Detecting time series anomalies at Uber scale with recurrent neural networks

Andrea Pasqua

Anny Chen

IDS Team - Uber

Observe, Relate, Predict

UBER

Anomaly Detection at Uber: the Business Angle

Our mission

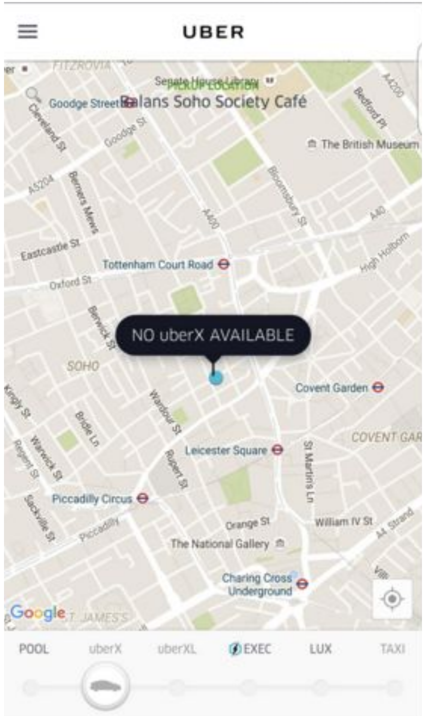
More reliable and safer transportation everywhere,

for everyone

Anomaly Detection at Uber: the Business Angle

An important component

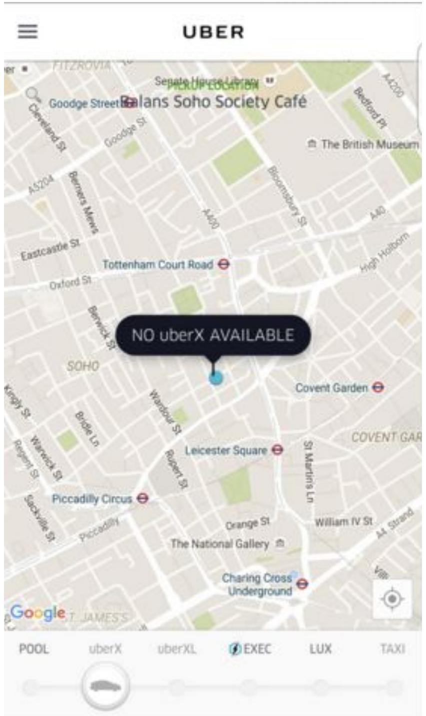
Reliability of the App



Anomaly Detection at Uber: the Business Angle

Uber's app is different

Nobody is "just browsing"



Unusually high cost of outages

- Transactions permanently lost
- Costs magnified by the scale of the business

Anomaly Detection at Uber: the Business Angle

Great opportunity for cost saving

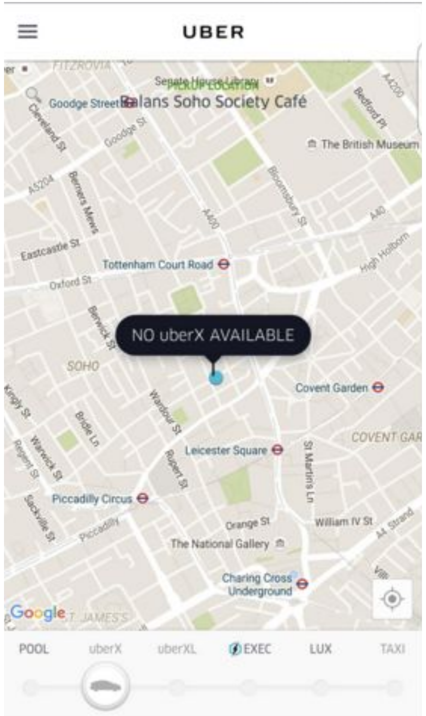
About $8M saved last year

Through intelligent, automated on-call alerting

- Conservative estimate

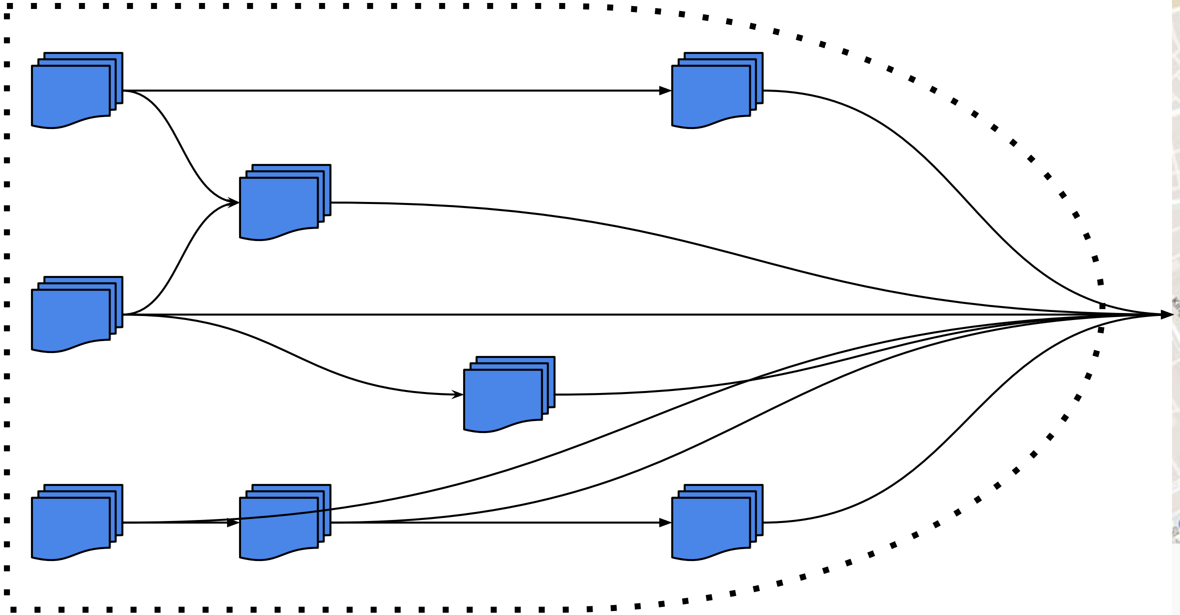
The Scale of the Problem

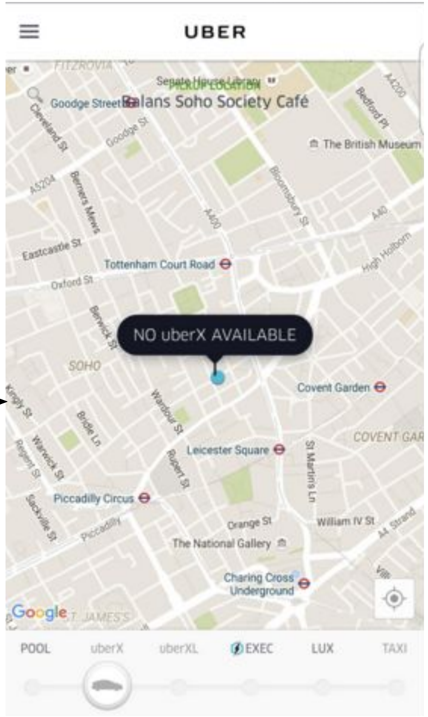
What does it take to ensure a reliable app?



The Scale of the Problem

An ecosystem of microservices





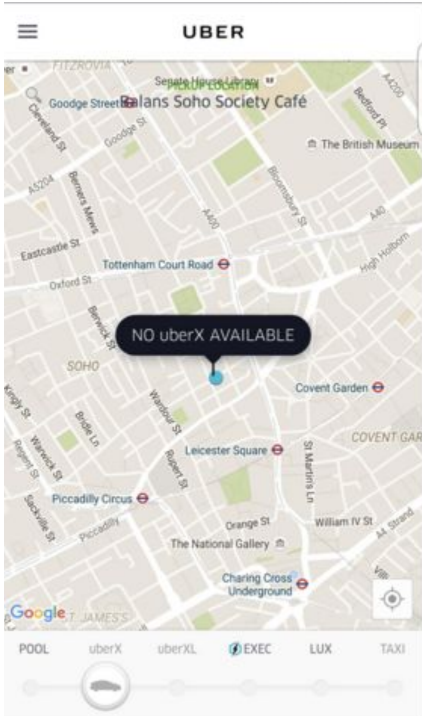
The Scale of the Problem

An ecosystem of microservices

Each service has multiple traces to monitor

Powerful combinatorics: x geo x product

more than 1 billion traces



The Scale of the Problem

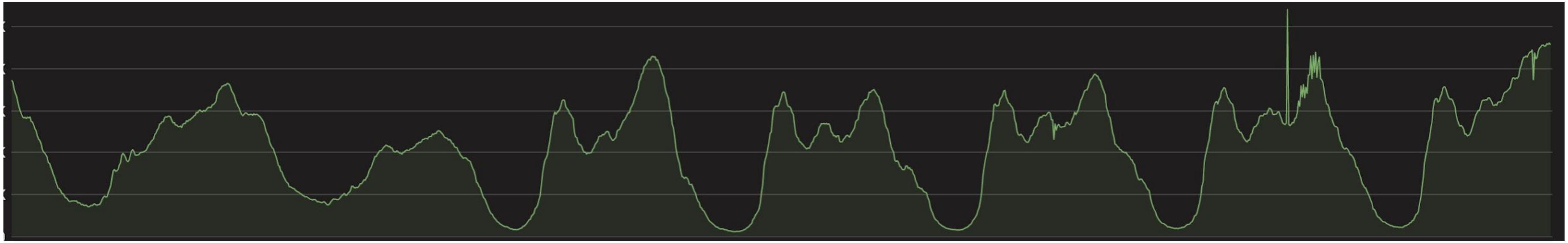
Compounded challenges

High Cardinality

The Scale of the Problem

Compounded challenges

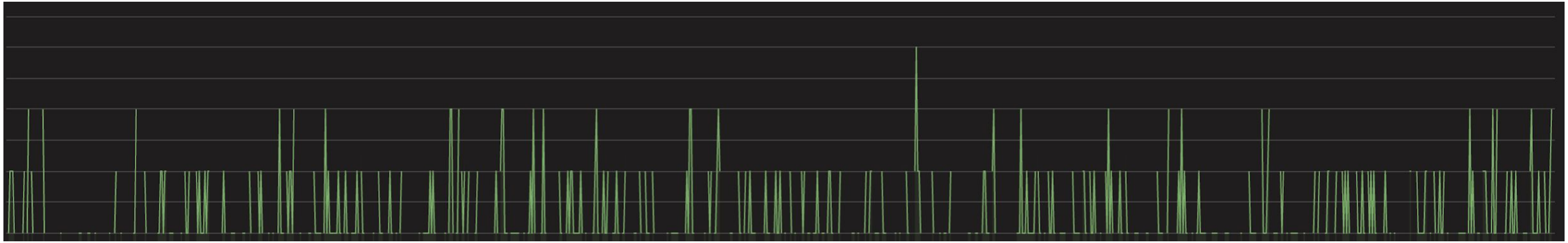
Variety of patterns



The Scale of the Problem

Compounded challenges

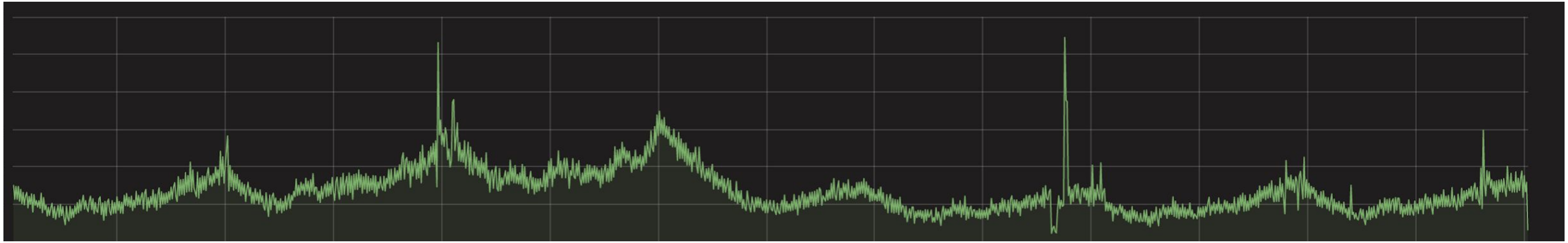
Variety of patterns



The Scale of the Problem

Compounded challenges

Variety of patterns



The Scale of the Problem

Compounded challenges

Variety of patterns

… and others

The Scale of the Problem

Compounded challenges

Speed of detection

1-minute granularity in most situations and whenever possible

Our Solution

The nature of the problem calls for …

Rationale

- Data rich situation
- Complex patterns
- Interrelated inputs
- Necessity of automation and speed

Our Solution

The nature of the problem calls for …

A Machine Learning Platform

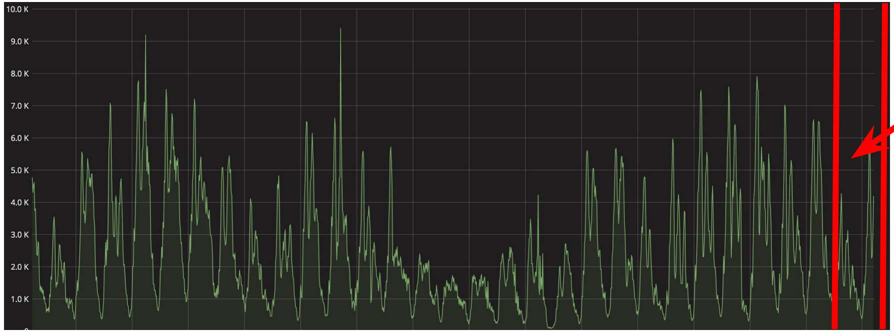
Rationale

- Data rich situation
- Complex patterns
- Interrelated inputs
- Necessity of automation and speed

Anomaly Detection Platform

- At the core, the platform implements a stream of binary classifiers

Detection Window



Is It Anomalous?

Yes

No

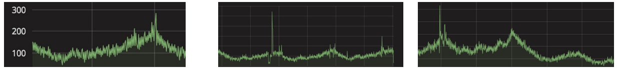
History

+

Additional Inputs…

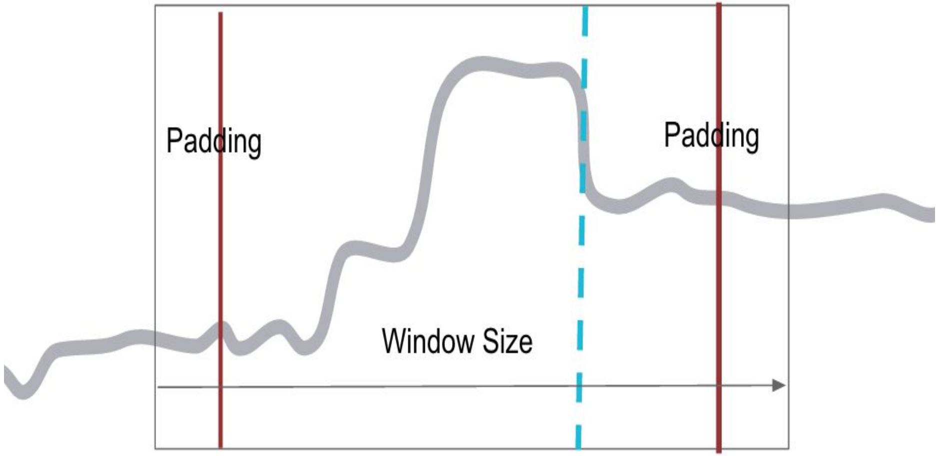
Metainfo (e.g. events)

… or …



Anomaly Detection Stack

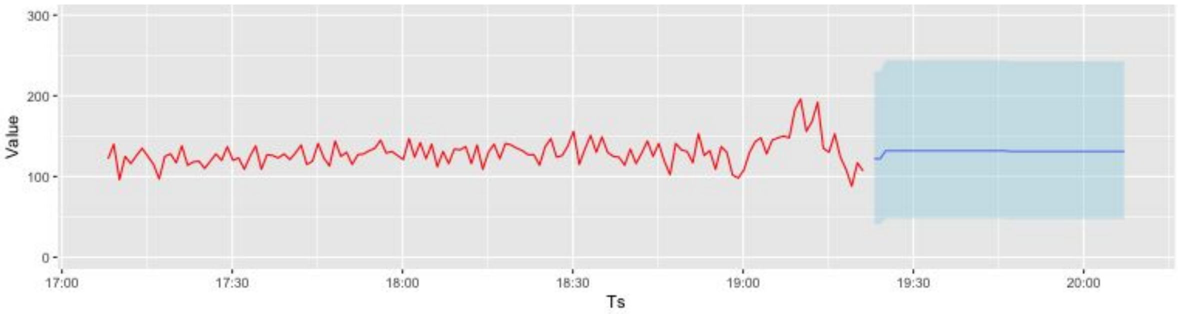
- Some models are indeed waveform binary classifiers



- Backward looking
- Good for new traces
- Does not rely on meta info

Anomaly Detection Stack

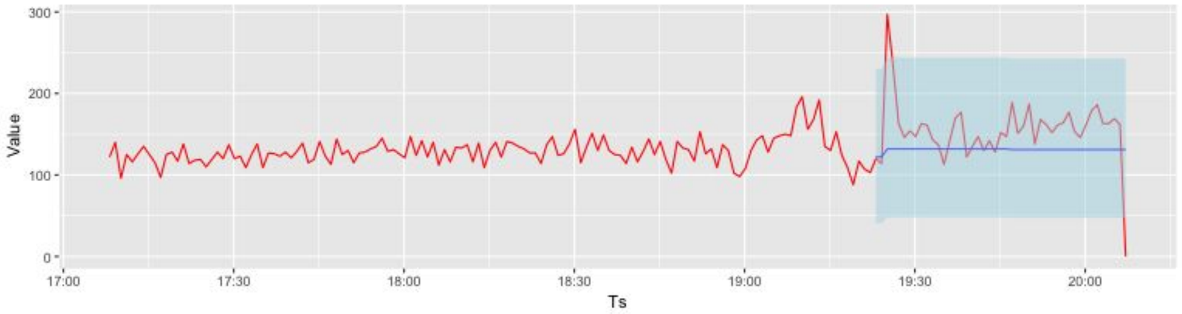
- But most carry out a density forecast behind the scenes



- Learn from the past
- Forecast our expectations…
- … and our uncertainty

Anomaly Detection Stack

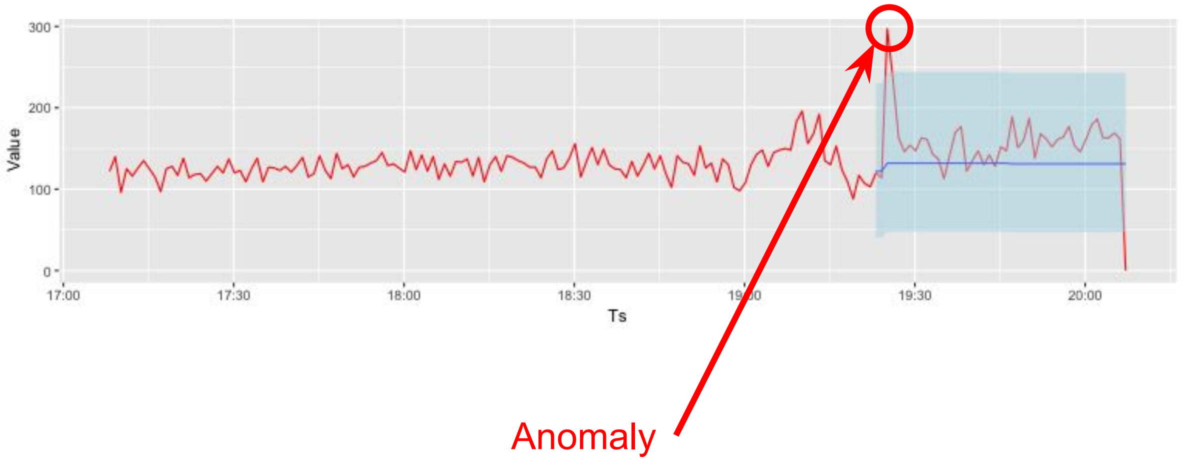
- But most carry out a density forecast behind the scenes



- Learn from the past
- Forecast our expectations…
- … and our uncertainty
- Compare with the actuals

Anomaly Detection Stack

- But most carry out a density forecast behind the scenes



- Learn from the past
- Forecast our expectations…
- … and our uncertainty
- Compare with the actuals

Anomaly Detection Stack

- Two types of forecasting models

Distinguished by type of input and by how they learn:

- Single time-series models
  - Trained online
- Models that learn across multiple time series
  - Training is slower

The Serving Layer

Even when the models require extensive training, serving needs to be rapid

A Golang Serving Layer

for speed and maximum integration with Uber's stack.

Review of Forecasting Methods

Many methodologies for time series forecasting

- Traditional models:
  - Moving Average (MA),
  - Autoregression (AR),
  - ARMA,
  - Etc.
- Exponential smoothing family:
  - Exponential smoothing
  - Holt-Winters
- Decomposition-based models:
  - Theta method
  - Spline regression
  - Prophet
- Proprietary models

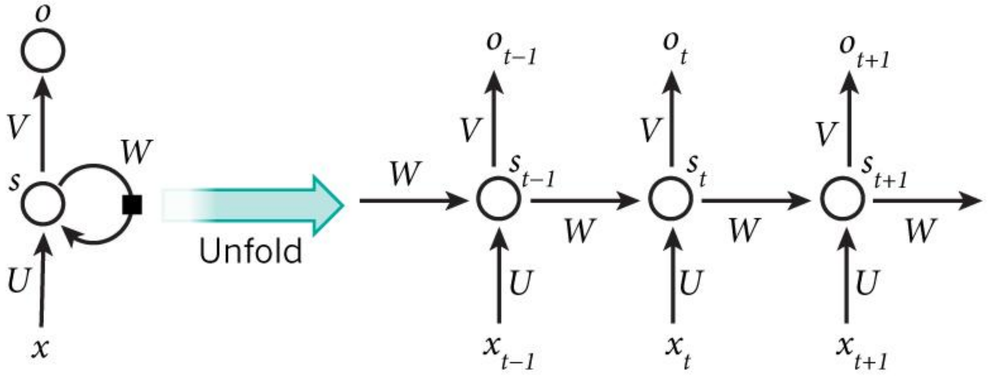
Forecasting with Neural Networks

Use recurrent neural network forecasting

- Capable of dealing with huge amounts of data
- Has some memory of the past
- Not just univariate, could make use of other features
- Neural network could adopt many model shapes

Recurrent Neural Networks

- Inputs are sequential
  - Apply to cases like language processing, time series, etc
- Model has some memory of the past
  - Remember previous look-back steps



x_t : input

s_t : hidden state,

which is usually function of x_t and s_{t-1}

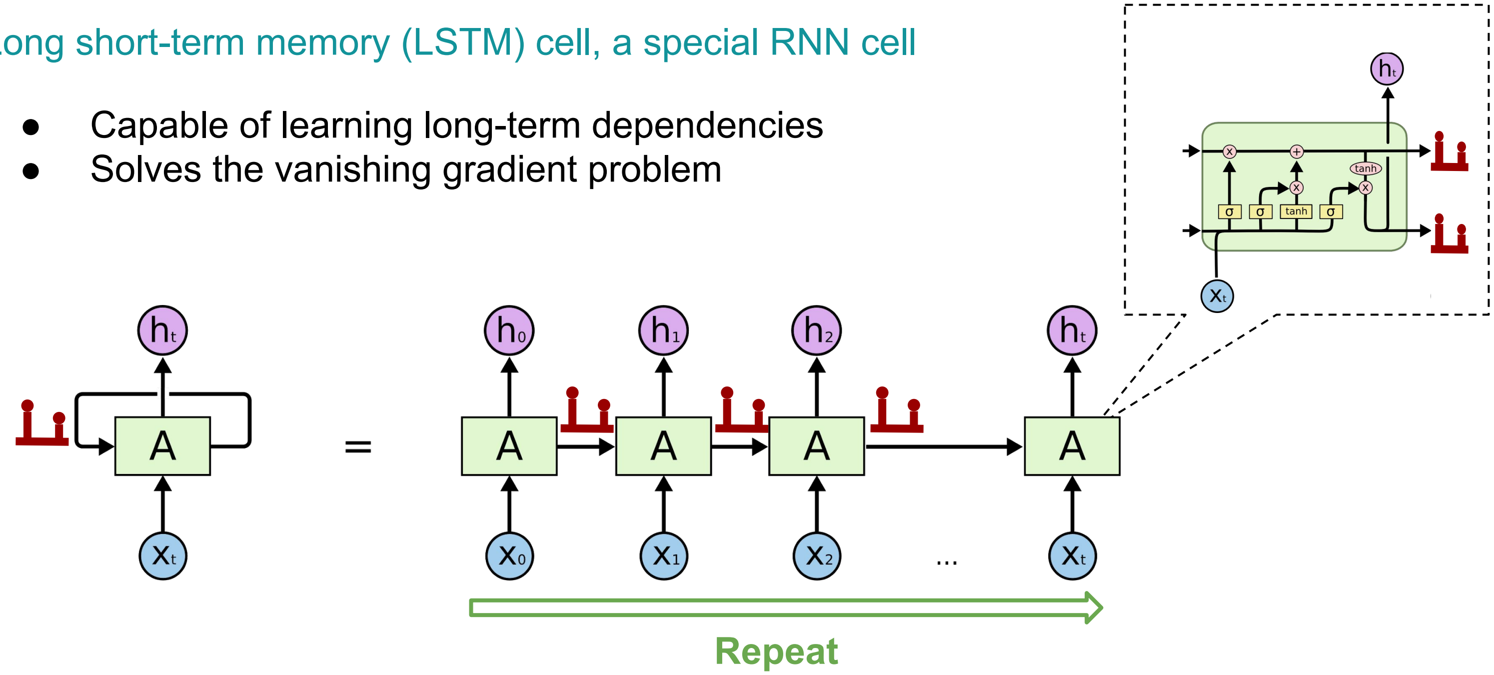
o_t : output

Plots from: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

Recurrent Neural Networks

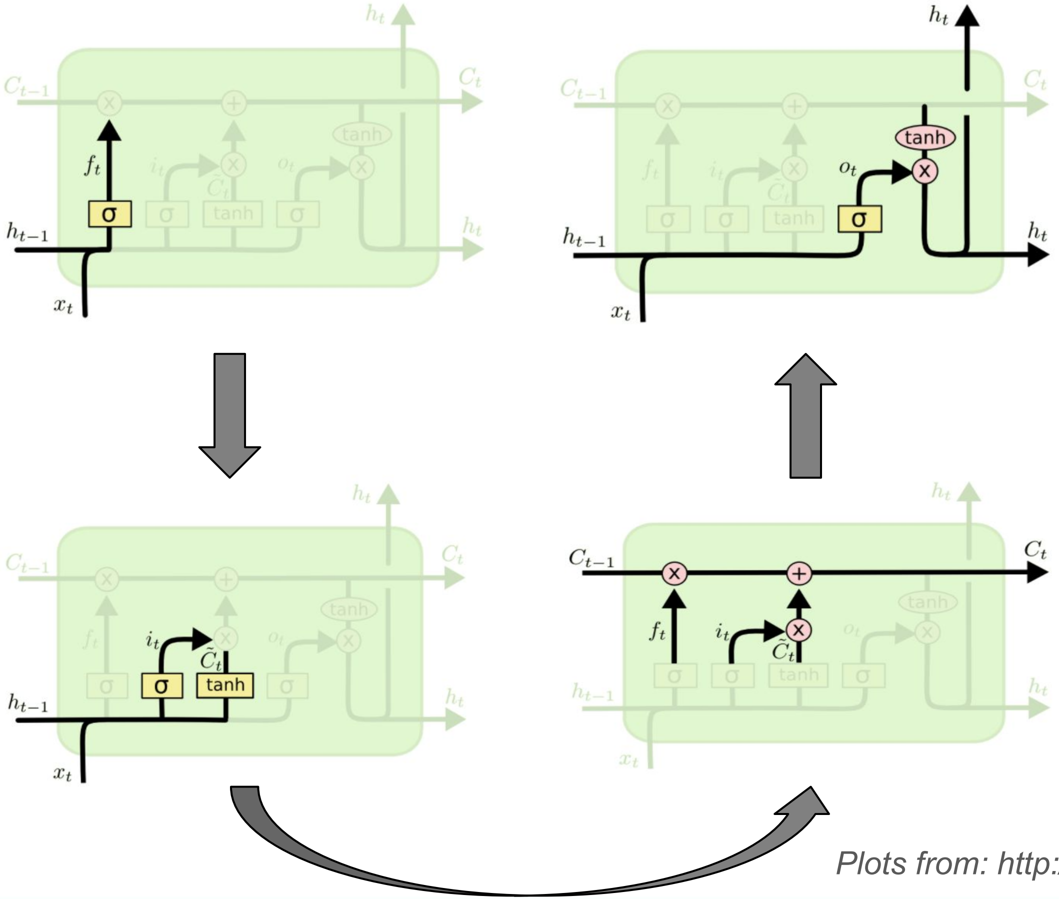
Long short-term memory (LSTM) cell, a special RNN cell

- Capable of learning long-term dependencies
- Solves the vanishing gradient problem



Plots from: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Inside the LSTM Cell



Three gates:

- Forget gate
- Input gate
- Output gate

Can accommodate both long and short term memory

- Selective memory

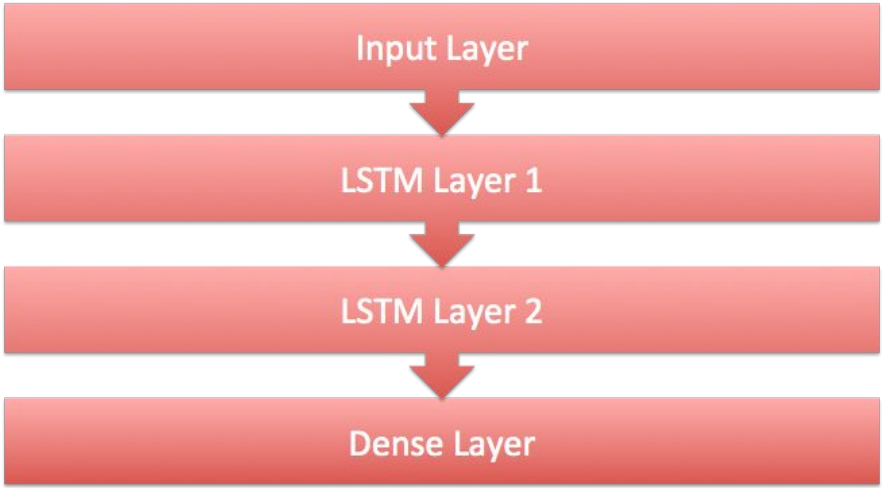
Plots from: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Strata
DATA CONFERENCE

Forecasting with Recurrent Neural Networks

Model

- Two LSTM layers and one dense layer
- Window-wide scaling of input and output
- Adam optimization
- Minimizing absolute error instead of squared error
- Decaying learning rate



Scaling Inputs and Outputs

Window-wide scaling of input and output

Min-max range scale

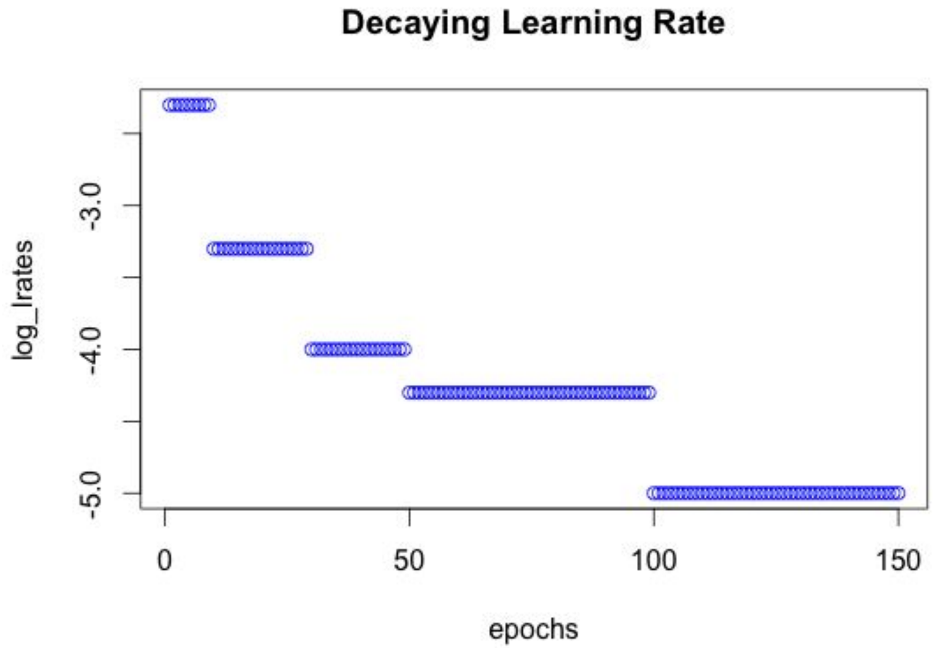
Single window

The entire time series

Learning Rate

Decaying learning rate

- Learning rate
  - Decay by epoch
  - Decay rate becomes constant after 100 epochs



Training Input and Output

V0

Input

- Multiple time series
  - Time series of different topics
  - Minute tile
  - Treated as different samples
- Look back
  - One day
- Features
  - Last 30 minutes

Output

- Next 30 minutes

V1

Input

- Same as before except for
- Features
  - Last 30 minutes **+ last week same time as prediction window**

Output

- Next 30 minutes

Model Performance

- Performance measured out-of-sample
- Each example predicts 30 minutes ahead

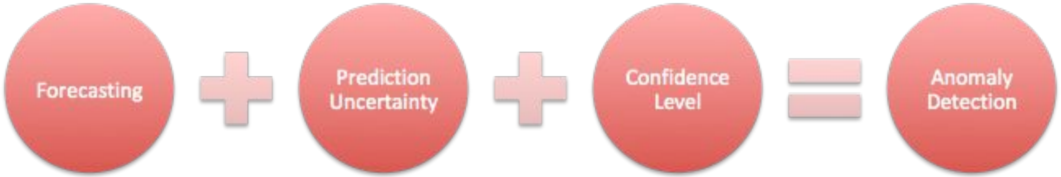
V0	wMAPE	sMAPE
Median	7.18	6.98
Mean	27.64	18.99

V1	wMAPE	sMAPE
Median	7.38	6.60
Mean	25.06	18.02

wMAPE: weighted mean absolute percentage error
sMAPE: symmetric mean absolute percentage error

Anomaly Detection using RNN

- With forecasting, we still need to
  - Decide on the desired level of confidence
  - Estimate prediction interval at a given confidence level
- Choose confidence level to adjust sensitivity
- Next let's focus on prediction interval at a given confidence level



What's the Prediction Interval?

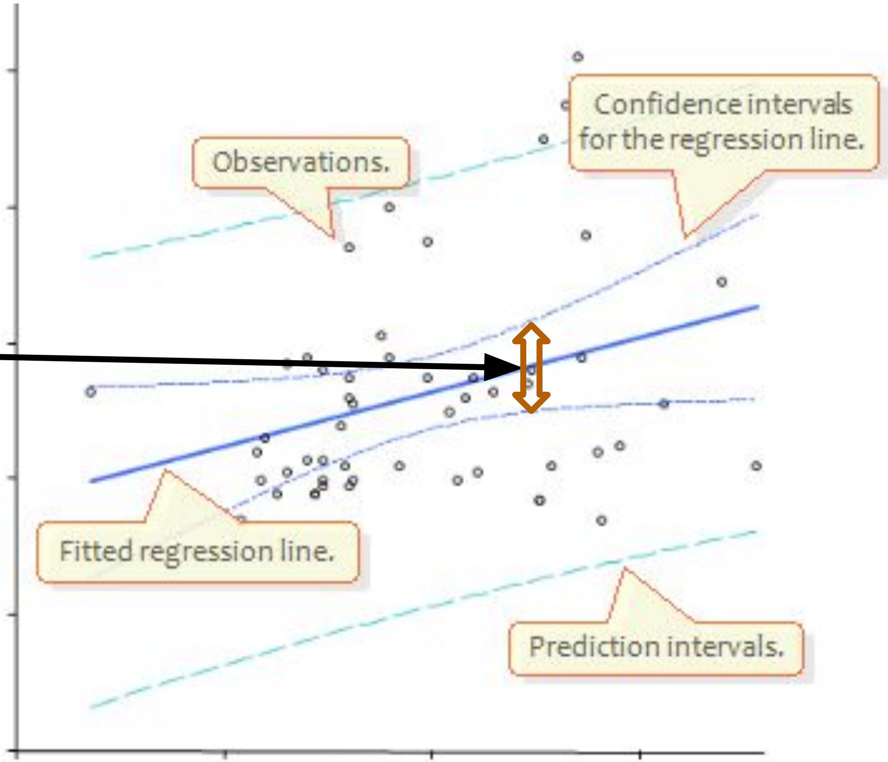
Prediction intervals quantify prediction uncertainty. What do we mean by uncertainty?

- Model uncertainty
  - Our ignorance of the model parameters
- Inherent noise
  - Irreducible noise level from the random process

Prediction uncertainty

Prediction Uncertainty

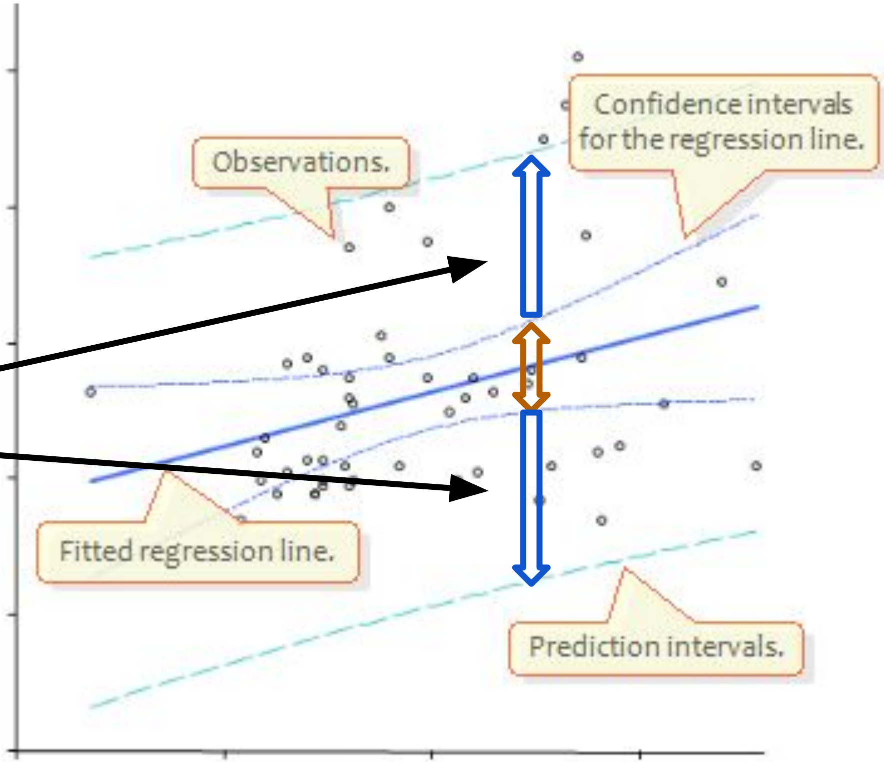
- ❑ Model uncertainty
- ❑ Inherent noise



Plot from: https://analyse-it.com/docs/220/standard/multiple_linear_regression.htm

Prediction Uncertainty

- ❑ Model uncertainty
- ❑ Inherent noise



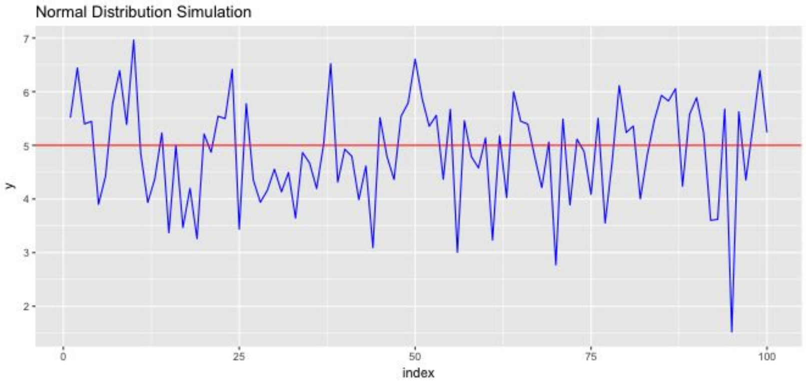
Plot from: https://analyse-it.com/docs/220/standard/multiple_linear_regression.htm

Strata DATA CONFERENCE

Estimating Inherent Noise

What does this noise mean?

- Uncertainty produced even if we know the true underlying distribution
- Generate 100 data from normal (5, 1) distribution
    - Y = 5 + ε where ε is normal(0, 1)
    - Model is identity * 5 and no variance
    - There’s still inherent noise in ε



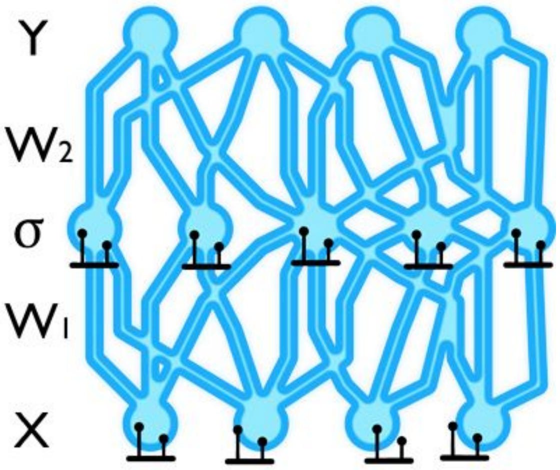
How to estimate noise?

- One possible way: compute residual sum of squares (RSS) to estimate noise

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2$$

Estimating Model Uncertainty

Random dropout during serving



Model uncertainty

Input:

X*, dropout probability p, repetition T=500

Algorithm:

1. Repeat T stochastic feed-forward passes
2. Collect predictions $Y_1, \ldots, Y_T$

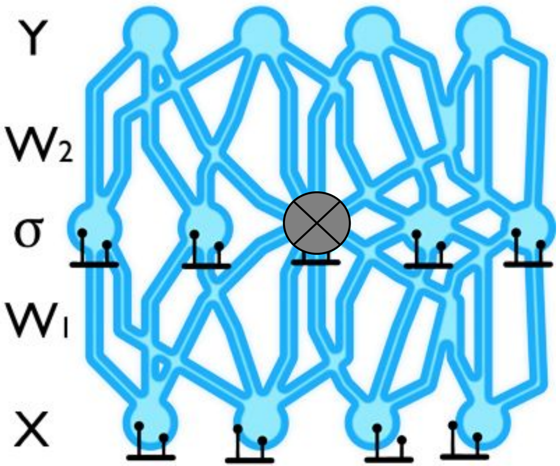
Output:

Sample variance $\boldsymbol{\sigma}_M^2$

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plot: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Estimating Model Uncertainty

Pass 1



Model uncertainty

Input:

X*, dropout probability p, repetition T=500

Algorithm:

1. Repeat T stochastic feed-forward passes
2. Collect predictions $Y_1, \ldots, Y_T$

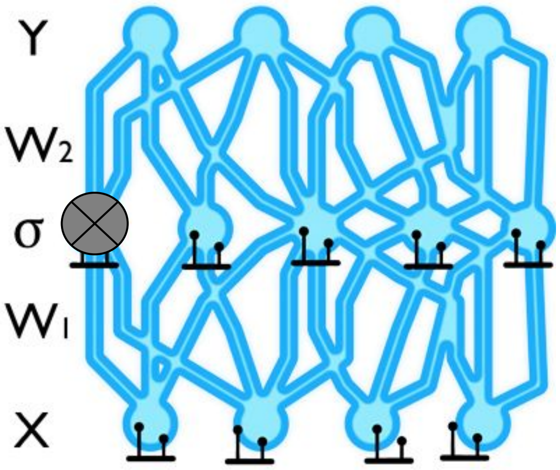
Output:

Sample variance $\boldsymbol{\sigma}_M^{\ 2}$

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plot: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Estimating Model Uncertainty

Pass 2



Model uncertainty

Input:

X*, dropout probability p, repetition T=500

Algorithm:

1. Repeat T stochastic feed-forward passes
2. Collect predictions $Y_1, \dots, Y_T$

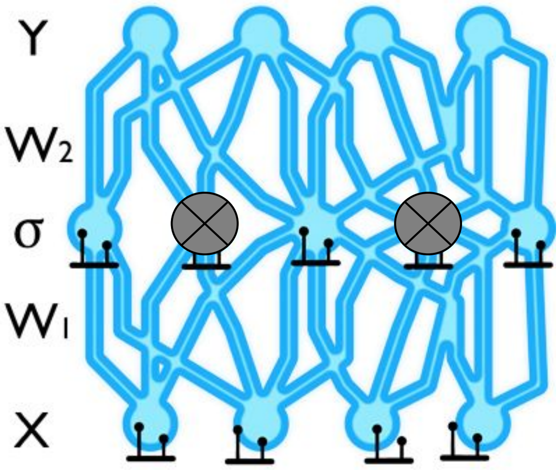
Output:

Sample variance $\boldsymbol{\sigma}_M^2$

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plot: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Estimating Model Uncertainty

Pass 3



Model uncertainty

Input:

X*, dropout probability p, repetition T=500

Algorithm:

1. Repeat T stochastic feed-forward passes
2. Collect predictions $Y_1, \ldots, Y_T$

Output:

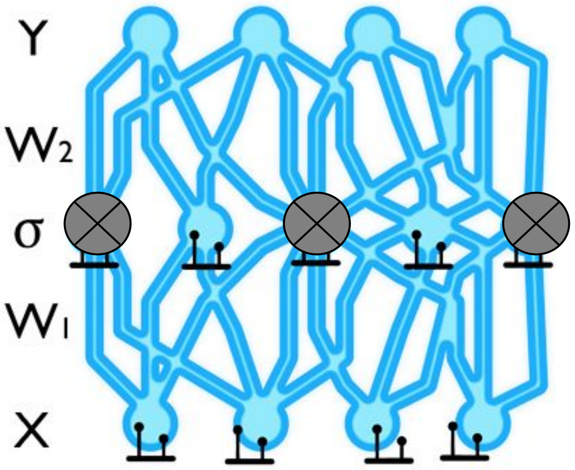
Sample variance $\boldsymbol{\sigma}_M^2$

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plot: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Strata DATA CONFERENCE

Estimating Model Uncertainty

Pass 4



Model uncertainty

Input:

X*, dropout probability p, repetition T=500

Algorithm:

1. Repeat T stochastic feed-forward passes
2. Collect predictions $Y_1, \ldots, Y_T$

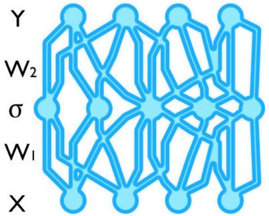
Output:

Sample variance $\boldsymbol{\sigma}_M^2$

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plot: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Flow of Forecasting and Uncertainty Estimation

Step 1 — **Training data** ⇒ **Train** the neural network

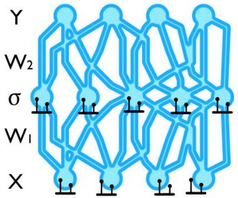


Fit the weights

Step 2 — **Validation data** ⇒ Use the trained model to estimate **noise level**

Inherent noise estimate

Step 3 — **New data** ⇒ Obtain **prediction** and **model uncertainty**



Repeat random dropout

Prediction uncertainty

Methodology: Gal (2016), Uncertainty in Deep Learning, PhD Thesis
Plots: http://mlg.eng.cam.ac.uk/yarin/blog_3d801aa532c1ce.html

Forecasting Daily Trips with Uncertainty

Input

- Look back
  - 28 days
- Features
  - Trip value
  - Holiday info
  - Calendar features

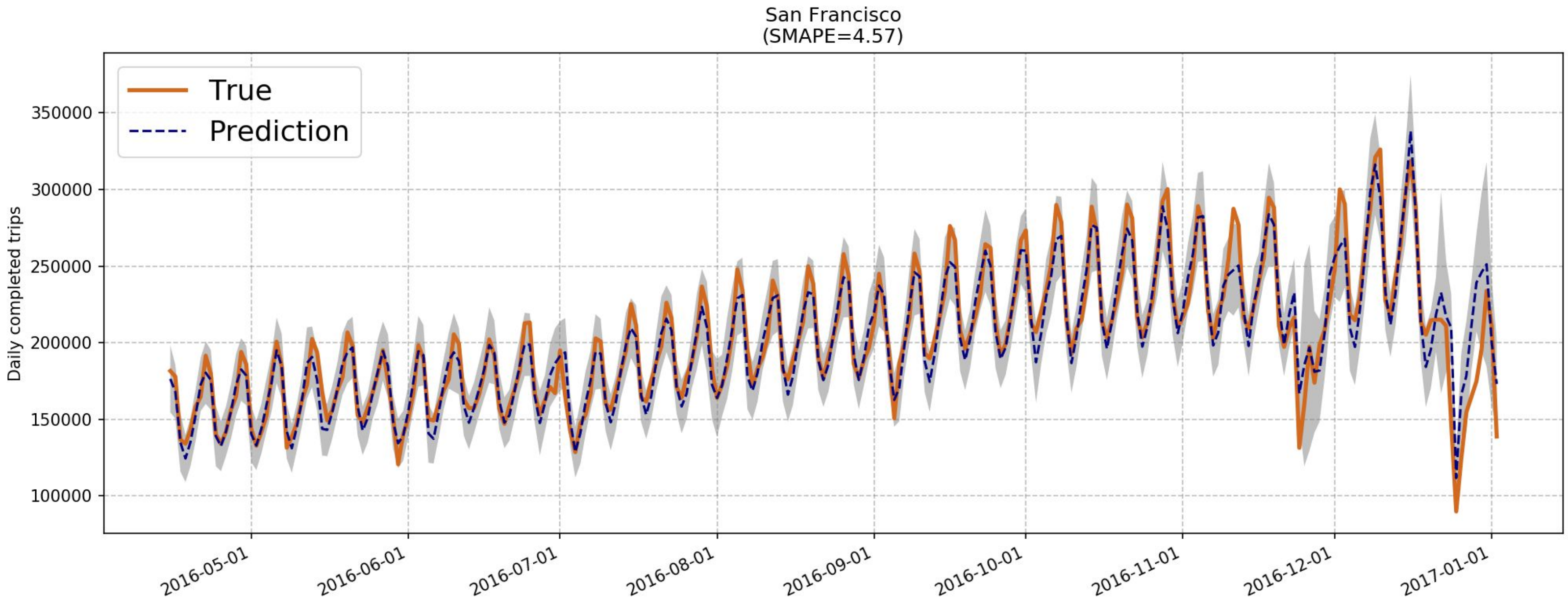
Output

  - Next 5 days

Uber Blog: Engineering Uncertainty Estimation in Neural Networks for Time Series Prediction at Uber

Forecasting Daily Trips with Uncertainty

Prediction with 95% prediction interval



Uber Blog: Engineering Uncertainty Estimation in Neural Networks for Time Series Prediction at Uber

Future Developments

Model Improvements

- Truncated backpropagation through time
  - Longer memory without vanishing gradients
- More feature engineering
  - Summary features: e.g. mean or quantiles
- Additional methods to deal with seasonality within NNs
  - Calendar features: hour of day, day of week
  - Per hour of day/week models
- Transfer learning

Thank you!

Any questions?

Learn more about Anomaly Detection at UBER!

- Engineering Uncertainty Estimation in Neural Networks for Time Series Prediction at Uber
- Engineering Extreme Event Forecasting at Uber with Recurrent Neural Networks
- Anomaly Detection
- Identifying Outages with Argos, Uber Engineering's Real-Time Monitoring and Root-Cause Exploration Tool

Proprietary and confidential © 2017 Uber Technologies, Inc. All rights reserved. No part of this document may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval systems, without permission in writing from Uber. This document is intended only for the use of the individual or entity to whom it is addressed and contains information that is privileged, confidential or otherwise exempt from disclosure under applicable law. All recipients of this document are notified that the information contained herein includes proprietary and confidential information of Uber, and recipient may not make use of, disseminate, or in any way disclose this document or any of the enclosed information to any person other than employees of addressee to the extent necessary for consultations with authorized personnel of Uber.