

Duplicate Detection and Linking with Spark (Discussion)

Oliver Meyn

<https://elephant.tech>



Global Biodiversity Information Facility

Free and Open Access to Biodiversity Data

639,562,662
OCCURRENCES

1,611,321
SPECIES

15,260
DATASETS

776
DATA PUBLISHERS

Sharing biodiversity
data for re-use

[Learn about GBIF](#)
[Publish your data through GBIF](#)
[Technical infrastructure](#)

Providing evidence for
research and decisions

[Using data through GBIF](#)
[Enabling biodiversity science](#)
[Supporting global targets](#)

Collaborating as a
global community

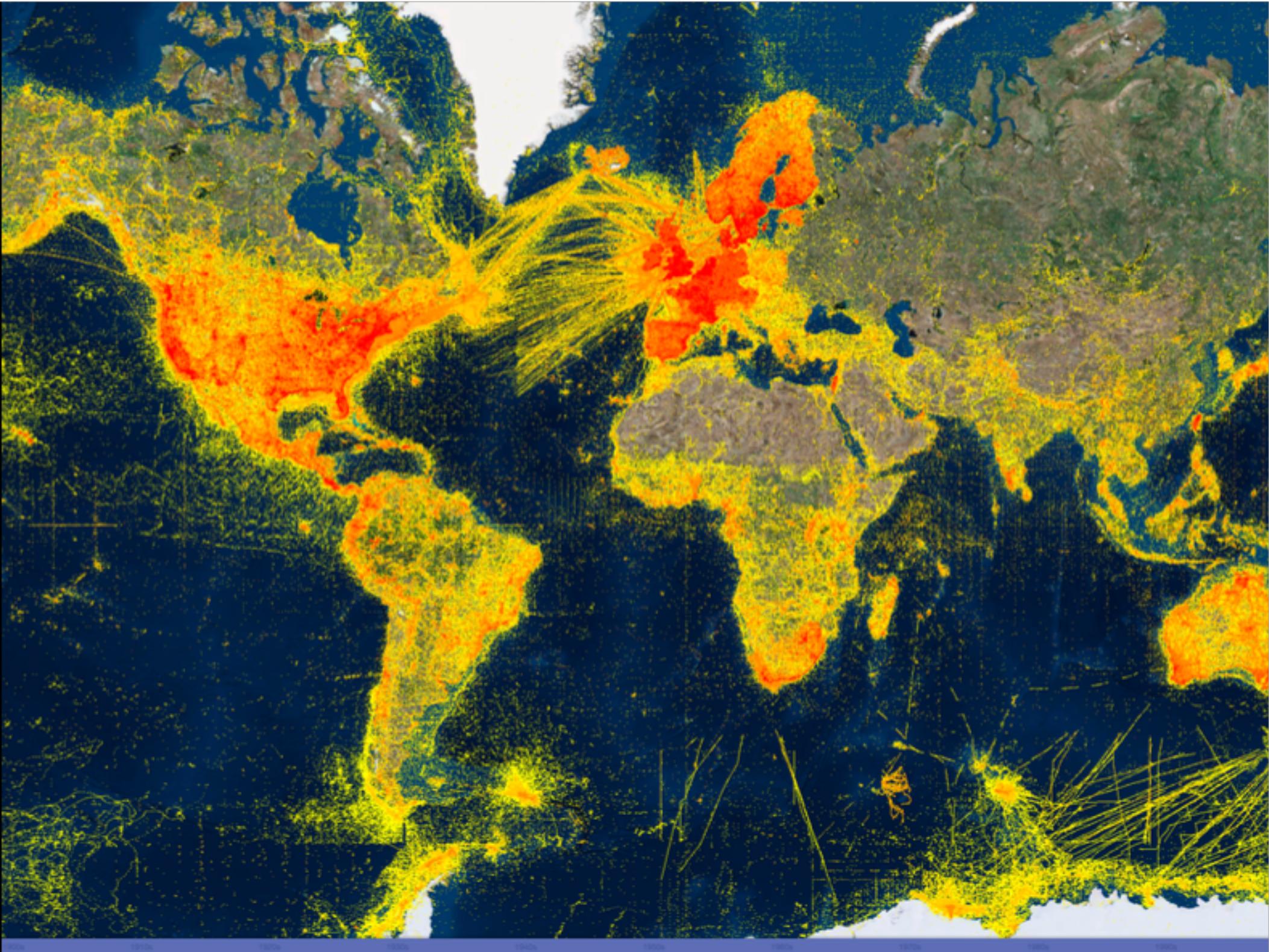
[Current Participants](#)
[How GBIF is funded](#)
[Enhancing capacity](#)

Search news items and information pages...

Search

gbif.org

~600 Million “occurrence records”





Occurrence Records

What Where When Who (a lot of birds)

Cleaned up occurrences

ID	Species	Latitude	Longitude	Date	Collector
1065606016	<i>Passer domesticus</i>	43.64529	-79.3807	2015-01-26	ashleytruong
1065606044	<i>Passer domesticus</i>	43.63946	-79.4371	2015-02-05	alexandragrayson
1065613620	<i>Sciurus carolinensis</i>	43.64465	-79.46762	2015-01-28	Leonardo Navarro

Raw and dirty occurrences

ID	Species	Latitude	Longitude	Date	Collector
1065606016	"Passer domesticus "	43.6452900 0000003	-79.380799 9996	Jan 26	Toronto
1065606044	Passer Domestiucs	43,63946	79,4371	10/11/12	alexandragrayson
1065613620	N/A	43.6	-79.4	2015-01-28	Leonardo Navarro



Flickr @marcobellucci



Flickr @Jacob Pilich

Duplicate occurrences?

ID	Species	Latitude	Longitude	Date	Collector
1065606016	Passer domesticus	43.64529	-79.3807	2015-01-26 11:45am	ashleytruong
1065606044	Passer domesticus	43.64529	-79.3807	2015-01-26 11:50am	null
1065613620	Passer domesticus	43.645	-79.38	2015-01-26	ashleytruong

Ideal: links with confidence

ID	123456	7890123	1065606044
1065606016	0.7	0.85	0.98
1065606044	0.6	0.75	
1065613620	0.8	0.91	0.75

Spark to the rescue?

- Straight SQL won't solve this
- Iteration over sets
- Clustering within each species?
- Minimize an error function of lat/lng, date, collector?



Thanks!

Oliver Meyn
oliver@elephant.tech
<https://elephant.tech>