SPARK+AI
SUMMIT EUROPE

# Great Models with Great Privacy

## Optimizing ML & AI Under GDPR

Sim Simeonov, CTO, Swoop
sim at swoop.com / @simeons

Slater Victoroff, CTO, Indico
slater at indico.io / @sl8rv

#SAISDD13

swoop

INTELLIGENCE BY IPM.ai

omni-channel marketing for your ideal population supported by privacy-preserving ML/AI

e.g., we improve health outcomes by increasing the diagnosis rate of rare diseases through doctor/patient education indico

Intelligent Process Automation for Unstructured Content using ML/AI with strong data protection guarantees

e.g. we automate the processing of loan documents using a combination of NLP and Computer Vision

# Does GDPR destroy data utility?

- GDPR Recital 26
  - "…personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable"
  - Not "likely to be no longer identifiable"

accuracy vs. privacy is a **false dichotomy**

(if you are willing to invest in privacy infrastructure)

# Privacy-preserving computation frontiers

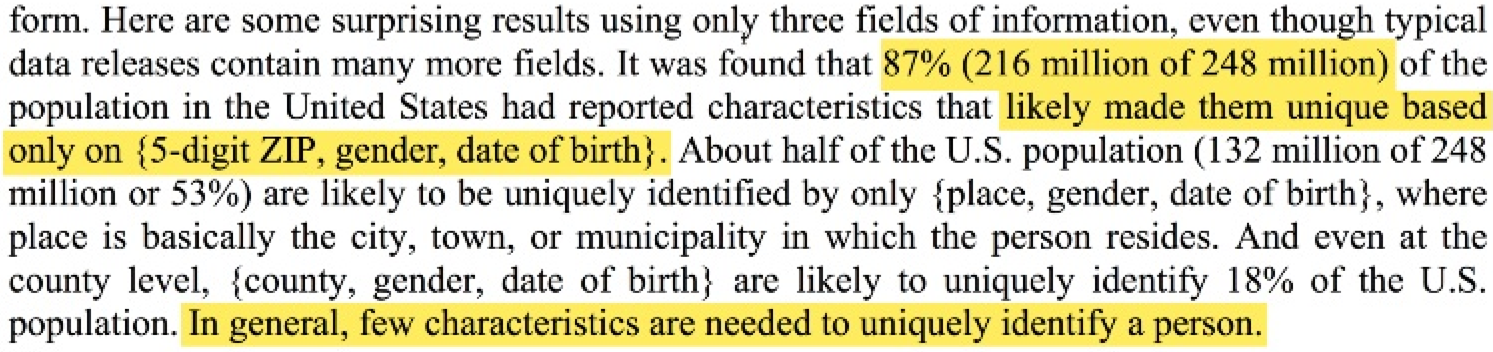
- Stochastic
  - Differential privacy (DP)
- Encryption-based
  - Fully homomorphic encryption (FHE)
- Protocol-based
  - Secure multi-party computation (SMC)

When privacy-preserving algorithms are immature,

sanitize the data the algorithms are trained on

# Sources of identifiability

- Direct (via personally-identifiable information)
- Indirect (via quasi-identifiers)



form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

- In structured or unstructured data

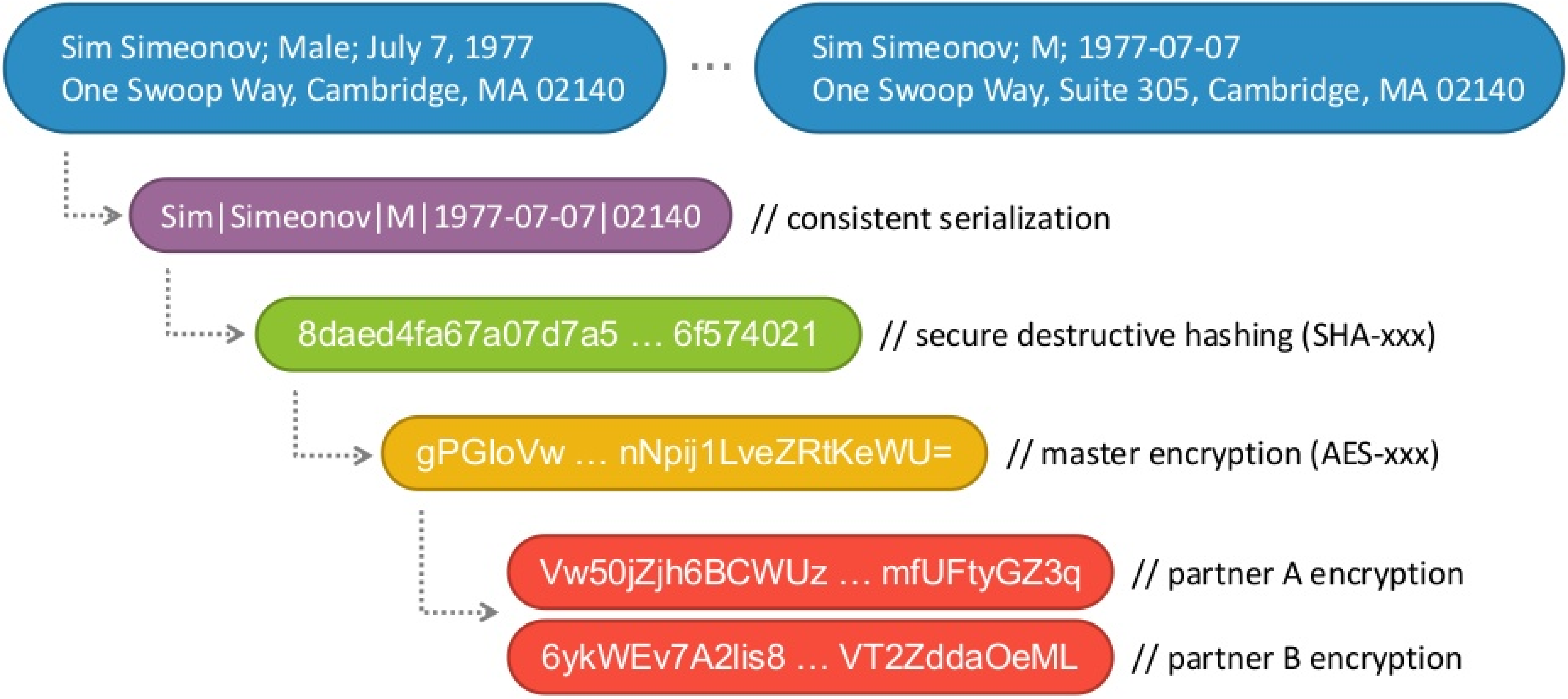
# Addressing identifiability

- Direct
  - Generate secure pseudonymous identifiers
  - Often uses clean room to process PII
- Indirect
  - Sanitize quasi-identifiers to desired privacy level
  - Control data enhancement to maintain anonymity

# Session roadmap

- The rest of this session
  - Building pseudonymous IDs with Spark
- After the break
  - Sanitize quasi-identifiers
  - Single vs. multiple dataset sanitization
  - Handling unstructured data via embeddings (cool!)

# Secure pseudonymous ID generation



Sim Simeonov; Male; July 7, 1977
One Swoop Way, Cambridge, MA 02140

…

Sim Simeonov; M; 1977-07-07
One Swoop Way, Suite 305, Cambridge, MA 02140

Sim|Simeonov|M|1977-07-07|02140 // consistent serialization

8daed4fa67a07d7a5 … 6f574021 // secure destructive hashing (SHA-xxx)

gPGloVw … nNpij1LveZRtKeWU= // master encryption (AES-xxx)

Vw50jZjh6BCWUz … mfUFtyGZ3q // partner A encryption

6ykWEv7A2lis8 … VT2ZddaOeML // partner B encryption

# Tip: generate multiple IDs

Sim|Simeonov|M|1977-07-07|02140 // full entry when data is clean

**S|S551**|M|1977-07-07|02140 // fuzzify names to handle limited entry & typos

Sim|Simeonov|M|**1977-07**|02140 // also may reduce dob/geo accuracy

tune fuzzification to use cases & desired FP/FN rates

# Building pseudonymous IDs with Spark

# Enjoy the break!

Come back for sanitizing

quasi-identifiers & unstructured data

# Sanitizing quasi-identifiers

- Deterministic
  - Generalize or suppress quasi-identifiers
  - *k*-anonymity + derivatives (a record maps to *k-1* others)
- Stochastic
  - Add noise to data
  - (*k*, Ɛ)-anonymity
- Domain-specific

# Sanitizing quasi-identifiers in Spark

- Optimal *k*-anonymity is an NP-hard problem
  - Mondrian algorithm: greedy O(nlogn) approximation
    - https://github.com/eubr-bigsea/k-anonymity-mondrian
- Active research
  - Locale-sensitive hashing (LSH) improvements
  - Risk-based approaches (e.g., LBS algorithm)

# Identifiability across datasets

- Centralized approach
  - Join all data + sanitize the whole
  - Big increase in dimensionality
- Federated approach
  - Keep data separate + sanitize operations across data
  - Smallest possible increase in dimensionality

# Centralized sanitization hurts ML/AI accuracy

We show that when the data contains a large number of attributes which may be considered quasi-identifiers, it becomes difficult to anonymize the data without an unacceptably high amount of information loss. ... we are faced with ... either completely suppressing most of the data or losing the desired level of anonymity.

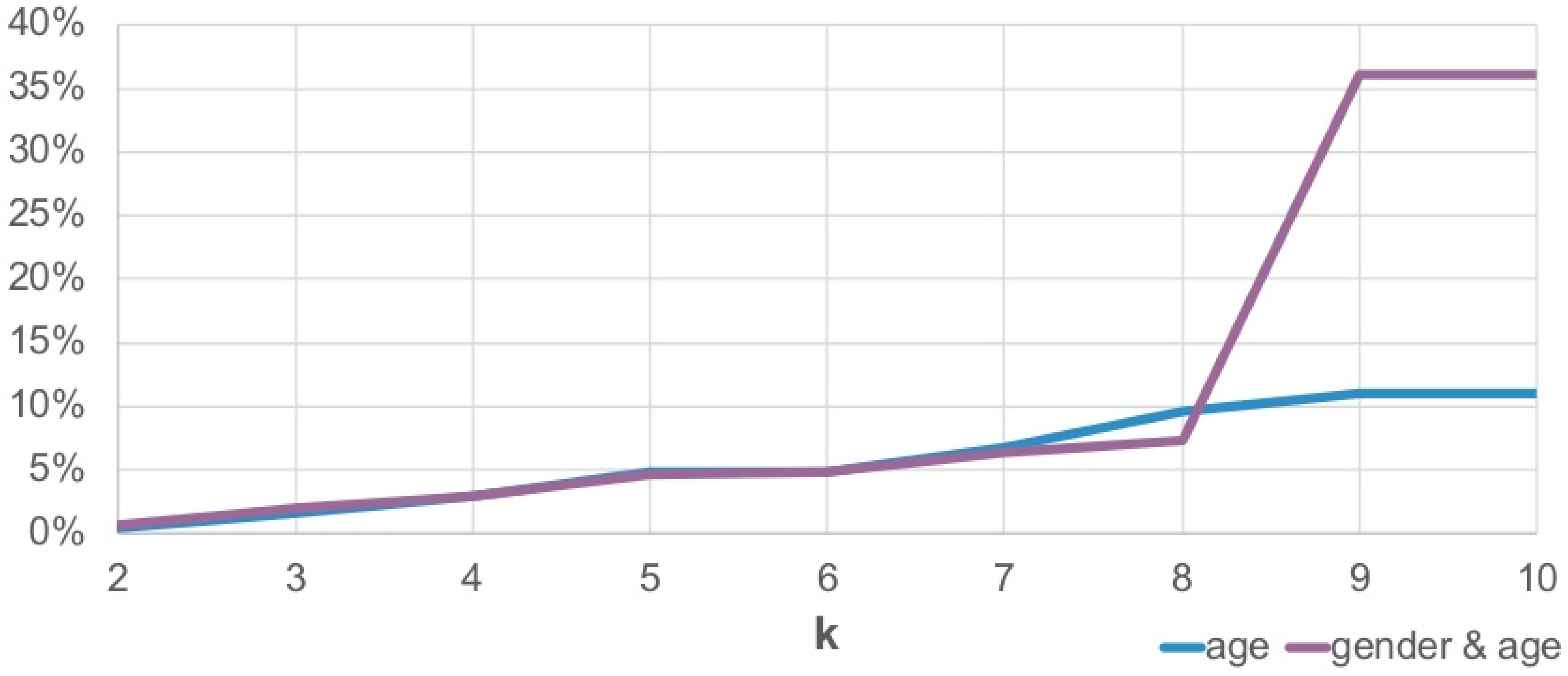
On *k*-Anonymity and the Curse of Dimensionality
2005 Aggarwal, C. @ IBM T. J. Watson Research Center

# Centralized sanitization increases risk

We find that for privacy budgets effective at preventing attacks, *patients would be exposed to increased risk of stroke, bleeding events, and mortality.*

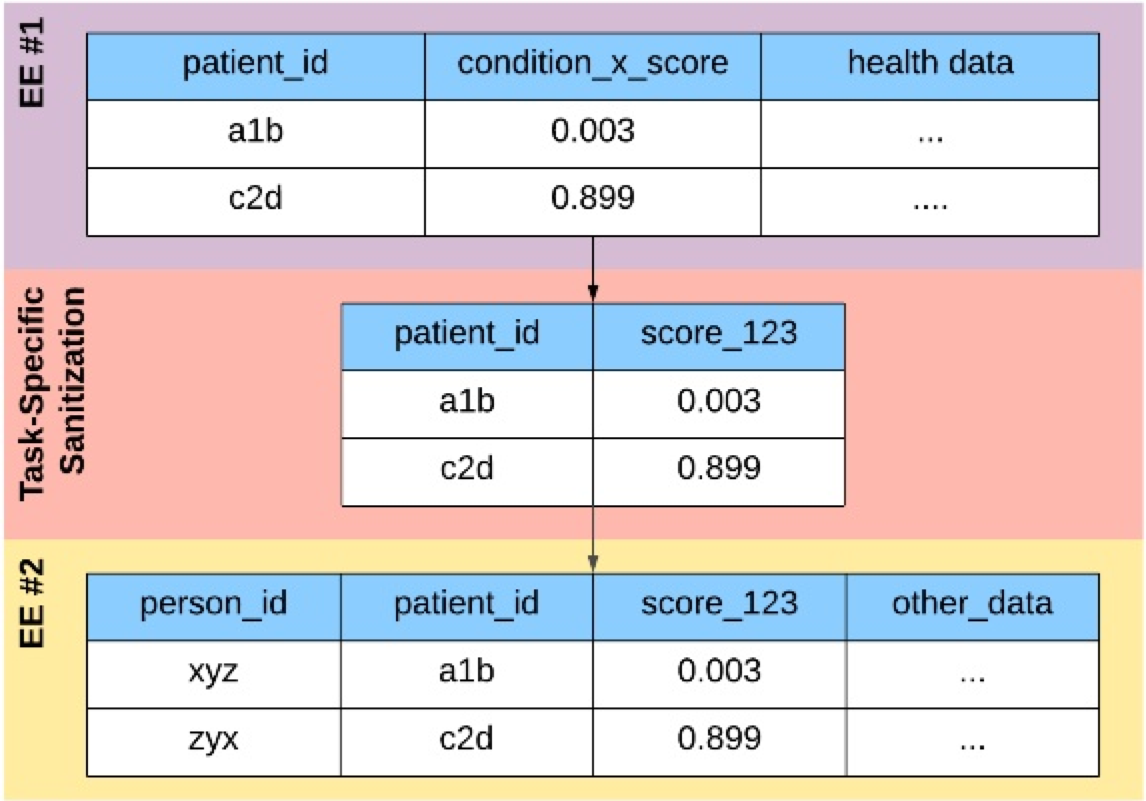
Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing
2014 Fredrikson, M. et. al. @ UW Madison and Marshfield Clinic Research Foundation

# Normalized Certainty Penalty (NCP)



*k*-anonymizing Titanic passenger survivability

# Federated sanitization (Swoop’s prAIvacy™)



EE #1

| patient_id | condition_x_score | health data |
|---|---|---|
| a1b | 0.003 | ... |
| c2d | 0.899 | .... |

Task-Specific Sanitization

| patient_id | score_123 |
|---|---|
| a1b | 0.003 |
| c2d | 0.899 |

EE #2

| person_id | patient_id | score_123 | other_data |
|---|---|---|---|
| xyz | a1b | 0.003 | ... |
| zyx | c2d | 0.899 | ... |

- Secure, isolated data pools
- Automated sanitization
- Min dimensionality growth
- Deterministic + stochastic
- Optimal + often lossless

Model condition X score on other data

no anonymization framework for unstructured data:

suppress or structure

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
PII

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
PII

***Solution(s)***

- Remove common names?
- Tell Doctors to stop using names in their notes?
- Lookup patient information in notes and intentional remove

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Quasi-Identifiers

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Quasi-Identifiers

***Solution(s)***

- Remove all locations?
- Remove all gendered pronouns?
- Remove all numbers?

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Weak Identifiers

# The Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Weak Identifiers

***Solution(s)***

- Remove all text

# The Real Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Predictive/Clinical Value

# The Real Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Predictive/Clinical Value

***Accuracy Matters***

# The Real Problem With Text

John Malkovitch plays tennis in Winchester. He has been reporting soreness in his elbow. His 60th birthday is in two weeks. After he returns from his birthday trip to Casablanca we will recommend a steroid shot to reduce inflammation.

***Problem***
Predictive/Clinical Value

***Accuracy Matters***

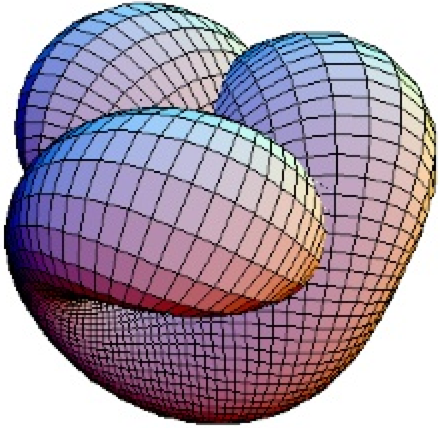
***Solution(s)***

- Embeddings

# What is an Embedding?

***Text Space***
*(e.g. English)*

***Embedding Method***
*(e.g. Word2Vec)*



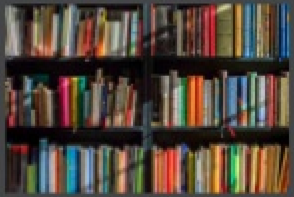
***Embedding Space***
*(e.g. $R^{300}$)*

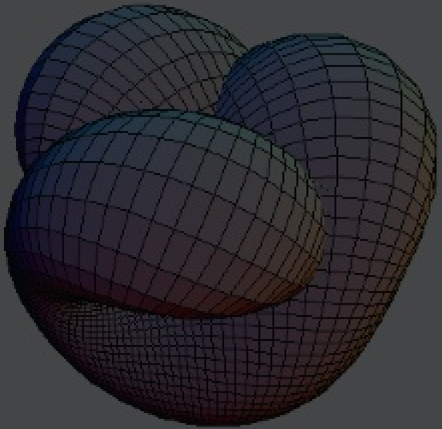
0.1
0.2
0.8
0.1
0.3
0.6
0.8
0.3
…

# What is an Embedding?

***Text Space***
*(e.g. English)*

***Embedding Method***
*(e.g. Word2Vec)*





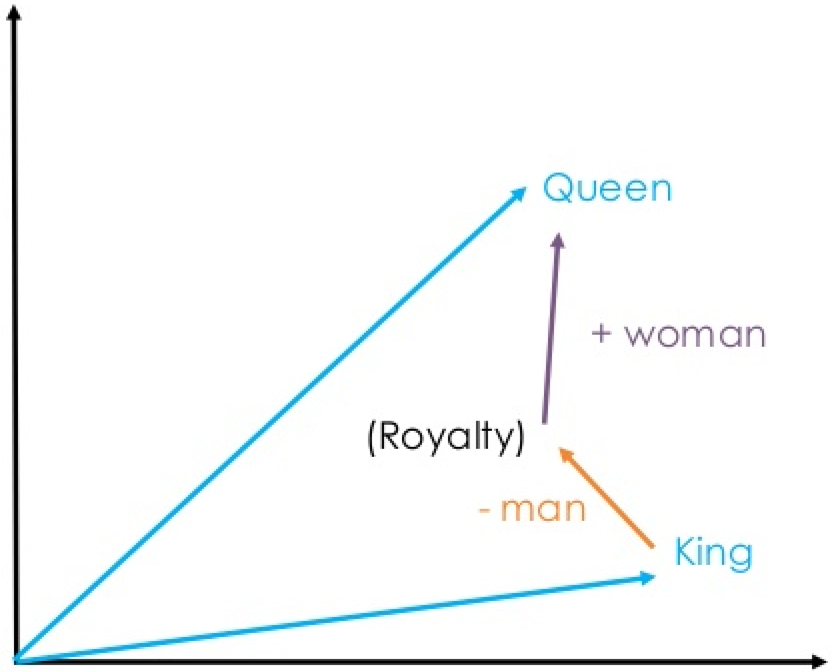
***Embedding Space***
*(e.g. $R^{300}$)*

0.1
0.2
0.8
0.1
0.3
0.6
0.8
0.3
…

# What is an Embedding?

- Text Space
  - Full Algorithmic Utility
  - Limited to No Guarantees on Privacy
  - Sparse, Inherently Brittle
- Embedding Space
  - Very High Algorithmic Utility
  - Strong Privacy Guarantees
  - Dense, Generically More Generalizable

# How do Embeddings Work?



- Meaning is “encoded” into the embedding space
- Individual dimensions are not human interpretable
- Embedding method learns by examining large corpora of generic language
- Goal is accurate language representation as a proxy for downstream performance

# “Word” Embeddings

***Examples***

Word2vec, GloVe,

***In Practice***

| Token | Value |
|---|---|
| “great” | [0.1, 0.3, …] |
| … | … |

***Training***

***CBOW***

The quick brown fox _____ over the lazy dog

***Skip Gram***

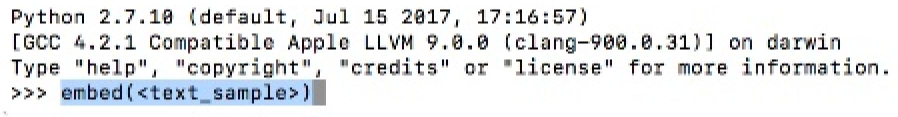
___ ___ ____ ___ jumps ___ __ ___ ___

# Text Embeddings

***Examples***

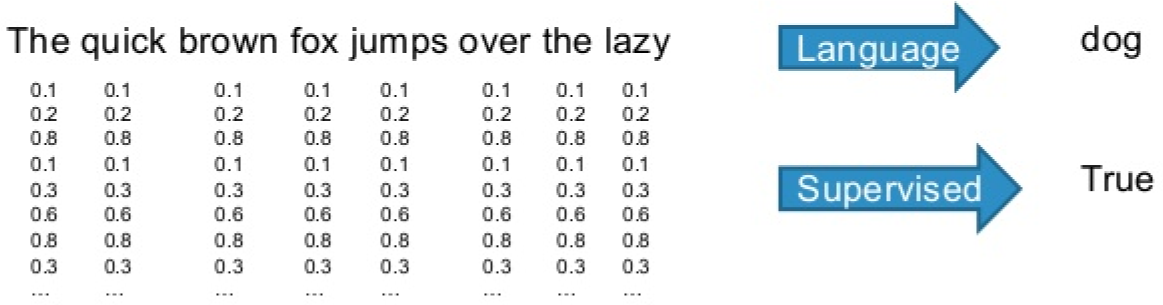
doc2vec, Elmo, ULMFiT

***In Practice***



```
Python 2.7.10 (default, Jul 15 2017, 17:16:57)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.31)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> embed(<text_sample>)
```

***Training***

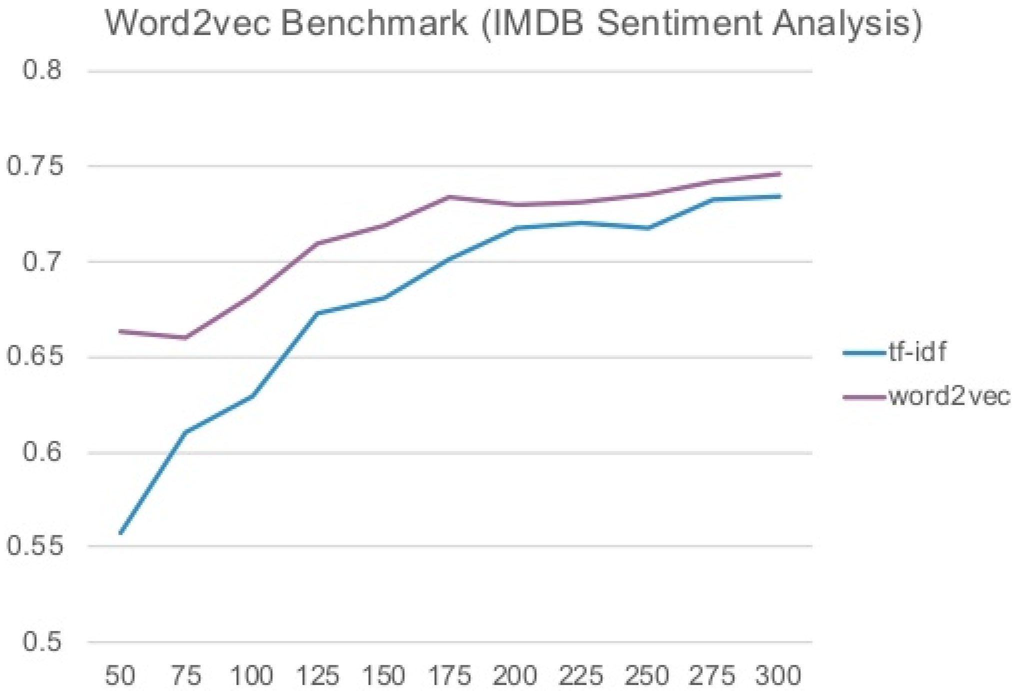


| The | quick | brown | fox | jumps | over | the | lazy |
|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 |
| 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| ... | ... | ... | ... | ... | ... | ... | ... |

# Additional Notes

- “Word” Embeddings
  - May not correspond directly to words
  - Many excellent public implementations
  - Vulnerable to rainbow-table attacks (Do not store)
- Text Embeddings
  - Often built on top of word embeddings
  - K-anonymity can be directly assessed
  - Naively implemented as mean of word embeddings

# Do They Really Preserve Algorithmic Value?



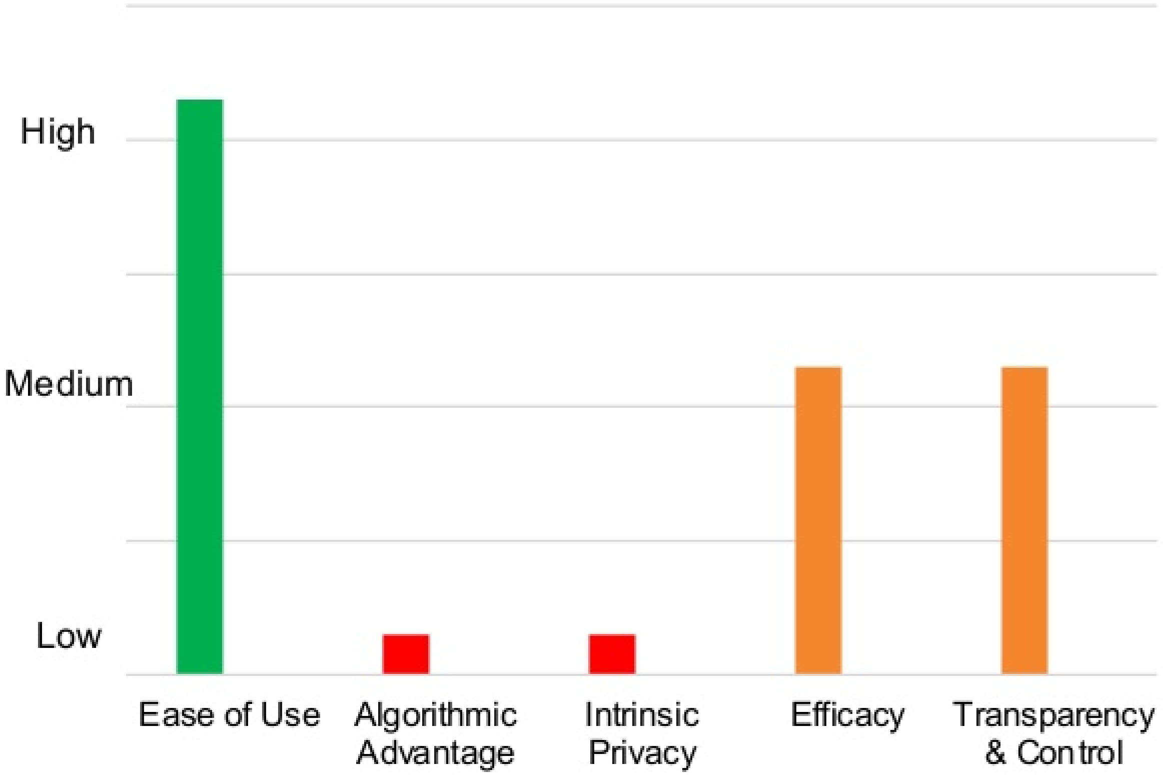
Reported numbers are the average of 5 runs of randomly sampled test/train splits each reporting the average of a 5-fold cv, within which Logistic Regression hyperparameters are optimized. Generated using Enso

- Embeddings generally outperform raw text at low data volumes
- Leveraging large, generic text corpora improves generalizability
- This is 6 year old tech. Embeddings have improved drastically. Text has not.

# Embedding Options

- Public
  - Ready-to-use downloadable word vectors
  - Trained on public datasets (e.g. word2vec trained on CommonCrawl)
- Homegrown
  - Embedding technique trained in-house
  - Typically a straightforward algorithm on a novel data corpus
- Third Party
  - Sourced from vendor responsible for updating data sources and algorithms

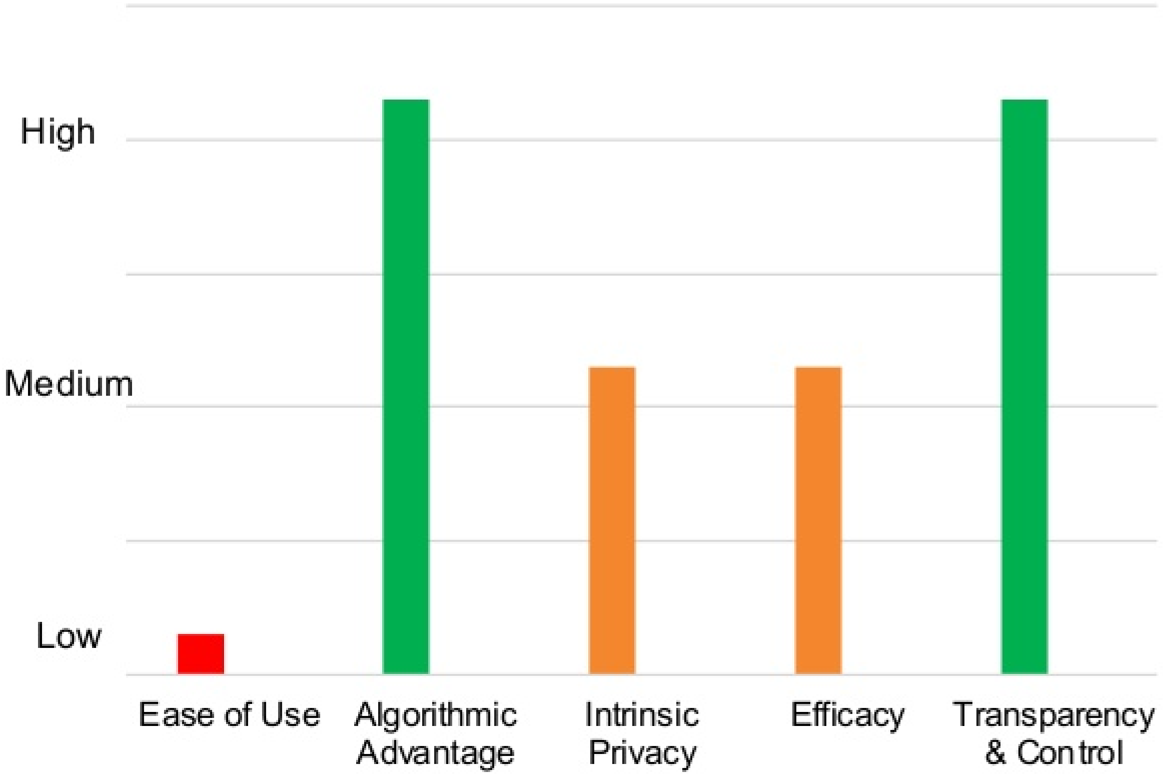
# Embedding Options (Public)



***Notes***

- Public embeddings effectively mean public rainbow tables
- Licensing is often ambiguous or incompatible with commercial use

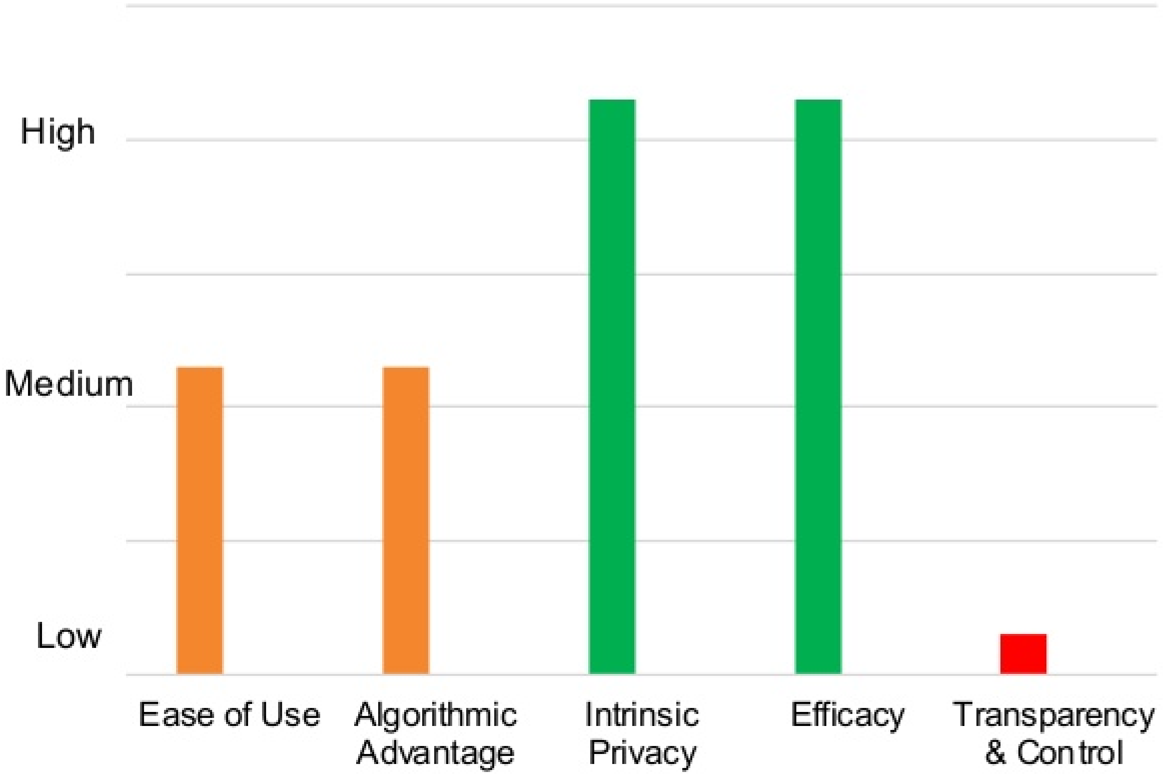
# Embedding Options (Homegrown)



***Notes***

- Deceptively easy to start, extremely difficult to effectively use in production
- While this can theoretically be a good approach, for most organizations this is a bad idea.

# Embedding Options (Third party)



***Notes***

- Data leakage can result in handing away competitive edge
- Efficacy highly variable. Ensure robust benchmarks before working with any third party
- Privacy-controlled environment makes involving a third party contractually difficult

# State of Embeddings in Spark

- Training yourself
  - Reasonable word2vec implementations
  - Proper evaluation requires two nested CV loops + a private holdout at a minimum
- Basic NLP Functionality
  - John Snow NLP Library (https://nlp.johnsnowlabs.com/)
- Pretrained Embeddings
  - Compute offline in python (Gensim, Spacy, indico)

swoop

INTELLIGENCE BY IPM.ai

indico

Privacy matters. Thank you for caring.

Interested in challenging data engineering, ML & AI on big data?
We'd love to hear from you. sim at swoop.com / slater at indico.io