Global Big Data Conference

www.globalbigdataconference.com

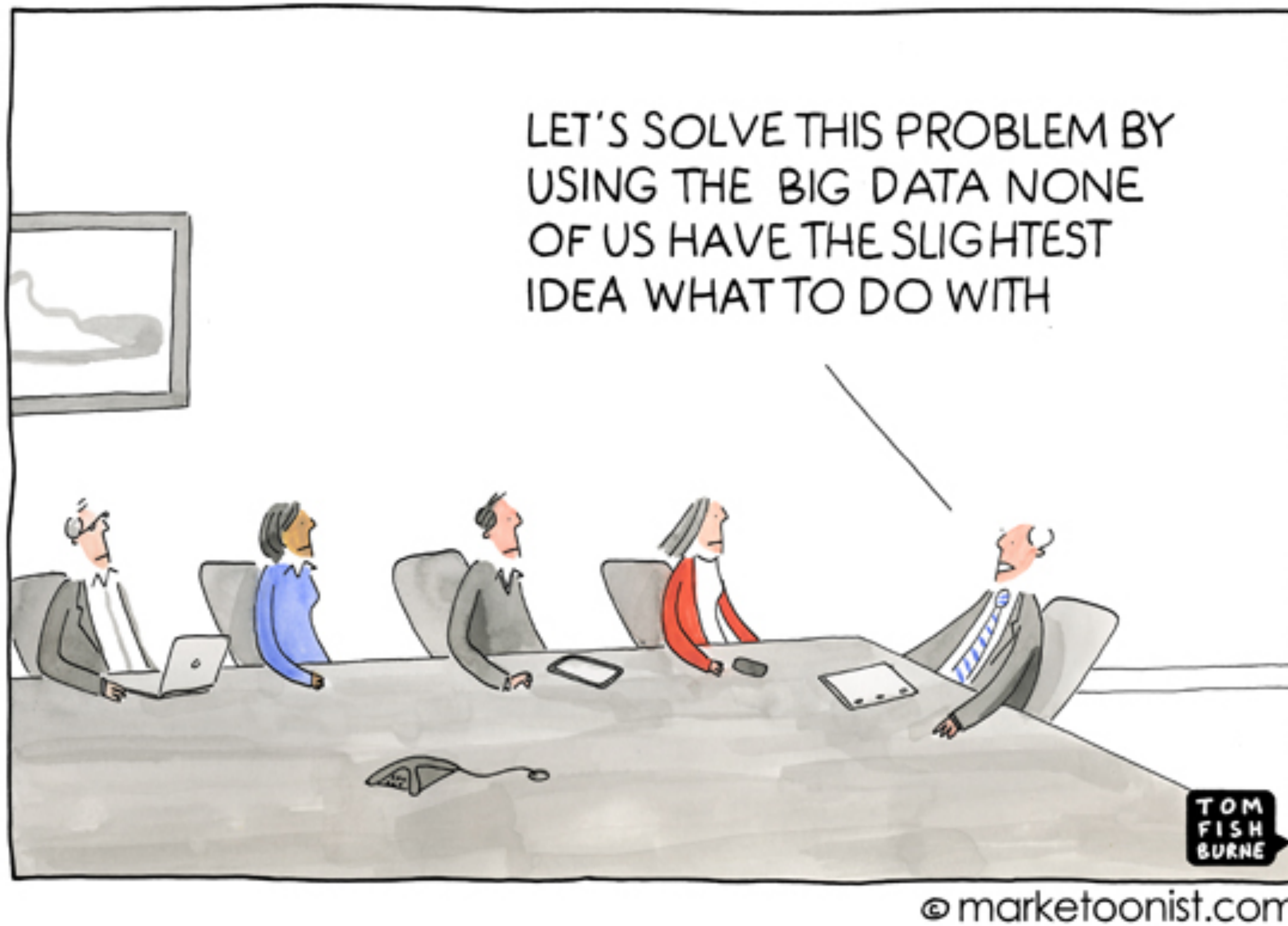
Twitter : @bigdataconf

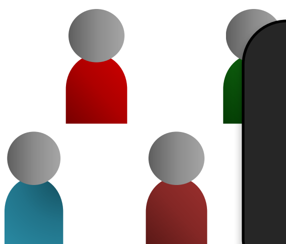
Enabling a dialog between **People** & **Data**

Lessons in Designing for Big Data

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Danielle Albers Szafir
Assistant Professor
Department of Information Science
Danielle.Szafir@Colorado.edu


LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com



Why don't we just compute the answer?

More Analysts
(& Fewer Experts)

More Heterogeneity

More Questions

Data Sample 1:

Mean(x) = 9

Variance(x) = 11

Correlation(x, y) = 0.816

Regression: y = 3 + 0.5x

Data Sample 2:

Mean(x) = 9

Variance(x) = 11

Correlation(x, y) = 0.816

Regression: y = 3 + 0.5x

Data Sample 3:

Mean(x) = 9

Variance(x) = 11

Correlation(x, y) = 0.816

Regression: y = 3 + 0.5x

Data Sample 4:

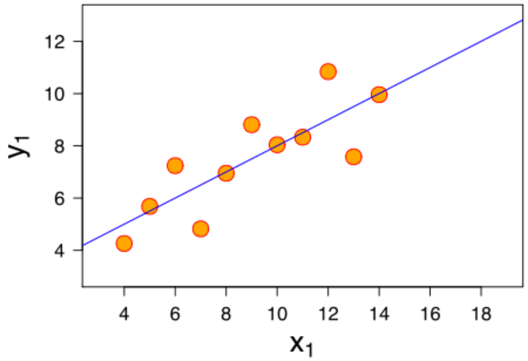
Mean(x) = 9

Variance(x) = 11

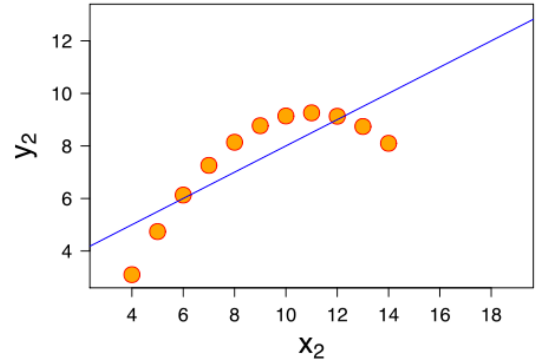
Correlation(x, y) = 0.816

Regression: y = 3 + 0.5x

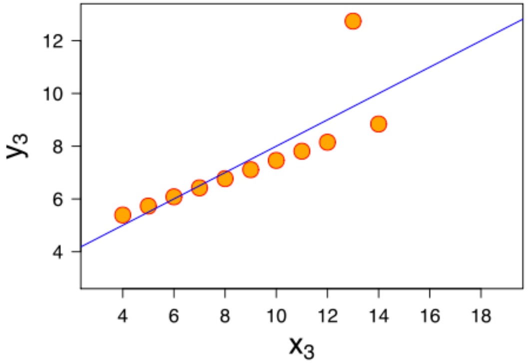
Data Sample 1:



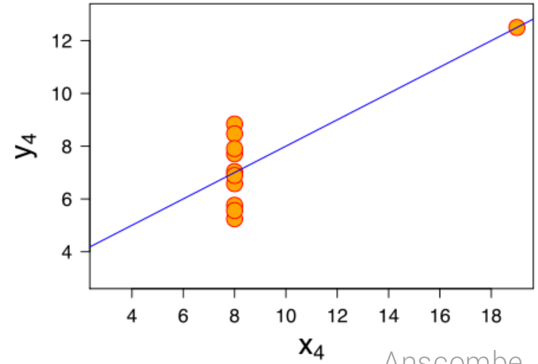
Data Sample 2:



Data Sample 3:

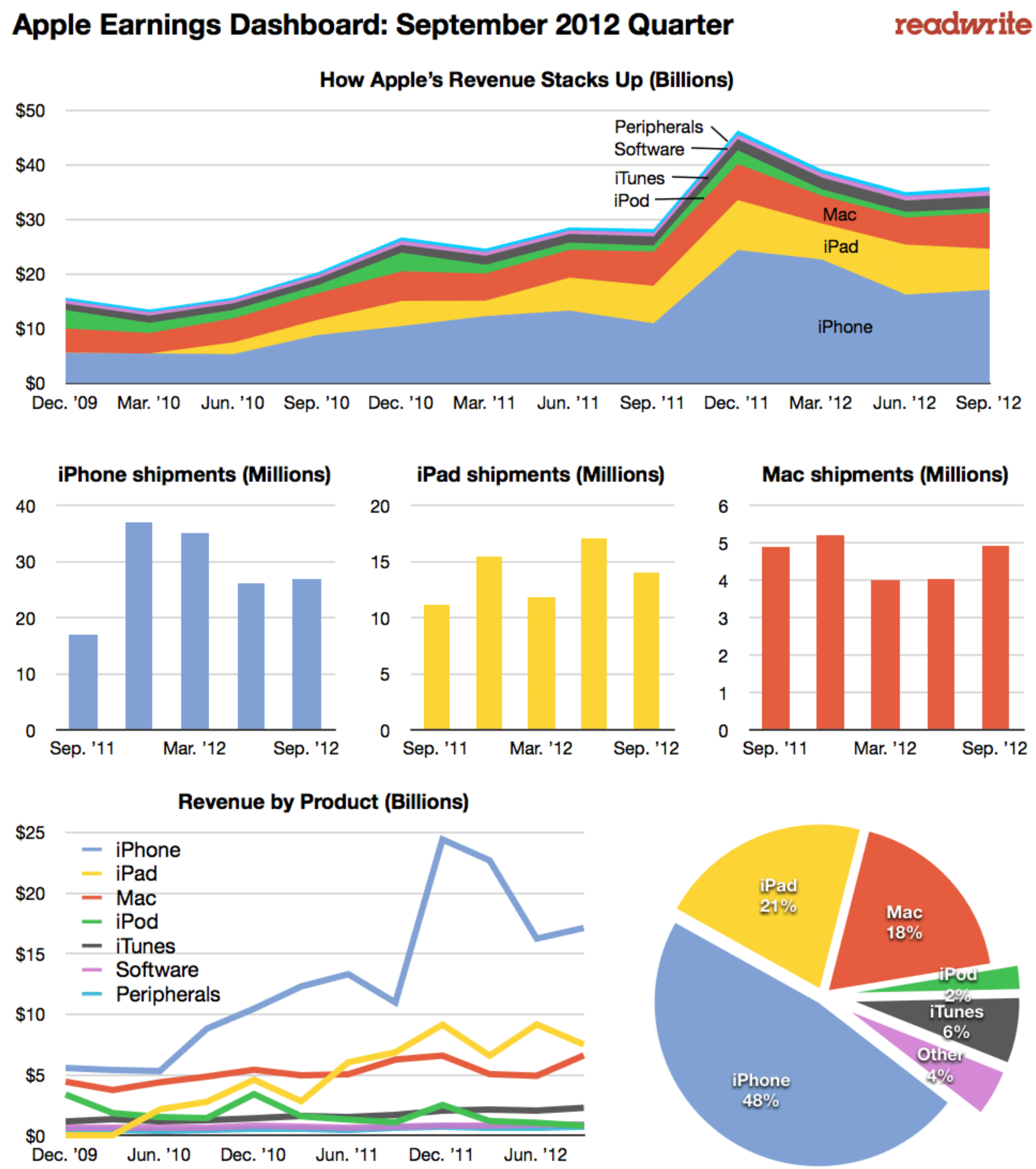


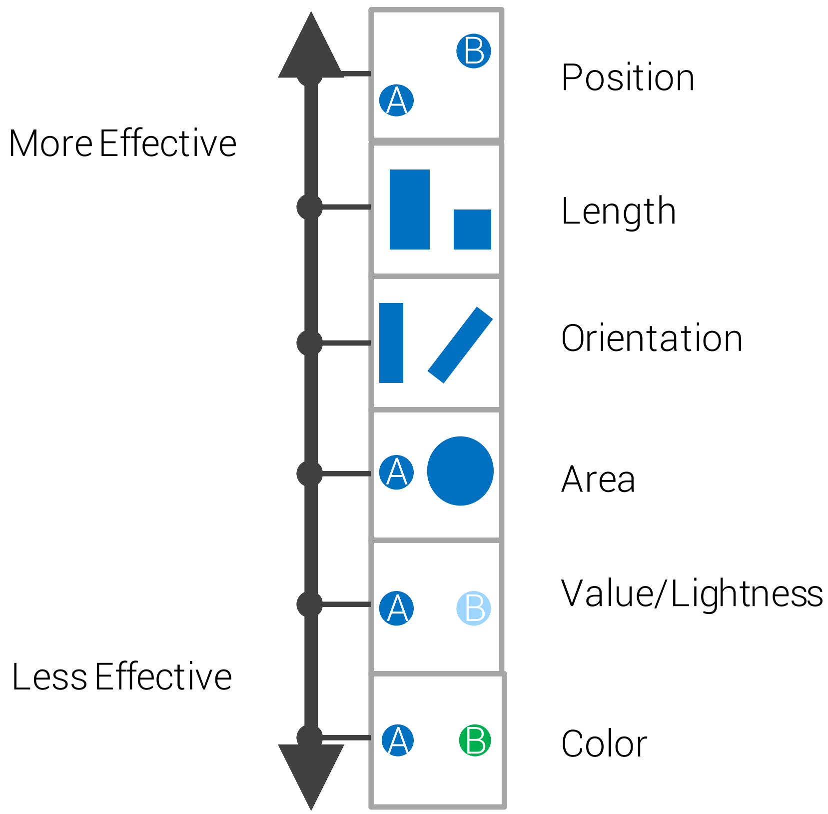
Data Sample 4:



Anscombe, American Statistician, 1973

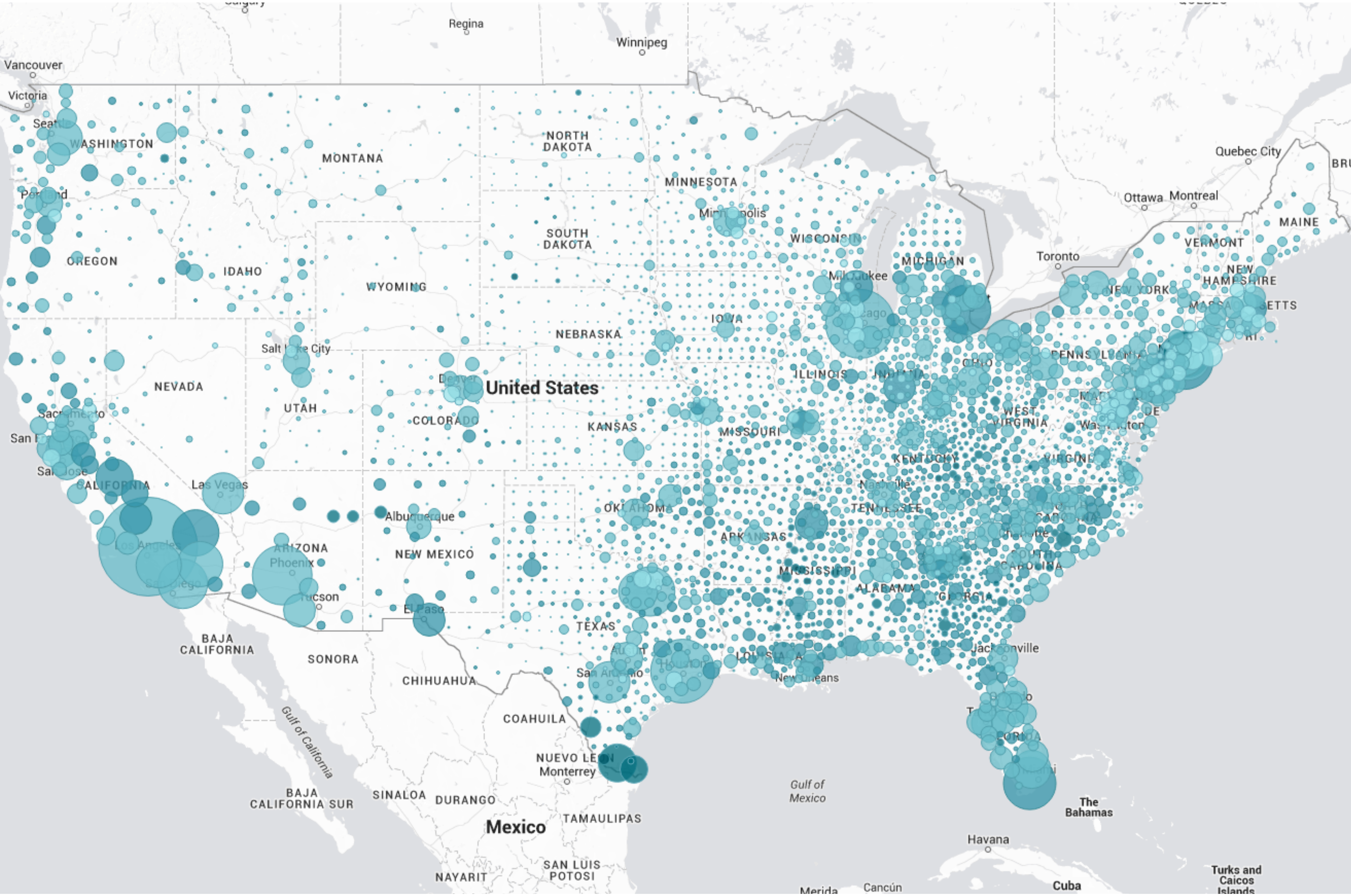
Statistical tools are powerful, but the human brain understands patterns


Apple Earnings Dashboard: September 2012 Quarter
readwrite
How Apple's Revenue Stacks Up (Billions)
$50
$40
$30
$20
$10
$0
Peripherals
Software
iTunes
iPod
Mac
iPad
iPhone
Dec. '09
Mar. '10
Jun. '10
Sep. '10
Dec. '10
Mar. '11
Jun. '11
Sep. '11
Dec. '11
Mar. '12
Jun. '12
Sep. '12
iPhone shipments (Millions)
40
30
20
10
0
Sep. '11
Mar. '12
Sep. '12
iPad shipments (Millions)
20
15
10
5
0
Sep. '11
Mar. '12
Sep. '12
Mac shipments (Millions)
6
5
4
3
2
1
0
Sep. '11
Mar. '12
Sep. '12
Revenue by Product (Billions)
$25
$20
$15
$10
$5
$0
iPhone
iPad
Mac
iPod
iTunes
Software
Peripherals
Dec. '09
Jun. '10
Dec. '10
Jun. '11
Dec. '11
Jun. '12
iPad 21%
Mac 18%
iPod 2%
iTunes 6%
Other 4%
iPhone 48%



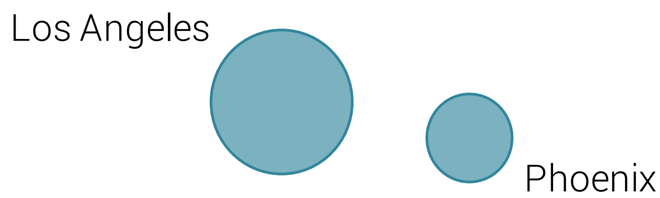
Cleveland & McGill, 1985

What happens when our tools don’t suit our data?



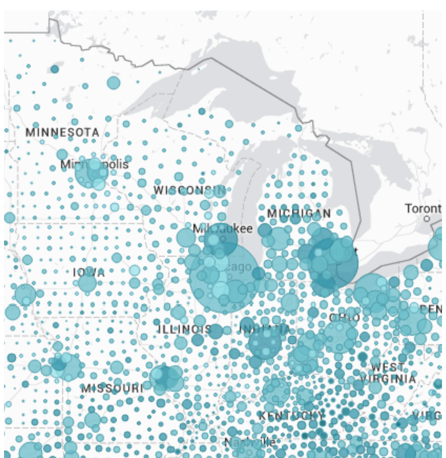
http://www.nytimes.com/newsgraphics/2014/01/05/poverty-map/?ref=multimedia

Low-Level Tasks → Individual Values

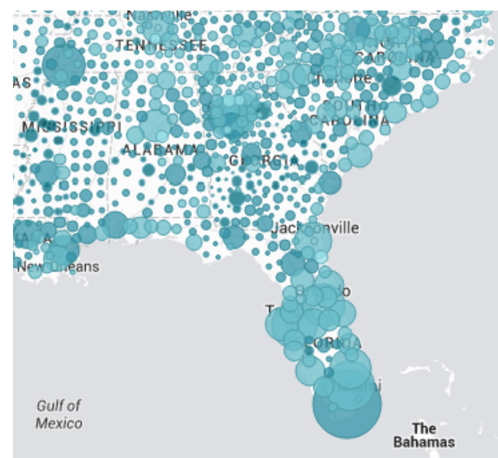


High-Level Tasks → Combine Many Values

Midwest



Southeast



Journal of Vision (2016) 16(5):11, 1–19 1

Four types of ensemble coding in data visualizations

Danielle Albers Szafir Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA

Steve Haroz Department of Psychology, Northwestern University, Evanston, IL, USA

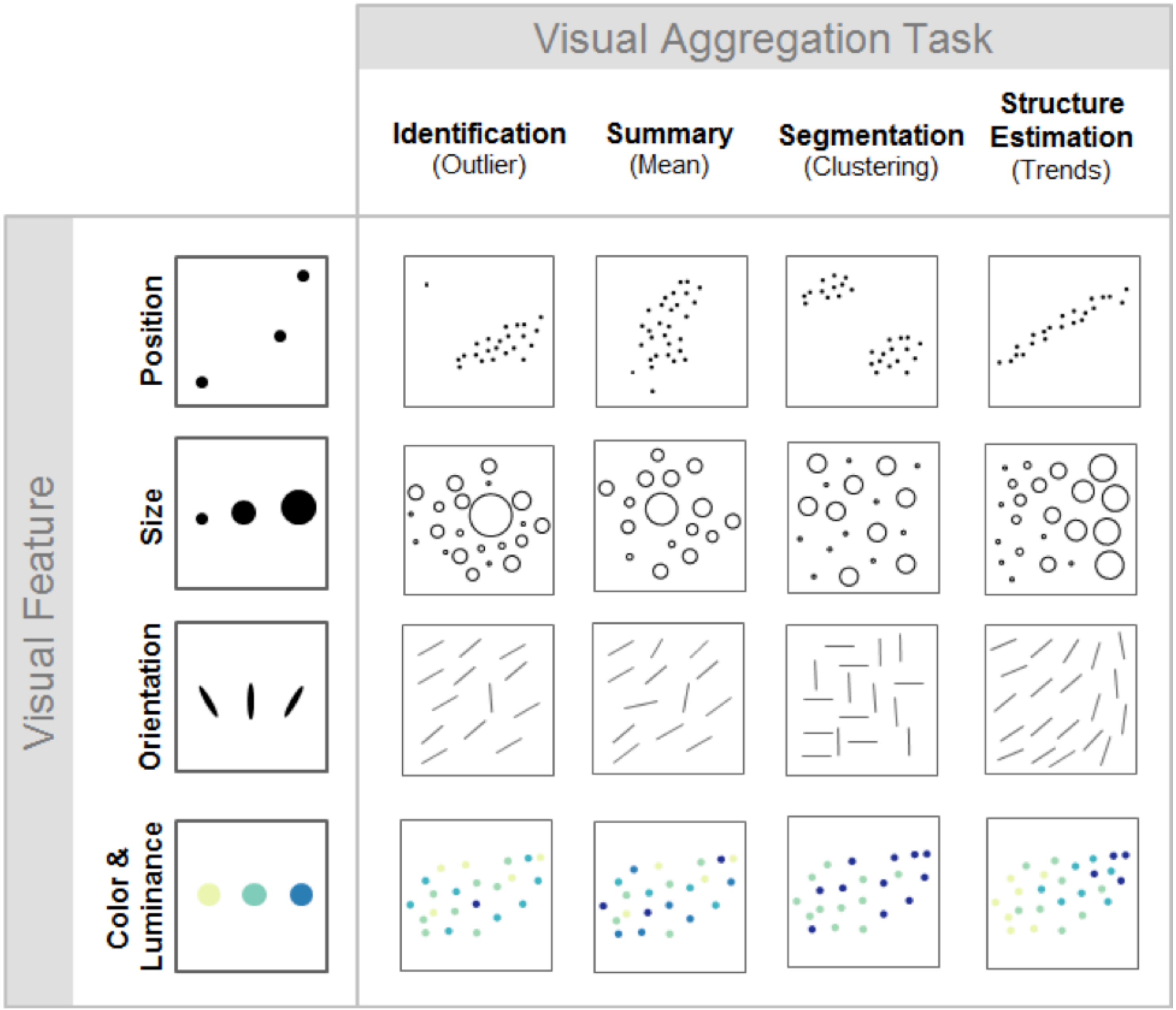
Michael Gleicher Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA

Steven Franconeri Department of Psychology, Northwestern University, Evanston, IL, USA

Ensemble coding supports rapid extraction of visual statistics about distributed visual information. Researchers typically study this ability with the goal of drawing conclusions about how such coding extracts

Kahn, 2012). Other types of information can be extracted and combined in parallel from large numbers of objects at once, such as the average object size (Ariely, 2001). A growing body of work seeks to

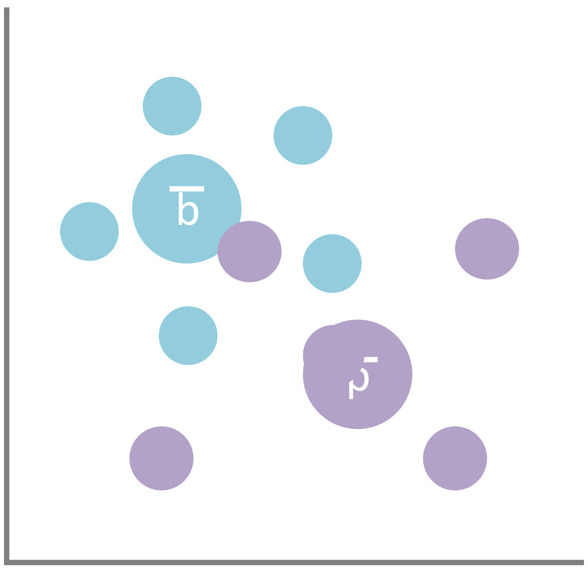
Binary Comparisons don't scale!



Visual Aggregation Task
Identification
(Outlier)
Summary
(Mean)
Segmentation
(Clustering)
Structure
Estimation
(Trends)
Visual Feature
Position
Size
Orientation
Color &
Luminance

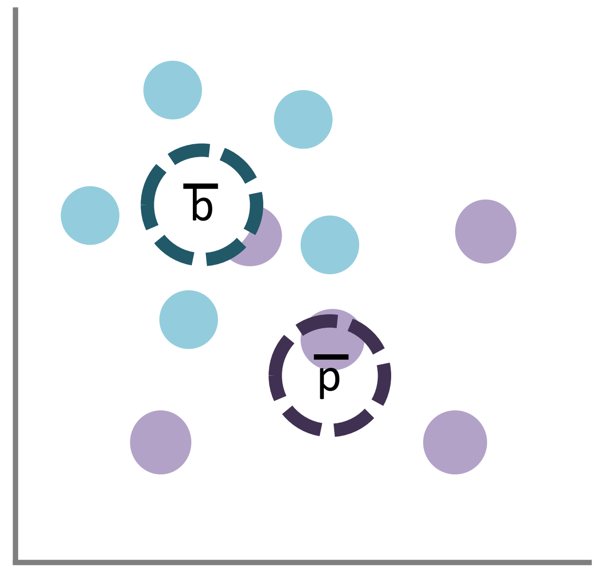
Big Picture Analyses

Computational Aggregation:



Compute the answer then visualize it

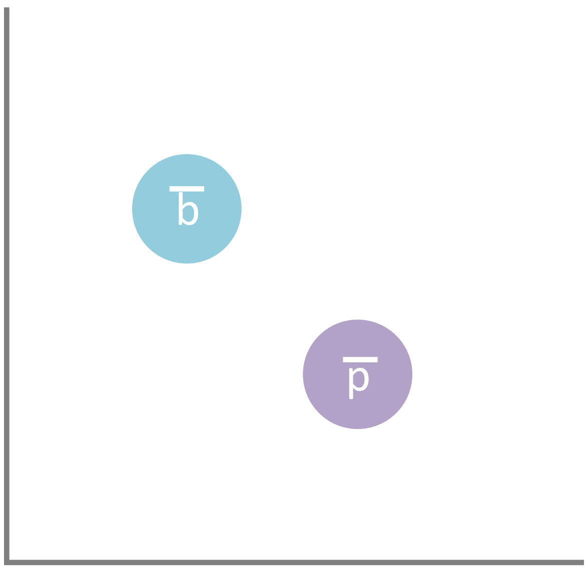
Visual Aggregation:



Use the visual system to estimate the answer

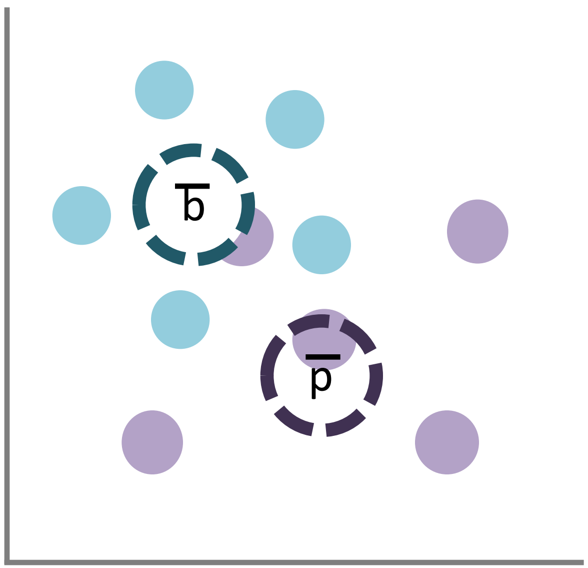
Big Picture Analyses

Computational Aggregation:



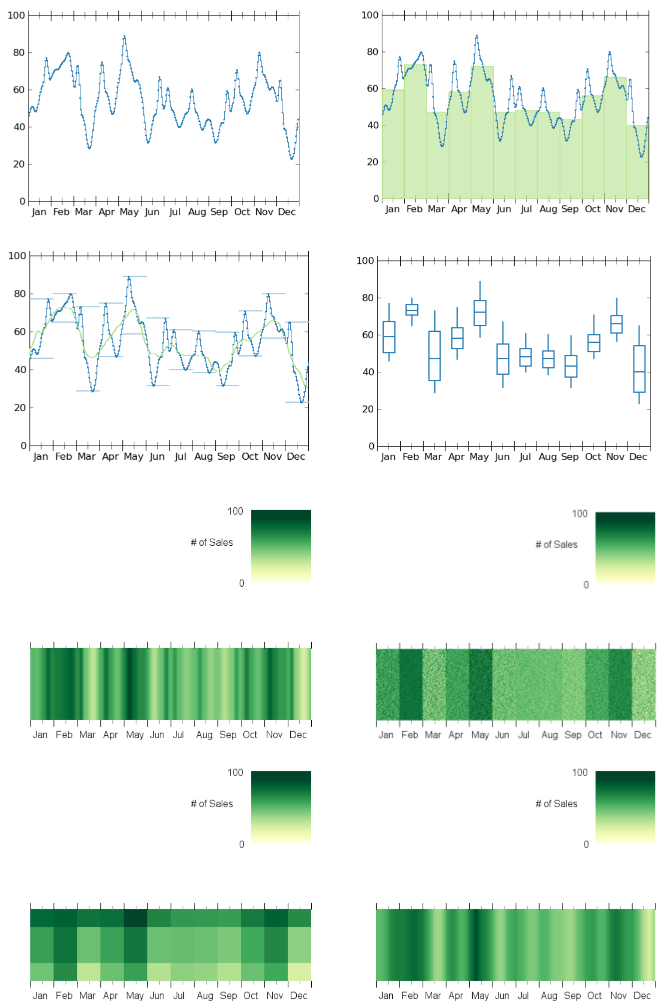
Compute the answer then visualize it

Visual Aggregation:



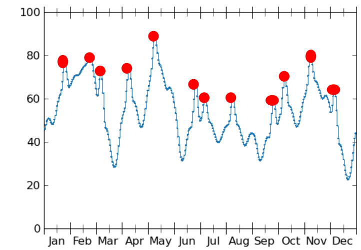
Use the visual system to estimate the answer

Encodings

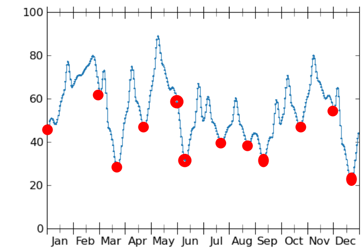


X

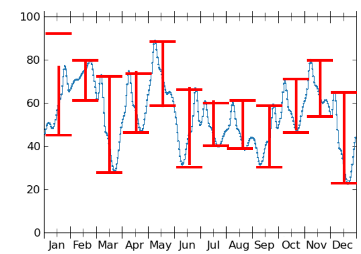
Tasks



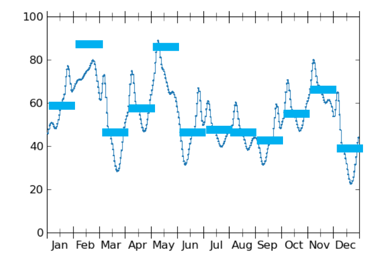
Maxima



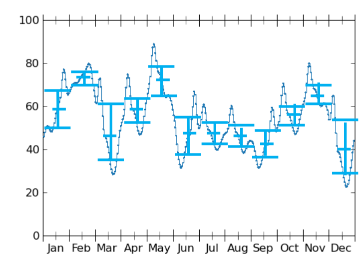
Minima



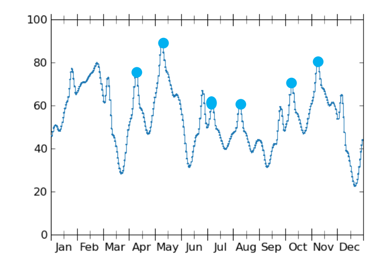
Range



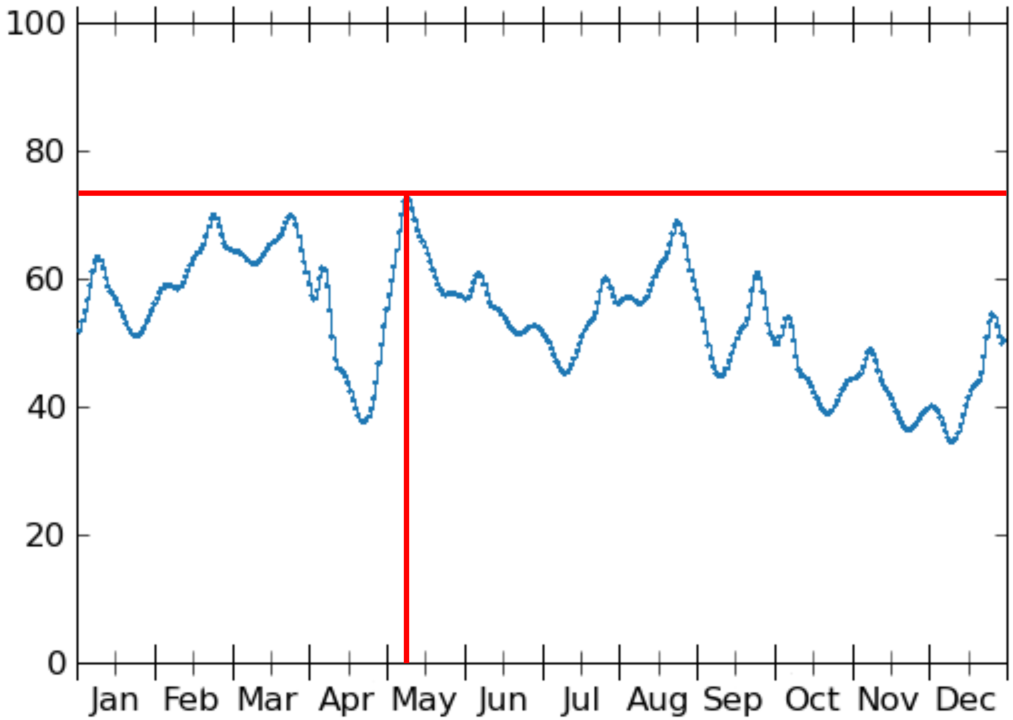
Average



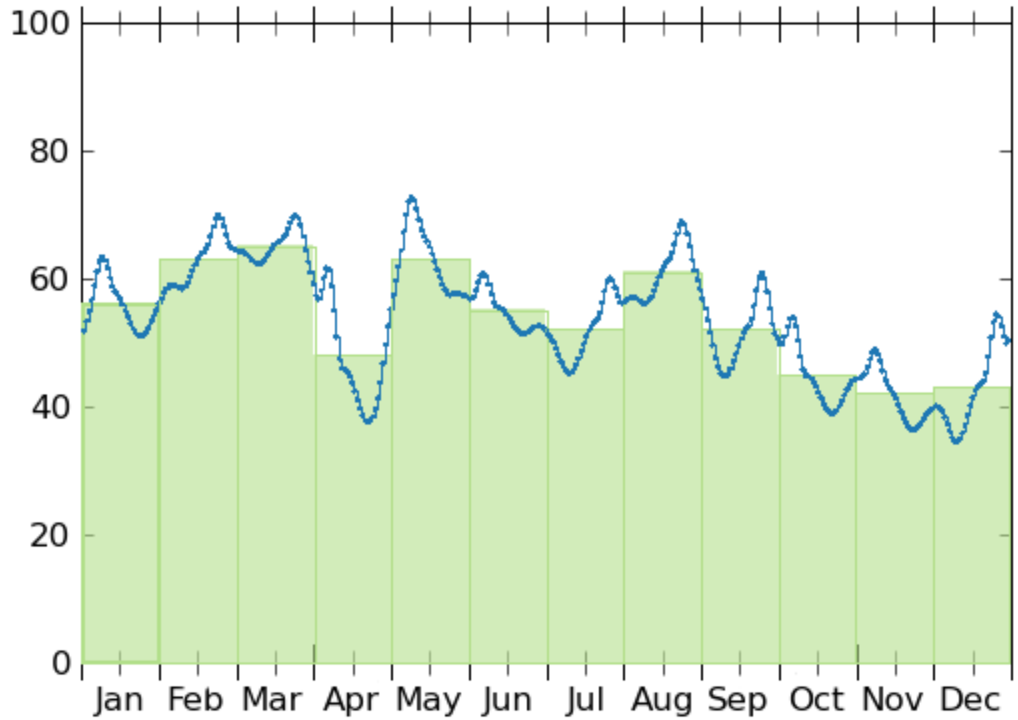
Variance



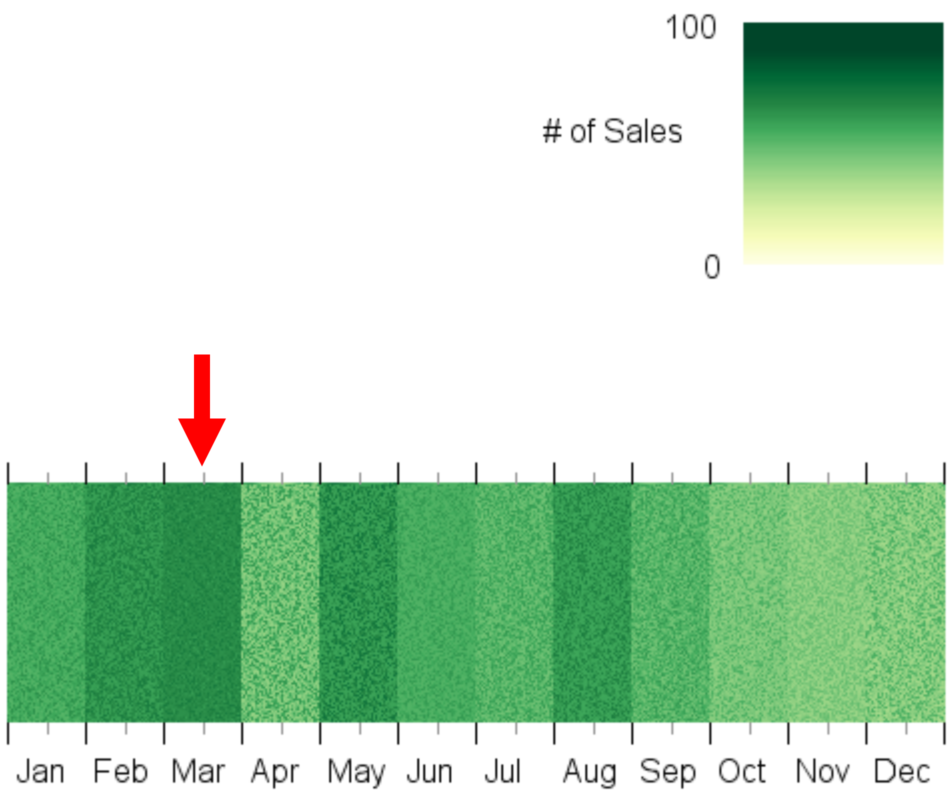
Outliers



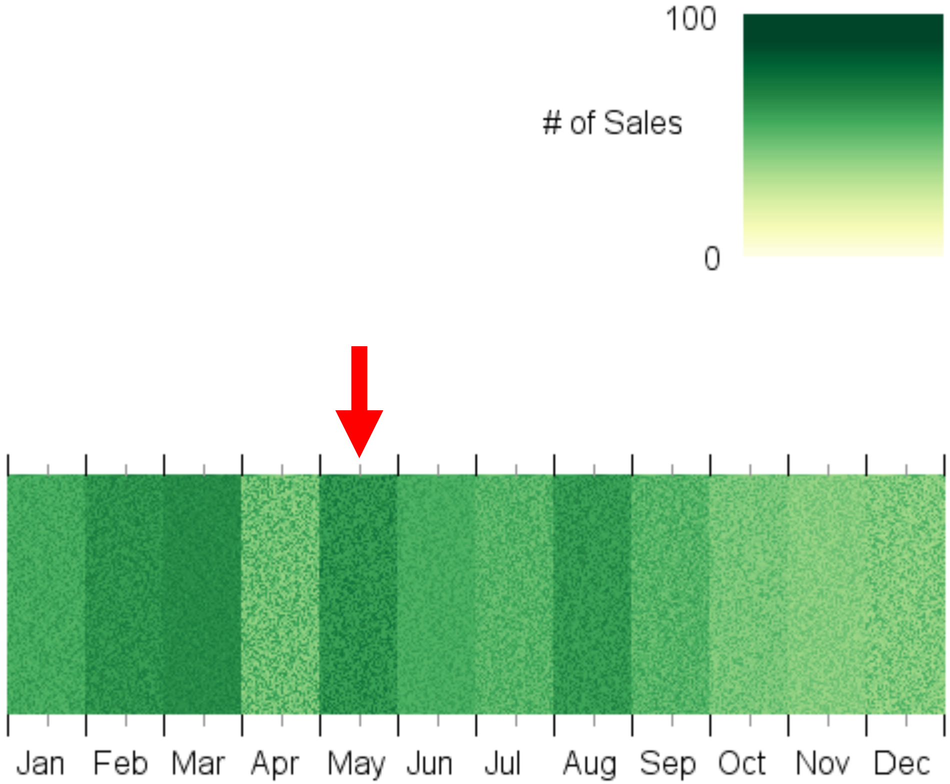
What month has the **highest sales day**?



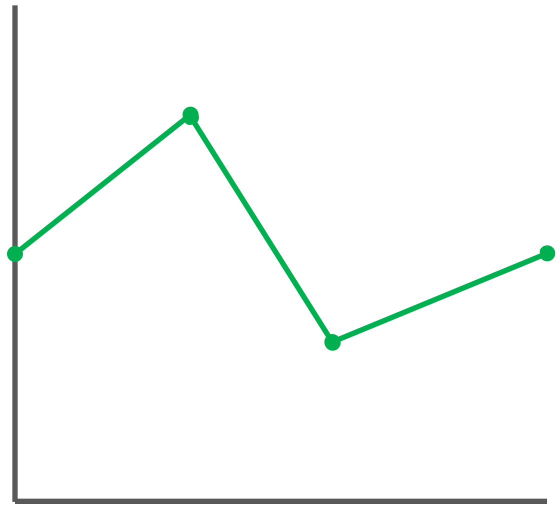
What month has the **highest sales on average**?



What month has the **highest sales on average**?



What month has the **highest sales day**?



Position for
Point Tasks

Color for
Summary Tasks

How you map the data impacts what information is readily extracted

Can we design better visualization systems that do support these analyses?

Two Challenges for Visualization

Scalability

How can we support insight across larger numbers and higher complexity?

Comprehensibility

How can we ensure estimates from a visualization are accurate?

Visualization in the Age of Big Data

Understand limits in current tools

Large Scale Sequence Alignment

Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules

Visualization in the Age of Big Data

Understand limits in current tools

What does the data look like?

Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules

ACGTTT CGATGC TGCCTC AAGCTA CAACGA

Organism

CGATGC ACGTTT TGCATA CAACGA CGATGC

Population

TGCATA ACGTTT CGATGC AAGCTA CGATGC

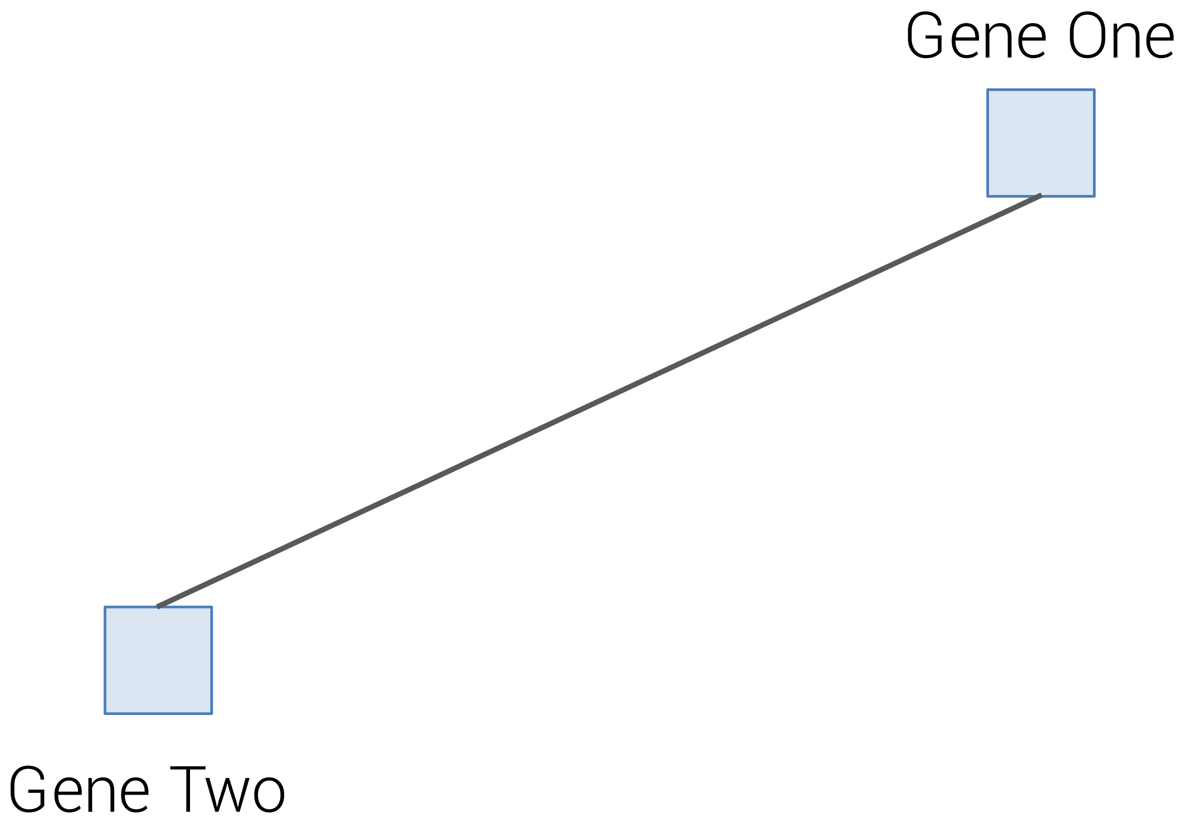
TGCCTC CAACGA ACGTTT AAGGAA CGATGC

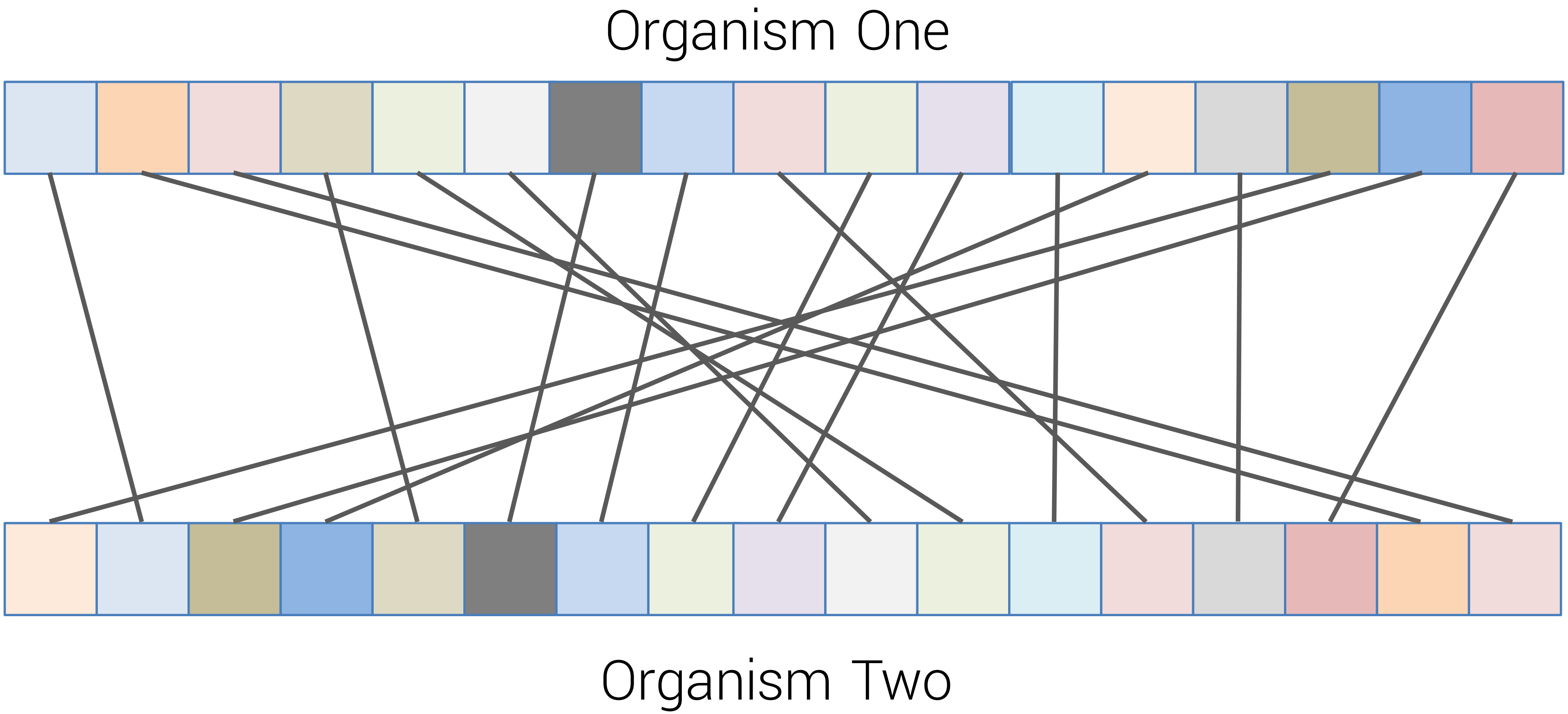
CGATGC CAACGA CGATGC AAGCTA ACGTTT

Albers, Dewey, & Gleicher, 2011

ACGTTT	CGATGC	TGCCTC	AAGCTA	CAACGA
CGATGC	ACGTTT	TGCATA	CAACGA	CGATGC
TGCATA	ACGTTT	CGATGC	AAGCTA	CGATGC
TGCCTC	CAACGA	ACGTTT	AAGGAA	CGATGC
CGATGC	CAACGA	CGATGC	AAGCTA	ACGTTT

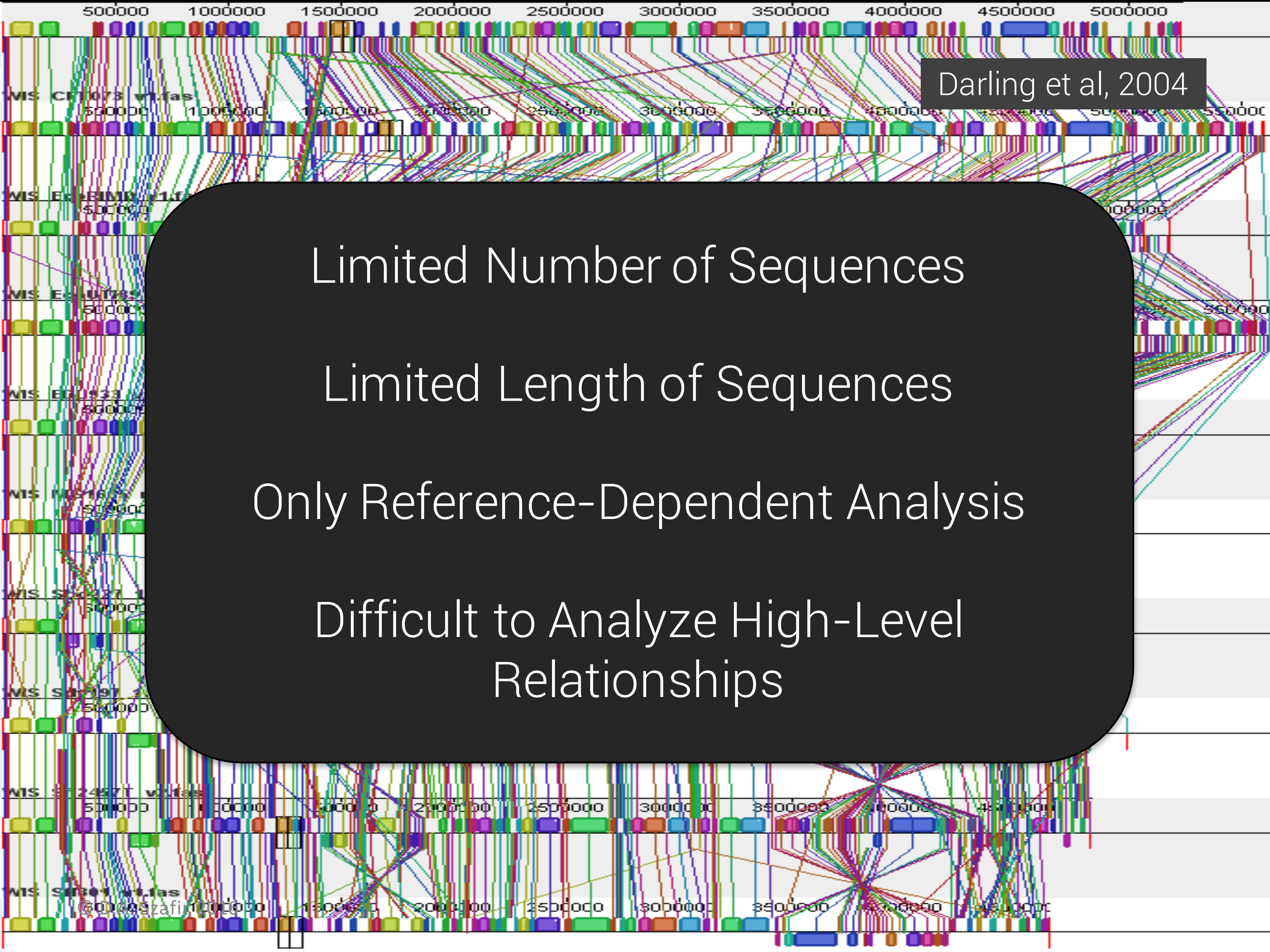
Albers, Dewey, & Gleicher, 2011


Gene One
Gene Two



Albers, Dewey, & Gleicher, 2011

Albers, Dewey, & Gleicher, 2011



Darling et al, 2004

Limited Number of Sequences

Limited Length of Sequences

Only Reference-Dependent Analysis

Difficult to Analyze High-Level Relationships

Visualization in the Age of Big Data

Understand limits in current tools

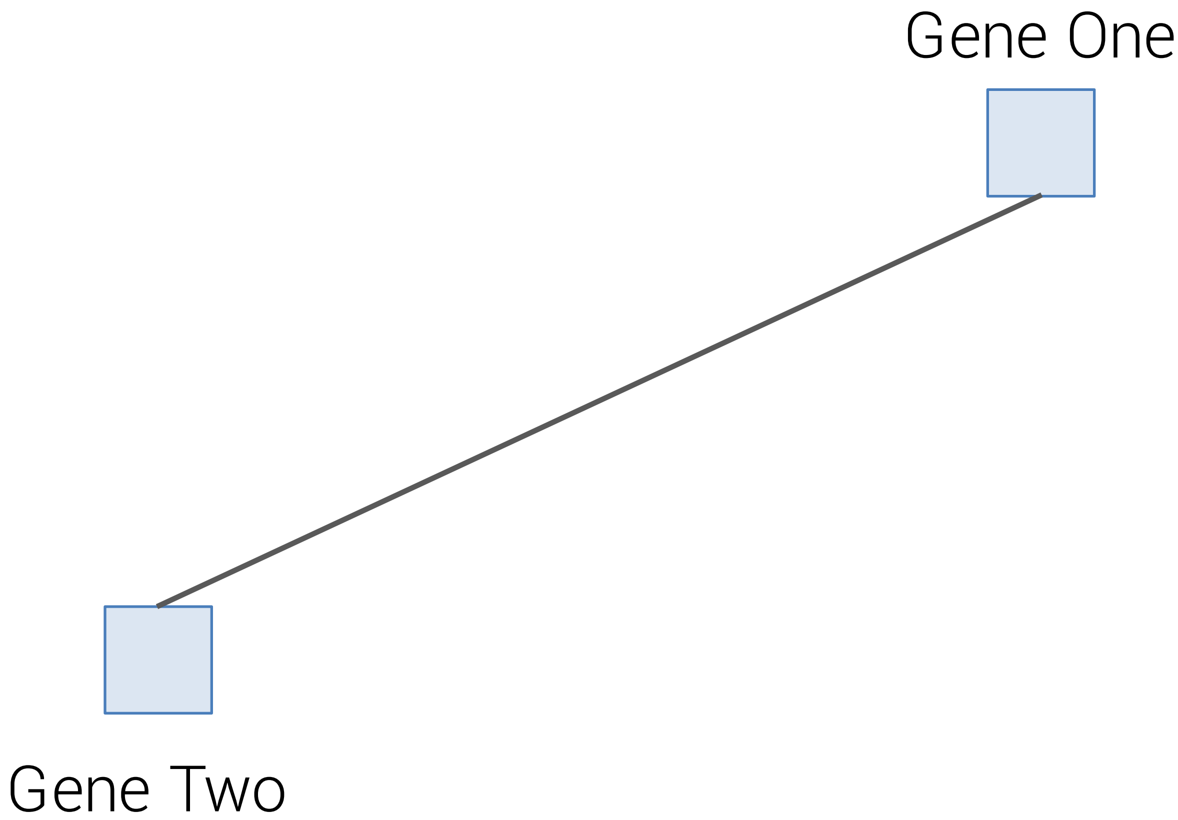
What does the data look like?
The Fix: Aligning patterns with tasks

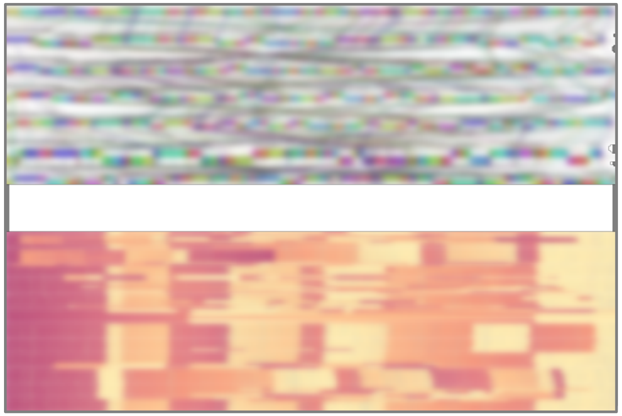
Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules


Gene One
Gene Two



Summarization

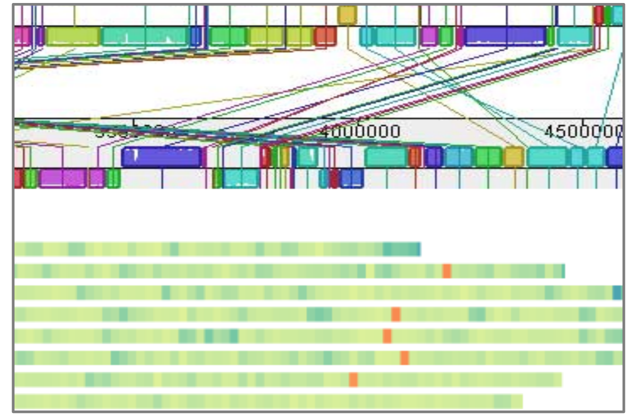
Pop-Out

Visual Search

Visual Clutter

Color better supports visual processing at scale

Summarization



Pop-Out

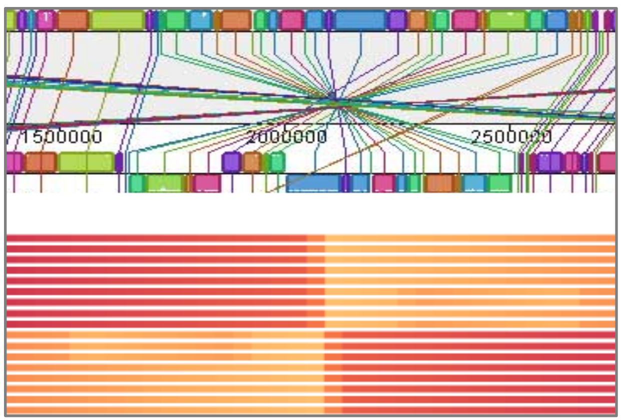
Visual Search

Visual Clutter

Color better supports visual processing at scale

Summarization

Pop-Out



Visual Search

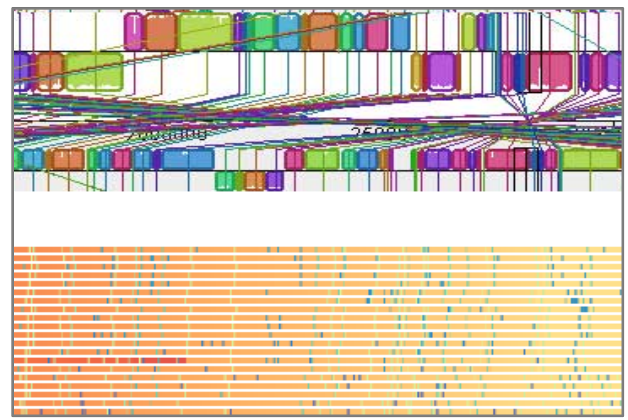
Visual Clutter

Color better supports visual processing at scale

Summarization

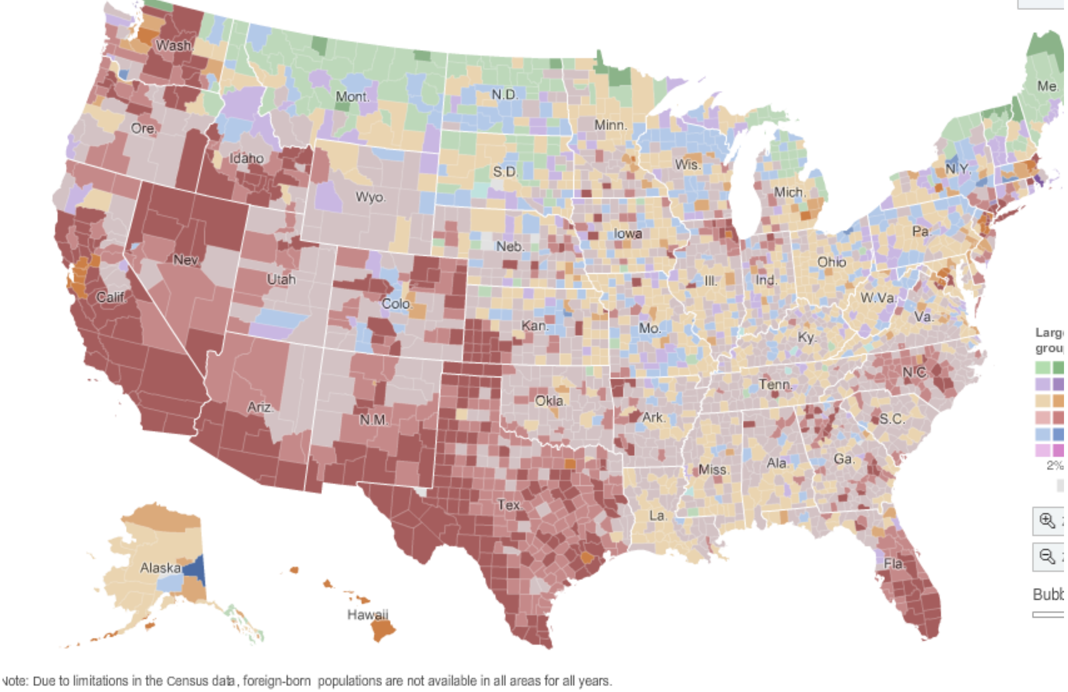
Pop-Out

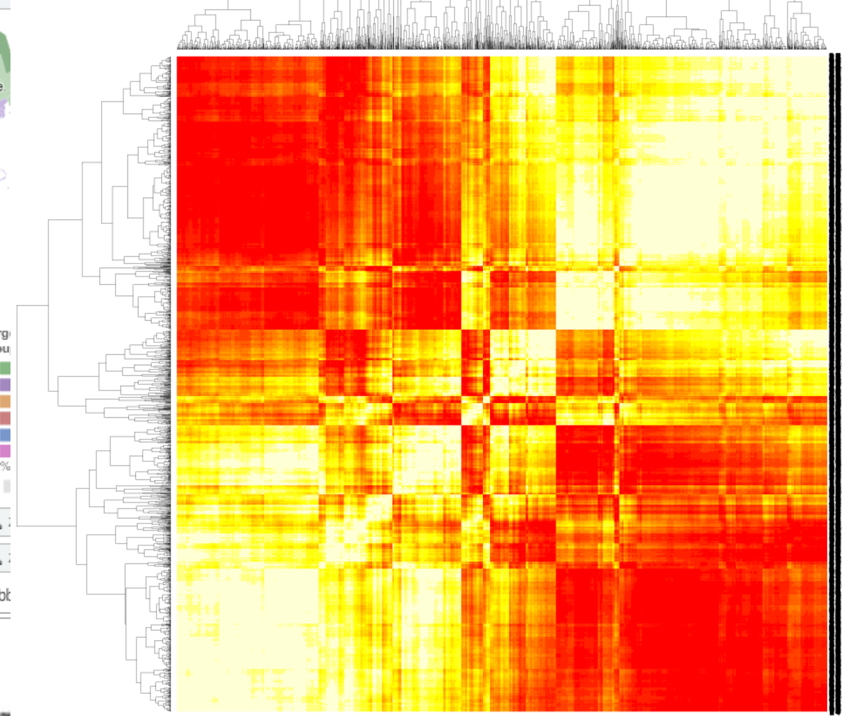
Visual Search

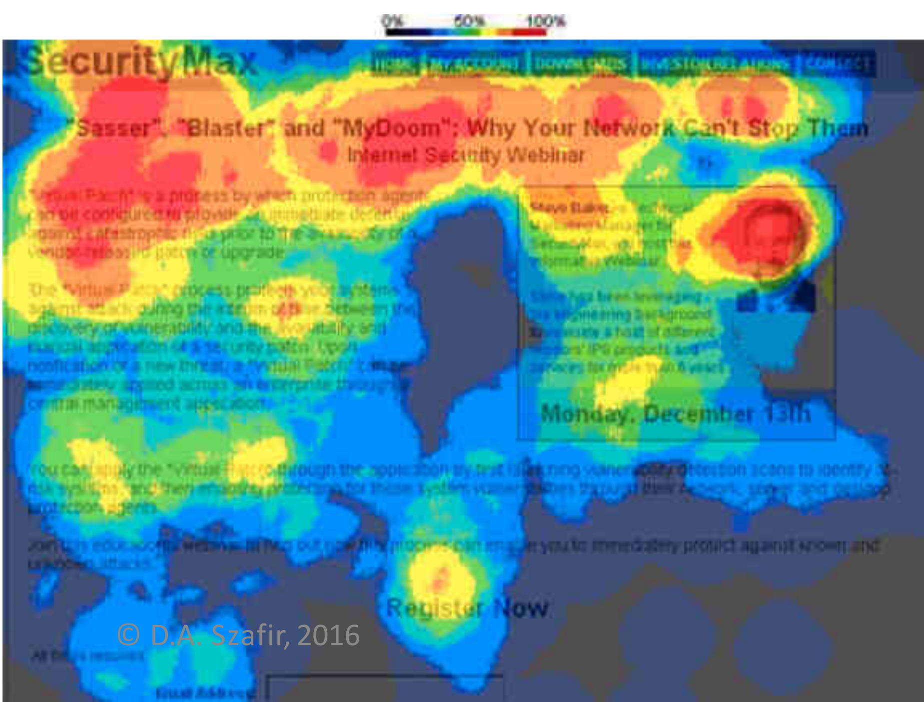


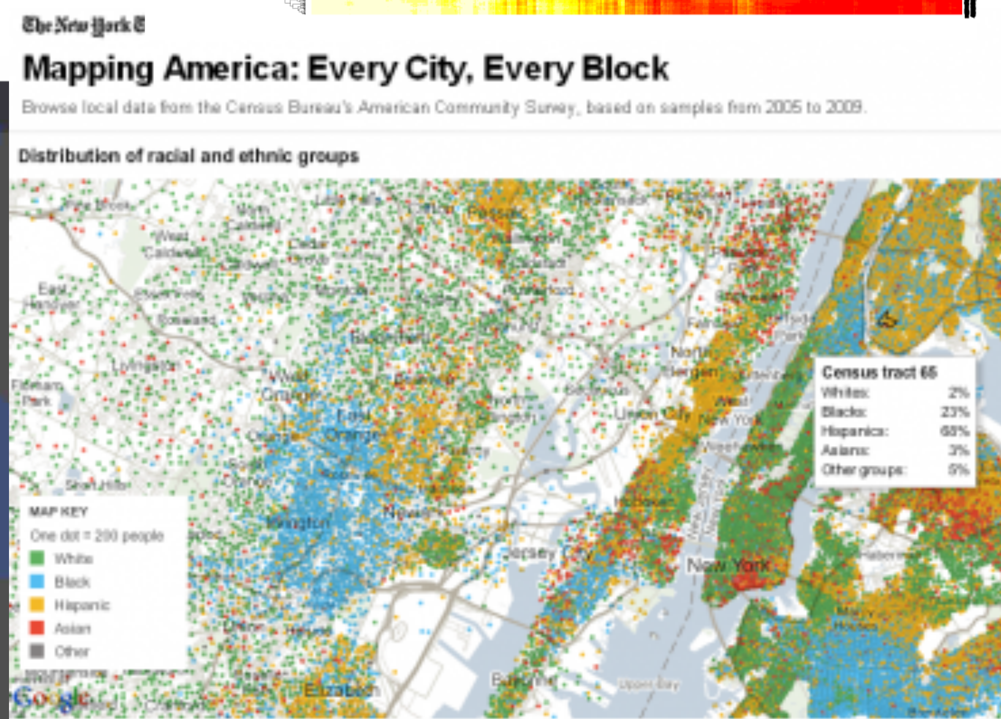
Visual Clutter

Color better supports visual processing at scale

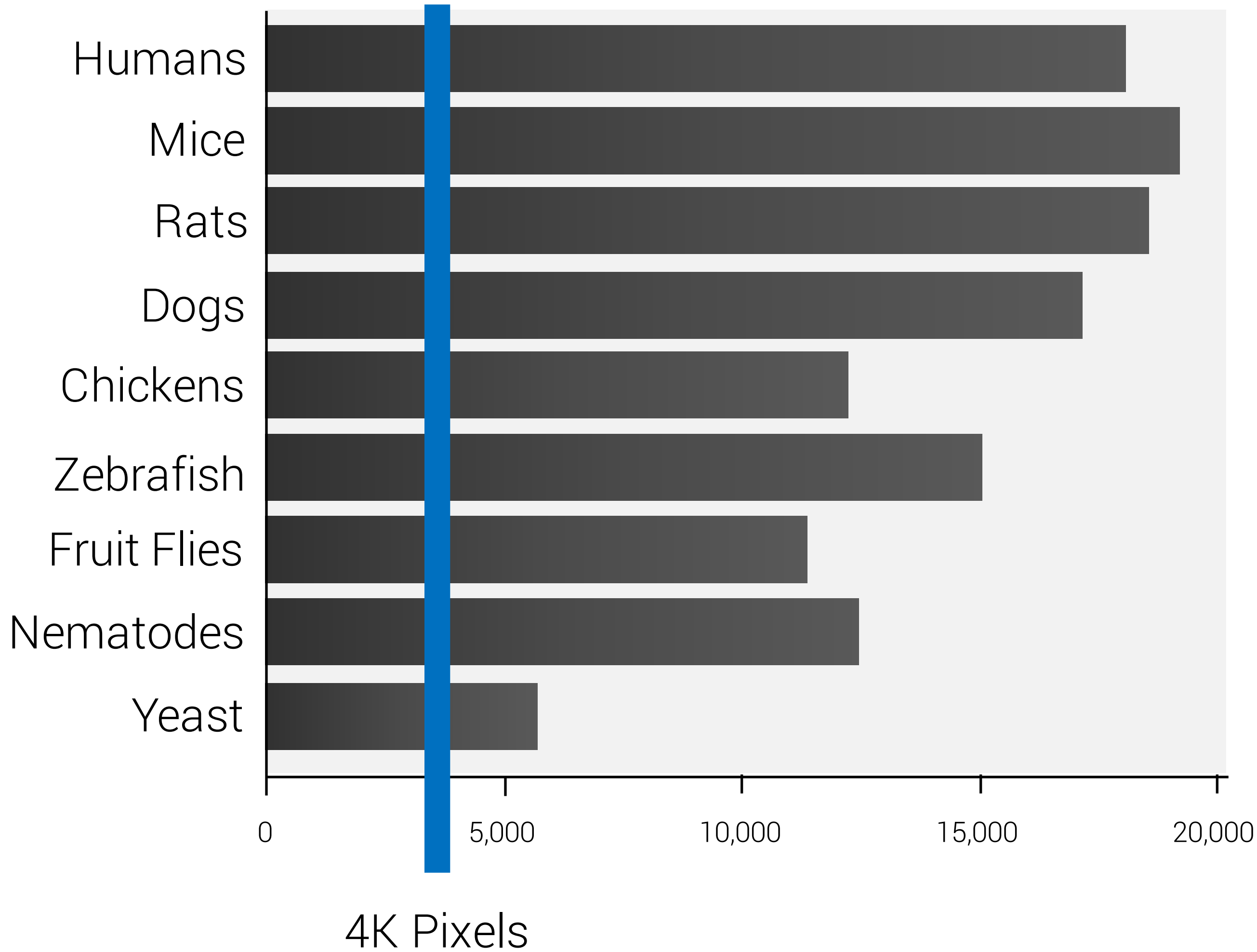

Wash.
Mont.
N.D.
Minn.
Ore.
Idaho
S.D.
Wis.
Wyo.
Mich.
N.Y.
Me.
Iowa
Neb.
Pa.
Nev.
Utah
Ill.
Ind.
Ohio
Calif.
Colo.
W.Va.
Va.
Kan.
Mo.
Ky.
N.C.
Tenn.
Okla.
Ariz.
N.M.
Ark.
S.C.
Ala.
Ga.
Miss.
Tex.
La.
Fla.
Alaska
Hawaii
ote: Due to limitations in the Census data, foreign-born populations are not available in all areas for all years.




0% 50% 100%
SecurityMax
"Sasser", "Blaster" and "MyDoom": Why Your Network Can't Stop Them
Internet Security Webinar
Monday, December 13th
Register Now

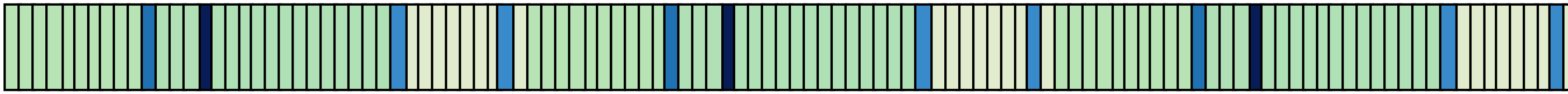

Mapping America: Every City, Every Block
Browse local data from the Census Bureau's American Community Survey, based on samples from 2005 to 2009.
Distribution of racial and ethnic groups
Census tract 65
Whites: 2%
Blacks: 23%
Hispanics: 68%
Asians: 3%
Other groups: 5%
MAP KEY
One dot = 200 people
White
Black
Hispanic
Asian
Other
New York

Average Number of Genes per Genome

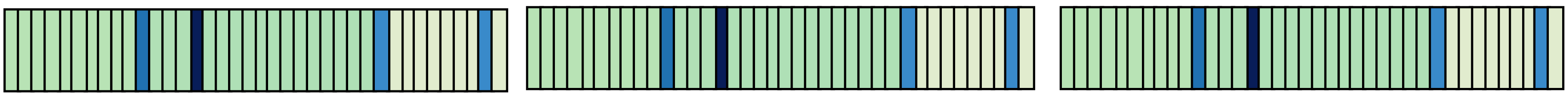


Often too many genes to display on the monitor

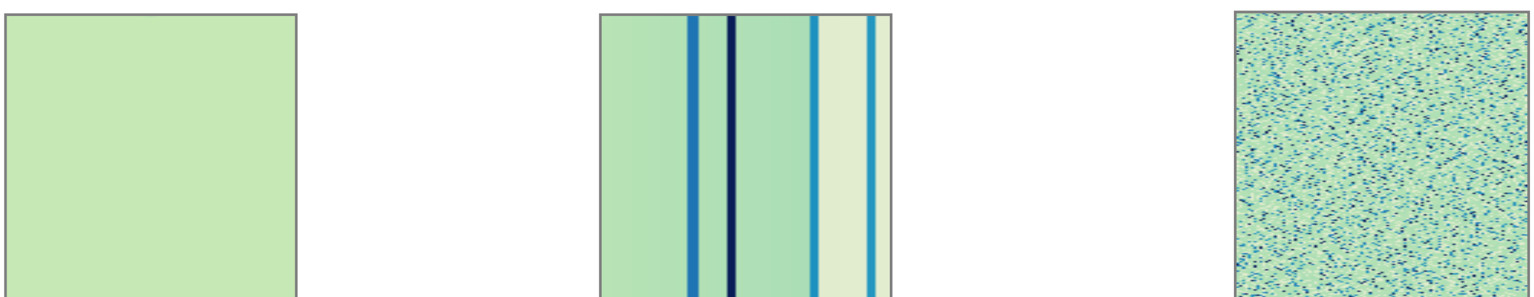
Raw Sequence

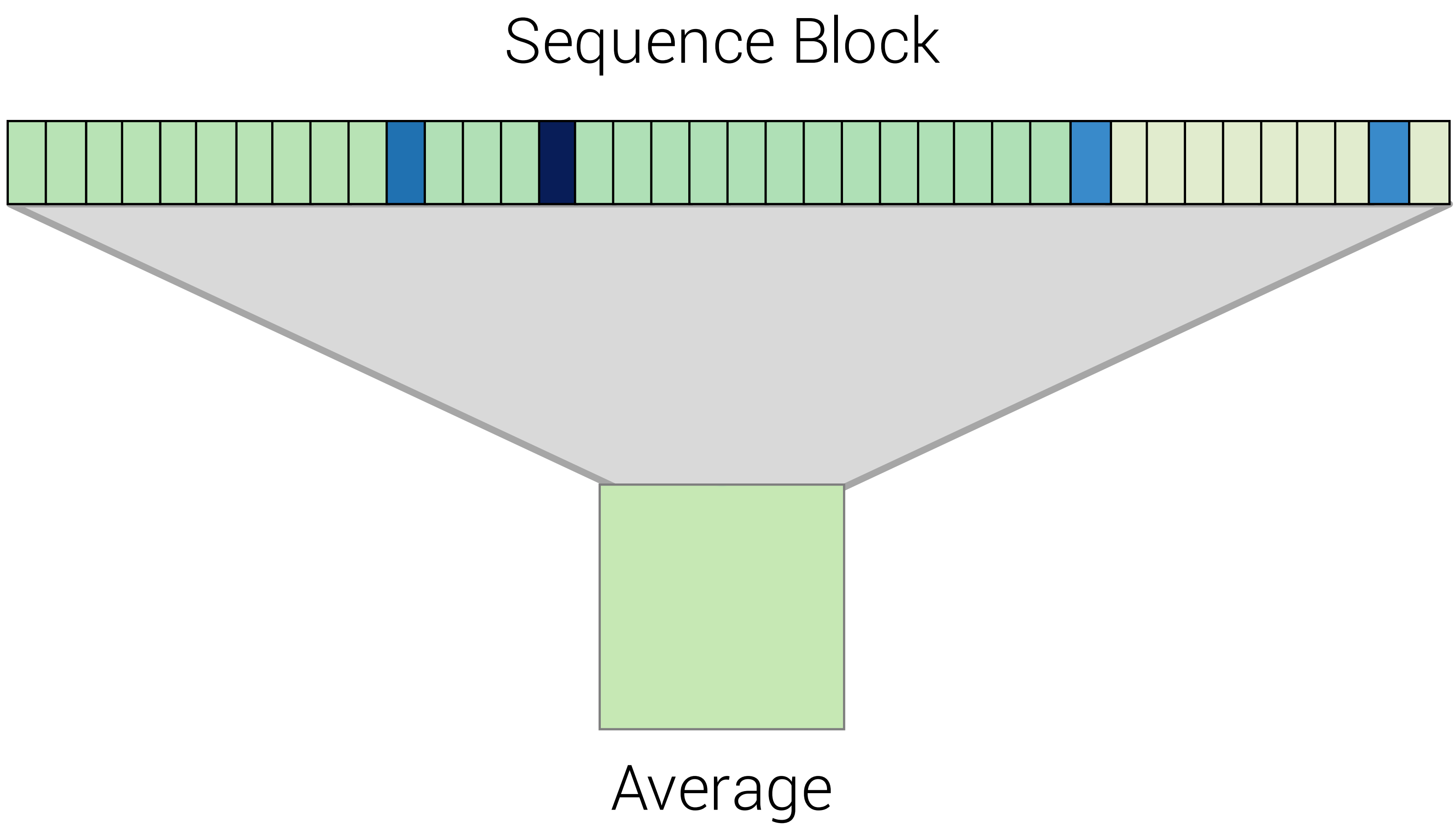


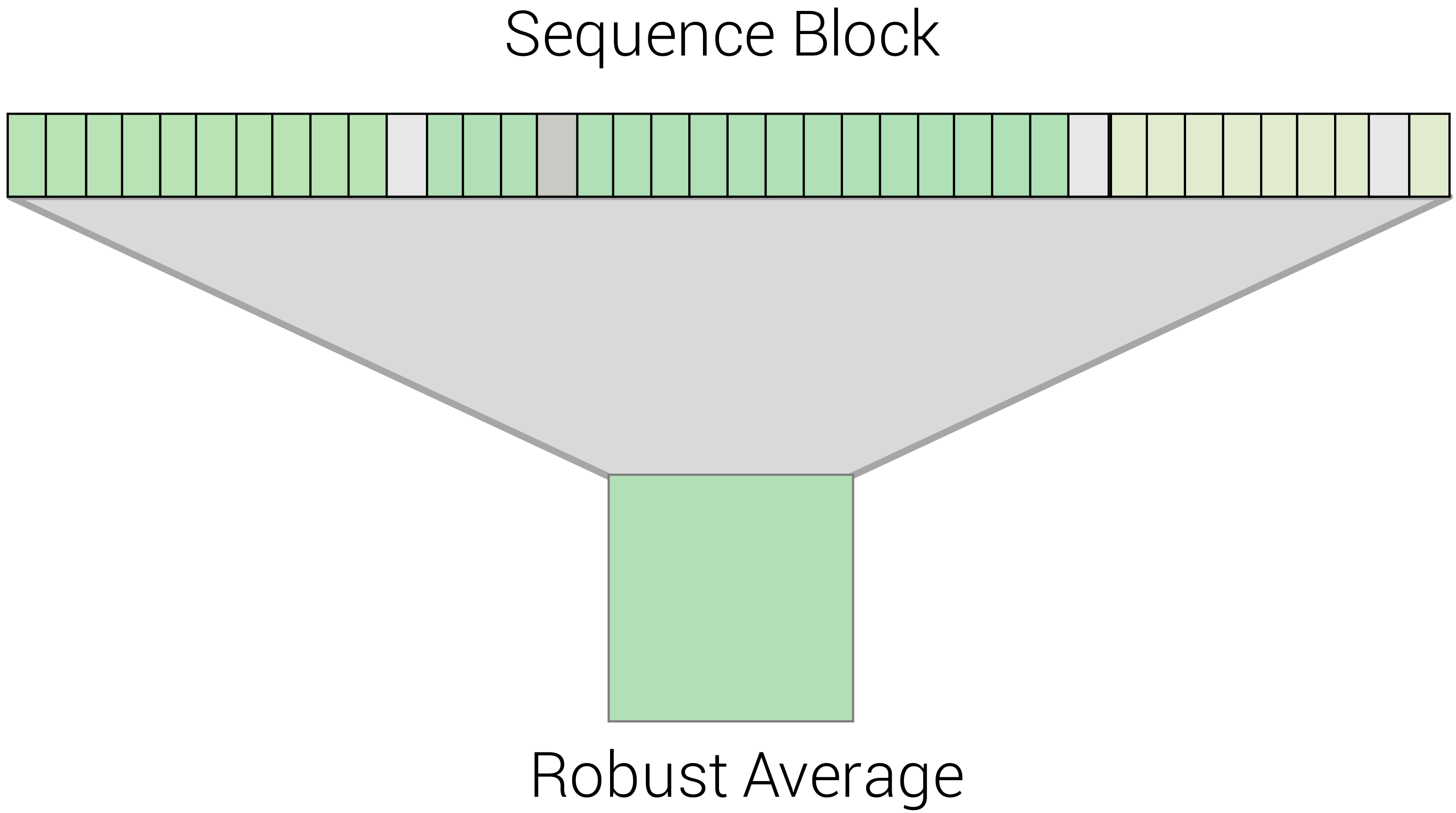
Sequence Blocks

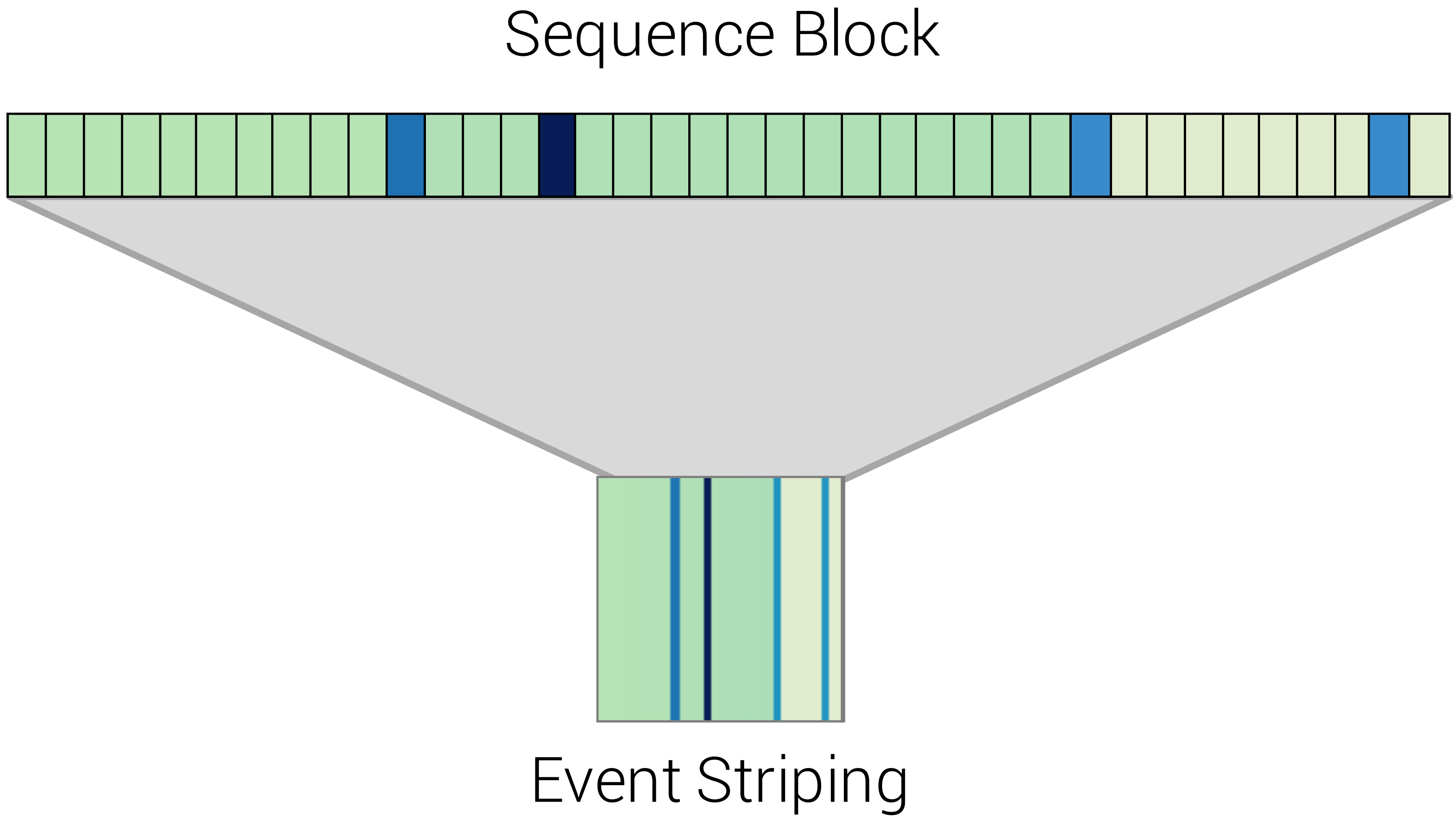


Aggregate Representation

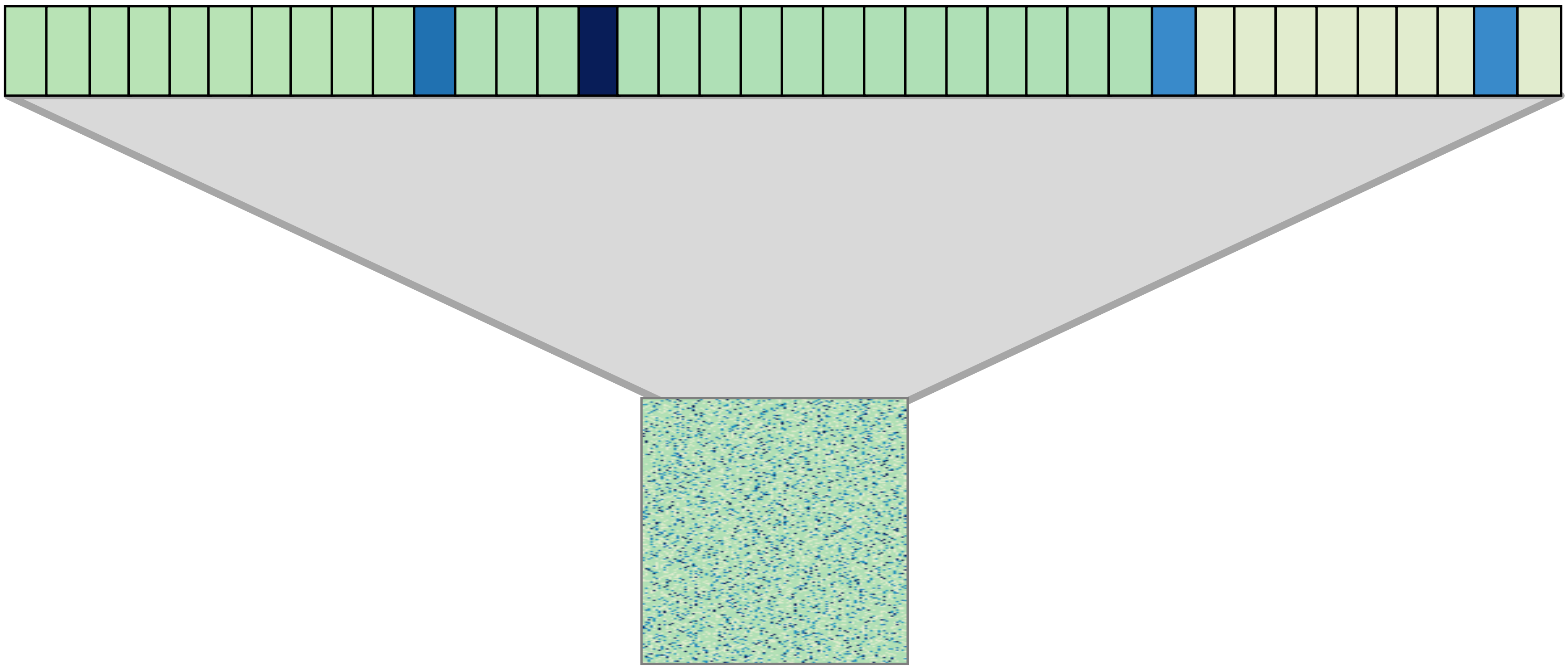



Sequence Block
Average


Sequence Block
Robust Average


Sequence Block
Event Striping

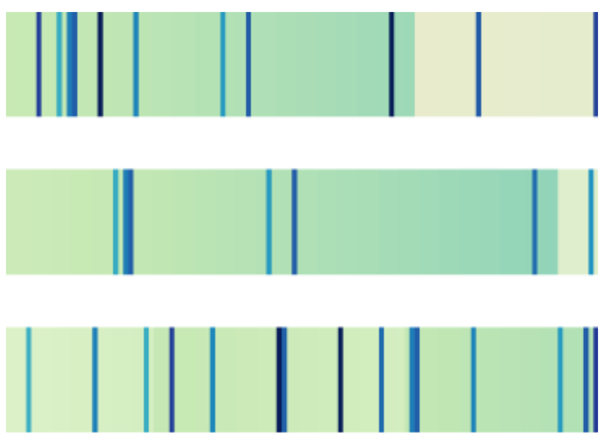
Sequence Block



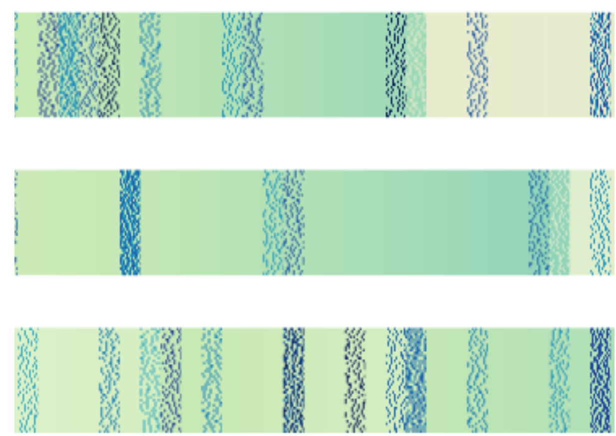
Color Weaving

Average

Robust Average



Event Striping



Color Weaving

Visualization in the Age of Big Data

Understand limits in current tools

- What does the data look like?
- The Fix: Aligning patterns with tasks
- Building The System

Derive inspiration across domains

- Literary Patterns

Link big and small

- Machine Learning & Molecules

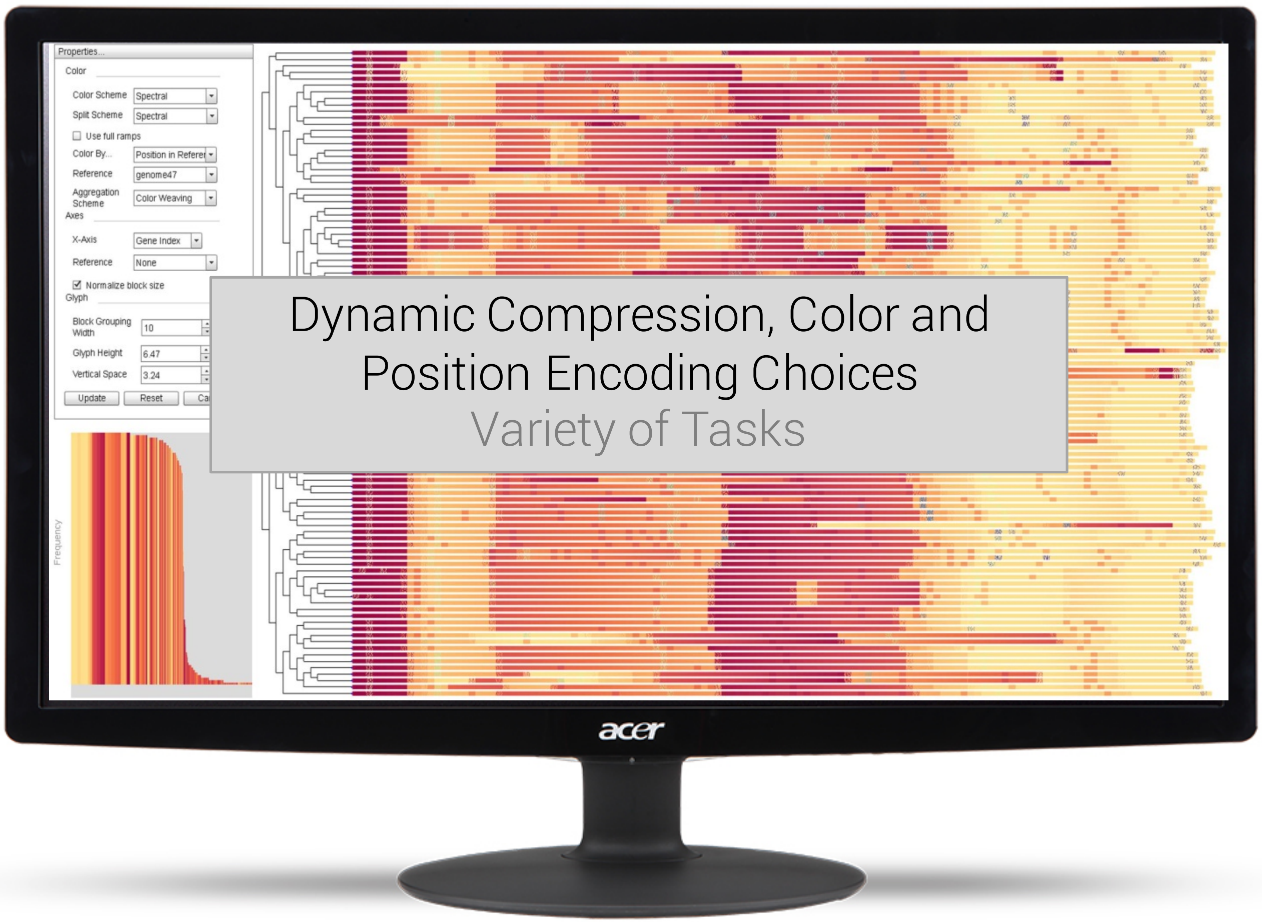
Sequence Surveyor

Task-Driven Sequence Aggregation

Length of Sequences

Perceptually-Driven Encoding

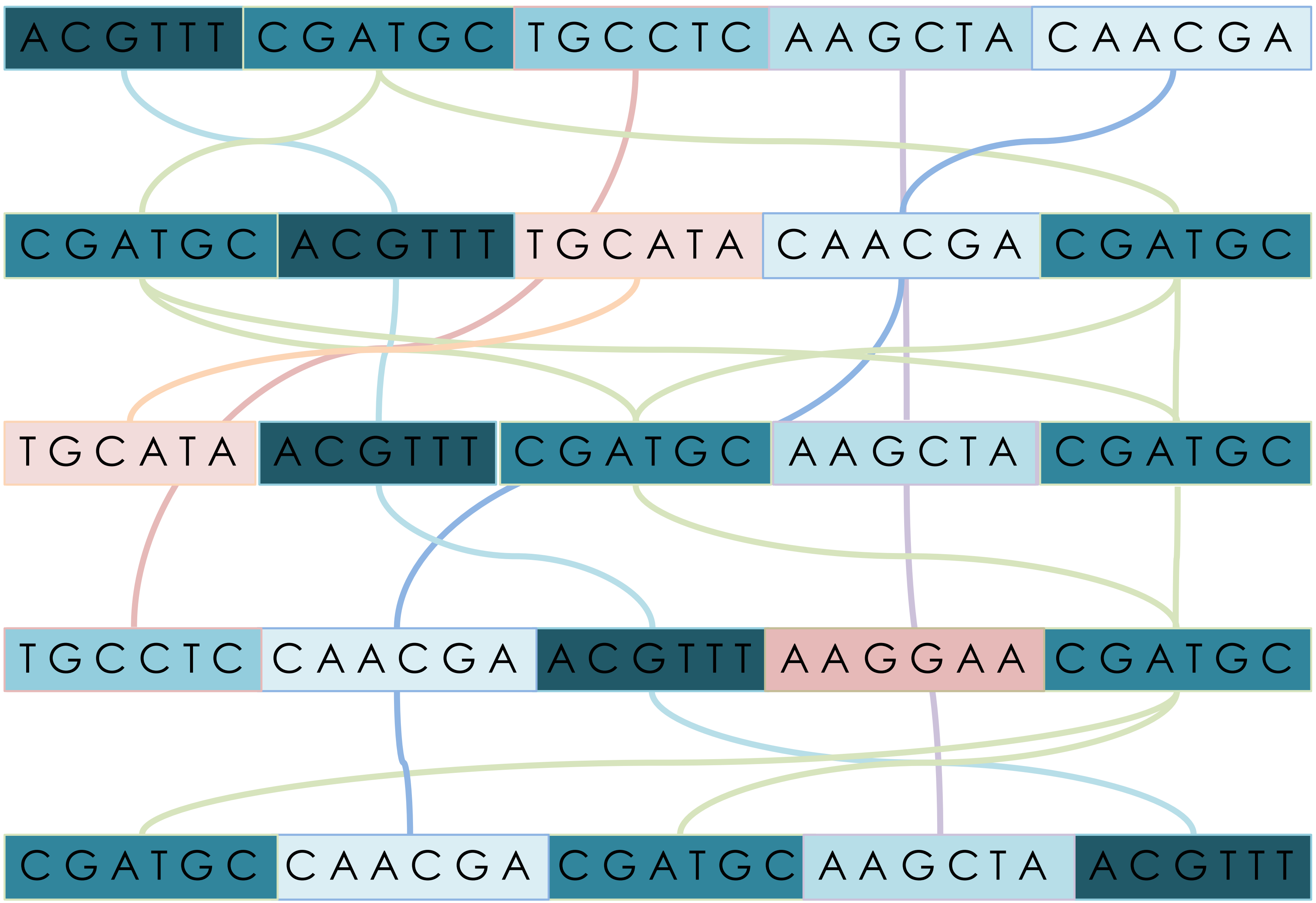
Number of Sequences



Dynamic Compression, Color and Position Encoding Choices

Variety of Tasks


ACGTTT CGATGC TGCCTC AAGCTA CAACGA
CGATGC ACGTTT TGCATA CAACGA CGATGC
TGCATA ACGTTT CGATGC AAGCTA CGATGC
TGCCTC CAACGA ACGTTT AAGGAA CGATGC
CGATGC CAACGA CGATGC AAGCTA ACGTTT


ACGTTT CGATGC TGCCTC AAGCTA CAACGA
CGATGC ACGTTT TGCATA CAACGA CGATGC
TGCATA ACGTTT CGATGC AAGCTA CGATGC
TGCCTC CAACGA ACGTTT AAGGAA CGATGC
CGATGC CAACGA CGATGC AAGCTA ACGTTT


CGATGC
ACGTTT
CAACGA
AAGCTA
CGATGC
CGATGC
ACGTTT
CAACGA
TG
CGATGC
CGATGC
ACGTTT
AAGCTA
CGATGC
ACGTTT
CAACGA
T
CGATGC
CGATGC
ACGTTT
CAACGA
AAGCTA

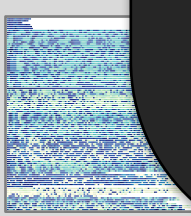
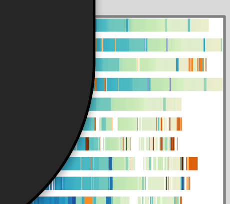

Index
Membership Freq
Grouped Freq
Pos in Reference
Index
Grouped Freq
Pos in Reference

10x More Sequences

100x Longer Sequences

Reference-Dependent, Independent, and Metadata-Based Analyses

Explicit Support for High-Level and Low-Level Relationships

100 Bacteria
6,000 genes

50 Bacteria
5,000 genes

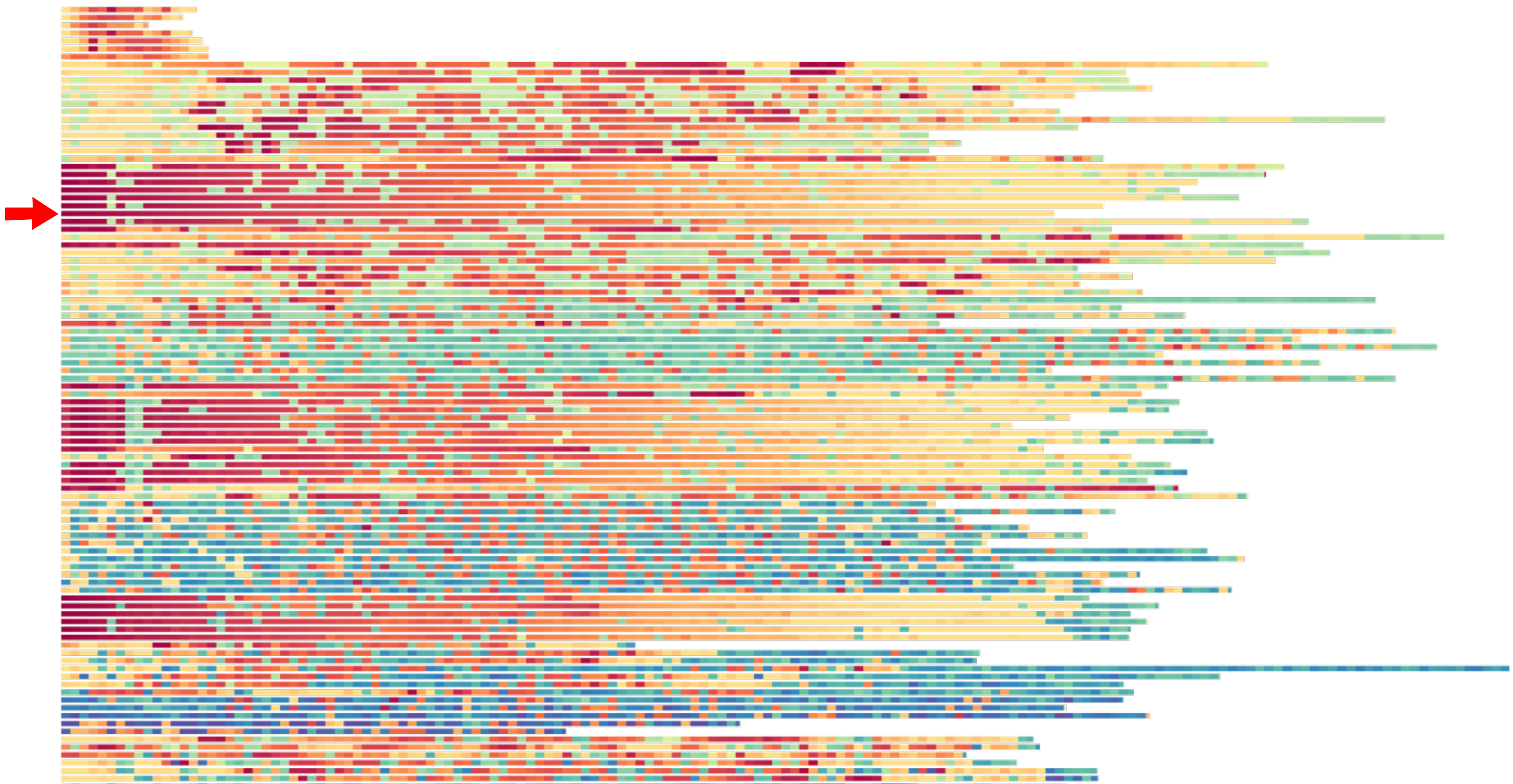
35 Fungi
17,000 genes

14 Pathogens
4,000 genes

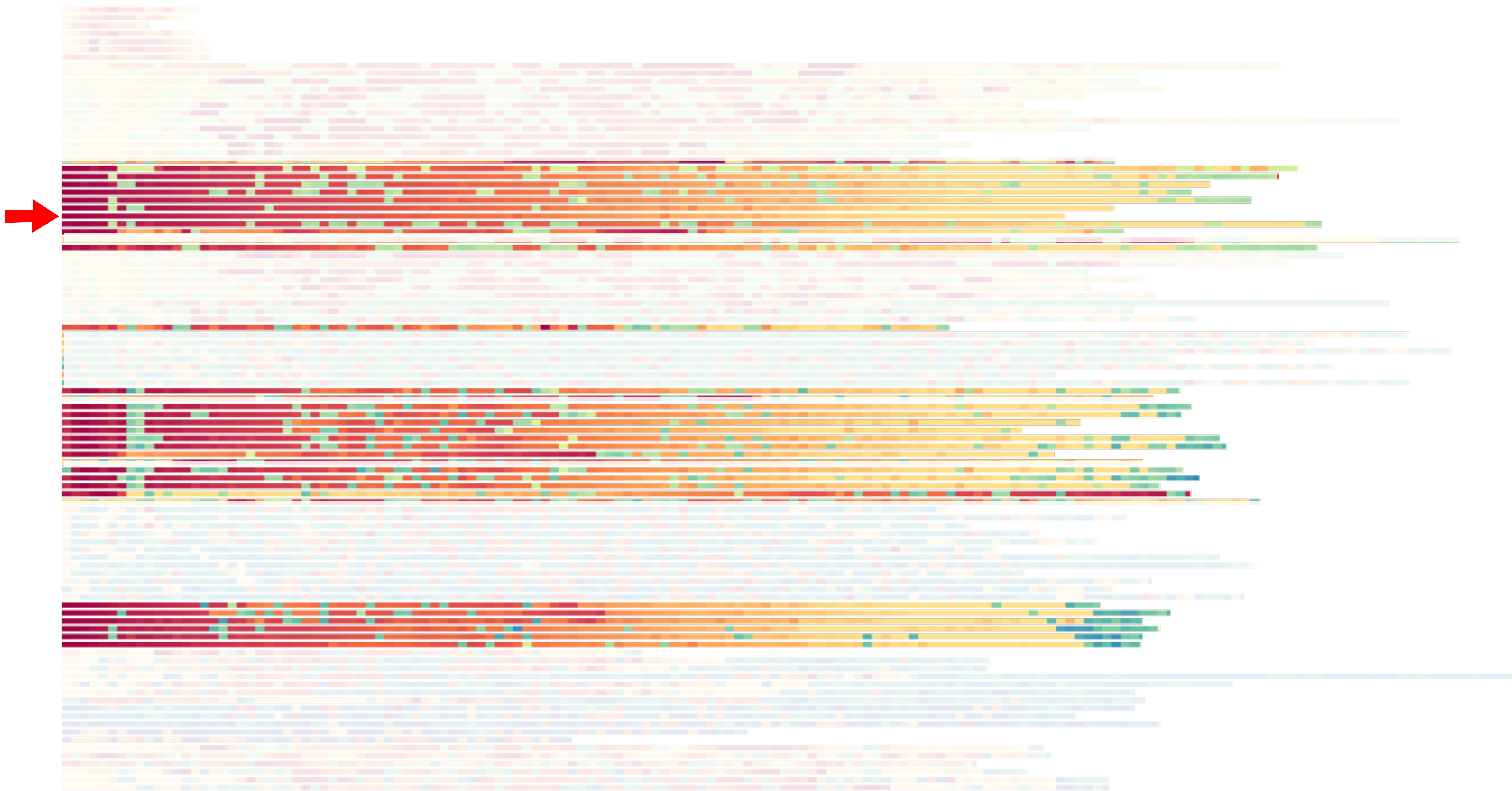
8 Partial *E. Coli*
300 genes

Explore Evolutionary Patterns in Organisms

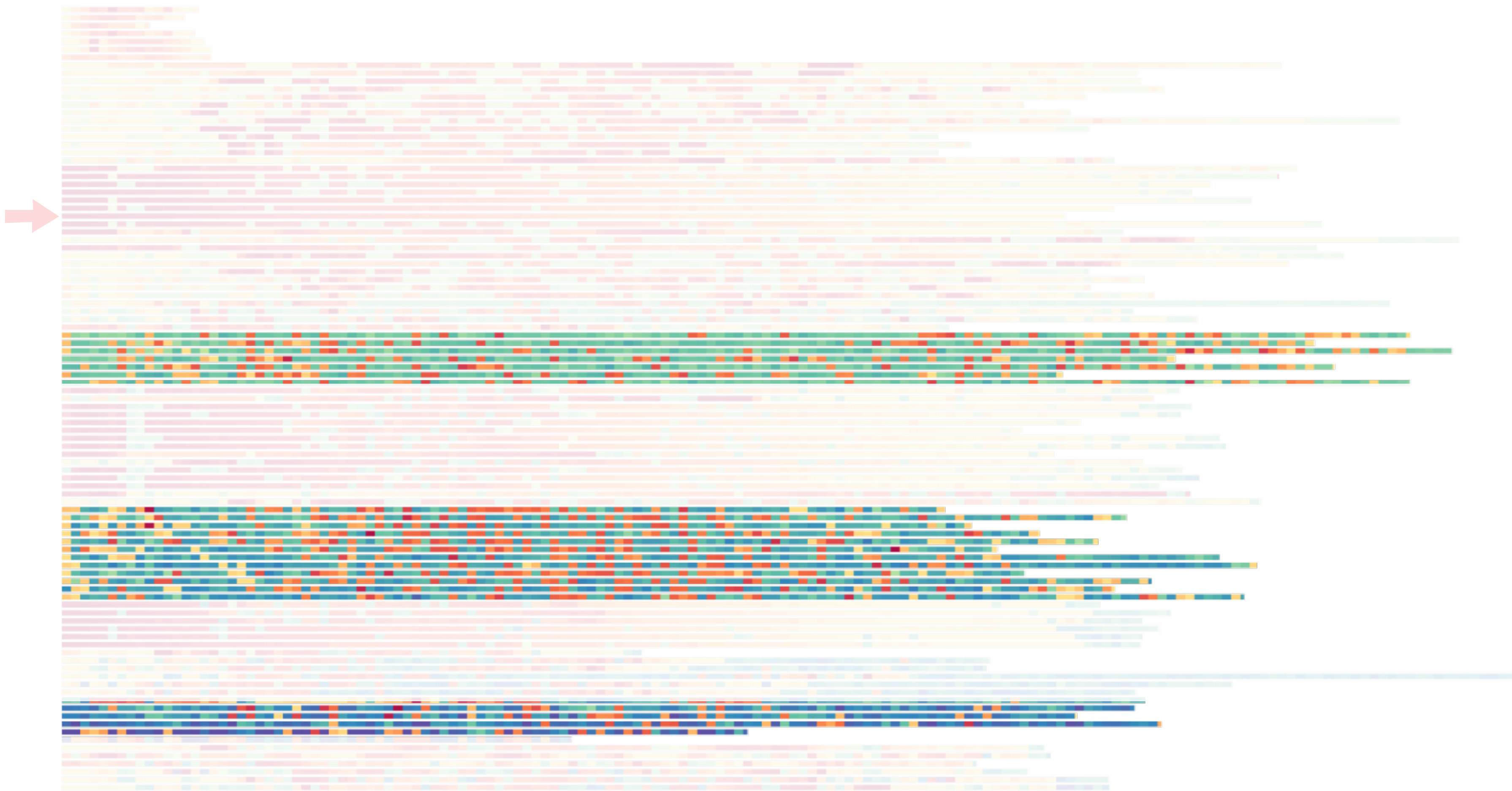
Explore Phylogenetic Relationships



Explore Phylogenetic Relationships



Explore Phylogenetic Relationships



"At a Glance" Algorithm Debugging

Color-based aggregation better supports analyses at scale

Visualization in the Age of Big Data

Understand limits in current tools

Large Scale Sequence Alignment

Derive inspiration across domains

Literary Patterns

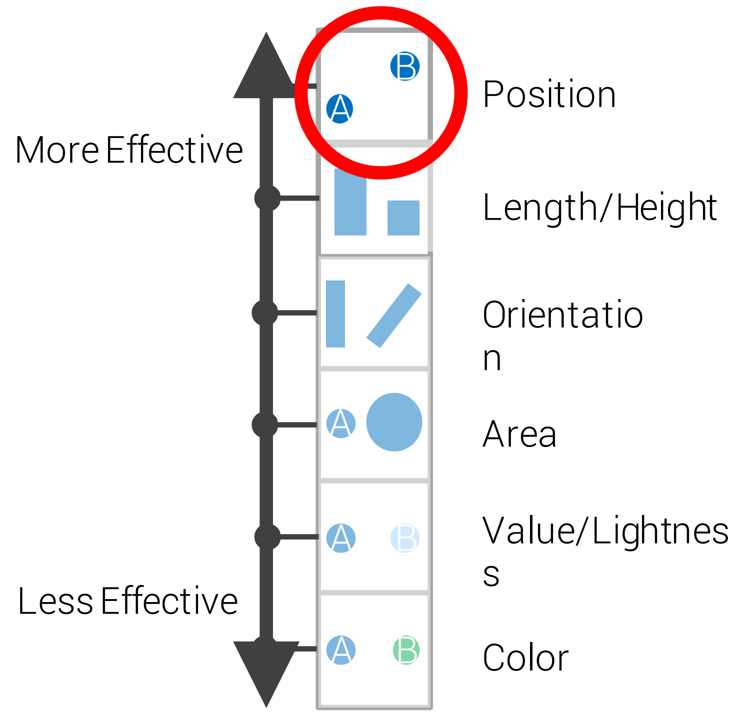
Link big and small

Machine Learning & Molecules

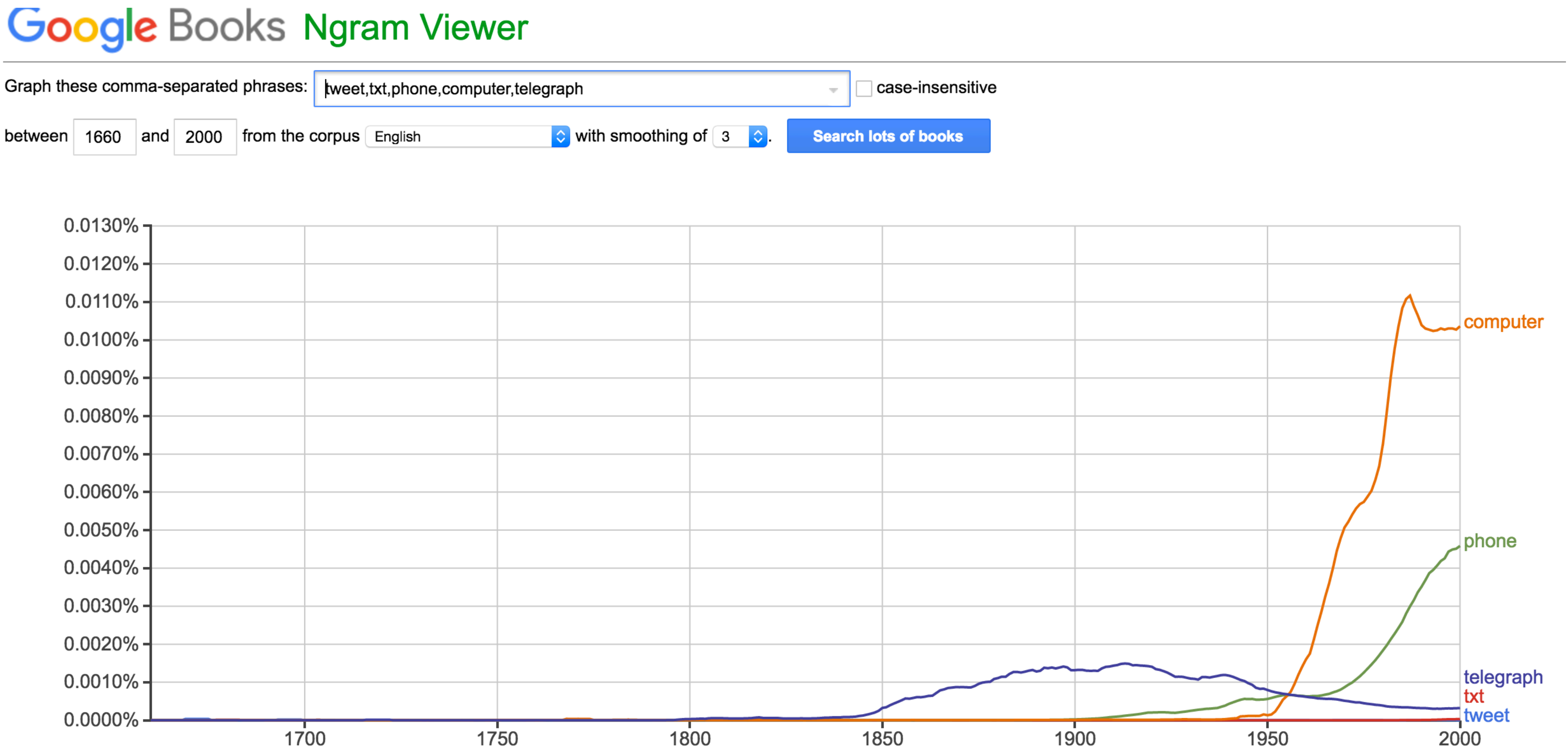
All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances;
And one man in his time plays many parts,
His acts being seven ages. At first the infant,
Mewling and puking in the nurse's arms.
And then the whining school-boy, with his satchel
And shining morning face, creeping like snail
Unwillingly to school. And then the lover,
Sighing like furnace, with a woeful ballad
Made to his mistress' eyebrow. Then a soldier,
Full of strange oaths and bearded like the pard,
Jealous in honour, sudden and quick in quarrel,
Seeking the bubble reputation
Even in the cannon's mouth.

Large Digitized Collections

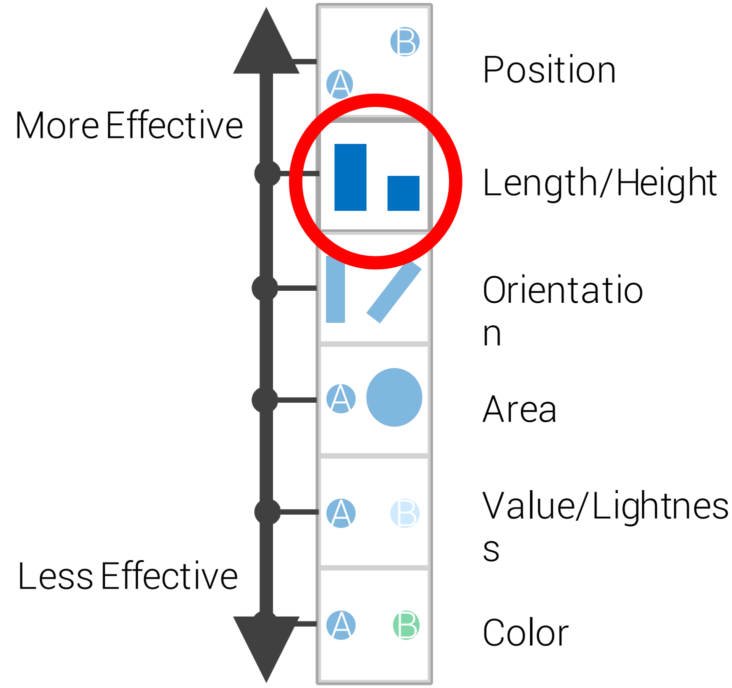
Google N-Grams: 5,195,769 books



Cleveland & McGill, 1985


Google Books Ngram Viewer
Graph these comma-separated phrases:
tweet,txt,phone,computer,telegraph
case-insensitive
between
1660
and
2000
from the corpus
English
with smoothing of
3
Search lots of books
0.0130%
0.0120%
0.0110%
0.0100%
0.0090%
0.0080%
0.0070%
0.0060%
0.0050%
0.0040%
0.0030%
0.0020%
0.0010%
0.0000%
1700
1750
1800
1850
1900
1950
2000
computer
phone
telegraph
txt
tweet

Please don't use wordclouds



Cleveland & McGill, 1985


Tree
Measurements
Trunk
Diameter
Ground
Species
Retrieved
Plant
Redwood
Gymnosperm
Sequoia
Growth
Height
Conifers
Tallest
Wikipedia
Eucalyptus
Construction
Damage
National
Arboriculture

Word Usage Analysis Tasks

Characterize and compare authors

Measure shifts in an author's writing over time

Evolution of language over time

Evolution of cultural influences over time

Indicate recurring themes and topics

Characterize typographic conventions

Word Usage Analysis Tasks

Characterize and compare organisms

Measure shifts in organisms over species

Evolution of organisms over time

Evolution of cultural influences over time

Indicate recurring genetic material

Characterize typographic conventions

Turning texts into sequences

```
All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances,
```

```
all the world a stage
and all the men and women merely players
they have their exits and their entrances
```

all – the – world – a – stage – and – all – the – men – and – women – merel

Text Sequence:

Present words in their original reading order

Highlight word locations

Precise analysis for single texts

Ranked Count:

Order words by how often they occur in a text collection

Highlight word frequency

Aggregate multiple texts

A Midsummer Night's Dream

Text Sequence:	now	fair	Hippolyta	our	nuptial	hour	draws	on	apace	four
Ranked Count:	the	and	to	I	you	of	a	in	my	is
Position	1	2	3	4	5	6	7	8	9	10

Text Sequence:

Present words in their original reading order

Highlight word locations

Precise analysis for single texts

Ranked Count:

Order words by how often they occur in a text collection

Highlight word frequency

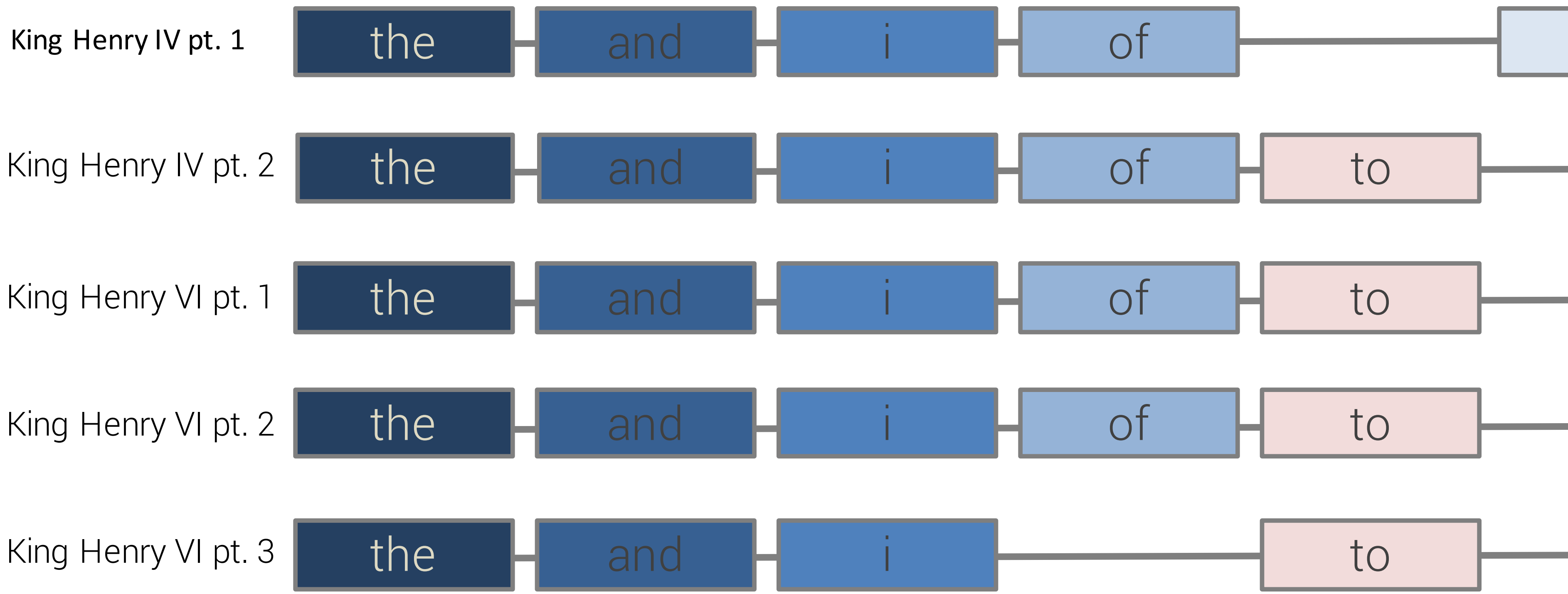
Aggregate multiple texts

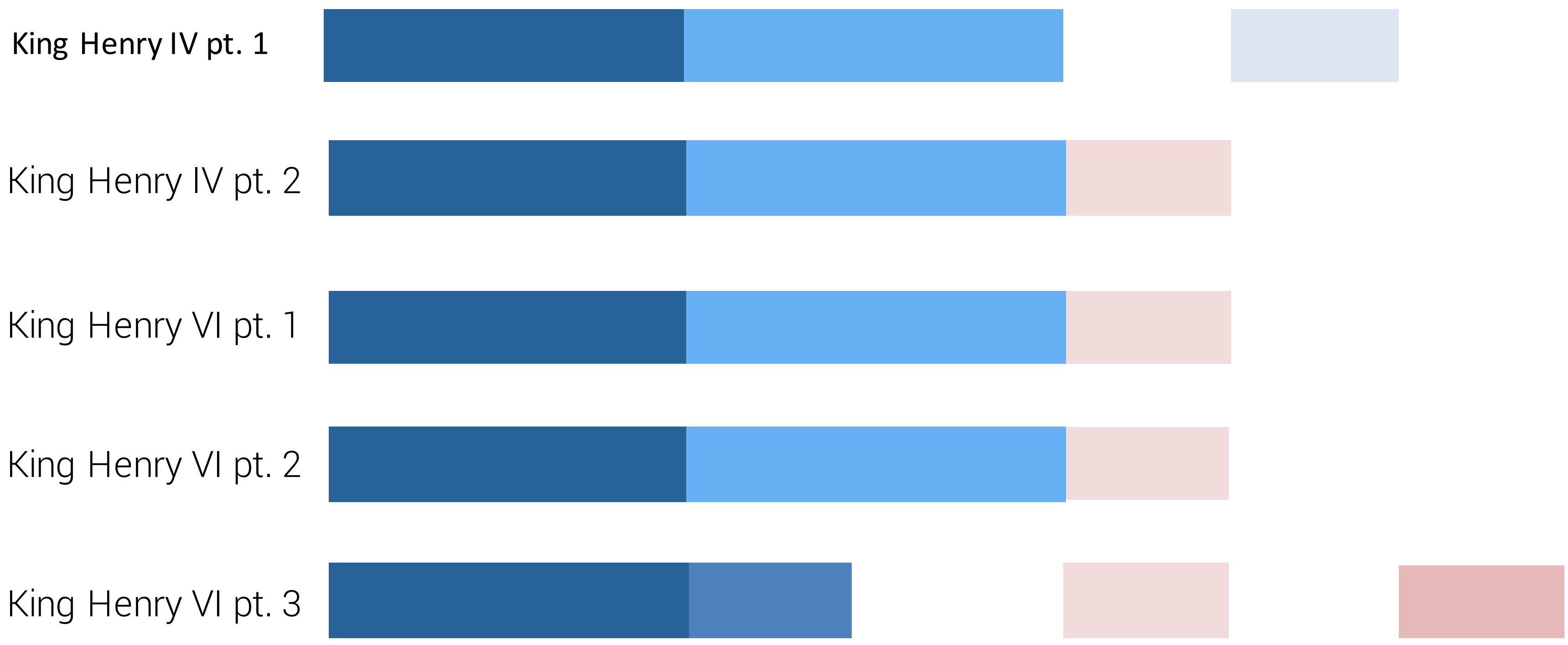
A Midsummer Night’s Dream

Text Sequence:	now	fair	Hippolyta	our	nuptial	hour	draws	on	apace	four
Ranked Count:	the	and	to	I	you	of	a	in	my	is
Position	1	2	3	4	5	6	7	8	9	10

King Henry IV pt. 1	the	and	i	of	a
King Henry IV pt. 2	the	and	i	of	to
King Henry VI pt. 1	and	the	of	to	i
King Henry VI pt. 2	the	and	to	i	of
King Henry VI pt. 3	and	the	to	i	my


King Henry IV pt. 1
the
and
i
of
a
King Henry IV pt. 2
the
and
i
of
to
King Henry VI pt. 1
and
the
of
to
i
King Henry VI pt. 2
the
and
to
i
of
King Henry VI pt. 3
and
the
to
i
my


King Henry IV pt. 1
the
and
i
of
King Henry IV pt. 2
the
and
i
of
to
King Henry VI pt. 1
the
and
i
of
to
King Henry VI pt. 2
the
and
i
of
to
King Henry VI pt. 3
the
and
i
to

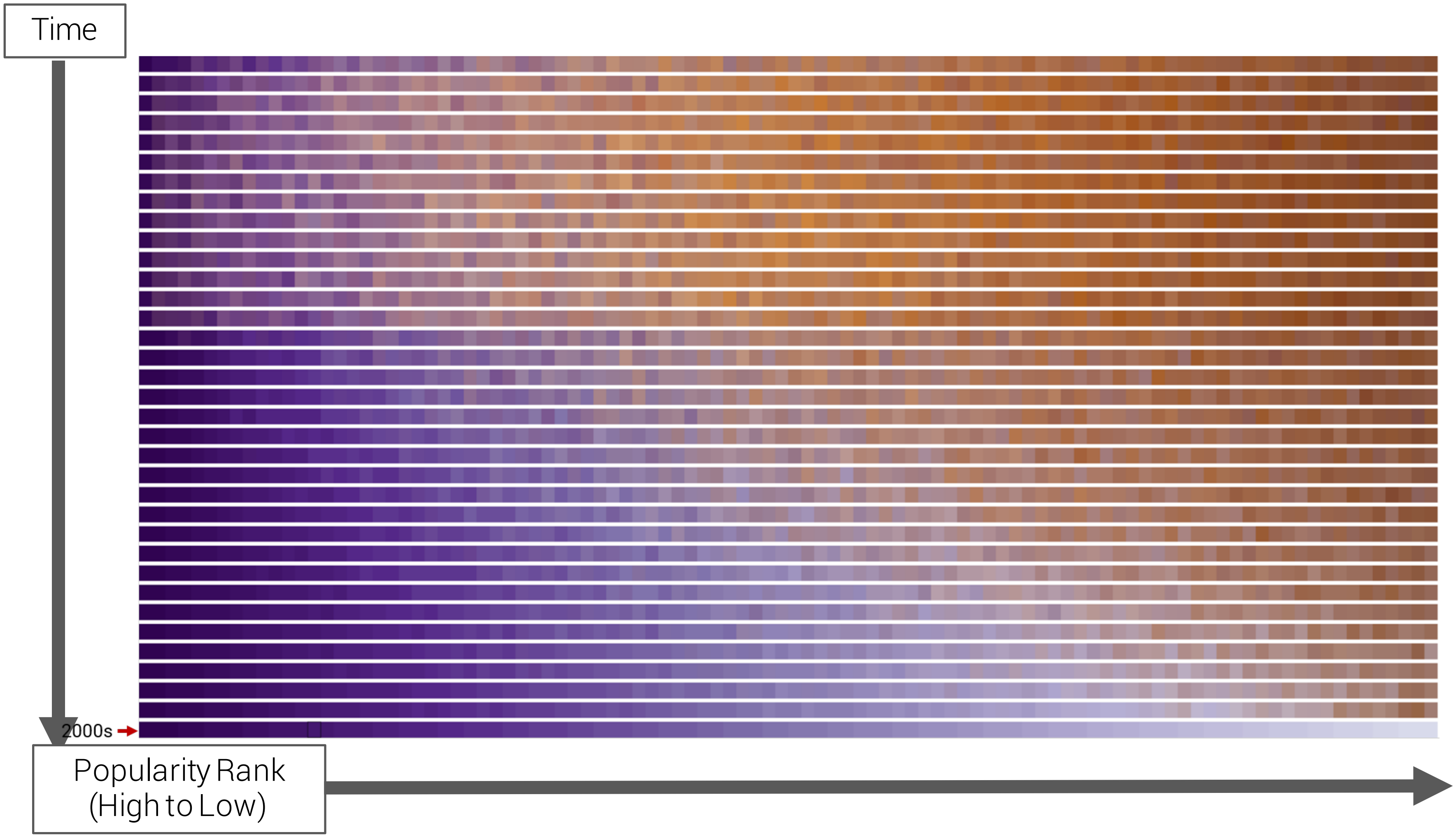

King Henry IV pt. 1
King Henry IV pt. 2
King Henry VI pt. 1
King Henry VI pt. 2
King Henry VI pt. 3

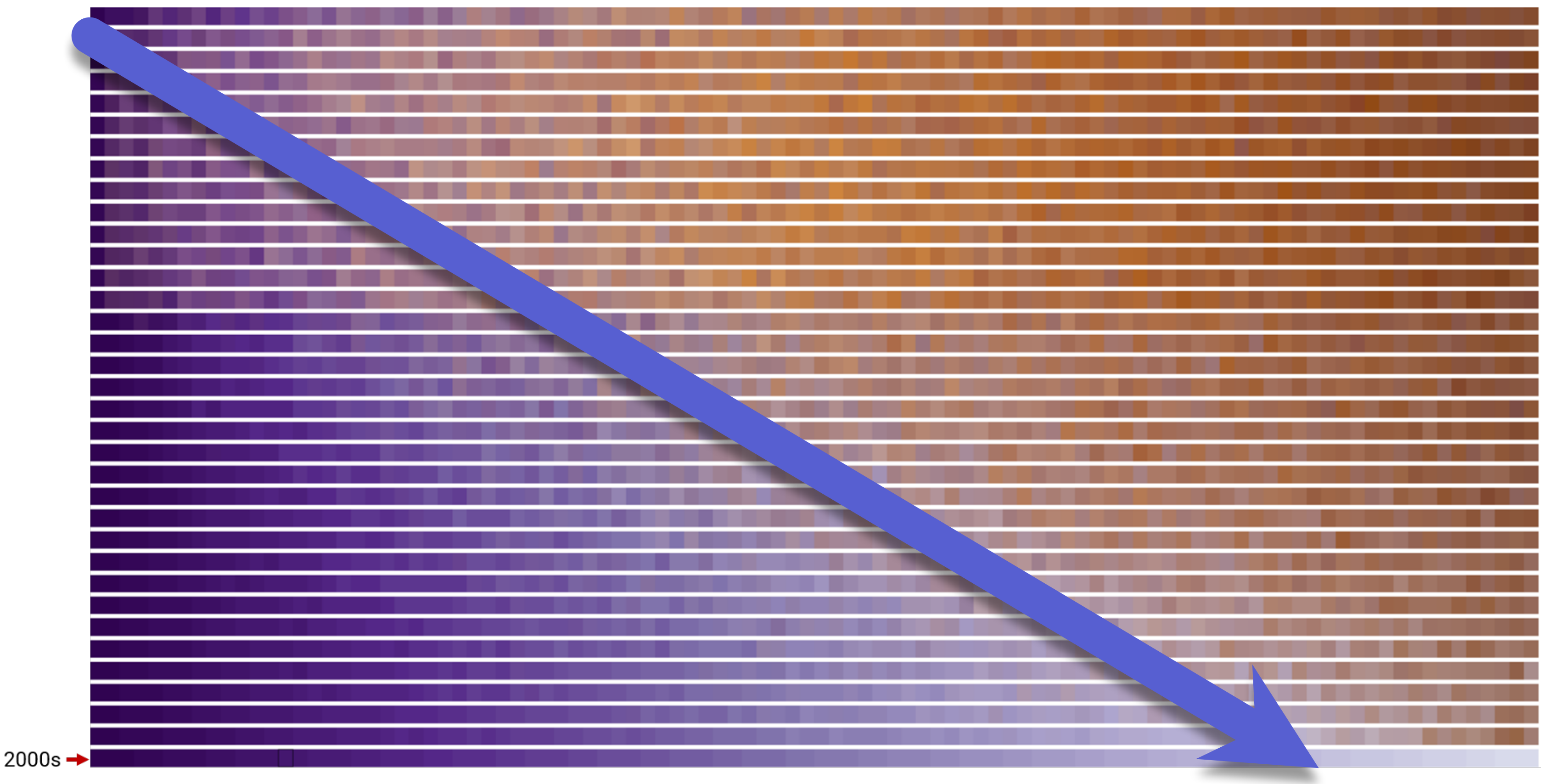
5.2 million books from 1660-2009

Mitchel et al, 2011

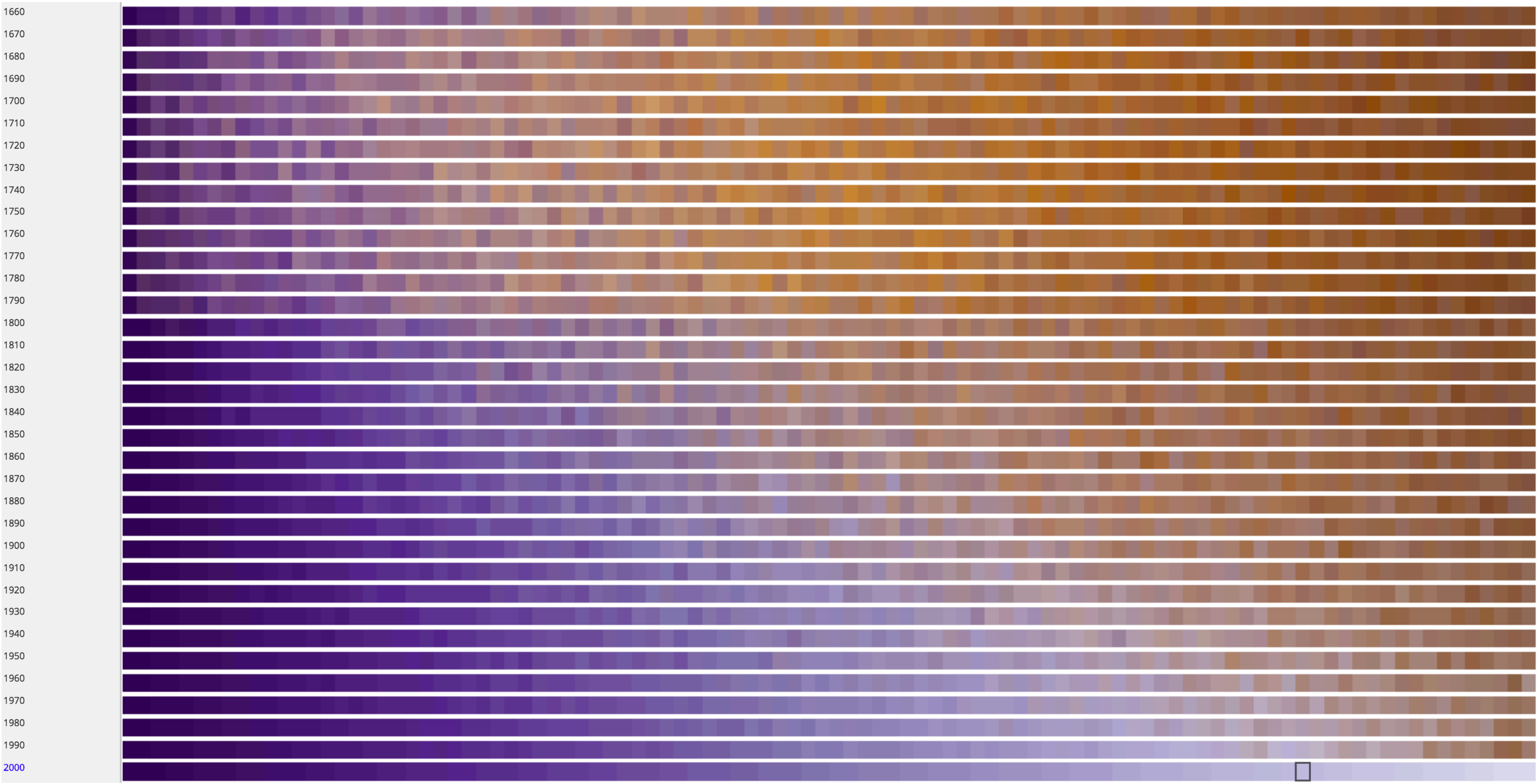
175,000 words over 35 decades

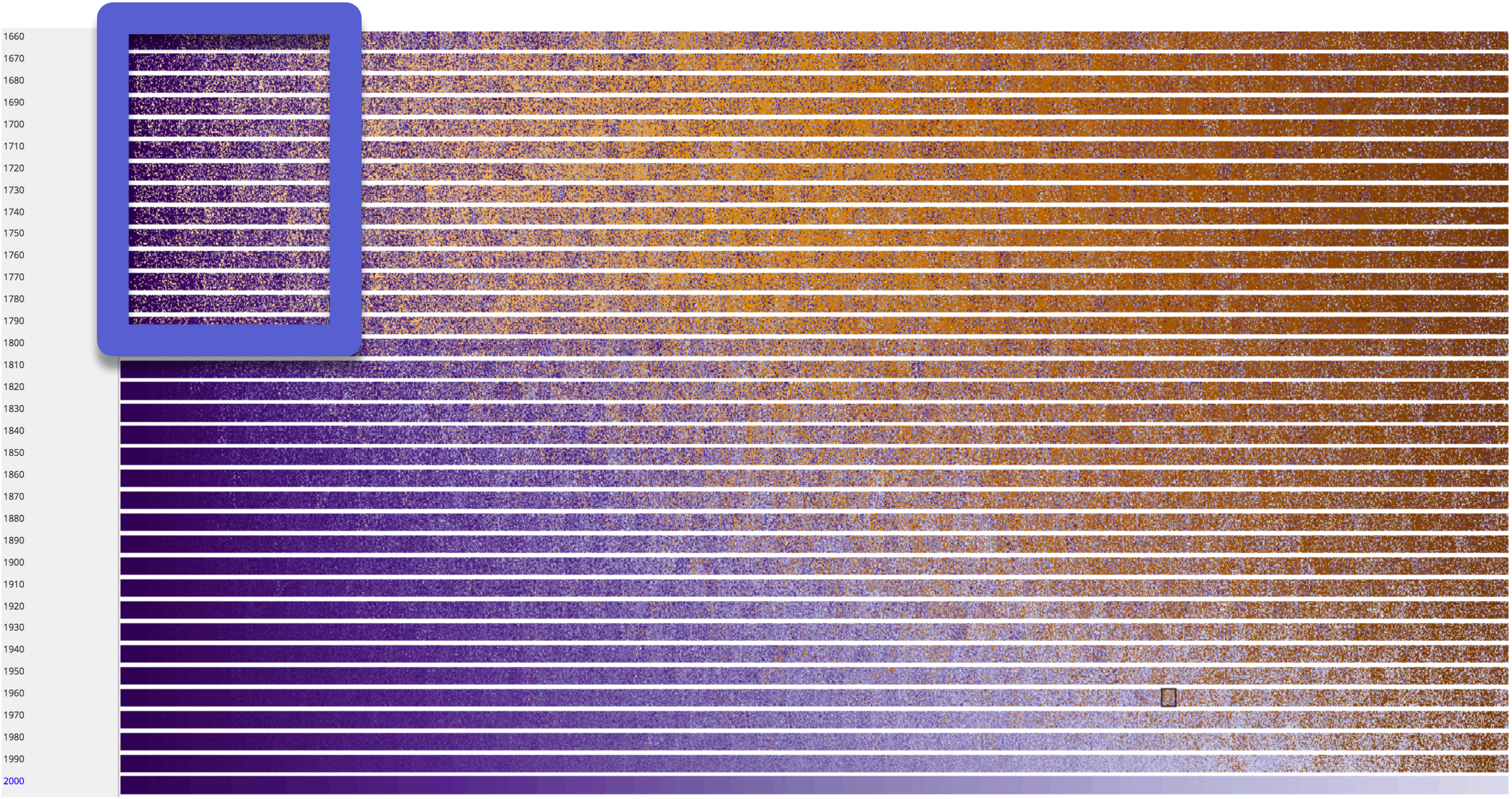
Explore Evolutionary Patterns in Writing

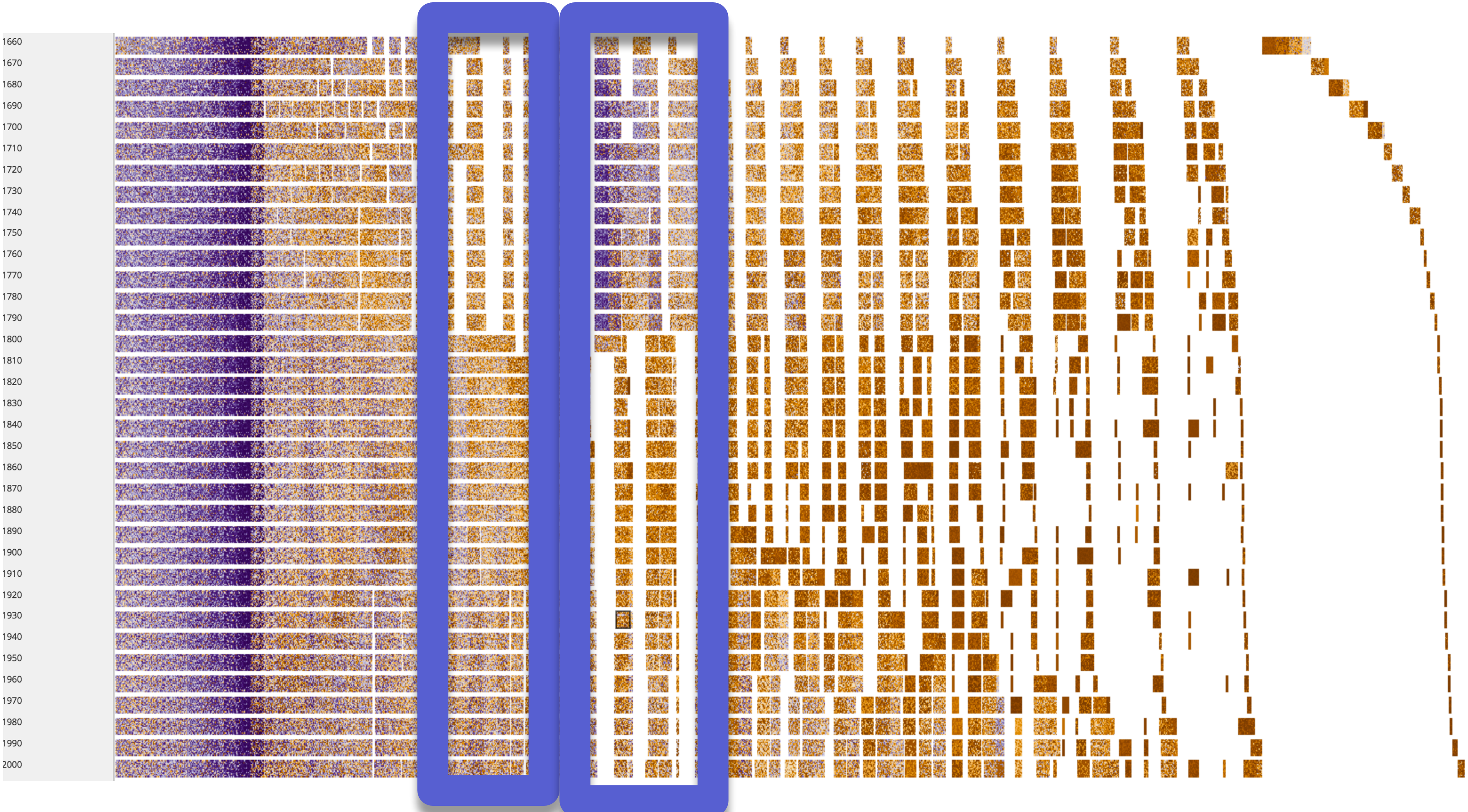

Time
2000s
Popularity Rank
(High to Low)


2000s

Confirm Prior Hypotheses


1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000


1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

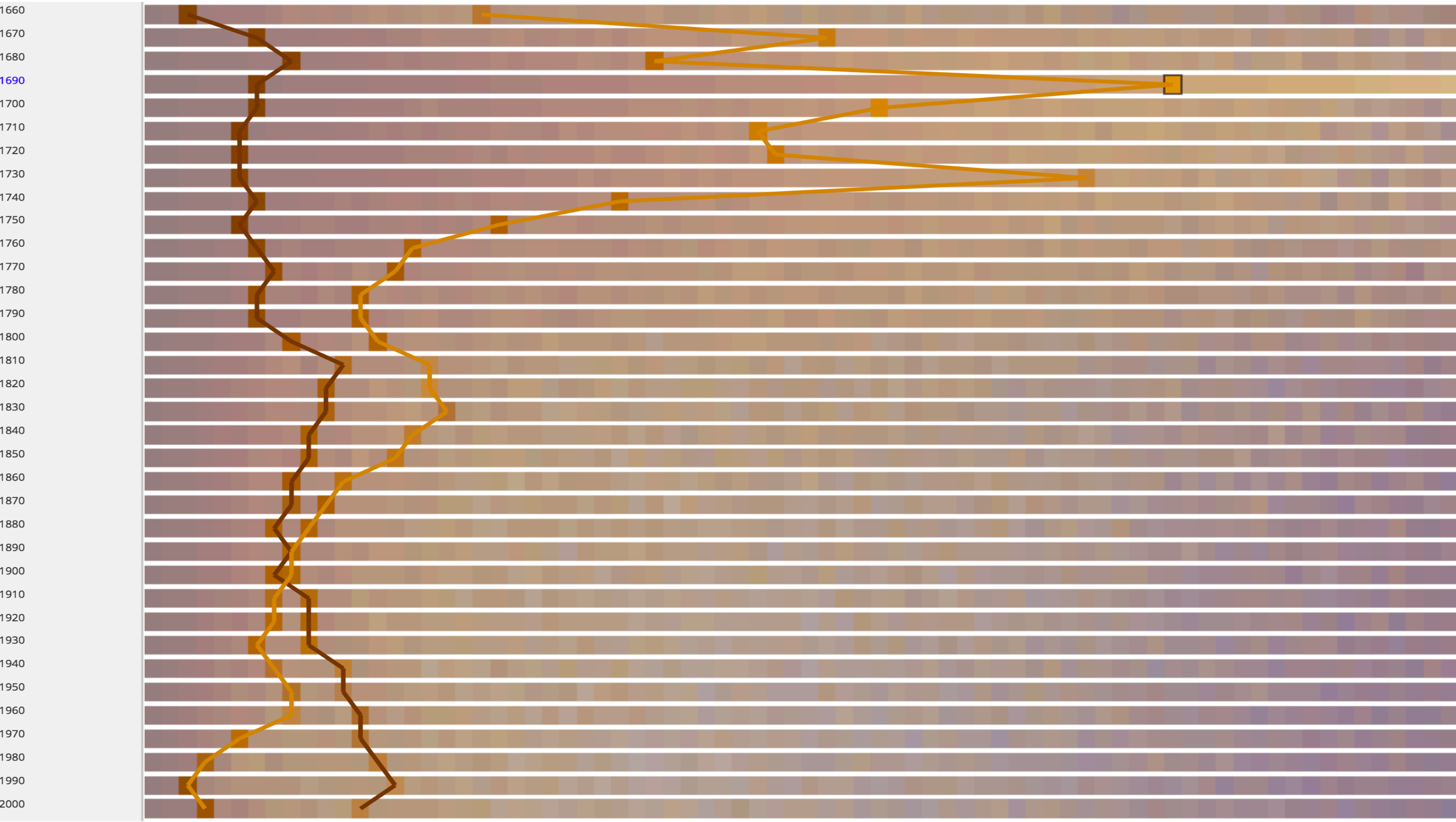

1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

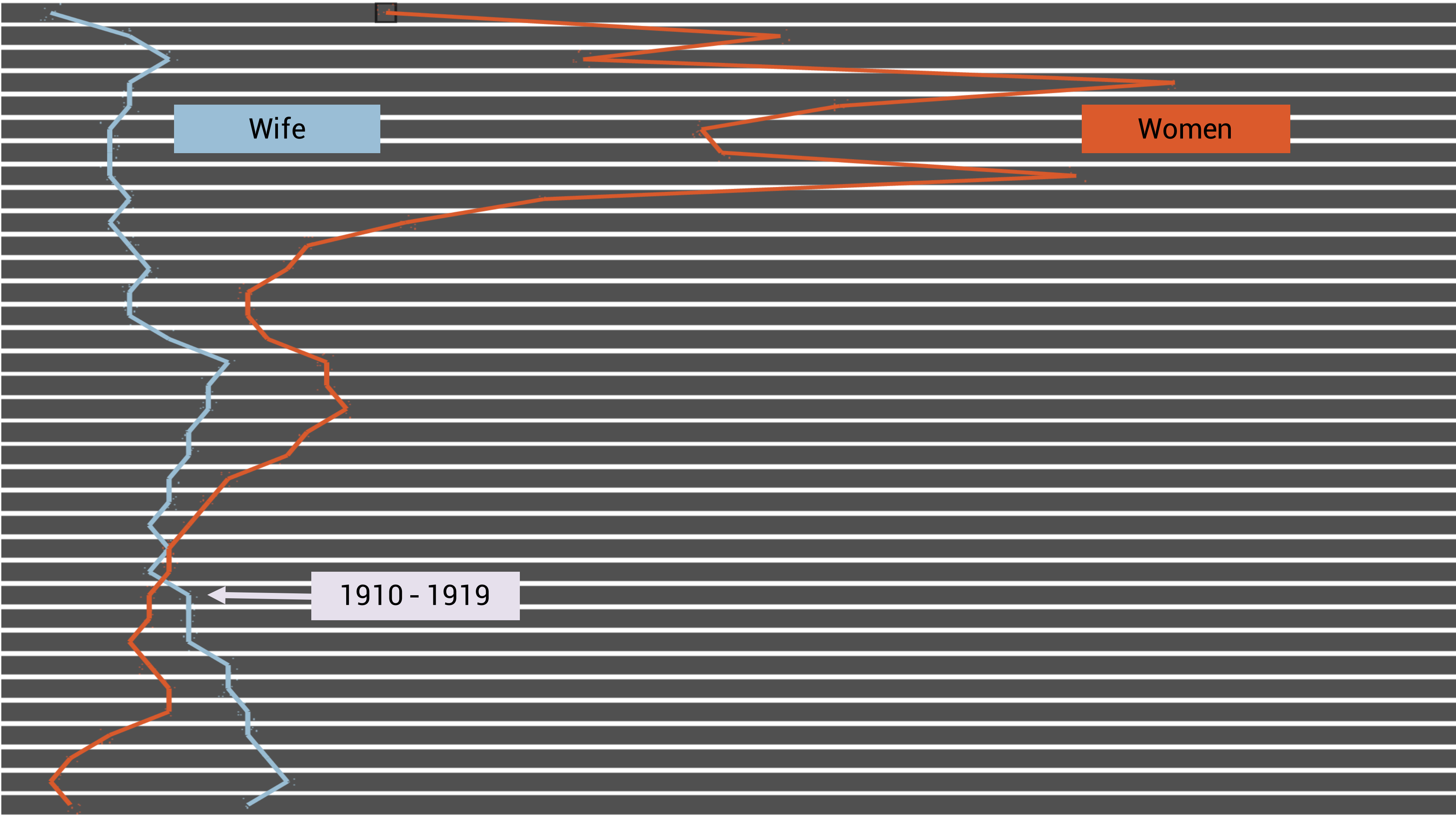
Sequences

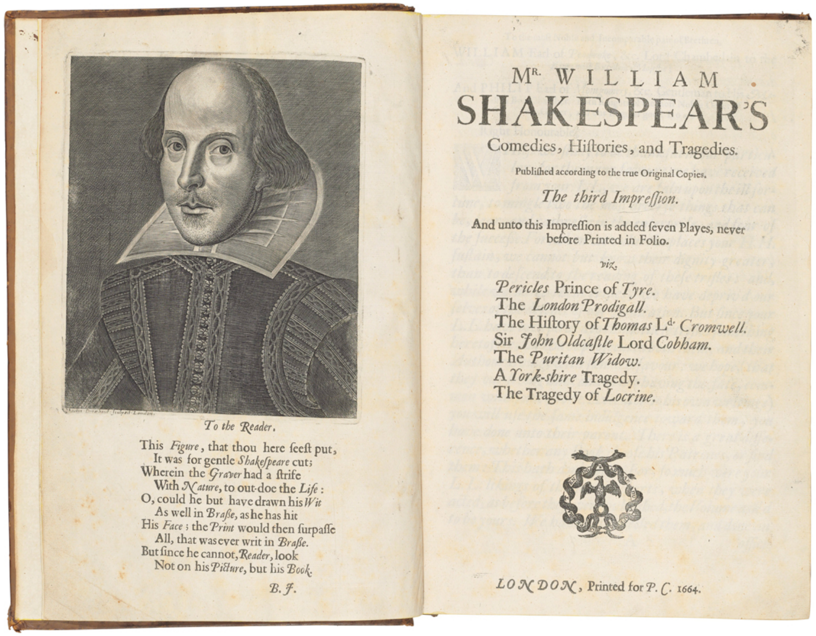
1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830

exercife	rank : 935, frequency : 15
fi	rank : 1877, frequency : 15
fecond	rank : 328, frequency : 15
defire	rank : 600, frequency : 15
pafs	rank : 556, frequency : 15
unlefs	rank : 826, frequency : 15
obfervation	rank : 1116, frequency : 15
pleafure	rank : 415, frequency : 15
neceflary	rank : 353, frequency : 15
ferve	rank : 894, frequency : 15
faw	rank : 484, frequency : 15
fent	rank : 256, frequency : 15
obferve	rank : 695, frequency : 15
fupply	rank : 851, frequency : 15
reafon	rank : 202, frequency : 15
truft	rank : 1052, frequency : 15
raifed	rank : 575, frequency : 15
confent	rank : 1151, frequency : 15
fuffer	rank : 981, frequency : 15
folid	rank : 1766, frequency : 15

Identify Cultural Shifts


1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000


Wife
Women
1910 - 1919



The Plays of William Shakespeare

961,304 words over 36 plays

Author Attribution

Look for inspiration in other data domains

Visualization in the Age of Big Data

Understand limits in current tools

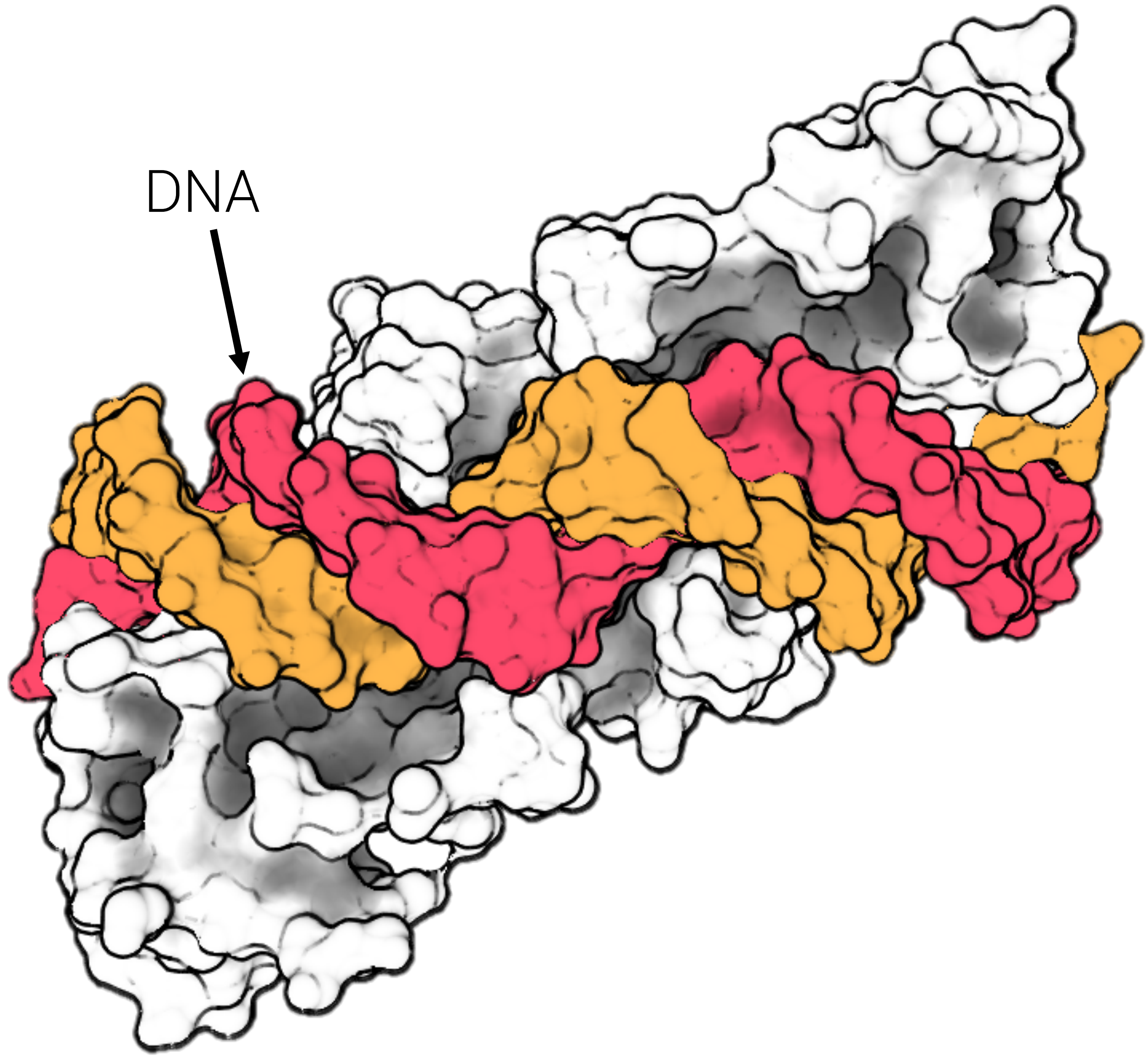
Large Scale Sequence Alignment

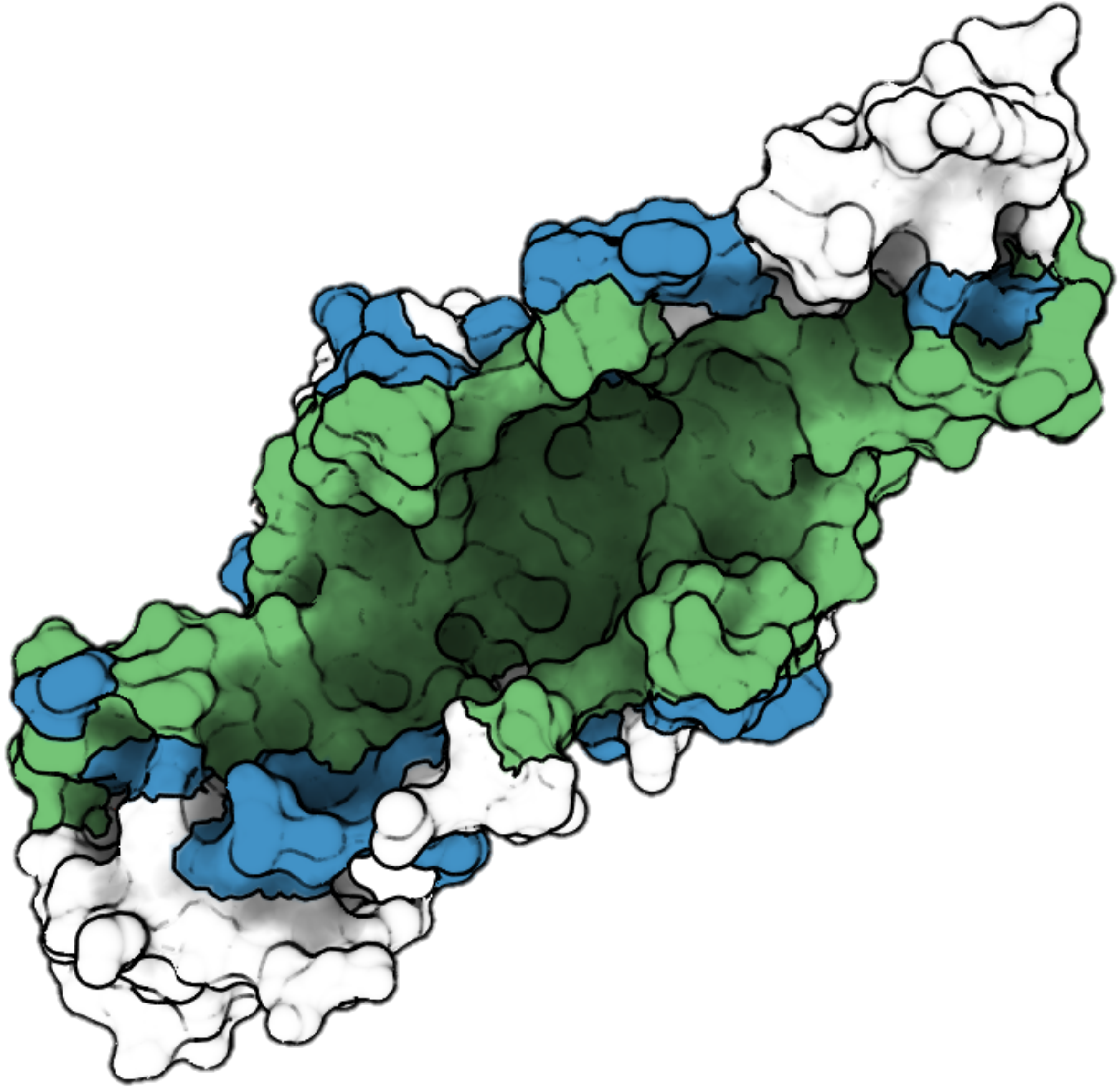
Derive inspiration across domains

Literary Patterns

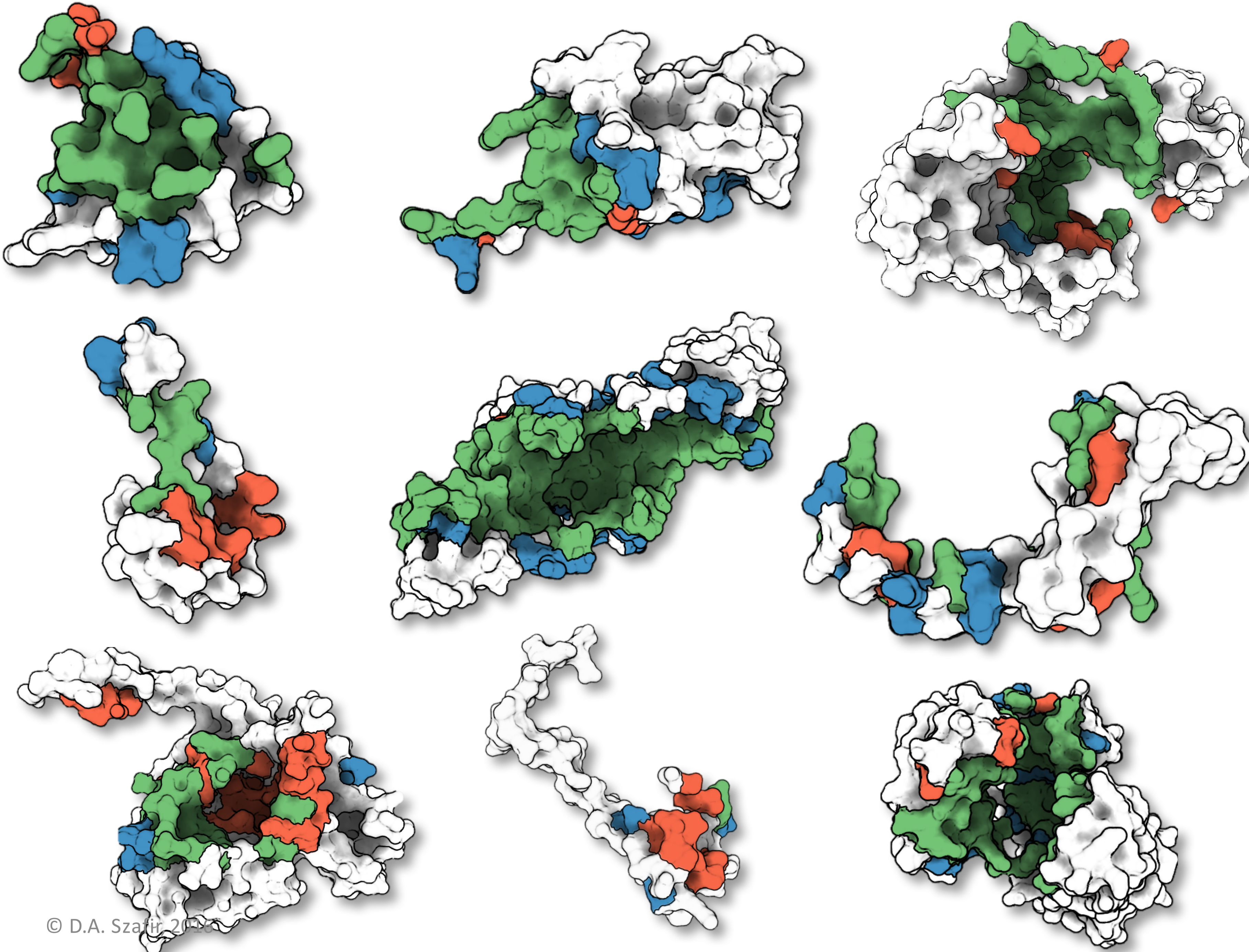
Link big and small

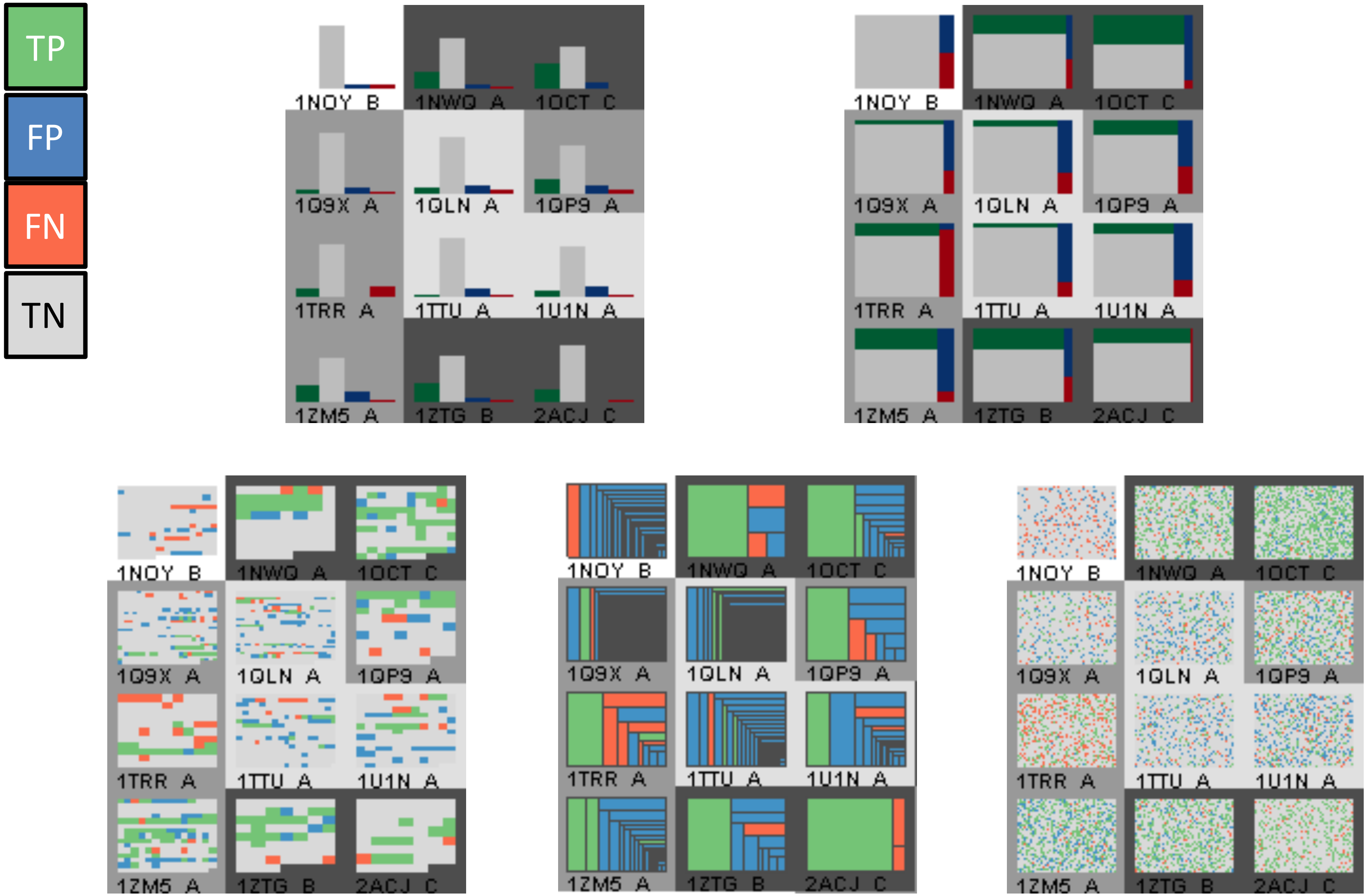
Machine Learning & Molecules


DNA

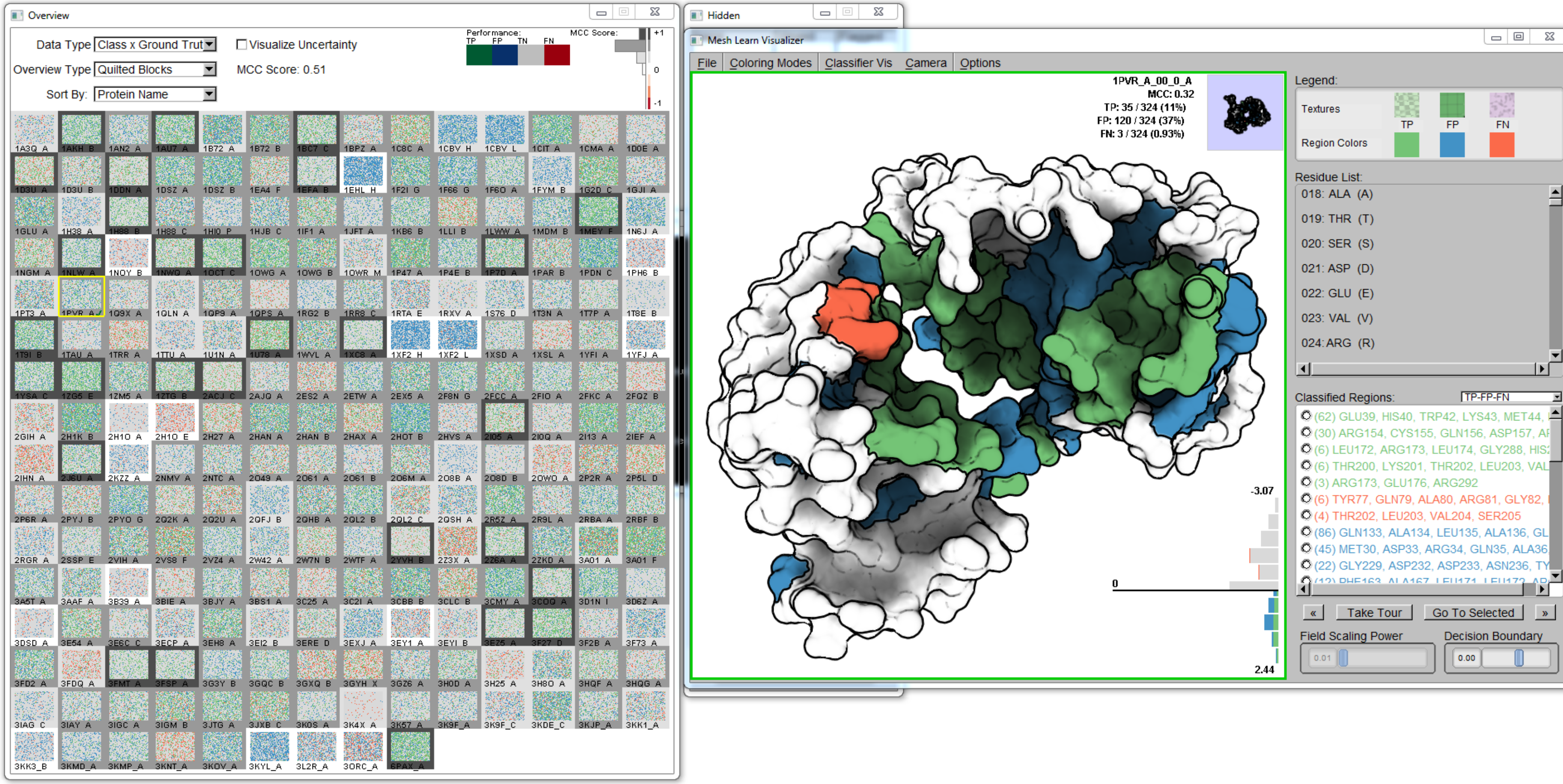


Hundreds of proteins with binding site predictions computed over hundreds of ligands

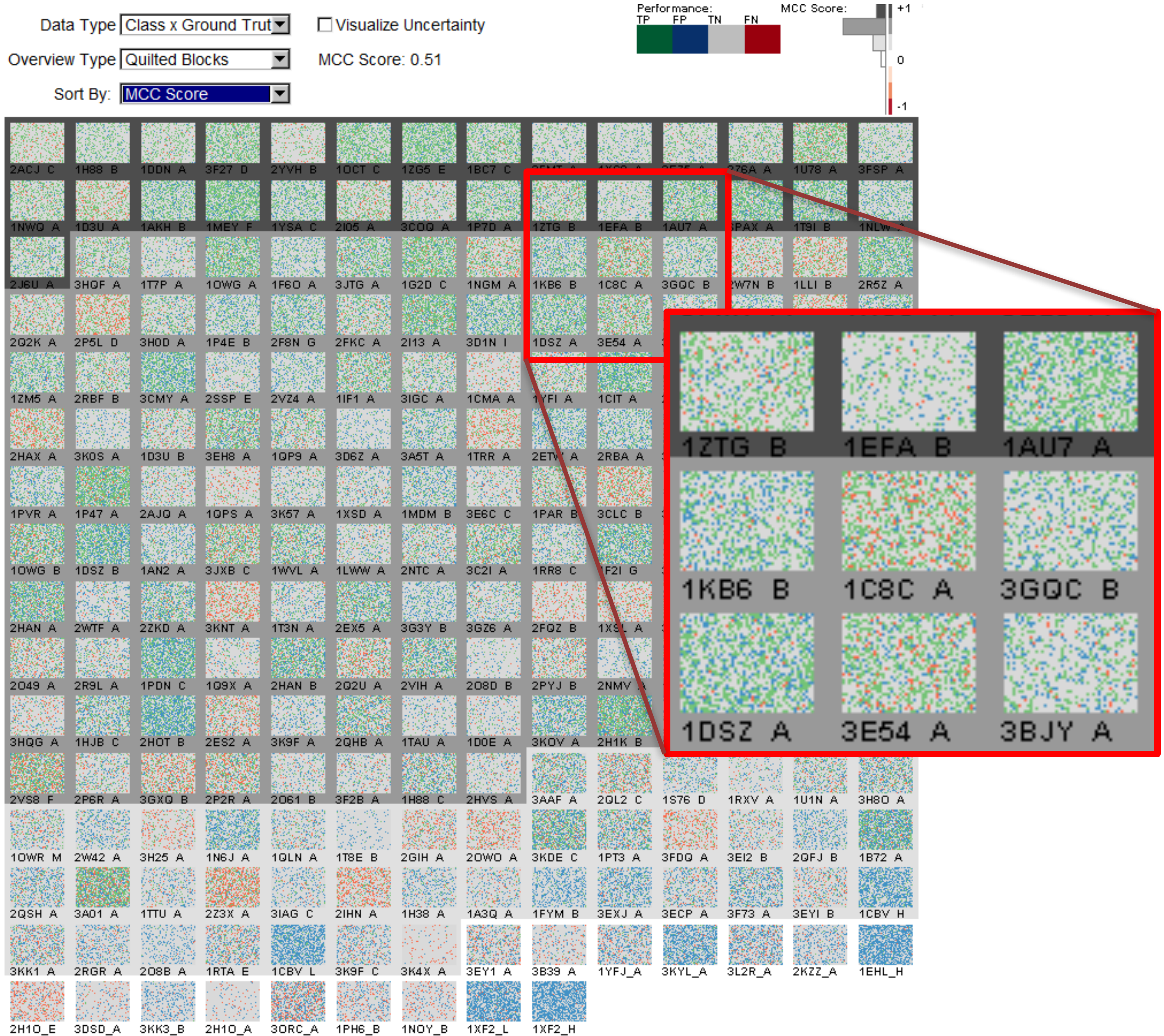


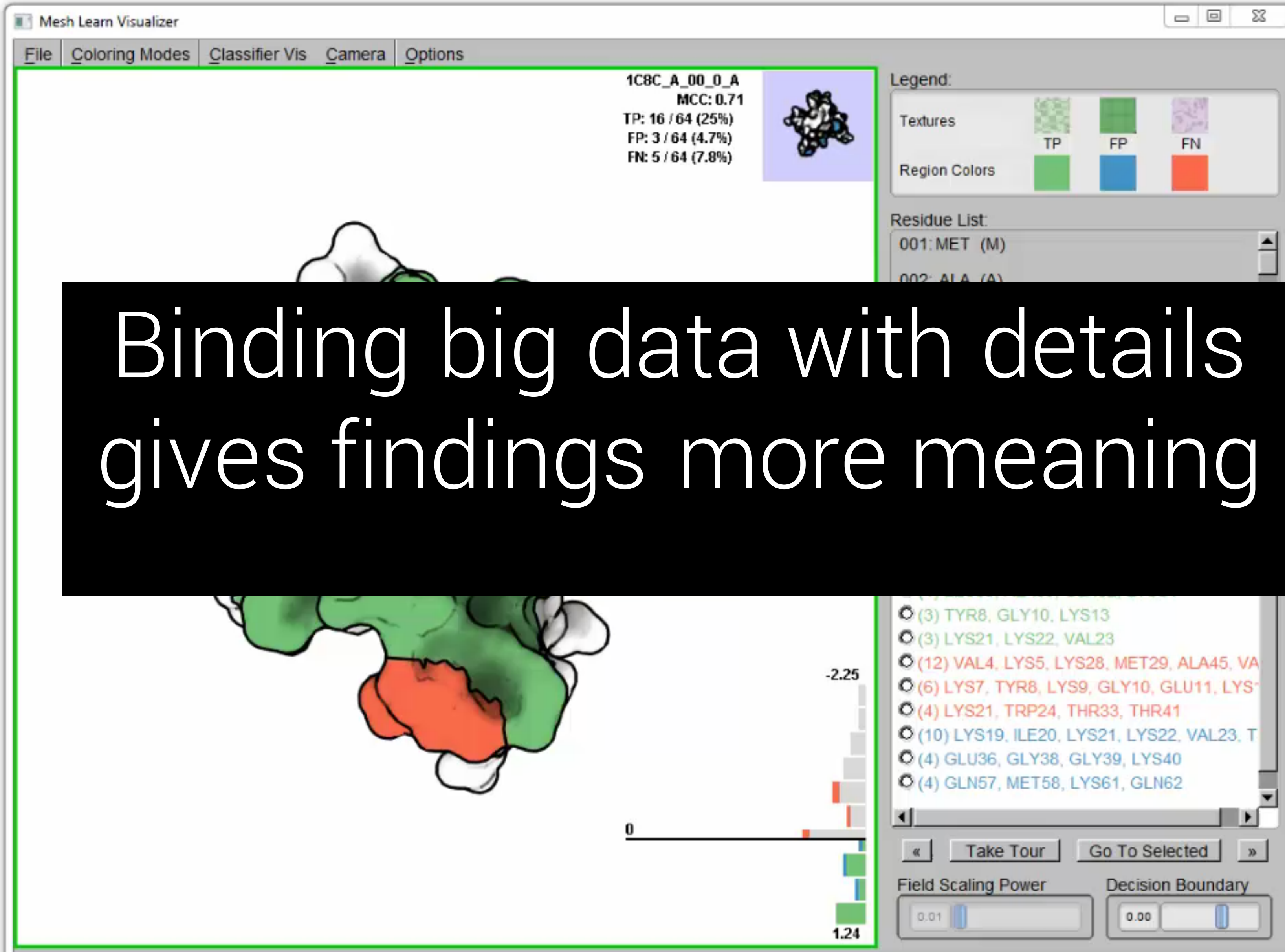


Task-driven overviews of large-scale machine learning performance data



DNA binding predictions over 216 proteins with 40 to 800 residues per protein


Data Type Class x Ground Trut
Visualize Uncertainty
Overview Type Quilted Blocks
MCC Score: 0.51
Sort By: MCC Score
Performance:
TP FP TN FN
MCC Score:
+1
0
-1
1ZTG B
1EFA B
1AU7 A
1KB6 B
1C8C A
3GQC B
1DSZ A
3E54 A
3BJY A



Binding big data with details gives findings more meaning

Visualization in the Age of Big Data

Understand limits in current tools

Large Scale Sequence Alignment

Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules

Designing for Big Data

Consider how the ways we communicate data support **high-level tasks**.

Look at **parallels** in the data structure and tasks associated with your data.

Don't lose sight of the **details**.

Thank You!

Danielle Albers Szafir
danielle.szafir@colorado.edu
@dalbersszafir

Funding by BACTER and the NSF

Demos & Papers at:
http://danielleszafir.com

Extra Slides