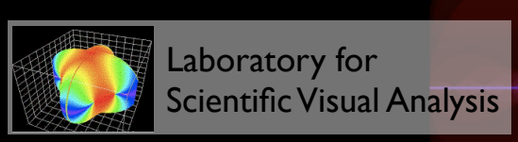
Comparative Genomics Through Visual Analytics

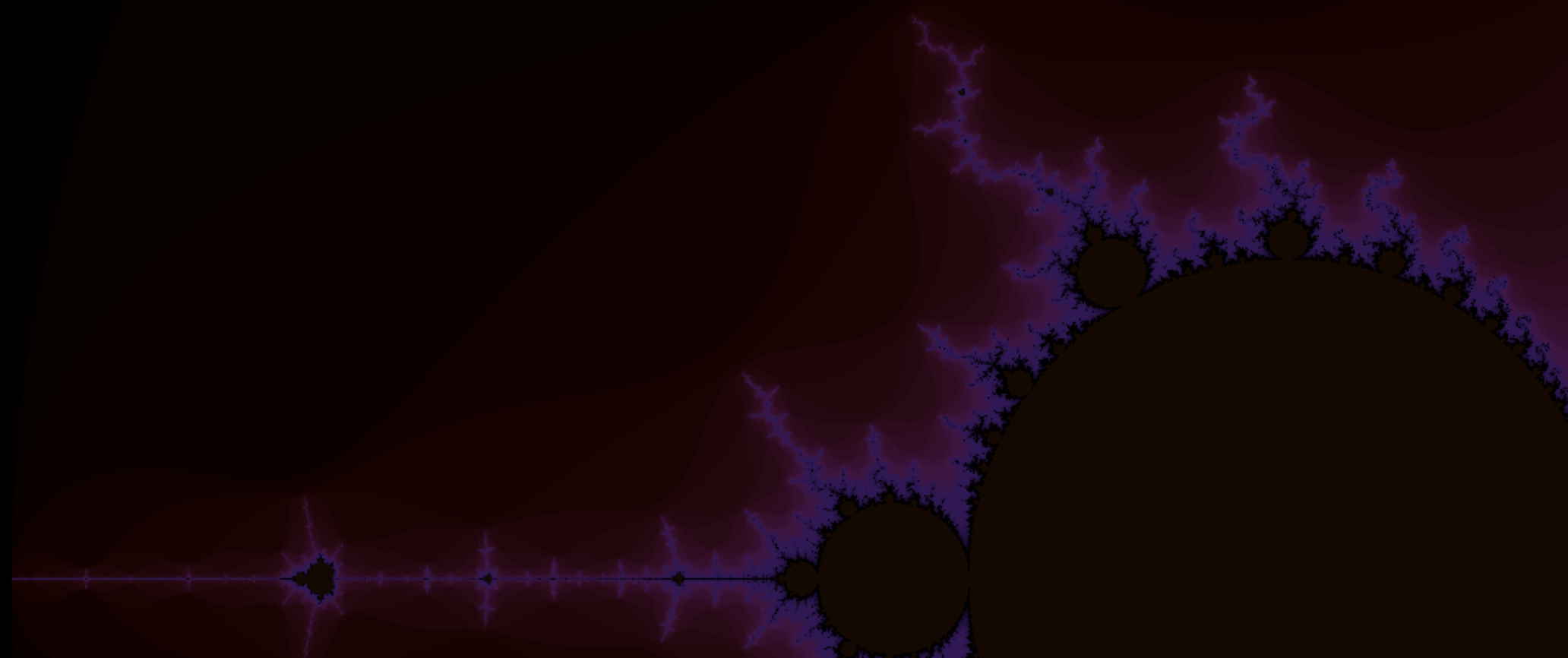
Project Isis (Integrated spatial information system)

Andrew Warren[1] and Timothy Driscoll[2]

Department of Computer Science[1] and
Graduate Program in Genetics, Bioinformatics, and Computational Biology[2]
Virginia Polytechnic Institute and State University, Blacksburg, VA



Apply Visual Analytics to Life Sciences



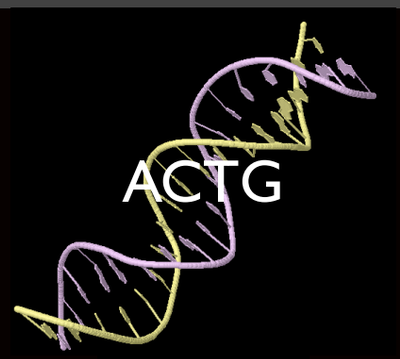
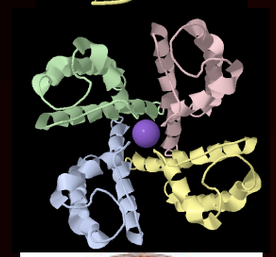
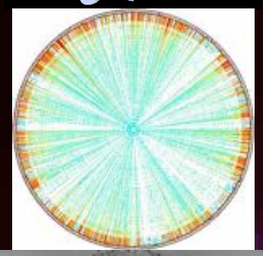
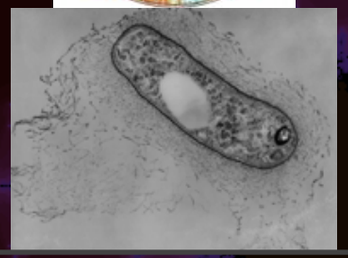
Apply Visual Analytics to Life Sciences

Create an analytical method

Extend concept into different realms

- Genomics
- Proteomics
- Metabolomics
- Regulatory systems
- Ecosystems

A Byte of Biology

	Life Scientist		Computer Scientist
DNA	Genetic material.		Alphabet for information storage.
RNA	Transcribes DNA into protein. Some characteristics of both.		Alphabet for information transfer and application.
Protein	Perform vast majority of cellular functions.		Alphabet for information application.
Genome	Total genetic content of an individual organism.		Total available alphabet.
Organism	Intricate system of interwoven components that we call life.		End result of information transfer from DNA.

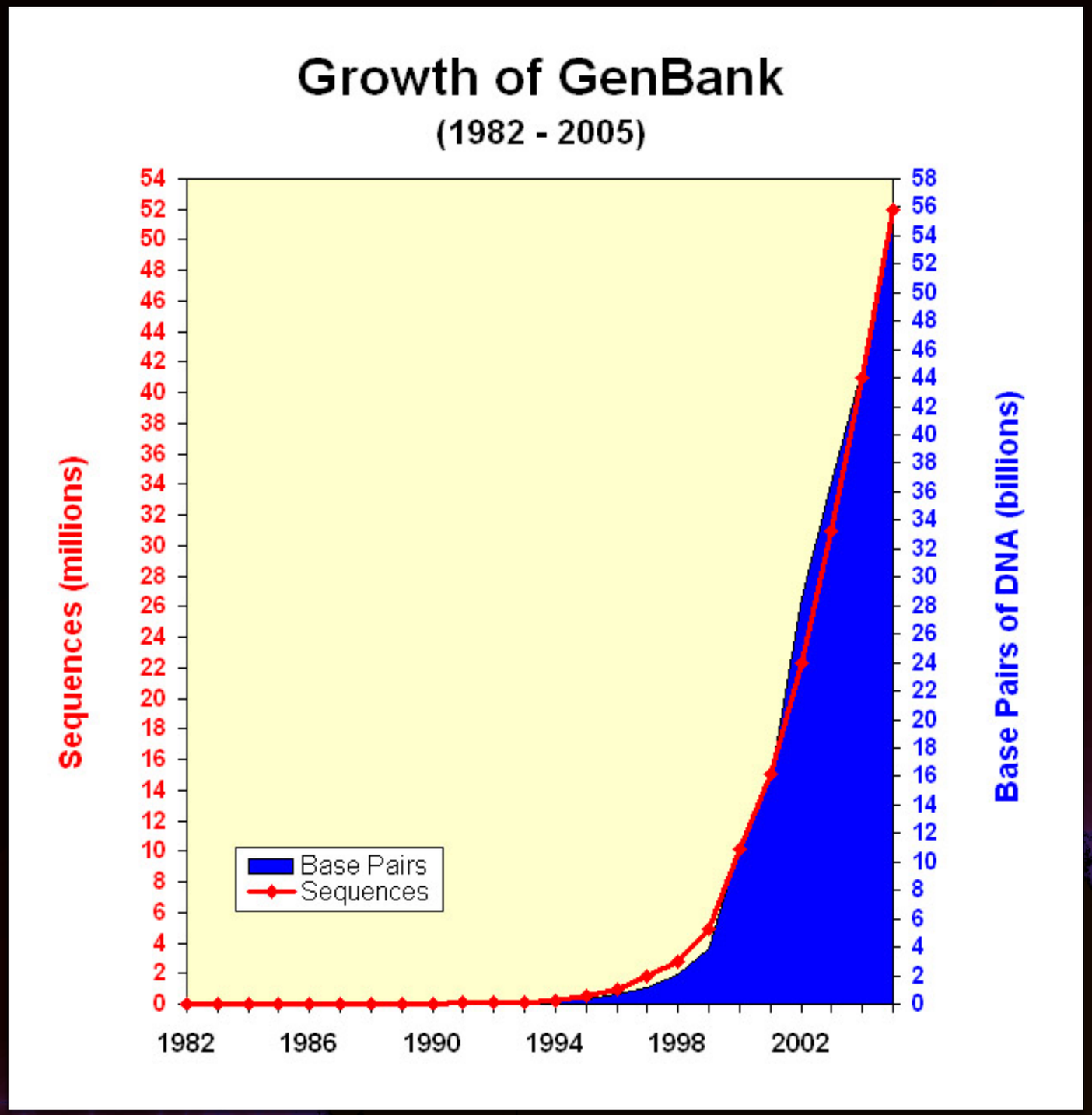
Project Rationale

Number of sequenced bacterial genomes

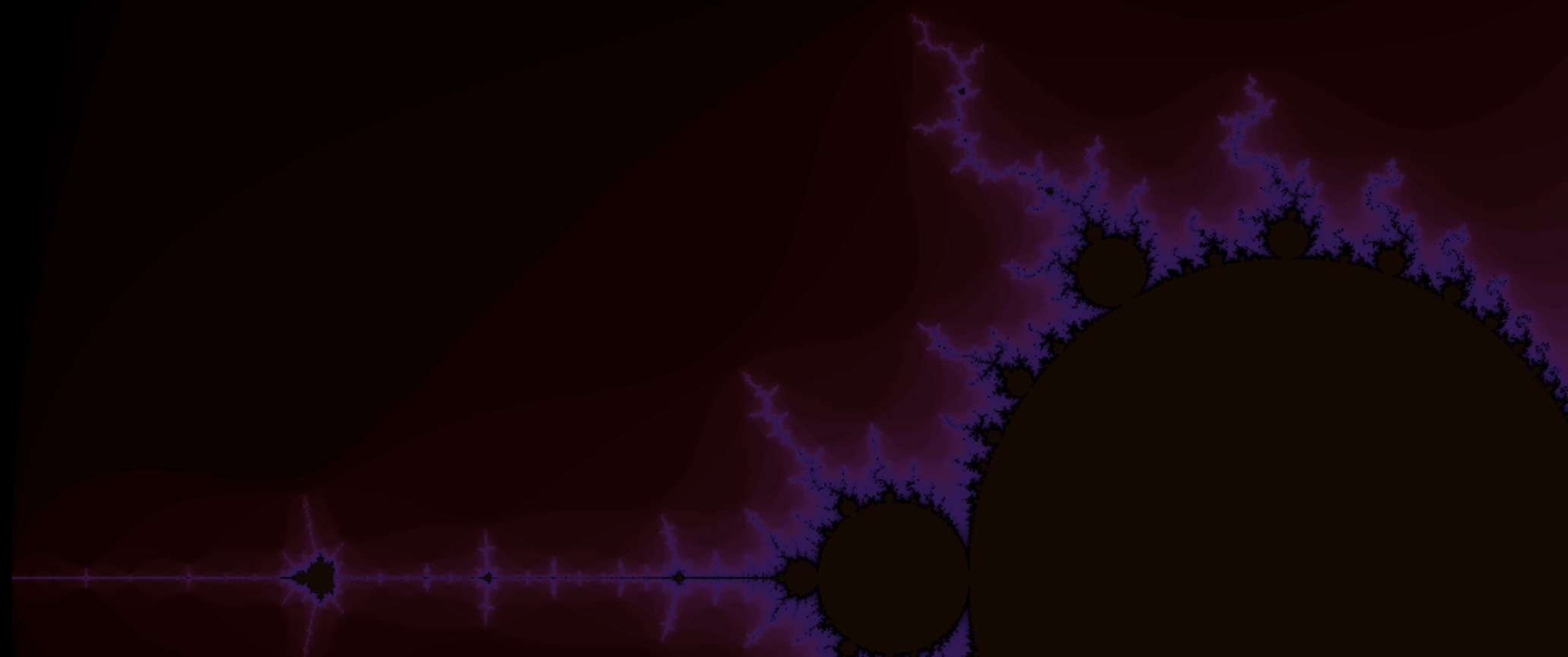
Need to compare on a genomic (not genetic) level

Gain insight through visualization

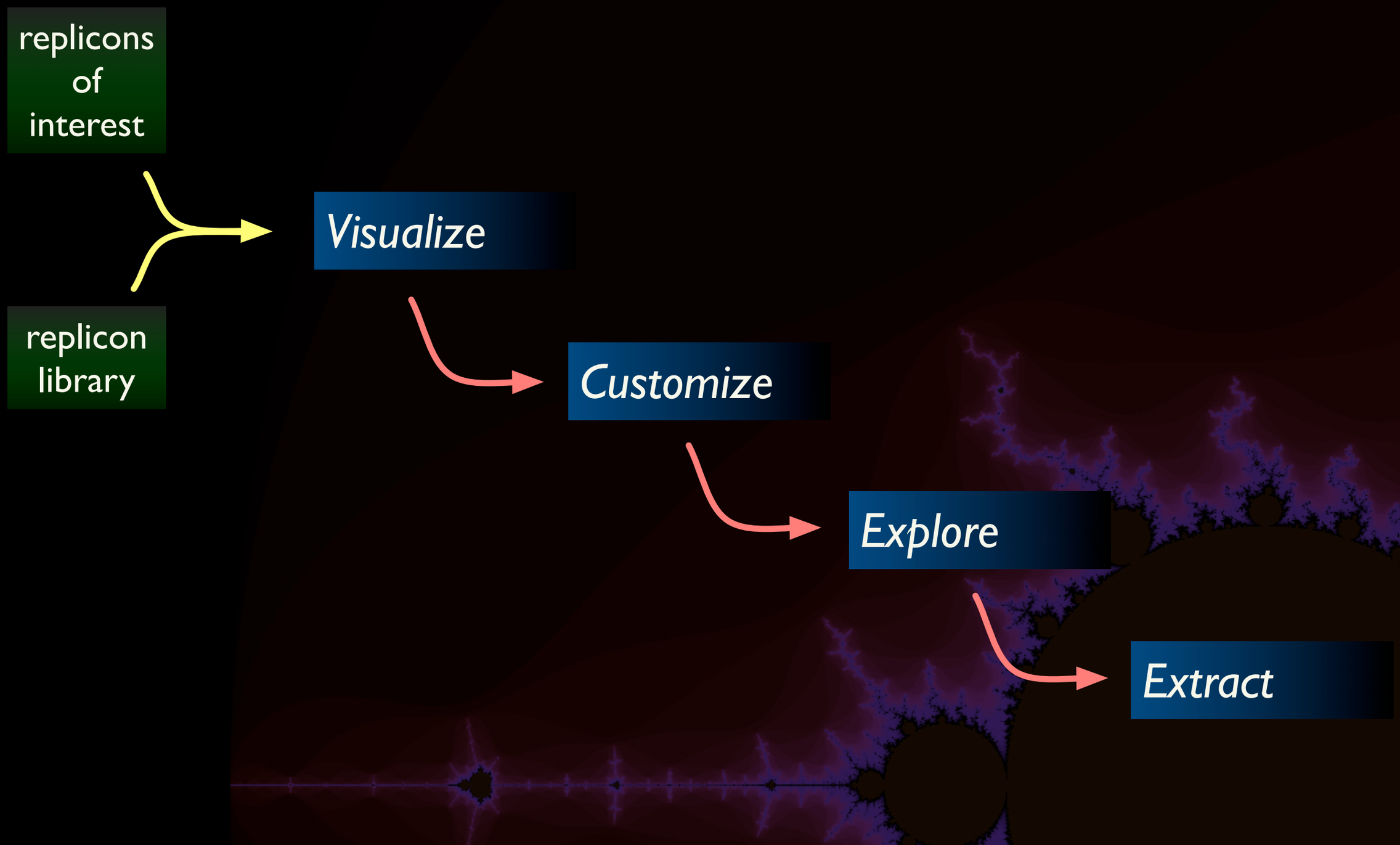
Now: data rich, tool poor



User Workflow



User Workflow



User Workflow

replicons of interest

replicon library

Visualize

Customize

Explore

Extract

Genome landscape: a 4D environment constructed from comparative metrics for a set of replicons.

Software Implementation

Visual ToolKit (VTK) for visualization

Java for interface

Benefits:

- Cross-platform
- Web friendly
- Extensive pre-built libraries

Drawbacks:

- Difficult setup and compile environment
- Combining C++ and Java may limit flexibility

Project Management Scheme

Primary input: table of *profiles*

- header row specifies metric names
- each data row gives values for a single profile

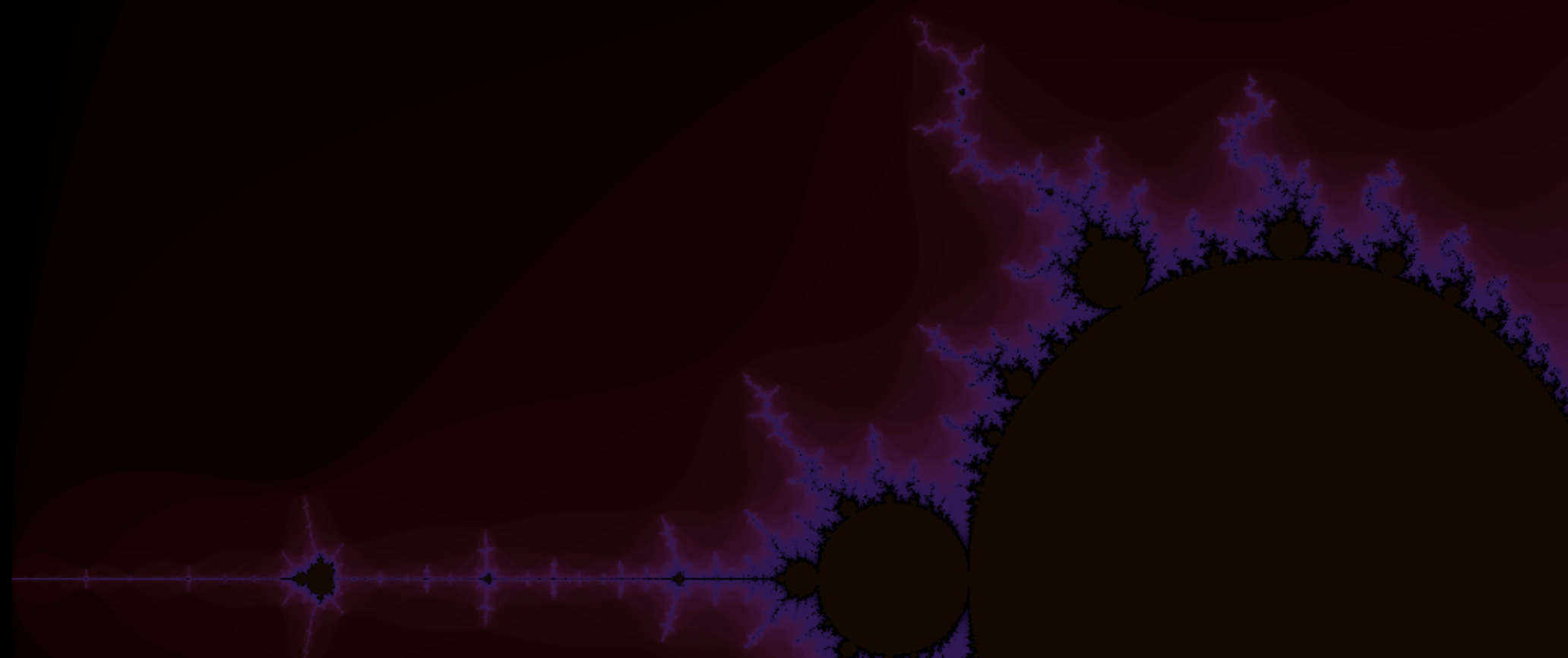
Project properties

- Assignment of metrics to axes and color dimensions
- List of all metrics that can be used for analysis (scalar values)
- Name and location of other project files

Project libraries

- *Landscape lib* stores profiles used in building the landscape
- *Sources lib* stores profiles of special interest (user-defined)

Metrics for Replicon Comparison



Metrics for Replicon Comparison

Global Calculations

- GC content
- Replicon length
- multi-ANI
- multi-AAI
- HGT events
- Extent of paralogy

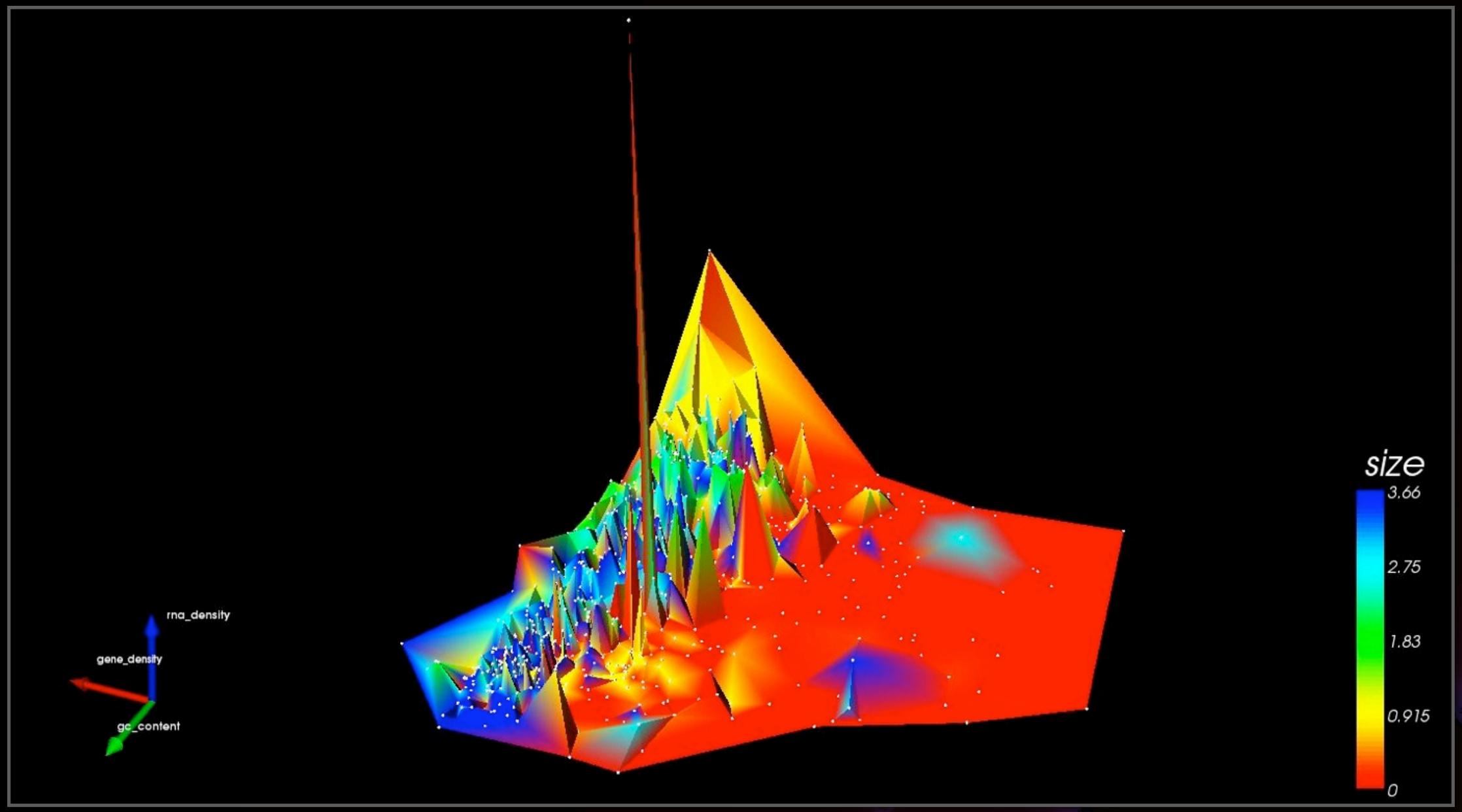
Density of Genetic Elements

- Coding regions (genes)
- RNA genes
- Promoters
- Pseudo-genes
- Transposons
- sRNA elements
- Repeat regions

Correlative & Statistical

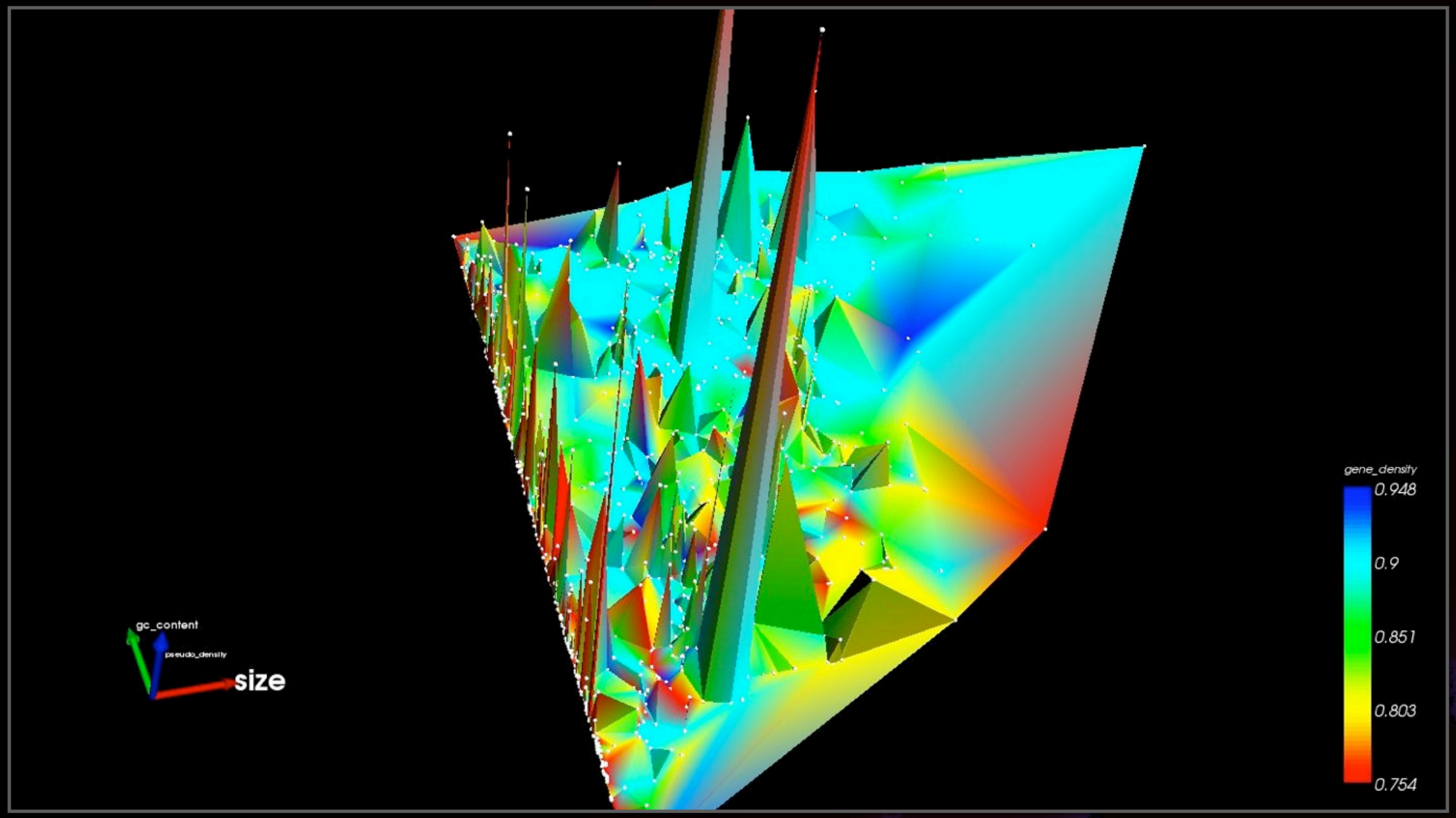
- Functional enrichment score
- Entropy density function
- Entropy distance ratio
- Phylogenetic distance
- Regulatory index
- Pathogenicity score
- Extent of orthology

Visualization Example - Genomic data



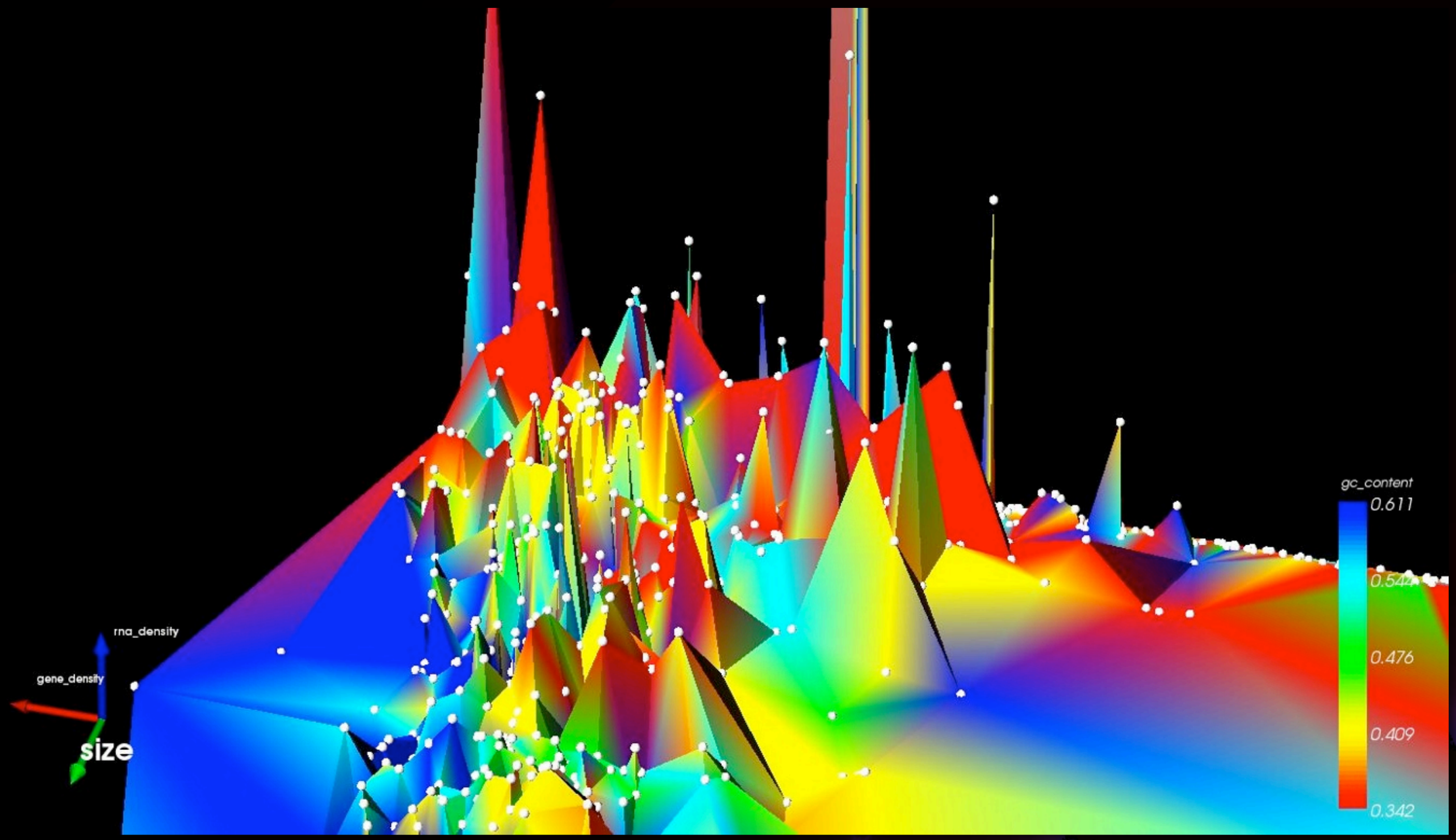
X-axis: Gene density. *Y-axis*: GC content. *Z-axis*: RNA density. Colored by replicon size (Mb).

Visualization Example - Genomic data



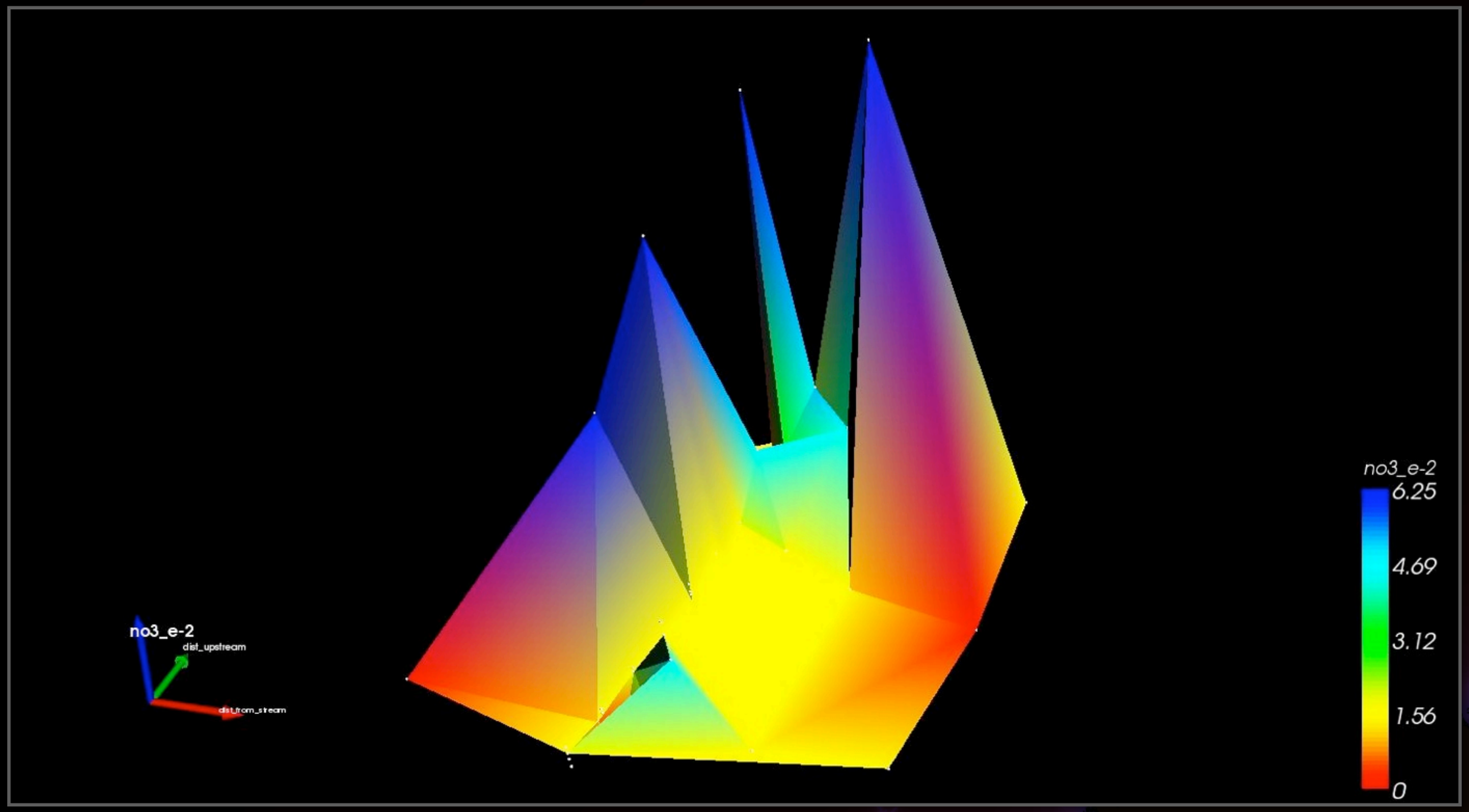
X-axis: Replicon size (Mb). *Y-axis*: GC content. *Z-axis*: Pseudo-gene density. Colored by gene density.

Visualization Example - Genomic data



X-axis: Gene density. *Y-axis*: Replicon size (Mb). *Z-axis*: RNA density. Colored by GC content.

Visualization Example - Watershed data



X-axis: Distance from stream. *Y-axis*: Distance upstream. *Z-axis*: Nitrogen content. Colored by Nitrogen content.

Project Status (1)

Interactive 4D landscape

- Full control with mouse
- Automatic data scaling and color assignment

Metrics Palette

- Shows data for all metrics in the dataset
- Highlights metrics used for XYZ axes and color
- Bindings in place for interactivity with the landscape

Project management

- New projects start with empty files, ready for data import
- Existing projects constantly updated during use
- All files are human-readable and accessible

Project Status (2)

Data formats

- Full implementation of (internal) data handling
- Standard tab-delimited format for data storage
- Standard key-value pairs for preference storage
- Easy to read, export-friendly, minimal redundancy

Being applied to disparate datasets

- Plug-in scripts for converting GenBank data to Isis format
- Metrics for comparative genomics chose; some applied
- Straight export of ecosystem data (nitrogen content)

Future work: Project Development

Selection widget ("Mother Ship")

- Graphic element to focus on groups of points in the landscape
- User controls size and shape of widget

Dynamic scale indicators

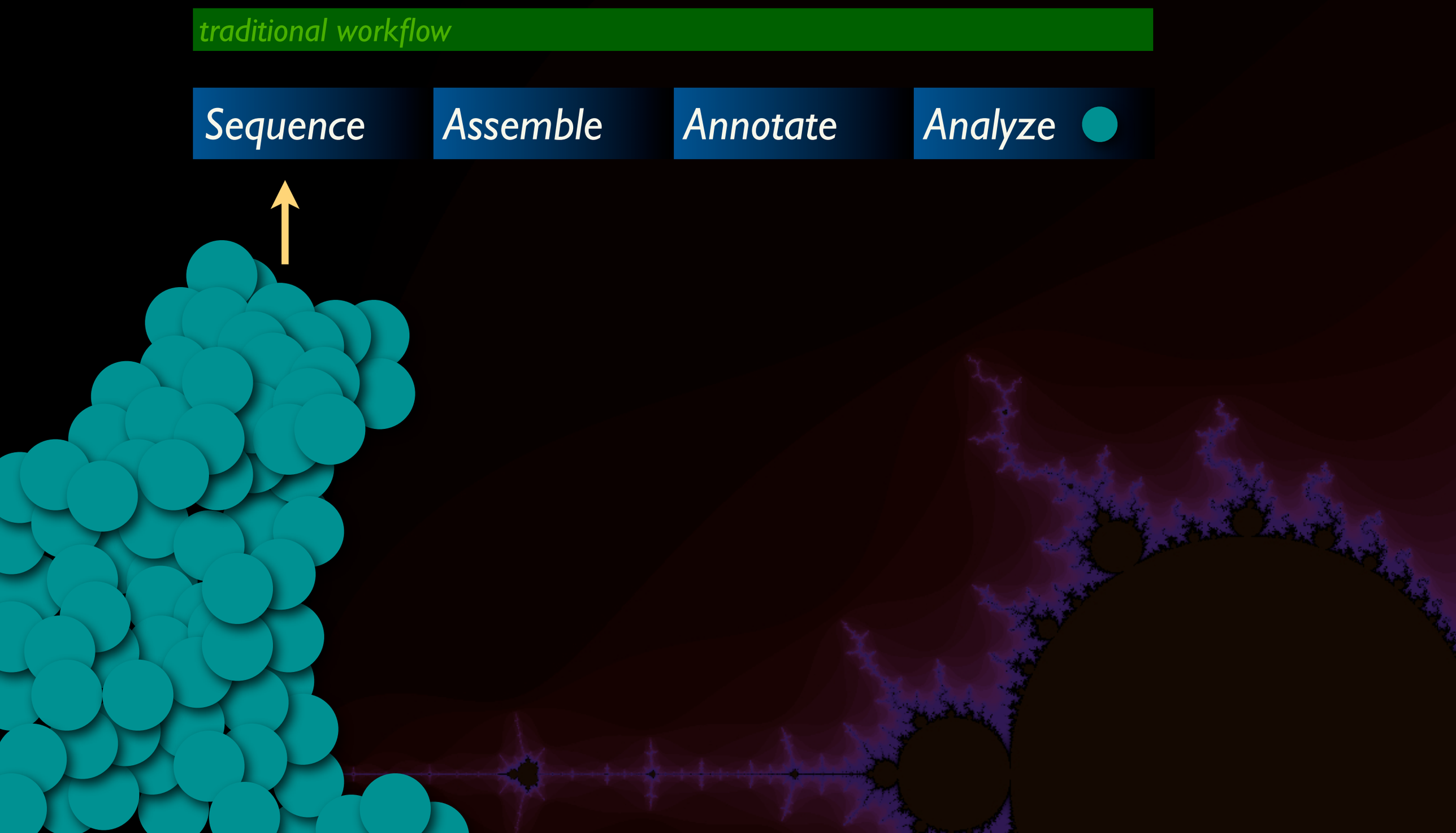
- "Towers of Power"
- Show average of values for points within the selection widget
- Dynamic color bar for visual comparison of selection to total

Axis reassignment by drag-and-drop

Control over landscape display

- Transparency and smoothness of surface
- Axis scaling
- Color bar scaling

Future work: VBI and 454 Life Sciences



Future work: VBI and 454 Life Sciences