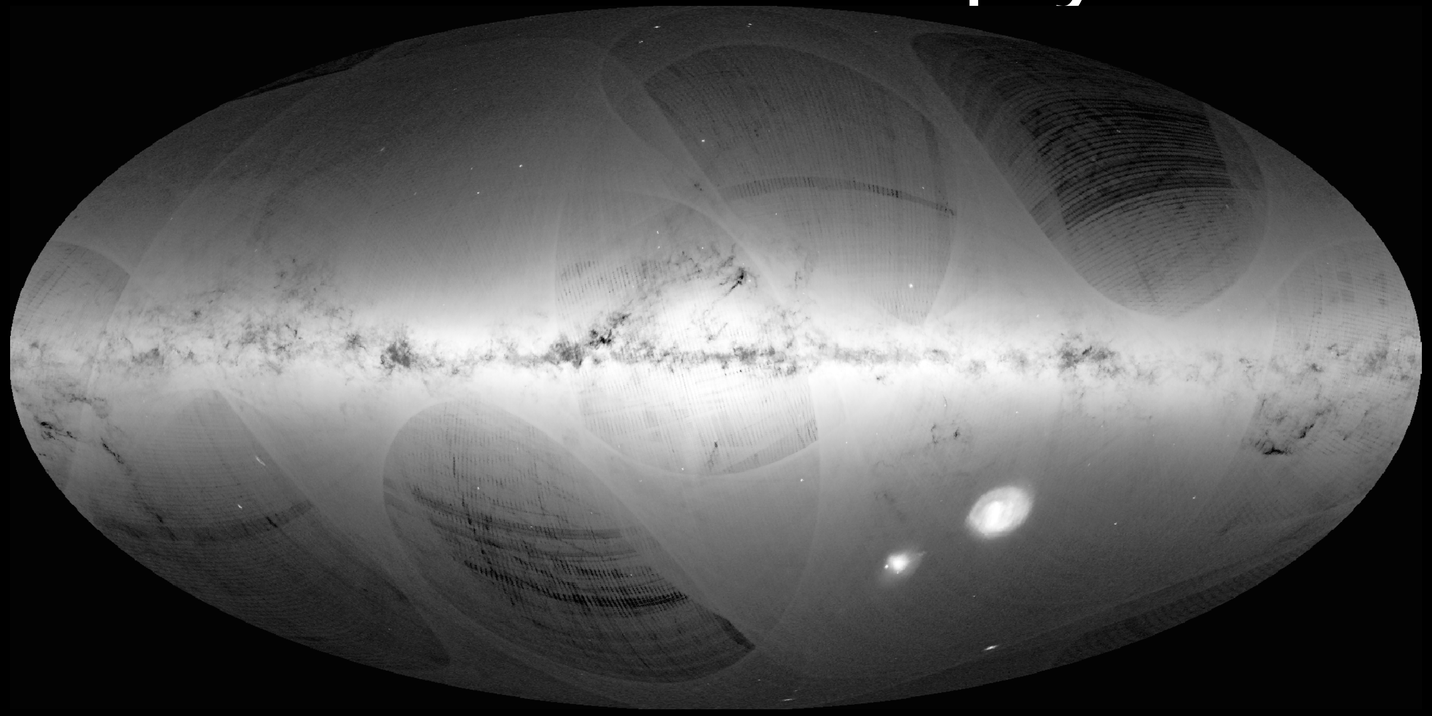
A Billion Stars in the Jupyter Notebook



https://arxiv.org/abs/1801.02638

Maarten A. Breddels
Jovan Veljanoski

Kapteyn - Lunchtalk - 12/March/2018

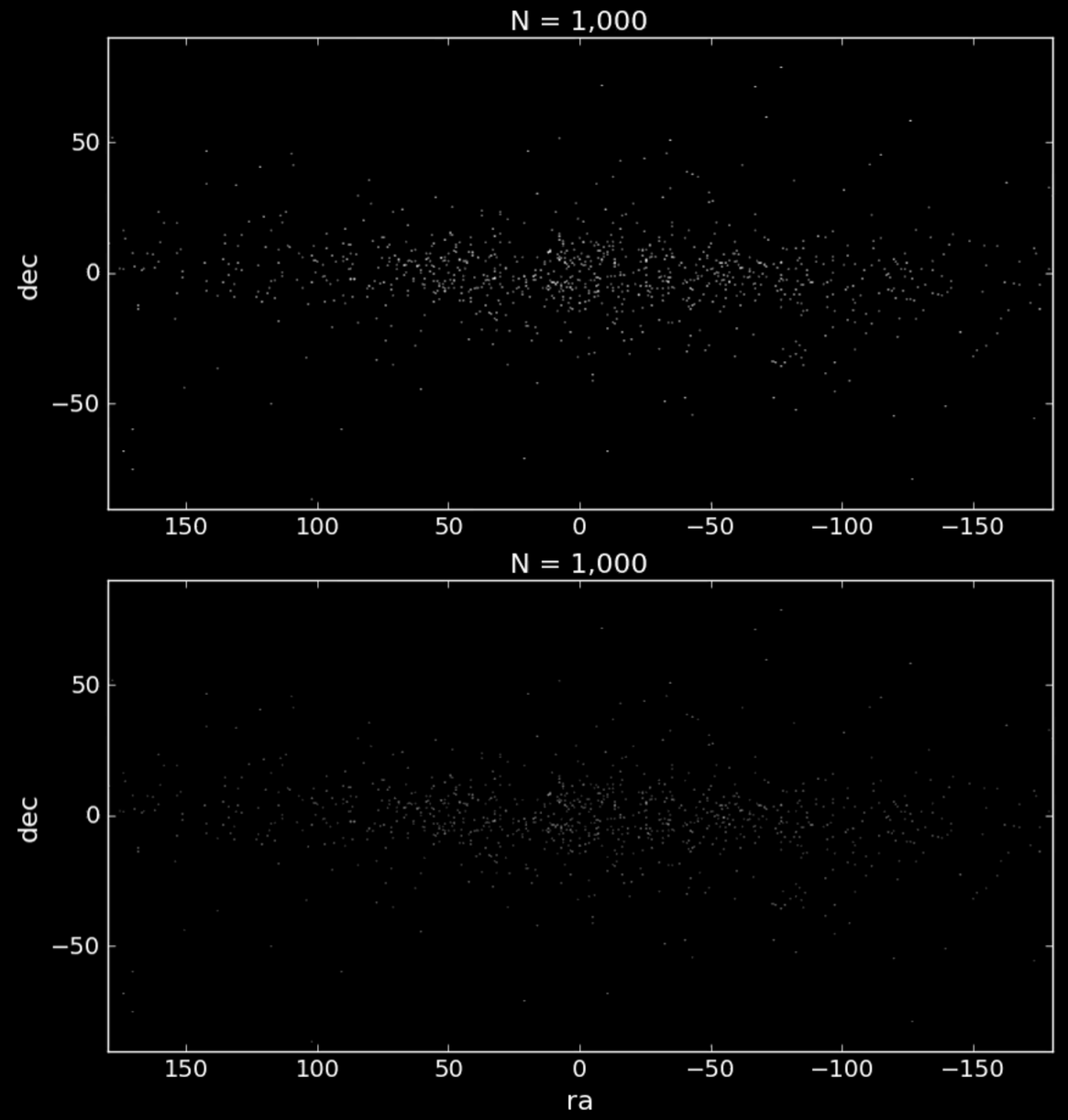
university of groningen / faculty of mathematics and natural sciences / kapteyn astronomical institute

Agenda

- Show how to deal with a billion objects/rows/stars?
- Why use/move to the notebook?
- How can we make the notebook interactive?

Motivation: Gaia

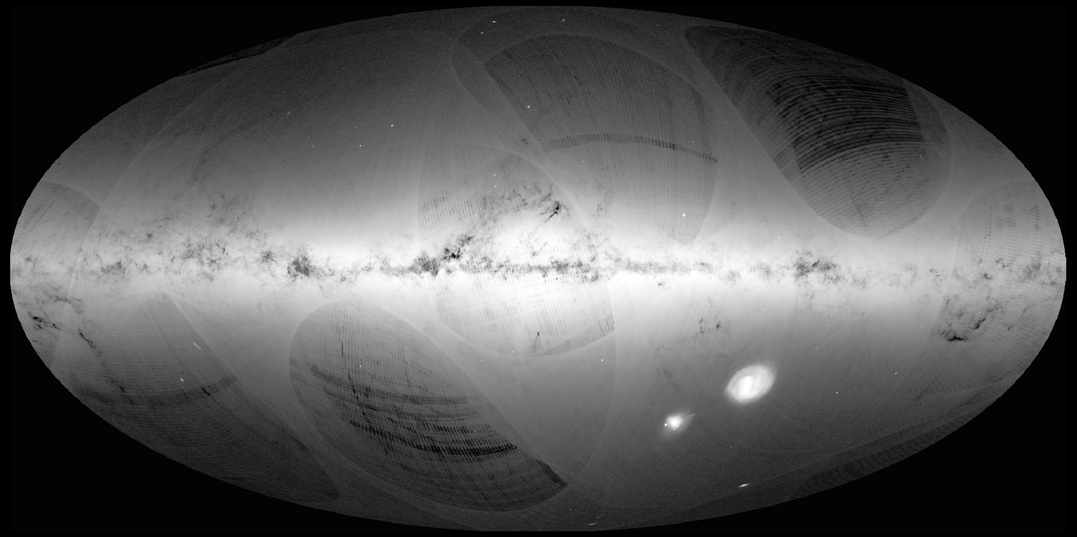
- > 1 billion stars
- DR1 (2016)
  - G
  - RA, Dec
- DR2 (April 25)
  - G, G_{BP}, G_{RP}
  - RA, Dec
  - Proper motion
  - Parallax
  - Radial vel. (6 million)



scatter

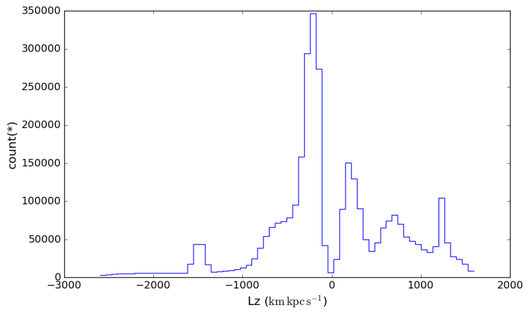
density

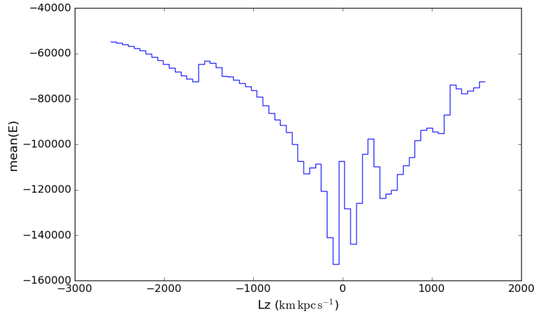
- How fast can it be done?
  - 10^9 * 2 * 8 bytes = 15 GiB (double is 8 bytes)
  - Memory bandwidth: 10-30 GiB/s: ~1 second
  - CPU: 3 Ghz (but multicore, say 4-8): 12-24 cycles/second
  - Few cycles per row/object, simple algorithm
  - Histograms/Density/Statistics grids



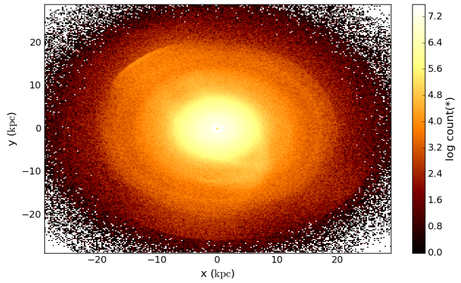
0d 330,000 rows mean: -0.083

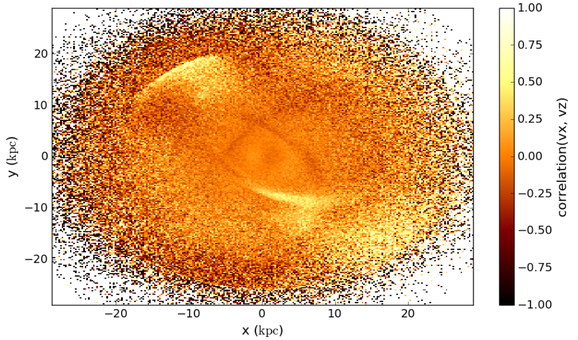
1d



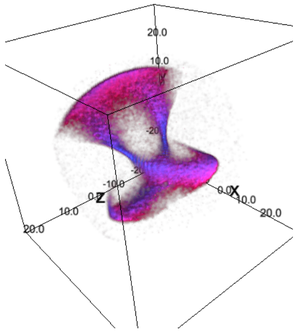


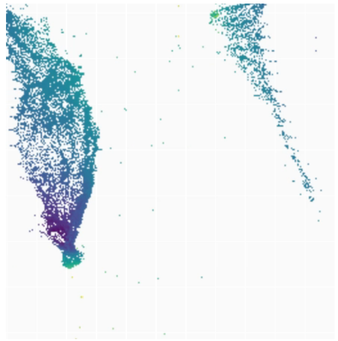
2d





3d

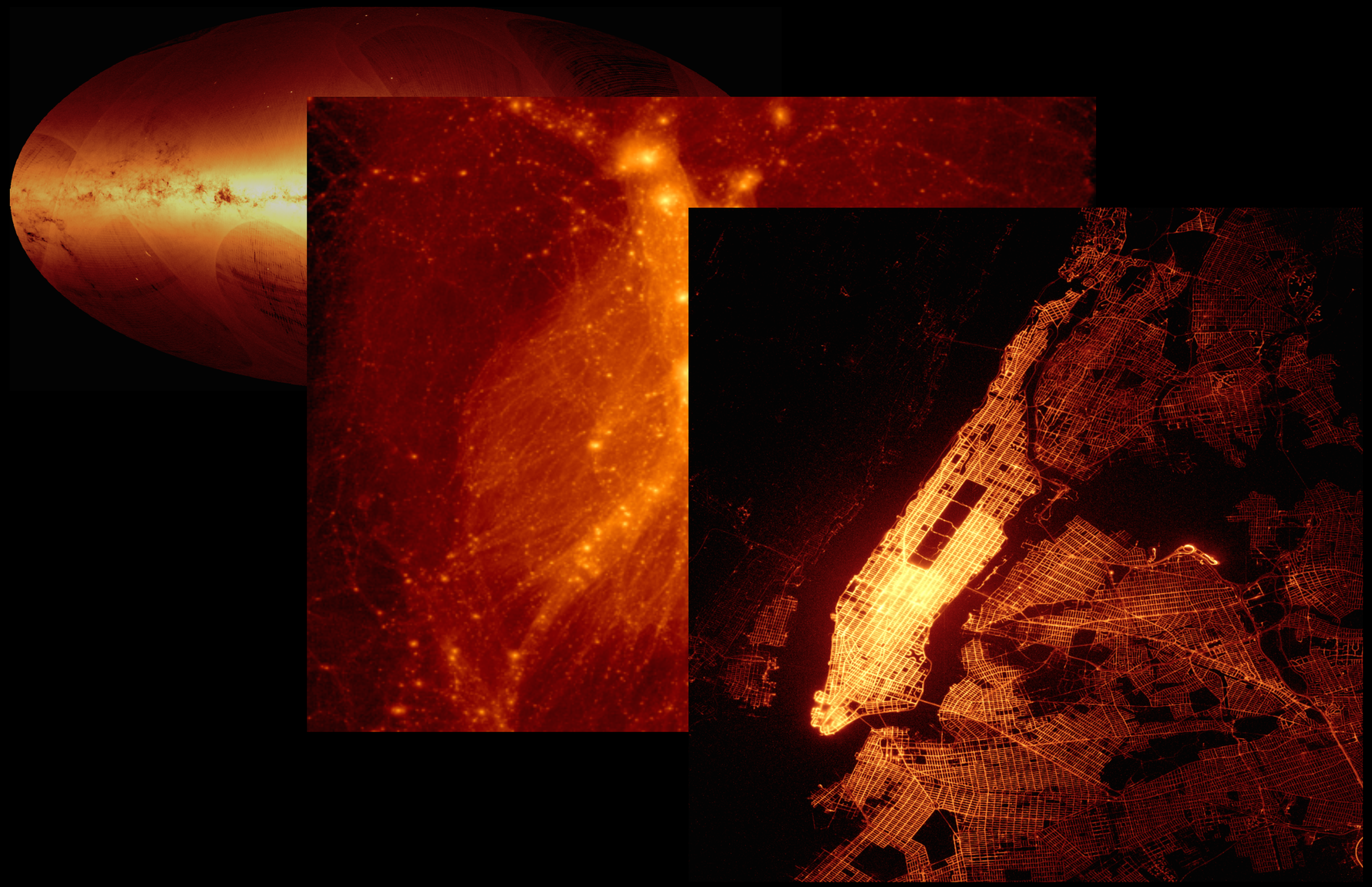




vaex

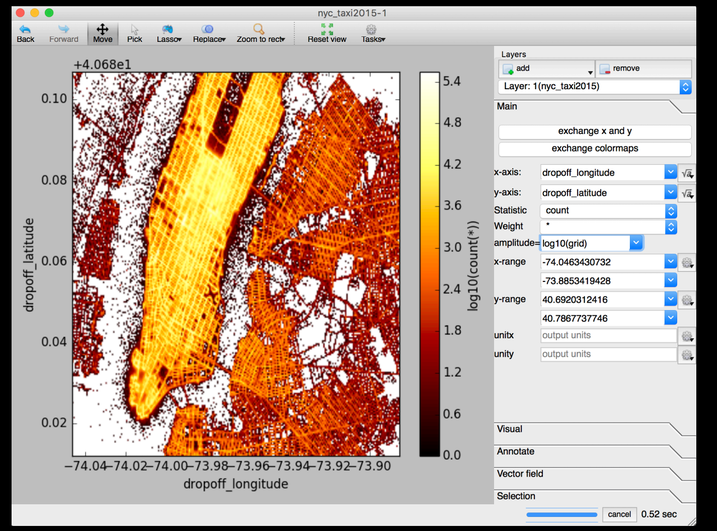
- Python library (conda/pip installable)
- pandas-like but for large datasets
  - Out-of-core, lazy expression, works on chunks
- Focusses mostly on statistics on N-d grids (count/mean/max/std/…)
- >1 billion rows / sec on a `decent` desktop (quad core 3Gz)
  - >50x faster than scipy.stats.binned_statistic_2d
- Does visualisation / matplotlib / bqplot / ipyvolume / ipyleaflet
- More
  - GUI
  - Machine learning (K-means, PCA, ..)
  - Distributed computing (>10^{10} rows)

What kind of data?



Why the Jupyter notebook?

- GUI (vaex has/had this)
  - Limited in options, cannot replace a programming language
- Scripts
  - Anything possible
  - Slow in exploration phase
- Jupyter Notebook
  - Quick exploration/iterating
  - What about interactive plots?



“Never do a live demo”

-Many people

Answers

- How to deal with a billion objects/rows/stars?
  - statistics in N-d grids / vaex
- Why use/move to using the notebook?
  - Quick exploration + interactivity
- How can we make the notebook interactive?
  - ipywidgets+bqplot+ipyleaflet+ipyvolume

- maartenbreddels@gmail.com
  - www.maartenbreddels.com
  - Twitter @maartenbreddels
- vaex
  - https://vaex.io
  - https://github.com/maartenbreddels/vaex
  - pip install —pre vaex / conda install -c conda-forge vaex
- ipyvolume
  - https://ipywidgets.readthedocs.io
  - https://github.com/maartenbreddels/ipyvolume
  - pip install ipyvolume / conda install -c conda-forge ipyvolume