Election Day

Communicating data science through visualization

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Due next week

- Midpoint presentation video
  - See template on website under deliverables
  - 10 min 0 sec max.

- Think of this as a draft of your final project presentation but without major results.
  - We expect that you have completed 50% of the project.
  - We would like to see your data and some initial results
  - Provide a complete picture of your project even if certain key parts have not yet been implemented/analyzed/solved.
- We grade based on the quality, as well as the completion of sections described on the next slide.

- Reminder: Now is a good time to start planning for your final report writing as well.
  - Midpoint includes briefly highlighting two similar research papers. Start early!

Acknowledgements

- Covered: Visualization for data science focused research papers, typically 2D in PDF

- Not covered here: Building interactive visualizations, focus on web
  - CSE512 Data Visualization by Jeff Heer
  - Interactive Data Visualization for the Web, 2nd Edition. Scott Murray. Read online for free.

- Slides based on Tutorial by Marinka Zitnik (Harvard University)

What is visualization?

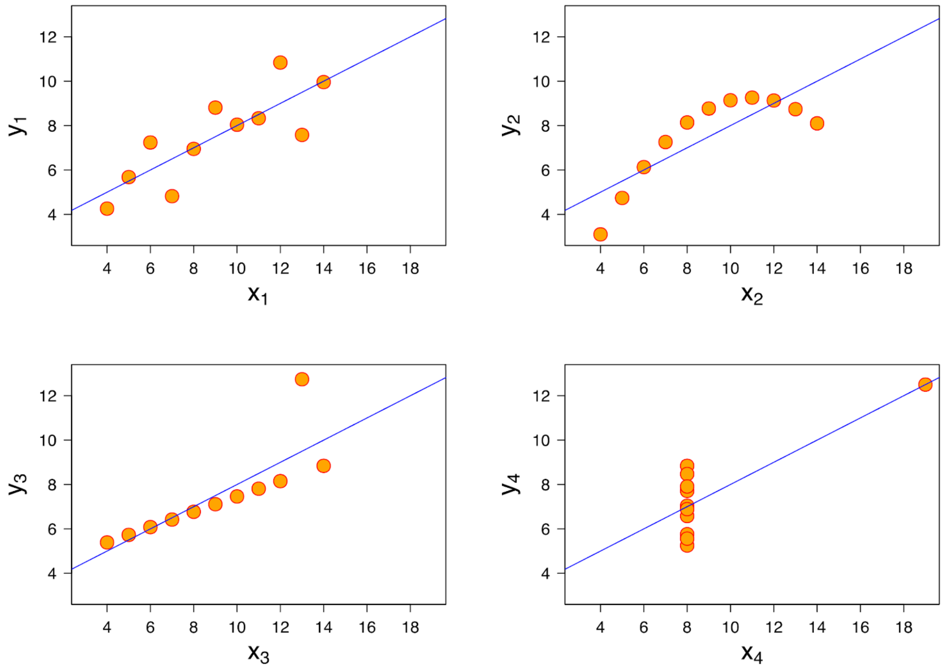
“The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]

“... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]

“Transformation of the symbolic into the geometric” [McCormick et al. 1987]

Anscombe's Quartet

four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



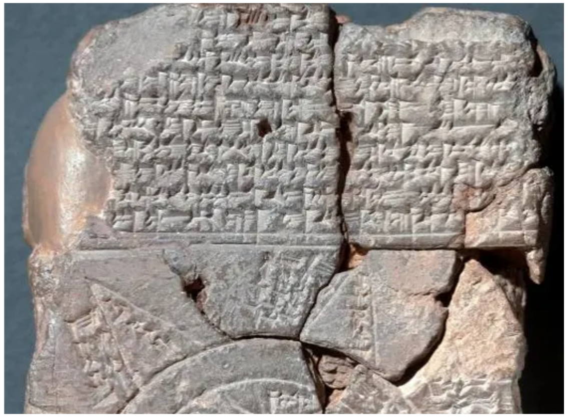
What do you use visualizations for?

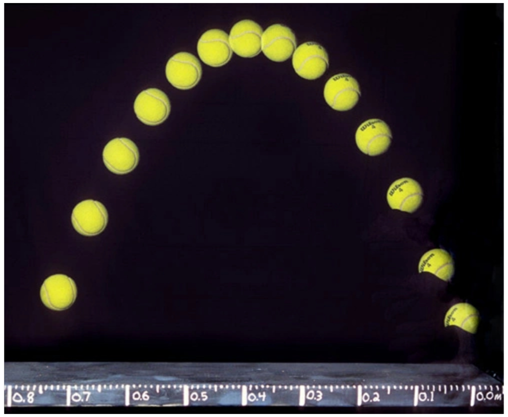
- Understand or make sense of data
- Discover insights
- Answer questions
- Make decisions (high or low stakes)
- Augment or extend our capabilities (memory, estimations, patterns)
- Tell a story
- Convince or inspire

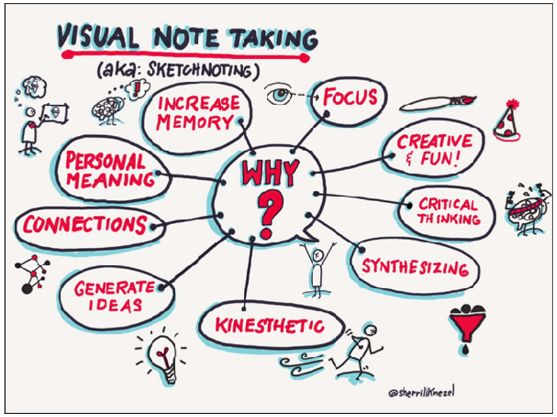
What does visualization do?

Record information

About the world

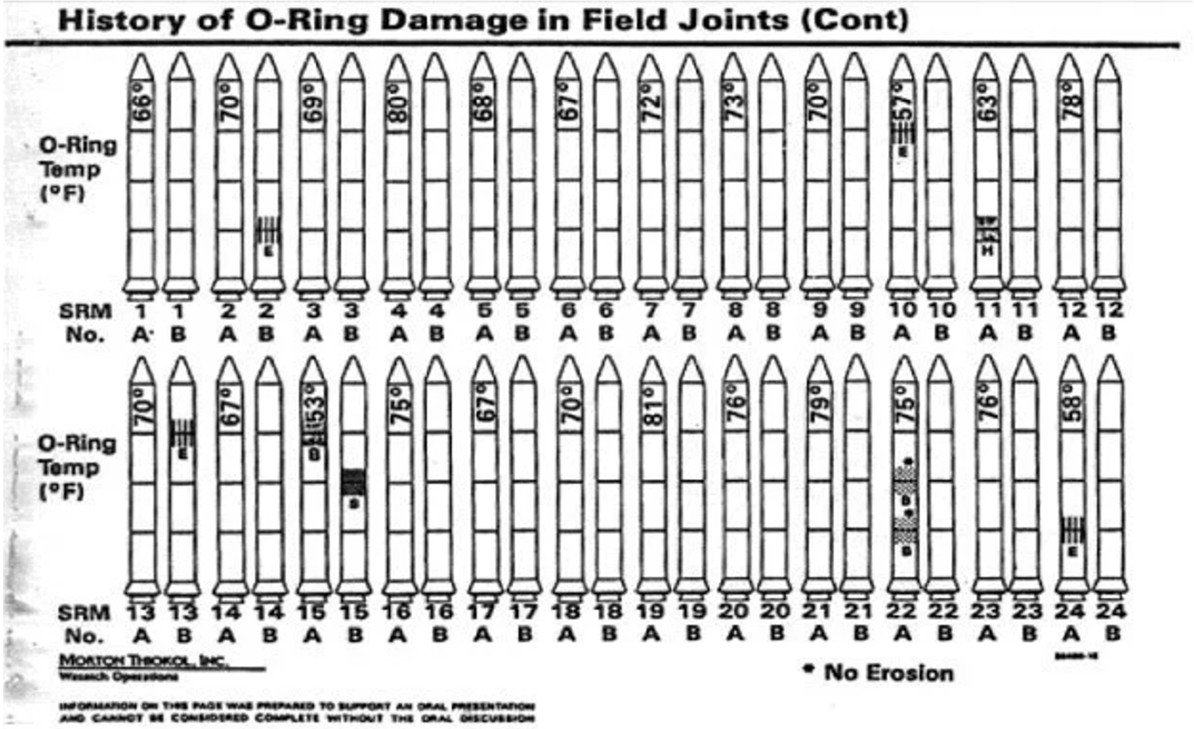






Make sense of information

Make (bad) decisions



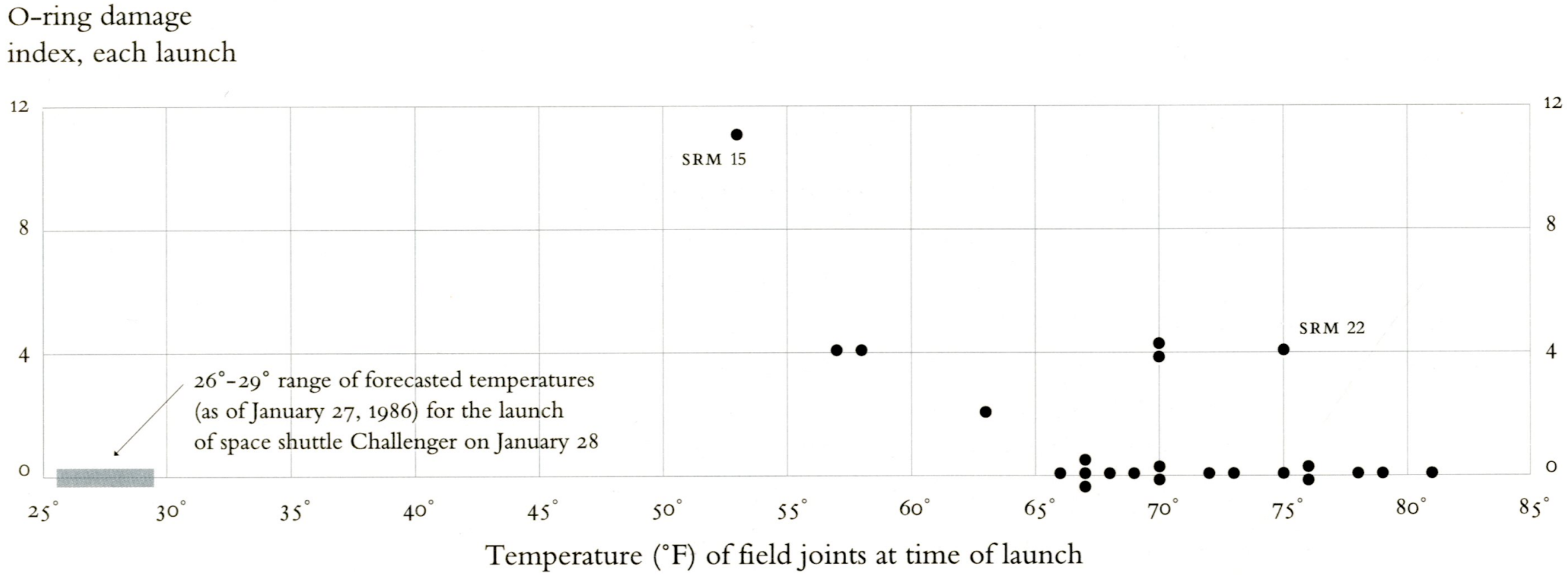
History of O-Ring Damage in Field Joints (Cont)

Morton Thiokol, Inc.
Wasatch Operations

Information on this page was prepared to support an oral presentation and cannot be considered complete without the oral discussion

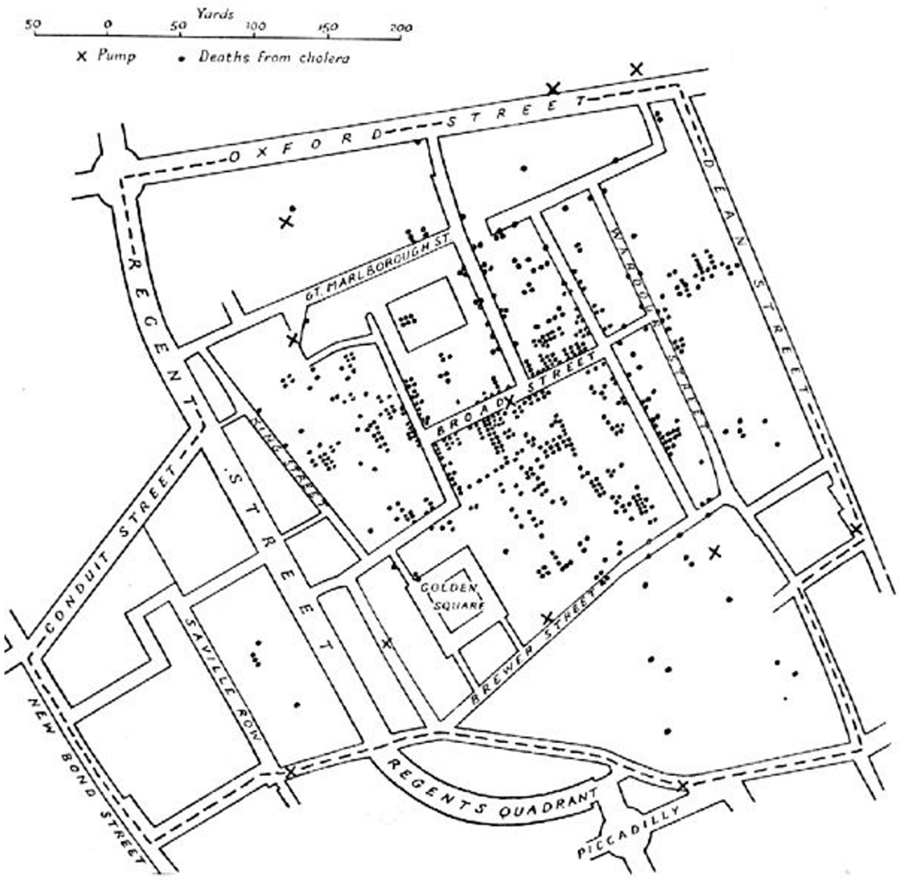
Make sense of information

Make (good) decisions



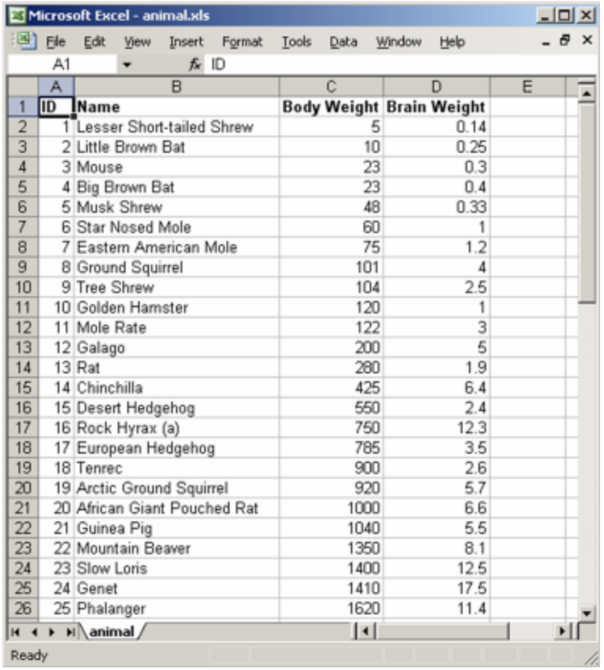
Make sense of information

See patterns

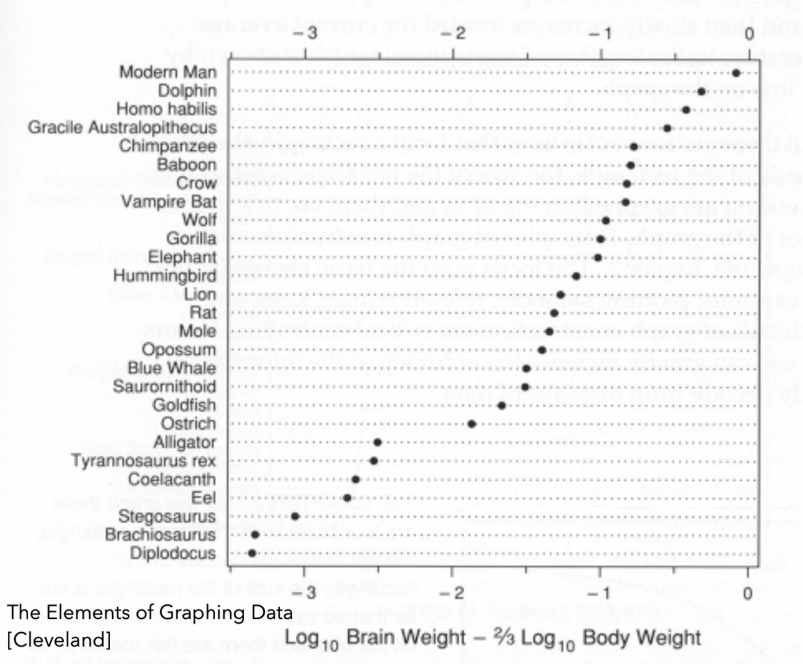


Make sense of information

Answer questions



ID	Name	Body Weight	Brain Weight
1	Lesser Short-tailed Shrew	5	0.14
2	Little Brown Bat	10	0.25
3	Mouse	23	0.3
4	Big Brown Bat	23	0.4
5	Musk Shrew	48	0.33
6	Star Nosed Mole	60	1
7	Eastern American Mole	75	1.2
8	Ground Squirrel	101	4
9	Tree Shrew	104	2.5
10	Golden Hamster	120	1
11	Mole Rate	122	3
12	Galago	200	5
13	Rat	280	1.9
14	Chinchilla	425	6.4
15	Desert Hedgehog	550	2.4
16	Rock Hyrax (a)	750	12.3
17	European Hedgehog	785	3.5
18	Tenrec	900	2.6
19	Arctic Ground Squirrel	920	5.7
20	African Giant Pouched Rat	1000	6.6
21	Guinea Pig	1040	5.5
22	Mountain Beaver	1350	8.1
23	Slow Loris	1400	12.5
24	Genet	1410	17.5
25	Phalanger	1620	11.4

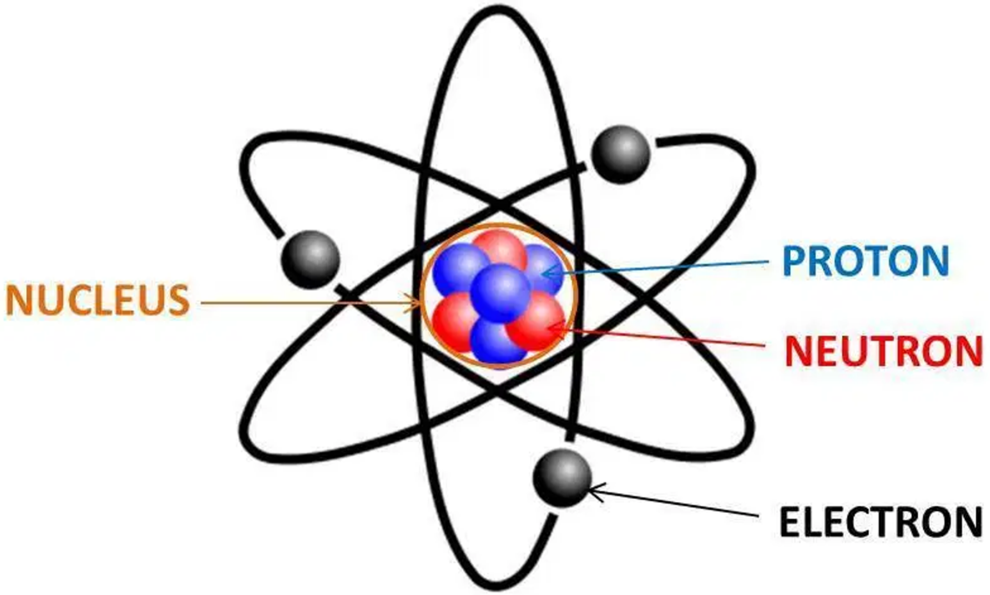


The Elements of Graphing Data
[Cleveland]

Communicate information

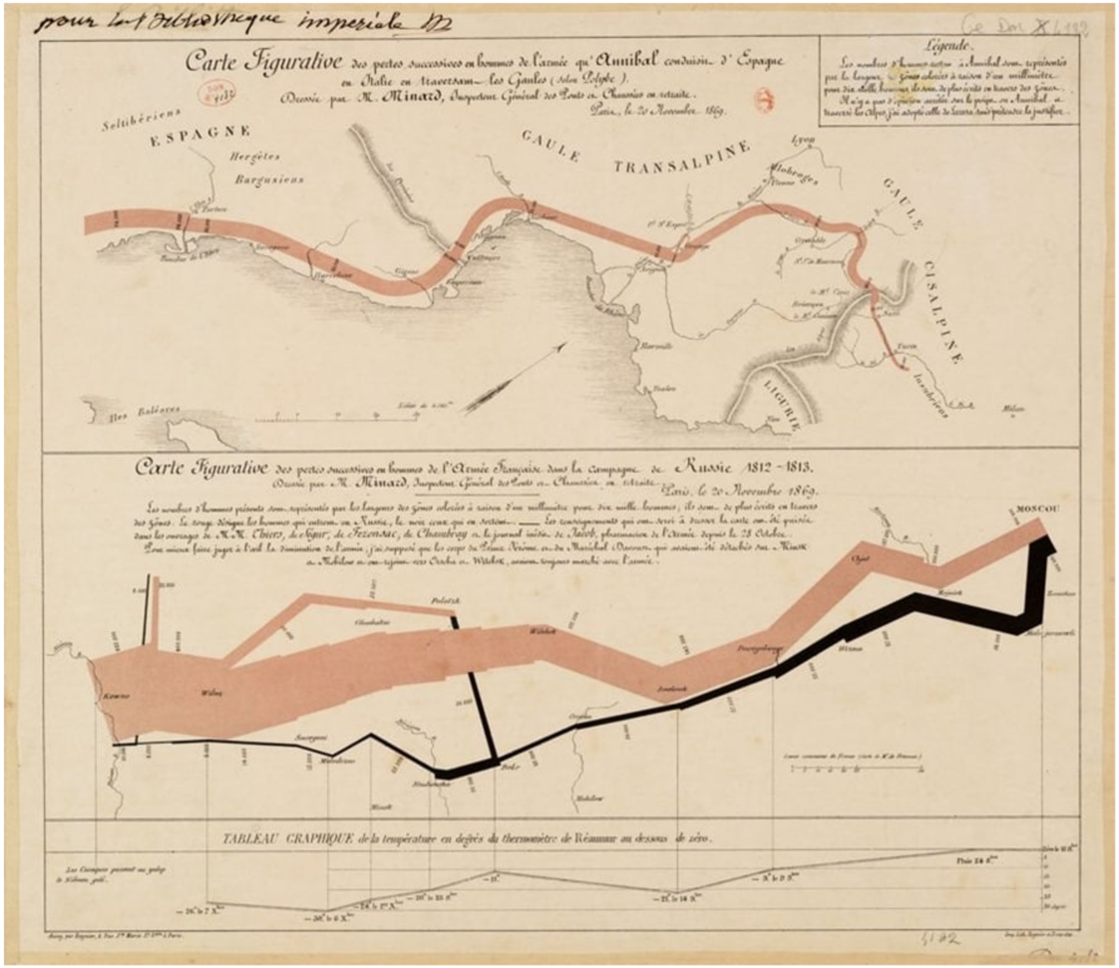
Describe the world

© 2006 Welsch & Partner, Tübingen
scientific multimedia



Communicate information

Prove a point


pour la Bibliothèque impériale
Carte Figurative des pertes successives en hommes de l'armée qu'Annibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe).
Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Légende.
Seltibériens
ESPAGNE
Ilergètes
Barguziens
GAULE TRANSALPINE
Lyon
GAULE CISALPINE
LIGURIE
Iles Baléares
Milan
Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
MOSCOU
Polotzk
Witebsk
Smolensk
Kowno
Wilna
Bobr
Mohilow
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Visualization is a tool

Analysis tool

Communication tool

→ Today we focus primarily on the communication aspect, i.e. “figures for papers”

Today's Lecture

1) Why figures matter
2) Figures in science
3) How to design effective figures
4) Tools, tips, and guidelines

Today’s Lecture

1) Why figures matter ☜

2) Figures in science

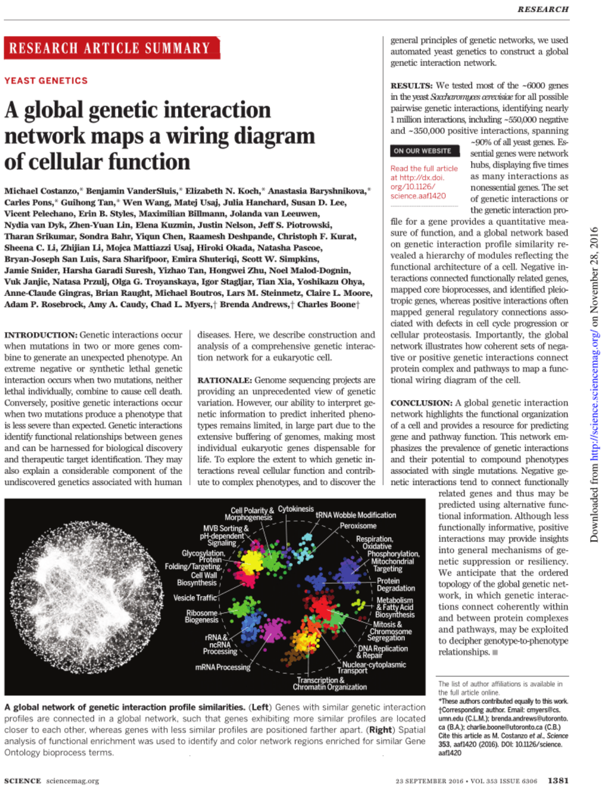
3) How to design effective figures

4) Tools, tips, and guidelines

Why do Figures Matter?

- Figures are often the first part of research papers examined by editors and your peers
- Informative and well-designed figures:
  - Convey facts, ideas, and relationships far more clearly and concisely than text
  - Provide a means for discovering/quantifying patterns, trends, and comparisons
  - Help the audience better understand the objective and results of your research

Design once, reuse many times: Reuse figures from papers for posters, talks, proposals, etc.



RESEARCH

RESEARCH ARTICLE SUMMARY

YEAST GENETICS

A global genetic interaction network maps a wiring diagram of cellular function

Michael Costanzo,* Benjamin VanderSluis,* Elizabeth N. Koch,* Anastasia Baryshnikova,* Carles Pons,* Guihong Tan,* Wen Wang, Matej Usaj, Julia Hanchard, Susan D. Lee, Vicent Pelechano, Erin B. Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S. Piotrowski, Tharan Srikumar, Sondra Bahr, Yiqun Chen, Raamesh Deshpande, Christoph F. Kurat, Sheena C. Li, Zhijian Li, Mojca Mattiazzi Usaj, Hiroki Okada, Natasha Pascoe, Bryan-Joseph San Luis, Sara Sharifpoor, Emira Shuteriqi, Scott W. Simpkins, Jamie Snider, Harsha Garadi Suresh, Yizhao Tan, Hongwei Zhu, Noel Malod-Dognin, Vuk Janjic, Natasa Przulj, Olga G. Troyanskaya, Igor Stagljar, Tian Xia, Yoshikazu Ohya, Anne-Claude Gingras, Brian Raught, Michael Boutros, Lars M. Steinmetz, Claire L. Moore, Adam P. Rosebrock, Amy A. Caudy, Chad L. Myers,† Brenda Andrews,† Charles Boone†

INTRODUCTION: Genetic interactions occur when mutations in two or more genes combine to generate an unexpected phenotype. An extreme negative or synthetic lethal genetic interaction occurs when two mutations, neither lethal individually, combine to cause cell death. Conversely, positive genetic interactions occur when two mutations produce a phenotype that is less severe than expected. Genetic interactions identify functional relationships between genes and can be harnessed for biological discovery and therapeutic target identification. They may also explain a considerable component of the undiscovered genetics associated with human diseases. Here, we describe construction and analysis of a comprehensive genetic interaction network for a eukaryotic cell.

RATIONALE: Genome sequencing projects are providing an unprecedented view of genetic variation. However, our ability to interpret genetic information to predict inherited phenotypes remains limited, in large part due to the extensive buffering of genomes, making most individual eukaryotic genes dispensable for life. To explore the extent to which genetic interactions reveal cellular function and contribute to complex phenotypes, and to discover the general principles of genetic networks, we used automated yeast genetics to construct a global genetic interaction network.

RESULTS: We tested most of the ~6000 genes in the yeast *Saccharomyces cerevisiae* for all possible pairwise genetic interactions, identifying nearly 1 million interactions, including ~550,000 negative and ~350,000 positive interactions, spanning ~90% of all yeast genes. Essential genes were network hubs, displaying five times as many interactions as nonessential genes. The set of genetic interactions or the genetic interaction profile for a gene provides a quantitative measure of function, and a global network based on genetic interaction profile similarity revealed a hierarchy of modules reflecting the functional architecture of a cell. Negative interactions connected functionally related genes, mapped core bioprocesses, and identified pleiotropic genes, whereas positive interactions often mapped general regulatory connections associated with defects in cell cycle progression or cellular proteostasis. Importantly, the global network illustrates how coherent sets of negative or positive genetic interactions connect protein complex and pathways to map a functional wiring diagram of the cell.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaf1420

CONCLUSION: A global genetic interaction network highlights the functional organization of a cell and provides a resource for predicting gene and pathway function. This network emphasizes the prevalence of genetic interactions and their potential to compound phenotypes associated with single mutations. Negative genetic interactions tend to connect functionally related genes and thus may be predicted using alternative functional information. Although less functionally informative, positive interactions may provide insights into general mechanisms of genetic suppression or resiliency. We anticipate that the ordered topology of the global genetic network, in which genetic interactions connect coherently within and between protein complexes and pathways, may be exploited to decipher genotype-to-phenotype relationships. ■

Cell Polarity & Morphogenesis
Cytokinesis
tRNA Wobble Modification
Peroxisome
MVB Sorting & pH-dependent Signaling
Respiration, Oxidative Phosphorylation, Mitochondrial Targeting
Glycosylation, Protein Folding/Targeting, Cell Wall Biosynthesis
Protein Degradation
Vesicle Traffic
Metabolism & Fatty Acid Biosynthesis
Ribosome Biogenesis
Mitosis & Chromosome Segregation
rRNA & ncRNA Processing
DNA Replication & Repair
mRNA Processing
Nuclear-cytoplasmic Transport
Transcription & Chromatin Organization

A global network of genetic interaction profile similarities. (**Left**) Genes with similar genetic interaction profiles are connected in a global network, such that genes exhibiting more similar profiles are located closer to each other, whereas genes with less similar profiles are positioned farther apart. (**Right**) Spatial analysis of functional enrichment was used to identify and color network regions enriched for similar Gene Ontology bioprocess terms.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: cmyers@cs.umn.edu (C.L.M.); brenda.andrews@utoronto.ca (B.A.); charlie.boone@utoronto.ca (C.B.)
Cite this article as M. Costanzo *et al.*, *Science* 353, aaf1420 (2016). DOI: 10.1126/science.aaf1420

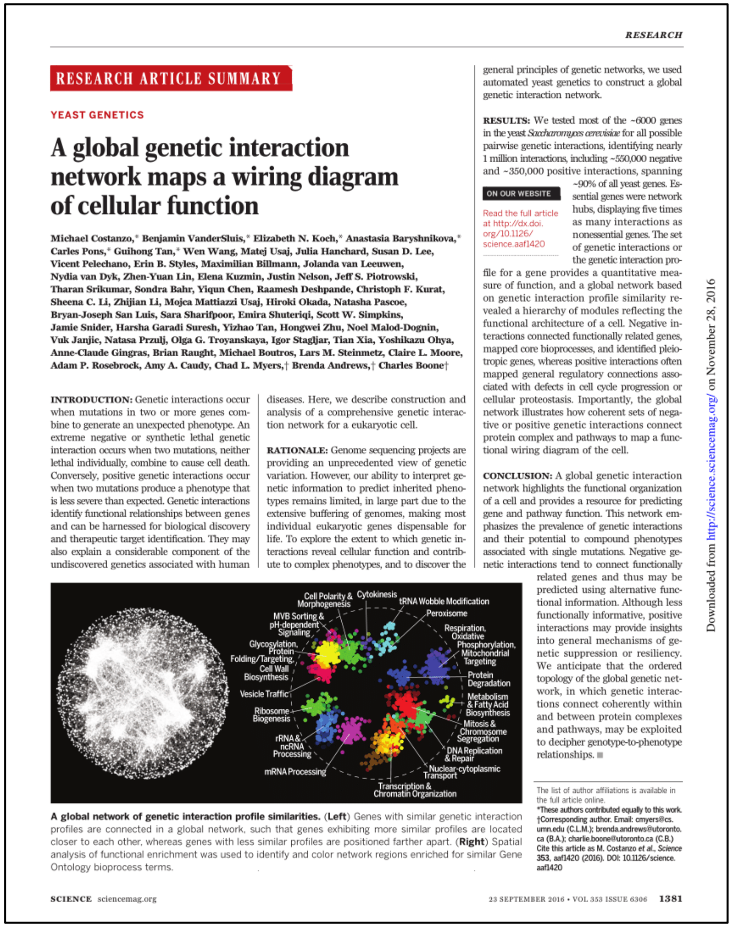
Downloaded from http://science.sciencemag.org/ on November 28, 2016

SCIENCE sciencemag.org

23 SEPTEMBER 2016 • VOL 353 ISSUE 6306 1381

Figure taken from: Costanzo et al. Science 353.6306 (2016).

Promote research ideas and make them accessible to other scientists



RESEARCH

RESEARCH ARTICLE SUMMARY

YEAST GENETICS

A global genetic interaction network maps a wiring diagram of cellular function

Michael Costanzo,* Benjamin VanderSluis,* Elizabeth N. Koch,* Anastasia Baryshnikova,* Carles Pons,* Guihong Tan,* Wen Wang, Matej Usaj, Julia Hanchard, Susan D. Lee, Vicent Pelechano, Erin B. Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S. Piotrowski, Tharan Srikumar, Sondra Bahr, Yiqun Chen, Raamesh Deshpande, Christoph F. Kurat, Sheena C. Li, Zhijian Li, Mojca Mattiazzi Usaj, Hiroki Okada, Natasha Pascoe, Bryan-Joseph San Luis, Sara Sharifpoor, Emira Shuteriqi, Scott W. Simpkins, Jamie Snider, Harsha Garadi Suresh, Yizhao Tan, Hongwei Zhu, Noel Malod-Dognin, Vuk Janjic, Natasa Przulj, Olga G. Troyanskaya, Igor Stagljar, Tian Xia, Yoshikazu Ohya, Anne-Claude Gingras, Brian Raught, Michael Boutros, Lars M. Steinmetz, Claire L. Moore, Adam P. Rosebrock, Amy A. Caudy, Chad L. Myers,† Brenda Andrews,† Charles Boone†

INTRODUCTION: Genetic interactions occur when mutations in two or more genes combine to generate an unexpected phenotype. An extreme negative or synthetic lethal genetic interaction occurs when two mutations, neither lethal individually, combine to cause cell death. Conversely, positive genetic interactions occur when two mutations produce a phenotype that is less severe than expected. Genetic interactions identify functional relationships between genes and can be harnessed for biological discovery and therapeutic target identification. They may also explain a considerable component of the undiscovered genetics associated with human diseases. Here, we describe construction and analysis of a comprehensive genetic interaction network for a eukaryotic cell.

RATIONALE: Genome sequencing projects are providing an unprecedented view of genetic variation. However, our ability to interpret genetic information to predict inherited phenotypes remains limited, in large part due to the extensive buffering of genomes, making most individual eukaryotic genes dispensable for life. To explore the extent to which genetic interactions reveal cellular function and contribute to complex phenotypes, and to discover the general principles of genetic networks, we used automated yeast genetics to construct a global genetic interaction network.

RESULTS: We tested most of the ~6000 genes in the yeast *Saccharomyces cerevisiae* for all possible pairwise genetic interactions, identifying nearly 1 million interactions, including ~550,000 negative and ~350,000 positive interactions, spanning ~90% of all yeast genes. Essential genes were network hubs, displaying five times as many interactions as nonessential genes. The set of genetic interactions or the genetic interaction profile for a gene provides a quantitative measure of function, and a global network based on genetic interaction profile similarity revealed a hierarchy of modules reflecting the functional architecture of a cell. Negative interactions connected functionally related genes, mapped core bioprocesses, and identified pleiotropic genes, whereas positive interactions often mapped general regulatory connections associated with defects in cell cycle progression or cellular proteostasis. Importantly, the global network illustrates how coherent sets of negative or positive genetic interactions connect protein complex and pathways to map a functional wiring diagram of the cell.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaf1420

CONCLUSION: A global genetic interaction network highlights the functional organization of a cell and provides a resource for predicting gene and pathway function. This network emphasizes the prevalence of genetic interactions and their potential to compound phenotypes associated with single mutations. Negative genetic interactions tend to connect functionally related genes and thus may be predicted using alternative functional information. Although less functionally informative, positive interactions may provide insights into general mechanisms of genetic suppression or resiliency. We anticipate that the ordered topology of the global genetic network, in which genetic interactions connect coherently within and between protein complexes and pathways, may be exploited to decipher genotype-to-phenotype relationships. ■

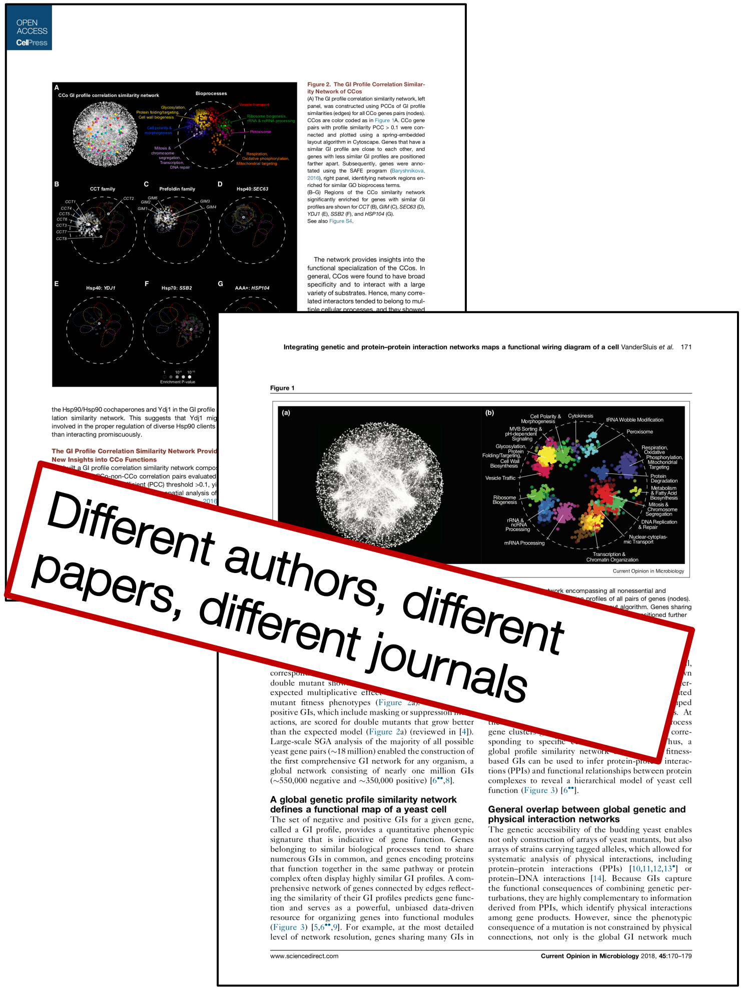
Cell Polarity & Morphogenesis
Cytokinesis
tRNA Wobble Modification
Peroxisome
MVB Sorting & pH-dependent Signaling
Glycosylation, Protein Folding/Targeting, Cell Wall Biosynthesis
Respiration, Oxidative Phosphorylation, Mitochondrial Targeting
Protein Degradation
Vesicle Traffic
Metabolism & Fatty Acid Biosynthesis
Ribosome Biogenesis
Mitosis & Chromosome Segregation
rRNA & ncRNA Processing
DNA Replication & Repair
mRNA Processing
Nuclear-cytoplasmic Transport
Transcription & Chromatin Organization

A global network of genetic interaction profile similarities. (**Left**) Genes with similar genetic interaction profiles are connected in a global network, such that genes exhibiting more similar profiles are located closer to each other, whereas genes with less similar profiles are positioned farther apart. (**Right**) Spatial analysis of functional enrichment was used to identify and color network regions enriched for similar Gene Ontology bioprocess terms.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: cmyers@cs.umn.edu (C.L.M.); brenda.andrews@utoronto.ca (B.A.); charlie.boone@utoronto.ca (C.B.)
Cite this article as M. Costanzo *et al*., *Science* 353, aaf1420 (2016). DOI: 10.1126/science.aaf1420

Downloaded from http://science.sciencemag.org/ on November 28, 2016

SCIENCE sciencemag.org 23 SEPTEMBER 2016 • VOL 353 ISSUE 6306 1381



OPEN ACCESS
CellPress

A
CCo GI profile correlation similarity network
Bioprocesses
Glycosylation, Protein folding/targeting, Cell wall biogenesis
Vesicle transport
Ribosome biogenesis, rRNA & ncRNA processing
Cell polarity & morphogenesis
Peroxisome
Mitosis & chromosome segregation, Transcription, DNA repair
Respiration, Oxidative phosphorylation, Mitochondrial targeting
B
CCT family
C
Prefoldin family
D
Hsp40: SEC63
E
Hsp40: YDJ1
F
Hsp70: SSB2
G
AAA+: HSP104
Enrichment P-value

Figure 2. The GI Profile Correlation Similarity Network of CCos

(A) The GI profile correlation similarity network, left panel, was constructed using PCCs of GI profile similarities (edges) for all CCo genes pairs (nodes). CCos are color coded as in Figure 1A. CCo gene pairs with profile similarity PCC > 0.1 were connected and plotted using a spring-embedded layout algorithm in Cytoscape. Genes that have a similar GI profile are close to each other, and genes with less similar GI profiles are positioned farther apart. Subsequently, genes were annotated using the SAFE program (Baryshnikova, 2016), right panel, identifying network regions enriched for similar GO bioprocess terms.

(B–G) Regions of the CCo similarity network significantly enriched for genes with similar GI profiles are shown for *CCT* (B), *GIM* (C), *SEC63* (D), *YDJ1* (E), *SSB2* (F), and *HSP104* (G).

See also Figure S4.

The network provides insights into the functional specialization of the CCos. In general, CCos were found to have broad specificity and to interact with a large variety of substrates. Hence, many correlated interactors tended to belong to multiple cellular processes, and they showed

the Hsp90/Hsp90 cochaperones and Ydj1 in the GI profile lation similarity network. This suggests that Ydj1 mig involved in the proper regulation of diverse Hsp90 clients than interacting promiscuously.

The GI Profile Correlation Similarity Network Provid New Insights into CCo Functions

built a GI profile correlation similarity network compo CCo-non-CCo correlation pairs evaluated ficient (PCC) threshold >0.1, yi patial analysis of 2016

Integrating genetic and protein–protein interaction networks maps a functional wiring diagram of a cell VanderSluis *et al.* 171

Figure 1

(a)
(b)
Cell Polarity & Morphogenesis
Cytokinesis
tRNA Wobble Modification
Peroxisome
MVB Sorting & pH-dependent Signaling
Glycosylation, Protein Folding/Targeting, Cell Wall Biosynthesis
Respiration, Oxidative Phosphorylation, Mitochondrial Targeting
Protein Degradation
Vesicle Traffic
Metabolism & Fatty Acid Biosynthesis
Ribosome Biogenesis
Mitosis & Chromosome Segregation
rRNA & ncRNA Processing
DNA Replication & Repair
mRNA Processing
Nuclear-cytoplasmic Transport
Transcription & Chromatin Organization

Current Opinion in Microbiology

twork encompassing all nonessential and profiles of all pairs of genes (nodes). t algorithm. Genes sharing sitioned further

correspo double mutant show expected multiplicative effec mutant fitness phenotypes (Figure 2a). positive GIs, which include masking or suppression actions, are scored for double mutants that grow better than the expected model (Figure 2a) (reviewed in [4]). Large-scale SGA analysis of the majority of all possible yeast gene pairs (~18 million) enabled the construction of the first comprehensive GI network for any organism, a global network consisting of nearly one million GIs (~550,000 negative and ~350,000 positive) [6••,8].

A global genetic profile similarity network defines a functional map of a yeast cell

The set of negative and positive GIs for a given gene, called a GI profile, provides a quantitative phenotypic signature that is indicative of gene function. Genes belonging to similar biological processes tend to share numerous GIs in common, and genes encoding proteins that function together in the same pathway or protein complex often display highly similar GI profiles. A comprehensive network of genes connected by edges reflecting the similarity of their GI profiles predicts gene function and serves as a powerful, unbiased data-driven resource for organizing genes into functional modules (Figure 3) [5,6••,9]. For example, at the most detailed level of network resolution, genes sharing many GIs in

l, wn er- ted ped s. At rocess gene clusters corre- sponding to specific Thus, a global profile similarity network fitness- based GIs can be used to infer protein-pro interactions (PPIs) and functional relationships between protein complexes to reveal a hierarchical model of yeast cell function (Figure 3) [6••].

General overlap between global genetic and physical interaction networks

The genetic accessibility of the budding yeast enables not only construction of arrays of yeast mutants, but also arrays of strains carrying tagged alleles, which allowed for systematic analysis of physical interactions, including protein–protein interactions (PPIs) [10,11,12,13•] or protein–DNA interactions [14]. Because GIs capture the functional consequences of combining genetic perturbations, they are highly complementary to information derived from PPIs, which identify physical interactions among gene products. However, since the phenotypic consequence of a mutation is not constrained by physical connections, not only is the global GI network much

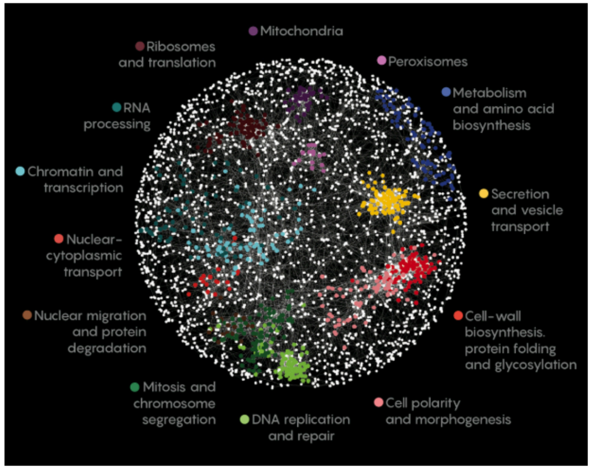
www.sciencedirect.com Current Opinion in Microbiology 2018, 45:170–179

Different authors, different papers, different journals

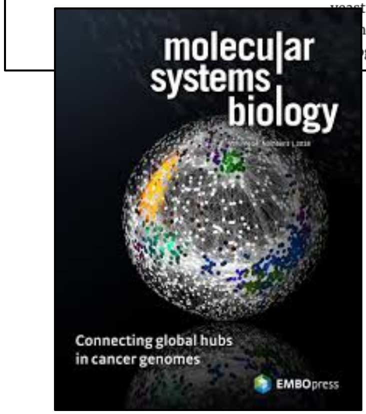
Promote your research among general audience and media

Quanta magazine

Physics Mathematics Biology Computer Science All Articles


Mitochondria
Ribosomes and translation
Peroxisomes
RNA processing
Metabolism and amino acid biosynthesis
Chromatin and transcription
Secretion and vesicle transport
Nuclear-cytoplasmic transport
Nuclear migration and protein degradation
Cell-wall biosynthesis, protein folding and glycosylation
Mitosis and chromosome segregation
DNA replication and repair
Cell polarity and morphogenesis

This figure maps the interactions among various genes (represented as dots) in the yeast genome. Genes with linked effects are connected by lines; genes with more ngly correlated effects are closer together. The color of the dots corresponds to the gical processes and organelles in which the genes are involved.



molecular systems biology

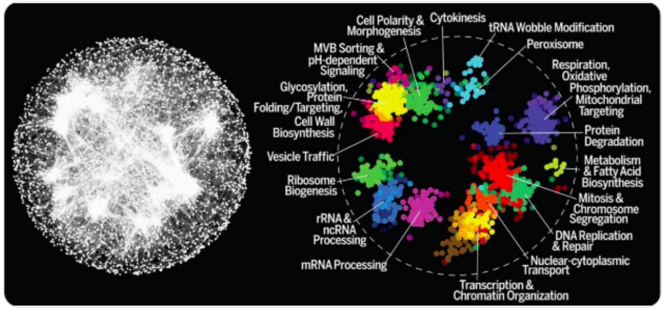
Connecting global hubs in cancer genomes

EMBOpress

DRYAD

Dryad @datadryad · 26 Dec 2016

Featured data from @sciencemagazine: **A global genetic interaction network maps a wiring diagram of cellular function** dx.doi.org/10.5061/dryad.…

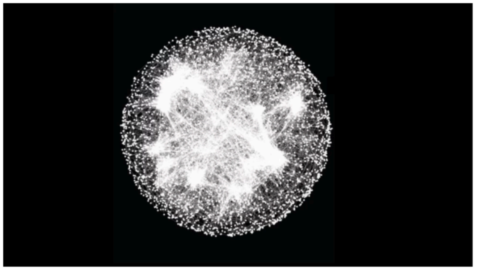

Cell Polarity & Morphogenesis
Cytokinesis
tRNA Wobble Modification
Peroxisome
MVB Sorting & pH-dependent Signaling
Respiration, Oxidative Phosphorylation, Mitochondrial Targeting
Glycosylation, Protein Folding/Targeting, Cell Wall Biosynthesis
Protein Degradation
Vesicle Traffic
Metabolism & Fatty Acid Biosynthesis
Ribosome Biogenesis
Mitosis & Chromosome Segregation
rRNA & ncRNA Processing
DNA Replication & Repair
mRNA Processing
Nuclear-cytoplasmic Transport
Transcription & Chromatin Organization

MOTHERBOARD
TECH BY VICE

It Took 15 Years to Map Every Gene Interaction in a Yeast Cell

Understanding how thousands of individual yeast genes interact in pairs could expose the underlying genetic bases of human diseases.

SHARE TWEET



Effective figures improve your papers

↓

Maximize impact, boost citation count, stand out among your peers

Today's Lecture

1) Why figures matter ✓
2) Figures in science ☚
3) How to design effective figures
4) Tools, tips, and guidelines

Two Types of Papers with Different Visual Structure

1) Core CS conference papers:
   - KDD, WebConf, NeurIPS, ICML, ICLR, AAAI, etc.

2) Interdisciplinary journal papers:
   - Nature, Science, PNAS, etc.

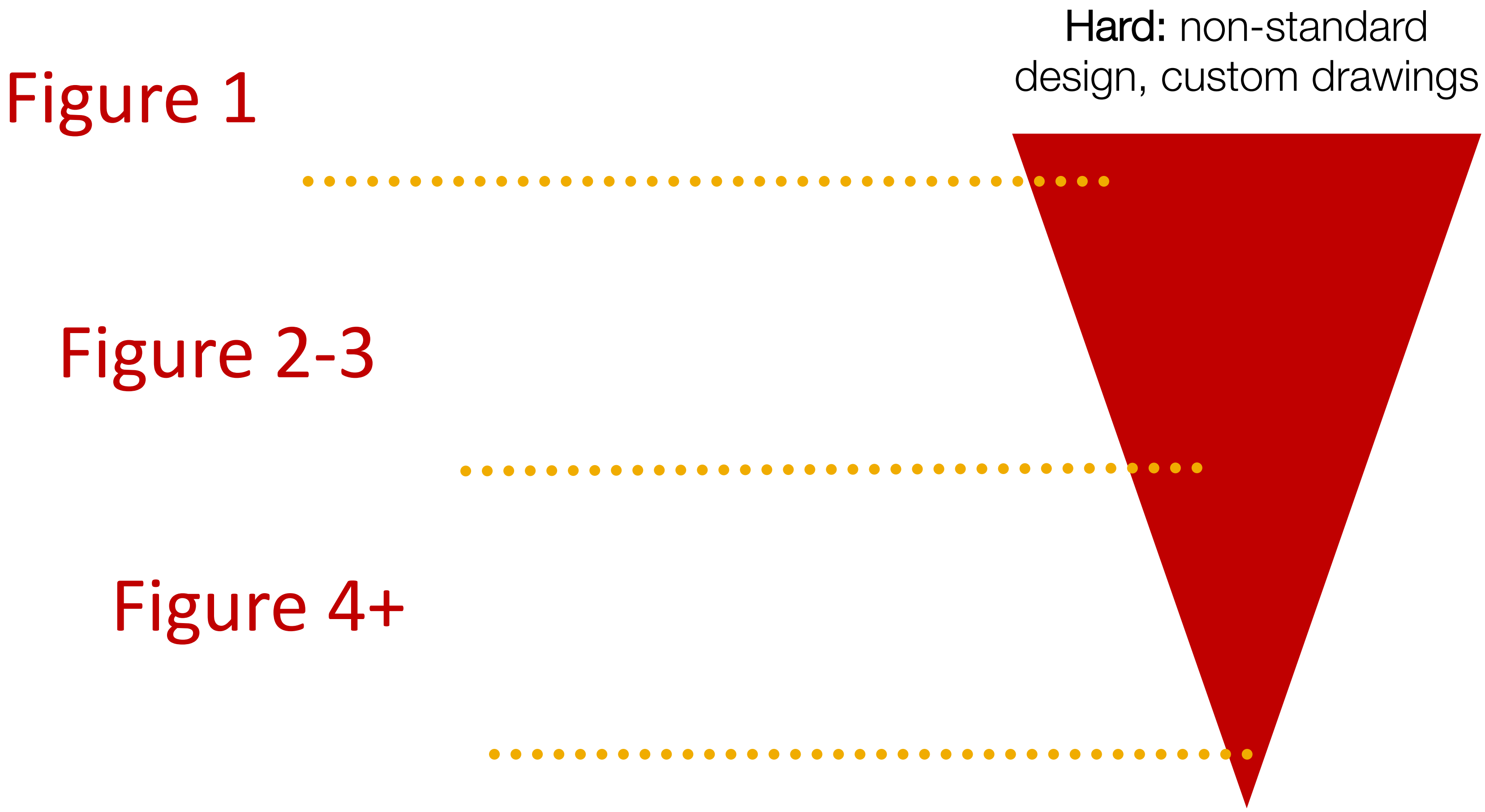
Core CS Conference Papers

The focus is on the development of new methods and their evaluation and comparison on benchmark datasets

Core CS Conference Papers: Visual Structure

- Figure 1: Key methodological contribution
  - Focus on most important information
  - **Impress your audience!**
    - Is your method/system the fastest, the largest, the most accurate?
    - What is the hard problem that your method solves?
    - What makes your method different from related work?
- Figure 2-3: Overview and algorithmic details
  - Inputs + Data transformation + Outputs
  - Show details about data transformations:
    - Graph convolutions, neural architectures, etc.
- Figure 4+: Results

Core CS Conference Papers: Visual Structure



Hard: non-standard design, custom drawings

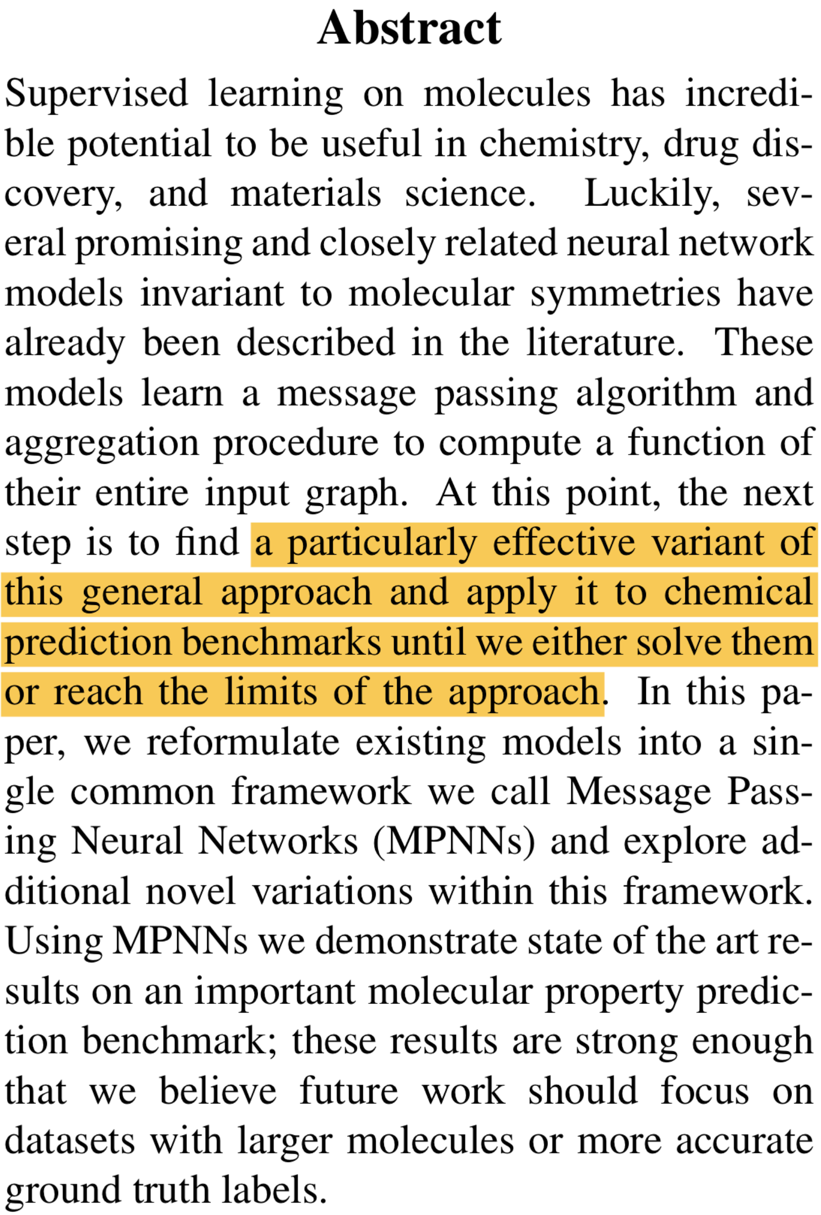
Figure 1

Figure 2-3

Figure 4+

Easy: standard design, visualization libraries like Matplotlib and Seaborn

Examples:
Core CS Conference Papers



Abstract

Supervised learning on molecules has incredible potential to be useful in chemistry, drug discovery, and materials science. Luckily, several promising and closely related neural network models invariant to molecular symmetries have already been described in the literature. These models learn a message passing algorithm and aggregation procedure to compute a function of their entire input graph. At this point, the next step is to find a particularly effective variant of this general approach and apply it to chemical prediction benchmarks until we either solve them or reach the limits of the approach. In this paper, we reformulate existing models into a single common framework we call Message Passing Neural Networks (MPNNs) and explore additional novel variations within this framework. Using MPNNs we demonstrate state of the art results on an important molecular property prediction benchmark; these results are strong enough that we believe future work should focus on datasets with larger molecules or more accurate ground truth labels.

Impress your audience! 😲

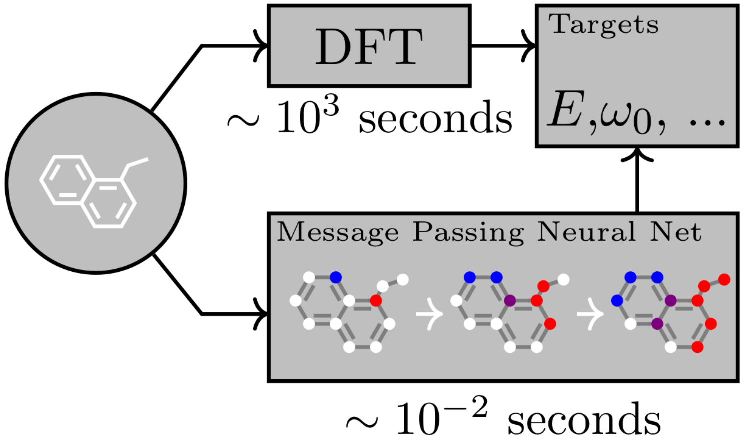
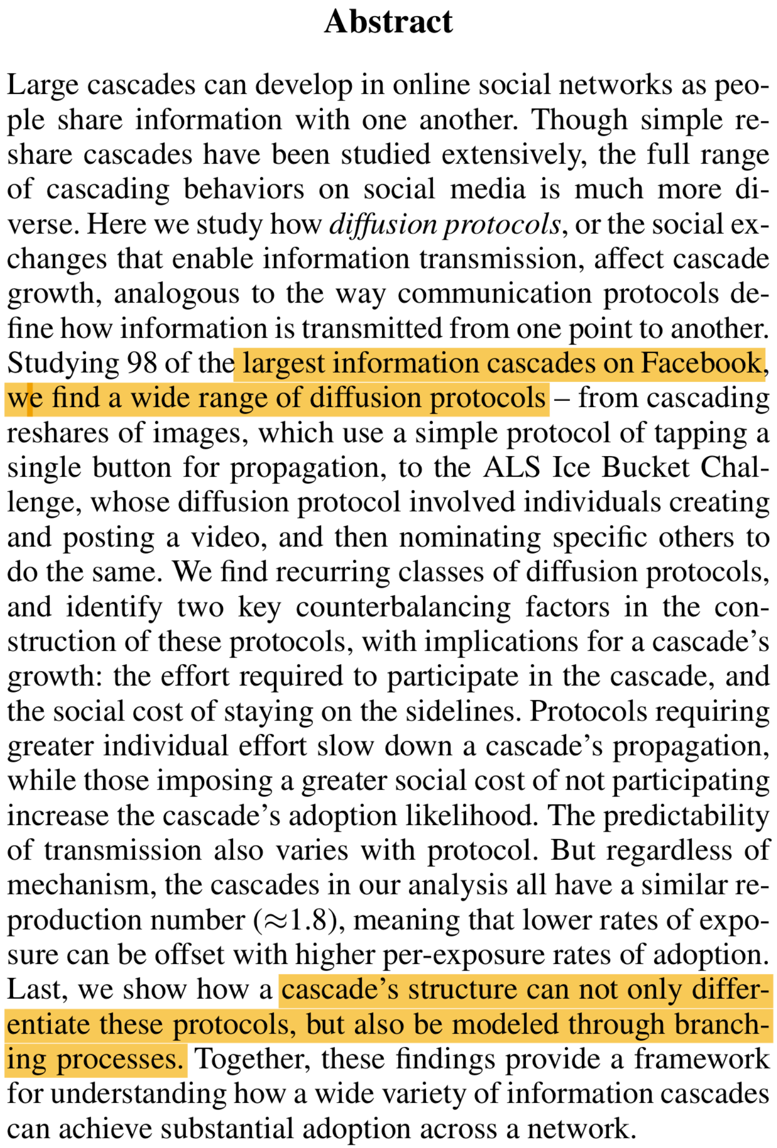


Figure 1. A Message Passin l Network predicts quantum properties of an organi y modeling a computationally expensive

Focus on key information: "Our method is so fast! Our paper should be published at ICML!"

Gilmer et al., Neural Message Passing for Quantum Chemistry, ICML, 2017.



Abstract

Large cascades can develop in online social networks as people share information with one another. Though simple reshare cascades have been studied extensively, the full range of cascading behaviors on social media is much more diverse. Here we study how *diffusion protocols*, or the social exchanges that enable information transmission, affect cascade growth, analogous to the way communication protocols define how information is transmitted from one point to another. Studying 98 of the largest information cascades on Facebook, we find a wide range of diffusion protocols – from cascading reshares of images, which use a simple protocol of tapping a single button for propagation, to the ALS Ice Bucket Challenge, whose diffusion protocol involved individuals creating and posting a video, and then nominating specific others to do the same. We find recurring classes of diffusion protocols, and identify two key counterbalancing factors in the construction of these protocols, with implications for a cascade's growth: the effort required to participate in the cascade, and the social cost of staying on the sidelines. Protocols requiring greater individual effort slow down a cascade's propagation, while those imposing a greater social cost of not participating increase the cascade's adoption likelihood. The predictability of transmission also varies with protocol. But regardless of mechanism, the cascades in our analysis all have a similar reproduction number (≈1.8), meaning that lower rates of exposure can be offset with higher per-exposure rates of adoption. Last, we show how a cascade's structure can not only differentiate these protocols, but also be modeled through branching processes. Together, these findings provide a framework for understanding how a wide variety of information cascades can achieve substantial adoption across a network.

Impress your audience! 😲

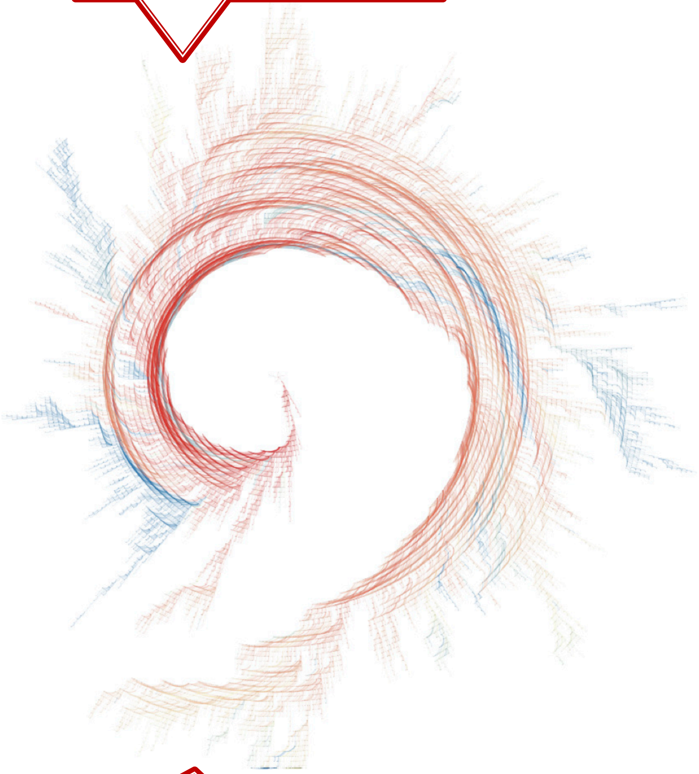
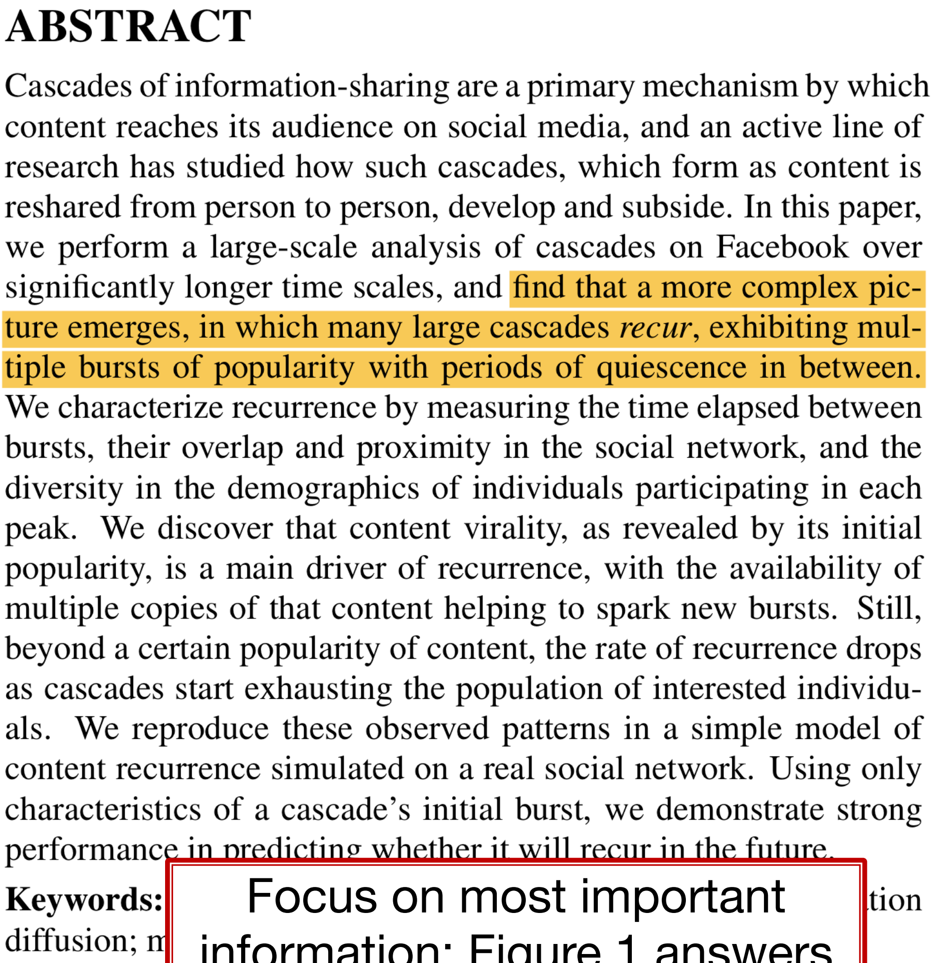


Figure 1: The d n tree of a cascade with a volunteer diffusion p individuals posted music from an

"Cascades can be so large! Despite that, we know how to study them! Our paper should be published at ICWSM!"

Cheng et al., Do Diffusion Protocols Govern Cascade Growth?, ICWSM, 2018.



ABSTRACT

Cascades of information-sharing are a primary mechanism by which content reaches its audience on social media, and an active line of research has studied how such cascades, which form as content is reshared from person to person, develop and subside. In this paper, we perform a large-scale analysis of cascades on Facebook over significantly longer time scales, and find that a more complex picture emerges, in which many large cascades *recur*, exhibiting multiple bursts of popularity with periods of quiescence in between. We characterize recurrence by measuring the time elapsed between bursts, their overlap and proximity in the social network, and the diversity in the demographics of individuals participating in each peak. We discover that content virality, as revealed by its initial popularity, is a main driver of recurrence, with the availability of multiple copies of that content helping to spark new bursts. Still, beyond a certain popularity of content, the rate of recurrence drops as cascades start exhausting the population of interested individuals. We reproduce these observed patterns in a simple model of content recurrence simulated on a real social network. Using only characteristics of a cascade's initial burst, we demonstrate strong performance in predicting whether it will recur in the future.

Keywords: tion diffusion; m

Focus on most important information: Figure 1 answers question asked by the title

Impress your audience! 😲

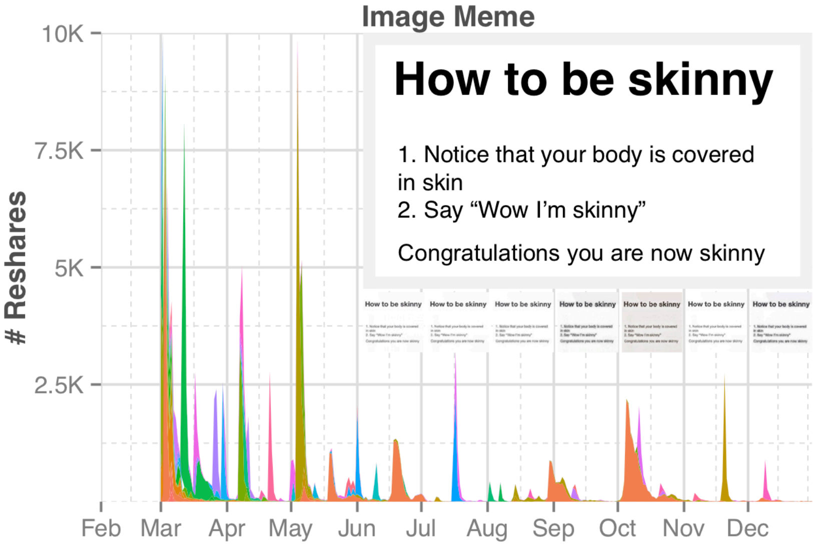
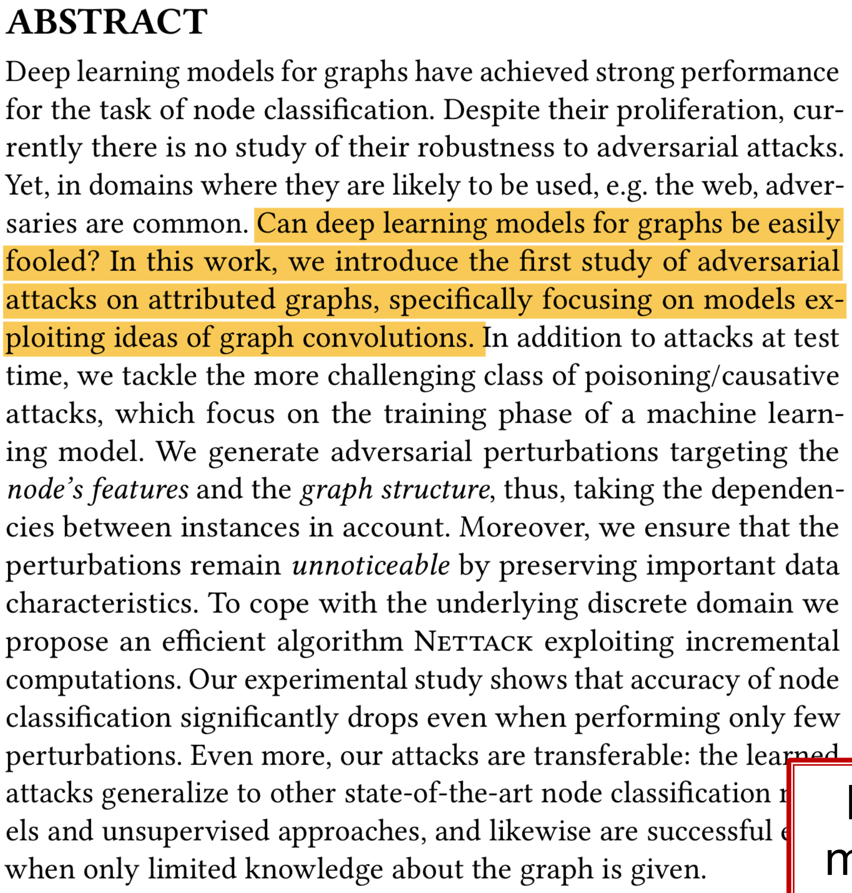


Figure 1: An example of a image meme that has recurred, or resurfaced in popularity multiple times, sometimes as a continuation of the same copy, and sometimes as a new copy of the same meme (example copies are shown as thumbnails). This recurrence appears as multiple peaks in the plot of reshares as a function of time.

"Cascades can be so complex! Despite that, we know how to study them! Our paper should be published at WWW!"

Cheng et al., Do Cascades Recur?, WWW, 2016.



ABSTRACT

Deep learning models for graphs have achieved strong performance for the task of node classification. Despite their proliferation, currently there is no study of their robustness to adversarial attacks. Yet, in domains where they are likely to be used, e.g. the web, adversaries are common. Can deep learning models for graphs be easily fooled? In this work, we introduce the first study of adversarial attacks on attributed graphs, specifically focusing on models exploiting ideas of graph convolutions. In addition to attacks at test time, we tackle the more challenging class of poisoning/causative attacks, which focus on the training phase of a machine learning model. We generate adversarial perturbations targeting the *node's features* and the *graph structure*, thus, taking the dependencies between instances in account. Moreover, we ensure that the perturbations remain *unnoticeable* by preserving important data characteristics. To cope with the underlying discrete domain we propose an efficient algorithm Nettack exploiting incremental computations. Our experimental study shows that accuracy of node classification significantly drops even when performing only few perturbations. Even more, our attacks are transferable: the learned attacks generalize to other state-of-the-art node classification m els and unsupervised approaches, and likewise are successful e when only limited knowledge about the graph is given.

Impress your audience! 😲

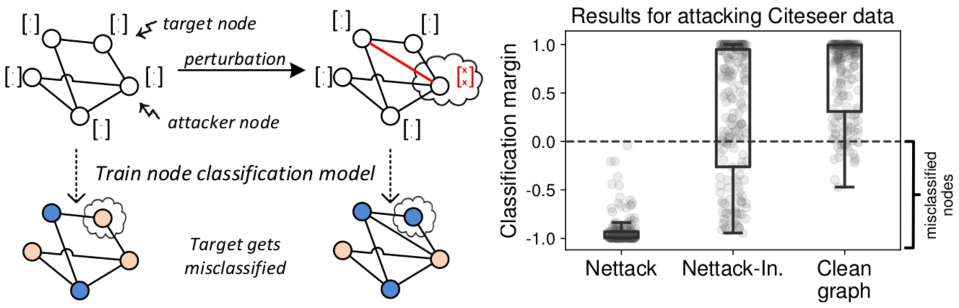


Figure 1: Small perturbations of the graph structure and node features lead to misclassification of the target.

Focus on key information: Yes, graph-based models for deep learning can be easily fooled. Here we show how devastating attacks can be.

Zugner et al., Adversarial Attacks on Neural Networks for Graph Data, KDD, 2018. *(Best paper award)*

Interdisciplinary Journal Papers

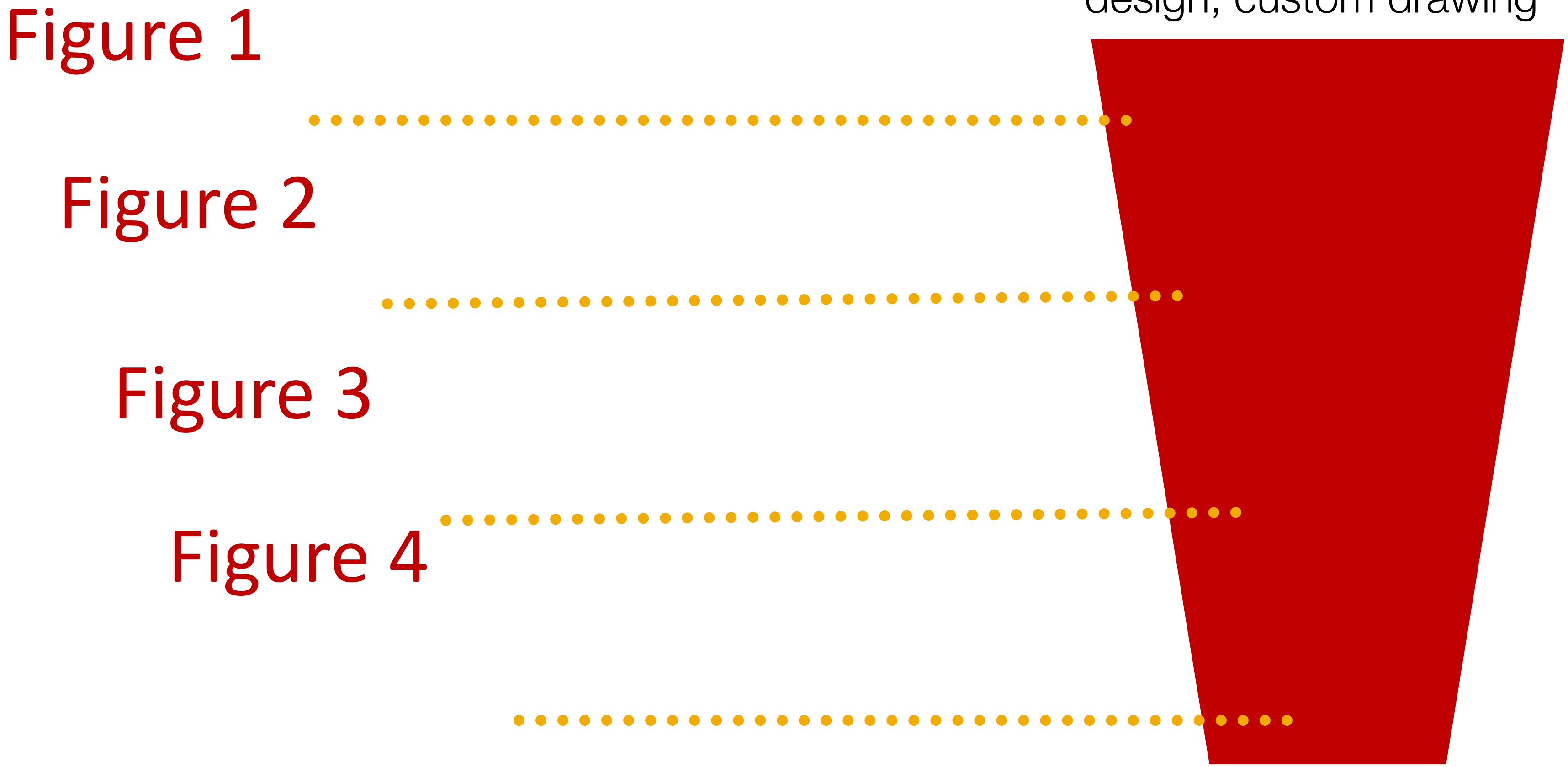
The focus is on new scientific insights and demonstrating the importance of those insights to advance science

Interdisciplinary Journal Papers: Visual Structure

- Figure 1: Dataset, approach and key result
  - **Impress your audience!**
- Figure 2: Key result, detailed and unpacked
- Figure 3: Orthogonal evidence supporting results
- Figure 4: Orthogonal evidence supporting results
- Supplementary Figures: Methodological contributions, algorithms, robustness analyses

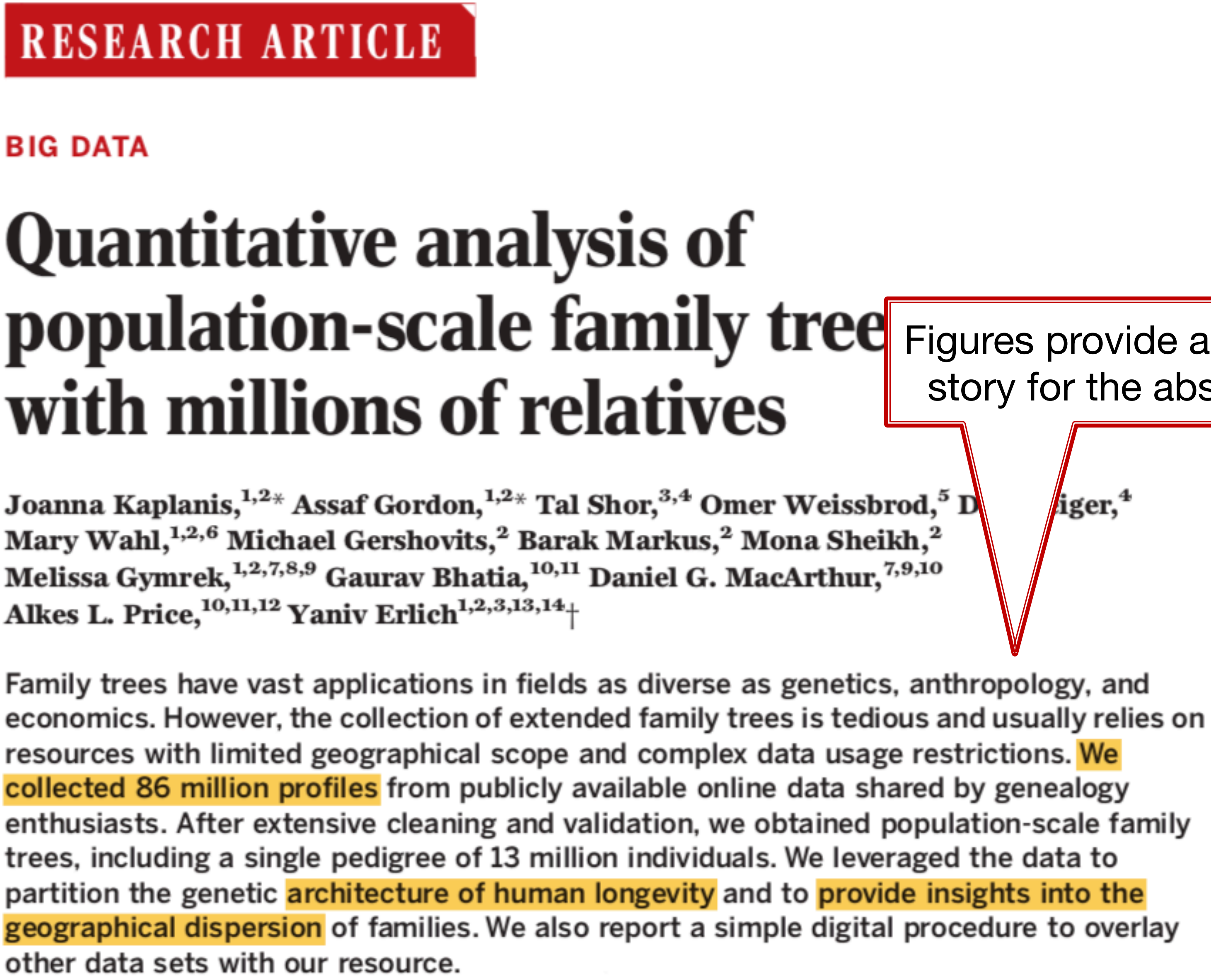
Interdisciplinary Journal Papers: Visual Structure

Very hard: non-standard design, custom drawing



Hard: non-standard design, mixture of custom drawings and standard visualization libraries

Examples: Interdisciplinary Journal Papers



RESEARCH ARTICLE

BIG DATA

Quantitative analysis of population-scale family tree with millions of relatives

Joanna Kaplanis,[1,2*] Assaf Gordon,[1,2*] Tal Shor,[3,4] Omer Weissbrod,[5] D iger,[4] Mary Wahl,[1,2,6] Michael Gershovits,[2] Barak Markus,[2] Mona Sheikh,[2] Melissa Gymrek,[1,2,7,8,9] Gaurav Bhatia,[10,11] Daniel G. MacArthur,[7,9,10] Alkes L. Price,[10,11,12] Yaniv Erlich[1,2,3,13,14]†

Family trees have vast applications in fields as diverse as genetics, anthropology, and economics. However, the collection of extended family trees is tedious and usually relies on resources with limited geographical scope and complex data usage restrictions. We collected 86 million profiles from publicly available online data shared by genealogy enthusiasts. After extensive cleaning and validation, we obtained population-scale family trees, including a single pedigree of 13 million individuals. We leveraged the data to partition the genetic architecture of human longevity and to provide insights into the geographical dispersion of families. We also report a simple digital procedure to overlay other data sets with our resource.

Figures provide a visual story for the abstract

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, *Science*, 2018.

Figure 1

Dataset

Approach

Impress your audience! 😲

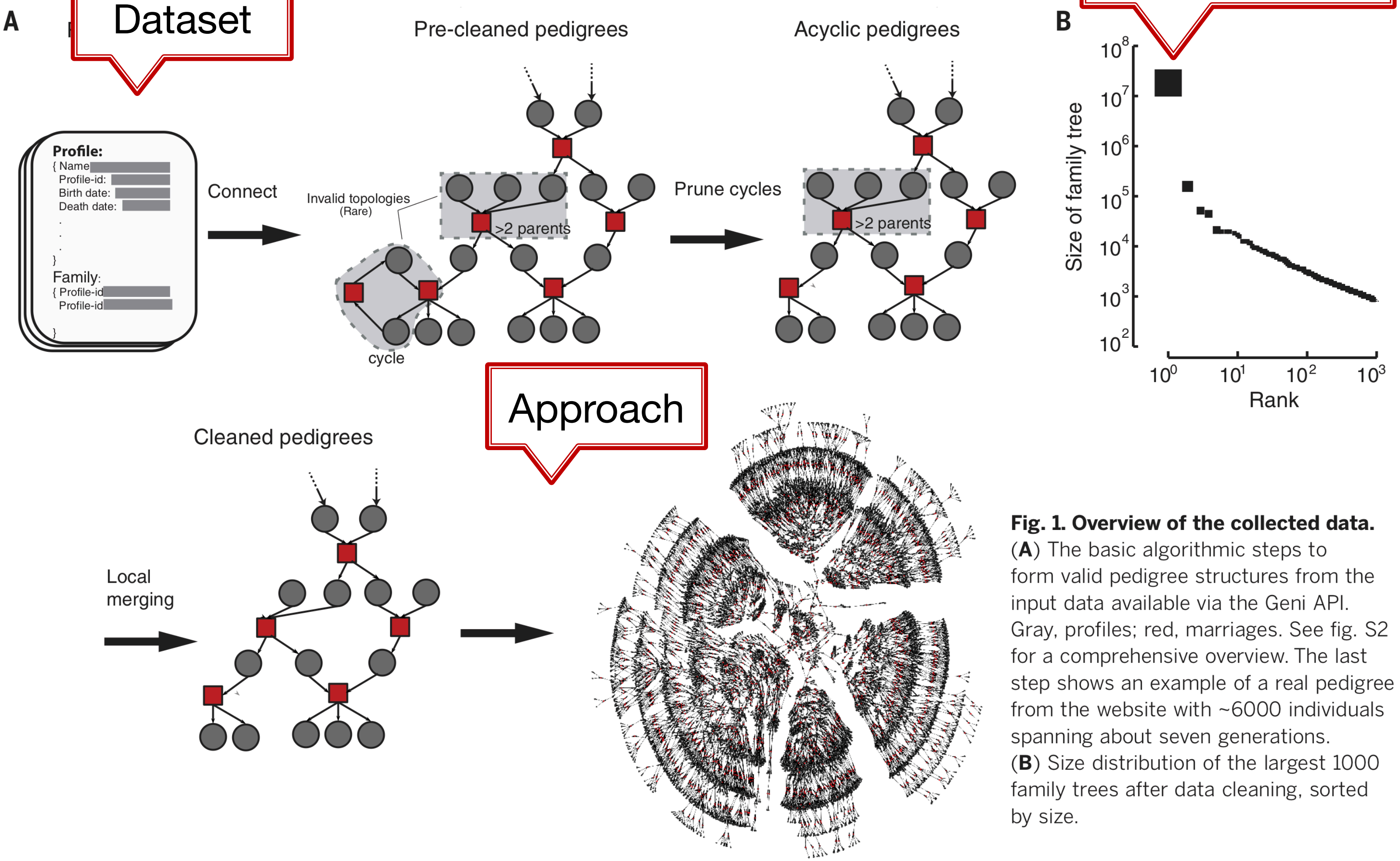
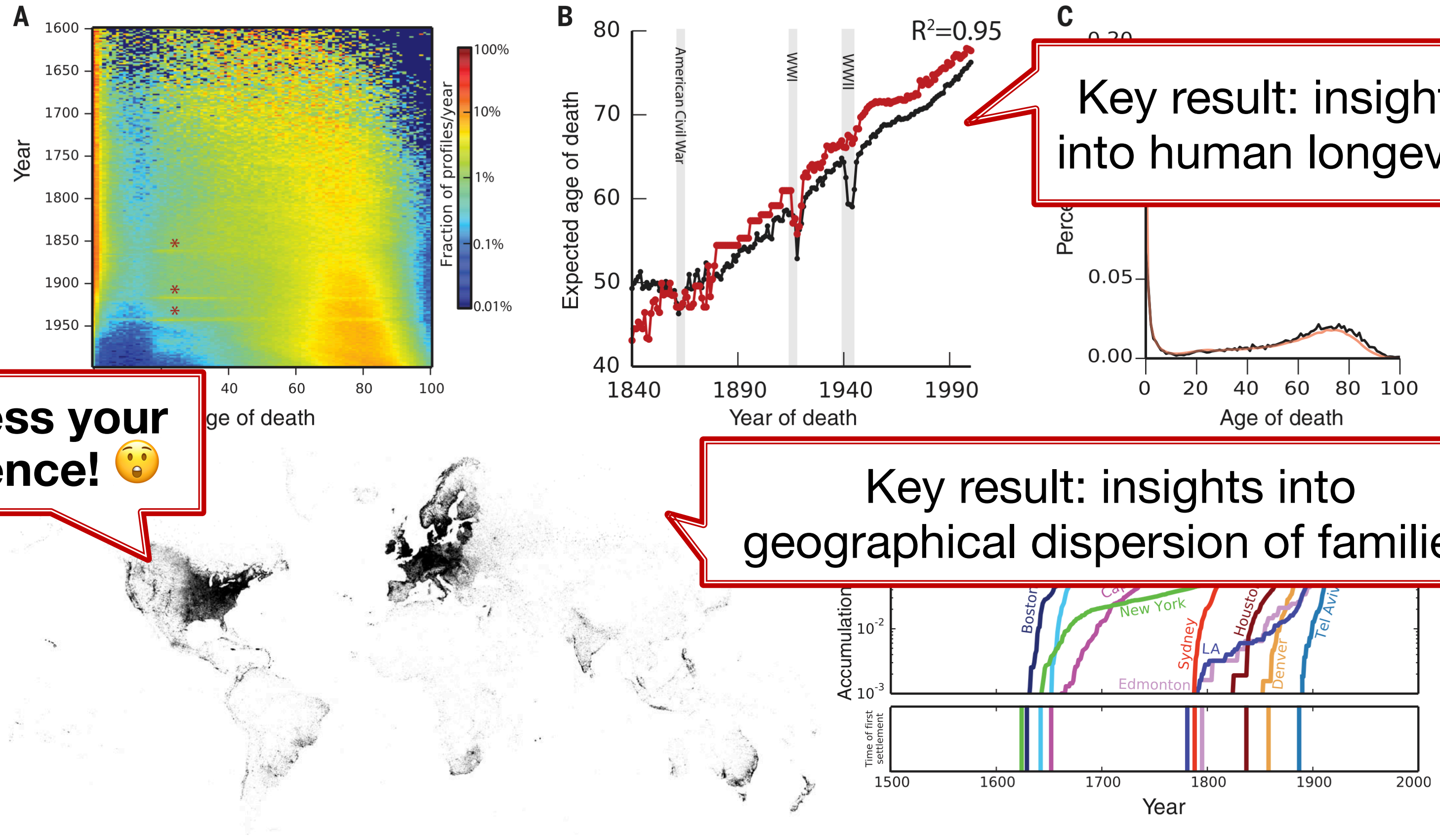


Fig. 1. Overview of the collected data. (**A**) The basic algorithmic steps to form valid pedigree structures from the input data available via the Geni API. Gray, profiles; red, marriages. See fig. S2 for a comprehensive overview. The last step shows an example of a real pedigree from the website with ~6000 individuals spanning about seven generations. (**B**) Size distribution of the largest 1000 family trees after data cleaning, sorted by size.

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, *Science*, 2018.

Figure 2



Key result: insights into human longevity

Impress your audience! 😲

Key result: insights into geographical dispersion of families

Fig. 2. Analysis and validation of demographic data. (**A**) Distribution of life expectancy per year. Colors correspond to the frequency of profiles of individuals who died at a certain age for each year. Asterisks indicate deaths at military age in the Civil War and First and Second World Wars. (**B**) Expected life span in Geni (black) and the Oeppen and Vaupel study [red (*27*)] as a function of year of death. (**C**) Comparison of the life-span distributions versus Geni (black) and HMD (red). See also fig. S5A. (**D**) Geographic distribution of the annotated place-of-birth information. Every pixel corresponds to a profile in the data set. (**E**) Validation of geographical assignment by historical trends. Top: Cumulative distribution of profiles since 1500 for each city on a logarithmic scale as a function of time. Bottom: Year of first settlement in the city.

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, *Science*, 2018.

Figure 3

Further analyses supporting key result

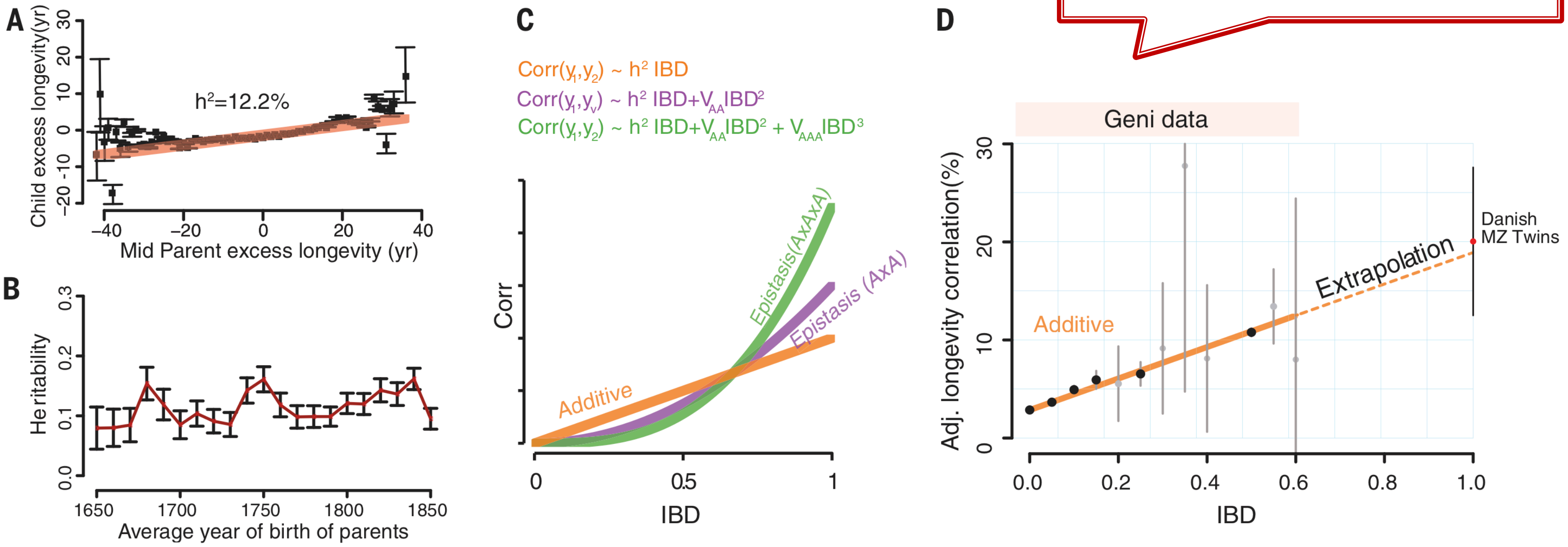
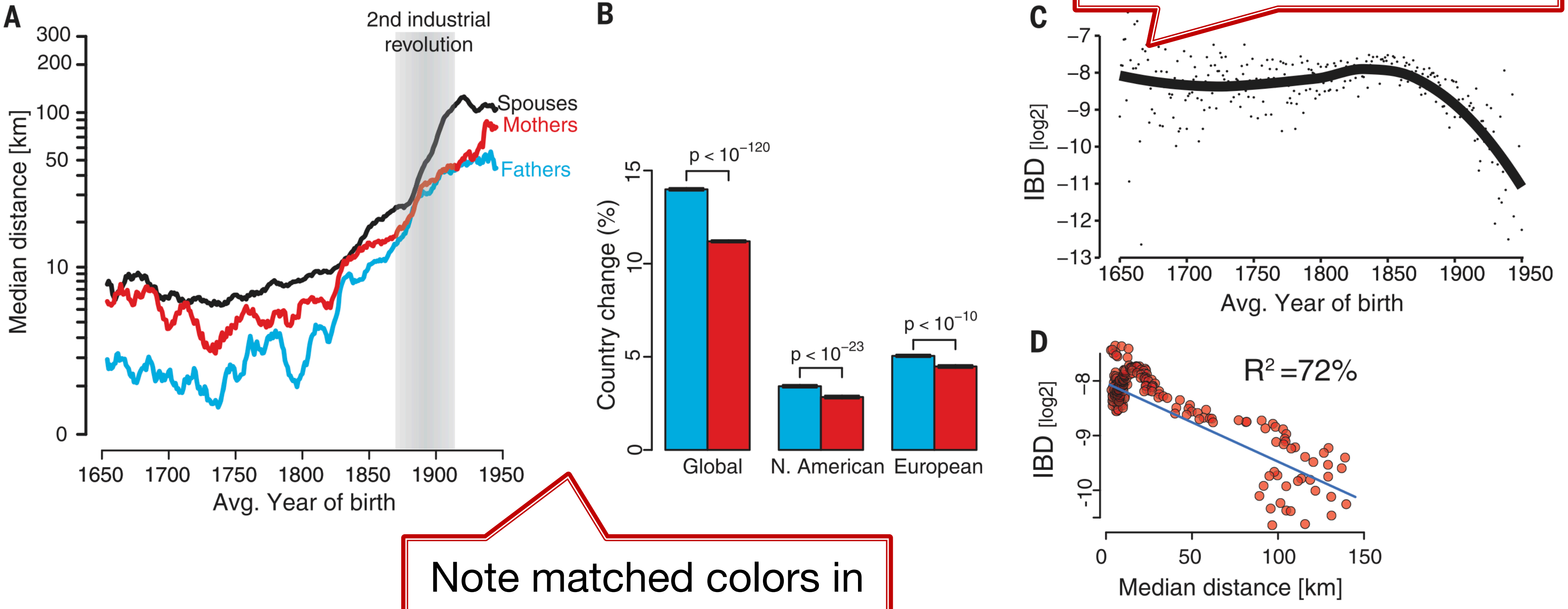


Fig. 3. The genetic architecture of longevity. (**A**) Regression (red) of child longevity on its mid-parent longevity (defined as difference between age of death and expected life span). Black squares, average longevity of children binned by the mid-parent value; gray bars, estimated 95% confidence interval (CI). (**B**) Estimated narrow-sense heritability (red) with 95% confidence intervals (black bars) obtained by the mid-parent design stratified by the average decade of birth of the parents. (**C**) Correlation of a trait as a function of IBD under strict additive (*h^2*, orange), squared (V_{AA}, purple), and cubic (V_{AAA}, green) epistasis architectures after dormancy adjustments. (**D**) Average longevity correlation as a function of IBD (black circles) grouped in 5% increments (gray: 95% CI) after adjusting for dominancy. A dashed line denotes the extrapolation of the models toward monozygotic twins from the Danish Twin Registry (red circle).

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, *Science*, 2018.

Figure 4

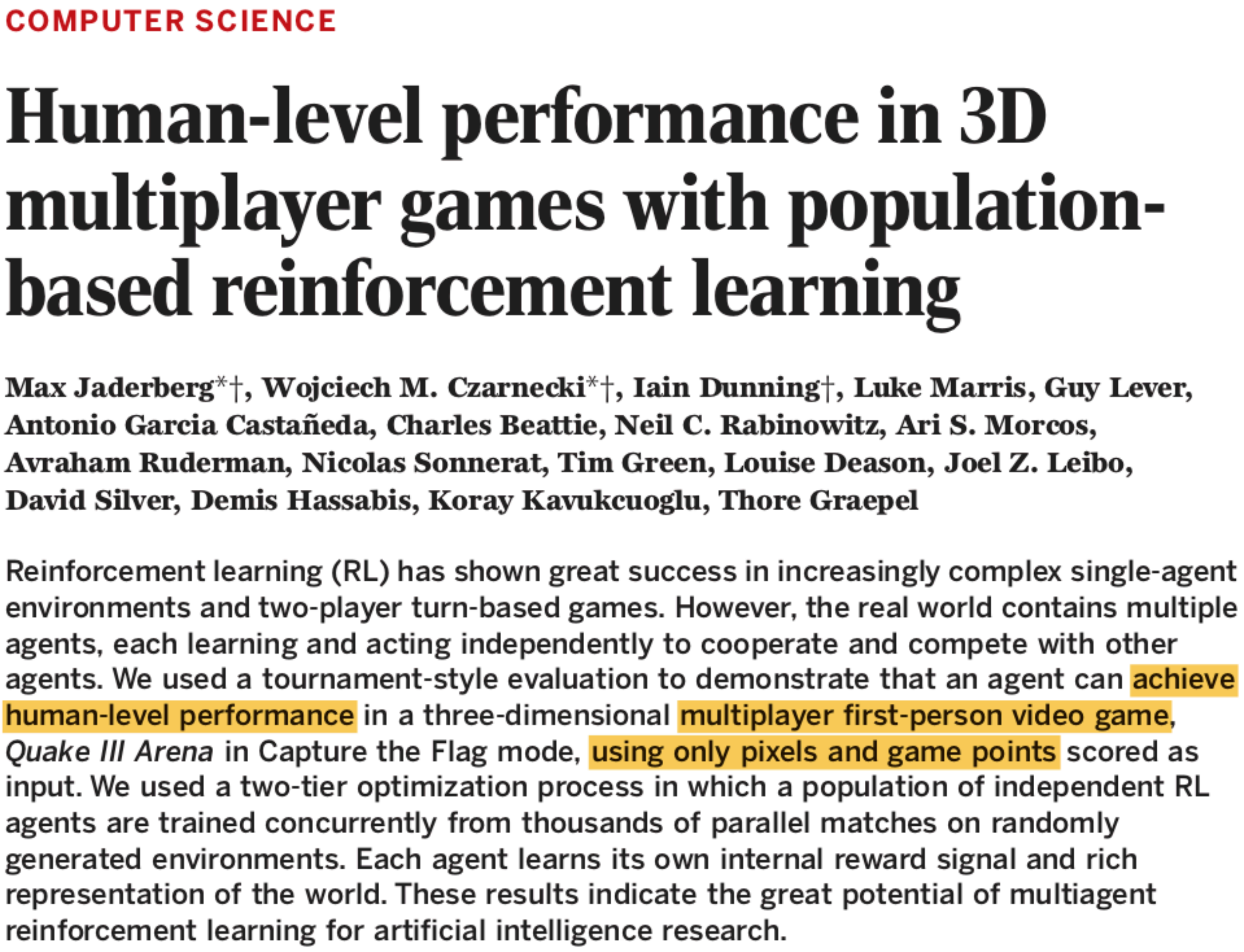
Further analyses supporting key result



Note matched colors in panels A and B

Fig. 4. Analysis of familial dispersion of father-offspring places of birth (cyan), mother-offspring (red), and marital radius (black) as a function of time (average year of birth). (**B**) Rate of change in the country of birth for father-offspring (cyan) or mother-offspring (red) stratified by major geographic areas. (**C**) Average IBD ($\log_2$) between s a function of average year of birth. Individual dots represent the measured average per year; the black line denotes the smooth trend using locally weighted regression. (**D**) IBD of couples as a function of marital radius. Each dot represents a year between 1650 to 1950. The blue line denotes the best linear regression line in log-log space.

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, *Science*, 2018.



COMPUTER SCIENCE

Human-level performance in 3D multiplayer games with population-based reinforcement learning

Max Jaderberg*†, Wojciech M. Czarnecki*†, Iain Dunning†, Luke Marris, Guy Lever, Antonio Garcia Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, Thore Graepel

Reinforcement learning (RL) has shown great success in increasingly complex single-agent environments and two-player turn-based games. However, the real world contains multiple agents, each learning and acting independently to cooperate and compete with other agents. We used a tournament-style evaluation to demonstrate that an agent can achieve human-level performance in a three-dimensional multiplayer first-person video game, *Quake III Arena* in Capture the Flag mode, using only pixels and game points scored as input. We used a two-tier optimization process in which a population of independent RL agents are trained concurrently from thousands of parallel matches on randomly generated environments. Each agent learns its own internal reward signal and rich representation of the world. These results indicate the great potential of multiagent reinforcement learning for artificial intelligence research.

Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019.

Figure 1

Dataset

Approach

Impress your audience! 😲

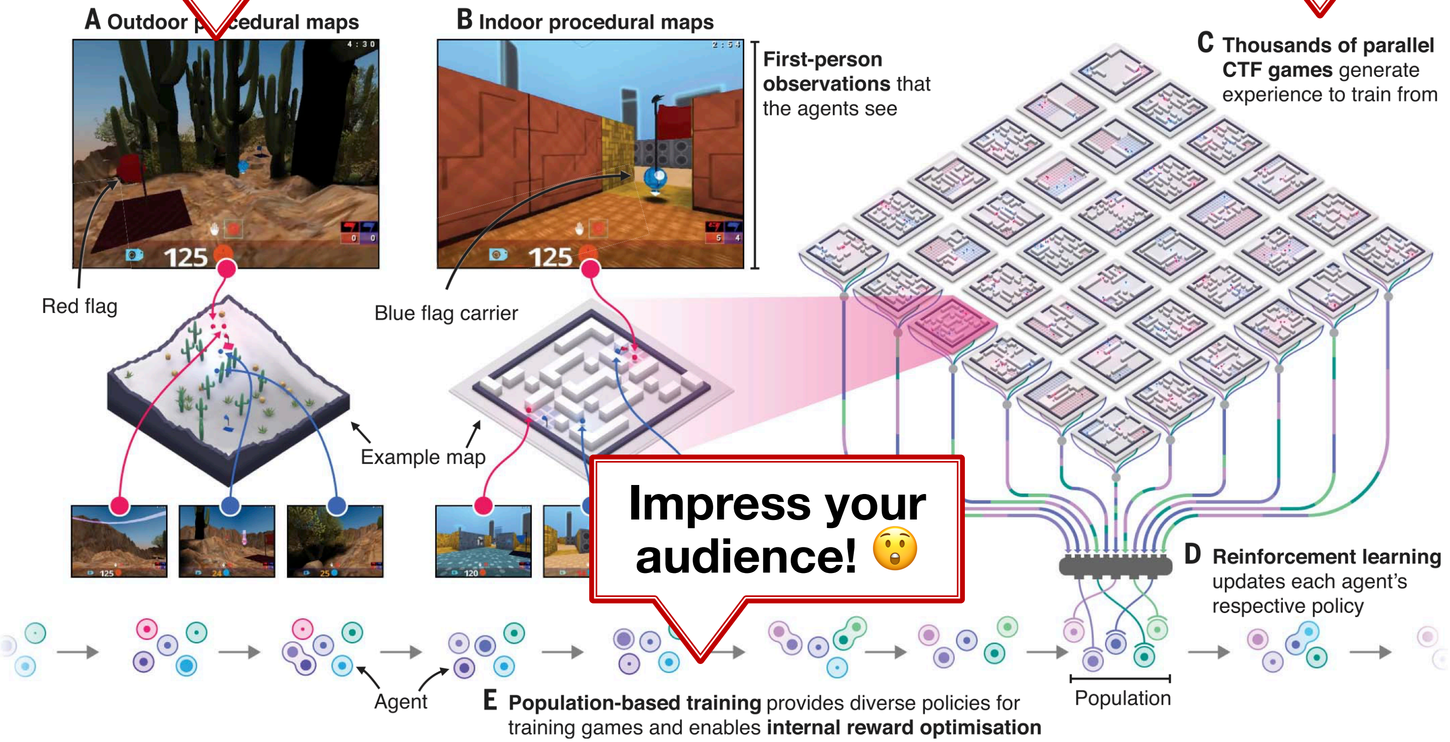

A Outdoor procedural maps
4:30
125
Red flag
B Indoor procedural maps
2:54
125
Blue flag carrier
First-person observations that the agents see
C Thousands of parallel CTF games generate experience to train from
Example map
D Reinforcement learning updates each agent's respective policy
Agent
E Population-based training provides diverse policies for training games and enables internal reward optimisation
Population

Fig. 1. CTF task and computational training framework. (**A** and **B**) Two example maps that have been sampled from the distribution of (A) outdoor maps and (B) indoor maps. Each agent in the game sees only its own first-person pixel view of the environment. (**C**) Training data are generated by playing thousands of CTF games in parallel on a diverse distribution of procedurally generated maps and (**D**) used to train the agents that played in each game with RL. (**E**) We trained a population of 30 different agents together, which provided a diverse set of teammates and opponents to play with and was also used to evolve the internal rewards and hyperparameters of agents and learning process. Each circle represents an agent in the population, with the size of the inner circle representing strength. Agents undergo computational evolution (represented as splitting) with descendents inheriting and mutating hyperparameters (represented as color). Gameplay footage and further exposition of the environment variability can be found in movie S1.

Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019.

Figure 2

Approach: details

Impress your audience! 😲

Key result: human-level performance

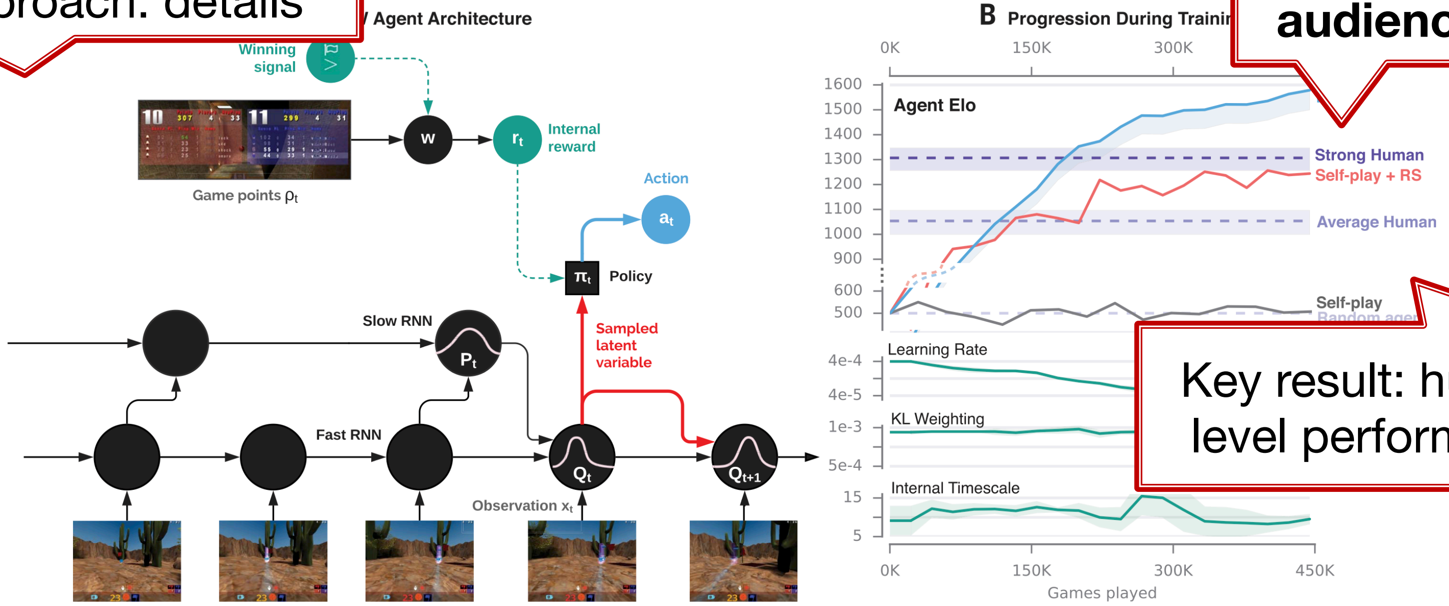


Fig. 2. Agent architecture and benchmarking. (**A**) How the agent processes a temporal sequence of observations x_t from the environment. The model operates at two different time scales, faster at the bottom and slower by a factor of τ at the top. A stochastic vector-valued latent variable is sampled at the fast time scale from distribution $\mathbb{Q}_t$ on the basis of observations x_t. The action distribution π_t is sampled conditional on the latent variable at each time step t. The latent variable is regularized by the slow moving prior $\mathbb{P}_t$, which helps capture long-range temporal correlations and promotes memory. The network parameters are updated by using RL according to the agent's own internal reward signal r_t, which is obtained from a learned transformation **w** of game points ρ_t. **w** is optimized for winning probability through PBT, another level of training performed at yet a slower time scale than that of RL. Detailed network architectures are described in fig. S11. (**B**) (Top) The Elo skill ratings of the FTW agent population throughout training (blue) together with those of the best baseline agents by using hand-tuned reward shaping (RS) (red) and game-winning reward signal only (black), compared with human and random agent reference points (violet, shaded region shows strength between 10th and 90th percentile). The FTW agent achieves a skill level considerably beyond strong human subjects, whereas the baseline agent's skill plateaus below and does not learn anything without reward shaping [evaluation procedure is provided in (*28*)]. (Bottom) The evolution of three hyperparameters of the FTW agent population: learning rate, Kullback-Leibler divergence (KL) weighting, and internal time scale τ, plotted as mean and standard deviation across the population.

Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019.

Figure 3

Further analyses supporting key result

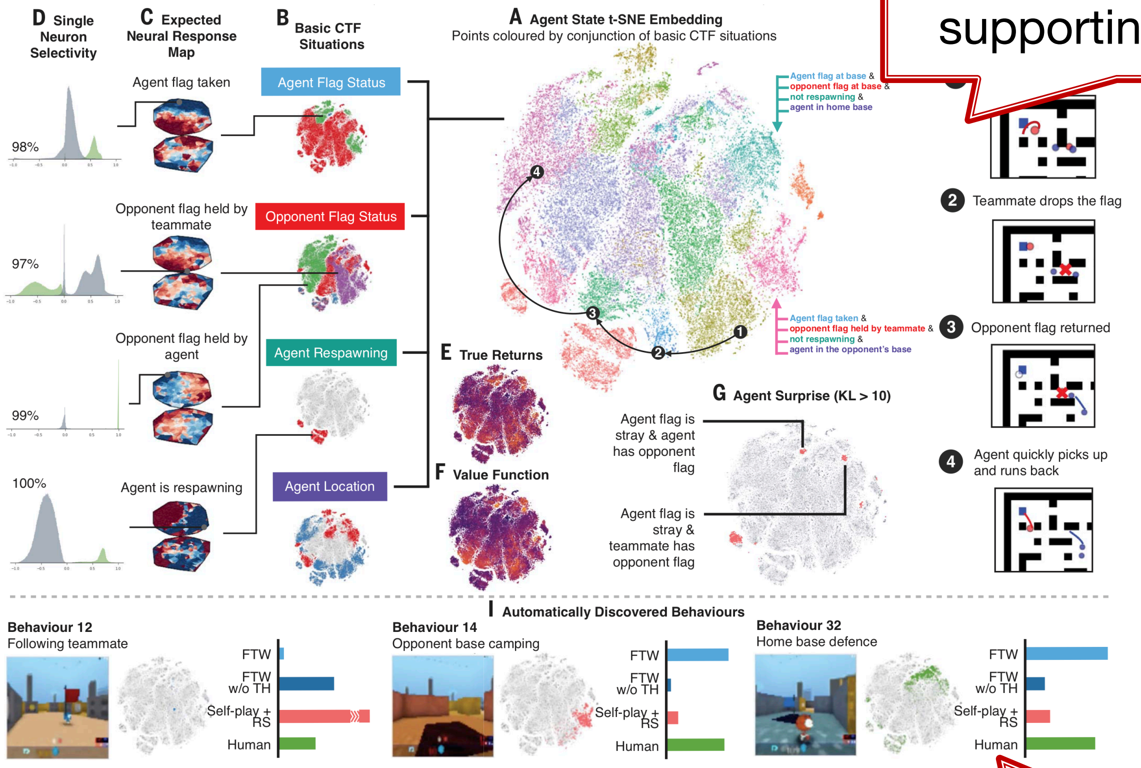


Fig. 3. Knowledge represen… behavioral analysis. (**A**) The 2D t-SNE embeddi… al states during gameplay. …oint … at this … of … by … the same … state arranged in a similarity-preserving topological embedding and colored according to activation (fig. S5). (**D**) Distributions of situation conditional activations (each conditional distribution is colored gray and green) for particular single neurons that are distinctly selective for these CTF situations and show the predictive accuracy of this neur… true return of the agent's internal reward sig… prediction, its value function (orange denote… denotes low value). (**G**) Regions where the ag… representation diverges (red), the agent's su… KL between the agent's slow– and fast–tim… (*28*). (**H**) The four-step temporal sequence … "opponent base camping." (**I**) Three automa… behaviors of agents and corresponding regions in the t-SNE embedding. (Right) Average occurrence per game of each behavior for the FTW agent, the FTW agent without temporal hierarchy (TH), self-play with reward shaping agent, and human subjects (fig. S9).

Note how observations x_t are visualized, i.e., maps of base camps

Note the use of matched colors

Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019.

Figure 4

Further analyses supporting key result

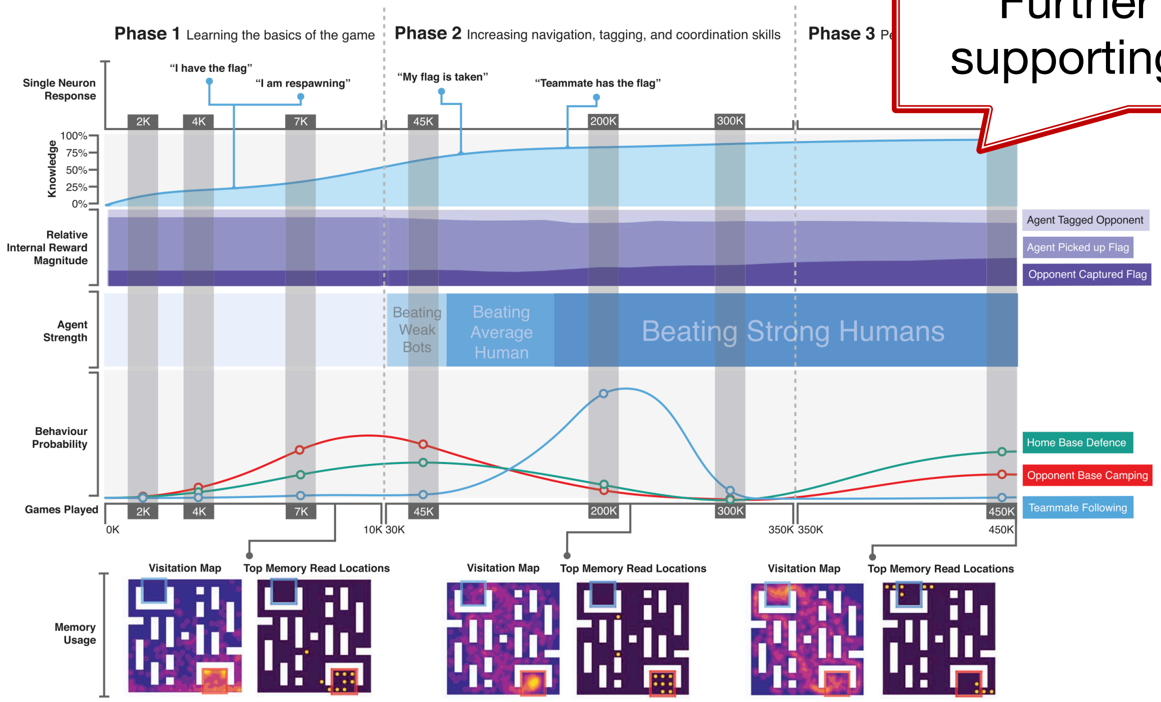


Fig. 4. Progression of agent during training. Shown is the development of knowledge representation and behaviors of the FTW agent over the training period of 450,000 games, segmented into three phases (movie S2). "Knowledge" indicates the percentage of game knowledge that is linearly decodable from the agent's representation, measured by average scaled AUCROC across 200 features of game state. Some knowledge is compressed to single-neuron responses (Fig. 3A), whose emergence in training is shown at the top. "Relative internal reward magnitude" indicates the relative magnitude of the agent's internal reward weights of 3 of the 13 events corresponding to game points ρ. Early in training, the agent puts large reward weight on picking up the opponent's flag, whereas later, this weight is reduced, and reward for tagging an opponent and penalty when opponents capture a flag are increased by a factor of two. "Behavior probability" indicates the frequencies of occurrence for 3 of the 32 automatically discovered behavior clusters through training. Opponent base camping (red) is discovered early on, whereas teammate following (blue) becomes very prominent midway through training before mostly disappearing. The "home base defense" behavior (green) resurges in occurrence toward the end of training, which is in line with the agent's increased internal penalty for more opponent flag captures. "Memory usage" comprises heat maps of visitation frequencies for (left) locations in a particular map and (right) locations of the agent at which the top-10 most frequently read memories were written to memory, normalized by random reads from memory, indicating which locations the agent learned to recall. Recalled locations change considerably throughout training, eventually showing the agent recalling the entrances to both bases, presumably in order to perform more efficient navigation in unseen maps (fig. S7).

Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019.

Today's Lecture

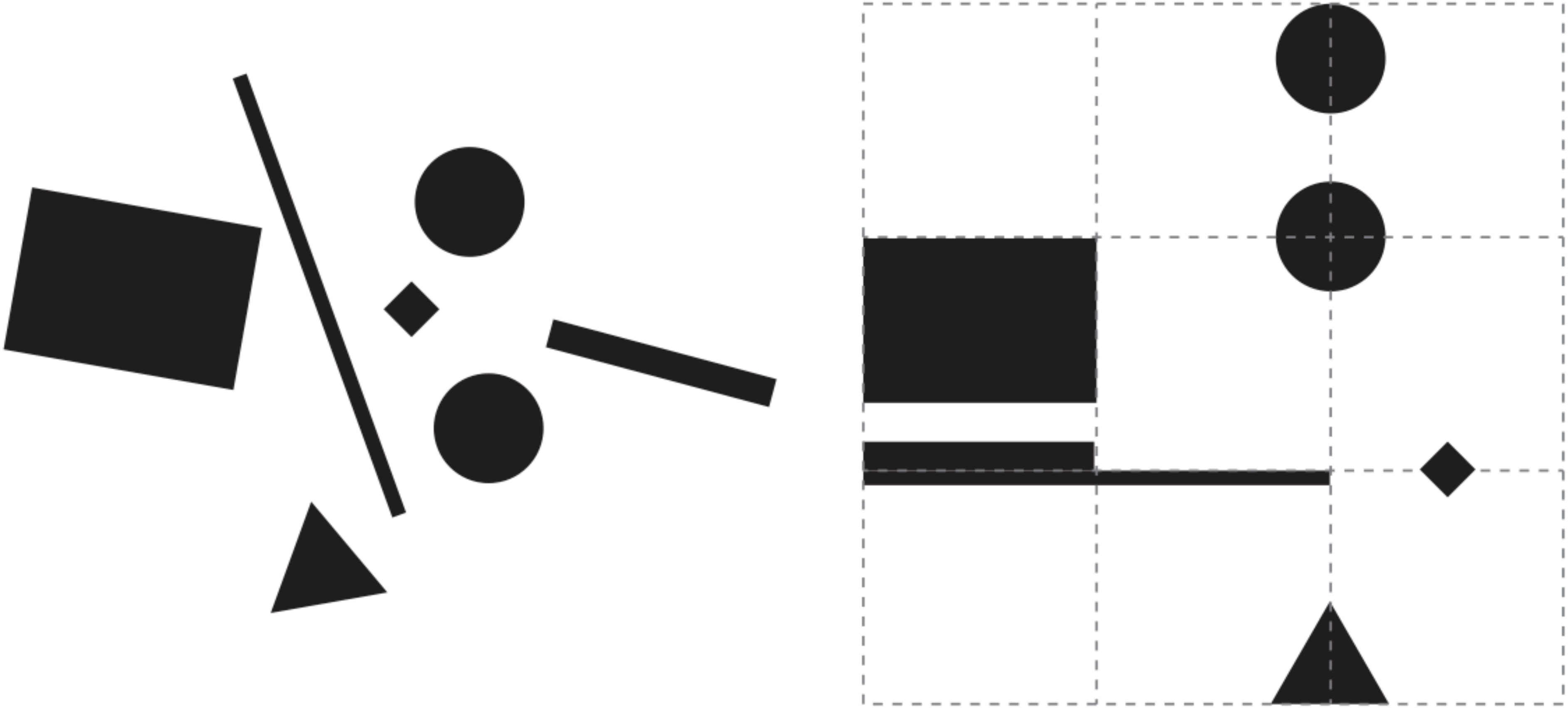
1) Why figures matter ✓
2) Figures in science ✓
3) How to design effective figures
4) Tools, tips, and guidelines ☜

Principle #1: Design figures for the audience (not for you)

Before your design figures think about:

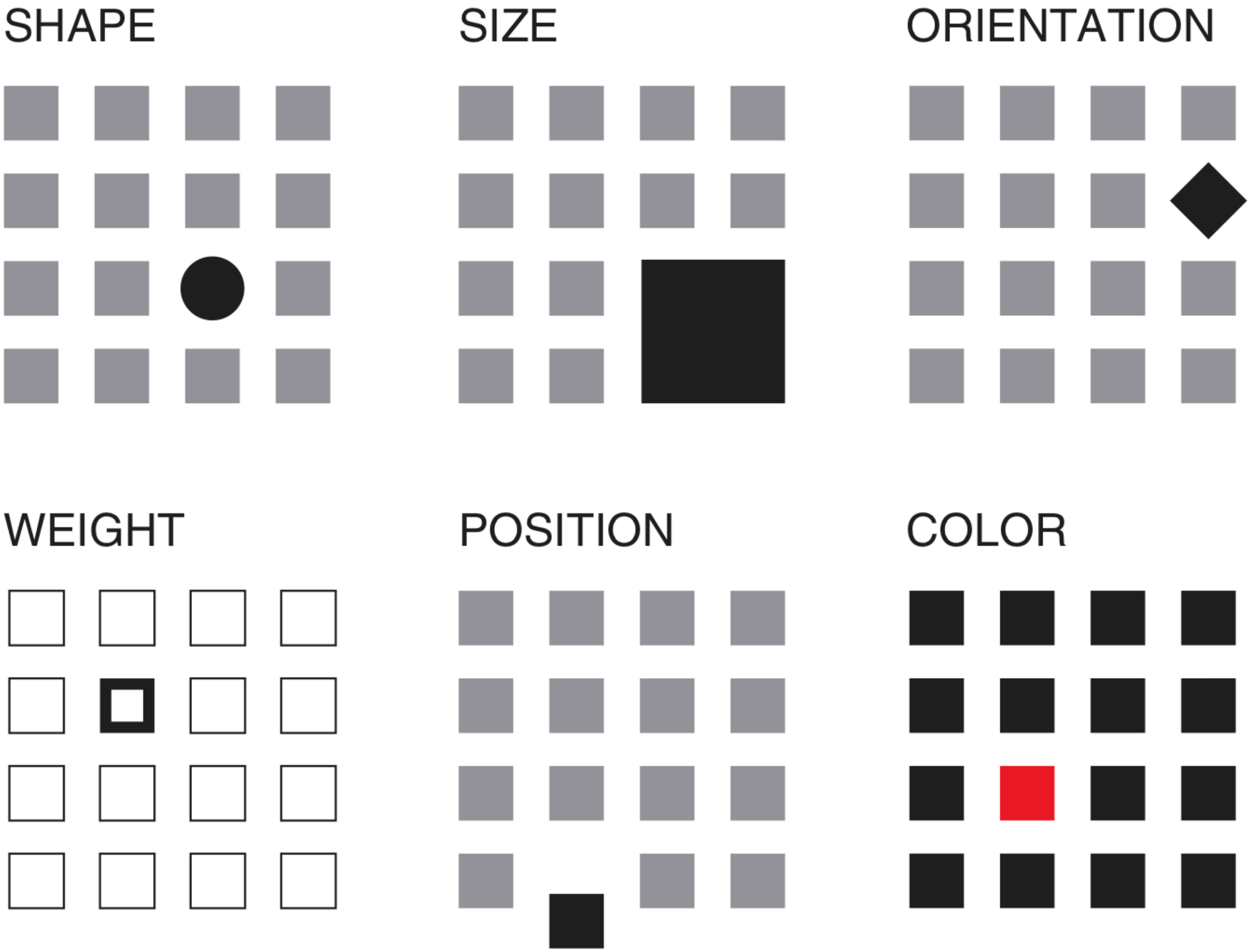
1) Make-up of the audience:
   - Will a figure appear in a specialized journal?
   - Is a figure aimed at a broad readership?
2) Background knowledge of the audience:
   - Audience may not know what you know
   - Figures should provide all the information necessary for the audience to fully comprehend them
3) Disciplinary conventions:
   - Graphical conventions and norms exist in each field

Principle #2: Design a clear visual structure with pleasant symmetries



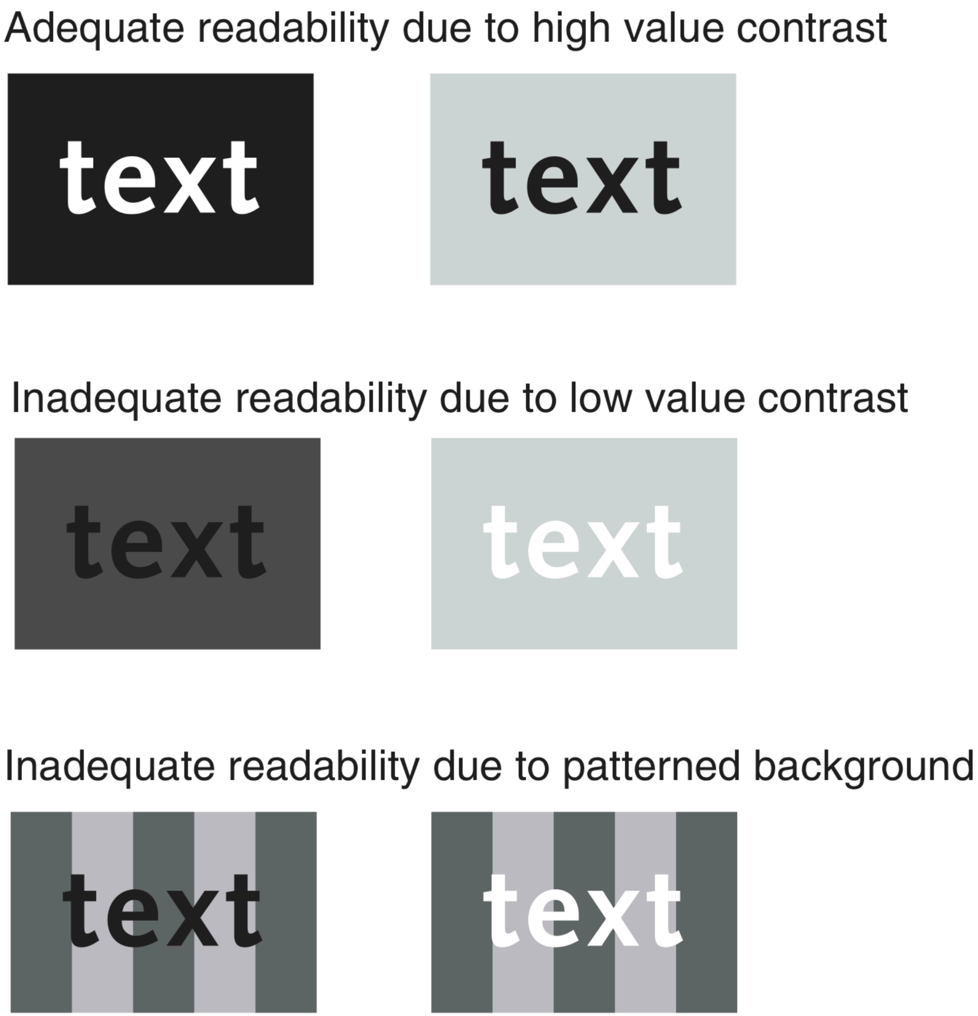
Rolandi *et al.*, A brief guide to designing effective figures for the scientific paper. Advanced Materials 23.38 (2011)

Principle #3: Use visual contrast, but keep figures simple



Rolandi *et al.*, A brief guide to designing effective figures for the scientific paper. Advanced Materials 23.38 (2011)

Principle #4: Use readable and legible typography



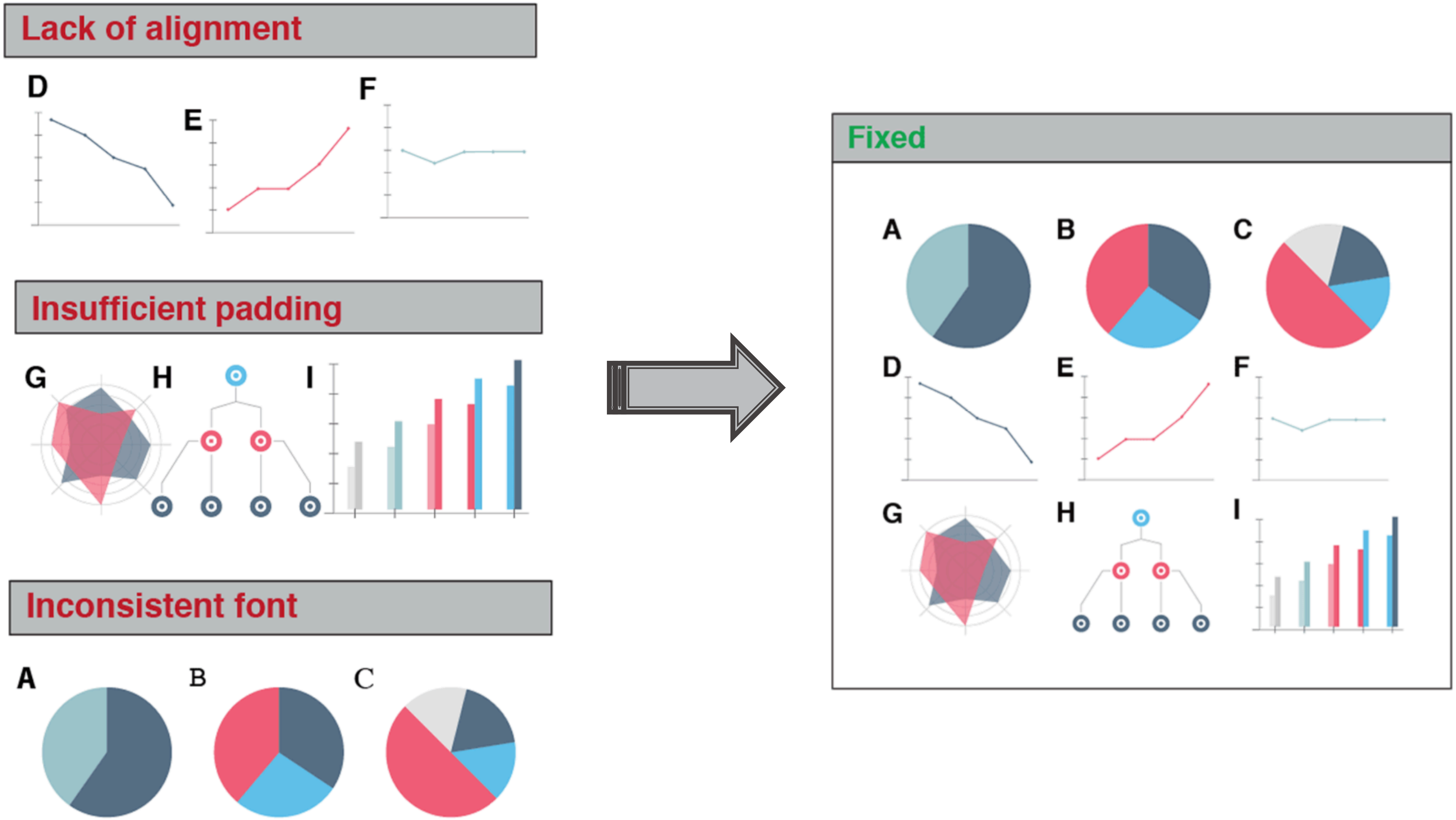
Adequate readability due to high value contrast

Inadequate readability due to low value contrast

Inadequate readability due to patterned background

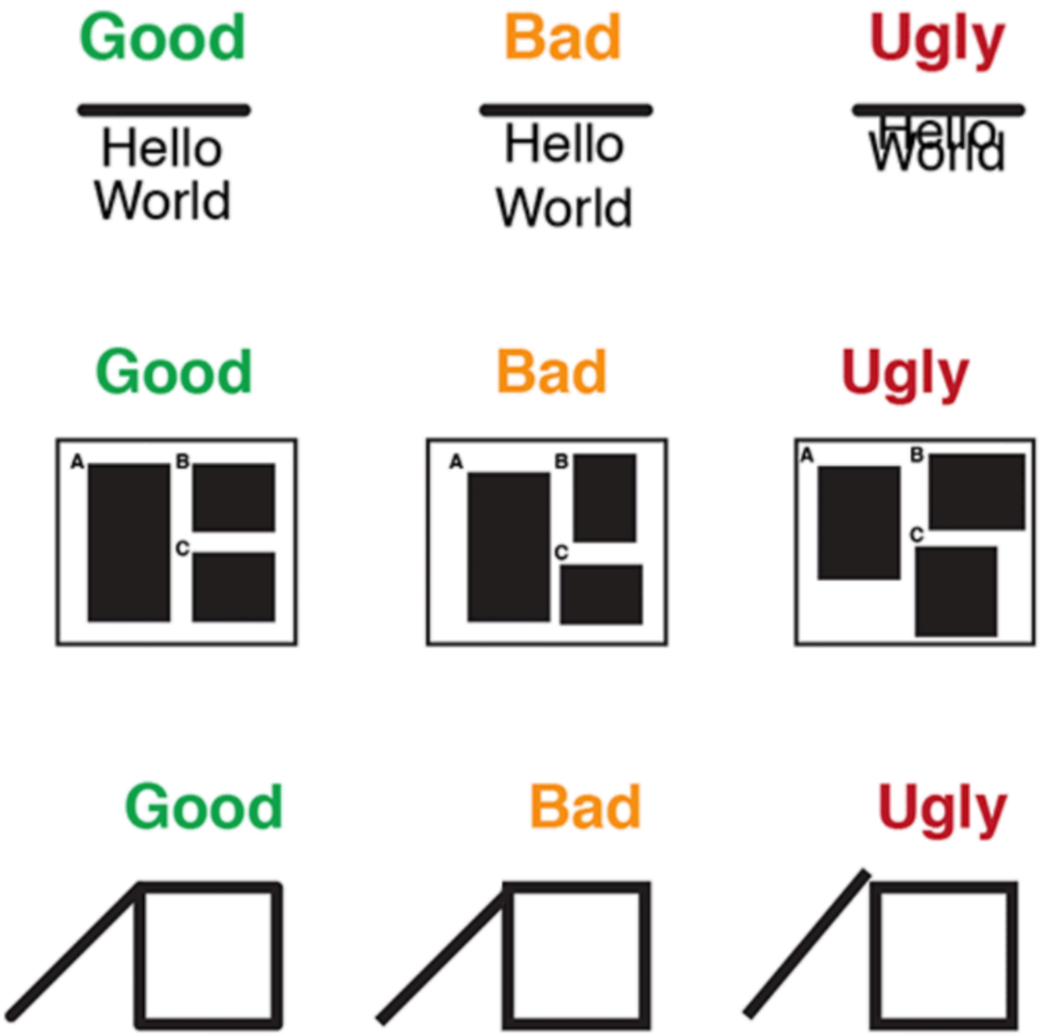
Rolandi *et al.*, A brief guide to designing effective figures for the scientific paper. Advanced Materials 23.38 (2011)

Principle #5: Be consistent, align panels and use sufficient padding



Source: Jean Fan, Harvard

The Good, Bad, and Ugly



Source: Jean Fan, Harvard

Today's Lecture

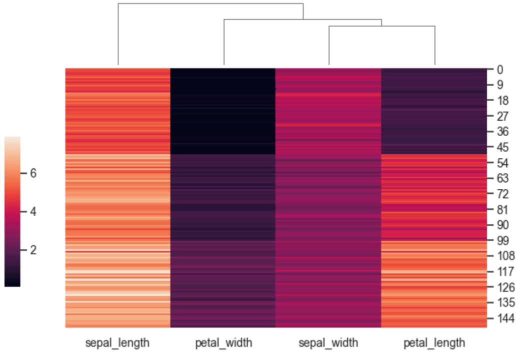
1) Why figures matter ✓
2) Figures in science ✓
3) How to design effective figures ✓
4) Tools, tips, and guidelines ☛

Key Rules to ALWAYS Follow

1) Save raw data and results to a tsv/csv/binary file:
   - Your figures will need multiple rounds of editing
2) Read in the data and design figures

Important: Save figures as PDF or other vector format:

- You might need to use multiple tools to draw a figure
  - Example:
    1. First, use seaborn to draw a clustermap
    2. Then, export clustermap as PDF
    3. Finally, use Adobe Illustrator to annotate the clustermap
  - Example:
    1. First, use D3.js to layout a network
    2. Then, export the network as PDF
    3. Finally, use Adobe Illustrator to show node features and node labels

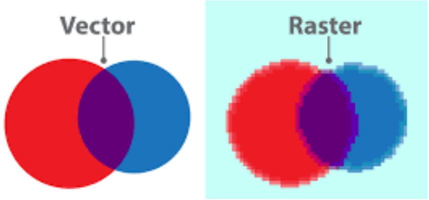


Why shouldn’t you use raster formats (e.g., JPG, GIF, PNG, TIF)?

Raster images:

- Use a fixed number of colored pixels and can’t be dramatically resized (pixilation, distortion issues)
- When saved, they cannot be reopened and edited!

Vector images (e.g., PDF, EPS, AI, SVG):

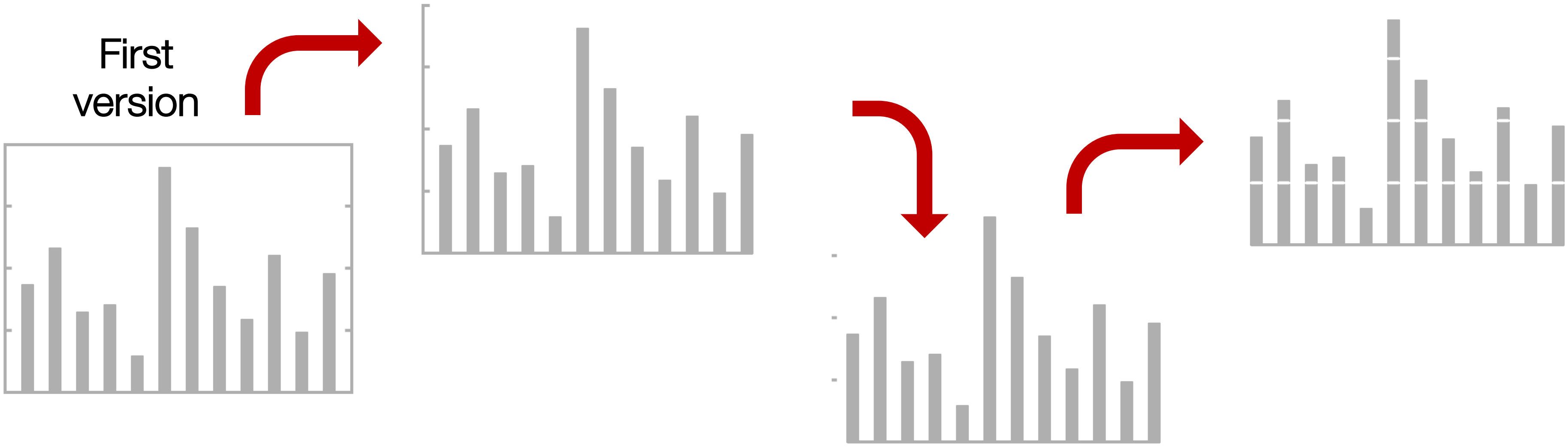


- Remain editable!
- You can open them in Illustrator and edit text or any other element within the graphic
- Can be converted to a raster image but not vice-versa
- plt.savefig(‘myfig.pdf’)

Only use raster format for web, Github repo, etc.

Visualization Design Guidelines #1

1) Tufte's design rules:
   - sealthreinhold.com/school/tuftes-rules
   - Data-to-ink-ratio: Maximize data-ink and erase as much non-data-ink as possible (avoid chart junk)
2) Art is science is art, mkweb.bcgsc.ca



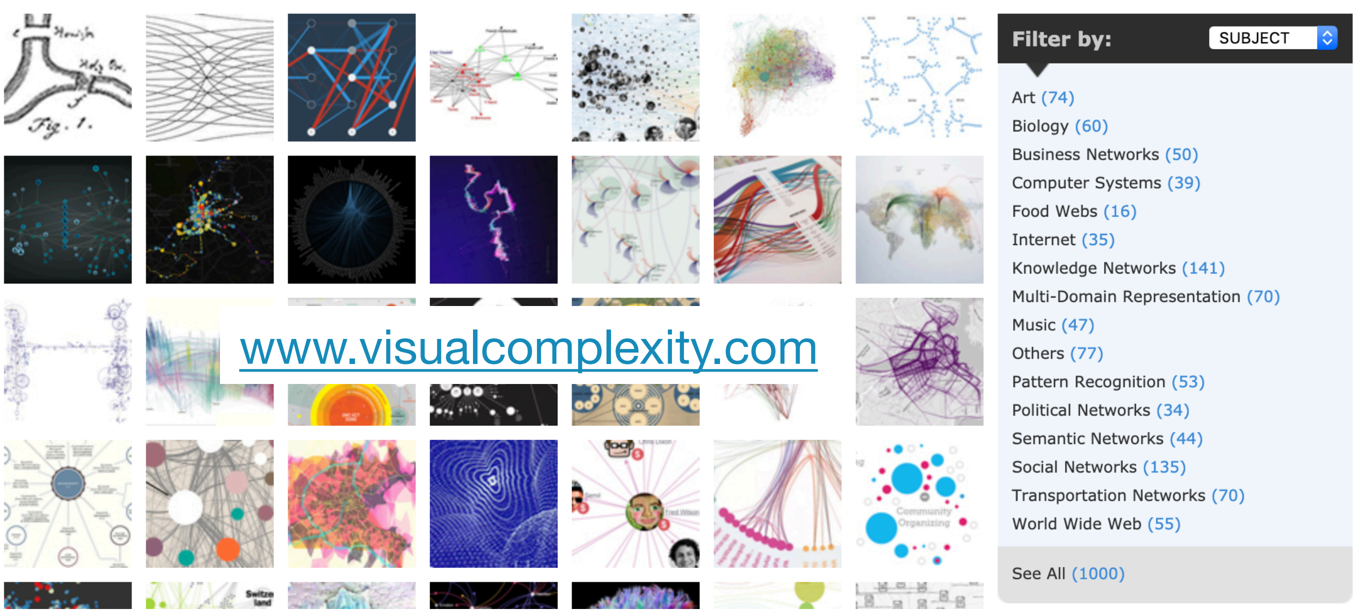
First version

Final version

Visualization Design Guidelines #2

3) Google's principles for designing charts:
- material.io/design/communication/data-visualization.html
  - Principles: Be honest, Lend a helping hand, Delight users, Give clarity of focus, Embrace scale, Provide structure

4) Manuel Lima, Design Lead @ Google:



www.visualcomplexity.com

Tools, Software & Frameworks

Tools, Software, and Frameworks

- Adobe Illustrator
  - Adobe Creative Cloud
- LaTeXiT
  - chachatelier.fr/latexit
- Matplotlib
  - matplotlib.org
- Seaborn
  - seaborn.pydata.org
- Bokeh
  - bokeh.pydata.org
- D3.js
  - d3js.org
- GeoPandas
  - geopandas.org

- Google Charts
  - developers.google.com/chart
- Circos
  - circos.ca
- gnuplut
  - gnuplot.info
- TikZ
  - texample.net/tikz
- Plotly
  - plot.ly/python
- missingno
  - github.com/ResidentMario/missingno
- billboard.js
  - naver.github.io/billboard.js
- Squaire.js
  - wsj.github.io/squaire

Adobe Illustrator and Alternatives

- Where to get on campus:
  - For purchase: https://itconnect.uw.edu/wares/uware/adobe-creative-cloud/
  - **Use for Free:** UW Library https://www.lib.washington.edu/media/software

- Free alternatives:
  - Inkscape, https://inkscape.org
  - GIMP, https://www.gimp.org
  - Boxy-SVG, https://boxy-svg.com

How to get from a JS vis to an effective figure?

Three steps:

1) Use a JS library from two slide ago and generate a visualization

2) Generate a PDF file from HTML:
   - stackoverflow.com/questions/18191893/generate-pdf-from-html-in-div-using-javascript

3) Open the PDF in Illustrator and make further edits:
   - Change colors
   - Add labels and annotations
   - Add new visual elements, e.g., insets, logos
   - Combine with other graphics to get a multi-panel figure

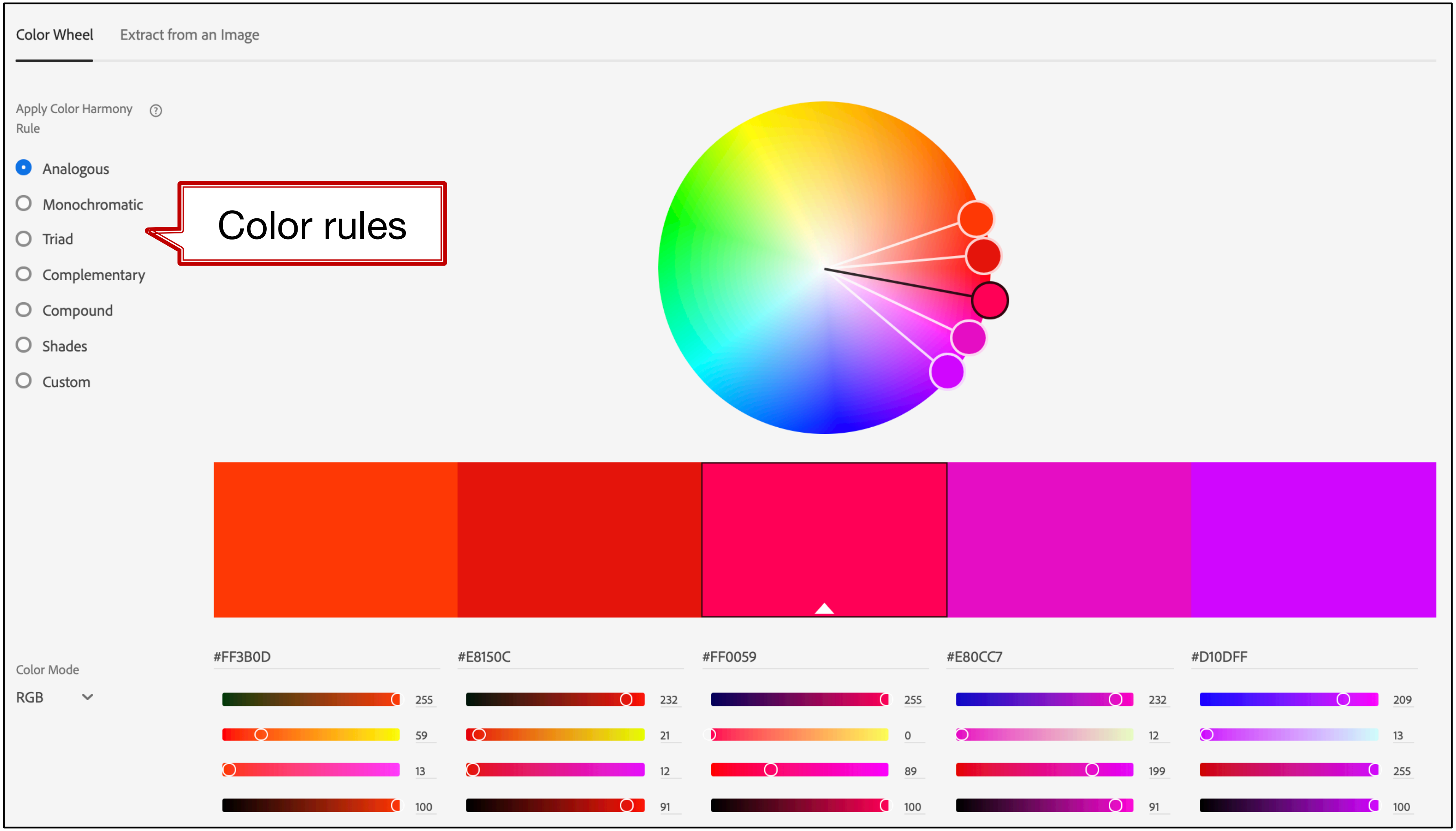
Tools for Network & Relational Data

- Gephi, gephi.org
- Graphviz, graphviz.org
- NetworkX, networkx.github.io
- JSNetworkX, jsnetworkx.org
- igraph, igraph.org/python
- sigma.js, sigmajs.org
- Cytoscape, cytoscape.org
- Hive plots, hiveplot.com

Colors

Color Advice

Adobe color, https://color.adobe.com



Color Wheel Extract from an Image

Apply Color Harmony Rule

- ● Analogous
- ○ Monochromatic
- ○ Triad
- ○ Complementary
- ○ Compound
- ○ Shades
- ○ Custom

Color rules

Color Mode

RGB

#FF3B0D	#E8150C	#FF0059	#E80CC7	#D10DFF
255	232	255	232	209
59	21	0	12	13
13	12	89	199	255
100	91	100	91	100

Color Advice: Brewer Palettes

Brewer palettes: Color combinations selected for their special properties for use in data visualization

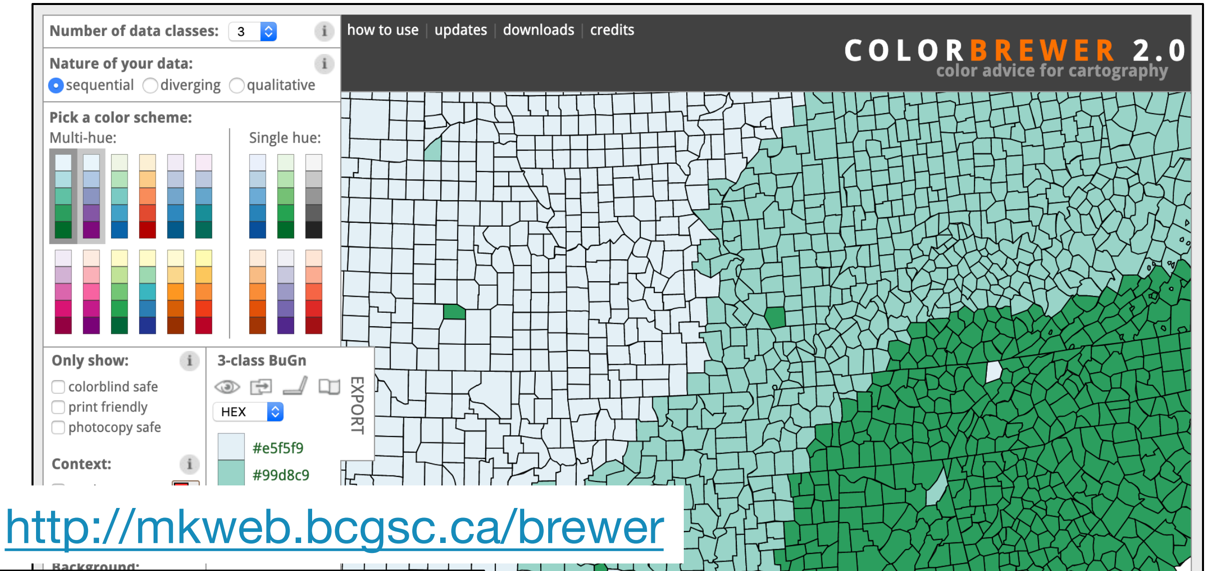
3 types of palettes:

qualitative — colors do not have a perceived order

sequential — colors have a perceived order and perceived difference between successive colors is uniform

diverging — two back-to-back sequential palettes starting from a common color

Color Brewer, http://colorbrewer2.org



Number of data classes: 3

Nature of your data: ● sequential ○ diverging ○ qualitative

Pick a color scheme:

Multi-hue: | Single hue:

Only show:

□ colorblind safe
□ print friendly
□ photocopy safe

Context:

3-class BuGn

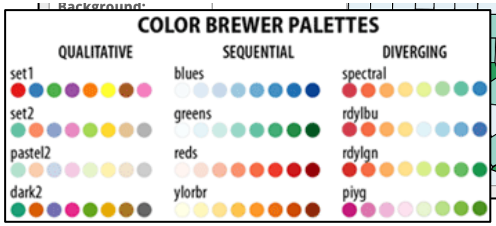
HEX

EXPORT

#e5f5f9
#99d8c9

how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography
Background:

http://mkweb.bcgsc.ca/brewer



Color palettes for color blindness, http://mkweb.bcgsc.ca/colorblind

Where to Get Ideas for Effective Figures?

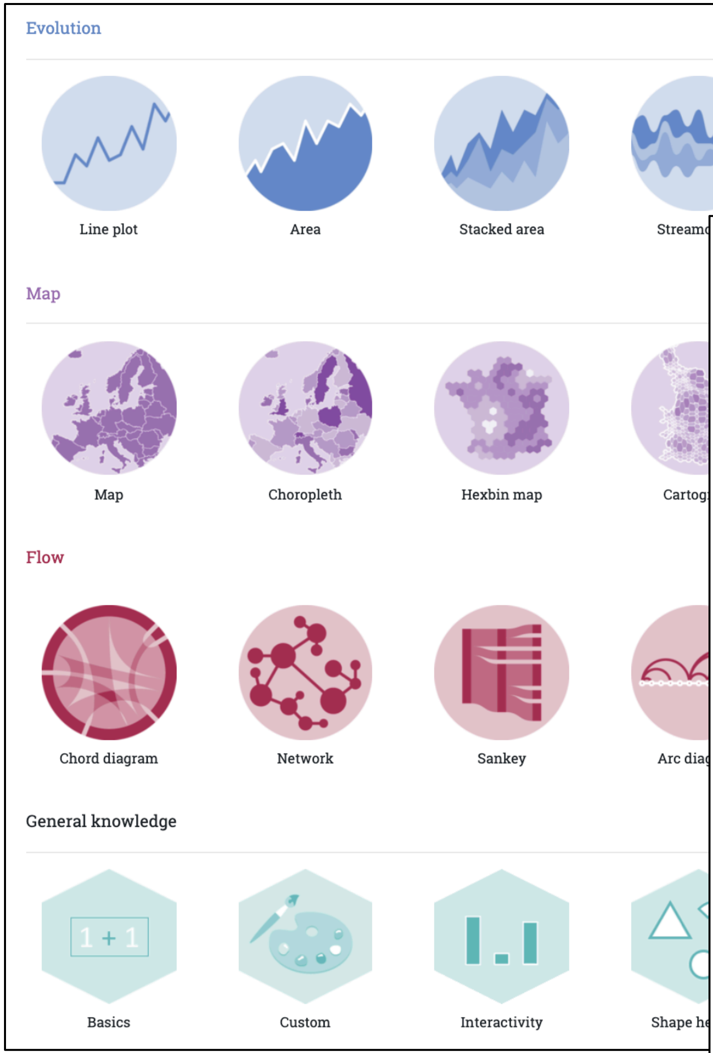
Where to get ideas for figures?

1) Papers published in last issues of Nature, Science, PNAS, Nature Methods, Nature Biotech, etc.
   - No need to read the papers, just look at figures!
2) Martin Krzywinski, mkweb.bcgsc.ca
   - Inventor of several popular visualization tools
   - Designed many Nature, Science, etc. covers
3) www.d3-graph-gallery.com
   - Gallery with hundreds of chart, graphs, geo, part-of-whole
   - Reproducible & editable source code!
4) developers.google.com/chart/interactive/docs/gallery
   - Over 30 chart types, including many non-standard ones
   - Tutorials and source code for every chart type!

Where to get ideas for figures?

www.d3-graph-gallery.com

Many non-standard, but highly effective chart types. Source code!



Evolution

Line plot

Area

Stacked area

Streamg

Map

Map

Choropleth

Hexbin map

Cartog

Flow

Chord diagram

Network

Sankey

Arc diag

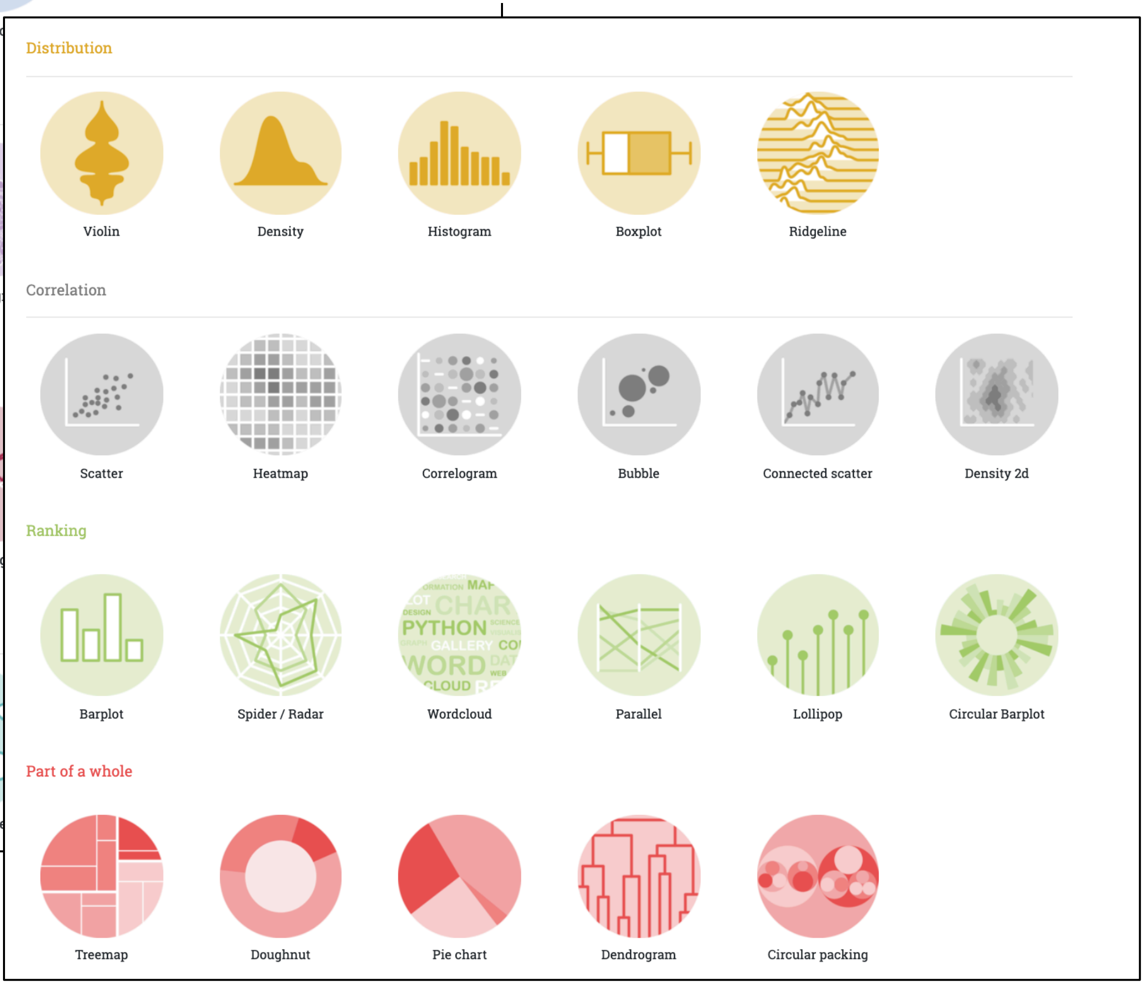
General knowledge

Basics

Custom

Interactivity

Shape he



Distribution

Violin

Density

Histogram

Boxplot

Ridgeline

Correlation

Scatter

Heatmap

Correlogram

Bubble

Connected scatter

Density 2d

Ranking

Barplot

Spider / Radar

Wordcloud

Parallel

Lollipop

Circular Barplot

Part of a whole

Treemap

Doughnut

Pie chart

Dendrogram

Circular packing

Where to get ideas for figures?

https://developers.google.com/chart with source code!

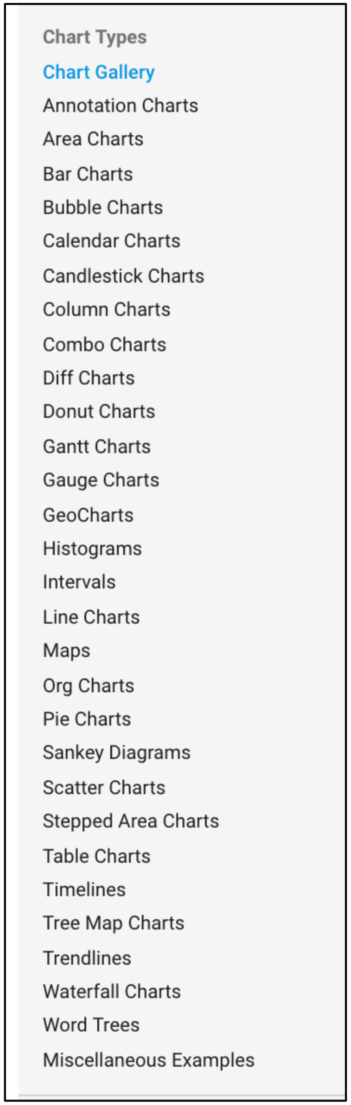


Chart Types

Chart Gallery

Annotation Charts

Area Charts

Bar Charts

Bubble Charts

Calendar Charts

Candlestick Charts

Column Charts

Combo Charts

Diff Charts

Donut Charts

Gantt Charts

Gauge Charts

GeoCharts

Histograms

Intervals

Line Charts

Maps

Org Charts

Pie Charts

Sankey Diagrams

Scatter Charts

Stepped Area Charts

Table Charts

Timelines

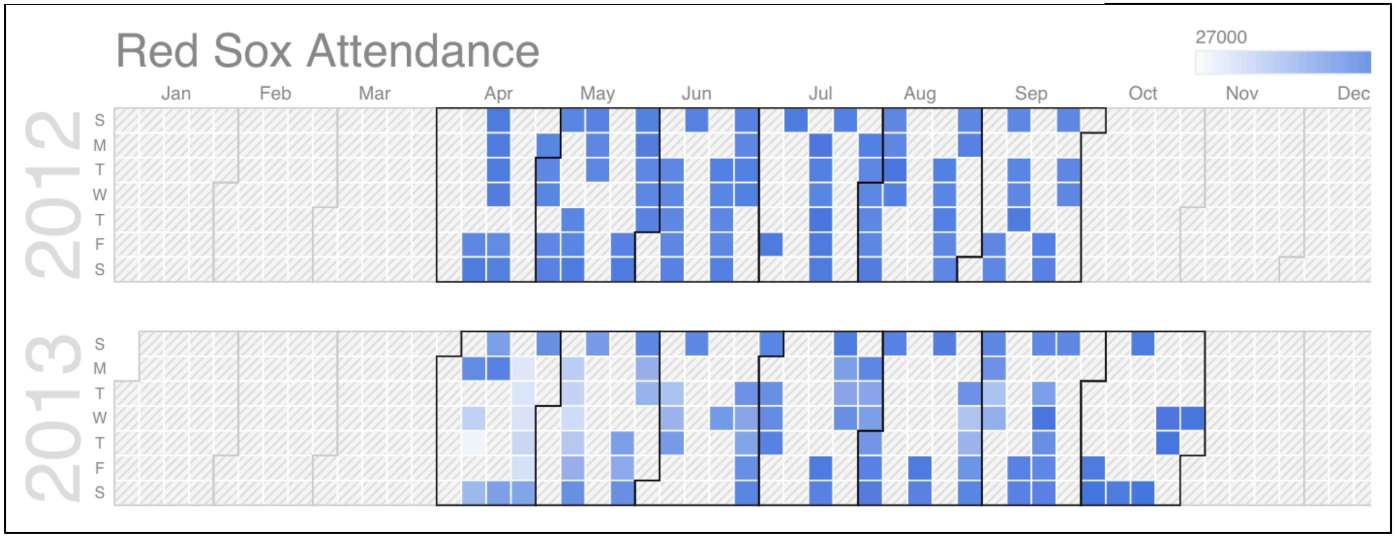
Tree Map Charts

Trendlines

Waterfall Charts

Word Trees

Miscellaneous Examples



Where to get ideas for figures?

https://developers.google.com/chart with source code!

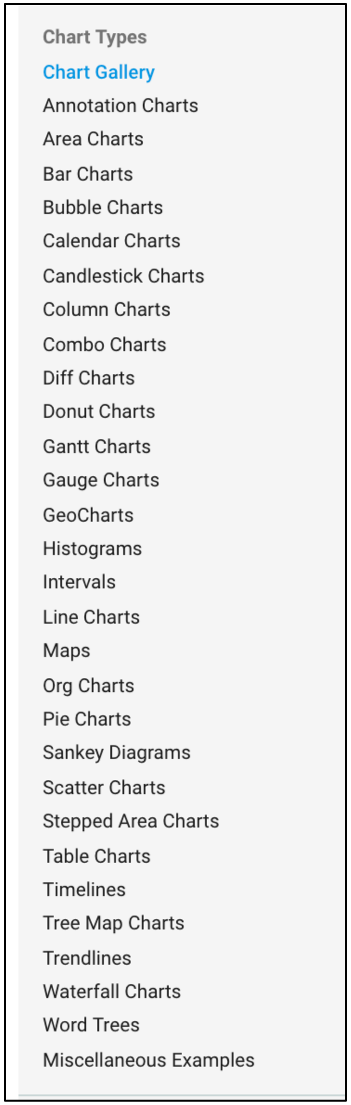


Chart Types

Chart Gallery

Annotation Charts

Area Charts

Bar Charts

Bubble Charts

Calendar Charts

Candlestick Charts

Column Charts

Combo Charts

Diff Charts

Donut Charts

Gantt Charts

Gauge Charts

GeoCharts

Histograms

Intervals

Line Charts

Maps

Org Charts

Pie Charts

Sankey Diagrams

Scatter Charts

Stepped Area Charts

Table Charts

Timelines

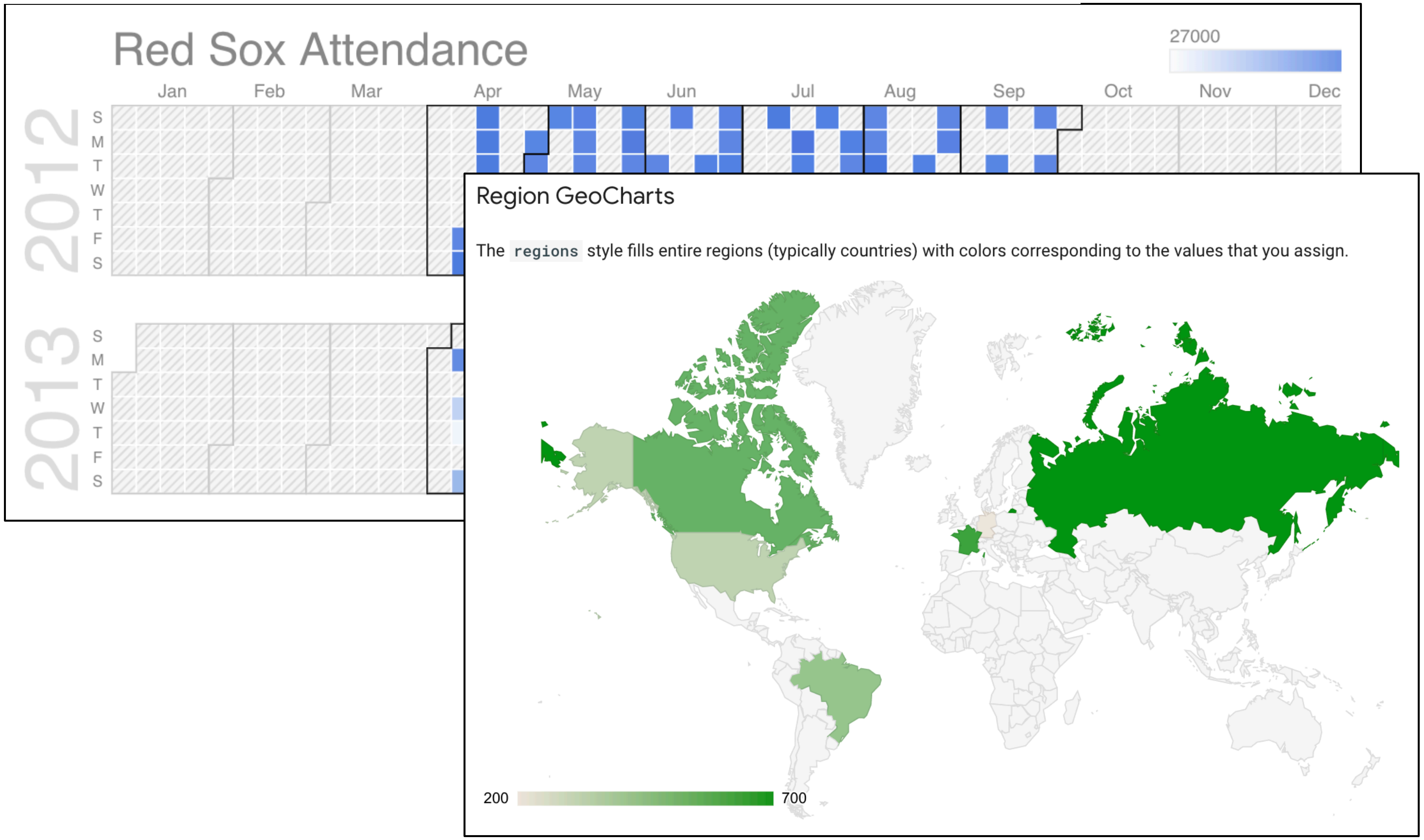
Tree Map Charts

Trendlines

Waterfall Charts

Word Trees

Miscellaneous Examples



Region GeoCharts

The `regions` style fills entire regions (typically countries) with colors corresponding to the values that you assign.

Where to get ideas for figures?

https://developers.google.com/chart with source code!

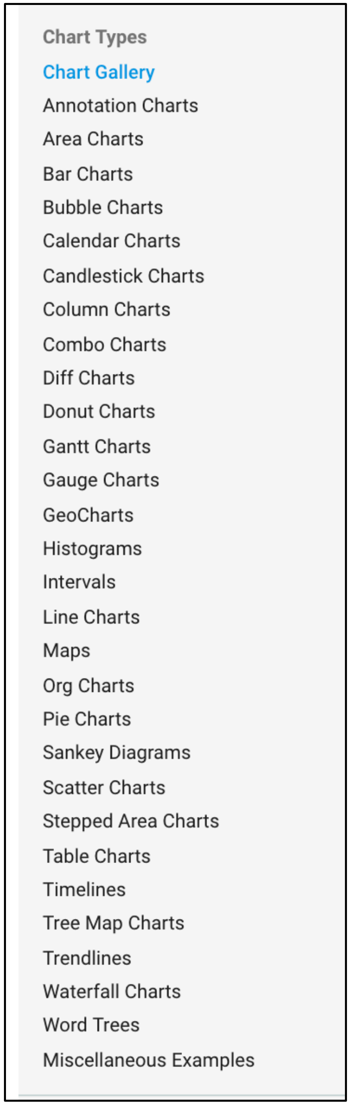
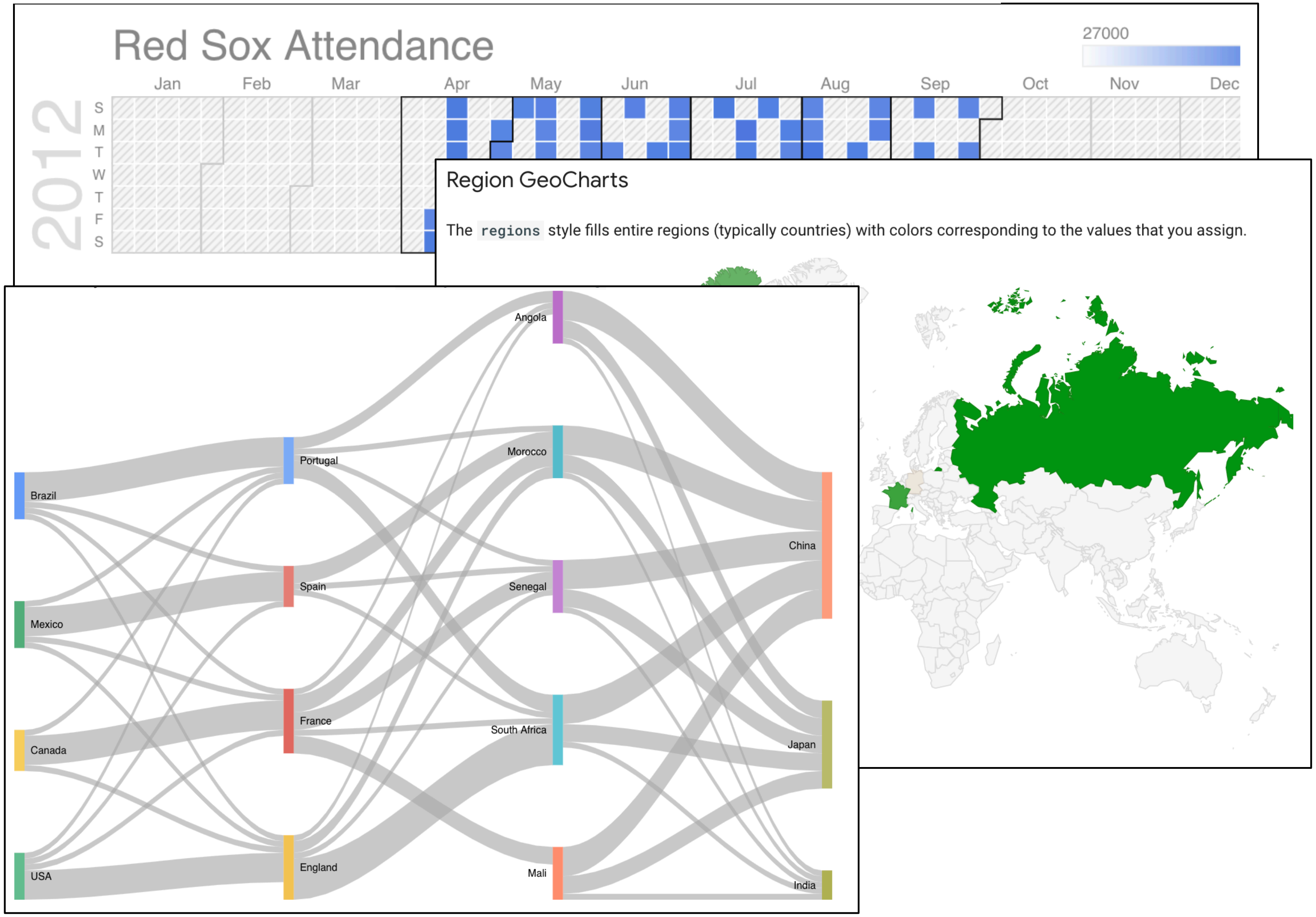


Chart Types
Chart Gallery
Annotation Charts
Area Charts
Bar Charts
Bubble Charts
Calendar Charts
Candlestick Charts
Column Charts
Combo Charts
Diff Charts
Donut Charts
Gantt Charts
Gauge Charts
GeoCharts
Histograms
Intervals
Line Charts
Maps
Org Charts
Pie Charts
Sankey Diagrams
Scatter Charts
Stepped Area Charts
Table Charts
Timelines
Tree Map Charts
Trendlines
Waterfall Charts
Word Trees
Miscellaneous Examples



Region GeoCharts

The `regions` style fills entire regions (typically countries) with colors corresponding to the values that you assign.

Where to get ideas for figures?

https://developers.google.com/chart with source code!

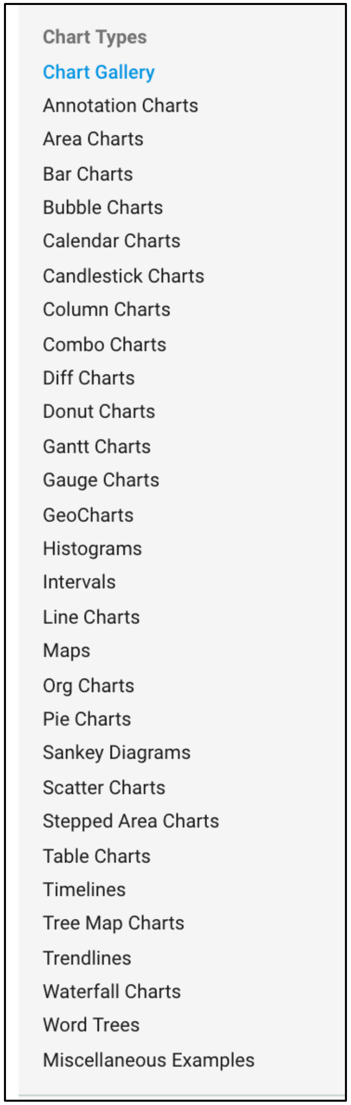
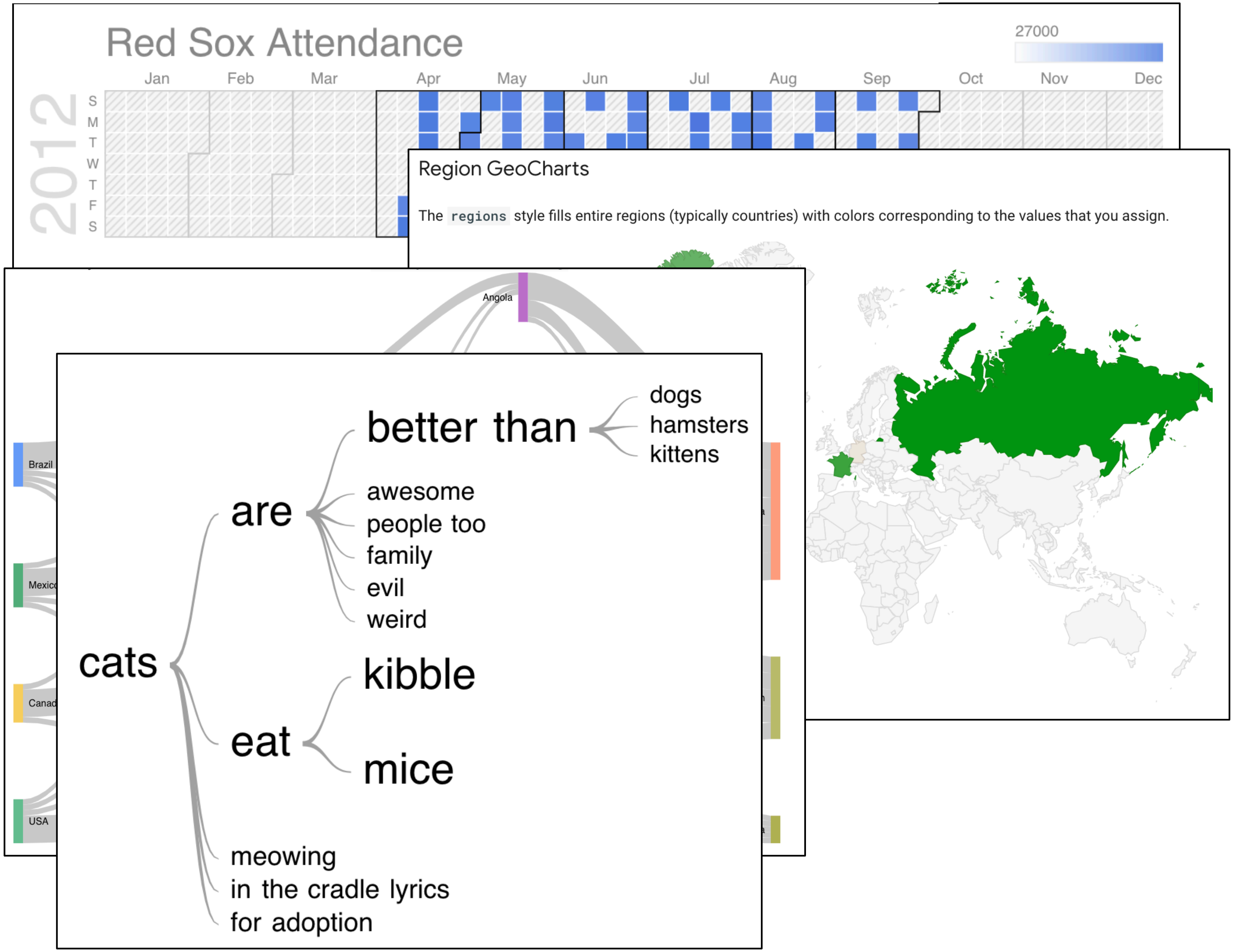


Chart Types
- Chart Gallery
- Annotation Charts
- Area Charts
- Bar Charts
- Bubble Charts
- Calendar Charts
- Candlestick Charts
- Column Charts
- Combo Charts
- Diff Charts
- Donut Charts
- Gantt Charts
- Gauge Charts
- GeoCharts
- Histograms
- Intervals
- Line Charts
- Maps
- Org Charts
- Pie Charts
- Sankey Diagrams
- Scatter Charts
- Stepped Area Charts
- Table Charts
- Timelines
- Tree Map Charts
- Trendlines
- Waterfall Charts
- Word Trees
- Miscellaneous Examples



Region GeoCharts

The `regions` style fills entire regions (typically countries) with colors corresponding to the values that you assign.

Where to get ideas for figures?

https://developers.google.com/chart with source code!

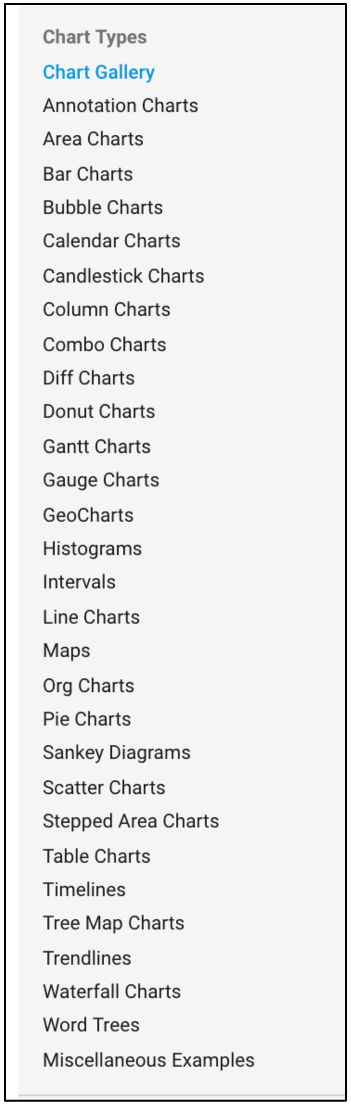


Chart Types

Chart Gallery

Annotation Charts

Area Charts

Bar Charts

Bubble Charts

Calendar Charts

Candlestick Charts

Column Charts

Combo Charts

Diff Charts

Donut Charts

Gantt Charts

Gauge Charts

GeoCharts

Histograms

Intervals

Line Charts

Maps

Org Charts

Pie Charts

Sankey Diagrams

Scatter Charts

Stepped Area Charts

Table Charts

Timelines

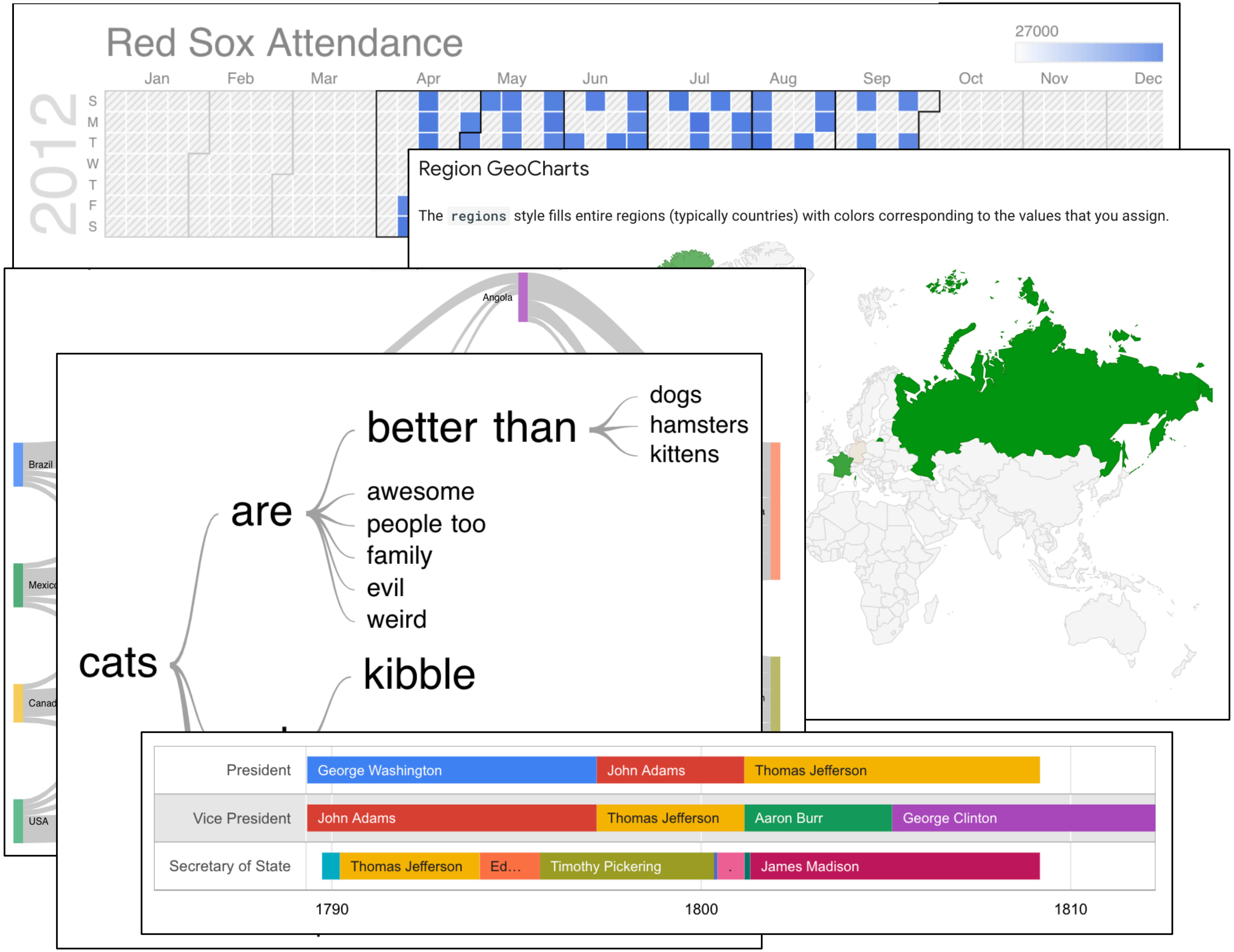
Tree Map Charts

Trendlines

Waterfall Charts

Word Trees

Miscellaneous Examples



Region GeoCharts

The `regions` style fills entire regions (typically countries) with colors corresponding to the values that you assign.

Today's Lecture

1) Why figures matter ✓
2) Figures in science ✓
3) How to design effective figures ✓
4) Tools, tips, and guidelines ✓

Three Takeaway Messages

1) Figures are often the first part of research papers examined by editors and your peers

2) Well-designed figures convey facts, ideas, and relationships far more clearly/concisely than text

3) Focus on effectively conveying complex information rather than on attention-getting decoration

Break

Your Turn ☺

Visualization Prototyping Activity

- Goal: Design and prototype the key figure(s) for your project!
- Take a piece of paper and draw out what you might see in your data and how you'd like to communicate [5 min]
  - What do you want to communicate?
  - Who is your audience?
  - Prototype and iterate! Great figures take many iterations.
- Iterate with your partner [10 min]
  - Could you communicate the same idea more concisely and effectively?
  - How could you impress your audience?
- Share your insights in class [10 min]

Visualization Lab