive surface charge whose descriptors contain the string "gold" were selected for further consideration.

After the selections are made, the user must click the step 1 button to save and continue.

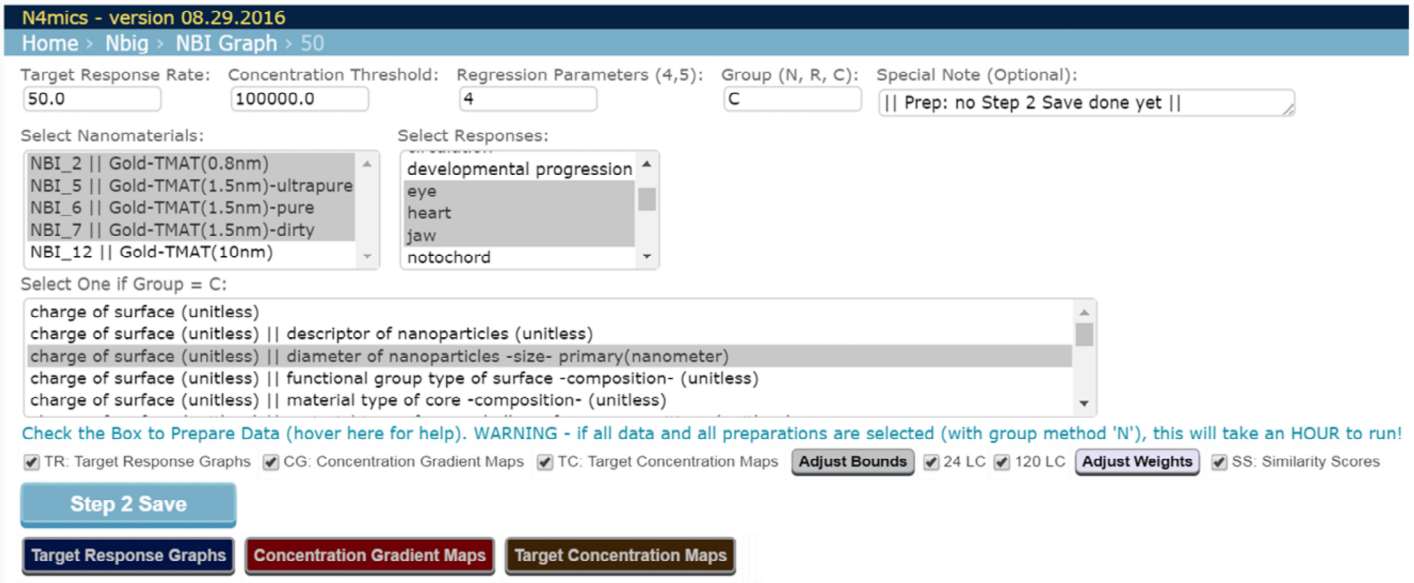
Customizing the Instance – Step 2



After the step 1 save is completed, the step 2 customization form is displayed. The top boxes of the step 2 form are the same as in the step 1 form. If desired, the response rate, concentration threshold, parameter number, grouping method, and the end portion of the special note fields can be updated here in the step 2 customization form.

Note that the nine characteristic boxes are gone and that new pick boxes have appeared. The pick boxes provide a means of further excluding nanomaterials and excluding specific biological responses from the current instance. Notice that only nanomaterials with "gold" in the descriptor are included in the nanomaterial pick box options, and, although not evident by the information provided in the pick boxes, the nanomaterials available in the list have a positive surface charge (because entering "gold" in the step 1 Descriptor box and "positive" in the step 1 Surface Charge box preselected only gold nanomaterials with a positive charge for display in the step 2 nanomaterial pick box).

Pick Box Selections

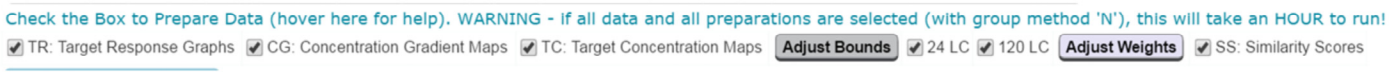


- Pick boxes designed so that selected items remain, unselected items are removed UNLESS, none are selected, then they all remain in
- If grouping method "C" is selected, one of the combinations of characteristics must be selected

The pick boxes are designed so that, if any of the nanomaterials or responses are selected, the ones that are not selected will be removed from further analysis. If none are selected, all that are displayed will remain for analysis in the current instance. With regard to the characteristic, when the group method "C" is selected, one of the combinations of characteristics must be selected. If an "N" or "R" is selected as the grouping method, the user is not required to select a characteristic.

Data Preparation Options

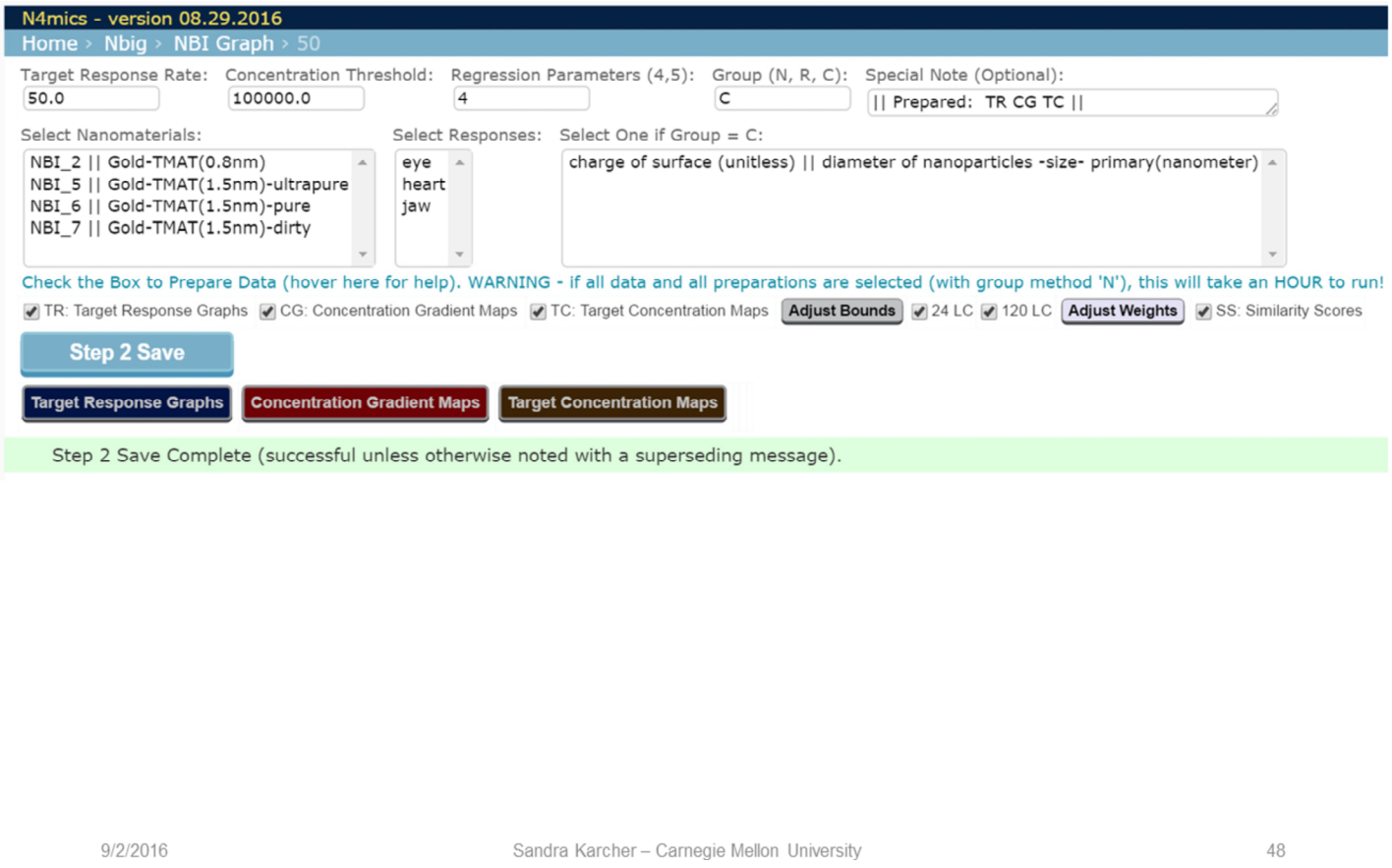
- Time consuming preparations can be bypassed if not needed.
- Preparing the concentration maps and the similarity scores can be time consuming, depending on the amount of data included and the number of bins data are being sorted into.
- New users are encouraged to read all the hover help tips. The data preparation help tip informs the user that, if they are new to the tool, start with a very small subset of nanomaterials and review all results carefully.



The next section of the form allows the user to select what visualization data preparation the tool should perform. Running the step 2 save will completely overwrite any previous saves. So, if all visualization methods are wanted, they must be run at the same time or run in multiple instances. The special note field can be used to keep track of which preparations were prepared in each instance.

If the selected data are to be grouped into a lot of “bins”, it is best to run the CG, TC, and SS preparations in separate instances. When data are processed into a lot of bins, these preparation processes take a long time (up to 30 minutes each) and processes that take over 40 minutes can cause errors in the nanoHUB framework.

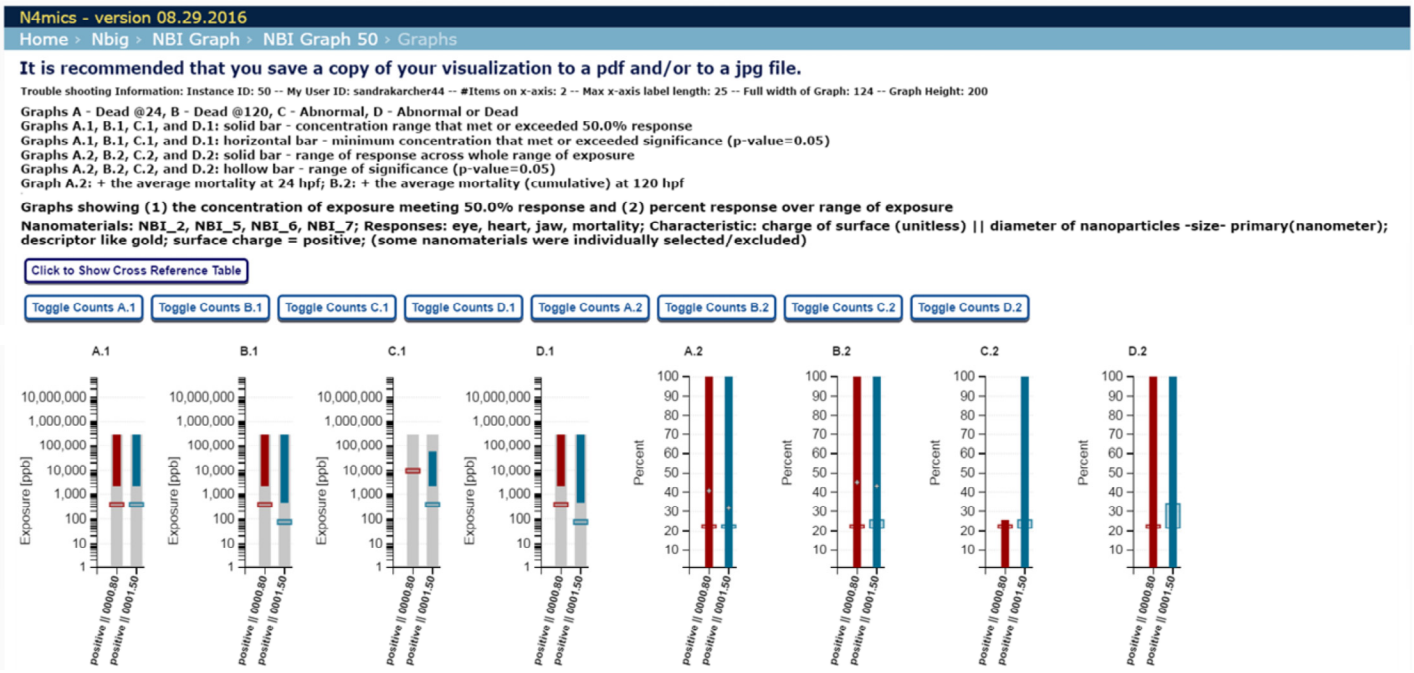
Successful Step 2 Save



After the user has completed the customization of the instance, the step 2 save must be performed. Running the step 2 save will overwrite any previous saves associated with this instance.

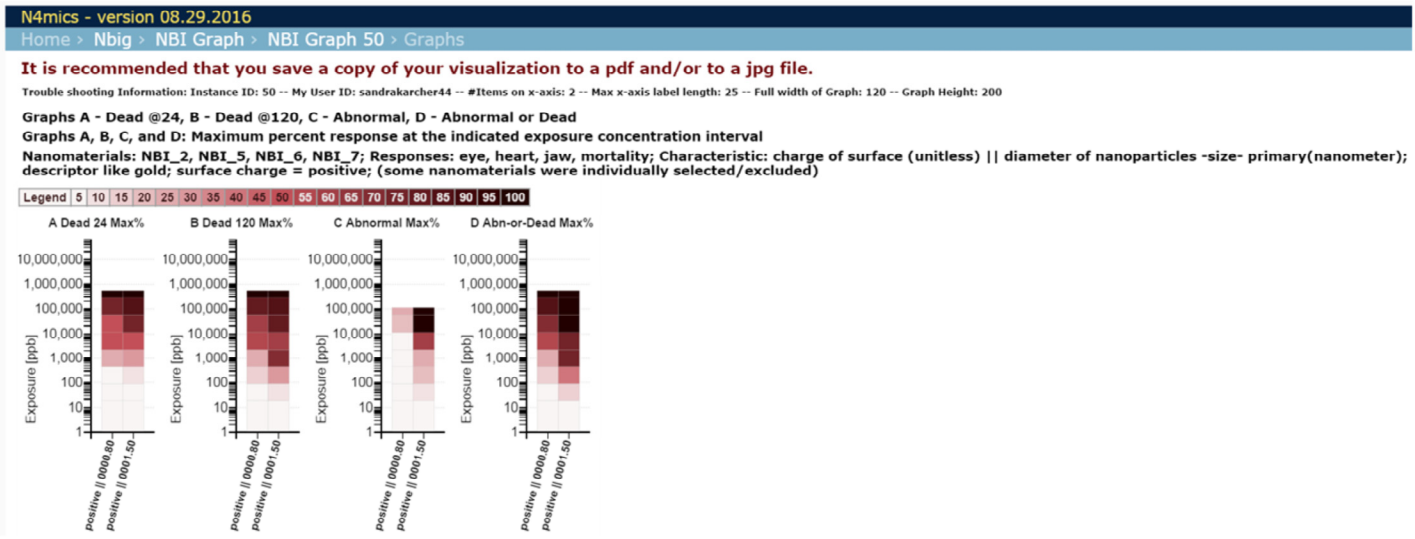
Once the step 2 save has been completed, the user can review the results by clicking on one of the click buttons near the bottom of the form.

Result of Visualization Option 1



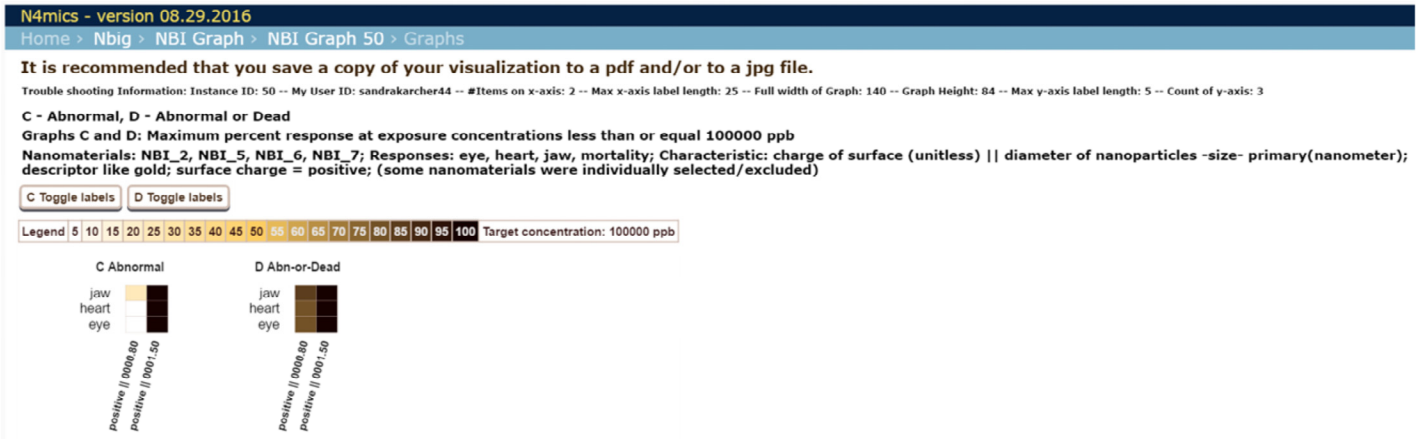
These are the results for visualization option 1. See the paper published with the release of the tool, part 2 of this presentation, or the supporting information (https://nanohub.org/resources/23991) to learn more about interpreting these graphs. Our focus here in part 3 of the series is on how to use the tool, so we will move on.

Result of Visualization Option 2



These are the results of visualization option 2. See the paper published with the release of the tool, part 2 of this presentation, or the supporting information (https://nanohub.org/resources/23991) to learn more about interpreting these graphs.

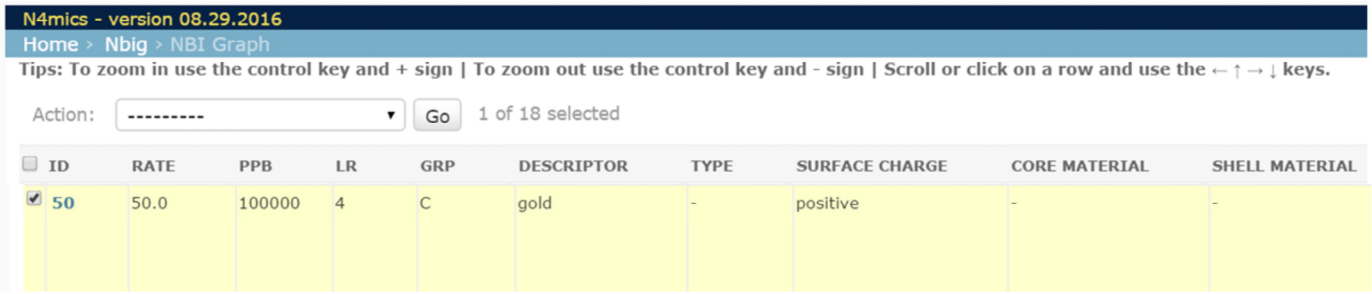
Result of Visualization Option 3



These are the results of visualization option 3. See the paper published with the release of the tool, part 2 of this presentation, or the supporting information (https://nanohub.org/resources/23991) to learn more about interpreting these graphs.

Grouping by "N" and the Bonus Features

- In the list of instances, check the box of instance (50) and in the Action box, select Duplicate selected NBI Graph and click Go.



- Then, click to the new instance (51) and Adjust the Bounds or Weights as desired.

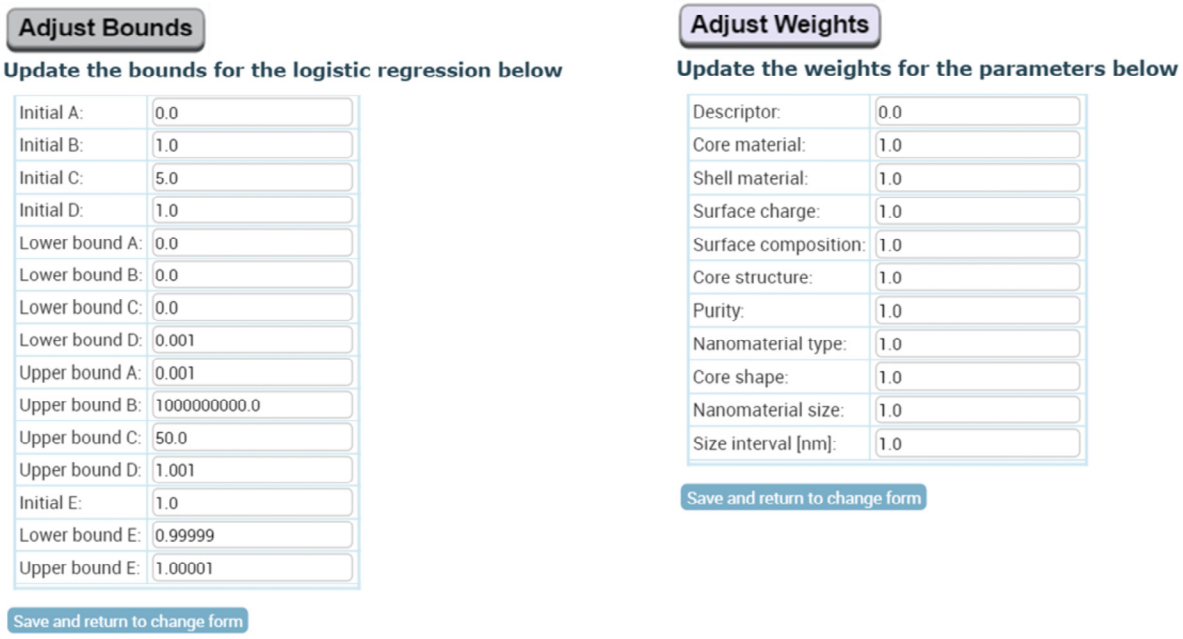
9/2/2016 Sandra Karcher – Carnegie Mellon University 52

Recall that there are two bonus features that are active when the grouping method "N" is selected. We can duplicate the instance we were just working on (instance 50) by selecting the instance (check the box next to it), then selecting "Duplicate selected NBI Graph" from the Action options, and then clicking the Go button. Once duplicated, it can be edited by clicking on the instance number. In this case, click the 51 to go to the step 2 change form.

Once in the form, the user can change the parameters used in the bonus feature calculations by clicking on the buttons labeled "Adjust Bound" and "Adjust Weights".

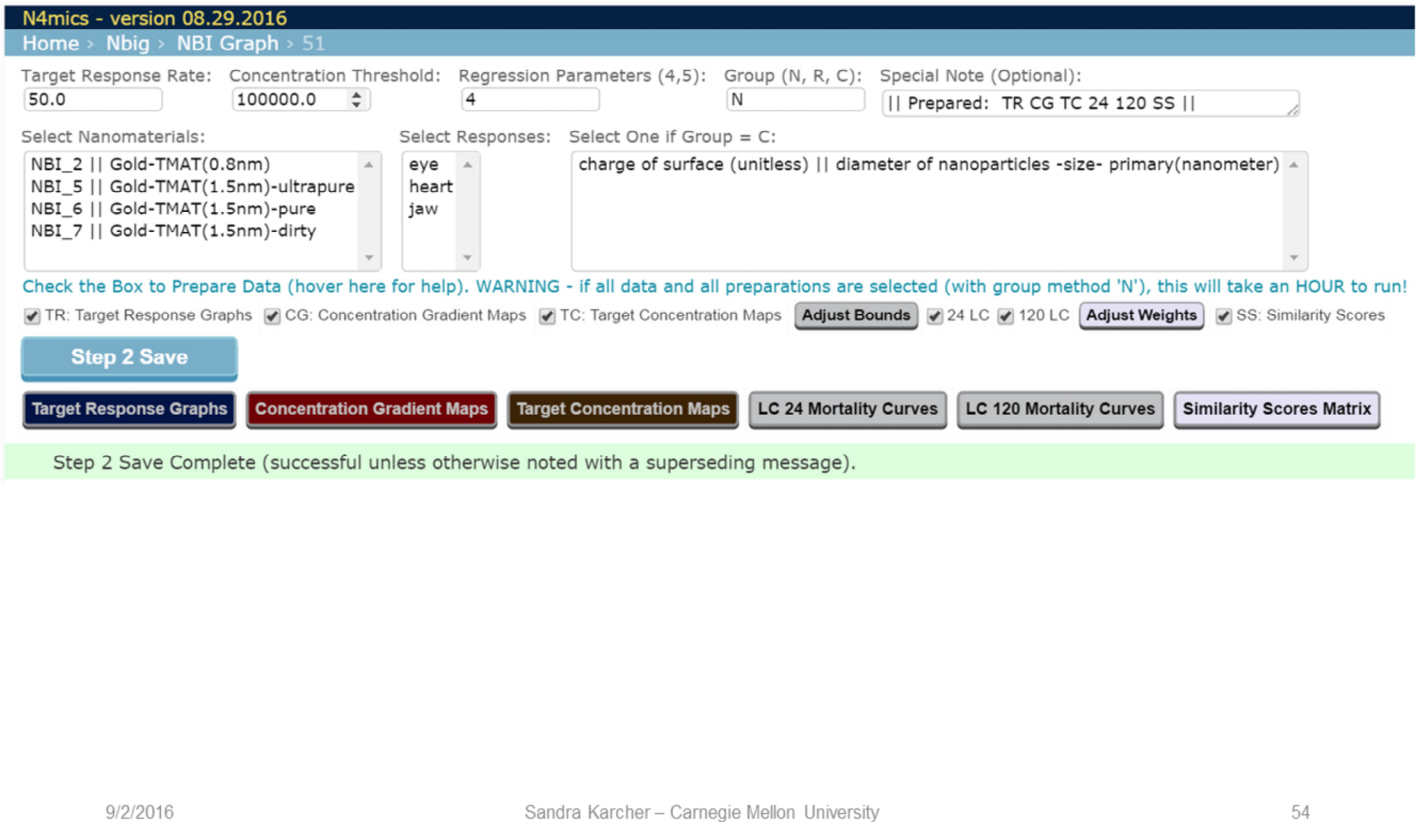
Grouping by "N" and the Bonus Features

- The "Adjust" button routes the user to forms where bonus feature parameters can be adjusted



Clicking on the button to adjust the bounds of the logistic regression routes the user to a form like that shown on the left. Clicking on the button to adjust the characteristic weights routes the user to a form like that shown on the right. The user is free to edit these parameters as they deem appropriate, then save their changes and return to the step 2 customization form.

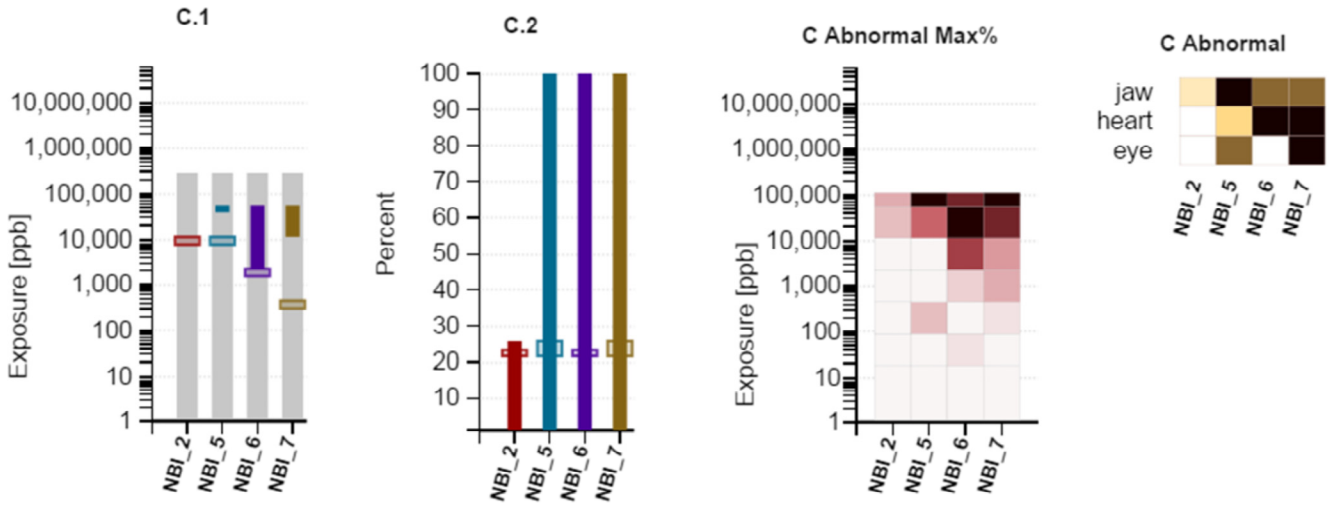
Grouping by “N” and the Bonus Features



Once back in the step 2 form, the group method can be changed to "N" and then a step 2 save completed.

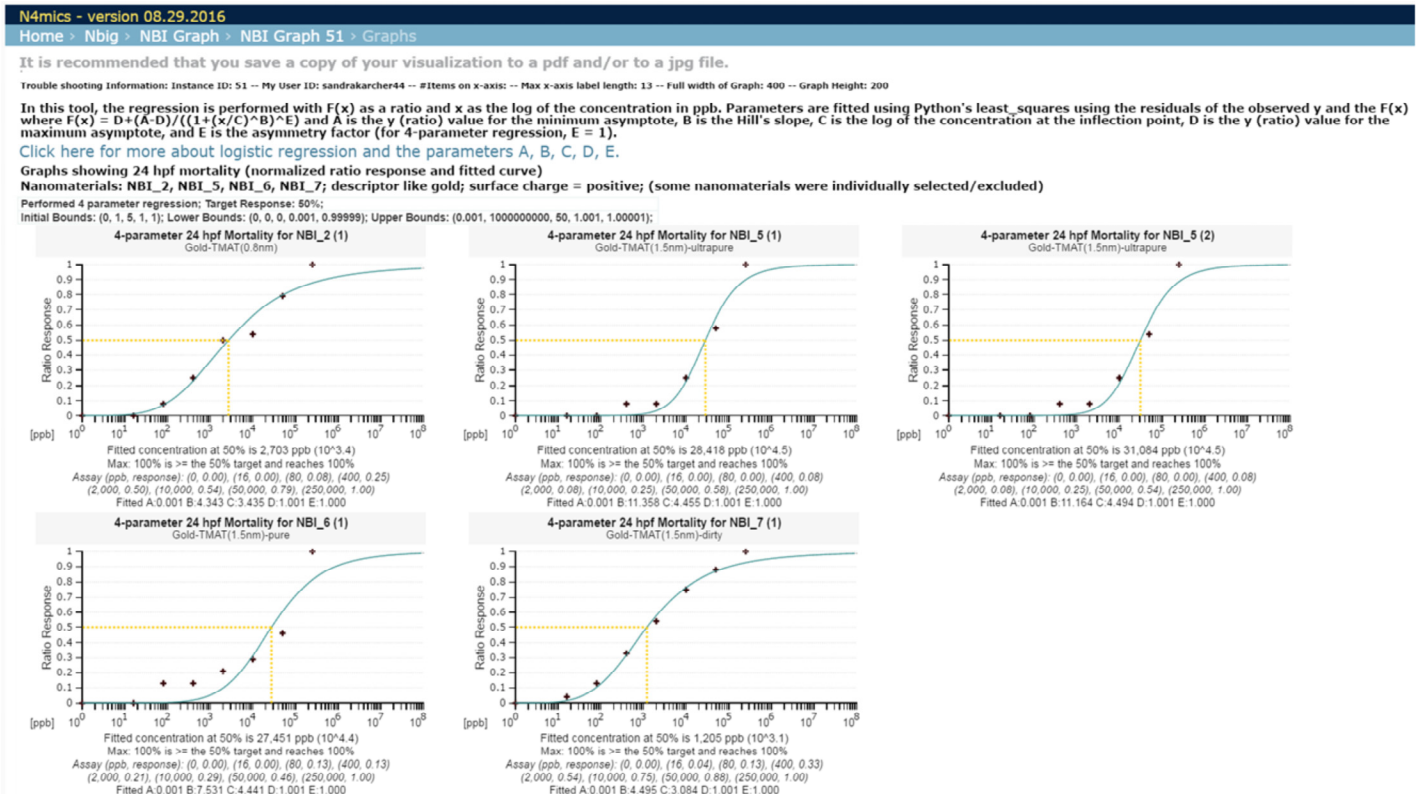
To review the results, the user clicks on the appropriate visualization button. Recall that performing a step 2 save overwrites all previous saves for this instance. This means that the graphs for the primary visualizations are now different because the group method was changed to “N” prior to performing the step 2 save.

Quick Look at Revised Results



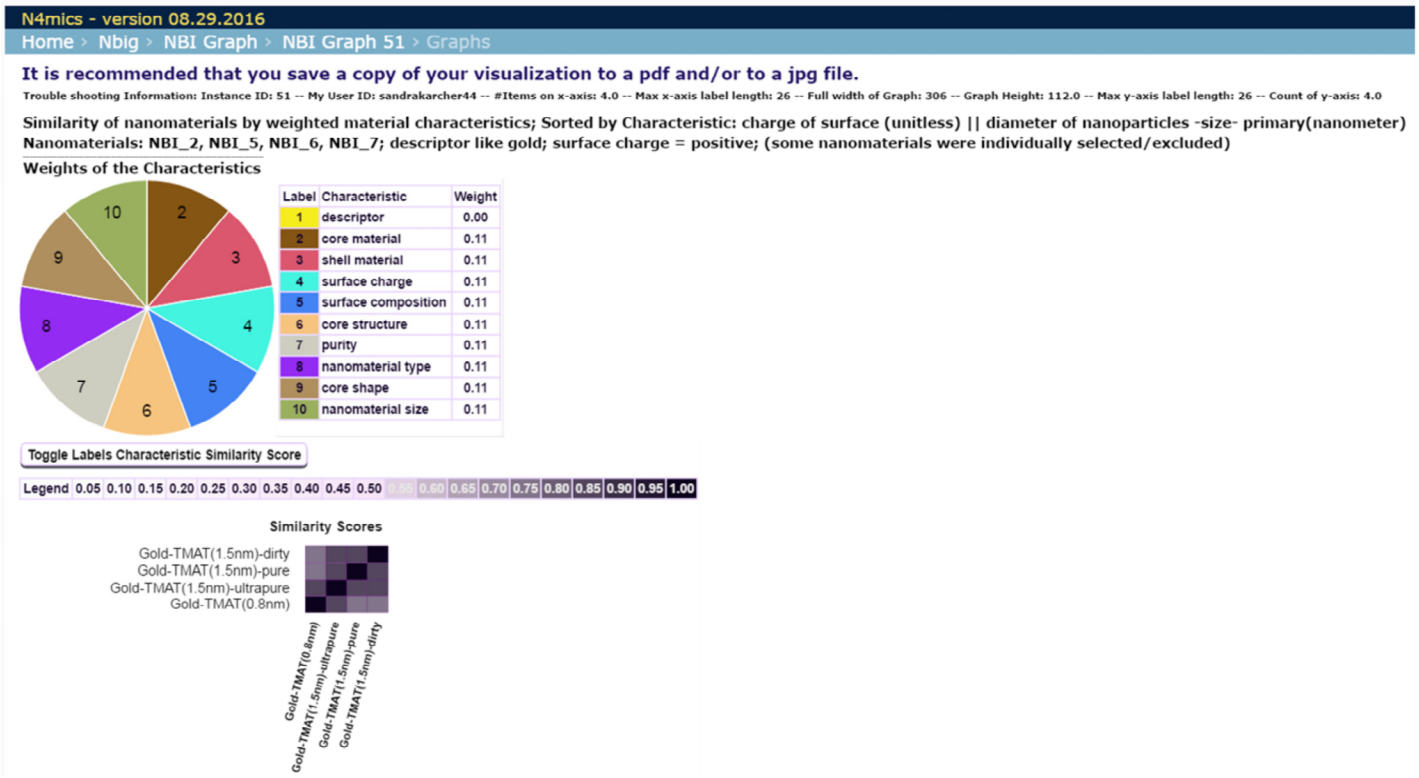
If we take a quick look at the revised results (for the Abnormal (C) graphs only), we see that the x-axis now shows the nanomaterial instead of the combination of characteristics.

Result of 24 Hour Mortality Curve Fitting



The results of the 24 hour mortality curve fitting are shown here. Notice that four nanomaterials were selected for inclusion in the analysis but five graphs are shown. This is because nanomaterial NBI_5 was reported in the NBI as having been used in two different assays. Keep in mind that the shape of these curves is highly dependent on the bounds used in the fitting.

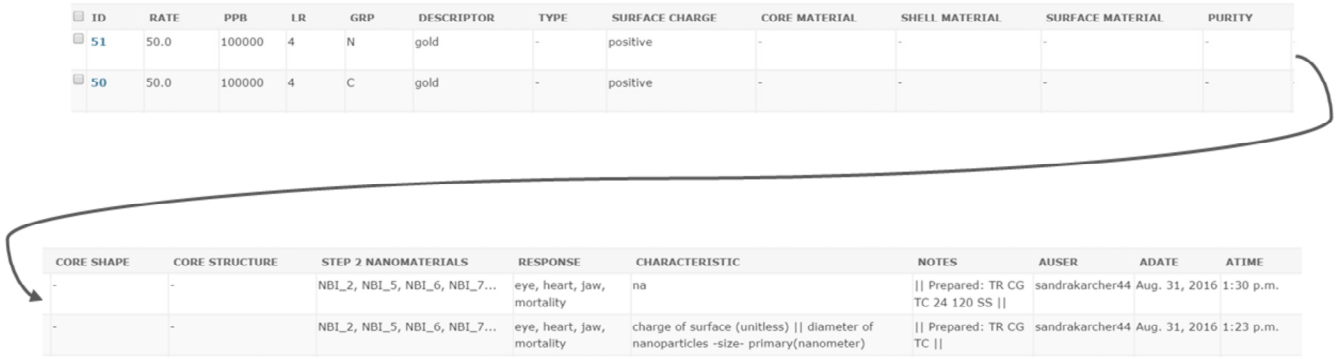
Result of Similarity Score



These are the results of the similarity score calculations. As indicated in part 2 of this presentation series, the similarity matrix is symmetric about the diagonal that goes from the bottom left to the top right. The similarity feature was developed as a first step in defining a path to scoring nanomaterial similarity based on biological response and on patterns of biological response. Unfortunately, available funding for the project did not support continued development in this area. Should more funding become available, development on this path could be further explored.

The Instance List

- Once instances have been added, the instance list can be used to review the selections



ID	RATE	PPB	LR	GRP	DESCRIPTOR	TYPE	SURFACE CHARGE	CORE MATERIAL	SHELL MATERIAL	SURFACE MATERIAL	PURITY
51	50.0	100000	4	N	gold	-	positive	-	-	-	-
50	50.0	100000	4	C	gold	-	positive	-	-	-	-

CORE SHAPE	CORE STRUCTURE	STEP 2 NANOMATERIALS	RESPONSE	CHARACTERISTIC	NOTES	AUSER	ADATE	ATIME
-	-	NBI_2, NBI_5, NBI_6, NBI_7...	eye, heart, jaw, mortality	na	|| Prepared: TR CG TC 24 120 SS ||	sandrakarcher44	Aug. 31, 2016	1:30 p.m.
-	-	NBI_2, NBI_5, NBI_6, NBI_7...	eye, heart, jaw, mortality	charge of surface (unitless) || diameter of nanoparticles -size- primary(nanometer)	|| Prepared: TR CG TC ||	sandrakarcher44	Aug. 31, 2016	1:23 p.m.

Now that we have created two instances, we can use the instance list to review the choices made in generating them. Recall that the home screen provided links to two parts of the tool. The user is routed to the instance list by clicking on the NBI Graph active link. Understanding how to read the information in the instance list enables appropriate use of the associated visualizations, and also prevents unnecessary duplication of the exact same selection criteria being used in multiple instances.

Reading across the instance list from left to right, the ID number is automatically generated by the tool and is provided for reference and for sorting purposes. The IDs are assigned independent from the user name, thus, an individual user may see large gaps in their ID number list. The RATE is the target response criterion used for the visualization option 1 graphs. The PPB is the threshold concentration of exposure used for visualization option 3. The LR indicates the selection of 4 or 5 parameter logistic regression. The GRP is the data grouping method. The next nine fields directly match the ones displayed in the step 1 "like" and "equal" boxes. The STEP 2 NANOMATERIALS indicates which, if any, nanomaterials were specifically selected using the pick box on the step 2 customization form. The RESPONSE indicates which, if any, responses were specifically selected using the pick box on the step 2 customization form. The CHARACTERISTIC indicates the characteristics or characteristic combination that was selected, if one was selected, using the pick box on the step 2 customization form. The NOTES indicates the data preparations checked by the user and, if provided, information typed in by the user. AUSER, ADATE, and ATIME are populated by the tool and cannot be changed by the user. The date and time are populated to indicate the last step 2 save.

End of Using N4mics

Part 3

More information, including a user's guide, is provided in the supporting information on nanoHUB (https://nanohub.org/resources/23991).

This concludes my three part series on nanoinformatics and the N4mics tool.

I hope you will explore the N4mics tool and provide your feedback in the nanoHUB group "Exploring the NBI with N4mics".