Mens Machina: Using Data to Train Machines
Today’s Presenters

Judy Jennison
Trademark Litigation

Nicola Menaldo
Privacy & Security

James G. Snell
Privacy & Security
Presentation Overview

- What is Artificial Intelligence (AI) and Machine Learning (ML)?
- Obtaining Datasets for ML
- The Problem of Biometrics
The War Against the Machines

AND THEN DMITRI NOTICED SOMETHING THAT WOULD HAVE A PROFOUND EFFECT ON THE HUMAN/ROBOT WARS.

© Guy & Rodd/Distributed by Universal Uclick via CartoonStock.com
What is AI and Machine Learning?

**ARTIFICIAL INTELLIGENCE**
Ai is the capability of a machine to imitate intelligent human behavior

**TRADITIONAL AI SYSTEMS**
Traditional AI systems are programmed to simulate human intelligence (e.g., IBM’s Deep Blue)

**MACHINE LEARNING**
ML is a subset of AI involving a system that learns from data without relying on rules-based programming (e.g., Google Deepmind’s AlphaGo)
Machine Learning
How Does it Differ from Traditional Software?

DISTINCTIONS

• Traditional software requires hand-coding with specific instructions to complete a task

• An ML system learns to recognize patterns and make predictions using large amounts of data

EXAMPLE:

• Spam the old way: “if the email contains the word ‘viagra,’ then …”

• Spam the new way: ML system learns from training data to identify if email is spam
How is a Machine Learning Model Developed?

**IDENTIFIES FEATURES**
(e.g., Country of origination, time sent...)

- Input
  (e.g., Emails as training data, ...)

- Learner - Algorithm

- Parameters
  (distilled learnings)

- Output
  Machine Model
How does a Supervised Machine Learning System Operate?

**INPUT**
(e.g., Emails as training data)

**Machine Model**

**PREDICTION**
Is the email spam?

**Parameters**

**Learner - Algorithm**

**OUTPUT – TRUTH**
Emails marked as spam
Europe’s GDPR: requires providing meaningful information about the “logic involved” → can only be achieved by disclosing what data is considered

Algorithmic Discrimination: Biased Datasets = Biased Products, so understanding bias requires understanding the data

• Example
Machine Learning Applications Across Industries

**Manufacturing**
- Predictive maintenance or condition monitoring
- Warranty reserve estimation
- Propensity to buy
- Demand forecasting
- Process optimization
- Telematics

**Retail**
- Predictive inventory planning
- Recommendation engines
- Upsell and cross-channel marketing
- Market segmentation and targeting
- Customer ROI and lifetime value

**Healthcare and Life Sciences**
- Alerts and diagnostics from real-time patient data
- Disease identification and risk stratification
- Patient triage optimization
- Proactive health management
- Healthcare provider sentiment analysis

**Travel and Hospitality**
- Aircraft scheduling
- Dynamic pricing
- Social media – consumer feedback and interaction analysis
- Customer complaint resolution
- Traffic patterns and congestion management

**Financial Services**
- Risk analytics and regulation
- Customer Segmentation
- Cross-selling and up-selling
- Sales and marketing campaign management
- Credit worthiness evaluation

**Energy, Feedstock, and Utilities**
- Power usage analytics
- Seismic data processing
- Carbon emissions and trading
- Customer-specific pricing
- Smart grid management
- Energy demand and supply optimization

Source: Tata Consultancy Services, Using Big Data for Machine Learning Analytics in Manufacturing – TCS
How are Machine Learning Issues Emerging in the Law?

- Algorithmic Discrimination
- Intellectual Property, e.g., Copyright
- Privacy
- Automated Decision Making
- Antitrust: Algorithmic Collusion
- Product Liability
- Law Enforcement Access
Machine Learning

"I’m sure it’s got a mind of it’s own!"
Ways to Obtain Data for Machine Learning

- License the Datasets
- Create a New Dataset Based on Others’ Work (Scraping or Web Crawling)
- Create a New Dataset Based on New Works
Licensing Data
Example of Datasets Available for License

Imagenet.org

Spirula, Spirula peronii
A small tropical cephalopod of the genus Spirula having prominent eyes and short arms and a many-chambered shell coiled in a flat spiral.

Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (32328)
  - plant, flora, plant life (4496)
  - geological formation, formation (17)
  - natural object (1112)
  - sport, athletics (176)
  - artifact, artefact (10504)
  - fungus (308)
  - person, individual, someone, some
  - animal, animate being, beast, brute
  - invertebrate (706)
  - arthropod (579)
  - zoophyte (0)
  - sponge, poriferan, parazoan
  - coelenterate, cnidian (28)
  - ctenophore, comb jelly (4)
  - worm (38)
  - woodborer, borer (0)
  - rotifer (0)
  - mollusk, mollusc, shellfish (-)
  - scaphopod (1)
  - gastropod, univalve (33)
  - chiton, coat-of-mail shell
  - bivalve, pelecypod, lance
  - cephalopod, cephalopod
  - chambered nautilus, \textit{Dibranchiate, dibranchiate}
Licensing Data
Example of Datasets Available for License

Mapillary Vistas Dataset
Copyrightable

Stuff You Want

Not Copyrightable
Fair Use

1. Nature of D’s use
2. Nature of original
3. Amount taken
4. Effect on Market

Examples:
- Time Shifting
- Image Search
- Google Books

Not Fair Use
But remember...
Does the model trainer have a license to the dataset?

DATASETS ARE DOWNLOADED FROM THE INTERNET

Types of license grants include:

- **Creative Commons broad license grant** ("...Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below...")

- **No license but statement that “This dataset is for non-commercial use only.”**

- **Copyleft license** (GPL v.3) which requires disclosure of the source code of any program linked to the GPL licensed code and that does not obviously apply to use of a dataset
Does the license permit the proposed use?

Problem

Most online datasets are licensed for “non-commercial” use. Is training a machine learning model a “non-commercial” use?

Question has not yet reached the courts.

• May turn on licensor’s intent: Indirectly receiving compensation for custom-made ML models, or ML models embedded in hardware may not be what licensor intended for “non-commercial” license
• Using data for internal research is less clear
Does the licensor have the rights to the underlying data?

Licensors of datasets made from data scraped from the Internet may have a copyright in the compilation, but may not own or license the underlying data.

Due diligence is important

- Promises to remove content from datasets upon request suggests that the datasets include other people’s original works
- Check whether the licensor has assets (an academic researcher may not be able to defend and indemnify in the event of a copyright dispute)
LITIGATION

Distributing a model that was trained on unlicensed data is unlikely to be an act of copyright infringement because the model does not include a representation of the creative elements of the works. But the acts of storing copyrighted content for use in training the data model may be alleged to be acts of “copying” for copyright purposes and to support a claim of copyright infringement.

RECOVERY UNDER THE ACT

Under the Copyright Act, a successful plaintiff may recover:

• actual damages;
• disgorgement of profits attributable to the infringement;
• statutory damages of between $750 and $30,000 per work infringed; up to $150,000 per work if the infringement is willful;
• attorneys’ fees; and
• an injunction preventing further infringement, potentially including use of the algorithm.
Mitigation for Licensing Issues

• Due Diligence

• Consider clarifying with copyright owner where there is uncertainty (such as whether training a model is “commercial use”)

• Pay for data where necessary (contract supersedes copyright law)

• Develop own datasets if possible (and if collecting data from consumers/employees, disclose that data may be used to improve products and services)
Creating a Dataset Based on Others’ Data (Scraping or Web Crawling)

3 Ways Scraped Data Can Build a Model

- License a dataset based on scraped data (e.g., IMDb dataset)
- Hire a third party to scrape data
- Scrape data yourself
Is Web Crawling and Scraping Legal?

**Highly fact-specific**, including based on the type of data scraped, applicable contract terms, whether it was public or protected by password or another measure, whether the scraping caused any harm apart from the data use, and the purpose for scraping/ the public interest.

**Main types of claims brought against scrapers:**

- Breach of contract (based on Terms of Use)
- Copyright infringement
- Digital Millennium Copyright Act (DMCA) (based on circumventing technological barriers)
- Computer Fraud and Abuse Act (CFAA)
- Trespass to chattels
Is Web Crawling and Scraping Legal?

**Legal risk can be minimized by:**

- Scraping only public data (do not scrape from password-protected, private sources or circumvent CAPTCHAs and other password-protected sources)
- Avoid scraping copyrighted works (factual, unoriginal textual elements—e.g., raw data—is not covered by copyright in the United States)
- Cease scraping if asked to stop
- Do not burden or damage websites when scraping
Scraping in a Changing Legal Landscape


(granting preliminary injunction against LinkedIn to permit continued scraping by hiQ and noting that the CFAA was intended to cover hacking, not access to public data)


(ACLU challenge to the constitutionality of the CFAA as applied to scraping and web crawling by academics, researchers, and journalists testing for housing discrimination on the Internet) (motion to dismiss pending)

**Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058 (9th Cir. 2016) and United States v. Nosal, 676 F.3d 854 (9th Cir. 2012)**

Certiori petitions that ask the Supreme Court to interpret the term “authorization” under the CFAA
Biometrics Legal Landscape

Existing Laws

Illinois
740 Ill. Comp. Stat. 14/10

Texas
Tex. Bus. & Com. Code § 503.001

Washington
H.B. 1493

Proposed Legislation 2017

Alaska
Connecticut
Michigan
Montana
New Hampshire
**Biometric identifier** defined as “data generated by automatic measurements of an individual's biological characteristics, such as a fingerprint, voiceprint, eye retinas, irises, or other unique biological patterns or characteristics that is used to identify a specific individual” (not including data generated from video or audio recording)

**Washington**

“Biometric identifier” defined as “data generated by automatic measurements of an individual's biological characteristics, such as a fingerprint, voiceprint, eye retinas, irises, or other unique biological patterns or characteristics that is used to identify a specific individual” (not including data generated from video or audio recording)

**Illinois and Texas**

“Biometric identifiers” defined as fingerprints, retina and iris scans, scans/records of face geometry, scans/records of hand geometry, and voiceprints

**Illinois**

“Biometric information” is “any information . . . based on an individual's biometric identifier used to identify an individual”

**What Are Biometrics?**
Biometrics in Texas and Illinois

Notice and consent before capturing or otherwise obtaining a scan of face geometry. In Illinois, consent must be written.

Very limited disclosure. In Illinois, prohibited to sell, lease, trade, or “otherwise profit” from an individual's biometric data. Otherwise, sharing permitted only with the data subject’s consent or other narrow exceptions. Texas prohibits the disclosure of biometric data to third parties, even with the data subject’s consent, unless the disclosure is related to the data subject’s disappearance or death or one of the other permitted purposes above.

Retention and Security. Generally, data must only be kept for as long as necessary for the purpose collected (other periods may apply if shorter, e.g., 3 years from last interaction with company). Illinois requires a publicly available retention schedule. Data must be secured like company confidential information.
ILLINOIS
Private right of action, class actions allows, up to $5,000 in liquidated damages per violation

TEXAS
No private right of action, enforceable only by the Attorney General, up to $25,000 in civil penalties per violation

WASHINGTON
No private right of action, enforceable only by the Attorney General under the Consumer Protection Act
Wave of biometrics litigation in Illinois began in 2015 and continues today.

- Fingerprint collection, largely in the employment context
- Facial recognition related to companies’ core products
  - Rivera v. Google
  - In re Facebook Biometrics
  - Monroy v. Shutterfly

No cases yet involving:
- “voiceprints”, iris/retina scans
- Enterprise use case
- “Research” use case
Extraterritorial Application?

- Generally, states cannot regulate outside their borders

- How to apply this concept with respect to biometric data derived from photos remains unresolved in ongoing litigation

- Where biometrics derived from photos, likely a multi-factored analysis
RISKS

Using data to train algorithms or provide an ML product or service involves risk.

UNDERSTAND YOUR RIGHTS

Minimize risk by understanding the rights you have to the underlying data. Understand that “public” data may not be free of risk.

TRACK YOUR DATA

Implement systems to inventory datasets and track which data drives which algorithms. Correcting a problem with a ML model will require correcting the data.

KNOW LAWS BY GEOGRAPHY

Avoid biometric litigation risk by considering geography, contractually allocating risk, implementing privacy controls, and being sensitive to potentially high-risk features like facial recognition.
Questions?
Email us at techlaw@perkinscoie.com.