How the Statistical Learning Supports the Technical Service

A Brief Discussion

Background: some definitions

Technical Service

- Subjects: the most specific word or group of words that captures the essence of the subject or one of the subjects of a book or other library material
- Controlled vocabularies: Library of Congress Subject Headings (**LCSH**), Medical Subject Headings (MeSH), etc.
- MAchine-Readable Cataloging (MARC) & other metadata formats

Statistical Learning

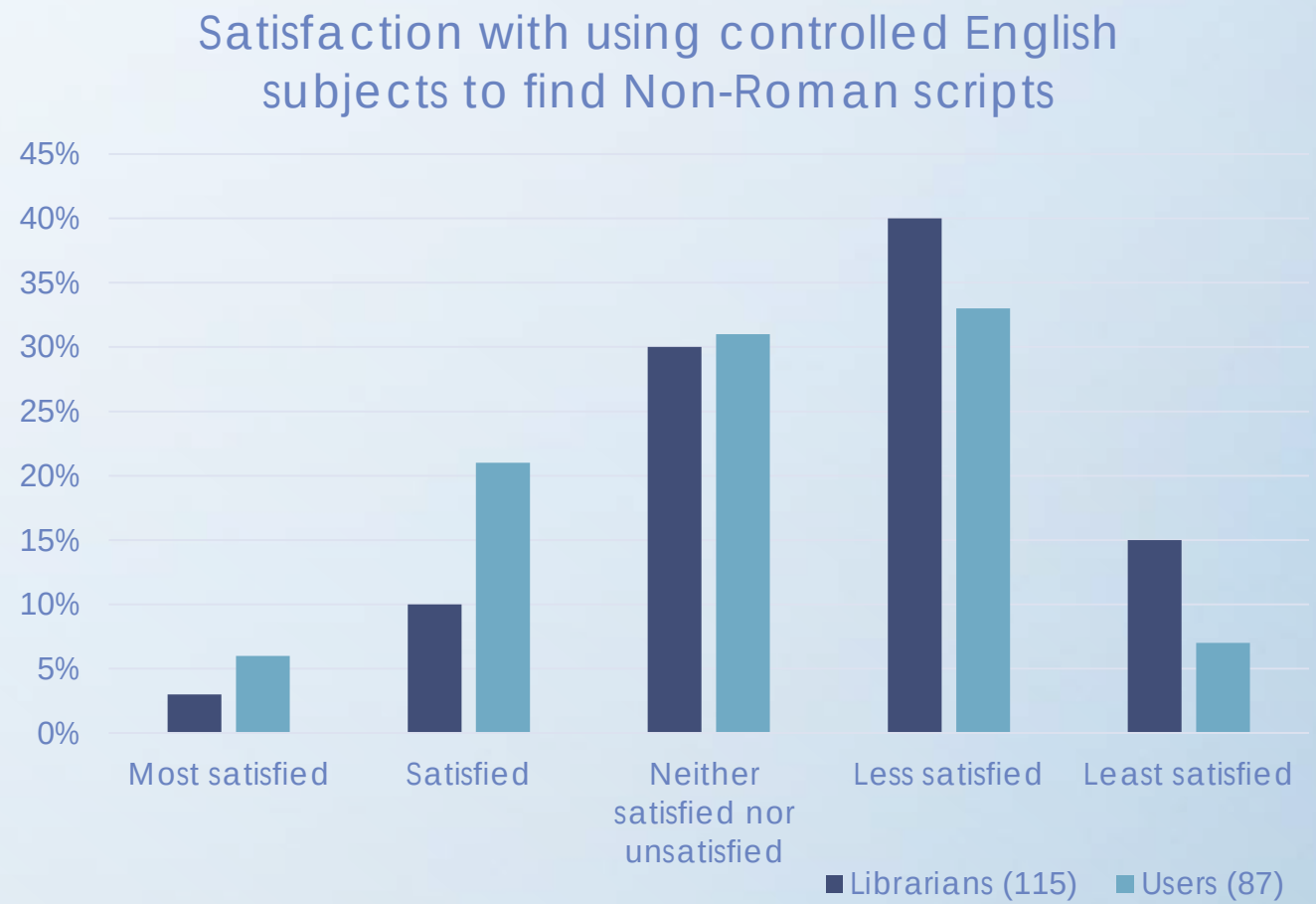
- Objects: incl. Natural Language Processing (**NLP**), Network Analysis, Ordinal data, etc.
- Modellings: incl. **Deep learning**, Classical ML, etc.
- Purposes: incl. **Classification**, **Prediction**, Identification, etc.
- Platforms: TensorFlow, Keras, Coffee, CNTK, PyTorch (mixed program with Python/R/C++)

Source: https://www.librarianshipstudies.com/2015/09/subject-heading.html

Technical Service

The biggest difficulty we have: artificiality

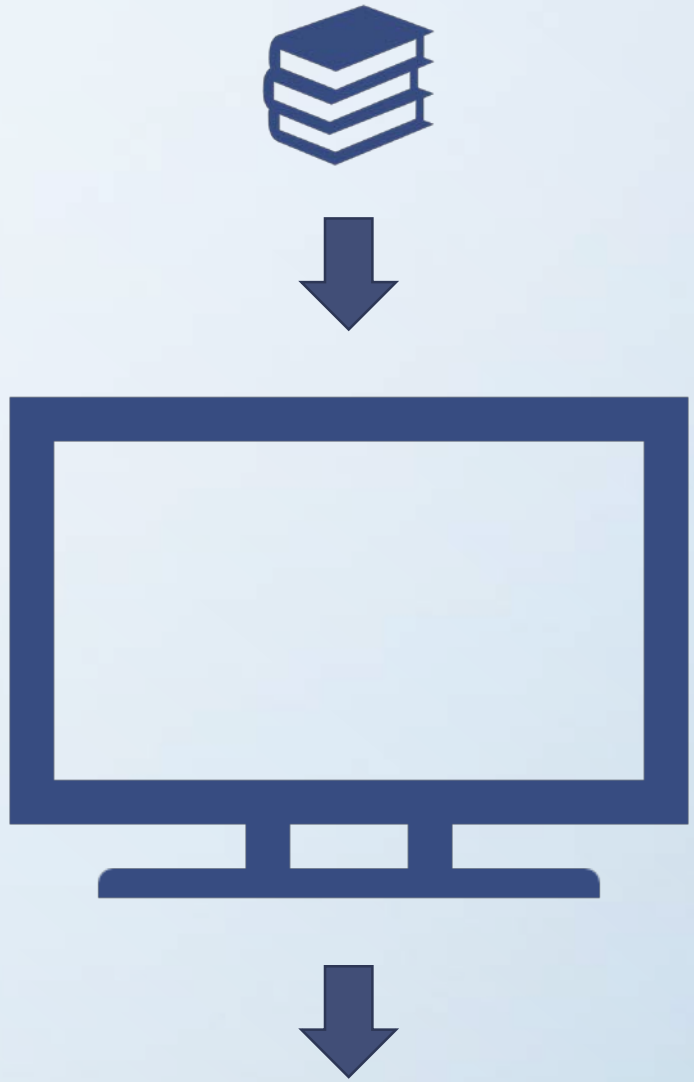
- “There are only two kinds of people who believe themselves able to read a MARC record without referring to a stack of manuals: a handful of our top cataloger and those on serious drugs. “ (Roy Tennant, 2002)



Source:
https://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/
El-Sherbini, M., & Chen, S. (2011). An assessment of the need to provide non-Roman subject access to the library online catalog. Cataloging & Classification Quarterly, 49(6), 457-483.

Statistical Learning

The greatest expectation we seek



Subjects

Statistical Learning

The greatest expectation we seek: deep learning

A Naïve Example on IMDB dataset

- Sub-dataset of movie reviews (25,000) from Internet Movie Database (IMDB), labeled by sentiment (positive/negative);
- They are split into two sets, namely 15,000 reviews for training and 10,000 reviews for testing;
- Which means building the model via 15,000 reviews and verifying/evaluating via the rest reviews to check whether they are applied in the correct classification label (positive/negative).

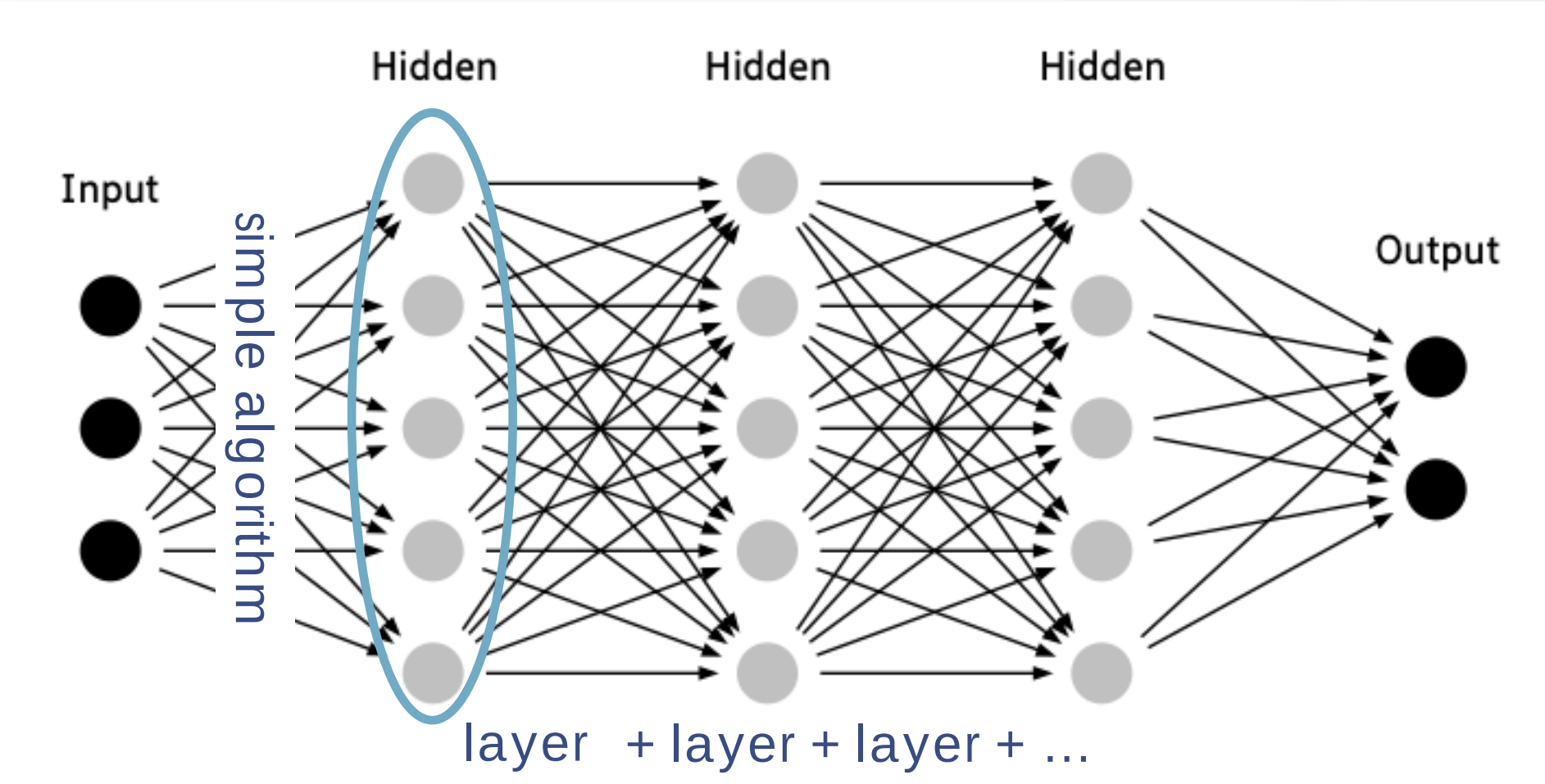
Source: Keras/Rstudio [https://keras.rstudio.com]

```
> decode_review(train_data[[1]])
[1] "<START> this film was just brilliant casting location scenery story
direction everyone's really suited the part they played and you could jus
t imagine being there robert <UNK> is an amazing actor and now the same b
eing director <UNK> father came from the same scottish island as myself s
o i loved the fact there was a real connection with this film the witty r
emarks throughout the film were great it was just brilliant so much that
i bought the film as soon as it was released for <UNK> and would recommen
d it to everyone to watch and the fly fishing was amazing really cried at
 the end it was so sad and you know what they say if you cry at a film it
 must have been good and this definitely was also <UNK> to the two little
 boy's that played the <UNK> of norman and paul they were just brilliant
children are often left out of the <UNK> list i think because the stars t
hat play them all grown up are such a big profile for the whole film but
these children are amazing and should be praised for what they have done
don't you think the whole story was so lovely because it was true and was
 someone's life after all that was shared with us all"
```

```
> train_data[[1]]
  [1]    1   14   22   16   43  530  973 1622 1385   65  458 4468   66 3941
 [15]    4  173   36  256    5   25  100   43  838  112   50  670    2    9
 [29]   35  480  284    5  150    4  172  112  167    2  336  385   39    4
 [43]  172 4536 1111   17  546   38   13  447    4  192   50   16    6  147
 [57] 2025   19   14   22    4 1920 4613  469    4   22   71   87   12   16
 [71]   43  530   38   76   15   13 1247    4   22   17  515   17   12   16
 [85]  626   18    2    5   62  386   12    8  316    8  106    5    4 2223
 [99] 5244   16  480   66 3785   33    4  130   12   16   38  619    5   25
[113]  124   51   36  135   48   25 1415   33    6   22   12  215   28   77
[127]   52    5   14  407   16   82    2    8    4  107  117 5952   15  256
[141]    4    2    7 3766    5  723   36   71   43  530  476   26  400  317
[155]   46    7    4    2 1029   13  104   88    4  381   15  297   98   32
[169] 2071   56   26  141    6  194 7486   18    4  226   22   21  134  476
[183]   26  480    5  144   30 5535   18   51   36   28  224   92   25  104
[197]    4  226   65   16   38 1334   88   12   16  283    5   16 4472  113
[211]  103   32   15   16 5345   19  178   32
```

- The ranking of frequency for each word
- "14" = "this" (the number = the word)
- Attitude: positive/negative
- Inference of reviews' attitudes on positive/negative

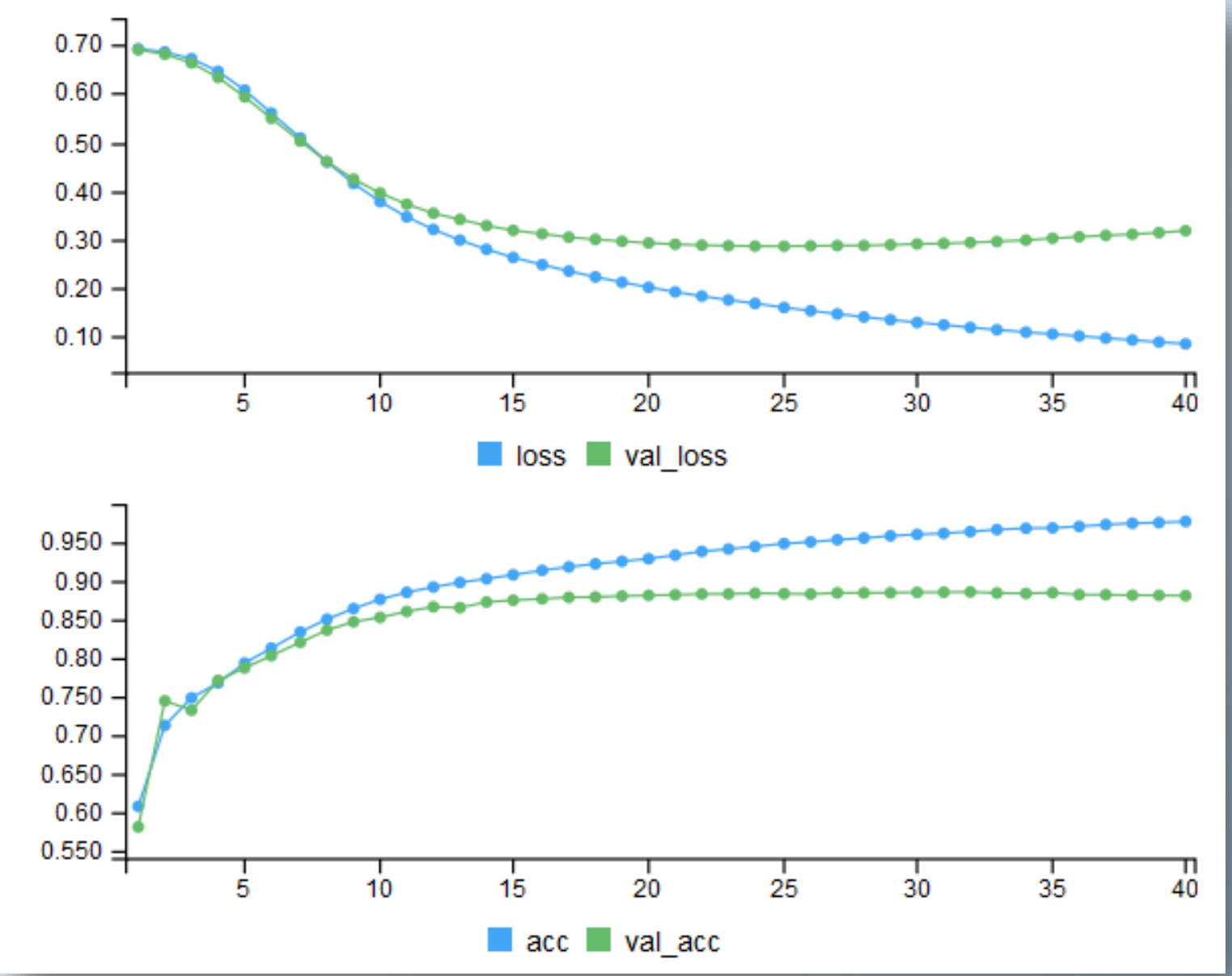
What's inside of deep learning?



Source: https://github.com/martisak/dotnets

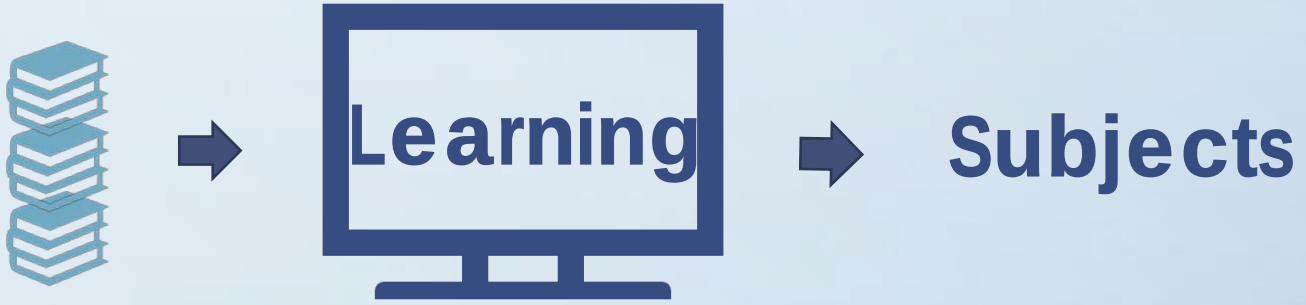
- Under the No. of layer = 3 with 16 hidden units, the accuracy rate reaches 98% for the training set, and **85%** for the validation set for the rest **10,000 reviews** with just 20 iterations (approx. 1 min)

- Modifying the layer structure might lead to a higher accuracy but higher consumptions on time and energy



Q1: Can we build the MODEL via the book dataset/**dictionary (e.g. frequency)** ?

Q2: Can we extract the **subjects** from **bibliographies** via the MODEL (?)



Q1: Can we build the MODEL via the book dataset/**dictionary (frequency)** ?

- For alphabet-based languages such as English, many successful methods have been reviewed in Cambria and White, 2014.
- For character-based languages such as Chinese, the effective learning algorithms are limited, since there is no space between Chinese characters in each sentence (Deng et al. 2016, PNAS). Challenges:
  - Word segmentation
  - Word and phrase discovery

Theoretically YES

Dictionary Construction

Unsupervised Model

- Without prior information
- TopWORDS(Deng et al. 2016); WDW(Ge et al. 1999)

Supervised Model

- Work well based on the prior information
- Fail when conflicting with the prior information(ancient Chinese grammar rules)

- Autonomous dictionary
- Potential keywords

Statistical Learnings Candidates

- Classical ML (incl. Logistic Regression)
- Adaboost/ADMM (based on weak classifiers)
- **Deep Learning**
- etc.

An experiment: finding subjects

- Use the full-text open source with MARC records (including LCSH) to train the model.
- Sample: Project Gutenberg (R package "gutenbergr" is available)

	gutenberg_id	subject_type	subject
	<int>	<chr>	<chr>
1	1	lcsh	United States. Declaration of Independence
2	1	lcsh	United States -- History -- Revolution, 1775-1~
3	2	lcsh	Civil rights -- United States -- Sources
4	2	lcsh	United States. Constitution. 1st-10th Amendmen~
5	3	lcsh	Presidents -- United States -- Inaugural addre~
6	3	lcsh	United States -- Foreign relations -- 1961-1963

Q2: Can we extract the **subjects** from **bibliographies** via the MODEL?

Challenges: budget

- High demands on servers
- Tremendous needs on electricity

Consumption evaluation

Challenges: accuracy

- Various types of languages
- Different types of record formats

Model bias judgement

Challenges: ethics

- Copyright for digitalizing/scanning the whole text;
- etc.

Opportunities

- Try to train models based on current digital collection;
- Cooperate with statistical experts;
- Our new horizon, such as: IG in LITA established in 2018.

Machine and Deep Learning Research Interest Group

lita Empowering libraries through technology
The Library and Information Technology Association
A DIVISION OF THE AMERICAN LIBRARY ASSOCIATION

Charge

The Machine and Deep Learning Research Interest Group is a forum for researching potential applications of Machine and Deep Learning in library science, including discussions, publications and outreach to the wider Library community. Its goal is to educate librarians on uses of the complex techniques of machine learning and to provide a space for critically thinking both about new applications, and about the ethical and social impact of these technologies , as the field rapidly expands in the coming decade.

Thank you!

Anlin Yang

East Asian Cataloging Librarian
University of Iowa

In collaboration with Dr. Meng Xu
Department of Statistics, University of Haifa, Israel