Identifying novel features from specimen data for the prediction of valuable collection trips

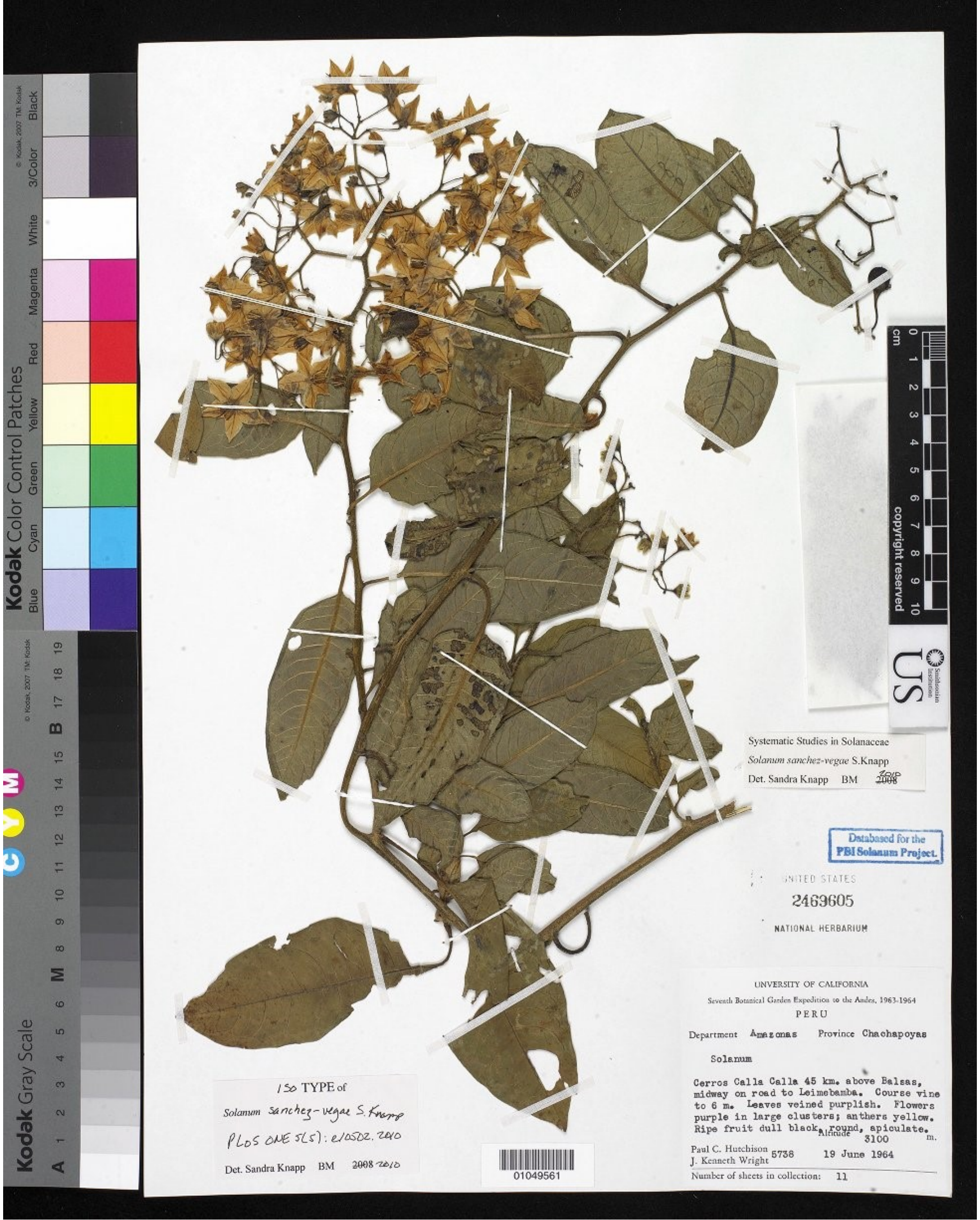
Nicky Nicolson[1,2], Allan Tucker[2]

1. Biodiversity Informatics & Spatial Analysis, Royal Botanic Gardens, Kew (UK). 2. Department of Computer Science, Brunel University London (UK).

Intelligent Data Analysis XVI, 26-28th October 2017, London (UK)

Outline

1. About the scientific domain:
    1. Examine a specimen
    2. Motivation for research
2. Method:
    1. Data exploration
    2. Data mining
    3. Applications:
        1. Data abstraction (grouping & feature definition)
        2. Classifier construction
3. Results
4. Conclusions



Systematic Studies in Solanaceae
Solanum sanchez-vegae S.Knapp
Det. Sandra Knapp BM ~~2008~~ 2010

Databased for the
PBI Solanum Project.

UNITED STATES
2469605
NATIONAL HERBARIUM

UNVERSITY OF CALIFORNIA
Seventh Botanical Garden Expedition to the Andes, 1963-1964
PERU
Department Amazonas Province Chachapoyas

Solanum

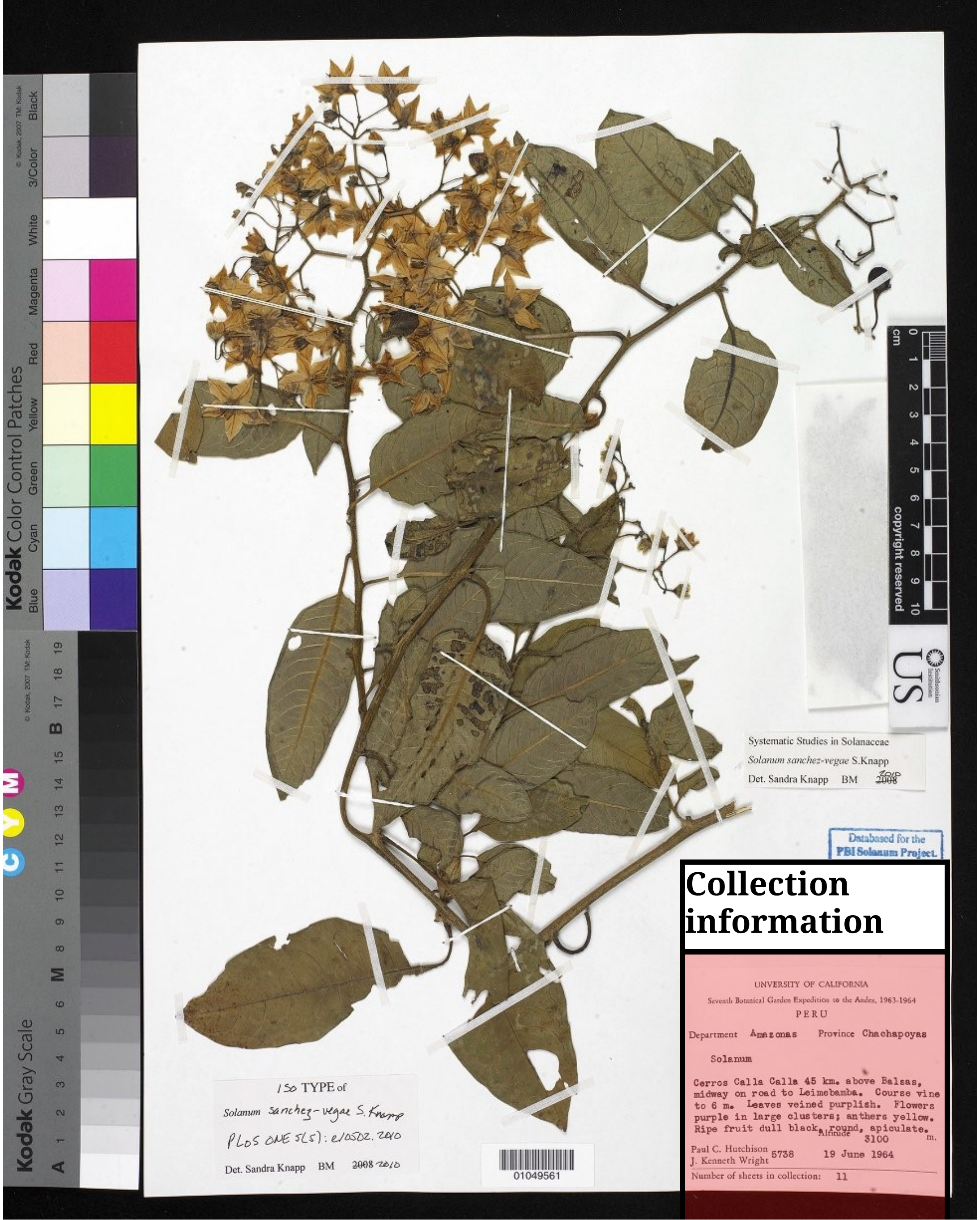
Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.
Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11

ISO TYPE of
Solanum sanchez-vegae S. Knapp
PLoS ONE 5(5): e10502. 2010
Det. Sandra Knapp BM ~~2008~~ 2010

01049561



Systematic Studies in Solanaceae
Solanum sanchez-vegae S.Knapp
Det. Sandra Knapp BM ~~2008~~ 2010

Databased for the
PBI Solanum Project.

ISO TYPE of
Solanum sanchez-vegae S.Knapp
PLOS ONE 5(5): e10502. 2010
Det. Sandra Knapp BM ~~2008~~ 2010

01049561

Collection information

UNVERSITY OF CALIFORNIA
Seventh Botanical Garden Expedition to the Andes, 1963-1964
PERU

Department Amazonas Province Chachapoyas

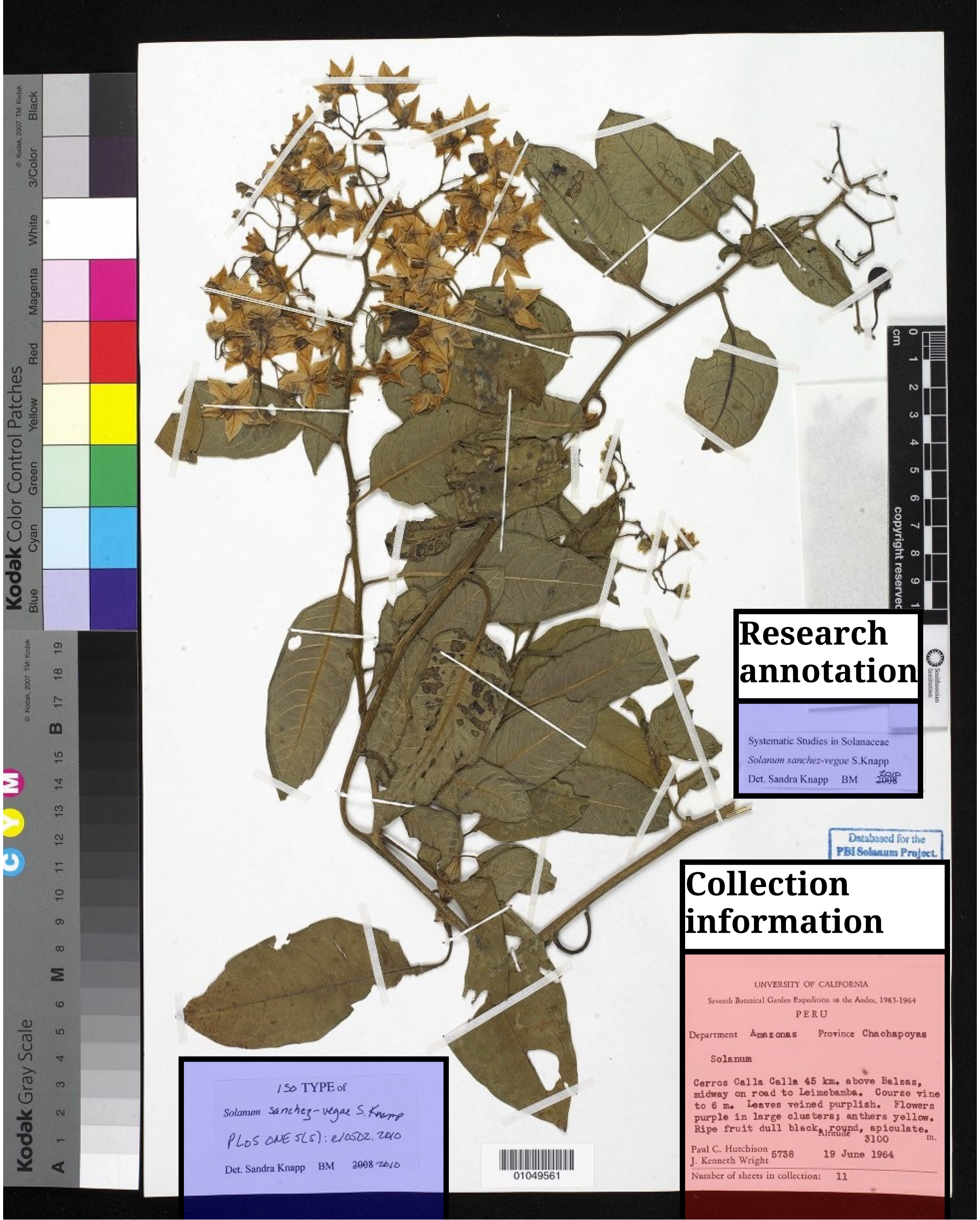
Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11



Research annotation

Systematic Studies in Solanaceae

Solanum sanchez-vegae S.Knapp

Det. Sandra Knapp BM ~~2008~~ 2010

Databased for the PBI Solanum Project.

Collection information

UNIVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11

ISO TYPE of

Solanum sanchez-vegae S. Knapp

PLoS ONE 5(5): e10502. 2010

Det. Sandra Knapp BM ~~2008~~ 2010

UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

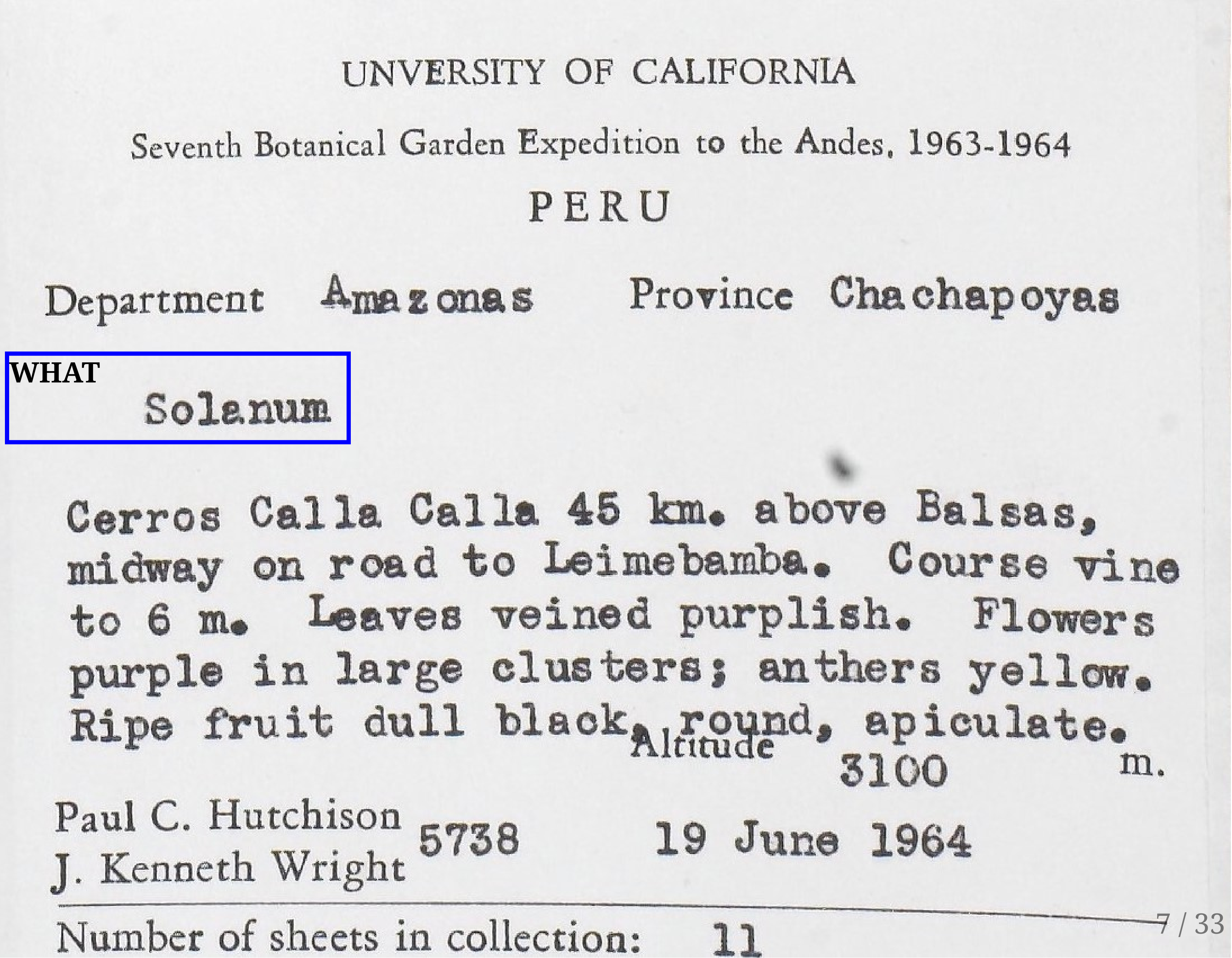
Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

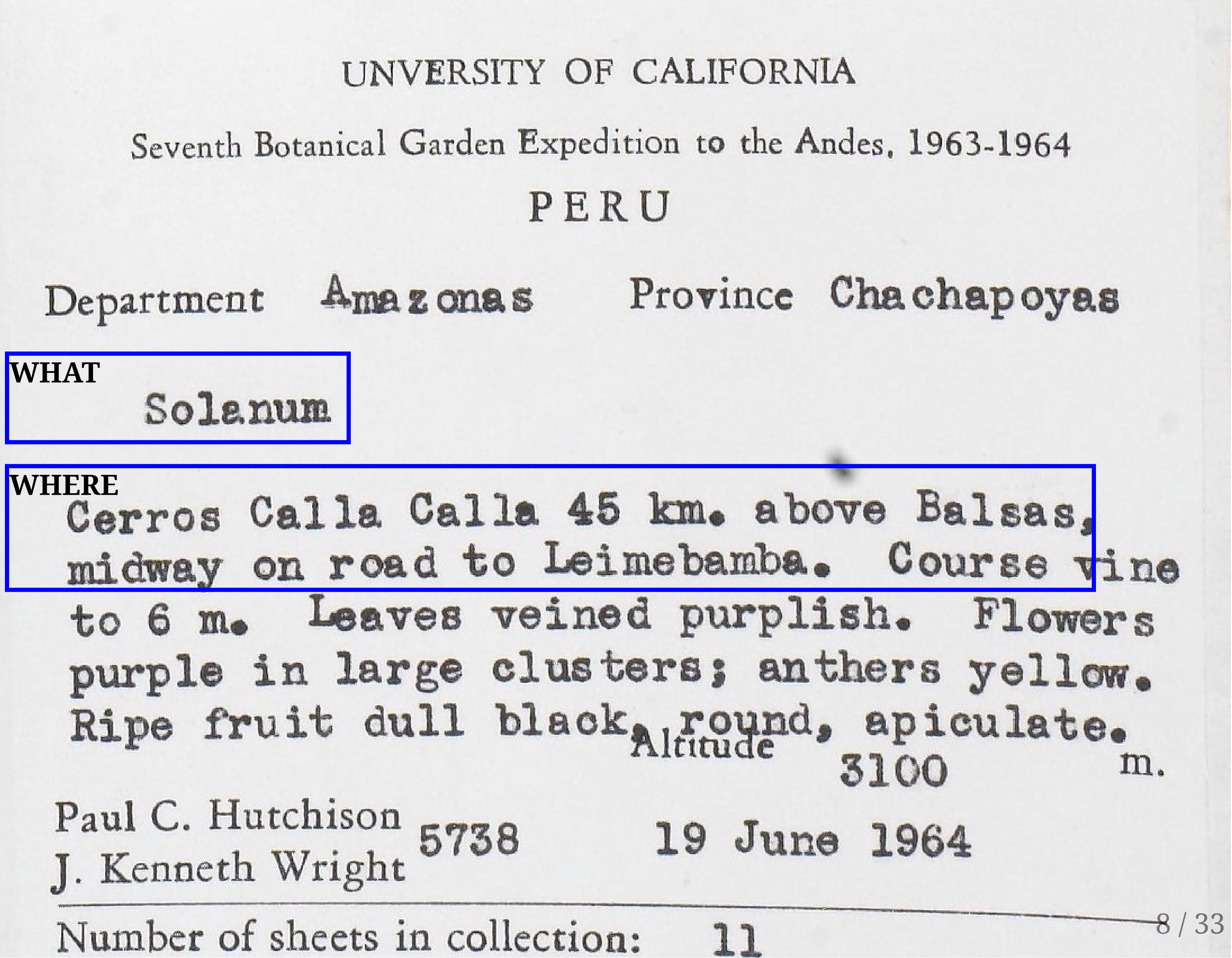
Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

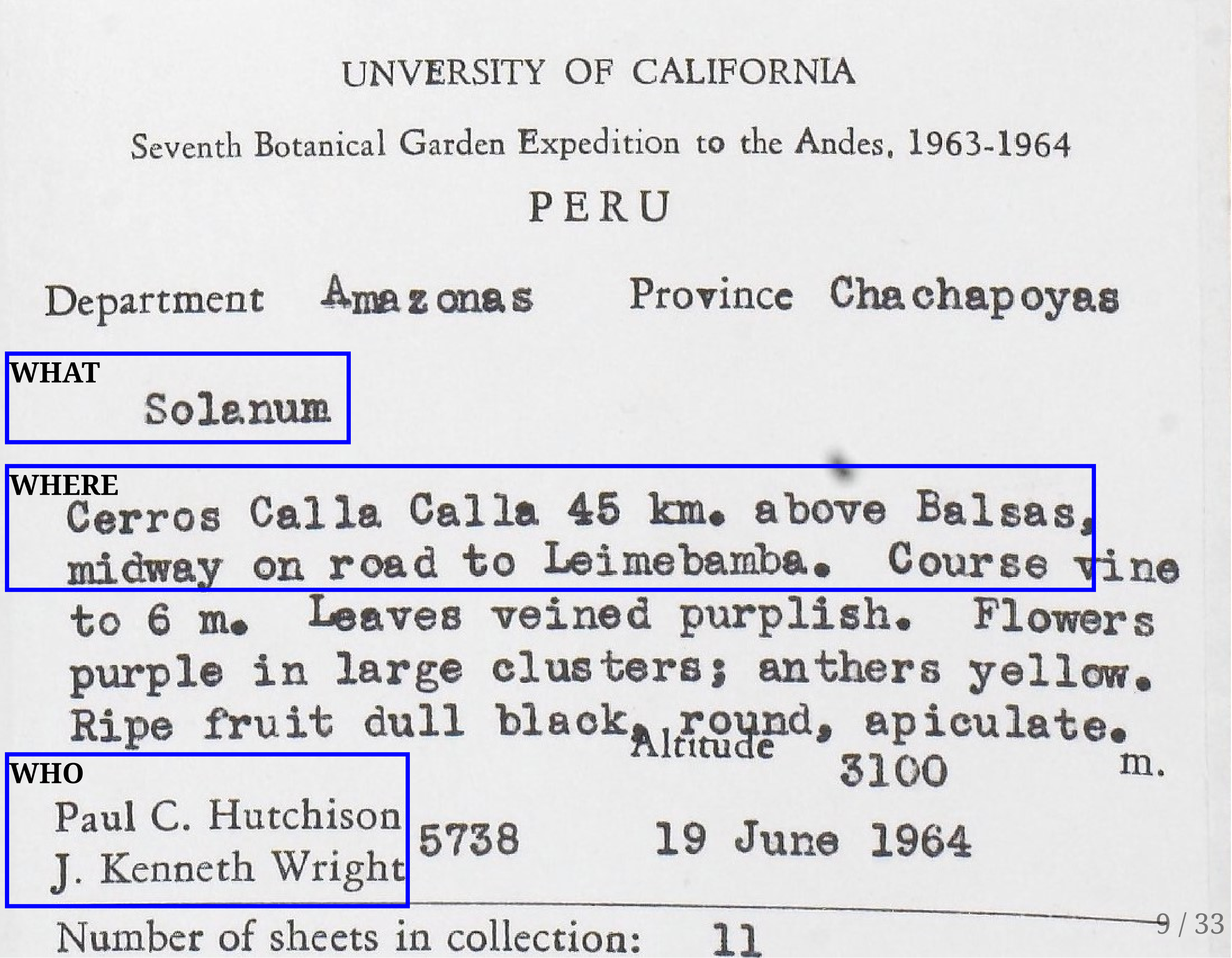
WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

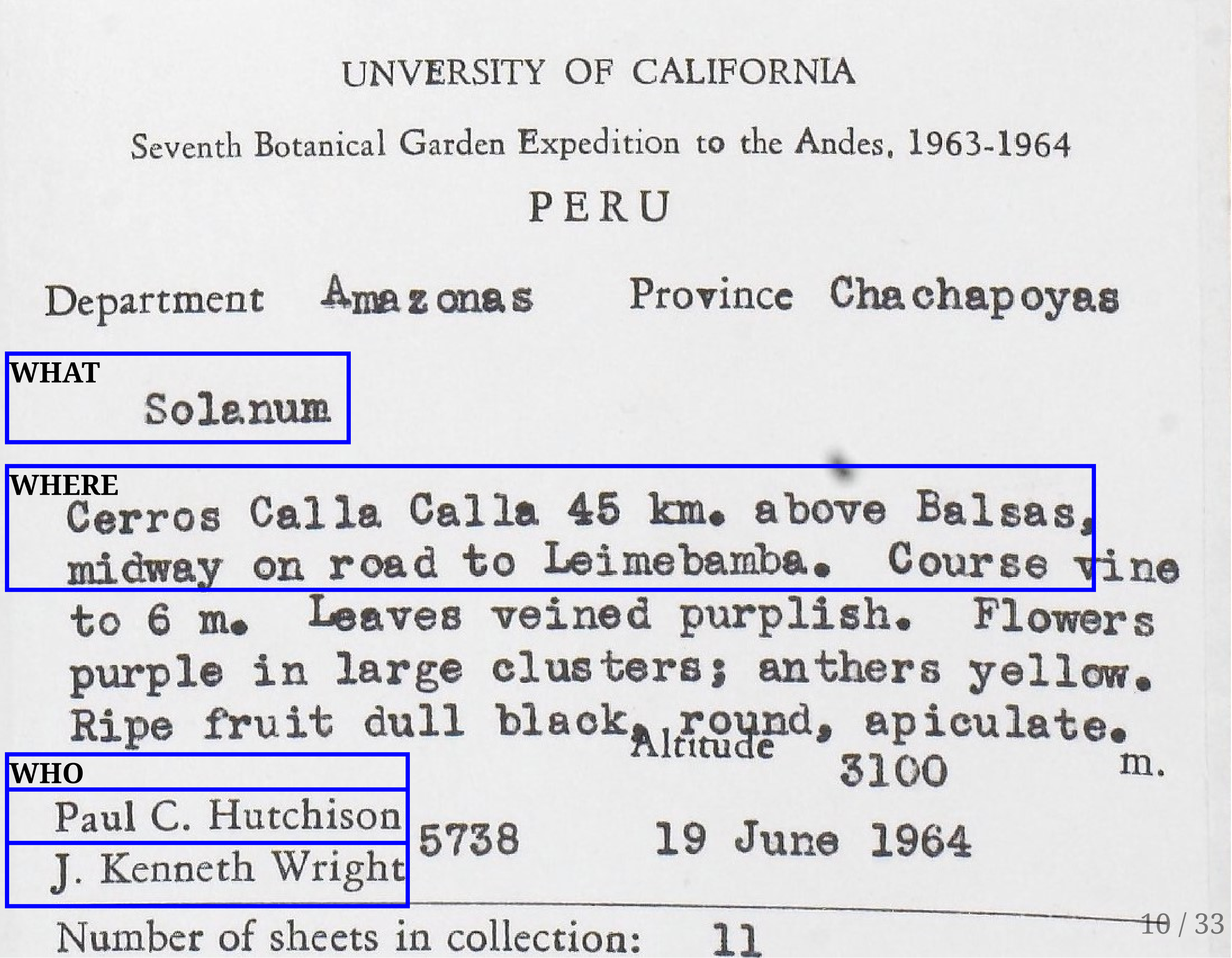
Altitude 3100 m.

WHO

Paul C. Hutchison
J. Kenneth Wright

5738 19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

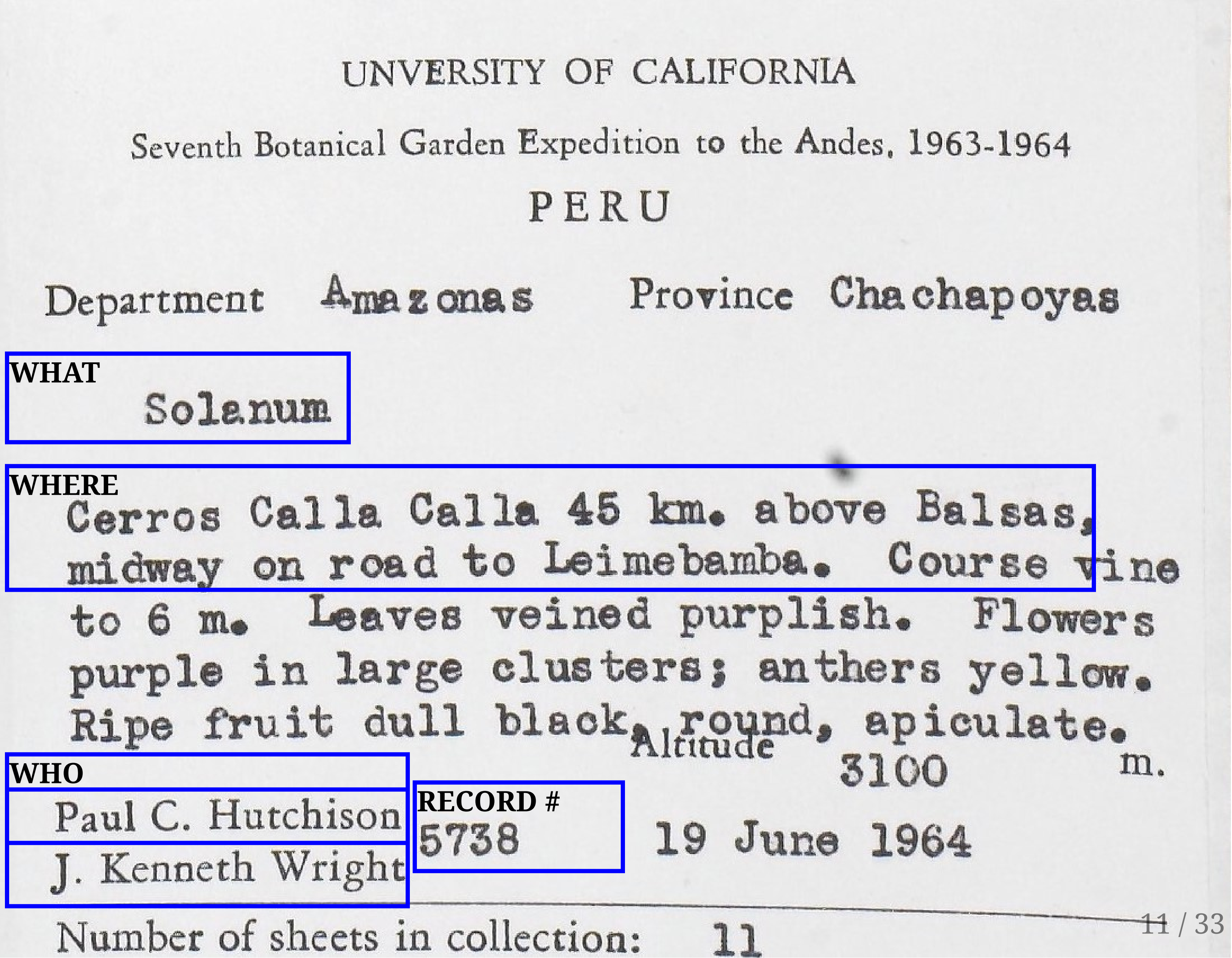
Altitude 3100 m.

WHO

Paul C. Hutchison 5738 19 June 1964

J. Kenneth Wright

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

WHO

Paul C. Hutchison

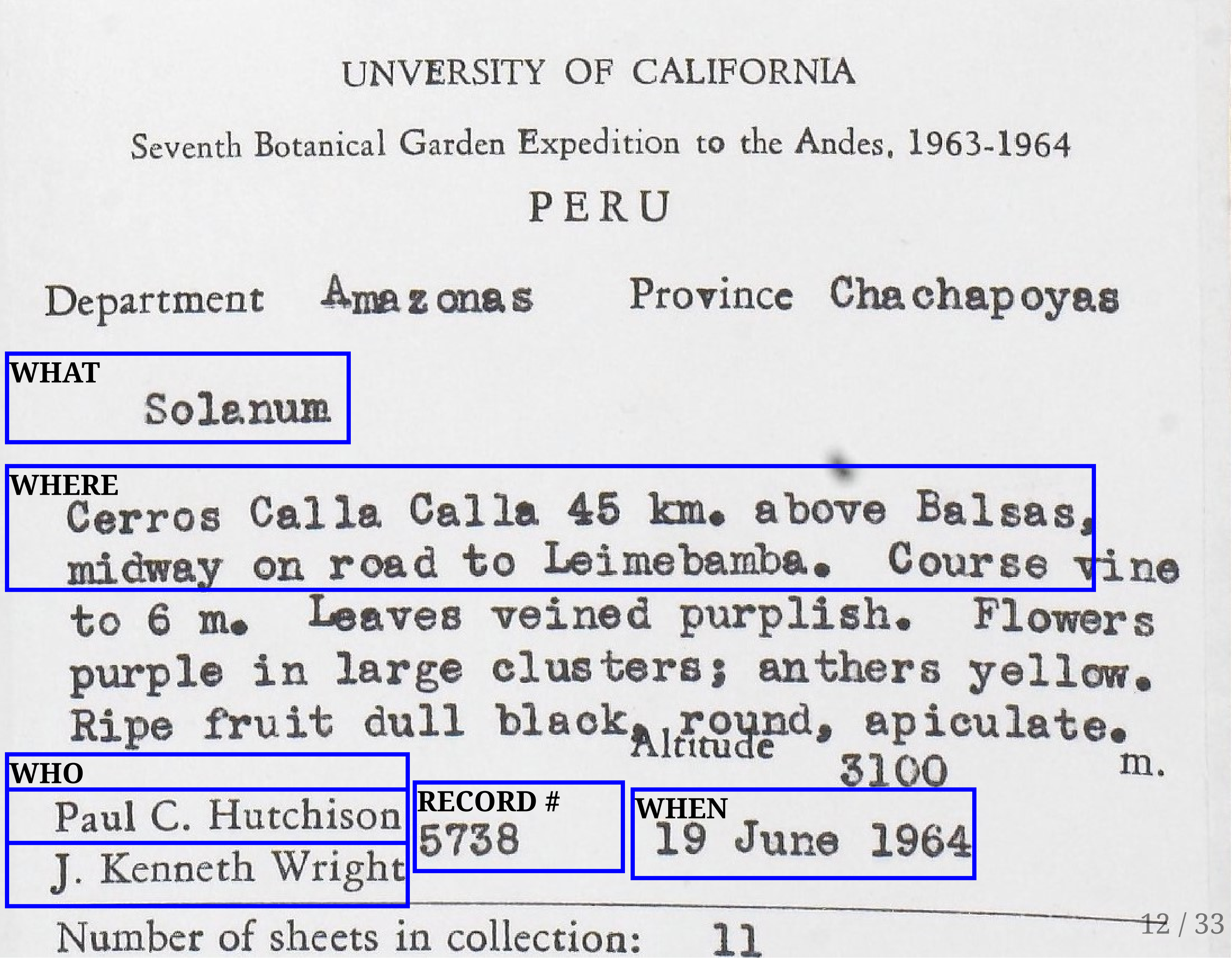
J. Kenneth Wright

RECORD #

5738

19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

WHO

Paul C. Hutchison

J. Kenneth Wright

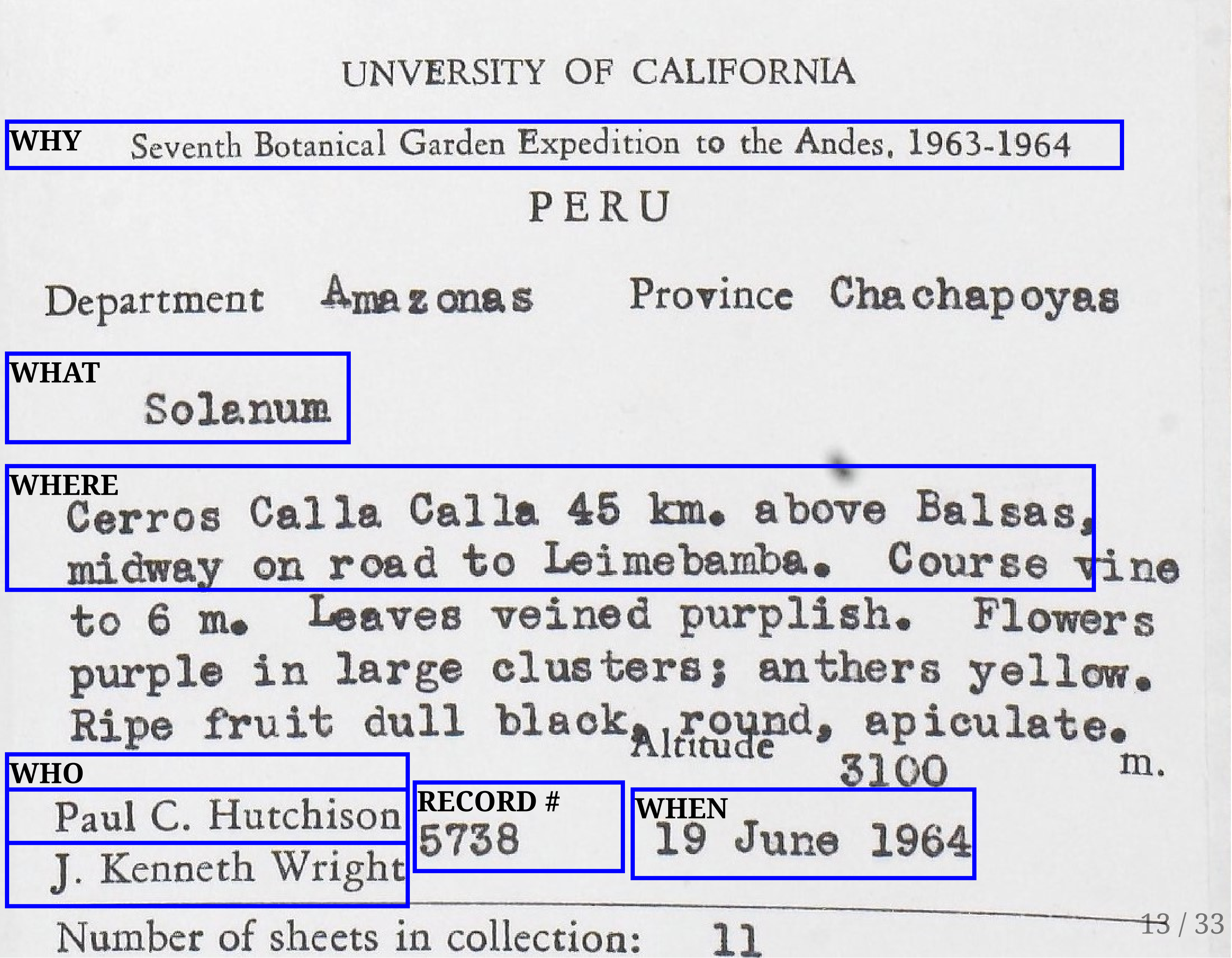
RECORD #

5738

WHEN

19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

WHY Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT Solanum

WHERE Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

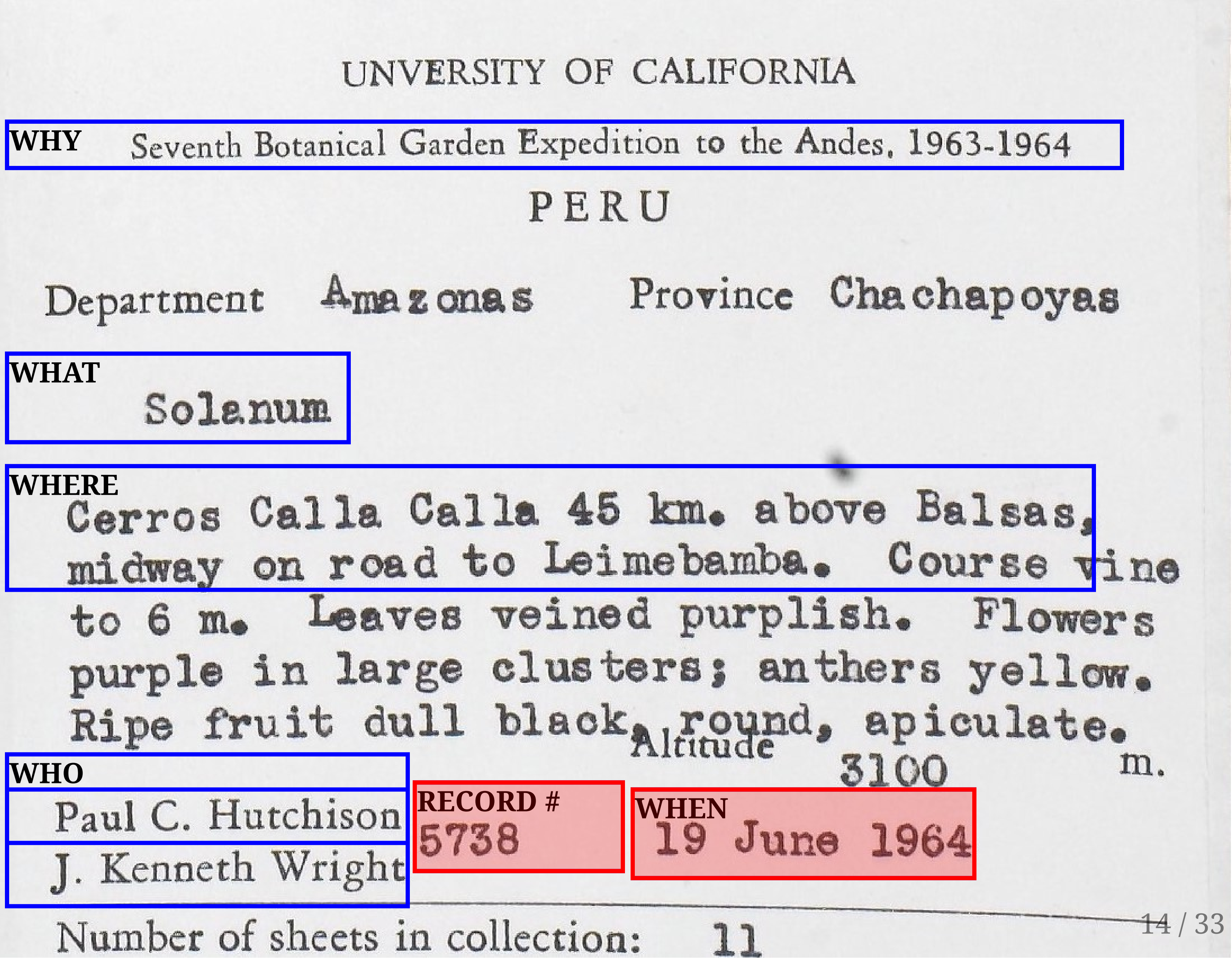
WHO Paul C. Hutchison

J. Kenneth Wright

RECORD # 5738

WHEN 19 June 1964

Number of sheets in collection: 11


WHY
WHAT
WHERE
WHO
RECORD #
WHEN

UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

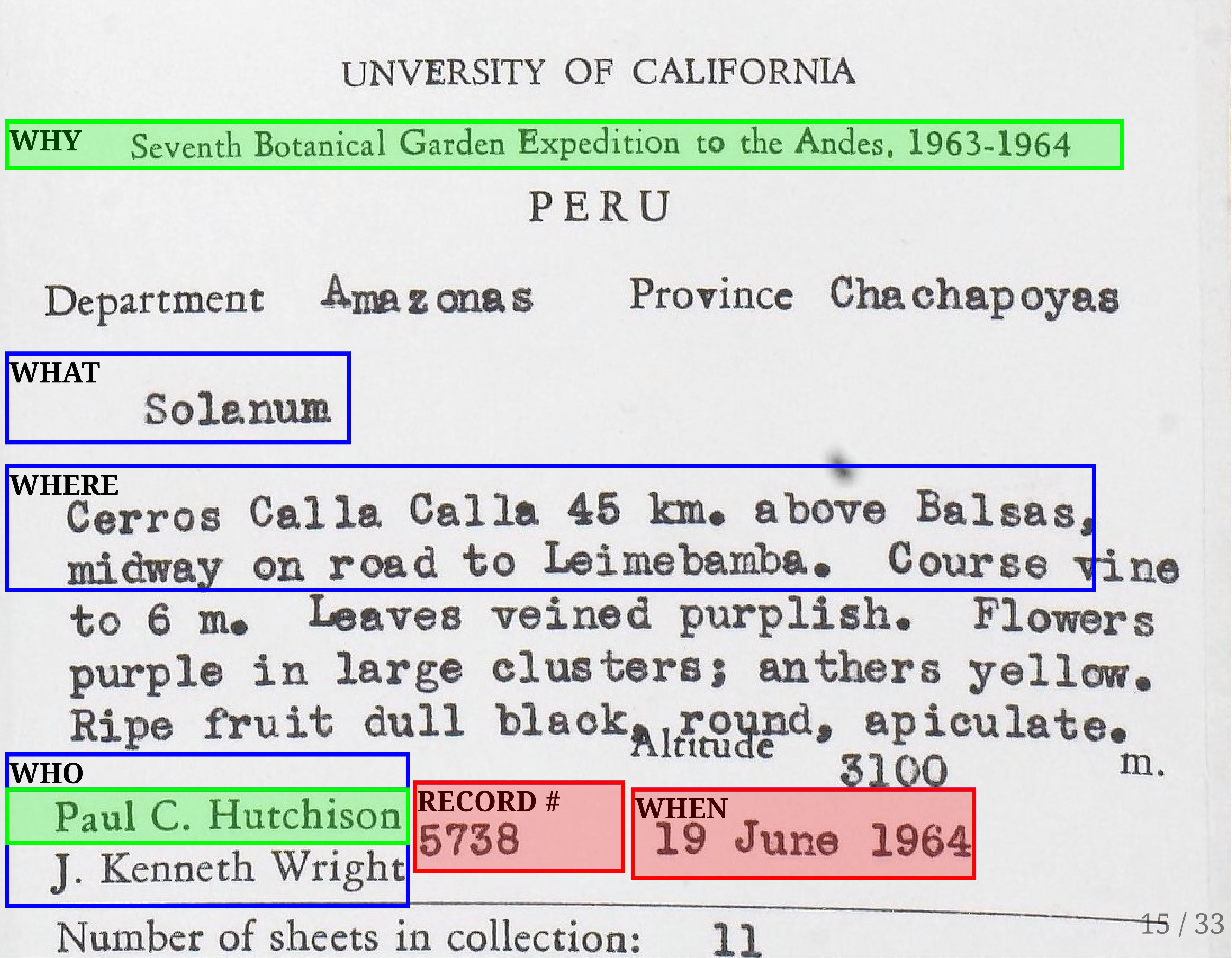
Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright

5738

19 June 1964

Number of sheets in collection: 11



UNVERSITY OF CALIFORNIA

WHY Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

WHAT

Solanum

WHERE

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

WHO

Paul C. Hutchison

J. Kenneth Wright

RECORD #

5738

WHEN

19 June 1964

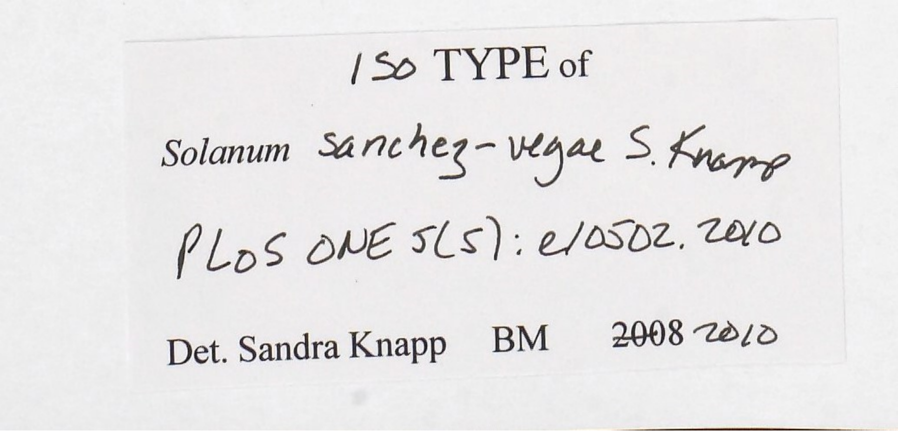
Number of sheets in collection: 11

Motivation: exploring species discovery

Species discovery on-going: 2000 new species described / year in higher plants.

Reviewing example specimen (via collection & research annotations):

- Early 1960s: collected from field
- 2010: recognised as a species new to science (and published)
- Specimen annotated, digital record flagged as a "type"



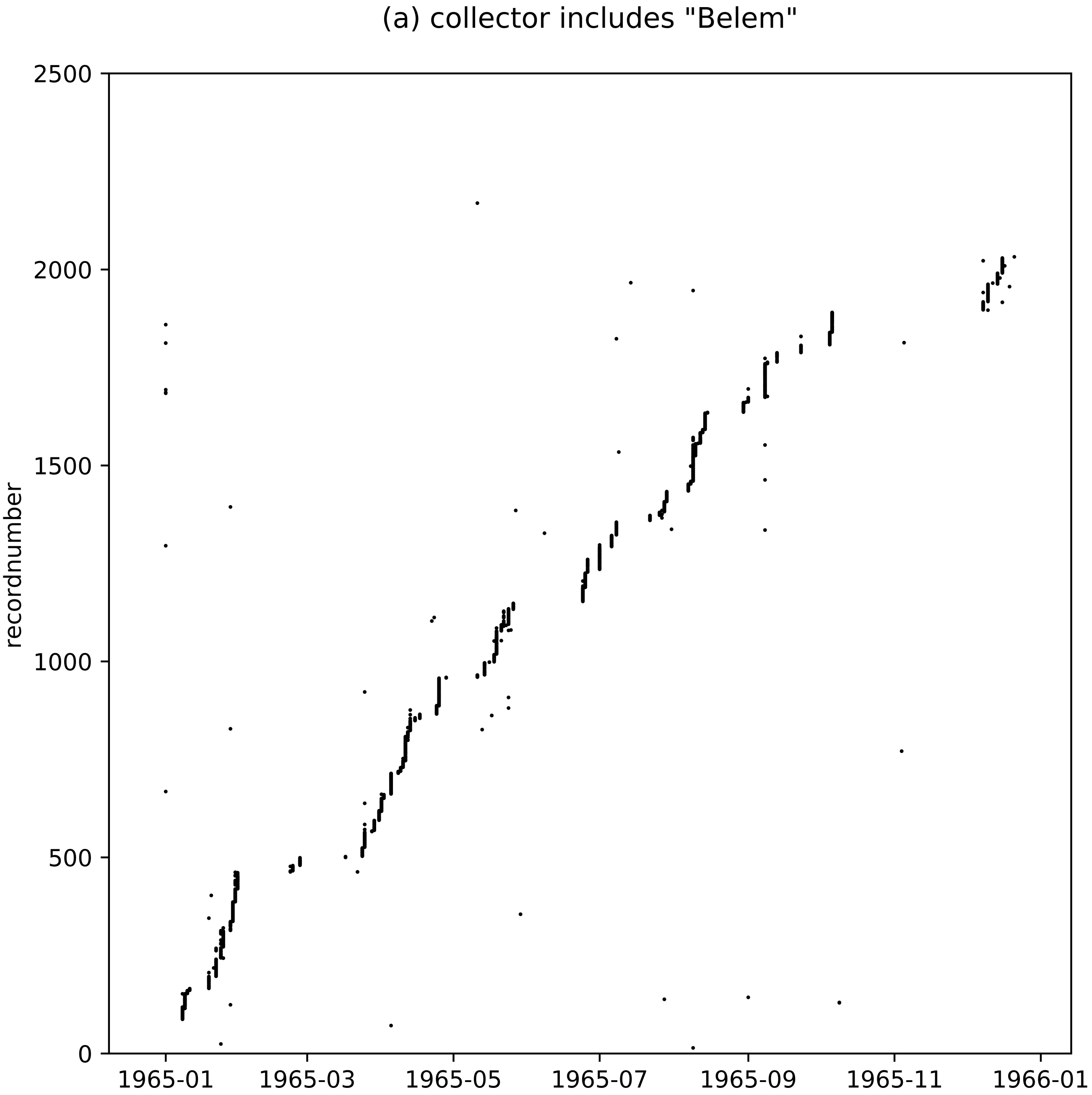
ISO TYPE of

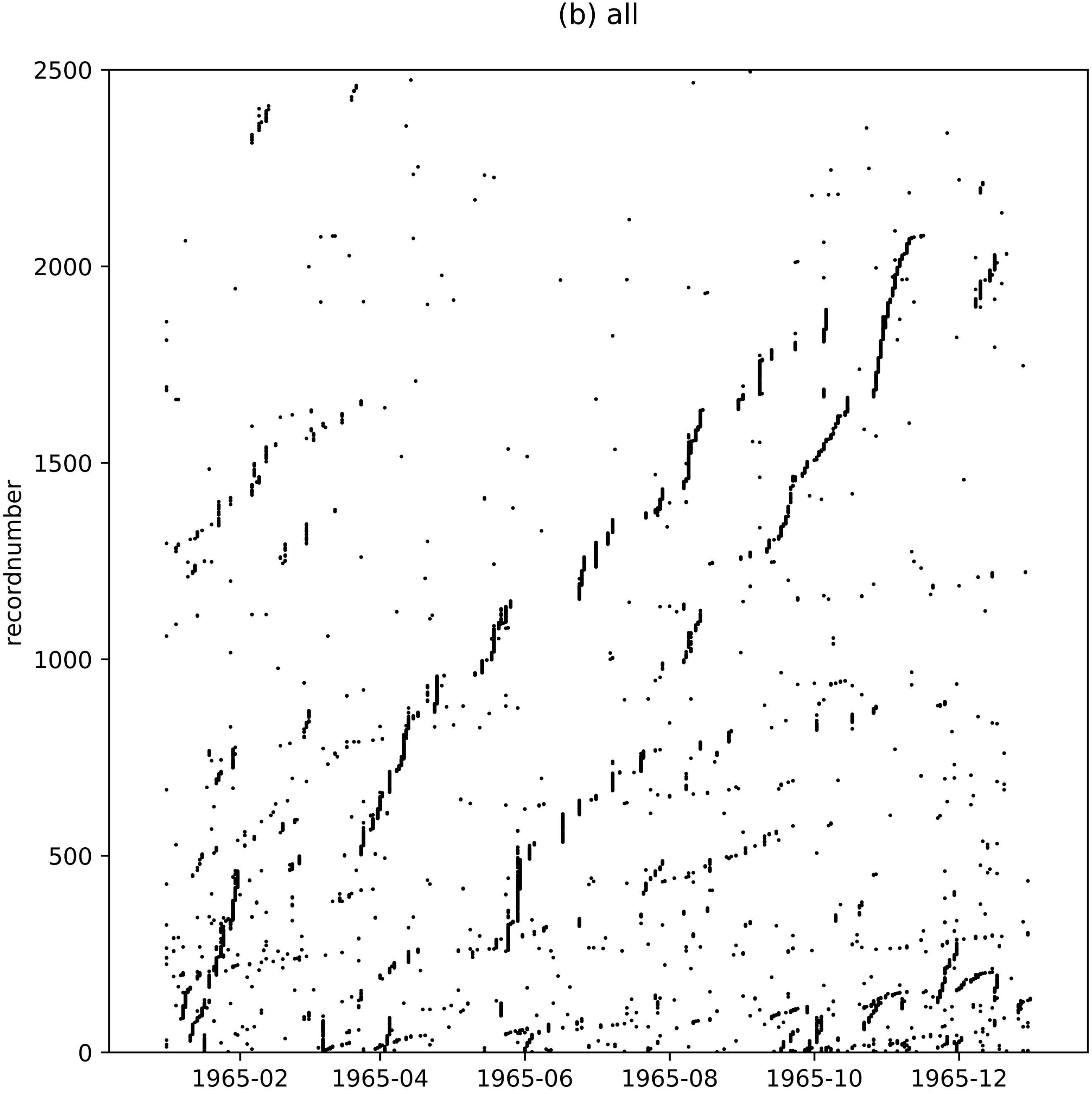
Solanum sanchez-vegae S. Knapp

PLoS ONE 5(5): e10502. 2010

Det. Sandra Knapp BM ~~2008~~ 2010

Can we direct effort to field collection localities / institutional collections that will yield new species?


(a) collector includes "Belem"
recordnumber
2500
2000
1500
1000
500
0
1965-01
1965-03
1965-05
1965-07
1965-09
1965-11
1966-01


(b) all
recordnumber
2500
2000
1500
1000
500
0
1965-02
1965-04
1965-06
1965-08
1965-10
1965-12

Data mining: preparation

Data

- c 3.5 million specimens from a single political country (Brazil)
- data downloaded from Global Biodiversity Information Facility (GBIF)

Feature definition

- Numeric feature-set:
    - eventdate - days since 1970-01-01
    - recordnumber - sequential and unique in context of particular primary collector
- Collector name transcription, e.g. Gert Hatcshbach
- Lexical feature-set:
    - First initial
    - First upper-case of surname
    - First lower-case of surname
    - Last lower-case of surname
    - e.g. Gert Hatcshbach -> **G**ert **Ha**tcshbac**h** -> G, H, a, h

Data mining: process (1/4)

Cluster

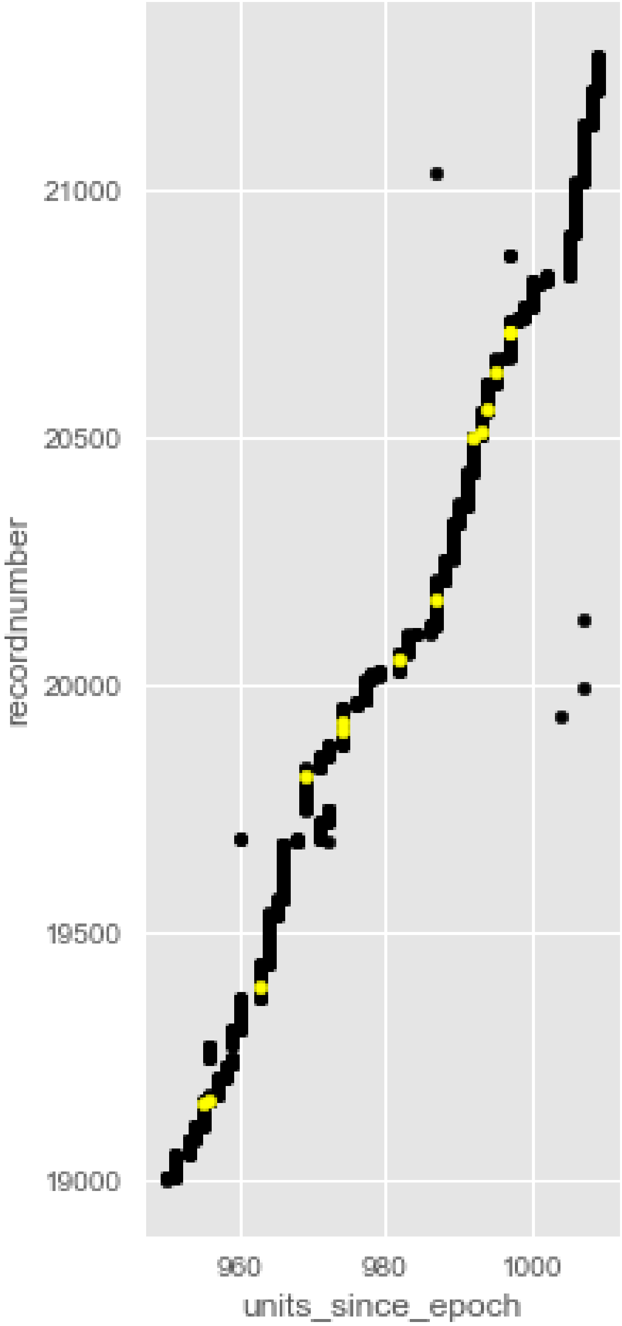
- DBSCAN: selected as we want to detect elongated clusters
  - featuresets: lexical & numeric
  - episilon: 300
  - min_samples: 2
- Expert analysis
  - variable transcriptions -> multiple logical collectors assigned to a single cluster
- Computational post-processing
  - clusters pessimistically broken into subclusters, based on lexical examination of transcriptions

Data mining: process (2/4)

Cluster

Classify

- Expert analysis identified a common problem:
  - variation in transcription results in variation in lexical featureset -> logical collectors assigned to separate clusters
  - visualisation using scatter plot show this
- Classify:
  - train decision tree on numeric featureset, to predict cluster identifier
  - commonly confused classes candidates for joining
  - computationally assessed for lexical similarity
  - iterative process (join affects overlap calculation)


21000
20500
20000
19500
19000
recordnumber
960
980
1000
units_since_epoch

Data mining: process (3/4)

Cluster

Classify

Join

- Aim to gather all data to get a career grouping for a single logical collector
- Two stage process, clusters are joined if:
  - Most frequently occurring transcription is shared and all variant transcriptions agree
  - Clusters share external identifier in bibliographic author dataset

Data mining: process (4/4)

Cluster

Classify

Join

Detect collection trips

- For each collector's career, pass all samples into DBSCAN to detect collecting trips
- Create and apply a trip identifier to each "collecting trip" cluster

Application: grouping

Grouping

1. Baseline - grouped by transcribed primary collector name
2. Collector - grouped by data-mined collector entity
3. Trip - grouped by data-mined collecting trip entity

Application: feature definition

Grouping

Feature definition

- Temporal:
  - Start year
- Scale:
  - # specimens
  - Range of numbers allocated
- Rate:
  - Slope of line of best fit
  - Correlation score
- Character:
  - Specialist (T/F)
  - Generalist (T/F)
- Experience:
  - # previous collections
- Class: species discovery value:
  - Does the grouping include material used as a type (T/F)

Application: classifier construction

Grouping

Feature definition

Classifier construction

- Decision tree classifier used to predict species discovery value.
- Datasets downsampled to balance class variable.
- 10-fold stratified cross-validation.
- Feature selection.

Results: data mining

Raw data:

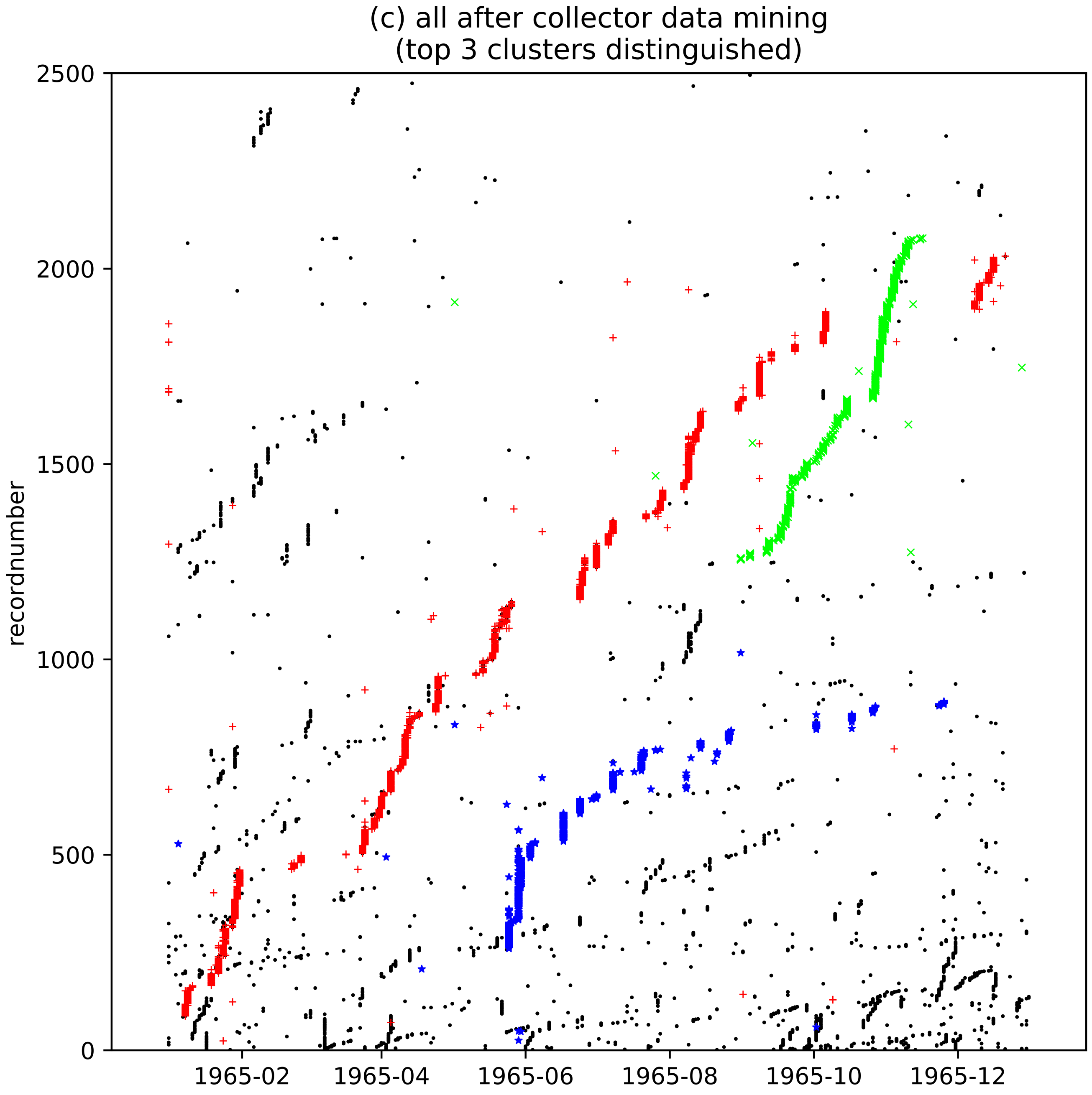
- 131582 unique collector team transcriptions
- 41511 unique primary collector name transcriptions

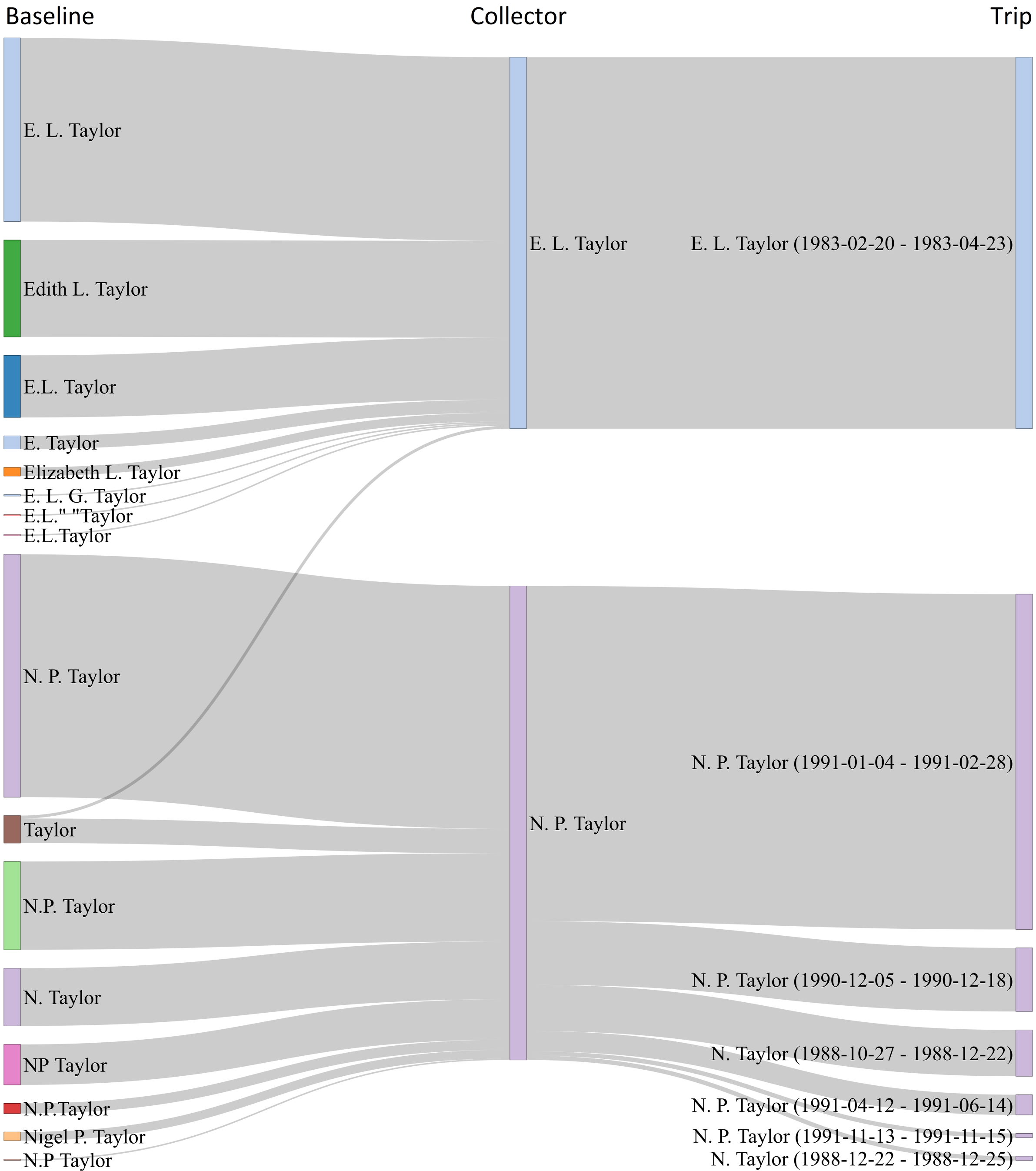
Data mining process:

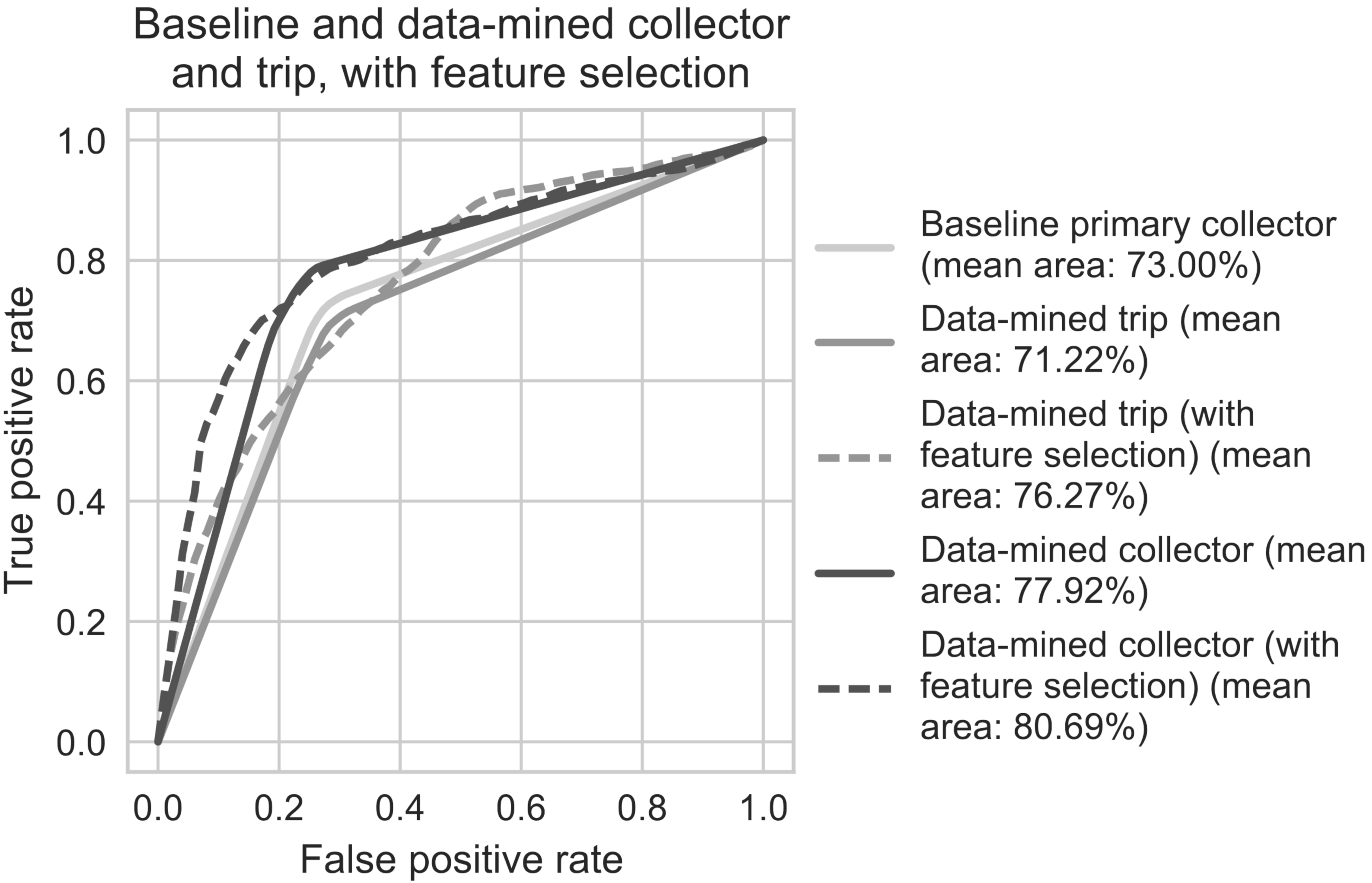
- Step 1: DBSCAN identified 42096 clusters; lexically post-processed to 51192 clusters
- Step 2: Resolved via decision-tree classifier to 44768 clusters
- Step 3: Joined to 19706 clusters representing collector entities
- Step 4: 79012 different collecting trips were identified

Species discovery value:

- 1127 (5.7%) of collectors and 3412 (4.3%) of trips collected specimens later labelled as type specimens.


(c) all after collector data mining
(top 3 clusters distinguished)
recordnumber
0
500
1000
1500
2000
2500
1965-02
1965-04
1965-06
1965-08
1965-10
1965-12


Baseline
Collector
Trip
E. L. Taylor
Edith L. Taylor
E.L. Taylor
E. Taylor
Elizabeth L. Taylor
E. L. G. Taylor
E.L." "Taylor
E.L.Taylor
N. P. Taylor
Taylor
N.P. Taylor
N. Taylor
NP Taylor
N.P.Taylor
Nigel P. Taylor
N.P Taylor
E. L. Taylor
N. P. Taylor
E. L. Taylor (1983-02-20 - 1983-04-23)
N. P. Taylor (1991-01-04 - 1991-02-28)
N. P. Taylor (1990-12-05 - 1990-12-18)
N. Taylor (1988-10-27 - 1988-12-22)
N. P. Taylor (1991-04-12 - 1991-06-14)
N. P. Taylor (1991-11-13 - 1991-11-15)
N. Taylor (1988-12-22 - 1988-12-25)


Baseline and data-mined collector and trip, with feature selection
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Baseline primary collector (mean area: 73.00%)
Data-mined trip (mean area: 71.22%)
Data-mined trip (with feature selection) (mean area: 76.27%)
Data-mined collector (mean area: 77.92%)
Data-mined collector (with feature selection) (mean area: 80.69%)

Conclusions

- Specimens visible end point of a hidden collecting process
- Machine learning techniques help to uncover the hidden processes
- Data mining results reshape the data, build models - steps towards understanding species discovery
- Techniques also have practical applications - efficiencies in data mobilisation

Further information

Identifying novel features from specimen data for the prediction of valuable collection trips

Nicky Nicolson[1,2], Allan Tucker[2]

1. Biodiversity Informatics & Spatial Analysis, Royal Botanic Gardens, Kew (UK). 2. Department of Computer Science, Brunel University London (UK).

`n.nicolson@kew.org / @nickynicolson`

https://nickynicolson.github.io/ida-2017-specimen-features/presentation.html

Acknowledgements

Data providers for sharing their specimen data

Reviewers for valuable comments