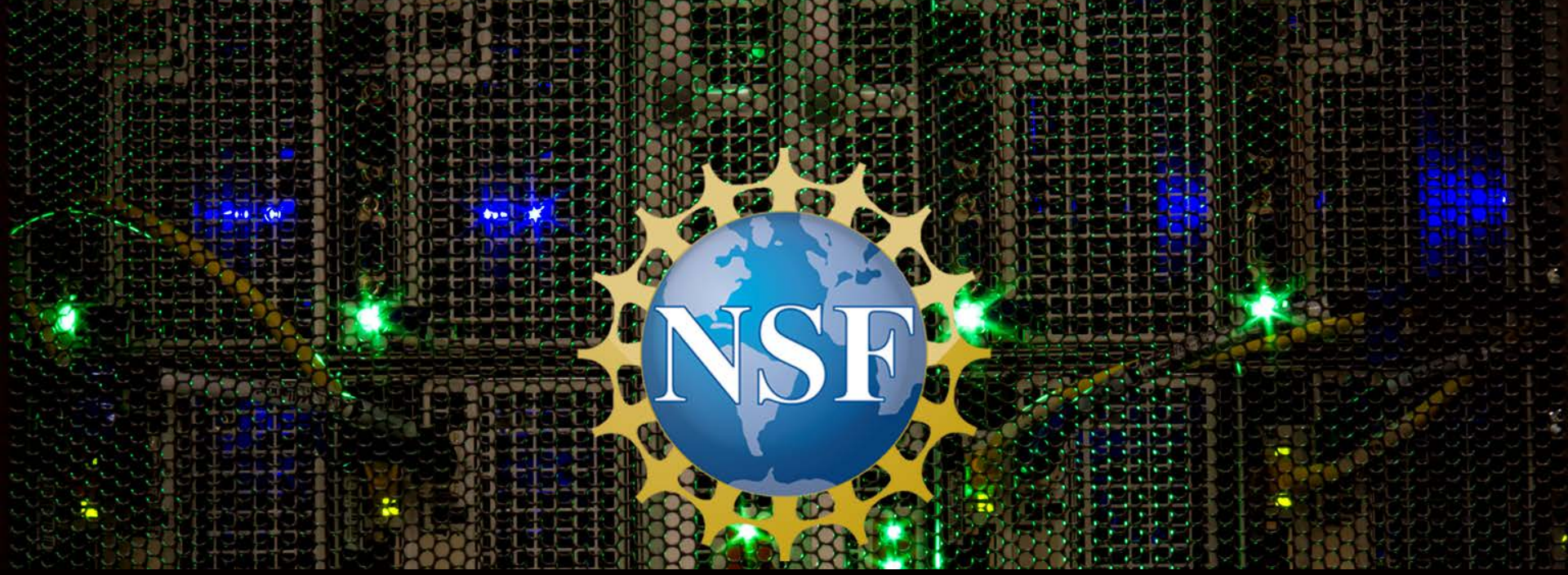


Data Management and the Research Record in Research Misconduct Investigations

Kenneth L. Busch, Ph.D.
National Science Foundation, Office of Inspector General

National Data Integrity Conference May 2015

Who is NSF OIG?

- Independent office reporting to the Congress and NSB.
- **Promote economy, efficiency, and effectiveness.**
- **Prevent and detect fraud, waste, and abuse.**
- Accomplishes mission through:
  - Audits
  - Investigations
    - Criminal and Civil (*e.g.*, false claims, false statements, embezzlement).
    - Administrative (*e.g.*, regulatory and policy violations).

Where does research misconduct fit in?

OIG is delegated the responsibility for investigating research misconduct allegations involving NSF programs.

Research Misconduct (RM)

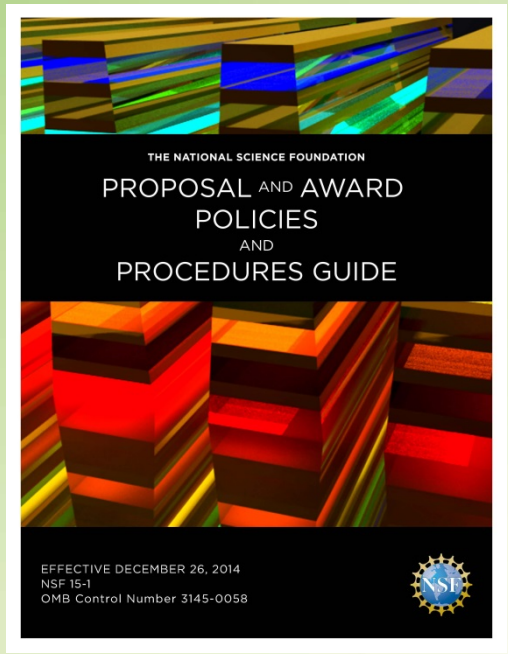
- Federal definition and procedural framework (OSTP Dec. 2000).
- RM means "fabrication, falsification, or plagiarism in proposing or performing research … , reviewing research proposals … or in reporting research funded by [the agency]." (45 C.F.R. 689.1(a))
- Not honest error or differences of opinion.
- Must be "reckless, knowing, or intentional" and not careless.
- Must be "a significant departure from the accepted practices of the relevant research community."

NSF Grant Policy Requirements

- In submitting a proposal to NSF, a grantee certifies to the accuracy and truthfulness of the materials in the proposal
- In accepting an award, a grantee accepts to the terms of NSF policy, and any special conditions that apply to that particular program or award.
- Many of the general provisions apply directly to financial administration of the award, such as time and effort reporting, auditability, and specific requirements such as the use of U.S. flag carriers for travel supported by NSF. The list is long.
- However, many provisions apply directly to the conduct of research -- research misconduct, IRB, IACUC, and presentation and publication of results
- AND specific policies apply to issues of **DATA MANAGEMENT**.

NSF Grant Policies

- **Data management**
  - Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled “Data Management Plan”. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include: 1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project; 2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies); 3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements; 4. policies and provisions for re-use, re-distribution, and the production of derivatives; and 5. plans for archiving data, samples, and other research products, and for preservation of access to them.



Data management plans are reviewed as part of merit review.

NSF Grant Policies

- **Data sharing**
- **734 Dissemination and Sharing of Research Results**
- Investigators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work conducted under NSF grants. Grantees are expected to permit and encourage such publication by those actually performing that work, unless a grantee intends to publish or disseminate such findings itself.
- Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division for a particular field or discipline to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators. A grantee or investigator also may request a particular adjustment or exception from the cognizant NSF Program Officer.
- Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.
- NSF normally allows grantees to retain principal legal rights to intellectual property developed under NSF grants to provide incentives for development and dissemination of inventions, software and publications that can enhance their usefulness, accessibility and upkeep. Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers.

NSF Grant Policies

- **Data retention**
  - Chapter II.E., Record Retention and Audit, states that financial records, supporting documents, statistical records and all other records pertinent to the NSF grant must be retained by the grantee for a period of three years from award financial closeout described in AAG Chapter III.E.3, except as noted in 2 CFR § 200.333.

Not from expiration date
Not from date of final report
Not from NSF approval of final report

This is a grantee responsibility.

NSF Grant Policies

- **America Competes Act (2007)**

America Competes Act PUBLIC LAW 110–69—AUG. 9, 2007
SEC. 7011. SHARING RESEARCH RESULTS. An investigator supported under a Foundation award, whom the Director determines has failed to comply with the provisions of section 734 of the Foundation Grant Policy Manual, shall be ineligible for a future award under any Foundation supported program or activity. The Director may restore the eligibility of such an investigator on the basis of the investigator's subsequent compliance with the provisions of section 734 of the Foundation Grant Policy Manual and with such other terms and conditions as the Director may impose.

This is an individual responsibility.

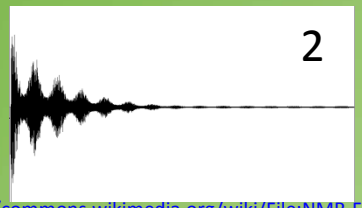
In simple words . . .

- The research record tells others what you did, how you did it, and what you observed, so that others can assess it, reproduce it, and build on it.
- The research record generally includes:
  - the data in the most original form reasonable.
  - the data and analysis for the specific results you chose to publish, and for what you excluded.
- Of course, the research record is amorphic and growing rapidly.
- Increased expectations (and requirements) for data sharing
- Increased expectations (and requirements) for data mining

Every job is easy if it's someone else doing the work

Raw vs. Processed Data

UNITY Plus 600
Emory Univ.
NMR Center



http://commons.wikimedia.org/wiki/File:NMR-FID.png

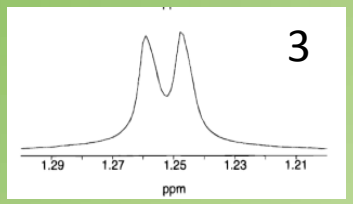


Table 1. NMR Data in $CDCl_3$ for $LCo(DH)_2CH_3$ and Steric Parameters for L

4

	L	Co−CH_3 $^1J_{CH}$ (Hz)	TCA (deg)[a]	CCA (deg)[b]	E_L (kcal/mol)[c]
1	Me_3P	137.06	118	115	39
2	Et_3P	137.82	132	130	61
3	n-Bu_3P	137.70	132	131	64
4	i-Pr_3P	139.01	160	164	109
5	Cy_3P	139.06	170	171	116
6	Bz_3P	138.38	165	136	82
7	$(NCCH_2CH_2)_3P$	139.70	132	134	
8	Ph_3P	139.51	145	160	75
9	$(p$-$ClPh)_3P$	140.26	145[d]	160[d]	74
10	$(p$-$FPh)_3P$	139.98	145[d]	160[d]	74
11	$(p$-$MePh)_3P$	139.47	145[d]	160[d]	74
12	$(p$-$MeOPh)_3P$	139.28	145[d]	157	76
13	$(p$-$Me_2NPh)_3P$	138.58	145[d]	160[d]	
14	$EtPh_2P$	138.64	140	145	66
15	i-$PrPh_2P$	139.26	150	166	75
16	Et_2PhP	137.82	136	137	57
17	Me_2PhP	137.55	122	125	44
18	$(NCCH_2CH_2)_2PhP$	139.24	136	142	
19	$(PhO)_3P$	139.00	128	119	65
20	$(MeO)_3P$	138.15	107	115	52
21	$(i$-$PrO)_3P$	137.96	130	121	74

[a] Tolman cone angle (ref 21). [b] Calculated cone angle (ref 17). [c] Ligand repulsive energy (ref 27). [d] Value estimated from Ph_3P.

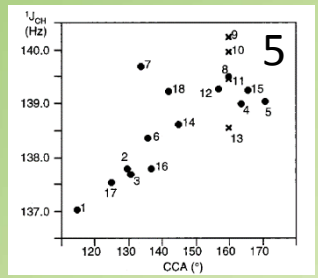


Table 3. Regression Analysis of $^1J_{CH}$ with Steric Parameters

6

	without $(p$-$RPh)_3P$ estimates		with $(p$-$RPh)_3P$ estimates	
	r_A^2	p	r_A^2	p
CCA	0.8784	<0.0001	0.7778	<0.0001
TCA	0.5357	0.0066	0.2935	0.0216
E_L	0.5108	0.0081	0.2729	0.0266

Raw

Processed

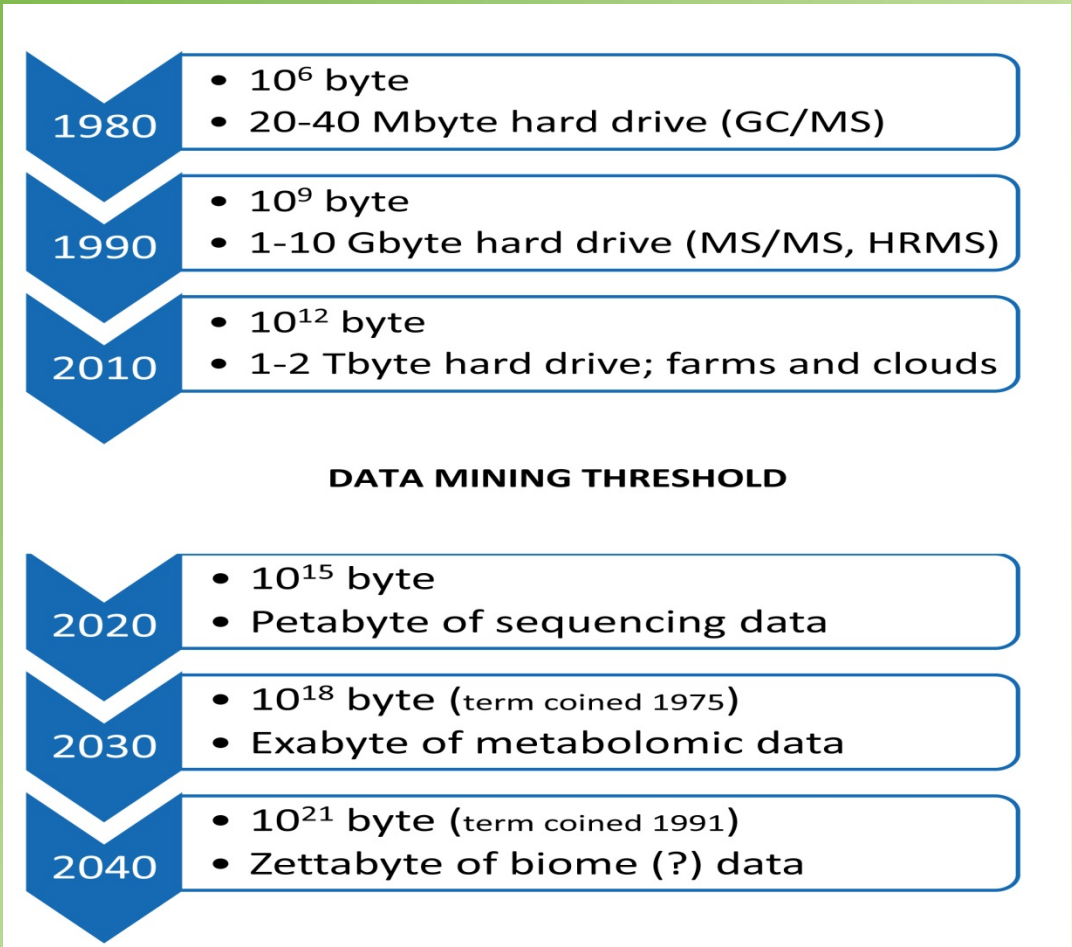
1. Digitized analog data in table of numerical values.
2. Values in table are plotted in graph.
3. Mathematical operations convert underlying numbers into useable form.

Publication or Report

4. Measurements between features become the data of interest.
5. Tables of this data are plotted to show trends.
6. Statistical calculations lead to results.
7. Results combine to produce conclusions.

Data scales in mass spectrometry

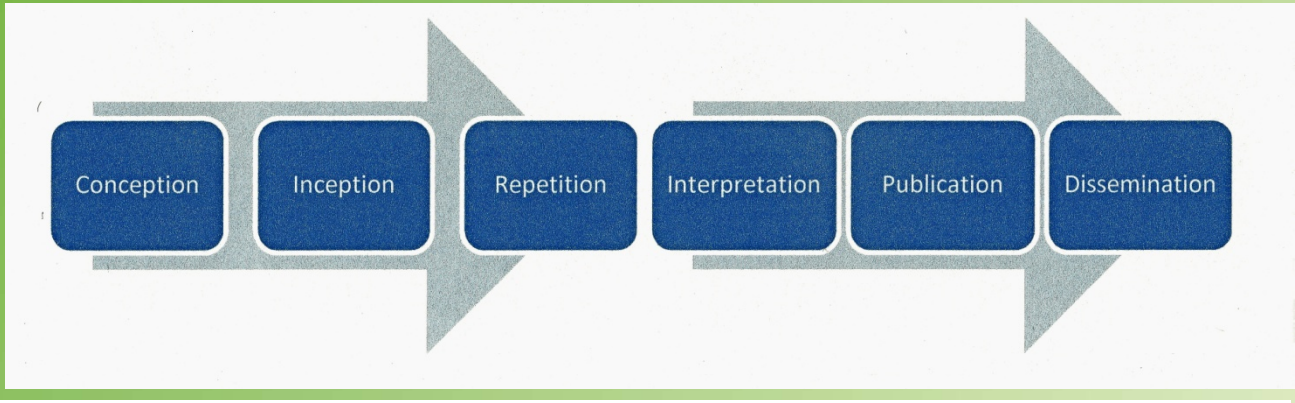
(Mass Spectrometry Forum, *Spectroscopy*, January 2012)

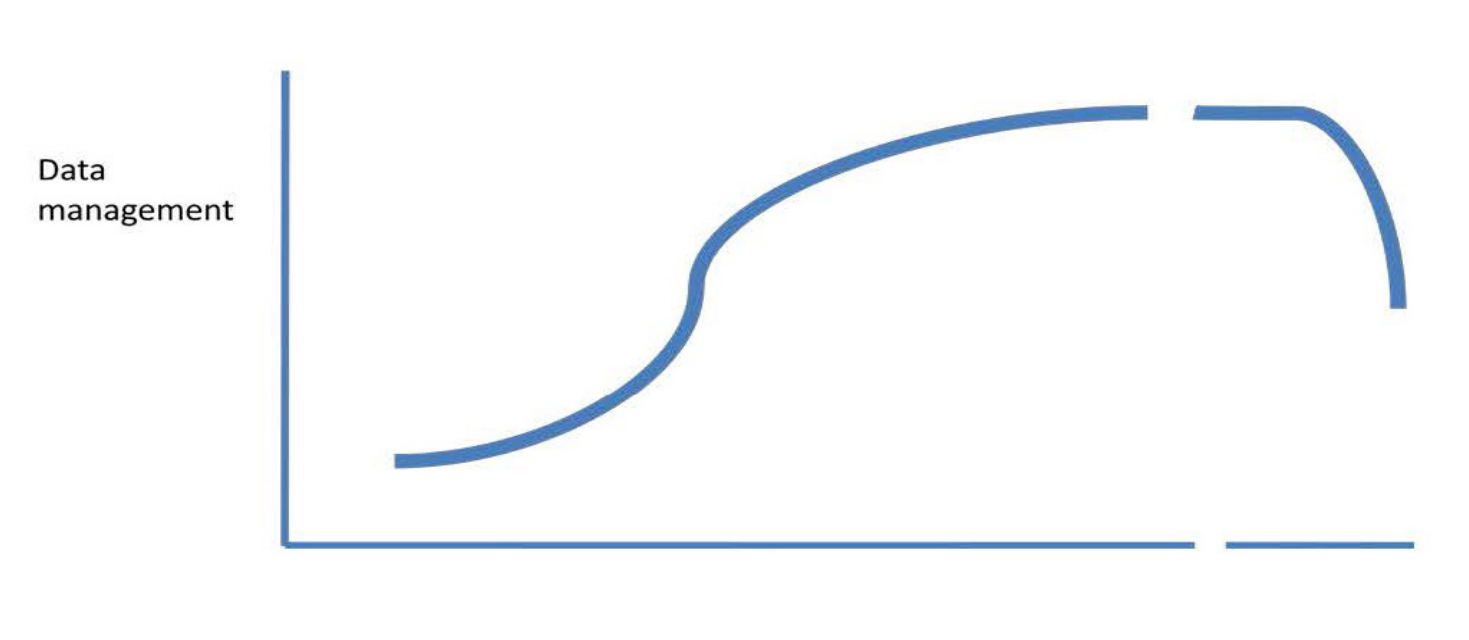


Data flow in mass spectrometry

(and everything else, probably)

(Mass Spectrometry Forum, *Spectroscopy*, January 2012)





The research record and RM Investigations

- Wilson et al. report that at least 25% of RM investigations at the university level encounter significant problems related to the research record (K. Wilson, A. Schreier, A, Griffin, and D. Resnick, "Research Records and the Resolution of Misconduct Allegations at Research Universities," *Accountability in Research, 14:1, 57-71* (2006).
- At NSF OIG, research record issues arise in investigations of allegations of fabrication and falsification (and more rarely in plagiarism issues), and in evolving issues of data sharing and data management

Case 1

- Fabricated replicate data was reported in a table appearing in a publication

- The fabricated data was created in Excel (through simple multiplication of the first data set) by a postdoc under time and effort pressure
- The data was never carefully examined by the senior author of the publication, who was the the PI of the NSF award
- The publication was retracted, and the postdoc fired by the grantee and ultimately debarred by NSF. The grantee promised better oversight of research records and RM training.
- **BUT** a subsequent case revealed that the grantee made no substantive changes to its data management practices or RCR training.

Case 2

- A graduate student supported for a full four years on an NSF graduate fellowship manipulated date in Excel to support favored hypotheses

- The student was quite collaborative. When he became involved, somehow things "worked."
- Two weeks before his Ph.D. defense (and after multiple publications), his research advisor finally examined some of the original data.
- The university investigated (and notified ORI but not NSF). The university supported its finding of research misconduct primarily by the student's admission, because the relevant data was solely in the possession of the student, and there was no provenance.
- NSF debarred the Subject. The PI asserted that his trust in the student was misplaced, that he checked data to the best of his ability, and he is giving **serious consideration** to some change in his laboratory practices.

Case 3

- A reader of a scientific publication noticed apparent data fabrication in the supplemental data, and contacts the journal
  - The journal contacts the PI and corresponding author, who had never reviewed or examined the original data (contrary to specific journal policy)
  - The PI begins an assessment himself, in apparent contradiction of university policy. One of the authors (a graduate student who had since graduated) conveyed the issue to the university, which starts an investigation. The PI does not cooperate fully with the investigation, and leaves the university
  - The university investigation committee completes its work without ever examining the data, relying instead on the admission of the graduate student
  - **Eight** publications are retracted. The university made no RM findings against the PI or the graduate student The PhD. Dissertation of the graduate student may be in jeopardy
  - In public statements, the grantee university does not address any changes or commitment to its data management or research misconduct policies

Case 4

- An NSF grantee was never in possession of a research database that was instead held in the sole possession of the Co-PI at another institution. An independent researcher requested a copy of the data directly from the Co-PI

- The Co-PI refused to share the data with the researcher, claiming it was still being used as a research result for potential publications
- After a "reasonable" time, the researcher requested the data from the grantee institution. But the grantee did not have the data, and in fact, had never examined it
- The Co-PI claims that the data was eventually lost in a personal computer crash, and admitted the loss was due to his own negligence
- The now-lost data was previously used as evidence in controversial Congressional testimony
- The data management responsibilities of collaborators are proscribed by the **policies of the lead institution** (collaborative awards vs subcontracts)

Case 5

- A former university postdoc used data which he helped produce under an NSF award, and which was in a public database, in a subsequent publication for which there was an authorship dispute, and an interpretive difference of opinion

- One of the postdoc's former collaborators, still at the grantee university, made an allegation of research misconduct (based on authorship) against the postdoc
- After an internal investigation, the university asserted that it "owned" the data, and convinced the journal to retract the publication over the objections of the author

- **Is there a basis for an NSF finding or action?**
- **Was there a basis for university action?**

Case 6

- Blatant image manipulation was evident in an online research publication that acknowledged NSF and NIH support (discovered by readers and not by reviewers)

- The publication was retracted; the university began an investigation.
- The publication's three authors were a postdoc, an undergraduate student, and the PI of an NSF award. Only the postdoc was involved with the creation of the image. He denied any improper manipulation, and soon left the university.
- The PI never examined any of the data despite the data management plan submitted to NSF with the original proposal that promised that he would do so.
- The university asserted that neither NSF nor NIH funds were used to support this particular research, despite the published acknowledgments.
- The grantee university took no action against the postdoc or the faculty member.

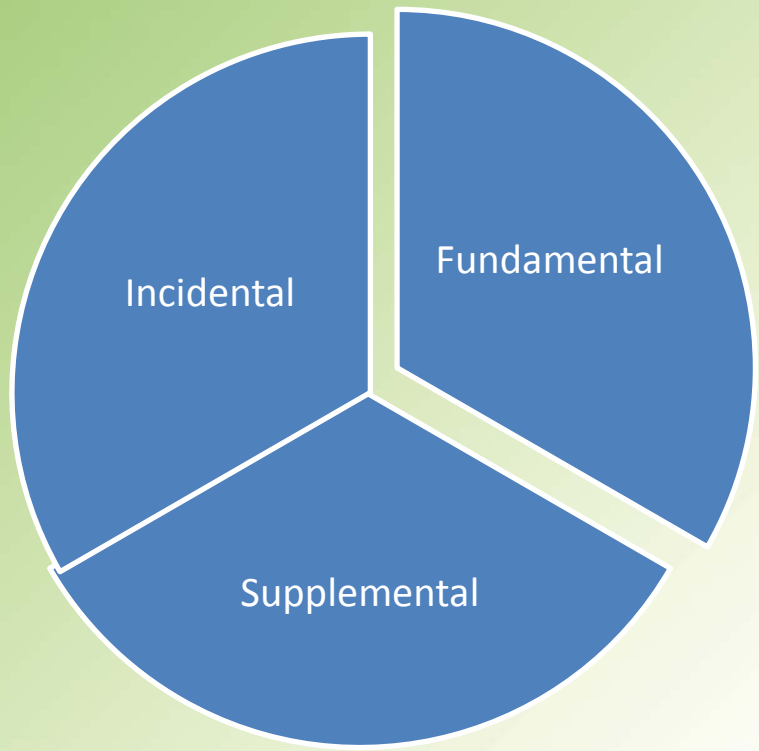
- **Is there a basis for an NSF finding or action?**

Data Domains

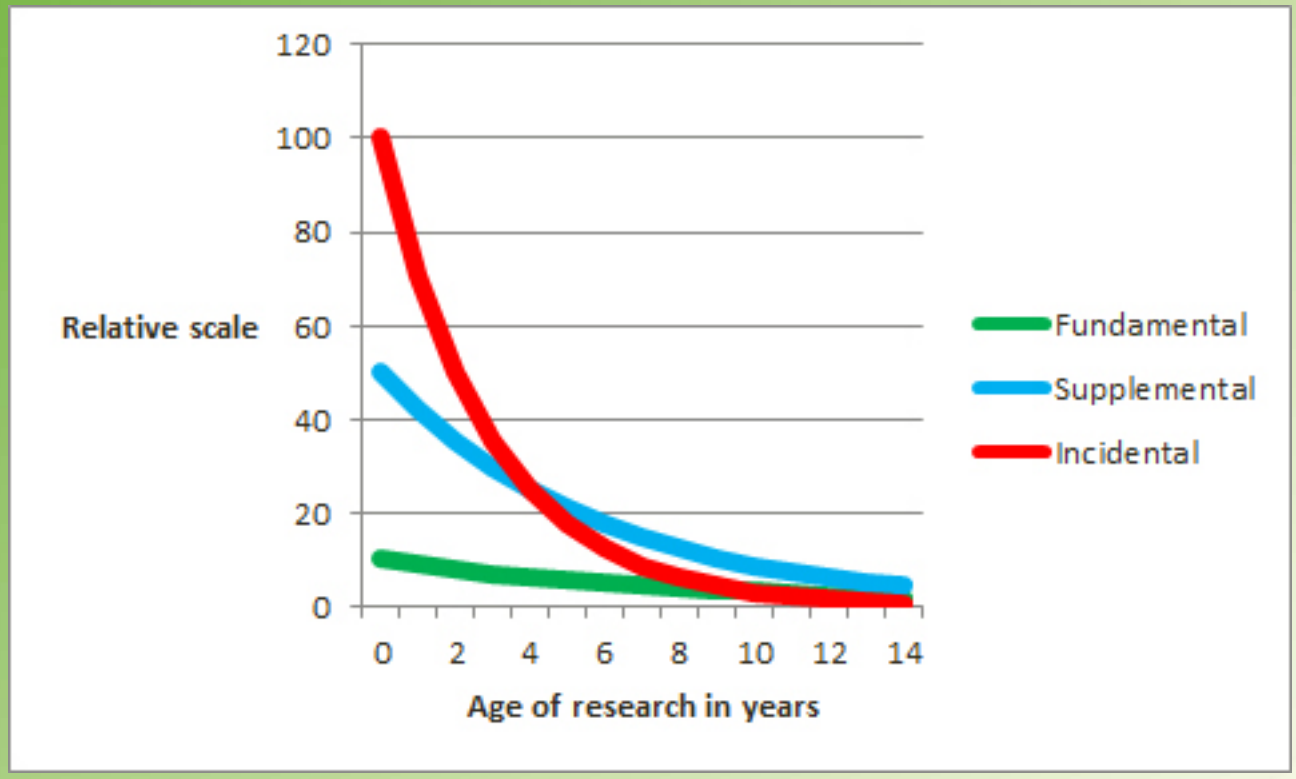
- The research record is populated with three types of data
  - Fundamental data
    - Primary data that decisively supports the conclusion
    - Replicated, confirmed, assessed, and archived
  - Supplemental data
    - Secondary data that corroborates a research conclusion
    - Confirmed, assessed, and archived
  - Incidental data
    - Concurrent but non confirmed data
    - Exploratory data

Data Domains

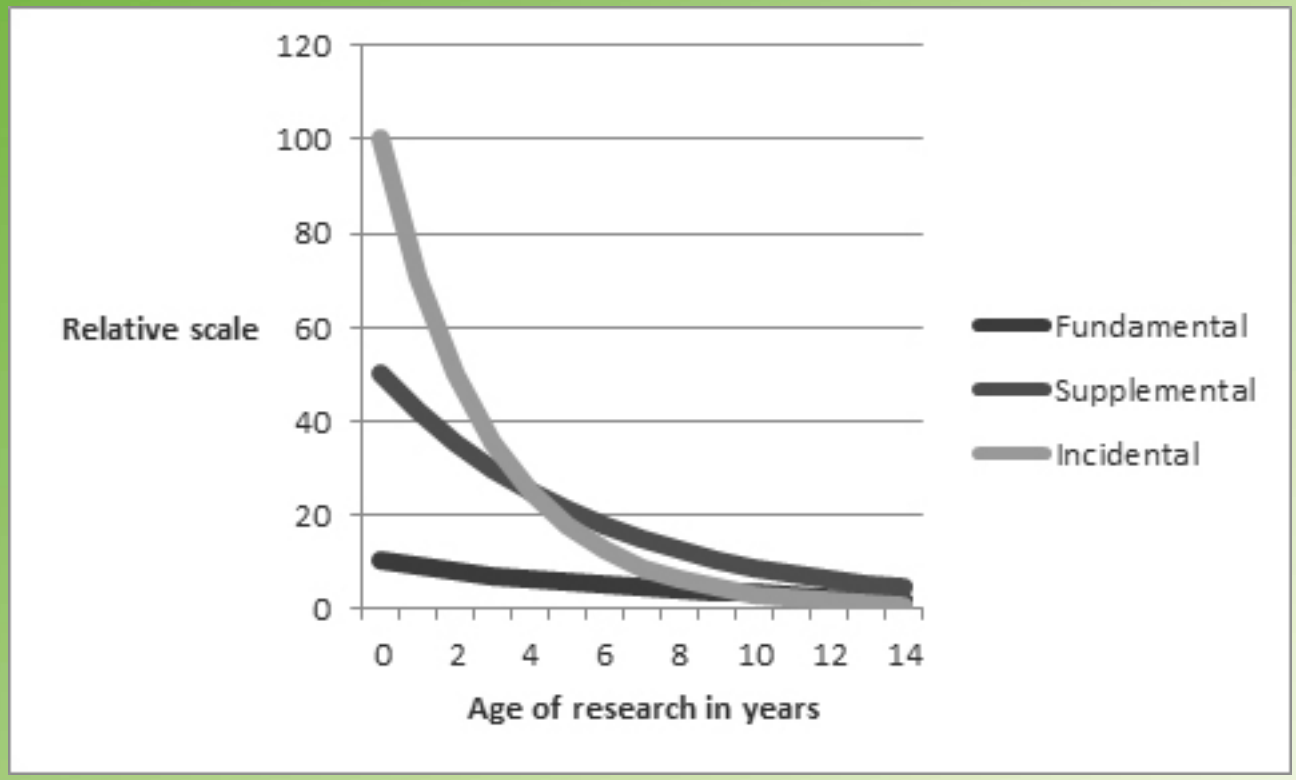
- Fundamental data
  - Scale = 10
  - Half-life = six years
- Supplemental data
  - Scale = 50
  - Half-life = four years
- Incidental data
  - Scale = 100
  - Half-life = two years



Data Domains



Data Domains



After a few years, distinction between significant and non-significant data becomes more difficult, and plays against common timing for receipt and resolution of allegations.

DATA POLLUTION

Areas of Concern for Data Management with Implications for NSF Investigations

Subjects (at least the ones who are caught) tend to use (Excel, Photoshop) in non-sophisticated ways. And when they get better?

Even non-sophisticated fabrications/falsifications get past those who should have responsibility for oversight and assessment of the data (e.g., PIs who do not examine data). Trust but verify.

Areas of Concern for Data Management with Implications for NSF Investigations

Many of the traditional methods for ensuring the integrity of data—whether universal or discipline specific—are being modified as digital technologies alter capabilities and procedures. Because of the huge quantities of data generated by digital technologies, an increasing fraction of the processing and communication of data is done by computers, sometimes with relatively little human oversight. If this processing is flawed or misunderstood, the conclusions can be erroneous. Documenting work flows, instruments, procedures, and measurements so that others can fully understand the context of data is a vital task, but this can be difficult and time-consuming. Furthermore, digital technologies can tempt those who are unaware of or dismissive of accepted practices in a particular research field to manipulate data inappropriately.

National Academy of Sciences (2009): Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age
http://www.nap.edu/catalog/12615.html

Areas of Concern for Data Management with Implications for NSF Investigations

From the grantee and community perspective

- Robust and enforced data management plans
- Cohesive data plans for collaborative efforts
- Reasoned plans for data archiving and its costs
- Coherent plans for residual data
- Integrated data management policies
- Certification of CVs and collaborations

New “stories” every six months

kbusch@nsf.gov