Interactive (statistical) visualisation and exploration of the full Gaia catalogue with vaex.

Maarten Breddels
&
Amina Helmi
WP985/WP945

Vaex demo / Gaia DR1 workshop ESAC 2016

university of groningen / faculty of mathematics and natural sciences / kapteyn astronomical institute

Outline

- Motivation
- Technical
- Demo
- Conclusions

Motivation

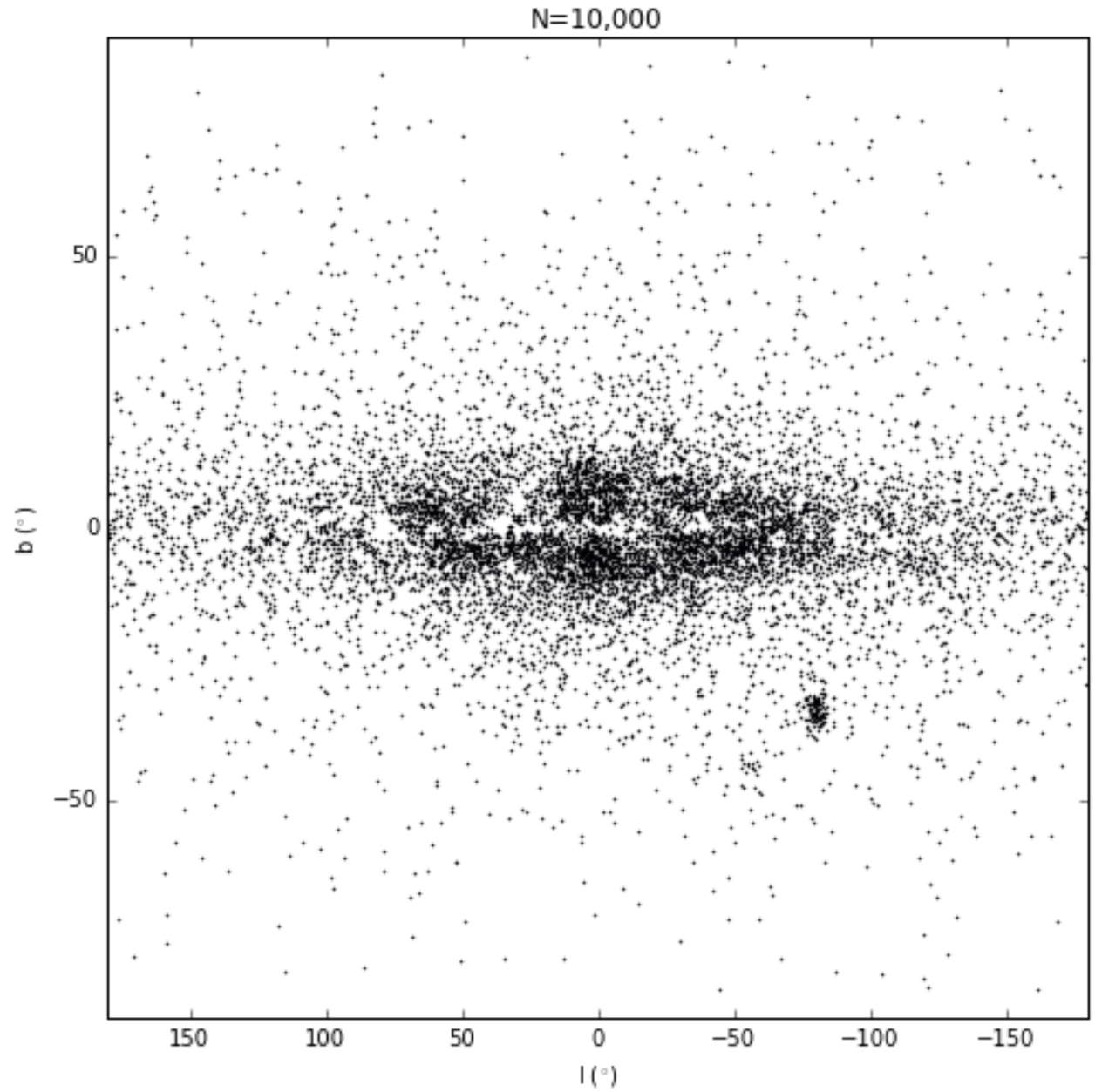
- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to ‘see’ the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network

Motivation

- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to 'see' the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network
- Problem
  - Scatter plots do not work well for 10^9 rows/objects (like Gaia)
  - Work with densities/statistics in 0,1,2 and 3d
  - Interactive?
    - Zoom, pan etc
    - Explore: selections/queries

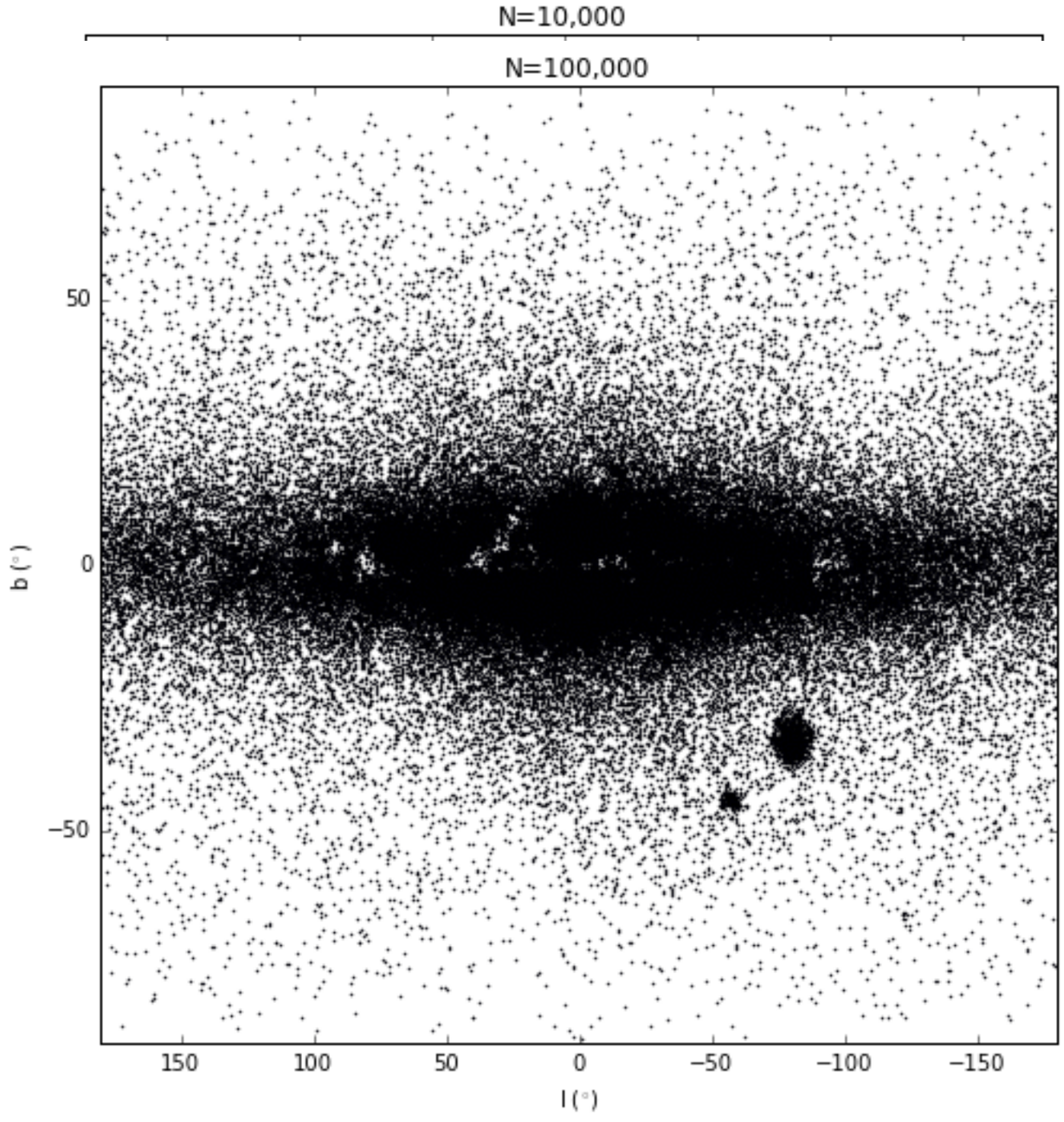
Motivation

- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to 'see' the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network
- Problem
  - Scatter plots do not work well for 10^9 rows/objects (like Gaia)
  - Work with densities/statistics in 0,1,2 and 3d
  - Interactive?
    - Zoom, pan etc
    - Explore: selections/queries



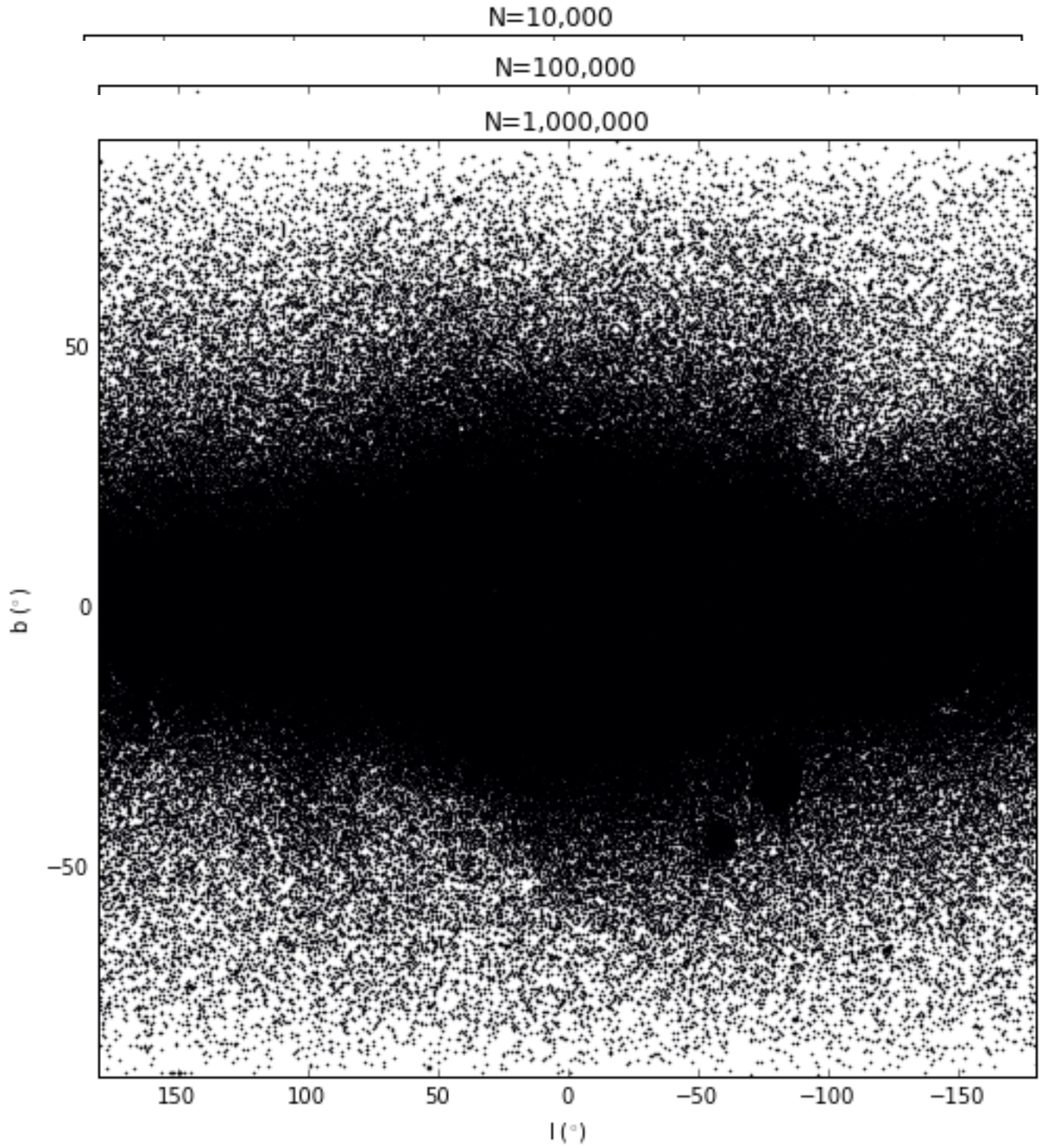
Motivation

- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to 'see' the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network
- Problem
  - Scatter plots do not work well for 10^9 rows/objects (like Gaia)
  - Work with densities/statistics in 0,1,2 and 3d
  - Interactive?
    - Zoom, pan etc
    - Explore: selections/queries



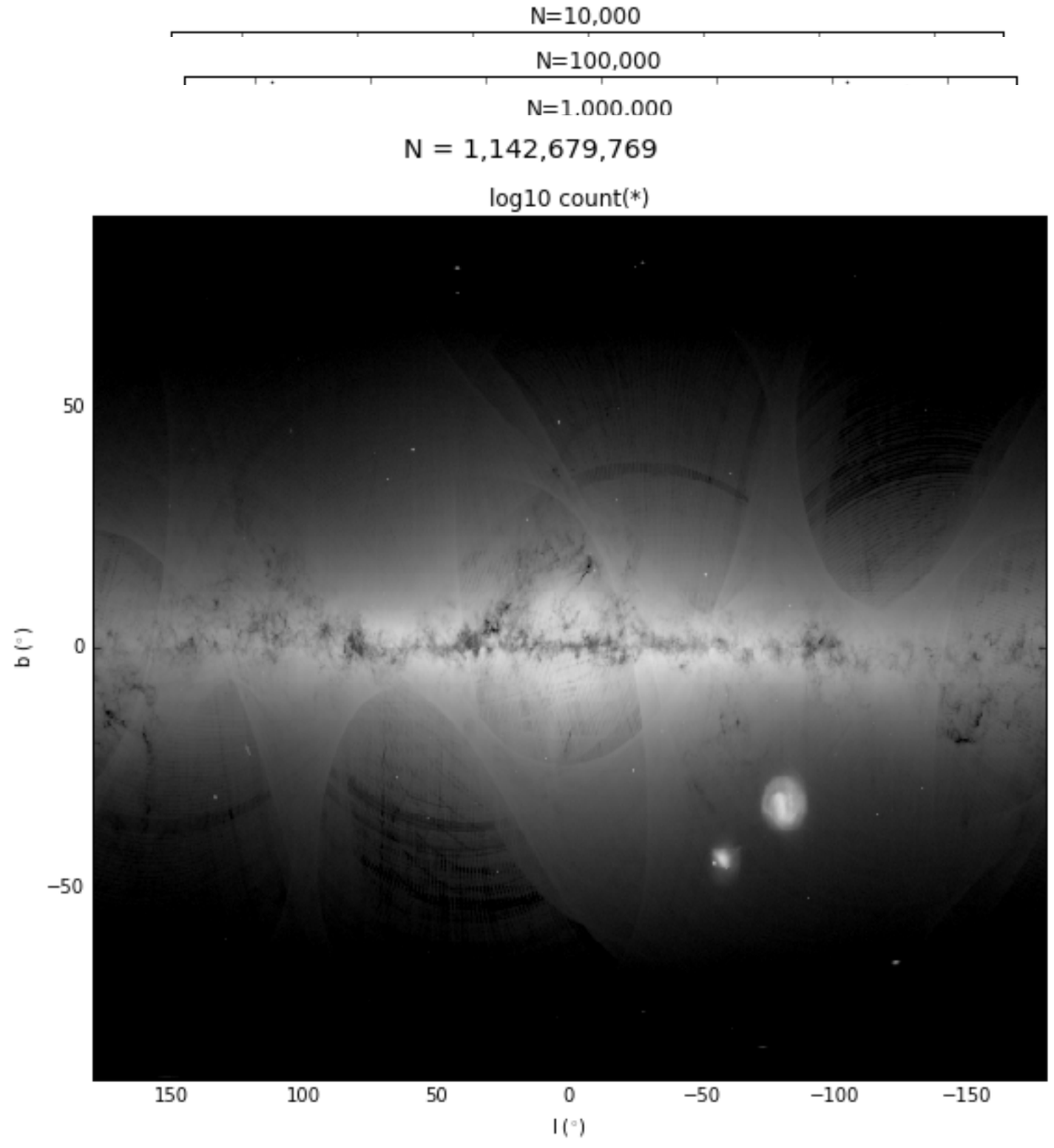
Motivation

- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to 'see' the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network
- Problem
  - Scatter plots do not work well for 10^9 rows/objects (like Gaia)
  - Work with densities/statistics in 0,1,2 and 3d
  - Interactive?
    - Zoom, pan etc
    - Explore: selections/queries



Motivation

- We have Gaia DR1
  - > 10^9 objects/stars
- Can we visualise and explore this?
  - We want to ‘see’ the data
  - Data checks/(Post) validation
  - Science: trends, relations, clustering
    - You are the (biological) neutral network
- Problem
  - Scatter plots do not work well for 10^9 rows/objects (like Gaia)
  - Work with densities/statistics in 0,1,2 and 3d
  - Interactive?
    - Zoom, pan etc
    - Explore: selections/queries



Situation

- TOPCAT comes close, not fast enough, works with individual rows/particles (written in Java)
- Your own IDL/Python code: a lot to consider to do it optimal (multicore, efficient storage, efficient algorithms, interactive becomes complex)
- DataShader: only visualisation of 2d and slower
- We want something to visualize 10^9 rows/objects in ~1 second
- Do we need to resort to Big Data solutions, Hadoop?

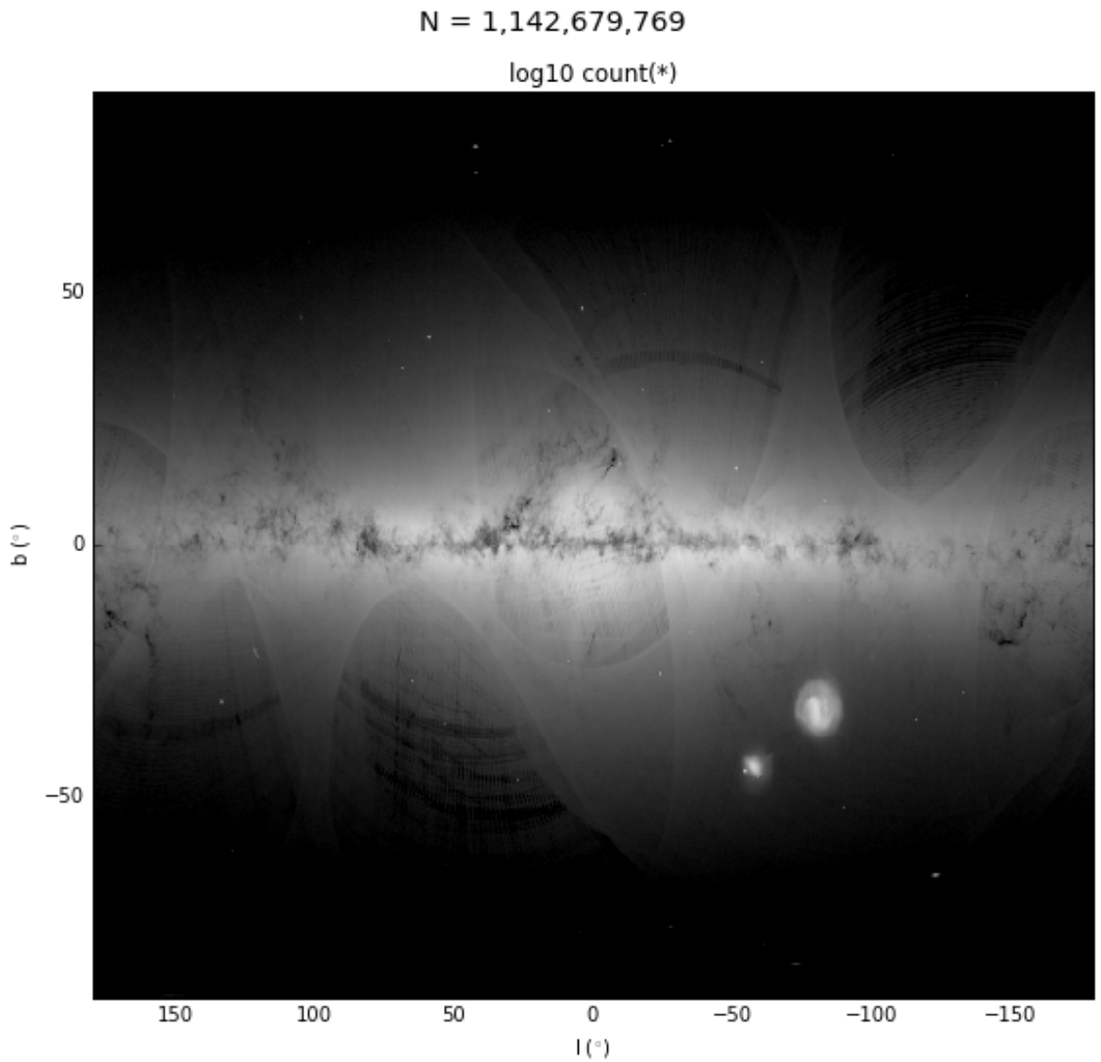
How fast can it be processed?

- What can be done?
  - 10^9 * 2 * 8 bytes = 15 GiB (double is 8 bytes)
  - Memory bandwidth: 10-20 GiB/s: ~1 second
  - CPU: 3 Ghz (but multicore, say 4): 12 cycles/ second
    - Few cycles per row/object, simple algorithm
      - Histograms/Density grids
- Yes, but
  - If it fits/cached in memory, otherwise ssd/hdd speeds (10-100 seconds)
  - proper storage and reading of data
  - simple and fast algorithm for binning

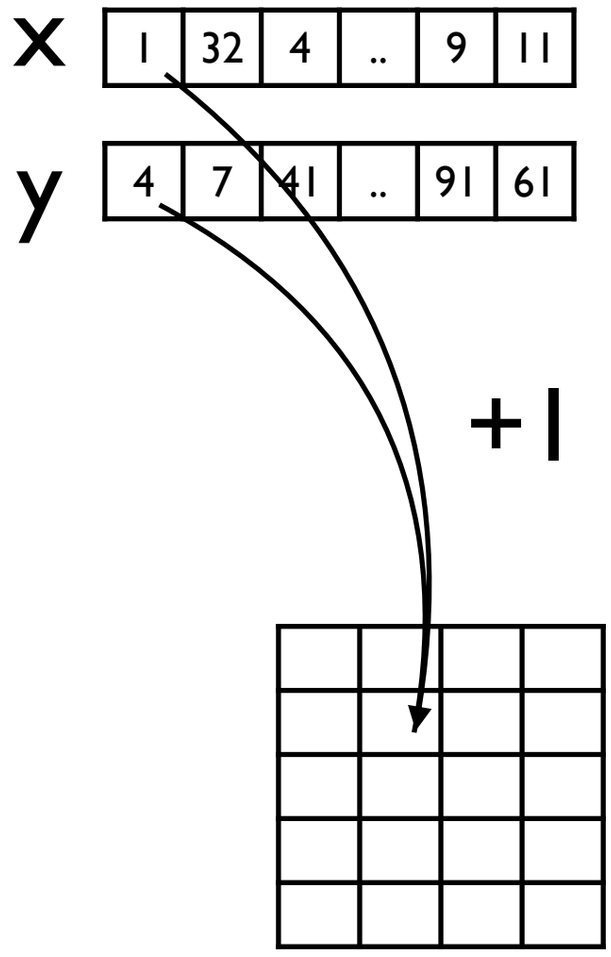
How fast can it be processed?

- What can be done?
  - 10^9 * 2 * 8 bytes = 15 GiB (double is 8 bytes)
  - Memory bandwidth: 10-20 GiB/s: ~1 second
  - CPU: 3 Ghz (but multicore, say 4): 12 cycles/second
  - Few cycles per row/object, simple algorithm
    - Histograms/Density grids
- Yes, but
  - If it fits/cached in memory, otherwise ssd/hdd speeds (10-100 seconds)
  - proper storage and reading of data
  - simple and fast algorithm for binning

•~1 second

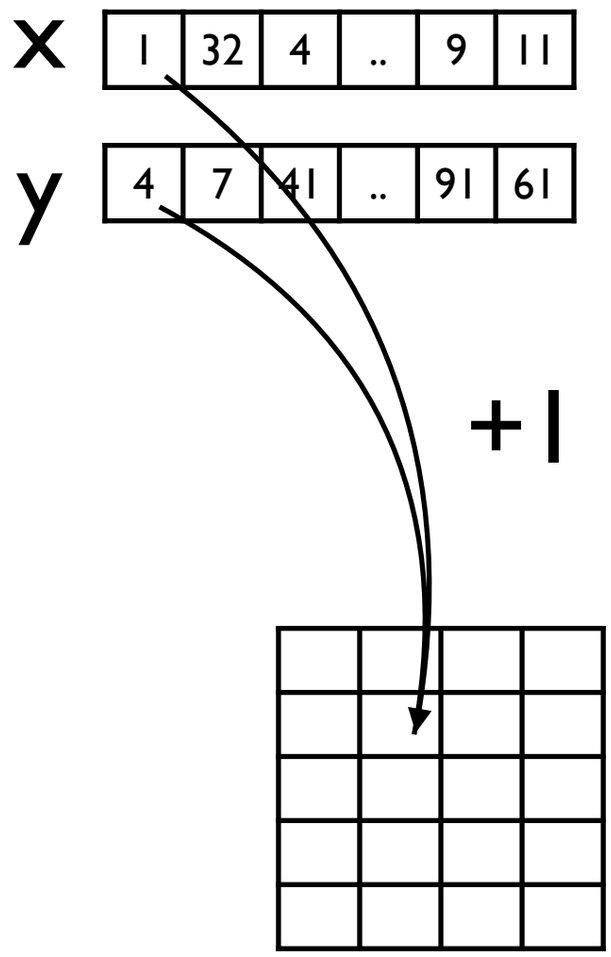


Statistics in N-d



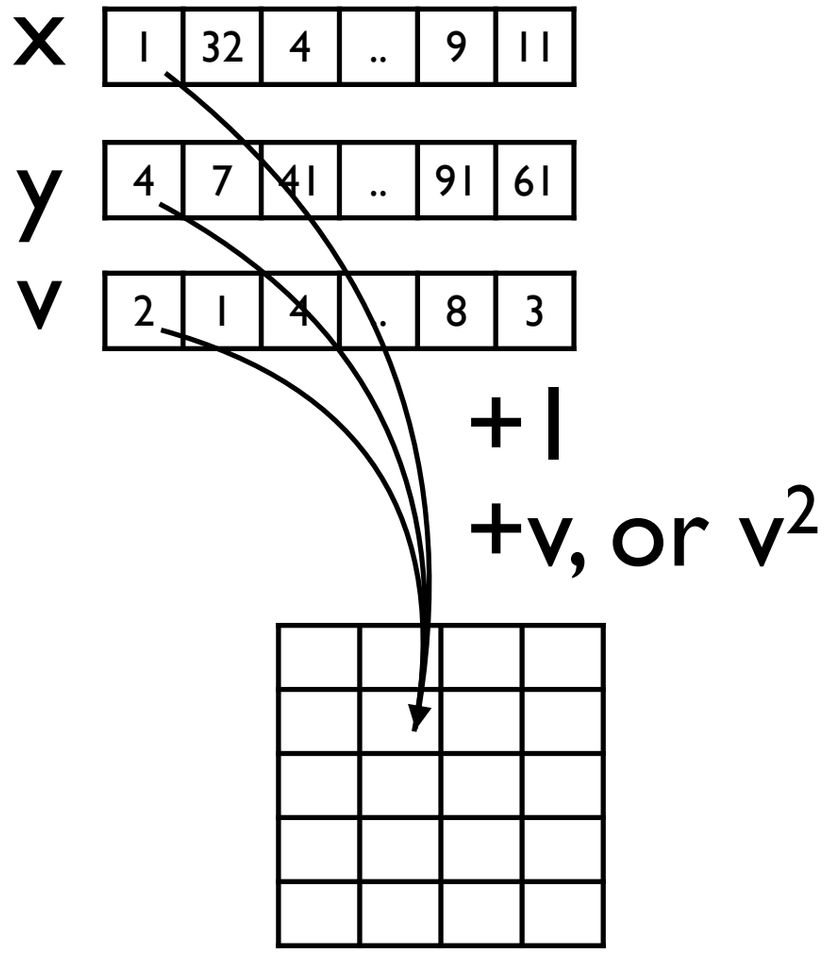
Statistics in N-d

- count
- sum values
- min
- max
- moments



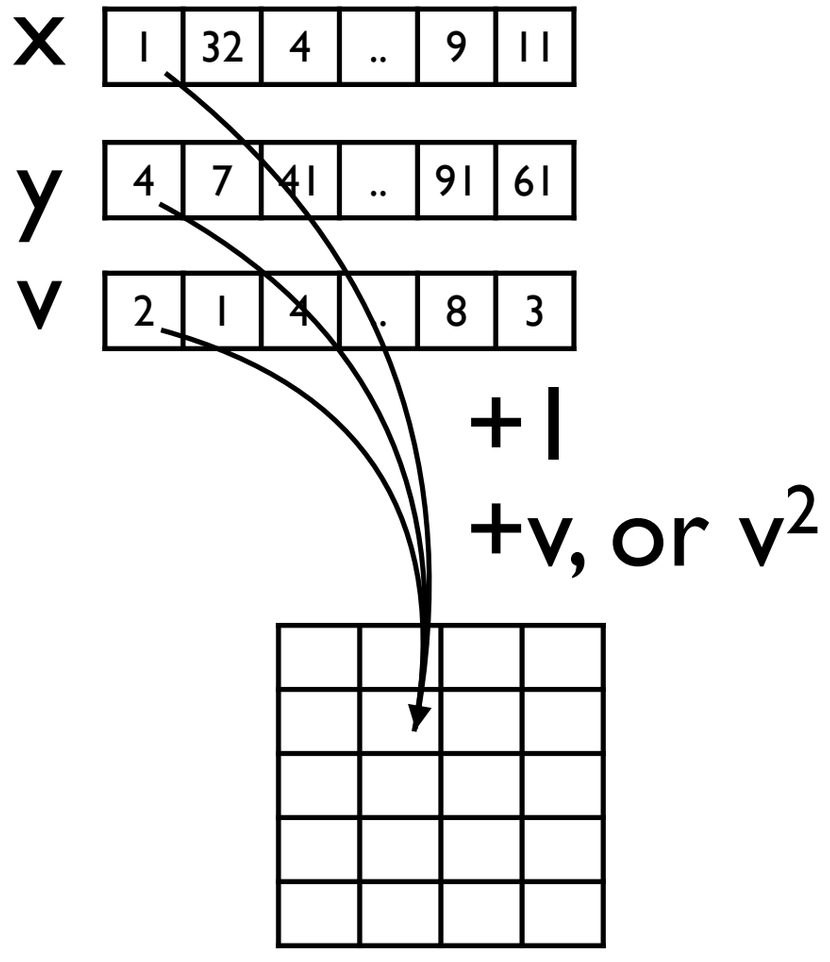
Statistics in N-d

- count
- sum values
- min
- max
- moments



Statistics in N-d

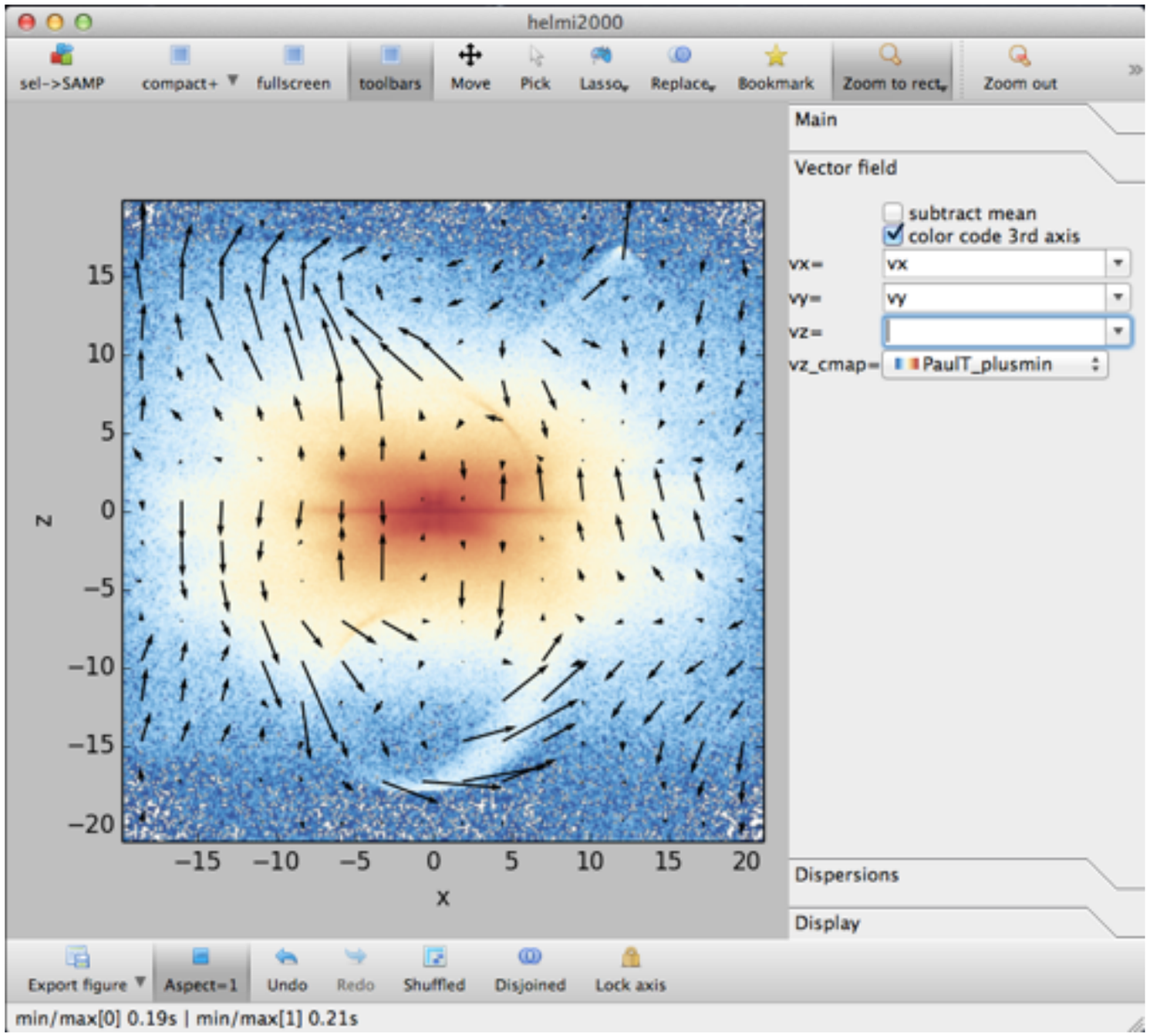
- count
- sum values
- min
- max
- moments



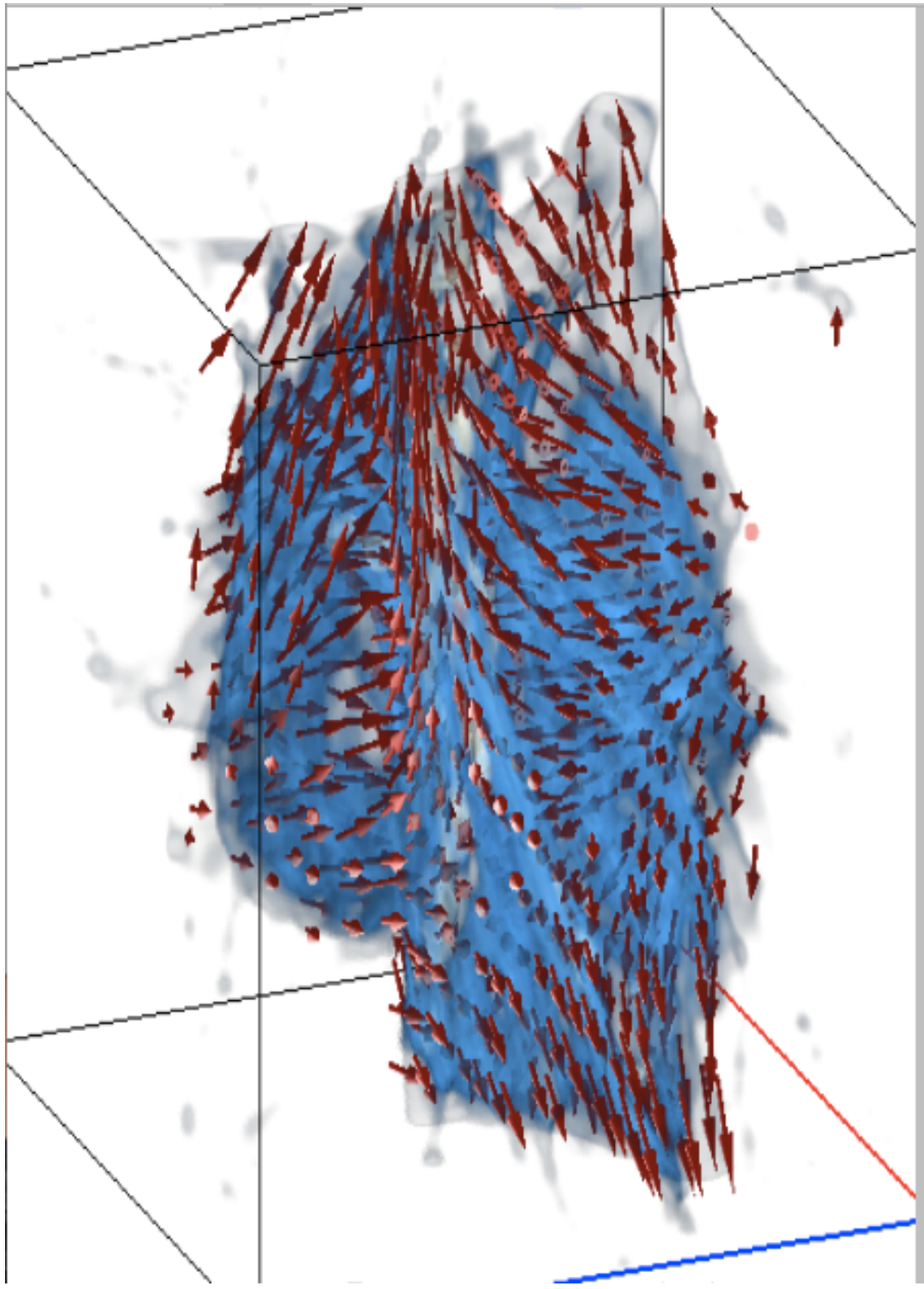
- Possibilities
  - Total: flux, mass
  - Mean: velocity, metallicity
  - Dispersions: velocity…
  - Correlation
- Statistics on a (N dim) grid
  - (And visualize them)

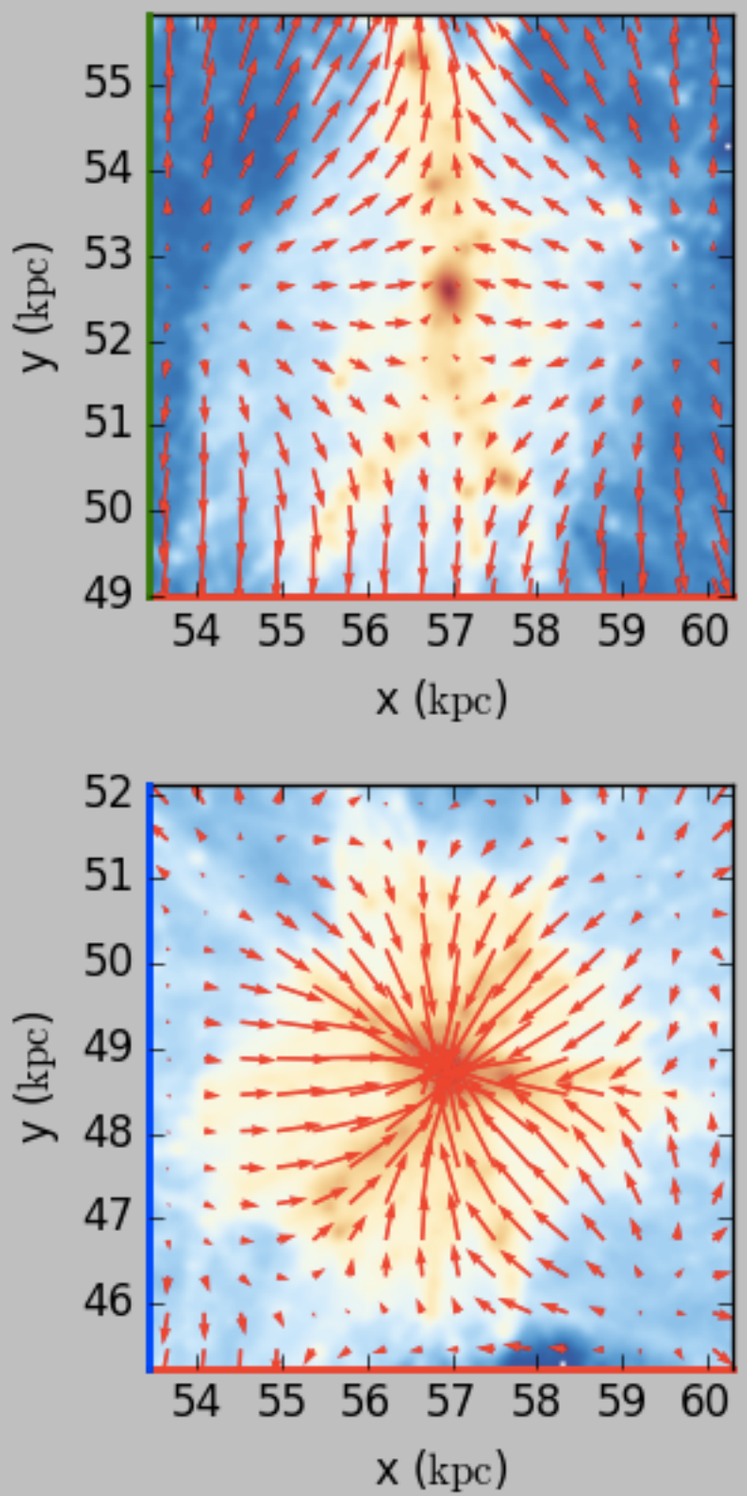
Examples

Examples


helmi2000
sel->SAMP
compact+
fullscreen
toolbars
Move
Pick
Lasso
Replace
Bookmark
Zoom to rect
Zoom out
Main
Vector field
subtract mean
color code 3rd axis
vx=
vx
vy=
vy
vz=
vz_cmap=
PaulT_plusmin
Dispersions
Display
Export figure
Aspect=1
Undo
Redo
Shuffled
Disjoined
Lock axis
min/max[0] 0.19s | min/max[1] 0.21s
x
z

Examples




55
54
53
52
51
50
49
y (kpc)
54 55 56 57 58 59 60
x (kpc)
52
51
50
49
48
47
46
y (kpc)
54 55 56 57 58 59 60
x (kpc)

Vaex: **V**isualization **A**nd **EX**ploration

- A library
  - python package
  - ‘import vaex’
- reading of data
- multithreading
- statistics/binning (0,1,2,3, Nd)
- selections/queries
- server/client
- integrates with IPython notebook
- Installation:
  - pip install —user —pre vaex
  - conda install -c conda-forge vaex
- open source / MIT License
- www.github.com/maartenbreddels/vaex

Vaex: Visualization And EXploration

- A library
  - python package
  - ‘import vaex’
- reading of data
- multithreading
- statistics/binning (0,1,2,3, Nd)
- selections/queries
- server/client
- integrates with IPython notebook
- Installation:
  - pip install —user —pre vaex
  - conda install -c conda-forge vaex
- open source / MIT License
- www.github.com/maartenbreddels/vaex

- A GUI program
- Gives interactive navigation, zoom, pan
- interactive selection (lasso, rectangle)
- client
- undo/redo
- Standalone binary
  - http://vaex.astro.rug.nl/

Vaex: Visualization And EXploration

- A library
  - python package
  - ‘import vaex’
- reading of data
- multithreading
- statistics/binning (0,1,2,3, Nd)
- selections/queries
- server/client
- integrates with IPython notebook
- Installation:
  - pip install —user —pre vaex
  - conda install -c conda-forge vaex
- open source / MIT License
- www.github.com/maartenbreddels/vaex

- A GUI program
- Gives interactive navigation, zoom, pan
- interactive selection (lasso, rectangle)
- client
- undo/redo
- Standalone binary
  - http://vaex.astro.rug.nl/

Demo program

- Basics (Helmi de Zeeuw 2000)
- Full Gaia DR1
- SAMP

- Laptop:
  - Macbook Air 13”, 8BG ram, ssd
- Server (gaia):
  - 2x8 cores (32 hyperthreading)
  - 256 GB RAM
  - 24 RAID
  - ~12 kEUR

Demo library

- Basics (statistics)
- Healpix/Full Gaia DR1
  - Saggitarius stream
- SAMP
- (Interactive)

Get vaex

- Standalone binary (OS X , Linux) (just download and start)
  - http://vaex.astro.rug.nl/#download
- Python library (superset of above)
  - Quick look / independent Python tree
    - curl http://vaex.astro.rug.nl/install_conda.sh | bash -
  - Anaconda (‘Python + package manager’ / recommended)
    - conda install -c conda-forge vaex
  - Vanilla Python (PyQt may be a challenge to get installed)
    - pip install —user —pre vaex

How to get (Gaia DR1) data in vaex

- See vaex.astro.rug.nl/latest/getting_data_in_vaex.html
- Download from archive
  - convert all fits files to one big colfits file
    - (convert with vaex to hdf5 for better performance)
- http://vaex.astro.rug.nl/#gaia (Affiliate Data, Groningen, NL)
  - Full download (or 10% / 1%)

Workflows

- Data local
  - vaex program
  - python script
  - Jupyter notebook
  - Laptop:
    - 1-10% random subset of the data

Workflows

- Data local
  - vaex program
  - python script
  - Jupyter notebook
  - Laptop:
    - 1-10% random subset of the data
- Data remote
  - vaex program (remote X11)
  - **vaex server**
    - vaex program
    - python script
    - Jupyter notebook
  - **Remote Jupyter notebook server**

Future plans

- Paper and 1.0.0 release ‘soon’
- Jupyter notebook
  - Interactive/ipywidgets has huge potential
- Distributed vaex (>10^10/sec)

Conclusions

- Vaex can handle 10^9 rows to compute N dimensional statistics
  - In order of ~1 second, interactive
  - which can be used for visualisation, in 1, 2 and 3d (vaex program)
  - Integrates with SAMP/TOPCAT
  - Publication quality plots with matplotlib
  - Even more relevant for DR2, Euclid, LSST, Pan-STARRS, others?