Machine Learning
(theory and practice)

Charles Sutton
Introduction to Research in Data Science
University of Edinburgh

Microsoft Research

THE UNIVERSITY of EDINBURGH informatics

EPSRC
Engineering and Physical Sciences Research Council

New methodology

New applications

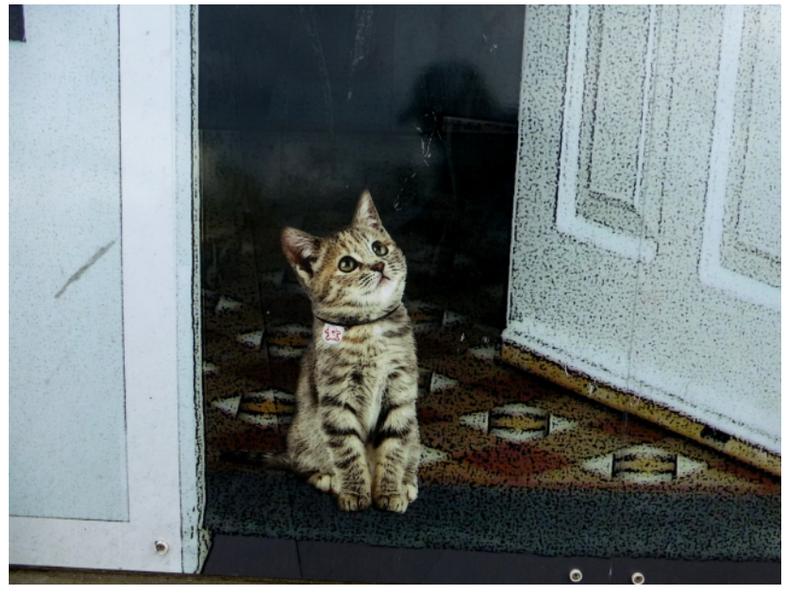
- New model architectures
- Inference algorithms
  (e.g., high dimensional, streaming)
- Approximate learning methods

- Analyzing computer programs
- Data mining
- Exploratory data analysis
- Home energy demand
- Computer security

Interactive machine learning

Data analysts are like cats.

1. Want to explore their data
2. Don’t know what they want.



Interactive machine learning for analysts

Whose information need is not explicit

Whose domain knowledge is difficult to encode

Allow analysts to explore intermediate results

… not just for dummies!

Per clustering accept / reject

13	M	Red
17	M	Red
25	F	Blue
23	M	Yellow
19	F	Yellow
35	F	Yellow
57	M	Red
60	M	Red
61	F	Blue
31	F	Red

Data

clustering

Clustering:
Partition of data

13	M	Red
17	M	Red
25	F	Blue
23	M	Yellow
19	F	Yellow
35	F	Yellow
57	M	Red
60	M	Red
61	F	Blue
31	F	Red

Clustered data

Reject

Accept

Reject

User feedback

TINDER

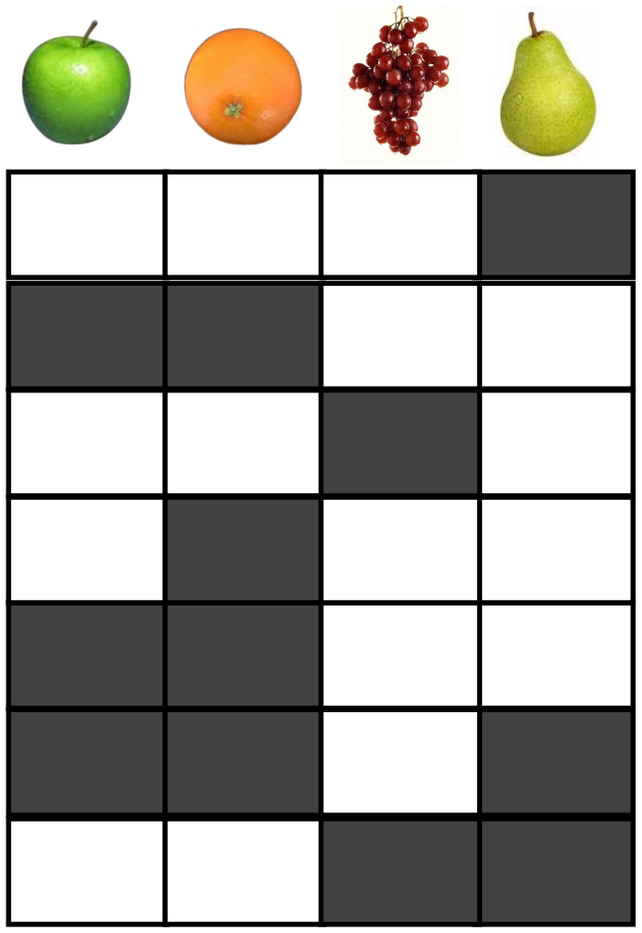
Technique for INteractive Data Exploration via Rejection

13	M	Red
17	M	Red
25	F	Blue
23	M	Yellow
19	F	Yellow
35	F	Yellow
57	M	Red
60	M	Red
61	F	Blue
31	F	Red

Re-clustered data

[Srivastava, Zou, Sutton, ICML W/S 2016]

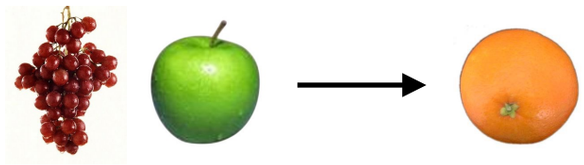
Association Rules



Database of transactions

Association rule mining

Find set of all rules



that have

$$\text{Prob}\left(\quad \middle| \quad \right) \geq \alpha$$

$$\text{Count}\left\{ \quad \right\} \geq M$$

Frequent itemsets

$$\text{Count}\left\{ \quad \right\} \geq M$$

Why? Exploratory data analysis

Probabilistic Itemset Mining

Generative Model

To sample a transaction,

1. For each itemset, sample

$$z_S \sim \mathbf{Bernoulli}(\pi_S).$$

2. Deterministically set

$$X = \bigcup_{z_s=1} S.$$

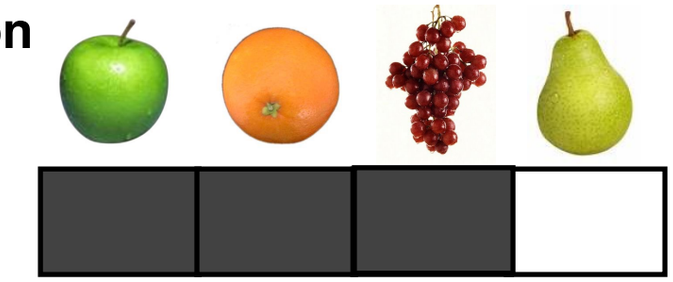
Patterns

0.0001

0.02

0.5

Transaction



Inference

View pattern finding as set cover

Alternate set cover and parameter inference (structural EM)

Itemsets: E-step is submodular set cover

Sequences: Interleaving model for patterns with gaps

[Fowkes and Sutton, KDD 2016, PKDD 2016]

API Call Patterns: “Big Code”

Twitter4j Java Library

ISM Variant *[Fowkes & Sutton, FSE ‘16]*	**MAPO** *[Zhong et al, ‘09]*	**UPMiner** *[Wang et al, ‘13]*
`TwitterFactory.<init>` `TwitterFactory.getInstance`	`TwitterFactory.<init>` `TwitterFactory.getInstance`	`TwitterFactory.<init>` `TwitterFactory.getInstance`

`TwitterFactory.<init>`
`TwitterFactory.getInstance`
`Twitter.setOAuthConsumer`
`Twitter.setOAuthAccessToken`
	`Status.getUser`
	`Status.getText`
		`TwitterFactory.getInstance`
		`Twitter.setOAuthConsumer`

`Status.getUser`
`Status.getText`
	`ConfigurationBuilder.<init>`
	`ConfigurationBuilder.build`
		`TwitterFactory.<init>`
		`TwitterFactory.getInstance`
		`Twitter.setOAuthConsumer`

`AccessToken.getToken` `AccessToken.getTokenSecret`	`ConfigurationBuilder.<init>` `TwitterFactory.<init>`	`Status.getUser` `Status.getText`

`ConfigurationBuilder.<init>`
`ConfigurationBuilder.build`
`TwitterFactory.<init>`
`TwitterFactory.getInstance`
	`ConfigurationBuilder.<init>`
	`ConfigurationBuilder.setOAuthCon`
	`sumerKey`
		`Twitter.setOAuthConsumer`
		`Twitter.setOAuthAccessToken`

: *two main types of twitter initialization call*

[Fowkes and Sutton, FSE 2016]

Syntactic Idioms in Code

```
...
if (c != null) {
  try {
   if (c.moveToFirst()) {
     number = c.getString(
              c.getColumnIndex(
               phoneColumn));
   }
  } finally {
   c.close();
  }
}
...
```

(a)

```
try {
  if ($(Cursor).moveToFirst()) {
     $BODY$
  }
} finally {
   $(Cursor).close();
}
```

```
IfStatement
  expression:
    c!=null
  then:Block
    TryStatement
      body:IfStatement
        expr:MethodInvocation
          expr:var%android.database.Cursor%
            name:c
          name:moveToFirst
        then:Block
          number = c.getString(c.getColumnIndex(phoneColumn));
      finally:Block
        ExpressionStatement
          MethodInvocation
            expr:var%android.database.Cursor%
              name:c
            name:close
```

E

Allamanis and Sutton, FSE 2014

Example Idioms

From: Nonparametric Bayesian Tree Substitution Grammar

[Post and Gildea, 2009; Cohn et al, 2010]

```
channel=connection.
  createChannel();
```

(a)

```
Elements $name=$(Element).
  select($StringLit);
```

(b)

```
Transaction tx=ConnectionFactory.
  getDatabase().beginTx();
```

(c)

```
catch (Exception e){
  $(Transaction).failure();
}
```

(d)

```
SearchSourceBuilder builder=
  getQueryTranslator().build(
    $(ContentIndexQuery));
```

(e)

```
LocationManager $name =
  (LocationManager)getSystemService(
    Context.LOCATION_SERVICE);
```

(f)

```
Location.distanceBetween(
    $(Location).getLatitude(),
    $(Location).getLongitude(),
    $...);
```

(g)

```
try{
  $BODY$
}finally{
  $(RevWalk).release();
}
```

(h)

```
try{
  Node $name=$methodInvoc();
  $BODY$
}finally{
  $(Transaction).finish();
}
```

(i)

```
ConnectionFactory factory =
  new ConnectionFactory();
$methodInvoc();
Connection connection =
  factory.newConnection();
```

(j)

```
while ($(ModelNode) != null){
  if ($(ModelNode) == limit)
    break;
  $ifstatement
  $(ModelNode)=$(ModelNode)
      .getParentModelNode();
}
```

(k)

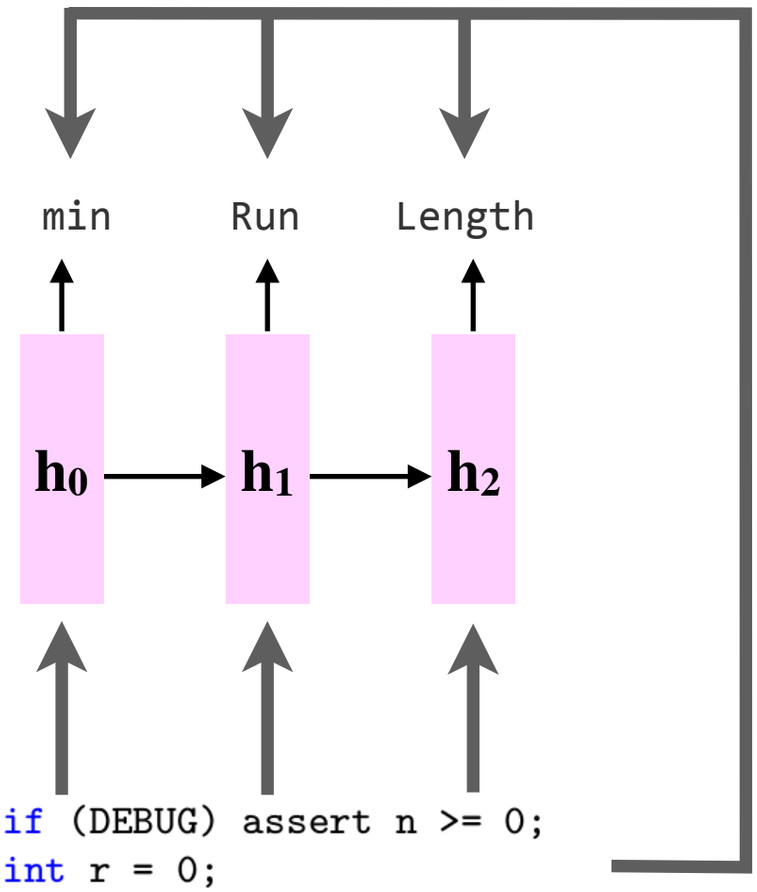
```
Document doc=Jsoup.connect(URL).
    userAgent("Mozilla").
    header("Accept","text/html").
    get();
```

(l)

Allamanis and Sutton, FSE 2014

Predicting Names of Methods

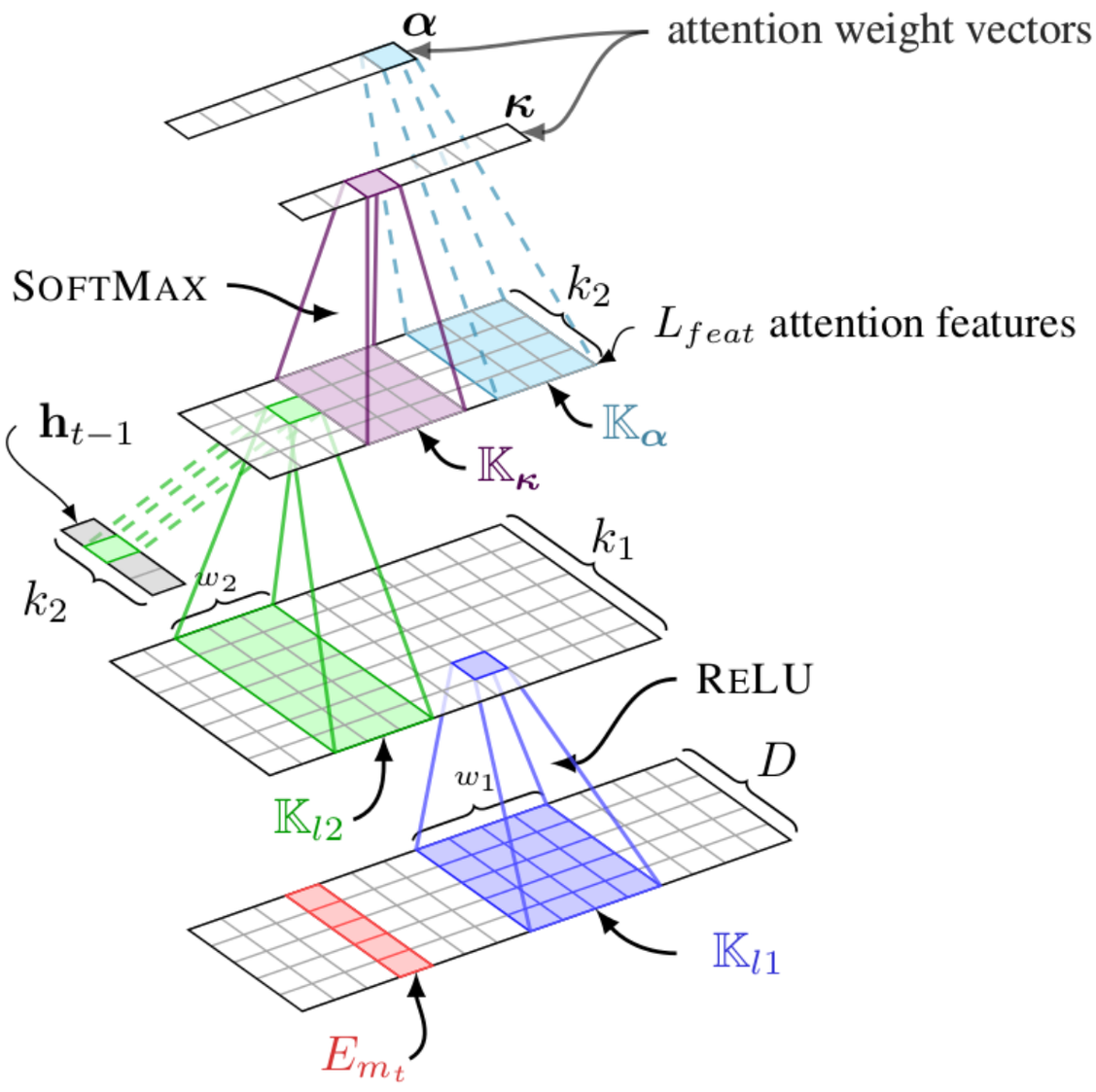
Name



Code

```
if (DEBUG) assert n >= 0;
int r = 0;
while (n >= MIN_MERGE) {
  r |= (n & 1);
  n >>= 1;
}
return n + r;
```

RNN for generating summary



convolutional attention mechanism

[Allamanis, Peng, and Sutton, ICML 2016]

- ***Machine learning for software engineering***
  - ML / NLP for programming languages
  - Combining program analysis with probabilistic machine learning
  - Find patterns in program executions: debugging
- ***Machine learning for data science***
  - Deep learning: Combining neural networks with prior knowledge
    - “interpretability bias”
  - Learning how to clean data
  - Interactive machine learning
  - Tools for monitoring models over time
  - Unsupervised and weakly supervised learning
- ***Deep learning: Unsupervised, structured, transfer learning***
  - ML for computer security, NLP, sustainable energy…

CUP, Wed and Fri 4pm

https://wiki.inf.ed.ac.uk/ANC/CharlesUncertainPeople