Machine Learning and Dewey Decimal Classification

Freddy Wetjen
National Library of Norway

Session 115 Transforming Libraries via Automatic Indexing – Subject Analysis and Access

Nasjonalbiblioteket

Outline

Machine learning and Dewey classification attempts in the National Library of Norway (NLN)

- Why?
- How ?
- Results

Nasjonalbiblioteket

What is Machine Learning at NLN?

- NLN has a machine learning lab
- Hands-on experiences with AI technology
- We work with AI and ML on different fields and media types
- AI and ML are tested with all major media types (Film,photo,text,sound..)
- Used for categorization, classification,recognition and discovery
- Build small applications to show the power of machine learning
- Identify strengths and weaknesses of the technology
- Close cooperation with Stanford University Library

AI is not a new technology and certainly not a new way of problem solving.

Machine learning models have improved much in the last five years

The concept of manual knowledge modelling in AI systems is almost gone

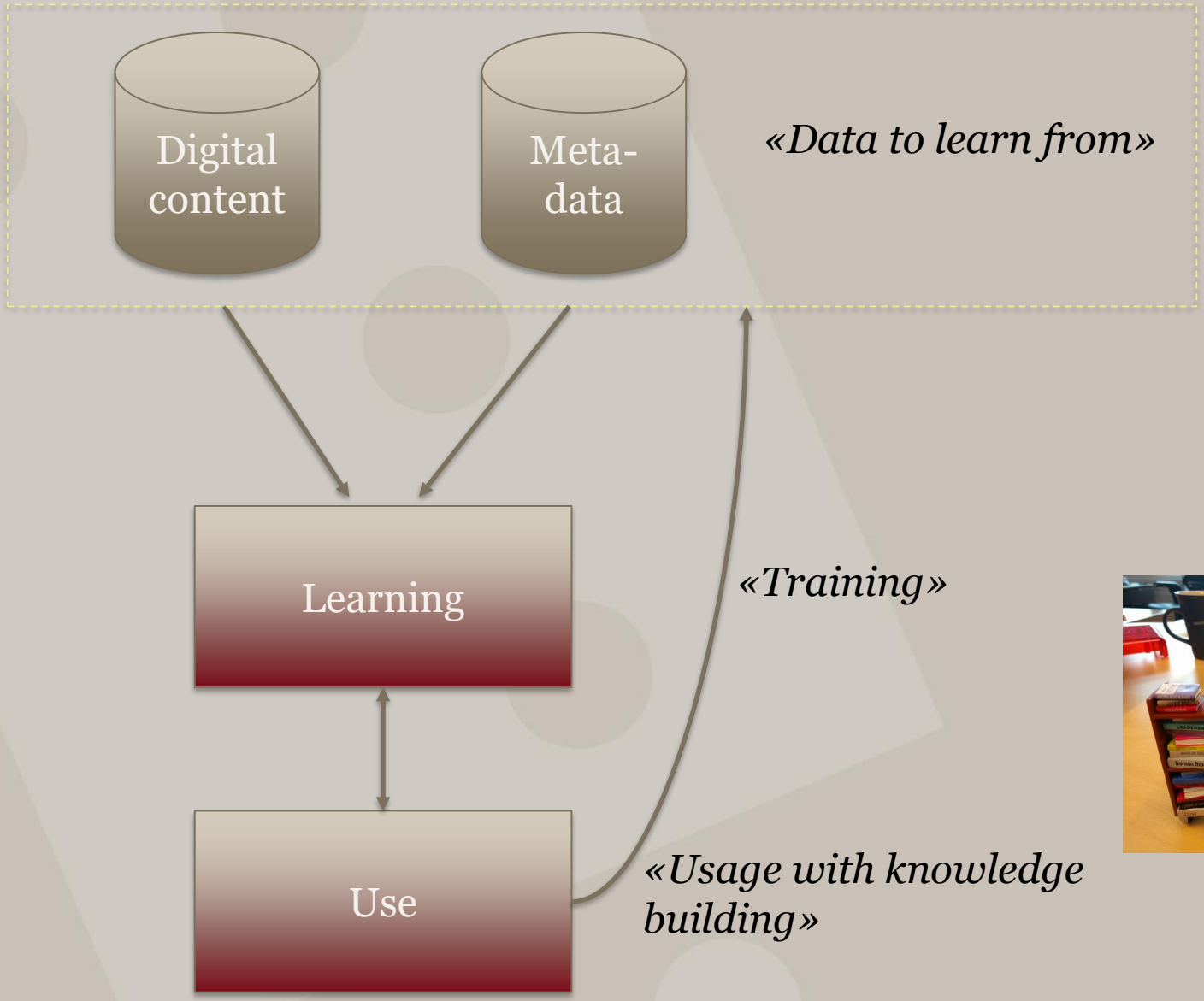
Instead, we have introduced the data science concept into machine learning and AI; we let the system build its own knowledge model although carefully selecting the «learning material».

AI methods gets widely available through open frameworks such as Tensorflow,Pytorch, gensim etc.

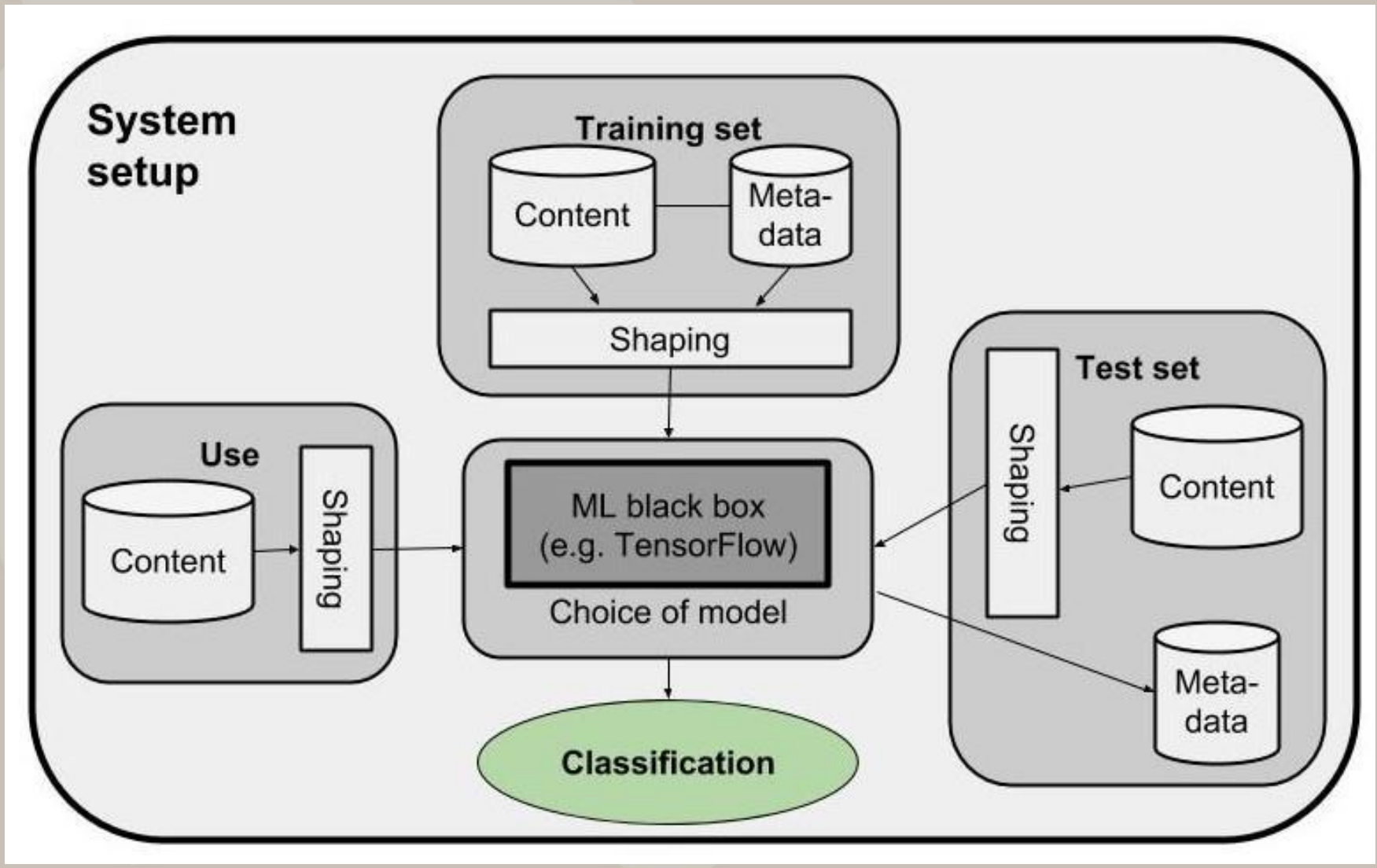
Increasing demand for data science specialists and programmers with knowledge and understanding of ML algorithms

From *programs* to *rules* to *learning*

- Tradition in programming
  - If-then-else
  - Control and precision
  - Deterministic
- Machine Learning
  - Learning from example data
  - Learning as an automatized task
  - Approximate
  - Non deterministic



Nasjonalbiblioteket



Prerequisites

- Computing power
  - Less power, more time
- Software
  - Mature open-source community
- Training and test data
  - Supervised learning requires high quality labeled data
  - Digital content with metadata (libraries)
- Skills in ML

Why ML at NLN?

NLN going digital - ambition

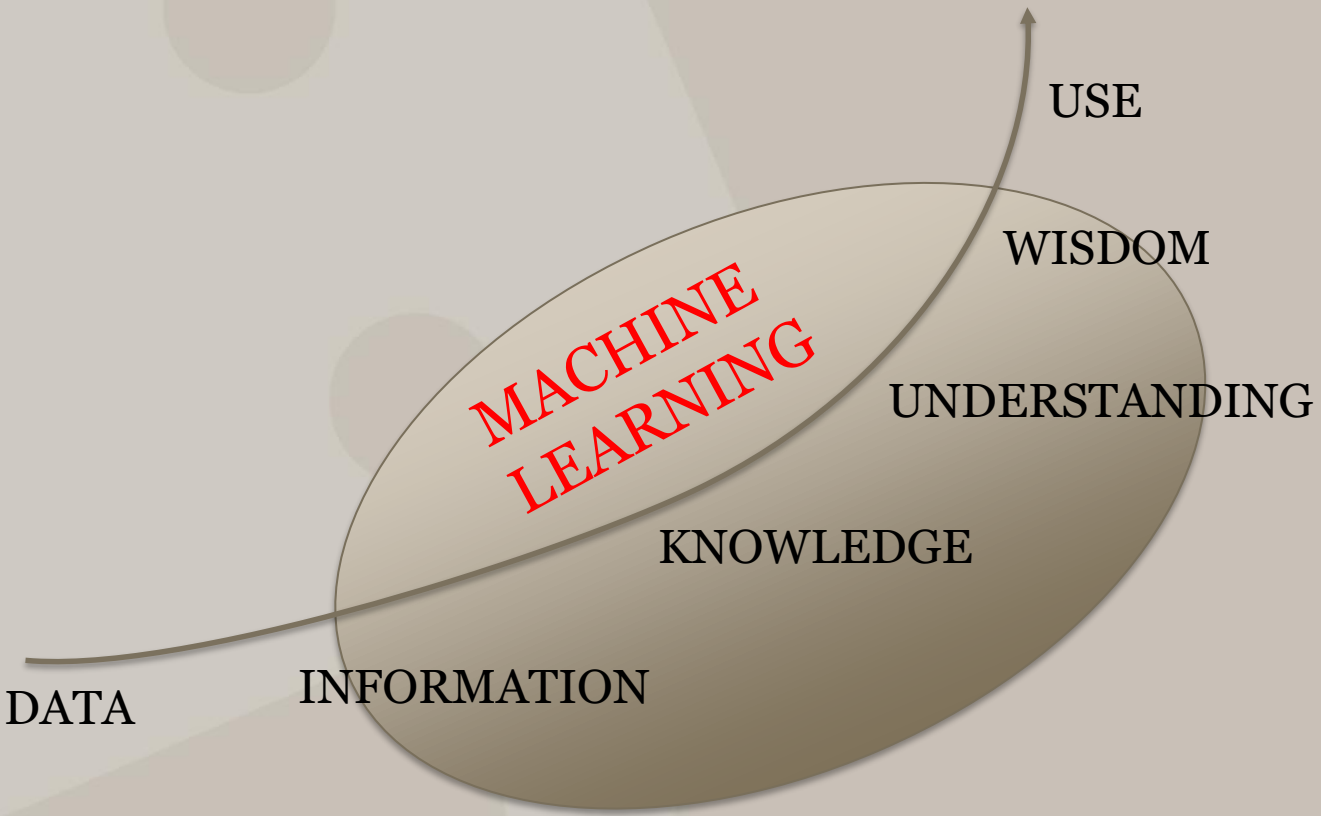
- Mass digitization
  - The complete collection is supposed to be digitized (2006)
  - Most of the published books close to 50 % of all newspaper editions are digitized
- Digital library
  - A complete library at the user's fingertips
  - Search in everything, access to everything
  - UX improvements wanted

Nasjonalbiblioteket

NLN is the perfect playground

- Massive digital content in all forms
- Good metadata for some data
- User data (user behaviour)
- Good domain understanding, high level of digital skills
- Mature digitalisation technology

ML helps us being a library

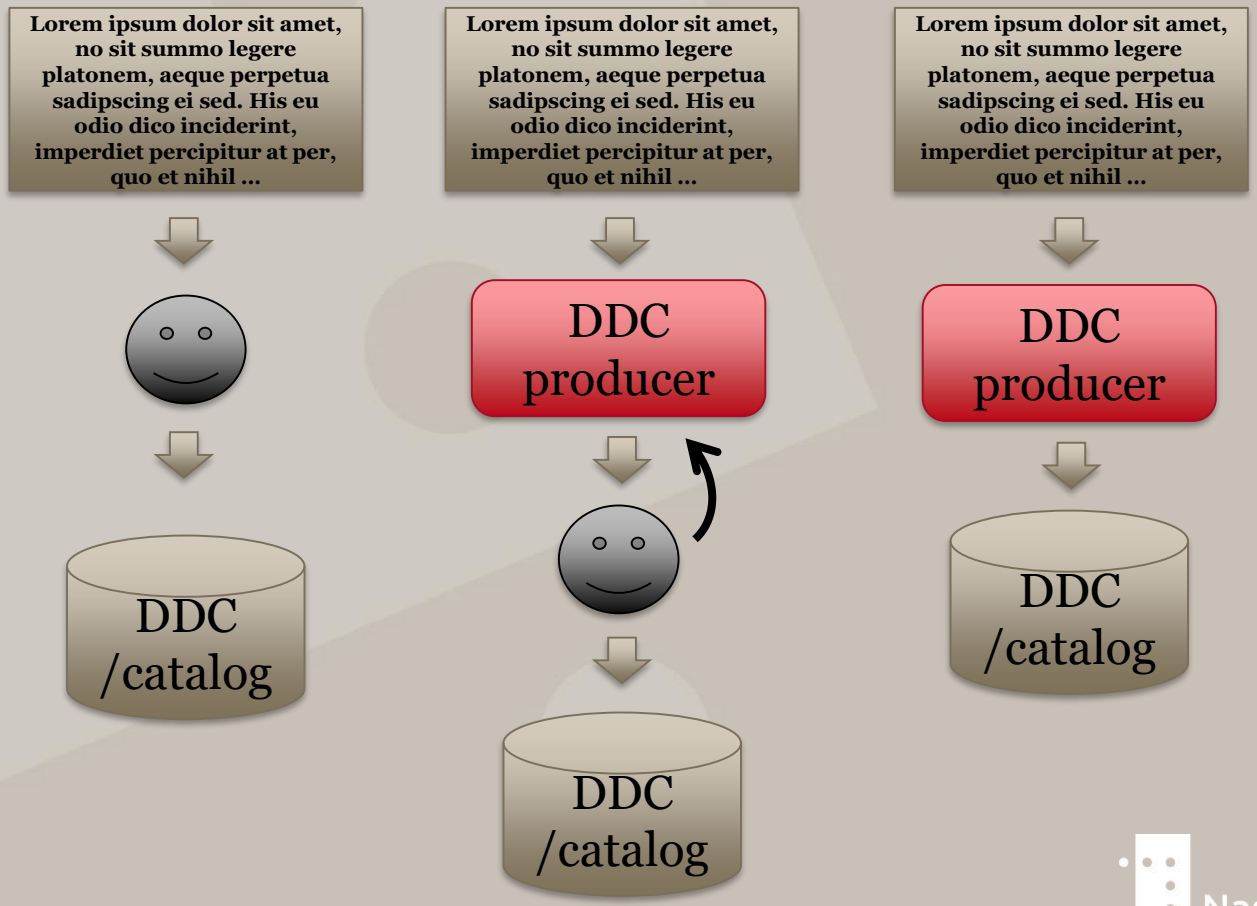


Nasjonalbiblioteket

Various experiments carried out

- Grouping of litterature
  - Poetry, Cooking, Sci-Fi, Crime…
- Identifying grey litterature
- Speech to text
- Analyzing still images and moving images (video), identifying objects
- Image and video search and identification
- Finding persons, places, organizations and more in text – and relationships between those
- Speaker identification
- Sound fingerprinting

Ambition: Alternative workflows



Nasjonalbiblioteket

Nasjonalbiblioteket

Dewey Decimal Classification experiments with their results

Nasjonalbiblioteket


LIBRARY MAP
http://publ.ac.uk/link/ddc.html
Mapped by rwild@1-900-870-6235.com
Knowledge is not delivered; it grows in unpredictable ways!
Home MAP
000 Generalities
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Natural sciences & mathematics
600 Technology (applied sciences)
700 The arts
Fine & decorative arts
Art galleries & museums
800 Literature & rhetoric
900 Geography & history
History of
Classics
Philosophy
Tools
See the Big Picture
Books Online
Zoom In
Contact Us
Help
Start
Here be Dragons
Feedback
080 General collections
060 General organisations & museology
020 Library and information sciences
004-006 Computing and the Internet
003 Systems
002 The book
090 Manuscripts, rare books, other rare printed materials
070 News media, journalism, publishing
030 General encyclopaedic works
010 Bibliography
001 Knowledge, humanities & research
107 education
100 societies
100 journals
100 departments
100 general resources
120 Epistemology, causation, humankind
130 Paranormal phenomena
150 Psychology
180 Ancient, medieval, oriental philosophy
190 Modern Western philosophy
170 Ethics (moral philosophy)
160 Logic
140 Specific philosophical schools and viewpoints
110 Metaphysics
200 Religion: education & research
230 Christianity: general resources
230 Christianity: Journals
238 Christian creeds & catechisms
242 Christian writings
268 Christian education
280 Christian denominations
294.4 Jainism
294.6 Sikhism
296 Judaism
297.93 Baha'i Faith
299 Other religions
299.514 Taoism
297 Islam
295 Zoroastrianism
294.5 Hindusm
294.3 Buddhism
270 History of Christianity
252 Texts of sermons
246 Christian art
234.161 Baptism
220 Bible
210 Philosophy & theory of religion
200 Religion: general resources
300 Social sciences: general resources
310 Collections of general statistics
320 Political science
340 Law
350 Public administration
355 Military science
370 Education
390 Customs, etiquette, folklore
360 Social problems & services, associations
350 Commerce, communications, transportation
301-306 Sociology
307 Urban communities
330 Economics
401 Language: philosophy and theory
403 Dictionaries & encyclopaedias
407 Language education & research
409 Languages spoken around the world
411 Writing systems
415 Grammar
416 Sign language
459.7 Swedish language
459.82 Norwegian language
459 Romanian language
469 Portuguese language
490 Other languages
480 Greek language
470 Latin language
460 Spanish language
450 Italian language
440 French language
430 German language
420 English language
400 Language: general resources
410 Linguistics
414 Phonology & phonetics
500 Natural sciences
510 Mathematics
520 Astronomy & allied sciences
530 Physics
540 Chemistry
550 Earth sciences
560 Palaeontology
570 Life sciences, biology
580 Plants
590 Animals
600 Technology
610 Medical sciences, medicine
620 Engineering & allied operations
630 Agriculture & related technologies
640 Home economics & family living
650 Management & auxiliary services
660 Chemical engineering
670 Manufacturing
680 Manufacture for specific uses
690 Buildings
700 The arts: general resources
700 general resources
700 artists
701 philosophy
703 dictionaries & encyclopaedias
705 journals & magazines
706 organisations
707 art schools
709 History of art
730 Plastic arts, sculpture
750 Painting and paintings
770 Photography & photographs
790 Recreational & performing arts
796 Sport & outdoor activities
780 Music
780.9 Music of specific periods
760 Graphic arts, printmaking & prints
740 Drawing and decorative arts
720 Architecture
710 Civic and landscape art
708 World-wide
708 In the US
708 In the UK
800 Literature: general resources
808 General rhetoric
808.83 Fiction
809 Literary study & criticism
808.02 Authorship, writing & editorial techniques
808.838 Science fiction
808.88 Quotations
808.899 Children's literature
809.89 Women writers
809 Poetry: general resources
808.8 Literature: general collections
810 Canadian literature
830 German literature
840 French literature
860 Spanish & Portuguese literature
880 Classical Greek literature
890 Literature in other languages
870 Latin literature
850 Italian literature
820 English, Scottish & Irish literature
810 American literature in English
900 Geography & history
909 World history
900 History: journals
910 Geography and travel
980 South America
960 Africa
940 Europe
930 ancient world
941-2 the British Isles
950 Asia and the Far East
970 North America
990 other parts of the world
920 Biography
900 History: departments
900 History: general resources
929 Genealogy, names, insignia
Nasjonalbiblioteket

Using NORART as an example..

- NORART is a hub for access to published Nordic and Norwegian scientific articles
- All articles have dewey classification assigned
- Librarians are classifying all articles
- Time consuming intellectual work
- Carefully selecting publications of particular dewey classification to create train and test sets.
- Working with carefully selected data and testing
- Design of algorithms, parameters, data sets

Approach

- Define scope for DDC
  - Classes, layers
- Define training set
  - Size
  - Content (articles)
  - Existing metadata
- Define test set
  - Size
  - Content (articles)
  - Existing metadata

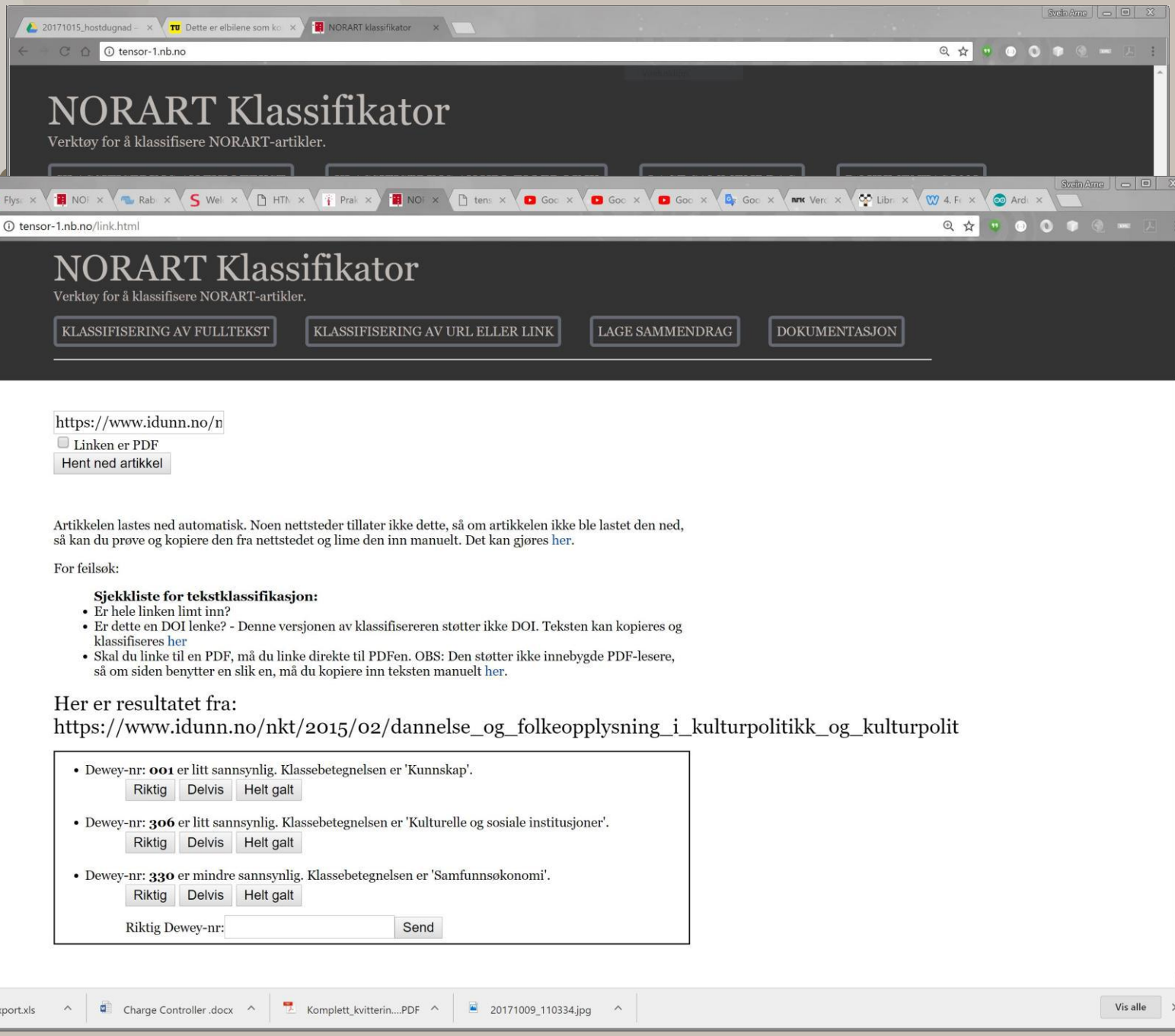
Constraints

- Limited no of DDC classes
- Only 3, 4, 5 and 6 levels
- More levels, less content per class

- Focus example: Automatic DDC identification of NORART scientific articles and content terms

Example of learning/test definition

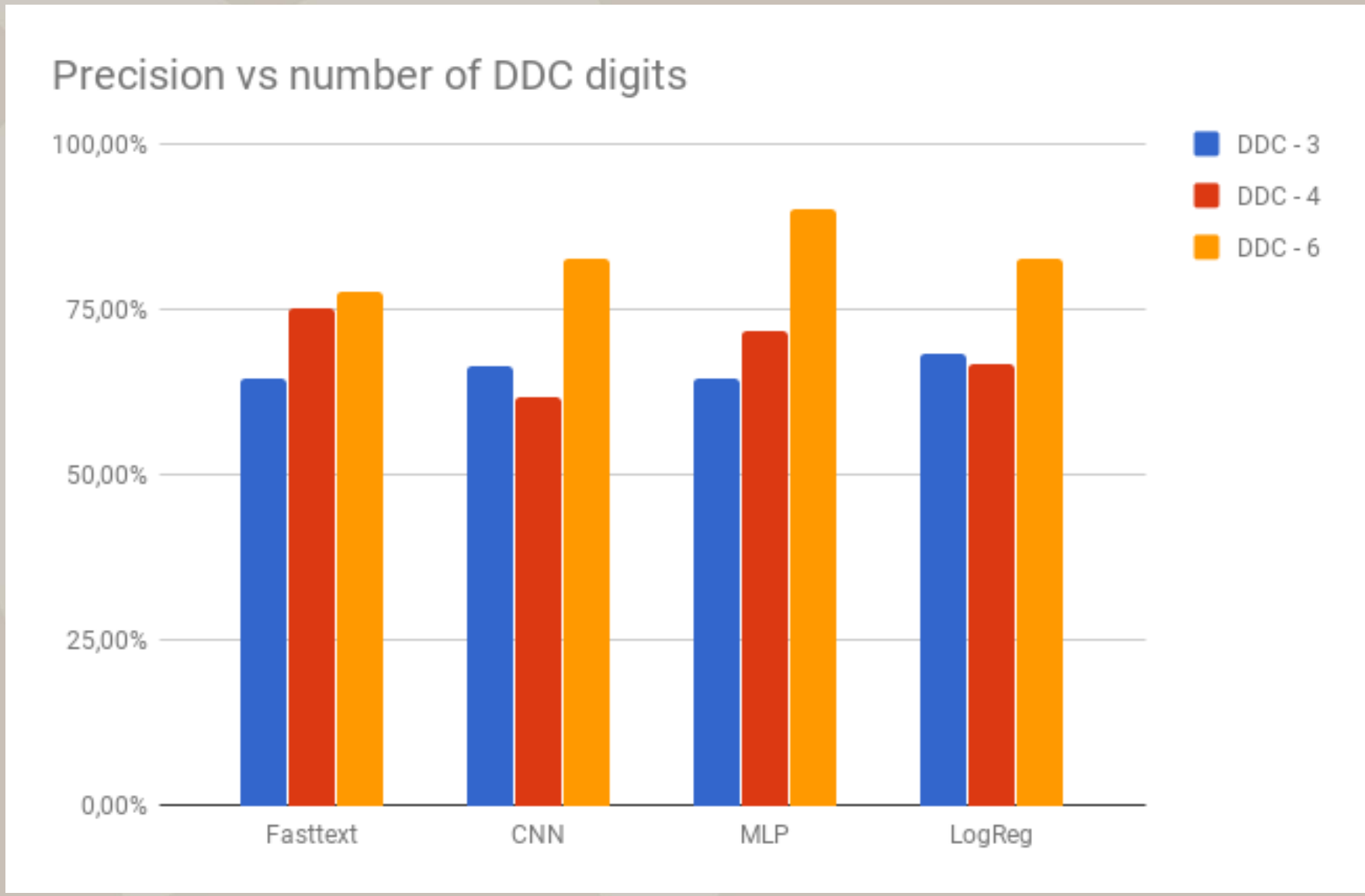
L=3	50	100	200	400
Test size	10	20	30	40
Real content only	Yes	Yes	Yes	Yes
Size of artifical content	5/10	10/20	20/40	40/80


20171015_hostdugnad
Dette er elbilene som ko
NORART klassifikator
tensor-1.nb.no
NORART Klassifikator
Verktøy for å klassifisere NORART-artikler.
tensor-1.nb.no/link.html
NORART Klassifikator
Verktøy for å klassifisere NORART-artikler.
KLASSIFISERING AV FULLTEKST
KLASSIFISERING AV URL ELLER LINK
LAGE SAMMENDRAG
DOKUMENTASJON
https://www.idunn.no/n
Linken er PDF
Hent ned artikkel
Artikkelen lastes ned automatisk. Noen nettsteder tillater ikke dette, så om artikkelen ikke ble lastet den ned, så kan du prøve og kopiere den fra nettstedet og lime den inn manuelt. Det kan gjøres her.
For feilsøk:
Sjekkliste for tekstklassifikasjon:
Er hele linken limt inn?
Er dette en DOI lenke? - Denne versjonen av klassifisereren støtter ikke DOI. Teksten kan kopieres og klassifiseres her
Skal du linke til en PDF, må du linke direkte til PDFen. OBS: Den støtter ikke innebygde PDF-lesere, så om siden benytter en slik en, må du kopiere inn teksten manuelt her.
Her er resultatet fra:
https://www.idunn.no/nkt/2015/02/dannelse_og_folkeopplysning_i_kulturpolitikk_og_kulturpolit
Dewey-nr: 001 er litt sannsynlig. Klassebetegnelsen er 'Kunnskap'.
Riktig
Delvis
Helt galt
Dewey-nr: 306 er litt sannsynlig. Klassebetegnelsen er 'Kulturelle og sosiale institusjoner'.
Riktig
Delvis
Helt galt
Dewey-nr: 330 er mindre sannsynlig. Klassebetegnelsen er 'Samfunnsøkonomi'.
Riktig
Delvis
Helt galt
Riktig Dewey-nr:
Send
transactionExport.xls
Charge Controller .docx
Komplett_kvitterin....PDF
20171009_110334.jpg
Vis alle

Btw: Artificial documents

- Used to improve the size of the training set
- «New» articles are produced by interchanging words between articles with the same DDC, or by replacing words/terms with synonyms
- Care taken not to insert bias; Not an easy task to avoid. Using artificial documents has its downside


Precision vs size of training set
100,00%
75,00%
50,00%
25,00%
0,00%
60
80
100
200
400
Fasttext
LogReg
Fasttext+
LogReg+


Precision vs number of DDC digits
100,00%
75,00%
50,00%
25,00%
0,00%
Fasttext
CNN
MLP
LogReg
DDC - 3
DDC - 4
DDC - 6

Improvements

- Reinforced learning
  - Continous improvement
  - Corrections from skilled librarians
  - Use of user behaviour
- Change of models

Conclusions

- Supervised learning on text and metadata from libraries works
- Relatively high precision in prediction of DDC
- Artificial documents helps
- Need for more training data
- Overall, modern ML will play a major role in digital libraries

Thanks for listening

freddy.wetjen@nb.no

Nasjonalbiblioteket