Data Curation @ UCSB:
First year progress

Greg Janée

October 3, 2013

Background

- *Genesis*
  - Organized by Library, ERI, Office of Research
  - Funded by EVC's office
- *Goals*
  - Characterize data campus production processes and data curation practices
  - Identify curation needs, solutions, gaps
  - Give Library guidance on future staffing, services
    - (to address curation outside the Library)
- *Strategy*
  - Campus-wide survey
  - In-depth case studies

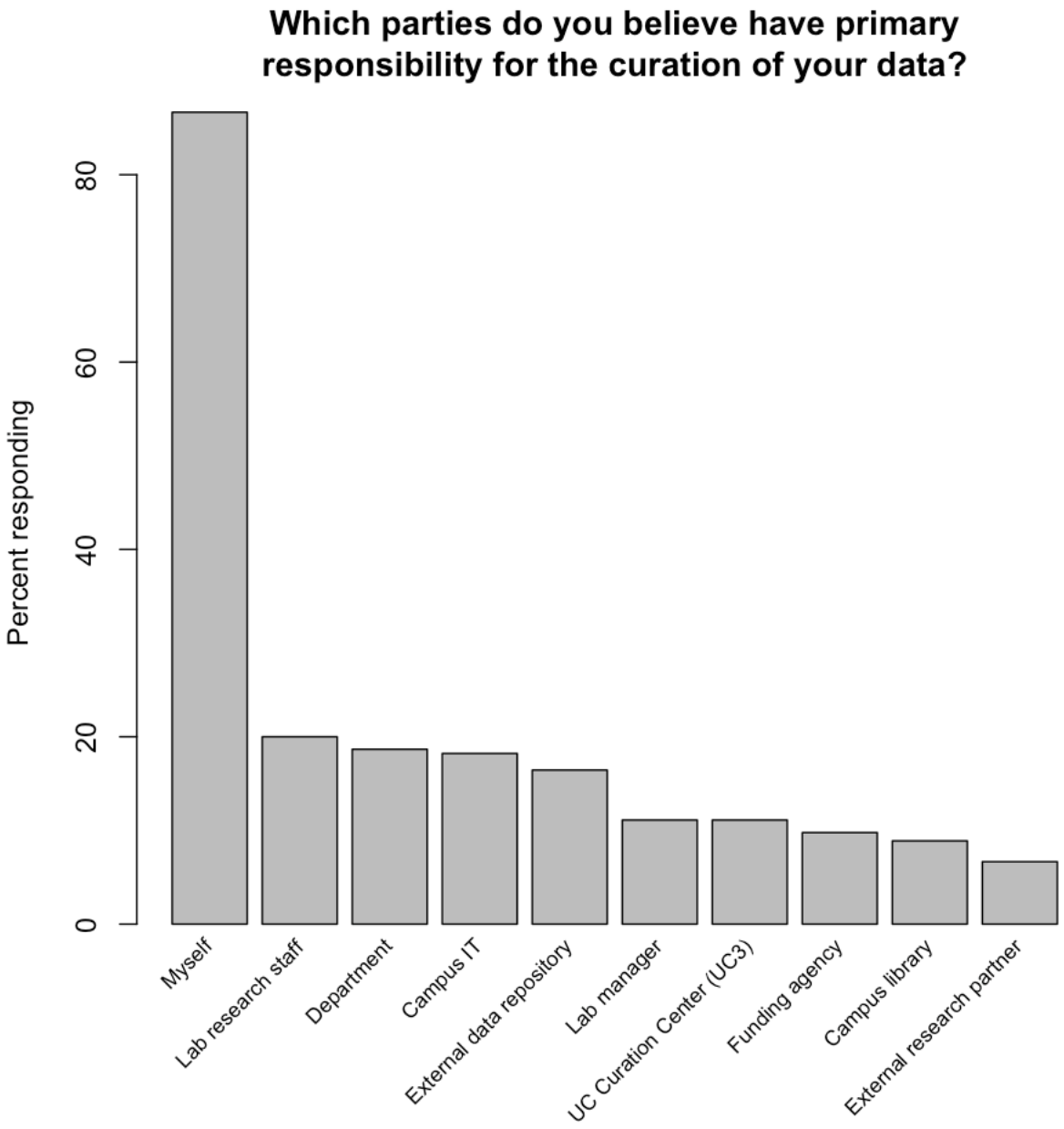
Survey

- *5 questions, 5 minutes*
  - In the course of your research or teaching, do you produce digital data that merits curation?
  - Which parties do you believe have primary responsibility for the curation of your data?
  - What data management activities could you use help with?
  - Are you mandated to provide for (or otherwise participate in) the curation of your data, and if so, by which agencies?
  - With which departments are you affiliated?

Survey response

- *294 responses*
  - 1/3 of estimated 900 faculty, researchers
  - Representing 90% of departments

- *Curation applicability question*
  - 77% answered “yes”
    - (all subsequent percentages relative to this population)
  - Up to 60% of all researchers on campus


Which parties do you believe have primary responsibility for the curation of your data?
Percent responding
0
20
40
60
80
Myself
Lab research staff
Department
Campus IT
External data repository
Lab manager
UC Curation Center (UC3)
Funding agency
Campus library
External research partner

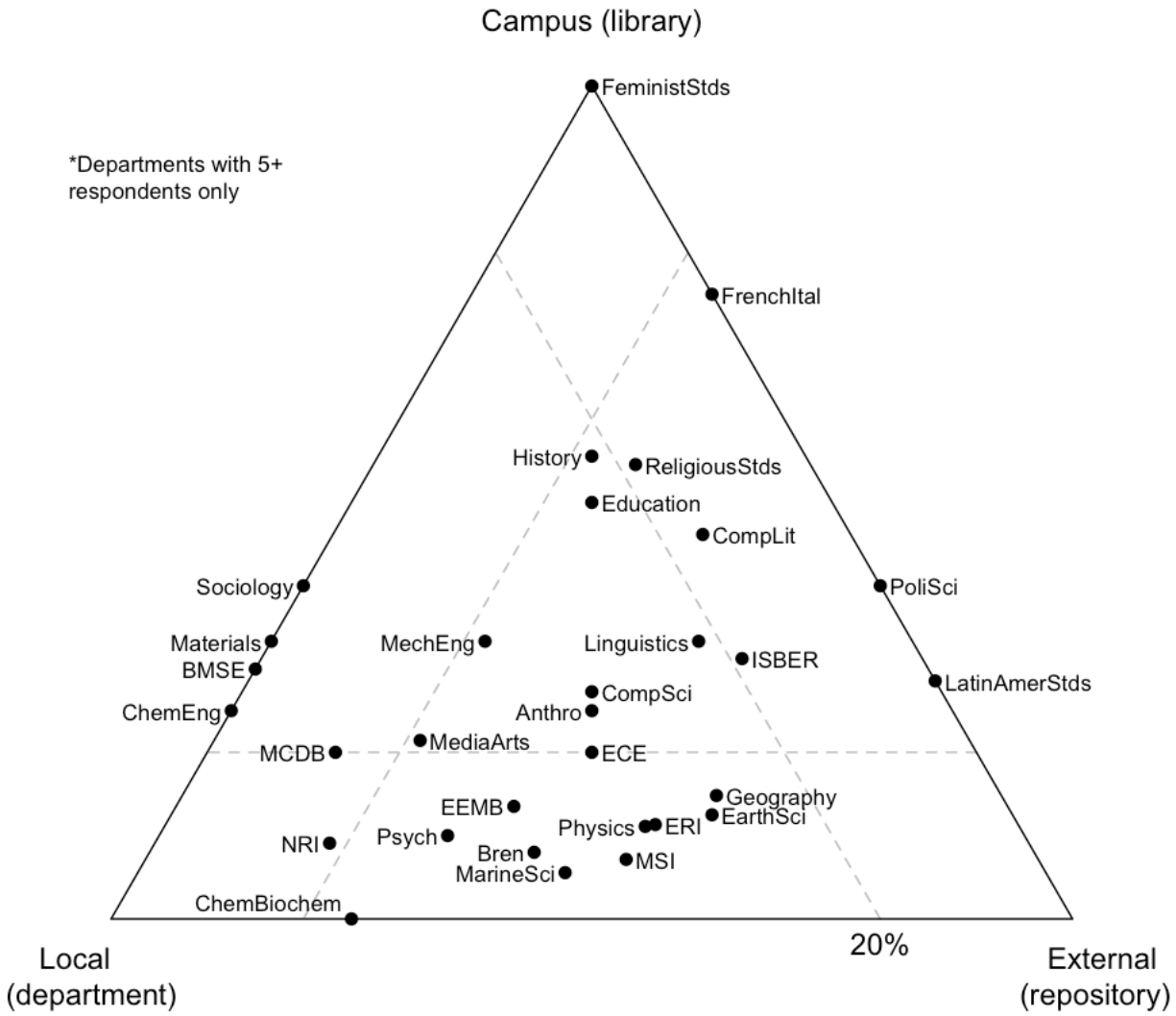
Other responses

- *Established solutions*
  - "Journals where we publish our data"
  - "Professional societies"
  - "Collections repository within my department"
  - "my e-mail group [...] has archives saved at UCSB"
  - "The research program with which the project is affiliated"
  - "publicity office"
- *Somebody else*
  - "PI / adviser"
  - "Co_PIs in my research group"

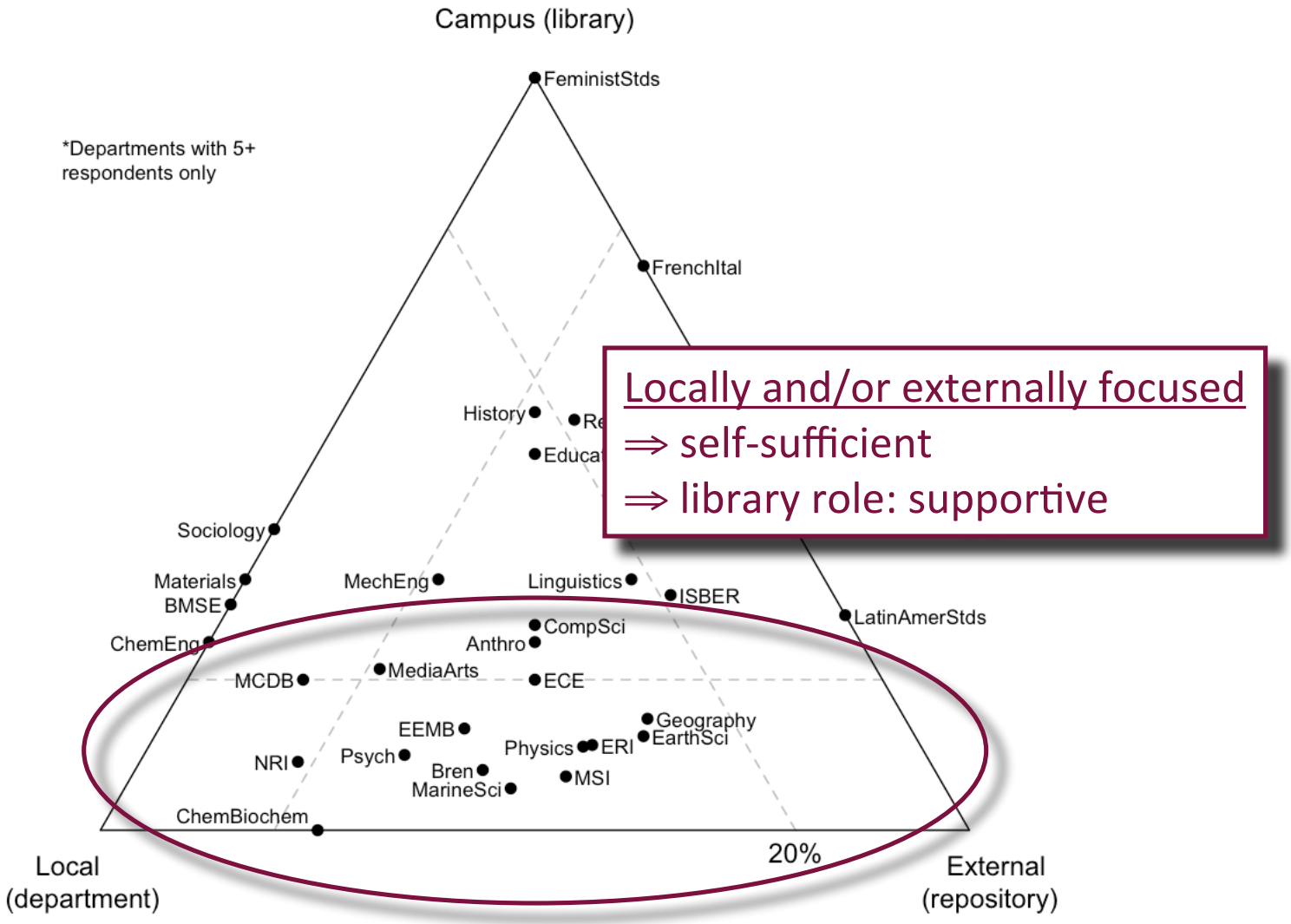
Responsibility spheres

- *"Local"*
  - Lab manager
  - Lab research staff
  - Department
- *"Campus"*
  - Campus IT
  - Campus library
- *"External"*
  - External research partner
  - External data repository
  - Funding agency
  - UC Curation Center (UC3)

Distribution of departments* with respect to responsibility spheres

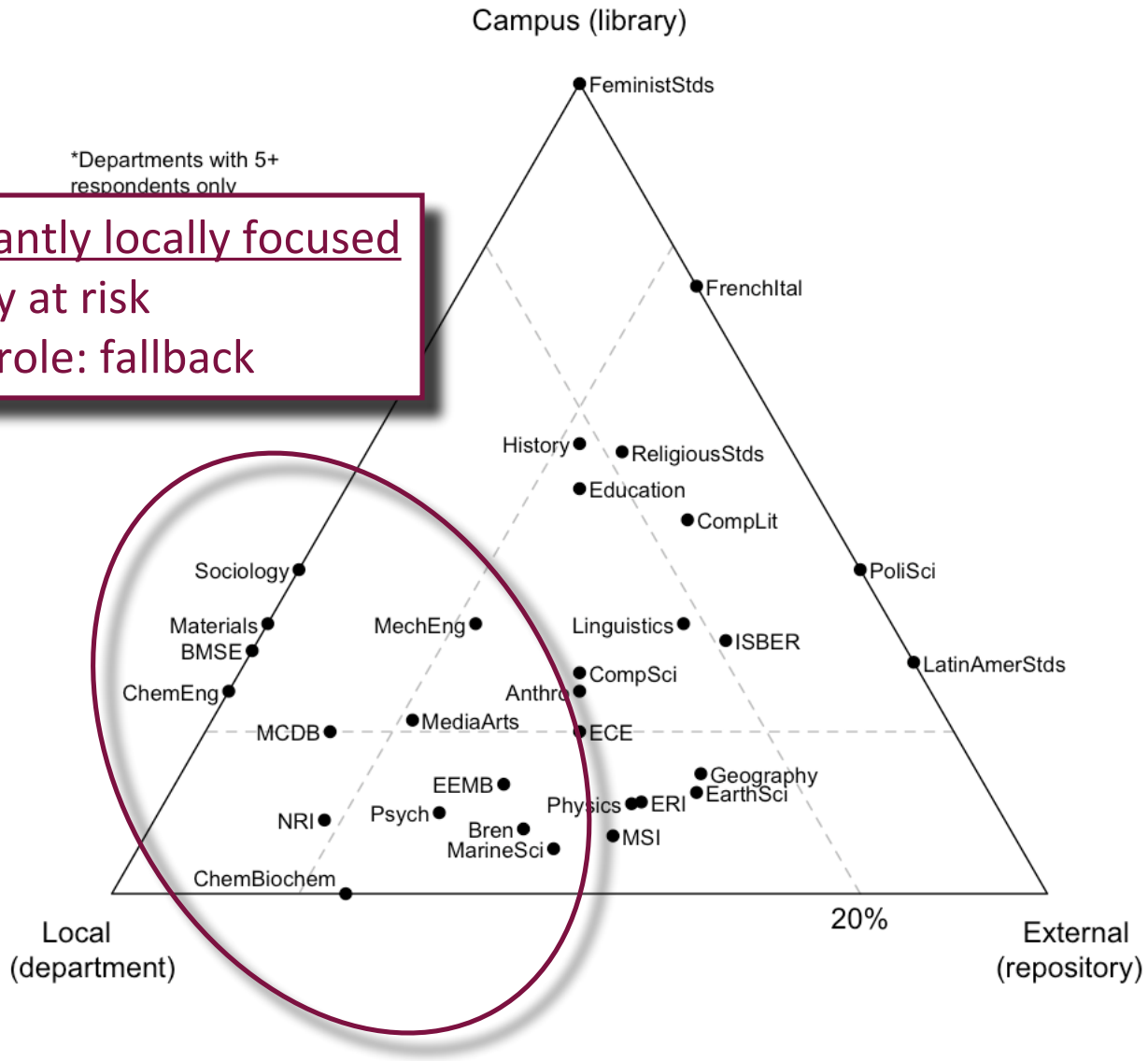


Distribution of departments* with respect to responsibility spheres



Locally and/or externally focused
⇒ self-sufficient
⇒ library role: supportive

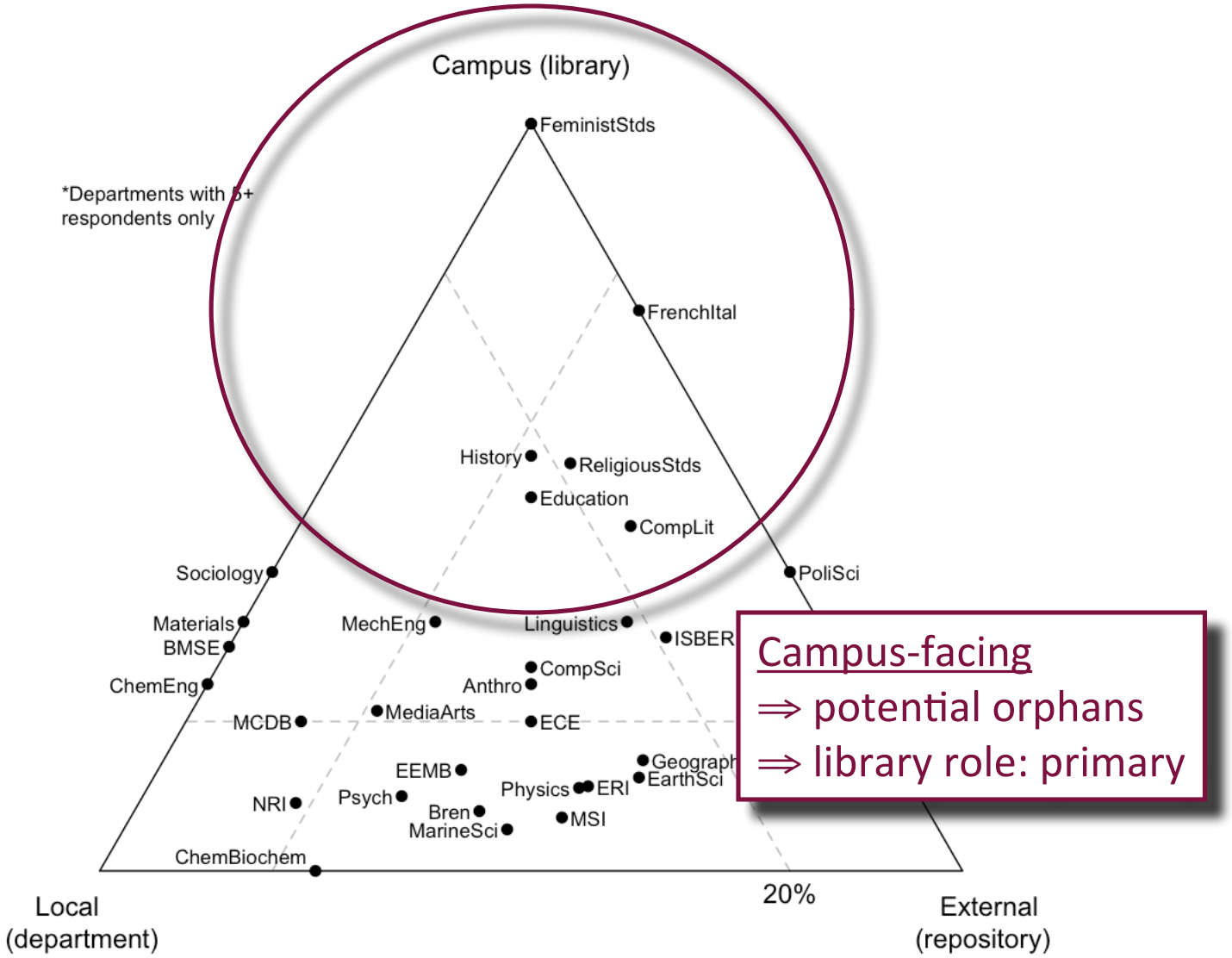
Distribution of departments* with respect to responsibility spheres



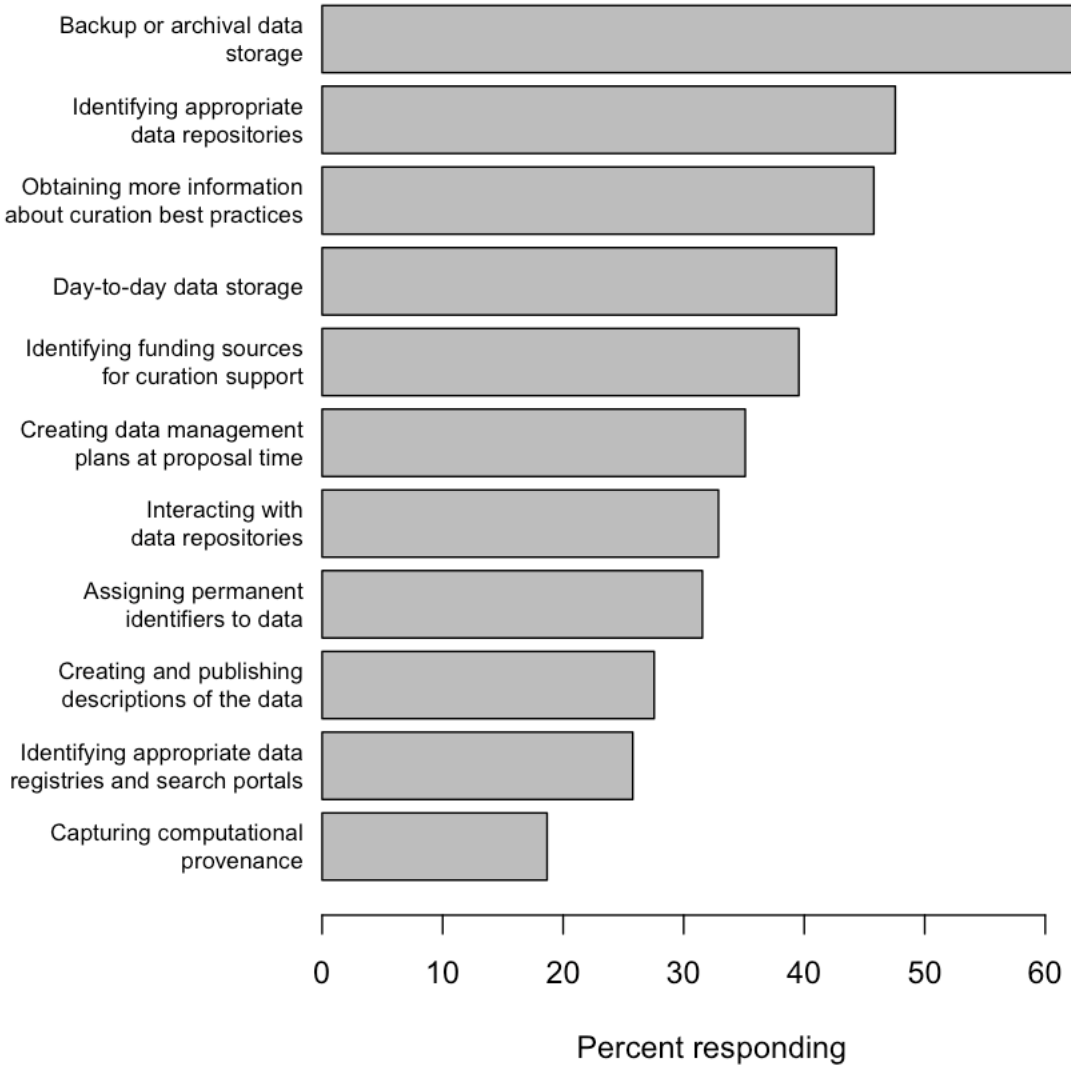
*Departments with 5+ respondents only

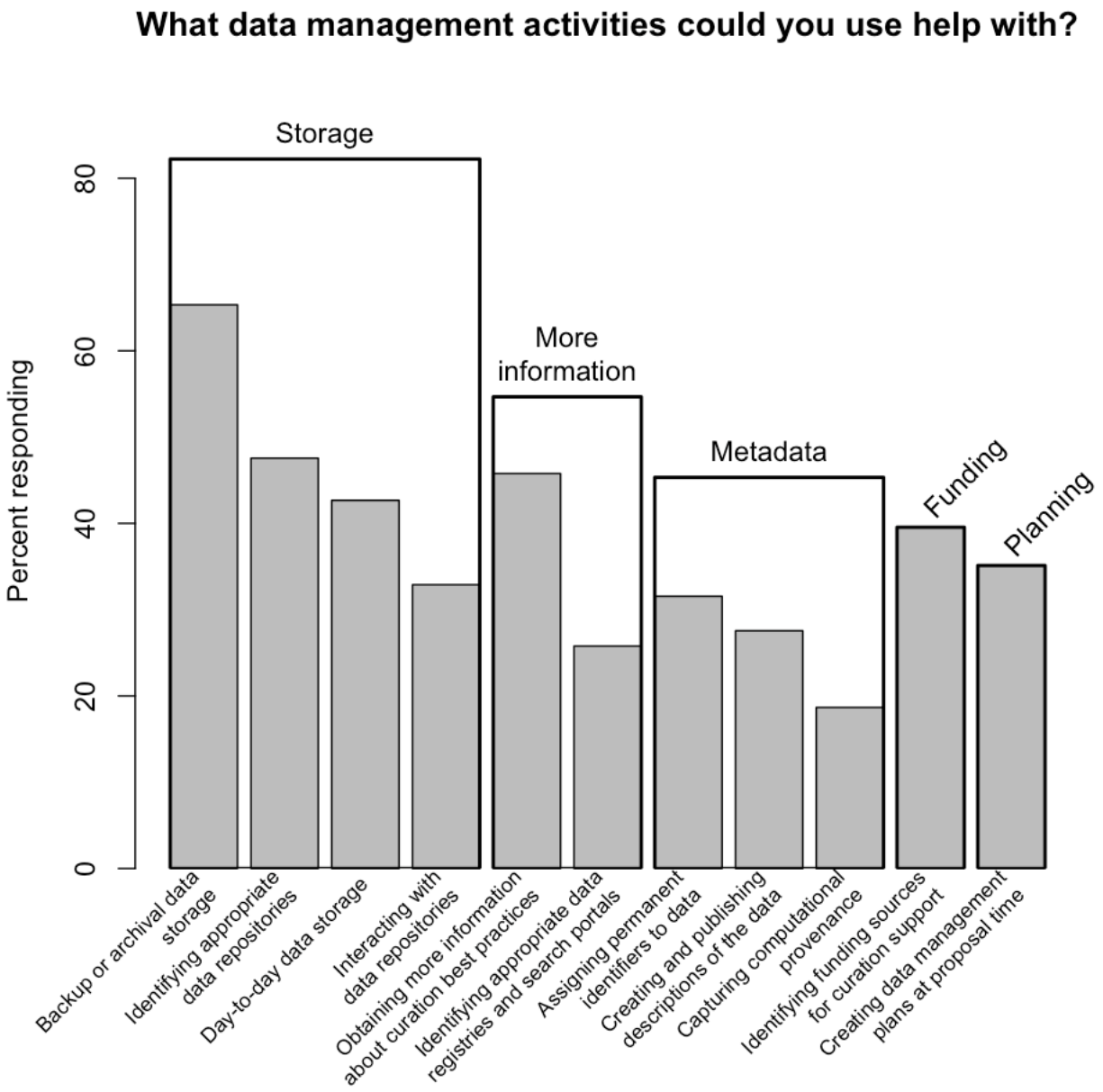
Predominantly locally focused
⇒ possibly at risk
⇒ library role: fallback

Distribution of departments* with respect to responsibility spheres



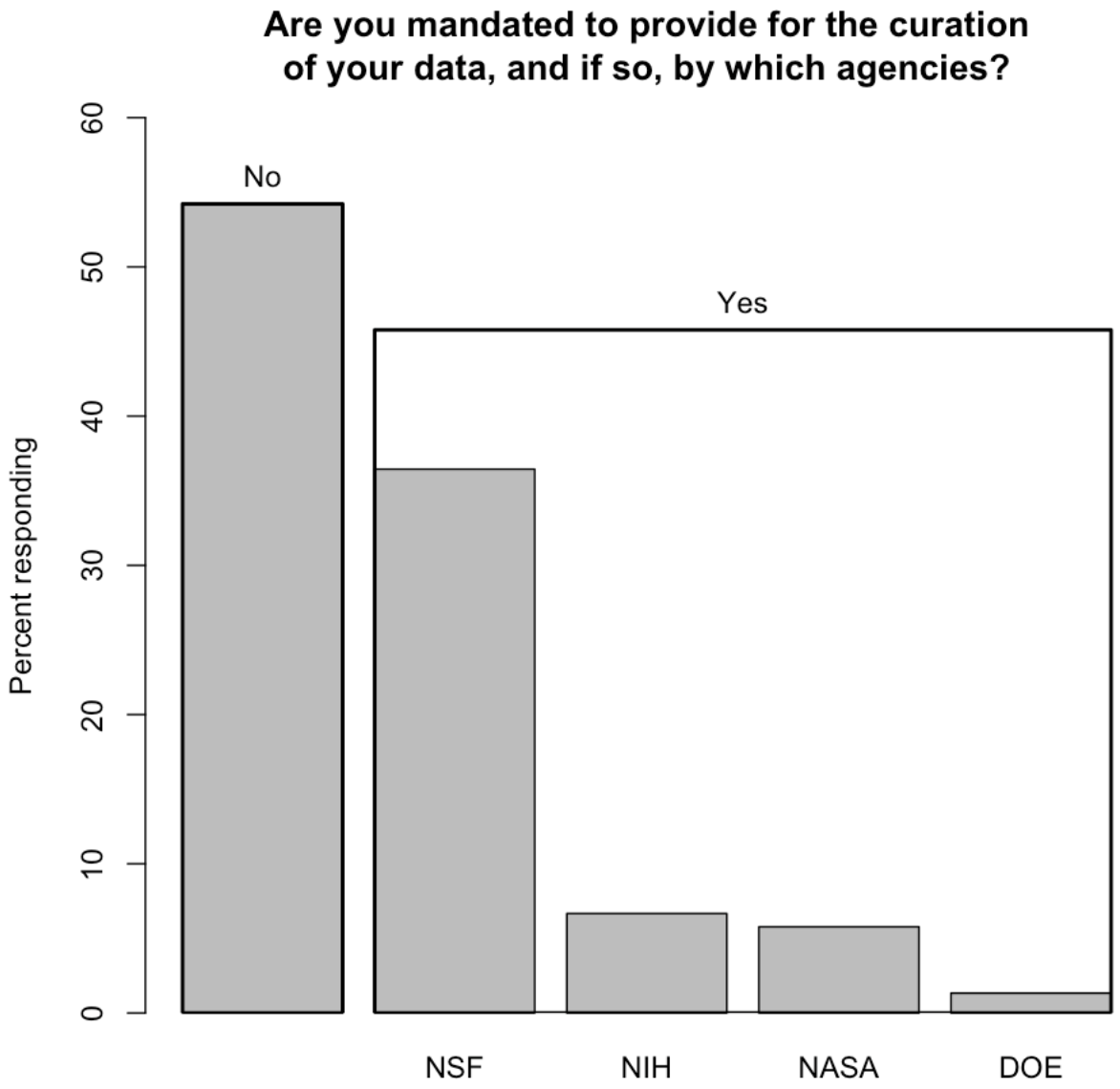
What data management activities could you use help with?




What data management activities could you use help with?
Percent responding
0
20
40
60
80
Storage
More information
Metadata
Funding
Planning
Backup or archival data storage
Identifying appropriate data repositories
Day-to-day data storage
Interacting with data repositories
Obtaining more information about curation best practices
Identifying appropriate data registries and search portals
Assigning permanent identifiers to data
Creating and publishing descriptions of the data
Capturing computational provenance
Identifying funding sources for curation support
Creating data management plans at proposal time

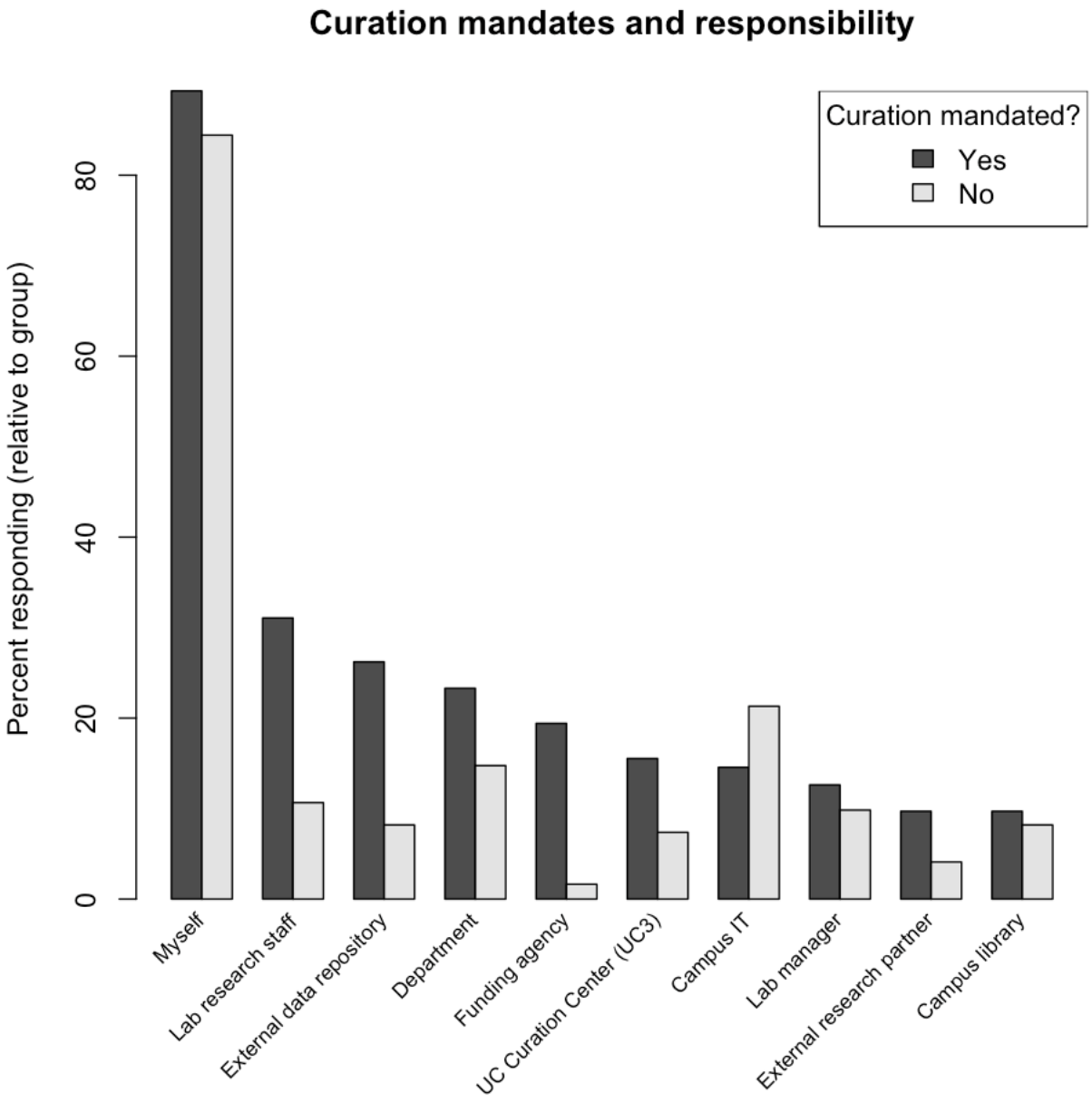
Other responses

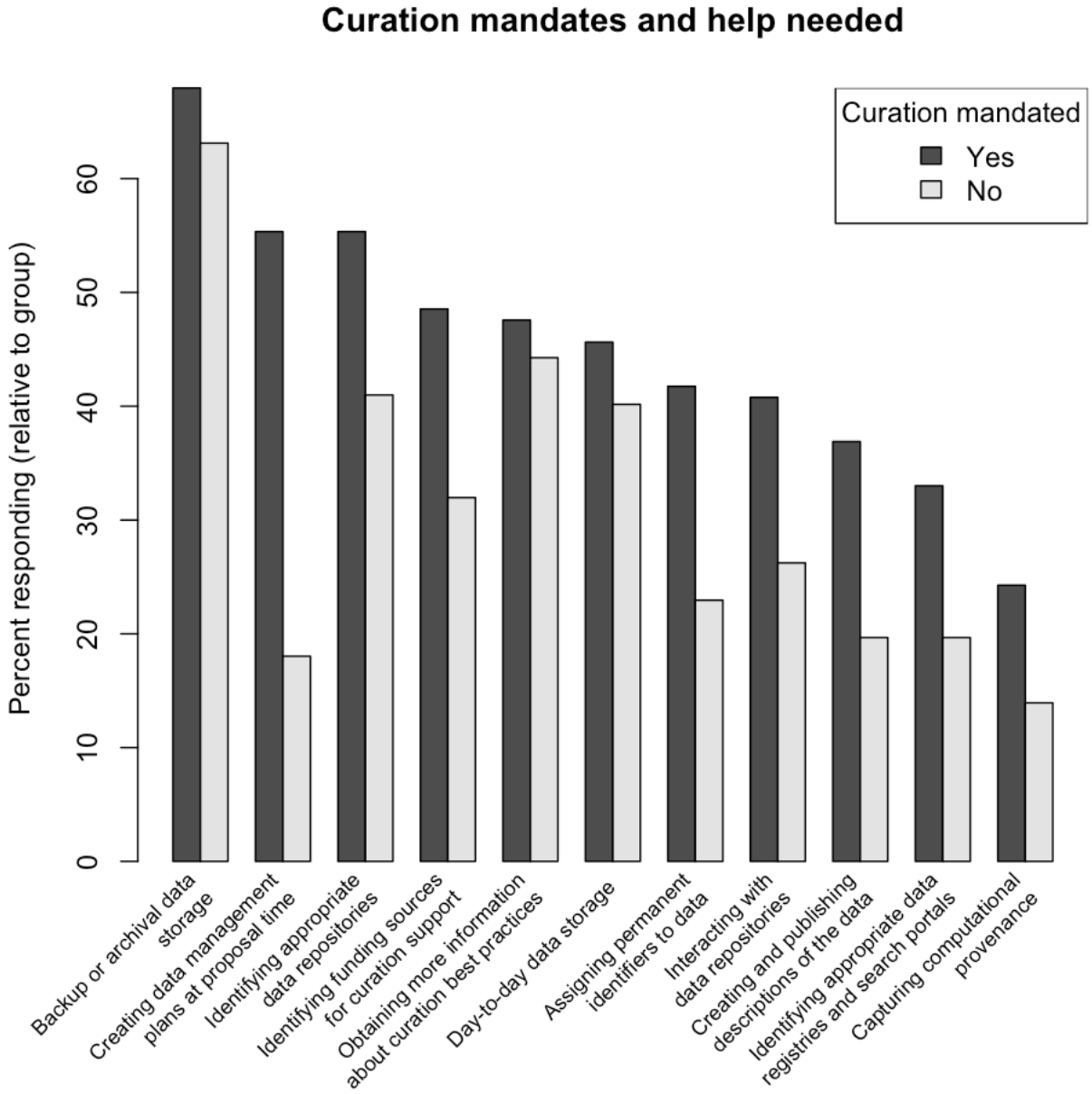
- *Digitization*
  - "interviews still on cassette tape"
  - "floppy to flashdrive"
- *Education*
  - "most people will not understand what all of the above are. […] There has to be a simpler way to describe."
  - "info about best practices and standards for metadata"
  - "tech info on suitable data storage"
  - "managing ethical issues"
- *Access*
  - "Develop data access tools for the users (search portals,...)"
- *Storage*
  - "We need a modern cross-platform file repository"
- *Management*
  - "Maintenance of data connected with publications - articles"


Are you mandated to provide for the curation of your data, and if so, by which agencies?
Percent responding
0
10
20
30
40
50
60
No
Yes
NSF
NIH
NASA
DOE

Additional mandate sources

- *U.S. federal agencies*
  - DOD, Department of the Interior, EPA, NEH, USGS
- *Grant funding organizations*
  - NARSAD, ACLS
- *Archives*
  - Social Science Data Archive, Moscow, Russia
- *Corporations*
  - Pharmaceutical companies
- *University units*
  - Chancellor's Outreach Advisory Board


Curation mandates and responsibility
Curation mandated?
Yes
No
Percent responding (relative to group)
0
20
40
60
80
Myself
Lab research staff
External data repository
Department
Funding agency
UC Curation Center (UC3)
Campus IT
Lab manager
External research partner
Campus library


Curation mandates and help needed
Curation mandated
Yes
No
Percent responding (relative to group)
0
10
20
30
40
50
60
Backup or archival data storage
Creating data management plans at proposal time
Identifying appropriate data repositories
Identifying funding sources for curation support
Obtaining more information about curation best practices
Day-to-day data storage
Assigning permanent identifiers to data
Interacting with data repositories
Creating and publishing descriptions of the data
Identifying appropriate data registries and search portals
Capturing computational provenance

Comments received

- *Kudos & interest*
  - "I welcome this initiative and for me it is very timely."
  - "Very important. One can only speculate at the lost research insights and lost innovations that have occurred."
  - "I got some help on my last grant from the office of research on a data management plan. Much appreciated."
  - "Thank you for this study. I, and probably other faculty and staff, would benefit from learning more about data storage/dissemination options."
  - "I […] want to know what I can do to help preserve and migrate the digital data being produce by our organization."

Comments received

- *Strategies*
  - "I save all my de-identified data in Gmail. I simply send anything I want to save to myself…"
  - "I use datadryad.org to publish my data with my papers and I am quite happy with their service."
- *Content types*
  - "…digital data and physical collections…"
  - "…microfilmed archival material…"
  - "…video..."
  - "…e-mail…"
  - "…VHS collection…"
  - "…some rare […] cartoons in french…"
  - "…'data bank' type data…"

Comments received

- *Commentary*
  - "For the most part, […] computers and storage are not a problem, however, software tools are in short supply. One of the largest hurdles is finding tools to efficiently gather and assemble metadata into prescribed models. Libraries have been engaged with these issues (for print resources) for many years, whereas many fields of research are just starting this process. It would benefit those fields greatly to have libraries apply their experience and resources to the curation of digital resources."

Survey conclusion

- *Survey largely confirmational*
  - Anecdotal observations representative
  - But, value in having evidence, numbers
- *Takeaway points*
  - Data curation is problem for entire campus
  - Common model: researcher takes responsibility, but works with partner(s)
  - Many sources of curation mandates
  - Researchers want help with everything
    - But might underestimate need for help using archival storage

In-depth case studies

- *Common themes*
  - Surface simplicity belies deeper curation challenges
  - Outreach needed
  - Education needed

- *Two sample cases*
  - Image collection
  - Longitudinal field study

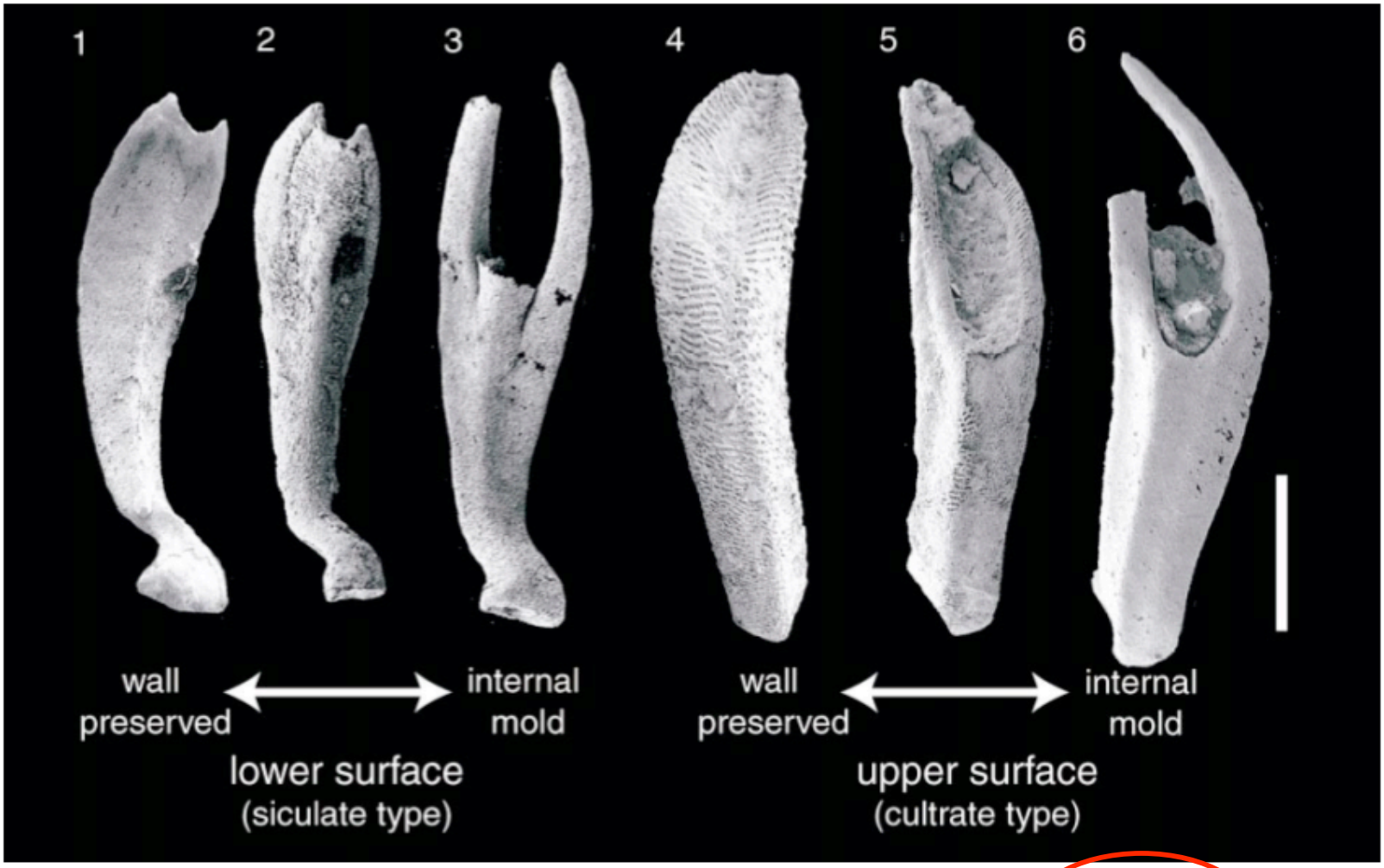
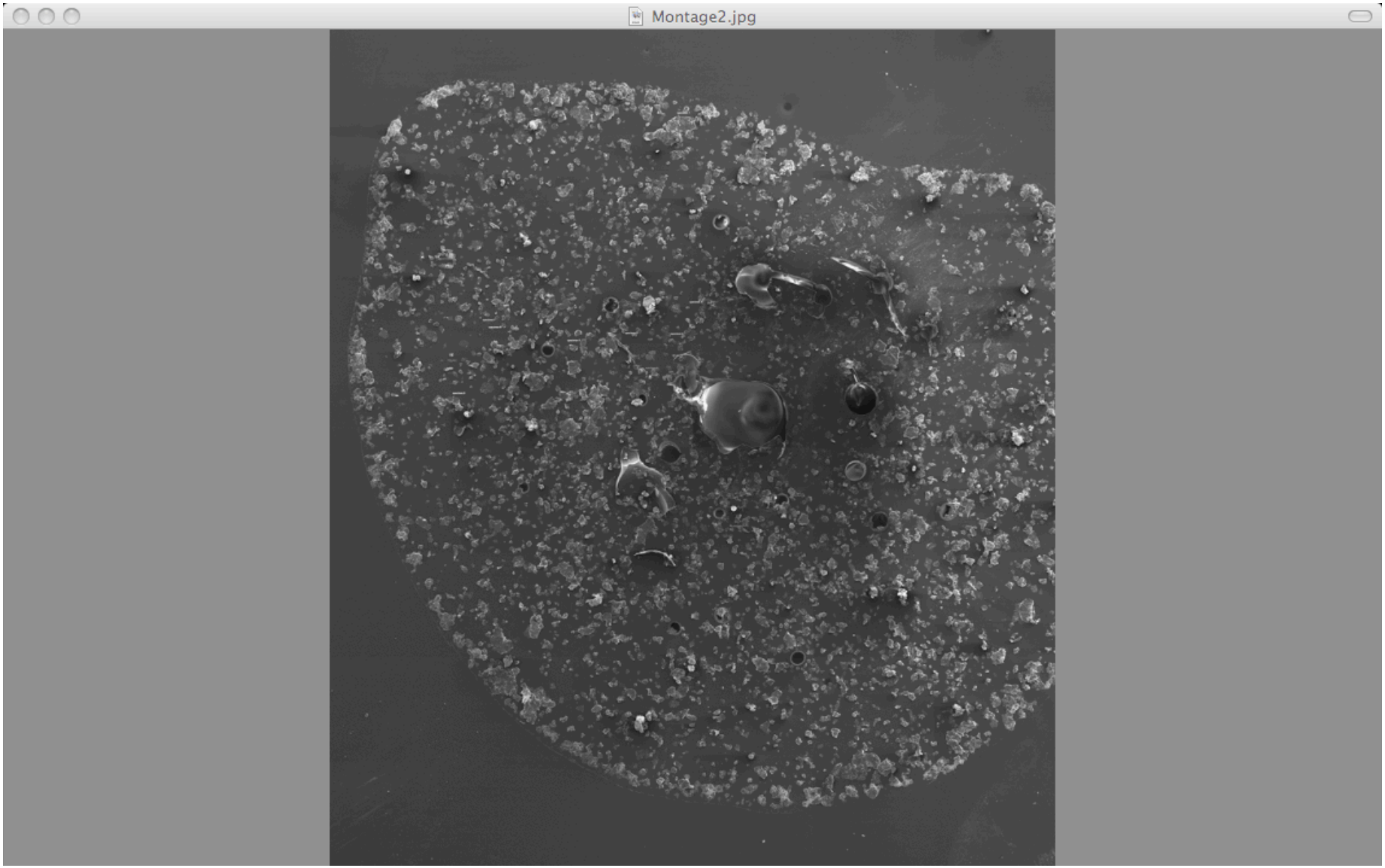
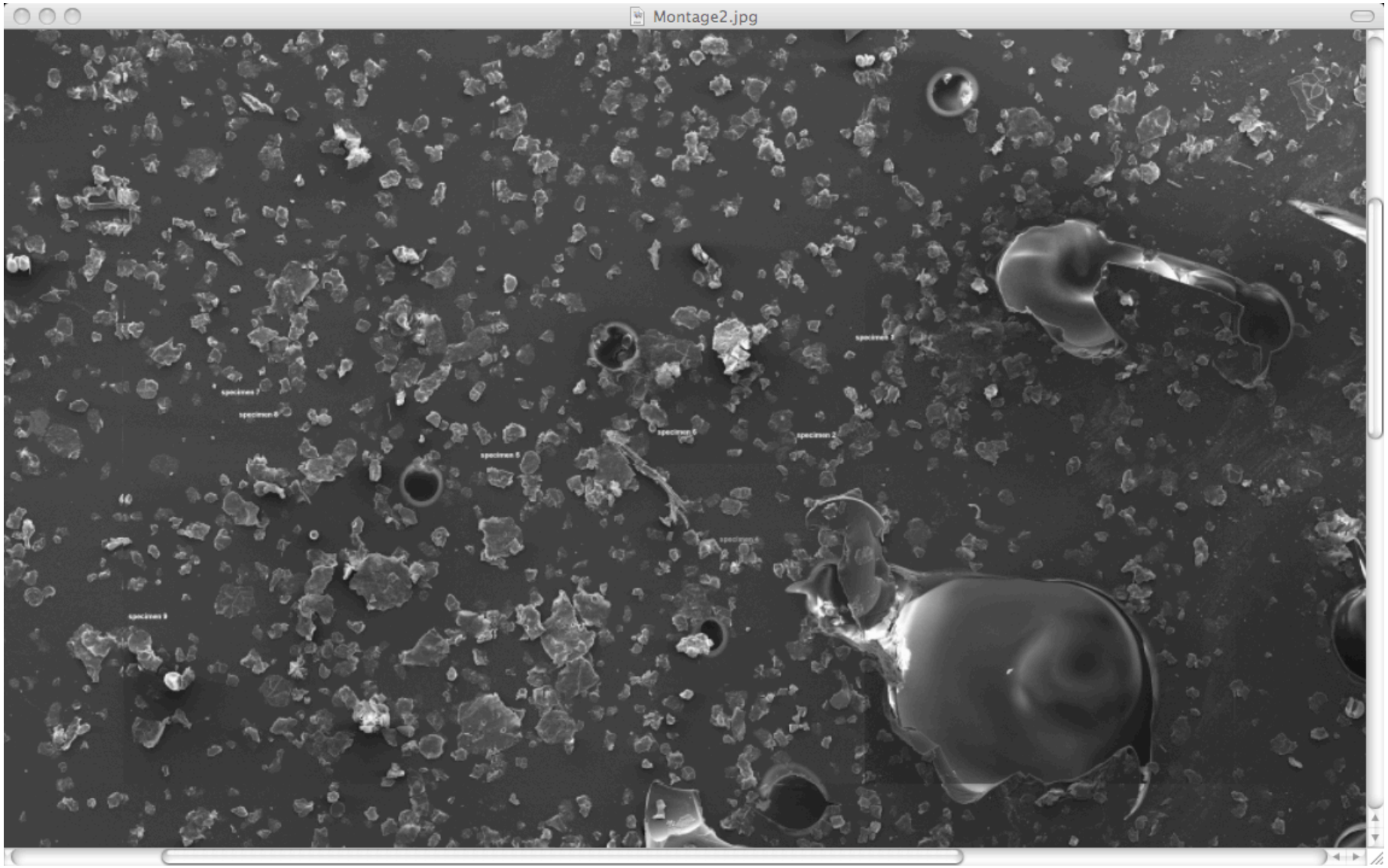
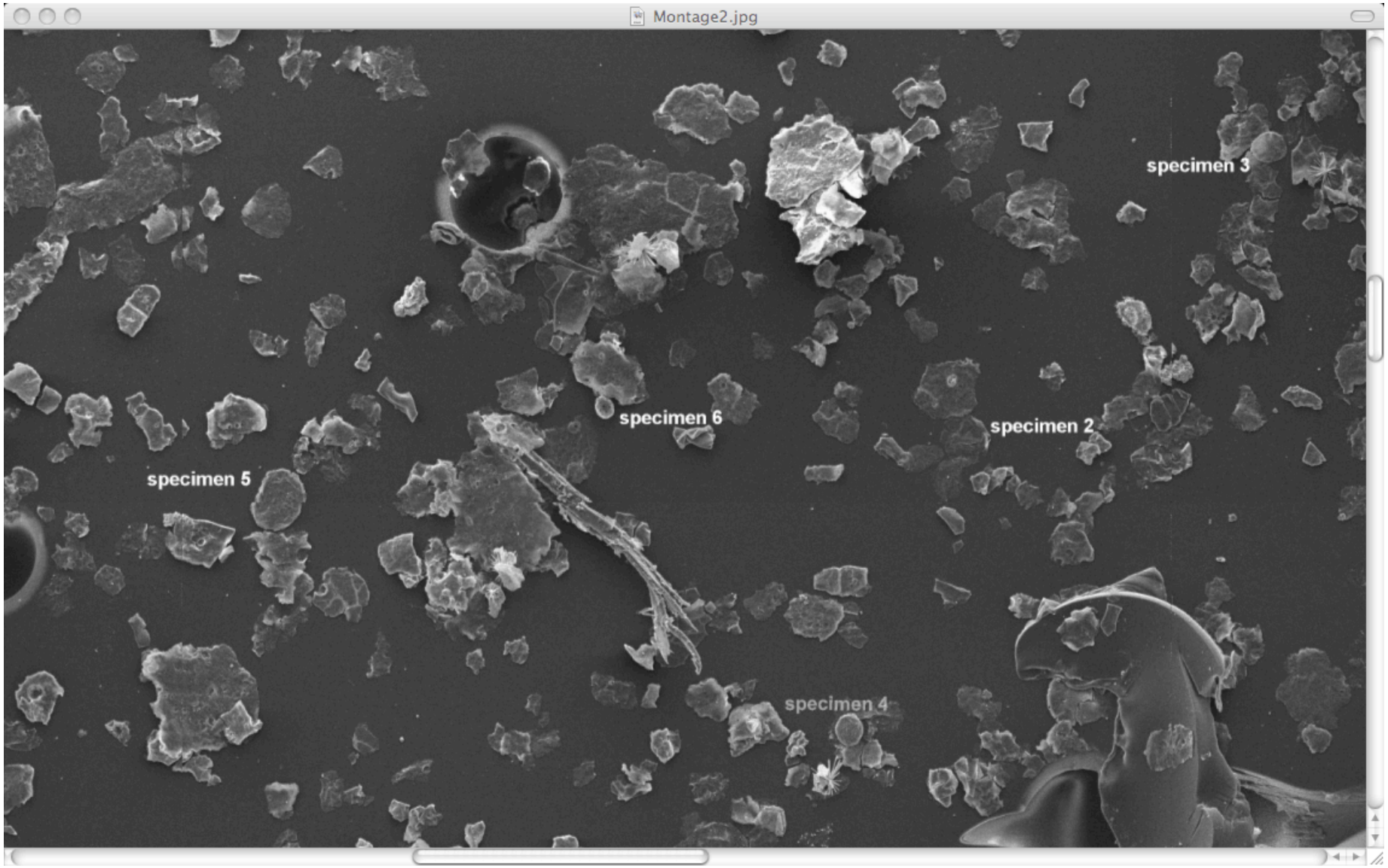
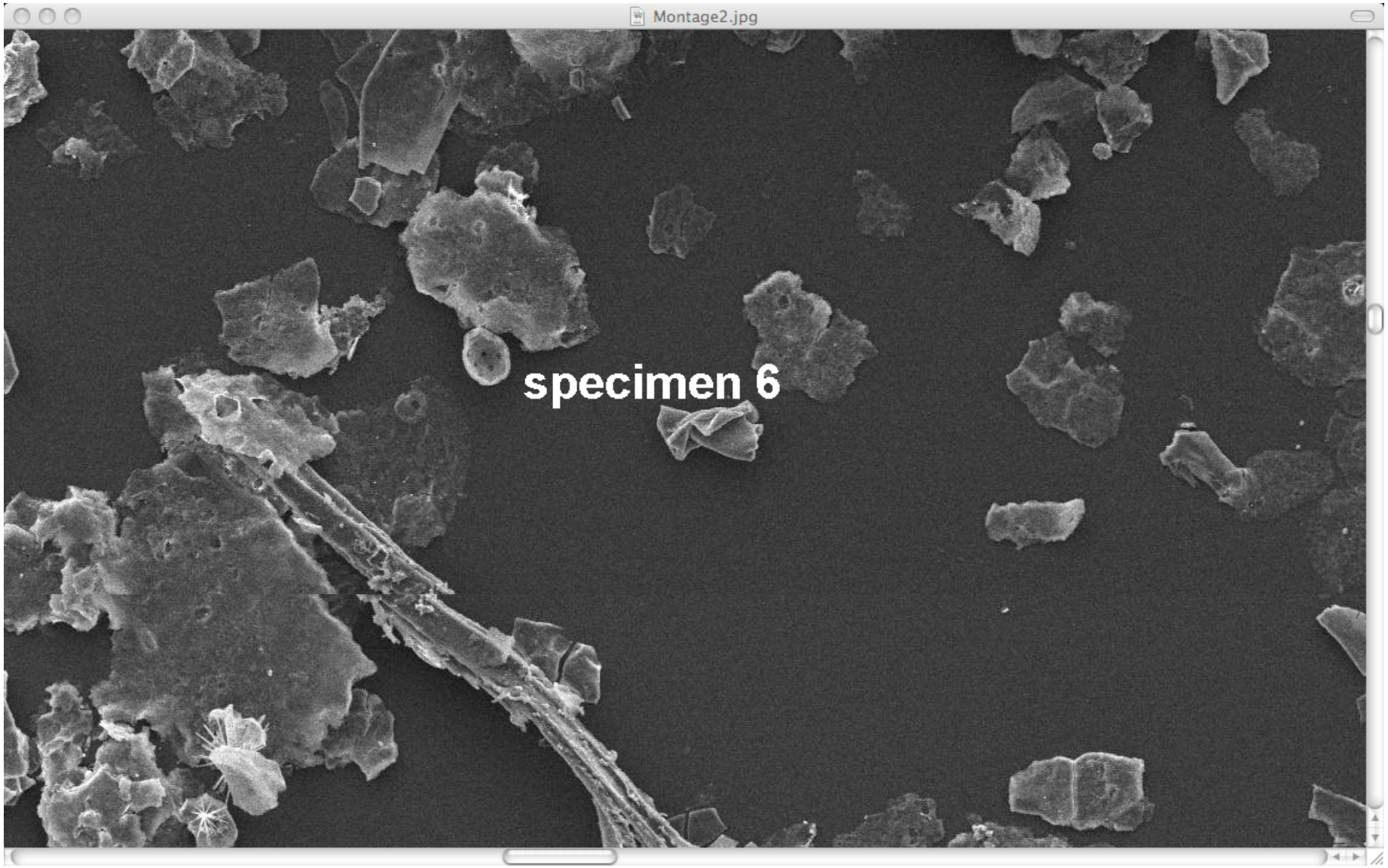


FIGURE 4—Examples of *A. superstes* blade structure. *1–3*, views of the lower surface of siculate sclerites with varying degree of wall preservation; *1*, CPC 37211; *2*, CPC 37212; *3*, CPC 37170. *4–6*, as in *1–3*, but for the upper surface of cultrate sclerites; *4*, CPC 37189; *5*, CPC 37190; *6*, CPC 37171. Scale bar represents 150 μm for *1*, *2*, *4*, *5*; 110 μm for *3*; and 90 μm for *6*.


Montage2.jpg

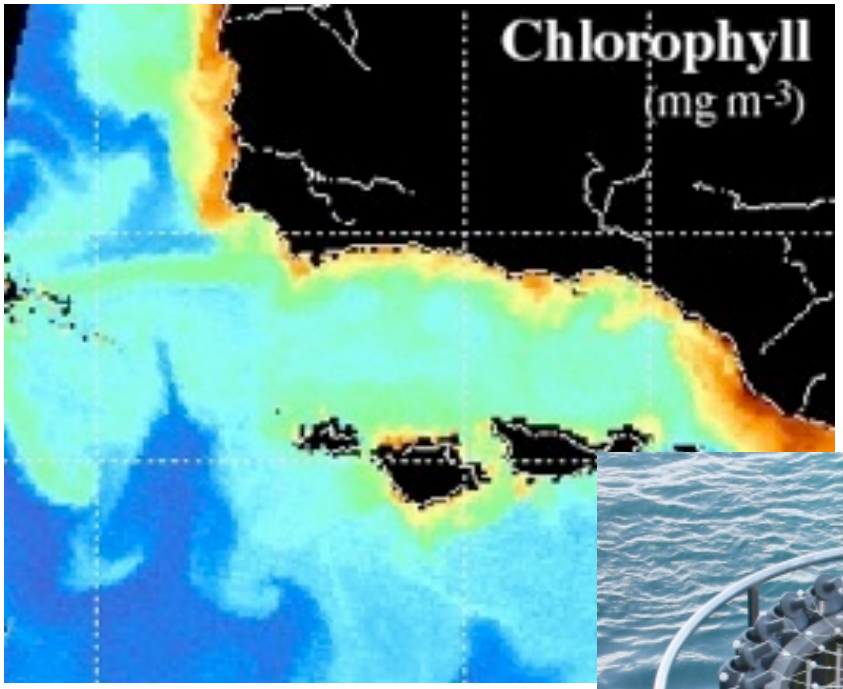

Montage2.jpg

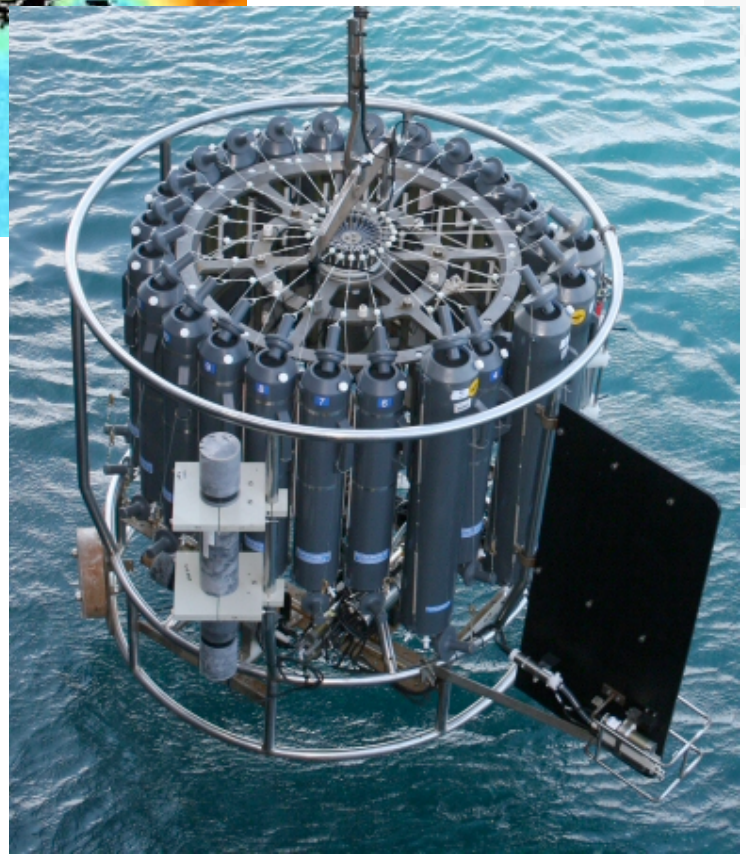

Montage2.jpg
specimen 3
specimen 6
specimen 2
specimen 5
specimen 4


Montage2.jpg
specimen 6

Deeper challenges

- *Map documents*
- *Consistent, unambiguous, persistent identification*
- *Image-specimen-map-slide linkages*
- *Additional images*
- *Intellectual privacy*
  - Researcher gets to name new species!
- *Copyright*
- *Lack of metadata (of course)*
- *Lots of contextual semantics*





http://noc.ac.uk/research-at-sea/nmfss/nmep/ctd

Calibration problems!

- *"…we are using long-term averages of calibration coefficients whenever possible."*
- *"…these detectors may have begun to deteriorate…"*
- *"…we are unsure if this is a trend or merely intercalibration variability."*
- *"…frequent transfer calibrations between lamps…"*

Deeper challenges

- *Complex workflows*
- *Dynamic data operation*
  - Constantly growing
  - Concurrent algorithm development, refinement
- *Retrospective reprocessing*
- *Versioning, identifiers, provenance*
- *Proprietary formats (and software)*
- *Multiple data representations*
- *Lack of automation*

What's the answer?

- *(I don't know!)*

- *Demonstrated needs:*
  - Support for unsupported researchers
    - DataShare, PURR
    - (but getting researchers to use tools may require compelling functionality + outreach)
  - Education in data management
    - Data managers are scientists first, programmers second
    - Curation concepts, services, tools are new anyway
    - (educate how?)

Parting thought

- *Are technological changes spurring new demands for transparency into research methods?*

The good old days

private | public

http://thevintagestandard.com/?p=1296

Zur affinen Feldtheorie.

Von A. Einstein.

Weiteres Nachdenken führte mich zu einer Vervollkommnung der in zwei früheren Mitteilungen behandelten Theorie des Feldes der Gravitation und Elektrizität. Ich will die Theorie in ihrer neuen Form im folgenden kurz darstellen.

Der affine Zusammenhang sei durch die 40 Funktionen $\Gamma^{\mu}_{\nu\sigma}$ dargestellt. Der Riemannsche Krümmungstensor zweiten Ranges $R_{\mu\nu}$ sei in einen symmetrischen Teil $\gamma_{\mu\nu}$ und einen antisymmetrischen Teil $\phi_{\mu\nu}$ gespalten, so daß man hat

$$\gamma_{\mu\nu} = -\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\alpha} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} + \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\nu} + \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right) - \Gamma^{\alpha}_{\nu\alpha}\Gamma^{\beta}_{\mu\beta}, \quad (1)$$

$$\phi_{\mu\nu} = \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\alpha} - \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right). \quad (2)$$

Die Hamiltonsche Funktion $\mathfrak{H}$ (skalare Dichte) sei eine vorläufig unbekannte Funktion[1] der $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$. Das Hamiltonsche Integral ist zu variieren nach den $\Gamma^{\mu}_{\nu\sigma}\,(= \Gamma^{\mu}_{\sigma\nu})$. Man erhält zunächst

$$\int (\mathfrak{g}^{\mu\nu}\delta\gamma_{\mu\nu} + \mathfrak{f}^{\mu\nu}\delta\phi_{\mu\nu})\,d\tau = 0, \quad (3)$$

wobei zur Abkürzung gesetzt ist

$$\left.\begin{aligned} \frac{\partial \mathfrak{H}}{\partial \gamma_{\mu\nu}} &= \mathfrak{g}^{\mu\nu} \\ \frac{\partial \mathfrak{H}}{\partial \phi_{\mu\nu}} &= \mathfrak{f}^{\mu\nu}. \end{aligned}\right\} \quad (3a)$$

$\mathfrak{g}^{\mu\nu}$ und $\mathfrak{f}^{\mu\nu}$ seien als Tensordichten des metrischen und elektrischen Feldes aufgefaßt. Setzt man die durch (1) und (2) gegebenen Ausdrücke von $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$ in (3) ein, so erhält man durch Variation die Gleichungen

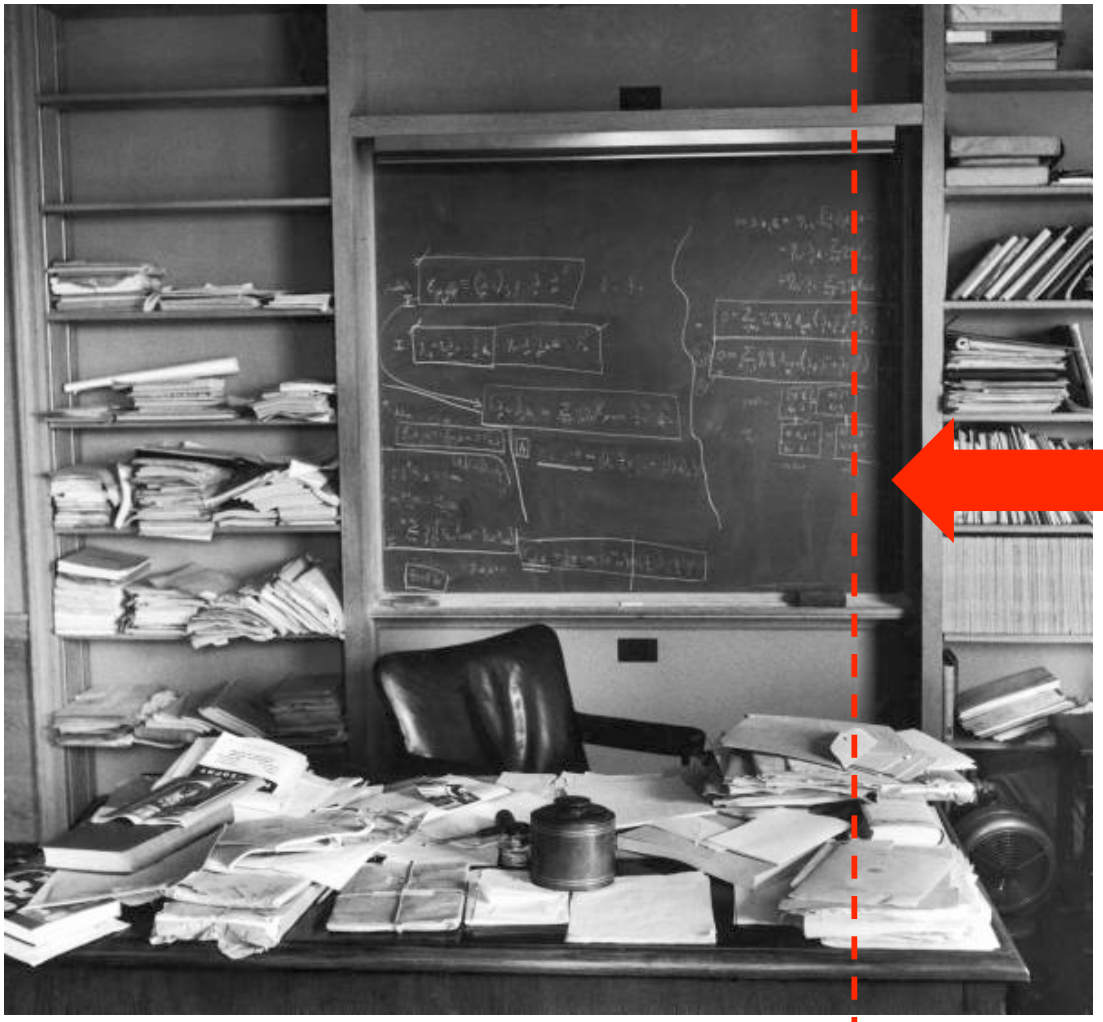
$$\mathfrak{g}^{\mu\nu}{}_{;\alpha} - \frac{1}{2}\mathfrak{g}^{\mu\tau}{}_{;}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{g}^{\nu\tau}{}_{;}\delta^{\mu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\mu\tau}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\nu\tau}\delta^{\mu}_{\alpha} = 0, \quad (4)$$

[1] Die Voraussetzung, daß $\mathfrak{H}$ nur von $\gamma_{\mu\nu}+\phi_{\mu\nu}$ abhänge, wird aufgegeben. Ich möchte ferner bemerken, daß Hr. Donder in Leiden schon vor zwei Jahren ähnliche Überlegungen angestellt aber nicht publiziert hat.

http://www.aip.org/history/einstein/images/ae70.jpg

Today

private | public



http://thevintagestandard.com/?p=1296

Zur affinen Feldtheorie.

Von A. Einstein.

Weiteres Nachdenken führte mich zu einer Vervollkommnung der in zwei früheren Mitteilungen behandelten Theorie des Feldes der Gravitation und Elektrizität. Ich will die Theorie in ihrer neuen Form im folgenden kurz darstellen.

Der affine Zusammenhang sei durch die 40 Funktionen $\Gamma^{\nu}_{\mu\sigma}$ dargestellt. Der Riemannsche Krümmungstensor zweiten Ranges $R_{\mu\nu}$ sei in einen symmetrischen Teil $\gamma_{\mu\nu}$ und einen antisymmetrischen Teil $\phi_{\mu\nu}$ gespalten, so daß man hat

$$\gamma_{\mu\nu} = -\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\alpha} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} + \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} + \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right) - \Gamma^{\alpha}_{\mu\nu}\Gamma^{\beta}_{\alpha\beta}, \quad (1)$$

$$\phi_{\mu\nu} = \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} - \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right). \quad (2)$$

Die Hamiltonsche Funktion $\mathfrak{H}$ (skalare Dichte) sei eine vorläufig unbekannte Funktion[1] der $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$. Das Hamiltonsche Integral ist zu variieren nach den $\Gamma^{\nu}_{\mu\sigma}(= \Gamma^{\nu}_{\sigma\mu})$. Man erhält zunächst

$$\int (\mathfrak{g}^{\mu\nu}\delta\gamma_{\mu\nu} + \mathfrak{f}^{\mu\nu}\delta\phi_{\mu\nu})\,d\tau = 0, \quad (3)$$

wobei zur Abkürzung gesetzt ist

$$\left.\begin{aligned} \frac{\partial \mathfrak{H}}{\partial \gamma_{\mu\nu}} &= \mathfrak{g}^{\mu\nu} \\ \frac{\partial \mathfrak{H}}{\partial \phi_{\mu\nu}} &= \mathfrak{f}^{\mu\nu}. \end{aligned}\right\} \quad (3a)$$

$\mathfrak{g}^{\mu\nu}$ und $\mathfrak{f}^{\mu\nu}$ seien als Tensordichten des metrischen und elektrischen Feldes aufgefaßt. Setzt man die durch (1) und (2) gegebenen Ausdrücke von $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$ in (3) ein, so erhält man durch Variation die Gleichungen

$$\mathfrak{g}^{\mu\nu}{}_{;\alpha} - \frac{1}{2}\mathfrak{g}^{\mu\tau}{}_{;}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{g}^{\nu\tau}{}_{;}\delta^{\mu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\mu\tau}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\nu\tau}\delta^{\mu}_{\alpha} = 0, \quad (4)$$

[1] Die Voraussetzung, daß $\mathfrak{H}$ nur von $\gamma_{\mu\nu}+\phi_{\mu\nu}$ abhänge, wird aufgegeben. Ich möchte ferner bemerken, daß Hr. Donder in Leiden schon vor zwei Jahren ähnliche Überlegungen angestellt aber nicht publiziert hat.

http://www.aip.org/history/einstein/images/ae70.jpg