Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation

Nicola Ferro, and *Gianmaria Silvello*

University of Padua, Italy

Marco Angelini, Giuseppe Santucci

"La Sapienza" University of Rome, Italy

Information Interaction in Context (IIiX 2012)
Nijmegen, the Netherlands, August 21-24, 2012

Outline

- Motivations
- Models for Interaction
- The Prototype and Application Examples
- The Domino Effect
- Final Remarks and On-Going Works

Motivation

Reduce the time needed to understand and analyze the behavior of an IR system providing a **visual interactive tool for evaluation**

Motivation

Reduce the time needed to understand and analyze the behavior of an IR system providing a **visual interactive tool for evaluation**

Failure Analysis

or "Understand What is Wrong".

Is it better to re-rank or to re-query?

Motivation

Reduce the time needed to understand and analyze the behavior of an IR system providing a **visual interactive tool for evaluation**

Failure Analysis

or "Understand What is Wrong".

Is it better to re-rank or to re-query?

What-if Analysis

"Try to anticipate the effect of a modification of the system".

The Underlying Idea

IR Evaluation + Visual Analytics

Visual Analytics is not only a mean to improve the presentation of results, but also a mean allowing the users to analyze and interact with data

Models for Interaction

Rank Gain/Loss Model

Clustering via Learning to Rank

What-If Analysis Model

Rank Gain/Loss Model

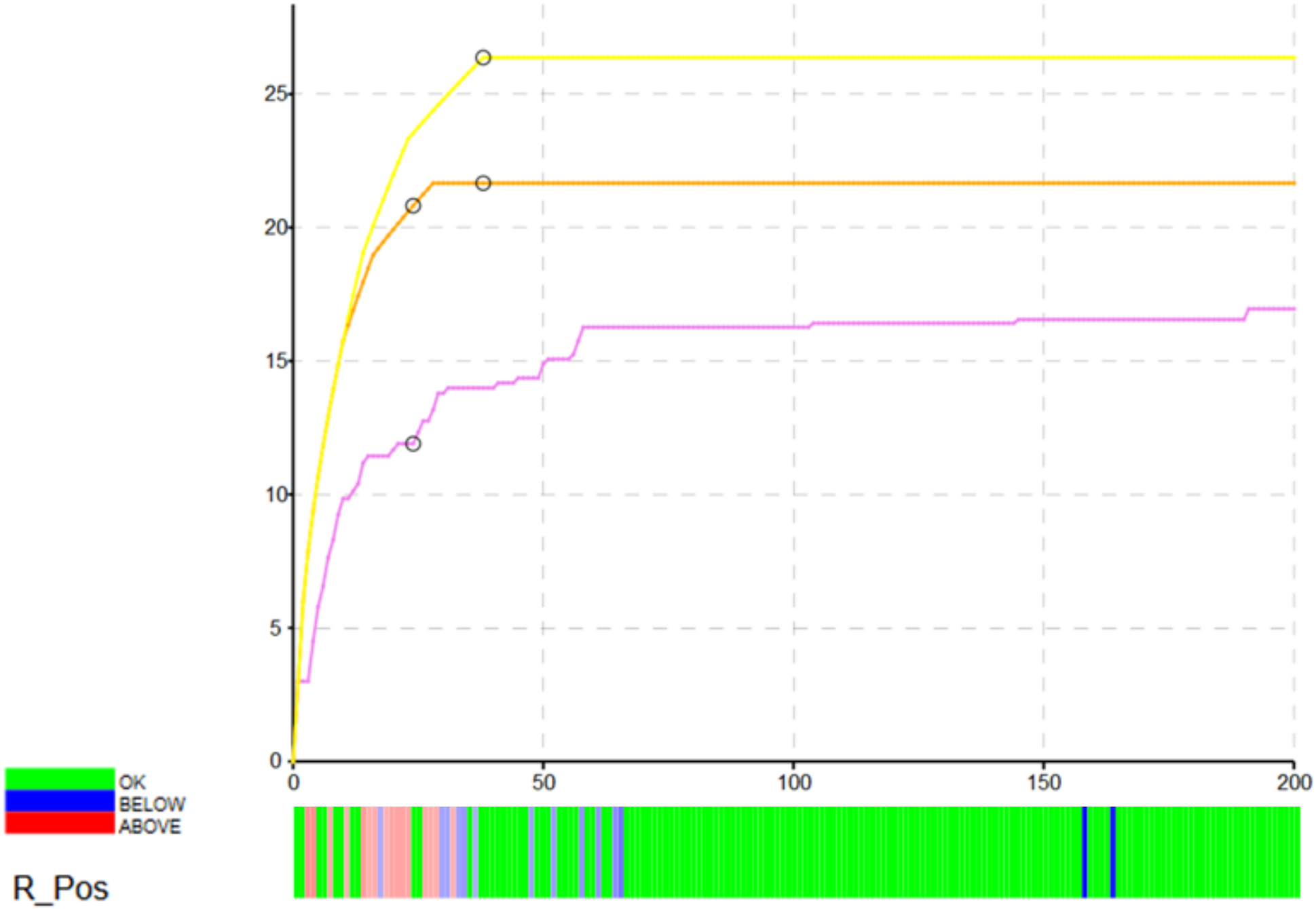
DCG is good if we want to compare performances between different systems but **what if we want to look at misplaced documents?**

Rank Gain/Loss Model

DCG is good if we want to compare performances between different systems but **what if we want to look at misplaced documents?**

Relative Position = RPos let us determine how much a document is misplaced with respect to its ideal rank.

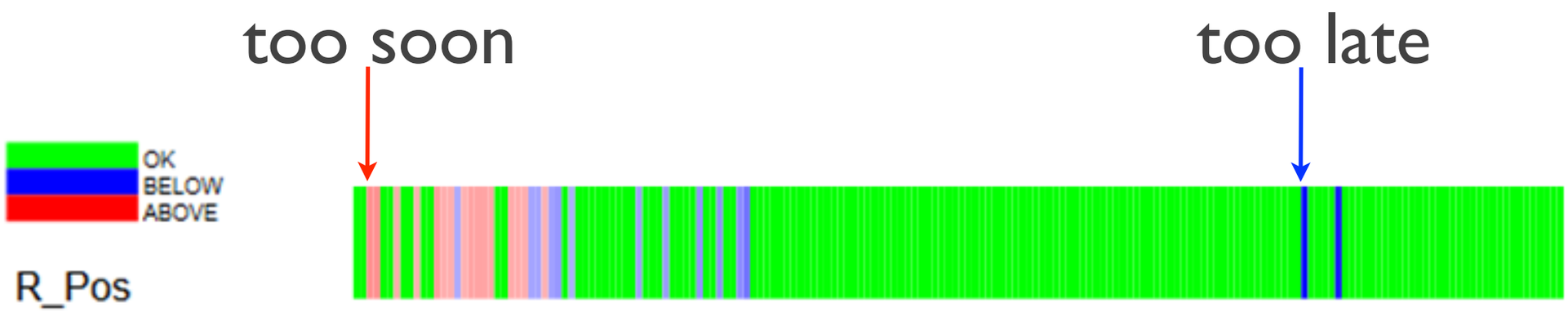
Rank Gain/Loss Model: R_Pos



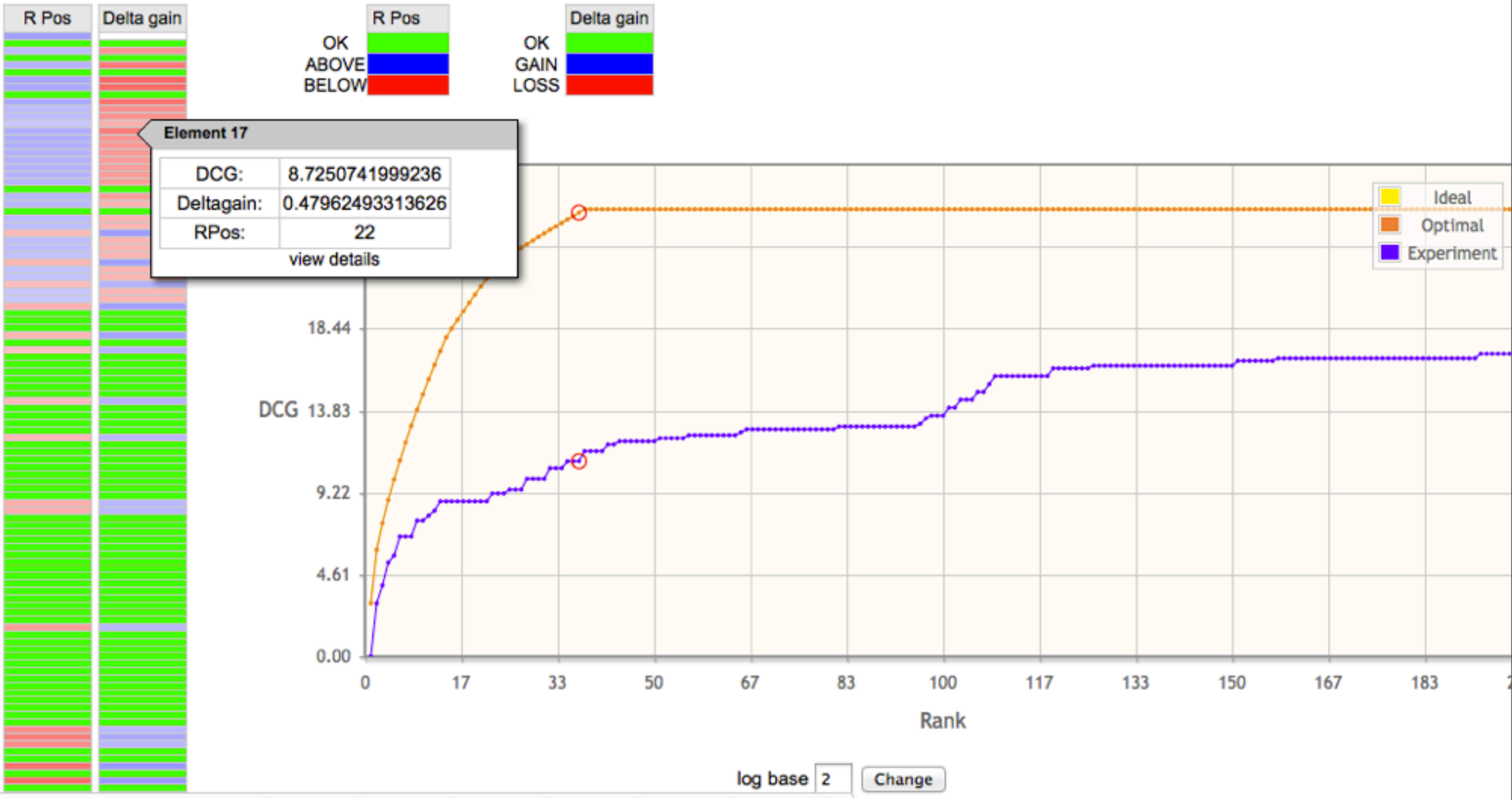
Rank Gain/Loss Model: R_Pos

By means of R_Pos we can say if a document is ranked too soon or too late. But**, can we quantify "too soon" and "too late"**?

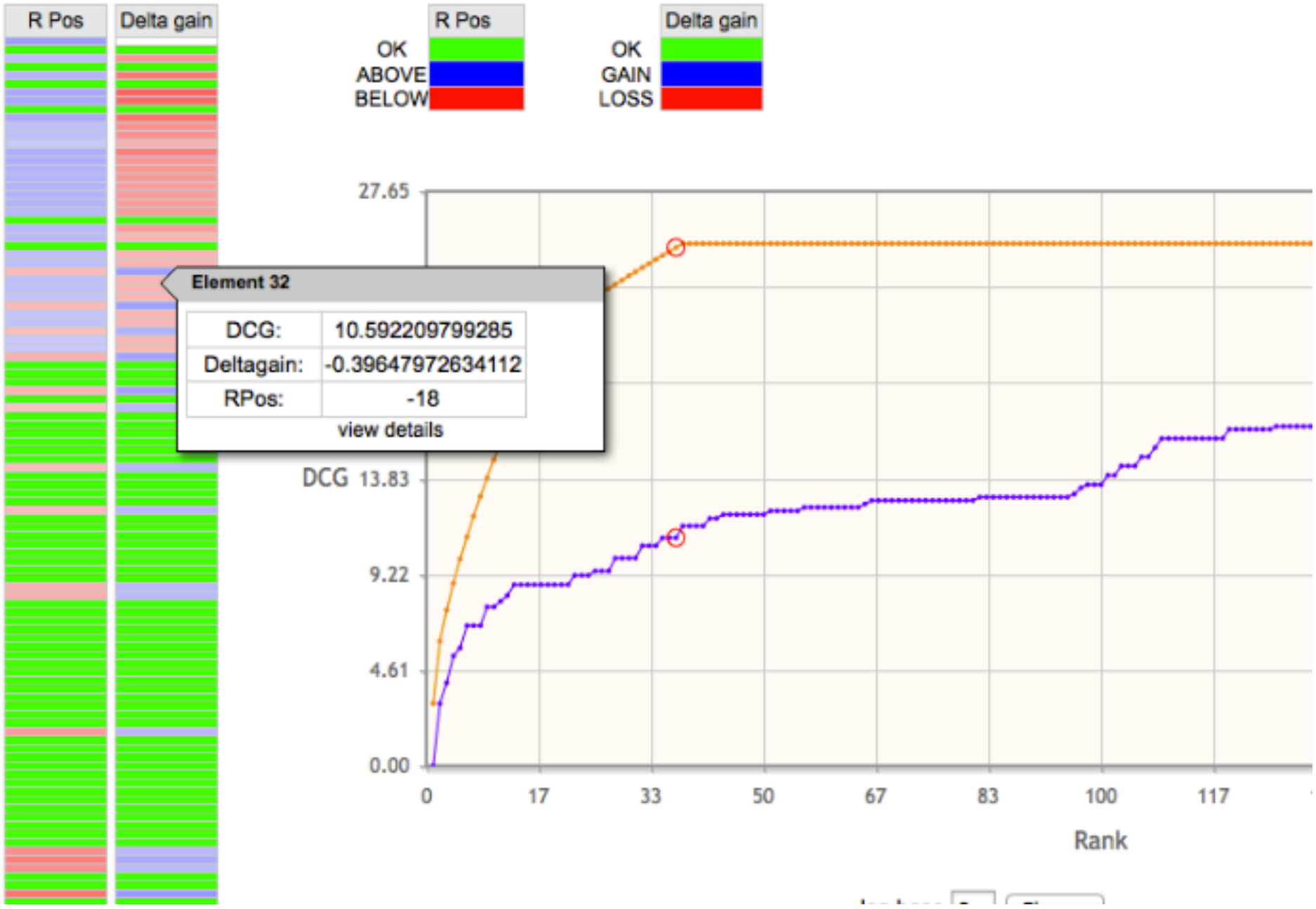
How much these misplaced documents impact on the **dcg**?



Rank Gain/Loss Model : Delta Gain

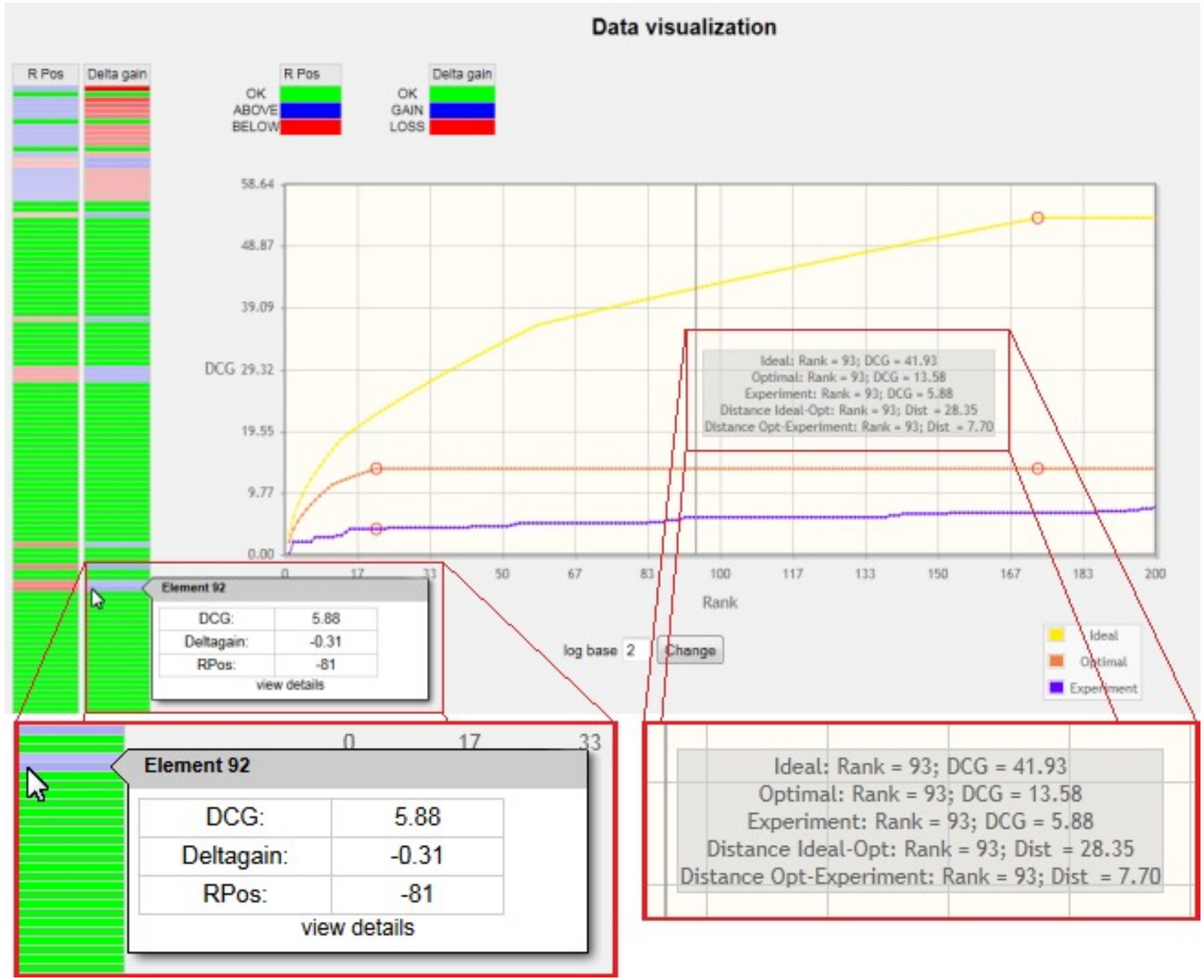


Rank Gain/Loss Model : Delta Gain



Models for Interaction

Failure Analysis

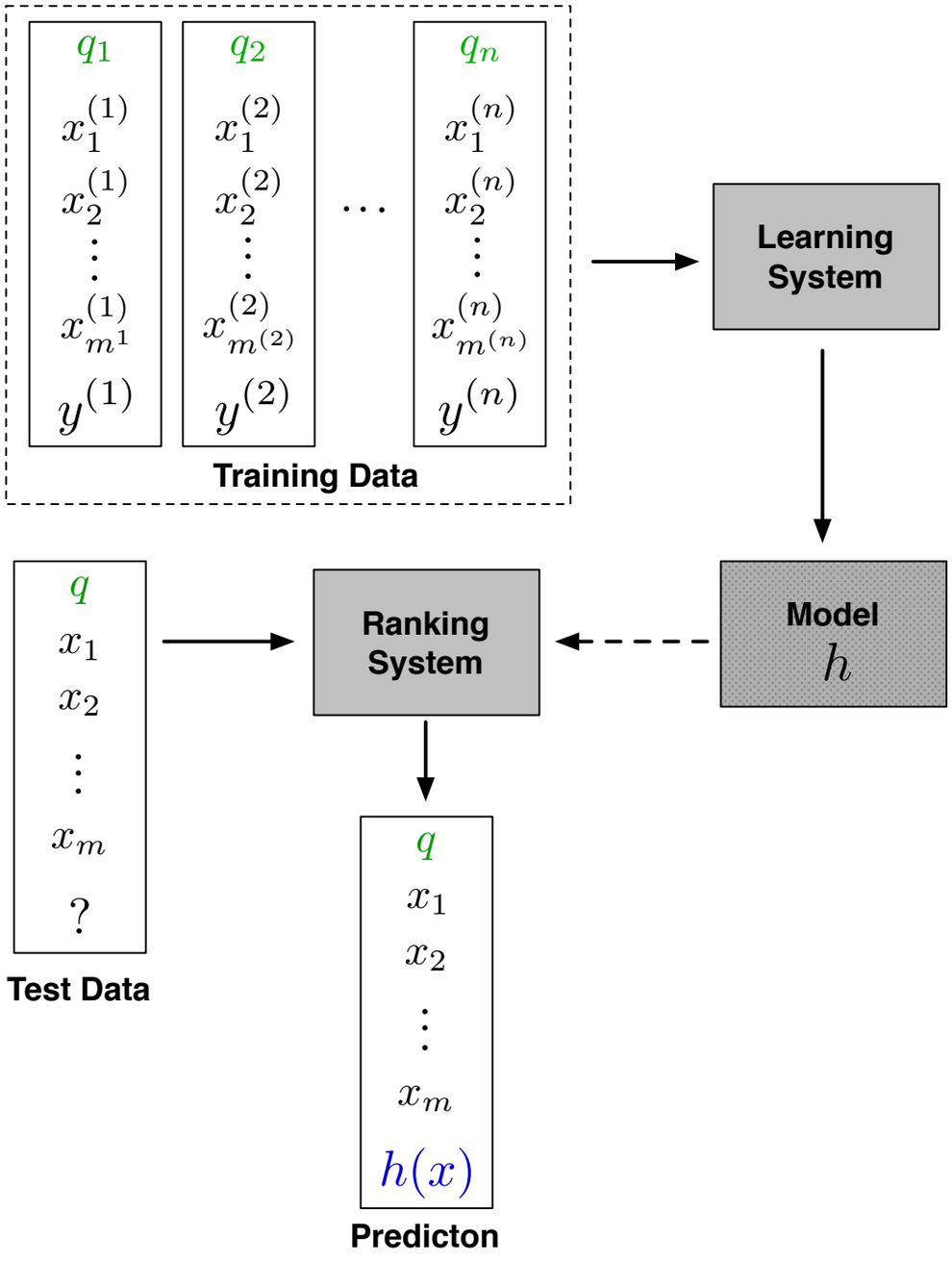


Clustering via Learning to Rank

Learn the ranking model of the IR system under investigation in order to simulate the way in which it ranks the documents

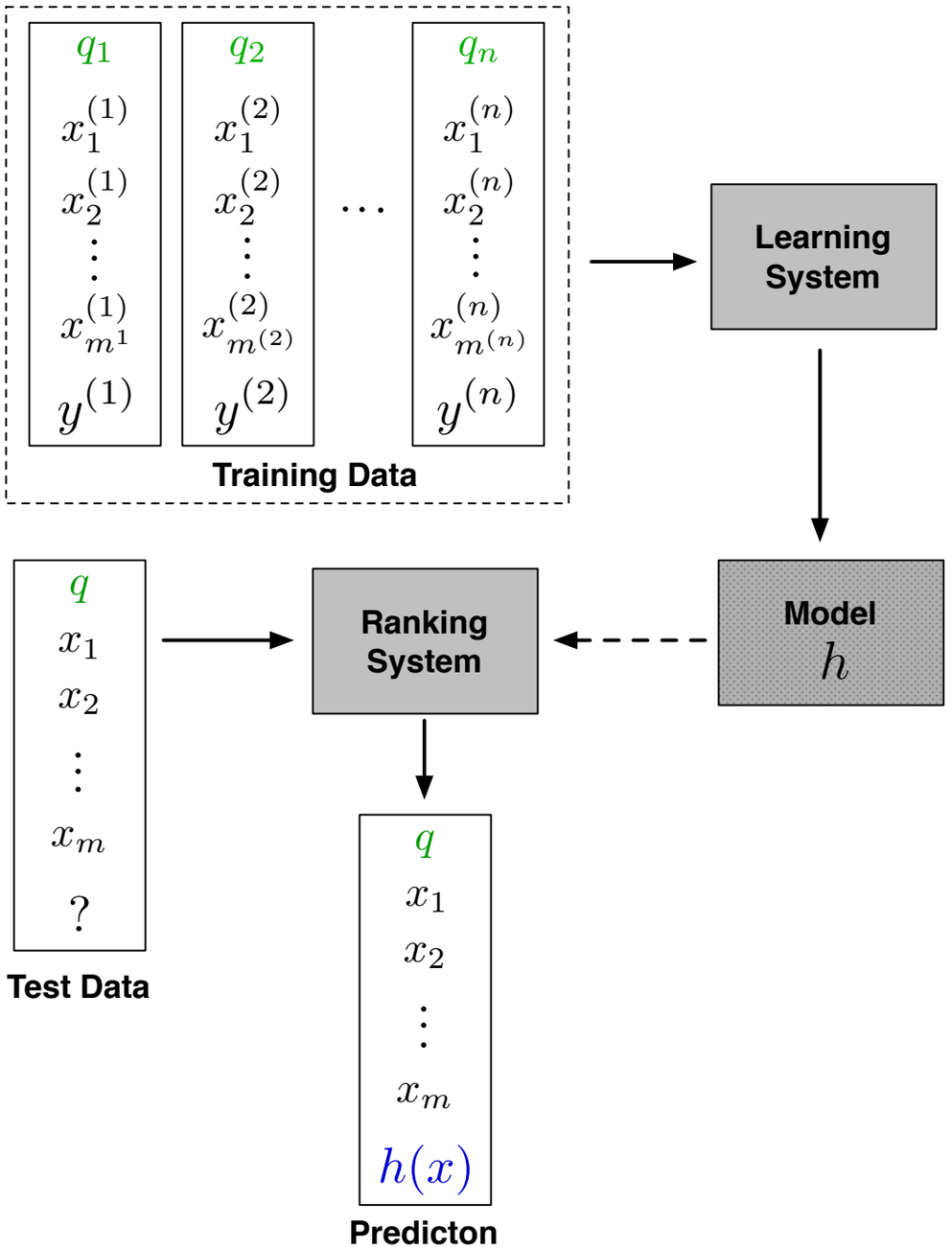
Models for Interaction

Clustering via Learning to Rank



Models for Interaction

Clustering via Learning to Rank



- Lever on the clustering hypothesis
- Group together the documents which are similar from the considered ranking model point-of-view

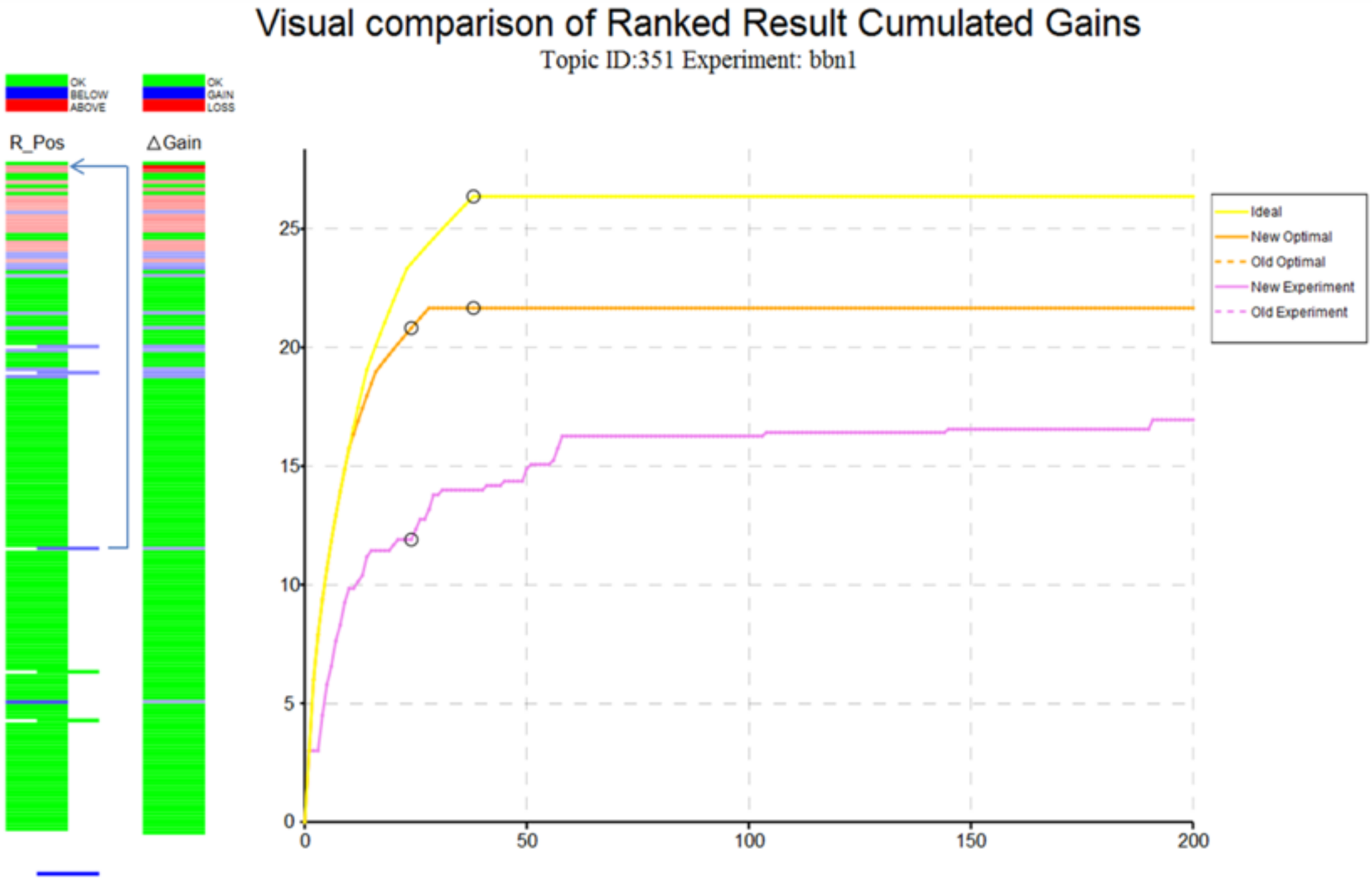
Generating the clusters

(1) Submit each doc in Dj as a query and retrieve a set of docs Di;

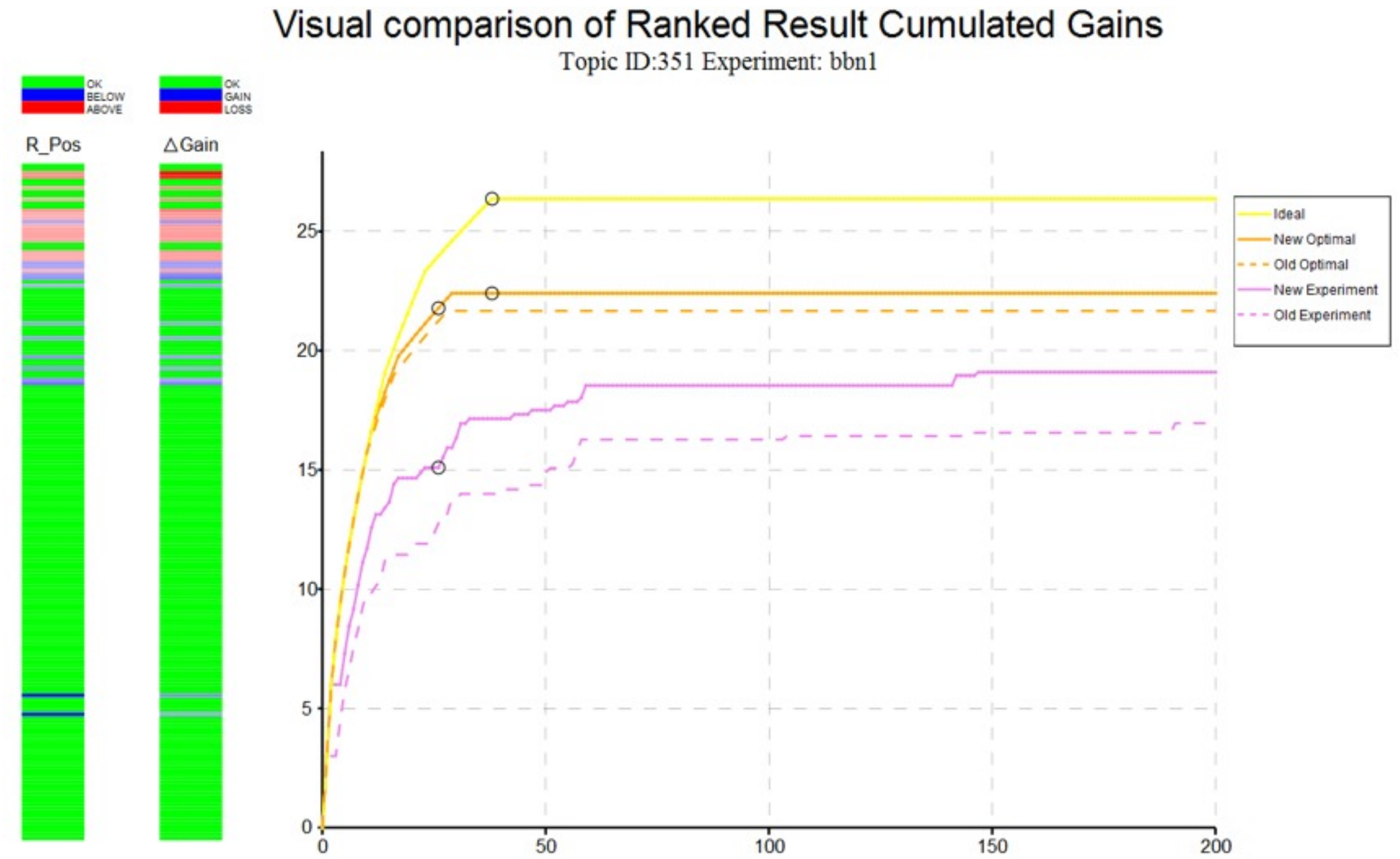
(2) determine Ci = Dj \ Di;

(3) ranking the documents in Ci by employing the learned ranking model

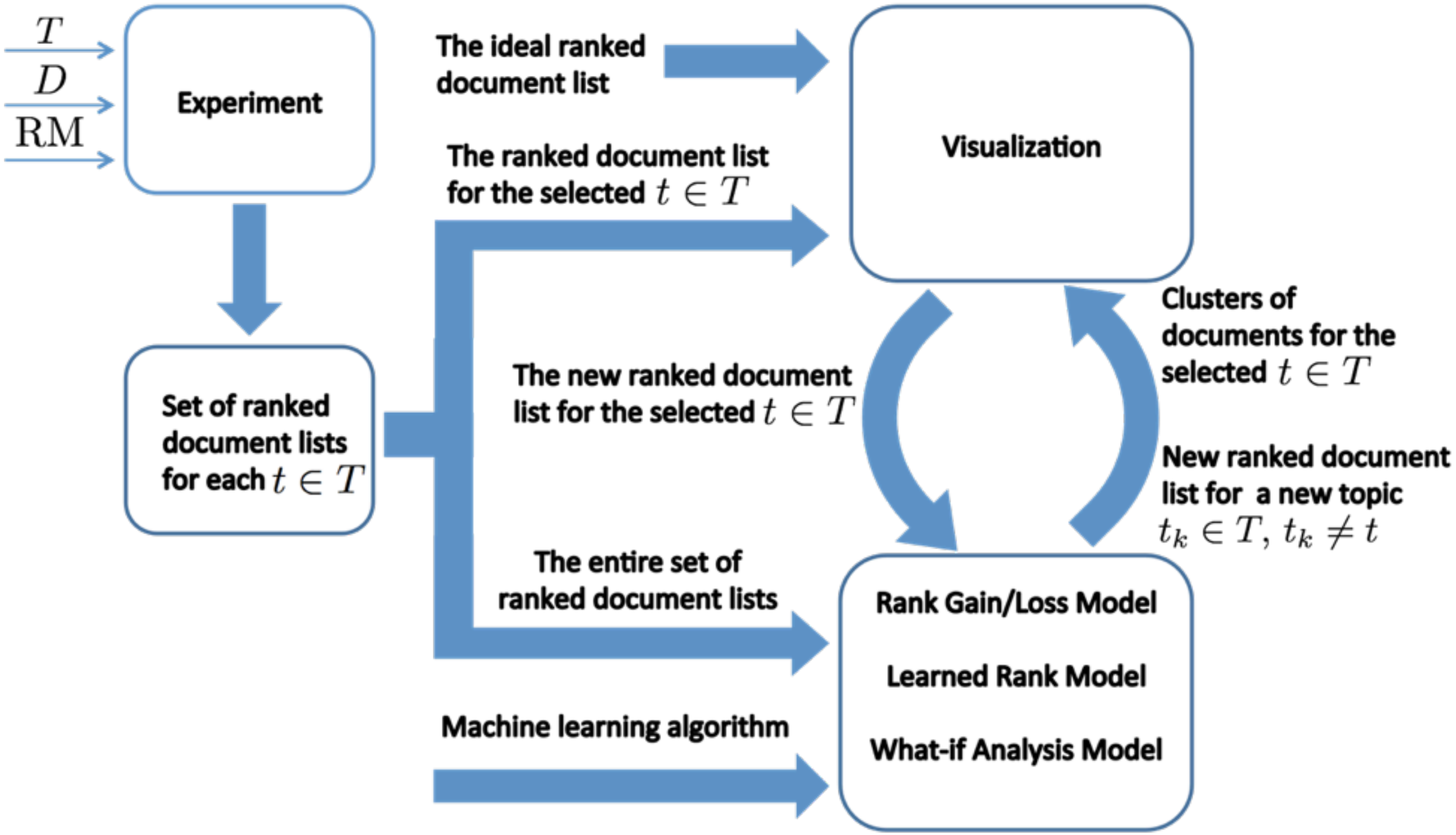
What-if Analysis



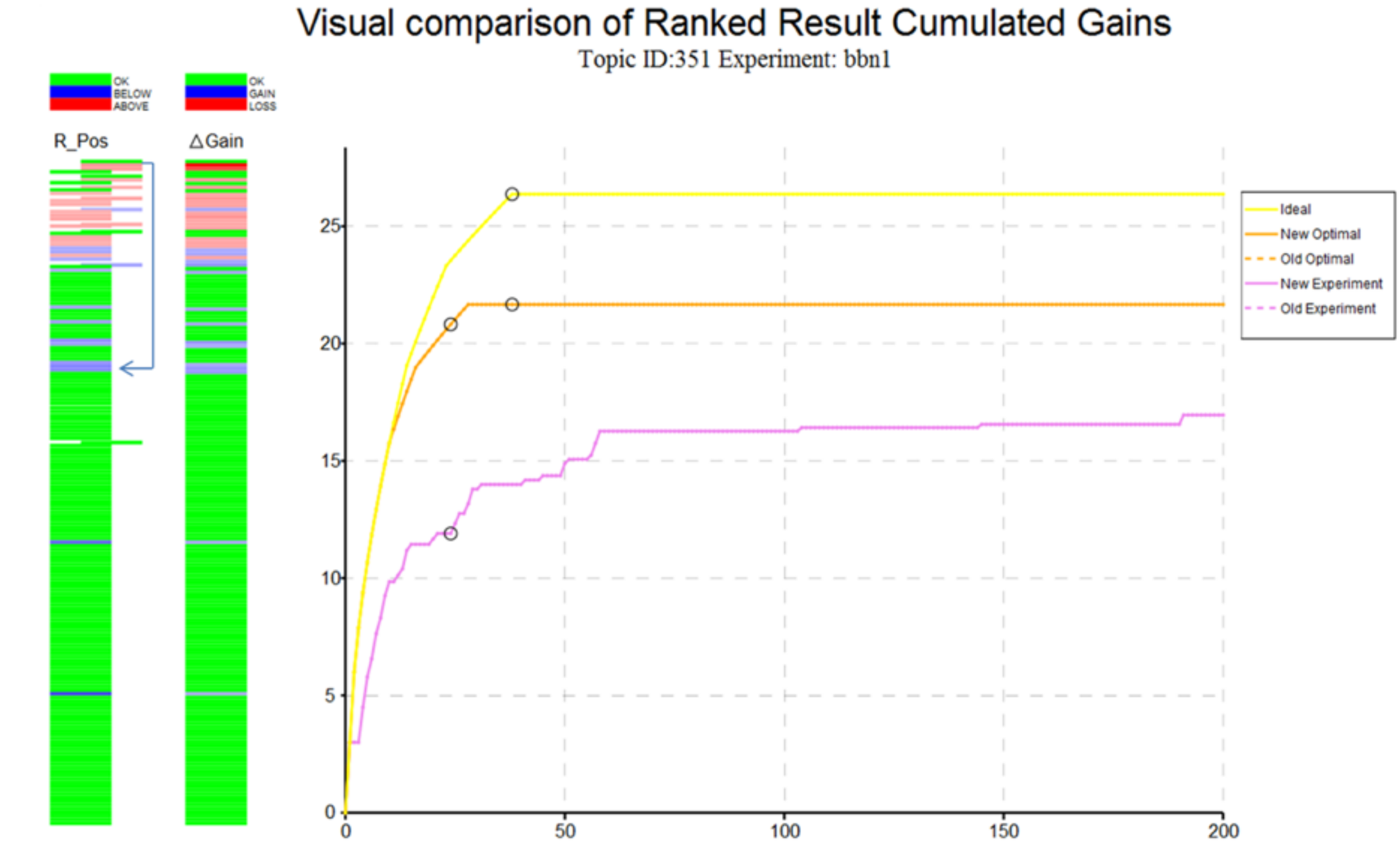
What-if Analysis



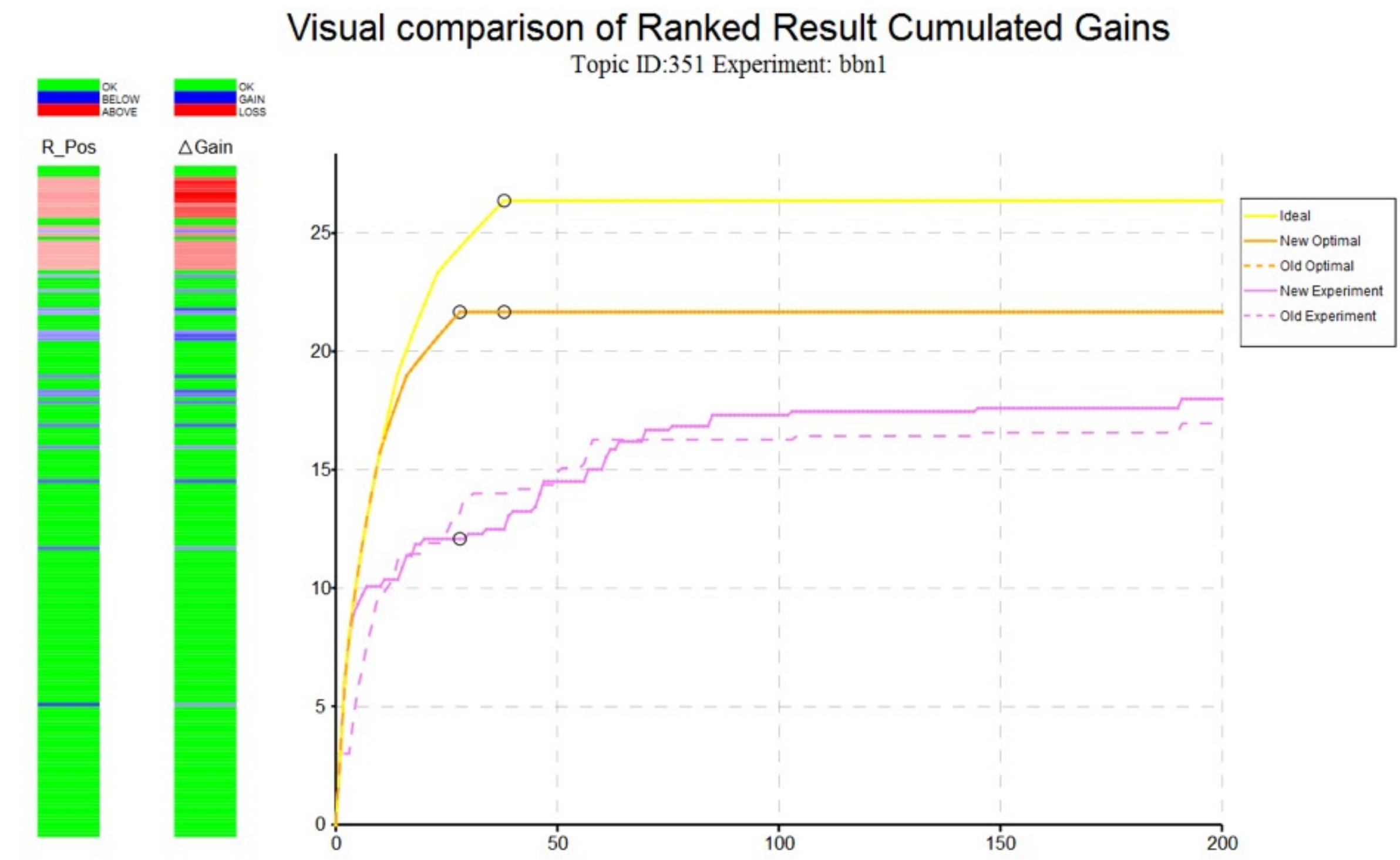
To Summarize: Data Pipeline



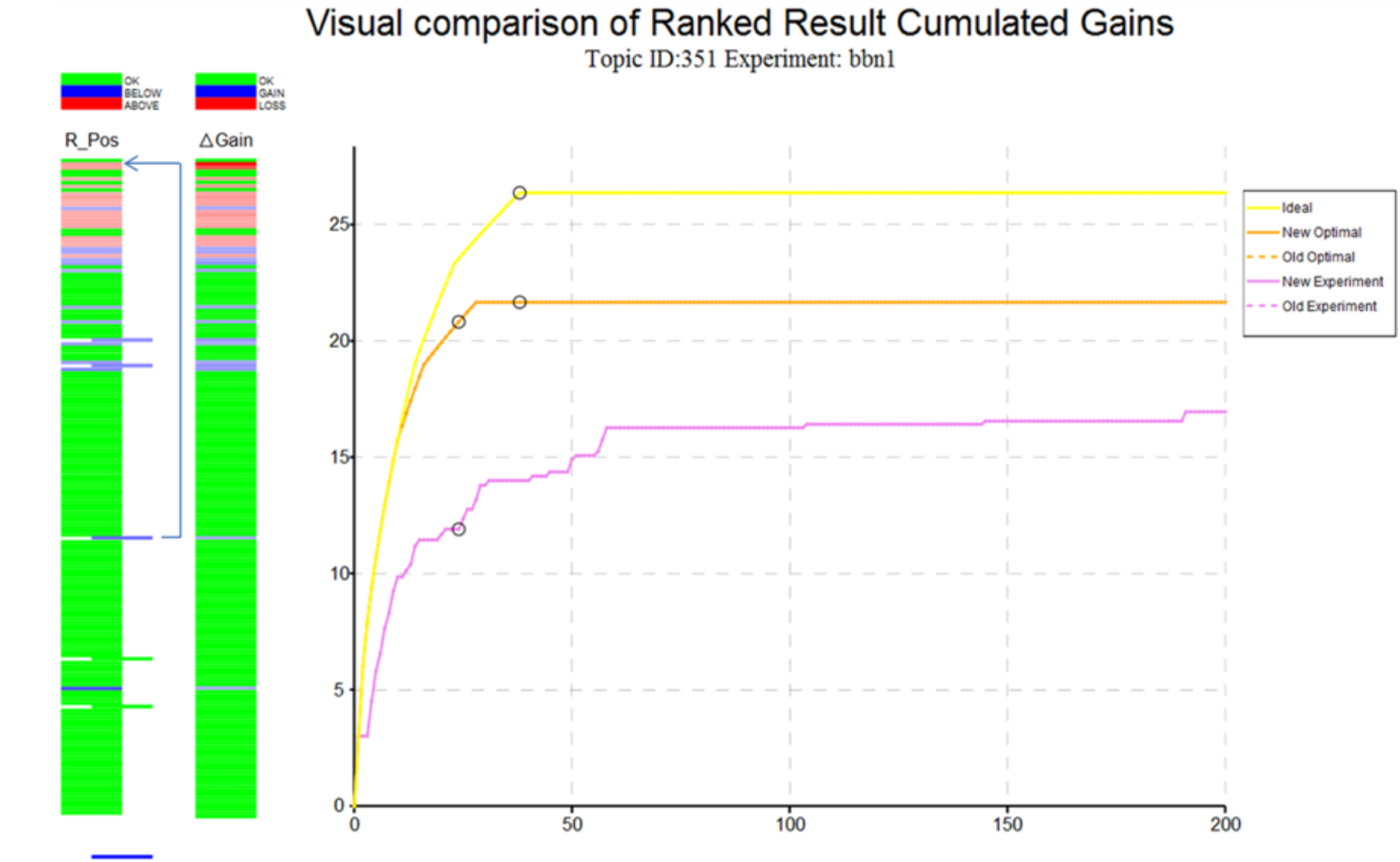
Application Example: Free Movement



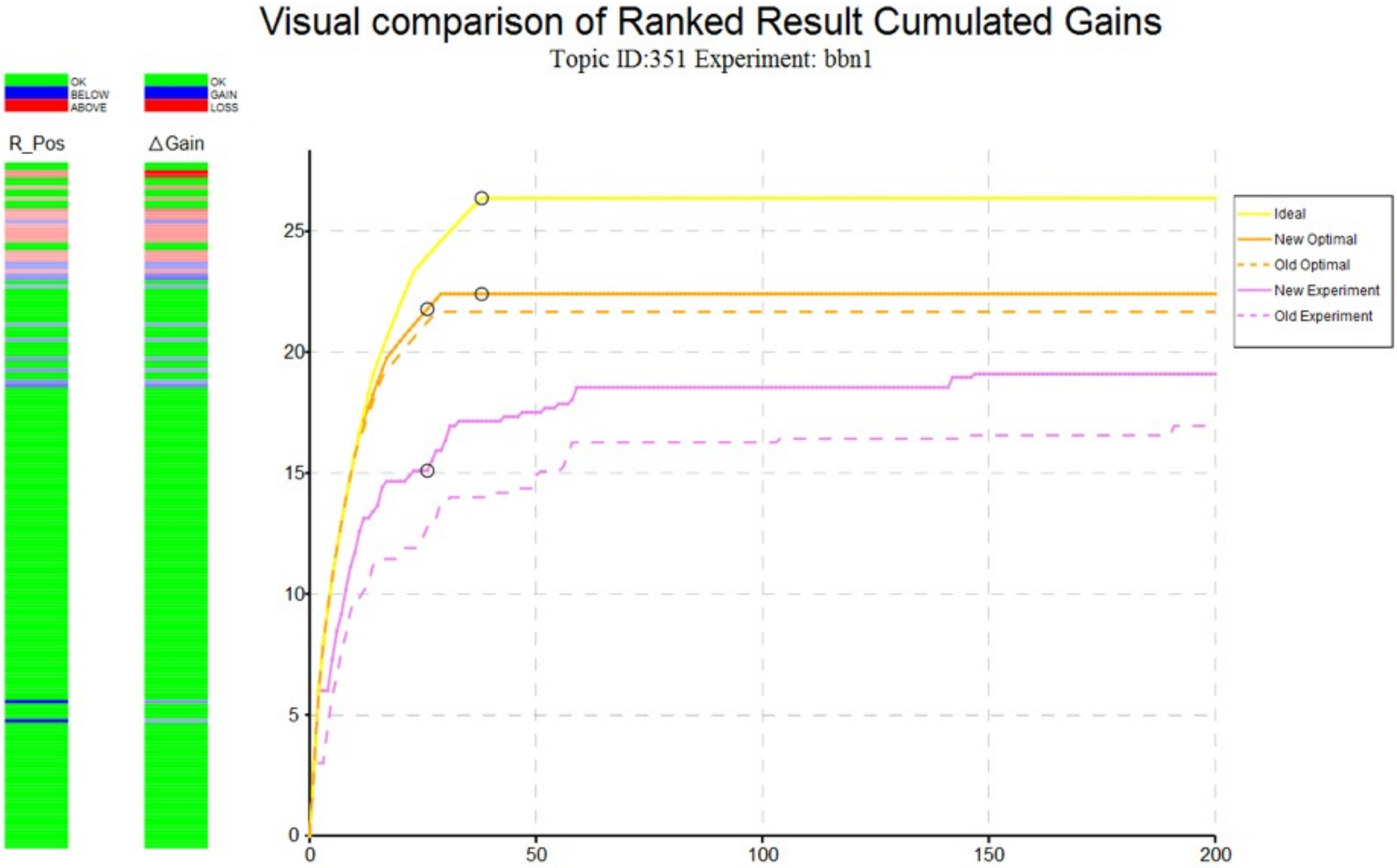
Application Example: Free Movement



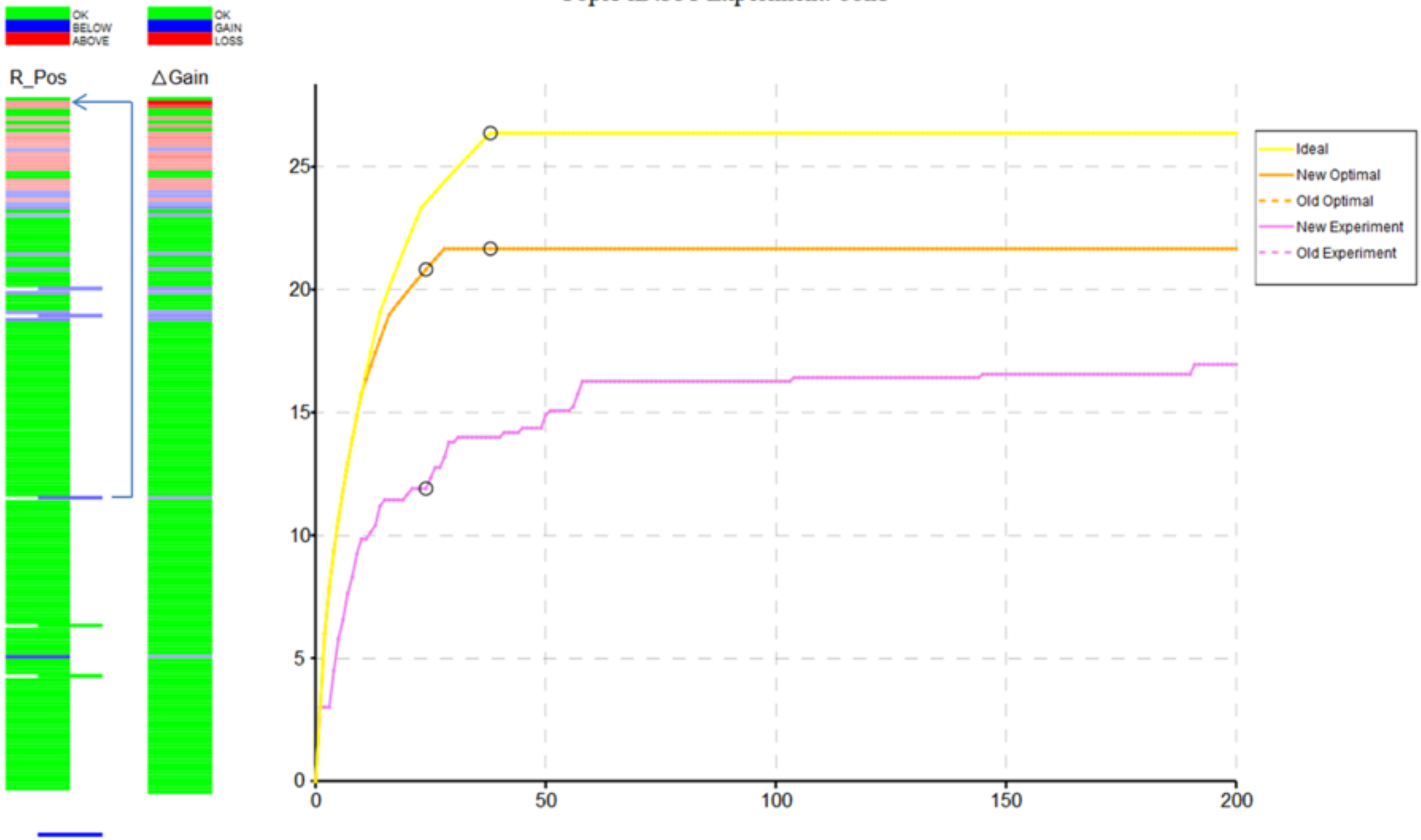
Application Example: Capped Movement



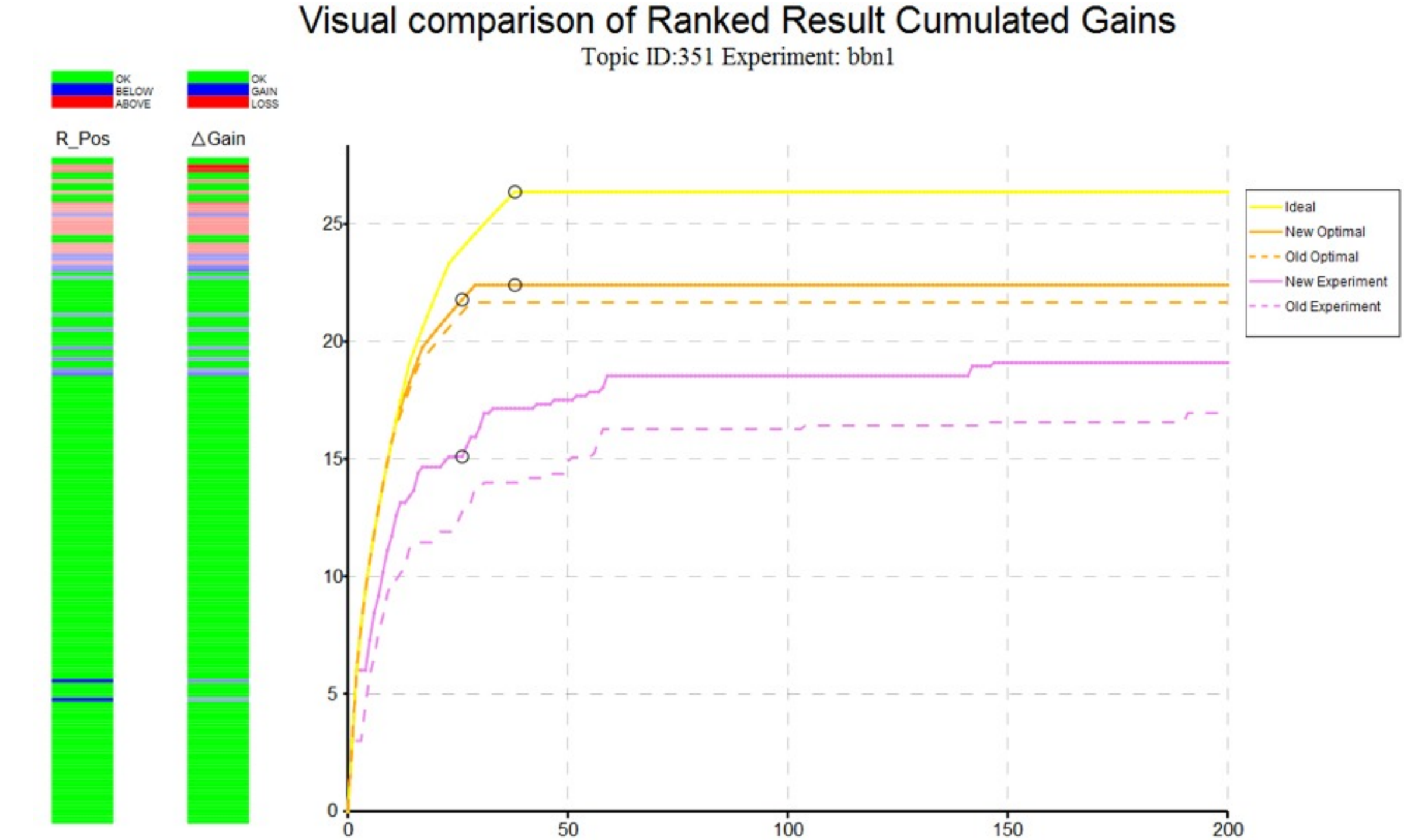
Application Example: Capped Movement



Application Example: New Entry Evolution

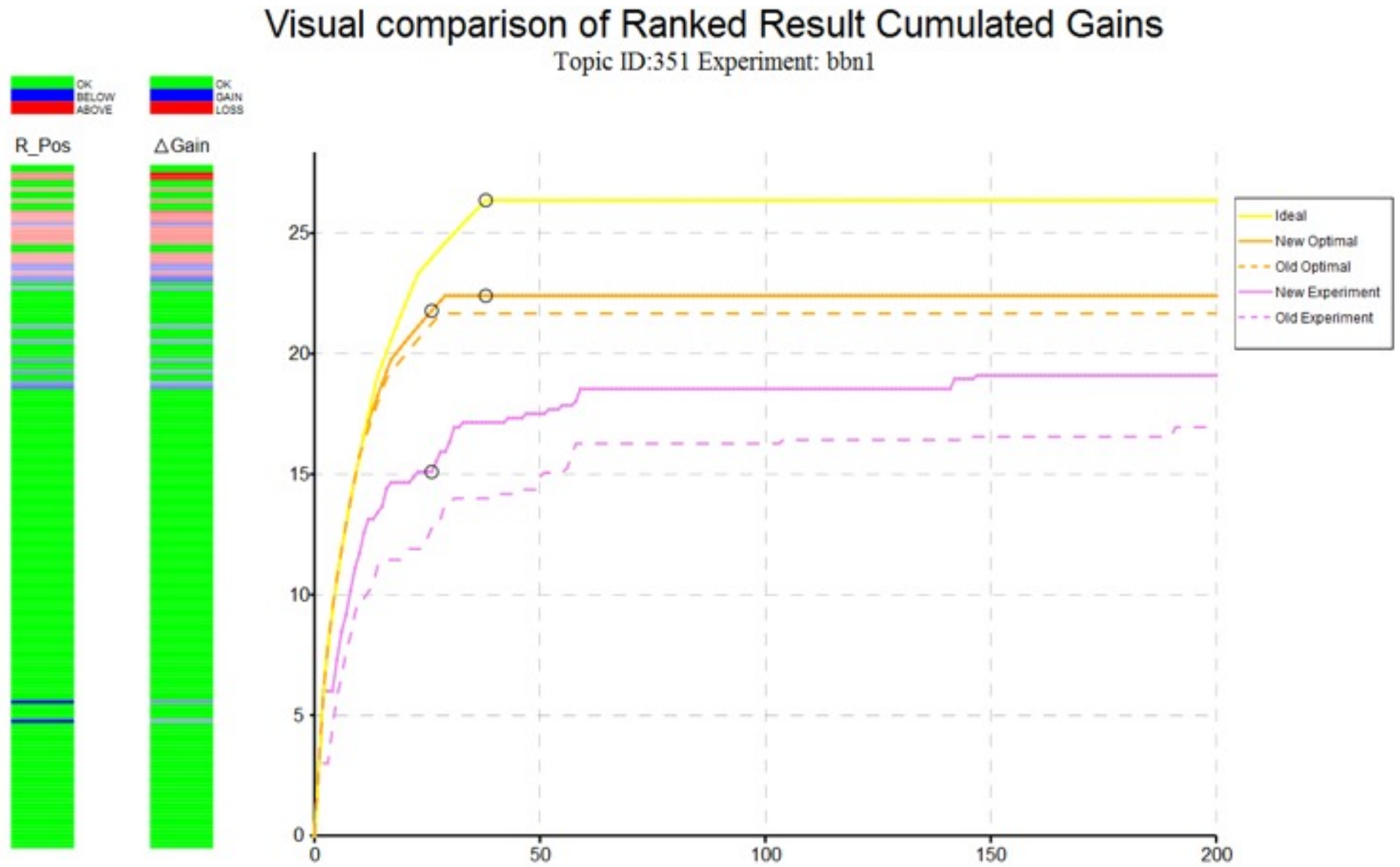


Application Example: New Entry Evolution



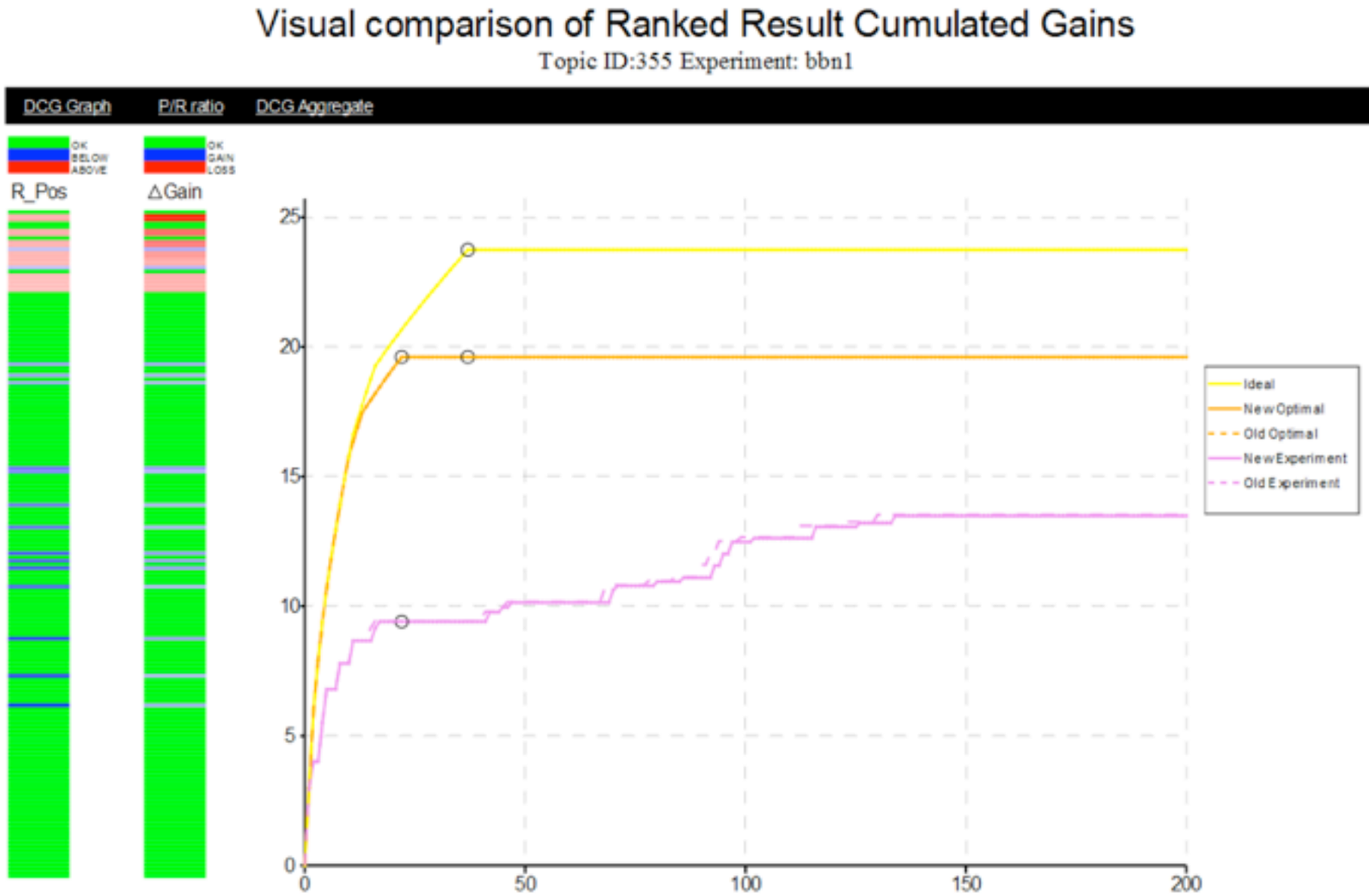
The Domino Effect

How does the change for topic 351 affect the other topics?



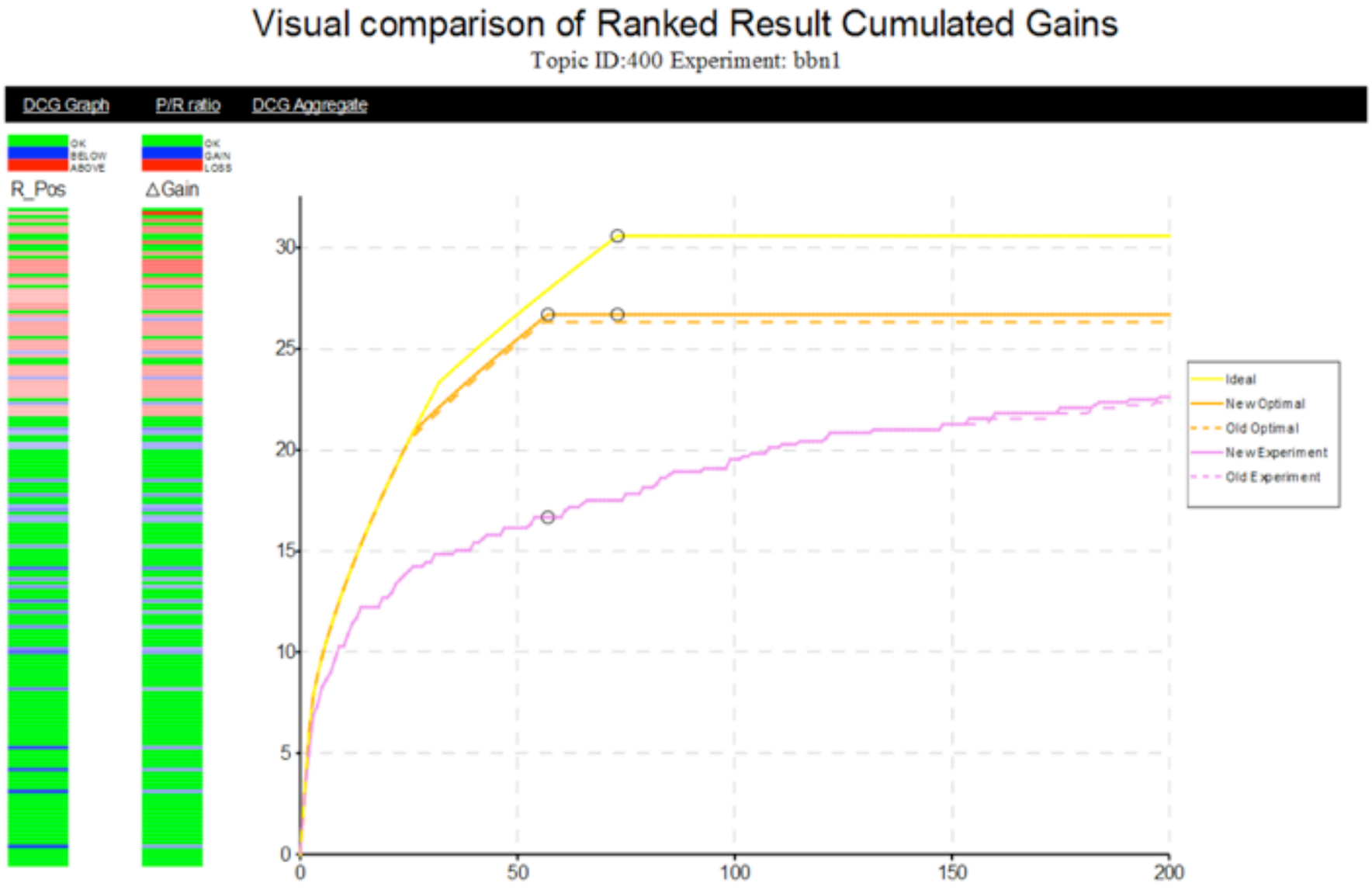
The Domino Effect

The change for 351 worsens the DCG curve of topic 355



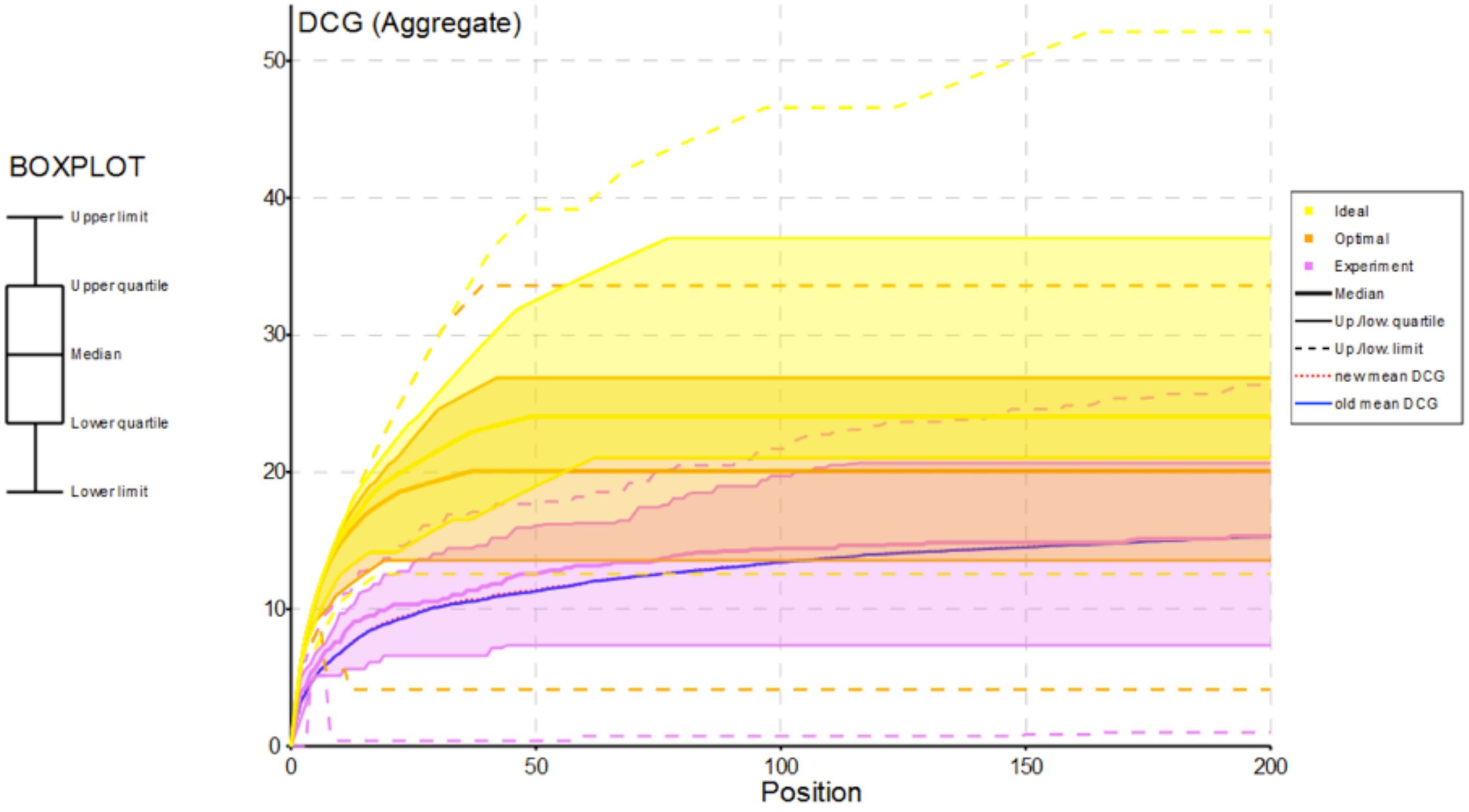
The Domino Effect

The change for 351 improves the DCG curve of topic 400



The Domino Effect

Aggregate view for the whole set of documents before and after the movement



Final Remarks and On-Going Work

- We presented the visual interactive tool allowing analysts to perform failure and what-if analyses
- We described the prototype implementing an actual fusion between IR evaluation and visual analytics tested on the TREC7 collection
- We are performing additional tests employing different learning to rank algorithms to construct the clusters
- We are investigating whether (and how) custom features extraction and selection may allow us to understand on which component of the IR system a change of ranking has an impact on