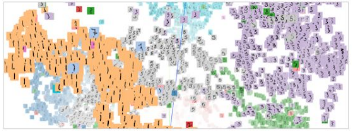
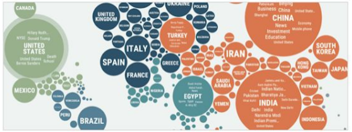
Visualization for Machine Learning

Fernanda Viégas @viegasf
Martin Wattenberg @wattenberg
Google Brain



Embedding Projector

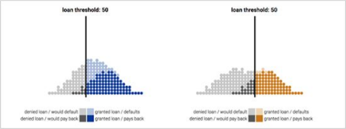
an open source, visualization tool for high-dimensional data



Unfiltered News

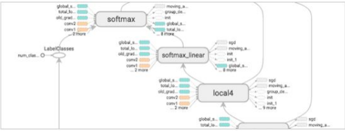
see news coverage around the world and spot underreported stories

(a collaboration with Jigsaw)



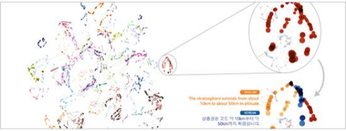
Fairness in ML

Try different tradeoffs yourself to understand issues around fairness and machine learning.



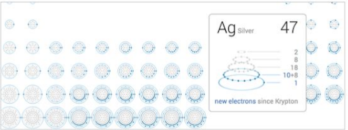
TensorFlow Graph Visualizer

an open source, high-level view of TensorFlow computation graphs



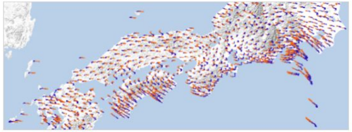
Machine Translation

Visualizing hints that a translation network learns an "interlingua", or universal language.



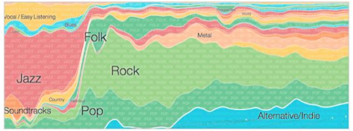
Periodic Table

a twist on the classic visualization of the atomic elements



Geodetic Velocities Visualization

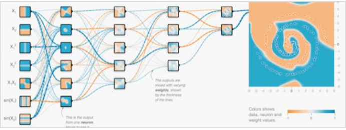
an open source visualization of earthquake cycle physics



Music Timeline

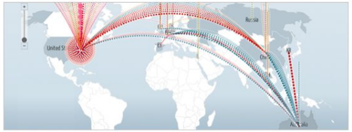
see how different musical genres became popular over time, and discover artists in each genre

UPDATED WEEKLY



TensorFlow Playground

an open source, transparent neural net you can play with in your browser



Digital Attack Map

see live data on denial-of-service attacks across the world, and observe historical patterns

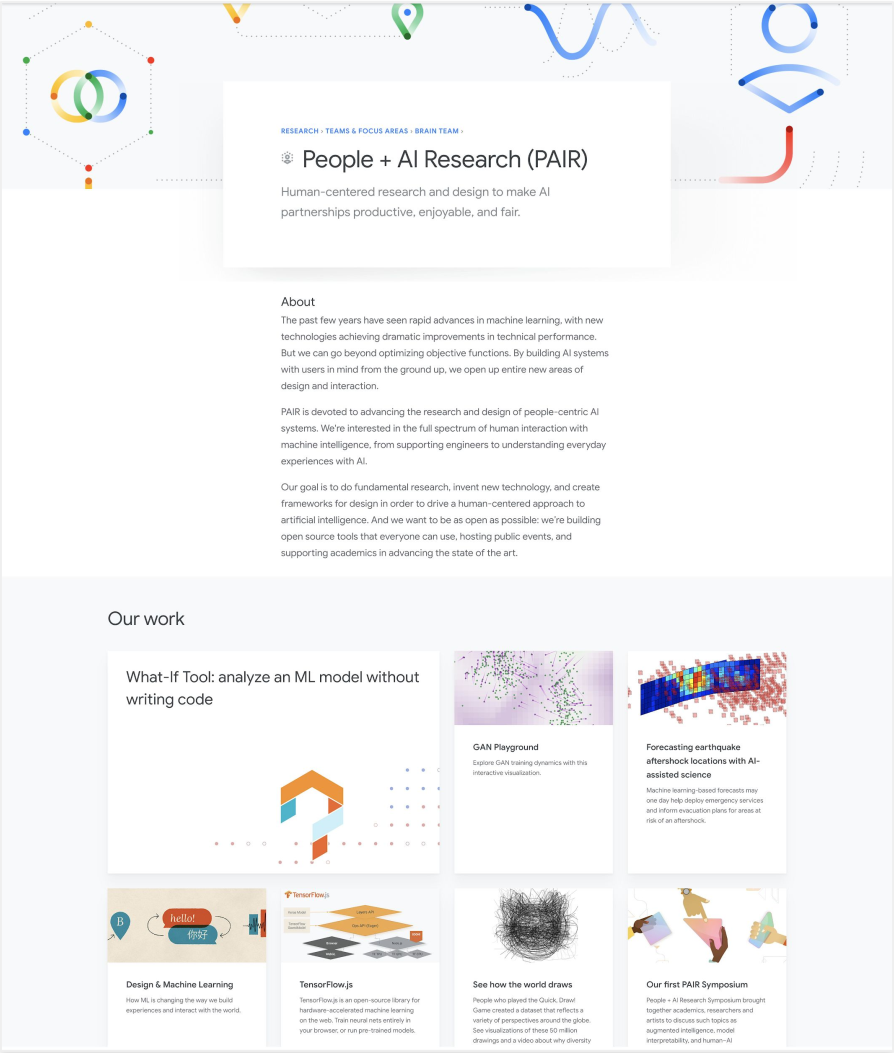
UPDATED DAILY

PAIR

People + AI Research

Bringing Design Thinking and HCI to Machine Learning

google.ai/pair


RESEARCH › TEAMS & FOCUS AREAS › BRAIN TEAM ›

People + AI Research (PAIR)

Human-centered research and design to make AI partnerships productive, enjoyable, and fair.

About

The past few years have seen rapid advances in machine learning, with new technologies achieving dramatic improvements in technical performance. But we can go beyond optimizing objective functions. By building AI systems with users in mind from the ground up, we open up entire new areas of design and interaction.

PAIR is devoted to advancing the research and design of people-centric AI systems. We're interested in the full spectrum of human interaction with machine intelligence, from supporting engineers to understanding everyday experiences with AI.

Our goal is to do fundamental research, invent new technology, and create frameworks for design in order to drive a human-centered approach to artificial intelligence. And we want to be as open as possible: we're building open source tools that everyone can use, hosting public events, and supporting academics in advancing the state of the art.

Our work

What-If Tool: analyze an ML model without writing code

GAN Playground

Explore GAN training dynamics with this interactive visualization.

Forecasting earthquake aftershock locations with AI-assisted science

Machine learning-based forecasts may one day help deploy emergency services and inform evacuation plans for areas at risk of an aftershock.

Design & Machine Learning

How ML is changing the way we build experiences and interact with the world.

TensorFlow.js

TensorFlow.js is an open-source library for hardware-accelerated machine learning on the web. Train neural nets entirely in your browser, or run pre-trained models.

See how the world draws

People who played the Quick, Draw! Game created a dataset that reflects a variety of perspectives around the globe. See visualizations of these 50 million drawings and a video about why diversity

Our first PAIR Symposium

People + AI Research Symposium brought together academics, researchers and artists to discuss such topics as augmented intelligence, model interpretability, and human–AI

Today's Agenda

What is data visualization?

How does it work? What are some best practices?

How has visualization been applied to ML?

Overview of the landscape

Special case: high-dimensional data

Goals

Understand state of the art

Known best practices in visualization
Broad survey of existing applications to ML

Apply visualizations in your own situation

References to tools and libraries
References to literature

What is data visualization?

Transform data into visual encodings

What is it good for?

Data exploration
Scientific insight
Communication
Education

How to ensure it works well?

Engage the visual system in smart ways
Take advantage of pre-attentive processing

What is data visualization?

Transform data into visual marks

What is it good for?

Data exploration
Scientific insight
Communication
Education

How is it different from statistics?

Vis: no specific question necessary
Classic Stats: you investigate a specific question*
Vis & Stats: wonderful, complementary partners

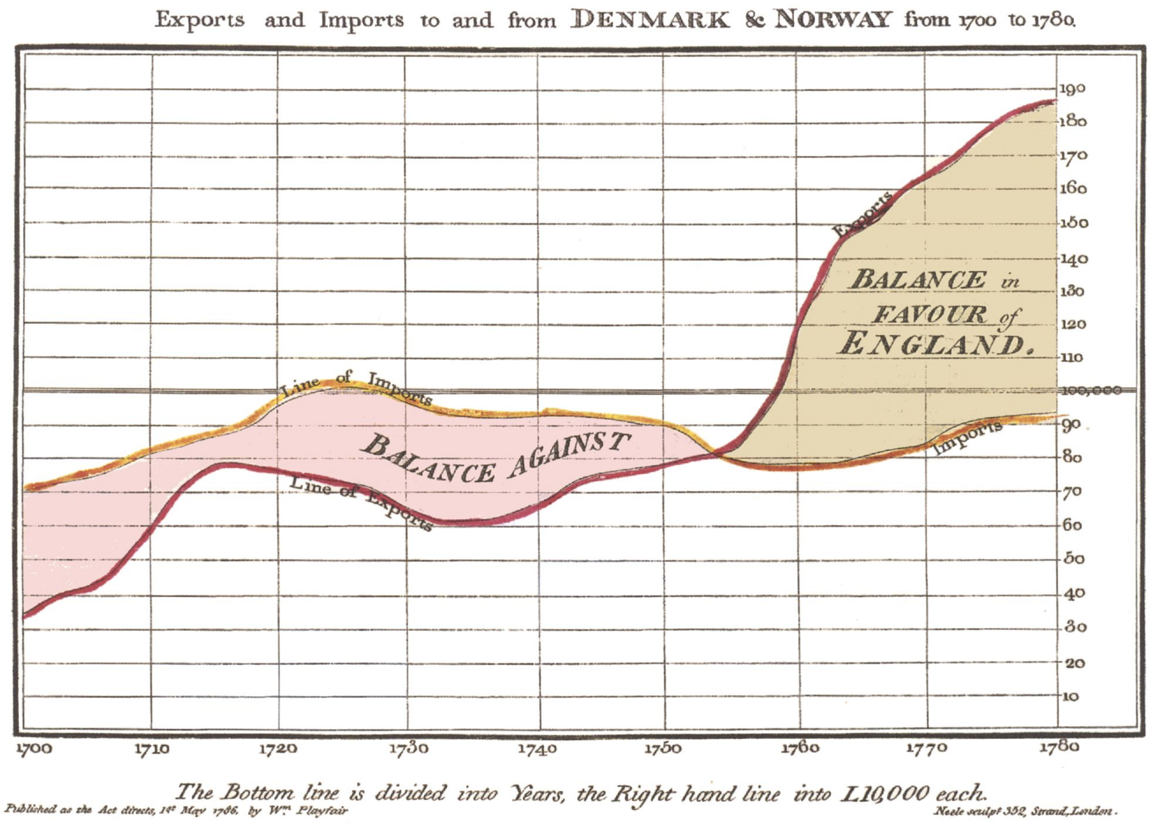
How to ensure it works well?

Engage the visual system in smart ways
Take advantage of pre-attentive processing

*OK, maybe not in EDA, but visualization is the key technique there anyway!

Predates computers...

William Playfair (1786)

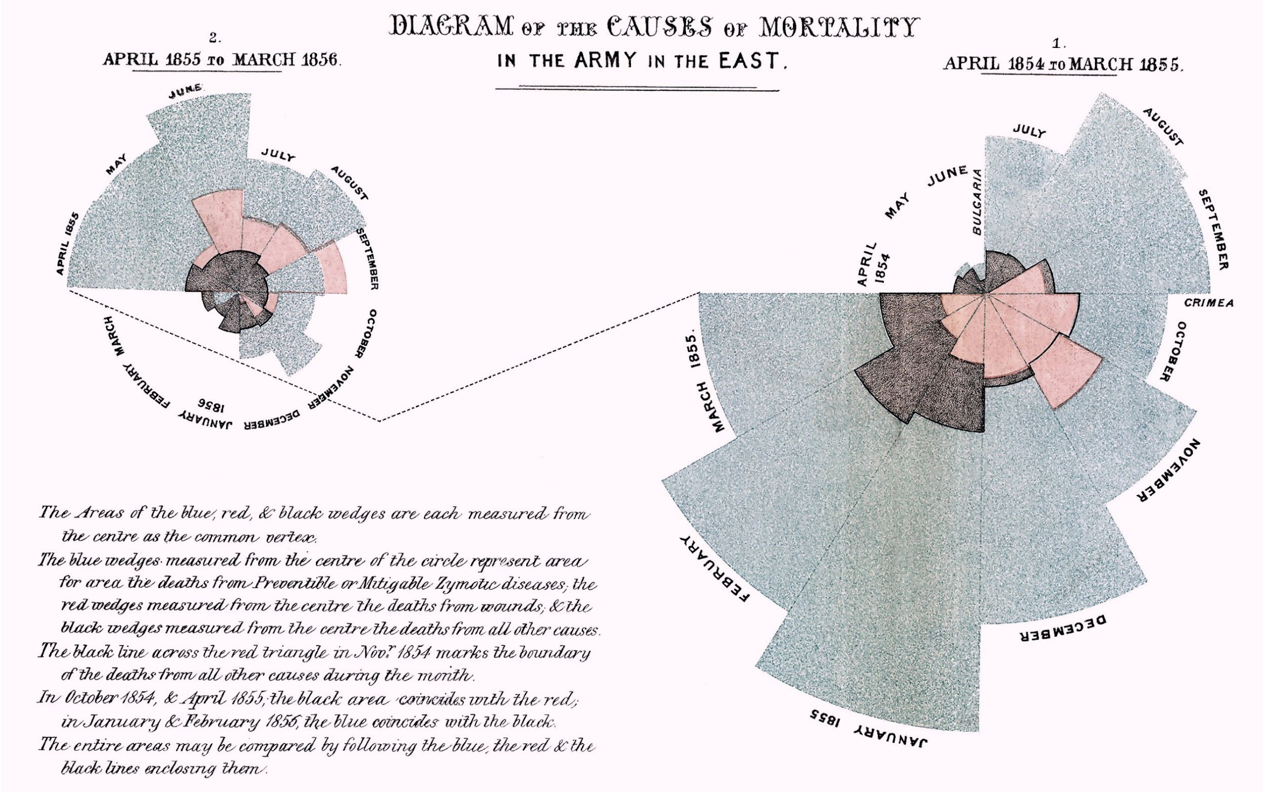


Line, bar, pie charts were all invented by the same person!

Aside from revolutionizing graphics, Playfair was an economist, engineer, and even a secret agent.

(Image: Wikipedia)

Florence Nightingale (1858)



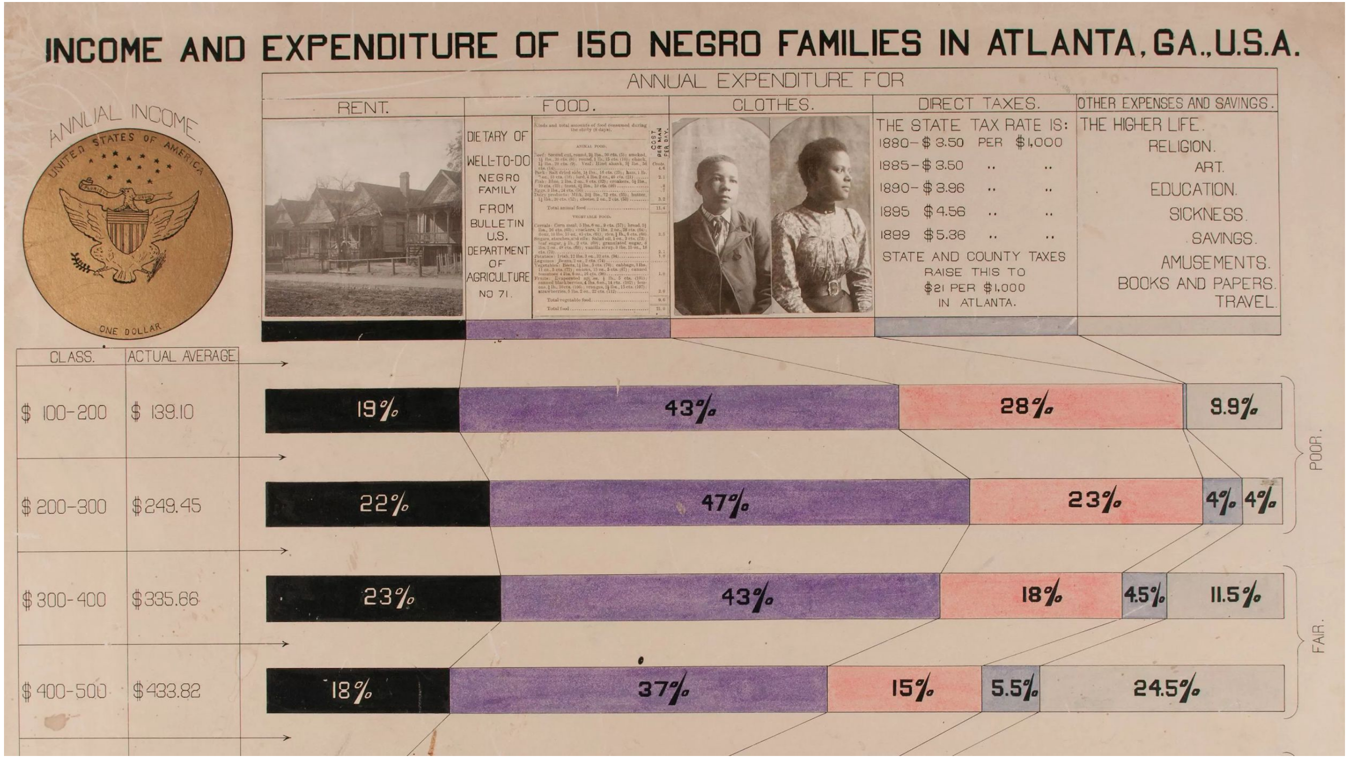
These charts led to the adoption of better hygiene / sanitary practices in military medicine, saving millions of lives.

Arguably the most effective visualization ever!

This particular visualization technique would be frowned on today. Lesson: technique is less important than having the right data and right message.

(Image: Wikipedia)

W. E. B. Du Bois (1900)



For 1900 World's Fair, a compendium of visualizations. Many new chart types!

Excellent example of visualization aimed at political change.

(Quartz)

What do these have in common?

Using special properties of the visual system to help us think.

Our visual system is like a GPU

- Incredibly good at a few special tasks
- With work, can be repurposed for more general situations

All visualizations are made from a series of compromises.

How do visualizations work?

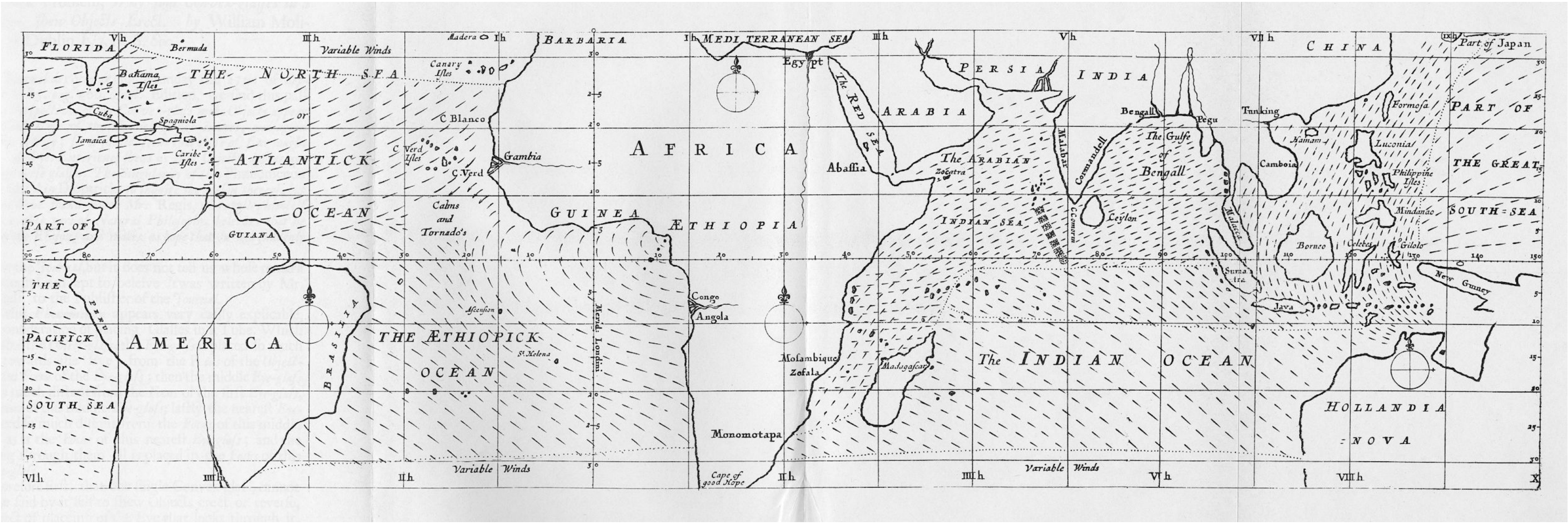
Find visual encodings that

- Guide viewer's attention
- Communicate data to the viewer
- Let viewer calculate with data

On computer

- Interactive exploration

Encodings: some examples



Edmund Halley, 1686

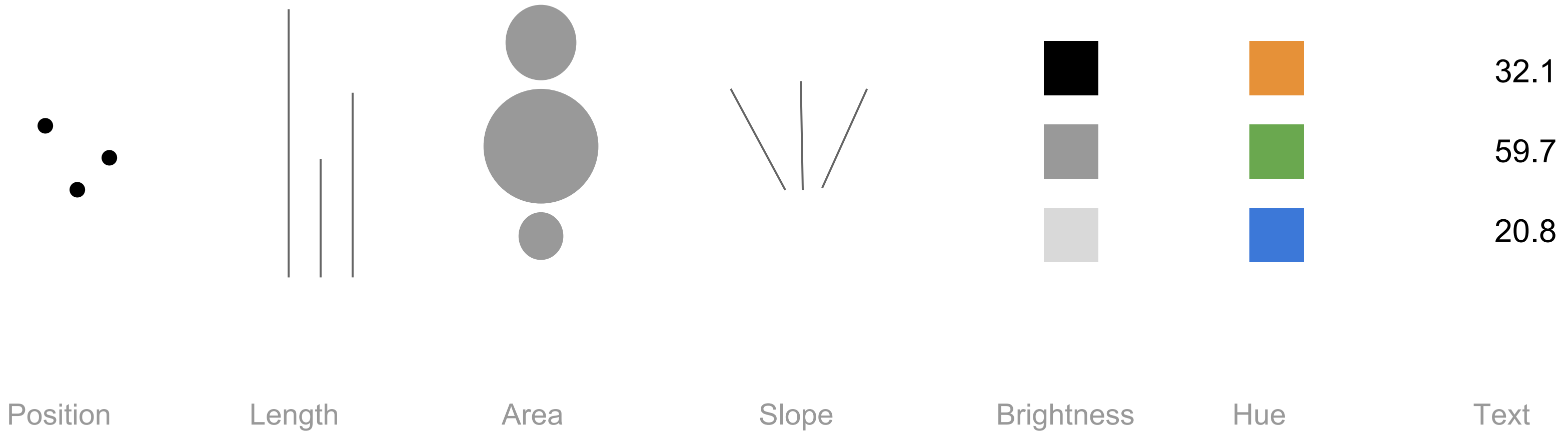
[Comparison A (2012)]: US Wind Map

[Comparison B (2013)]: Earth.nullschool

Encodings: some theory

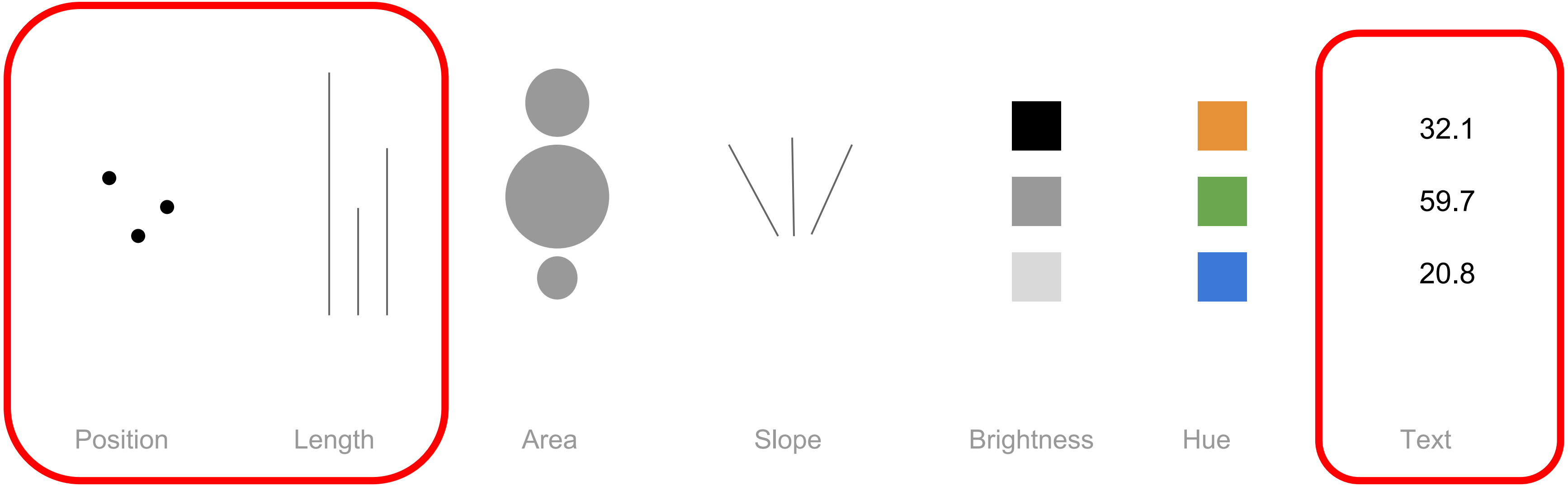
From perceptual psychology:

different encodings have different properties.



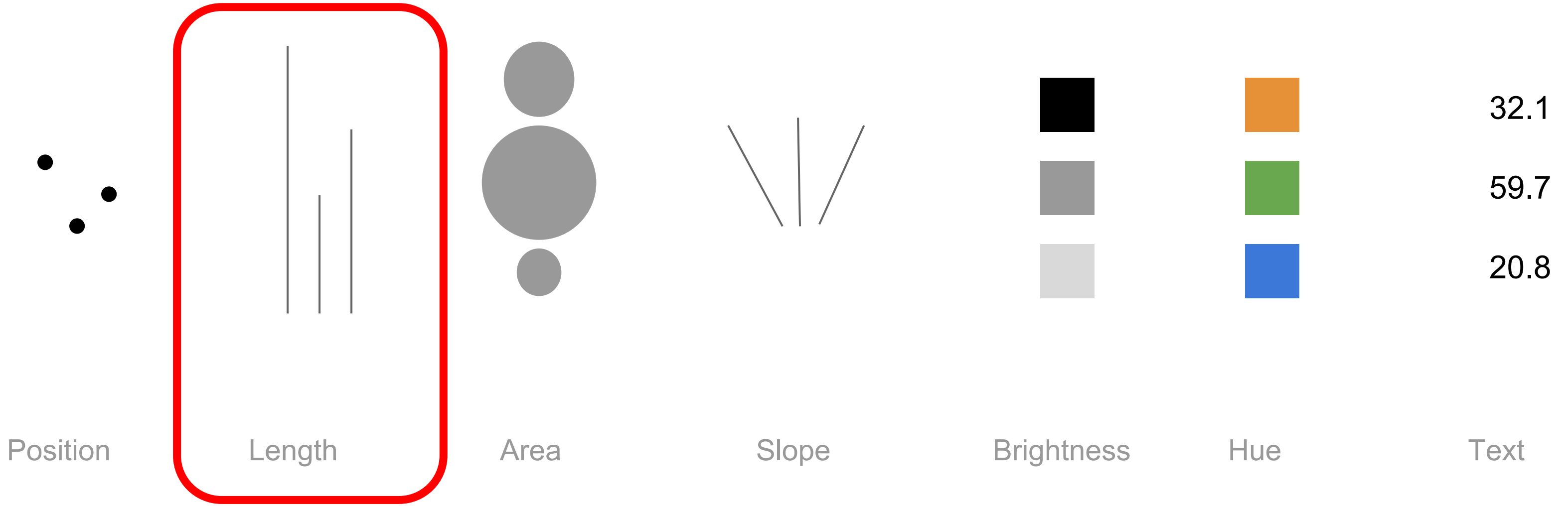
Encodings: some theory

Good for communicating exact values...



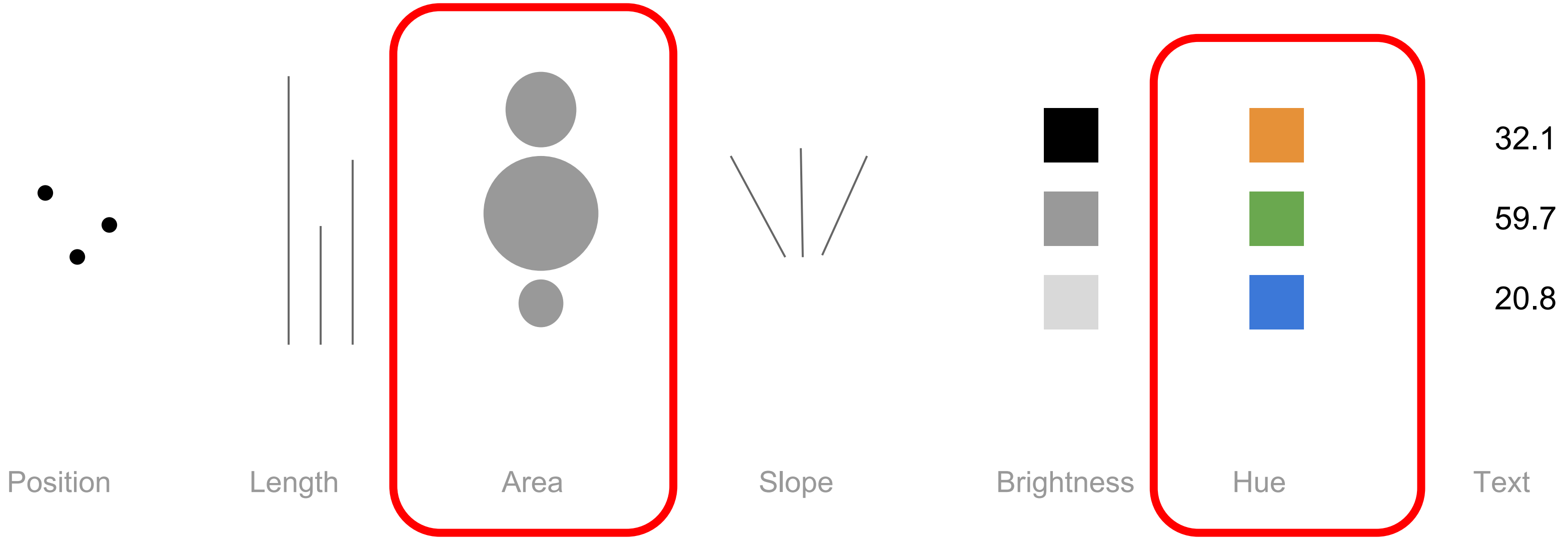
Encodings: some theory

Good for communicating ratios...



Encodings: some theory

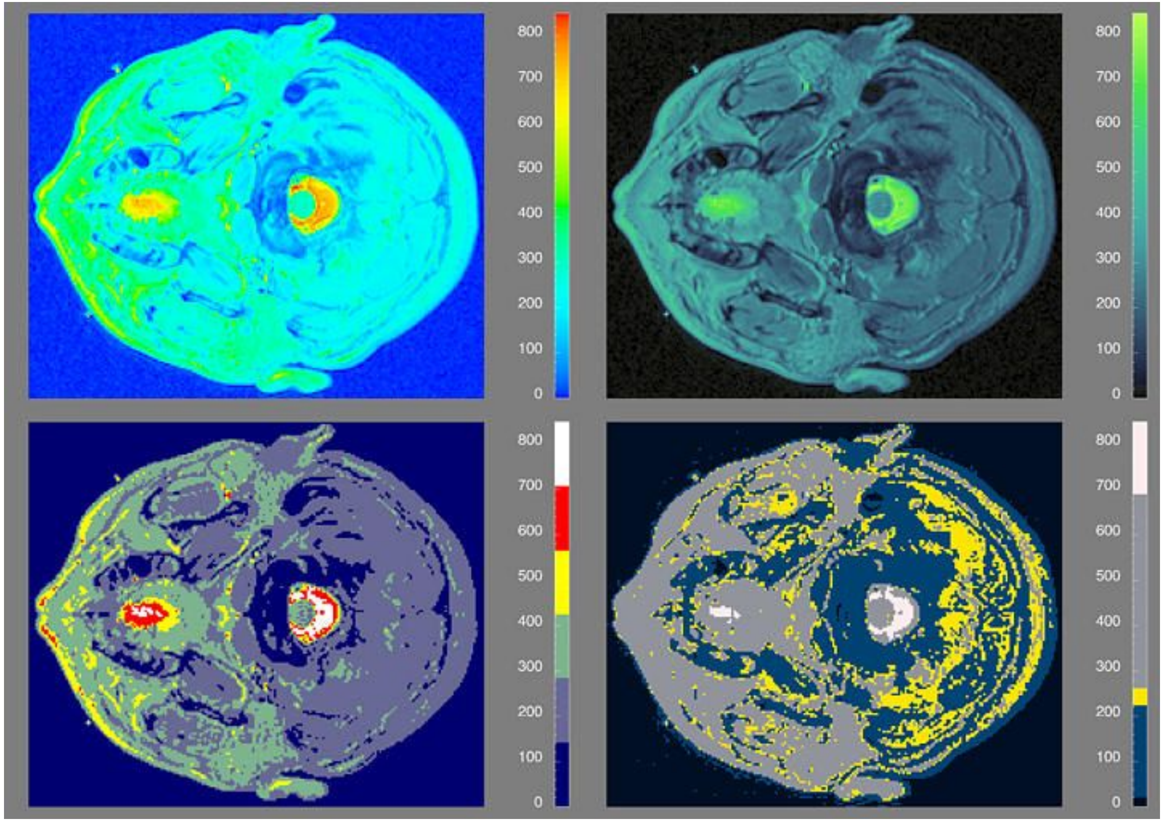
Good for drawing attention…



Special case: color scales

Intensively studied for decades...

Rogowitz & Treinish (1996)



Web article:

"Why Should Engineers and Scientists Be Worried About Color?"

Conclusions:

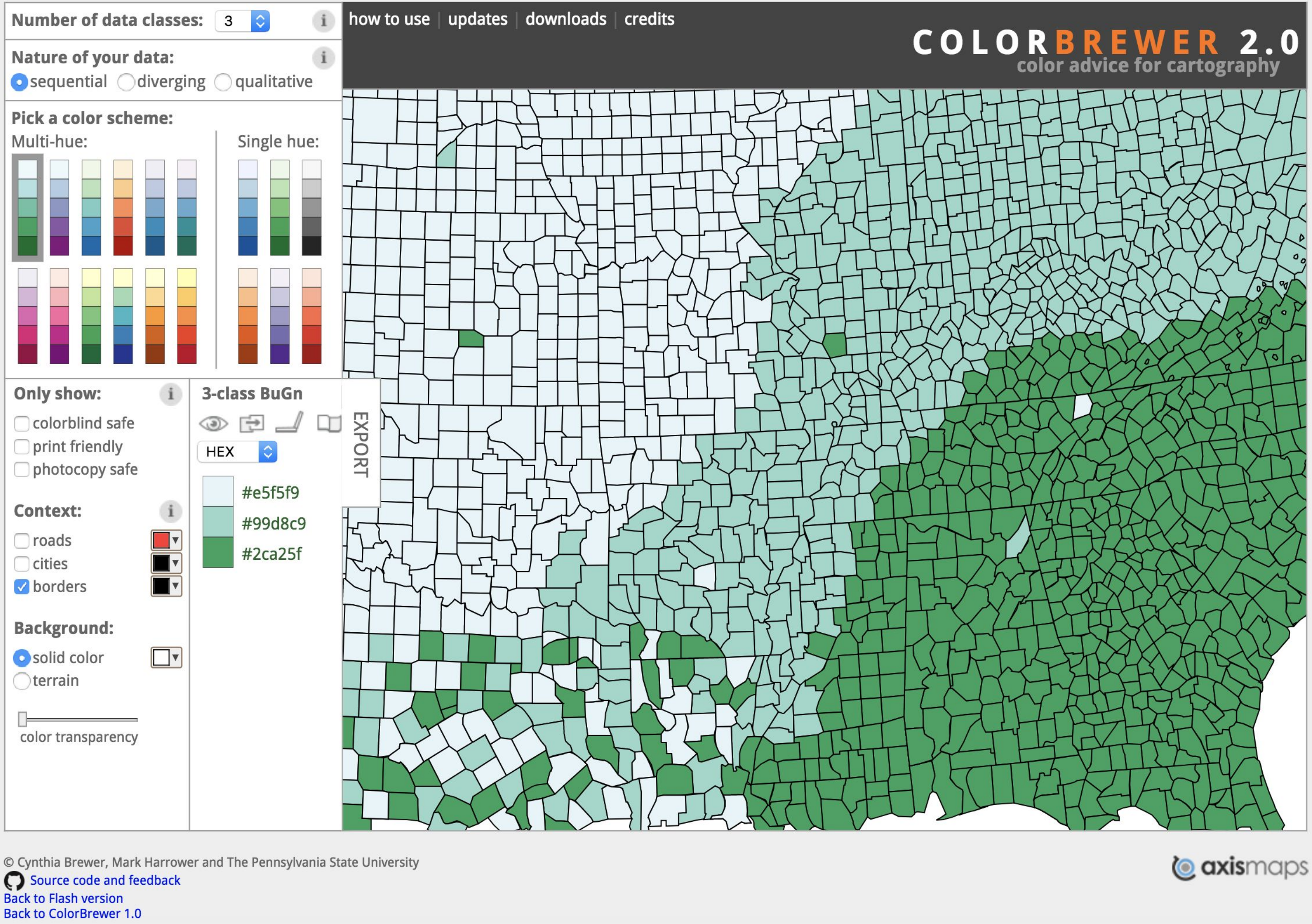
- Rainbow scales: bad
- There is no "best" scale

Practically speaking...

When in doubt, use the "Color Brewer" site:

http://colorbrewer2.org

(Built by Cynthia Brewer, a cartographer)


Number of data classes: 3
Nature of your data:
sequential diverging qualitative
Pick a color scheme:
Multi-hue:
Single hue:
how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
3-class BuGn
HEX
EXPORT
#e5f5f9
#99d8c9
#2ca25f
© Cynthia Brewer, Mark Harrower and The Pennsylvania State University
Source code and feedback
Back to Flash version
Back to ColorBrewer 1.0
axismaps

And study continues to this day…

A dive into a very recent paper (CHI 2018)

Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps

Yang Liu
University of Washington
Seattle, WA, USA
yliu0@cs.washington.edu

Jeffrey Heer
University of Washington
Seattle, WA, USA
jheer@uw.edu

Color scales

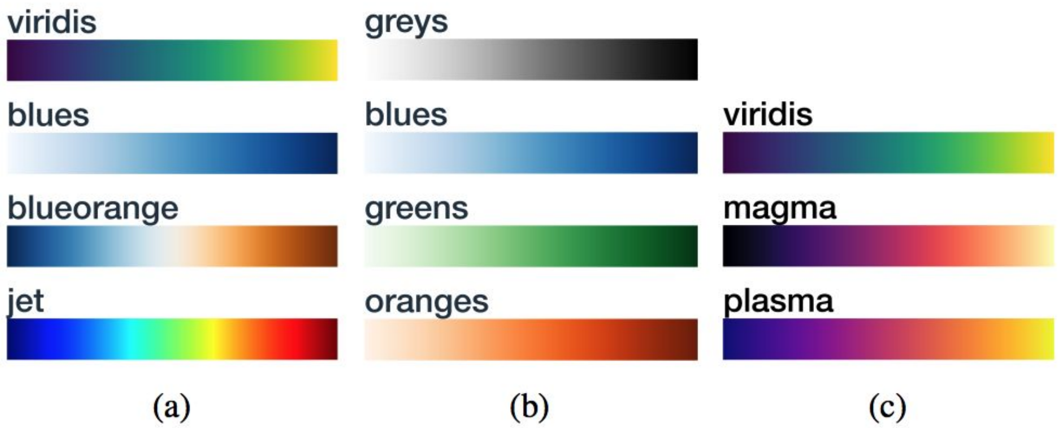


Figure 1: **Colormaps under study.** We evaluate four single-hue, three perceptually-uniform multi-hue, a diverging, and a rainbow colormap(s). We divide them into (a) assorted, (b) single-hue and (c) multi-hue groups, with two colormaps repeated across groups for replication.

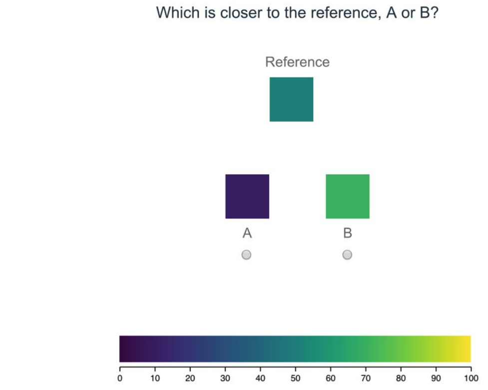
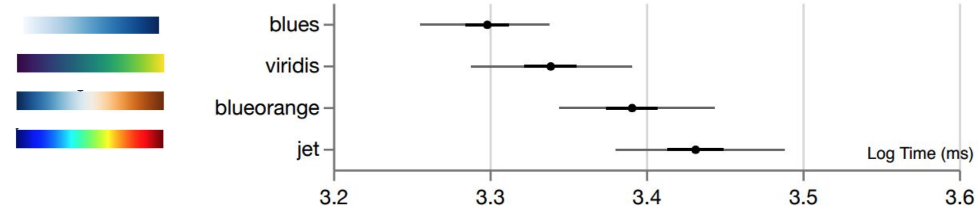
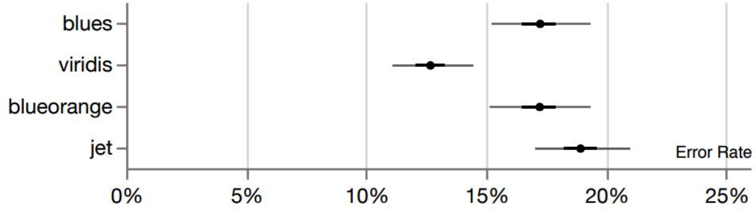


Figure 2: **Experiment interface.** Participants see a reference stimulus along with two choices, and pick which of these alternatives is closer in distance to the reference.

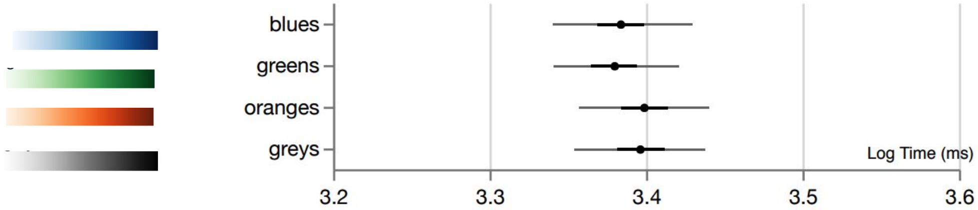
Color scales



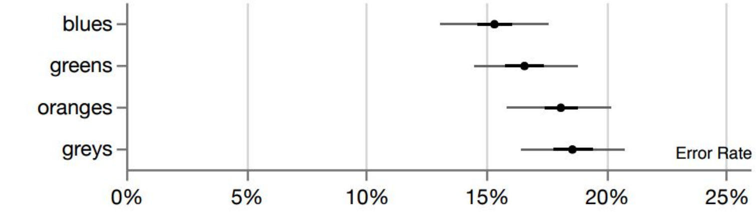
(a) Assorted Colormaps



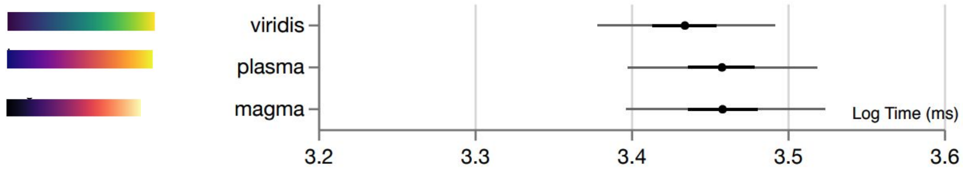
(a) Assorted Colormaps



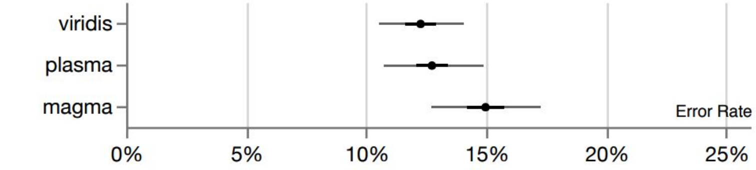
(b) Single-Hue Colormaps



(b) Single-Hue Colormaps



(c) Multi-Hue Colormaps



(c) Multi-Hue Colormaps

Uh oh, colorblindness... (very common!)

Red-blind protonopia. See http://www.color-blindness.com/coblis-color-blindness-simulator/

Guiding attention

Pre-attentive processing

Count the 5s

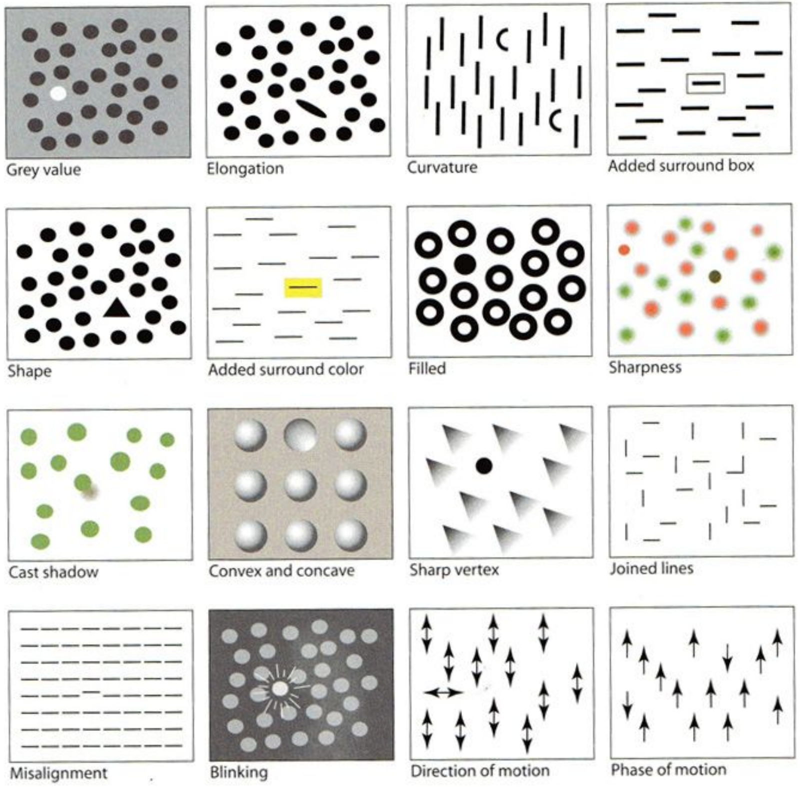
98734979027564790289472862409240603707057027907 2
80320802900730250127023700837408207872027200708 3
24780260270379377570970737797066746209709470278 0
92797970972309723097959275092727979873497260802 7

Count the 5s

987349790275647902894728624092406037070570279072
803208029007302501270237008374082078720272007083
247802602703793775709707377970667462097094702780
927979709723097230979592750927279798734972608027

987349790275647902894728624092406037070570279072
803208029007302501270237008374082078720272007083
247802602703793775709707377970667462097094702780
927979709723097230979592750927279798734972608027

Theory: attention



(Colin Ware, Visual Thinking for Design)

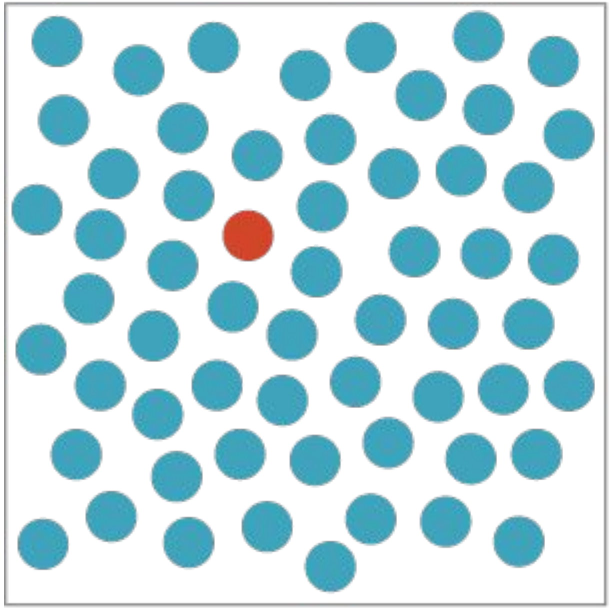
Pre-attentive processing / "popout"

Under the right circumstances, visual search can be parallel, rather than serial

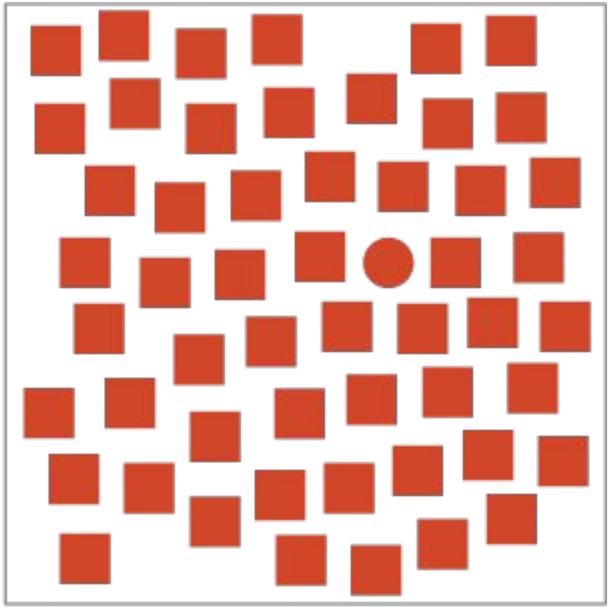
Time to find target does not increase as number of distractors increases

Pre-Attentive Processing

Color



Shape

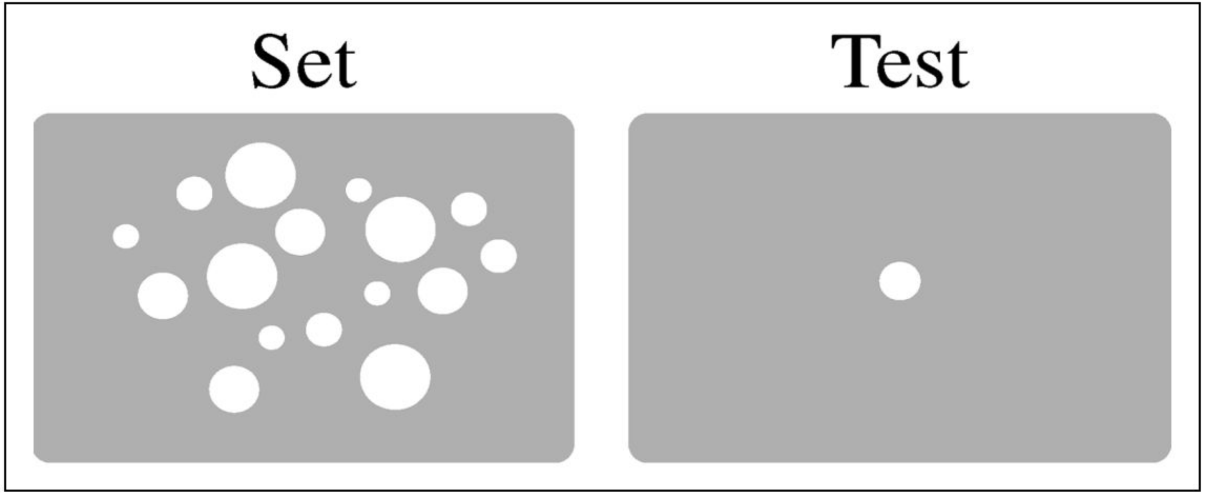


Theory: calculation

Calculation

Example: we naturally average sizes.

"Seeing Sets: Representation by Statistical Properties." Dan Ariely (2001)



Calculation

We can do weighted averages, too!

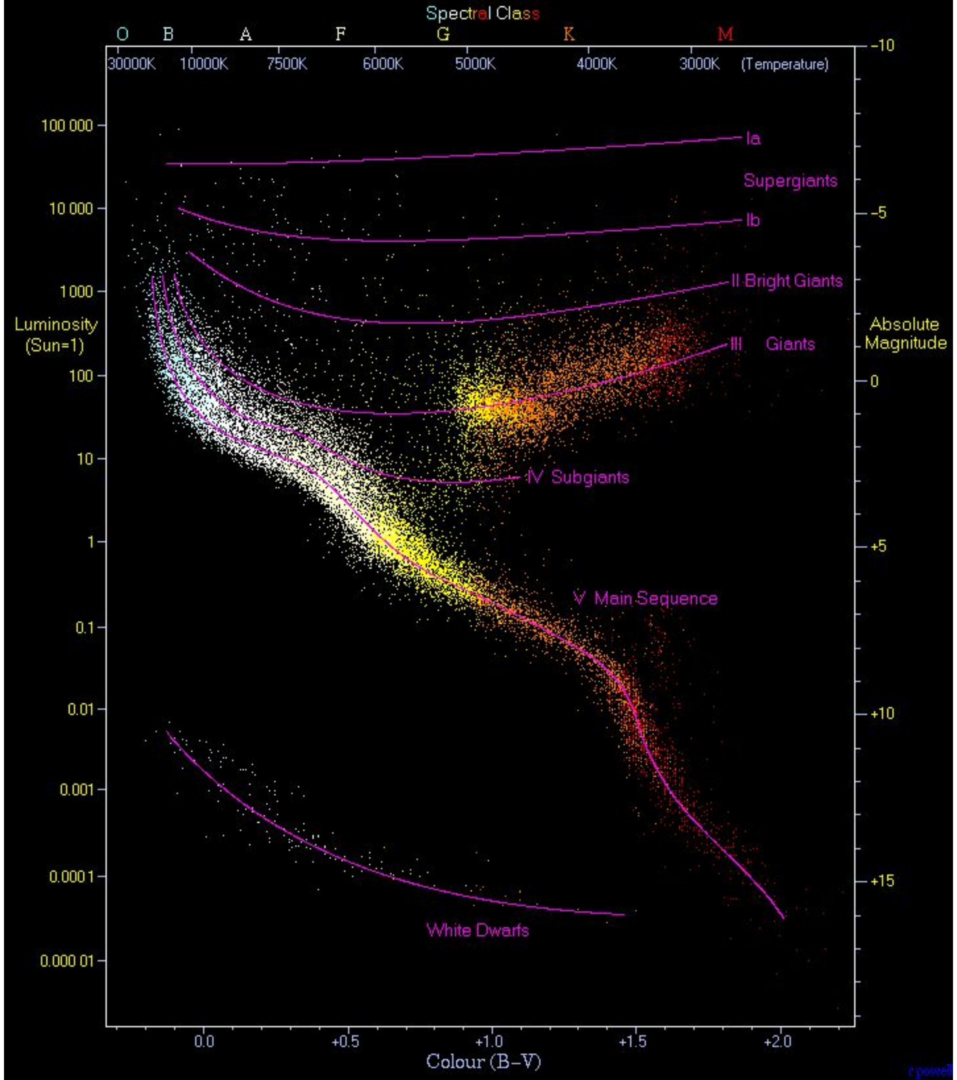
Example

Calculation

Hertzsprung-Russell diagram (via Wikipedia)

Your eye is doing something like kernel density estimation…

Source: Wikipedia



How do visualizations work - on computers?

Beyond static representations

- Interaction
- Conversation and collaboration

Theory: interaction

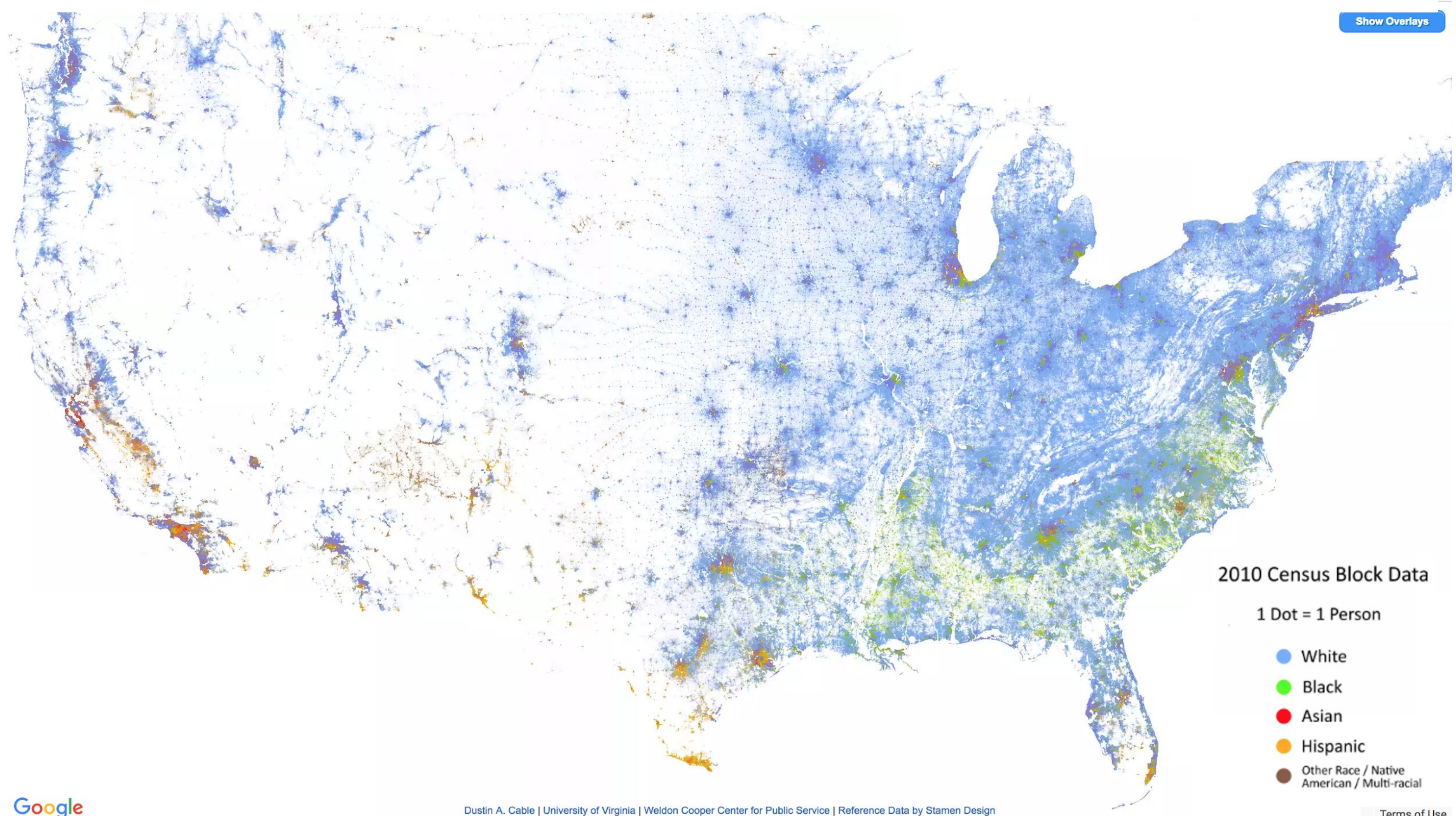
Shneiderman “mantra”:

(1996: “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”)

- Overview first
- Zoom and filter
- Details on demand

The Racial Dot Map: One Dot Per Person for the Entire U.S.

demographics.virginia.edu/DotMap/



Case study: the humble table

We've talked to many, many ML teams

Every one of them displayed data in tables

Good design can make a huge difference

Design thinking in action, a little movie:

Remove to improve data tables
Joey Cherdarchuk
DarkHorse Analytics

Key points

- Structure & hierarchy
- Alignment
- Typography
- Color

These all apply to more complicated visualizations!

Some common techniques

That could help in the ML context...

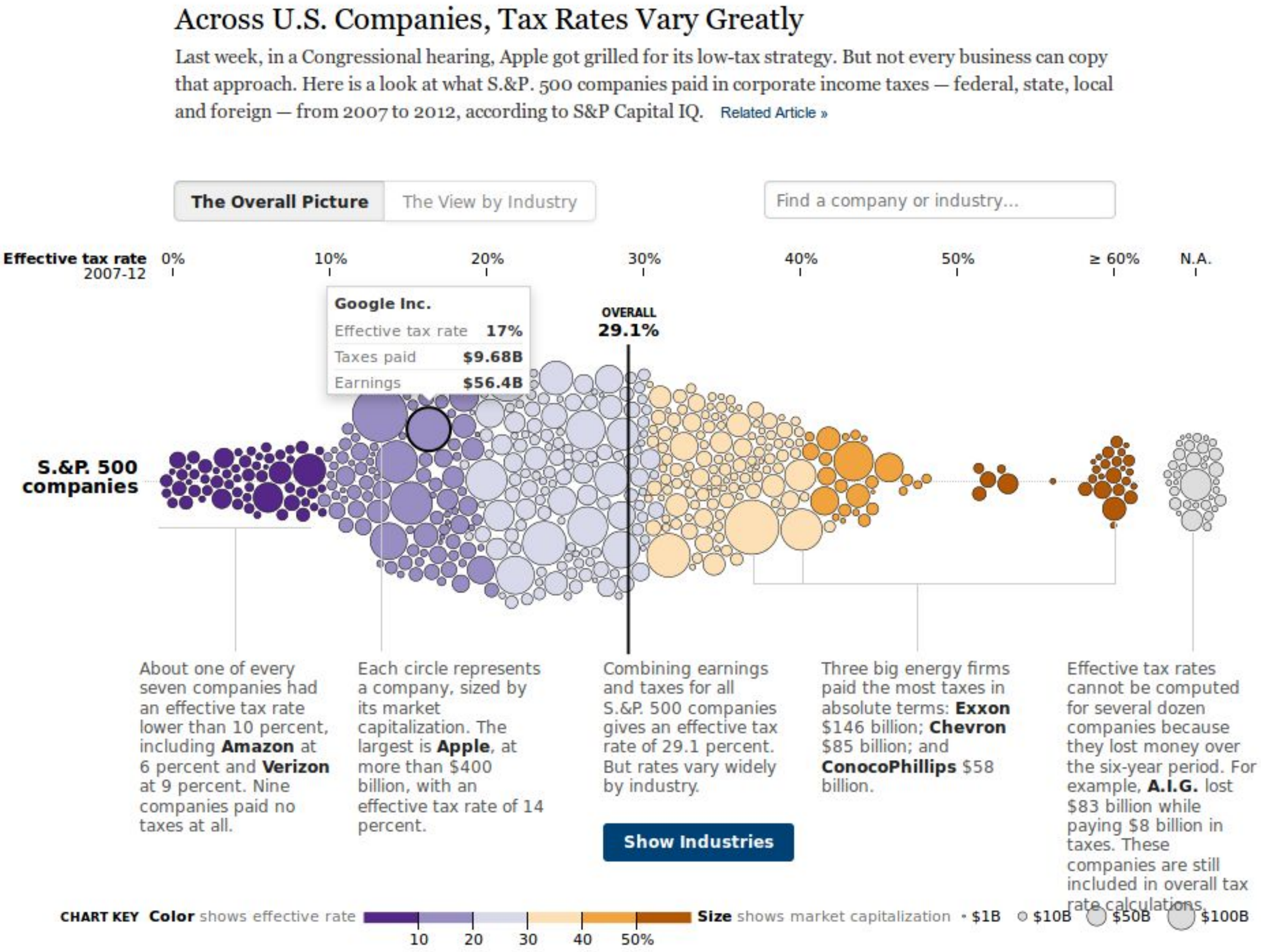
Data density: small multiples

[Drought's Footprint]

Haeyoun Park, Kevin Quealy

NY Times

Data faceting



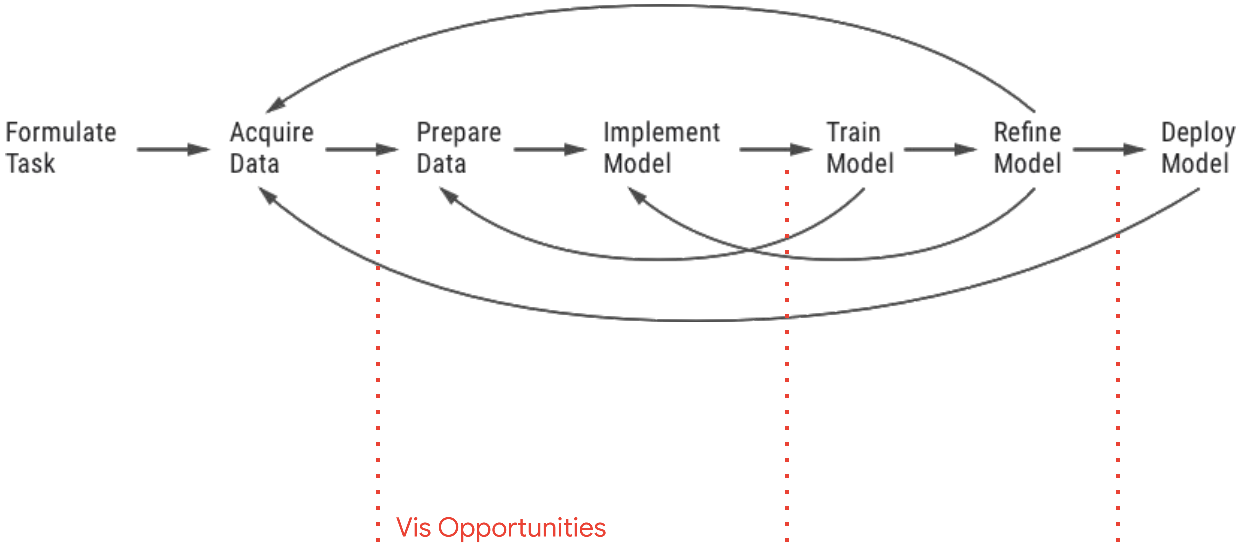
Across U.S. Companies, Tax Rates Vary Greatly

M. Bostock, M. Ericson, D. Leonhardt, B. Marsh

NY Times

Back to machine learning!

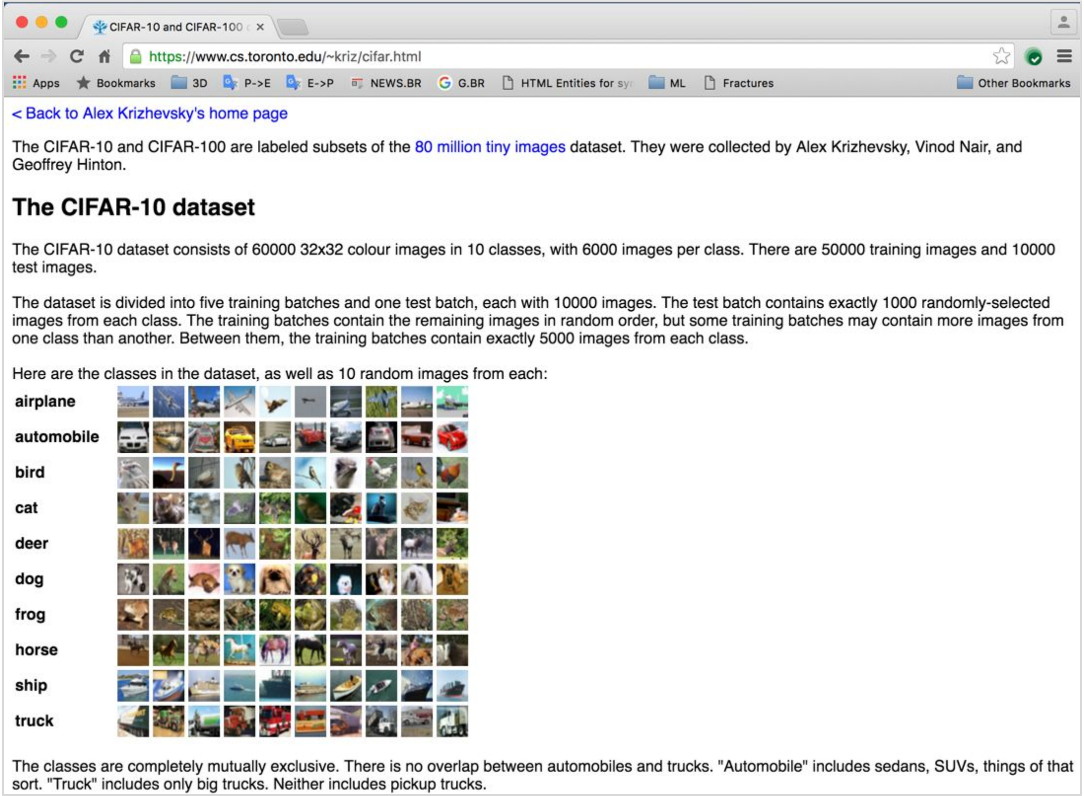
Opportunities for Vis



Source: Yannick Assogba

1. Visualizing training data

Visualizing CIFAR-10



CIFAR-10 and CIFAR-100

https://www.cs.toronto.edu/~kriz/cifar.html

Apps Bookmarks 3D P->E E->P NEWS.BR G.BR HTML Entities for sy ML Fractures Other Bookmarks

< Back to Alex Krizhevsky's home page

The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

The CIFAR-10 dataset

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

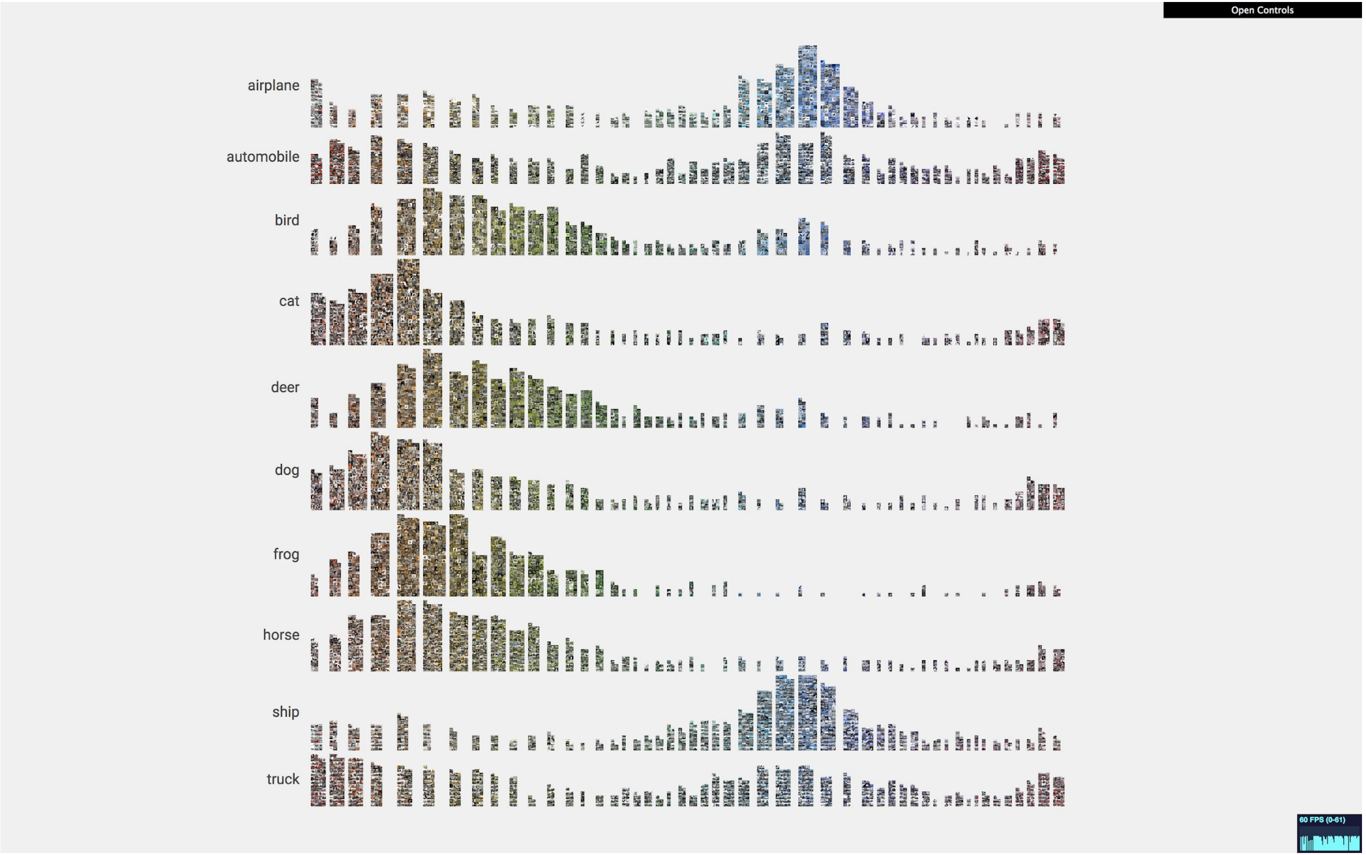
The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

Here are the classes in the dataset, as well as 10 random images from each:

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

The classes are completely mutually exclusive. There is no overlap between automobiles and trucks. "Automobile" includes sedans, SUVs, things of that sort. "Truck" includes only big trucks. Neither includes pickup trucks.

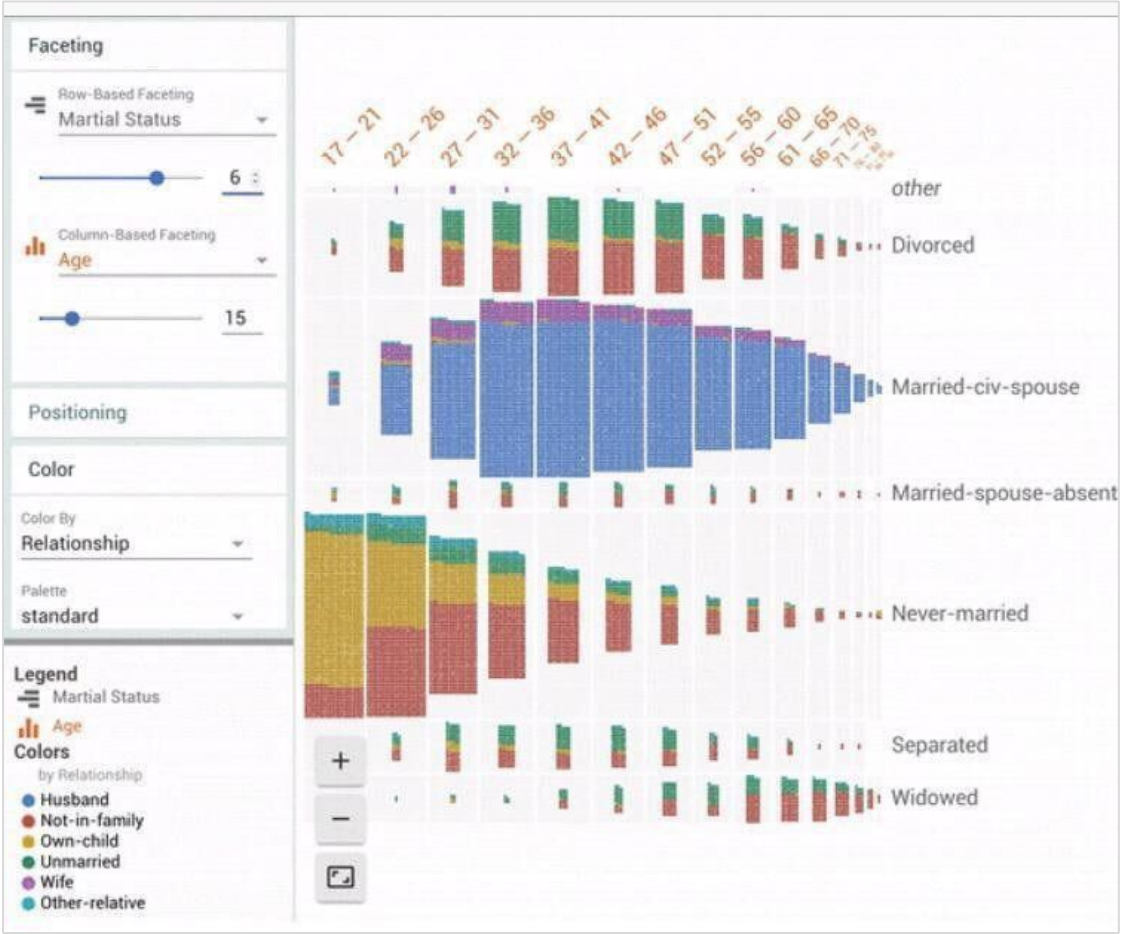
CIFAR-10 Facets Demo

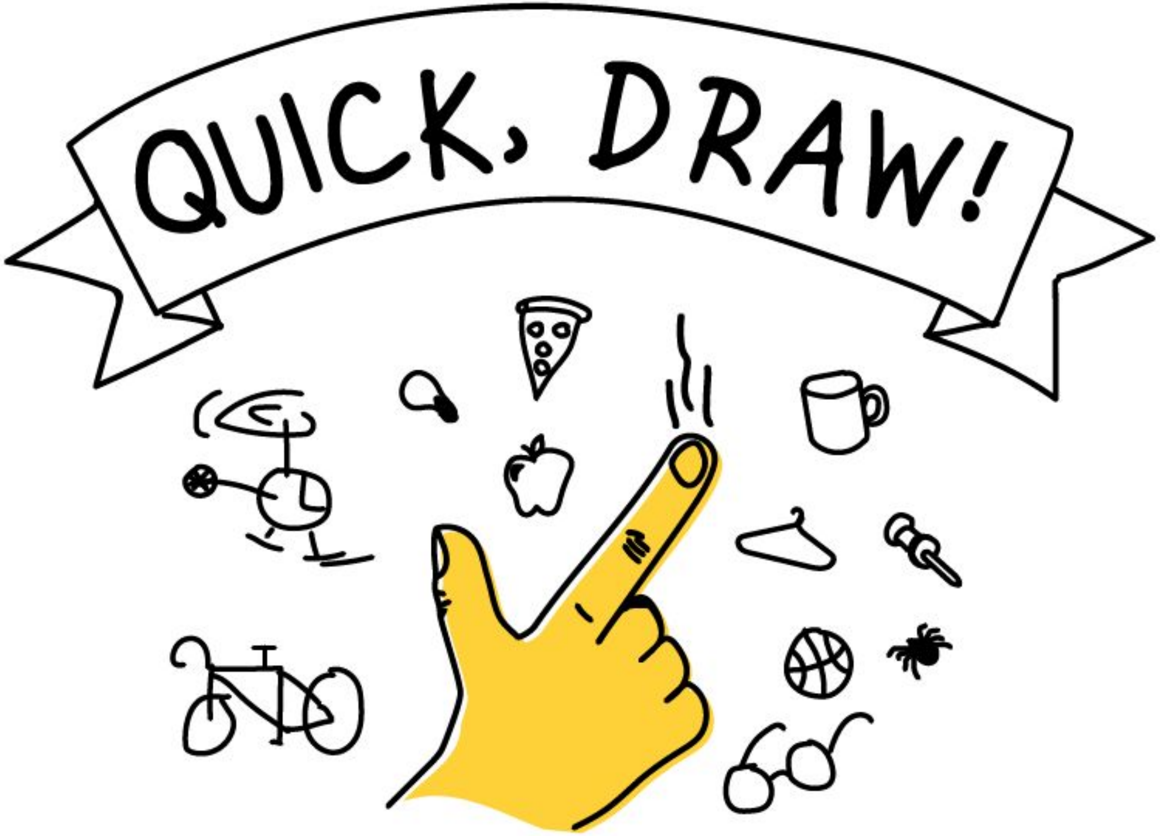


Facets

Open-source visualization

pair-code.github.io/facets



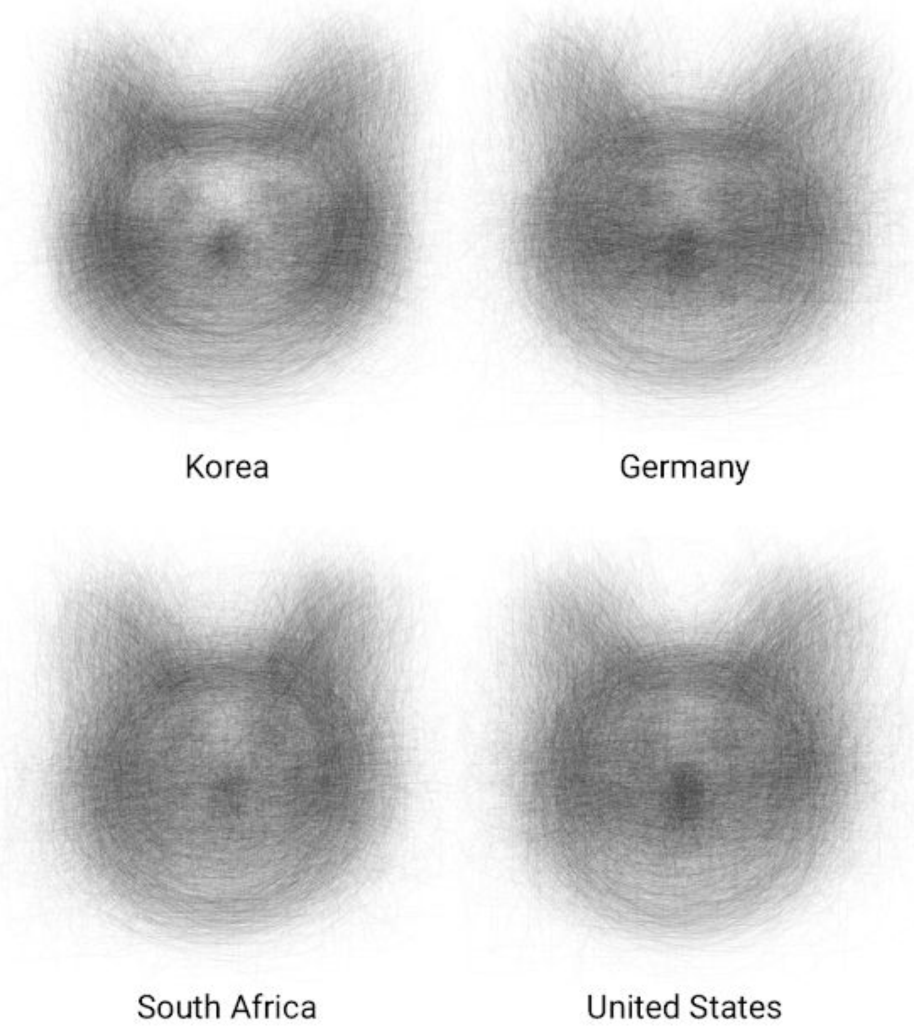


Google Creative Lab
https://quickdraw.withgoogle.com/

.Quick Draw, the data. .

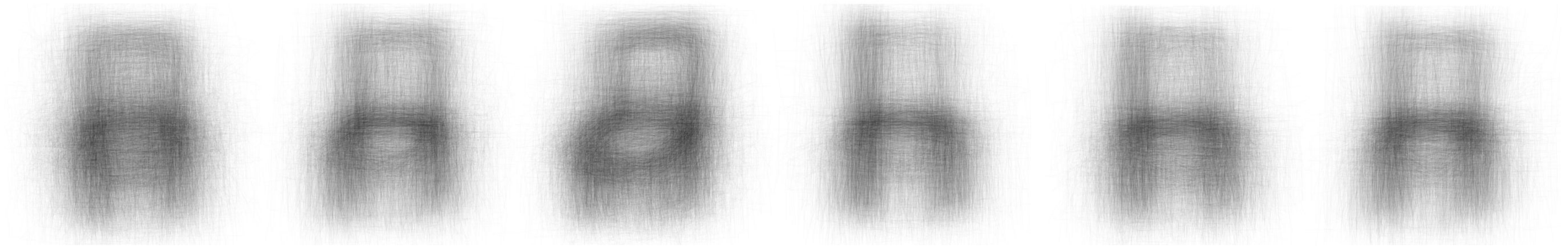
https://quickdraw.withgoogle.com/data. .

When things look alike across cultures



[Machine Learning for Visualization](#)
Let's Explore the Cutest Big Dataset
Ian Johnson

And when they don't

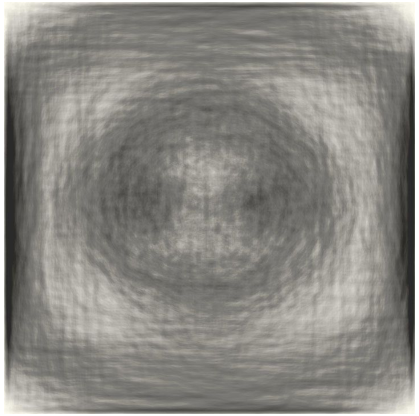


South Africa   Russia   Korea   Brazil   United States   Germany

[Visual Averages by Country]

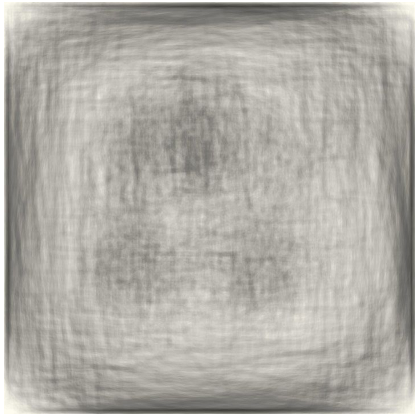
Kyle McDonald

Outlets

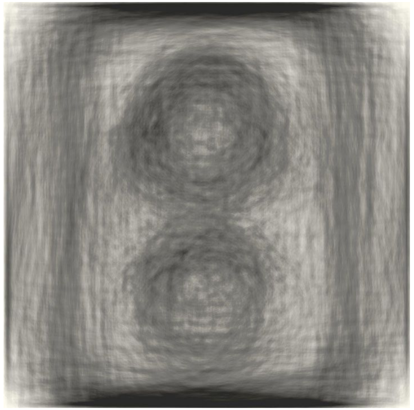


Germany

Japan



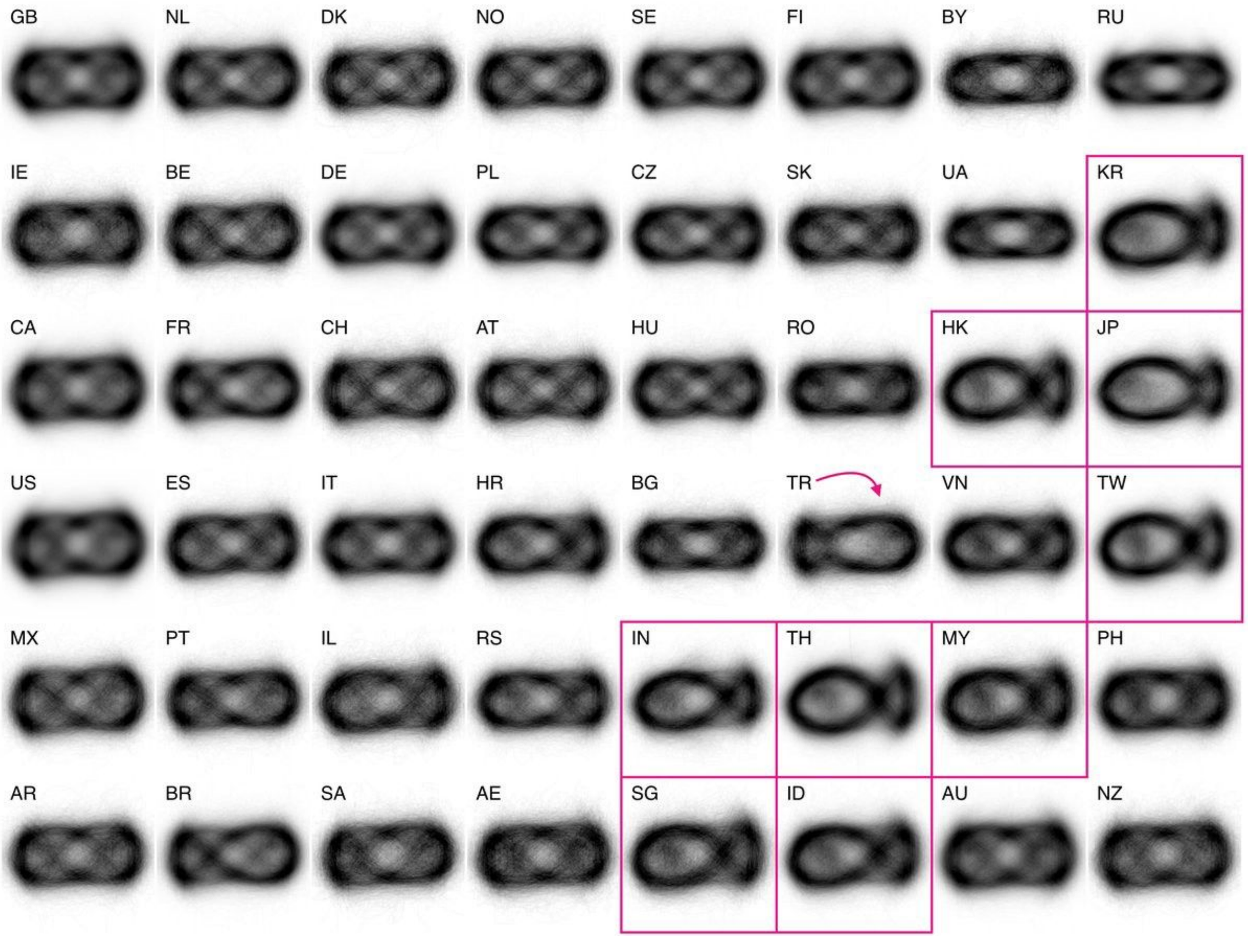
Malaysia



Sweden

[Visual Averages by Country](#)
Kyle McDonald

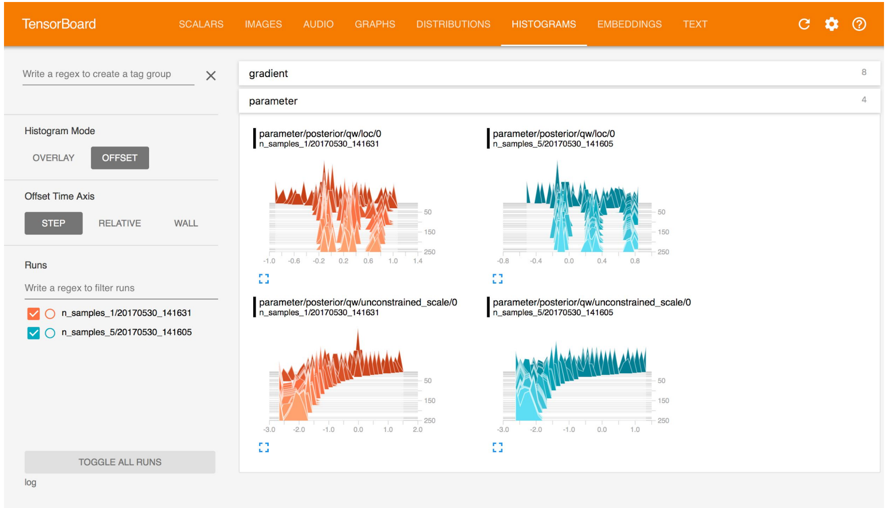
Finding nemo: small multiples



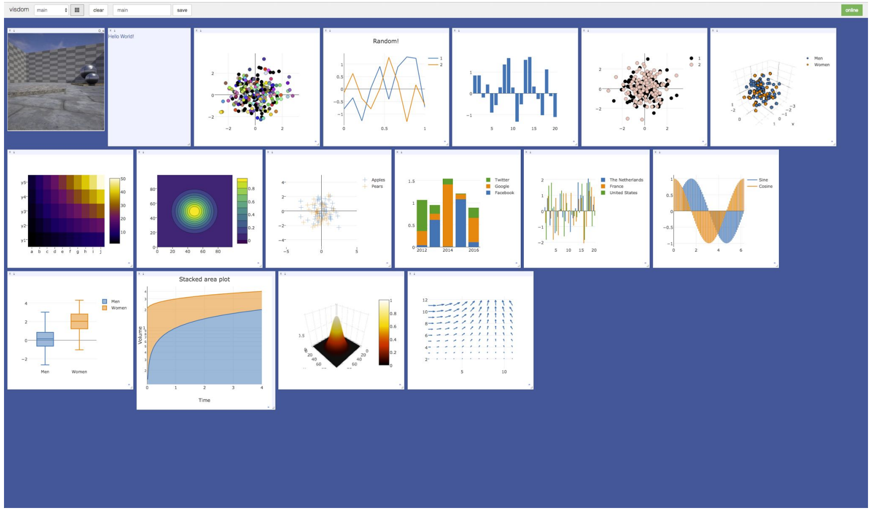
[Visual Averages by Country]
Kyle McDonald

2. Performance monitoring (very briefly!)

Monitoring dashboards - apply standard visualization tools!



TensorBoard



Visdom

Two examples among many...

3. Interpretability + model inspection

Convolutional NNs

Image classification: interpretability petri dish

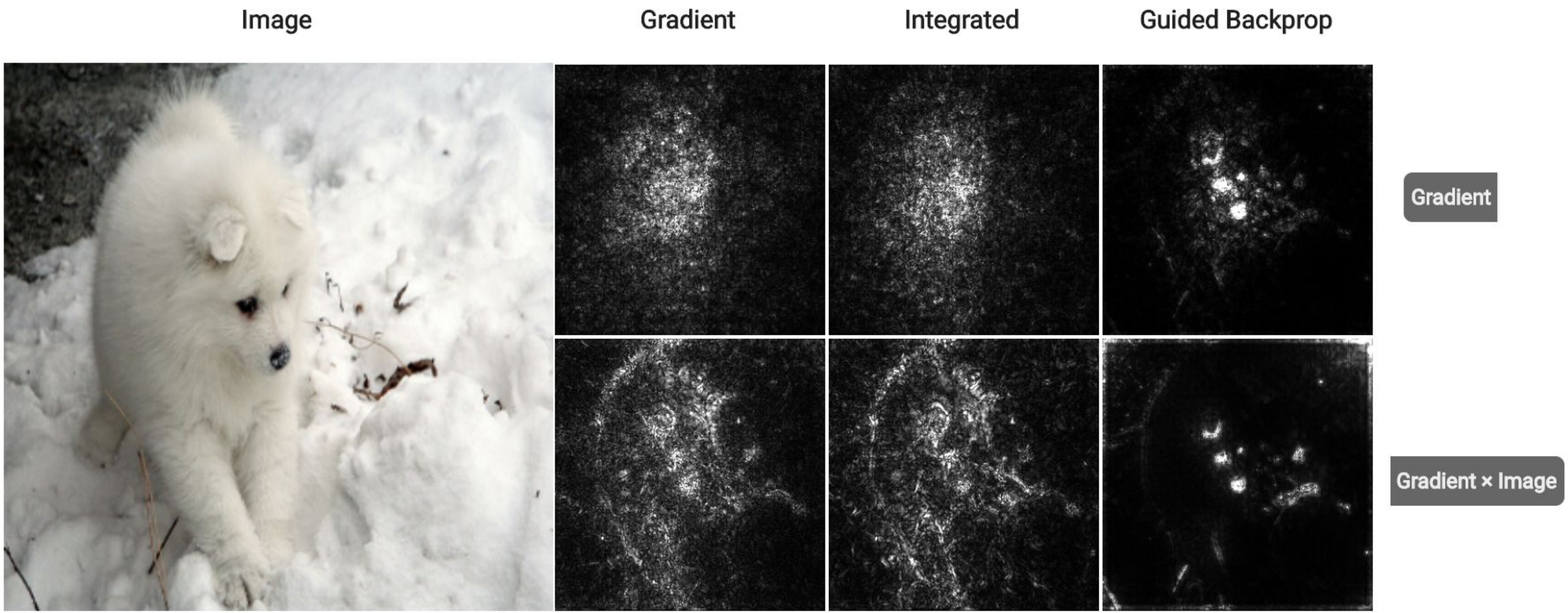
Image classifiers are effective in practice

Exactly what they're doing is somewhat mysterious
- And their failures (e.g. adversarial examples) add to mystery

But: Way easier to inspect what's going on in artificial classifiers than in human classifiers ;-)

Since these are visual systems, it's natural to use visualization to inspect them
- What features are these networks really using?
- Do individual units have meaning?
- What roles are played by different layers?
- How are high-level concepts built from low-level ones?

Saliency maps - examples



More comparisons: https://pair-code.github.io/saliency/

Saliency maps
(a.k.a. "Sensitivity maps")

Idea: consider sensitivity of class to each pixel
i.e. grad(f), where f is function from pixels to class score.

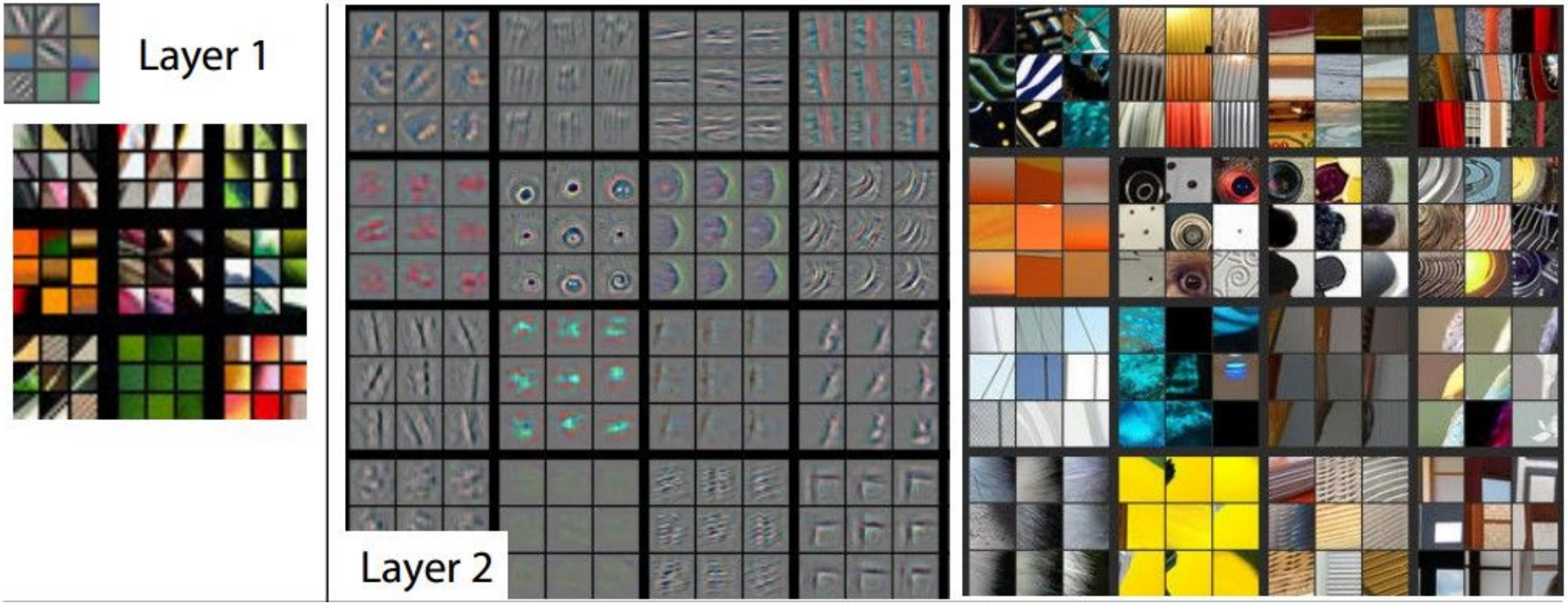
Many ways to extend basic idea!

- Layer-wise relevance propagation (Binder et al.)
- Integrated gradients (Sundararajan et al.)
- Guided backprop (Springenberg et al.)
- etc.

Yet interpretation is slippery (Adebayo et al., Kindermans et al.)

- Tend to be visually noisy. Are these sometimes Rorschach tests?
- Are some of these methods essentially edge detectors?

Visualizing arbitrary neurons along the way to the top...



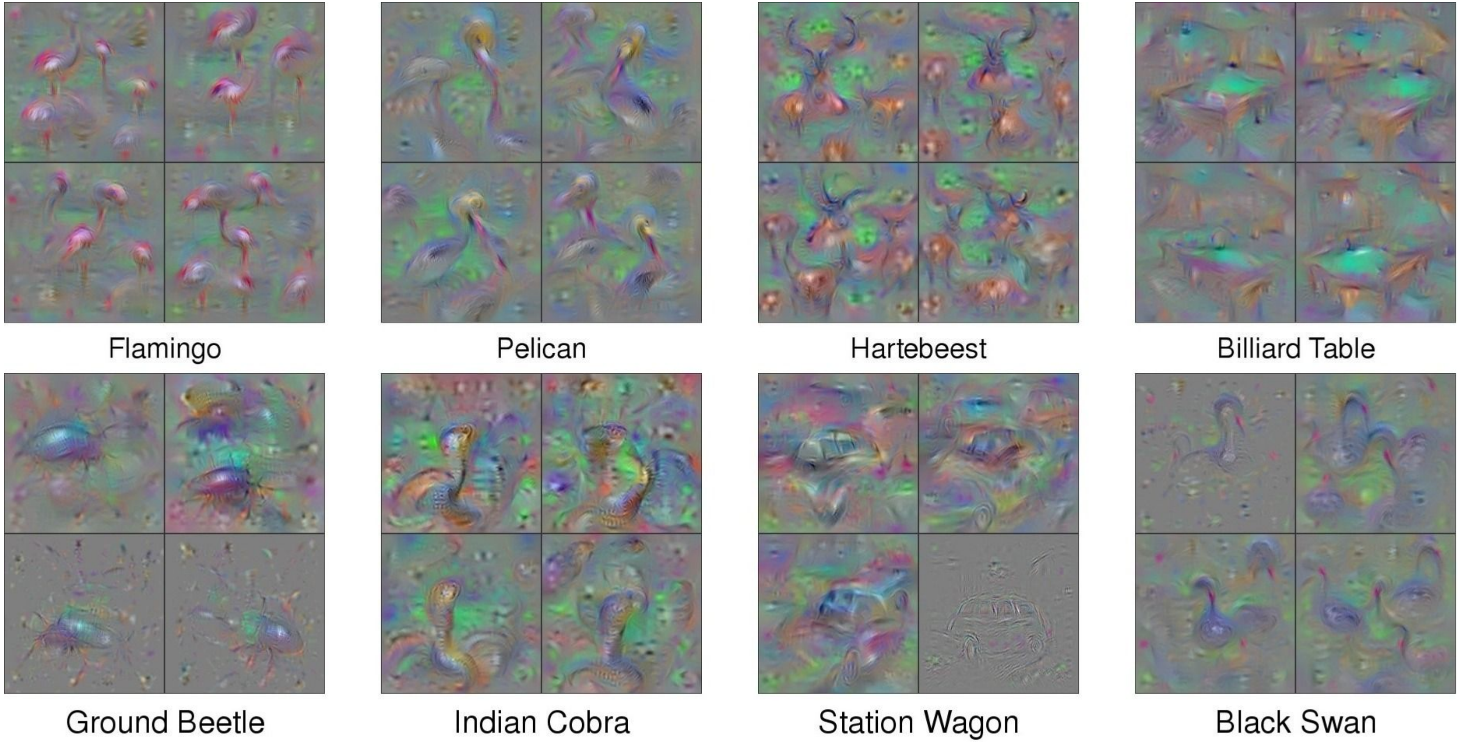
Gray: trying to maximize neural response. Colorful squares: maximal examples from an image data set

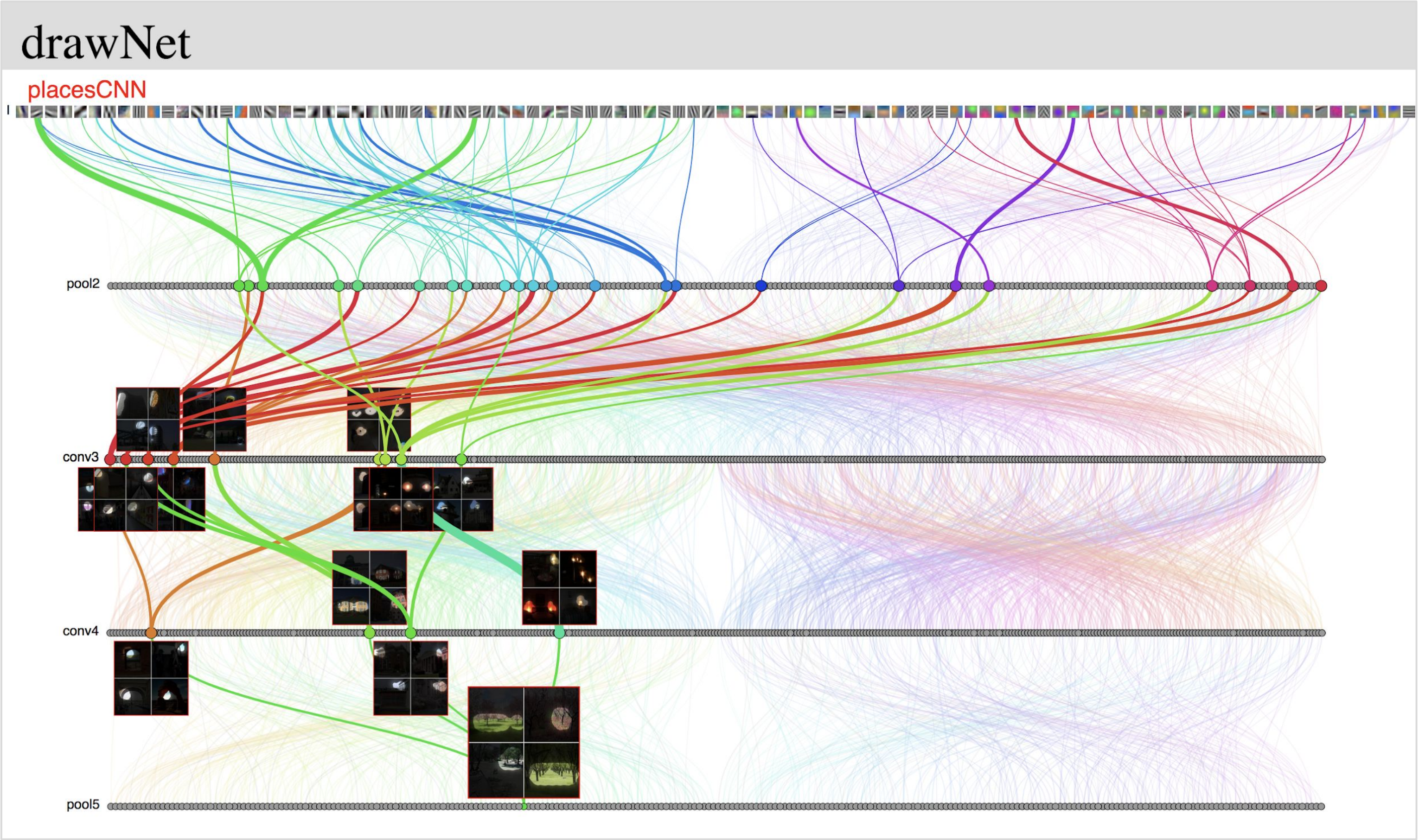
Visualizing and Understanding Convolutional Networks
Zeiler & Fergus, 2013

Understanding Neural Networks Through Deep Visualization

Yosinski et al. , 2015

http://yosinski.com/deepvis

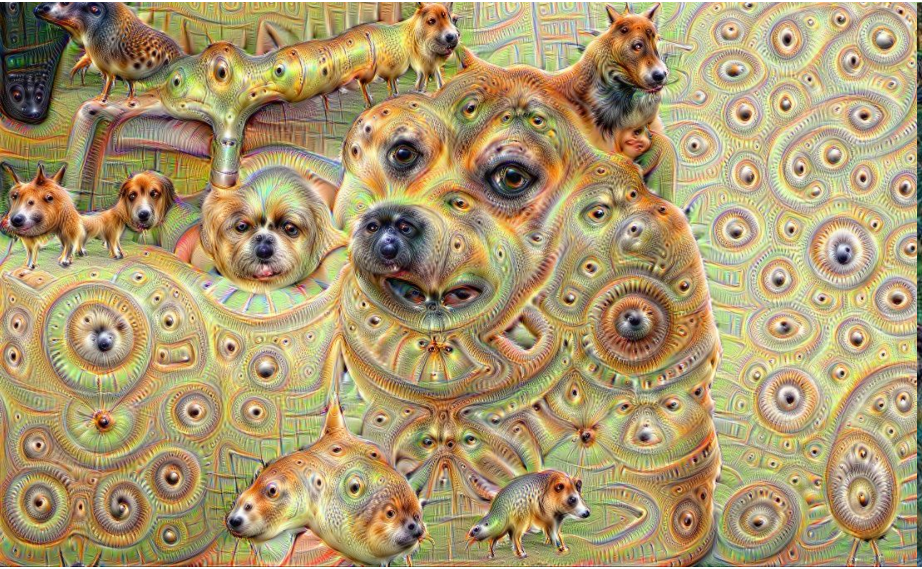
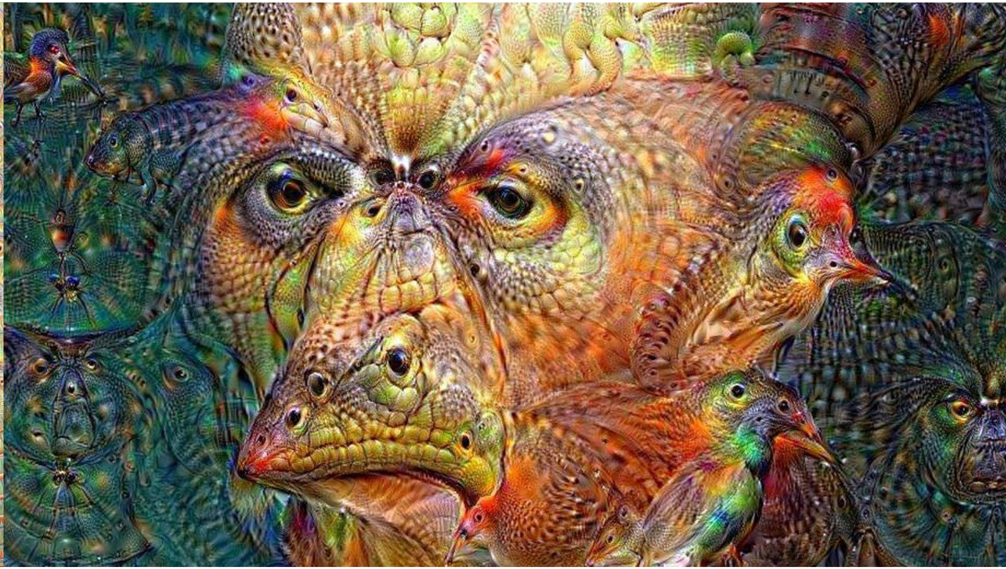




drawNet

Torralba

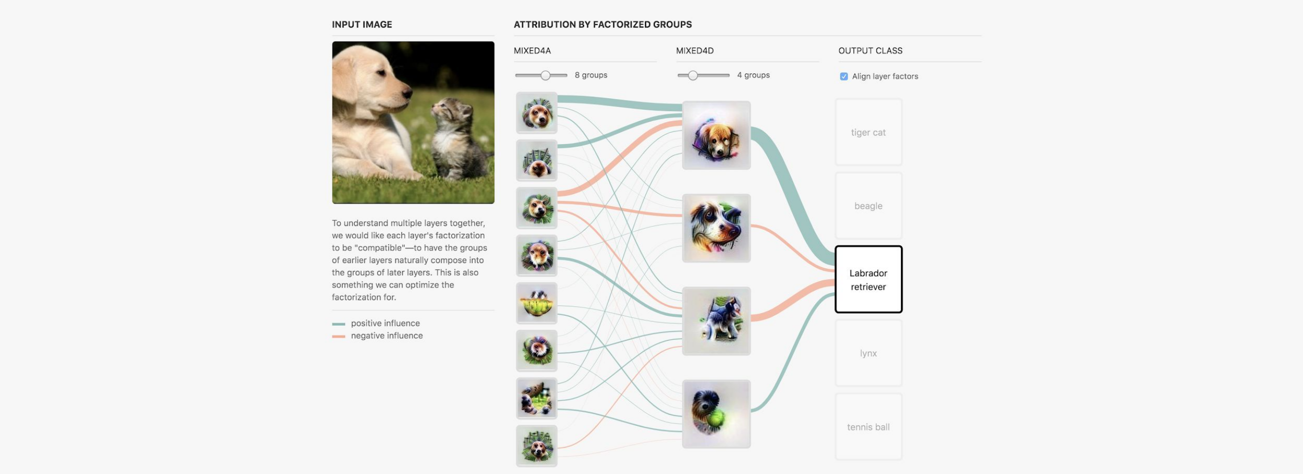
Deep Dream

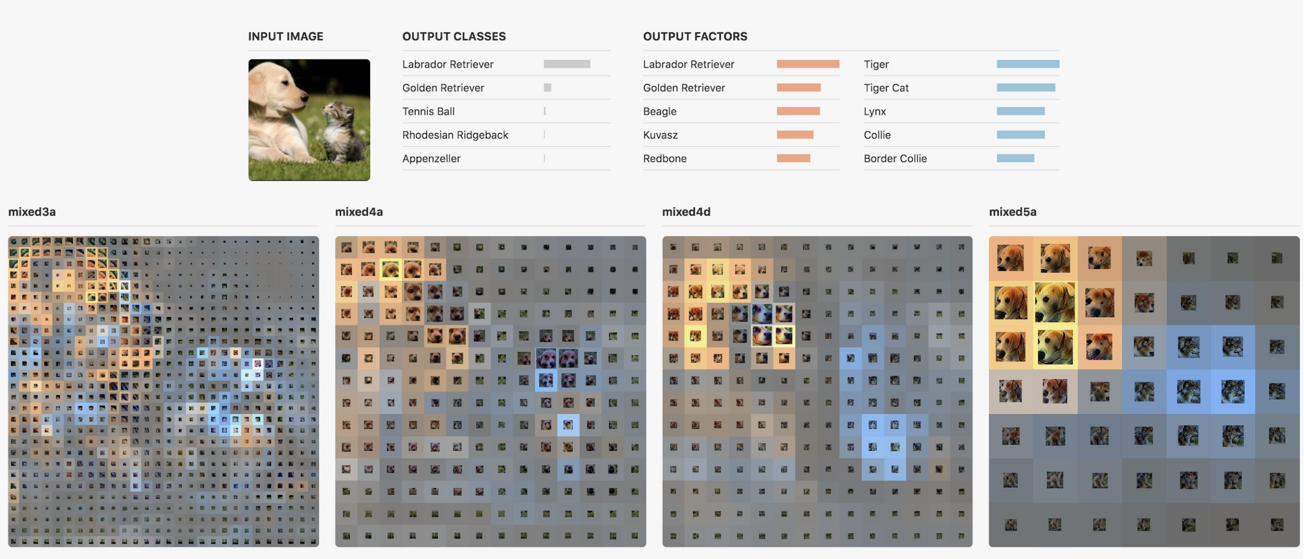



deepdream

Mordvintsev, Tyka, Olah

Combining these interpretability ideas to create new visualizations

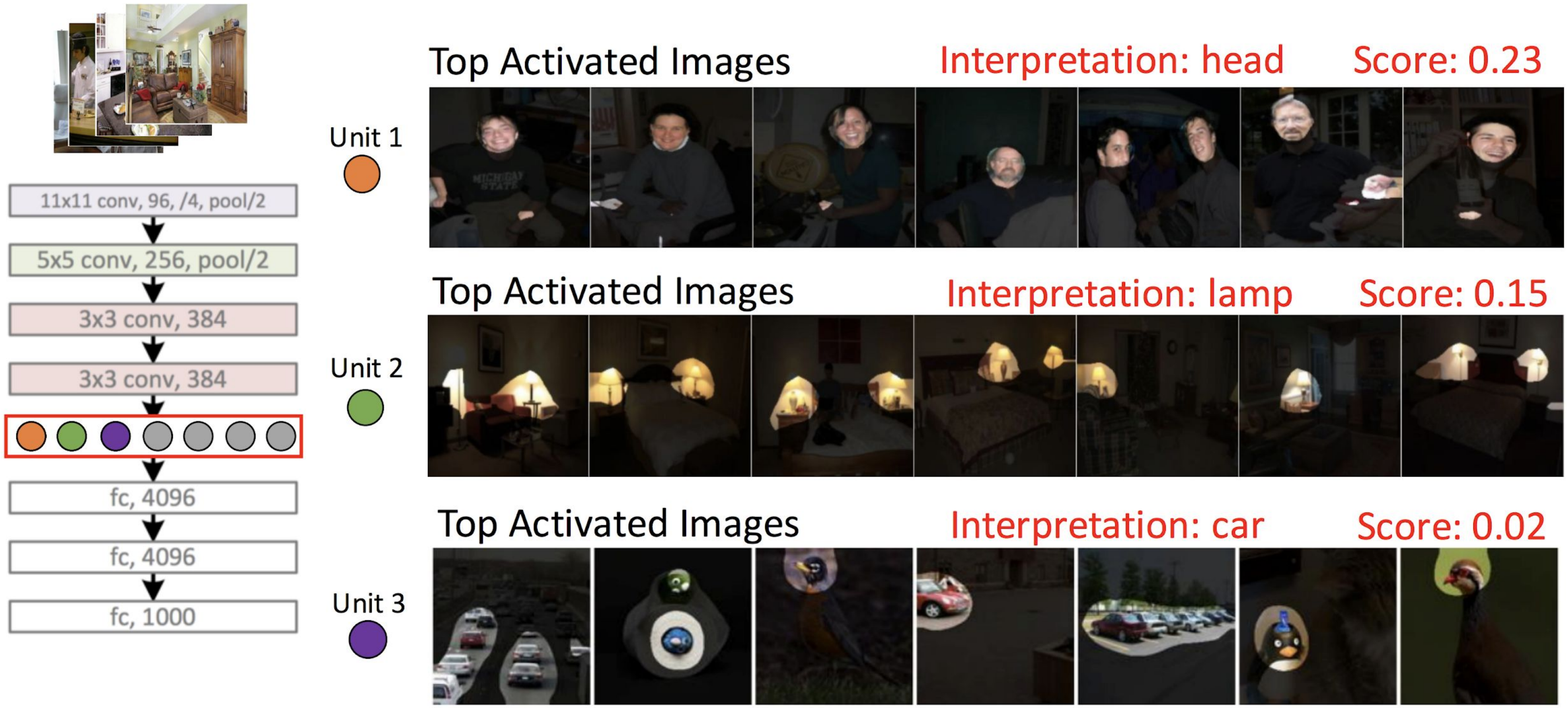




The Building Blocks of Interpretability

Olah, Satyanarayan, Johnson, Carter, Schubert, Ye, Mordvintsev

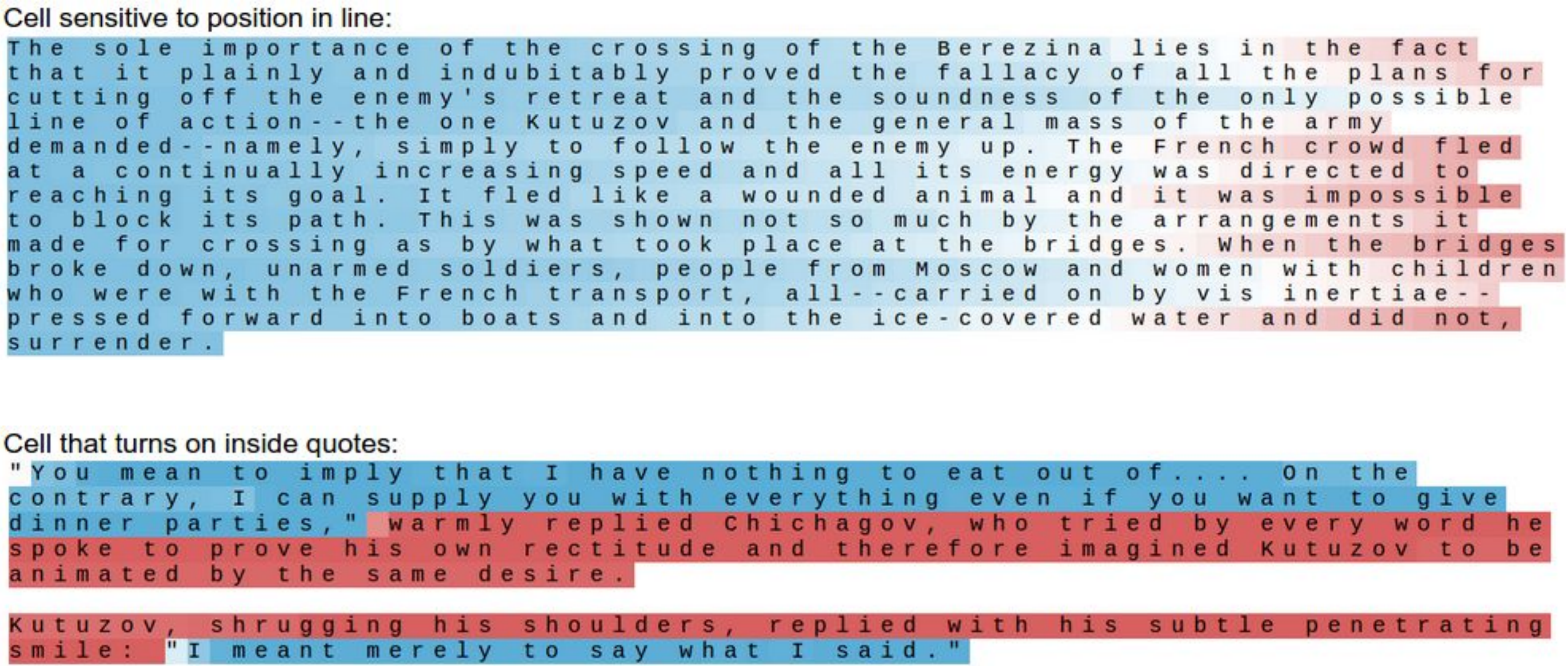
Going From Visualization to Interpretation



Interpreting Deep Visual Representations
Bau, Khosla, Oliva, Torralba

RNNs

Visualizing text sequences, colored by activations of a cell



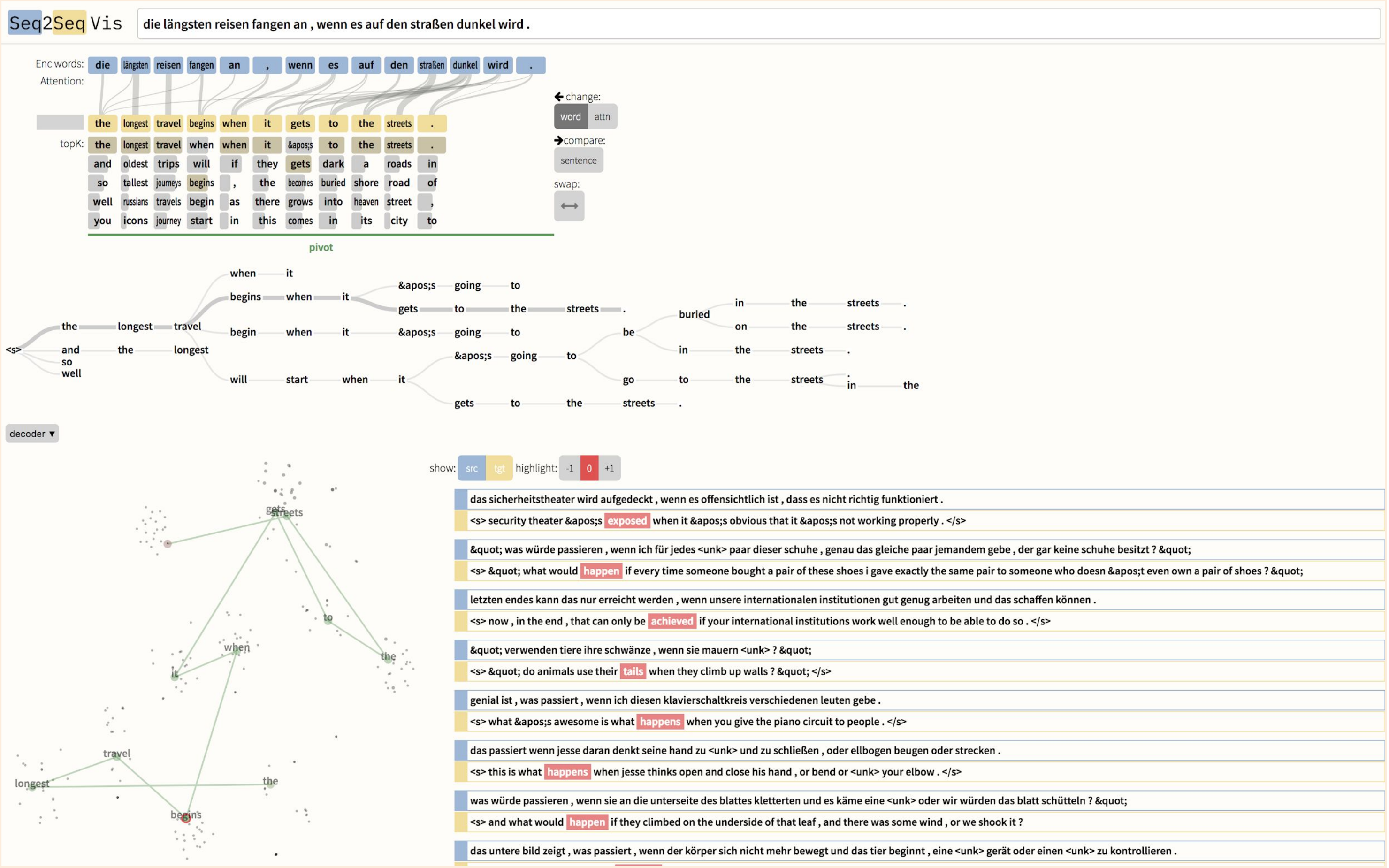
The Unreasonable Effectiveness of Recurrent Neural Networks
Karpathy, 2015

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
    siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

The Unreasonable Effectiveness of Recurrent Neural Networks

Karpathy, 2015



Seq2Seq-Vis: Visual Debugging Tool for Sequence-to- Sequence Models

Strobelt, 2018

Examine model decisions

Connect decisions to previous examples

Test alternative decisions

4. High-dimensional data

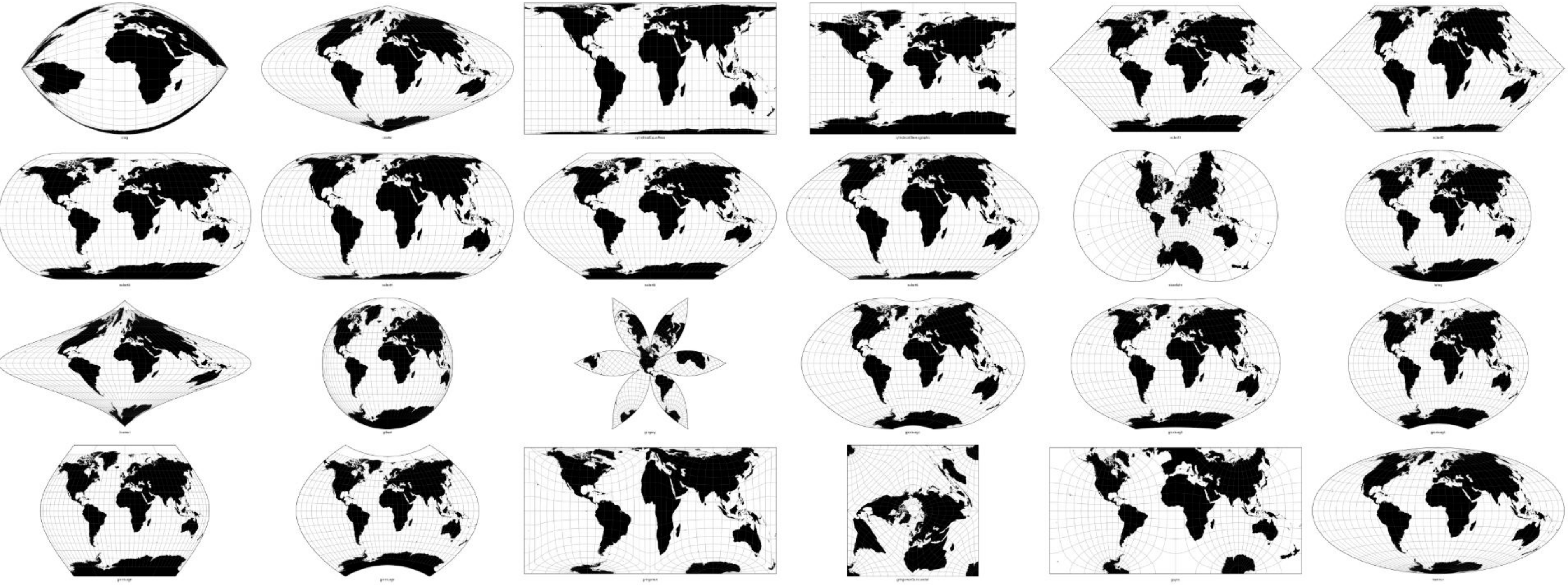
Why high-dimensional data?

Vectors spaces are the lingua franca of much of ML these days

- Data such as images, audio, video is naturally high-dimensional
- Dense representations of discrete data (e.g. word embeddings) have had major successes

Why is it hard? Because it's impossible

Why is it hard? Because it's impossible



See Every Map Projection, Bostock.

Main approaches

Linear

- Principal Component Analysis (show as much variation in data as possible)
- [Visualization of Labeled Data Using Linear Transformations]() (clusters match labels)

Non-linear (just a few of many)

- Multidimensional scaling
- Sammon mapping
- Isomap
- **t-SNE**
- **UMAP**

Minimize distortion, according to some metric

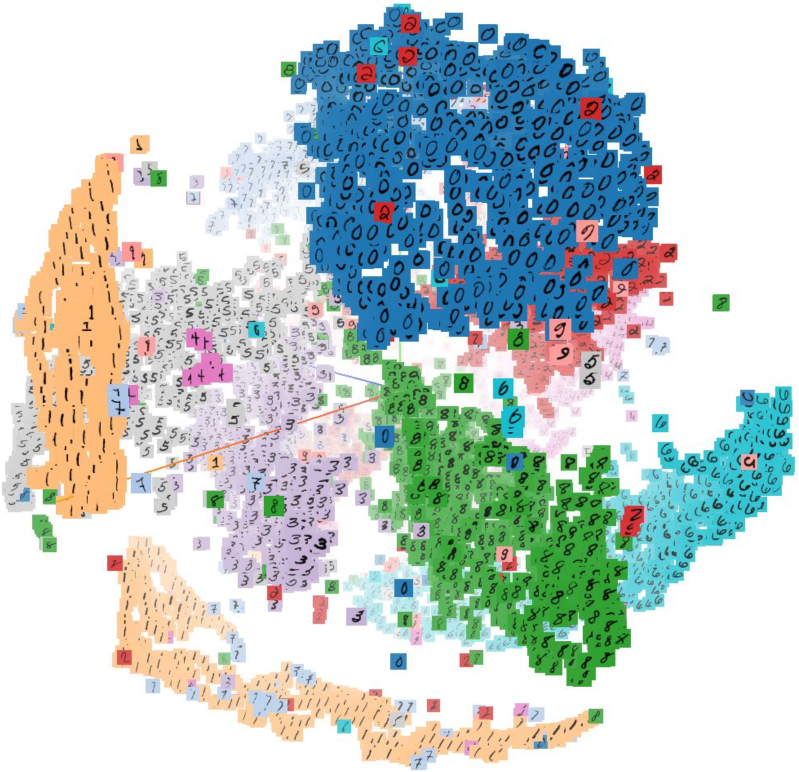
t-SNE

Fairly complex non-linear technique

Uses an adaptive sense of "distance." Translates well between geometry of high- and low-dimensional space

Has become a standard tool, so we'll spend some time discussing how to read it.

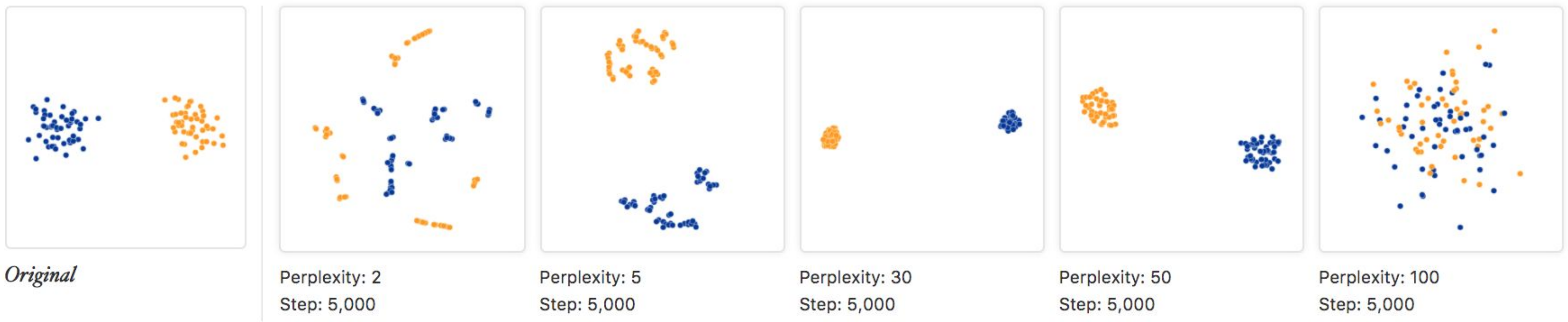
Demo: MNIST visualization



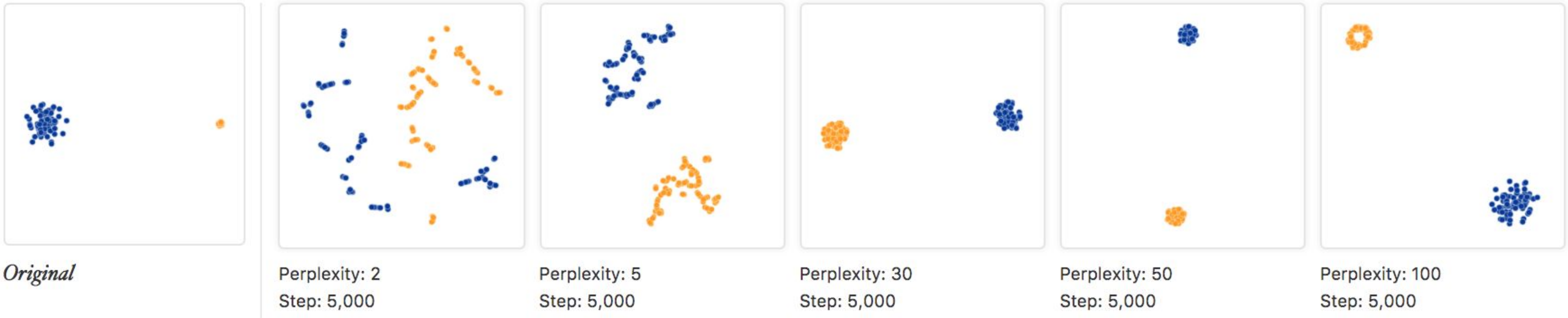
Embedding Projector

Open Source visualization tool
Also available on Tensorboard
projector.tensorflow.org/

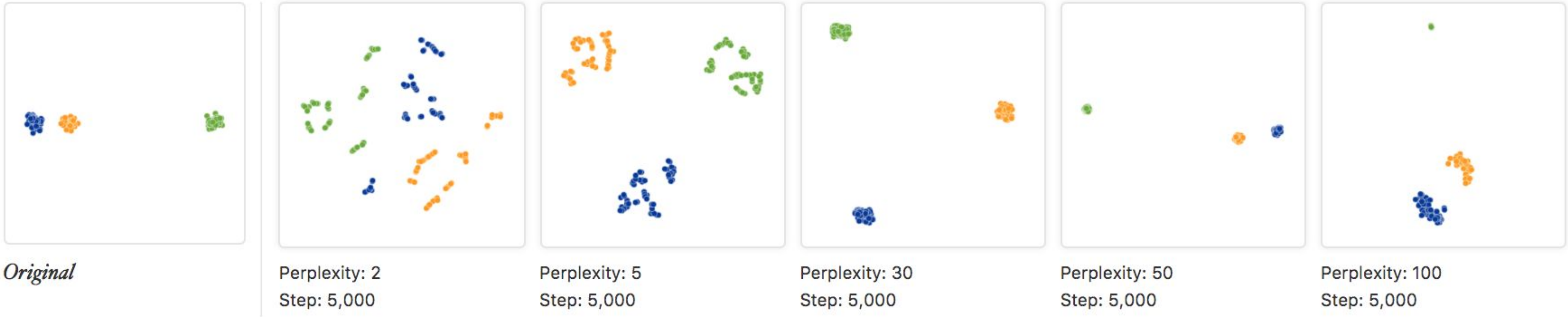
Those hyperparameters really matter



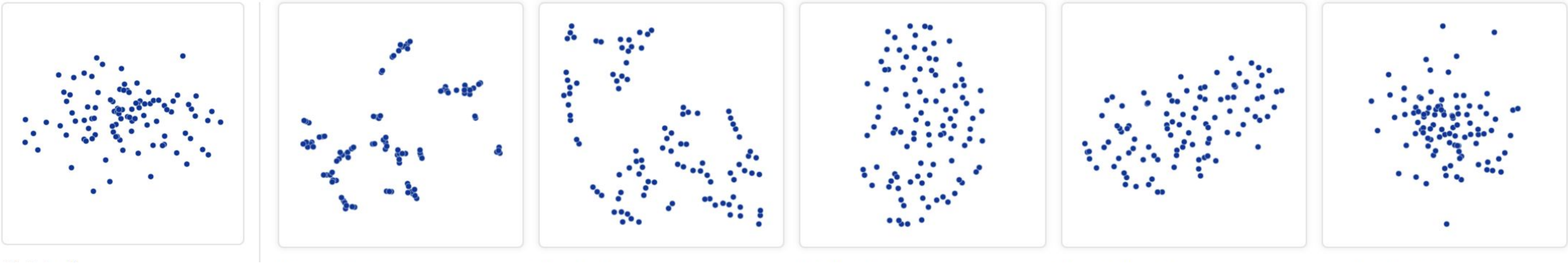
Cluster sizes in a t-SNE plot mean nothing

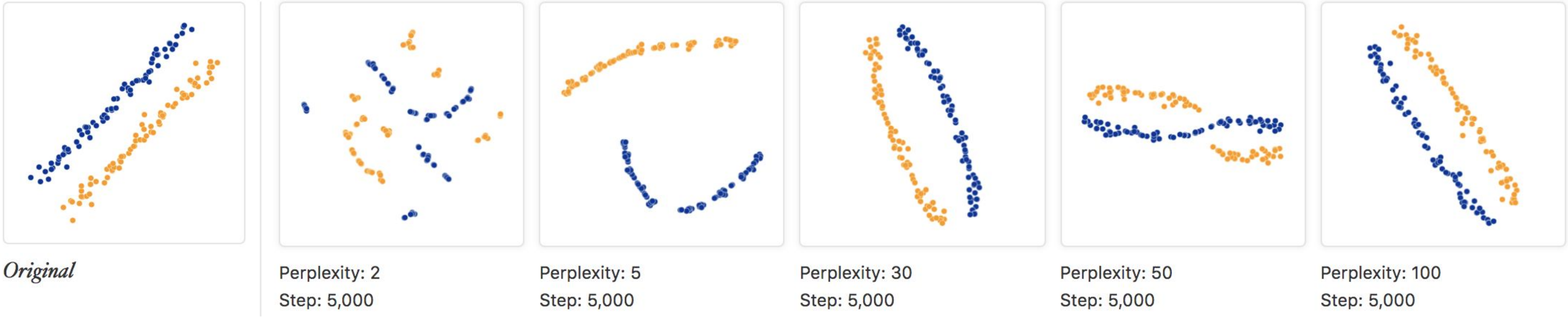


Distances between clusters may not mean much

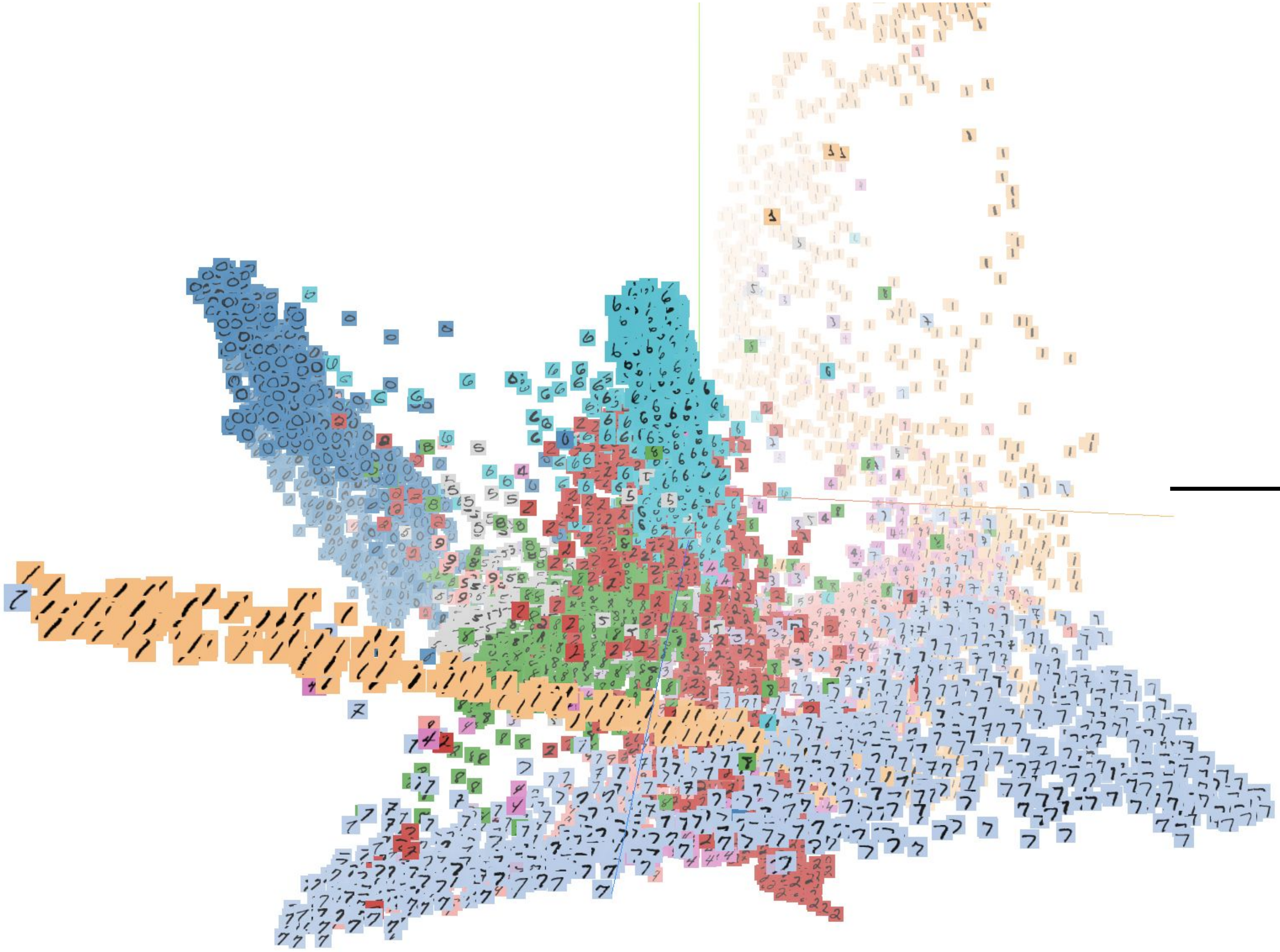


You can see some shapes, sometimes

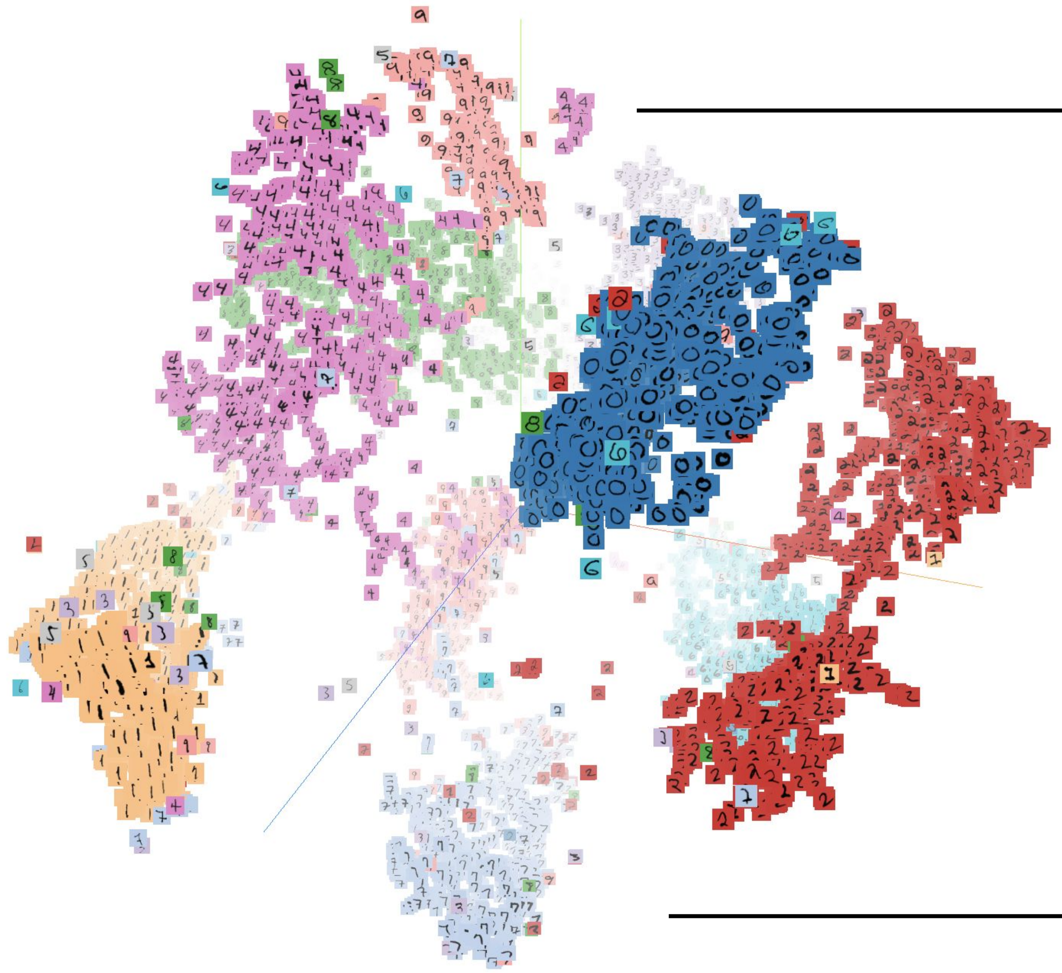




Let's try this out with MNIST

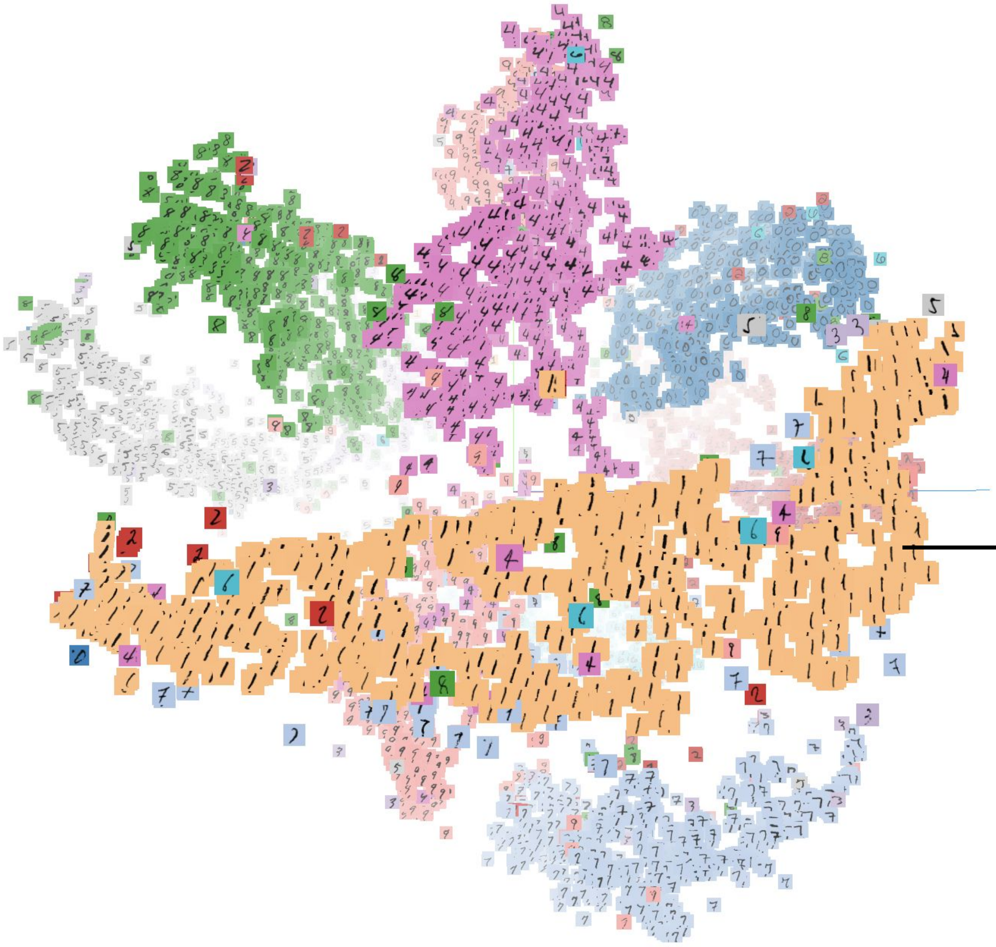


Stopping too soon yields weird artifacts.



The 4's may not be separated into two clusters.

Clusters seem about equally far apart in 3D; may not actually be.



The clusters of 1's probably is long and thin.

UMAP: New kid on the block

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes and John Healy
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com jchealy@gmail.com

February 13, 2018

UMAP: New kid on the block

Practical value

- Faster than t-SNE
- Can efficiently embed into high dimensions (i.e. useful not just for visualization)
- Often seems to capture global structure better

UMAP: New kid on the block

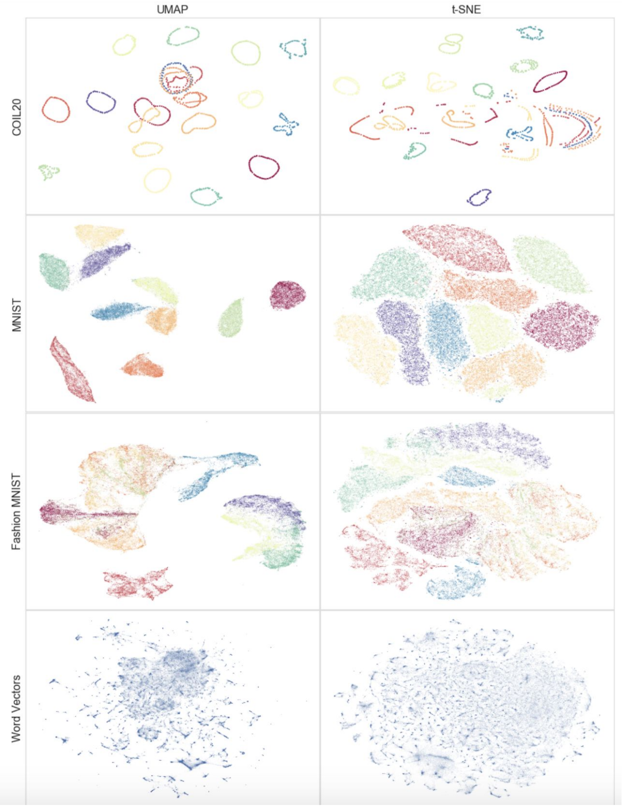
Practical value

- Faster than t-SNE
- Can efficiently embed into high dimensions (i.e. useful not just for visualization)
- Often seems to capture global structure better

Theory

- Roughly: manifold learning combined with explicit topology
- In detail: I don't completely understand the theory!
    - This note does an amazing job of extracting key bits of UMAP paper: https://www.math.upenn.edu/~jhansen/2018/05/04/UMAP/

UMAP: New kid on the block



Comparison of UMAP (left) and t-SNE (right) from McInnes & Healy.

Global structure does seem to emerge more in UMAP.

For more

Let's compare in real-time on an audio data set!

Comparative Audio Analysis With Wavenet, MFCCs, UMAP, t-SNE and PCA

(Leon Fedden)

Putting this together

The Beginner's Guide to Dimensionality Reduction
Matthew Conlen and Fred Hohman

https://idyll.pub/post/dimensionality-reduction-293e465c2a3443e8941b016d/
(just Google "Beginner's Guide to Dimensionality Reduction")

Pitfalls of high-dimensional space

Geometry of high-dimensional space holds many surprises...
Be careful about interpreting visualizations!

Adding "**usually**," "**most**," and "**approximately**" where appropriate:

- Two random vectors are perpendicular
- A standard Gaussian distribution is just a uniform distribution on a sphere
- A random matrix is a scalar multiple of an orthogonal matrix
- Random walks all have the same shape

Model interpretability example

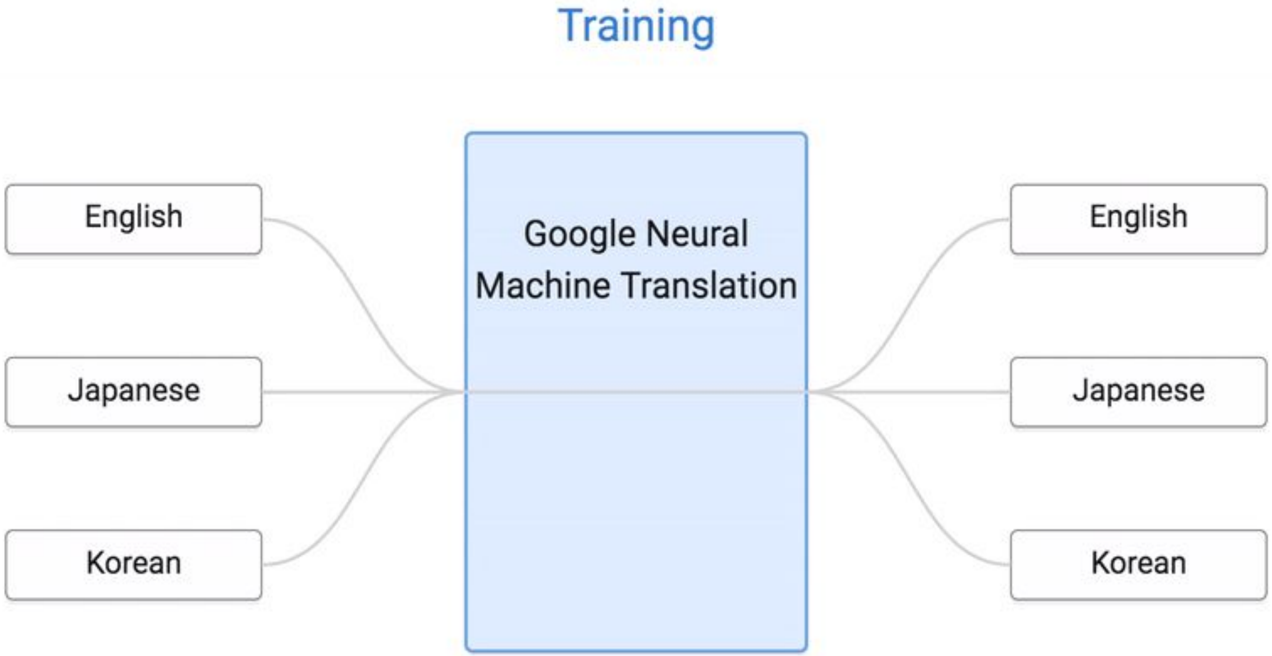
Multi-lingual translation

What does the language embedding space look like?

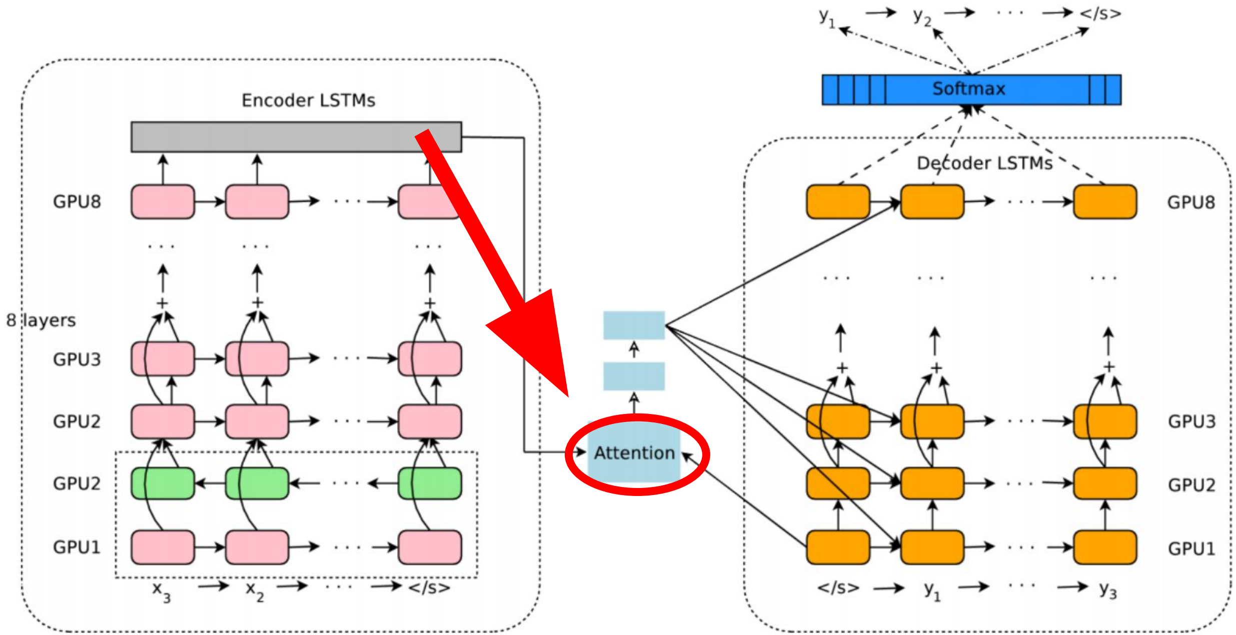
https://arxiv.org/abs/1611.04558
Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean

Training: English ⟵ ⟶ Japanese
English ⟵ ⟶ Korean
Japanese ⟵ ⟶ Korean (zero shot)

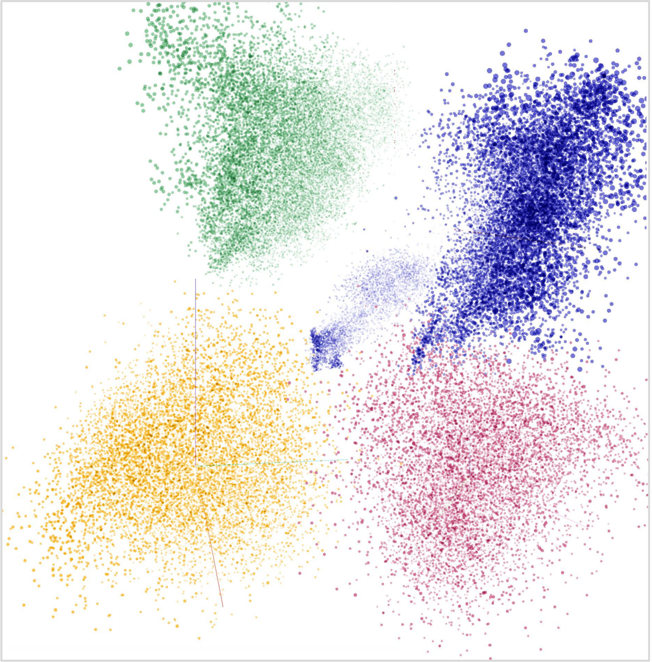


Visualize internal representation ("embedding space")

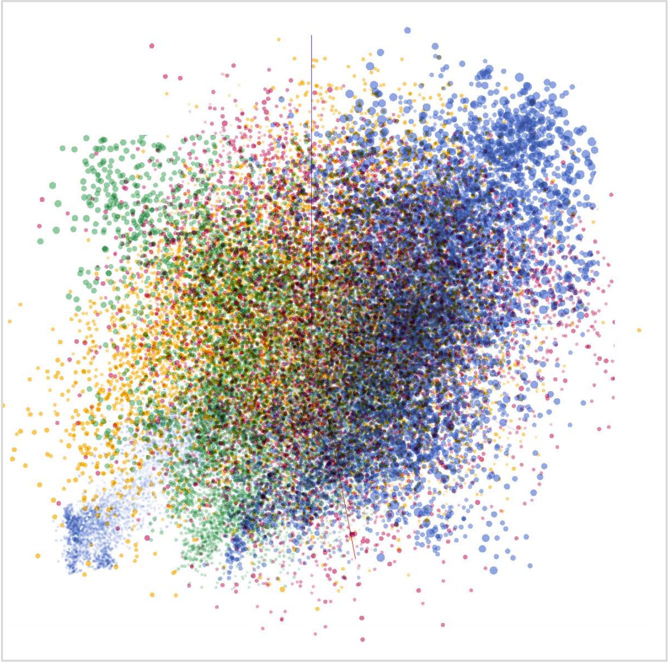


Research question

What does the multi language embedding space look like?



or



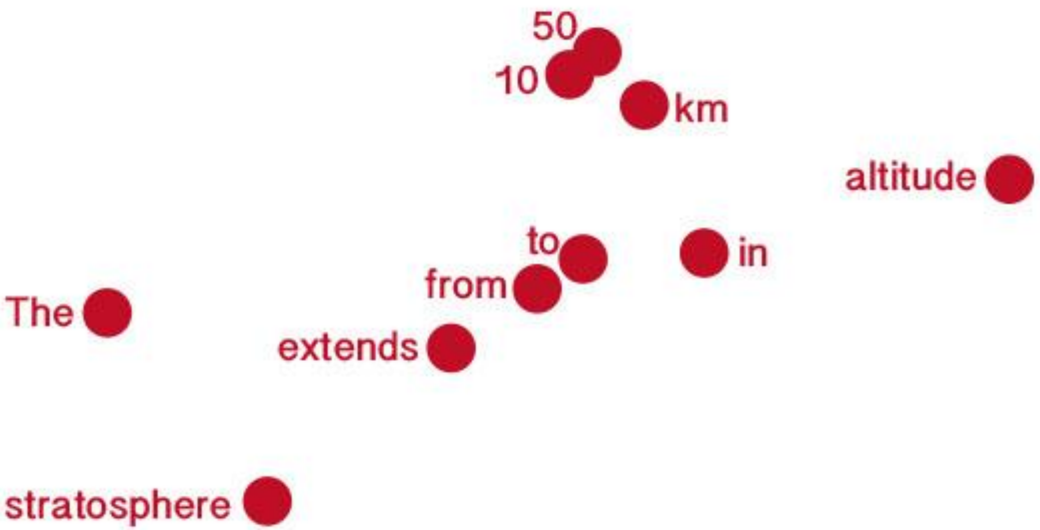
Note: not real data

What does a sentence look like in embedding space?

(points in 1024-dim space: the data that the decoder receives)

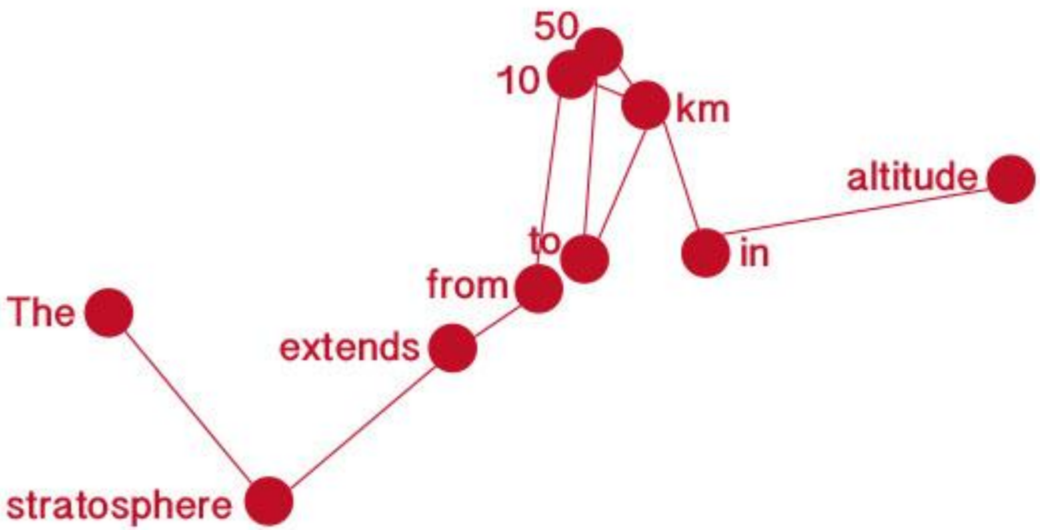
E.g. "*The stratosphere extends from 10km to 50km in altitude*"

What does a sentence look like in embedding space?



Note: simplification of real situation!

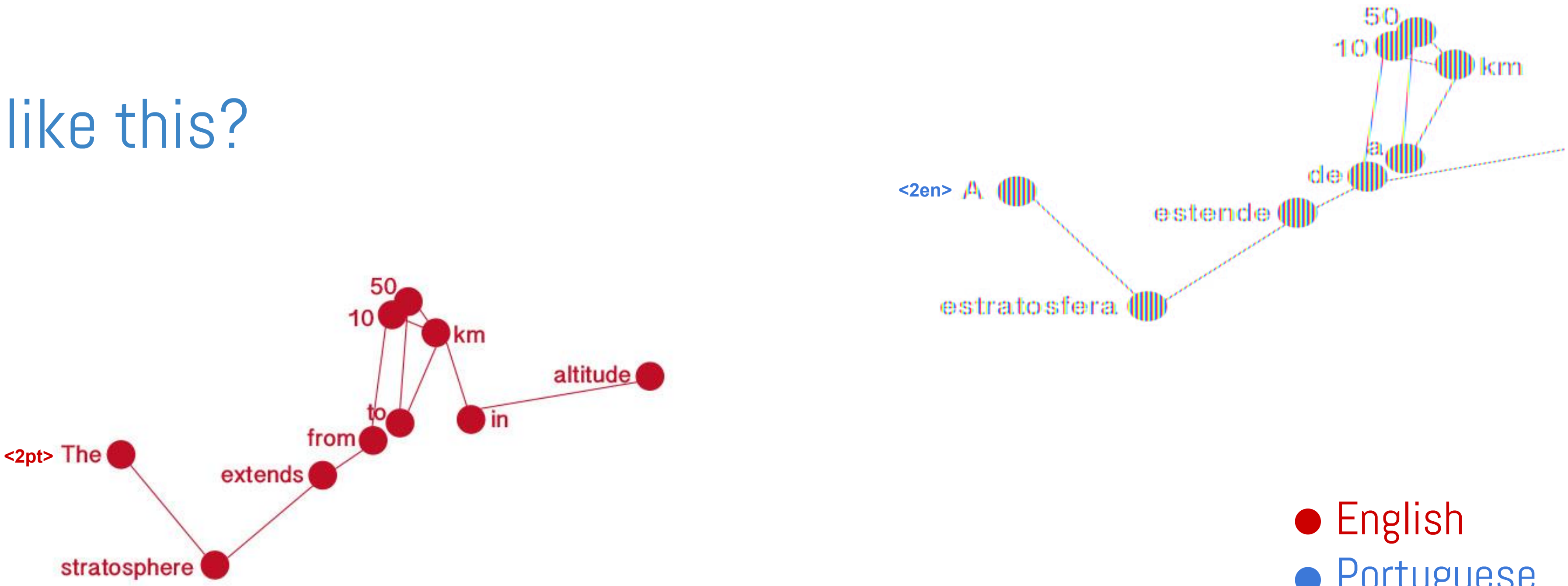
What does a sentence look like in embedding space?



What do parallel sentences look like in embedding space?

(same meaning, different language)

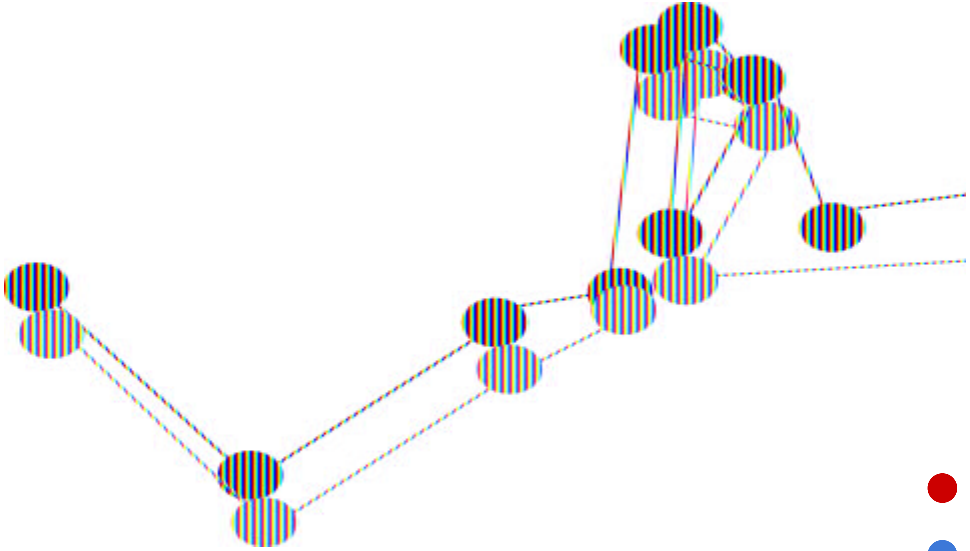
like this?



What do parallel sentences look like in embedding space?

(same meaning, different language)

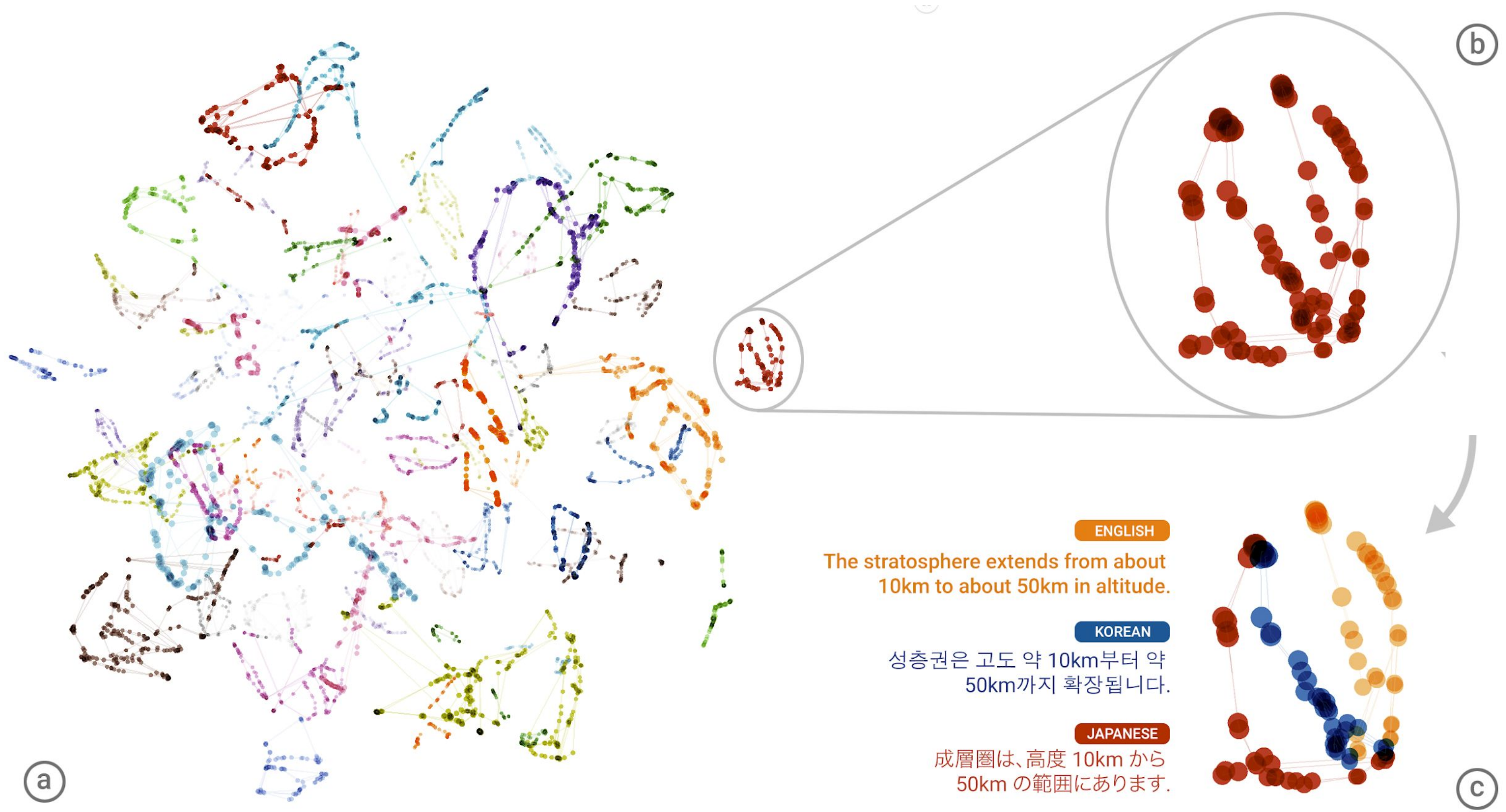
or like this?



● English
● Portuguese

Interlingua?

Sentences with the same meaning mapped to similar regions regardless of language!



Distance between bridge / non-bridge sentences is inversely related to translation quality

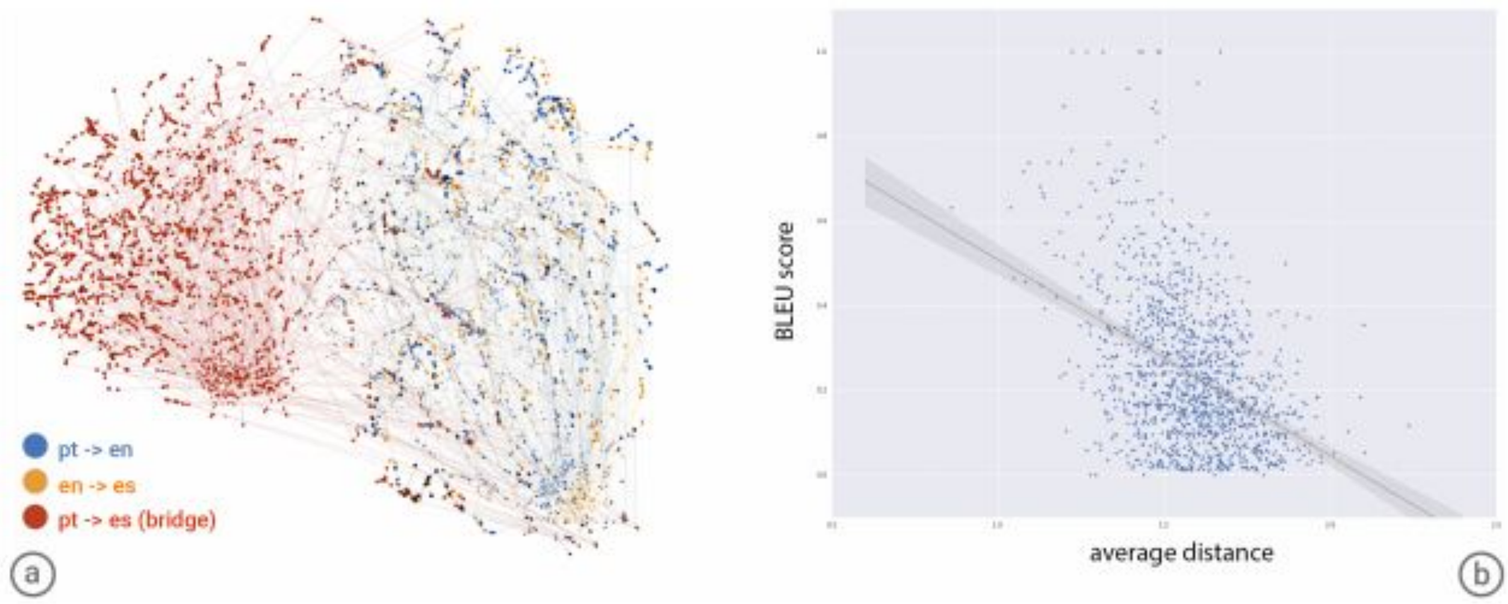


Figure 3: (a) A bird's-eye view of a t-SNE projection of an embedding of the model trained on Portuguese→English (blue) and English→Spanish (yellow) examples with a Portuguese→Spanish zero-shot bridge (red). The large red region on the left primarily contains the zero-shot Portuguese→Spanish translations. (b) A scatter plot of BLEU scores of zero-shot translations versus the average point-wise distance between the zero-shot translation and a non-bridged translation. The Pearson correlation coefficient is −0.42.