Machine Learning @ Amazon

Ralf Herbrich
Amazon

Background


1992 – 1997 (Berlin, Diploma)


1997 – 2000 (Berlin, PhD)


2000 – 2009 (Microsoft Research)


2009 – 2011 (Microsoft)


2011 – 2012 (Facebook)


2012 – Present (Amazon)

Overview

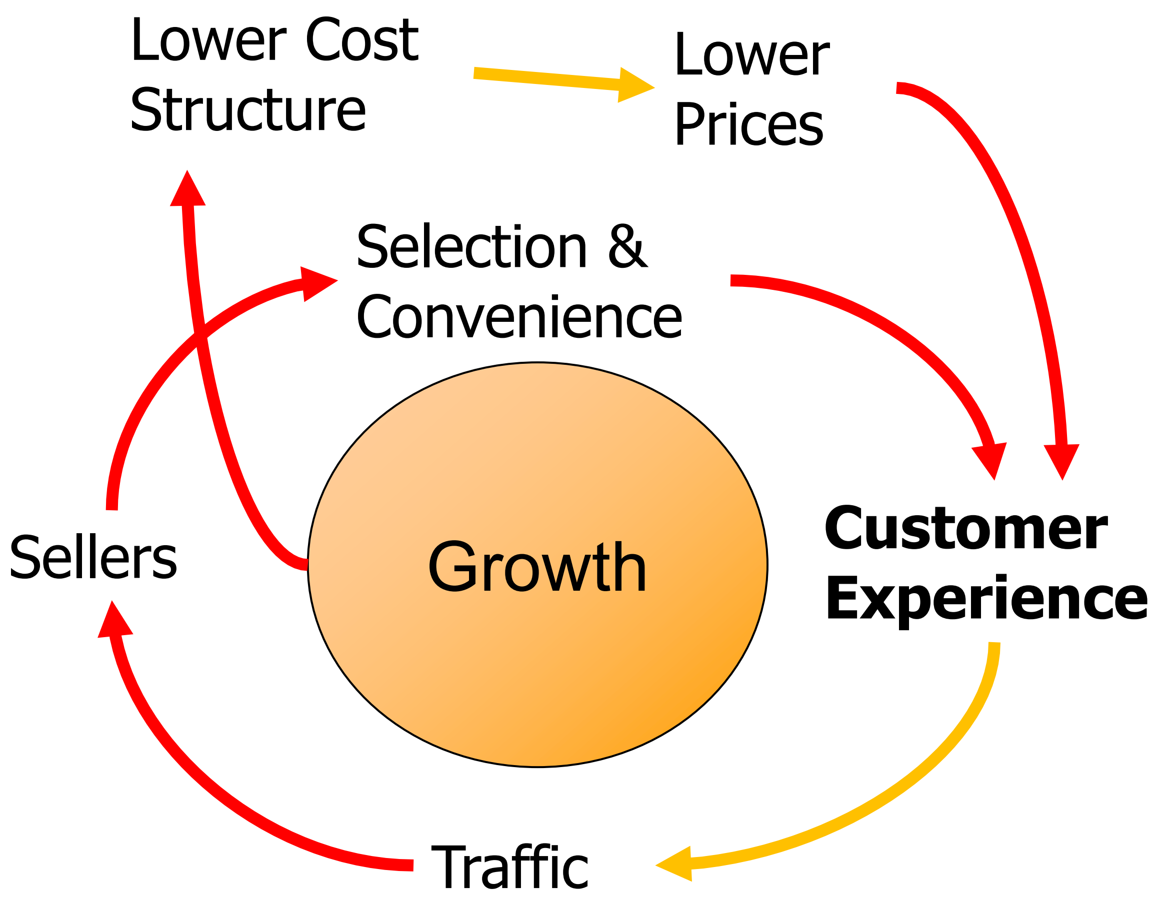
- **What is Amazon?**
- Machine Learning in Practise
  - Probabilities
  - Finite Resource
- Machine Learning @ Amazon
  - Forecasting
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Our Customers

Devices


amazon
kindle
kindle
Shop Games Apps Books Music Videos Newsstand Web Photos Docs
LIFE OF PI
amazon

Amazon's Virtuous Cycles



Overview

- What is Amazon?
- **Machine Learning in Practise**
  - Probabilities
  - Finite Resource
- Machine Learning @ Amazon
  - Forecasting
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Artificial Intelligence and Machine Learning

Science

- Computer Science
- Statistics
- Neuroscience
- Operations Research

Artificial Intelligence

- Knowledge Representation
- Knowledge Extraction
- Reasoning
- Planning

Machine Learning

- Rule Extraction from Past (Training)
- Forecast of Future (Prediction)
- Taking Actions Now (Decision Making)

Overview

- What is Amazon?
- Machine Learning in Practise
  - **Probabilities**
  - Finite Resource
- Machine Learning @ Amazon
  - Forecasting
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Machine Learning: Formal Definition

- Labelled Data

$$\{P(y|x,w)\}_{w\in\mathcal{W}} + \{(x_i,y_i)\}_{i=1}^{n} \mapsto (x \mapsto P(y|x))$$

- Unlabelled Data

$$\{P(x|z,w)\}_{w\in\mathcal{W}} + \{x_i\}_{i=1}^{n} \mapsto (x \mapsto P(z|x))$$

- Probability is a central concept in Machine Learning!

Why Probability?

1. **Mathematics of Uncertainty (Cox' axioms)**

Cox Axioms: Probabilities and Beliefs

- Design: System must assign degree of plausability $p(A)$ to each logical statement A.
- Axiom:
  - $p(A)$ is a real number
  - $p(A)$ is independent of Boolean rewrite
  - $$\begin{aligned} p(A|C') > p(A|C) \quad &\wedge \quad p(B|AC') = p(B|AC) \\ &\Rightarrow \quad p(AB|C') \geq P(AB|C) \end{aligned}$$

P must be a probability measure!

Why Probability?

1. **Mathematics of Uncertainty (Cox' axioms)**

2. **Variables and Factors map to Memory & CPU**

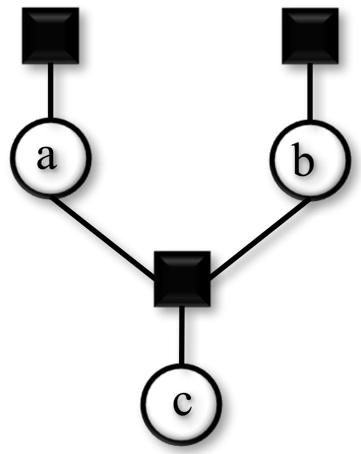
Factor Graphs

- **Definition:** Graphical representation of product structure of a function (Wiberg, 1996)
  - Nodes: = Factors = Variables
  - Edges: Dependencies of factors on variables.

- **Semantic:**

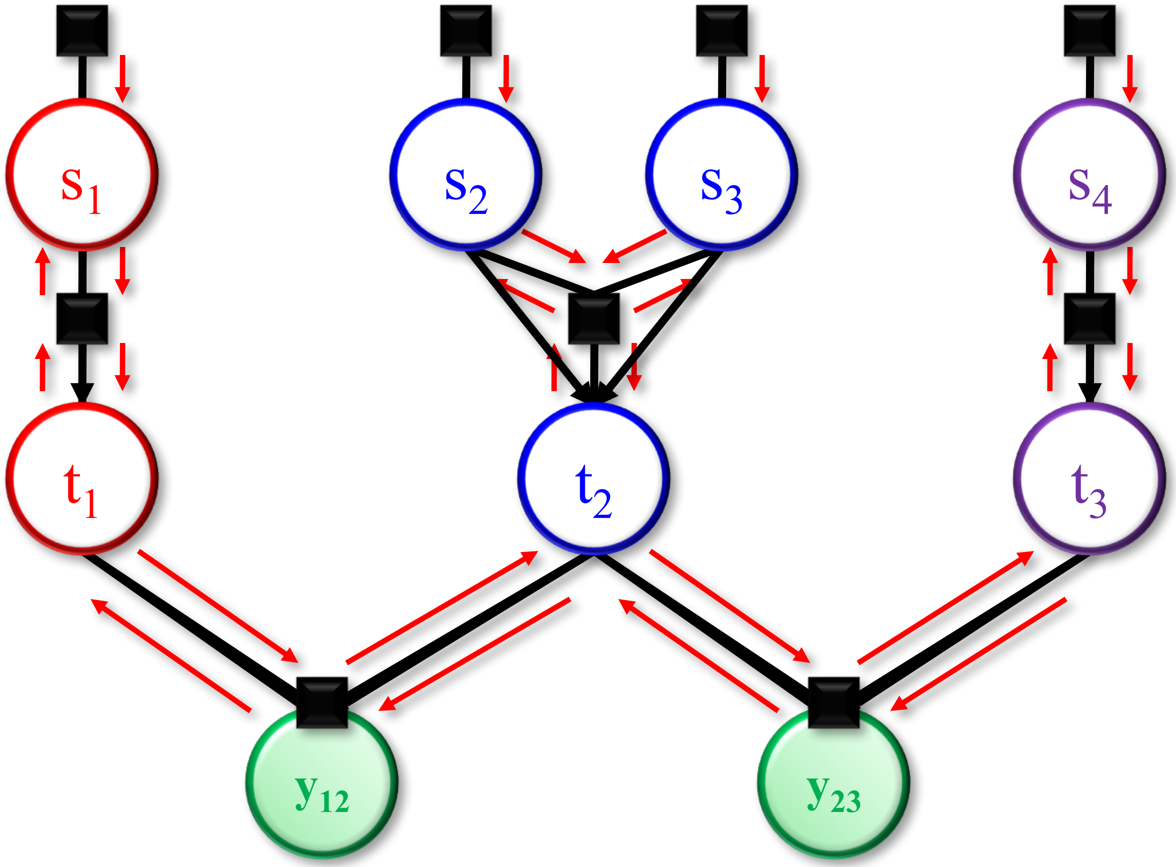
$$p(\mathbf{x}) = \prod_f f\left(\mathbf{x}_{V(f)}\right)$$

  - Local variable dependency of factors



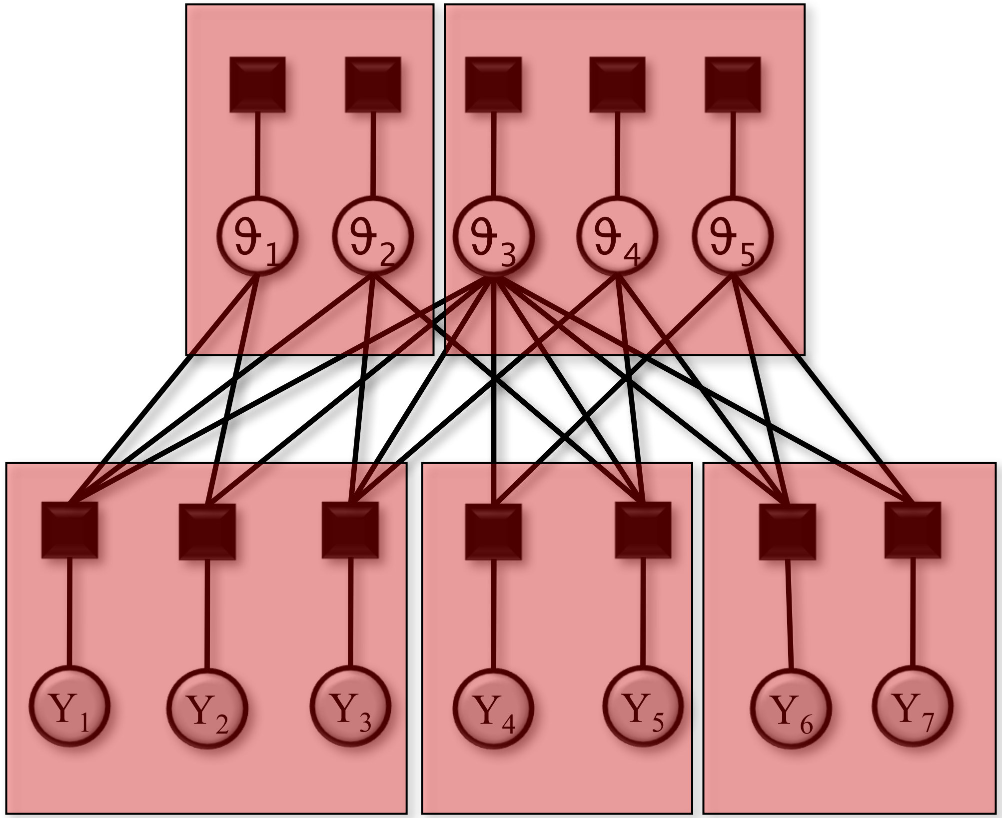
$$p(a,b,c) = f_1(a) \cdot f_2(b) \cdot f_3(a,b,c)$$

Inference in a Factor Graph



Factor Graphs and Cloud Computing

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto \prod_i p(y_i|\boldsymbol{\theta}, \mathbf{x}_i) \cdot \prod_j p(\theta_j)$$



Belief Store
("Memory")

Message Passing
("Communicate")

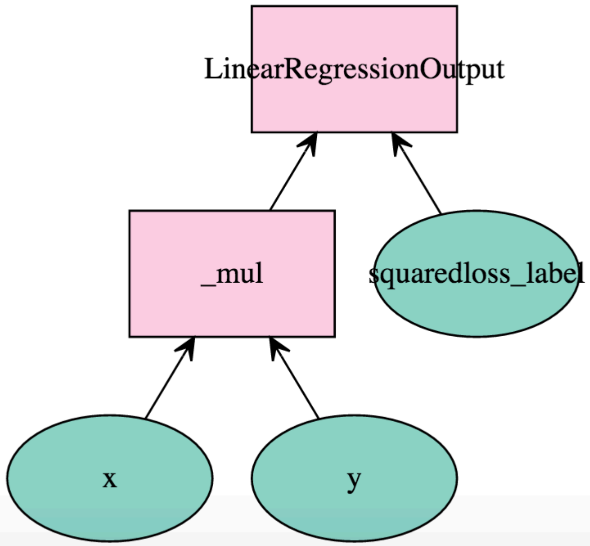
Data Messages
("Compute")

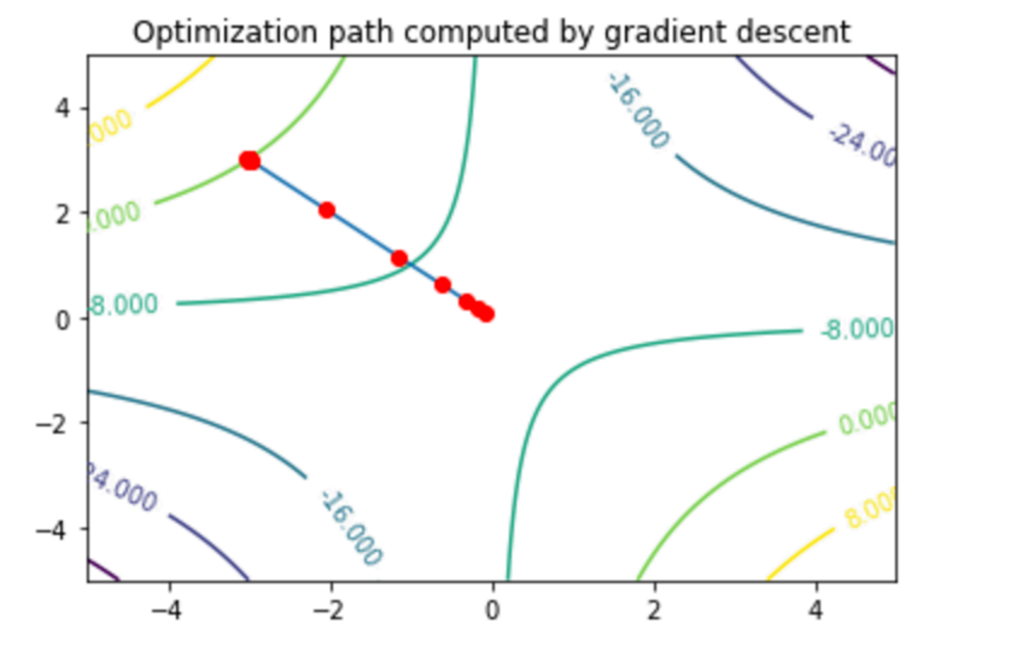
Factor Graphs and MXNet

In [118]:
```python
import mxnet as mx

# A simple network
x = mx.sym.Variable('x')
y = mx.sym.Variable('y')
z = x * y
net = mx.sym.LinearRegressionOutput(z, name='squaredloss')
mx.viz.plot_network(net)
```

Out[118]:





Why Probability?

1. **Mathematics of Uncertainty (Cox' axioms)**

2. **Variables and Factors map to Memory & CPU**

3. **Decouple Data Modeling and Decision Making**

Infer-Predict-Decide Cycle

Inference:
P(Parameters) + Data → **P**(Parameters|Data)

- Requires a (structural) model **P**(Data|Parameters)
- Allows to incorporate prior information **P**(Parameters|Data)

Prediction:
P(Parameters) + Data → **P**(Data)

- Requires integration/summation of parameter uncertainty
- Does not change state!

Decision Making:
Loss(Action,Data) + **P**(Data) → Action

- Business-loss not learning-loss!
- Often involves optimization!

Overview

- What is Amazon?
- Machine Learning in Practise
  - Probabilities
  - **Finite Resource**
- Machine Learning @ Amazon
  - Forecasting
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Finite Resources: Time

- **Real-Time Prediction Service: 5ms**
  - Number of search & ads candidates to rank: 10,000 → 500 ns/candidate
  - Time to read from main memory: 10ns → 50 variables
  - **L1/L2/L3 ranking pipelines to use more complex predictors**
- **Real-Time Learning Service: 1B examples/day**
  - Number of second per day: 86,400 → 86,400 ns/example
  - **Time to write to RAM: 100ns → 864 writes only (excluding disc access!)**
- Sparse models are the result of time constraints!

amazon

Finite Resource: Cost

Economics 101

- Profit = Revenue – Cost
- In the long run, a business that generates negative profits is not viable!

Facebook	2015
Annual Revenue	$17,928,000,000.00*
Daily Revenue	$49,117,808.22
Number of DAU	1,038,000,000**
Number of Story Candidates	1,500***
Number of Daily Stories	1.557E+12
Maximum Cost per Story Candidate	$0.0000315

*http://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/
**http://www.statista.com/statistics/346167/facebook-global-dau/
***https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed

It's power, stupid!

Some constraints might not be obvious: building new datacenters and **powering** them is non-trivial.

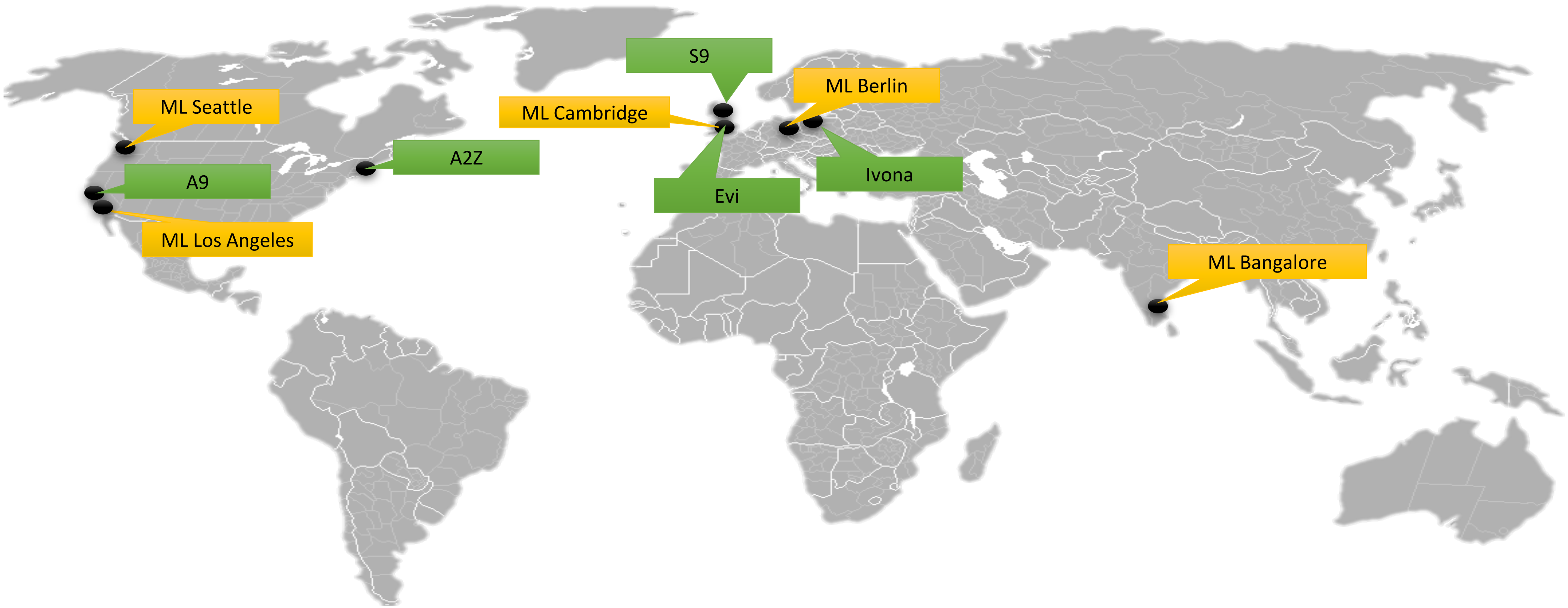
Example: 1 GPU box = 20 CPU boxes
(in terms of power consumption)

amazon

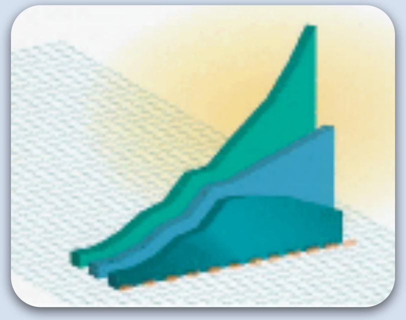
Overview

- What is Amazon?
- Machine Learning in Practise
  - Probabilities
  - Finite Resource
- **Machine Learning @ Amazon**
  - Forecasting
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Locations


S9
ML Berlin
ML Seattle
ML Cambridge
A2Z
A9
Evi
Ivona
ML Los Angeles
ML Bangalore

Machine Learning Opportunities @ Amazon



Retail

- Demand Forecasting
- Vendor Lead Time Prediction
- Pricing
- Packaging
- Substitute Prediction

Customers

- Product Recommendation
- Product Search
- Visual Search
- Product Ads
- Shopping Advice
- Customer Problem Detection

Seller

- Fraud Detection
- Predictive Help
- Seller Search & Crawling

Catalog

- Browse-Node Classification
- Meta-data validation
- Review Analysis
- Hazmat Prediction

Digital

- Named-Entity Extraction
- XRay
- Plagiarism Detection
- Echo Speech Recognition
- Knowledge Acquisiion

Overview

- What is Amazon?
- Machine Learning in Practise
  - Probabilities
  - Finite Resource
- Machine Learning @ Amazon
  - **Forecasting**
  - Machine Translation
  - Visual Systems
- Conclusions and Challenges

Fashion Forecasting (2014 – 2016)

Setting

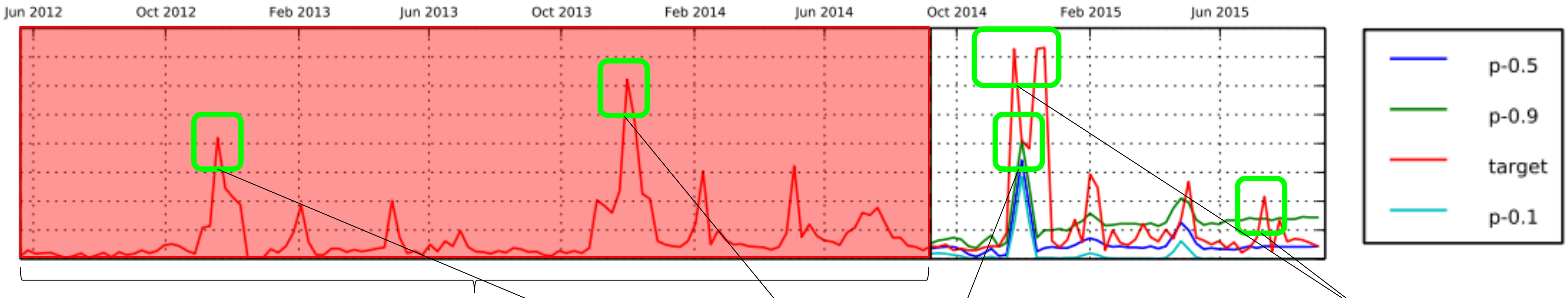
- *Given past sales of a fashion product, predict market demand up to 18 months into the future*

Challenges

- **Sparsity**: Huge skew – many products sell very few items
- **Size**: Variable for Amazon retail in buying but not a variable for end customers
- **Seasonal**: Most fashion products only run for one season – often 12 months in advance
- **Distributions**: Future is uncertain ➔ predictions must be distributions
- **Scale**: 5M+ fashion products in each market ➔ 10^9 training examples
- **Censored**: Past sales ≠ past demand (inventory constraint)

Demand Forecasting

Example fashion product to illustrate the challenges of forecasting.

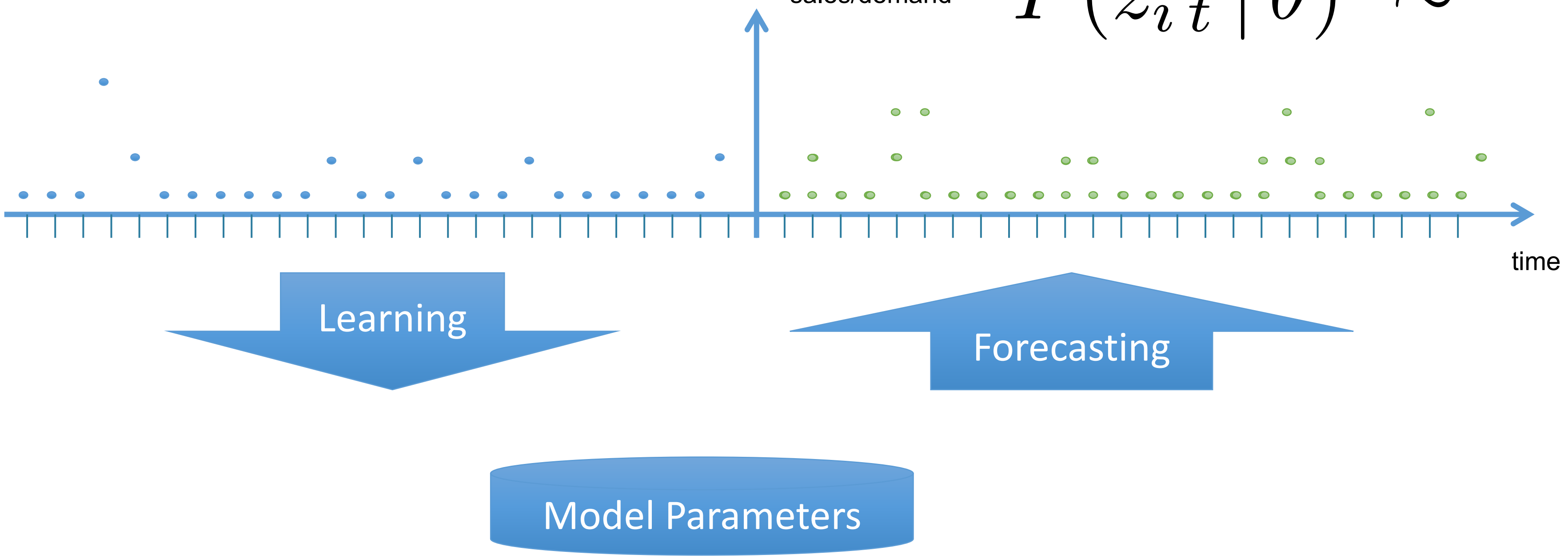


Training Range: Non-fashion items have longer training ranges that we can leverage. Need to information share across new and old products.

Missing Features or Input: Unexplained spikes in demand are likely caused by missing features or incomplete input data.

Seasonality: This item has Christmas seasonality with higher growth over time. This is where we need growth features in addition to date features.

Learning and Prediction



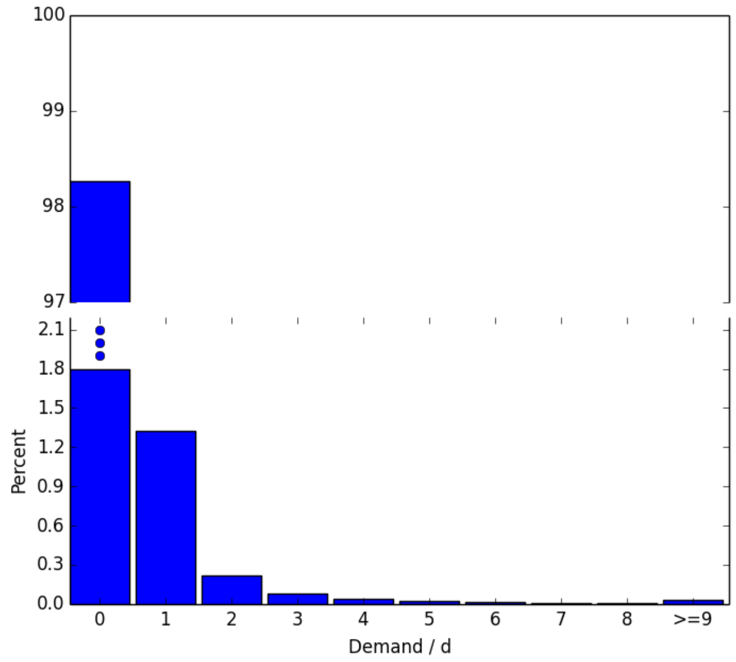
$$P(z_{i\,t} \mid \theta) \;\sim$$

amazon

Slow Moving Inventory

Typical midsize dataset:

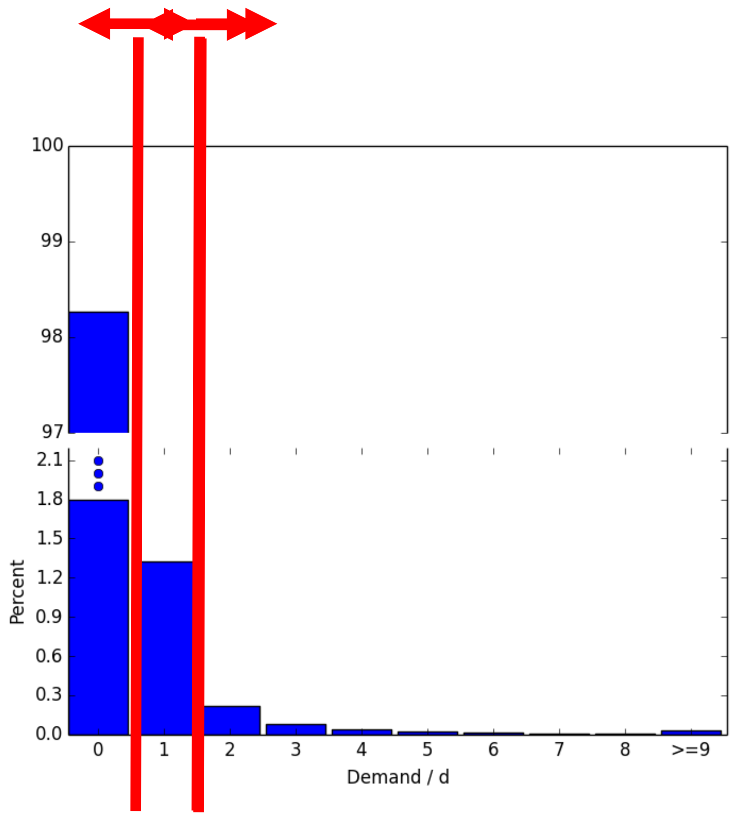
- About 5M items
- About 4.5B item-days
- About 98% zero demand



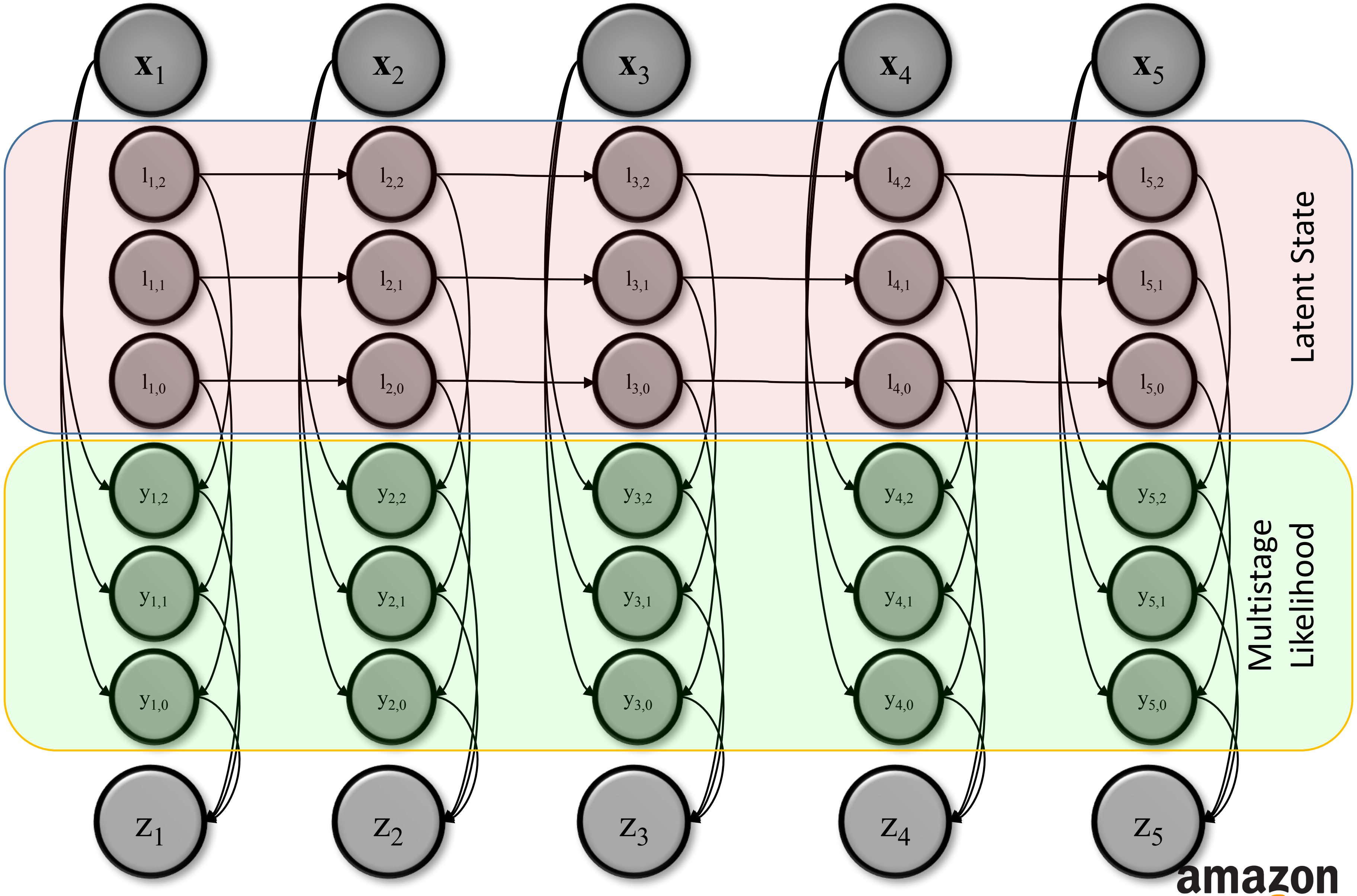
Sampling Predictions

$$P(z_{i\,t} \mid \theta) \;\sim$$

- 0 or ≥1 ?
  Binary classification #1
- 1 or ≥2 ?
  Binary classification #2
- If ≥2:
  Count regression z-2



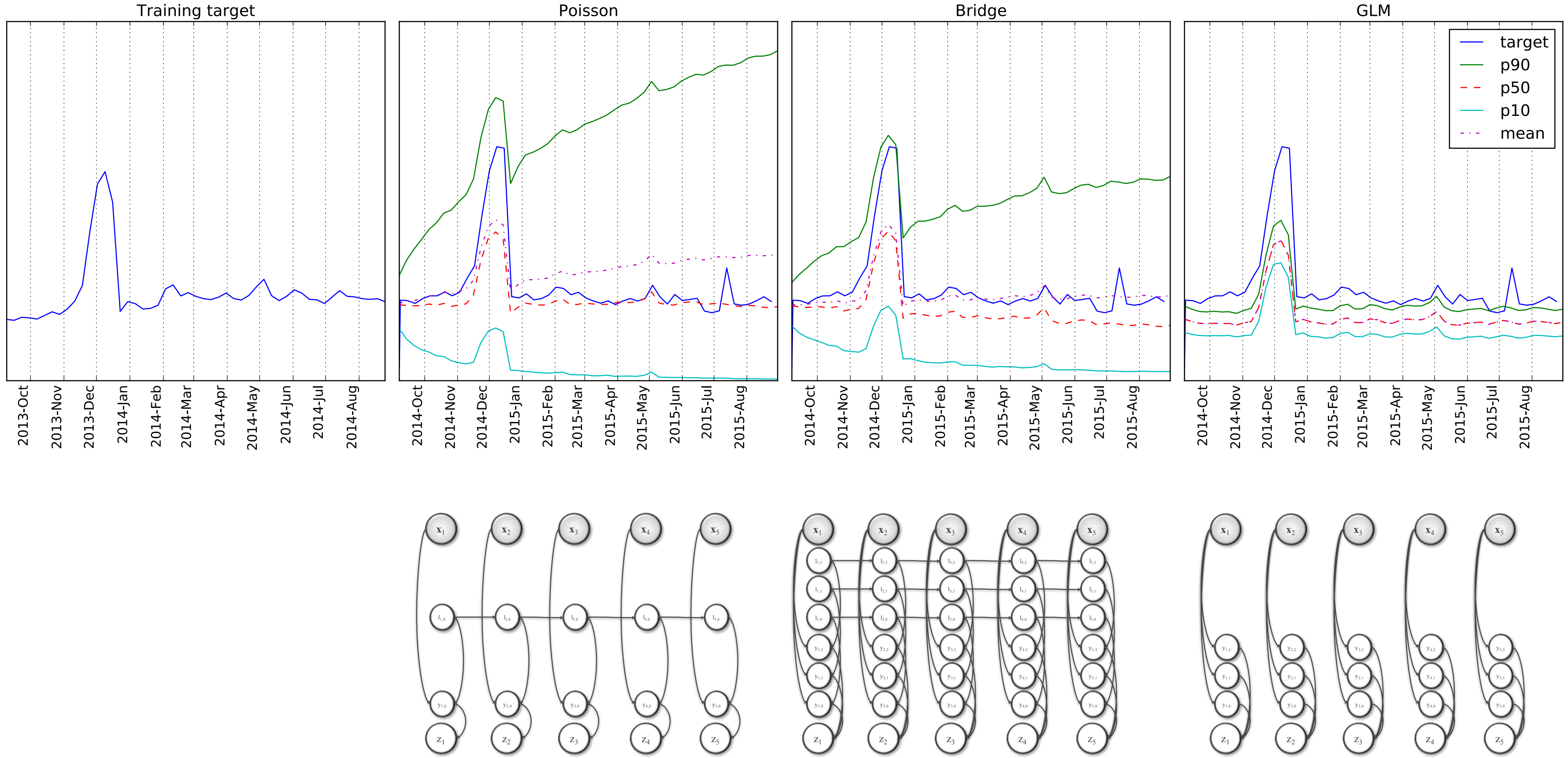
amazon



$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$
$l_{1,2}$ $l_{2,2}$ $l_{3,2}$ $l_{4,2}$ $l_{5,2}$
$l_{1,1}$ $l_{2,1}$ $l_{3,1}$ $l_{4,1}$ $l_{5,1}$
$l_{1,0}$ $l_{2,0}$ $l_{3,0}$ $l_{4,0}$ $l_{5,0}$
Latent State
$y_{1,2}$ $y_{2,2}$ $y_{3,2}$ $y_{4,2}$ $y_{5,2}$
$y_{1,1}$ $y_{2,1}$ $y_{3,1}$ $y_{4,1}$ $y_{5,1}$
$y_{1,0}$ $y_{2,0}$ $y_{3,0}$ $y_{4,0}$ $y_{5,0}$
Multistage
Likelihood
z_1 z_2 z_3 z_4 z_5

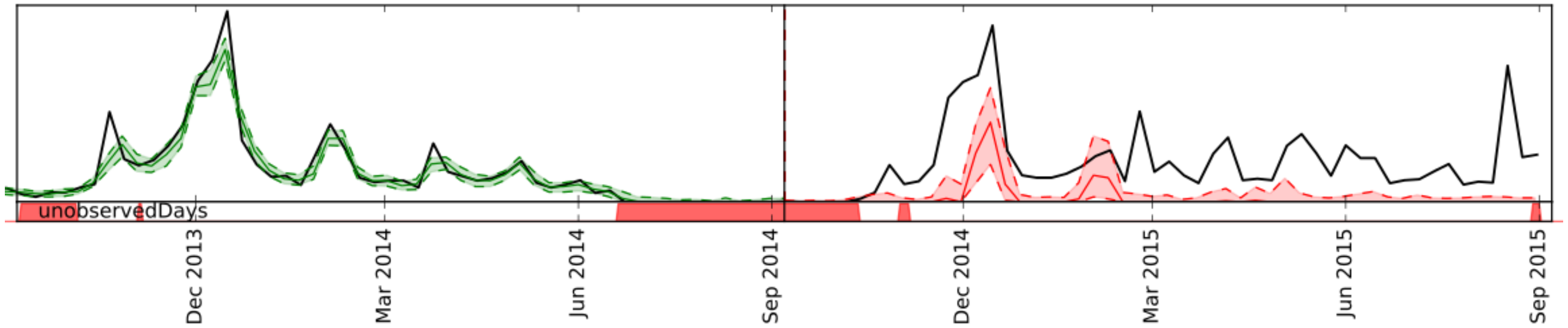
amazon

In Practice

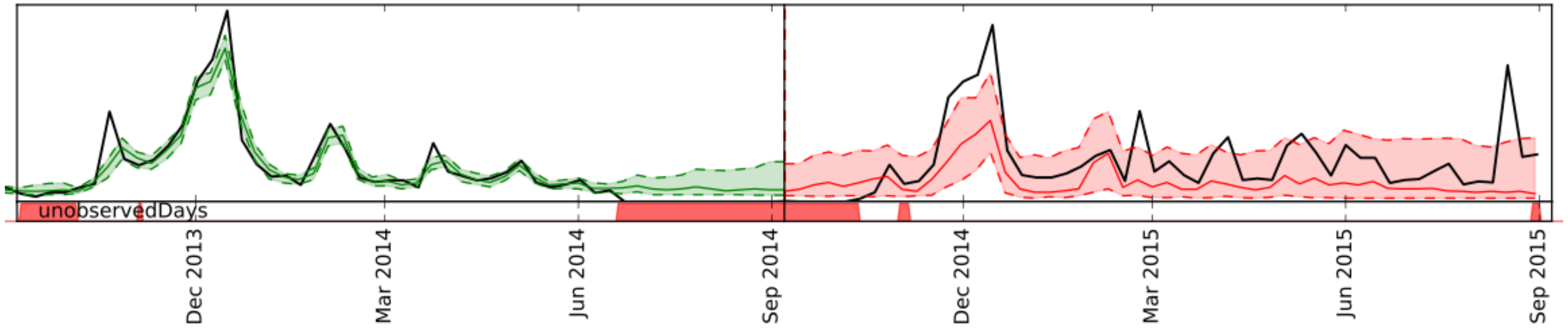

Training target
Poisson
Bridge
GLM
target
p90
p50
p10
mean
2013-Oct
2013-Nov
2013-Dec
2014-Jan
2014-Feb
2014-Mar
2014-Apr
2014-May
2014-Jun
2014-Jul
2014-Aug
2014-Oct
2014-Nov
2014-Dec
2015-Jan
2015-Feb
2015-Mar
2015-Apr
2015-May
2015-Jun
2015-Jul
2015-Aug

Modelling Out of Stock

GLM



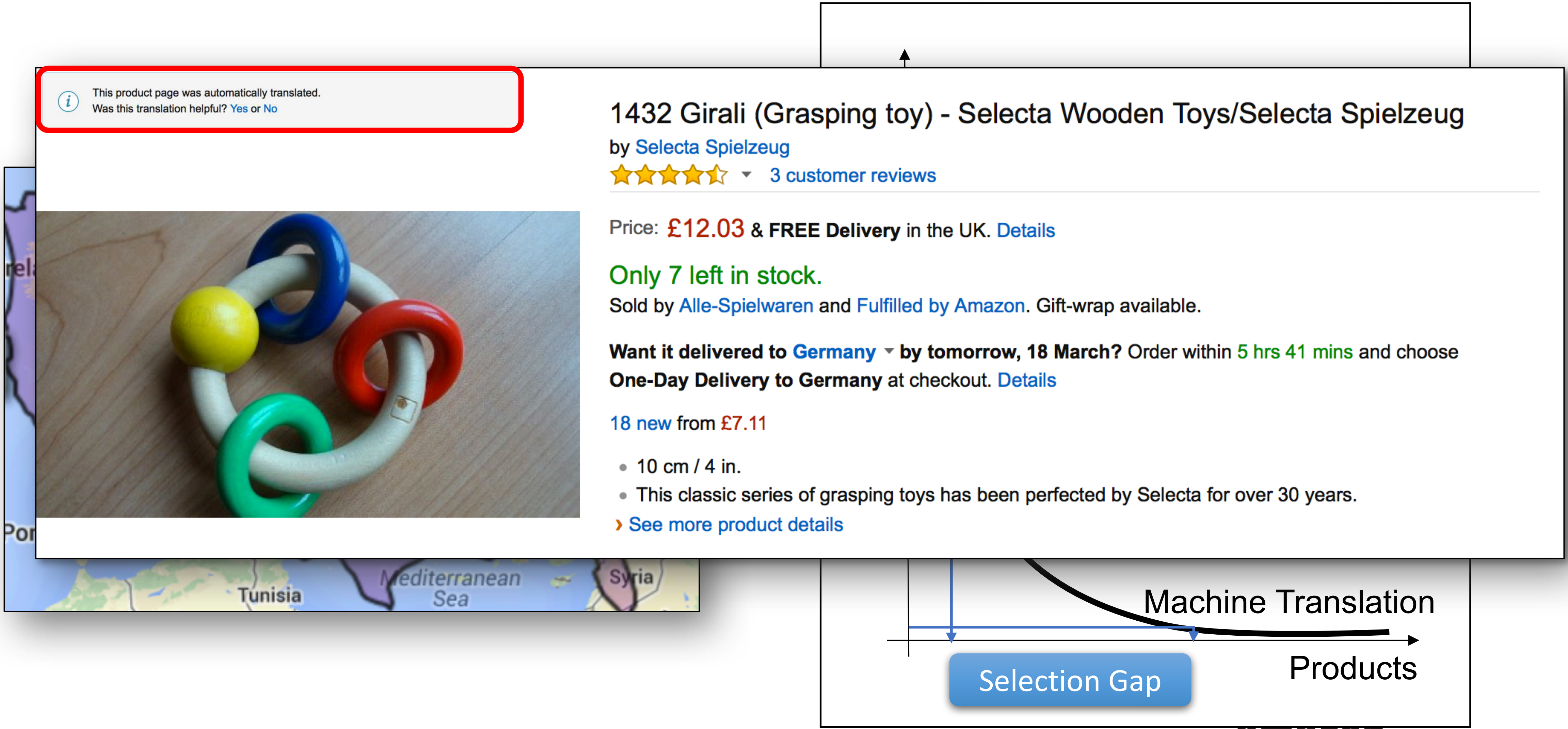
Bridge



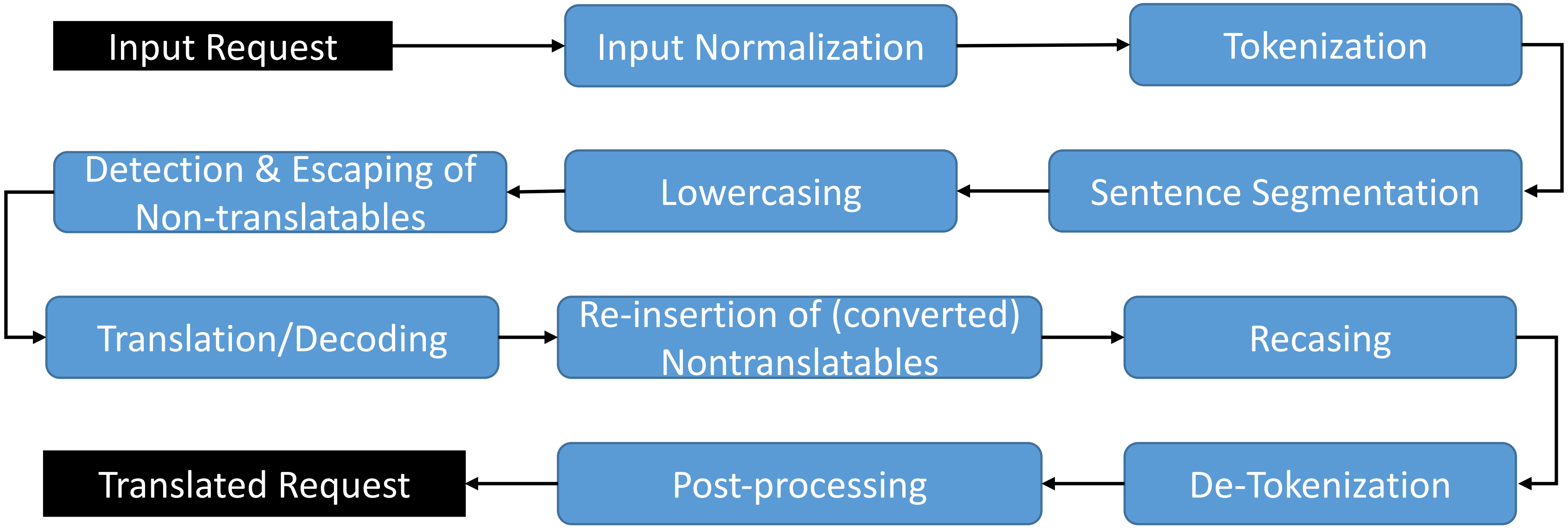
Overview

- What is Amazon?
- Machine Learning in Practise
  - Probabilities
  - Finite Resource
- Machine Learning @ Amazon
  - Forecasting
  - **Machine Translation**
  - Visual Systems
- Conclusions and Challenges

Product Machine Translation (2013 – 2015)



Machine Translation Pipeline



Machine Translation: Deep Dive

$$p(English \mid German) = \frac{p(English) \times p(German \mid English)}{p(German)}$$

$$\propto \underbrace{p(English)}_{\text{Language Model}} \times \underbrace{p(German \mid English)}_{\text{Translation Model}}$$

- **Language Model:** What are fluent English sentences?
- **Translation Model:** What English sentences account well for a given German sentence?

Overview

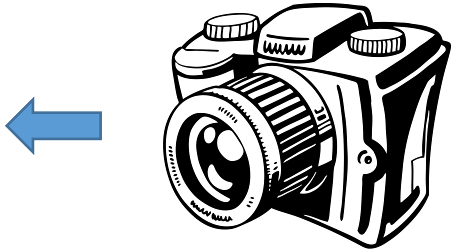
- What is Amazon?
- Machine Learning in Practise
    - Probabilities
    - Finite Resource
- Machine Learning @ Amazon
    - Forecasting
    - Machine Translation
    - **Visual Systems**
- Conclusions and Challenges

Automated Produce Inspection: The Goal

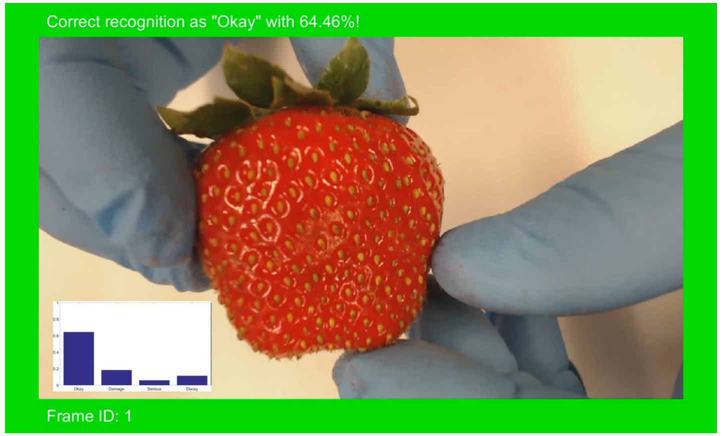
Current Inspection

	Defect	Decay	Bruising	Bruising	Overripe/Soft
Sample #	Defect Cat.	Serious	Damage	Serious	Serious
1	60	3	10	7	1
2	60	5	5	1	3
3	60	0	10	7	0
4	60	1	15	1	0
Total	240	9	40	16	4
% of Total	100%	4%	17%	7%	2%

New Automated Inspection

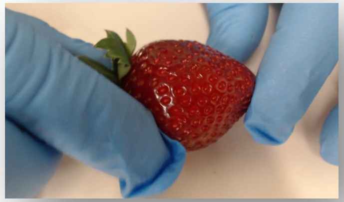


Computer Vision



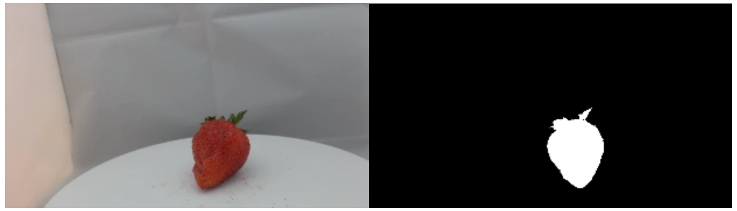
Challenges

- Illumination
- Clutter/Occlusions
- Viewpoint
- Scale
- Intra-class variability



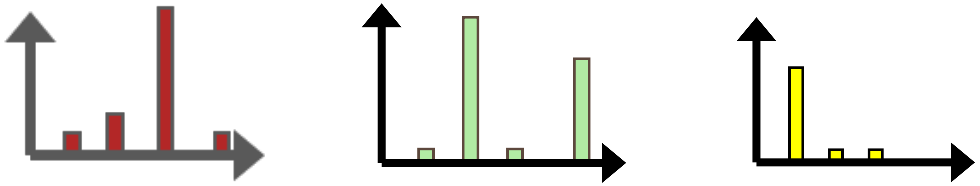
Computer Vision Pipeline

- Segmentation

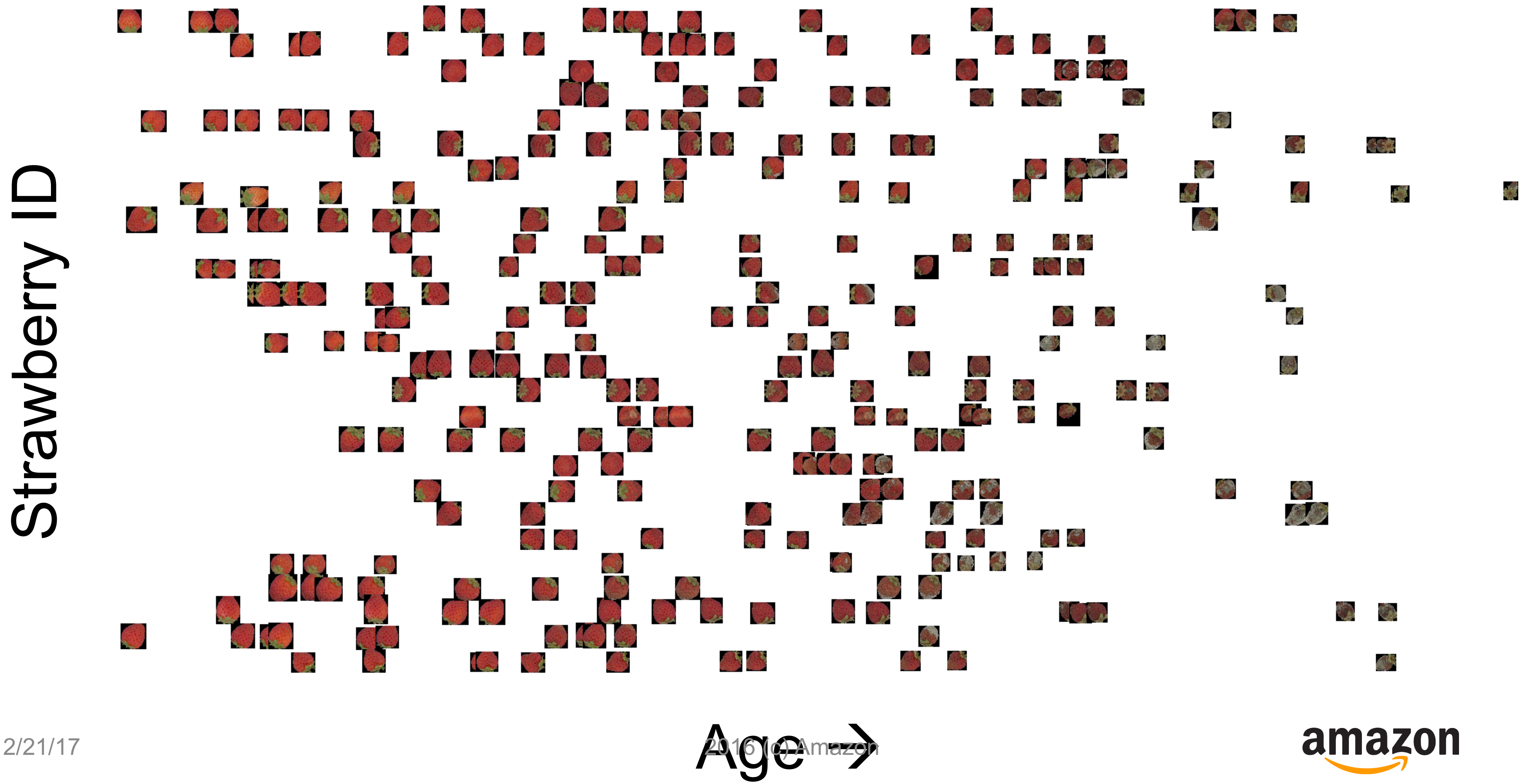


- Defect Mining

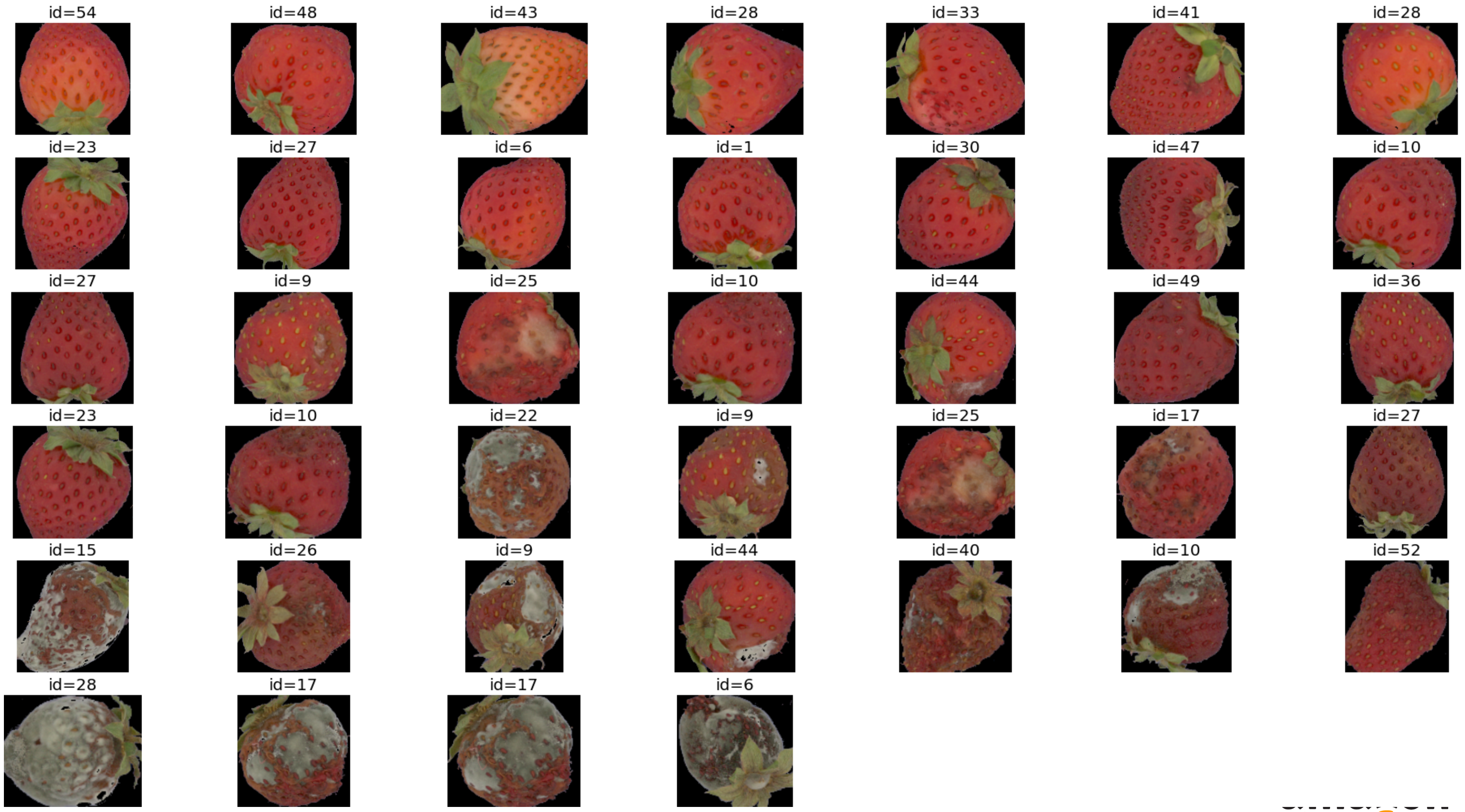
- Produce Classification



Predicting Longevity



Age Aligned Strawberries (Test Set)



Overview

- What is Amazon?
- Machine Learning in Practise
  - Probabilities
  - Finite Resource
- Machine Learning @ Amazon
  - Forecasting
  - Machine Translation
  - Visual Systems
- **Conclusions and Challenges**

Conclusions

- Machine Learning “translates” data from the past into accurate predictions about the future!
- In practice, probabilistic models and finite resources matter.
- Machine Learning helps to improve customer experience at Amazon!

Thanks!