Research Integrity: The Importance of Data Acquisition and Management.

Randall Reed, Ph.D.,
Director, Center for Sensory Biology
Professor, Department of Molecular Biology & Genetics, Neuroscience, and Otolaryngology, Head & Neck Surgery

Disclosures

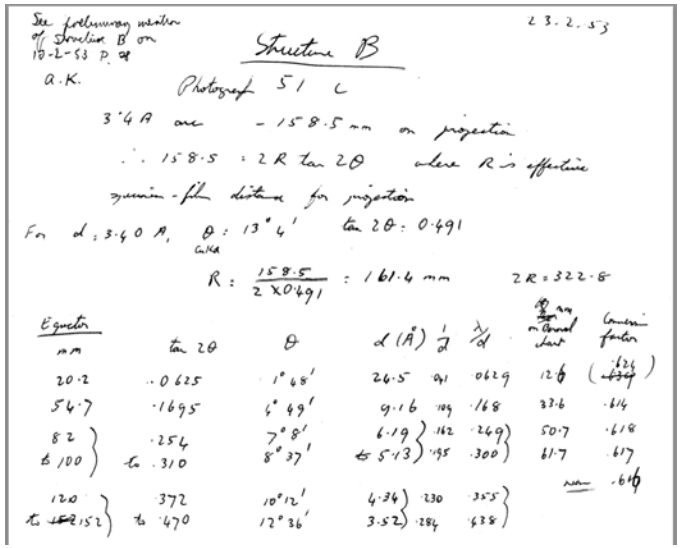
No Relevant Financial Relationships
with Commercial Interests

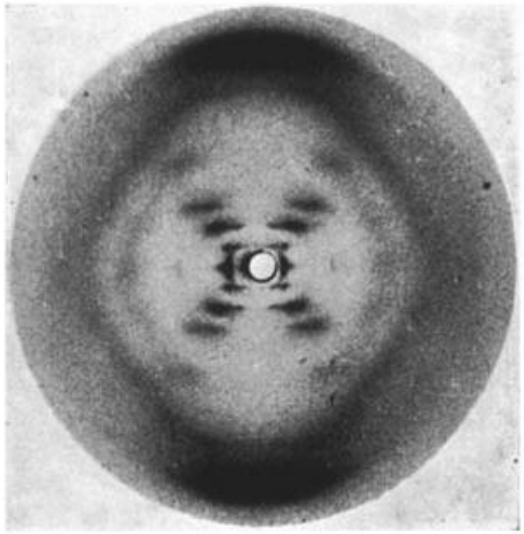
Data Acquisition and Management

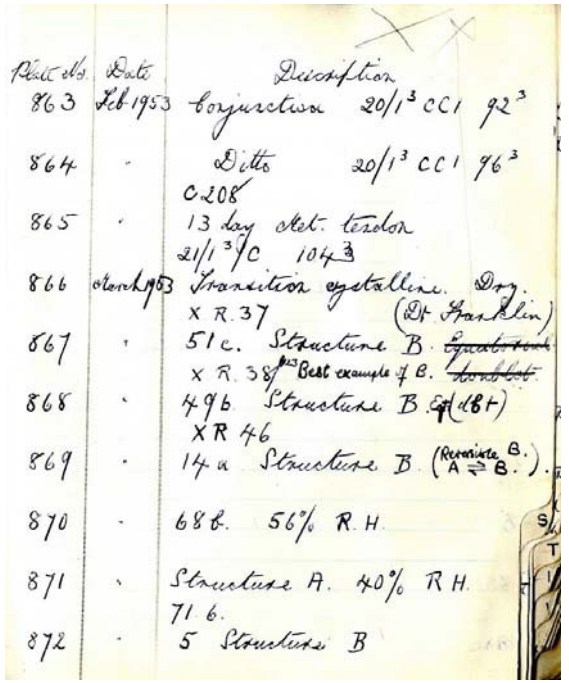
- The *Original Record*- What is it?
  - Contemporary Challenges of Primary data – *Nature/Volume*
- The Pathway to Publication
  - Perspectives from the participants
  - Data processing – *Wrongs and Rights*
- Data Retention and Organization
  - Requirements for *Retention*
  - The importance of *Organization*

Primary Data – A look back

- The Structure of DNA – Rosalind E. Franklin





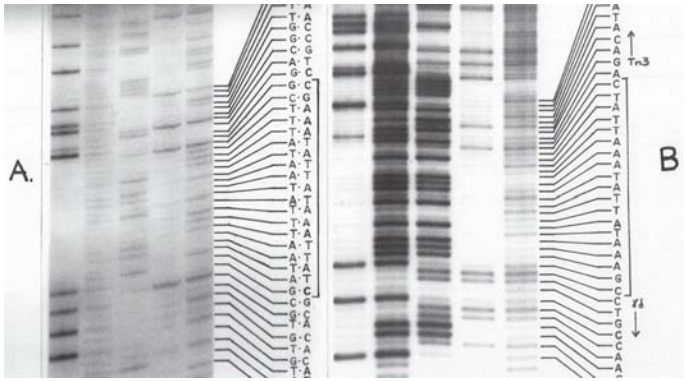


These reflections due to the bases. R.E.F. missed this. A.K.

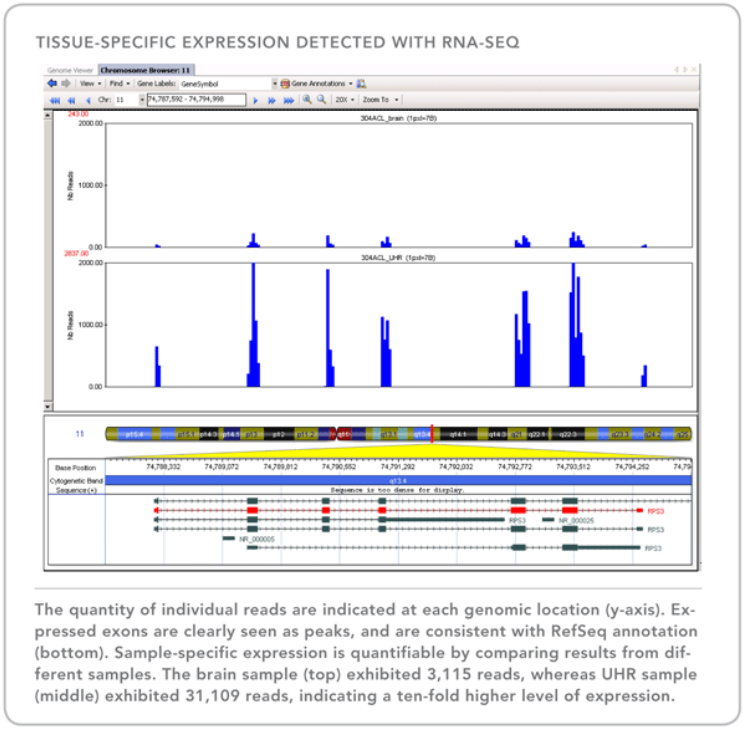
R.E.F. is at last making the correct connection between the A and B. A.K.

Data Explosion(s)

Genetic and Genomic Information



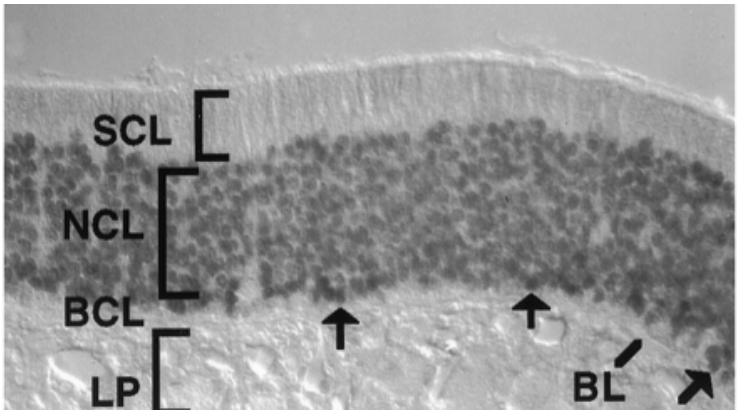
R. Reed PNAS 1981

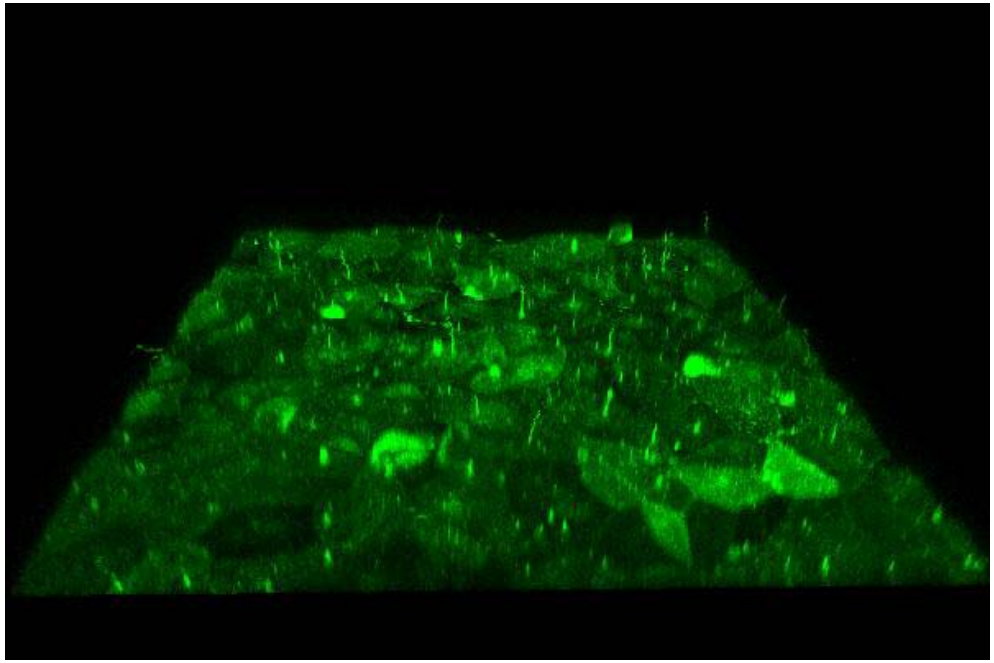


Illumina.com

Data Explosion(s)

Imaging and Visualized Information





The sheer **Volume** of data that supports/incorporated into publications has accelerated by many orders of magnitude!!

Similar Data Explosions are Occurring in almost all areas of Biomedical Research

- Complex multi-modal databases of Human Subjects information in Clinical Studies
  - Consider impact of "*Precision Medicine Initiative*"
    - a database of a 1 Million Patient Cohort

- Recordings of functional activity in brain by fMRI and multi-electrode

- Microbiome genomic information

The Current Reality -

- Bench/Bedside Biomedical Scientists responsible for data collection are more detached from the *Primary Data* underlying their research

- More analysis is done on a representation(s) of *Primary Data*

- PIs are especially distant because of volume and representations

Questions for Discussion:

- What is “Primary Data”?
- Do I really need to keep it all?
- Do I really need to use it all?

Data Acquisition and Management

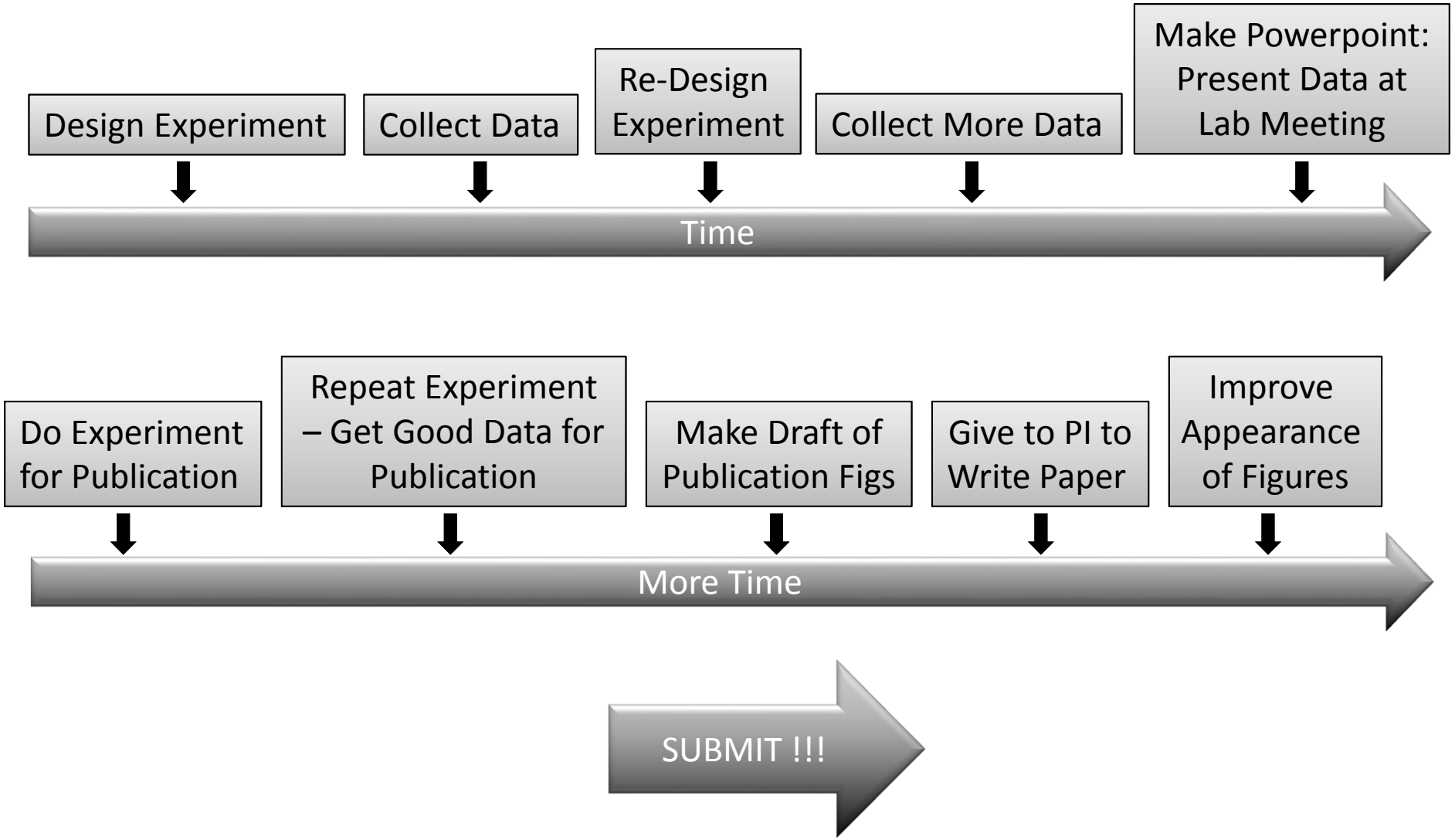
- The *Original Record*- What is it?
  - Contemporary Challenges of Primary data – *Nature/Volume*

- The Pathway to Publication
  - Perspectives from the participants
  - Data processing – *Wrongs and Rights*

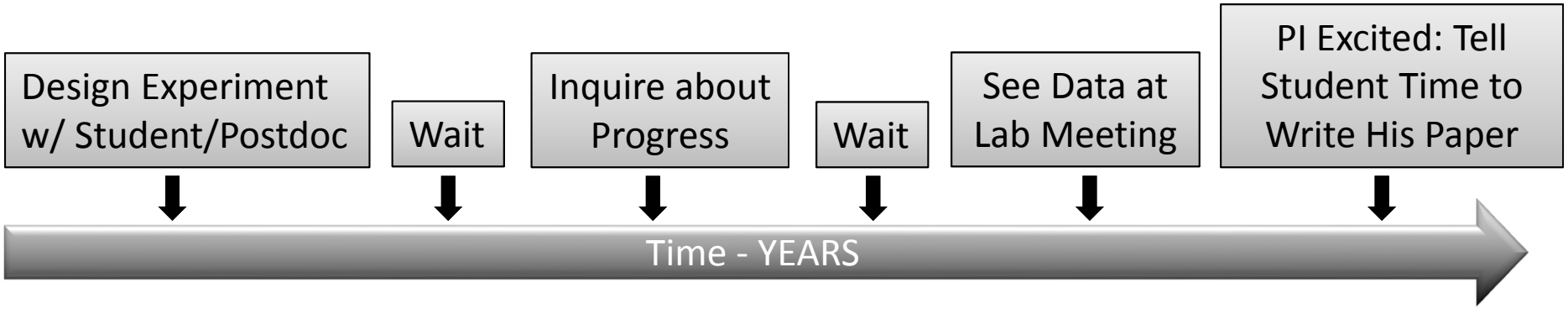
- Data Retention and Organization
  - Requirements for *Retention*
  - The importance of *Organization*

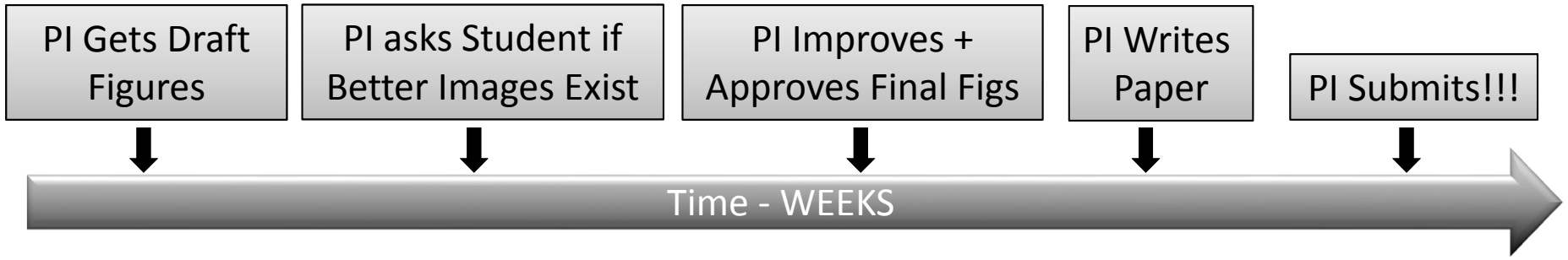
A Student's Perspective on Publication:

A Time Line….



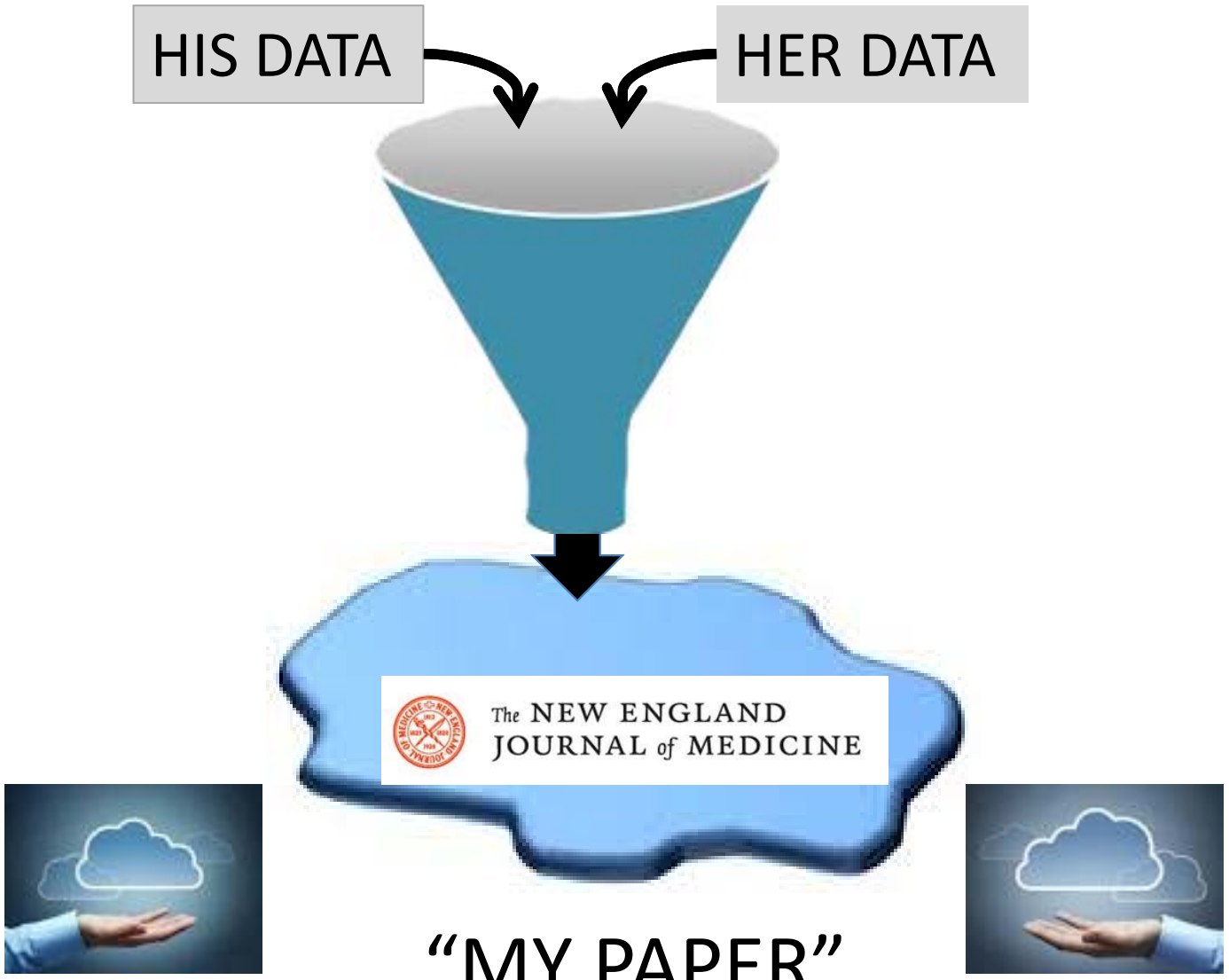
A PI's Perspective on Publication:





What's Wrong with these Scenarios??

Two Collaborators' Perspective on Publication:

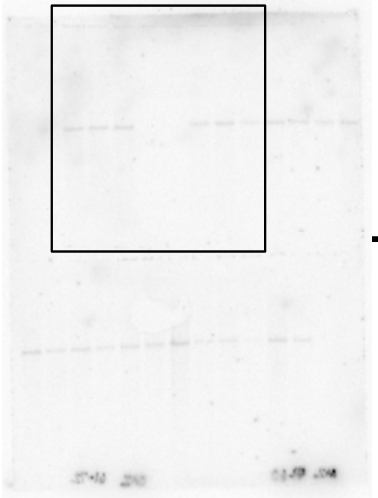


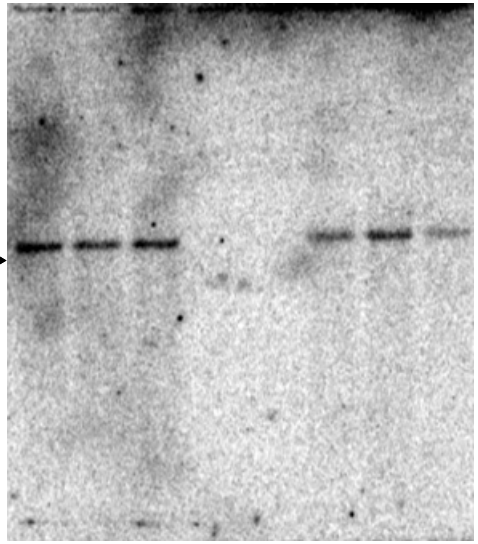
Data Processing and Manipulations

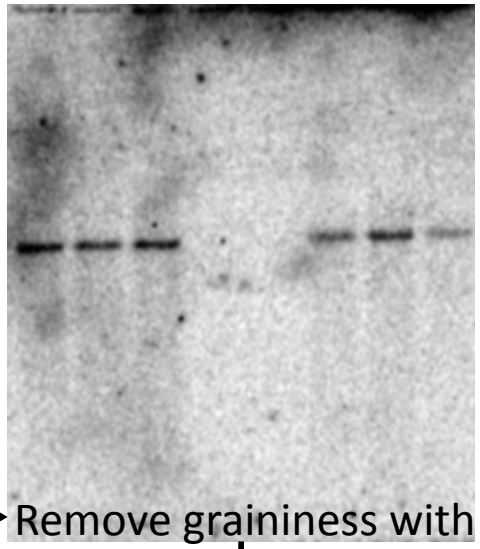
- Cases of *Fraud/Fabrication*
  - Probably more common than we appreciate
  - Still relatively rare
  - Recent changes are improving detection
- Excessive/Inappropriate Data manipulation
  - Done to “assist” the reader
  - Eliminate “distractions”
  - Make “Best ever” data appear “typical”

Data Processing: *Wrongs and Rights*

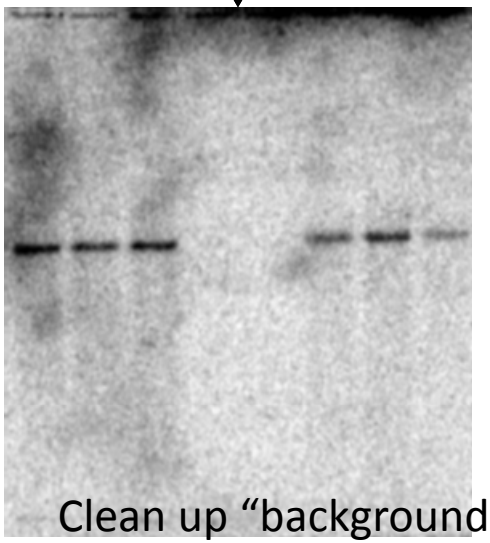
Original image



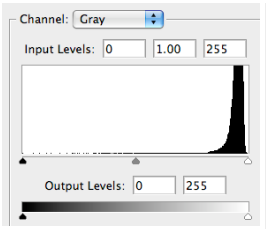


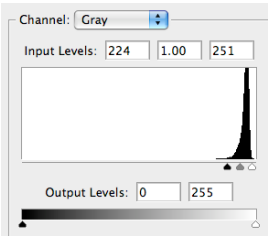


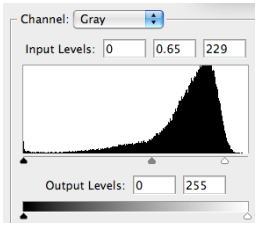
→ Remove graininess with blur



Clean up “background”

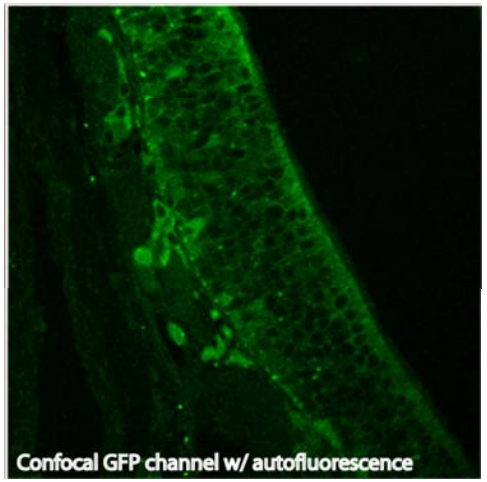






original → Adjust levels and contrast

Automated Image Analysis and Reconstruction



Traditional image

Automated “Pipelines” Exist for Many Analyses

- Raw Data Enters
- Manipulated for Uniformity/Conformity
- Filtered for Accuracy
- Processed for Representation
- Displayed/Analyzed

“A Pipeline is a long round black-box with holes at each end”

Evolving Standards for Data Presentation

- Journals have promulgated standards for acceptable image manipulations
  - not standardized across journals
  - Sometimes vague
  - Original data depository (Journal of Cell Biology)

- Journal Editors have become very proactive in "image forensics"

Data Acquisition and Management

- The *Original Record*- What is it?
  - Contemporary Challenges of Primary data – *Nature/Volume*

- The Pathway to Publication
  - Perspectives from the participants
  - Data processing – *Wrongs and Rights*

- Data Retention and Organization
  - Requirements for *Retention*
  - The importance of *Organization*

The Essentials of Data Retention

- For most NIH Grants need to retain Data for:
  - 3 years after filing of final reports at end of grant
  - Longer for clinical studies – even longer in some circumstances
- Data is retained by Institution *but… Responsibility of the PI for Compliance*
  - Accessible to the PI and other authors
    - Free to make copies
  - All **Original Data** must stay with host Institution
    - Includes notebooks, raw files, associated records

Requirements for Data Retention

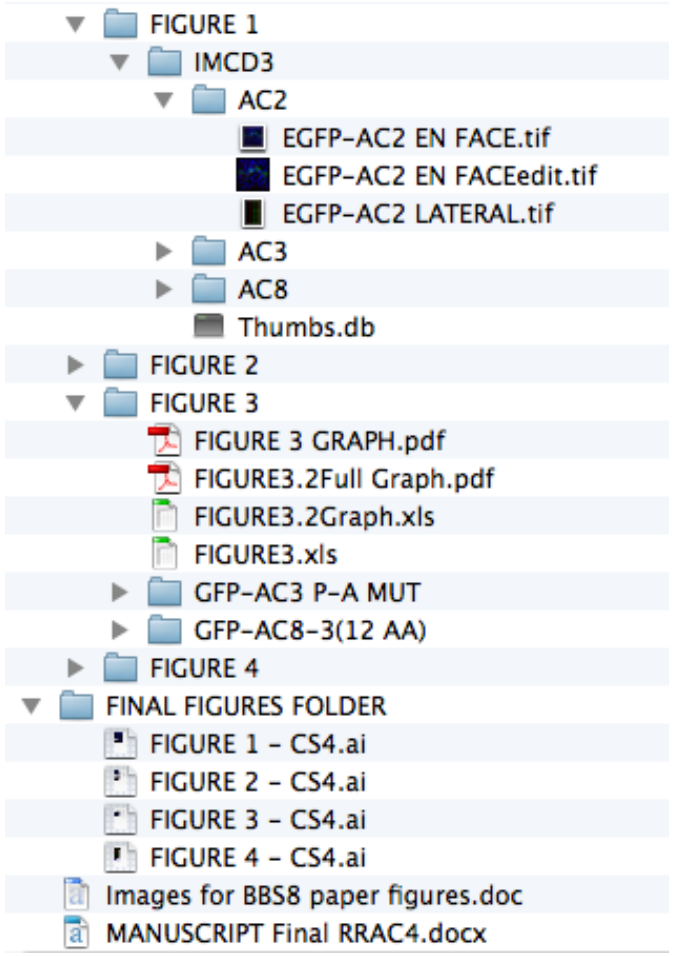
- Primary Data needs to be stored in a secure place
  - Not on a laptop or a USB memory stick
  - Permanent storage
  - Backed up
  - Secured against theft and tampering
- Data needs to be *Organized* to be effectively managed

Possible Solutions for Data Retention

- Laboratory/Group managed Central Server
- Institutional Data Maintenance, Back-up and Storage
  - JH IT $2500/TB per year
- Cloud Solutions
    - Amazon Glacier - 120$/TB per year <1GB minimum

Some Universities Evaluating Centralized Mandatory Institution-Supported Archiving…

Organizing Data from Acquisition to Publication



- Assemble raw data and finished figures in one location
- Retain steps in editing figures as they are assembled
- Prepare tables that indicate the permanent storage location of the original data including all identifiers
- **Compare final figure to assess whether is it an accurate representation of original data**

Organizing Data from Acquisition to Publication

An example - *a word document with the following info:*

Animal IDs and images used for figure 4		image size (microns)	scale bar
(AcTub wt)	AD994 AcTub single a.tif	92.1	10um
(AcTub KO)	AD996 AcTub single a.tif	92.1	10um
(i7/ac3 wt)	AD994 I7 AC3 proj b.tif	92.1	10um
(i7/ac3 KO)	AD996 I7 AC3 proj b.tif	92.1	10um
(i7/slp3 wt)	AD994 I7 SLP3 proj b.tif	92.1	10um
(i7/slp3 KO)	AD996 I7 SLP3 proj b.tif	92.1	10um

Original files located in D:\\reedlab\Images\BBS8

Allows *permanent* connection between published materials and the original source data!!

Best Practices: Data Acquisition and Management

- Data should be organized from earliest collection with cross-correlation regarding all identifiers
- Organization principles should be familiar to experimenter and PI.
- Everyone Involved in Publication should undertake Post-hoc review of final figure and original data and *ASK:*
  - *"Is this an accurate representation of our observations?"*
- All Data should be available to all authors on request

Panel Discussion

- Randal Reed, Ph.D.
- Sarah Wheelan, M.D., Ph.D.
- Irina Burd, M.D., Ph.D.
- Marquis Walker, Ph.D.

- Antony Rosen, M.B., Ch.B., B.Sc. - *Moderator*