

# German wings challenge

Kai Chen

Github repo: <https://github.com/ck-unifr/german-wings-airline-challenge>

# Resume

- Oct 2018 - present, Data Science Manager, Unitymedia, Germany
- May 2017 - Oct 2018, Senior Data Scientist, AGT International, Germany
- Sep 2012 - Jun 2017, PhD in CS, University of Fribourg, Switzerland
- Jun 2015 - Sep 2015, Visiting PhD, Chinese Academy of Sciences, China
- Sep 2009 - Sep 2012, Master in CS, University of Fribourg, Switzerland
- Oct 2005 - Feb 2008, Bachelor in CS, University of Applied Science, Switzerland

Github: [github.com/ck-unifr](https://github.com/ck-unifr)

Google scholar: [kai.chen.unifr](https://scholar.google.com/citations?user=kai.chen.unifr)

Linkedin: [linkedin.com/in/kai-chen-29503288/](https://www.linkedin.com/in/kai-chen-29503288/)

# Use Cases

- Case 1: Sentiment Analysis
  - Predict a review is positive or negative
- Case 2: Topic Modeling
  - Find topics in the reviews

# Methodology

- **Data Loading**
  - Load data from a txt file and save the data into a pandas dataframe
- **EDA (Exploratory Data Analysis)**
  - Plot distribution of variables
  - Show relationship between the variables
  - Text analysis: Plot word frequencies, Topic modelling with LDA
- **Feature Engineering**
  - TFIDF (term frequency–inverse document frequency)
  - Count Features
  - Dimensionality reduction using truncated SVD
  - Word Embedding: GloVe
- **Evaluation Metrics**
  - Cross-Entropy Loss
  - Precision and Recall
  - ROC (Receiver Operating Characteristics)
- **Modelling**
  - Logistic Regression, Naive Bayes, Gradient Boost Machine, Deep Learning
- **Error Analysis**

# Notebook

# Word Embedding

The slides are taken from, *Sequence Models, Deep Learning Specialization, Coursera* by Andrew Ng.

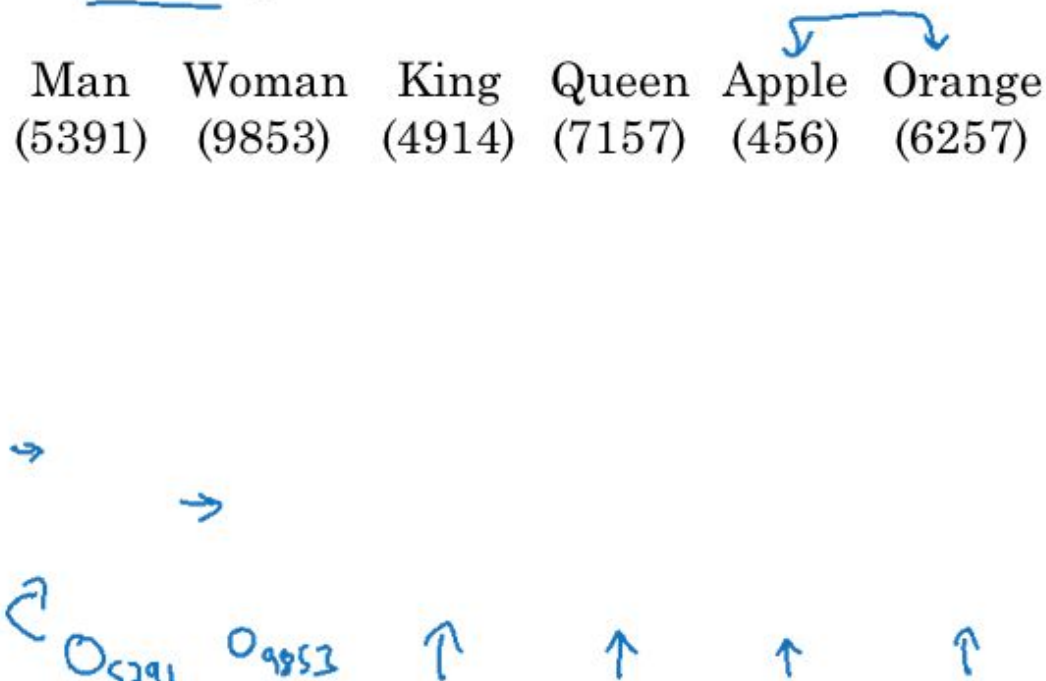
# Word representation

$V = [a, aaron, \dots, zulu, <UNK>]$

$|V| = 10,000$

1-hot representation

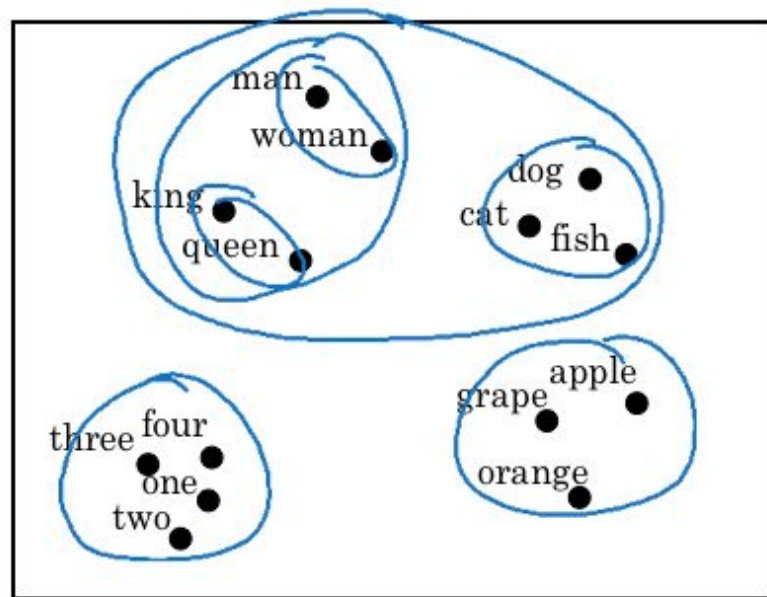
Man	Woman	King	Queen	Apple	Orange
(5391)	(9853)	(4914)	(7157)	(456)	(6257)



I want a glass of orange juice.

I want a glass of apple ?.

# Visualizing word embeddings

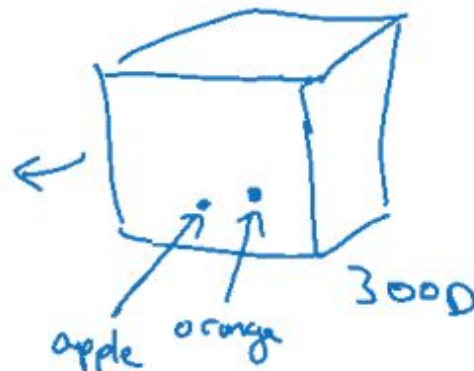


t-SNE

→ 3000



2D





# Featurized representation: word embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender <sup>←</sup>	-1	1	-0.95	0.97	0.00	0.01
Royal <sup>←</sup>	0.01	0.02	<u>0.93</u>	<u>0.95</u>	-0.01	0.00
Age <sup>←</sup>	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
size cost alt verb	⋮	⋮				

I want a glass of orange juice.

I want a glass of apple juice.

Andrew Ng

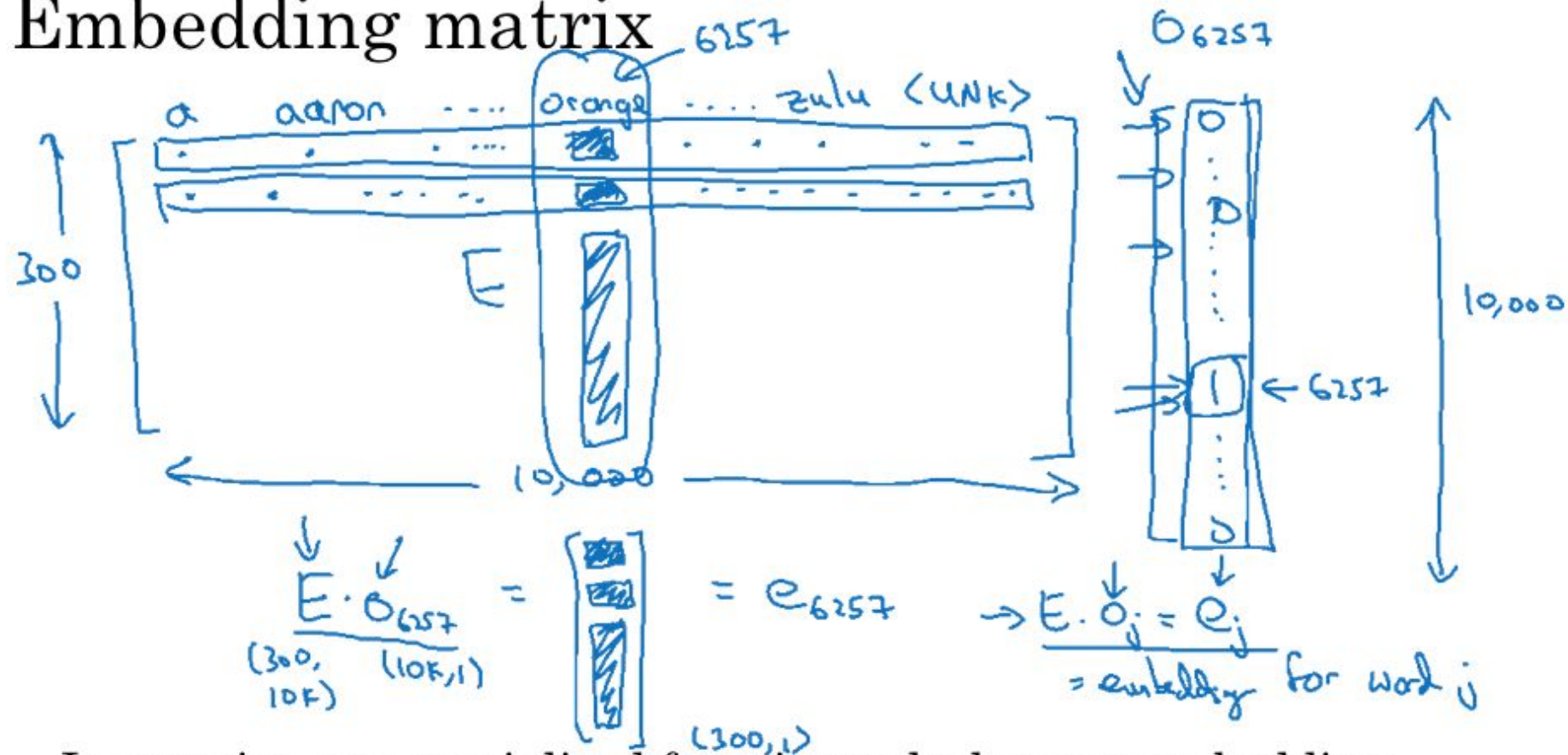
# Analogy

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

$\underbrace{e_{5391}}_{e_{\text{man}}} \rightarrow \underbrace{e_{9853}}_{e_{\text{woman}}} \quad \propto \quad \underbrace{e_{4914}}_{e_{\text{king}}} \rightarrow ? \quad \underbrace{e_{7157}}_{e_{\text{queen}}}$   
 $e_{\text{man}} - e_{\text{woman}} \approx e_{\text{king}} - e_{\text{queen}}$

$e_{\text{man}} - e_{\text{woman}} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$   
 $e_{\text{king}} - e_{\text{queen}} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

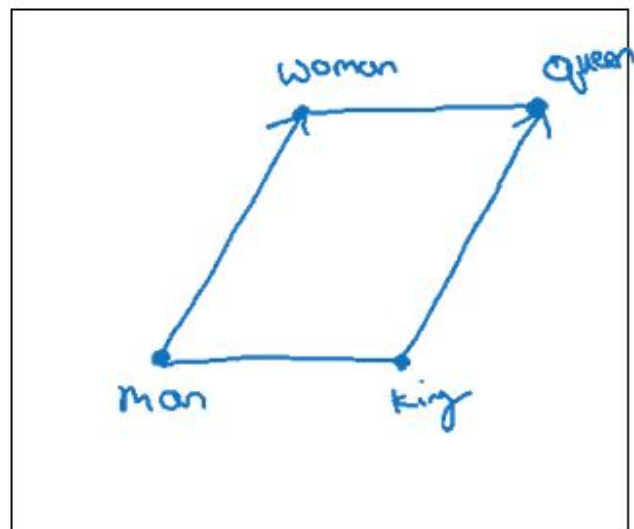
# Embedding matrix



In practice, use specialized function to look up an embedding.

$\rightarrow \text{Embedding}$

# Analogies using word vectors

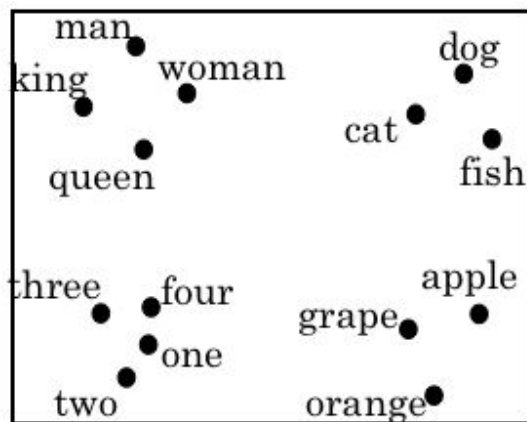


300 D

Find word  $w$ :  $\arg \max_w$

$$\text{Sim} \left[ \underset{\uparrow}{e_w}, \underbrace{e_{\text{king}} - e_{\text{man}} + e_{\text{woman}}}_{30-75\%} \right]$$

$3000 \rightarrow 20$   
 $\uparrow$



t-SNE

$$e_{\text{man}} - e_{\text{woman}} \approx e_{\text{king}} - e_?$$

$e_w$

# GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

$c, t$

$X_{ij}$  = # times  $i$  appears in context of  $j$ .

$\begin{matrix} \uparrow & \uparrow \\ c & t \end{matrix}$        $\begin{matrix} \uparrow \\ t \end{matrix}$        $\begin{matrix} \uparrow \\ c \end{matrix}$

$X_{ij} = X_{ji} \leftarrow$



# Model

Minimize

$$\sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(x_{ij}) \left( \underbrace{\Theta_i^T e_j}_{\substack{t \quad c \\ \text{"}\Theta_t^T e_c\text{"}}} + b_i + b_j' - \log x_{ij} \right)^2 \leftarrow$$

0?

weighting  
term

$f(x_{ij}) = 0$  at  $x_{ij} = 0$ .

" $0 \log 0$ " = 0

→ this is, at, a, ...  
→ derivation

$\Theta_i, e_j$  are symmetric

$$e_w^{(\text{final})} = \frac{e_w + \Theta_w}{2}$$

# Work Experience

# Work at Unitymedia

- Recommendation System
- Churn Prevention
- ETL in Hadoop

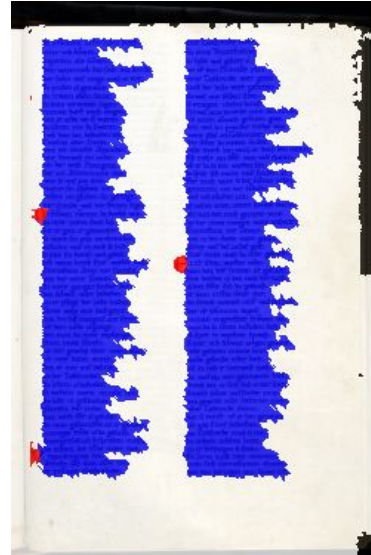
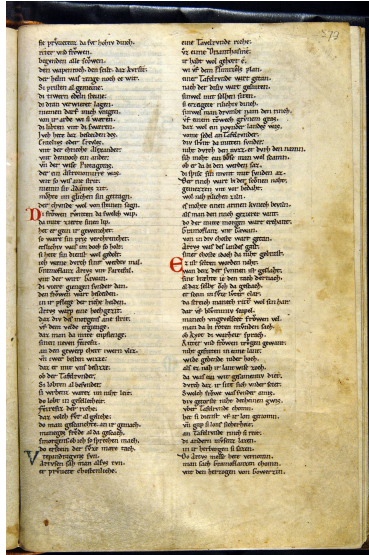


# Work at AGT International

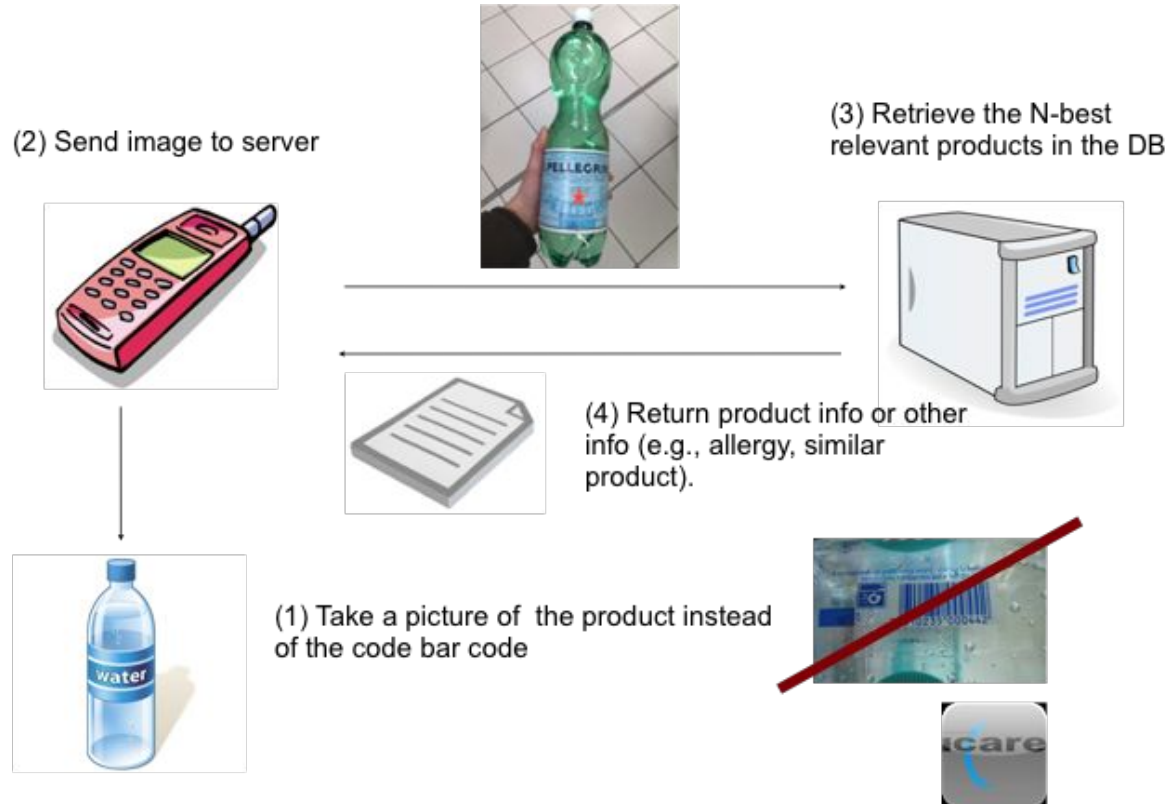
- Punch Recognition with Deep Learning
- A/B Testing for Punch Recognition Evaluation
- Anomaly Detection

# PhD thesis: Historical Document Layout Analysis with Machine Learning

- Goal
  - Developing a general page segmentation method with minimal prior knowledge.
- Basic Idea
  - Page Segmentation ➔ Pixel Labeling



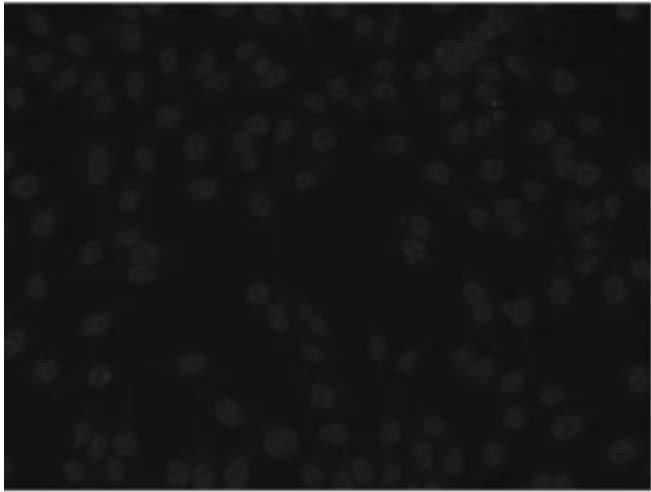
# Camera-based image retrieval (master project)



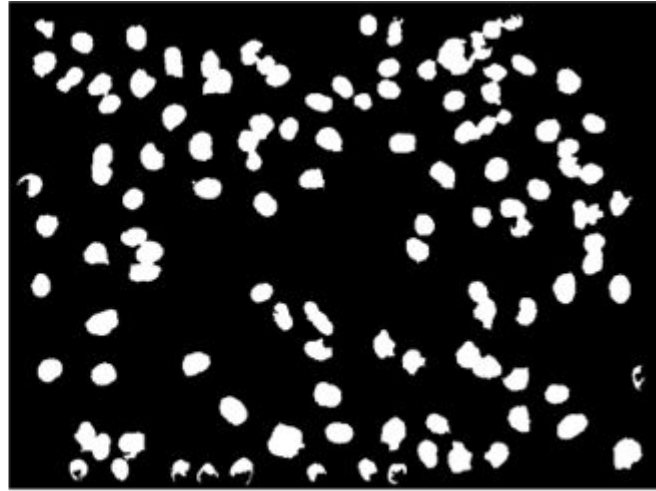
# Text line detection in historical document



# Cell image segmentation



input



output

# Other Projects

- Structured Data
  - Booking prediction
  - Revenue per click prediction
  - Consumer shopping prediction
  - Car price prediction
- Computer Vision (CV)
  - Product category classification
- Natural Language Processing (NLP)
  - Toxic Comment Classification (Kaggle challenge, top 16%)
- CV + NLP
  - Product description generation

Why I want to join zeroG?