(Innovative) methodologies to approach locational data quality issues

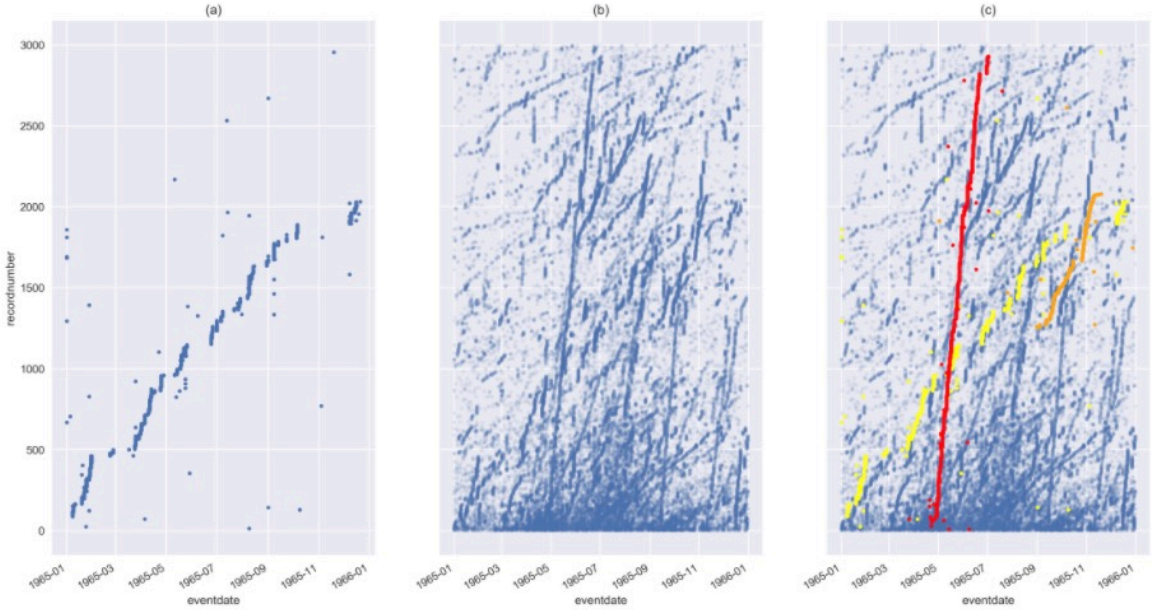
Nicky Nicolson

Senior Research Leader, Biodiversity Informatics, Royal Botanic Gardens, Kew

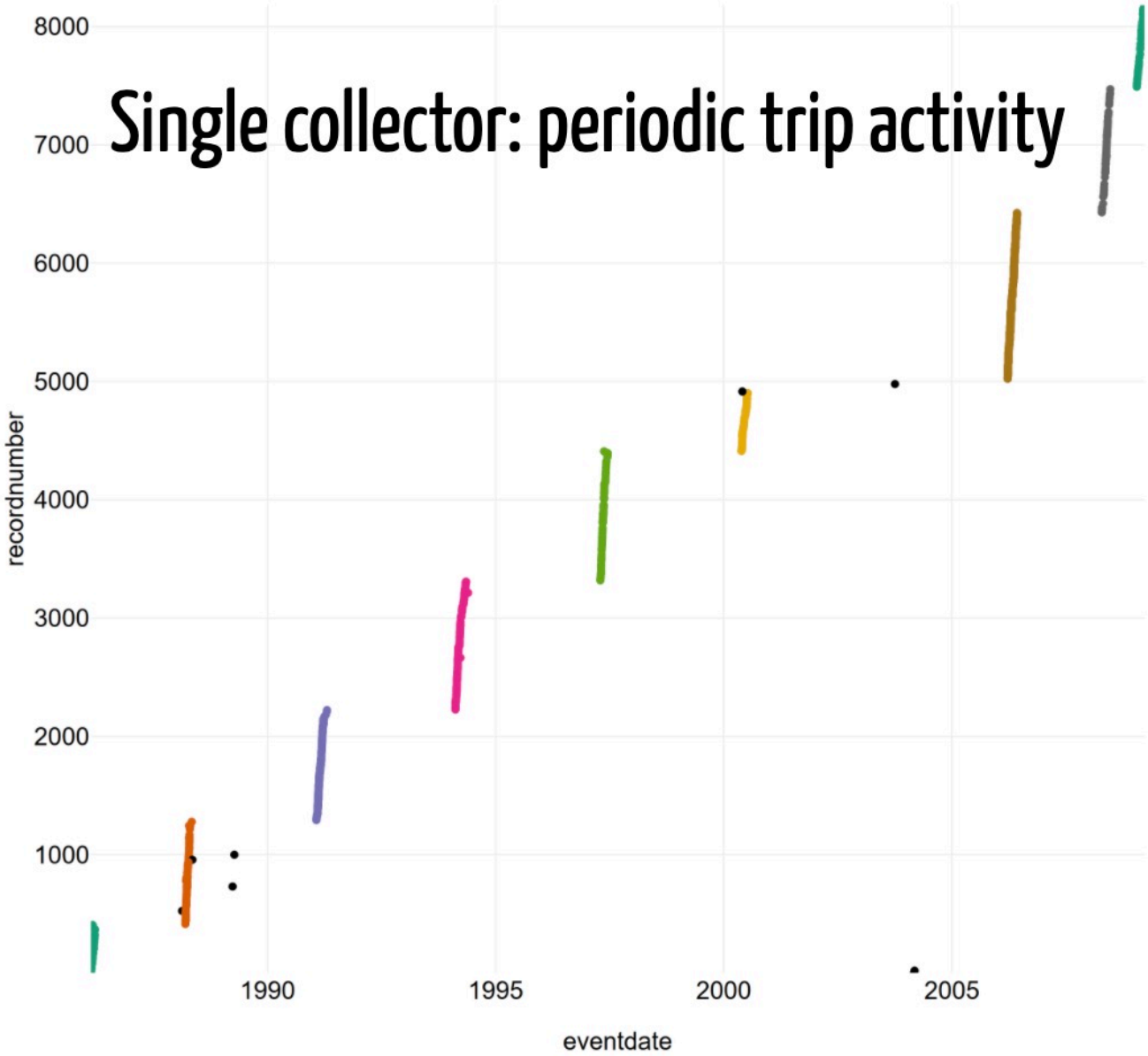
Background

- Data mining project using GBIF preserved specimen data to assert new entities:
    - Collector
    - Collecting trip
    - Collecting "state" run
- Exploit botanical collectors practice - used Darwin Core terms `recordedBy`, `eventDate` and `recordNumber` as input into clustering algorithm (`DBSCAN`) to detect collectors
- Iterative process
    - Once *collectors* established, further cluster their preserved specimens to detect *collecting trips*
    - Once *collecting trip* established, datamine specimens to detect *collecting state runs* - intense days of collecting activity, separated by days with lesser activity, (likely travelling) (Hidden Markov model)

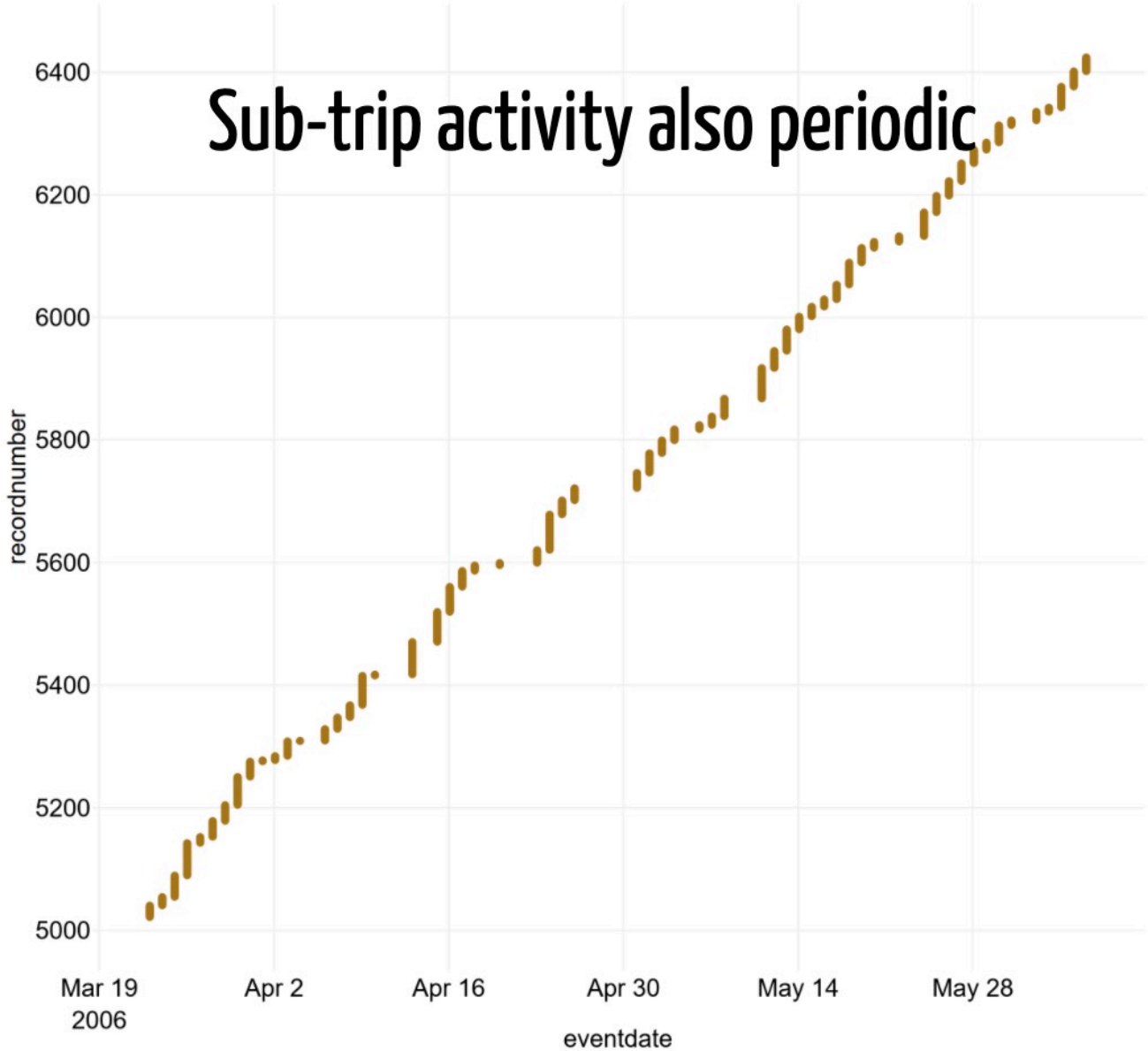
Data mining - collectors and trips



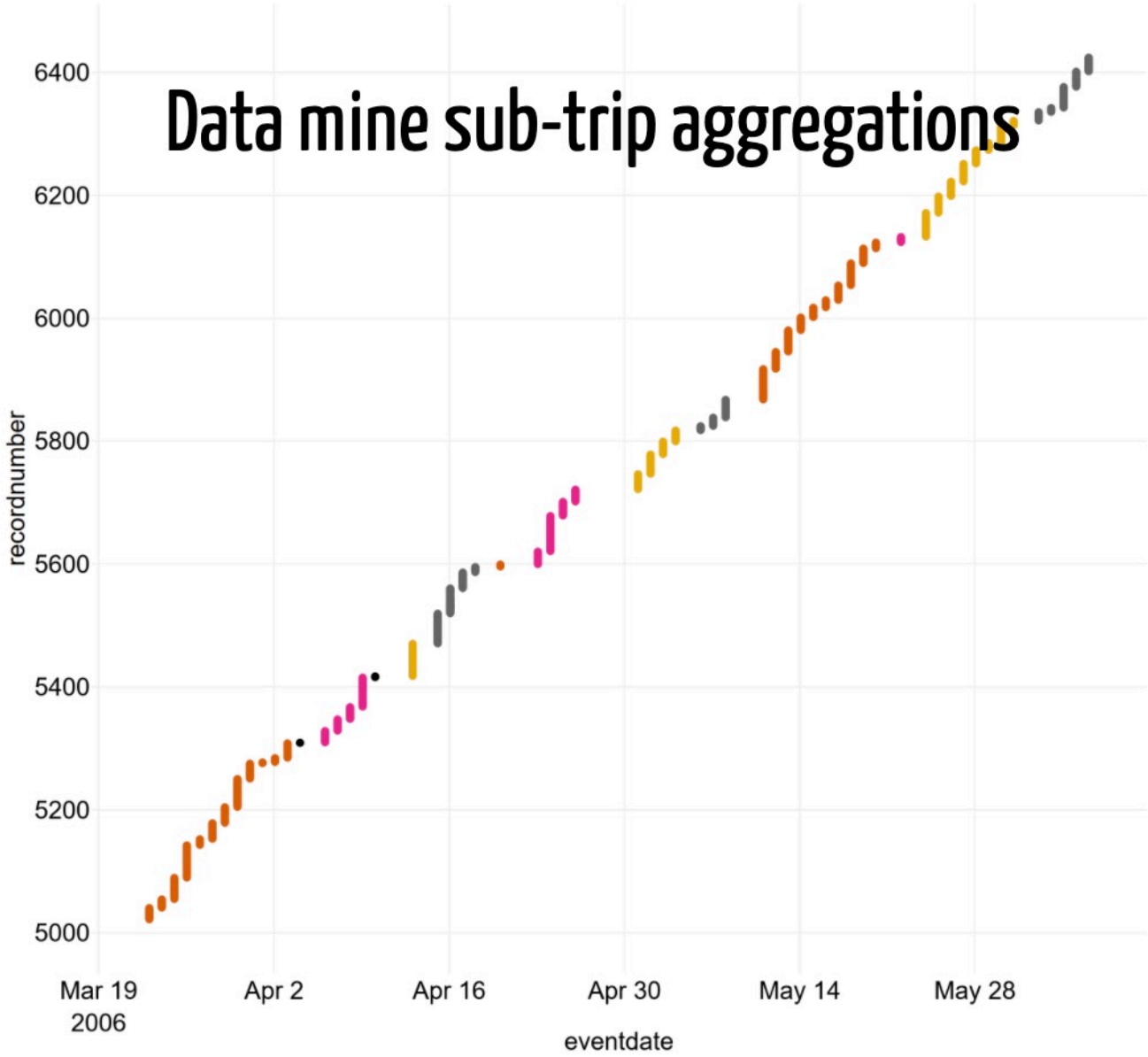
Single collector: periodic trip activity



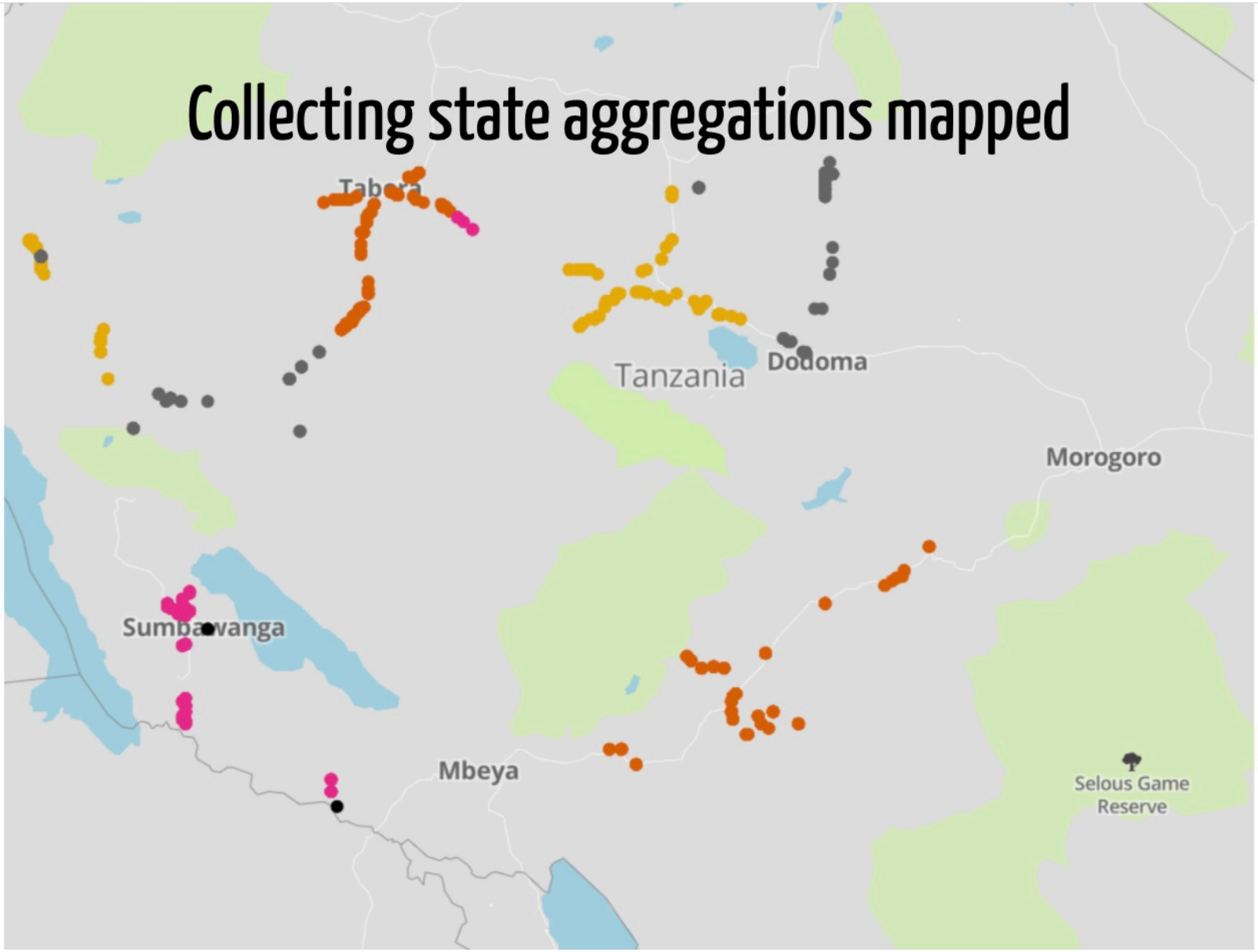
Sub-trip activity also periodic



Data mine sub-trip aggregations



Collecting state aggregations mapped



Applications: identifying duplicate specimens

Reading across

- Multiple specimens from a single collecting event common practice in botany
- Duplicate specimens managed in separate institutional repositories, georeferenced separately.
- Duplicates (traditionally) identified using collector, recordnumber and year
- Data mining shared collector identity simplifies duplicate resolution

Given duplicate resolution:

- Counted number of georeferences that could be mobilised via duplicate links (> 1.1 million)
- Identified institutions that could work together (network view of holdings and overlaps)

"Reading across": from collecting event to generated specimens

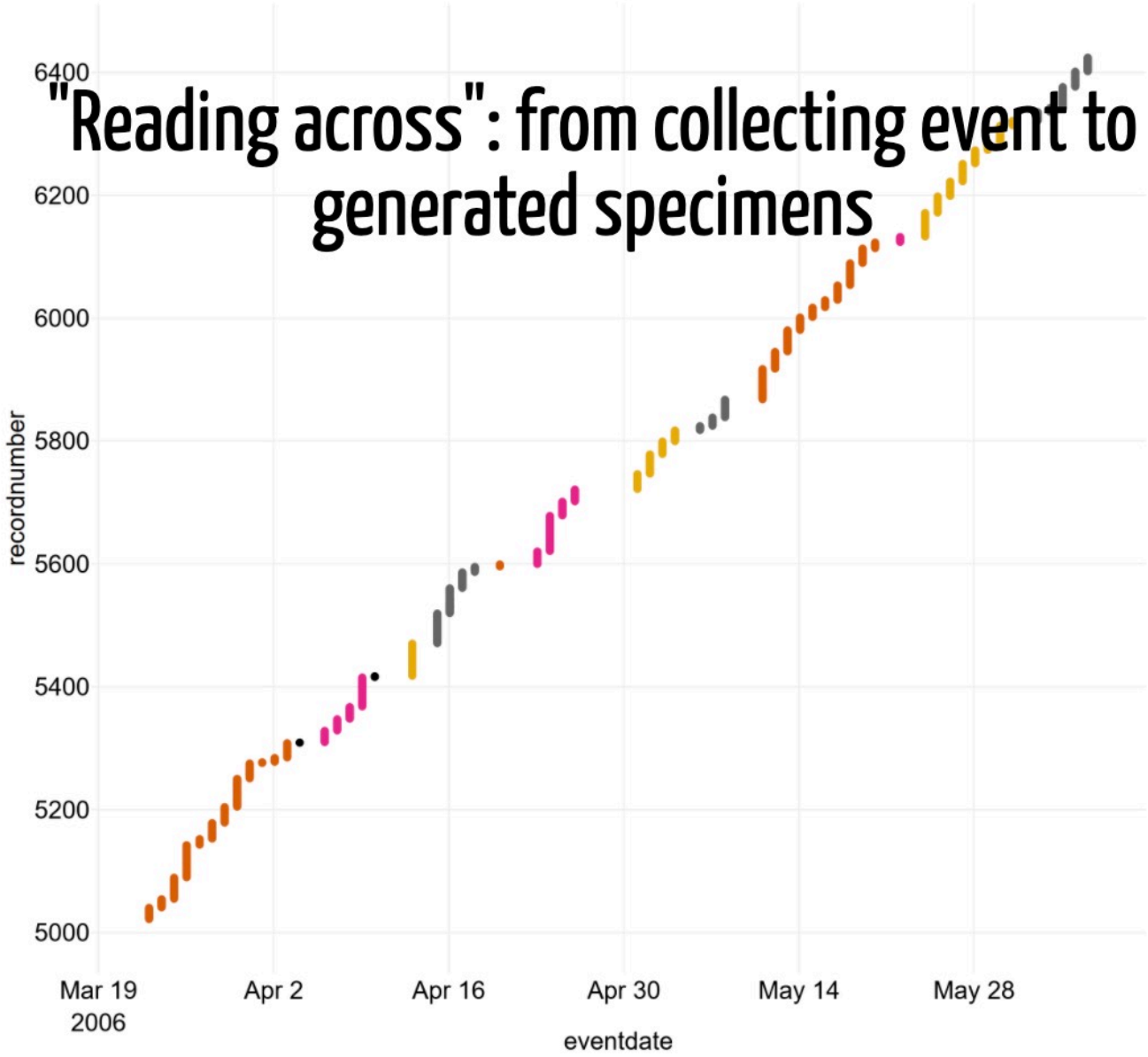
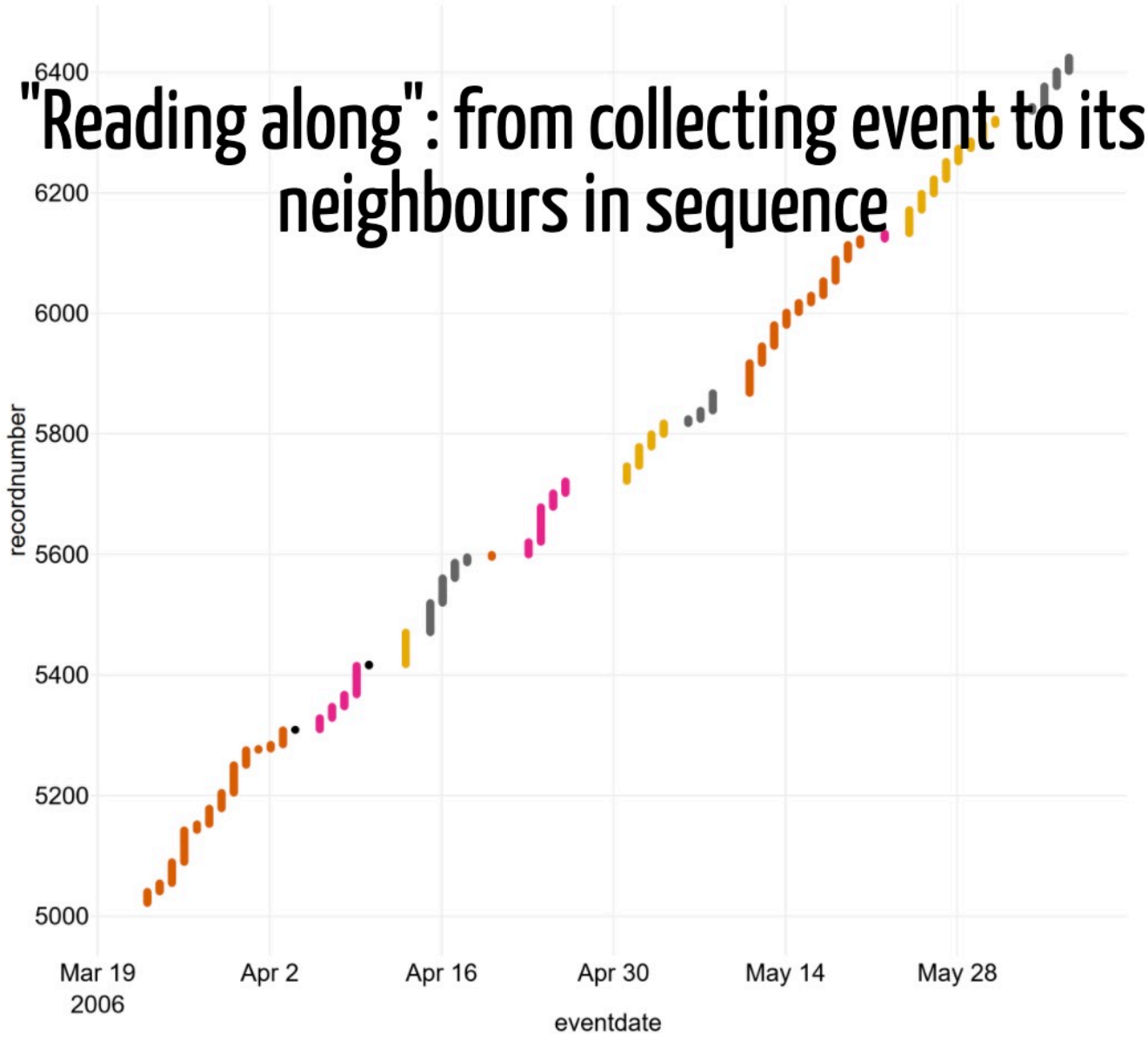


TABLE 6.2: Distributed curation of specimens arising from a common collection event, worked example (Hutchison 5738)

recordedBy	scientificName	held in	cited	digitised	type	georef'd	imaged
P. C. Hutchison & J. K. Wright	Solanum sanchez-vegae S.Knapp	F	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	Solanum aligerum Schltdl.	F	-	✔	-	-	-
Hutchison, P.C.	Solanum sanchez-vegae S.Knapp	K	✔	✔	✔	✔	✔
Paul C. Hutchison | J. Kenneth Wright	Solanum cutervanum Zahlbr.	MO	-	✔	-	-	-
P. C. Hutchison	Solanum sanchez-vegae S.Knapp	NY	-	✔	✔	✔	✔
P. C. Hutchison	Solanum sanchez-vegae S.Knapp	NY	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	Solanum sanchez-vegae S.Knapp	P	✔	-	-	-	-
P. C. Hutchison & J. K. Wright	Solanum sanchez-vegae S.Knapp	US	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	Solanum sanchez-vegae S.Knapp	USM	✔	-	-	-	-

"Reading along": from collecting event to its neighbours in sequence



Applications: exploiting collecting event sequence

Reading along

- Constrained clustering:
  - Dataset of specimens relating to single collecting trip
  - Use locality text and sequential record number
- Identify reused localities, even if variably recorded

Sequential ordering of locality texts

recordnumber	locality
10598.0	6-8 km below Mollepata on road to river valley of Río Tablachaca, R side of river
10602.0	just before Puente Chucusvalle over Río Tablachaca coming from Mollepata, R bank of river (other side of bridge in Ancash)
10603.0	Puente Chucusvalle over Río Tablachaca, L bank of river (other side of bridge in La Libertad)
10 604.0	Ancash; Pallasca; Pallasca; Puente Chucusvalle over Río Tablachaca, L bank of river (other side of bridge in La Libertad)
10611.0	NaN
10611.0	7-8 km above Puente Chucusvalle over Río Tablachaca, on road to Pallasca
10614.0	ca. 10 km above Puente Chucusvalle over Río Tablachaca, on road to Pallasca; small stream crossing road on steep slopes
10617.0	ca. 11 km above Puente Chucusvalle over Río Tablachaca, on road to Pallasca
10620.0	9 km from Pallasca on road to Cabanas, just outside village of Inaco
10624.0	11 km fr om Pallasca on road to Cabanas, past village of Huncachuqia
10625.0	19 km from Pallasca on road to Cabanas, 10 km beyon d Inaco
10628.0	22 km from Pallasca on road to Cabanas, 13-14 km beyond Inaco, just before village of Huandoval
10630.0	hillsides above Cabanas, on trail to small chapel

Conclusions

- Preserved specimens generated from human scale collection events
- Contextual information from collector, collecting trip etc relevant to georeferencing
- Large aggregated datasets enable contextual processing

References

- N. Nicolson, A. J. Paton, S. Phillips, and A. Tucker, “Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation” in 2018 IEEE 14th International Conference on e-Science (e-Science), 2018, pp. 125-135. doi:10.1109/eScience.2018.00028

- N. Nicolson "Automating the construction of higher order data representations from heterogeneous biodiversity datasets" (thesis) 2019 https://bura.brunel.ac.uk/handle/2438/19620

- GBIF dataset, Tracheophyta, preserved specimen doi:10.15468/dl.wjjrdk

Contact email: n.nicolson@kew.org