Performance Curves

A Methodology for Evaluation in Ad-Tech
Andres Corrada-Emmanuel
Head of Optimization @ Swoop

Swoop

- Founded 2011 by Ron Elwell and Sim Simeonov

- Ad-network:

  - Users/Publishers/Advertisers part of the eco-system.

  - Context = Search

  - No RTB, just-in-time ad insertion.

  - Strong view signal

- Mike Baker advisor.

A Personal Focus on Measurement not Models

- ICML 2008: Autonomous geometric precision error estimation in low-level computer vision tasks

  - We can measure precision (not accuracy!) without knowing ground truth of terrain elevations (real variable)

- Patent 2010: Polynomial approach to ground truth inference of classifiers (class labels).

- Patent 2014: Measuring Web Browser Tag Properties Without True Unique Tags

  - We can measure error rates of a unique id identifier system without knowing true identity of users.

What is mostly wrong with ROC/AUC?

- **Wrong**: Costs/Benefits are not included in the assessment.

- **Wrong**: Difficult to generalize to tasks that are not classification and/or do not have just false positive/ false negative errors (e.g. a unique id system, speech recognition).

- **More Wrong**: Papers by David Hand (see Wikipedia article on ROC).

- **Right**: It compares detectors via a ranking procedure.

The Problem of Offline Evaluation in Ad-Tech

- **Counterfactuals:**

  - You must pay for all your evidence.

  - You don't know what would have happened if a different course was taken. As opposed to, say, financial stock picking.

  - Machine Learning a small portion of what ad-serving engines do. Other parts of the system can help/hinder an ML model's job.

  - All offline evaluation methods have this problem.

  - Bottou et al: Counterfactual reasoning and learning systems (MLJ).

The Performance Curve Methodology I

- Inspired by comment by Beth.

- KPIs rule the land: $\sum_i x_i / \sum_i y_i$

- **Requirements**

  - x_i: Numerator of KPI for i-th decision.

  - y_i: Denominator of KPI for i-th decision

  - s_i: Ranking score

  - xAxis_i: KPI versus cumulative of x-axis variable

The Performance Curve Methodology II

- **Algorithm**

  - Sort by s_i, usually descending.

  - Start at the top and at each decision:

    - Compute KPI.

   - Accumulate x-axis variable

  - Plot

A Toy Example

n	score	cost	actions
1	0.75	10	4
2	0.55	5	2
3	0.8	3	3
4	0.3	7	5

n	score	cost	action	C	A	V	A/C
1	0.8	3	3	3	3	1	3/3=1
2	0.75	10	4	3+10=13	3+4=7	1+1=2	7/13
3	0.55	5	2	18	9	3	9/18
4	0.3	7	5	25	14	4	14/25

Actions / View Example

Auto Vertical, Action Campaigns Actions/View By LibFM Component Score:
2017–06–01 to 2017–06–04

Much better now

Auto Vertical, Action Campaigns Actions/View By LibFM Component Score:
2017–06–08 to 2017–06–11

- libFM model
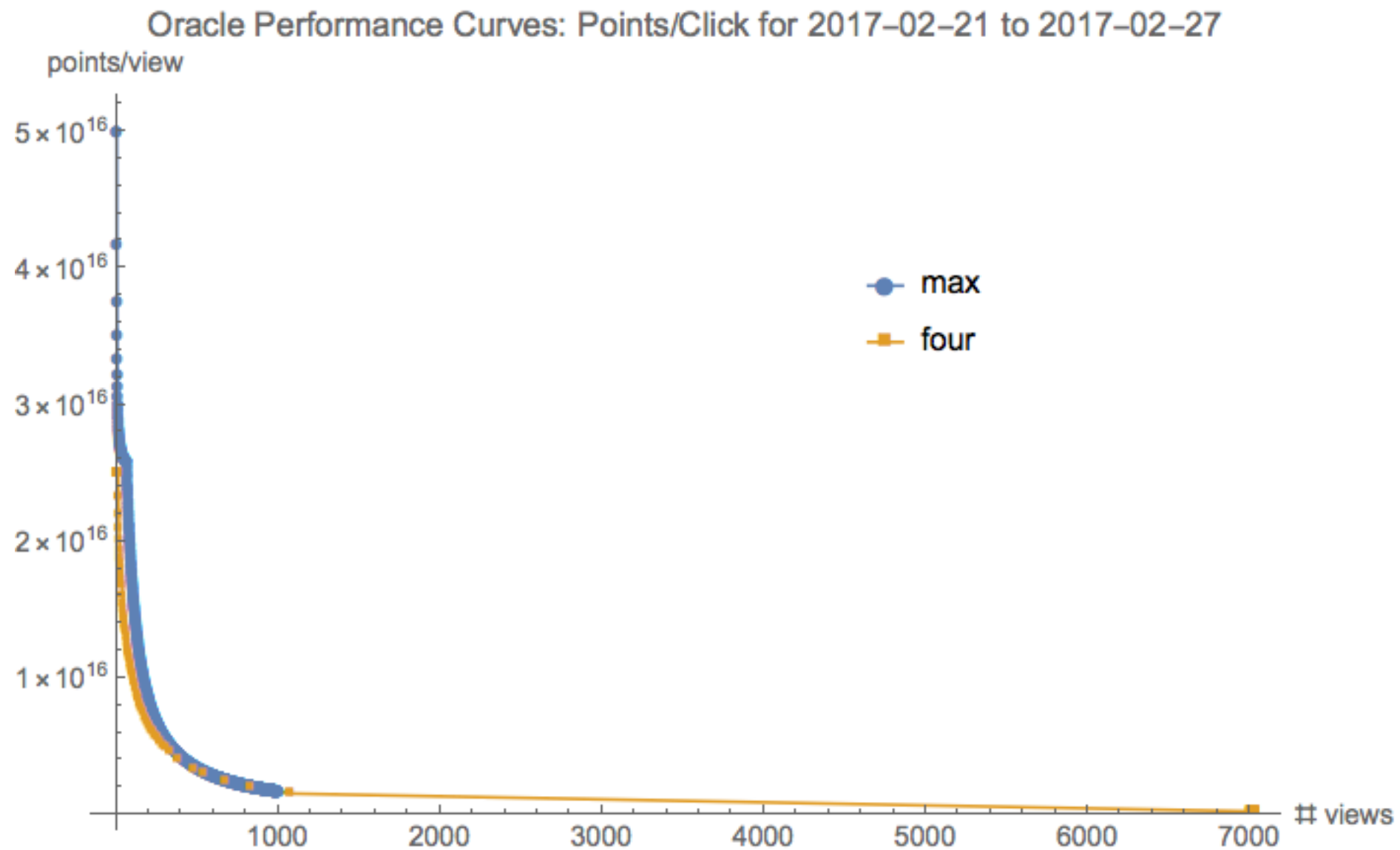- view–level oracle
- 2.5%
- 97.5%
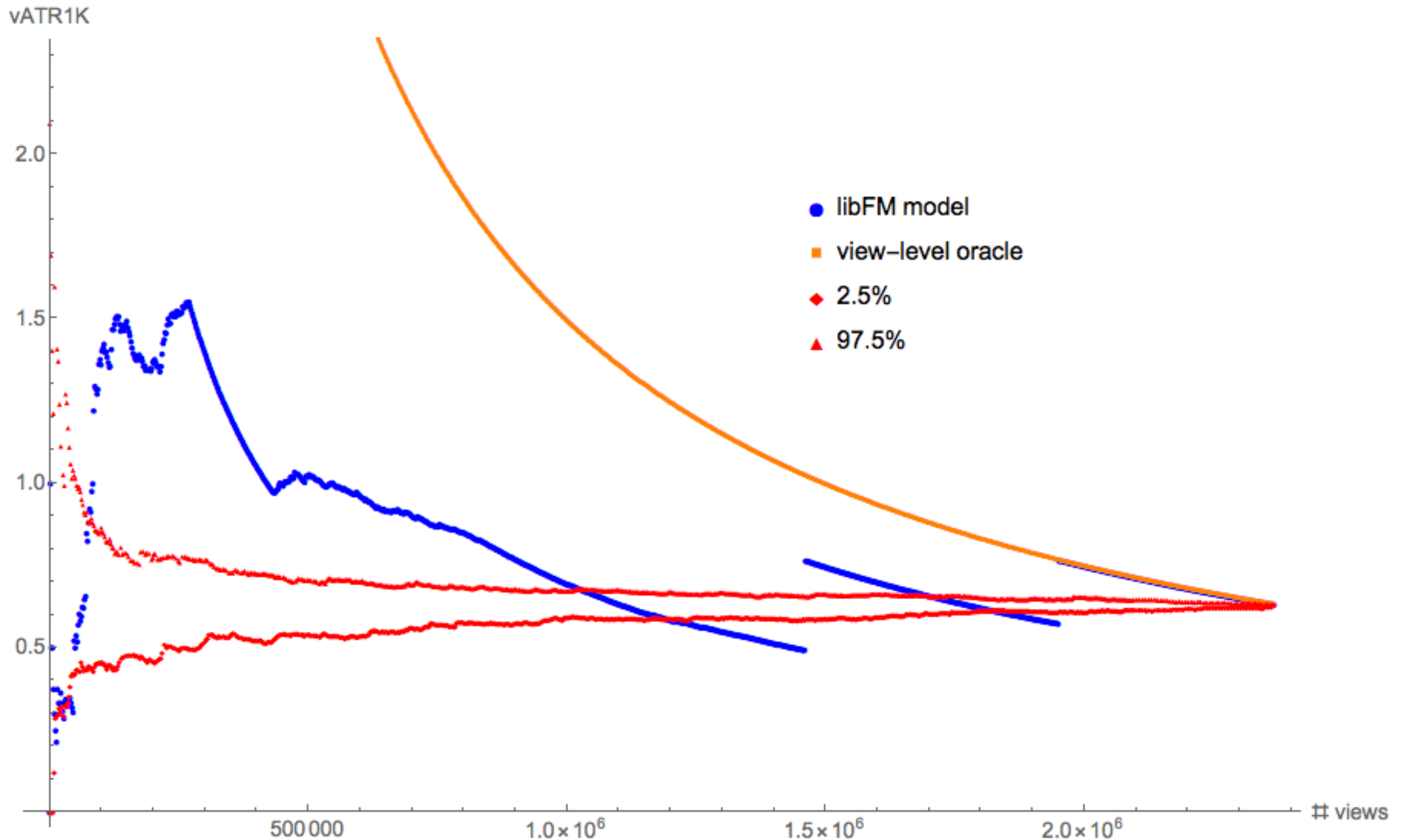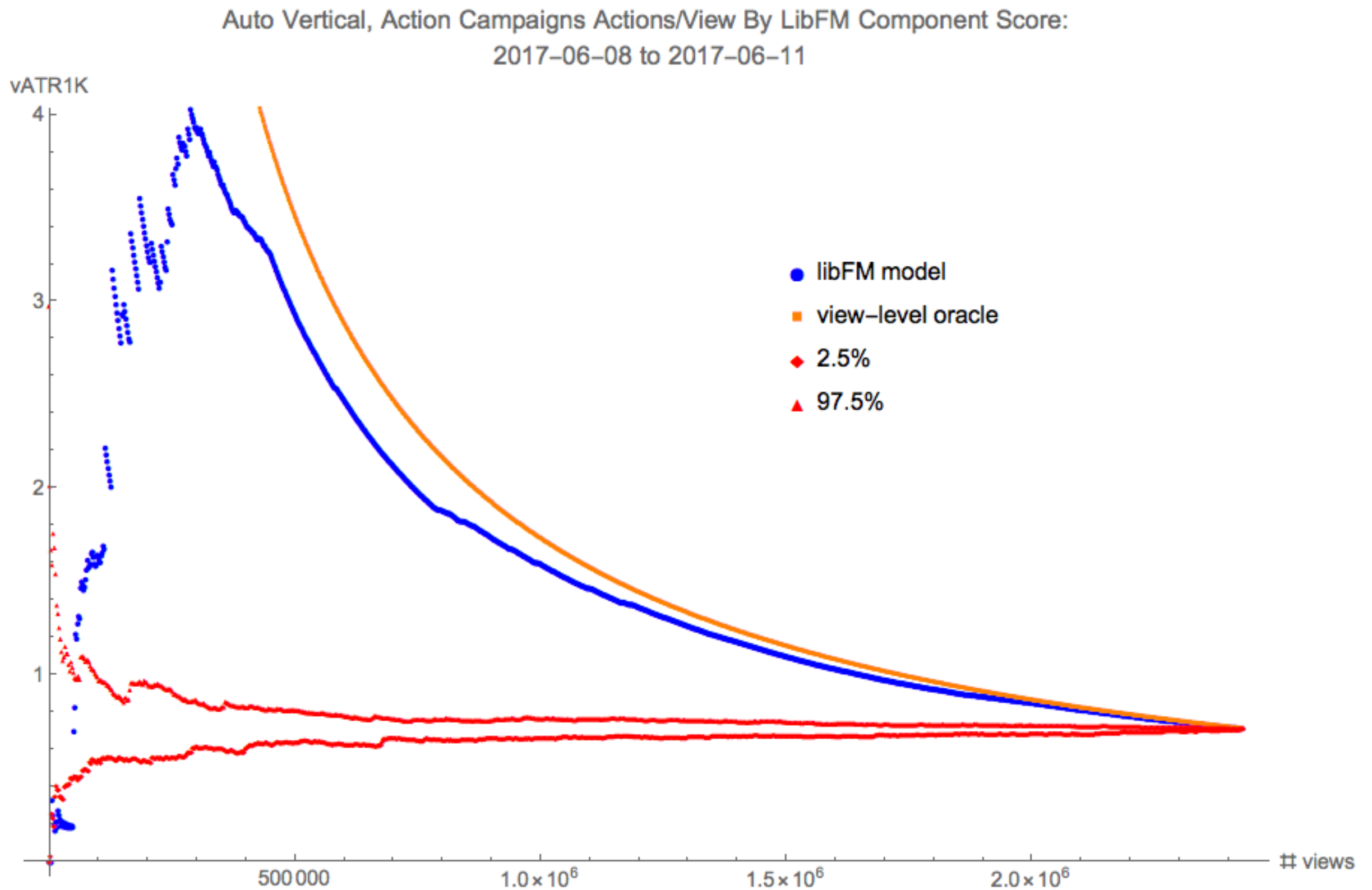
What can we learn from data?

- Evaluation data limits the conclusions you can make.

- Consider the time period for an evaluation.

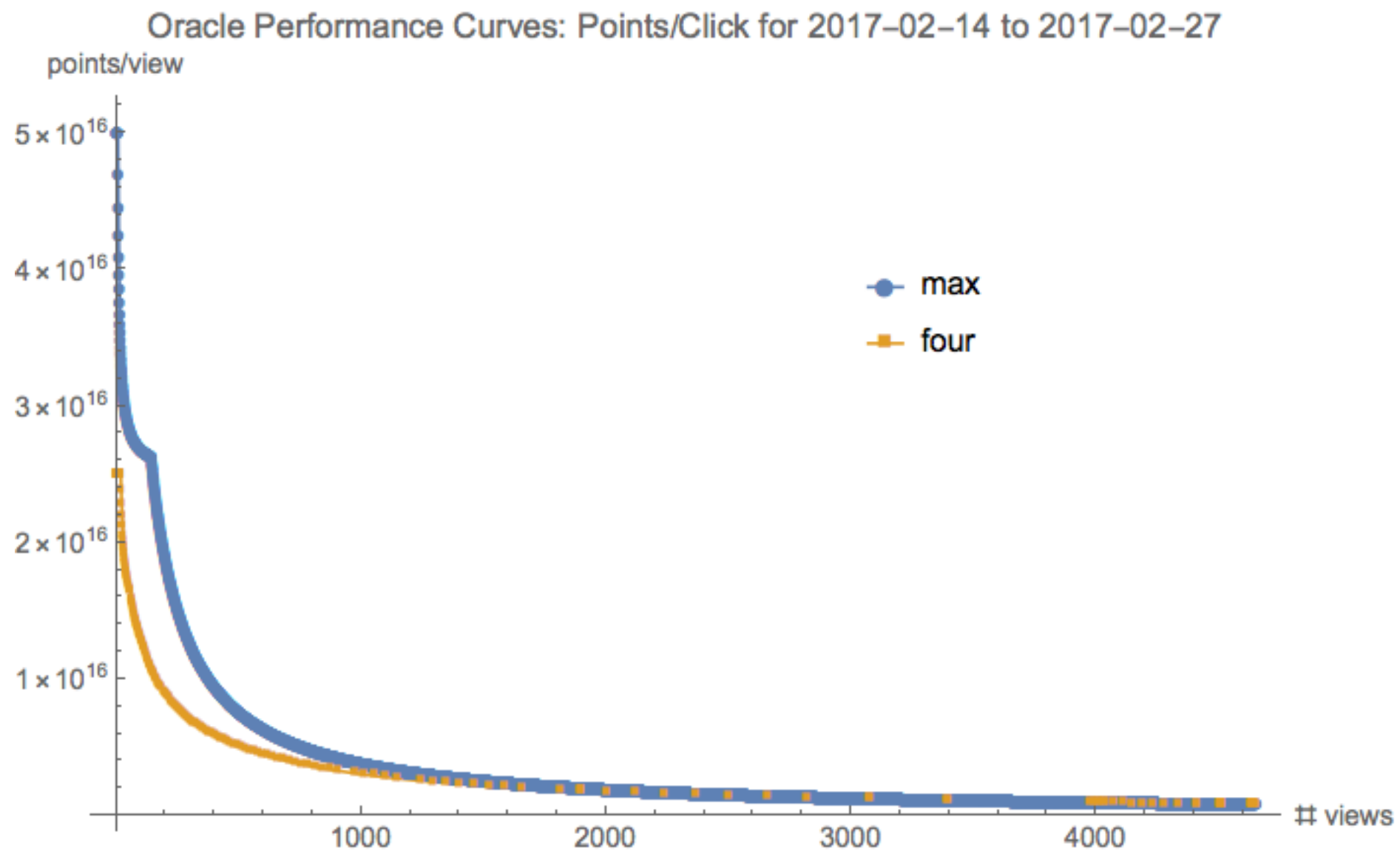# One Week of Data Is Not Enough

- Feature oracles below max oracle (view level granularity)

Oracle Performance Curves: Points/Click for 2017–02–21 to 2017–02–27

Two Weeks Seems Enough

- Feature oracles below max oracle (view level granularity)

Oracle Performance Curves: Points/Click for 2017–02–14 to 2017–02–27

Comparing Different Steps in Production

Auto Vertical, Action Campaigns Actions/View – Comparing Stages:
2017–06–08 to 2017–06–11

- 2nd stage: component
- 1st stage: creative
- view–level oracle
- 2.5%
- 97.5%
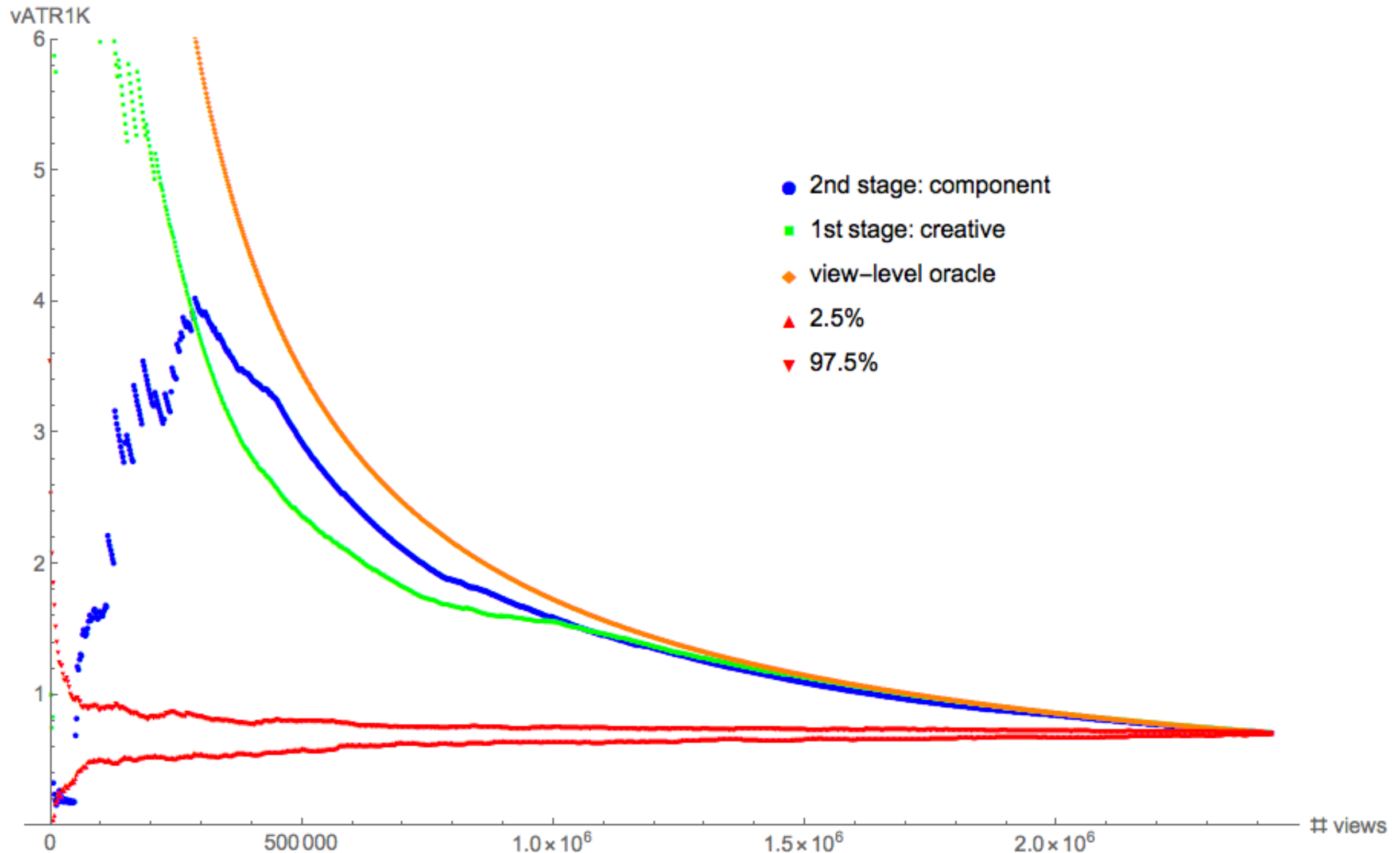
Topics not discussed

- **A single figure of merit can be derived as the average KPI over the performance curve**

  - Measured in units that make sense - KPI.

  - Can be used to automate model selection.

- **Precision of models**

  - Create random envelopes for equally scored decisions.

- **Sorting by 2nd price very illuminating in an RTB setting**

  - below random performance

  - not a "rational" market

Thank You

- github.com/andrescorrada/performance-curves