

The State of Galaxy

Dannon Baker
The Galaxy Team & Emory University

What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

<http://galaxyproject.org>

Vision

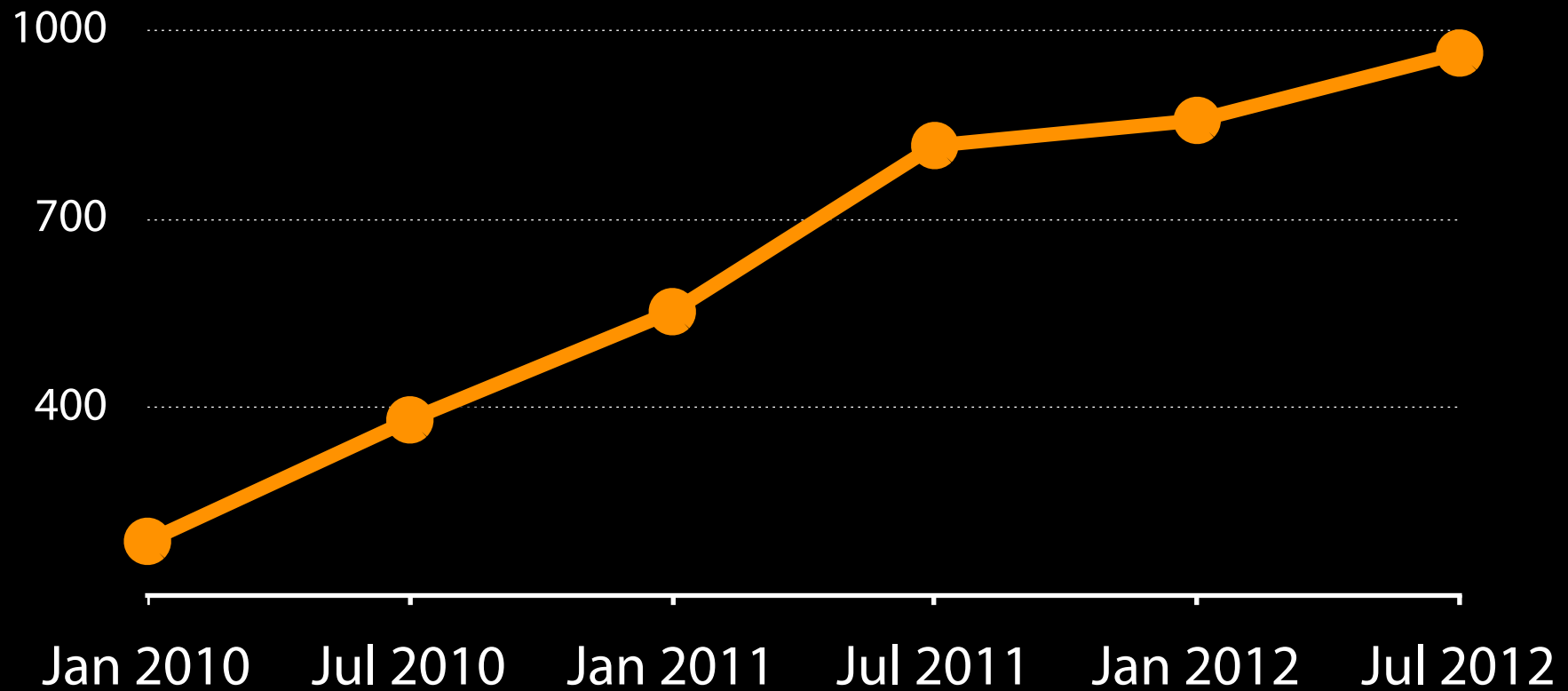
Galaxy is an **open, Web-based platform** for accessible, reproducible, and transparent computational biomedical research

How's Galaxy Doing?

- Growth
- Scaling
- Where we're headed

<http://galaxyproject.org>

New Users per Month

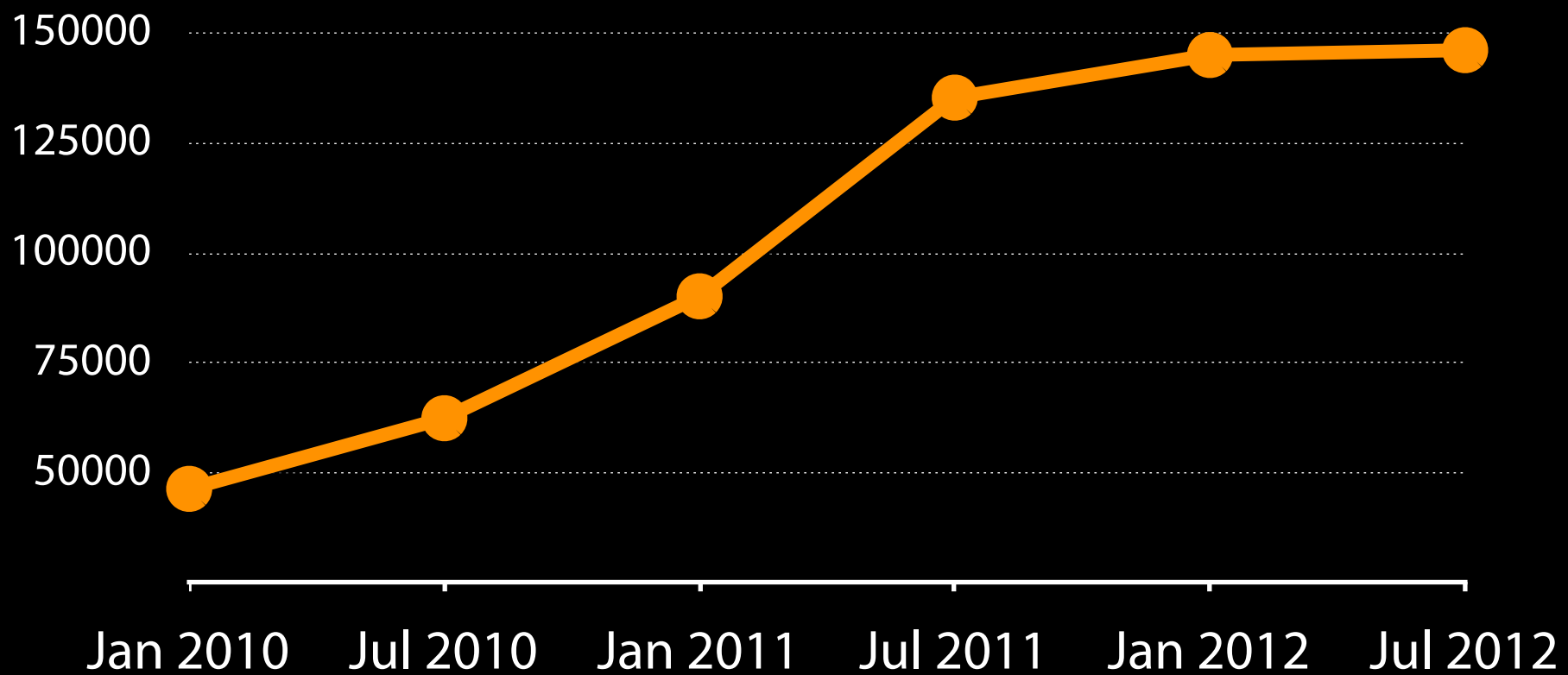


Users

25,768

total users registered

Jobs per Month



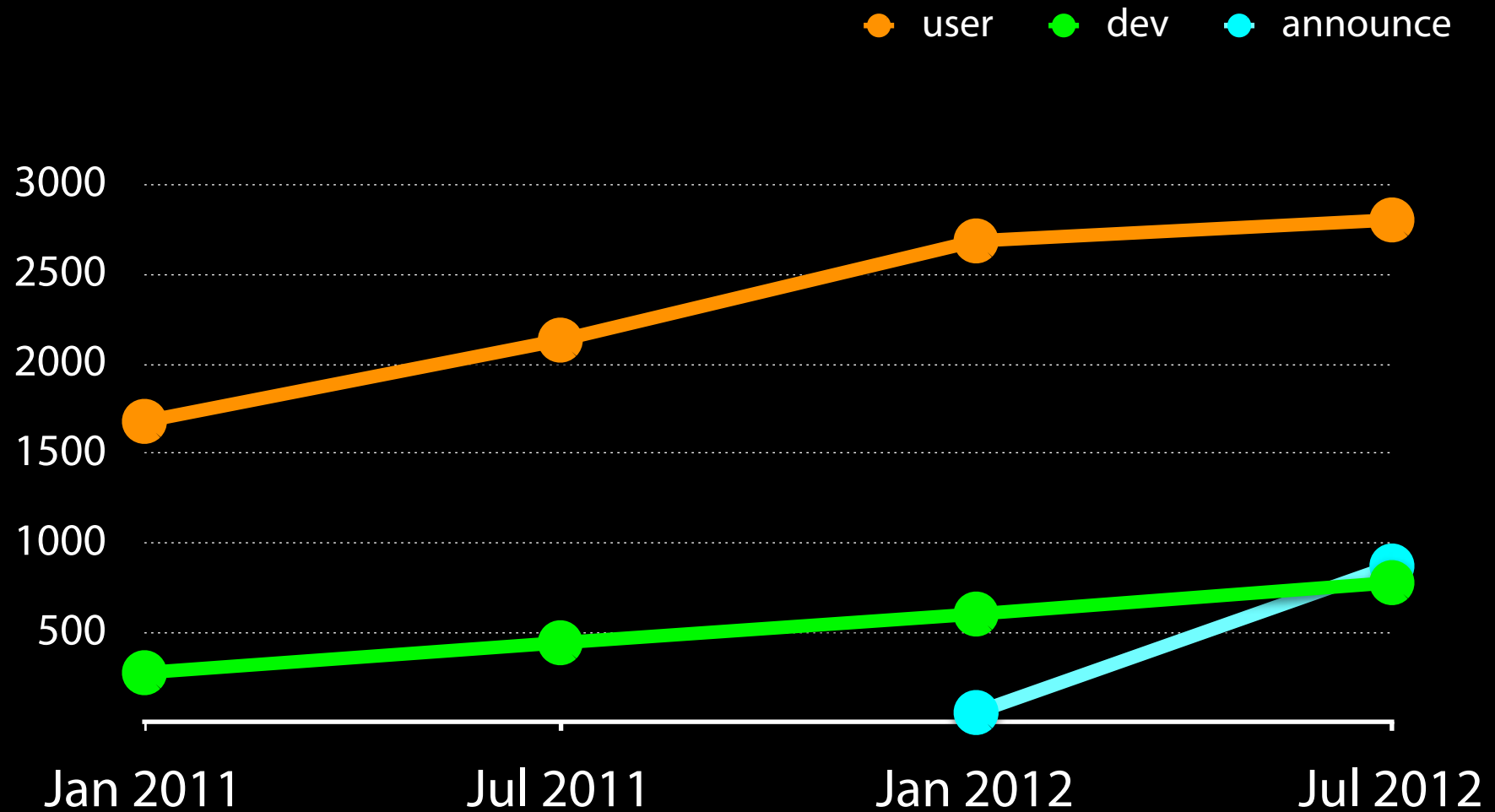
0.3 Pb

user data at the main site

11,442

jobs in a single day

Mailing List Membership



~180
papers using galaxy

STEM CELLS®

STEM CELL TECHNOLOGY: EPIGENETICS, GENOMICS, PROTEOMICS AND METABONOMICS

Stochastic Choice of Allelic Expression in Human Neural Stem Cells

Aaron R Jeffries¹, Leo W Perfect¹, Julia Ledderose¹, Leonard C Schalkwyk², Nicholas J Bray¹, Jonathan Mill², Jack Price

U. United Kingdom;

U.F. United Kingdom

NA Methylation •

© 2012 Wiley Periodicals, Inc.

Birth Defects Research (Part A) 00:000–000 (2012)

Responses of the Embryonic Epigenome to Maternal Diabetes

J. Michael Salbaum^{1*} and Claudia Kappen²

Open

Cell Death and Differentiation (2012), 1–11

© 2012 Macmillan Publishers Limited. All rights reserved. 1350-9047/12

www.nature.com/cdd

Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis

F Nikolenkov¹, C Spinner^{1,2}, H Li^{1,2}, C Tonelli¹, Y Shi¹, M Turunen^{2,3}, T Kivioja^{2,3}, I Ignatiev¹, A Kei^{4,5}, J Taipale^{2,3} and G Selivanova^{4,1}

The tumor-suppressor p53 can induce various biological responses. Yet, it is not clear whether it is p53 in vivo promoter is of genome-wide chromatin /, that is, the pattern of major tumor cell apoptosis (RITA) or of gene expression allowed is one of the p53 modulators, and that STAT3 antagonizes

is unclear port a ge- rith differ- t. We find cases and tion marks : data sug- ed risk for ind mater- nditions is lyses that le high-fat ges in his- nisms in rth Defects

Reorganization of the host epigenome by a viral oncogene

Roberto Ferrari,^{1,2} Trent Su,^{1,3} Bing Li,¹ Giancarlo Bonora,^{1,4} Amit Oberai,¹ Yvonne Chan,¹ Rajkumar Sasidharan,⁵ Arnold J. Berk,^{6,7} Matteo Pellegrini,^{2,5} and Siavash K. Kurdiani^{1,2,7,8,9}

California, Los Angeles, California 90095, USA; ²El- nity of California, Los Angeles, California 90095, ol of Medicine, University of California, Los Angeles, David Geffen School of Medicine, University of r, and Developmental Biology, University of nology and Molecular Genetics, University of rsity of California, Los Angeles, California 90095, of Medicine, University of California, Los Angeles,



GENOME
RESEARCH

Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis

Ozren Bogdanovic, Ana Fernandez-Minan, Juan J Tena, et al.

Genome Res. published online May 16, 2012
Access the most recent version at doi:10.1101/gr.134833.111

Specialized Public Galaxy Servers

<http://galaxyproject.org/wiki/PublicGalaxyServers>

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Sequence and tiling arrays?

✓ Oqtans

Text Mining?

✓ DBCLS Galaxy

Reasoning with ontologies?

✓ GO Galaxy

Internally symmetric protein structures?

✓ SymD

Other Known Celestial Bodies

<http://synd.nci.nih.gov/>

<http://galaxy.tuebingen.mpg.de/>

<http://preview.galaxy.dbcls.jp/>

<http://gwips.ucc.ie/>

<http://urgi.versailles.inra.fr/galaxy>

<http://mgtaxa.jcvi.org/>

<http://cistrome.org/>

<http://galaxy.genenetwork.org/>

<http://nebula.curie.fr/>

<http://www.genboree.org/galaxy/>

<http://galaxy.wur.nl/>

<http://galaxy.nmcdsp.org/>

<http://hyperbrowser.uio.no/hb/>

<http://galaxy.berkeleybop.org/>

<http://galaxy.nbic.nl/>

<http://discovery.hsci.harvard.edu/galaxy>

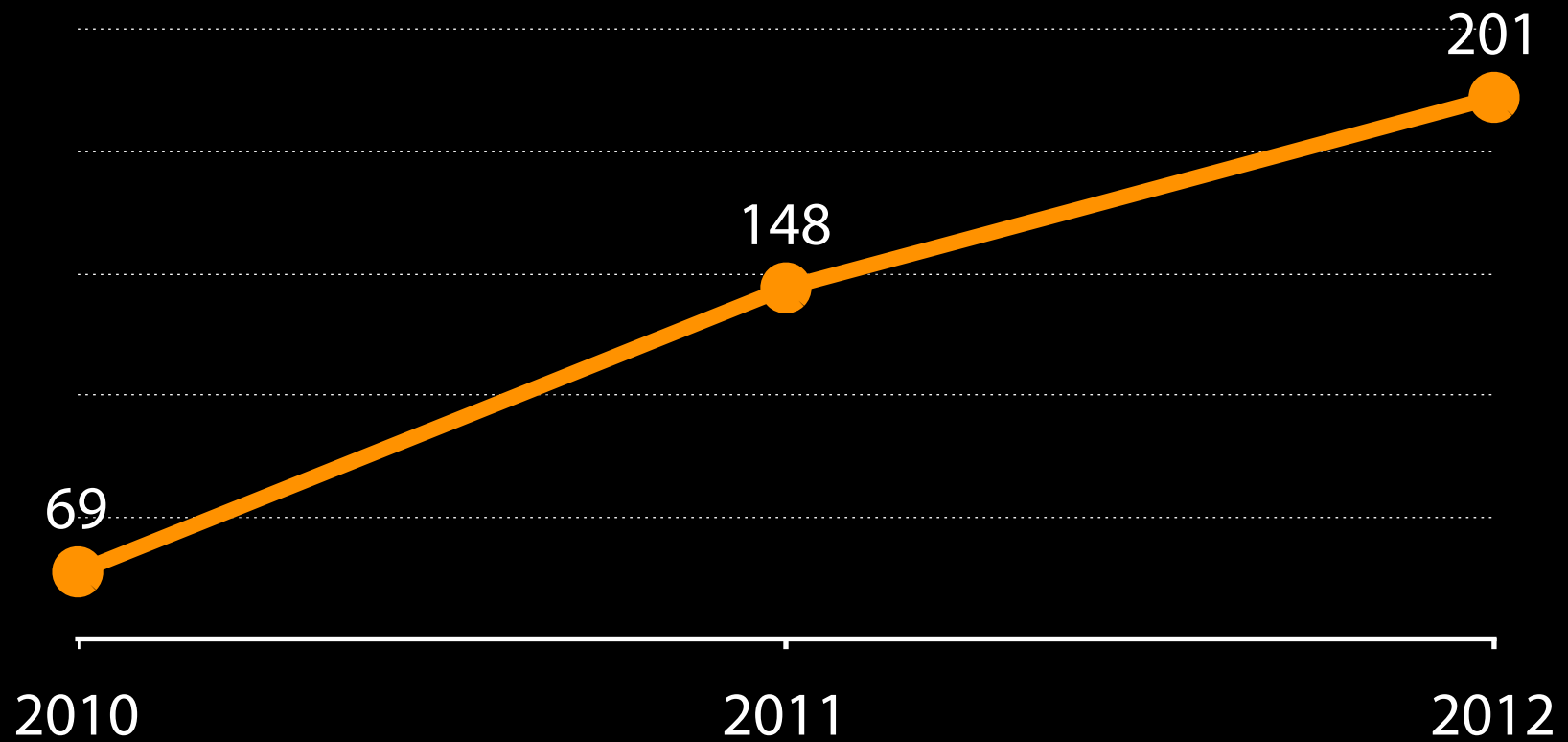
<http://regulatorygenomics.upf.edu/galaxy>

<http://huttenhower.sph.harvard.edu/lefse/>

~250

commits from the community

GCC attendance through the years



Scaling Galaxy: two distinct problems

- So much data, not enough infrastructure.
 - Solution, encourage local Galaxy instances, cloud Galaxy, support increasingly decentralized model, *improve access to exiting resources*
- So many tools and workflows, not enough manpower
 - Focus on building infrastructure to allow community to integrate and share tools, workflows, and best practices

MAJOR PUSHES

Wednesday, October 3, 12

2,124 Tools

Dependency
management

TOOLS SHED

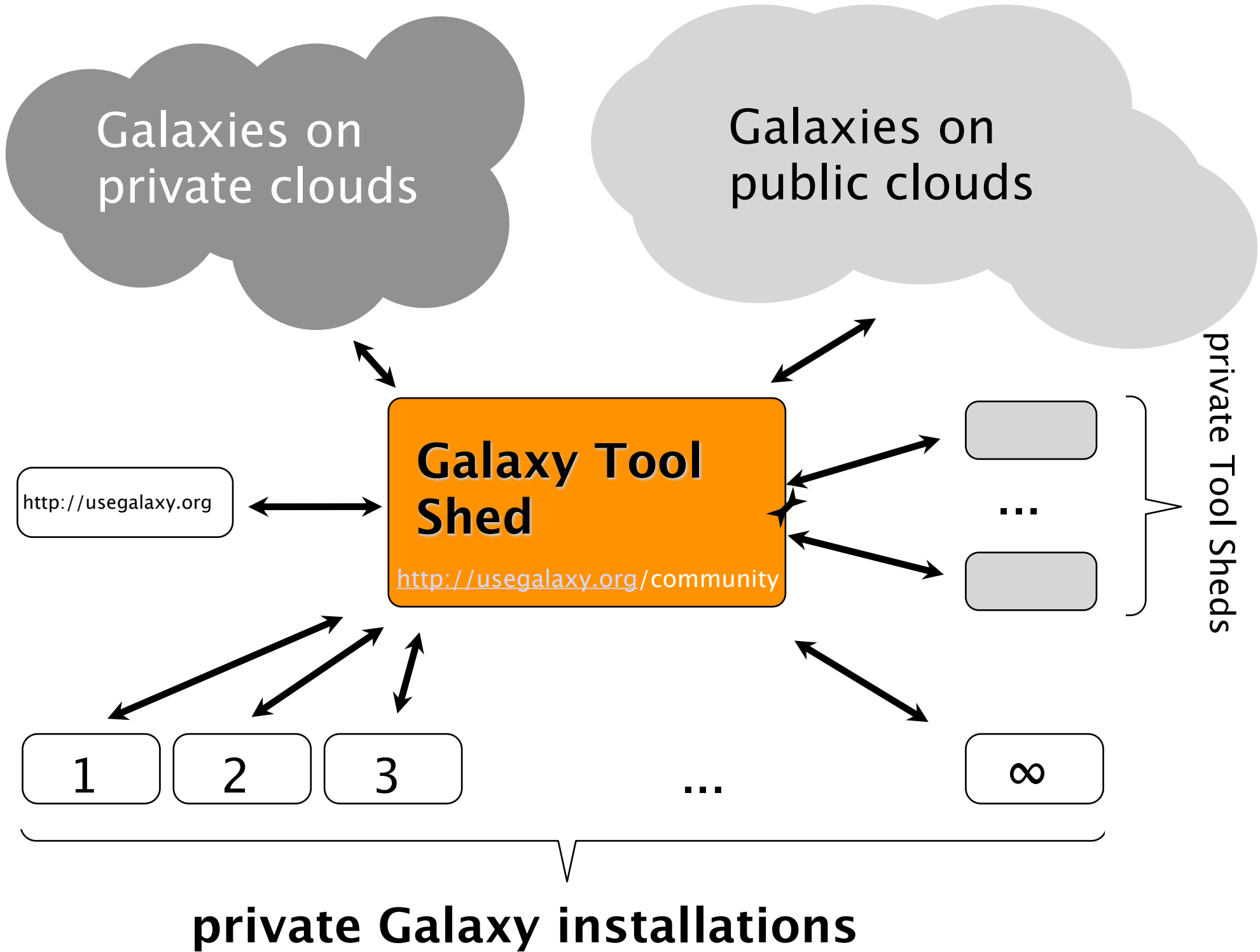
Intergalactic
HUB

Cite yourself

Galaxy Tool Shed

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Version controlled - multiple versions installed
- Community driven - rating, annotation, comments, review
- Integration with Galaxy instances to automate tool installation and updates

<http://usegalaxy.org/community>



Galaxy Tool Shed

- Multiple versions of the same tool installed
- Cite yourself - Now, you can cite the actual tool with the toolshed link and credit goes straight to the author

<http://usegalaxy.org/community>

Cloud

Intelligent parallelism

XSEDE

Object Store

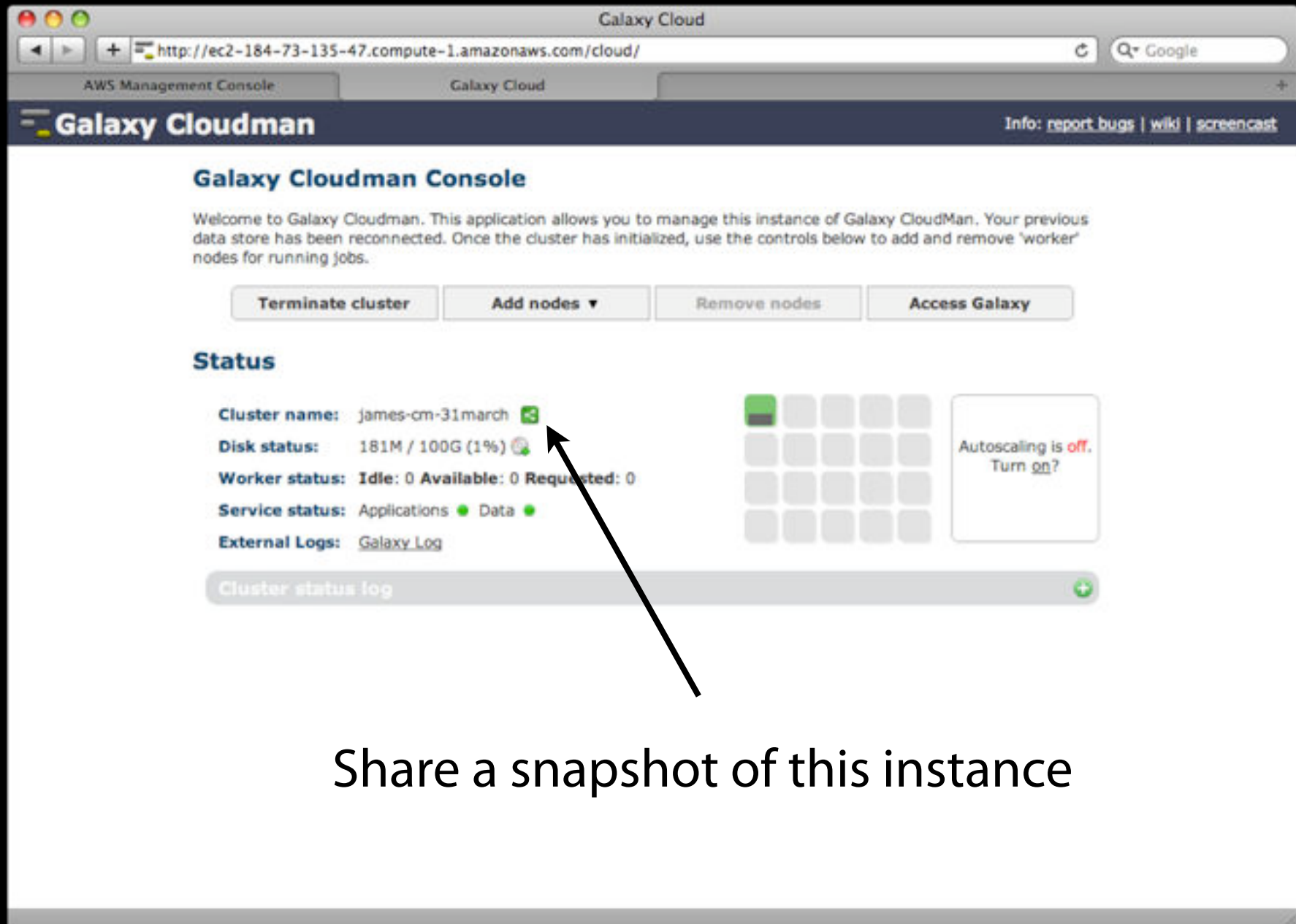
Galaxy CloudMan

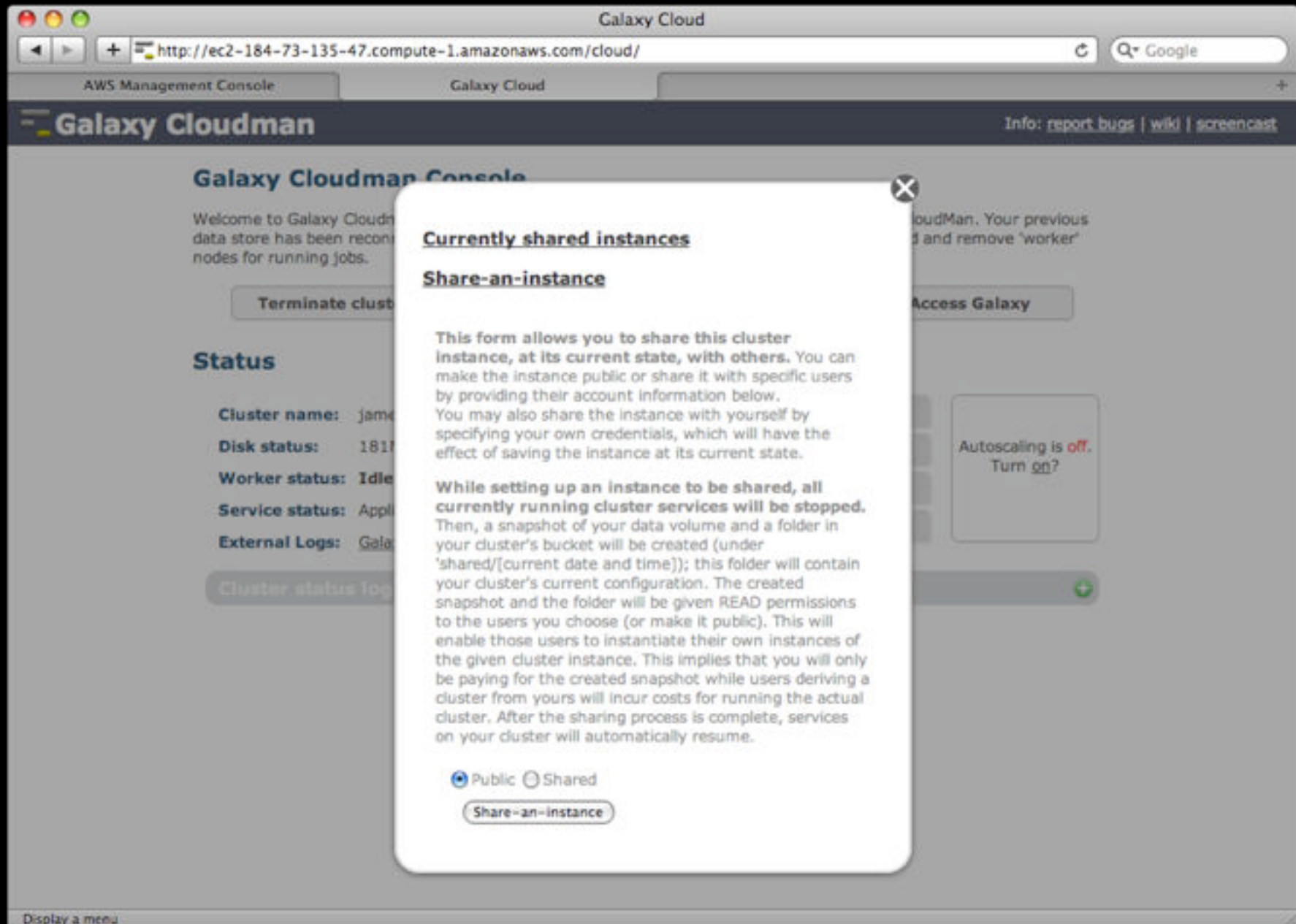
<http://usegalaxy.org/cloud>

- Start with a fully configured and populated (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.



<http://aws.amazon.com/education>





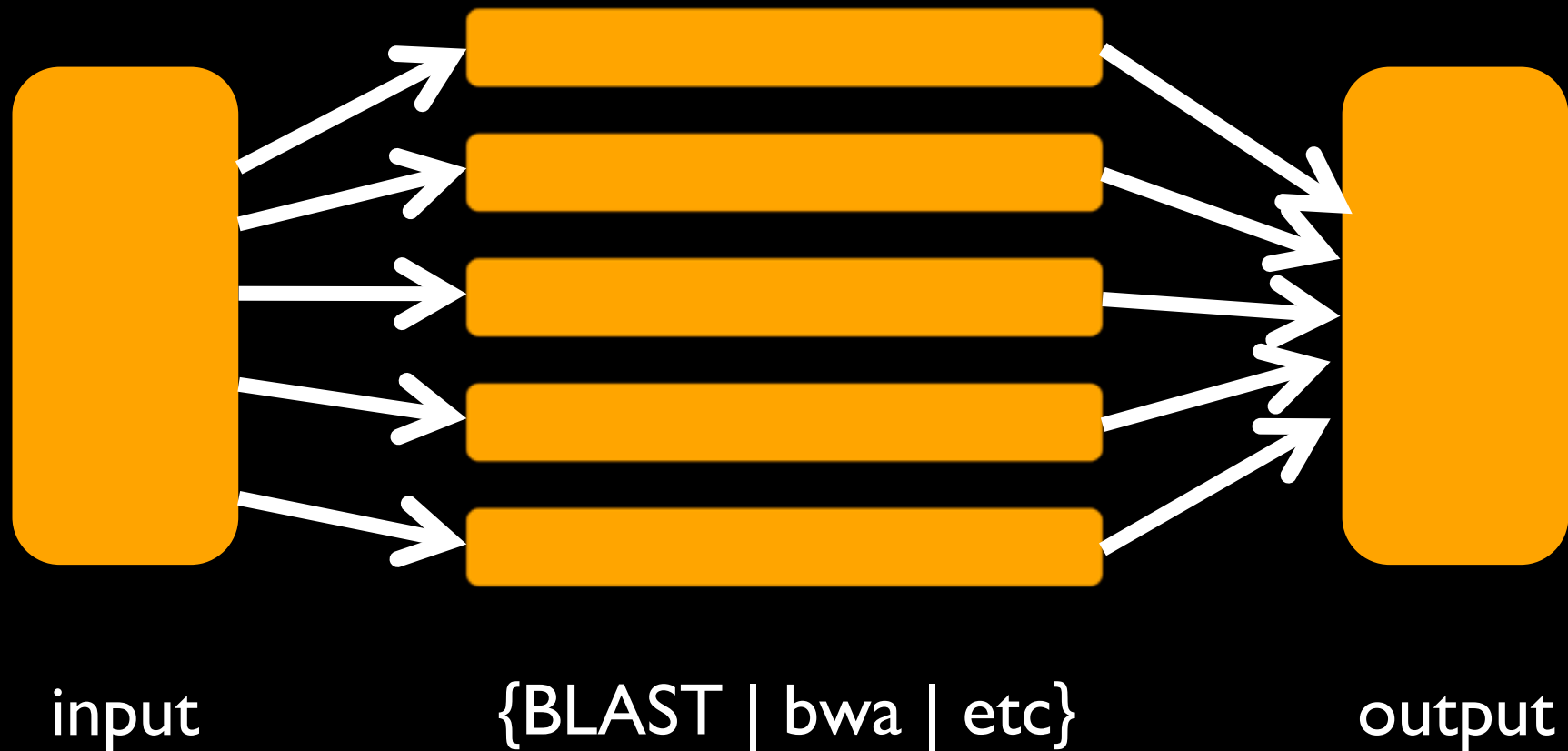
Parallelism

- As data sizes grow, increasingly important to be able to express within tool parallelism
- Naturally parallel (split/join) constructs can be specified in configuration
- Parallel environments (MPI) can be used, but management delegated to underlying resources

Parallelism - why?

- Take maximum advantage of available resources
- Less costly fault recovery (spot instances)
- Overhead in splitting time
- Increased temporary storage requirement

Automatic Parallelism



Use it now

`use_tasked_jobs = True`

Tools supported:

BLAST, BWA, Bowtie

Yours?

Implement for your tool

```
<parallelism
```

```
    method = "multi"
```

```
    split_inputs = "query"
```

```
    split_mode = "number_of_parts"
```

```
    split_size = "4"
```

```
    shared_inputs = "subject"
```

```
    merge_outputs = "output1"
```

```
/>
```

Advanced Splitting

- FUSE
 - No disk write required
 - But it is slightly slower to read
- More splitters/data providers:
 - Chromosome based?

```
yoplait@gizmo ~/work/mountpoint ls
all.seqs          sample.tabular
yoplait@gizmo ~/work/mountpoint ls all.seqs/
0  113 129 144 16 175 190 205 220 236 251 31 47 62 78 93
1  114 13 145 160 176 191 206 221 237 252 32 48 63 79 94
10 115 130 146 161 177 192 207 222 238 253 33 49 64 8 95
100 116 131 147 162 178 193 208 223 239 254 34 5 65 80 96
101 117 132 148 163 179 194 209 224 24 255 35 50 66 81 97
102 118 133 149 164 18 195 21 225 240 256 36 51 67 82 98
103 119 134 15 165 180 196 210 226 241 257 37 52 68 83 99
104 12 135 150 166 181 197 211 227 242 258 38 53 69 84
105 120 136 151 167 182 198 212 228 243 259 39 54 7 85
106 121 137 152 168 183 199 213 229 244 26 4 55 70 86
107 122 138 153 169 184 2 214 23 245 260 40 56 71 87
108 123 139 154 17 185 20 215 230 246 261 41 57 72 88
109 124 14 155 170 186 200 216 231 247 27 42 58 73 89
11 125 140 156 171 187 201 217 232 248 28 43 59 74 9
110 126 141 157 172 188 202 218 233 249 29 44 6 75 90
111 127 142 158 173 189 203 219 234 25 3 45 60 76 91
112 128 143 159 174 19 204 22 235 250 30 46 61 77 92
yoplait@gizmo ~/work/mountpoint cat all.seqs/99
>8
CGCAGGCCGCAATCTGGTGTGGATTATCTGCGTGCCCATCGTGCTCCTGTACTTCTTCTTCGGCC
GCTTCCTGCTGCTGCTGTTTCATGAATGACCCGCGAGGGTCCGGCCATCGACACCGGTATGCTGTTCT
CTGCGCATCCTGTCTCCCTTCTATTTGTGGTGTGCTCAAGCTGGTCACTGACGGCATCCTGCGC
GGCTGCGGCATGATGGTGCGC
>9
CAGCATCAGGCAGCCAACCAGGATGGCGGCATCGGGCAGGGTGAGGGAACGATGATGGTAACAA
TGATCGGGAAGAGGATCTTCTCAGTCTTGGAACCTTGCGCGGGCTTTCATCACGATTTGACGC
TCTTTTGTGGTG
>10
CAAGTACAGCAGATGCAGAAACAGCATGATACGCAAATCCACAATCTCAGAGAAATGCACAGGCA
GGAACCTTGACATGAAAGAAAAAGAACTGTCACGGCTCGCCAGAATCATAGACAAGGCTTTCAGGT
GGTTTCCGGTATGTTCAAGGAAA
>11
GTTGTTTGCCTATTTACAACATTCTATATTATGGTTGGATACCTGTAAGATTAGTTTTGCTAATT
TCCATCTGGTATTTCTTGATGCGTCTTTCTACGCTTTTTGGGTTCTGTTACCCCAAACCCCTAC
GGACGGAATT
>12
GTAATTGACATGGTTCGTGTGGCAACTTACGTGCACCAGATGCCTGCTGCAATCGAAATGATTGA
AGATGCCAAGGCACACAGGGGATAGG
>13
TGCGTCTGTACCTTTGCGGAATTTGGCTGCGTACTCCTTGATATACTTGTCGAACTTGTCAGG
TAGATGTTAGCCAATATGGGGCTGATTATTCCACCTTGCGGAGTTCTGAGTACGTCTTGTTGTA
CTTCCAGTCTCCACATACCCTGCATTGAAGAATTTACGGATAAGTCTTAGGAAGCGTTCGTCTG
ATATGCGTTCACGAAGAATGCTAATCAGTACGTTGTGGTCTATGTTGTCAAAGAA
```

Object Store

- All Galaxy history and library datasets are assigned an ID and stored in `galaxy-dist/database/files/`
- Single directory = single massive filesystem
- Galaxy now has a dataset abstraction layer to decouple from a single local filesystem: Object Store
 - Disk backend: single filesystem
 - Distributed backend: multiple filesystems
 - Amazon S3 backend - EBS as cache

ChIP-seq

BEST

RNA-seq

Proteomics

PRACTICES

Genotyping

Search parameter
space

VISUAL

Works for any
tool

ANALYSTS

Share & Publish

Galaxy

- ✦ tool integration framework
- ✦ heavy focus on usability
- ✦ sharing, publication framework

Genome Browser

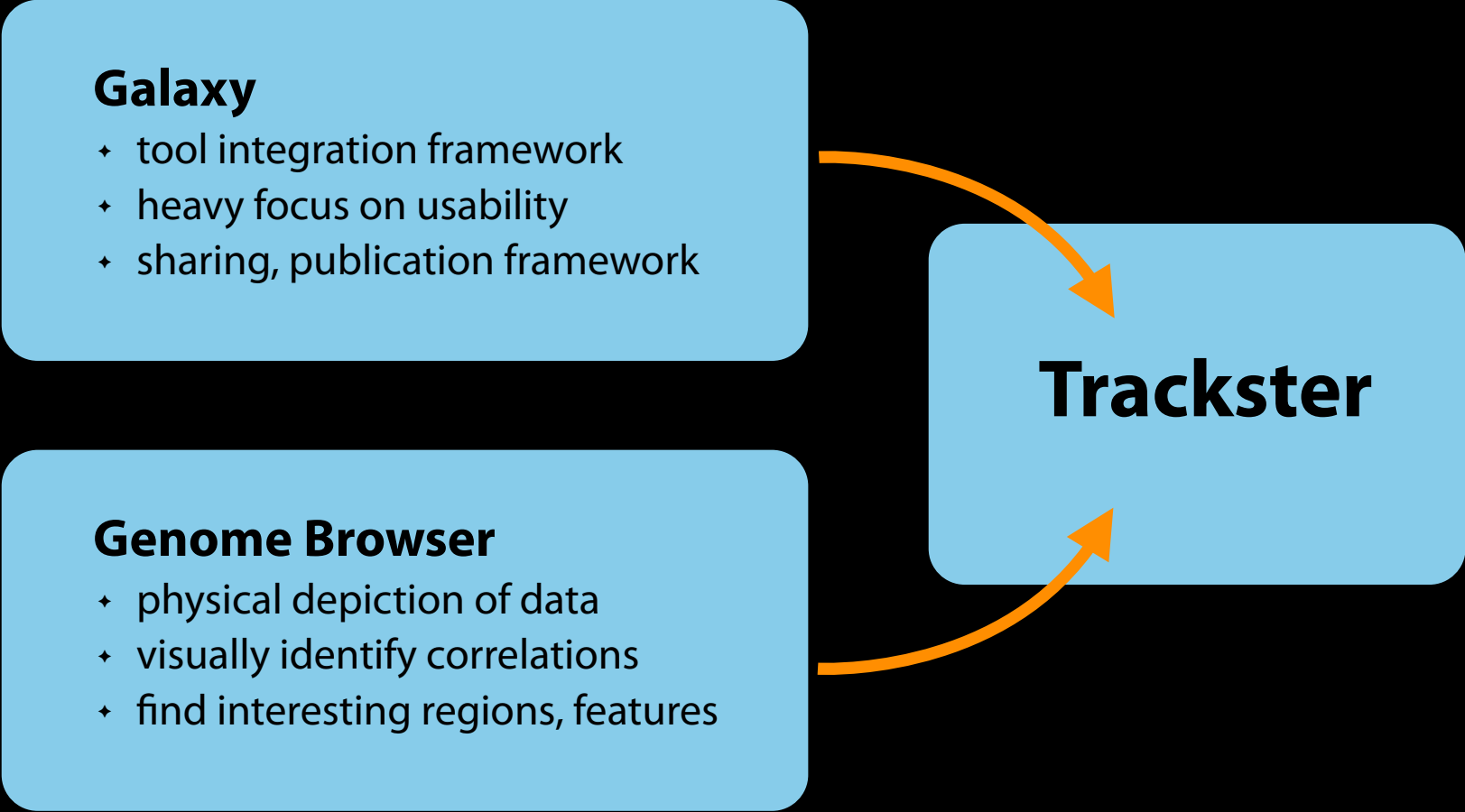
- ✦ physical depiction of data
- ✦ visually identify correlations
- ✦ find interesting regions, features

Galaxy

- ✦ tool integration framework
- ✦ heavy focus on usability
- ✦ sharing, publication framework

Genome Browser

- ✦ physical depiction of data
- ✦ visually identify correlations
- ✦ find interesting regions, features



```
graph LR; Galaxy[Galaxy] --> Trackster[Trackster]; GB[Genome Browser] --> Trackster;
```

Trackster

Trackster

View your data from within Galaxy

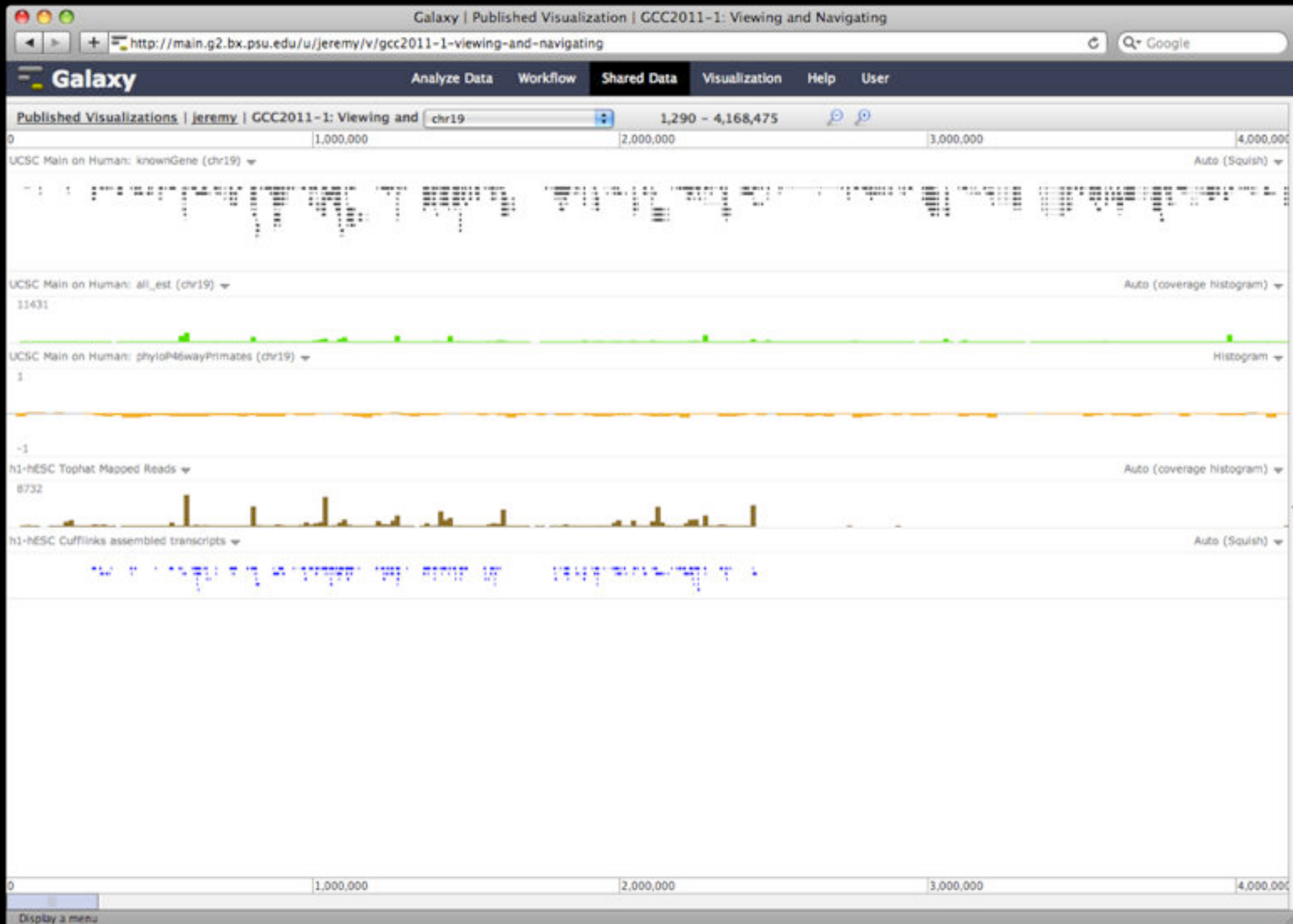
- ✦ No data transfers to external site
- ✦ Use it locally, even without internet access

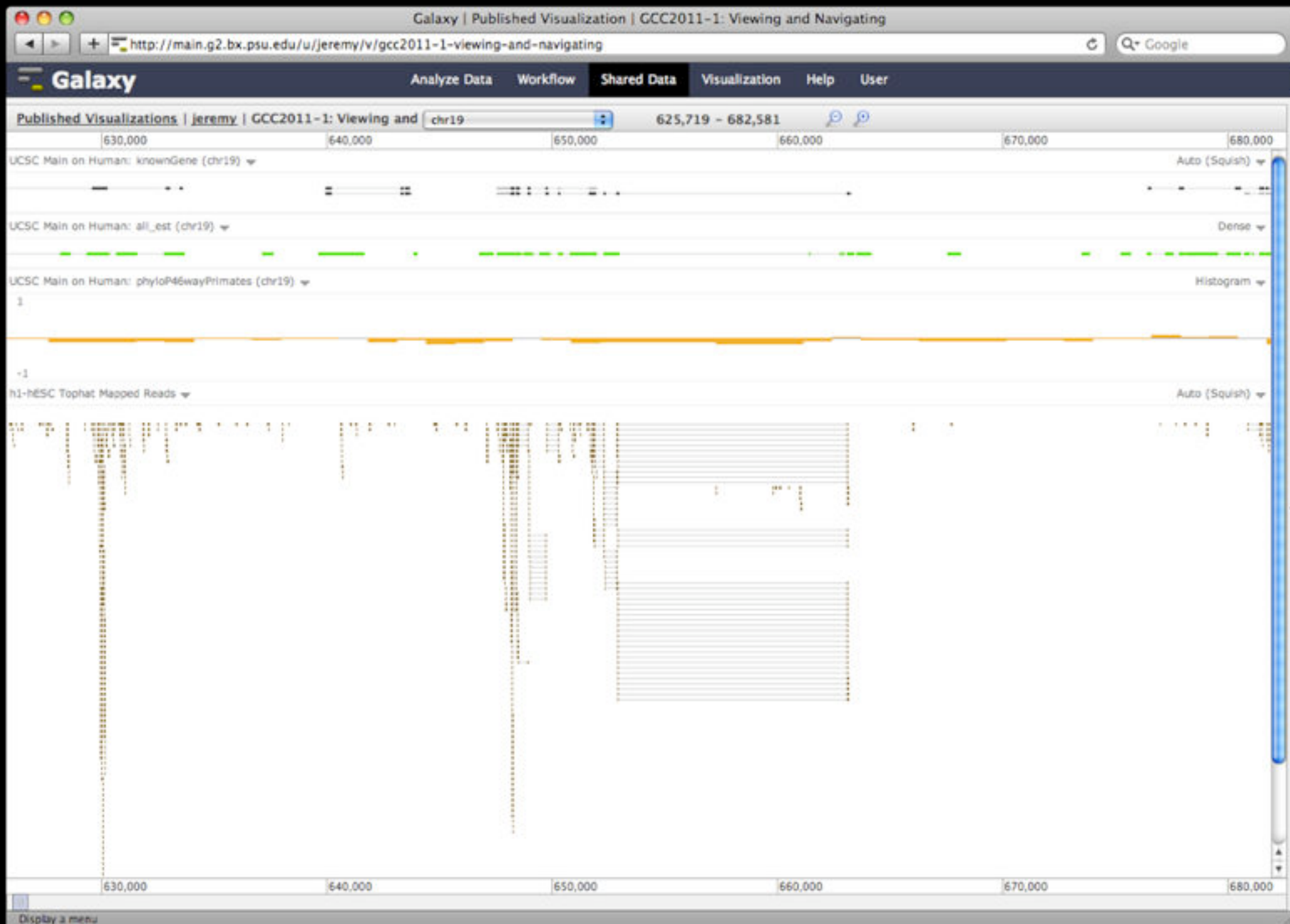
Supports common filetypes

- ✦ BAM, BED, GFF/GTF, WIG

Unique features

- ✦ custom genomes
- ✦ highly interactive





But really, why *another* genome browser

From static browsing to **visual analysis**

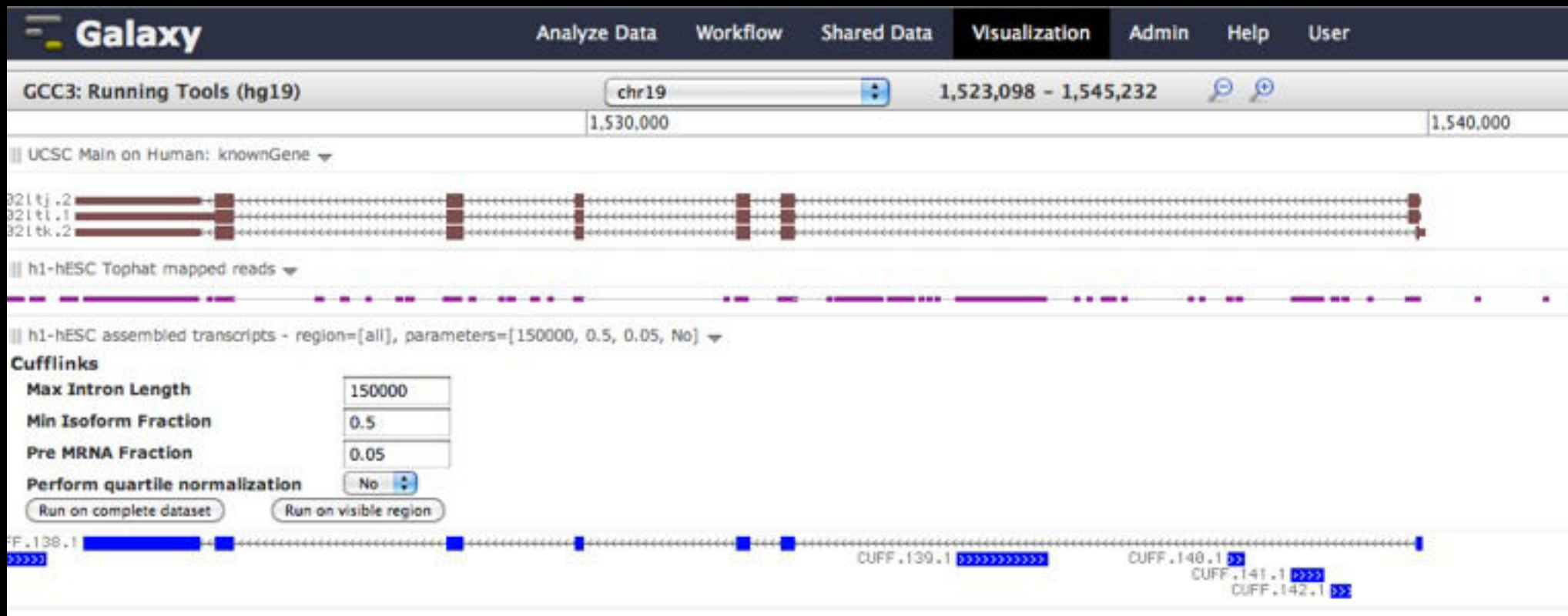
Visual feedback and experimentation needed for complex tools with many parameters

Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

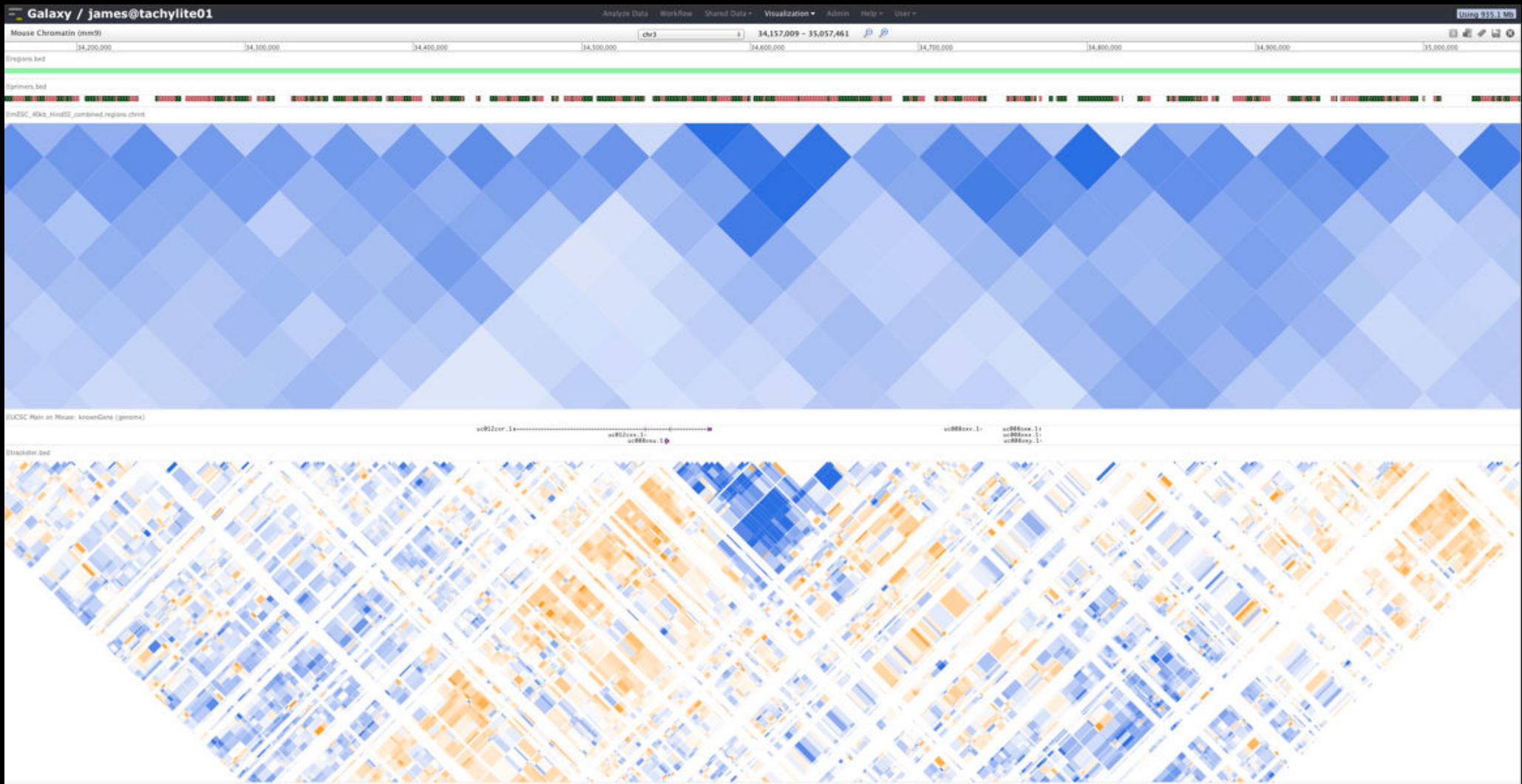
Dynamic Filtering



Integrating Tools and Visualization



5C data in Trackster



5C data in Trackster



Circster

Circos-like visualization that provides
genome-wide views

Complements Trackster

Very much a work in progress

Sweepster

Visualization for

- ✦ tool parameter space
- ✦ outputs from different settings

Can easily find good settings by visual inspection

- ✦ for many settings, across multiple regions

Can explore parameter space systematically or
ad-hoc

Architecture

**Web
browser**

Galaxy
HTML UI

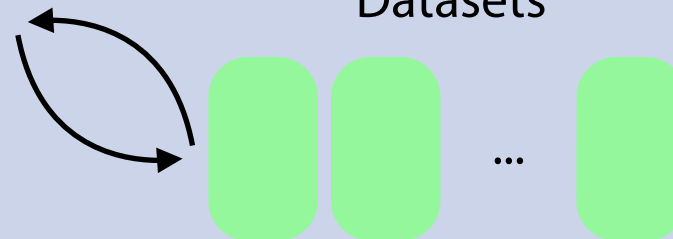


Galaxy

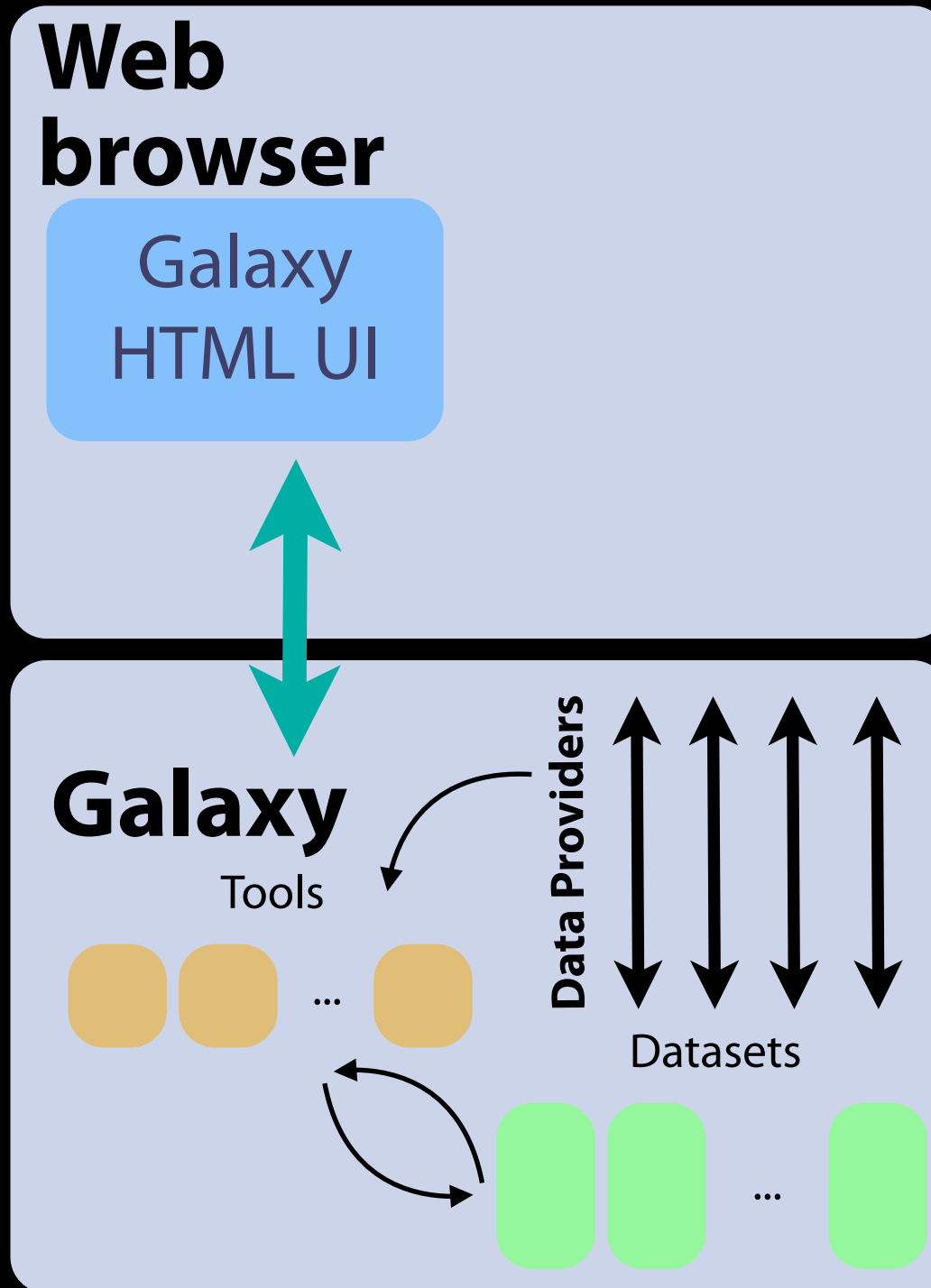
Tools



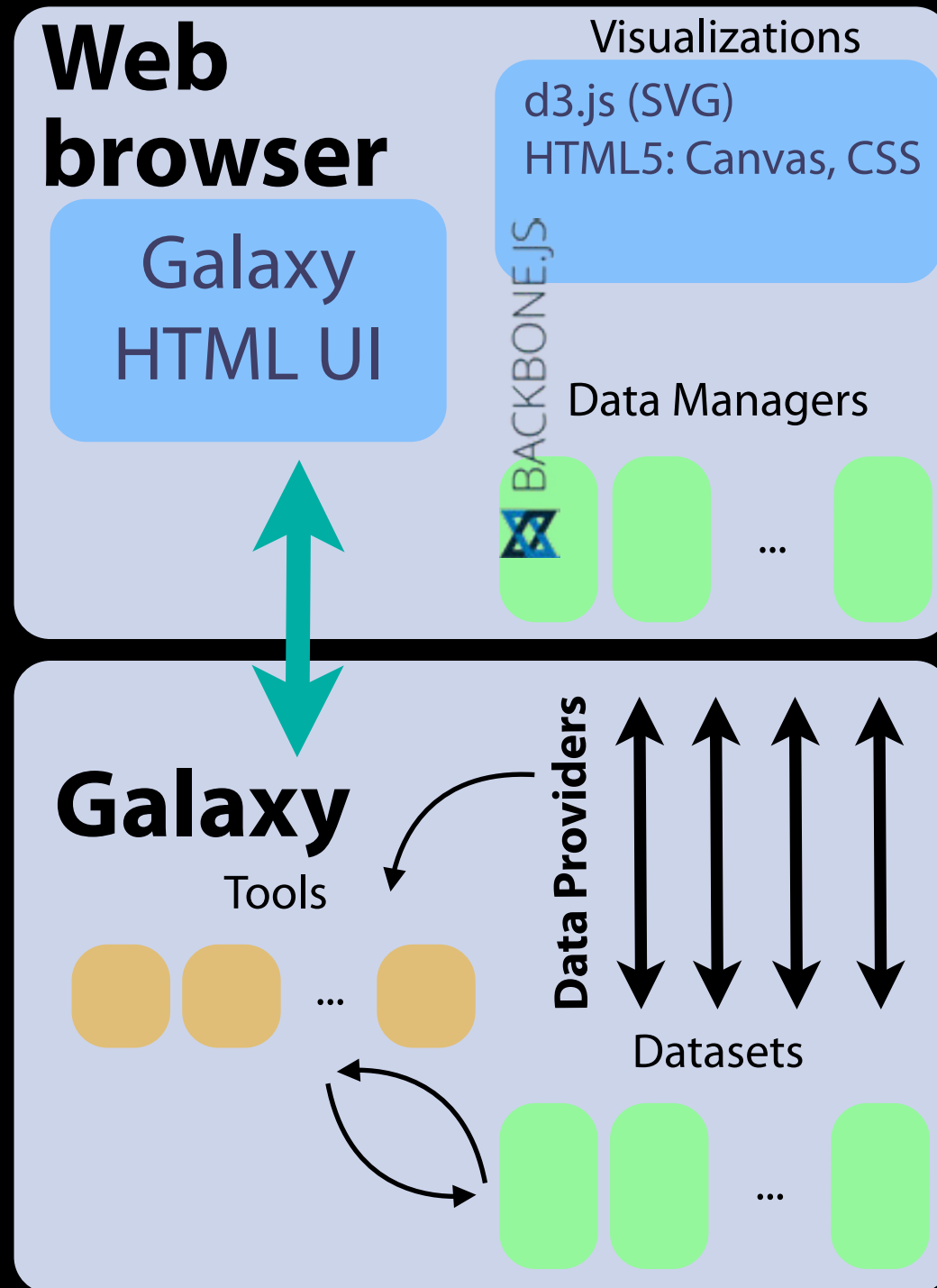
Datasets



Architecture



Architecture



Technical Highlights

Rough pluggable support for generic JavaScript visualizations + data providers

Data providers = fast, random access to data in Python, JS

API: run tools, run them on data subsets

Backbone + HTML5 objects for Web-based genomic visualizations

- ✦ e.g. data managers, linear and circular views

JS binding to Galaxy API (blendJS?)

- ✦ visualizations, tools, datasets
- ✦ custom Galaxy UIs

```

// -- Viz set up. --

var genome = new Genome(JSON.parse('${ h.to_json_string( genome ) }'))
    visualization = new GenomeVisualization(JSON.parse('${ h.to_json_string( viz_config ) }')),
    viz_view = new CircsterView({
        width: 600,
        height: 600,
        // Gap is difficult to set because it very dependent on chromosome size and organization.
        total_gap: 2 * Math.PI * 0.2,
        genome: genome,
        model: visualization,
        radius_start: 100,
        dataset_arc_height: 15
    });

// -- Render viz. --

viz_view.render();
$('#vis').append(viz_view.$el);

```

**Display
applications**

Datatypes

Datasources

Auxiliary data

Customization extends beyond tools

- Everything in the Galaxy framework is either configuration driven or pluggable (or both)
- Tools conventionally extended through configuration, but new tool types can be added
- Datatypes added through configuration, or plugin classes for advanced functionality
- *Nothing inherently specific to genomics!*



Enis Afgan
IRB



Guru Ananda
Penn State



Dannon Baker
Emory



Dan Blankenberg
Penn State



Dave Bouvier
Penn State



Dave Clements
Emory



Nate Coraor
Penn State



Carl Eberhard
Emory



Jeremy Goecks
Emory



Jennifer Hillman Jackson
Penn State



Greg von Kuster
Penn State



Ross Lazarus
Harvard & BakerIDI



Rémi Marenco
Emory



Scott McManus
Emory



Anton Nekrutenko
Penn State



James Taylor
Emory



EMORY
UNIVERSITY

Supported by the **NHGRI** (HG005542, HG004909, HG005133, HG006620), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

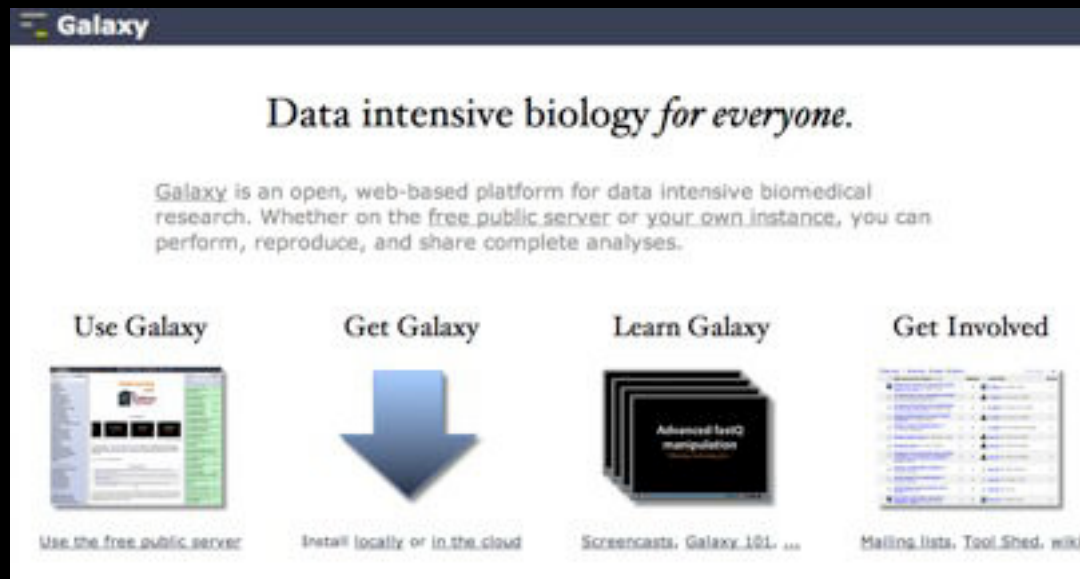
Galaxy Wants You!
jobs@galaxyproject.org



(and your pull requests)

Thanks! Questions?

<http://galaxyproject.org>



Galaxy publications: <http://galaxyproject.org/wiki/Citing>

dannon.baker@emory.edu