Summary report on survey responses to:
Survey of Computational Needs for Data and Modeling in Seismology
IRIS High-Performance Computing and Data Working Group
https://www.iris.edu/hq/about_iris/governance/hpcwg
chair: Carl Tape, ctape@alaska.edu
February 27, 2017

1 The survey

The High Performance Computing and Seismic Data Working Group created a survey to assess computational needs in seismology. The full survey is included in Appendix A. Our summary here is based on the responses in Appendices B and C. The motivation for the survey was listed at the top of the survey:

==========
Motivation: *High-performance computing (HPC) resources** can open doors to new scientific problems and can accelerate solving existing problems. Some key ingredients for seismological research are available to the community, for example, access to seismic waveforms via IRIS (http://ds.iris.edu/ds/nodes/dmc/), access to HPC codes via CIG (https://geodynamics.org/), and access to HPC resources via XSEDE allocations. However, the integration of data, processing, modeling, and science with HPC resources is challenging. We seek feedback from the seismological community—students, staff, faculty, scientists in research/academic settings—to assess the needs for HPC in training and research in seismology. The purpose is to provide this feedback to the U.S. National Science Foundation.

*For the sake of this survey, we define *HPC resources* as a system with at least 32 cores (1 core = an individual processor) that can interact with each other in parallel to solve large-memory problems. This could be a high-end laptop with 32 cores or a local cluster with 200 cores or a larger HPC facility with thousands of cores.
==========

On October 7, 2016, the survey was emailed to distribution lists from the following organizations: IRIS (https://www.iris.edu/), SCEC (https://www.scec.org/), CIG (https://geodynamics.org/), EarthScope (http://earthscope.org/), and QUEST (http://www.quest-itn.org/), a European training network in computational seismology.

2 Summary of responses

The survey received an impressive number of 348 responses. [Q1] Respondents were balanced among students, postdocs, faculty, and researchers. [Q2] Most respondents were from North America (66%) and Europe (17%). [Q4] 94% of the respondents indicated that they either currently use HPC resources (59%) or would like to do so in the future (35%).

A couple questions were asked to all respondents, including those not using HPC resources. [Q3] Most time is spent processing data, learning/compiling/running codes, validating/verifying results, and visualization. [Q5] Research among the respondents is distributed among five categories, with the three largest “major research” categories being Data Intensive, Forward Calculation of Wavefield, and Inverse modeling. [Q6] These same trends were also reflected in the question about interest in future research.

[Q8] Many respondents have access to multiple different HPC resources. In order of most to least common, these are: (1) shared facility/cluster at own institution, (2) single-group or departmental computing cluster, (3) own high-performance computer, (4) remote machine at national facility, (5) cluster at another institution, (6) Hadoop/Sparc framework for high-throughput computing. [Q9] Among HPC users, the top categories of “heavy use” HPC were Data Intensive, Forward Calculation of Wavefield, and Inverse modeling; the lowest was Real-time analysis.

Questions 11-18 pertain to seismic data. 94% of respondents use observational seismic data in their research. [Q12] 86% of the respondents obtained data from IRIS DMC, though most people also accessed data from other centers (NCEDC, SCEDC, EIDA, etc). [Q13] The dominant file formats in use are ASCII, SAC, and SEED, though some respondents use SEG formats, ASDF, or their own. [Q14] Respondents process data using multiple tools, with the most widely used being Matlab, SAC, and ObsPy. [Q15,Q16] Over half of the respondents use over 100,000 data files in their research; 34% or respondents use over 1 million data files in their research. “Typical sizes” of data sets are gigabytes to terabytes. [Q17] Half of the data sets are static, while half require time-dependent changes such as reprocessing. [Q18] Two-thirds of the current users of HPC resources said they would benefit from having their data in close proximity to HPC resources.

Questions 19-24 pertain to simulations (or “synthetic data”). [Q19] 90% of respondents use simulations of modeling in their research. [Q20] Most respondents use multiple file formats, with the most common being ASCII and SAC, followed by Matlab binary, SEG, netCDF, and ASDF. [Q21] Matlab, SAC, python, and ObsPy have comparable useage for processing synthetic time series. [Q22,Q23] 41% of the respondents use over 10,000 synthetic time series files in their research; 29% of respondents use over 100,000 files in their research. “Typical sizes” of synthetic data sets are gigabytes. [Q24] About half of the synthetic data sets are static, while half require time-dependent changes such as reprocessing.

[Q25] Approximately half of all current users of HPC resources identified “more HPC resources” as the greatest need. Each of the other 8 needs was identified as moderate-to-great by more than half of the respondents.

[Q26] Among current users of HPC resources, about 72% expressed moderate to great benefit to their research from additional computational training.

3 Take-home points

1. The number of respondents (348) provides an indication of the importance of the topic of HPC resources to the field of seismology. (By comparison, other surveys from the IRIS Data Services committee have received no more than 50 responses.)
2. Data processing, forward modeling of the seismic wavefield, and inverse modeling are the dominant research categories.
3. Most seismologists' time is spent processing data, learning/compiling/running codes, validating/verifying results, and visualization.
4. It seems possible that the large time spent processing data is related to responses showing a range of data formats in use, a lack of HPC resources, and a desire to locate data next to HPC resources.
5. Current users of HPC resources expressed a need for more HPC resources and for more training.

4 Miscellaneous points

Several respondents pointed out that the survey was focused on seismology, while the email announcement asked for feedback from communities in infrasound and geodesy. This is true, in the sense that seismology was the focus, but we were also interested in getting feedback from other communities. There is a good detailed comment in Appendix C about the needs for handling data volumes for earthquake studies using synthetic aperture radar data.

One comment was that two categories of Inverse Modeling and Bayesian / Uncertainty Quantification have significant overlap. Our examples distinguished between using HPC for computational intensive forward modeling ("Inverse Modeling") versus using HPC to obtain millions of evaluations of a computationally inexpensive model, toward quantifying uncertainties ("Bayesian / Uncertainty Quantification"). In the future we could consider revising the five categories.

Appendix A: The survey questions

The survey was constructed in Survey Monkey. In order to show exactly what the respondents saw, we screen-captured the questions. (Hence the images are a bit pixelated.)

IRIS HPC and Seismic Data Working Group
Survey of Computational Needs for Data and Modeling in Seismology

1.

Motivation: High-performance computing (HPC) resources* can open doors to new scientific problems and can accelerate solving existing problems. Some key ingredients for seismological research are available to the community, for example, access to seismic waveforms via IRIS (http://ds.iris.edu/ds/nodes/dmc/), access to HPC codes via CIG (https://geodynamics.org/), and access to HPC resources via XSEDE allocations. However, the integration of data, processing, modeling, and science with HPC resources is challenging. We seek feedback from the seismological community—students, staff, faculty, scientists in research/academic settings—to assess the needs for HPC in training and research in seismology. The purpose is to provide this feedback to the U.S. National Science Foundation.

*For the sake of this survey, we define HPC resources as a system with at least 32 cores (1 core = an individual processor) to solve large-memory problems. This could be a high-end laptop with 32 cores or a local cluster with 200 cores or a larger HPC facility with thousands of cores.

Survey Authors: IRIS HPC and Seismic Data Working Group
https://www.iris.edu/hq/about_iris/governance/hpcwg

* 1. What is your career level?

- ◯ Student
- ◯ Postdoc
- ◯ Researcher
- ◯ Faculty

Other (please specify)

* 2. In what region do you live and work?

- ◯ Africa
- ◯ Asia
- ◯ Australia/Oceania
- ◯ Europe
- ◯ Middle East
- ◯ North America
- ◯ South America
- ◯ Other (please specify)

* 3. Where do you spend most of your time doing your science? (What are you bottlenecks?)

	No time or not applicable	Little	Moderate	A lot of time
Establishing, using, and understanding workflows	○	○	○	○
Accessing data	○	○	○	○
Processing data	○	○	○	○
Learning concepts or theories, including finding relevant bibliography.	○	○	○	○
Learning, compiling, and running codes	○	○	○	○
Waiting for simulations to run (limited by not enough computational resources)	○	○	○	○
Transferring data or simulation output to and from machines	○	○	○	○
Validating or verifying results	○	○	○	○
Ensuring reproducibility	○	○	○	○
Visualization (for understanding or for publication-ready figures)	○	○	○	○

2.

The following categories of computational seismology problems will be referred to in some of the following questions.

Data intensive
Examples: massive waveform correlation, ambient noise cross-correlation, stacking, back projection of wavefield for source location

Forward calculation of wavefield in 1D or 3D Earth models
Examples: earthquake ground motion, array-based waveform modeling (e.g., crustal, mantle, core, D", volcano, planets)

Inverse modeling
Examples: data plus HPC forward/adjoint simulations for seismic imaging of structure and source (adjoint tomography, full waveform inversion)

Bayesian methods & uncertainty quantification
Examples: Monte Carlo, resolution/covariance, model testing, transdimensional Bayesian

Real-time analysis (real-time processing of streaming data)
Examples: seismic event detection, high-precision estimation of source parameters

* 4. Do you use HPC resources (defined above as >32 cores)?

- ◯ Yes I do now
- ◯ No, I don't use HPC resources now but I would like to use HPC resources
- ◯ No. And I do not have plans to use HPC resources. [Please explain why.]

3.

* 5. I would characterize my research in each of the following categories as:

	No research	Minor Research	Moderate Research	Major Research
Data Intensive	○	○	○	○
Forward calculation of wavefield in 1D or 3D Earth models	○	○	○	○
Inverse modeling	○	○	○	○
Bayesian methods & uncertainty quantification	○	○	○	○
Real-time analysis	○	○	○	○

* 6. In the future I might be interested in pursuing research in these categories:

	Not Interested	Slightly Interested	Moderately Interested	Highly Interested
Data intensive	○	○	○	○
Forward calculation of wavefield in 1D or 3D Earth models	○	○	○	○
Inverse modeling	○	○	○	○
Bayesian methods & uncertainty quantification	○	○	○	○
Real-time analysis	○	○	○	○

7. Sorry to ask again but which of the following is true?

- ○ I currently use HPC resources
- ○ I would like to use HPC resources in the future

5.

* 8. I have access to and/or use the following types of HPC resources and environments: [Select all that apply]

- [ ] high-performance, stand-alone computer (e.g., 32-core laptop)
- [ ] single-group or departmental computing cluster.
- [ ] shared facility/cluster at my institution.
- [ ] remote machine at a National Facility (e.g. XCEDE, INCITE)
- [ ] cluster at another institution
- [ ] Hadoop/Sparc framework / high-throughput computing

Other (please specify)

* 9. For each category, do you use HPC resources:

	Do Not Use	Little Use	Moderate Use	Heavy Use
Data Intensive	○	○	○	○
Forward Modeling	○	○	○	○
Inverse Modeling	○	○	○	○
Bayesian / UQ	○	○	○	○
Real-time analysis	○	○	○	○

* 10. What types of HPC do you use or want to use (check all that apply)?

	Do Not Use	Little Use	Moderate Use	Heavy Use
Large shared-memory, Linux cluster	○	○	○	○
CPU	○	○	○	○
GPU	○	○	○	○
Hadoop/Sparc framework	○	○	○	○
Other	○	○	○	○

* 11. Do you use observational data in your research

- ◯ Yes
- ◯ No

7.

Since you use observational data please answer the following questions.

* 12. I use the following seismological data resources (check all that apply):

- [ ] IRIS DMC
- [ ] NCEDC - Berkeley
- [ ] SCEDC - Caltech
- [ ] European Integrated Data Center (EIDA)
- [ ] Another FDSN Data Center
- [ ] Another non-FDSN Center

Another Data Center

* 13. I use the following time-series file formats (check all that apply):

- [ ] ASCII (text file)
- [ ] SAC
- [ ] SEED
- [ ] miniSEED and Dataless Seed or StationXML
- [ ] SEG formats (e.g. SEG-D, SEG-Y, SEG-B)
- [ ] ASDF
- [ ] My Own

Other (please specify) or if your own name it

* 14. I use the following tools for processing data (check all that apply):

- [ ] SAC
- [ ] my own python programs
- [ ] ObsPy
- [ ] Matlab
- [ ] Antelope
- [ ] Database (either relational or NoSQL)

Other Processing Software (please specify)

* 15. My research involves the following number of data files:

- ◯ < 1,000
- ◯ < 10,000
- ◯ < 100,000
- ◯ < 1,000,000
- ◯ > 1,000,000

* 16. The typical size of my dataset is

- ◯ megabytes
- ◯ gigabytes
- ◯ terabytes
- ◯ petabytes

* 17. In terms of time-dependent changes or frequent reprocessing, my datasets can be characterized as

- ◯ static
- ◯ requiring frequent changes or processing
- ◯ requiring on-demand updates (such as servicing and reprocessing data from a large field experiment)
- ◯ real-time (need for streaming continuous data and metadata)

18. Do you need your data close to HPC Resources?

	No	A little	Moderately	A great deal
I would benefit from having my data in close proximity to HPC resources.	○	○	○	○

9.

* 19. I use simulations or modeling in my research

- ◯ yes
- ◯ no

10.

* 20. I use the following seismic data file formats for simulation output (check all that apply):

- [ ] ASCII text file
- [ ] SAC
- [ ] SEG Formats (e.g. SEG-D, SEG-Y)
- [ ] ASDF
- [ ] netCDF
- [ ] Matlab binary

Your own or something else

* 21. I use the following tools for processing time-series simulation output (check all that apply):

- [ ] SAC
- [ ] python
- [ ] ObsPy
- [ ] MatLab
- [ ] Database (either Relational or NoSQL)

Other (please specify)

* 22. My research involves the following number of simulated time series files:

- ◯ <100
- ◯ < 1000
- ◯ < 10,000
- ◯ < 100,000
- ◯ > 100,000

* 23. My research involves simulated time series with a total file size of:

- ◯ megabytes
- ◯ gigabytes
- ◯ terabytes
- ◯ petabytes

* 24. In terms of time-dependent changes or frequent reprocessing, my simulated datasets can be characterized as

- ◯ static
- ◯ requiring frequent changes or processing
- ◯ requiring on-demand updates (e.g., new velocity model, new seismograms)

12.

* 25. Please rate your needs for the following:

	No Need	Small Need	Moderate Need	Greatest Need
more HPC resources	○	○	○	○
more short-term storage	○	○	○	○
more long-term storage	○	○	○	○
codes for modeling	○	○	○	○
codes for efficient processing of data or simulation output	○	○	○	○
codes to manage workflow	○	○	○	○
high-speed networks to download data and move simulation output files	○	○	○	○
training for using HPC resources	○	○	○	○
options to modify Earth models (e.g., velocity model, topography, internal boundaries)	○	○	○	○

Other (please specify)

* 26. My research would benefit from hands-on training in the following:

	No Benefit	Little Benefit	Moderate Benefit	Great Benefit
software engineering (e.g. principles of - github, unit testing, documentation, best practices, etc.)	○	○	○	○
coding (e.g. matlab, C, C++, python, scripting	○	○	○	○
specific software training (e.g., paraview, cubit/trelis)	○	○	○	○

Other (please specify)

27. The transition from training to research would be most effective if they were both done in the same computing environment.

	Strongly Disagree	Mildly Disagree	Agree	Strongly agree
Level of agreement	○	○	○	○

28. Please provide any comments or suggestions. If you have a research problem in seismology that is not addressed within the categories listed above, please describe it.

29. supply your contact information.

Name

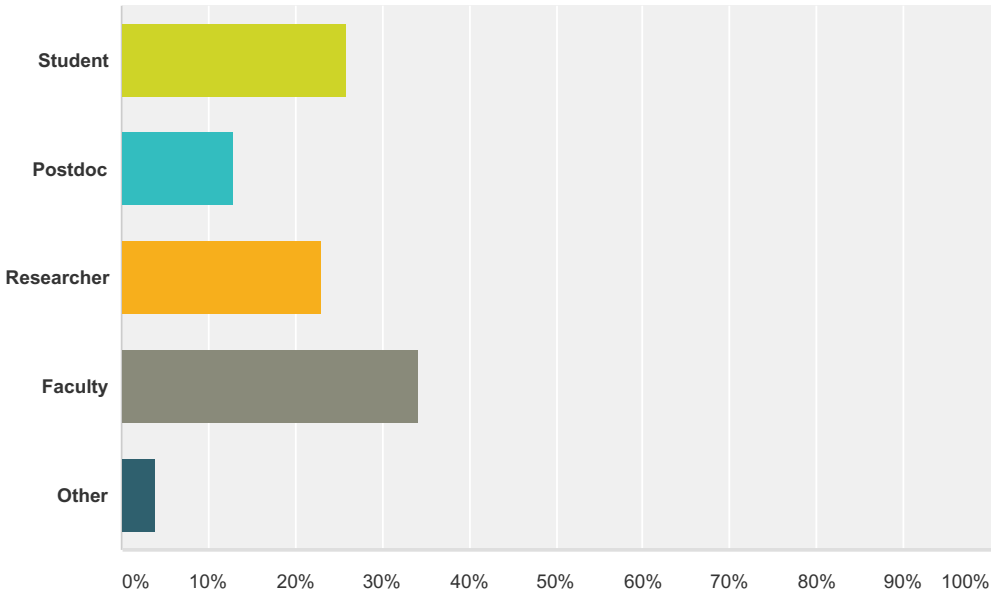
Email Address

Appendix B: Summary of responses

Survey Monkey provided a summary of all responses.

Q1 What is your career level?

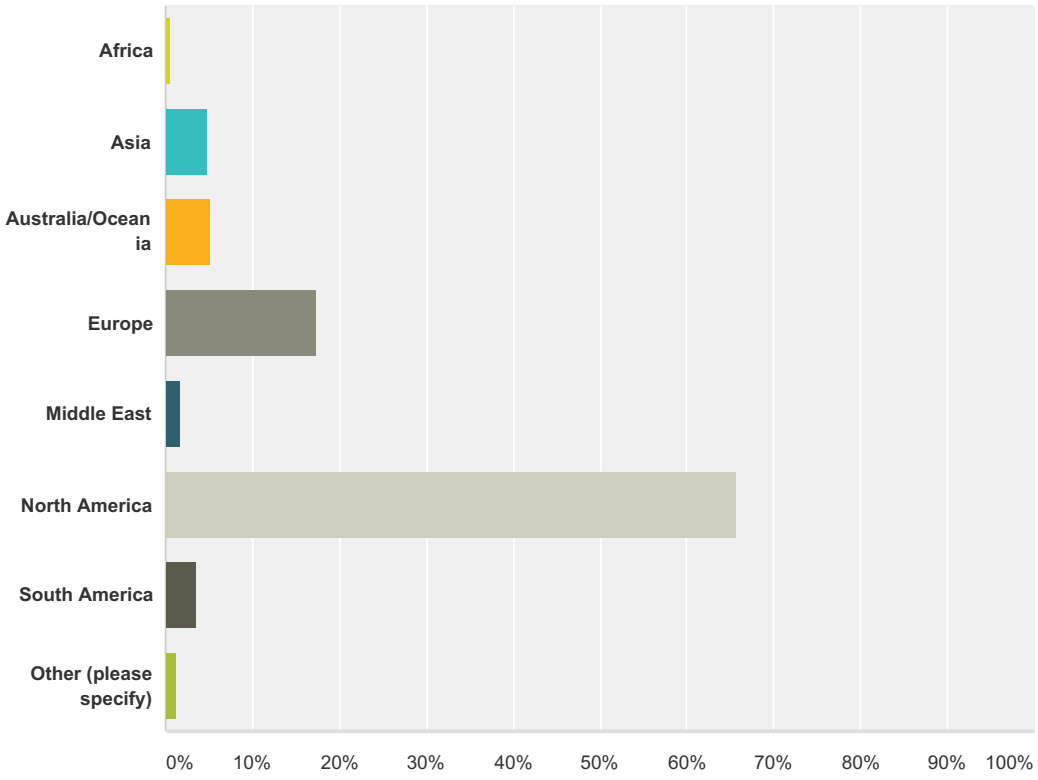
Answered: 348 Skipped: 0



Answer Choices	Responses	
Student	**25.86%**	90
Postdoc	**12.93%**	45
Researcher	**22.99%**	80
Faculty	**34.20%**	119
Other	**4.02%**	14
Total		**348**

Q2 In what region do you live and work?

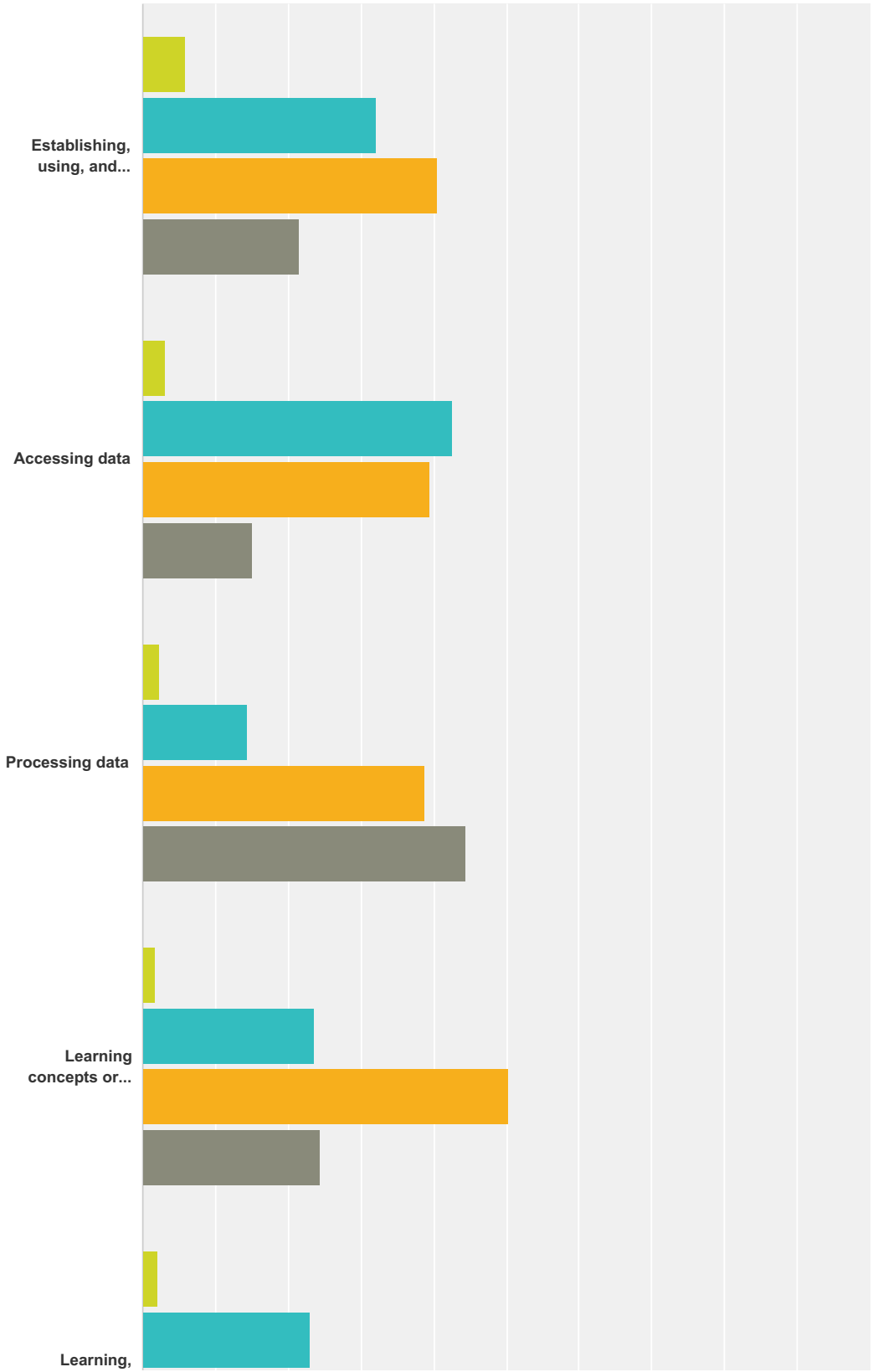
Answered: 348 Skipped: 0

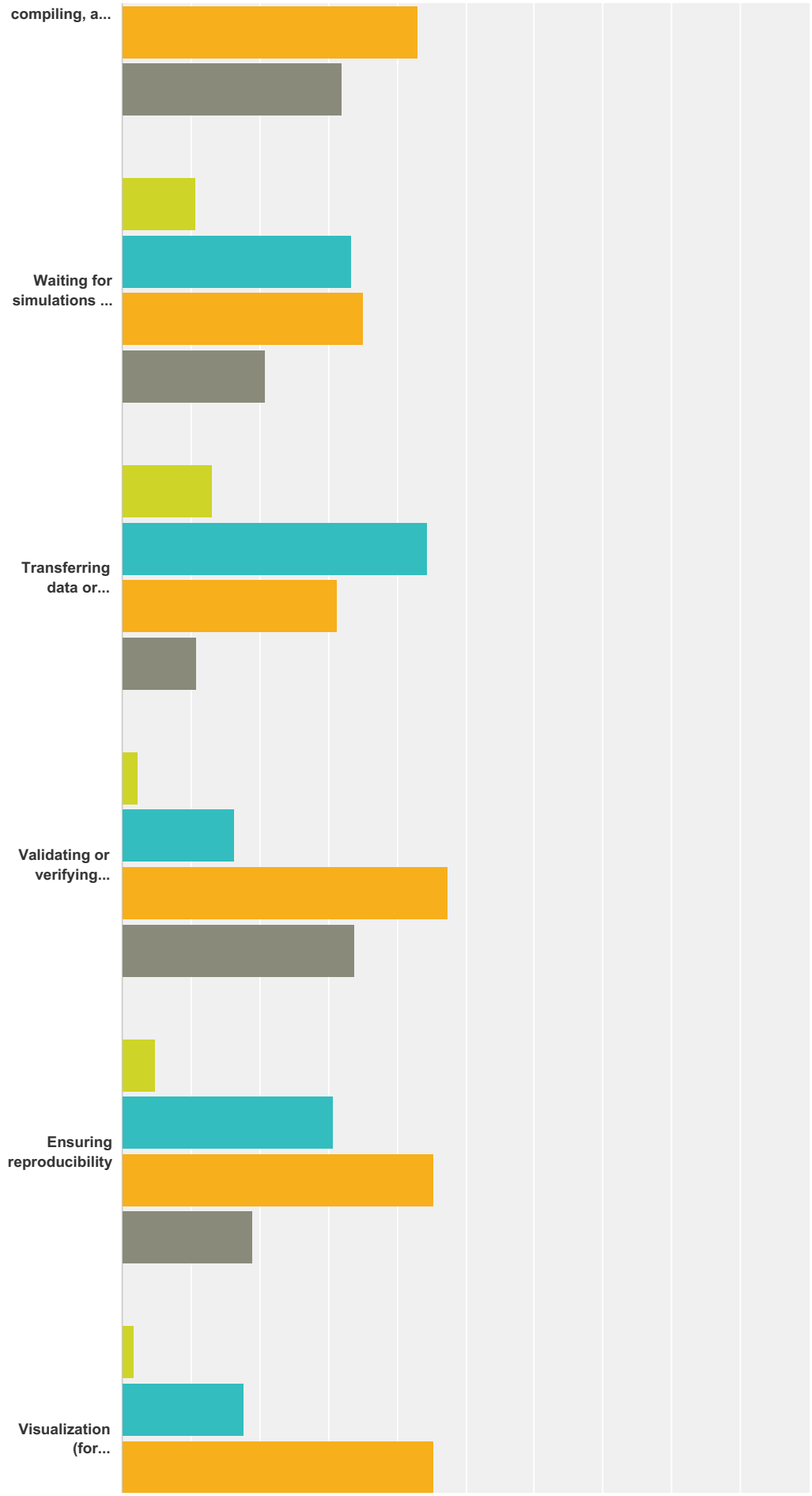


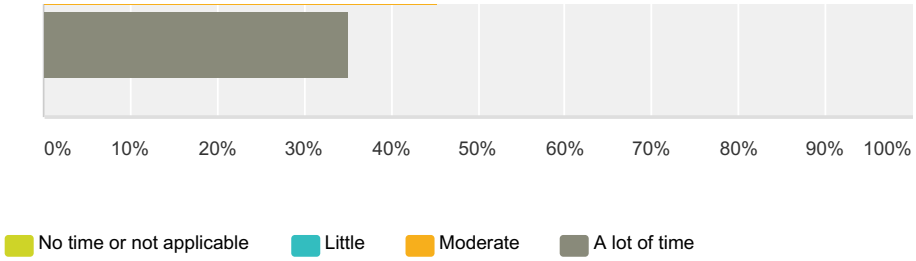
Answer Choices	Responses	
Africa	**0.57%**	2
Asia	**4.89%**	17
Australia/Oceania	**5.17%**	18
Europe	**17.24%**	60
Middle East	**1.72%**	6
North America	**65.80%**	229
South America	**3.45%**	12
Other (please specify)	**1.15%**	4
Total		**348**

Q3 Where do you spend most of your time doing your science? (What are you bottlenecks?)

Answered: 348 Skipped: 0



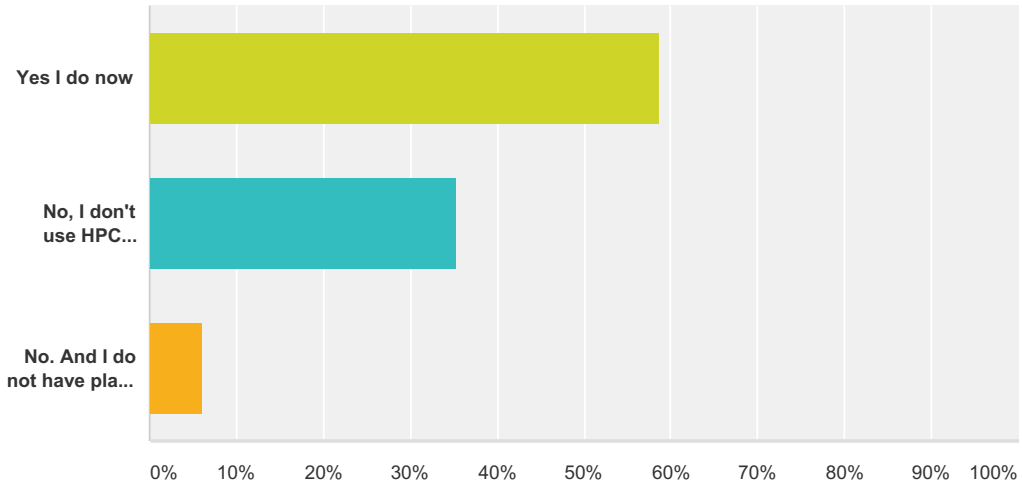

compiling, a...
Waiting for simulations ...
Transferring data or...
Validating or verifying...
Ensuring reproducibility
Visualization (for...



	No time or not applicable	Little	Moderate	A lot of time	Total
Establishing, using, and understanding workflows	**5.75%** 20	**32.18%** 112	**40.52%** 141	**21.55%** 75	348
Accessing data	**3.16%** 11	**42.53%** 148	**39.37%** 137	**14.94%** 52	348
Processing data	**2.30%** 8	**14.37%** 50	**38.79%** 135	**44.54%** 155	348
Learning concepts or theories, including finding relevant bibliography.	**1.72%** 6	**23.56%** 82	**50.29%** 175	**24.43%** 85	348
Learning, compiling, and running codes	**2.01%** 7	**22.99%** 80	**43.10%** 150	**31.90%** 111	348
Waiting for simulations to run (limited by not enough computational resources)	**10.63%** 37	**33.33%** 116	**35.06%** 122	**20.98%** 73	348
Transferring data or simulation output to and from machines	**13.22%** 46	**44.54%** 155	**31.32%** 109	**10.92%** 38	348
Validating or verifying results	**2.30%** 8	**16.38%** 57	**47.41%** 165	**33.91%** 118	348
Ensuring reproducibility	**4.89%** 17	**30.75%** 107	**45.40%** 158	**18.97%** 66	348
Visualization (for understanding or for publication-ready figures)	**1.72%** 6	**17.82%** 62	**45.40%** 158	**35.06%** 122	348

Q4 Do you use HPC resources (defined above as >32 cores)?

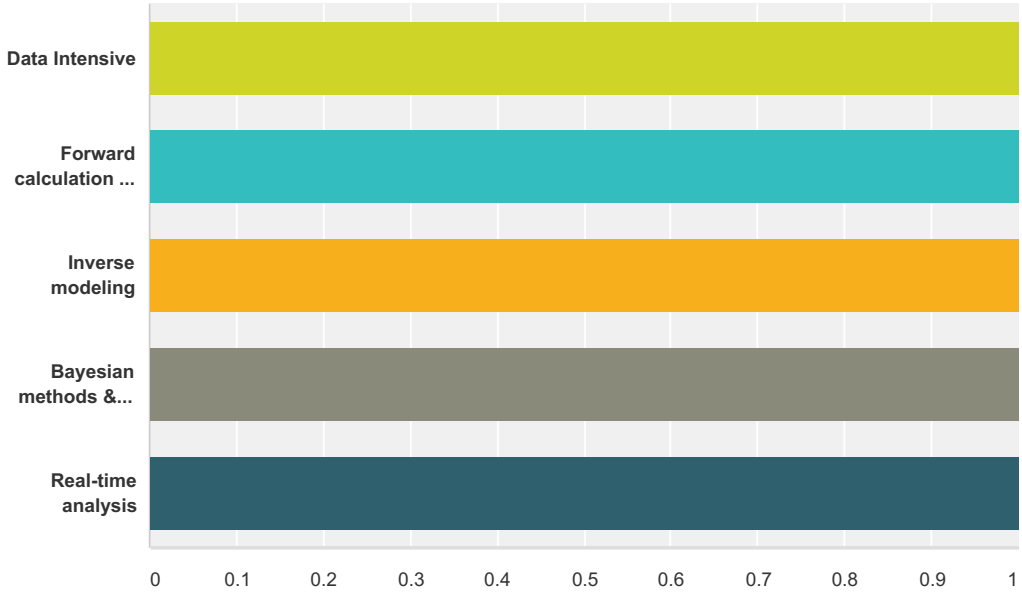
Answered: 341 Skipped: 7



Answer Choices	Responses	
Yes I do now	**58.65%**	200
No, I don't use HPC resources now but I would like to use HPC resources	**35.19%**	120
No. And I do not have plans to use HPC resources. [Please explain why.]	**6.16%**	21
Total		**341**

Q5 I would characterize my research in each of the following categories as:

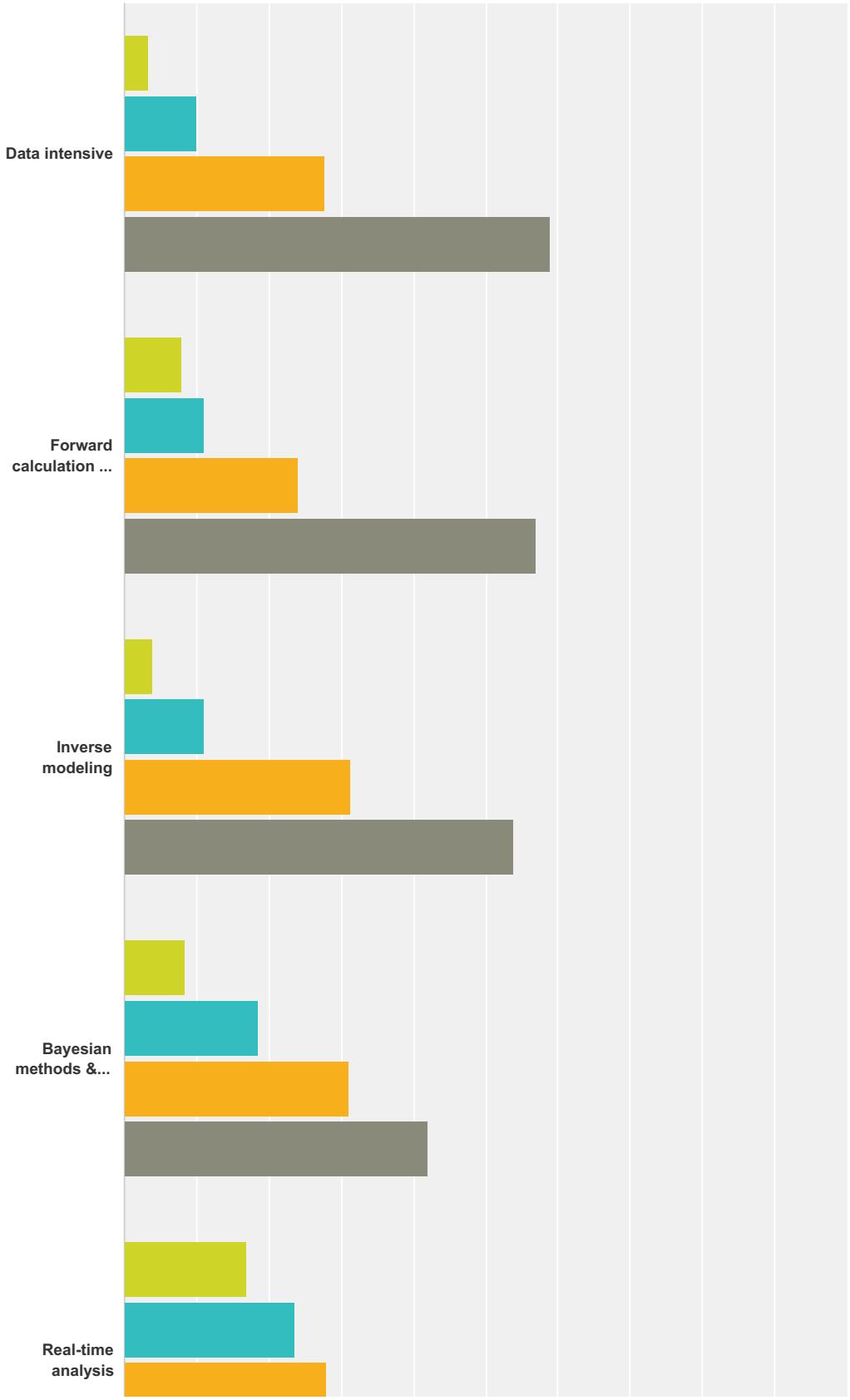
Answered: 310 Skipped: 38

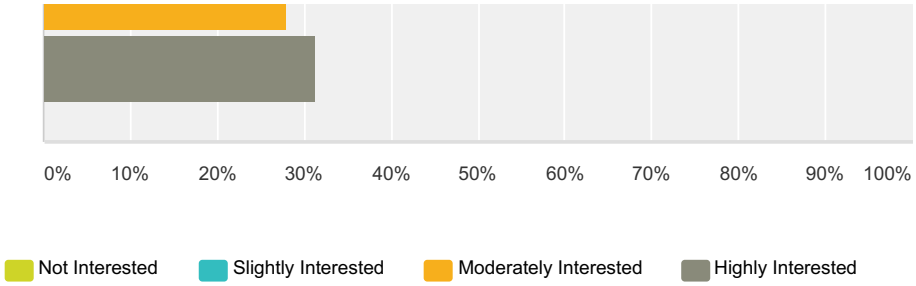


	No research	Minor Research	Moderate Research	Major Research	Total	Weighted Average
Data Intensive	**5.48%** 17	**19.68%** 61	**31.29%** 97	**43.55%** 135	310	1.00
Forward calculation of wavefield in 1D or 3D Earth models	**18.39%** 57	**22.90%** 71	**24.19%** 75	**34.52%** 107	310	1.00
Inverse modeling	**9.35%** 29	**26.13%** 81	**24.52%** 76	**40.00%** 124	310	1.00
Bayesian methods & uncertainty quantification	**27.42%** 85	**35.48%** 110	**24.84%** 77	**12.26%** 38	310	1.00
Real-time analysis	**44.52%** 138	**28.06%** 87	**15.48%** 48	**11.94%** 37	310	1.00

Q6 In the future I might be interested in pursuing research in these categories:

Answered: 308 Skipped: 40

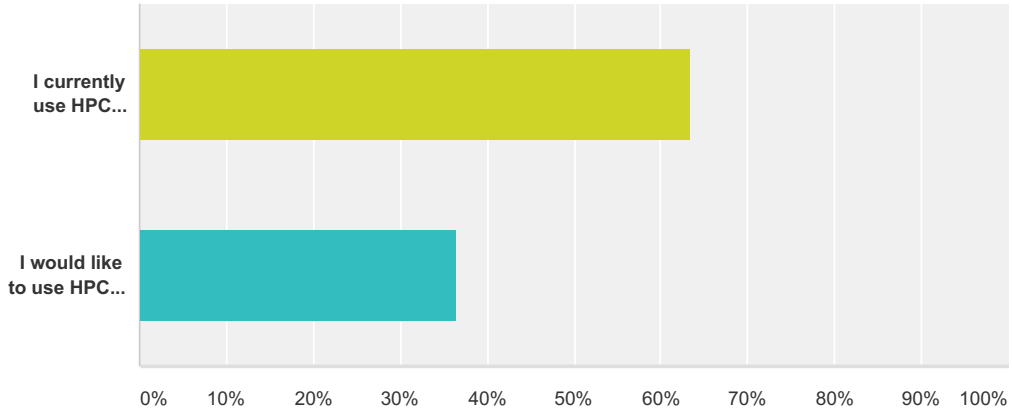




	Not Interested	Slightly Interested	Moderately Interested	Highly Interested	Total
Data intensive	**3.34%** 10	**10.03%** 30	**27.76%** 83	**58.86%** 176	299
Forward calculation of wavefield in 1D or 3D Earth models	**8.00%** 24	**11.00%** 33	**24.00%** 72	**57.00%** 171	300
Inverse modeling	**3.99%** 12	**10.96%** 33	**31.23%** 94	**53.82%** 162	301
Bayesian methods & uncertainty quantification	**8.28%** 25	**18.54%** 56	**31.13%** 94	**42.05%** 127	302
Real-time analysis	**17.00%** 51	**23.67%** 71	**28.00%** 84	**31.33%** 94	300

Q7 Sorry to ask again but which of the following is true?

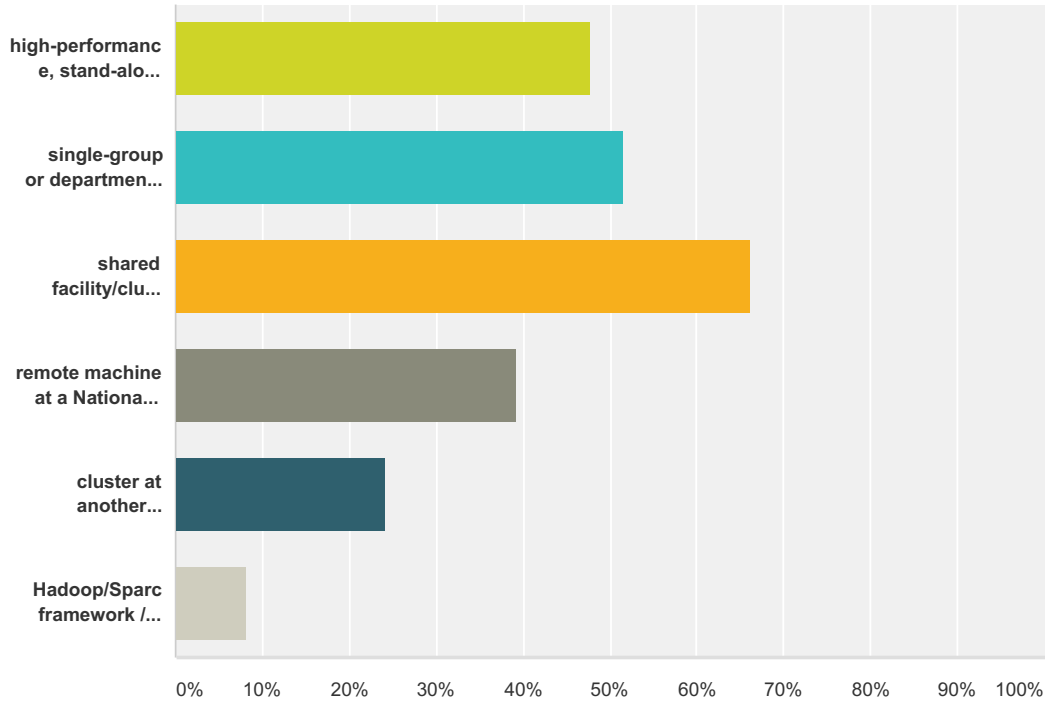
Answered: 301 Skipped: 47



Answer Choices	Responses	
I currently use HPC resources	**63.46%**	191
I would like to use HPC resources in the future	**36.54%**	110
Total		**301**

Q8 I have access to and/or use the following types of HPC resources and environments: [Select all that apply]

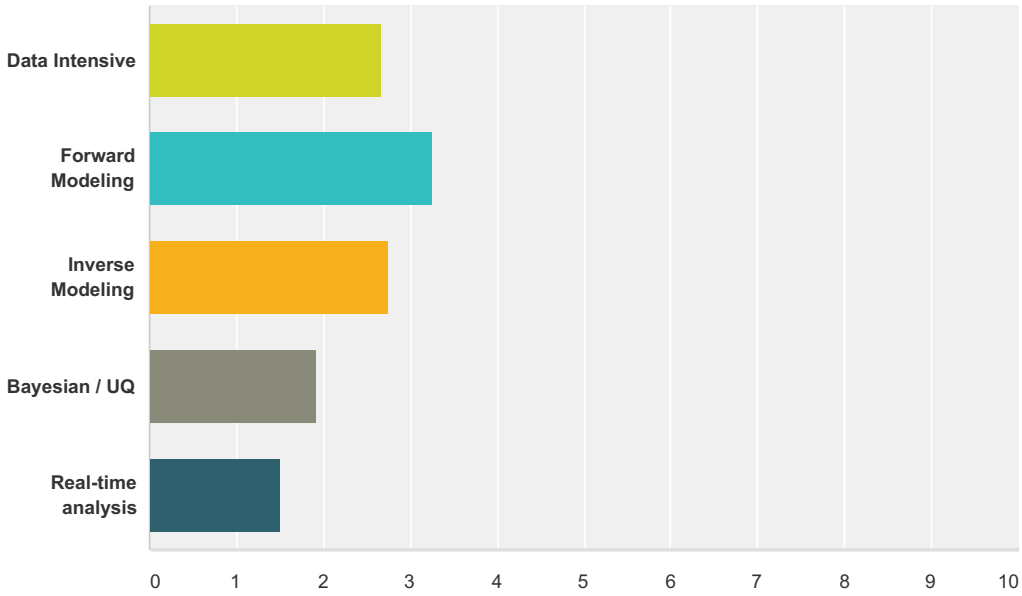
Answered: 186 Skipped: 162



Answer Choices	Responses	
high-performance, stand-alone computer (e.g., 32-core laptop)	**47.85%**	89
single-group or departmental computing cluster.	**51.61%**	96
shared facility/cluster at my institution.	**66.13%**	123
remote machine at a National Facility (e.g. XCEDE, INCITE)	**39.25%**	73
cluster at another institution	**24.19%**	45
Hadoop/Sparc framework / high-throughput computing	**8.06%**	15
Total Respondents: 186		

Q9 For each category, do you use HPC resources:

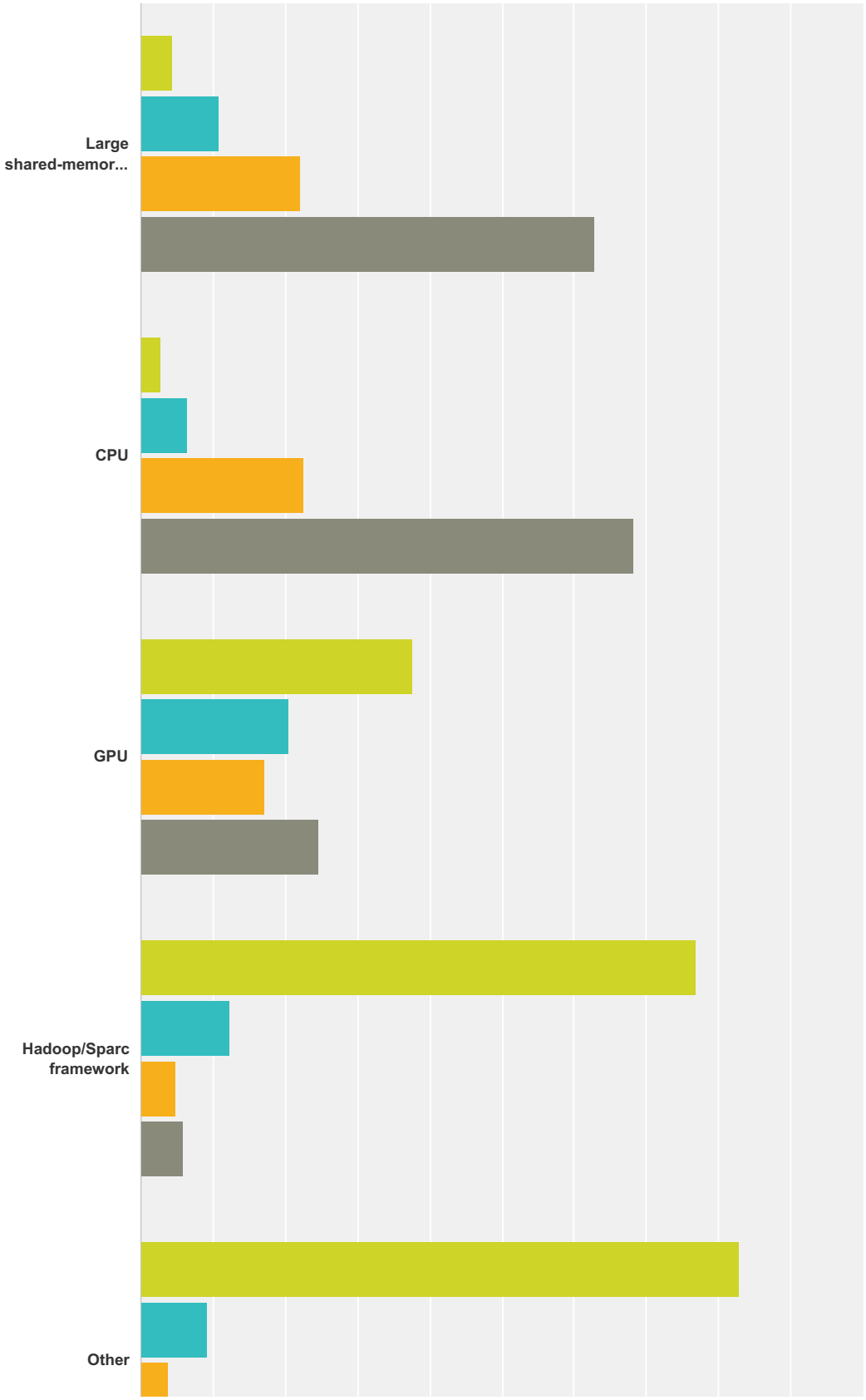
Answered: 186 Skipped: 162

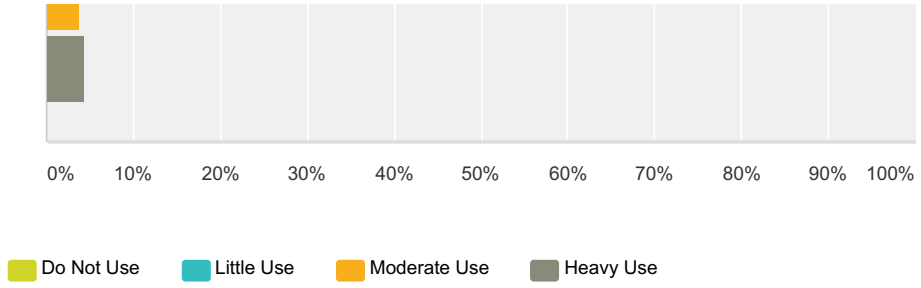


	Do Not Use	Little Use	Moderate Use	Heavy Use	Total	Weighted Average
Data Intensive	**22.58%** 42	**19.35%** 36	**25.27%** 47	**32.80%** 61	186	2.68
Forward Modeling	**7.53%** 14	**12.90%** 24	**25.27%** 47	**54.30%** 101	186	3.26
Inverse Modeling	**23.12%** 43	**16.67%** 31	**22.04%** 41	**38.17%** 71	186	2.75
Bayesian / UQ	**47.31%** 88	**24.73%** 46	**15.59%** 29	**12.37%** 23	186	1.93
Real-time analysis	**67.74%** 126	**19.35%** 36	**8.06%** 15	**4.84%** 9	186	1.50

Q10 What types of HPC do you use or want to use (check all that apply)?

Answered: 186 Skipped: 162

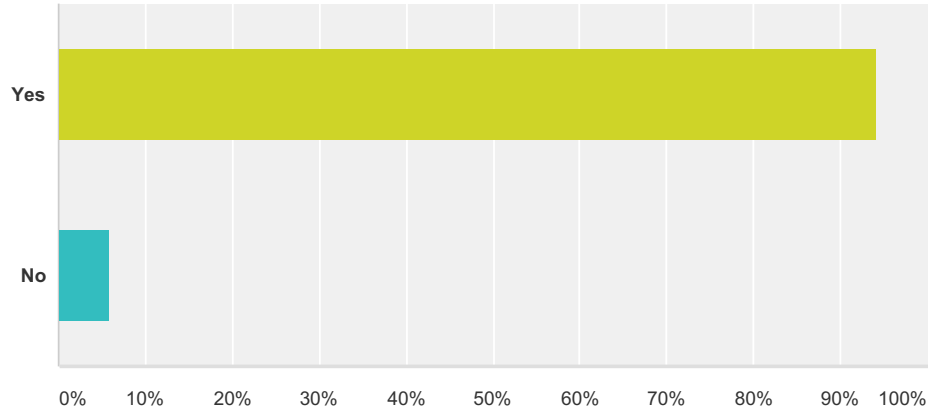




	Do Not Use	Little Use	Moderate Use	Heavy Use	Total
Large shared-memory, Linux cluster	**4.30%** 8	**10.75%** 20	**22.04%** 41	**62.90%** 117	186
CPU	**2.69%** 5	**6.45%** 12	**22.58%** 42	**68.28%** 127	186
GPU	**37.63%** 70	**20.43%** 38	**17.20%** 32	**24.73%** 46	186
Hadoop/Sparc framework	**76.88%** 143	**12.37%** 23	**4.84%** 9	**5.91%** 11	186
Other	**82.80%** 154	**9.14%** 17	**3.76%** 7	**4.30%** 8	186

Q11 Do you use observational data in your research

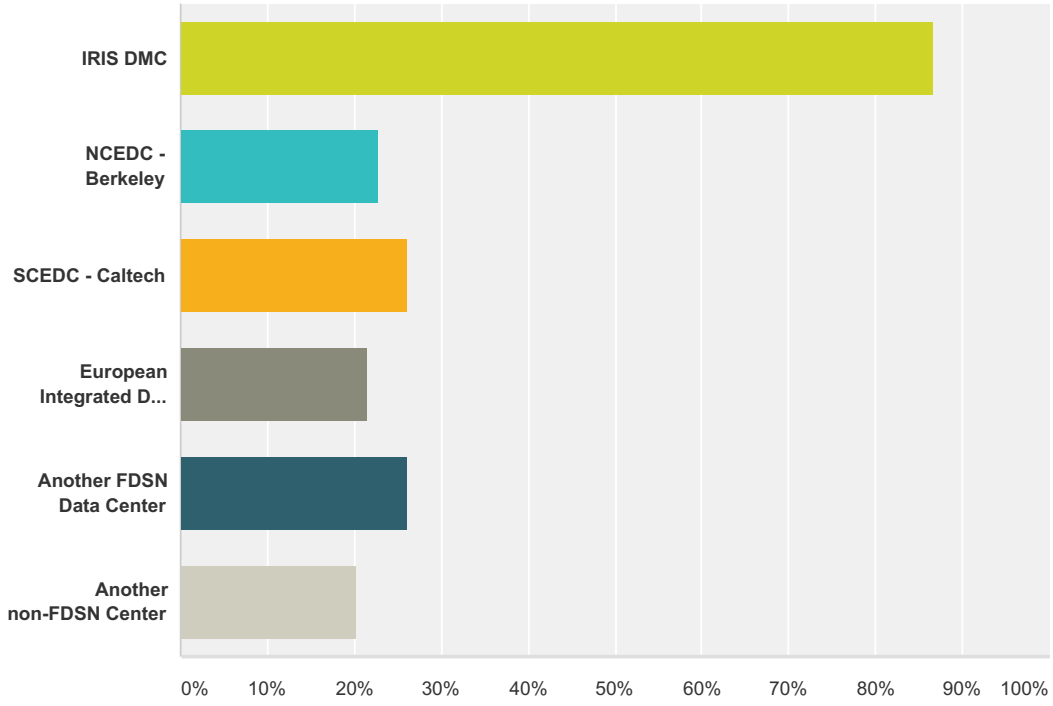
Answered: 186 Skipped: 162



Answer Choices	Responses	
Yes	**94.09%**	175
No	**5.91%**	11
Total		**186**

Q12 I use the following seismological data resources (check all that apply):

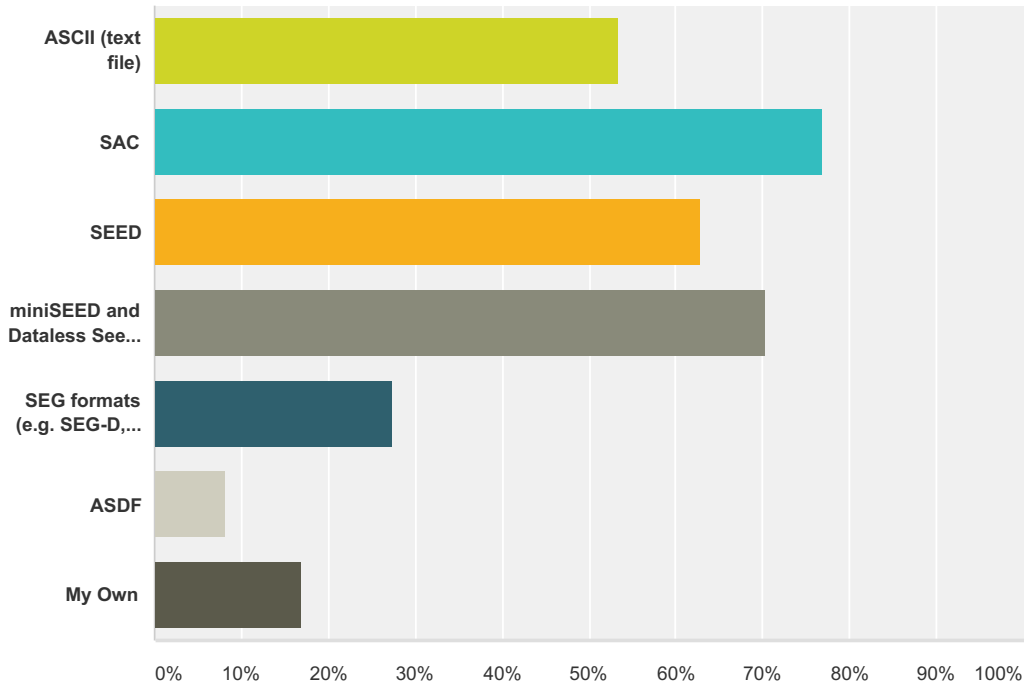
Answered: 172 Skipped: 176



Answer Choices	Responses	
IRIS DMC	**86.63%**	149
NCEDC - Berkeley	**22.67%**	39
SCEDC - Caltech	**26.16%**	45
European Integrated Data Center (EIDA)	**21.51%**	37
Another FDSN Data Center	**26.16%**	45
Another non-FDSN Center	**20.35%**	35
Total Respondents: 172		

Q13 I use the following time-series file formats (check all that apply):

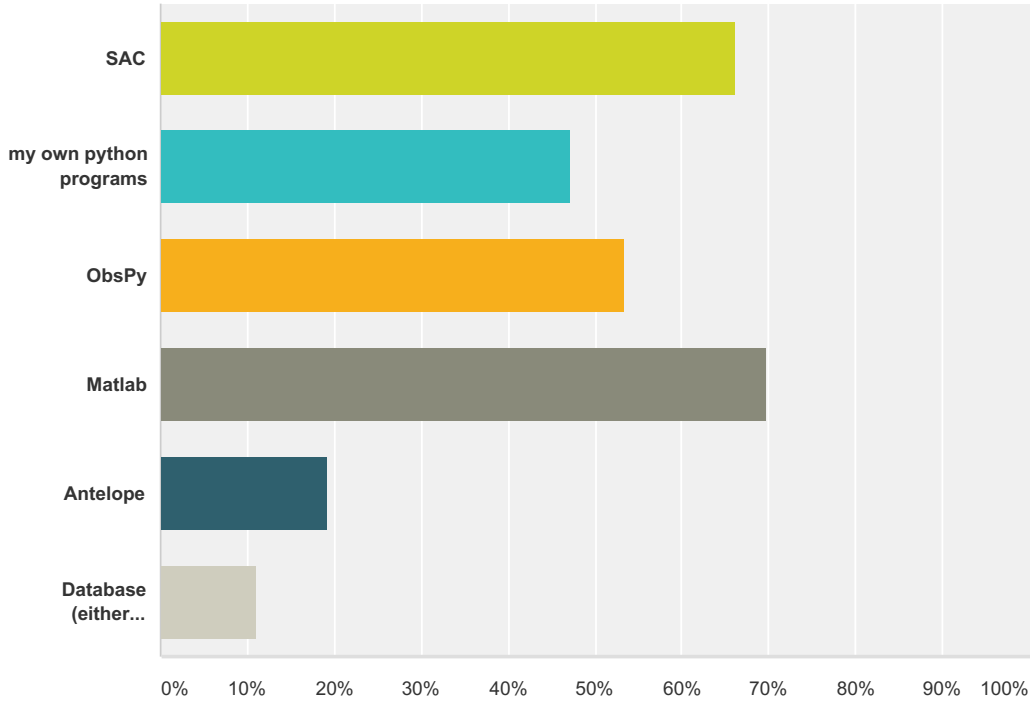
Answered: 172 Skipped: 176



Answer Choices	Responses	
ASCII (text file)	**53.49%**	92
SAC	**76.74%**	132
SEED	**62.79%**	108
miniSEED and Dataless Seed or StationXML	**70.35%**	121
SEG formats (e.g. SEG-D, SEG-Y, SEG-B)	**27.33%**	47
ASDF	**8.14%**	14
My Own	**16.86%**	29
Total Respondents: 172		

Q14 I use the following tools for processing data (check all that apply):

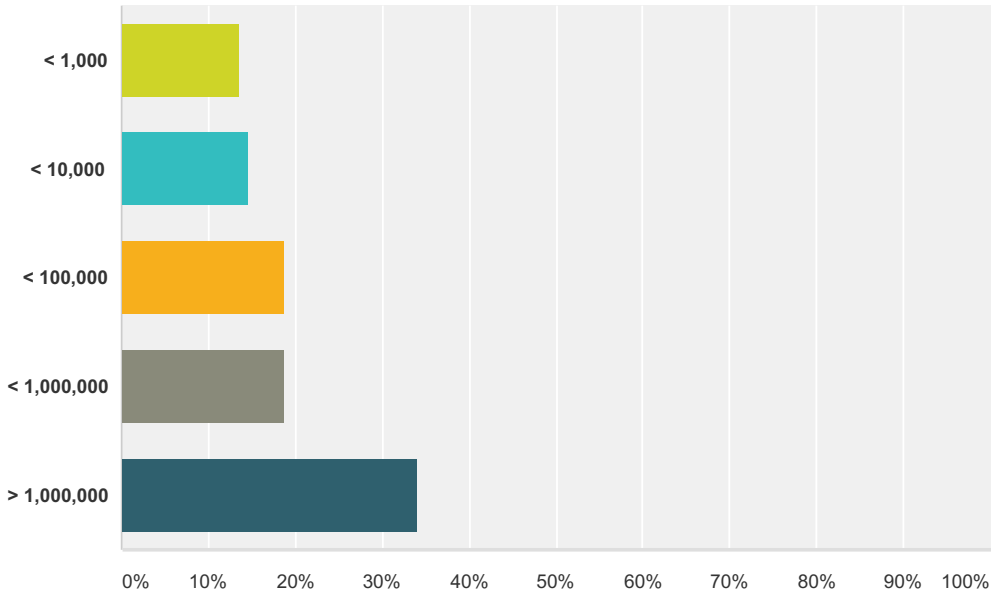
Answered: 172 Skipped: 176



Answer Choices	Responses	
SAC	**66.28%**	114
my own python programs	**47.09%**	81
ObsPy	**53.49%**	92
Matlab	**69.77%**	120
Antelope	**19.19%**	33
Database (either relational or NoSQL)	**11.05%**	19
Total Respondents: 172		

Q15 My research involves the following number of data files:

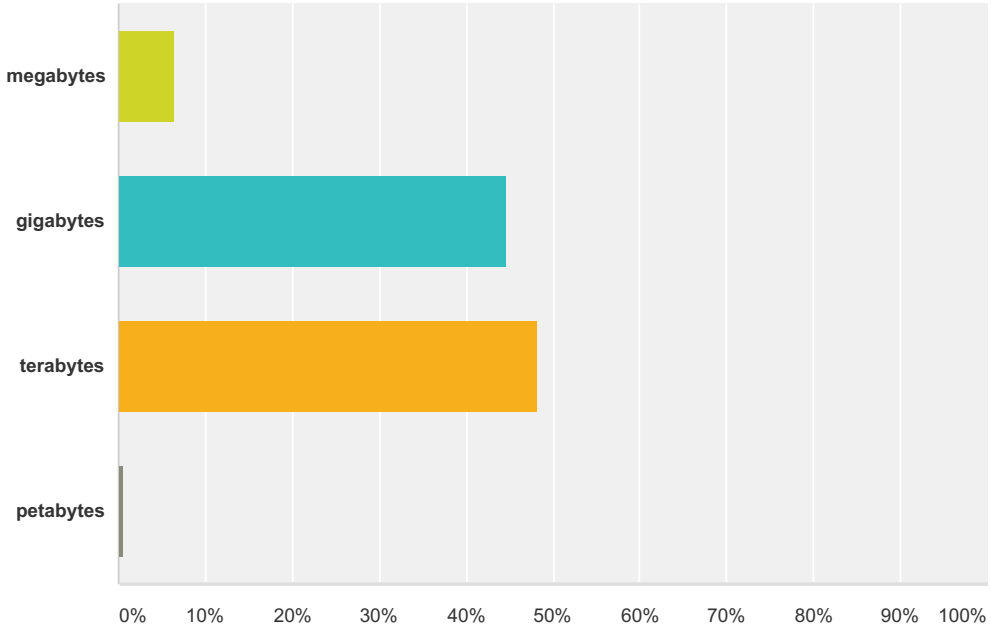
Answered: 170 Skipped: 178



Answer Choices	Responses	
< 1,000	**13.53%**	23
< 10,000	**14.71%**	25
< 100,000	**18.82%**	32
< 1,000,000	**18.82%**	32
> 1,000,000	**34.12%**	58
Total		**170**

Q16 The typical size of my dataset is

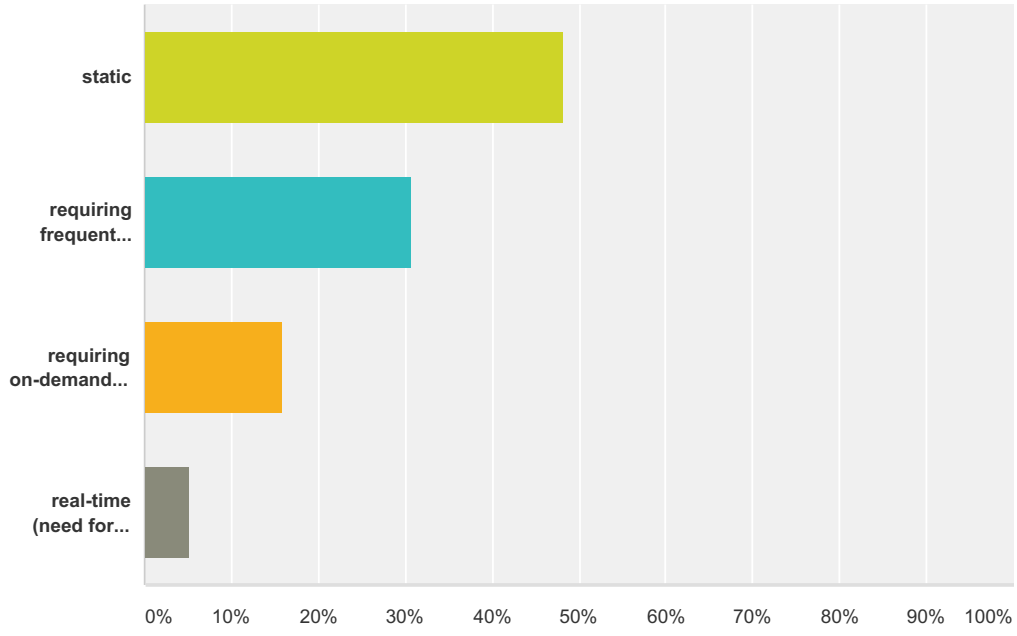
Answered: 170 Skipped: 178



Answer Choices	Responses	
megabytes	**6.47%**	11
gigabytes	**44.71%**	76
terabytes	**48.24%**	82
petabytes	**0.59%**	1
Total		**170**

Q17 In terms of time-dependent changes or frequent reprocessing, my datasets can be characterized as

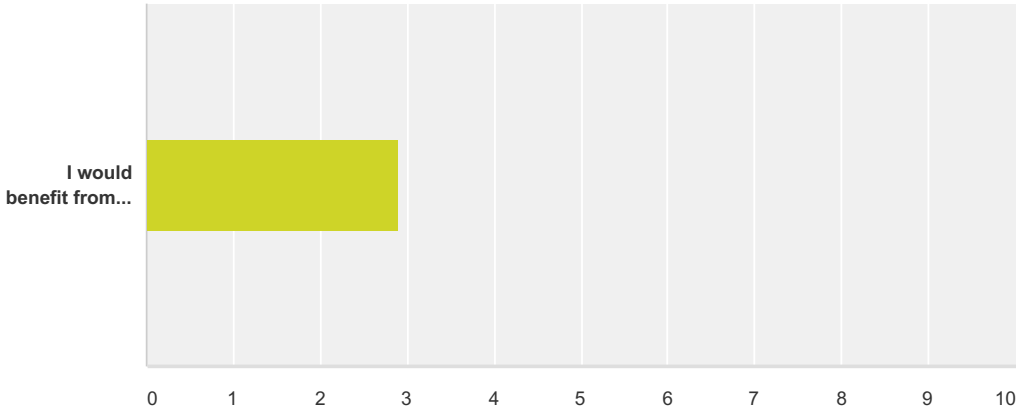
Answered: 170 Skipped: 178



Answer Choices	Responses	
static	**48.24%**	82
requiring frequent changes or processing	**30.59%**	52
requiring on-demand updates (such as servicing and reprocessing data from a large field experiment)	**15.88%**	27
real-time (need for streaming continuous data and metadata)	**5.29%**	9
Total		**170**

Q18 Do you need your data close to HPC Resources?

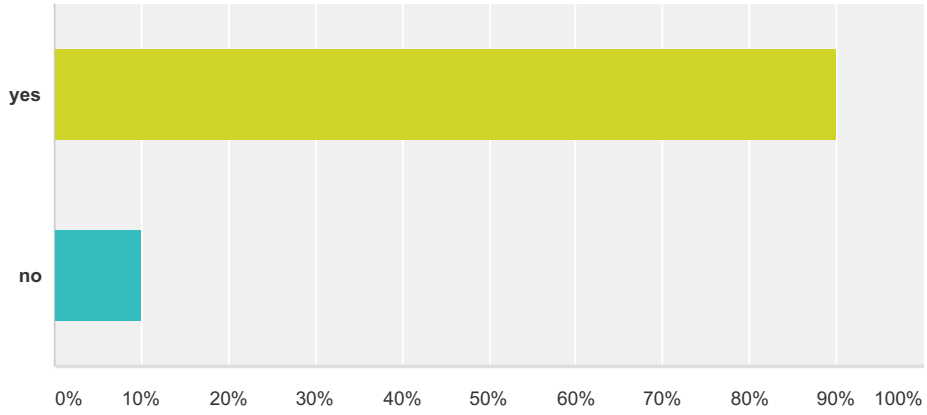
Answered: 170 Skipped: 178



	No	A little	Moderately	A great deal	Total	Weighted Average
I would benefit from having my data in close proximity to HPC resources.	**10.59%** 18	**21.76%** 37	**34.12%** 58	**33.53%** 57	170	2.91

Q19 I use simulations or modeling in my research

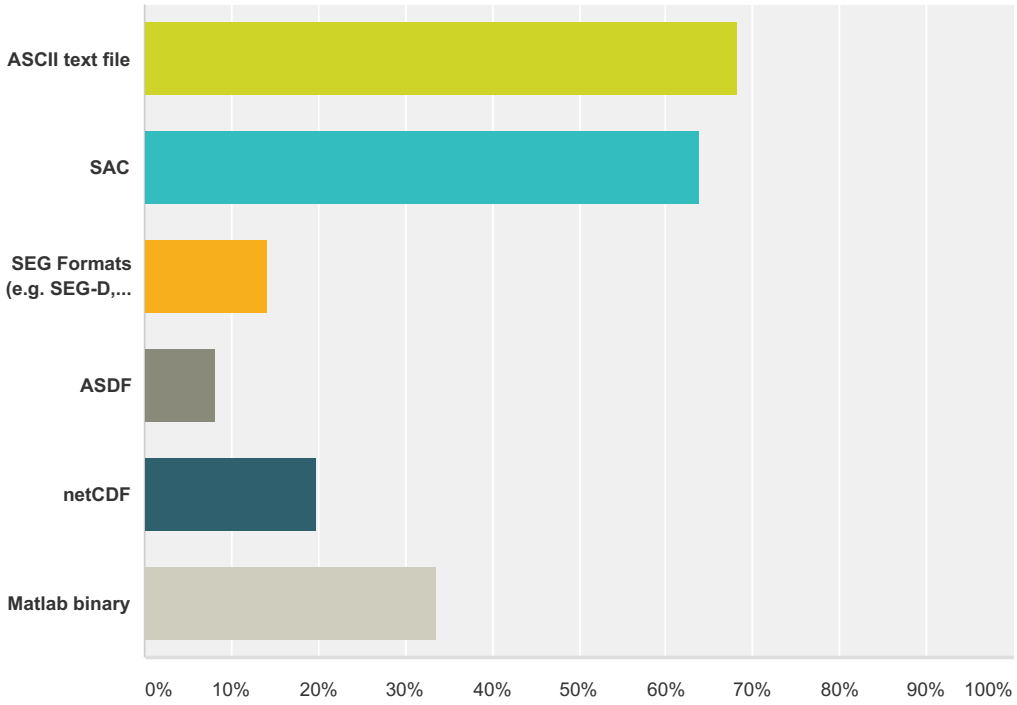
Answered: 181 Skipped: 167



Answer Choices	Responses	
yes	**90.06%**	163
no	**9.94%**	18
Total		**181**

Q20 I use the following seismic data file formats for simulation output (check all that apply):

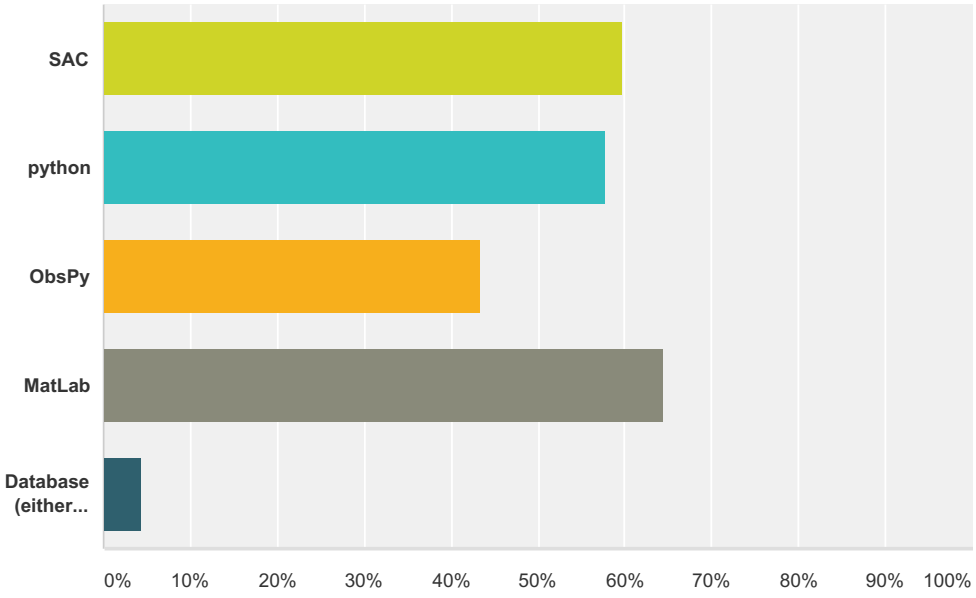
Answered: 161 Skipped: 187



Answer Choices	Responses	
ASCII text file	**68.32%**	110
SAC	**63.98%**	103
SEG Formats (e.g. SEG-D, SEG-Y)	**14.29%**	23
ASDF	**8.07%**	13
netCDF	**19.88%**	32
Matlab binary	**33.54%**	54
Total Respondents: 161		

Q21 I use the following tools for processing time-series simulation output (check all that apply):

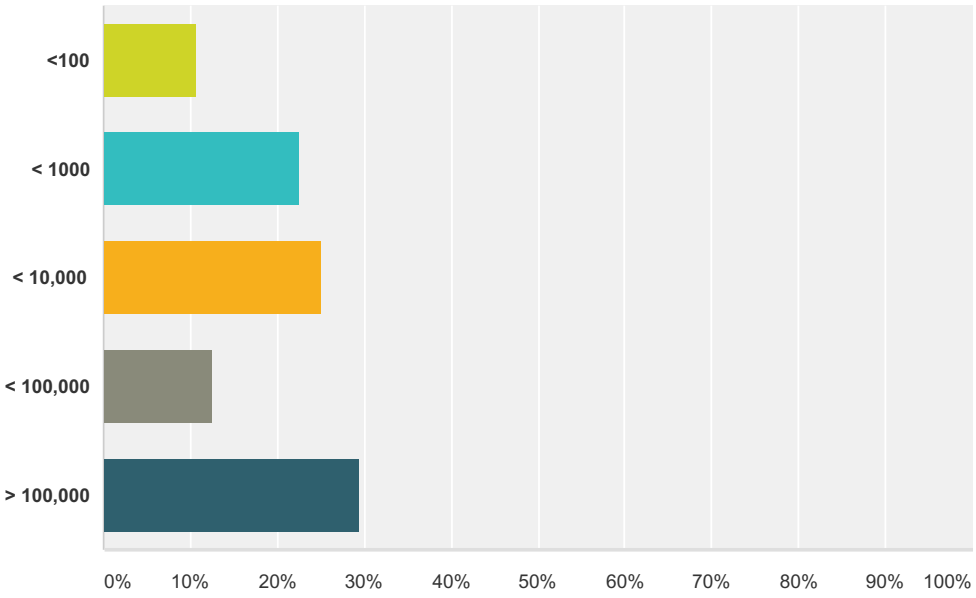
Answered: 161 Skipped: 187



Answer Choices	Responses	
SAC	**59.63%**	96
python	**57.76%**	93
ObsPy	**43.48%**	70
MatLab	**64.60%**	104
Database (either Relational or NoSQL)	**4.35%**	7
Total Respondents: 161		

Q22 My research involves the following number of simulated time series files:

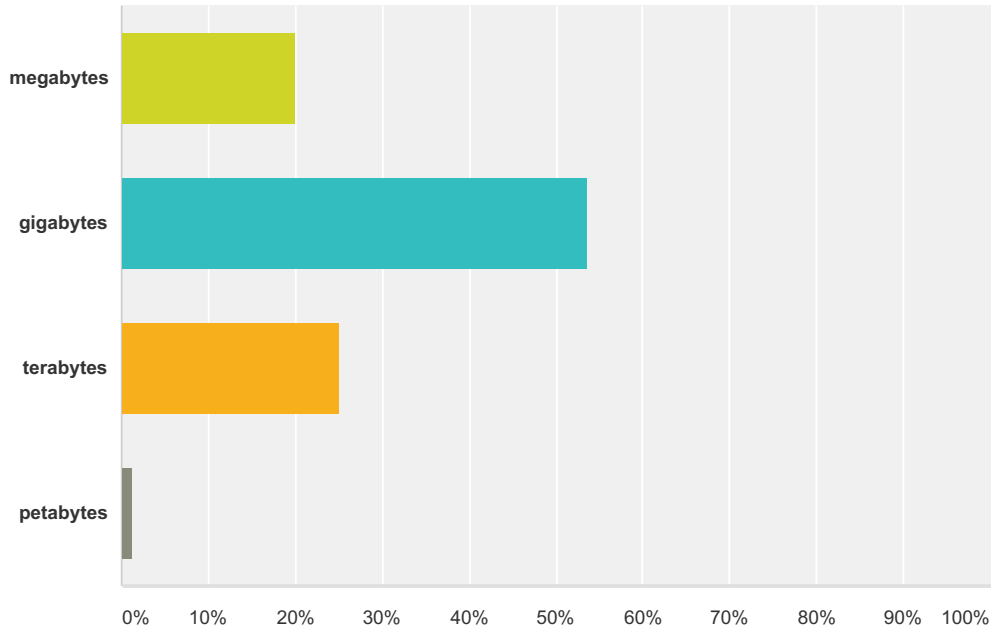
Answered: 160 Skipped: 188



Answer Choices	Responses	
<100	**10.63%**	17
< 1000	**22.50%**	36
< 10,000	**25.00%**	40
< 100,000	**12.50%**	20
> 100,000	**29.38%**	47
Total		**160**

Q23 My research involves simulated time series with a total file size of:

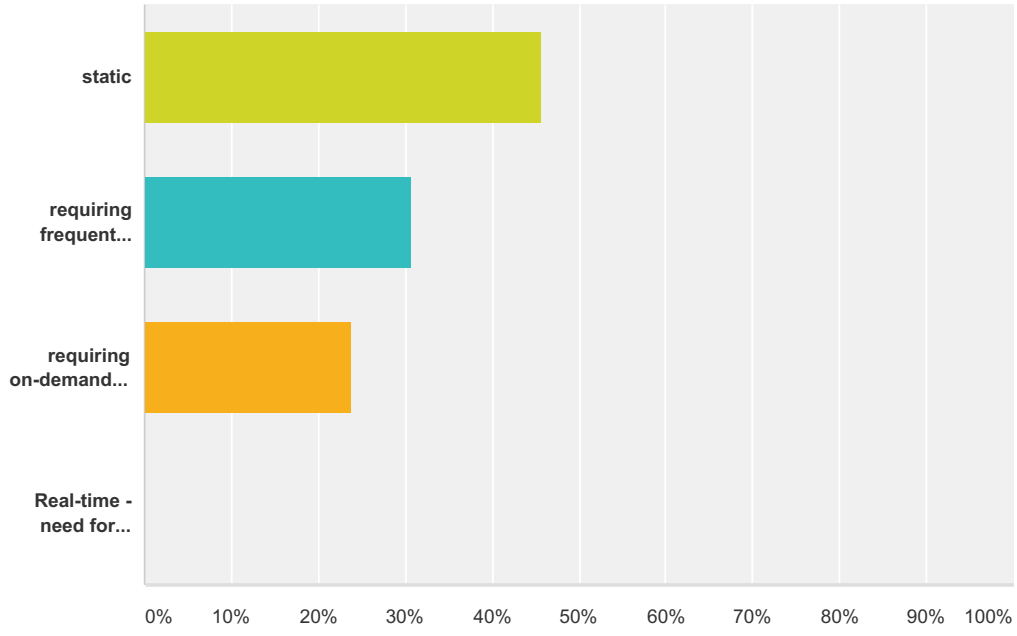
Answered: 160 Skipped: 188



Answer Choices	Responses	
megabytes	**20.00%**	32
gigabytes	**53.75%**	86
terabytes	**25.00%**	40
petabytes	**1.25%**	2
Total		**160**

Q24 In terms of time-dependent changes or frequent reprocessing, my simulated datasets can be characterized as

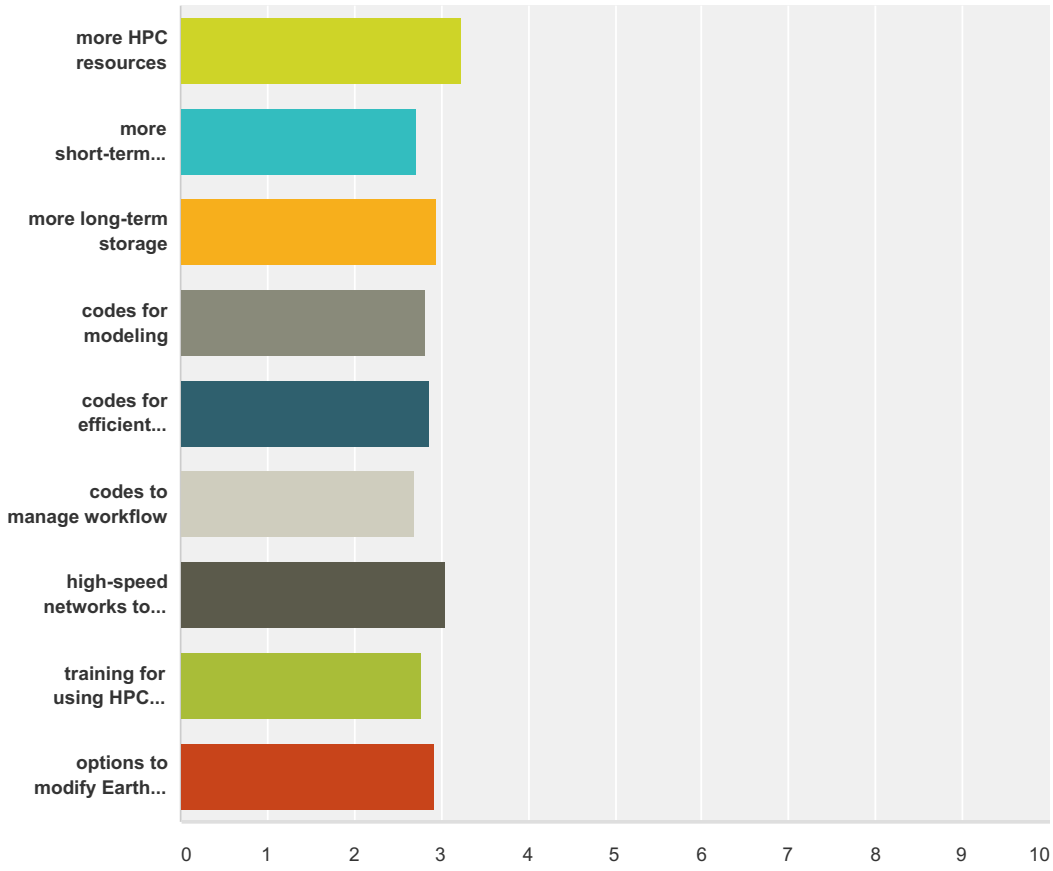
Answered: 160 Skipped: 188



Answer Choices	Responses	
static	**45.63%**	73
requiring frequent changes or processing	**30.63%**	49
requiring on-demand updates (e.g., new velocity model, new seismograms)	**23.75%**	38
Real-time - need for “streaming” or continuous data and metadata ingest and output	**0.00%**	0
Total		**160**

Q25 Please rate your needs for the following:

Answered: 177 Skipped: 171

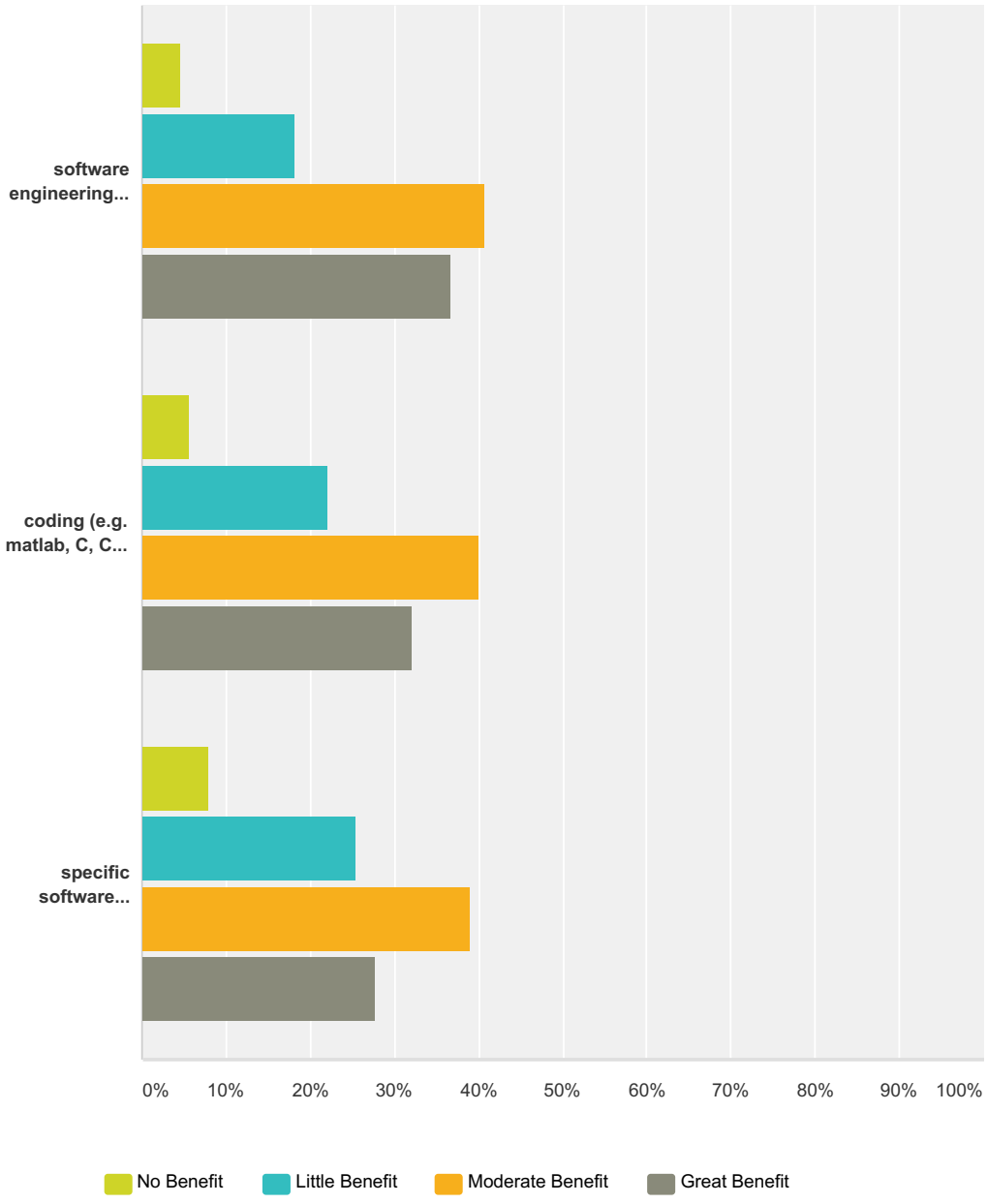


	No Need	Small Need	Moderate Need	Greatest Need	Total	Weighted Average
more HPC resources	**4.52%** 8	**16.38%** 29	**30.51%** 54	**48.59%** 86	177	3.23
more short-term storage	**6.21%** 11	**34.46%** 61	**41.81%** 74	**17.51%** 31	177	2.71
more long-term storage	**7.34%** 13	**24.29%** 43	**33.90%** 60	**34.46%** 61	177	2.95
codes for modeling	**14.12%** 25	**24.29%** 43	**27.68%** 49	**33.90%** 60	177	2.81
codes for efficient processing of data or simulation output	**11.30%** 20	**23.16%** 41	**33.33%** 59	**32.20%** 57	177	2.86
codes to manage workflow	**15.25%** 27	**28.25%** 50	**27.68%** 49	**28.81%** 51	177	2.70
high-speed networks to download data and move simulation output files	**6.21%** 11	**19.21%** 34	**38.98%** 69	**35.59%** 63	177	3.04
training for using HPC resources	**10.17%** 18	**28.25%** 50	**35.03%** 62	**26.55%** 47	177	2.78

options to modify Earth models (e.g., velocity model, topography, internal boundaries)	**8.47%** 15	**24.29%** 43	**33.90%** 60	**33.33%** 59	177	2.92

Q26 My research would benefit from hands-on training in the following:

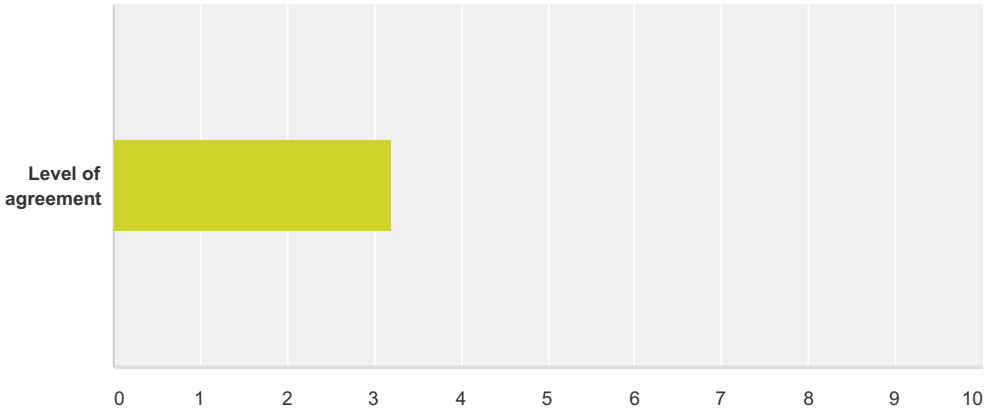
Answered: 177 Skipped: 171



	No Benefit	Little Benefit	Moderate Benefit	Great Benefit	Total
software engineering (e.g. principles of - github, unit testing, documentation, best practices, etc.)	**4.52%** 8	**18.08%** 32	**40.68%** 72	**36.72%** 65	177
coding (e.g. matlab, C, C++, python, scripting	**5.65%** 10	**22.03%** 39	**40.11%** 71	**32.20%** 57	177
specific software training (e.g., paraview, cubit/trelis)	**7.91%** 14	**25.42%** 45	**38.98%** 69	**27.68%** 49	177

Q27 The transition from training to research would be most effective if they were both done in the same computing environment.

Answered: 176 Skipped: 172



	Strongly Disagree	Mildly Disagree	Agree	Strongly agree	Total	Weighted Average
Level of agreement	**1.70%** 3	**10.80%** 19	**53.41%** 94	**34.09%** 60	176	3.20

Q28 Please provide any comments or suggestions. If you have a research problem in seismology that is not addressed within the categories listed above, please describe it.

Answered: 49 Skipped: 299

Q29 supply your contact information.

Answered: 224 Skipped: 124

Answer Choices	Responses	
Name	100.00%	224
Email Address	99.11%	222

Appendix C: Comments from respondents

Below are two detailed responses received over email. These are followed by all the responses to:
Q28. Please provide any comments or suggestions. If you have a research problem in seismology that is not addressed within the categories above, please describe it.

==========

I alluded to this in my formal online response to your Computational needs for seismology/geodesy/infrasound survey. If, as advertised, this survey is partially intended to capture the computational needs of the geodesy and crustal dynamics community, then I found this survey to be insufficient. It is clearly designed by folks doing tomography of various sorts, waveform x-correlations, receiver functions etc. I am concerned that NSF or others would not get a complete or representative assessment based on these questions as they are completely IRIS-centric.

The data volumes associated with the current (not even future) InSAR archives is becoming enormous and that is purely the raw data. If you start including the volumes and computational expense associated with InSAR processing, InSAR time series analysis and the associated modeling then the numbers explode. Further, these issues do not fit easily into the way many of the questions were posed in the survey. If you are interested in big earthquake modeling using geodesy and seismology using the rapidly developing field of Bayesian Inverse approaches (note Inverse Problems are not separate from Bayesian approaches as suggested in the survey), then you need large HPC resources. If you want realistic Green's functions, even for the static problem, you will want larger resources for simulations as well (this at least overlaps well with what the seismology community needs). Similarly, with both InSAR and GNSS data sets, people will want to routinely analyze these at ever increasing sample rates with more sophisticated corrections (e.g., tidal/non-tidal ocean loading effects, atmospheric propagation and loading effects, sidereal filters to reduce multi path etc.) - all of which will strain available resources. If you want to do transient event detection on large geodetic networks or deep InSAR stacks using modern methods, the computational requirements are equally demanding. Note that geodetic time series (GNSS and InSAR) are frequently a very different flavor (formats, file types, size, etc) than seismic time series. And then we have the issue of fully spun up earthquake cycle models that capture inter-,co-,post-seismic time periods with a self-consistent rheological parameterization and that are somehow constrained by real data. This kind of simulation and associated inverse problem is potentially a huge HPC hog.

Another component that is missing is the need to somehow document the importance of supporting mid-scale HPC resources within the universities - as you know, people need to develop tools before going on the large national machines. The national facilities are also frequently impacted such that the wall clock time from submittal to completion is long. Scientists also need to be able to get quick turn around on ideas without waiting for quarterly (at best) resource allotment proposals to be adjudicated by the national HPC facilities. Finally, from an education perspective, mid-tier machines are more conducive to encouraging folks to experiment with HPC. The hurdle of going to a large national resources (which require some kind of track record) is too daunting for many newcomers and thus stifles curiosity-driven developments. At present, getting support from NSF for such department scale resources has become nearly impossible.

I am sure I am missing issues, but I hope you consider the need for more complete representation as you go forwards with your community sounding.

==========

Hi. I've completed the survey but would like to note that while the survey instructions indicate it is to cover non-seismic areas such as infrasound or geodesy, all of the questions specifically pertain to seismic data and analyses. MT data acquisition, distribution and archival is a core activity of IRIS, and I have

responded to the survey but all of my responses relate to MT data and related processing/analysis rather than seismic. The survey may confuse people working with IRIS on non-seismic issues.

==========

HPC is important in seismological research in Indonesia, because there are so big an erthquake data

Quantifying and modeling of site effects can be computationally challenging, I suspect that those were included in the 1D & 3D waveform modeling

I have used HPC resources in the past but am not currently using them. I expect to use them again in the future.
I think these two are also useful: 1) A common place to share research codes that can reproduce results in publication 2) A forum to discuss the usages of these codes

Yes we need nodes to be able to access the internet "on the fly" during the processing. There is already enough clusters that cannot do it.

I am a seismologist from developing country. Through my research life, I had a limit in both computational hardware and software. If this HPC is international available it would be great for me.

I would be excited to attend any HPC workshop/training event specifically designed for the needs of the seismological community

I work on earthquake coseismic slip models using combined waveforms and geodetic data, especially InSAR and GPS. I also study interseismic and postseismic deformation with InSAR and GPS. The huge increase in data volumes available for InSAR in the last year with Copernicus Sentinel-1A and Sentinel-1B is already straining the InSAR analysis systems. The future data volume from NASA's NISAR mission will increase another order of magnitude to about 0.1 petabytes a day. Efficiently processing this amount of data to study global strain will be a challenging computational problem.

Training in HPC and mathematical concepts behind seismic wavefield simulation/inversion will be really helpful

A very computationally expensive process that would be ideal to perform is iterative methods using pre-stack depth migration to converge on the "true" velocity field.

If we're going to teach people HPC we should also teach them basic aspects of coding, algorithms, reproducibility, testing, etc., beyond the very few things that they learn in undergraduate courses.

"Inverse" and "modeling" don't fit well together. Modeling generally seeks solutions, which fit the data, but give little information about resolution. Inverse methods provide good information about uncertainty, but don't necessarily result in useful, specific models.
It would be nice to see more GPU, and GPU+CPU oriented trainings, webinars, tutorials and codes to speed up processing large seismic datas.

Looking forward to study phd in USA in Ambient noise

Large-N seismic (and infrasound) deployments is future.

Distributed Acoustic Sensing (DAS) Parallel codes for Markov Chain Monte Carlo (MCMC)

I recognize that there are problems that need HPC, and good people to pursue them. Often though HPC is used as a "bigger hammer" where thinking more carefully about the problem would be (and for me has been) more productive.

Methods to provide, access, and permanently archive (publish) code referred to in published papers at a permanent, library. Too many codes are ephemeral relative to the published results

Why did you ask twice if I use HPC?

translating from python to C in order to make faster correlations

There is a big hurdle when it comes to transferring code that is written/optimized for a single machine to that of HPC. Computational resources are often available but adapting research code to the HPC environment is a non trivial task that needs more attention in my opinion. For most of my research, I write my own code and I find that it is significantly faster to develop/prototype in Matlab. I've had some success scaling up embarrassingly parallel code to our cluster but I'm not sure I would be able to implement any software that truly required distributed computing.

The license for SAC will hopefully change soon, but it will not be completely open due to restrictions imposed by LLNL. The biggest issue will be a export restriction to embargoed countries. We are able, or willing, to construct an external library of SAC functions that would be worthwhile to HPC and non-HPC users.

I often waste time converting data formats. While SAC files are portable, opening and closing thousands of files and file headers is too inefficient, and I don't have the packages / knowledge to work directly out of SEED or Antelope databases. This usually means I convert my data to HDF5, though I would appreciate further packages or tools to work with flexible file formats and headers.

Safari on my Mac froze twice on me when I tried to fill in my contact information on this page! I had to redo the survey using Firefox. SurveyMonkey should avoid whatever special software they are using on this page.

InSAR processing has many computation limitations (time and storage) for new data sets. The same problems exist as the seismology needs.

I spend some time writing my own codes, it wasn't clear if that was included in 'learning and compiling code'.

I mostly perform forward modeling earthquake simulations that require access to HPC resources, long and short-term storage, and the ability to simulation results between computers. These topics remain issues even considering the access I have at a national lab. Additionally, I struggle with visualization of my results. A common problem I face is that with HPC, and highly parralized code, I am able to full explore a vast parameter space. However, understanding and visualizing the results in a meaningful way poses quite a challenge. The most important thing for me now is how to manage data (includes store, transportation, processing)

This survey is incomplete for the non-seismologist. While the email suggests it applies to geodesists, I dont think one was consulted in designing the questions.

I think that improving the ability for researchers at all sorts of institutions to do routine elastic 3-d simulations in realistic, complex structures (and for educational purposes) would one of the most important things that we can do to advance seismology.

One of the larger problems I have with HPC computing is adapting codes developed for smaller data sets that did not require HPC computing to larger data sets that do require these resources. Having access or, if access already exists, clearer access to information on how to update codes would be really really useful.

My main focus is infrasound, however many of the HPC problems we face are similar to those in seismology.

Infrasound

I am interested in cross-correlation and seismic interferometry and would like to test different codes to generate outputs for different applications. I would also like easy access to instrument corrected seismograms. Although IRIS provides all the tools necessary to down load raw data and remove instrument response, in codes like SAC there is more than one way to do this, and validating the response is correctly removed is one of my biggest challenges and prefer downloading from PEER or COSMOS as instrument corrected seismograms are available.

Agree that additional training on different tools, workflows, etc would be useful for the community, particularly for students and early career scientists.

First of all, thank you very much for all your efforts. In our workflow, we have "old" codes (Fortran 77/90), and now I am trying to implement/use modern data structures and scalable algorithms and routines for our applications. You already asked about "netCDF", but I am also thinking about libraries such as PETSc. I strongly believe that having well-tested codes and data structures can accelerate my research, but more importantly, having libraries of routines and scripts can be beneficial, for example, ObsPy and PETSc. Therefore, I would suggest well-defined work-groups to train and develop codes/workflows for specific applications based on these libraries. All the best, Kasra Hosseini

I have problem in obtaining data and software.

Great survey, thanks for doing it!
We don't have computer or laptop with high coprocessor in my Institution and the internet cannot work good

HPC training requires training, which is done for the most part at supervisor-student level. Late starters like myself (early careers) who are not trained in HPC at the PhD level do not have enough opportunities to learn these methods. This is important because these HPC codes (3D wave simulations) have become so very complicated over the years.

Training in HPC for people of all career levels would be great, along ways to build virtual communities for researchers that are the sole HPC user in their department or college.

IRIS DMC programmers should be part of the IRIS “instrument” pool, such that a researcher can "borrow" them for a project and programming knowledge transfer.

Good job in getting this survey up. Yes it would be great to put the massive data achieved at IRIS DMC (and other data centers) to easily accessible HPC so that lots of simple computation can be done.

Training in HPC needs to be sustained for support and construction of a strong cohort.

The survey ignored non-seismic data and codes that are supported by IRIS. Please replace all references to "seismic" in my responses with "magnetotelluric".

Development of new theory
Archiving and managing/working with huge datasets (~50 Terabytes) is a major challenge too.

A lot of time is spent just trying to figure out how best to tackle a problem - sorting through reams of literature on seismic methods.

==========

Appendix D: Webpage for the IRIS High Performance Computing and Seismic Data Working Group

https://www.iris.edu/hq/about_iris/governance/hpcwg (last accessed February 2017)

High Performance Computing and Seismic Data Working Group

Charge to the IRIS HPCWG

As the volume of archived seismic data increases, the need to have these data processed in new and more powerful computational systems has become more important. The High Performance Computing and Seismic Data Working Group (HPCWG) is a working group reporting to the IRIS Data Services Standing Committee (DSSC) that will focus on the use of the "Big Data" available in the IRIS DMC storage systems within high performance computing environments. The HPCWG will address data-driven seismological research requiring HPC resources, either for data processing or for simulation-based data assimilation.

HPCWG Tasks

1. Identify categories of data-driven seismological problems needing HPC
2. Identify data accessibility methods and the performance needed to support the processing of seismic data in an HPC environment
3. Define specific scenarios ("use cases") requiring the use of HPC environments on the data at the IRIS DMC
4. Recommend formats for observed seismic data that are suitable for use in an HPC environment.
5. Identify possible partnerships between IRIS and HPC environments within the NSF (Computer and Information Science and Engineering [CISE], Extreme Science and Engineering Discovery Environment [XSEDE]), Department of Energy (DoE), or other agencies that have resources that can be shared with the academic research community.
6. Report activities of the working group in advance of each DSSC meeting.

Frequently Asked Questions
Travel Tips
Image Gallery
Contact Us
Committee

Membership

- Carl Tape, Chair and member of the DSSC
- Heiner Igel, Ludwig Maximilian University of Munich (LMU)
- Weisen Shen, Washington University
- Jeroen Tromp, Princeton University
- Felix Waldhauser, Columbia University
- Omar Ghattas, University of Texas

Ex-Officio:

- Arthur Rodgers, Lawrence Livermore National Lab (LLNL)
- Stan Ruppert, Lawrence Livermore National Lab (LLNL)
- Carene Larmat, Los Alamos National Lab (LANL)
- Scott Klaske (ORNL)
- Lorraine Hwang, Computational Infrastructure for Geodynamics
- Tim Ahern, IRIS

Term: 01 April 2016 to 30 July 2017

All meetings will take place virtually, using phone or internet communication.