COMP GI13/M050
Deep Learning Lecture 2

Thore Graepel & Guest Lecturers from DeepMind

Thanks to Mark Herbster for providing some of the revision material

Administration

- Assignment deadlines are extended by 1 week, **new deadlines:**
  - Assignment 1: “TensorFlow and MNIST”, due 7th February 2017
  - Assignment 2: “Sequence Generation”, due 7th March 2017
  - Assignment 3: “Deep RL”, due 4th April 2017

- Office hour Tuesday 4- 5pm @ GS 66-72, First floor Hub room until week of last lecture

Overview

- Review of concepts from supervised learning
  - Generalisation, overfitting, Underfitting
  - Learning curves
  - Stochastic gradient descent
- Linear regression
  - Cost function
  - Gradients
- Logistic regression
  - Cost function
  - Gradients
- Case Study: Using Regression to Predict Private Attributes from Facebook Likes

Supervised Learning Problem

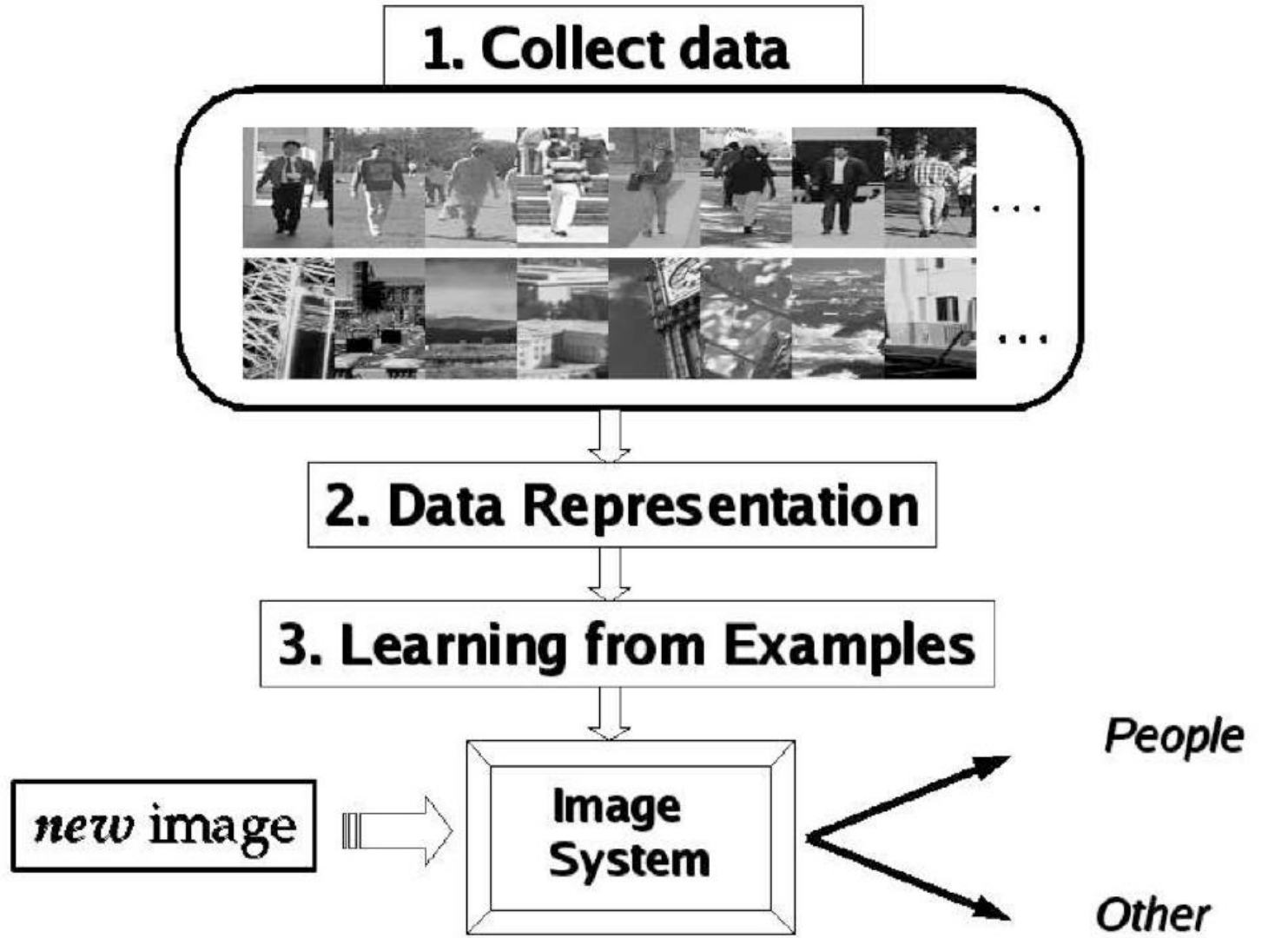
Given a set of **input/output** pairs (**training set**) we wish to compute the functional relationship between the input and the output

Example 1: (people detection) given an image we wish to say if it depicts a person or not. The output is one of two possible categories

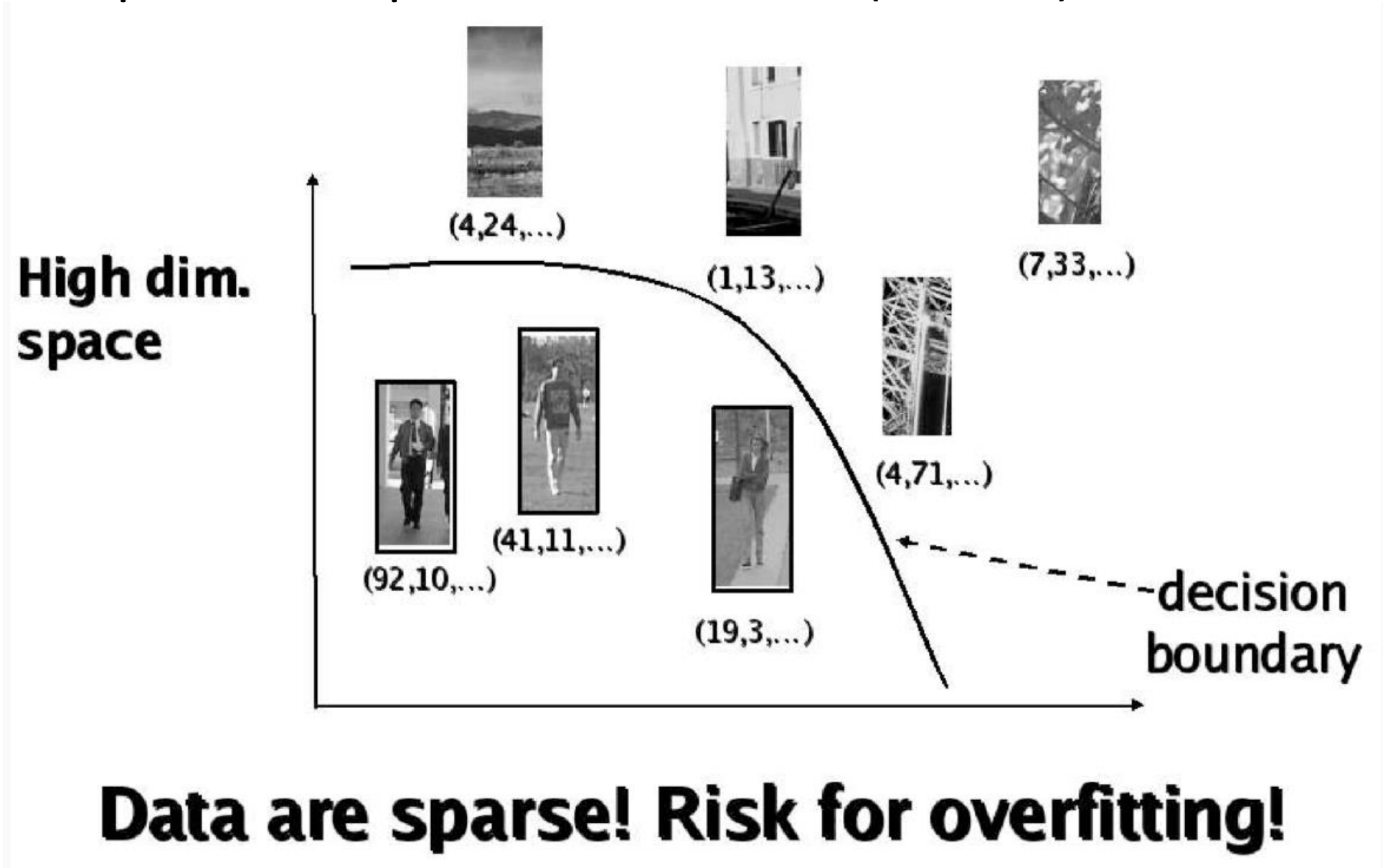
Example 2: (pose estimation) we wish to predict the pose of a face image The output is a continuous number (here a real number describing the face rotation angle)

In both problems the input is a high dimensional vector **x** representing pixel intensity/colour

Example: People Detection



Example: People Detection (cont.)



Supervised Learning Model

- Goal: Given training data (pattern,target) pairs

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$$

infer a function f_S such that

$$f_S(\boldsymbol{x}_i) \approx y_i$$

for the future data

$$S' = \{(\boldsymbol{x}_{m+1}, y_{m+1}), (\boldsymbol{x}_{m+2}, y_{m+2}), \ldots\}.$$

- Classification : $y \in \{-1, +1\}$; Regression : $y \in \mathbb{R}$
- $\mathcal{X}$: input space (eg, $\mathcal{X} \subseteq \mathbb{R}^d$), with elements $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{x}_i, \ldots$
- $\mathcal{Y}$: output space, with elements $y, y', y_i, \ldots$

Supervised Learning Problem: Compute a function which best describes I/O relationship

Learning Algorithm

- Training set: $S = \{(\boldsymbol{x}_i, y_i)_{i=1}^m\} \subseteq \mathcal{X} \times \mathcal{Y}$
- A **learning algorithm** is a mapping $S \mapsto f_S$
- A new input $\boldsymbol{x}$ is predicted as $f_S(\boldsymbol{x})$

- Example Algorithms:
  - Linear Regression
  - Logistic Regression
  - Neural Networks
  - Decision Trees
- In this lecture, we will revise linear and logistic regression

Key Questions for the ML Practitioner

- How is the data **collected**? (need assumptions!)
- How do we **represent** the inputs? (may require pre-processing step)
- How **accurate** is the learnt function on new data (study of **generalization error**)?
- Many algorithms may exist for a task. How do we choose?
- How "**complex**" is a learning task? (computational complexity, sample complexity)

Important Challenges for ML

- New inputs **differ** from the ones in the training set (look up tables do not work!)
- Inputs are measured with **noise**
- Output is **not deterministically** obtained by the input
- Input is often **high dimensional** but some components/variables may be irrelevant
- How can we incorporate **prior knowledge**?

Generalisation

Most important idea of machine learning:

Train models such that they correctly predict on unseen data
(from the same distribution)

- Empirical risk minimization: Minimise error on training sample
- Validation: Hold out data for testing to obtain unbiased estimator

$$\underbrace{\text{Validation Error}}_{\text{What we care about}} = \underbrace{\text{Training Error}}_{\text{What we optimise}} + \underbrace{(\text{Validation Error} - \text{Training Error})}_{\text{GeneralisationError}}$$

- When data is scarce, can use cross-validation

Cross Validation

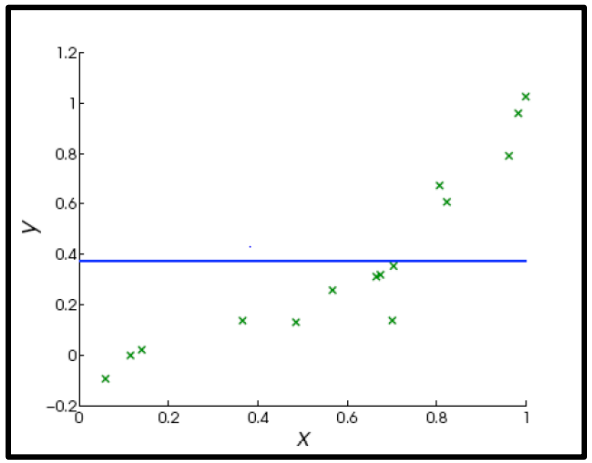
1. we split the data in K parts (of roughly equal sizes)
2. repeatedly train on $K - 1$ parts and test on the part "left out"
3. average the errors of K "validation" sets to give so-called cross-validation error
4. smaller K is less expensive but poorer estimate as size of training set is smaller and random fluctuations larger

For a dataset of size m, m -fold cross-validation is referred to as leave-one-out (LOO) testing

Underfitting and Overfitting

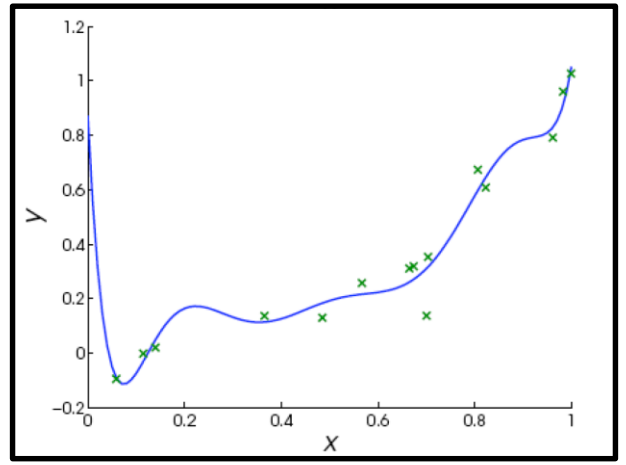
Underfitting

- Error driven by approximation
- High bias / low variance
- What to do?
  - Use more features
  - User more complex model
  - Reduce regularization
  - Train for longer



Overfitting

- Error driven by generalization
- Low bias / high variance
- What to do?
  - Use fewer features
  - Use simpler model
  - Increase regularization
  - Stop training early



More Data versus Better Algorithm

- In high-variance, overfitting situations more data helps
- Example: Confusion Set Disambiguation
- Banko and Brill 2001, “Scaling to Very Very Large Corpora for Natural Language Disambiguation”
- See also: “The Unreasonable Effectiveness of Data”, Pereira, Norvig, Halevy

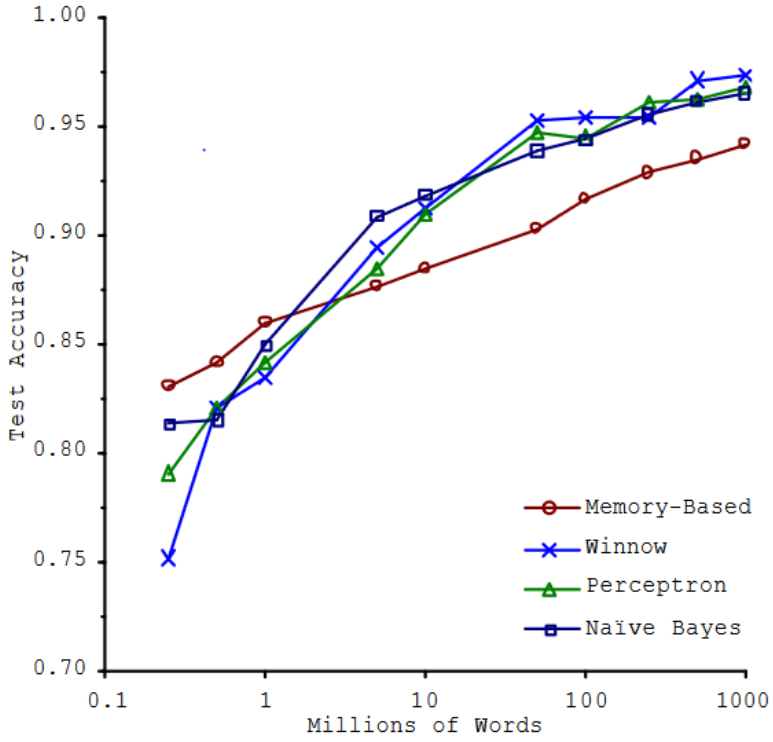
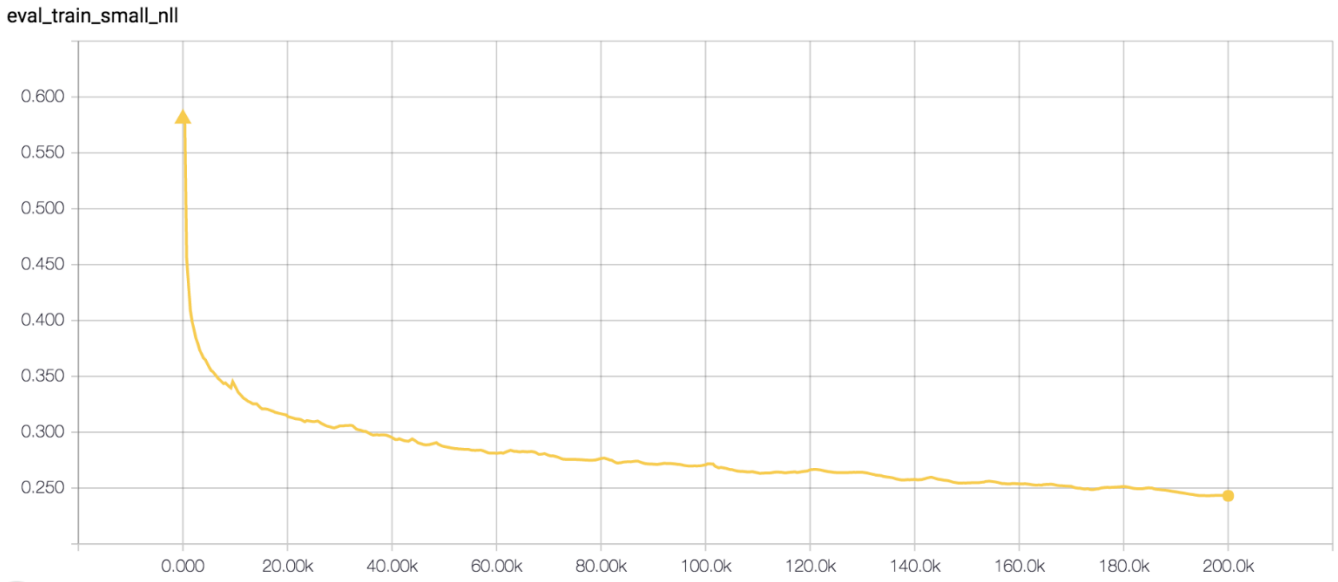


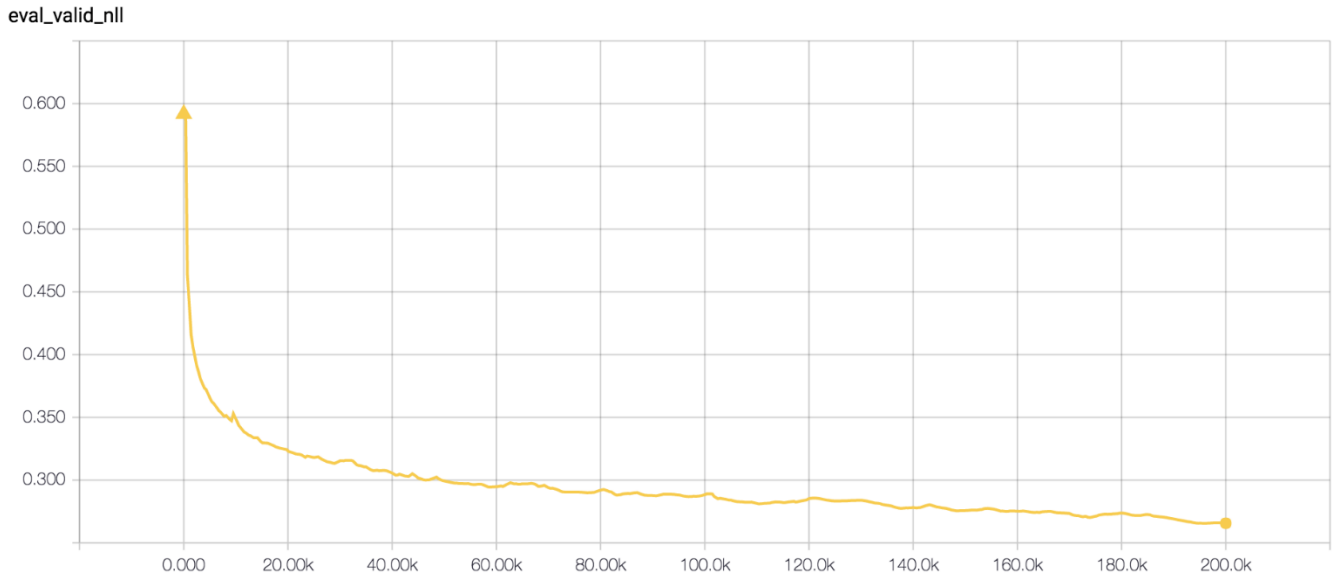
Figure 1. Learning Curves for Confusion Set Disambiguation

Real-World Learning Curves: Underfitting

Training Error

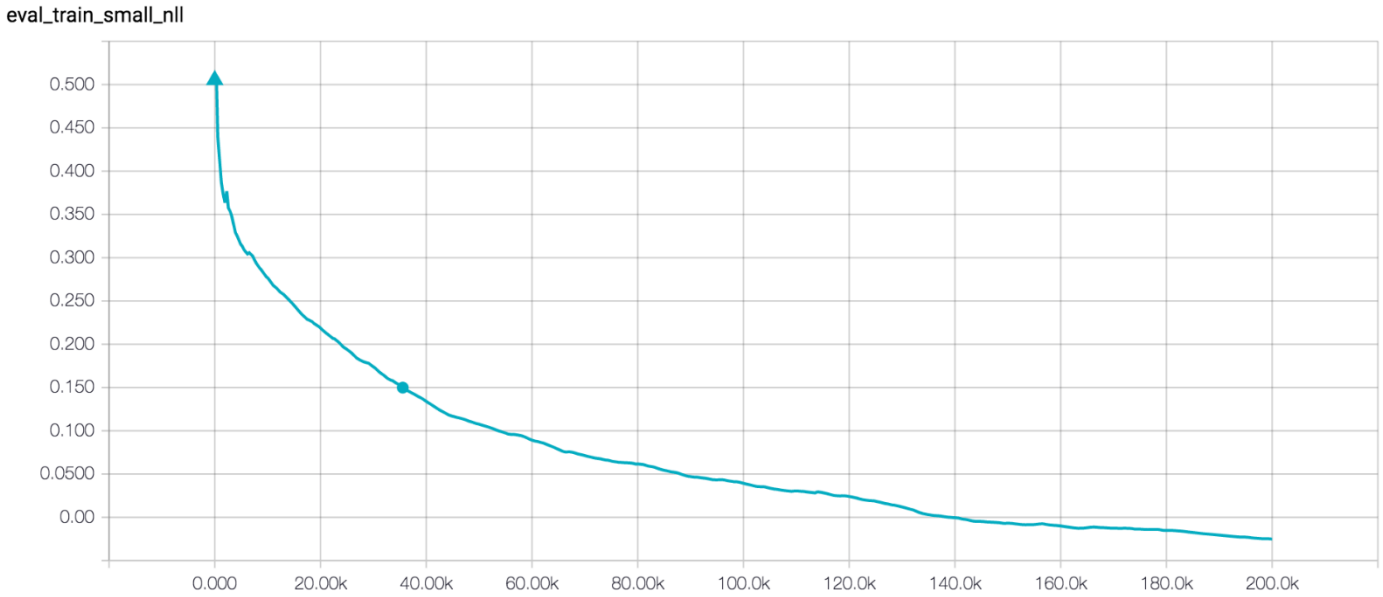


Validation Error

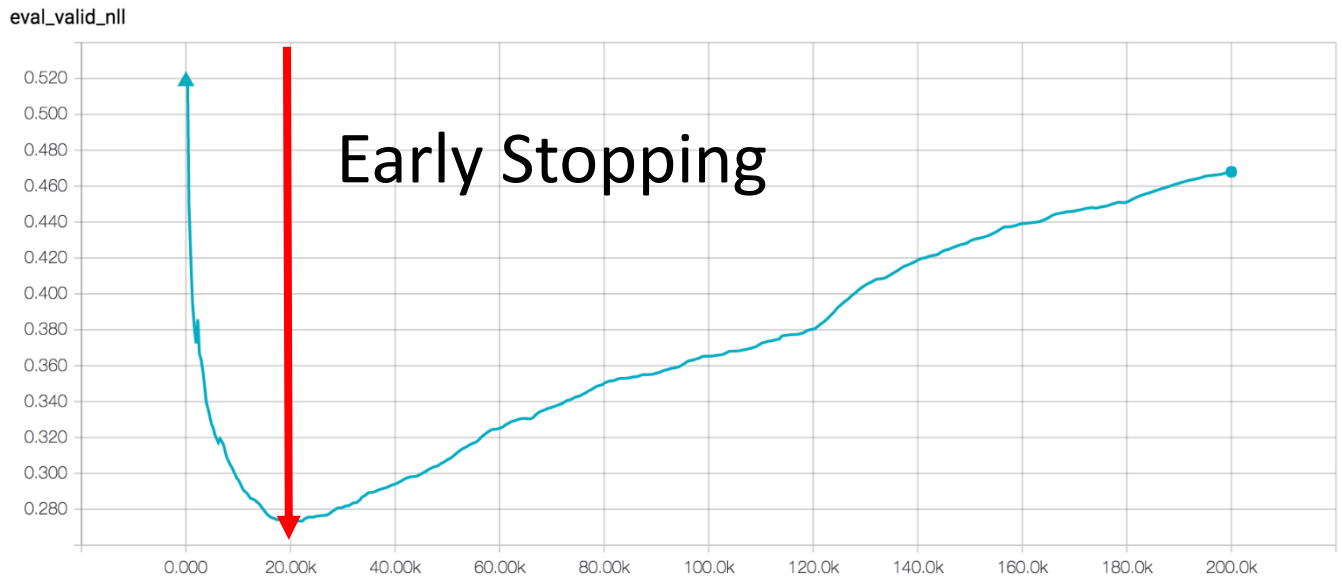


Real-World Learning Curves: Overfitting

Training Error

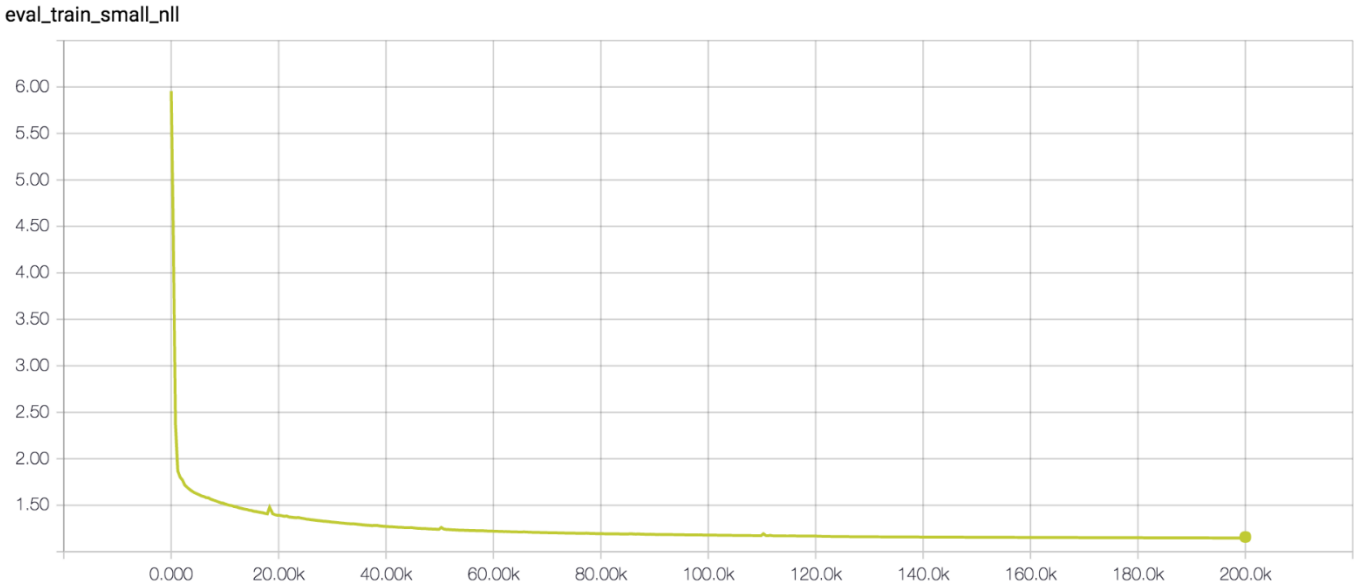


Validation Error

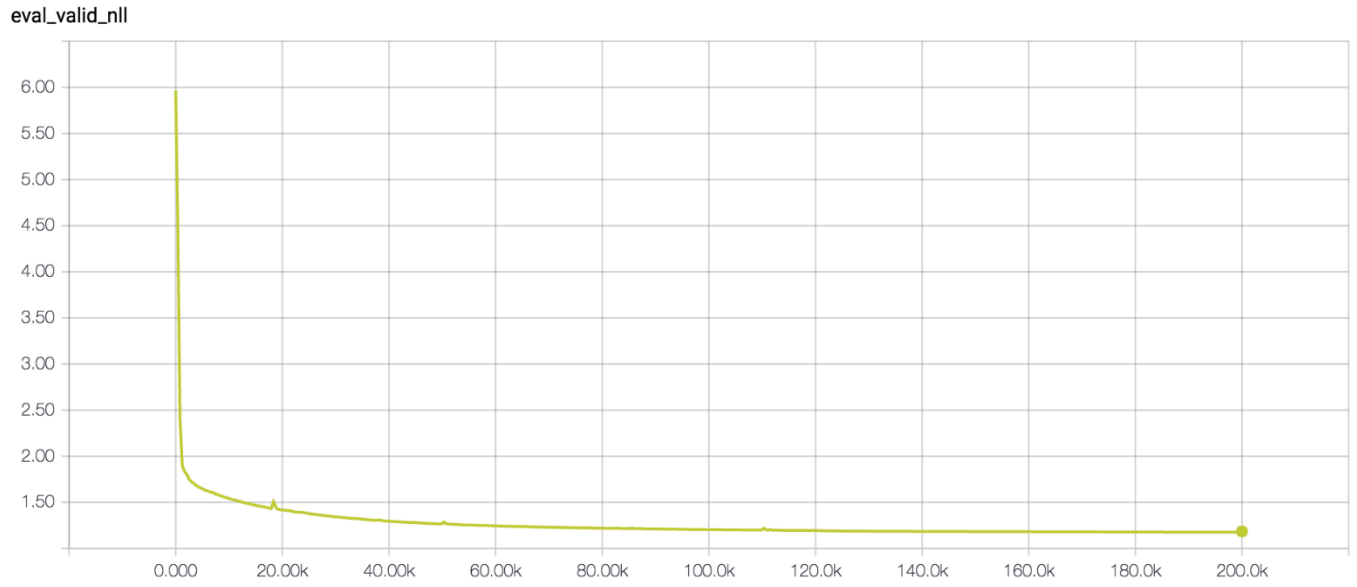


Real-World Learning Curves: Just Right

Training Error

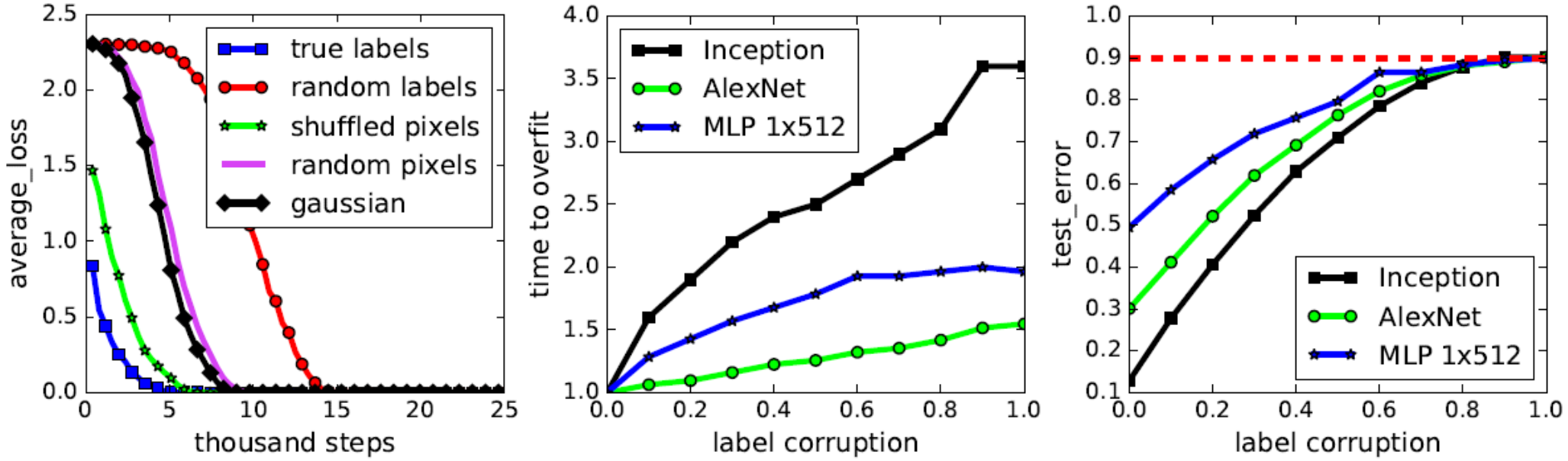


Validation Error



Generalisation in Deep Learning

- “Understanding Deep Learning requires rethinking generalization”, Zhang, S. Bengio, Hardt, Recht, Vinyals
- Deep Neural Networks easily fit random labels
- Generalization error varies from 0 to 90% without changes in model
- Deep NNs can even (rote) learn to classify random images



(a) learning curves (b) convergence slowdown (c) generalization error growth

(Stochastic) Gradient Descent

- For linear regression, we can find closed form solution using the (pseudo) matrix inverse (computationally expensive).
- With large data sets or more complex models, this may be impossible
- Batch gradient descent for loss $L(\mathbf{w})$ with learning rate η:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla L(\mathbf{w})$$

- Often better to use stochastic gradient descent for cost functions of the form $L(\boldsymbol{w}) = \sum_{i=1}^{m} f_i(\boldsymbol{w})$:
- For mini-batch S_j apply weight update: $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \sum_{i \in S_j} \nabla f_i(\mathbf{w})$
- Mini-batch size trades off computational cost and variance.
- Alternative optimization algorithms and analysis in James Martens' lecture

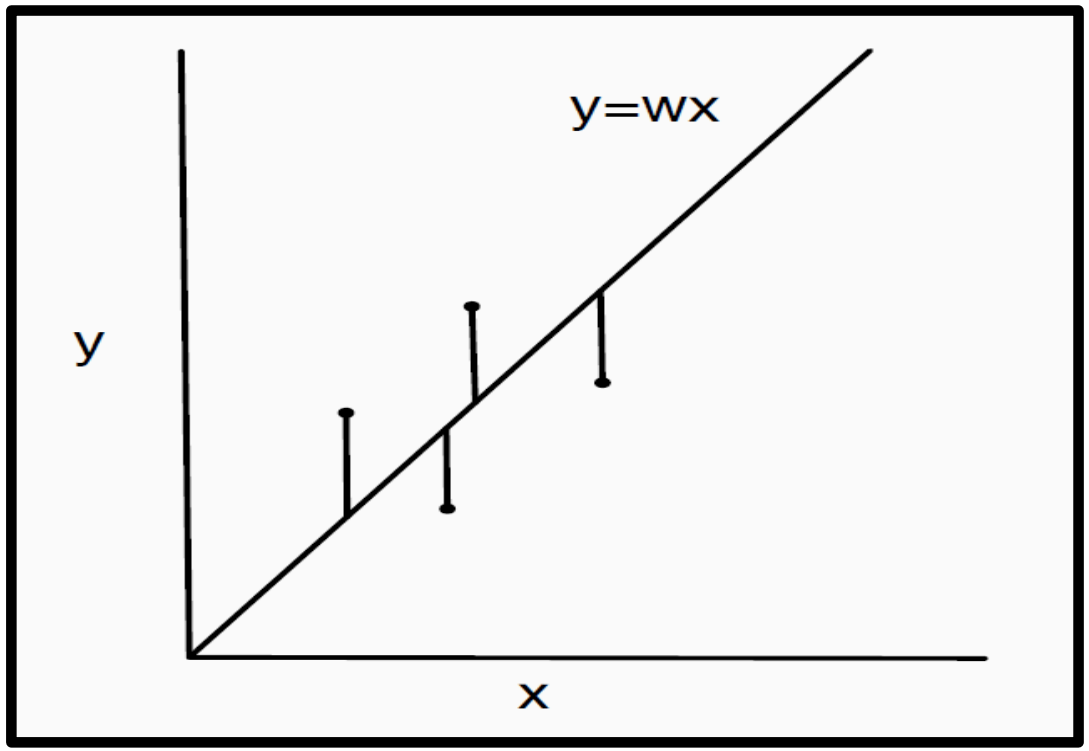
Generalisation from Stochastic Gradient Descent

- “Train faster, generalize better: Stability of Stochastic Gradient Descent”, Moritz Hardt, Benjamin Recht, Yoram Singer

Any model trained with stochastic gradient method in a reasonable amount of time attains small generalization error.

- A learning algorithm is called *stable* if it produces very similar results on two training samples S and S' differing only in one example
- If a learning algorithm is stable its learned model will exhibit good generalization.
- Under certain conditions stochastic gradient descent is stable, hence, the resulting model exhibits good generalisation
- The fewer steps stochastic gradient descent requires, the better the generalization of the resulting model!

Linear Regression



Find a linear predictor $\hat{y} = \boldsymbol{w} \cdot \boldsymbol{x}$ to minimize the square error over the data $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ thus

$$\text{Minimize: } \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - \boldsymbol{w} \cdot \boldsymbol{x}_i)^2$$

Linear Regression Cost Function

- Model: $\hat{y}(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$
- Example-wise loss function: $l(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$
- Total loss function:

$$L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) = \sum_{i=1}^{m} \frac{1}{2}(y_i - \hat{y}(\mathbf{w}, \mathbf{x}_i))^2$$

$$= \frac{1}{2}\sum_{i=1}^{m}(y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

- Minimising the squared error is equivalent to assuming Gaussian noise in a maximum likelihood estimation

Stochastic gradient descent for regression

- Total loss gradient: $\nabla_{\mathbf{w}} L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) = \sum_{i=1}^m \frac{\partial l(y_i, \hat{y})}{\partial \hat{y}} \times \nabla_{\mathbf{w}} \hat{y}(\mathbf{w}, \mathbf{x}_i)$
- Loss gradient: $\frac{\partial l(y, \hat{y})}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} \frac{1}{2}(y - \hat{y})^2 = y - \hat{y}$
- Model gradient: $\nabla_{\mathbf{w}} \hat{y}(\mathbf{w}, \mathbf{x}) = \nabla_{\mathbf{w}} (\mathbf{w} \cdot \mathbf{x}) = \mathbf{x}$
- Put together:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) = \sum_{i=1}^m (y_i - \hat{y}(\mathbf{w}, \mathbf{x})) \times \mathbf{x}_i$$

$$= \sum_{i=1}^m (y_i - \mathbf{w} \cdot \mathbf{x}_i) \times \mathbf{x}_i$$

Batch and stochastic gradient descent

- Batch gradient descent (entire batch of data):

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^{m}) \quad = \quad \mathbf{w} - \eta \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \mathbf{x}_i) \times \mathbf{x}_i$$

- Online gradient descent (one by one) :

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} l(\mathbf{w}, \mathbf{x}_j, y_j) \quad = \quad \mathbf{w} - \eta (y_j - \mathbf{w} \cdot \mathbf{x}_j) \times \mathbf{x}_j$$

- Often best to use mini-batch ($1 < k < m$ examples at a time)

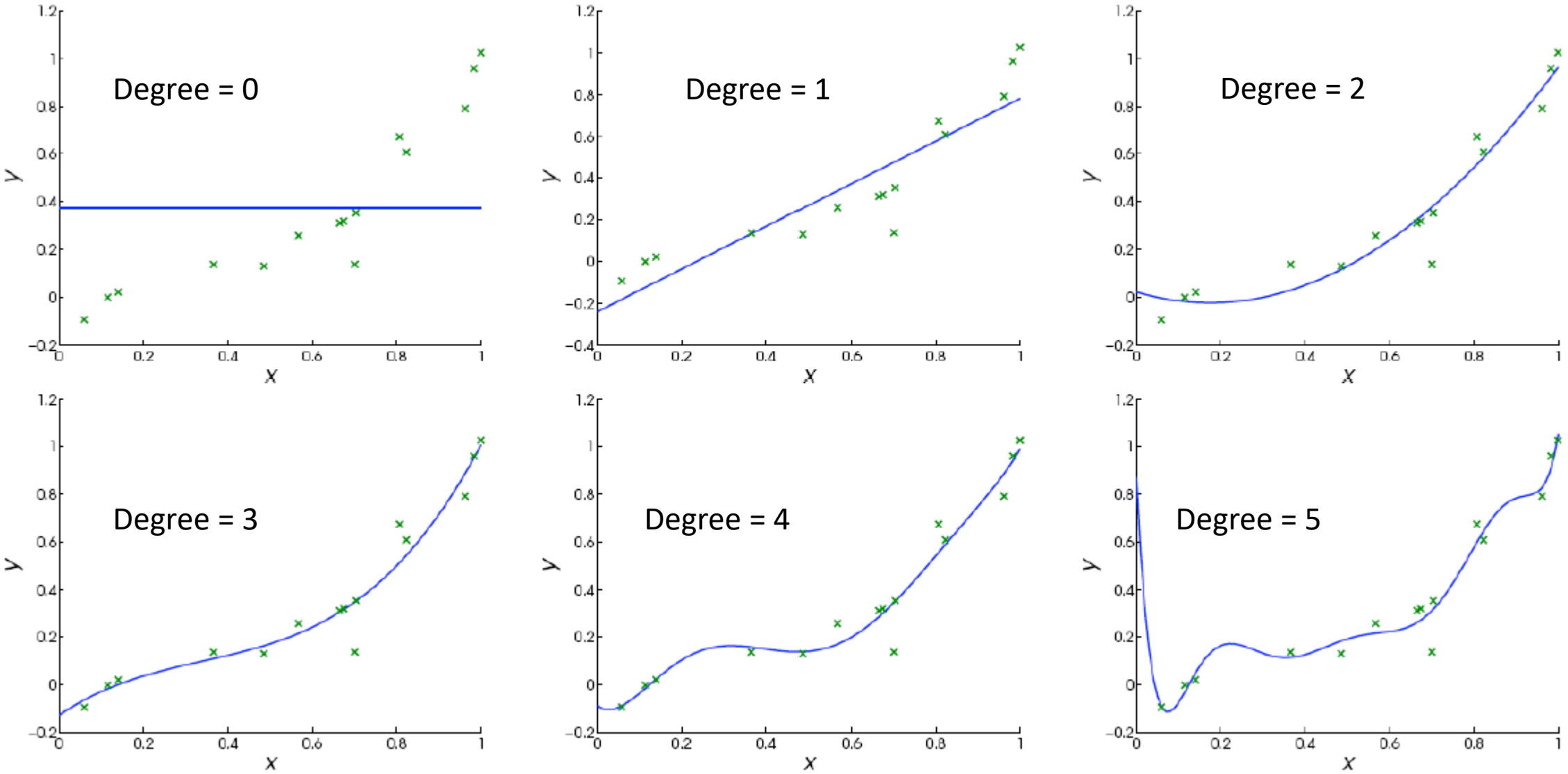
Regularisation

- One way of limiting the capacity of models is regularization
- The most popular regulariser for linear models and neural networks is called weight decay
- In the cost function, one adds a term $\lambda||\boldsymbol{w}||^2$ to penalize large weights
- In the gradient update rule, this leads to a “force” that aims to reduce the length of the weight vector
- Alternatively, in a Bayesian maximum aposteriori view, the same effect can be achieved with a zero mean Gaussian prior over weights
- Other regularisers can induce sparse structure in the weights (L1) or reduce co-dependence (dropout) → more in Simon’s, Karen’s lectures

Non-linear Basis Functions

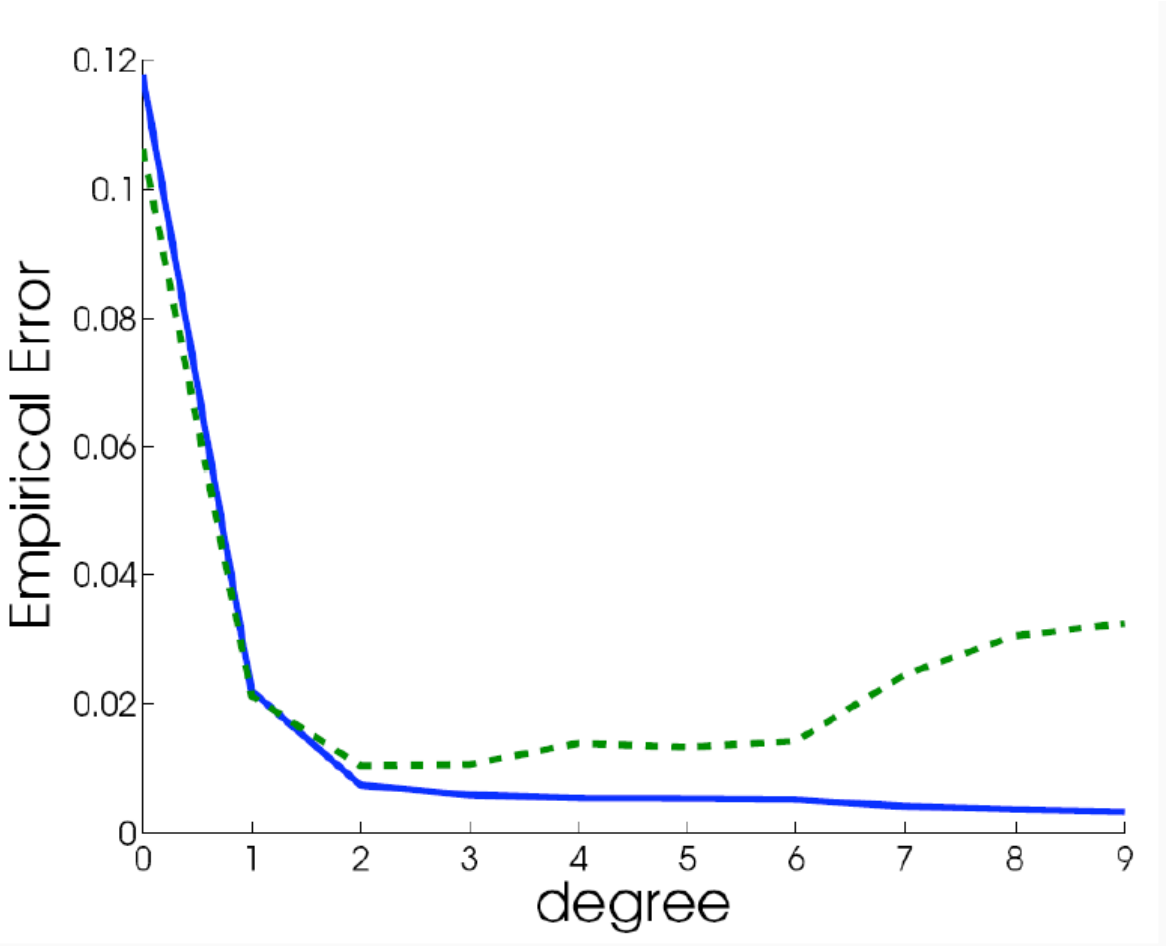
- Linear regression is linear in the parameters, not necessarily in the inputs!
- Can use basis functions $\phi(x)$ to map data into a different feature space representation.
- Example: $\phi_0(x) = 1, \phi_1(x) = (1, x), \phi_2(x) = (1, x, x^2)$, etc.
- Now linear model can be written as $\hat{y} = \boldsymbol{w} \cdot \phi(x)$ and is non-linear in the data
- This provides a richer hypothesis space for fitting the data
- With kernels $k(x, x')$ we can avoid explicit mapping (SVMs)

Regression with polynomial basis functions



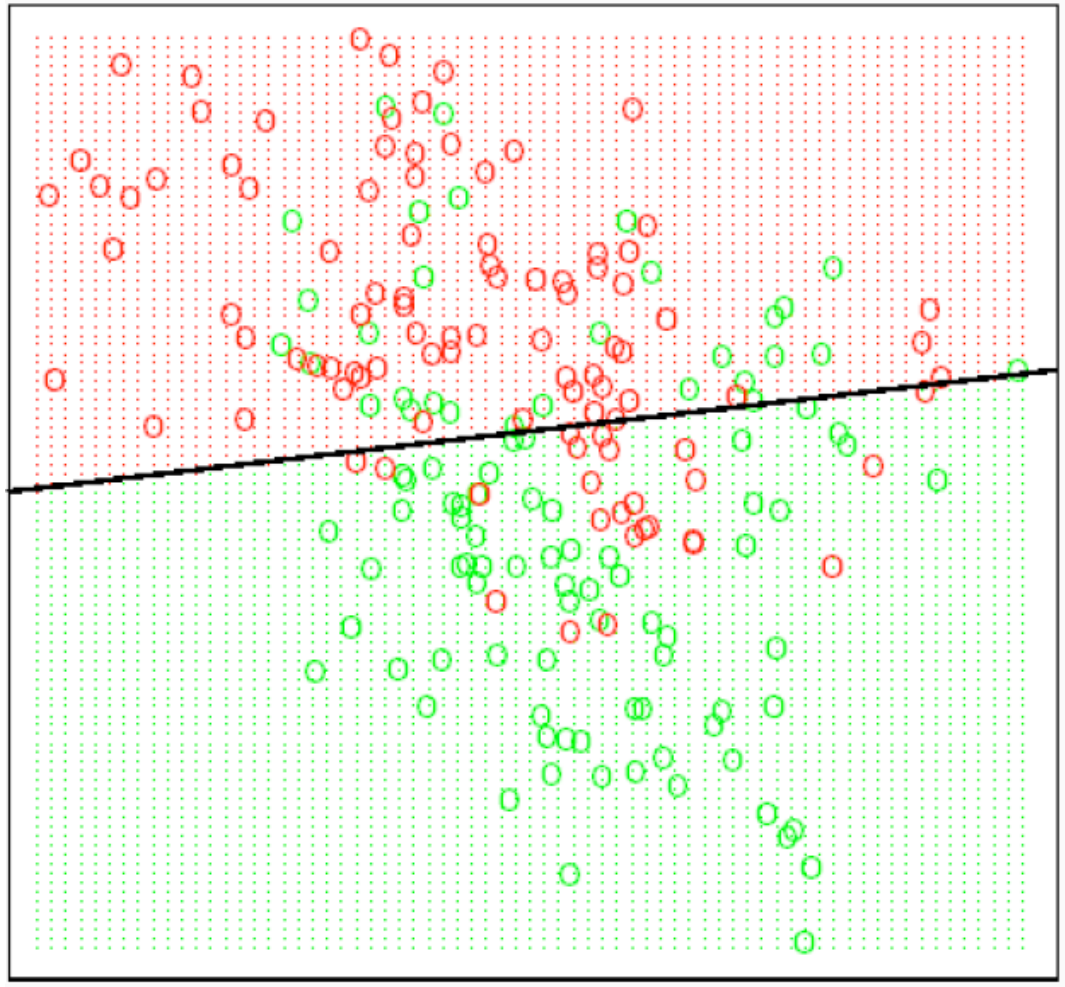
Polynomial Fit for different degrees

- Training error goes down with increasing degree (better fit)
- Test error is optimal at degree 2, and deteriorates for higher degrees
- Note the similarity to learning curves discussed earlier. The effective hypothesis class of neural networks becomes more complex with longer training



Logistic Regression for classification

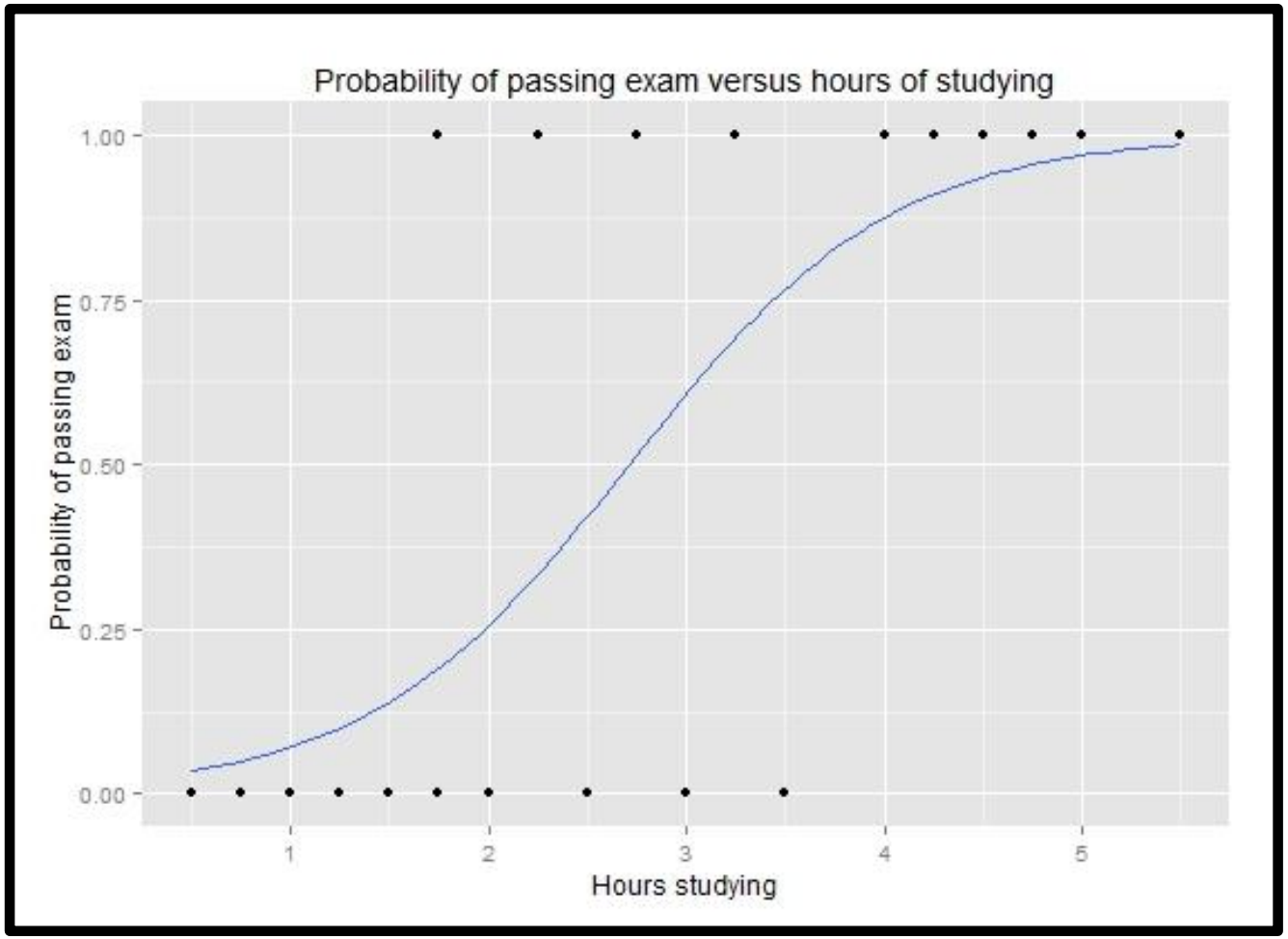
- Generalized linear model for binary classification
- Used, e.g., in click-through-rate prediction for search engine advertising
- Find linear hyperplane to separate the data
- Predict probability of class



Logistic Regression Cost Function

- Linear model: $z(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$
- (Inverse) Link function: $\hat{p}(z) = \frac{1}{1 + \exp(-z)}$
- Cross entropy loss: $l(y, \hat{p}) = y \log \hat{p} + (1 - y) \log(1 - \hat{p})$
- The regression loss is a composition of these three functions, aggregated over training examples

Logistic (Inverse) Link Function



Cross Entropy

- The general form of cross entropy between distributions p and q is given by $H(p,q) = E_p[-\log q] = H(p) + D_{KL}(p||q)$
- Cross entropy measures how many bits are needed to encode a message assuming that the letters are drawn from distribution q when the real distribution is p.
- In the case of two point discrete distributions with success probabilities p and $1-p$, cross entropy is given by

$$H(p,q) = -p\log q \ -(1-p)\log(1-q)$$

Logistic Regression Cost Function

$$
\begin{aligned}
L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) &= -\sum_{i=1}^m y_i \log \hat{p}(z_i) + (1 - y_i) \log(1 - \hat{p}(z_i)) \\
&= -\sum_{i=1}^m y_i \log \left(\frac{1}{1 + \exp(-z_i)} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + \exp(-z_i)} \right) \\
&= -\sum_{i=1}^m y_i \log \left(\frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x}_i)} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x}_i)} \right) \\
&= \sum_{i=1}^m \log(1 + \exp(\mathbf{w} \cdot \mathbf{x}_i)) - y_i \mathbf{w} \cdot \mathbf{x}_i
\end{aligned}
$$

Modular Gradients for Logistic Regression

- Total Gradient:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) = \sum_{i=1}^m \left.\frac{\partial l(y_i, \hat{p})}{\partial \hat{p}}\right|_{\hat{p}=\hat{p}(z_i)} \times \left.\frac{\partial \hat{p}(z)}{\partial z}\right|_{z=z_i} \times \nabla_{\mathbf{w}} z(\mathbf{w}, \mathbf{x})|_{\mathbf{x}=\mathbf{x}_i}$$

- Loss gradient:

$$\left.\frac{\partial l(y_i, \hat{p})}{\partial \hat{p}}\right|_{\hat{p}=\hat{p}(z_i)} = -\frac{y_i}{\hat{p}(z_i)} + \frac{1-y_i}{1-\hat{p}(z_i)}$$

- Link gradient:

$$\left.\frac{\partial \hat{p}(z)}{\partial z}\right|_{z=z_i} = \left.\frac{\exp(-z)}{(1+\exp(-z))^2}\right|_{z=z_i} = \hat{p}(z_i)(1-\hat{p}(z_i))$$

- Model gradient:

$$\nabla_{\mathbf{w}} z(\mathbf{w}, \mathbf{x})|_{\mathbf{x}=\mathbf{x}_i} = \nabla_{\mathbf{w}} \mathbf{w} \cdot \mathbf{x}|_{\mathbf{x}=\mathbf{x}_i} = \mathbf{x}_i$$

Putting the gradient back together

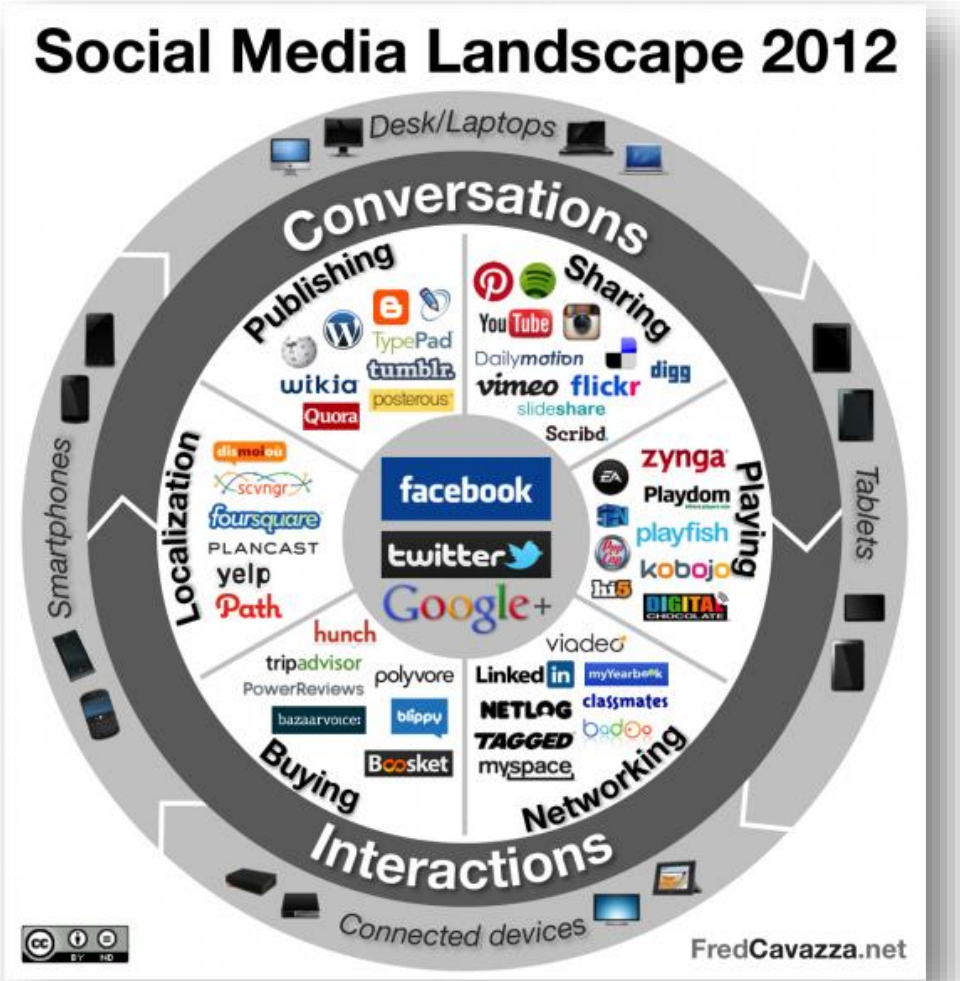
$$
\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i=1}^m) &= \sum_{i=1}^{m} \left(-\frac{y_i}{\hat{p}(z_i)} + \frac{1-y_i}{1-\hat{p}(z_i)} \right) \times \hat{p}(z_i)(1-\hat{p}(z_i)) \times \mathbf{x}_i \\
&= (\hat{p}(z_i) - y_i)\mathbf{x}_i \\
&= \left(\frac{1}{1+\exp(-z_i)} - y_i \right) \mathbf{x}_i \\
&= \left(\frac{1}{1+\exp(-\mathbf{w} \cdot \mathbf{x}_i)} - y_i \right) \mathbf{x}_i
\end{aligned}
$$

- Similarly, the backpropagation algorithm works through the layers of deeper neural networks to calculate error gradients w.r.t. to weights
- Simon's lecture will give more details, also relevant to Assignment 1

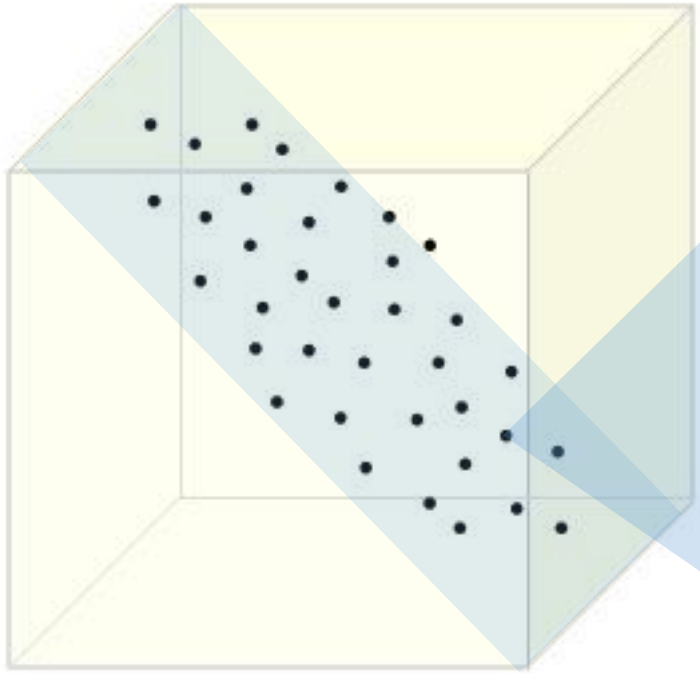
Case Study: Using Regression to Predict Private Attributes from Facebook Likes

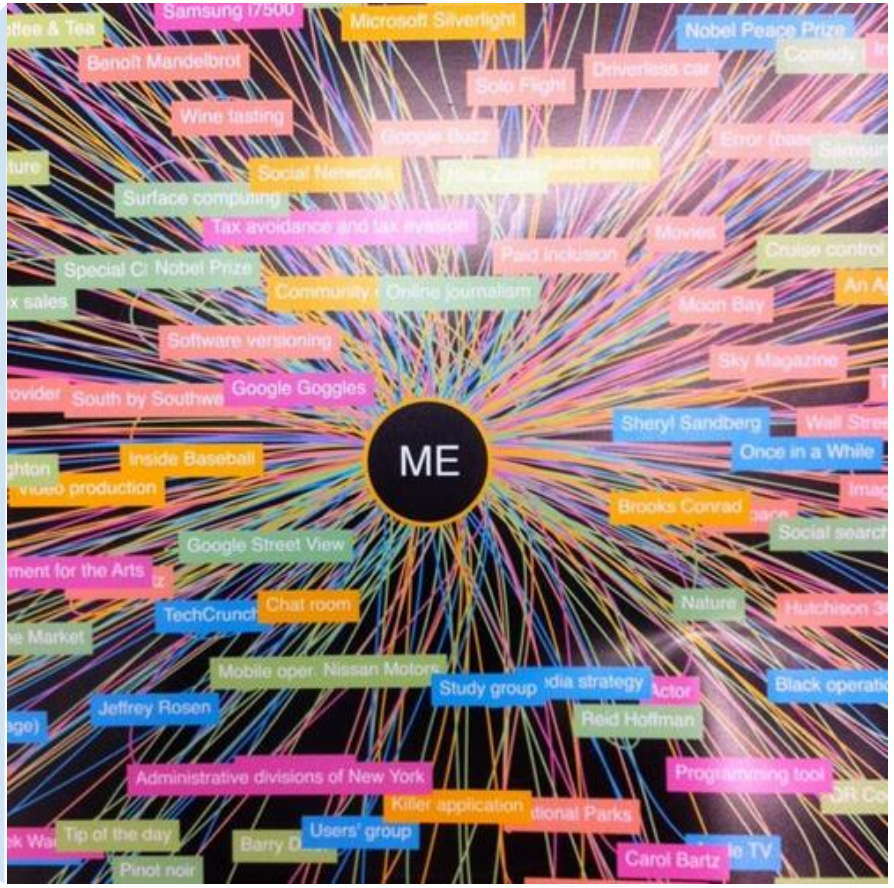
Based on joint work with:
MSR Cambridge: Yoram Bachrach and Pushmeet Kohli
Cambridge University: Michal Kosinski and David Stillwell

The Digital Traces: Online Data



High-Dimensional Observation Space





www.marketersstudio.com

Mapping the Human Manifold

- Scientific potential - Understanding:
  - Better understand human behaviour
  - Understand commonality and individual differences among people
  - Obtain psychometric measurements at an unprecedented scale
- Business potential - Predicting:
  - Develop fine-grained and predictive psycho-demographic user profiles
  - Increase user satisfaction by deep personalization for products and services
  - Increase revenue by providing more engaging ads and recommendations
- How: Find mapping to interpretable dimensions
  - Personality, Intelligence, Happiness, etc.

Big Five Personality traits

Openness

- **Appreciation of art, emotion, adventure, and variety of experience**

Conscientiousness

- **Self-discipline, act dutifully, and aim for achievement**

Extraversion

- **Energy, positive emotions, seek social stimulation**

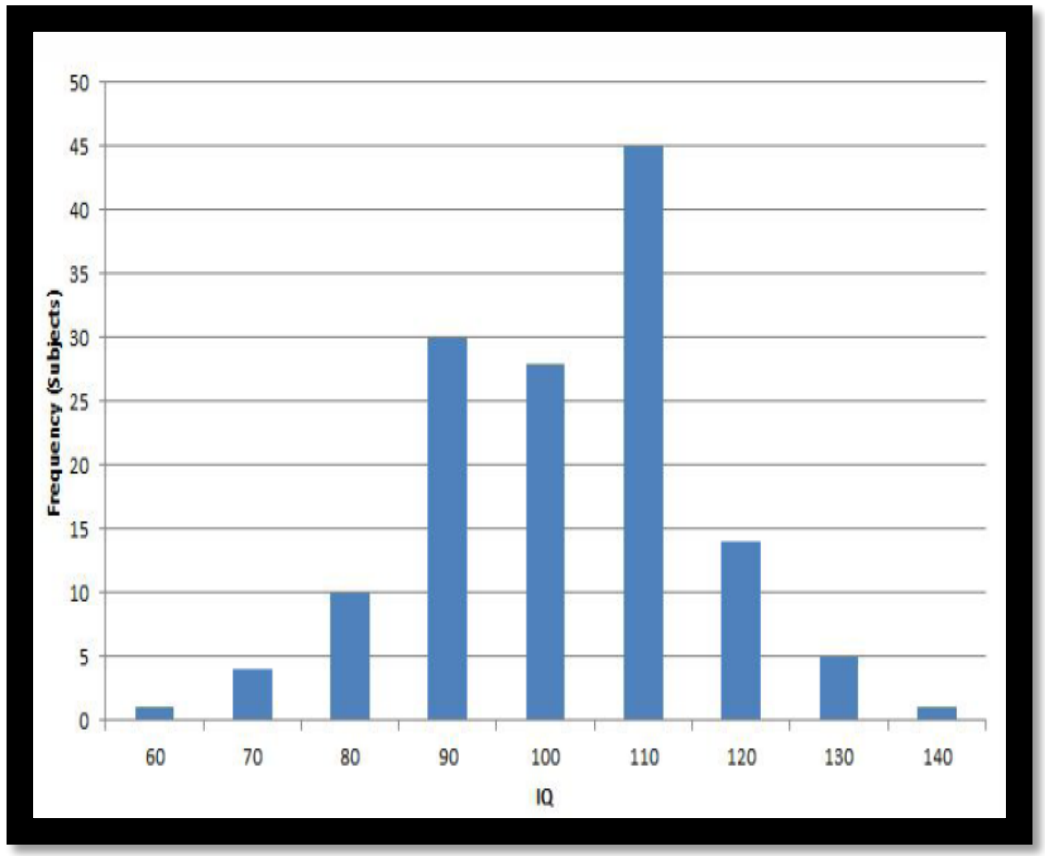
Agreeableness

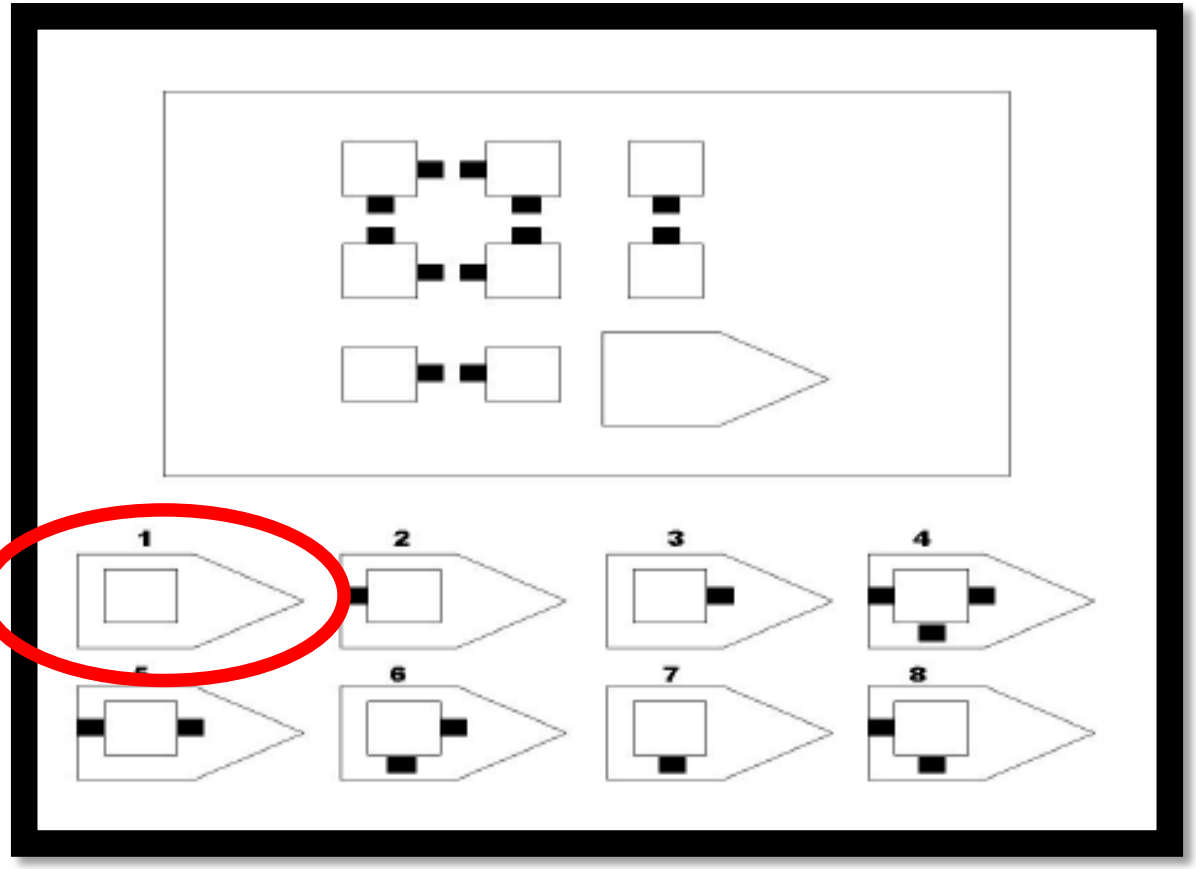
- **Compassionate and cooperative rather than suspicious**

Neuroticism

- **Experience unpleasant emotions easily, such as anger and anxiety**

General Intelligence - IQ





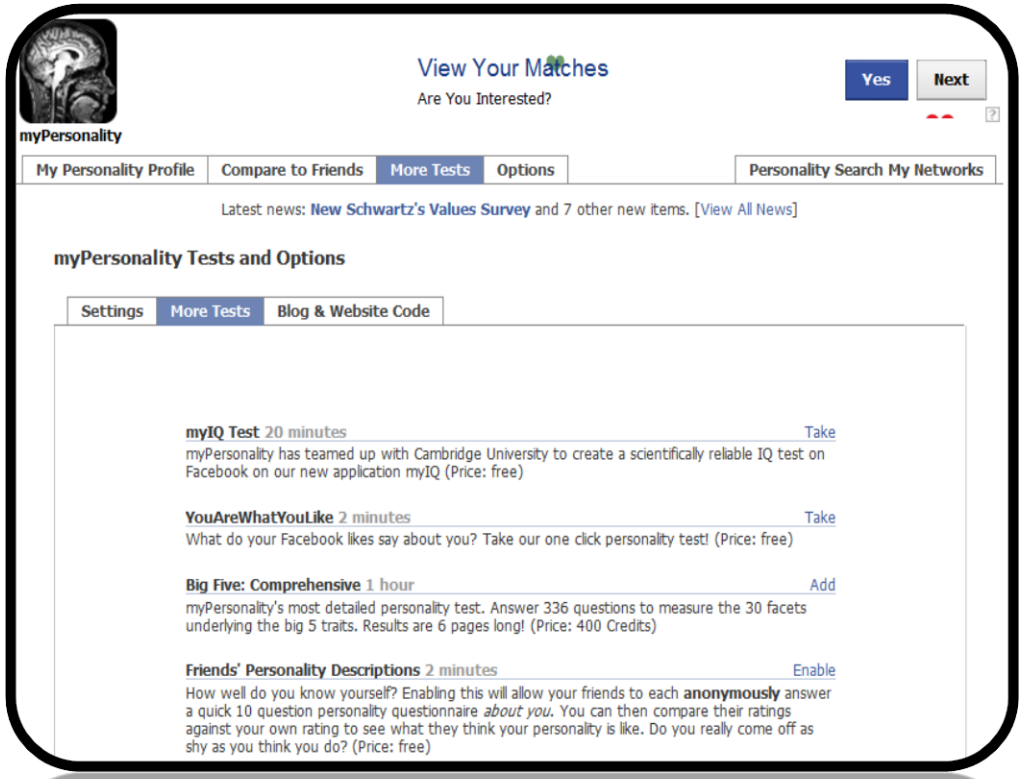
Satisfaction with Life Scale (Happiness)

To what degree to you agree with the following statements?

- In most ways my life is close to ideal.
- The conditions of my life are excellent.
- I am satisfied with my life.
- So far I have gotten the important things I want in life.
- If I could live my life over, I would change almost nothing.

A Treasure Chest of Data: MyPersonality

- Facebook App since 2008
- Over 8 Million psychometric test results
  - Personality
  - Intelligence
  - Happiness
- Volunteered user profiles
  - Relationship status, age, gender
  - Facebook Likes
  - Friendship network



Data: www.myPersonality.org

David Stillwell, Michal Kosinski

Cambridge Psychometrics Centre

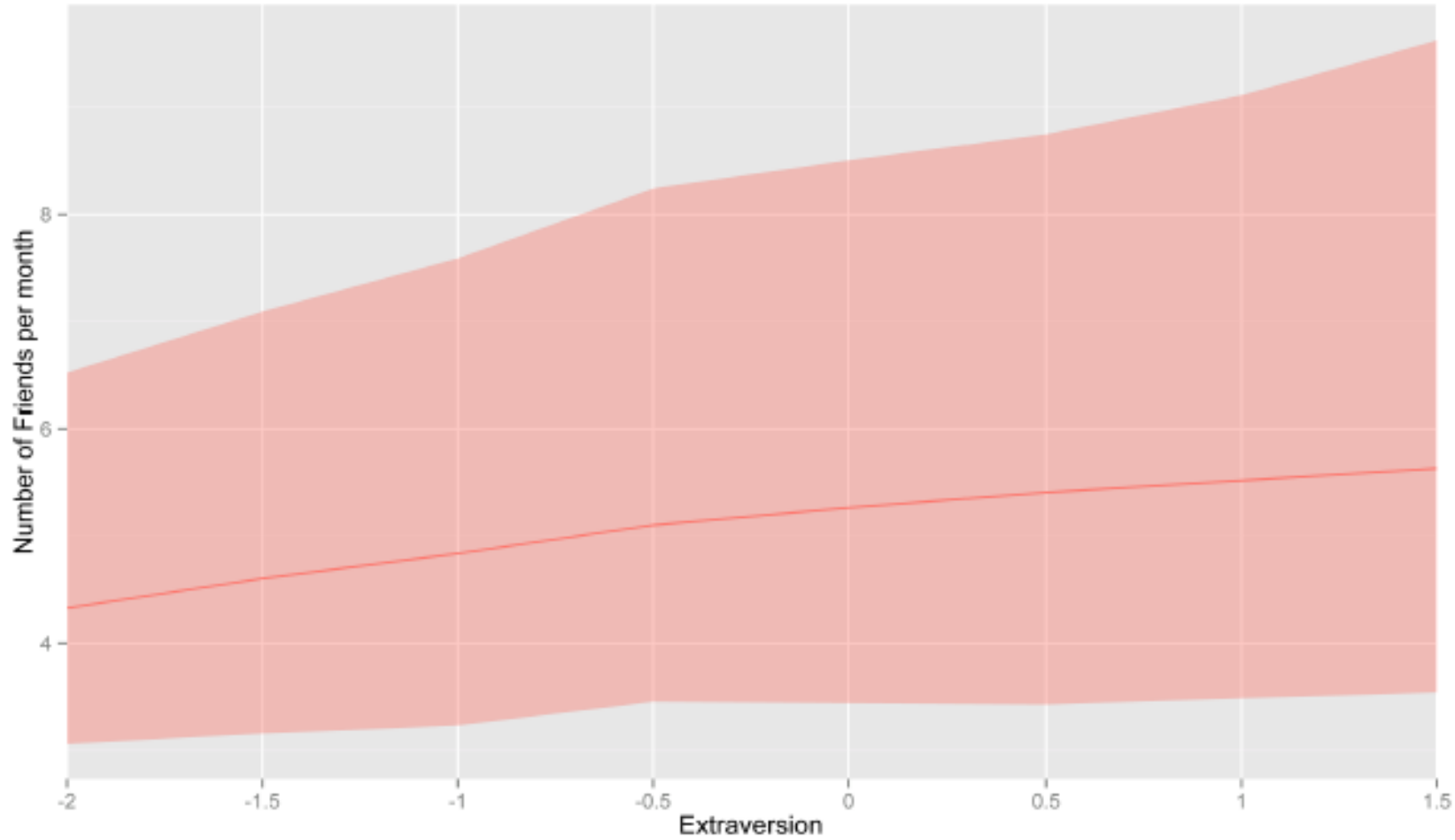


Fig. 6 Median of friends added per month by users characterized by different levels of Extroversion. Ribbon represents the interquartile range, or the middle 50 percentiles of the number of friends added per month

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]The Psychometrics Centre, University of Cambridge, Cambridge CB2 3EB United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization and privacy.

social networks | computational social science | machine learning | big data | data mining | psychological assessment

A growing proportion of human activities, such as social interactions, entertainment, shopping, and gathering information, are now mediated by digital services and devices. Such digitally mediated behaviors can easily be recorded and analyzed, fueling the emergence of computational social science (1) and new services such as personalized search engines, recommender systems (2), and targeted online marketing (3). However, the widespread availability of extensive records of individual behavior, together with the desire to learn more about customers and citizens, presents serious challenges related to privacy and data ownership (4, 5).

We distinguish between data that are actually recorded and information that can be statistically predicted from such records. People may choose not to reveal certain pieces of information about their lives, such as their sexual orientation or age, and yet this information might be predicted in a statistical sense from other aspects of their lives that they do reveal. For example, a major US retail network used customer shopping records to predict pregnancies of its female customers and send them well-timed and well-targeted offers (6). In some contexts, an unexpected flood of vouchers for prenatal vitamins and maternity clothing may be welcome, but it could also lead to a tragic outcome, e.g., by revealing (or incorrectly suggesting) a pregnancy of an unmarried woman to her family in a culture where this is unacceptable (7). As this example shows, predicting personal information to improve products, services, and targeting can also lead to dangerous invasions of privacy.

Predicting individual traits and attributes based on various cues, such as samples of written text (8), answers to a psychometric test (9), or the appearance of spaces people inhabit (10), has a long history. Human migration to digital environment renders it possible to base such predictions on digital records of human behavior. It has been shown that age, gender, occupation, education level, and even personality can be predicted from people's Web site browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases. For example, observing users' Likes related to music provides similar information to observing records of songs listened to online, songs and artists searched for using a Web search engine, or subscriptions to related Twitter channels. In contrast to these other sources of information, Facebook Likes are unusual in that they are currently publicly available by default. However, those other digital records are still available to numerous parties (e.g., governments, developers of Web browsers, search engines, or Facebook applications), and, hence, similar predictions are unlikely to be limited to the Facebook environment.

The design of the study is presented in Fig. 1. We selected traits and attributes that reveal how accurate and potentially intrusive such a predictive analysis can be, including "sexual orientation," "ethnic origin," "political views," "religion," "personality," "intelligence," "satisfaction with life" (SWL), substance use ("alcohol," "drugs," "cigarettes"), "whether an individual's parents stayed together until the individual was 21 y old," and basic demographic attributes such as "age," "gender," "relationship status," and "size and density of the friendship network." Five Factor Model (9) personality scores ($n = 54,373$) were established using the International Personality Item Pool (IPIP) questionnaire with 20 items (25). Intelligence ($n = 1,350$) was measured using Raven's Standard Progressive Matrices (SPM) (26), and SWL ($n = 2,340$) was measured using the SWL Scale (27). Age ($n = 52,700$; average, $\mu = 25.6$; SD = 10), gender ($n = 57,505$; 62% female), relationship status ("single"/"in relationship"; $n = 46,027$; 49% single), political views ("Liberal"/"Conservative", $n = 9,752$;

Author contributions: M.K. and T.G. designed research; M.K. and D.S. performed research; M.K. and T.G. analyzed data; and M.K., D.S., and T.G. wrote the paper.

Conflict of interest statement: D.S. received revenue as owner of the myPersonality Facebook application.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the myPersonality Project database (www.mypersonality.org/wiki).

[1]To whom correspondence should be addressed. E-mail: michalkosinski@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218772110/-/DCSupplemental.

PNAS

Inferring the Demographics of Search Users

When Social Data Met Search Queries

Bin Bi*
UCLA
United States
bbi@cs.ucla.edu

Milad Shokouhi
Microsoft Research
United Kingdom
milads@microsoft.com

Michal Kosinski
University of Cambridge
United Kingdom
michal@michalkosinski.com

Thore Graepel
Microsoft Research
United Kingdom
thoreg@microsoft.com

ABSTRACT

Knowing users' views and demographic traits offers a great potential for personalizing web search results or related services such as query suggestion and query completion. Such signals however are often only available for a small fraction of search users, namely those who log in with their social network account and allow its use for personalization of search results. In this paper, we offer a solution to this problem by showing how user demographic traits such as age and gender, and even political and religious views can be efficiently and accurately inferred based on their search query histories. This is accomplished in two steps; we first train predictive models based on the publically available myPersonality dataset containing users' Facebook Likes and their demographic information. We then match Facebook Likes with search queries using Open Directory Project categories. Finally, we apply the model trained on Facebook Likes to large-scale query logs of a commercial search engine while explicitly taking into account the difference between the traits distribution in both datasets. We find that the accuracy of classifying age and gender, expressed by the area under the ROC curve (AUC), are 77% and 84% respectively for predictions based on Facebook Likes, and only degrade to 74% and 80% when based on search queries. On a US state-by-state basis we find a Pearson correlation of 0.72 for political views between the predicted scores and Gallup data, and 0.54 for affiliation with Judaism between predicted scores and data from the US Religious Landscape Survey. We conclude that it is indeed feasible to infer important demographic data of users from their query history based on labelled Likes data and believe that this approach could provide valuable information for personalization and monetization even in the absence of demographic data.

*The work was done during Bin's internship at Microsoft Research Cambridge.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]:

General Terms

Algorithms, Human Factors

Keywords

User demographics, Personalized search, Social networks

1. INTRODUCTION

In recent years, we have been witnessing the rapid emergence of social networks and an increasing amount of user generated data. Meanwhile, it became apparent that the relevance of search results can be improved by personalization, i.e., by taking into account additional information about the user, such as interests, demographic and psychological traits, social background, or the context of the search. As a consequence, search engines have been evolving into *social-aware* platforms, Google's social layer (Google+), and Bing's *social pane* being perhaps the two most noteworthy examples.

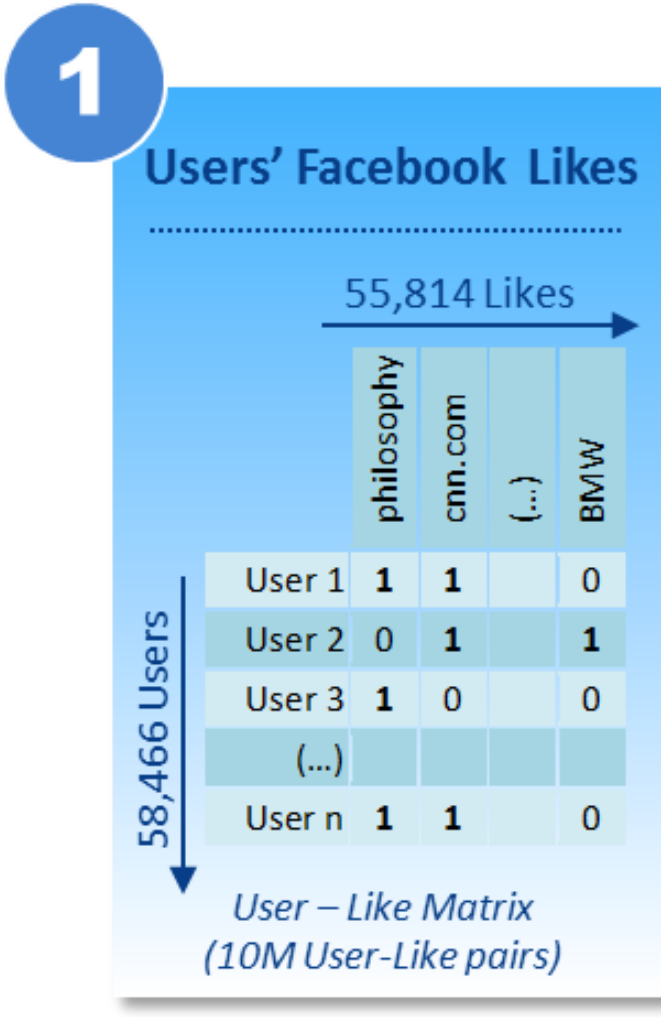
While leveraging the background information about the users in ranking models has shown significant promise in enhancing users' search experience both in academic [Carmel et al., 2009] and industrial[1] studies, obtaining such features for all users can be difficult. For instance, a recent study suggests that only about 22% of Bing users are logged into Facebook accounts while searching[2], and even them may have not given the search engine access to their profile information. It would therefore be useful to be able to infer characteristics of users relevant to their search experience from information more readily available in the context of a search engine, such as the search query histories.

This paper addresses the question of how demographic traits and users' views can be inferred based on the query histories. The main challenge, however, lies in the fact that only a very limited amount of data is available to allow training models for predicting such traits based on the search

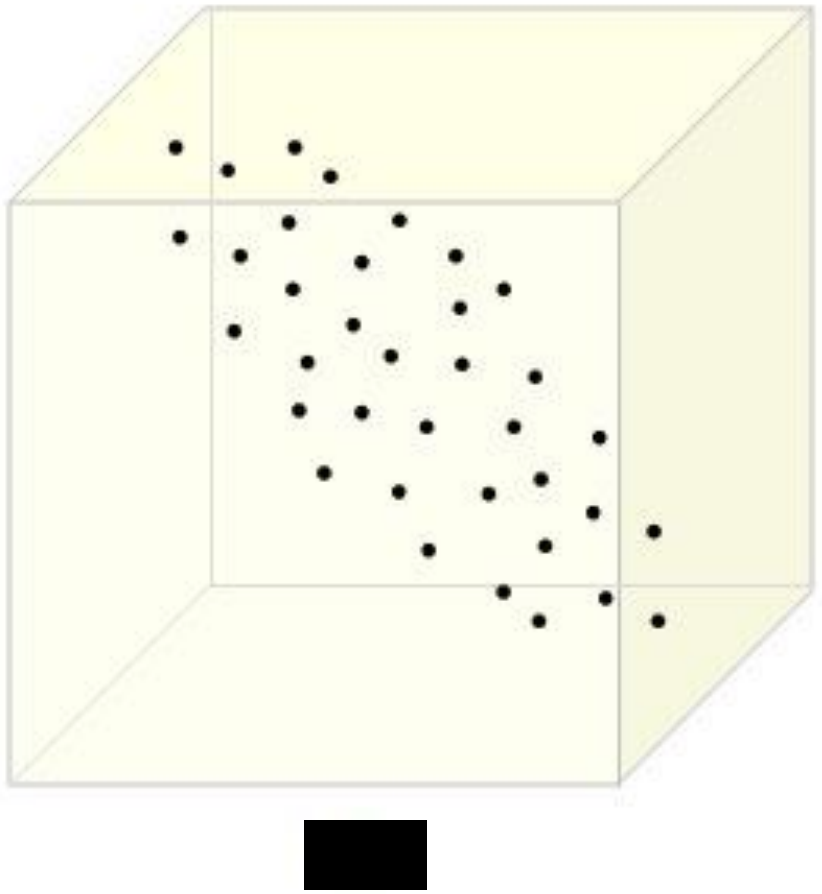
[1]Google blog, http://bit.ly/YaJvSm1
[2]Search Engine Land: http://selnd.com/R5dpTN

WWW 2013

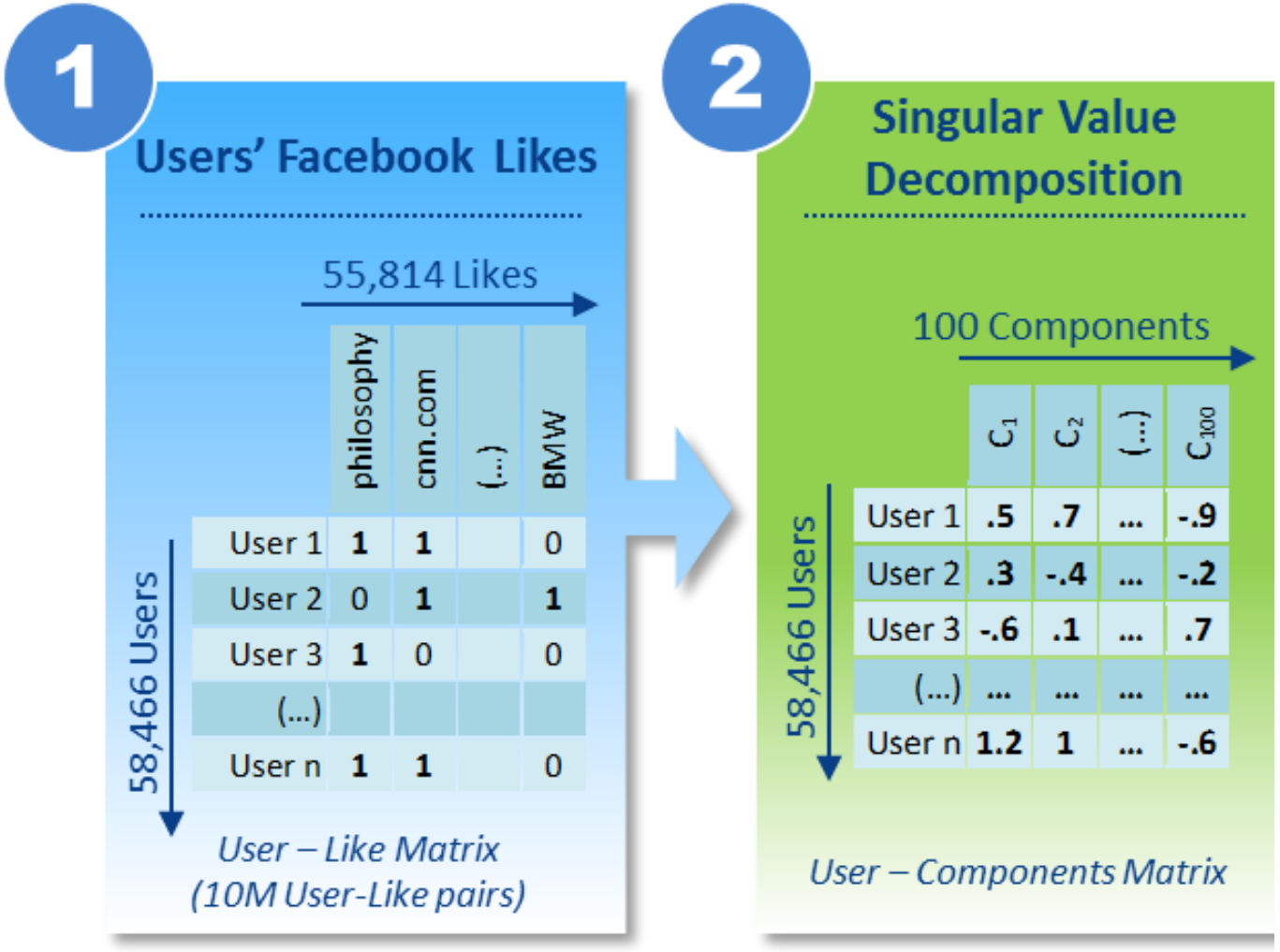
Statistical Methodology



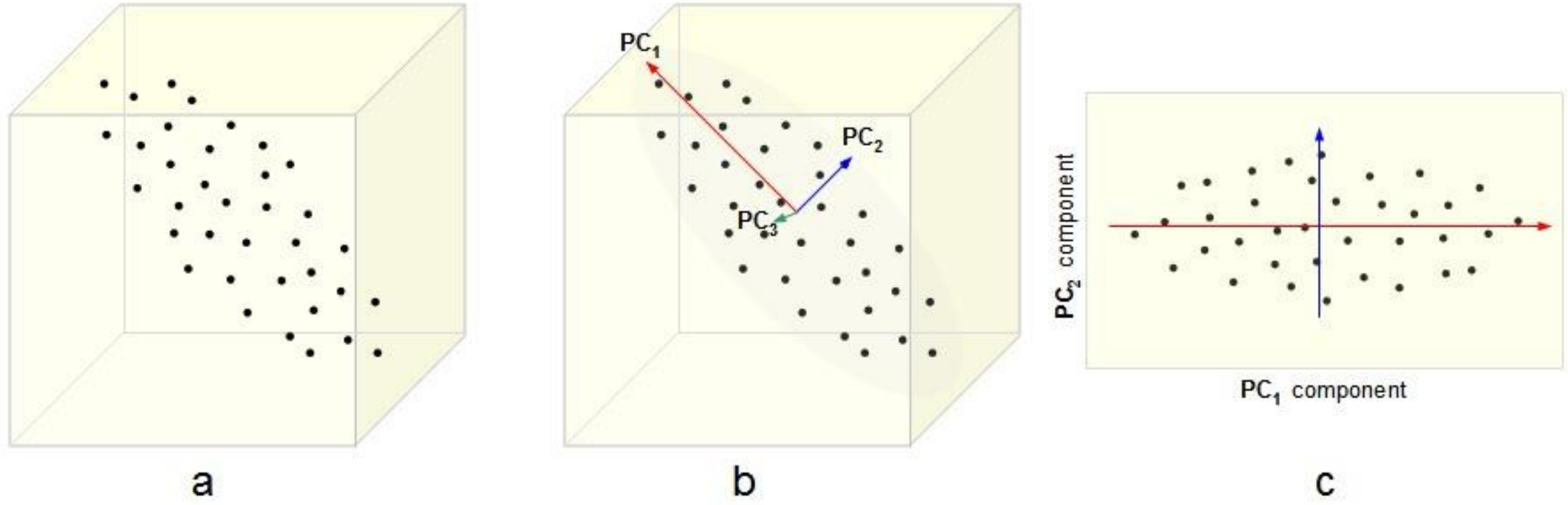
High-Dimensional Observation Space



Statistical Methodology

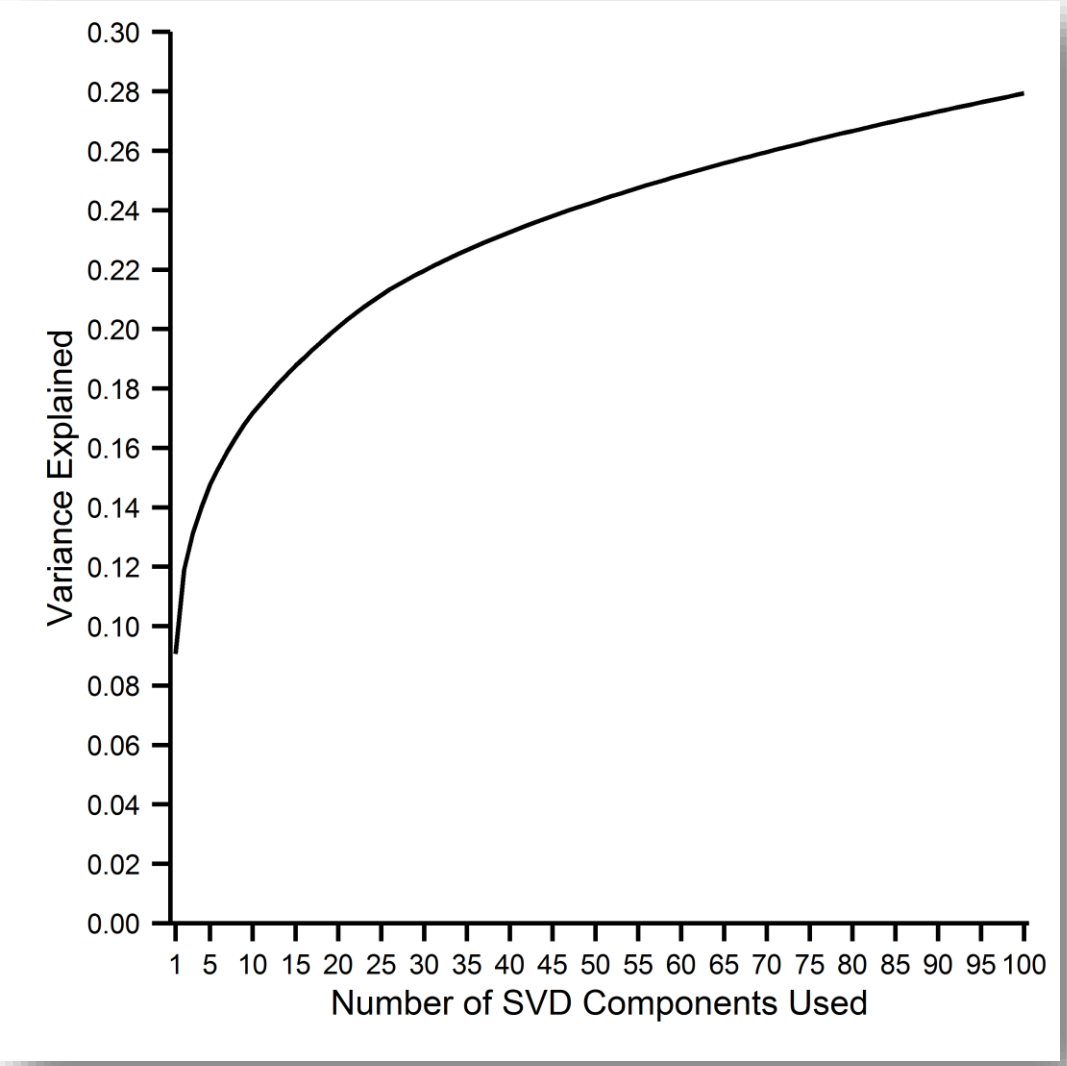


Mapping the Manifold

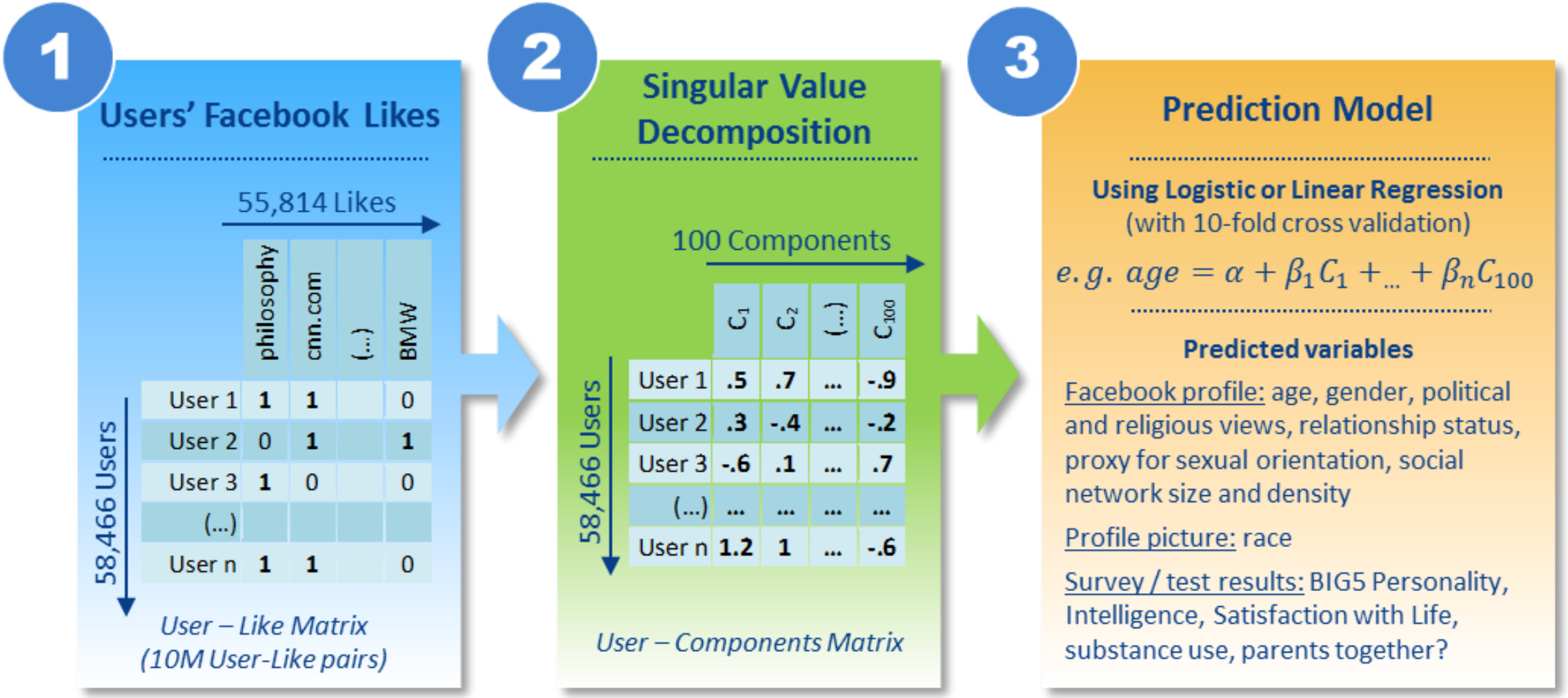


100 out of 55,814 dimensions explain 28% of variance

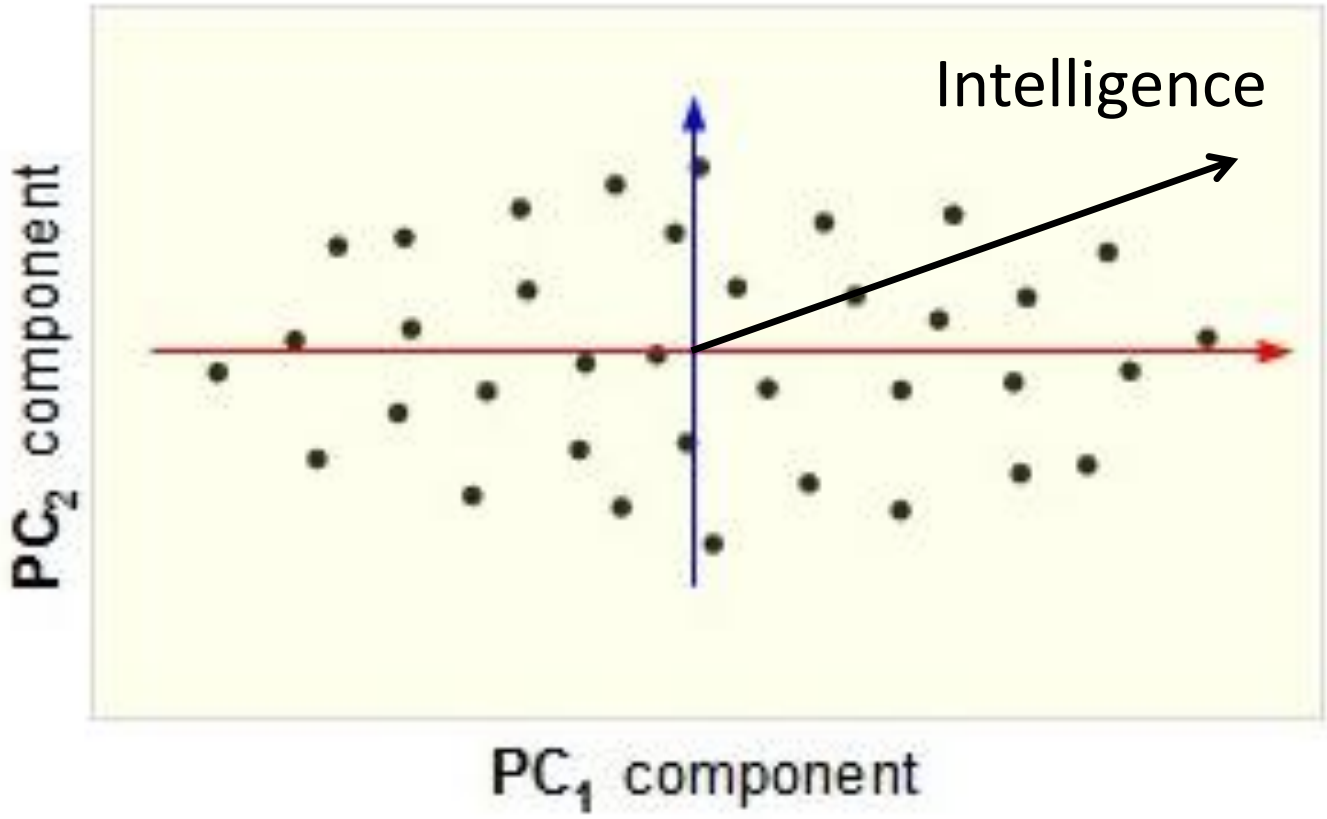
Scree Plot: How much variance explained?



Statistical Methodology

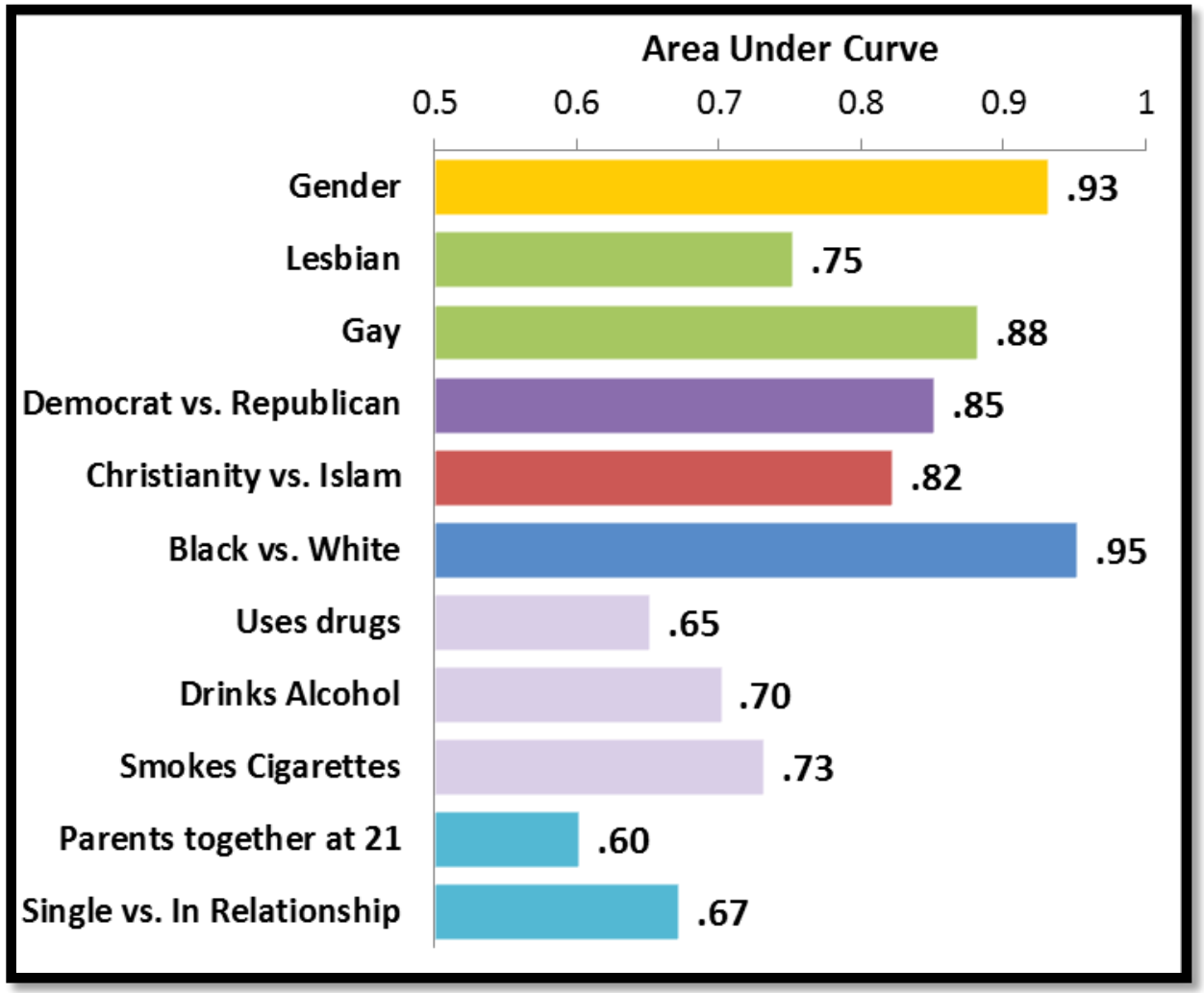


Regression:
Finding predictive directions in data space

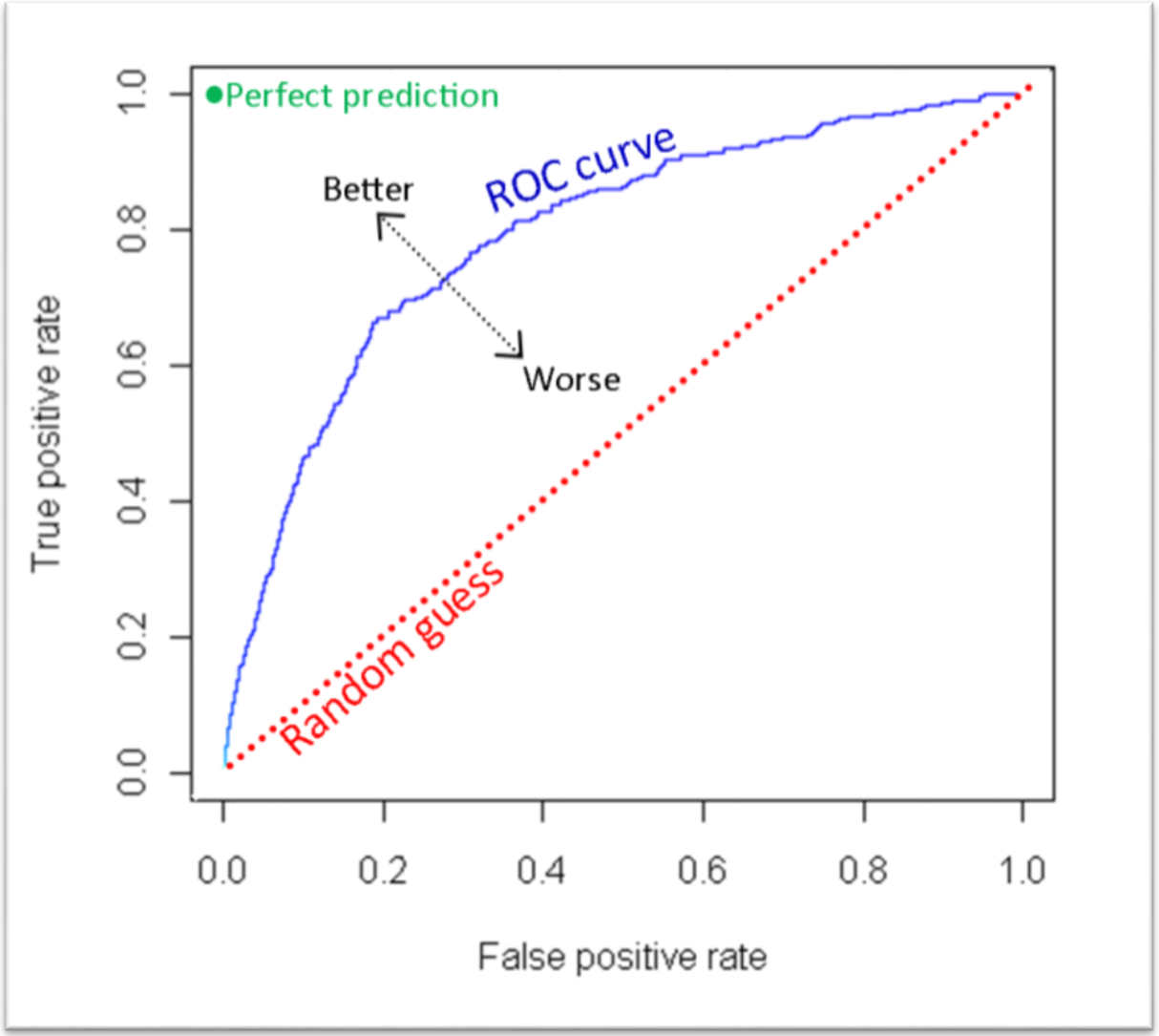


Source: Bloomberg

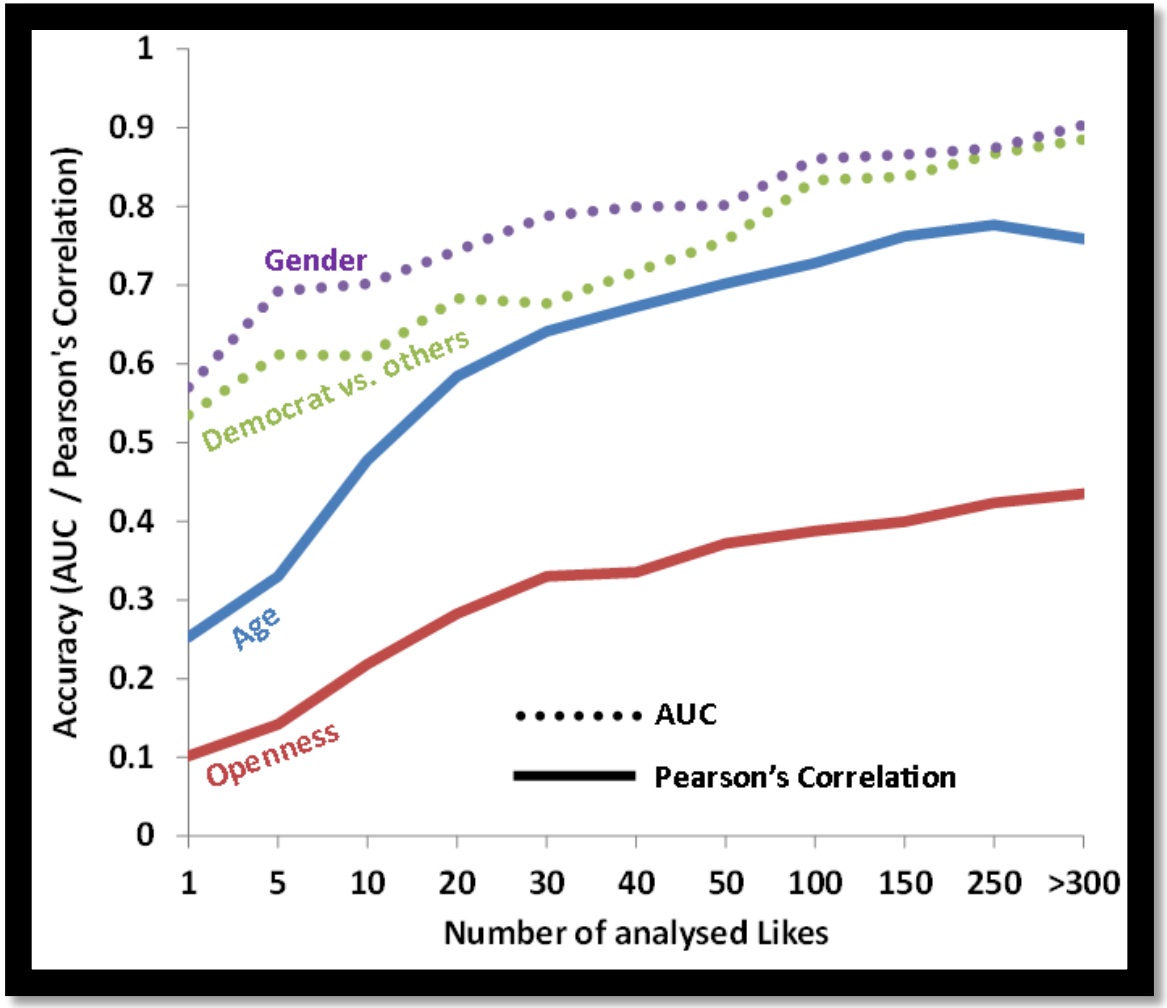
Prediction Accuracy: Binary variables



What is “Area under the Curve” (AUC)

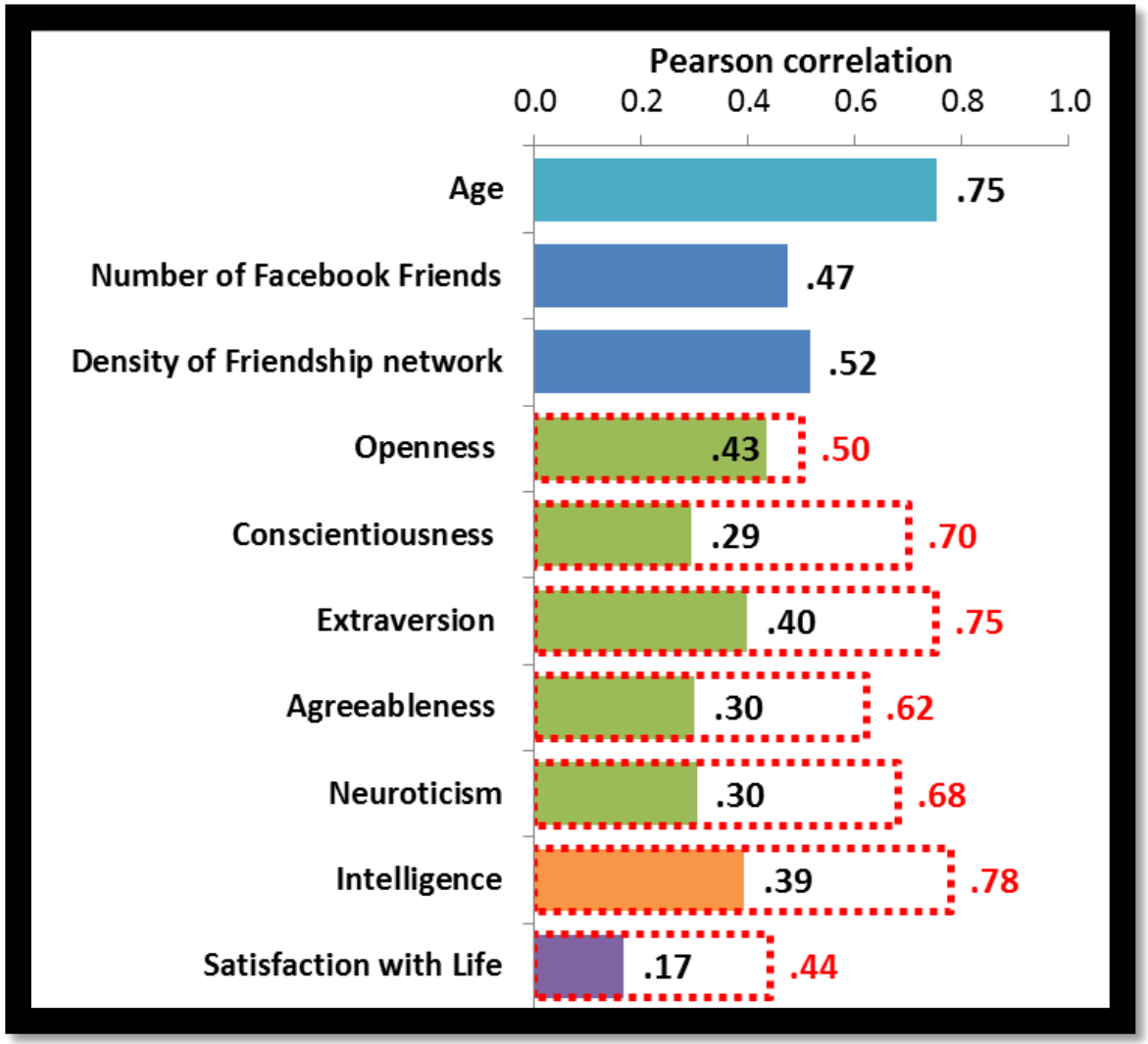


How many Likes for good accuracy?



- About 50% of users in this sample had at least 100 Likes and about 20% had at least 250 Likes.
- Baseline (random guessing) is 50% for binary variables – Gender and Political View

Prediction Accuracy: Numeric Variables



Source: The Colbert Report, Comedy Central

Correlation and Causality

If two quantities A and B appear to be correlated…

… there are only four possibilities:

1. It is a statistical fluke
2. A causes B
3. B causes A
4. There are one or more hidden causes that influence both A and B

What is there to like? Intelligence

IQ			
High	The Godfather	Jason Aldean	*Low*
	Mozart	Tyler Perry	
	Thunderstorms	Sephora	
	The Colbert Report	Chiq	
	Morgan Freemans Voice	Bret Michaels	
	The Daily Show	Clark Griswold	
	Lord Of The Rings	Bebe	
	To Kill A Mockingbird	I Love Being A Mom	
	Science	Harley Davidson	
	Curly Fries	Lady Antebellum	

Which Likes? Happiness

Satisfaction With Life	*Satisfied*	*Dissatisfied*
	Sarah Palin	Hawthorne Heights
	Glenn Beck	Kickass
	Proud To Be Christian	Atreyu (Metal Band)
	Indiana Jones	Lamb Of God
	Swimming	Gorillaz
	Jesus Christ	Science
	Bible	Quote Portal
	Jesus	Stewie Griffin
	Being Conservative	Killswitch Engage
	Pride And Prejudice	Ipod

Which Likes? Extraversion

Extraversion		
Outgoing & Active	Beerpong	RPGs
	Michael Jordan	Fanfiction.Net
	Dancing	Programming
	Socializing	Anime
	Chris Tucker	Manga
	I Feel Better Tan	Video Games
	Modeling	Role Playing Games
	Cheerleading	Minecraft
	Theatre	Voltaire
	Flip Cup	Terry Pratchet

Shy & Reserved

Which Likes? Agreeableness

Agreeableness		
Cooperative		*Competitive*
Compassion International	I Hate Everyone	
Logan Utah	I Hate You	
Jon Foreman	I Hate Police	
Redeeming Love	Friedrich Nietzsche	
Pornography Harms	Timmy South Park	
The Book Of Mormon	Atheism / Satanism	
Circles Of Prayer	Prada	
Go To Church	Sun Tzu	
Christianity	Julius Caesar	
Marianne Williamson	Knives	


BLOOMBERG DATA CENTER
What Your s Reveal About You
SINGLE VS. RELATIONSHIP:
0.67
PARENTS TOGETHER AT 21:
0.60
SMOKES CIGARETTES:
0.73
DRINKS ALCOHOL:
0.70
USES DRUGS:
0.65
STOCKS
TREASURIES
CURRENCIES
MARCH 12 11:51 AM
MARKET MAKERS
Bloomberg
SAMSUNG

Daily Mail

Facebook threat to users' privacy

By **Andrew Levy**

FACEBOOK users are at risk of unwittingly revealing personal details simply by 'liking' pages on the site dedicated to anything from celebrities to charities, researchers warn.

Sexuality, drug use, political views and religious beliefs can be accurately predicted by monitoring users' activity on the social networking website, they said.

The team from Cambridge University focused their research on Facebook's system of liking pages – the seemingly innocuous act of clicking a button illustrated with a thumbs up.

Worryingly, the researchers found that liking even apparently unrelated information still can be used to accurately predict personal details.

For example, the researchers found drug use is suggested by 'liking' milkshakes and swimming, while high IQs are indicated by showing a taste for curly fries, and the Godfather movies.

The study was carried out by Cambridge's Psychometrics Centre and based on the Facebook profiles of 58,000 people in the US.

Their 'likes' were fed into a computer algorithm which was used to predict a range of personality traits. Researchers predicted male sexuality with 88 per cent accuracy. They also had an 85 per cent success rate with political leanings and 82 per cent with religion.

Dr Gus Hosein, of campaigners Privacy International, said: 'It's a nightmare scenario that Facebook is entirely responsible for setting up. This information can be used to pre-categorise people.

'Banks could use it to decide who gets a loan. It also creates the perfect surveillance state for governments.'

Facebook declined to comment yesterday.

The Telegraph

Toronto Star

THE TIMES | Tuesday March 12 20

How F… your cl…

Murad Ahmed
Technology
Reporter

A4 | TORONTO STAR

ON ON0

News

Reactions

Sam Gosling, a psychologist at the University of Texas at Austin, calls it a "landmark study" because it illustrates "how things are no longer ephemeral." He has been studying Facebook behavior since 2006, and has seen this new study.

USA Today

that can be used maliciously. Graham Cluley, internet security expert at Sophos, says: "In isolation these bits of data are not very useful, but when it is combined with other information it can become a little stepping stone to identity theft. People need to be so careful about how information can be put together, assimilated and used."

The Telegraph

Dr Gus Hosein, of campaigners Privacy International, said: 'It's a nightmare scenario that Facebook is entirely responsible for setting up. This information can be used to pre-categorise people.

'Banks could use it to decide who gets a loan. It also creates the perfect surveillance state for governments.'

The Daily Mail

less impressed. You already have more sensitive information online, Dr. Nicholas Christakis, director of the Human Nature Lab at Harvard, tells *The Los Angeles Times*. "I think this paper is alarmist. We can go from curly fries to pogroms in a couple steps."

The Los Angeles Times

Questions about User Privacy


Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?
- If they were, would they make that data public?
- Do these predictions just summarize what users are signalling to their friends using Facebook Likes?
- Should companies/services be allowed to use this information for commercial purposes?

Neural Nets, Backprop, Automatic Differentiation

- Lecture topics:
  - Neural nets
  - Multi-class classification and softmax loss
  - Modular backprop
  - Automatic differentiation
- Guest Lecturer: **Simon Osindero**
  - Joined DeepMind in 2016.
  - Undergrad/Masters in Natural Sciences/Physics at University of Cambridge.
  - PhD in Computational Neuroscience from UCL (2004). Supervisor: Peter Dayan.
  - Postdoc at University of Toronto with Geoff Hinton. (Deep belief nets, 2006).
  - Started an A.I. company, LookFlow, in 2009. Sold to Yahoo in 2013.
  - Current research topics: deep learning, RL agent architectures and algorithms, memory, continual learning.