Search Engine Topics

Will Poynter • CLOSER

CLOSER:

Maximise the use, value and impact of the UK's longitudinal studies, both at home and abroad

CLOSER Discovery

- Launched January 2017
- DDI-3.2
- Colectica
- 9 longitudinal studies

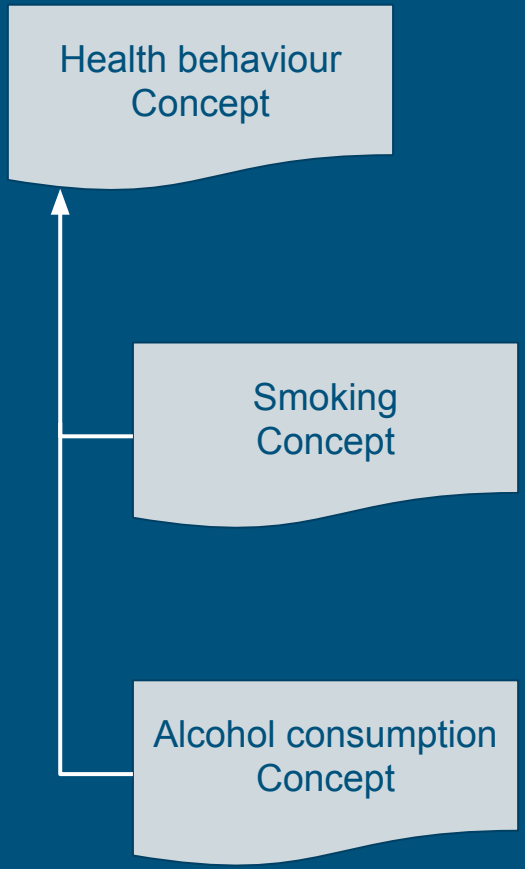
Searching

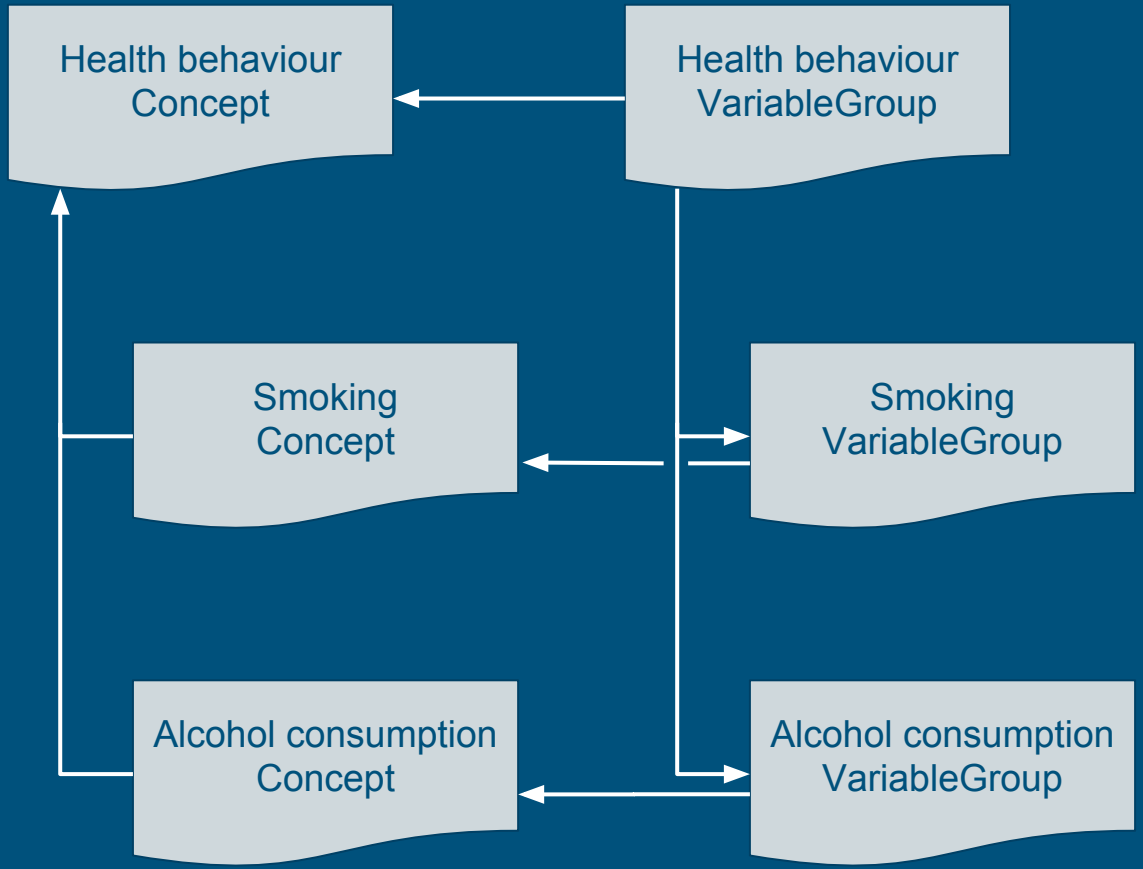
- [ ] Health behaviour (Health and lifestyle) (7,443)
  - [ ] Diet and nutrition (3,472)
  - [ ] Physical activity (888)
  - [ ] Sleep (380)
  - [ ] Smoking (445)
  - [x] Alcohol consumption (764)
  - [ ] Substance abuse (665)
  - [ ] Risk taking (88)

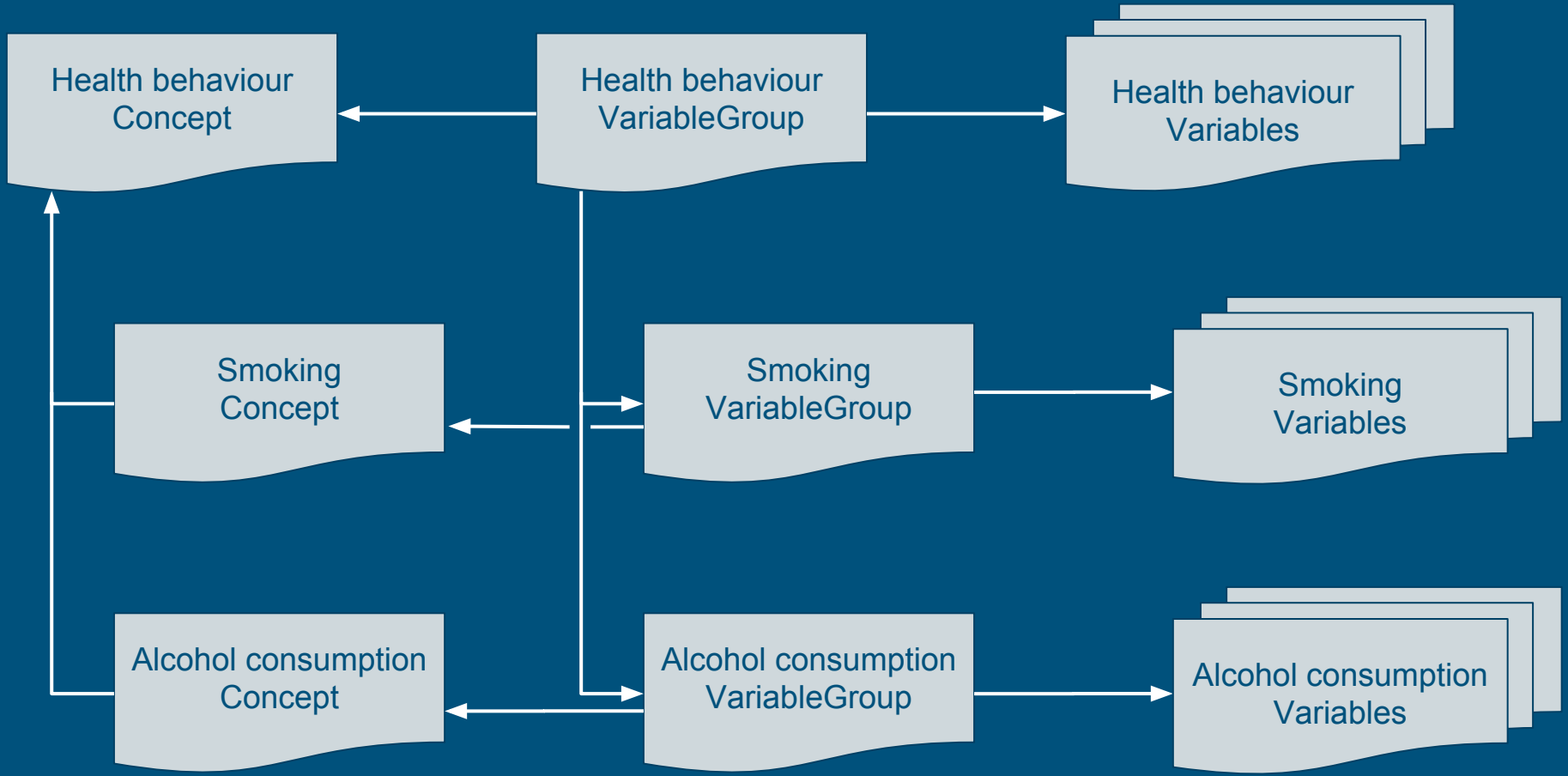
Filtering

DDI Concepts

Health behaviour Concept


Health behaviour Concept
Smoking Concept
Alcohol consumption Concept


Health behaviour Concept
Health behaviour VariableGroup
Smoking Concept
Smoking VariableGroup
Alcohol consumption Concept
Alcohol consumption VariableGroup


Health behaviour Concept
Health behaviour VariableGroup
Health behaviour Variables
Smoking Concept
Smoking VariableGroup
Smoking Variables
Alcohol consumption Concept
Alcohol consumption VariableGroup
Alcohol consumption Variables

Scale

118

CLOSER Topics

51,000

Questions Constructs

98,000

Variables

How?

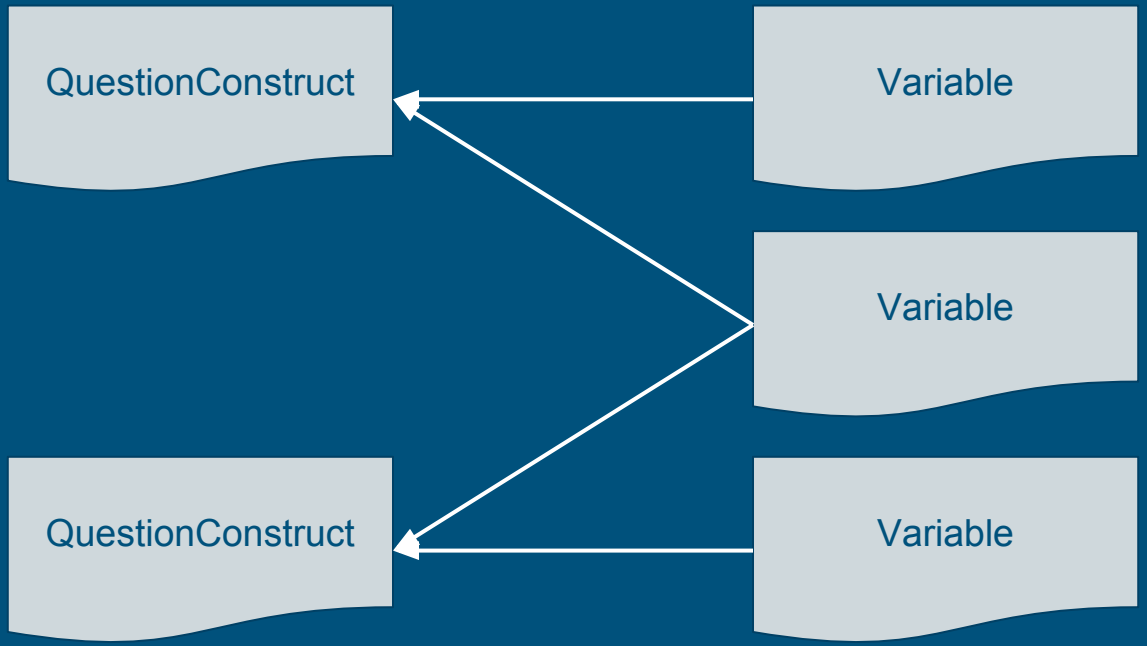
Strands

- Collection of QuestionConstructs and Variables
- Joined via **SourceQuestionReference**
- Must have the same topic

Strand

Must have the same topic

Strand

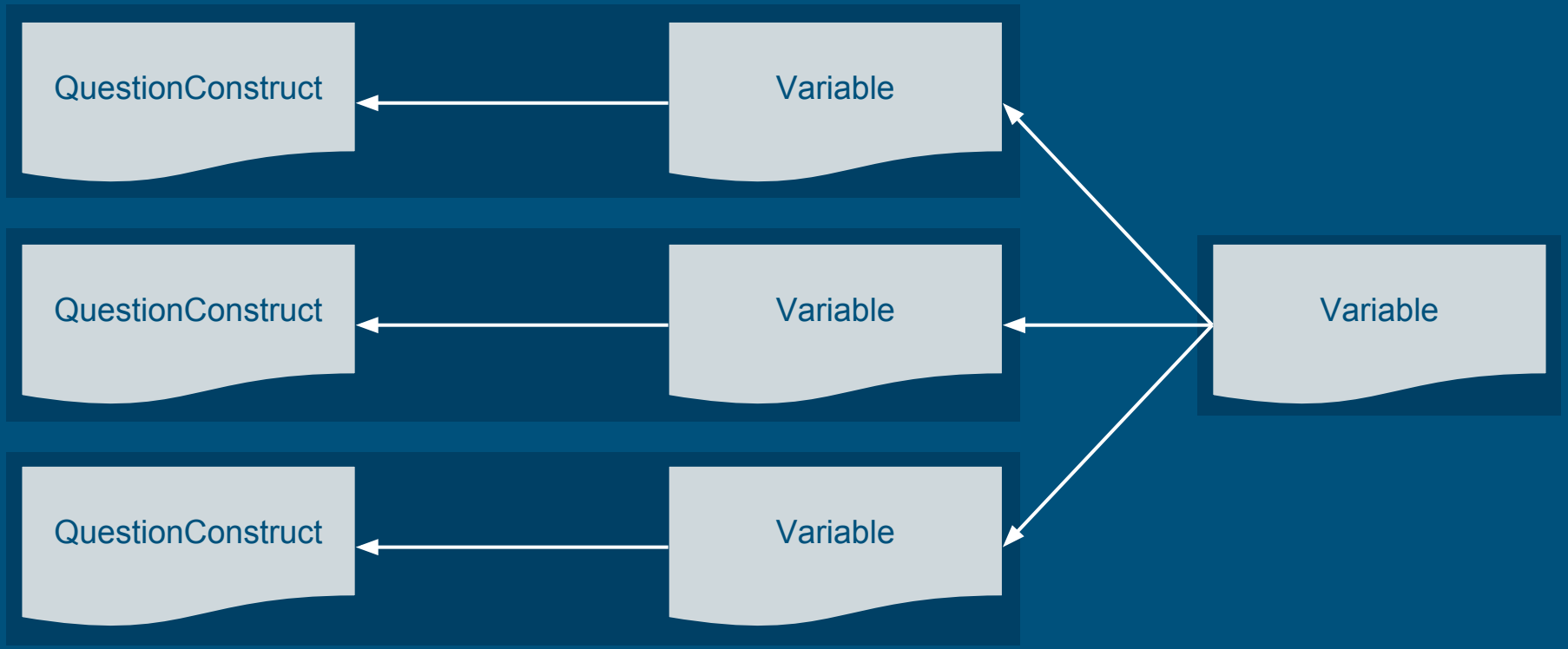


Must have the same topic

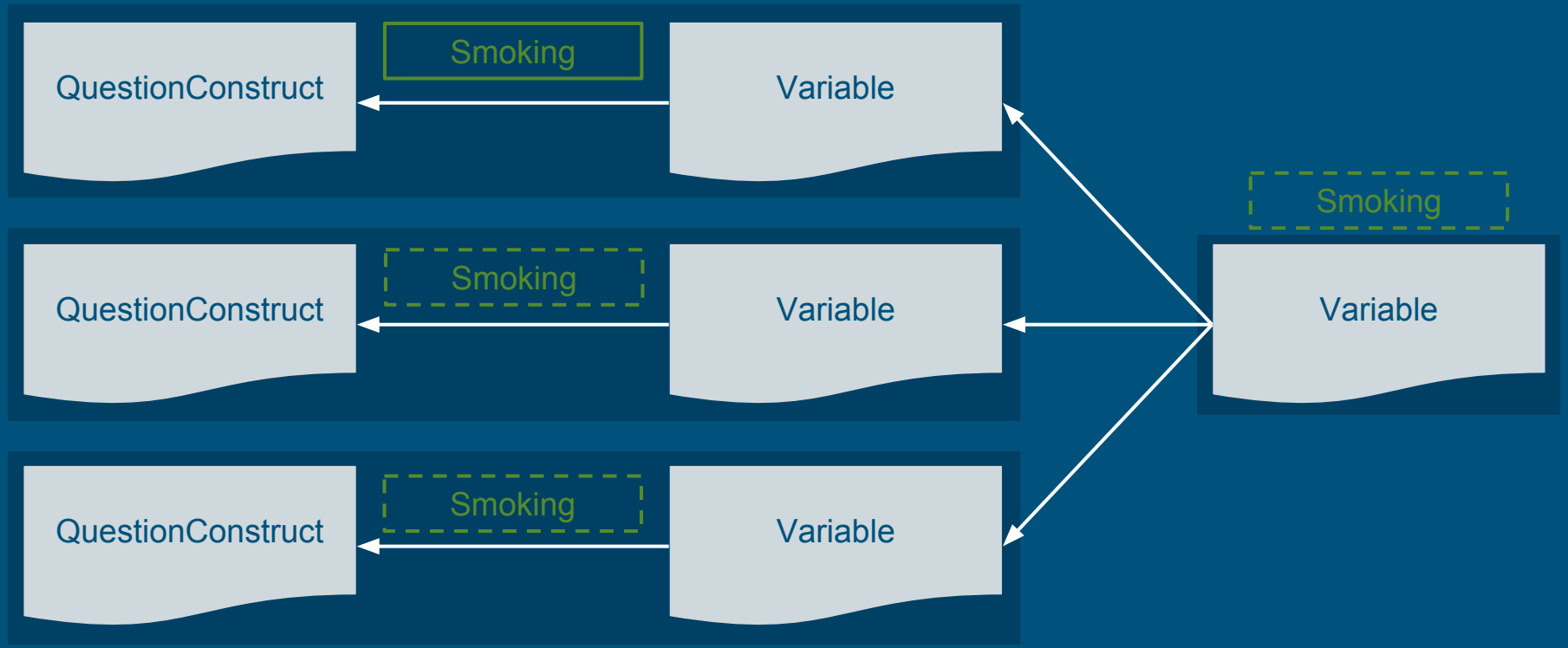
Clusters

- Collection of Strands
- Joined via **SourceVariableReference**
- Suggests the same topic

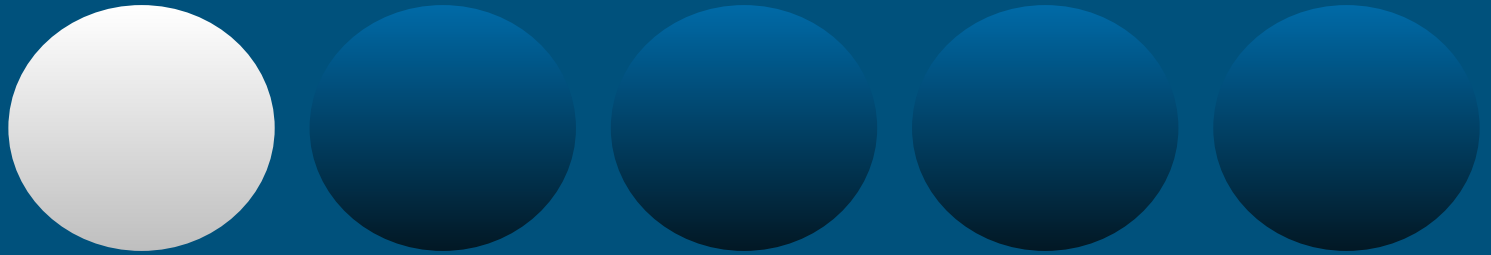
Cluster


QuestionConstruct
Variable
QuestionConstruct
Variable
QuestionConstruct
Variable
Variable

Cluster


QuestionConstruct
Smoking
Variable
QuestionConstruct
Smoking
Variable
QuestionConstruct
Smoking
Variable
Smoking
Variable

Savings



30,000

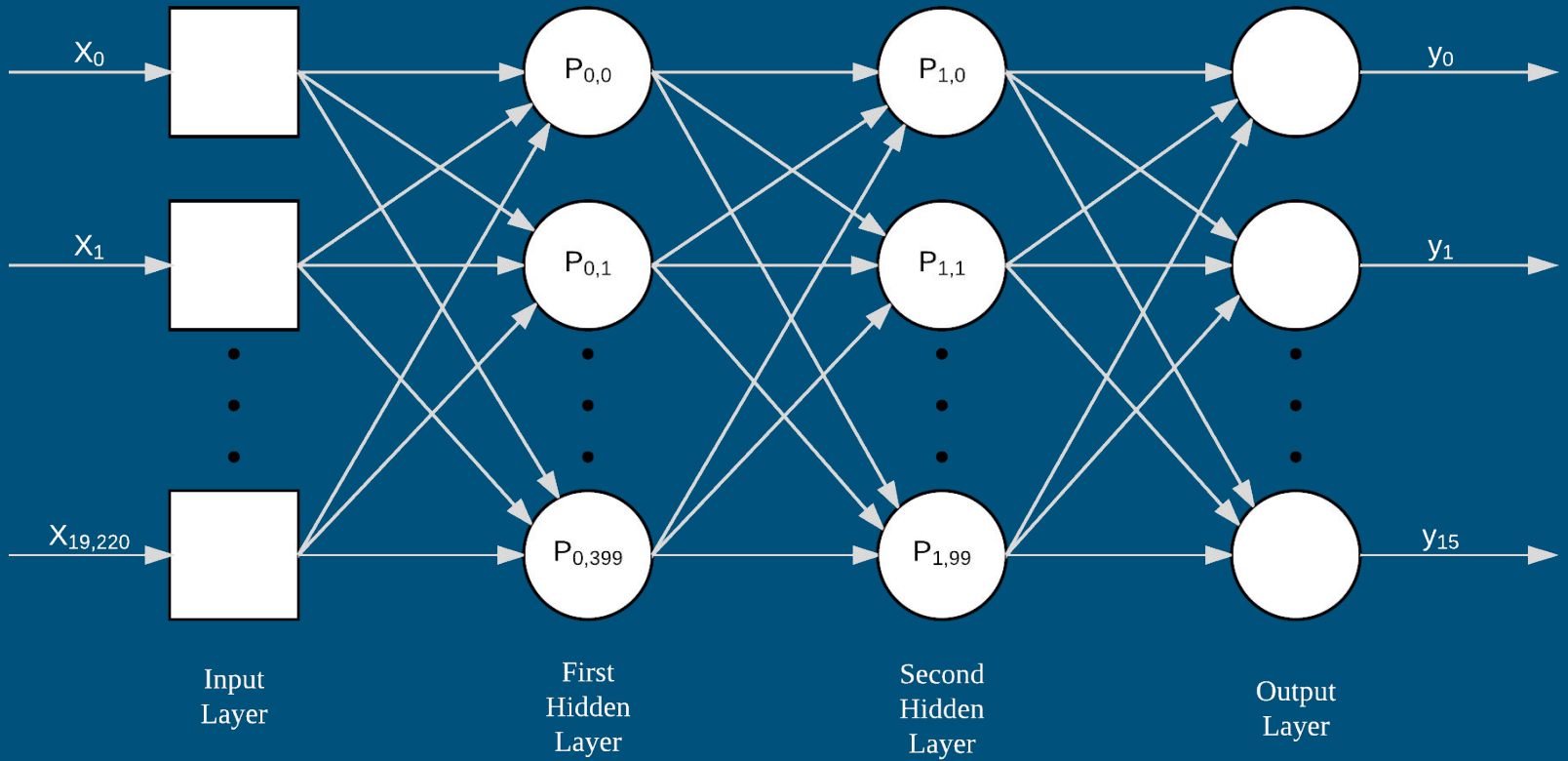
30,000

150,000


Machine learning

Model

- Multilayer Perceptron (MLP)
- Trained using variable labels
- Vocabulary of 19,221 words
    - Weighted using term-frequency inverse-document-frequency
- Predicting level-1 CLOSER Topics


X_0
X_1
$X_{19,220}$
$P_{0,0}$
$P_{0,1}$
$P_{0,399}$
$P_{1,0}$
$P_{1,1}$
$P_{1,99}$
y_0
y_1
y_{15}
Input Layer
First Hidden Layer
Second Hidden Layer
Output Layer

Entire Dataset

Study	No. of Variables Used
ALSPAC	46,787
BCS70	6,813
HCS	1,026
MCS	898
NCDS	4,431
NSHD	1,445
SWS	2,511
USoc	0
Total	**63,911**

Training

- 50,000 labels
- Batches of 200
- Single core i7-5775R Intel & 5GB RAM
- 23 iterations
- Approximately 7 hours

Evaluation

- 13,911 labels
- 93% predicted ‘correctly’
- Accuracy greatly depends on topic
- Unweighted random allocation accuracy 6.3%

Evaluation

- Most accurate: Omics
- Least accurate: Expectations, attitudes and beliefs

Future work

- Apply model to entire Strands
- Train for level-1 and level-2 topics
- Use study as a weighted input
- Integrate with current metadata enhancement tools aaS

Thank you

Any questions?

Special thanks you to

- ESRC
- MRC
- JetBrains

Contact me

- @willpoynter
- w.poynter@ucl.ac.uk

Contact CLOSER

- @CLOSER_UK
- closer@ucl.ac.uk

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MRC Medical Research Council

closer Cohort & Longitudinal Studies Enhancement Resources