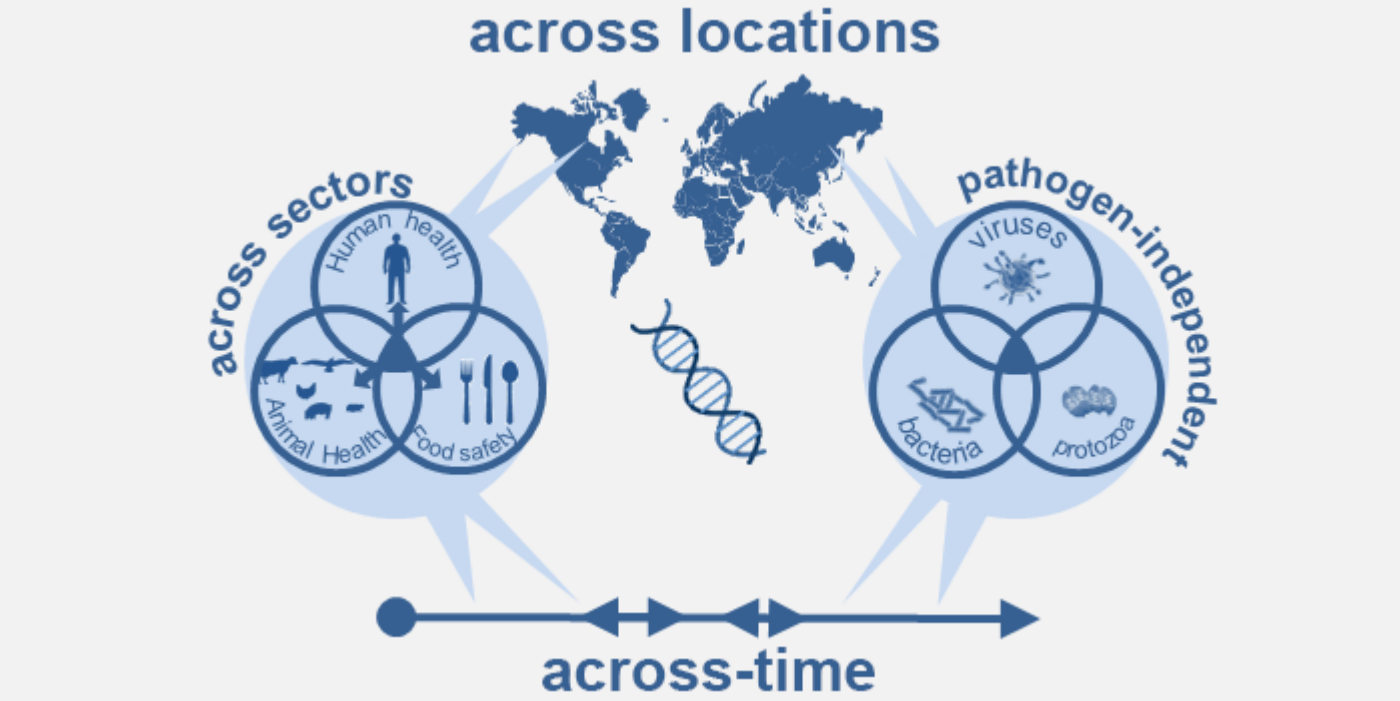
Collaborative Platform for Genomics Big Data Analysis @ MTA cloud

Istvan Csabai, Laszlo Oroszlany, Janos Szalai-Gindl, David Visontai, Laszlo Dobos, Dezso Ribli
Eötvös University Dept. of Physics of Complex Systems & Wigner RCP

MTA Cloud Symposium 2017.02.16


across locations
across sectors
Human health
Animal Health
Food safety
pathogen-independent
viruses
bacteria
protozoa
across-time

COMPARE ANALYTICAL FRAMEWORK

A RA Models and risk-based strategies

Samples and metadata

B Harmonized standards

comparable data

C Analytical workflows

actionable information

Risk communication tools **E**

D Data and information platform

for sample and data collection

for sample processing and sequencing

for generating actionable information

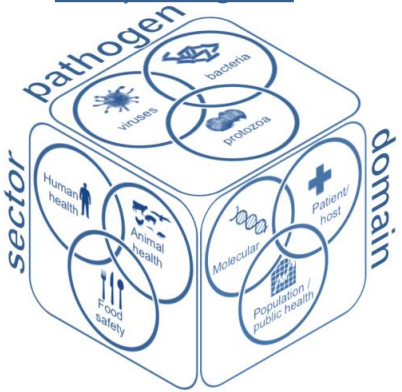
Supporting activities

- **F** Stakeholder Consultations
- **F** Studies into Barriers to open data sharing
- **F** Dissemination and Training
- **G** Cost Effectiveness studies

User-stakeholders:

- Public and Veterinary Health, and Food Safety Authorities, Wildlife/Nature management;
- Clinicians, General Practitioners, Veterinary practices;
- Scientific community, Research Labs (Microbiological, Clinical, Food safety labs);
- Food industry;

Across domains, sectors and pathogens



COMPARE ANALYTICAL FRAMEWORK

A RA Models and risk-based strategies

B Harmonized standards

C Analytical workflows

D Data and information platform

E Risk communication tools

Samples and metadata

comparable data

actionable information

for sample and data collection

for sample processing and sequencing

for generating actionable information

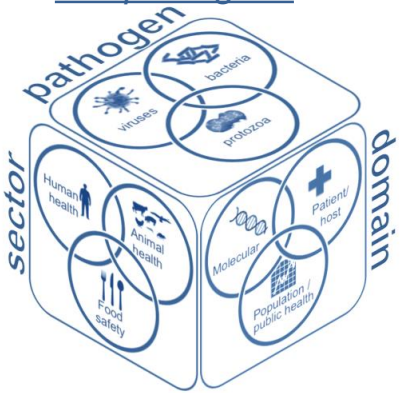
Supporting activities

- **F** Stakeholder Consultations
- **F** Studies into Barriers to open data sharing
- **F** Dissemination and Training
- **G** Cost Effectiveness studies

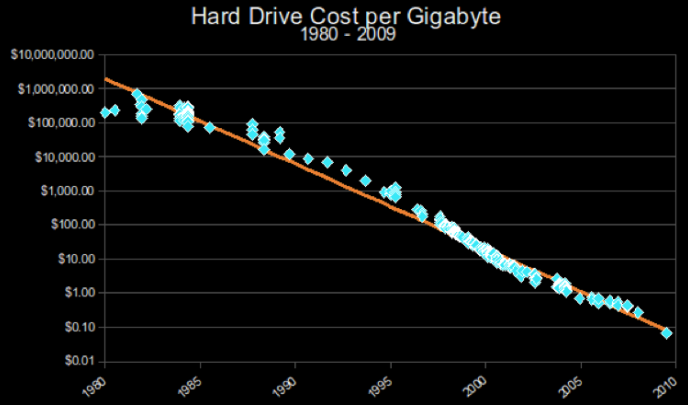
User-stakeholders:

- Public and Veterinary Health, and Food Safety Authorities, Wildlife/Nature management;
- Clinicians, General Practitioners, Veterinary practices;
- Scientific community, Research Labs (Microbiological, Clinical, Food safety labs);
- Food industry;

Across domains, sectors and pathogens

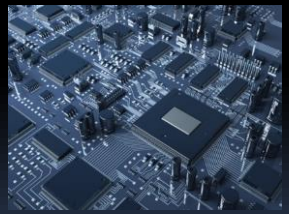


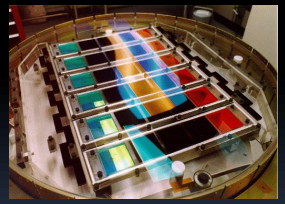
Exponentially cheaper devices



Moore's law

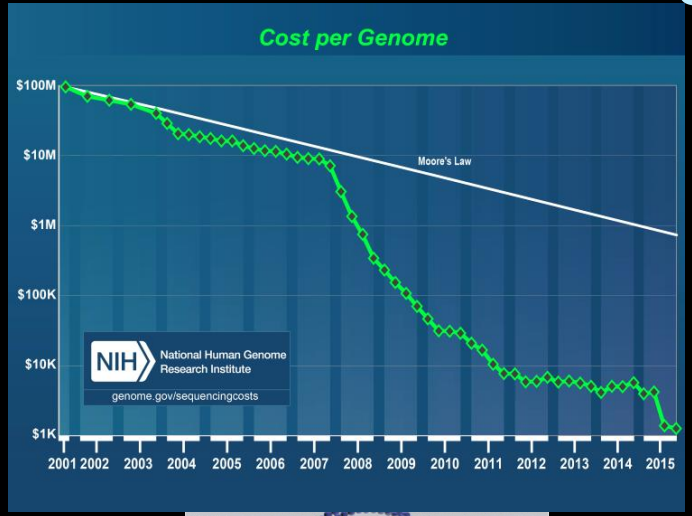
Exponential growth





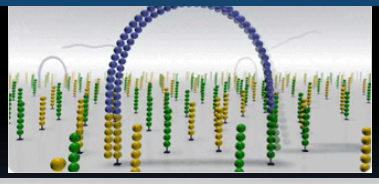
Elektronics → Sensors → Data

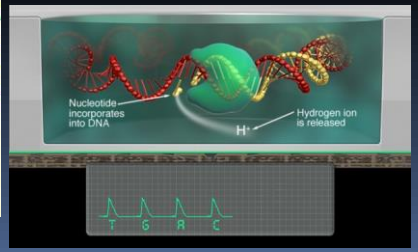
Moore’s law in genetics



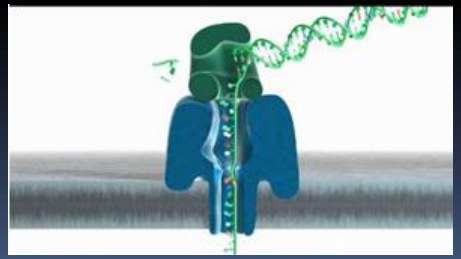
Human genome sequencing
1990-2003: 13yrs /2.7 Bn USD
2016: ~days/1000 USD
2020: ?????

CCD!
- 2006 X Prize 10M, 100 genoms 30 days, $10k - cancelled
- Microarray
- Mass spectroscopy
- Digital microscopy
- …





Oxford Nanopore
100Mb,$900



National Cancer Institute
National Human Genome Research Institute

The Cancer Genome Atlas Data:
Navigating the Data Portal and the Cancer Genomics Hub

The Cancer Genome Atlas
http://cancergenome.nih.gov/

3.2Bn nucleotides / human genome

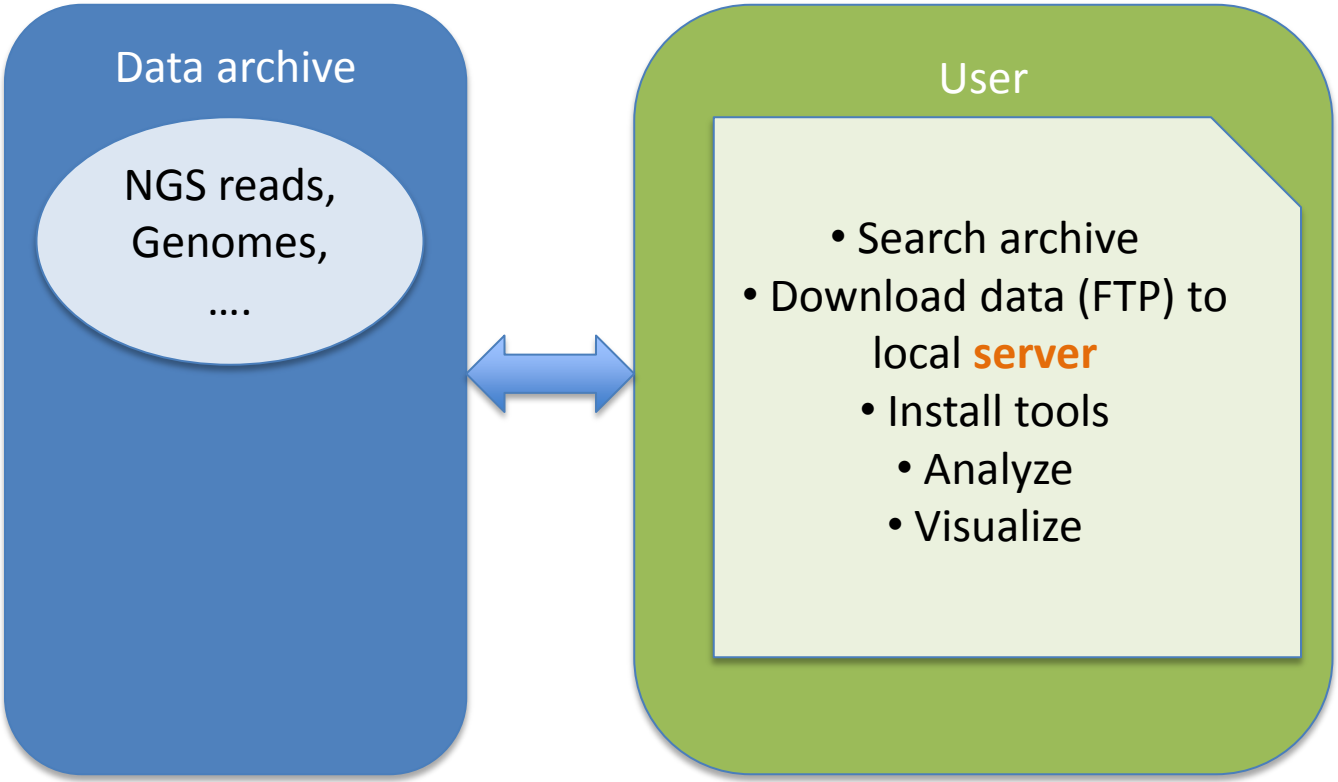
The Cancer Genome Atlas (TCGA) is a large-scale, collaborative effort led by the National Institutes of Health to map the genomic changes that occur in over 30 types of human cancer, including nine rare tumors. Its goal is to support new discoveries and accelerate the pace of research aimed at improving the diagnosis, treatment, and prevention of cancer.

TCGA is a community resource project. The information generated by TCGA is centrally managed and entered into databases as it becomes available, making the data rapidly accessible to the entire research community. By January 2014, TCGA had generated one petabyte of data for about 10,000 cases of tumor and matching normal tissue samples.

TCGA data are available in two data repositories: the TCGA Data Portal and the Cancer Genomics Hub. All data can be accessed directly from the TCGA Data Portal regardless of which repository houses the data file.

S.Spisak, K.Lawrenson,Y.Fu,I.Csabai, ... M. Freedmann. Nature Medicine doi:10.1038/nm.3975 (2015)

Data exploration: Traditional approach



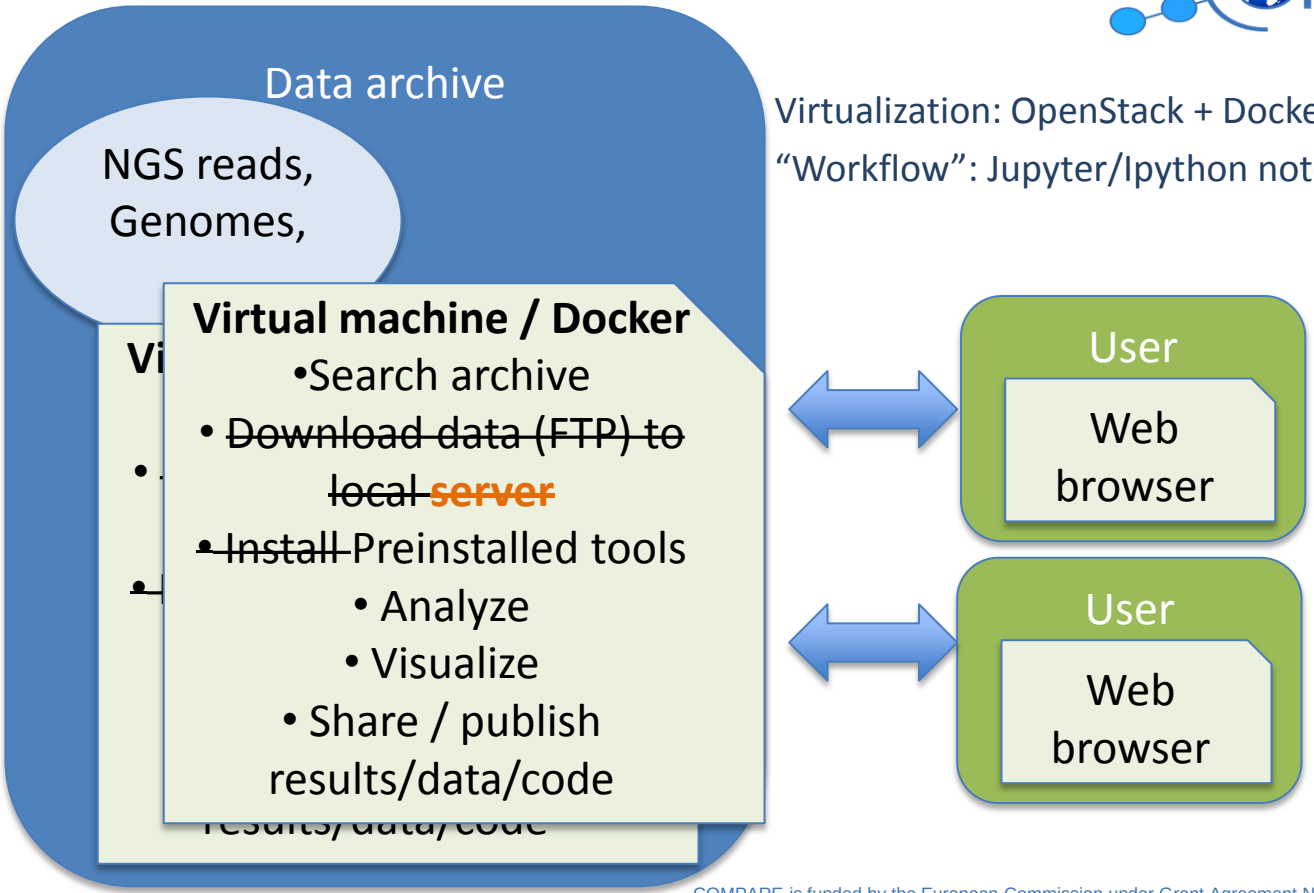
COMPARE is funded by the European Commission under Grant Agreement No 643476

Challenges

- Big Data – downloading data is not optimal/possible
- Data sharing is not enough – share data + complete processing pipeline + result figures, tables, …

COMPARE is funded by the European Commission under Grant Agreement No 643476

Cloud+notebook approach



Virtualization: OpenStack + Docker

"Workflow": Jupyter/Ipython notebook

COMPARE is funded by the European Commission under Grant Agreement No 643476

wignercloud.compare-europe.hu/hub/

Kooplex Hub | Hello icsabai! | Log off | Change Password

Kooplex

Infrastructure for flexible collaboration

Worksheets

Run pre-compiled worksheets to hide program code and focus on the problem.

Worksheets hide the complexity of notebooks from end users interested in scientific output.

➜ to worksheets

jupyter

Jupyter notebooks

Run existing or create Jupyter notebooks from existing projects to process your data or author your own projects, notebooks and share with others.

➜ Browse notebooks

Gitlab

Manage your project, add members to it, file issues.

➜ to Gitlab

owncloud

Owncloud

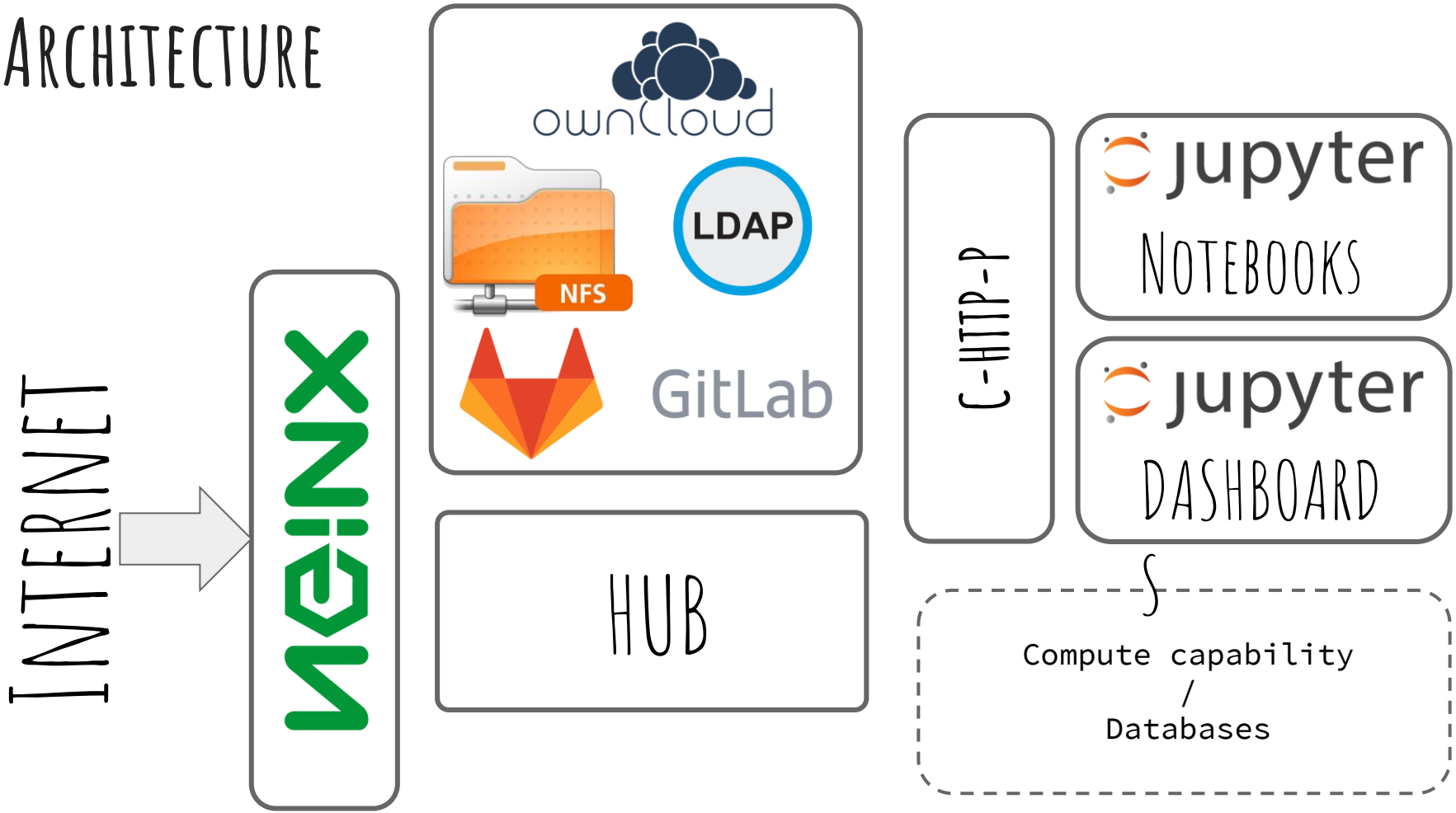
Manage easily your data files through the owncloud... (just as easy as in Dropbox! :))

➜ to Owncloud

- About
- Contact

Designed and built with all the love in the world by the **Kooplex Team**. We try to maintain it too! :)

Architecture

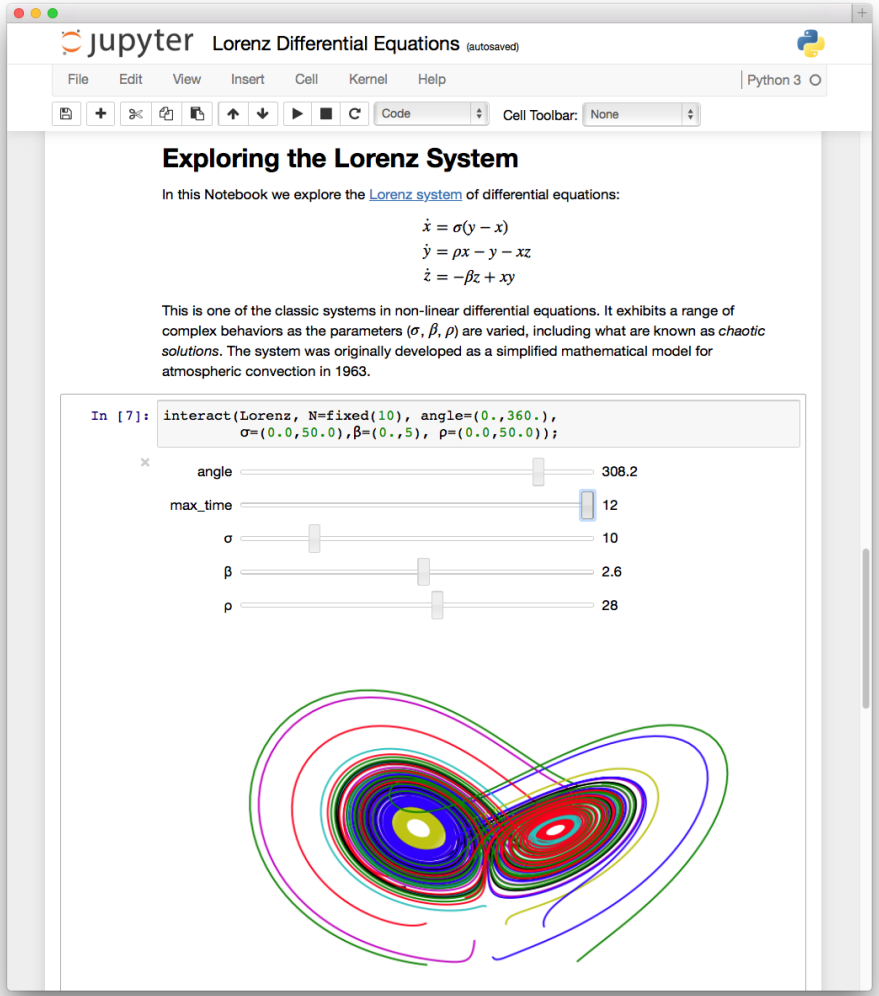


Components

- Nginx/C-HTTP-P [networking/routing, user⇔resources]
- Docker [compartmentalization of services and notebooks]
- LDAP [authentication and user database]
- NFS [persistent user workspace]
- HUB [front page + service spawner, django powered website]
- Gitlab [project management, versioning, collaborative forum]
- Jupyter [dev. env., sandbox, DASHBOARDS]
- Own Cloud [small data drag & drop, sharing]
- … + [workflow/pipeline services]

Jupyter

- Integrate:
  - Executable code (Python, R, Matlab, SQL, bash, …)
  - Notes, descriptions, math equations, …
  - Results, figures, interactions
- Accessible from anywhere through browser
- Runs in the cloud



Ebola

www.compare-europe.hu/ena_europe_loco.html

In [12]:
```
width, height = 650, 500
flu_map = folium.Map(location=[47, -17], zoom_start=3,
                     tiles='OpenStreetMap', width=width, height=height)
```

Add point to the map object

- Let's make point area proportional to number of cases
  - This is miseleading, beacuse somewhere all the cases around have the sample position (Europe), and somewhere the positions are more scattered (Shanghai)

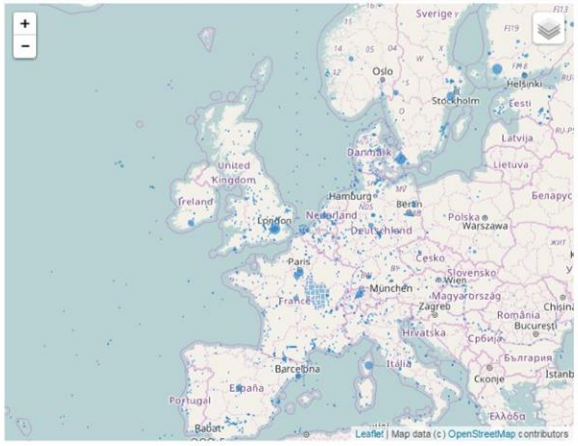
In [13]:
```
for i in xrange(len(uniq_locs_w_acc)):
    loc=(uniq_locs_w_acc.iloc[i]['lat'],uniq_locs_w_acc.iloc[i]['lon'] )
    name='Number of cases: '+str(uniq_locs_w_acc.iloc[i]['count'])
    name+='   Accesions: '+uniq_locs_w_acc.iloc[i]['acc_list']
    size=uniq_locs_w_acc.iloc[i]['count'] ** 0.5

    flu_map.circle_marker(location=loc, radius=1e3*size,
                          line_color='none',fill_color='#3186cc',
                          fill_opacity=0.7, popup=name)
```

And finally draw the map

In [15]:
```
inline_map(flu_map)
```

Out[15]:

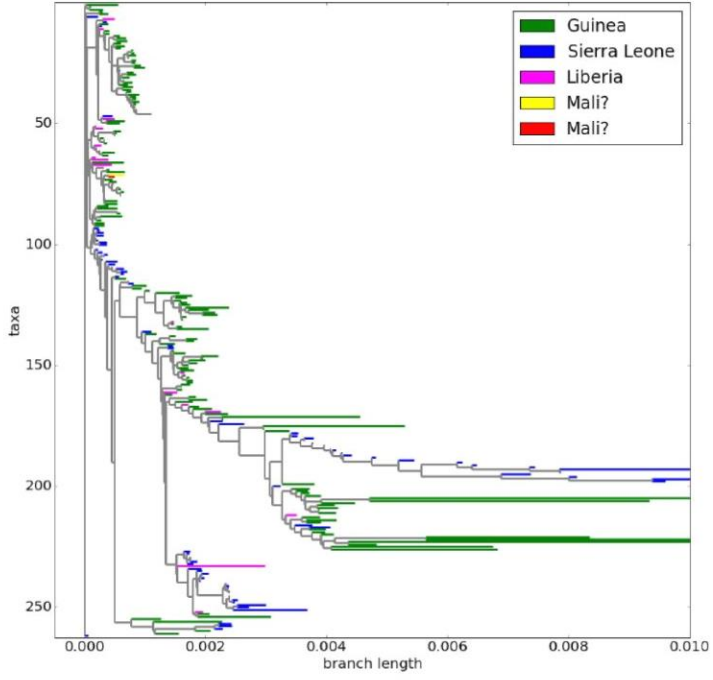



```
%matplotlib inline

#some settings
matplotlib.rc('font', size=20)
matplotlib.rcParams['lines.linewidth'] = 3
matplotlib.rcParams['figure.figsize'] =(16,16)

fig,ax=plt.subplots()
#custom Legend
gui_proxy = plt.Rectangle((0, 0), 1, 1, fc="green")
sle_proxy = plt.Rectangle((0, 0), 1, 1, fc="blue")
lib_proxy = plt.Rectangle((0, 0), 1, 1, fc="magenta")
dpr2_proxy = plt.Rectangle((0, 0), 1, 1, fc="yellow")
dpr1_proxy = plt.Rectangle((0, 0), 1, 1, fc="red")
ax.legend([gui_proxy,sle_proxy,lib_proxy,dpr2_proxy,dpr1_proxy],
          ['Guinea','Sierra Leone','Liberia','Mali?','Mali?'])

#draw tree
def my_label(clade):
        return None
Phylo.draw(tree,my_label,axes=ax,xlim=(-0.0005,0.01))
```



Global Sewage Metagenome

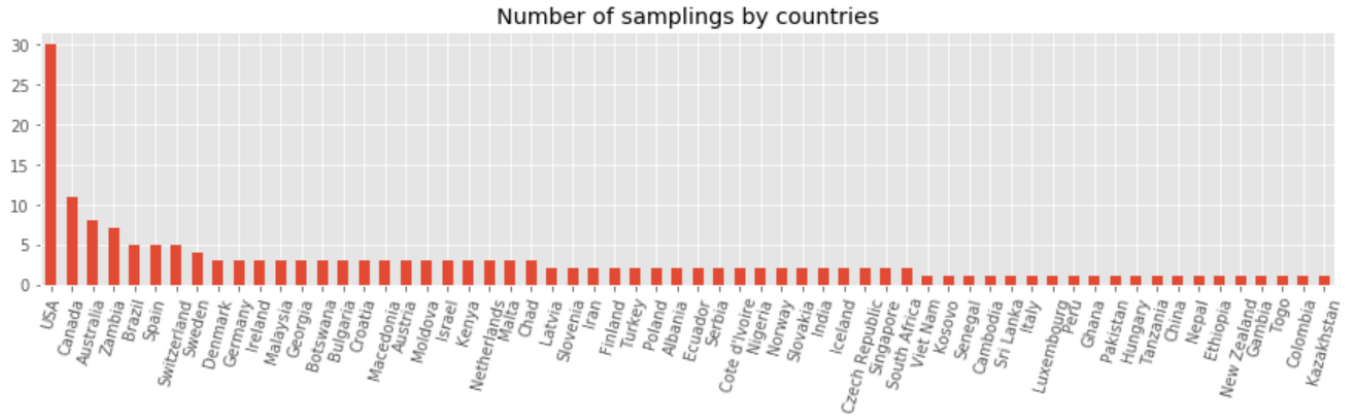
jupyter sewage-metadata Last Checkpoint: 16 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Markdown CellToolbar

Number of samplings

In [4]:
```
sewage_metadata['country'].value_counts().plot.bar(figsize=(15,3), title="Number of samplings by countries")
plt.xticks(rotation=75)
plt.show()
```



In [5]:
```
# For label purposes
sewage_metadata['date_fact']= sewage_metadata['collection_date']
sewage_metadata['type_fact']= sewage_metadata['sewage_type']
sewage_metadata = sewage_metadata.apply(lambda x: (pd.factorize(x)[0]) if x.name in ['date_fact', 'type_fact'] else x)
```

Cosmology without Dark energy

jupyter darkEnergy Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Help

Code CellToolbar

Calculate ISW for non-LCDM model without Dark Energy

Growth function

The definition if growth function:

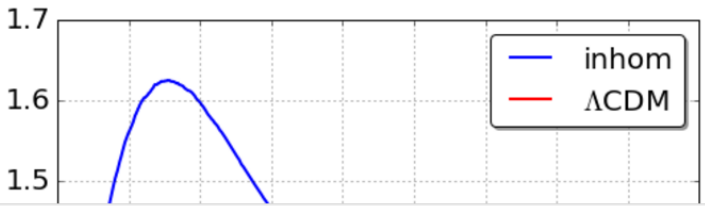
$$D_1(z) = \frac{H(z)}{H(0)} \int_z^\infty \frac{(1+z)}{H^3(z)} dz' \left[\int_0^\infty \frac{(1+z)}{H^3(z)} dz' \right]^{-1}$$

EdS:

$$\int_{9.0}^\infty \frac{(1+z)}{H^3(z)} dz' = \frac{27}{8} \int_{9.0}^\infty \frac{(1+z)}{(1+z)^{9/2}} dz' = \frac{27}{8} \left[\frac{-2}{5(1+x)^{5/2}} \right]_{9.0}^\infty = \frac{27}{8} \frac{1}{250\sqrt{10}} = 0.0042690375$$

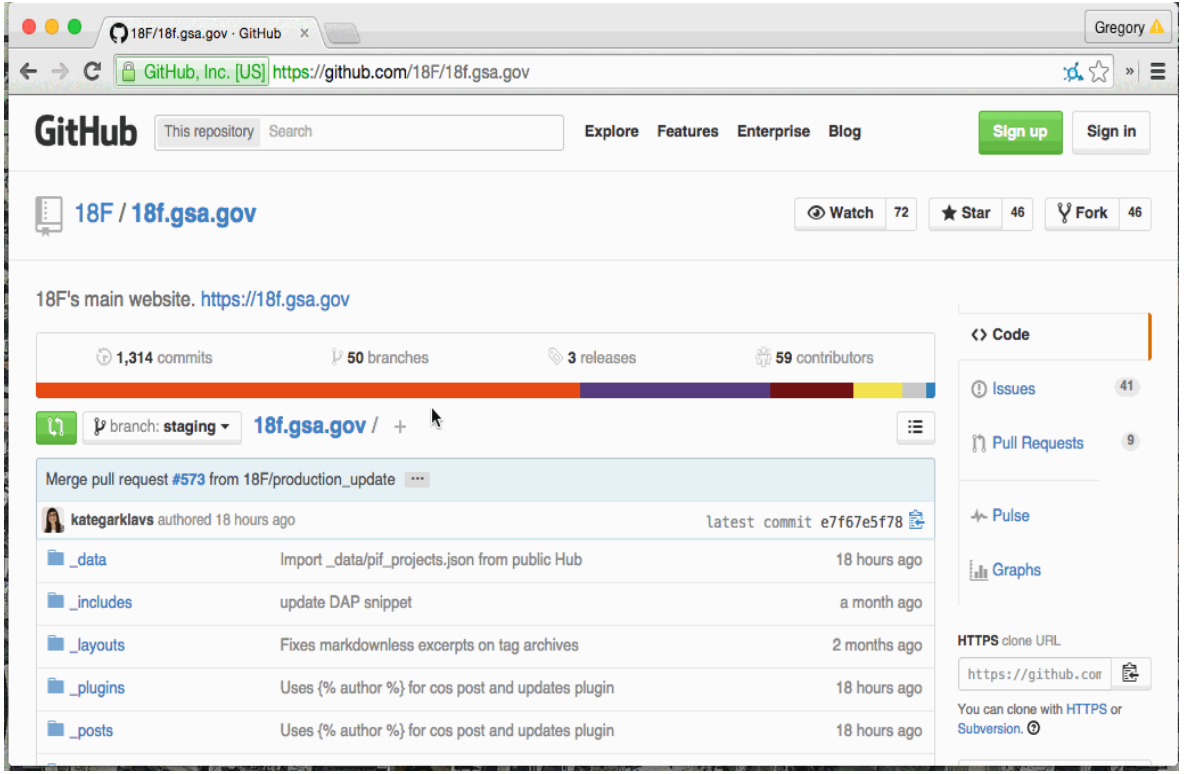
$$\int_{0.0}^\infty \frac{(1+z)}{H^3(z)} dz' = 1.35$$

In [36]: `plotModel(x)`

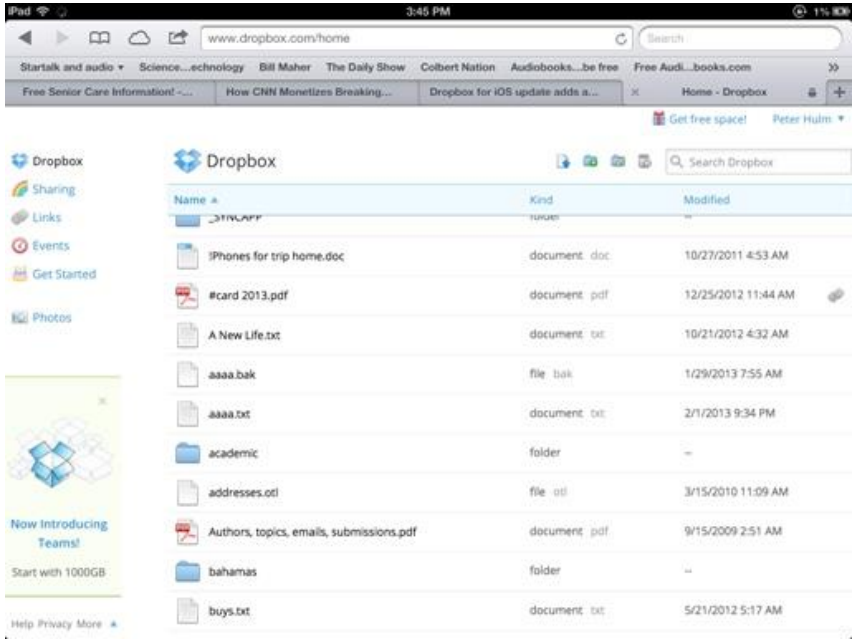


GitHub/GitLab

- project management, versioning,
- collaborative forum



Dropbox/OwnCloud



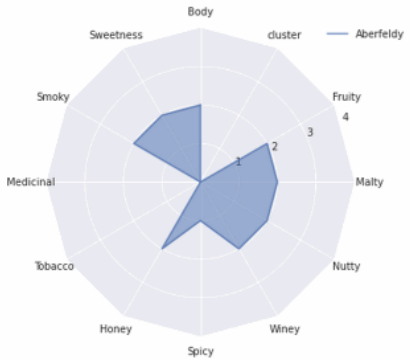
Worksheets/Dashboard – active services

Got Scotch?

If you like Aberfeldy you might want to try these five brands. Click one to see how its taste profile compares.

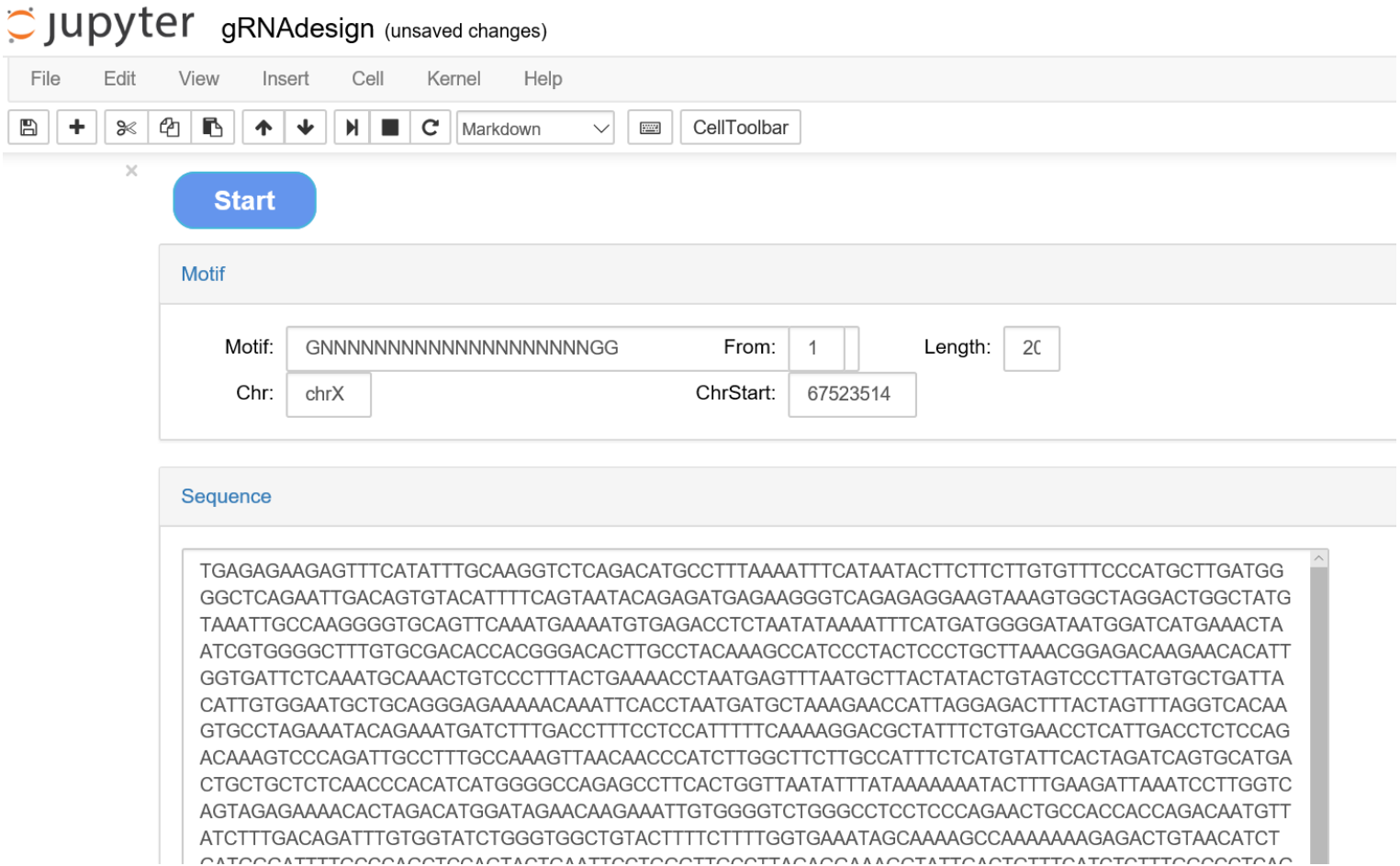
Scotch Aberfeldy

	Similarity
Benromach	0.972973
BlairAthol	0.972973
Benrinnes	0.962908
RoyalLochnagar	0.959202
Scapa	0.959043



Powered by data from https://www.mathstat.strath.ac.uk/outreach/nessie/nessie_whisky.html and inspired by analysis from http://blog.revolutionanalytics.com/2013/12/k-means-clustering-86-single-malt-scotch-whiskies.html. This dashboard originated as a Jupyter Notebook.

CRISPR DNA Editing – Guide RNA designer



Neural recordings - CRCNS

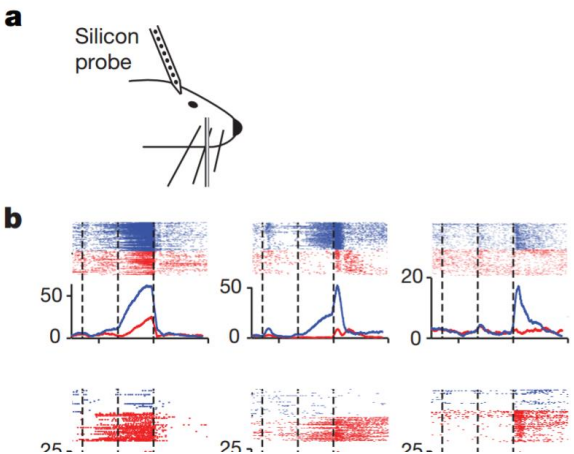
ARTICLE

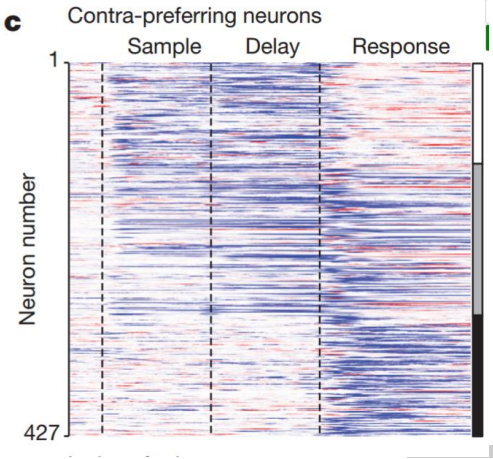
doi:10.1038/nature14178

A motor cortex circuit for motor planning and movement

Nuo Li[1], Tsai-Wen Chen[1], Zengcai V. Guo[1], Charles R. Gerfen[2] & Karel Svoboda[1]

Activity in motor cortex predicts specific movements seconds before they occur, but how this preparatory activity relates to upcoming movements is obscure. We dissected the conversion of preparatory activity to movement within a structured motor cortex circuit. An anterior lateral region of the mouse cortex (a possible homologue of premotor cortex in primates) contains equal proportions of intermingled neurons predicting ipsi- or contralateral movements, yet unilateral inactivation of this cortical region during movement planning disrupts contralateral movements. Using cell-type-specific electrophysiology, cellular imaging and optogenetic perturbation, we show that layer 5 neurons projecting within the cortex have unbiased laterality. Activity with a contralateral population bias arises specifically in layer 5 neurons projecting to the brainstem, and only late during movement planning. These results reveal the transformation of distributed preparatory activity into movement commands within hierarchically organized cortical circuits.





jupyter crcns_data1 Last Checkpoint: 17 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Code CellToolbar Dashboard View:

```
# plot the rasters
rasterplot('RealCueTimes',trials,
            spike_times,nid=neuron_n,
            hue='response',
            window=window,
            binsize=20)
```