Fair Use in Data Mining and Machine Learning

A Comparative Study between Mainland China and Taiwan

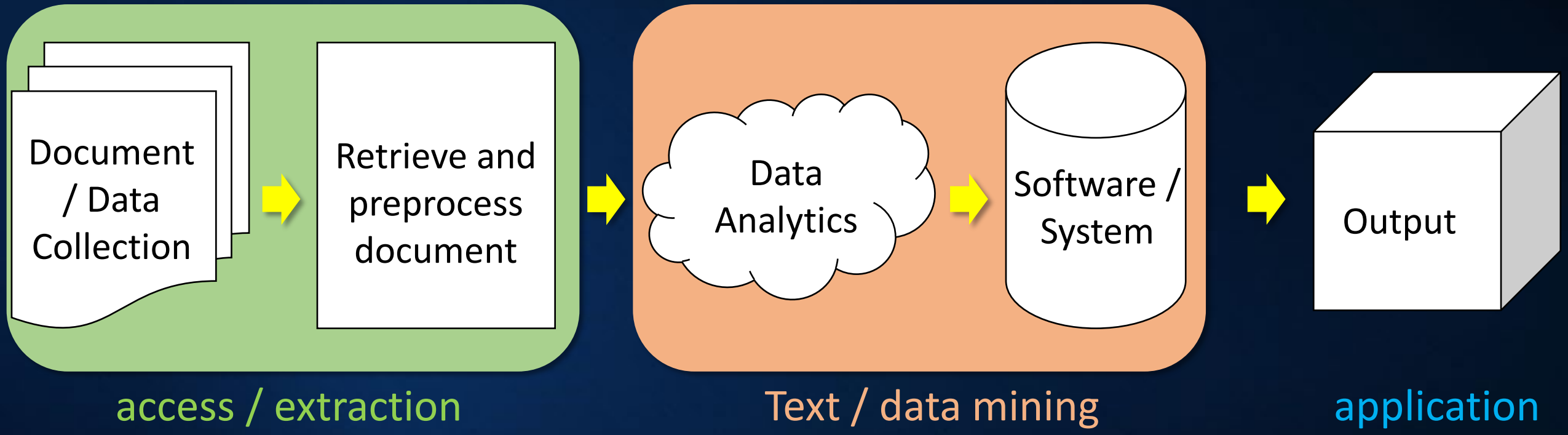
2019 IPSC

Assistant Professor Gi-Kuen Jacob Li
National Tsing Hua University, Taiwan
Institute of Law for Science and Technology

Outline

- Background
  - Technology
  - Legal
- Fair use in data mining & machine learning
  - US
  - Cross-strait comparison

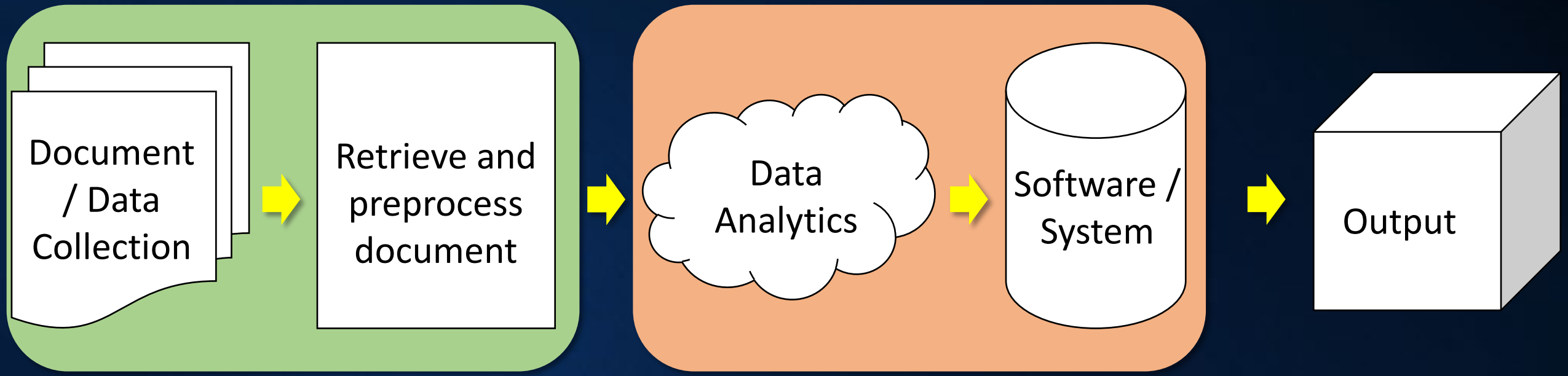
Rough Process of Data Mining



Data mining is the process of "finding valuable patterns in data."
SHOLOM M. WEISS ET AL., FUNDAMENTALS OF PREDICTIVE TEXT MINING 1 (2nd ed. 2015).

V. Gupta & G. S. Lehal, A survey of text mining techniques and applications, Journal of emerging technologies in web intelligence, Vol. 1, No. 1, pp. 60–76, 2009.
Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications, Vol. 39, Issue 12, pp. 11303-11311, 2012.

Rough Process of Data Mining



access / reproduce (copy)

Reproduce / prepare derivative works (?)

“Expression” / prepare derivative works (?)

Text Mining, Data Mining, and Machine Learning - Techniques and Applications

- Association rule
- Categorization / classification
- Neural network
- Decision tree
- Information retrieval
- Clustering and organizing
- Information extraction
- Prediction and evaluation

Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications, Vol. 39, Issue 12, pp. 11303-11311, 2012.
Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, *Overview of Text Mining,* in FUNDAMENTALS OF PREDICTIVE TEXT MINING (2015).

Legal Background

- Current US fair use doctrine could adept to resolve latent ambiguities, but there are still questions unanswered
  - Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. Copyright Soc'y USA (2019)
- Fair use for unauthorized, expressive uses of copyrighted works in commercial machine learning are undesirable
  - Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45 (2017)

Legal Background

- What about jurisdictions that adopt different fair use rules?
  - Mainland China followed the Berne three-step test with enumerated limitations
  - Taiwan blended categorical illustrations of limitations and US fair use factors into the Copyright Act

Fair Use for Data Mining and Machine Learning

General Issues and Background

“Expressive” and “Non-expressive” Use

- Foundation of Copyright Law
  - Idea-expression dichotomy
- Berne Convention Article 2 (1)
  - “The expression ‘literary and artistic works’ shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its **expression**, . . .”
- TRIPS Agreement Article 9 (2)
  - “Copyright protection shall extend to **expressions** and not to ideas, procedures, methods of operation or mathematical concepts as such.”

“Copying” in Data Mining and Machine Learning

- Can be “non-expressive”
  - Data mining in healthcare
  - Machine learning for precipitation forecast
  - Crime prediction

Original Contributions

Data Mining Applications in Healthcare

Hian Chye Koh and Gerald Tan

ABSTRACT

Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services.

The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making.

This article explores data mining applications in healthcare. In particular, it discusses data mining and its applications within healthcare in major areas such as the evaluation of treatment effectiveness, management of healthcare, customer relationship management, and the detection of fraud and abuse. It also gives an illustrative example of a healthcare data mining application involving the identification of risk factors associated with the onset of diabetes. Finally, the article highlights the limitations of data mining and discusses some future directions.

KEYWORDS

■ Healthcare management ■ Customer relationship management ■ Healthcare applications ■ Data mining methodology and techniques ■ Data mining applications ■ Predictive modeling

Introduction

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models.[1] Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns.[2]

Data mining is not new—it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling.

In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf

Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting

Xingjian Shi Zhourong Chen Hao Wang Dit-Yan Yeung
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{xshiab,zchenbb,hwangaz,dyyeung}@cse.ust.hk

Wai-kin Wong Wang-chun Woo
Hong Kong Observatory
Hong Kong, China
{wkwong,wcwoo}@hko.gov.hk

Abstract

The goal of precipitation nowcasting is to predict the future rainfall intensity in a local region over a relatively short period of time. Very few previous studies have examined this crucial and challenging weather forecasting problem from the machine learning perspective. In this paper, we formulate precipitation nowcasting as a spatiotemporal sequence forecasting problem in which both the input and the prediction target are spatiotemporal sequences. By extending the *fully connected LSTM* (FC-LSTM) to have convolutional structures in both the input-to-state and state-to-state transitions, we propose the *convolutional LSTM* (ConvLSTM) and use it to build an end-to-end trainable model for the precipitation nowcasting problem. Experiments show that our ConvLSTM network captures spatiotemporal correlations better and consistently outperforms FC-LSTM and the state-of-the-art operational ROVER algorithm for precipitation nowcasting.

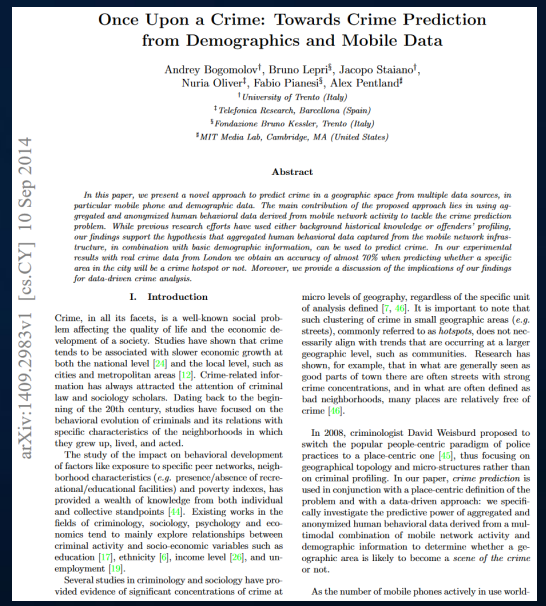
1 Introduction

Nowcasting convective precipitation has long been an important problem in the field of weather forecasting. The goal of this task is to give precise and timely prediction of rainfall intensity in a local region over a relatively short period of time (e.g., 0-6 hours). It is essential for taking such timely actions as generating society-level emergency rainfall alerts, producing weather guidance for airports, and seamless integration with a longer-term numerical weather prediction (NWP) model. Since the forecasting resolution and time accuracy required are much higher than other traditional forecasting tasks like weekly average temperature prediction, the precipitation nowcasting problem is quite challenging and has emerged as a hot research topic in the meteorology community [22].

Existing methods for precipitation nowcasting can roughly be categorized into two classes [22], namely, NWP based methods and radar echo[1] extrapolation based methods. For the NWP approach, making predictions at the nowcasting timescale requires a complex and meticulous simulation of the physical equations in the atmosphere model. Thus the current state-of-the-art operational precipitation nowcasting systems [19, 6] often adopt the faster and more accurate extrapolation based methods. Specifically, some computer vision techniques, especially optical flow based methods, have proven useful for making accurate extrapolation of radar maps [10, 6, 20]. One recent progress along this path is the *Real-time Optical flow by Variational methods for Echoes of Radar* (ROVER)

[1] In real-life systems, radar echo maps are often constant altitude plan position indicator (CAPPI) images [9].

http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf



arXiv:1409.2983v1 [cs.CY] 10 Sep 2014

Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data

Andrey Bogomolov[†], Bruno Lepri[§], Jacopo Staiano[†], Nuria Oliver[‡], Fabio Pianesi[§], Alex Pentland[♯]
[†] *University of Trento (Italy)*
[‡] *Telefonica Research, Barcellona (Spain)*
[§] *Fondazione Bruno Kessler, Trento (Italy)*
[♯] *MIT Media Lab, Cambridge, MA (United States)*

Abstract

In this paper, we present a novel approach to predict crime in a geographic space from multiple data sources, in particular mobile phone and demographic data. The main contribution of the proposed approach lies in using aggregated and anonymized human behavioral data derived from mobile network activity to tackle the crime prediction problem. While previous research efforts have used either background historical knowledge or offenders' profiling, our findings support the hypothesis that aggregated human behavioral data captured from the mobile network infrastructure, in combination with basic demographic information, can be used to predict crime. In our experimental results with real crime data from London we obtain an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not. Moreover, we provide a discussion of the implications of our findings for data-driven crime analysis.

I. Introduction

Crime, in all its facets, is a well-known social problem affecting the quality of life and the economic development of a society. Studies have shown that crime tends to be associated with slower economic growth at both the national level [24] and the local level, such as cities and metropolitan areas [12]. Crime-related information has always attracted the attention of criminal law and sociology scholars. Dating back to the beginning of the 20th century, studies have focused on the behavioral evolution of criminals and its relations with specific characteristics of the neighborhoods in which they grew up, lived, and acted.

The study of the impact on behavioral development of factors like exposure to specific peer networks, neighborhood characteristics (*e.g.* presence/absence of recreational/educational facilities) and poverty indexes, has provided a wealth of knowledge from both individual and collective standpoints [44]. Existing works in the fields of criminology, sociology, psychology and economics tend to mainly explore relationships between criminal activity and socio-economic variables such as education [17], ethnicity [6], income level [26], and unemployment [10].

Several studies in criminology and sociology have provided evidence of significant concentrations of crime at micro levels of geography, regardless of the specific unit of analysis defined [7, 46]. It is important to note that such clustering of crime in small geographic areas (*e.g.* streets), commonly referred to as *hotspots*, does not necessarily align with trends that are occurring at a larger geographic level, such as communities. Research has shown, for example, that in what are generally seen as good parts of town there are often streets with strong crime concentrations, and in what are often defined as bad neighborhoods, many places are relatively free of crime [46].

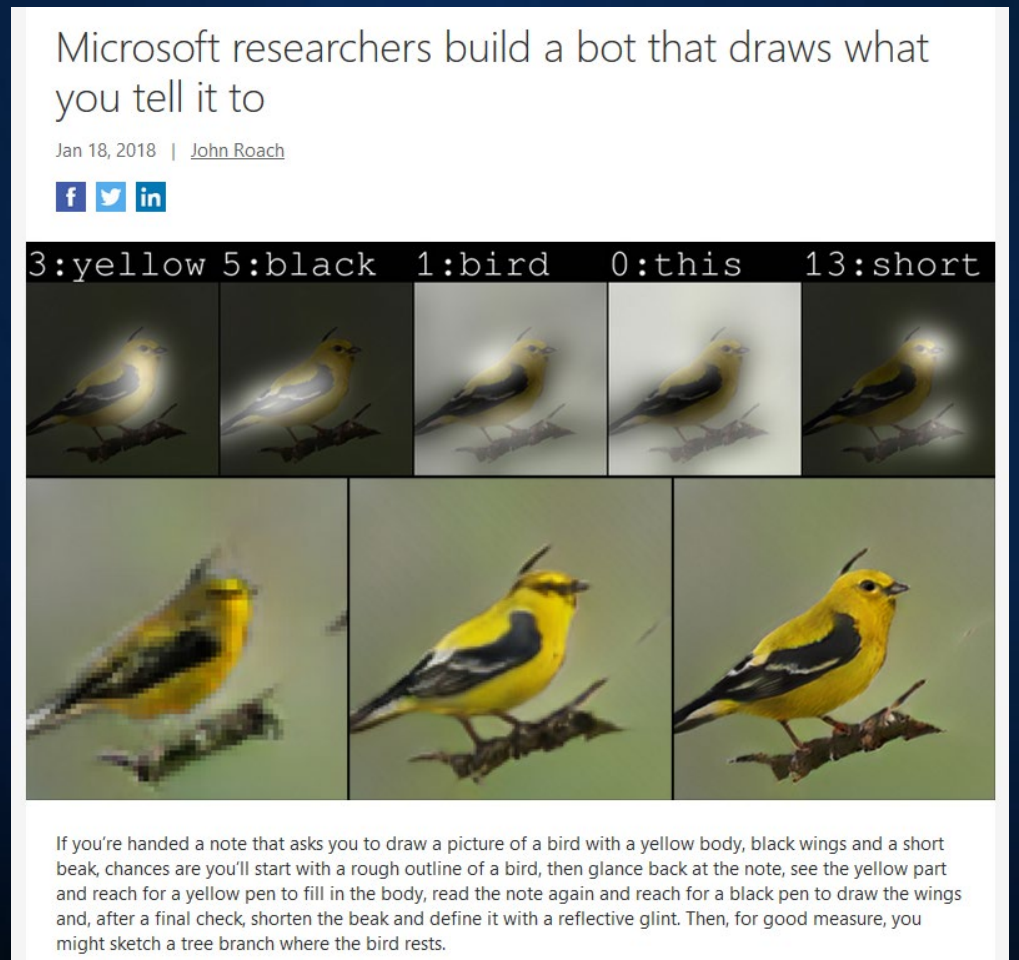
In 2008, criminologist David Weisburd proposed to switch the popular people-centric paradigm of police practices to a place-centric one [45], thus focusing on geographical topology and micro-structures rather than on criminal profiling. In our paper, *crime prediction* is used in conjunction with a place-centric definition of the problem and with a data-driven approach: we specifically investigate the predictive power of aggregated and anonymized human behavioral data derived from a multimodal combination of mobile network activity and demographic information to determine whether a geographic area is likely to become a *scene of the crime* or not.

As the number of mobile phones actively in use world-

https://arxiv.org/pdf/1409.2983.pdf

“Copying” in Data Mining and Machine Learning

- Copying lead to “expressive” results?



“The technology, which the researchers simply call the drawing bot, can generate images of everything from ordinary pastoral scenes, such as grazing livestock, to the absurd, such as a floating double-decker bus. **Each image contains details that are absent from the text descriptions**, indicating that this artificial intelligence contains an **artificial imagination**.”

https://blogs.microsoft.com/ai/drawing-ai/

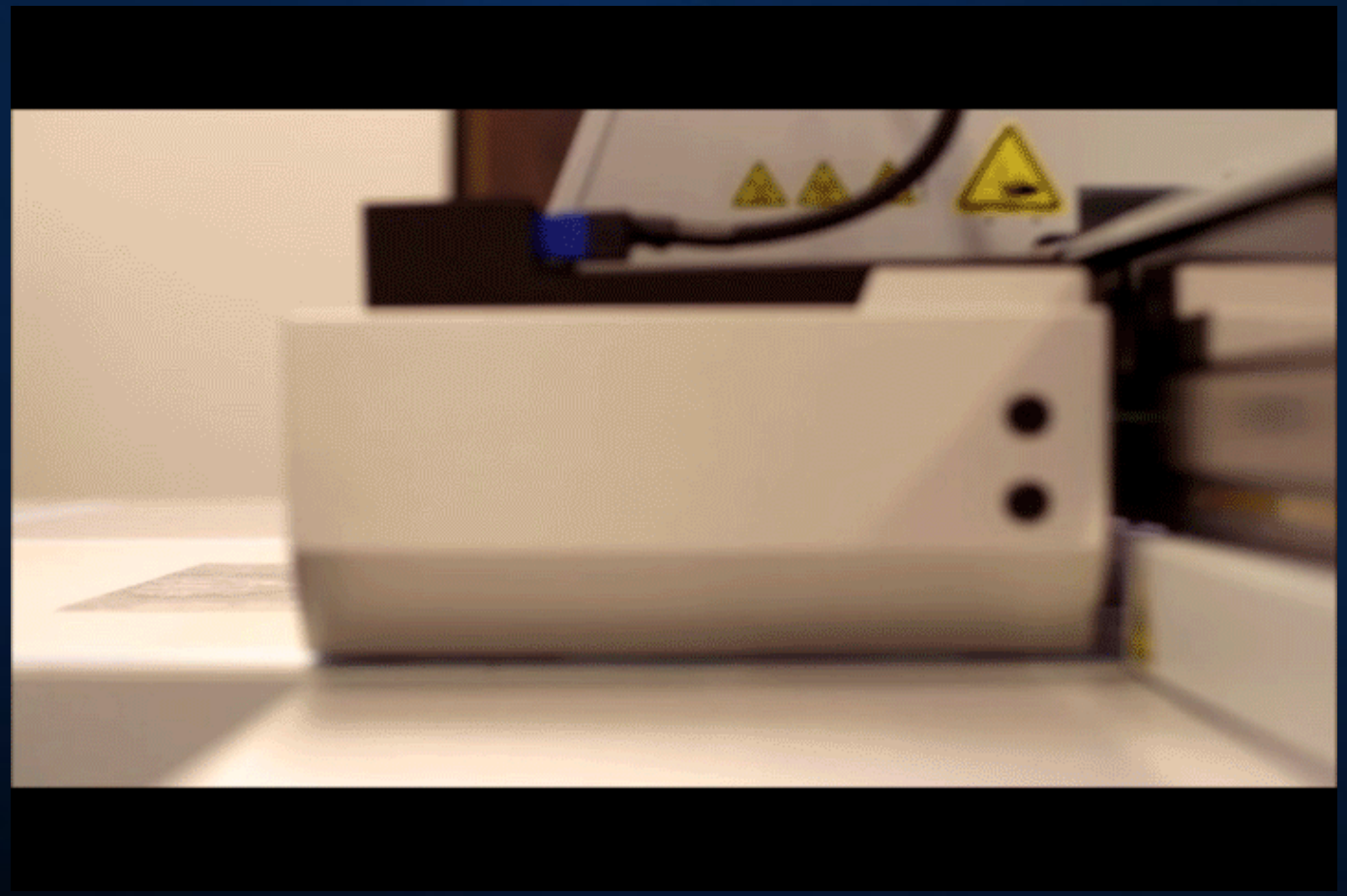
“Copying” in Data Mining and Machine Learning

- Copying lead to “expressive” results?

https://youtu.be/vUgUeFu2Dcw?t=54s

“Copying” in Data Mining and Machine Learning

- Copying lead to “expressive” results?



https://www.nextrembrandt.com/
https://news.microsoft.com/europe/features/next-rembrandt/

“Copying” in Data Mining and Machine Learning

- Copying lead to “expressive” results?
  - Computoser
    - http://computoser.com/
  - FLOW MACHINES
    - http://www.flow-machines.com/

Walking the Fine Line?

- How to differentiate “expressive” and “non-expressive” use
  - Research results can be utilized in very different way

Not Just A Copyright Issue

- Ethics guidelines for developing AI
- Breach of contract? (terms of use)
  - Binding or not
  - Scope
- Privacy

Fair Use for Data Mining and Machine Learning

Cross-strait Comparison

Limitations of Exclusive Rights (China)

- Article 22: Enumerated Fair Use
  - In the following cases, a work may be used without permission from, and without payment of remuneration to, the copyright owner, provided that the name of the author and the title of the work shall be indicated and the other rights enjoyed by the copyright owner by virtue of this Law shall not be prejudiced:

Limitations of Exclusive Rights (China)

- Article 22: Enumerated Fair Use (cont.)

  (1) use of a published work of others for the user's own private study, research or self entertainment;

  (2) appropriate quotation from a published work of others in one's work for the purposes of introduction to, or comment on, a work, or demonstration of a point;

  (3) use of a published work in newspapers, periodicals, radio programs, television programs or newsreels for the purpose of reporting current affairs;

  (4) reprinting by newspapers or periodicals, or rebroadcasting by radio stations or television stations, of editorials or commentator's articles published by other newspapers, periodicals, radio stations or television stations;

  (5) publication in newspapers or periodicals, or broadcasting by radio stations or television stations, of a speech delivered at a public gathering, except where the author has declared that the publication or broadcasting is not permitted;

  (6) translation or reproduction in a small quantity of copies, of a published work for use by teachers or scientific researchers, in classroom teaching or scientific research, provided that the translation or reproduction **shall not be published or distributed**;

Limitations of Exclusive Rights (China)

- Article 22: Enumerated Fair Use (cont.)

  (7) use of a published work by a state organ for the purpose of performing its official duties;

  (8) reproduction of a work in its collections by a library, an archives center, a memorial hall, a museum, an art gallery or a similar institution, for the purposes of display, or preservation of a copy, of the work;

  (9) free performance of a published work;

  (10) copying, drawing, photographing, or video recording of an artistic work located or on display in an outdoor public place;

  (11) translation of a published work from the language of the Han nationality into minority nationality languages for publication and distribution in the country;

  (12) transliteration of a published work into Braille and publication of the work so transliterated.

  The above limitations on rights shall be applicable also to the rights of publishers, performers, producers of sound recordings and video recordings, radio stations and television stations.

Limitations of Exclusive Rights (China)

- Article 23:

  Anyone who compiles or publishes textbooks for the purpose of implementing the nine-year compulsory education or State education planning may, without the permission from the copyright owner, except that the author has declared in advance that the exploitation is not permitted, compile published fragments of works, short written works or musical works, a single work of fine art, or photographic works into the textbooks, however, he shall pay the remuneration as provided, mention the name of the author and the title of the work, and shall not infringe upon other rights which the copyright owner shall enjoy in accordance with this Law.

  The provisions in the preceding paragraph shall be applicable to the limitations on the rights of publishers, performers, producers of sound recordings and video recordings, radio stations and television stations.

Are the Rules Applicable to DM or ML?

- Article 22: Enumerated Fair Use
  - (1) use of a published work of others for the user's own private study, research or self entertainment;
  - (2017) Zhe 0381 Mingchu 832 Hao civil decision:
    - Defendant "linmo" (learn to draw or write calligraphy by mimicking some famous works of art) Plaintiff's paintings. The court stated that "linmo" is fair use, and it is a common practice in the society. However, the user is required to (1) attribute work to the original author and indicate the name of the work; (2) the use must not infringe other's exclusive rights
  - (2000) HuErZhongZhiChuZiDi 109 Hao civil decision:
    - Defendant's website highly resembled Plaintiff's website in the look and feel, structure, context, color, and font. The court stated that fair use is limited to personal use, and it excludes any commercial use. The scope must be limited, internal, and not published.

Are the Rules Applicable to DM or ML?

- Article 22: Enumerated Fair Use
  - (6) translation or reproduction in a small quantity of copies, of a published work for use by teachers or scientific researchers, in classroom teaching or scientific research, provided that the translation or reproduction **shall not be published or distributed**;
  - Most cases concern unlicensed copy and distribution of teaching materials
  - Uncertain for research purposes
    - Even if it applies, the research outcome or result cannot be published or distributed

Limitations on Economic Rights (Taiwan)

- Article 44 ~ Article 63 illustrations of limitations
  - Article 44 & 45:
    - central or local government agencies may reproduce for the purpose of legislation, administration, and for judicial proceedings
  - Article 46 & 47:
    - for the educational purposes, including teaching in schools, preparing teaching materials
  - Article 48 & 48-1
    - for archival purposes of libraries, museums, and other cultural institutions
  - Article 49
    - for News report purposes

Limitations on Economic Rights (Taiwan)

- Article 44 ~ Article 63 illustrations of limitations (cont.)
  - Article 52:
    - Within a reasonable scope, works that have been publicly released may be quoted where necessary for reports, comment, teaching, research, or other legitimate purposes.
  - Article 59:
    - The owner of a legal copy of a computer program may alter the program where necessary for utilization on a machine used by such owner, or may reproduce the program as necessary for backup; provided, this is limited to the owner's personal use.
      If the owner referred to in the preceding paragraph loses ownership of the original copy for any reason other than the destruction or loss of the copy, all altered and backup copies shall be destroyed unless the economic rights holder grants its consent otherwise.

Limitations on Economic Rights (Taiwan)

- Article 65
  - Dual function
    - Standard for interpreting the illustrations of limitations – “within a reasonable scope”
    - Can be directly assert as a defense of an infringing act

Limitations on Economic Rights (Taiwan)

- Article 65
  - Fair use of a work shall not constitute infringement on economic rights in the work.

    In determining whether the exploitation of a work complies with the reasonable scope referred to in the provisions of Articles 44 through 63, or other conditions of fair use, all circumstances shall be taken into account, and in particular the following facts shall be noted as the basis for determination:

    1. The purposes and nature of the exploitation, including whether such exploitation is of a commercial nature or is for nonprofit educational purposes.

    2. The nature of the work.

    3. The amount and substantiality of the portion exploited in relation to the work as a whole.

    4. Effect of the exploitation on the work's current and potential market value.

    Where the copyright owner organization and the exploiter organization have formed an agreement on the scope of the fair use of a work, it may be taken as reference in the determination referred to in the preceding paragraph.

    In the course of forming an agreement referred to in the preceding paragraph, advice may be sought from the specialized agency in charge of copyright matters.

Are the Rules Applicable to DM or ML?

- Probably in similar situation to the US
- Could (or should) the illustrations or enumeration of limitations become the more feasible legislative solution?

Fair Use Factors / Illustrations / Enumerations

Fair Use Factors	
Pros	**Cons**
More flexible	Flexibility might create uncertainty, which is risky for AI researchers at schools, startups, and corporations

Illustrations / Enumerations	
Pros	**Cons**
For lay person, it seems to be more predictable and trustworthy	Rigid rules may create interpretation and application problems for courts

Questions? Advice?

Thank you for your attention!

Gi-Kuen Jacob Li 李紀寬
Assistant Professor of Law
National Tsing Hua University, Taiwan
Institute of Law for Science and Technology
gkli@mx.nthu.edu.tw