Effective Data Science Communication

ICS 491

Milestone #4: IRB Questions

- Your assignment will be to answer each question. Each question requires no more than a single sentence, though you are welcome to provide more than 1 sentence per question if you would like.

- If your project involves an already existing dataset, you should describe the process that you would follow if you were to have collected the data. Alternatively, you can describe your plan to collect follow-up data that would enhance the analyses you are performing this semester.

- ***Do not*** write your protocol about the publicly available dataset. Instead, you should describe either the process of collecting that dataset, or a similar dataset that you would collect for a follow-up study.

Why work on an IRB protocol in this class?

- The IRB protocol forces you to think about the Methods section of your project.
- Parts of the protocol related to analysis can be copy-and-pasted into the Methods section of your final paper.
- Other parts can be copy-and-pasted into the Discussion and Future Work section of your final paper.

Final Paper Progress So Far

Introduction	**Milestone #1 (with modifications and revisions based on comments + discussions)**
Related Work	**Milestone #2 (with modifications and revisions based on comments + discussions)**
Methods	**Milestones #3 and 4 (with modifications and revisions based on comments + discussions)**
Results	
Discussion	

Final Paper Progress So Far

Introduction	**Milestone #1 (with modifications and revisions based on comments + discussions)**
Related Work	**Milestone #2 (with modifications and revisions based on comments + discussions)**
Methods	**Milestones #3 and 4 (with modifications and revisions based on comments + discussions)**
Results	**Milestone #5 (with modifications and revisions based on comments + discussions)**
Discussion	

IRB Questions

Proposed Start Date: **Proposed End Date:**

1. **Summary**

a) **Provide a brief summary of the scope of work of this project, using non-technical terms that would be understood by a non-scientific reader. This summary should be no more than 200 words.**

2. **Purpose**

a) **Describe the purpose for the proposed project as well as the hypotheses/research questions to be examined.**

b) **What do the investigators hope to learn from this project?**

3. **Procedures**

a) **Describe in chronological order of event(s) how the activities will be conducted, providing information about all procedures (e.g. interventions/interactions with subjects, data collection, photographing, audio and video recording), including follow up procedures.**

i) **Be sure to identify what procedures are experimental and what are standard of care or established practice for the condition/situation.**

IRB Questions

b) **Explain who will conduct the procedures and where and when they will take place. Indicate the frequency and duration of visits/sessions as well as the subject's total time commitment for the study. Include how the data will be collected (i.e. in person or online).**

i) **Indicate that the instruments used are in the public domain or provide appropriate documentation of permission to use each scale.**

c) **For school-based activities where class time is used, describe in detail the activities planned for non-subjects and explain where both subjects and nonsubjects will be located during the activities.**

d) **State if deception will be used. If so, provide a rationale and describe debriefing procedures. Submit a debriefing script in attachments section**

e) **Will audio or video taping of individuals occur? Will photographs of individuals be taken? Describe what will become of the tapes/photographs (e.g., shown at scientific meetings, erased, etc.).**

f) **Will the proposed research involve the use of existing data/specimens? If so, check all that apply:**

- ☐ i. **The research involves data from publicly available sources**
- ☐ ii. **That data will be recorded by the investigator in such a manner that subjects cannot be identified.**
- ☐ iii. **Any link to identifying information has been destroyed**
- ☐ iv. **N/A**

IRB Questions

4 a. **Relevant Background: Discuss the present knowledge, appropriate literature and rationale for conducting the research. Include the rationale for the selected subject population.**

b. **Describe the statistical methods of the research and plans for analysis of the data (i.e. planned statistics, justification of sample size, etc.).**

c. **Alternative Procedures. Describe any alternatives to participating in the research. (e.g., standard of care treatment, etc.). Any standard treatment that is being withheld must be disclosed. This information must be included in the consent form.**

d. **Will subjects be followed after their active participation is complete?** Yes No

If yes, explain why and describe how:

e. **Will subjects have access to the study treatment/procedure after completing the study?** Yes No

If yes, explain why and describe how:

IRB Questions

5. **Subject Population**

a) **How many subjects do you intend to enroll and/or how many subject records do you intend to access?**

i. **At this site**

# of subjects

# of records

ii. **At all sites** ☐ **N/A**

# of subjects

# of records

b) **Inclusion and Exclusion Criteria (e.g., Participants must have 20/20 vision, Participants must be 30-45 years of age, etc.)**

i. **Identify inclusion criteria.**

ii. **Identify exclusion criteria.**

c) **What is the rationale for studying the requested group(s) of participants?**

d) **If your participants include pregnant women, human fetuses, neonates, children, adults with diminished capacity, and/or prisoners, describe the protocol-specific safeguards used to protect the rights and welfare of this study population:**

IRB Questions

e) **Provide a clear compelling rational for excluding women, minorities, or minors, if they are intentionally excluded from the research.** ☐ **N/A**

f) **State if any of the subjects are students, employees, or laboratory personnel. Please explain how subjects will be protected from coercion and undue influence** ☐ **N/A**

g) **Please describe the expertise you have, or have access to, which prepares you to conduct research in this location and/or with this subject population, including specific qualifications (e.g., relevant coursework, background, experience, and training). Also, explain your knowledge of local community attitudes and cultural norms and cultural sensitivities necessary to carry out the research (e.g., differences with U.S. culture).**

IRB Questions

6. Recruitment Process:

a) Describe the step-by-step procedures for identifying and recruiting potential research subjects or requesting pre-existing data or materials.

- **List any specific agencies or institutions that will provide access to prospective subjects.**
- **Identify who will contact prospective subjects and how.**

IRB Questions

c) Planned Recruitment Materials/Methods:

- ☐ N/A
- ☐ Phone Scripts
- ☐ Television ads
- ☐ Letters to prospective subjects
- ☐ Oral Scripts
- ☐ Internet ads/postings
- ☐ Face to face interactions
- ☐ Other (please specify):
- ☐ Flyers/posters
- ☐ Letters to providers/schools/organizations
- ☐ Newspaper ads
- ☐ Radio ads
- ☐ PowerPoint presentations
- ☐ Email
- ☐ UH Subject Pool

*(All advertising must be submitted for review in its final printed/recorded form)

Note: Attach copies of ALL recruitment materials in the Attachment Section

7. Subject Compensation and Costs:

a) Will subjects receive compensation for participation? Yes No

Total amount (in dollars or equivalent)

b) Form of Compensation:

- ☐ Cash
- ☐ Check
- ☐ Gift card/certificate
- ☐ Other (please specify):
- ☐ Voucher
- ☐ Course/extra credit
- ☐ Reimbursement only

c) Describe the remuneration plan (Include when subjects will be paid, whether payment will be prorated and whether a 1099 will be issued.)

d) If extra course credit is offered be sure to address the alternative means by which students can accrue extra course credit should they not wish to participate in the study.

IRB Questions

8. Risks

US Department of Health & Human Services (HHS) Regulations define a subject at risk as follows: "...any individual who may be exposed to the possibility of injury, including physical, psychological, or social injury, as a consequence of participation as a subject in any research, development, or related activity which departs from the application of those accepted methods necessary to meet his needs, or which increases the ordinary risks of daily life, including the recognized risks inherent in a chosen occupation or field of service."

a) PI's evaluation of the overall level of Risk. (Please check one: minimal or > minimal.)

- ☐ Minimal (everyday living)
- ☐ > Minimal (greater than everyday living)

b) Describe all known risks or discomforts associated with study procedures whether physical, psychological or social (e.g., pain, stress, invasion of privacy, breach of confidentiality) noting probability and magnitude of potential harm. Specify the risk(s) associated with each research procedure or test.

c) Describe the procedures or safeguards in place to protect against or minimize potential risks (e.g., referral to psychological counseling resources).

d) How will subjects be assessed for adverse events?

e) Is there a plan to monitor study data for subject safety? Yes No

If yes, discuss who will monitor the study data and describe the monitoring plan:

IRB Questions

9. Benefits

a) Discuss any potential benefits that would justify involvement of subjects in this study. Compensation is not considered a benefit.

i. Direct benefits to subjects (if applicable)

ii. Indirect benefits to society

b) Explain how the potential benefits justify the potential risks involved in participation in this research.

IRB Questions

10

a) If information derived from the study will be provided to the subject's personal physician, a government agency, or any other person or group (other than the research team), describe to whom the information will be given and the nature of the information, if applicable.

b) Explain how you will protect subjects' privacy.
Note: Privacy refers to persons and their interest in controlling the access of others to themselves. For example, based on their privacy interest's people want to control:
- The time and place where they give information.
- The nature of the information they give.
- The nature of the experiences that are given to them.
- Who receives and can use the information.

For example, persons might not want to be seen entering a place that might stigmatize them, such as a pregnancy-counseling center that is clearly identified as such by signs on the front of the building. Please keep this definition in mind as you respond to this item.

c) Describe how you will maintain the confidentiality of subjects' information.
Note: Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others (without permission) in ways that are inconsistent with the understanding of the original disclosure. Please keep this definition in mind as you respond to this item.

d) Who will have access to study records or specimens? (Please identify specific team members by name.)

IRB Questions

e) **If you plan to use existing data, records or specimens, what is the source of the data/records/specimens, and how will you access them?**
NOTE: "Existing" means data or specimens collected (i.e., on the shelf) prior to the IRB application submission. It includes data or specimens collected for research and non-research activities.

f) **How will subjects be asked to provide their permission for release of identifiable data collected as a part of this proposed research (e.g., pictures, recordings, responses to research questions), now or in future? Explain and include appropriate statements in consent materials.**

g) **If using existing data/biological specimens, will the researchers have access to a code linking the data to personally identifiable information?**

h) **If the data is coded, explain where the key to identifiers will be stored, how it will be protected, and who will have access to it.**

i) **Explain why, where, in what format, and for how long data/specimens will be retained.**

IRB Questions

11. **Consent Information**

11 a & b only apply to applications where no consent document is provided to research subjects.

a) **How will subjects be informed of procedures, intent of the study, and potential risks to them?**

b) **How will subjects be informed they may withdraw at any time without penalty?**

See sample consent forms at www.hawaii.edu/researchcompliance/templates

c) **Click Add to answer consent process questions and provide the consent forms.**

Note: Attach, in the Attachments Section, written and/or verbal instructions the subject will receive.

Effective Data Science Communication

Thoughts?

Effectively Visualizing Results and Analyses

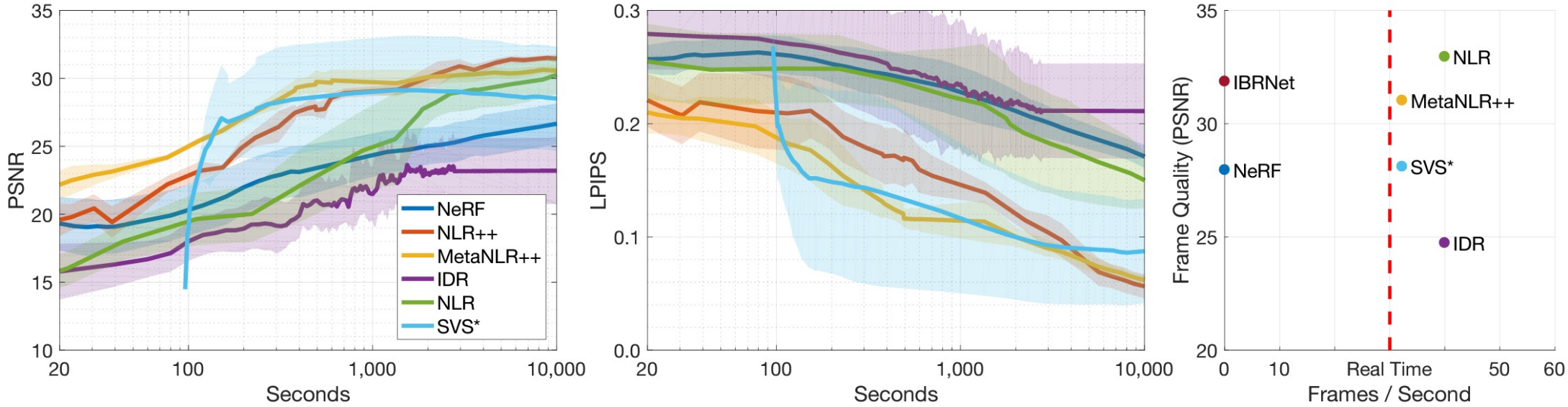


Figure 2: We demonstrate that at all training-times, MetaNLR++ is comparable to or outperforms all competitive representation learning methods, including both neural volumetric and surface representations in PSNR↑ (left) and LPIPS↓ [86] (center). We also plot the render time versus converged image quality, showing that MetaNLR++ can generate high-quality frames at real-time rates (right). The shaded area around each line represents the standard deviation of the method across three DTU scenes.

Figure 3: Novel views synthesized using various methods after a set training time. MetaNLR++ outperforms other surface and volume representation methods, especially for a training time budget on the order of minutes, and does not sacrifice quality of the final converged result.

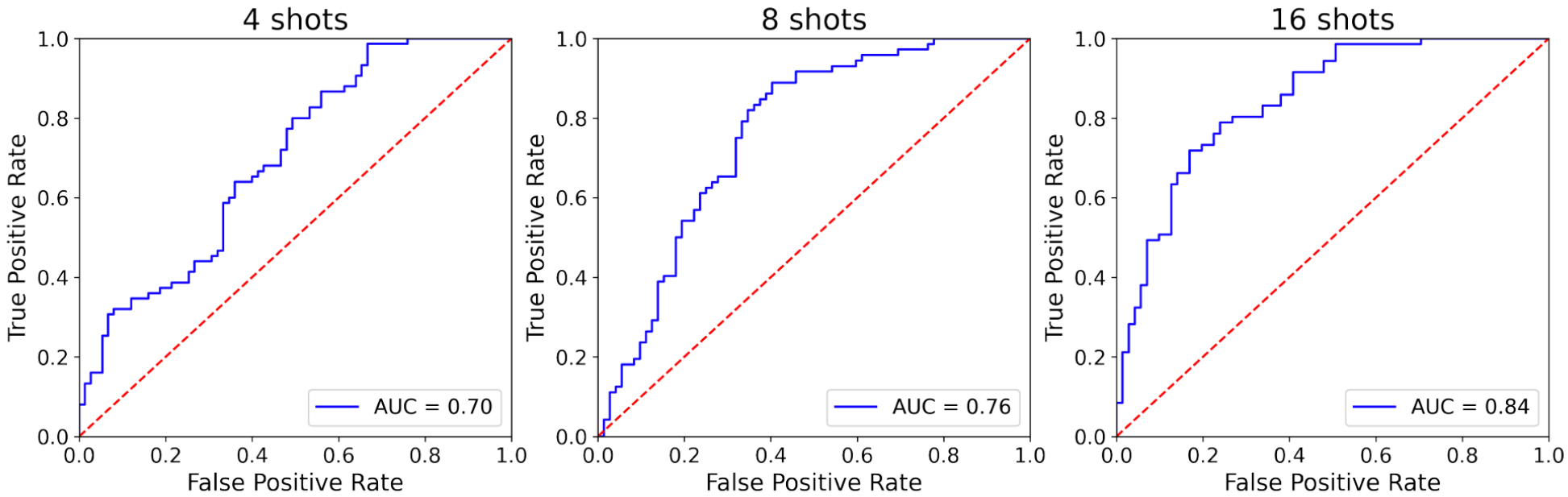


Figure 2: Results on the same/different experiments. Each column shows the results of classifying pairs of dataset with a certain number of shots.

Table 2: Ablation results.

Model	Mean Acc.
Grad2Task w/ Gradients	**45.99**
ProtoNet Longer Training	45.10
Grad2Task w/ X	45.66
Grad2Task w/ X&Y	45.16
Grad2Task Adapt All	44.57
Grad2Task w/ Pretrained TaskEmb	45.68
Hypernetwork	44.79

Furthermore, we find that deep projection head not only reduces the differences among different generalized contrastive losses, but has a similar effect for batch size. With proper learning rate scaling across batch sizes (e.g. square root scaling with LARS optimizer [21]), the impact of batch size on representation quality is small. Table 2 demonstrate this phenomenon for the standard contrastive loss, and more results on other losses can be found in Appendix A.3.

Table 2: Linear eval accuracy of ResNet-50 on ImageNet.

Projection head	Batch size	Epoch 100	200	400	800
2 layers	512	65.4	67.3	68.7	69.3
	1024	65.6	67.6	68.8	69.8
	2048	65.3	67.6	69.0	70.1
3 layers	512	66.6	68.4	70.0	71.0
	1024	66.8	68.9	70.1	70.9
	2048	66.8	69.1	70.4	71.3
4 layers	512	66.8	68.8	70.0	70.7
	1024	67.0	69.0	70.4	70.9
	2048	67.0	69.3	70.4	71.3

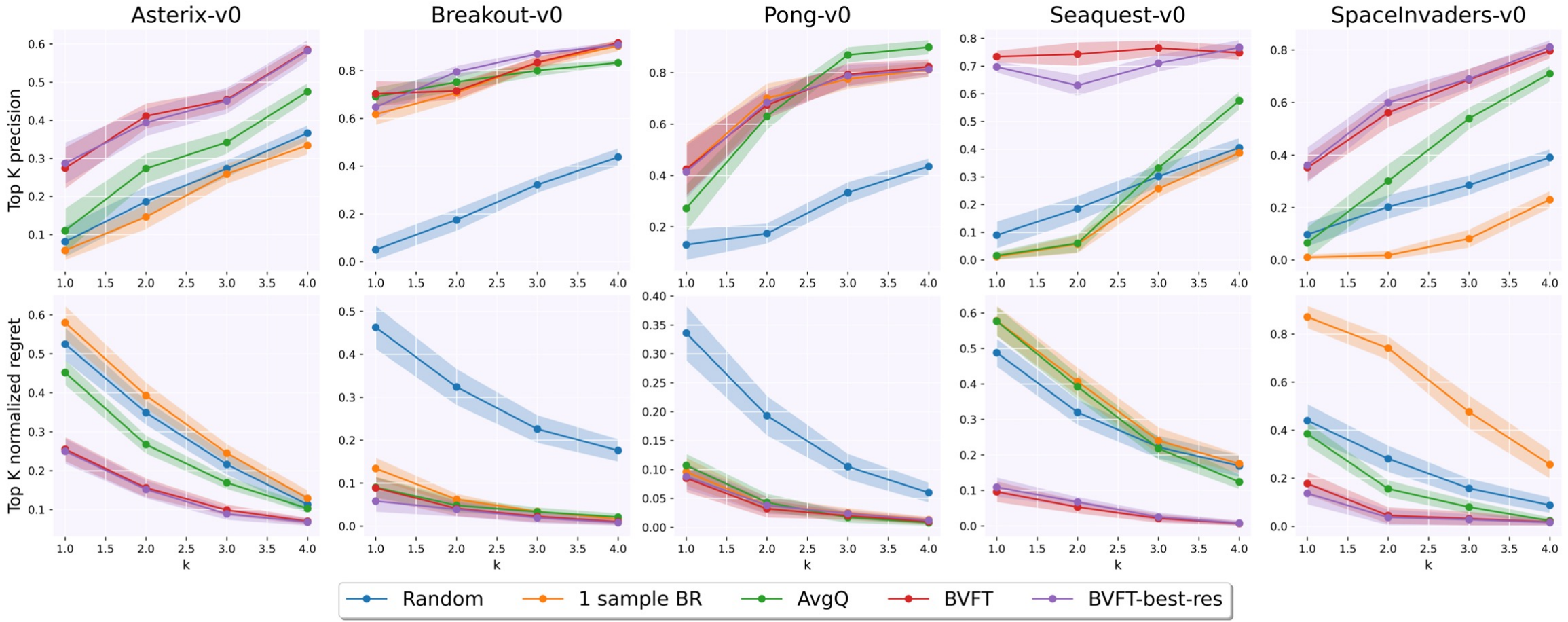
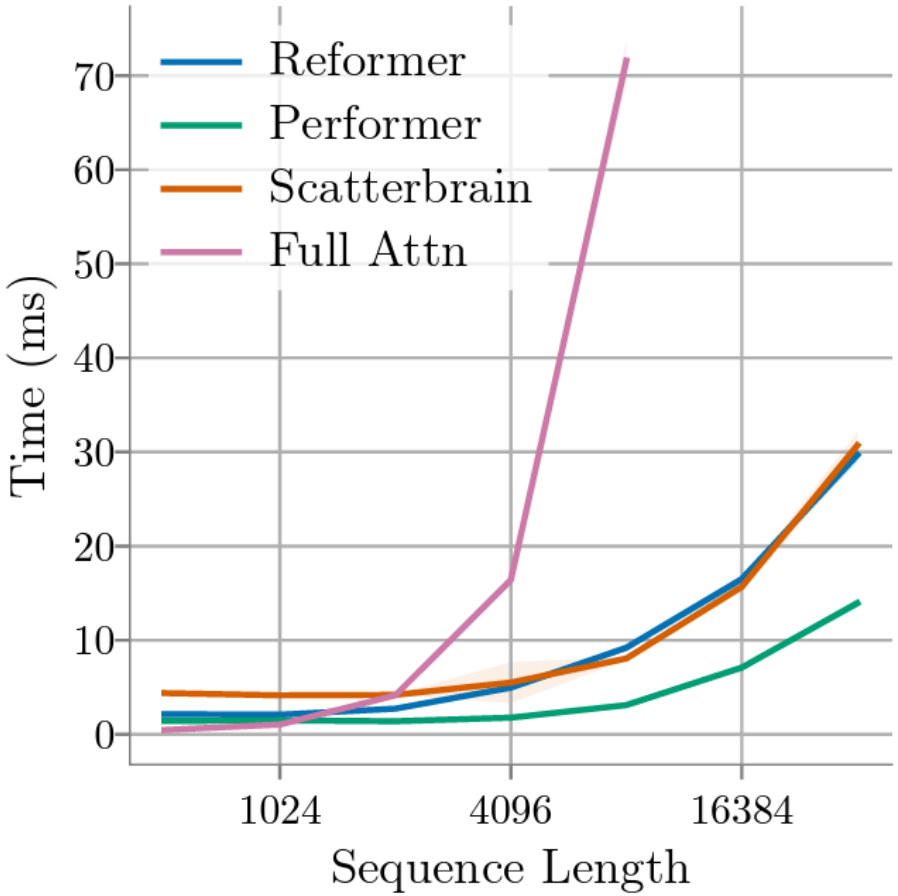


Figure 2: Top-k metrics of policy rankings vs. k in Atari. Row 1 shows top-k precision (the higher the better), and Row 2 shows top-k regret (the lower the better). Training algorithms are BCQ with different hyperparameters. The dataset for policy selection has 50,000 transition, which is an order of magnitude less than needed by FQE in Atari (see FQE in Enduro [VLJY19], as well as Figure 4).


Reformer
Performer
Scatterbrain
Full Attn
Time (ms)
0
10
20
30
40
50
60
70
1024
4096
16384
Sequence Length

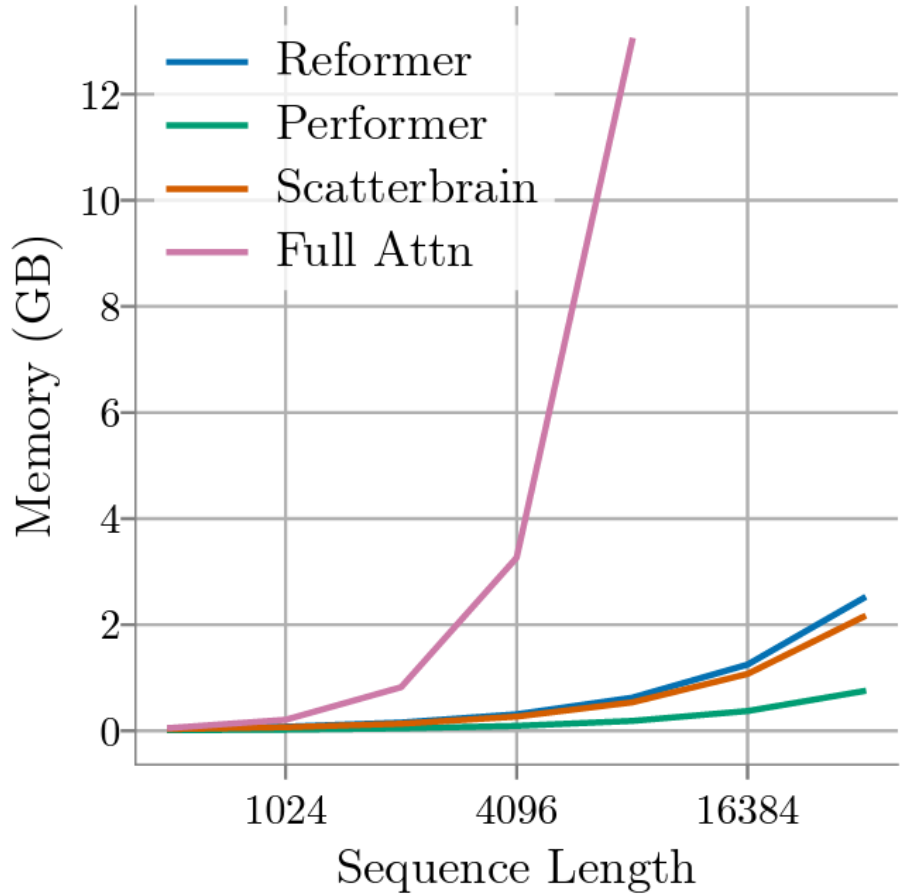
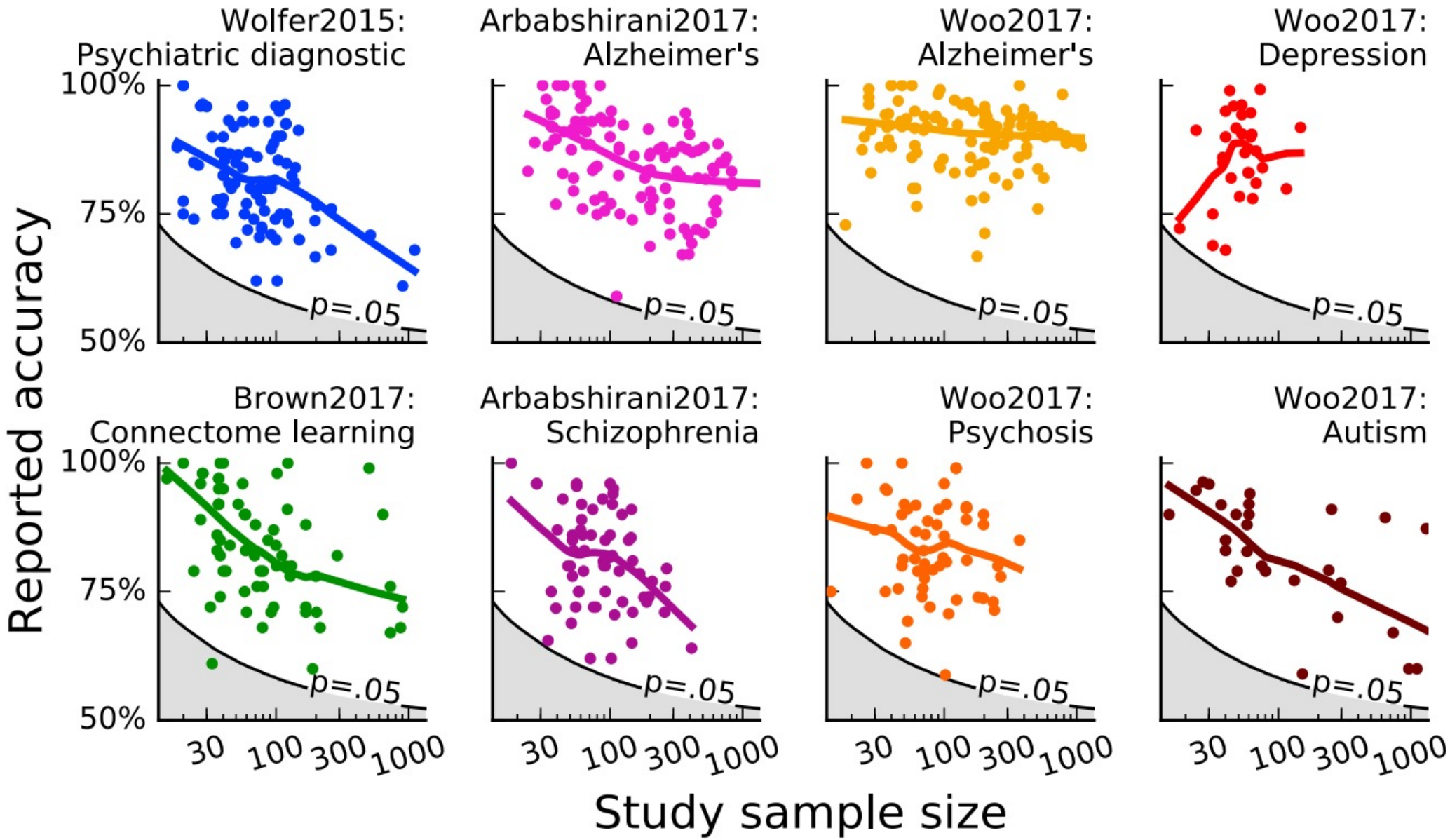

Reformer
Performer
Scatterbrain
Full Attn
Memory (GB)
0
2
4
6
8
10
12
1024
4096
16384
Sequence Length

Table 1: Top-1 Accuracy of pre-trained T2T Vision Transformer on ImageNet with different attention replacements. Error represents the average normalized approximation error to full attention.

Attention	Top-1 Acc	Error (avg)
Full Attention	81.7%	-
SMYRF	79.8%	11.4%
Performer	80.1%	7.5%
Baseline SMYRF + Performer	79.7%	12.6%
Scatterbrain	**80.7**%	**5.3**%

Table 1: Quantitative results of the evaluated methods in SGCls and PredCls tasks in recall, with and without graph constraint (GC). We additionally report our method with and without the meta-embedding (ME) intervention. The standard deviation is reported after the mean recall value.

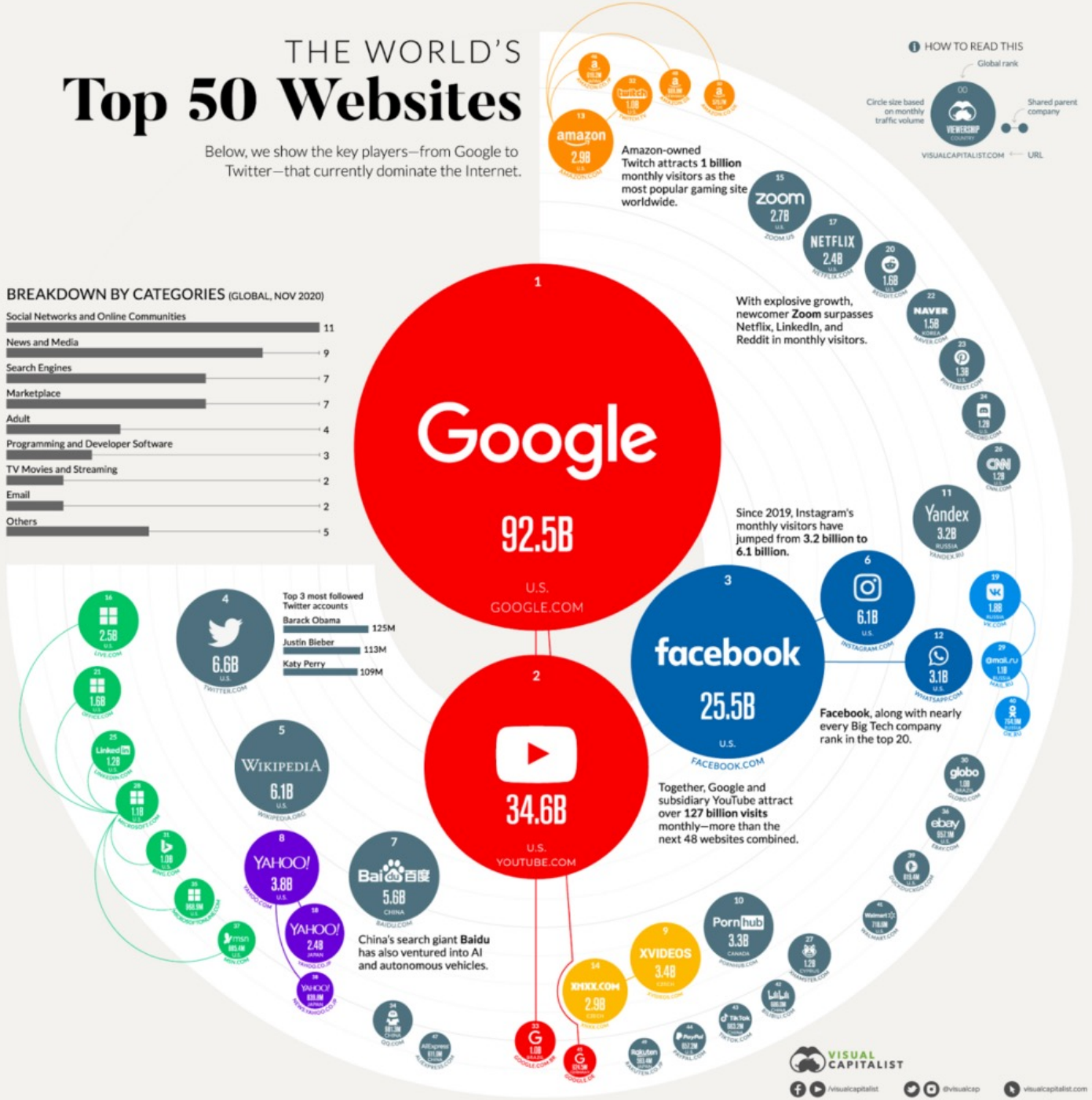
		SGCls			PredCls		
	Model	R@20	R@50	R@100	R@20	R@50	R@100
w/ GC	Co-Occurrence	0.148±0.000	0.197±0.000	0.199±0.000	0.347±0.000	0.474±0.000	0.479±0.000
	KERN [4]	0.203±0.007	0.224±0.008	0.227±0.008	0.468±0.004	0.557±0.007	0.565±0.007
	SGPN [29]	0.270±0.001	0.288±0.001	0.290±0.001	0.519±0.004	0.580±0.005	0.585±0.004
	Schemata [22]	0.274±0.003	0.292±0.004	0.294±0.004	0.487±0.004	0.582±0.007	0.591±0.006
	Ours (w/o ME)	0.282±0.002	0.299±0.001	0.301±0.001	0.529±0.004	0.592±0.004	0.598±0.005
	Ours	**0.285**±0.001	**0.300**±0.001	**0.301**±0.001	**0.593**±0.004	**0.650**±0.004	**0.653**±0.004
w/o GC	Co-Occurrence	0.141±0.000	0.202±0.000	0.258±0.000	0.351±0.000	0.556±0.000	0.706±0.000
	KERN [4]	0.208±0.007	0.247±0.007	0.276±0.005	0.483±0.003	0.648±0.006	0.772±0.011
	SGPN [29]	0.282±0.001	0.326±0.001	0.353±0.001	0.545±0.006	0.701±0.001	0.824±0.002
	Schemata [22]	0.288±0.001	0.335±0.003	0.363±0.002	0.496±0.002	0.671±0.003	0.802±0.009
	Ours (w/o ME)	0.293±0.001	0.338±0.003	0.367±0.003	0.549±0.004	0.716±0.005	0.824±0.008
	Ours	**0.298**±0.002	**0.343**±0.004	**0.370**±0.002	**0.622**±0.005	**0.784**±0.004	**0.883**±0.002


Wolfer2015:
Psychiatric diagnostic
Arbabshirani2017:
Alzheimer's
Woo2017:
Alzheimer's
Woo2017:
Depression
Brown2017:
Connectome learning
Arbabshirani2017:
Schizophrenia
Woo2017:
Psychosis
Woo2017:
Autism
100%
75%
50%
p=.05
30
100
300
1000
Reported accuracy
Study sample size

Good Data Visualizations

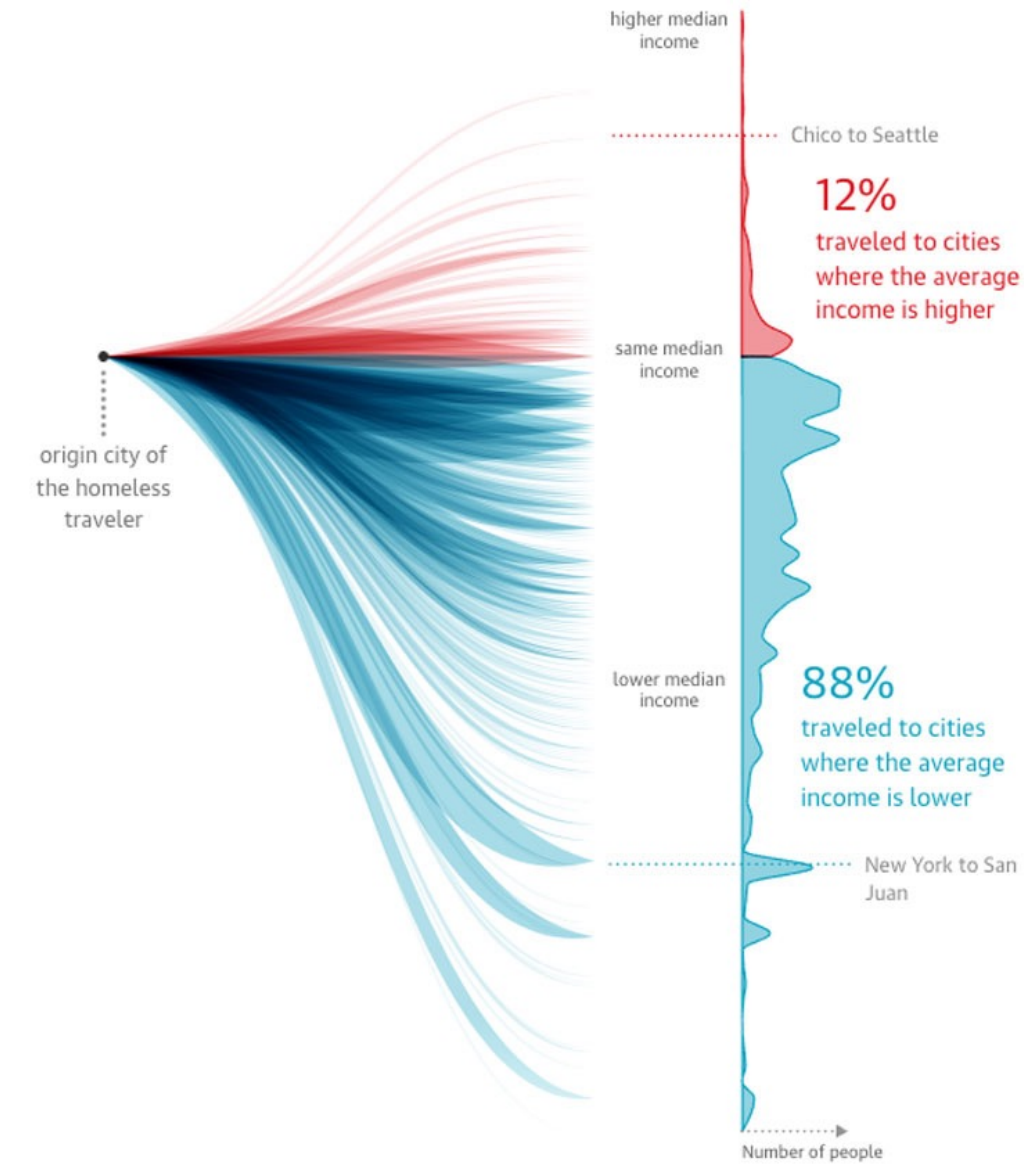
THE WORLD'S Top 50 Websites

Below, we show the key players—from Google to Twitter—that currently dominate the Internet.



SOURCE: SIMILARWEB

Most ticket recipients are relocated to places with a lower median income



World's Population at 8 BILLION PEOPLE

Around November 2022, the world will reach a pivotal milestone—8 billion global population. What is the distribution of this population, by region and country?

North America
Central America 52M
506M
Caribbean 44M
North America

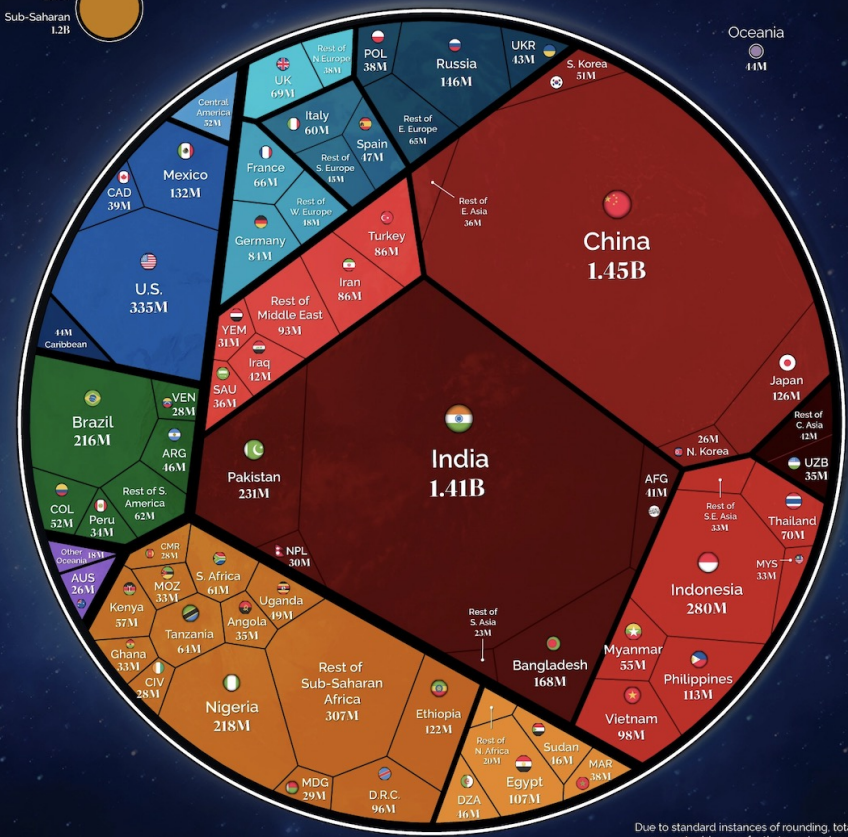
South America
439M

Africa
North Africa 257M
Sub-Saharan 1.2B

Europe
North Europe 107M
West Europe 198M
East Europe 292M
South Europe 152M

Asia
East Asia 1.7B
Central Asia 77M
Middle East 376M
South Asia 1.9B
S.E. Asia 684M

Oceania
44M

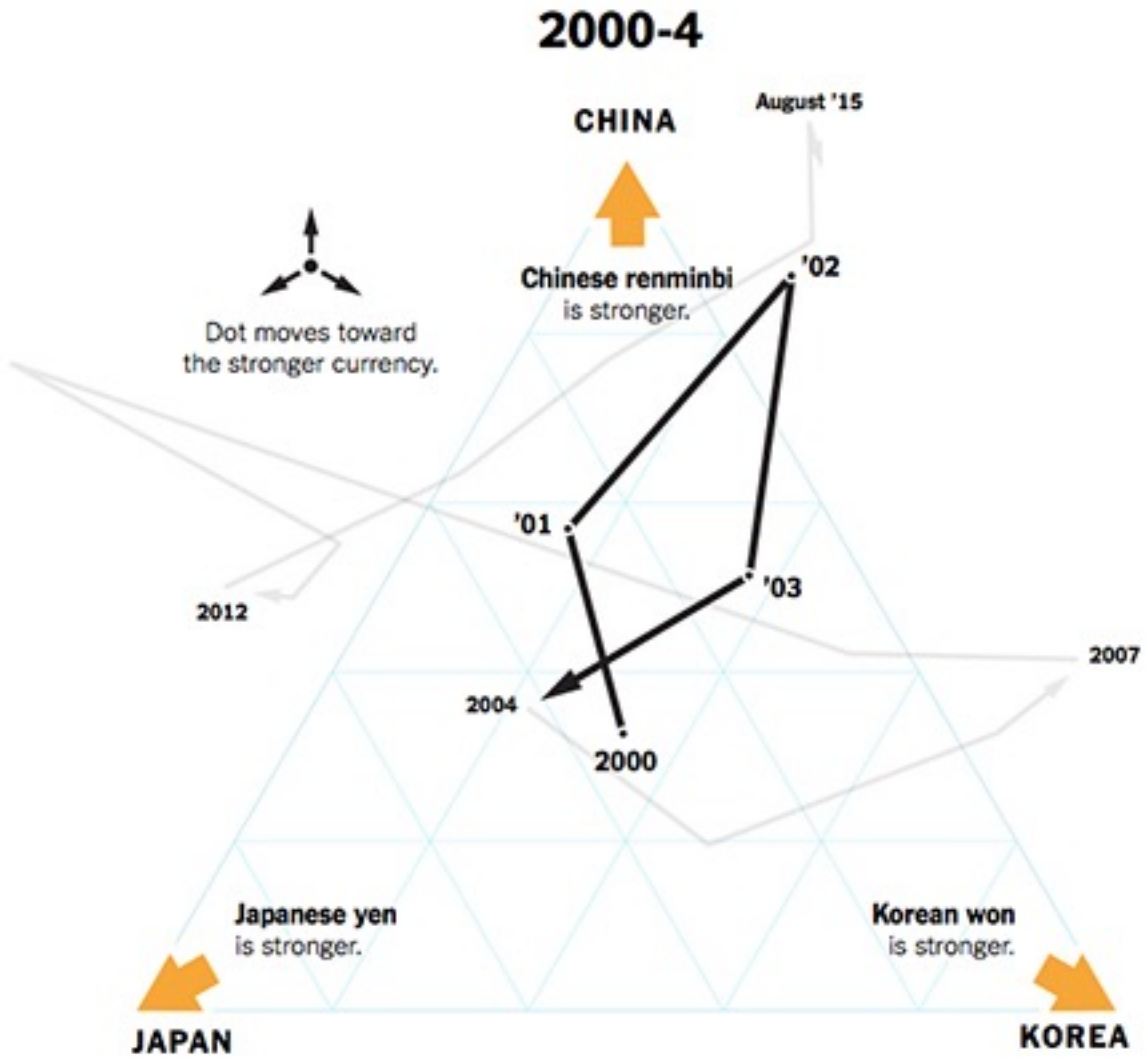


Due to standard instances of rounding, totals may not add up perfectly to regional sums.

Source: UN Population Division, 2022.

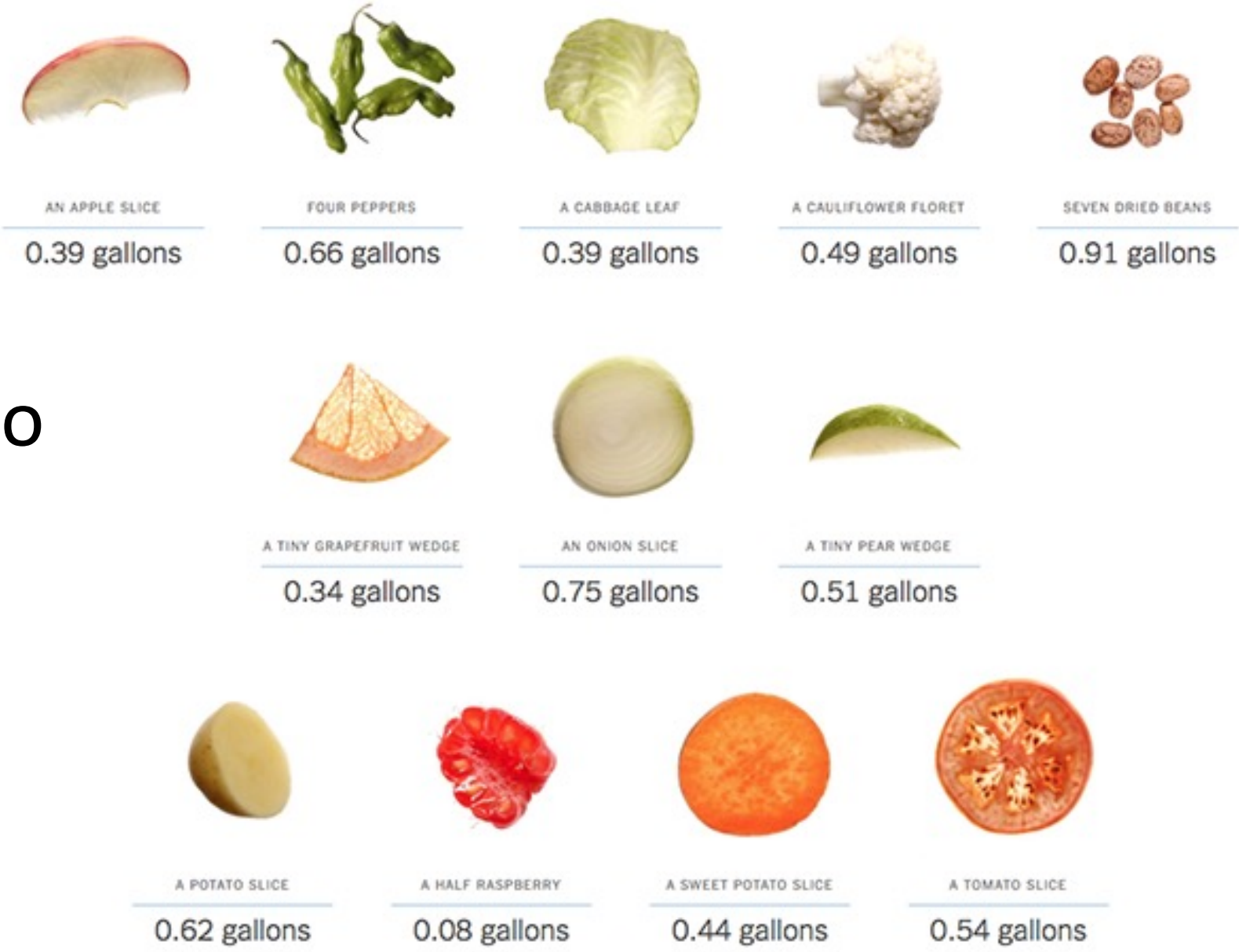
VISUAL CAPITALIST

/visualcapitalist @visualcap visualcapitalist.com



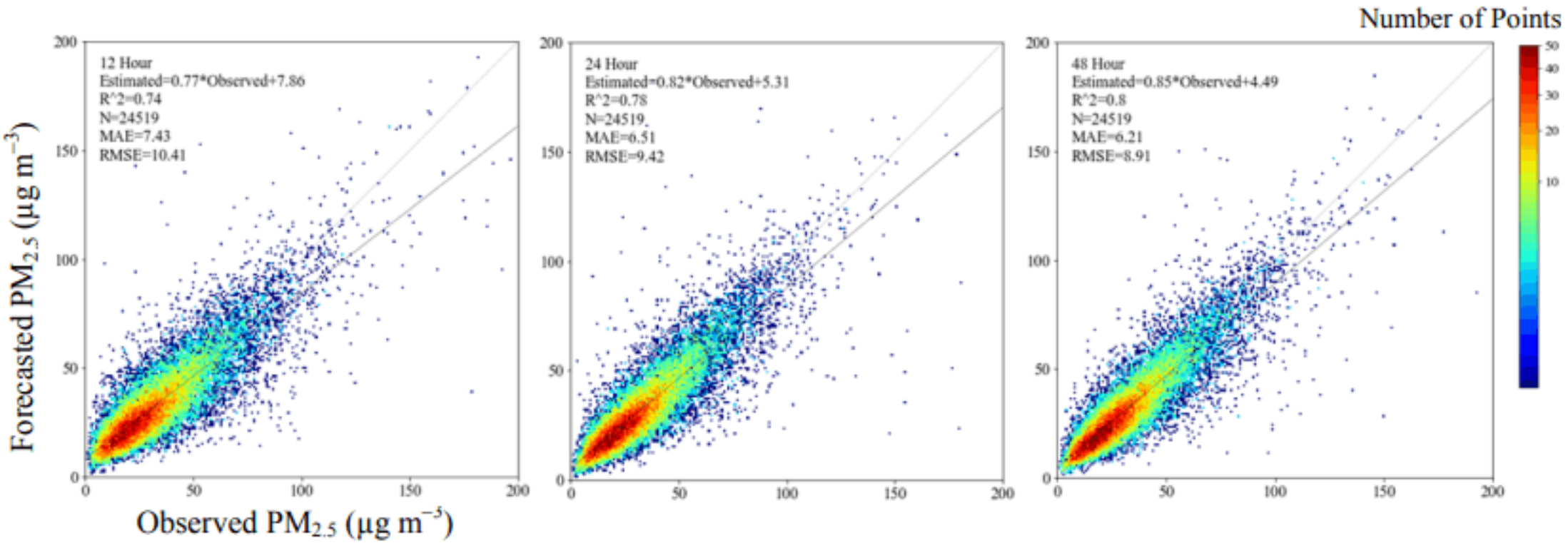
During a global economic slowdown in the early 2000s, Japan and Korea's currencies lost value relative to the dollar-pegged renminbi, but recovered by late 2004.

Contribution to the California drought

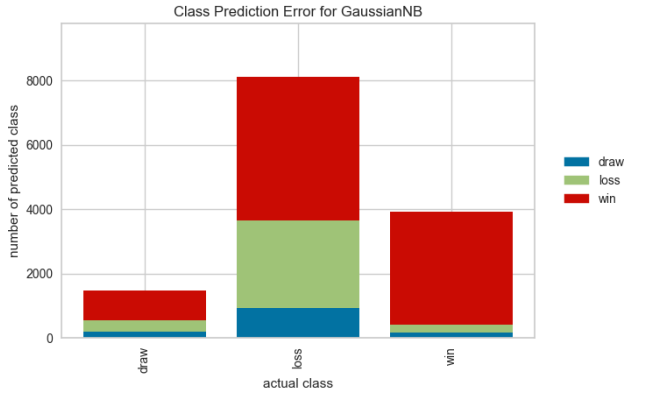


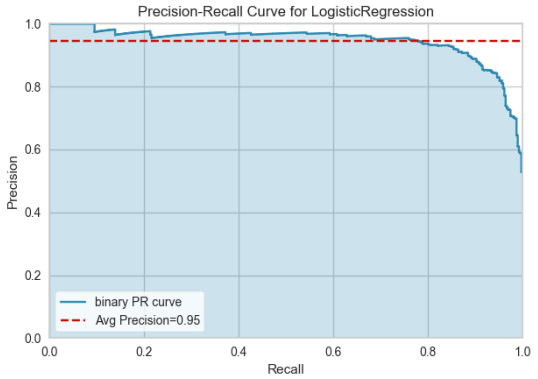
https://www.nytimes.com/interactive/2021/us/covid-cases.html

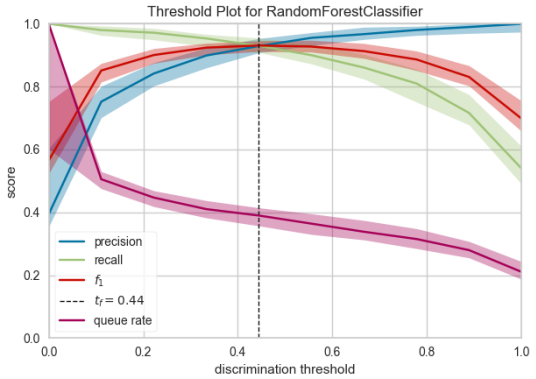
Keeping It Simple Is Okay Too

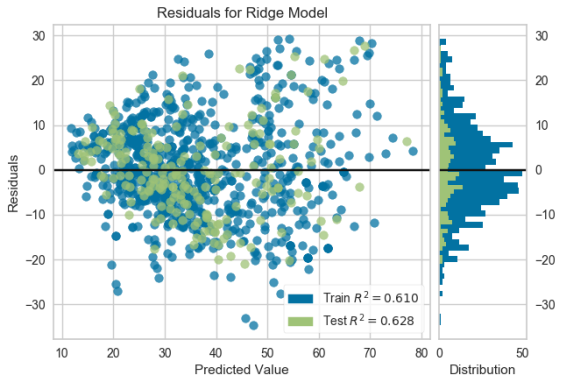


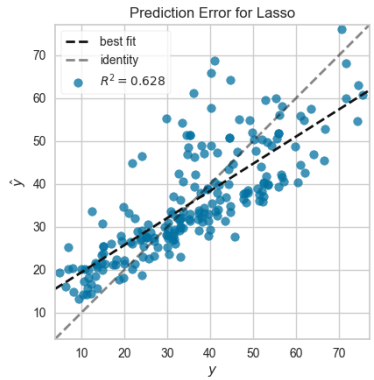
Keeping It Simple Is Good for AI/ML Evaluation

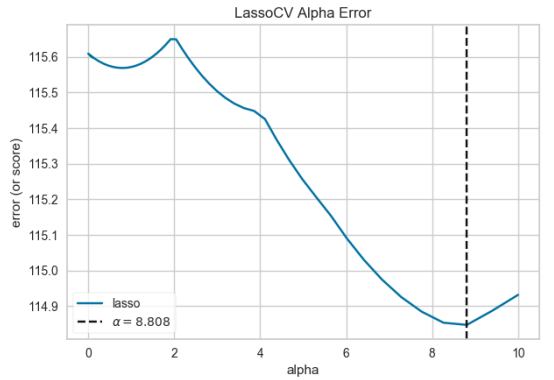


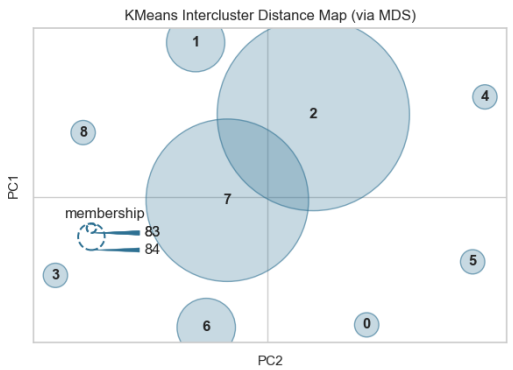


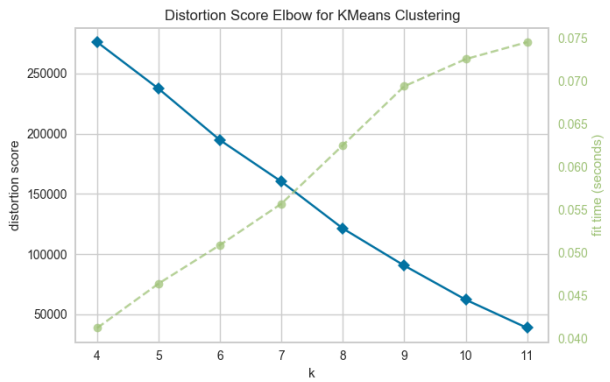


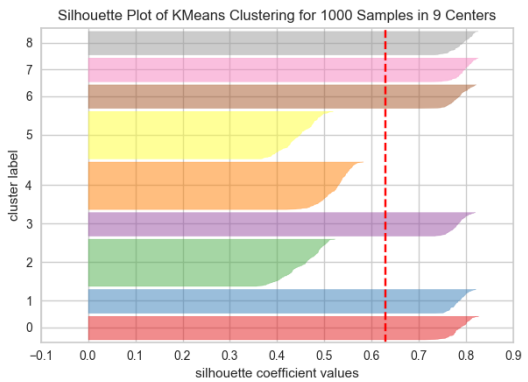


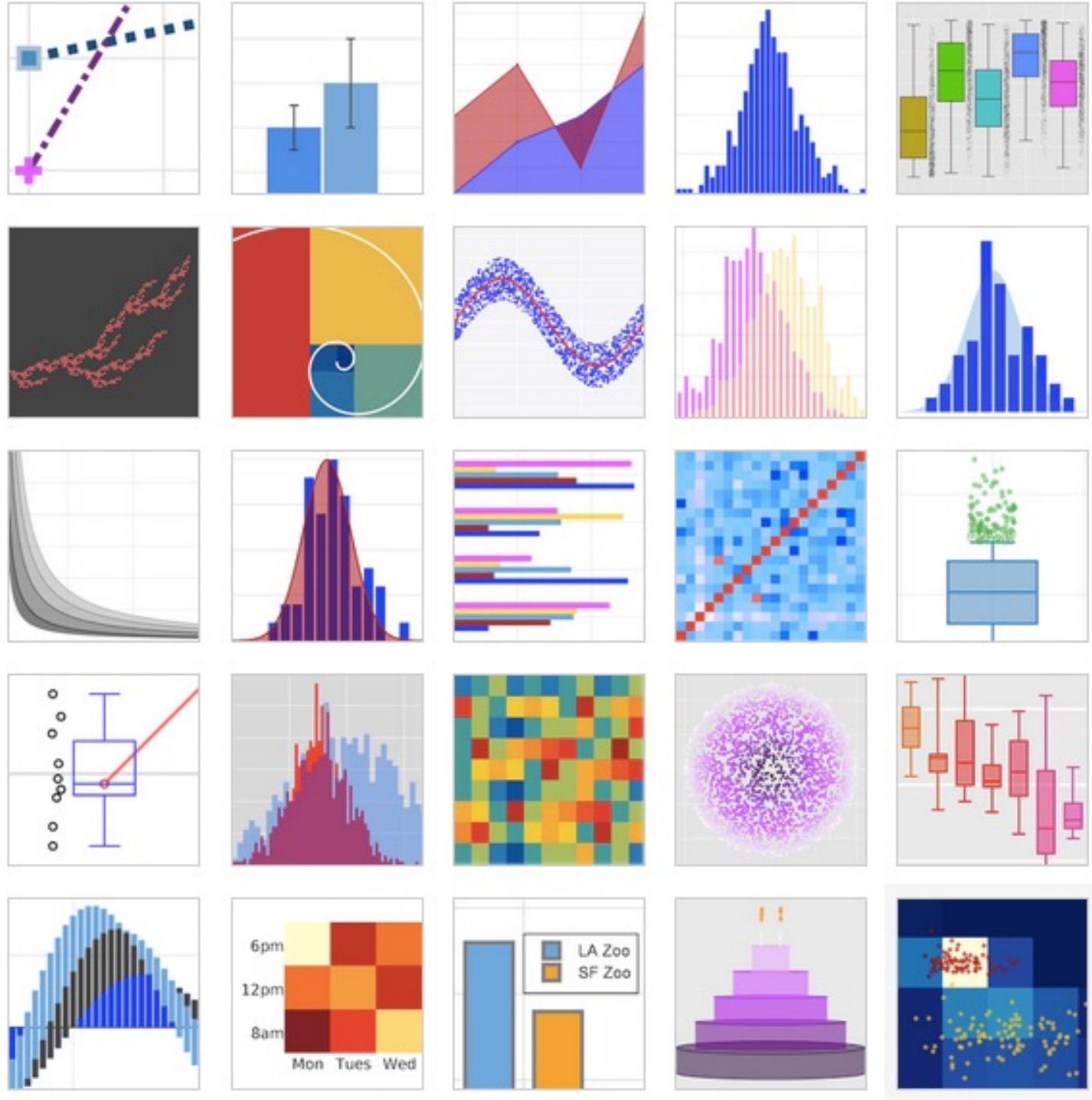











6pm
12pm
8am
Mon Tues Wed
LA Zoo
SF Zoo

General Data Visualization Rules

- Always have a title
- Always have axes and labels
- Keep it as simple as possible while still conveying the information you want to convey

When you show off your visualizations to your boss

Class Exercise

- Find a data visualization online that you'd like to share
- Email a link to: pyw@hawaii.edu
- Be prepared to discuss either something good or something bad about the visualization
- The more interesting the visualization, the better!