# ISAD353: Advanced Databases and Data Management

## Seminar 1: Searching

Marco Palomino

# Who am I?

Google marco palomino

All News Images Videos Maps More Search tools

About 509,000 results (0.56 seconds)

Marco Palomino: Research Fellow - University of Exeter Medical School
people.exeter.ac.uk/mp374/
**Marco Palomino** is a research fellow at the University of Exeter Medical School. **Marco Palomino** works at the European Centre for Environment and Human ...

Marco Palomino University of Exeter
medicine.exeter.ac.uk › Medical School › About us › Staff Profiles
Skip to main content Go back to the Medical School home page. University home | Studying | Research | Business and community | Working here | Alumni and ...

Marco Palomino - University of Nottingham
www.cs.nott.ac.uk/~pszmap/
**Marco Palomino**, Nottingham, Publications, Research.

Marco Palomino - Google Scholar Citations
scholar.google.co.uk/citations?user=nryRVlMAAAAJ
**Marco Palomino**. Follow ... MA Palomino, S Bardsley, K Bown, J De Lurio, P Ellwood, . ... MA Palomino, T Taylor, G McBride, H Mortimer, R Owen, M Depledge.

Marco Palomino | LinkedIn
https://uk.linkedin.com/in/**marcopalomino**
Aberdeen, United Kingdom - Research Fellow

Marco Palomino

Update your profile

Research Fellow at University of Exeter
Lives in Plymouth, UK

# Assessed Learning Outcomes

Critically evaluate current and emerging approaches for analysing and interpreting data using data mining and **business intelligence** techniques.

# Business intelligence

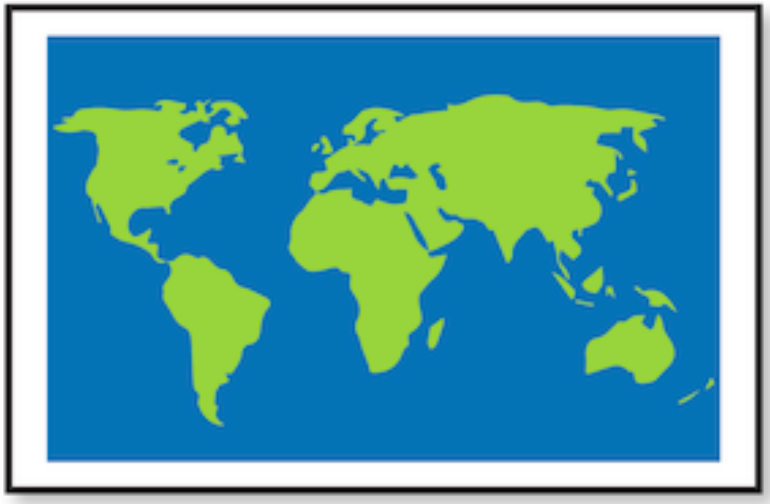
- The term **'business intelligence'** was coined by **Howard Dresner** in 1989 to mean "concepts and methods to improve business decision making by using fact-based support systems".

- Business intelligence (BI) can be described as "a set of techniques and tools for the **acquisition and transformation of raw data into meaningful and useful information** for business analysis purposes".

# Data vs information

DATA

INFORMATION



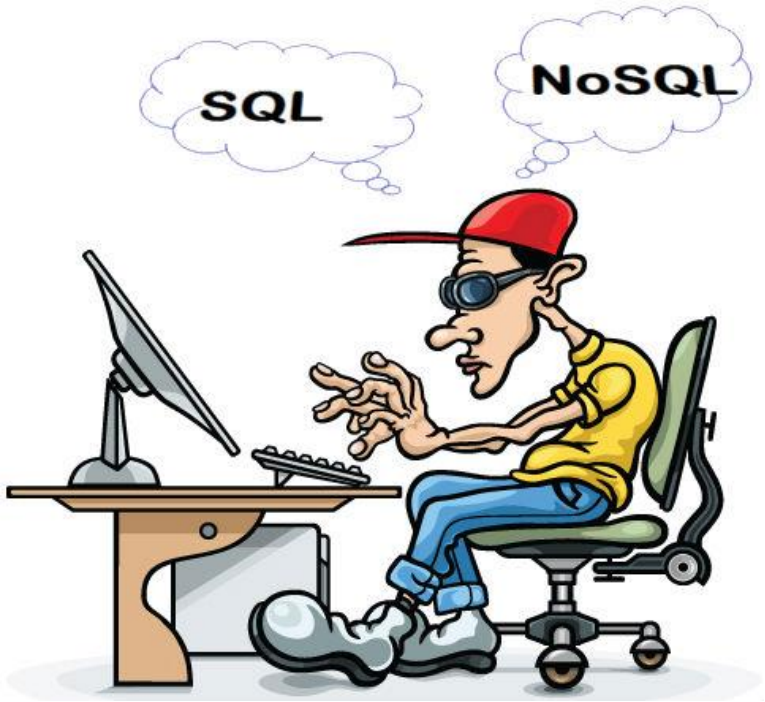
Data is raw, unorganised facts that need to be processed.

Information refers to the meaning of data as understood by some user.

# Module Aims

To expose students to the challenges of solutions for managing, processing, analysing and interpreting large amounts of unstructured data

within relational and **non-relational database** environments.



# Choosing your database (coursework)

# Today's outline

- My background
  - I spent ~3 years at the University of Exeter as a researcher.
  - I carried out a study in collaboration with Lloyd's of London.
  - I worked on a **risk discovery information system**.
- Search
  - **Search engines**
  - **An experiment using Google WUI and API**
  - Google search logs
- Coursework


# Group exercise

- **Search engines** represent one of the most obvious uses of **data mining** and **business intelligence** techniques, requiring the examination of a massive amount of data on the Web to identify and classify resources to answer user queries.

- What is a **search engine**?

- How do you think search engines will **change** in the next **5-10 years**?

    - What sort of new features/services do you envisage?

    - Will search engines disappear?

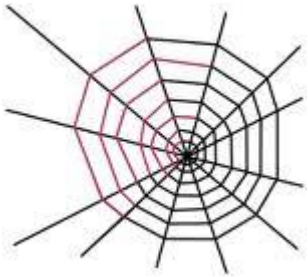
# Let’s begin

- What is a **search engine**?
- How do you think search engines will **change** in the next **5-10 years**?
  - What sort of new features/services do you envisage?
  - Will search engines disappear?

# Definition

- **Search engines** are complex aggregates of several pieces of software that allow users to enter a **query**—a combination of **keywords**—and receive in return a list of **links**, **ranked** according to their relevance to the query.

# Sketching a search engine



Link repository

Scoring results

Results

End user

Query engine

Query

Crawler

Document repository

Crawling Control

Indexer

Inverted index

**From Baldi, Frasconi and Smyth (2003)**
***"Modelling the Internet and the Web"***

# Impact

- Search engines rank **second** only to e-mail as the **most popular** online activity (*Pew Internet Project Report*).

- According to a McKinsey report, employees spend **1.8 hours every day** searching and gathering information—9.3 hours per week, on average.

# Engines into the deep Web?

# Engines into personalised news search?

# From search engines to shopping engines

## Google to deliver goods to online shoppers

USATODAY


*(Photo: Nicholas Kamm, AFP/Getty Images)*

**STORY HIGHLIGHTS**

- **New service called Google Shopping Express**
- **Express will provide same-day delivery of food and other products**

SAN FRANCISCO (AP) — Internet search leader Google is taking another step beyond information retrieval into grocery delivery.

The new service, called Google Shopping Express, will initially provide same-day delivery of food and other products bought online by a small group of consumers in San Francisco and suburbs located south of the city. The company, based in Mountain View, Calif., didn't say how many people will be part of the test.

If the pilot program goes well, Google plans to expand delivery service to other markets.

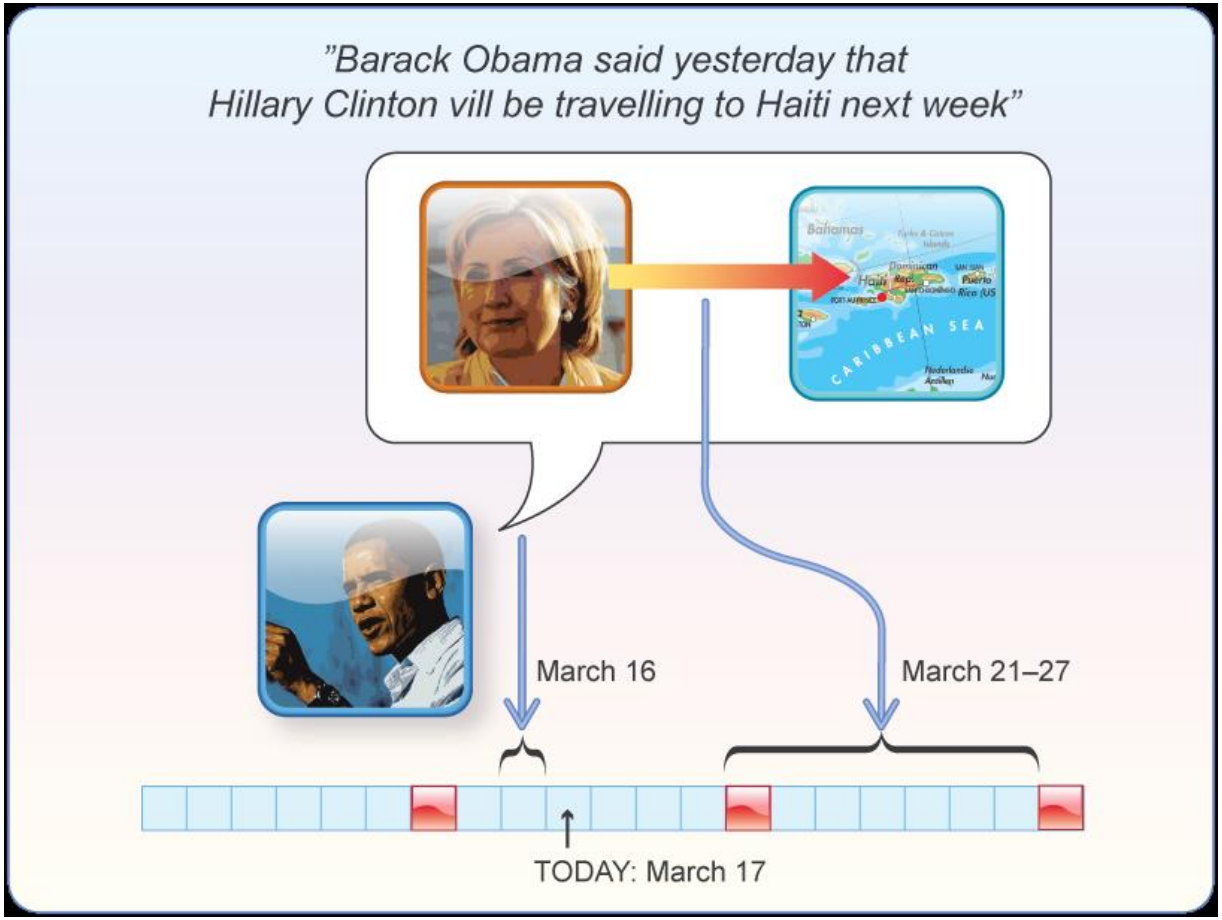
"We hope this will help users explore the benefits of a local, same-day delivery service, and help us kick

# Engines as cyber crystal balls?

- The future is being prepared now, so we may be able to tell where it's headed, **if we can learn to read the Web correctly...**

- Here is a sample text:

*"Barack Obama said yesterday that Hillary Clinton will be travelling to Haiti next week." – March 17th.*



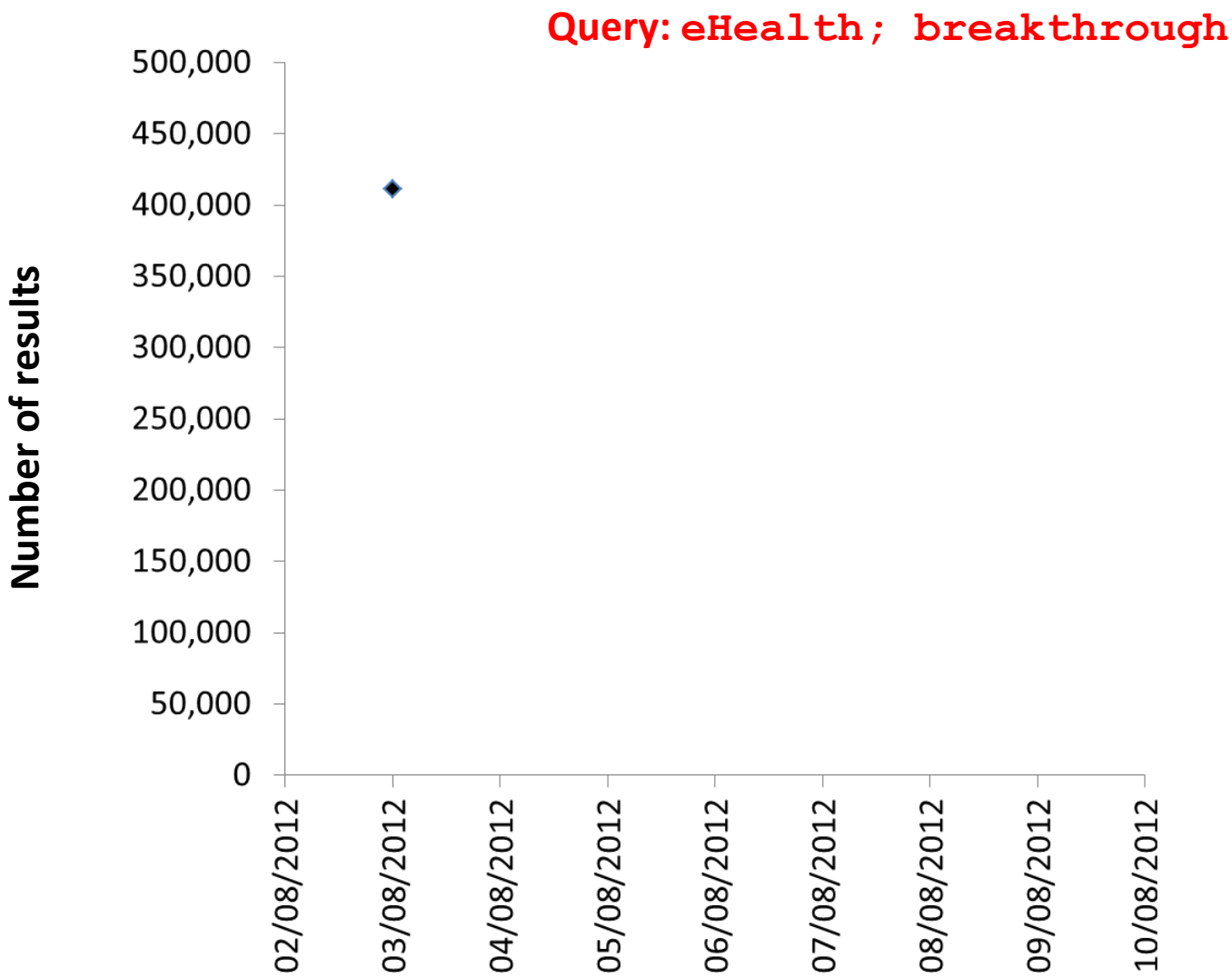

- This sentence can be represented as two events (**a quotation event in the past**, and **a travel event in the future**) on a timeline.

# Searching for future information: A search that lasted 100 days

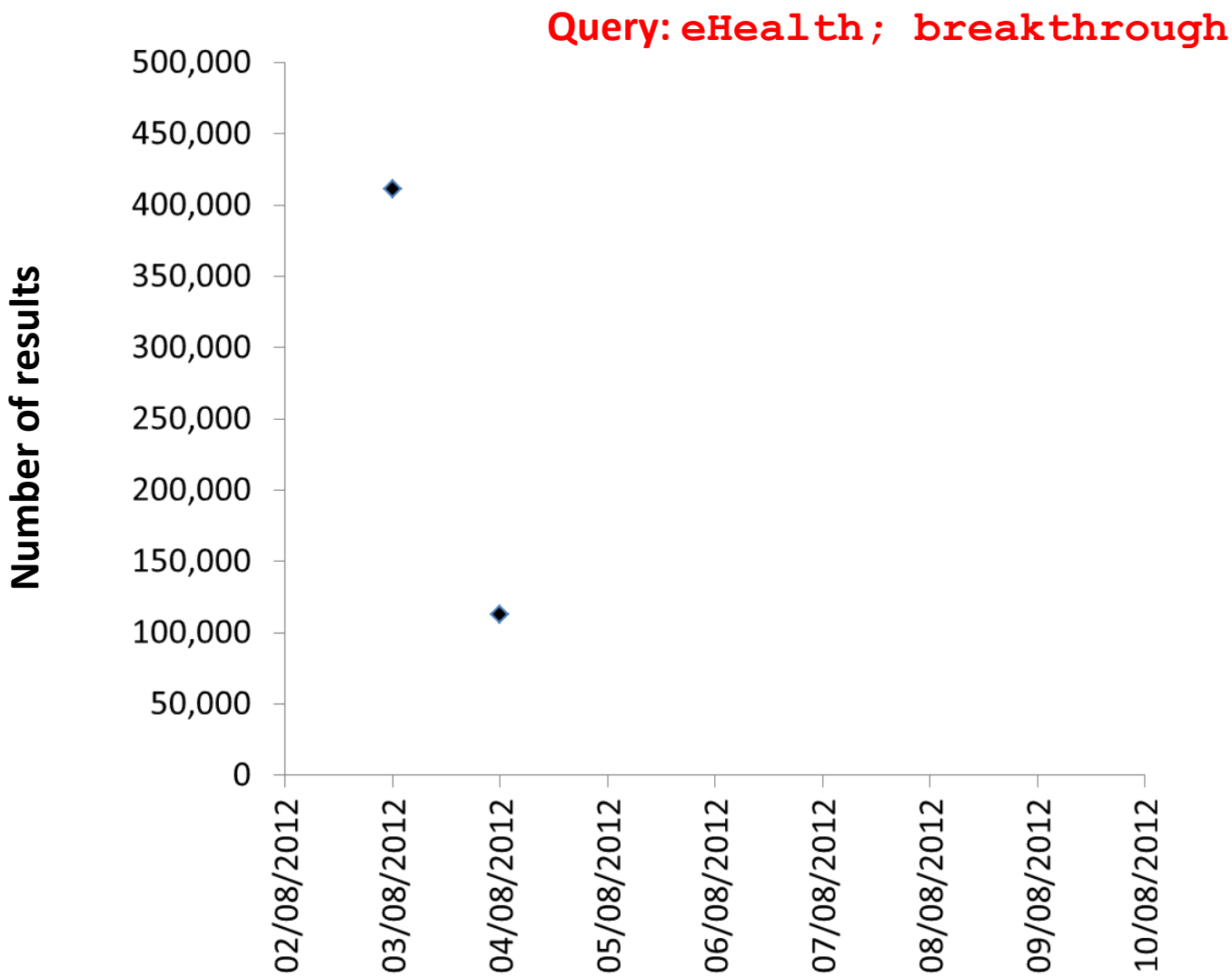
eHealth: A relatively recent term for healthcare practice supported by electronic processes and communication.

| Search terms (query) |
| --- |
| ehealth; **breakthrough** |
| ehealth; **"closer to reality"** |
| ehealth; **"paves the way"** |
| ehealth; **"previously impossible"** |
| ehealth; **revolutionary** |
| ehealth; **unprecedented** |

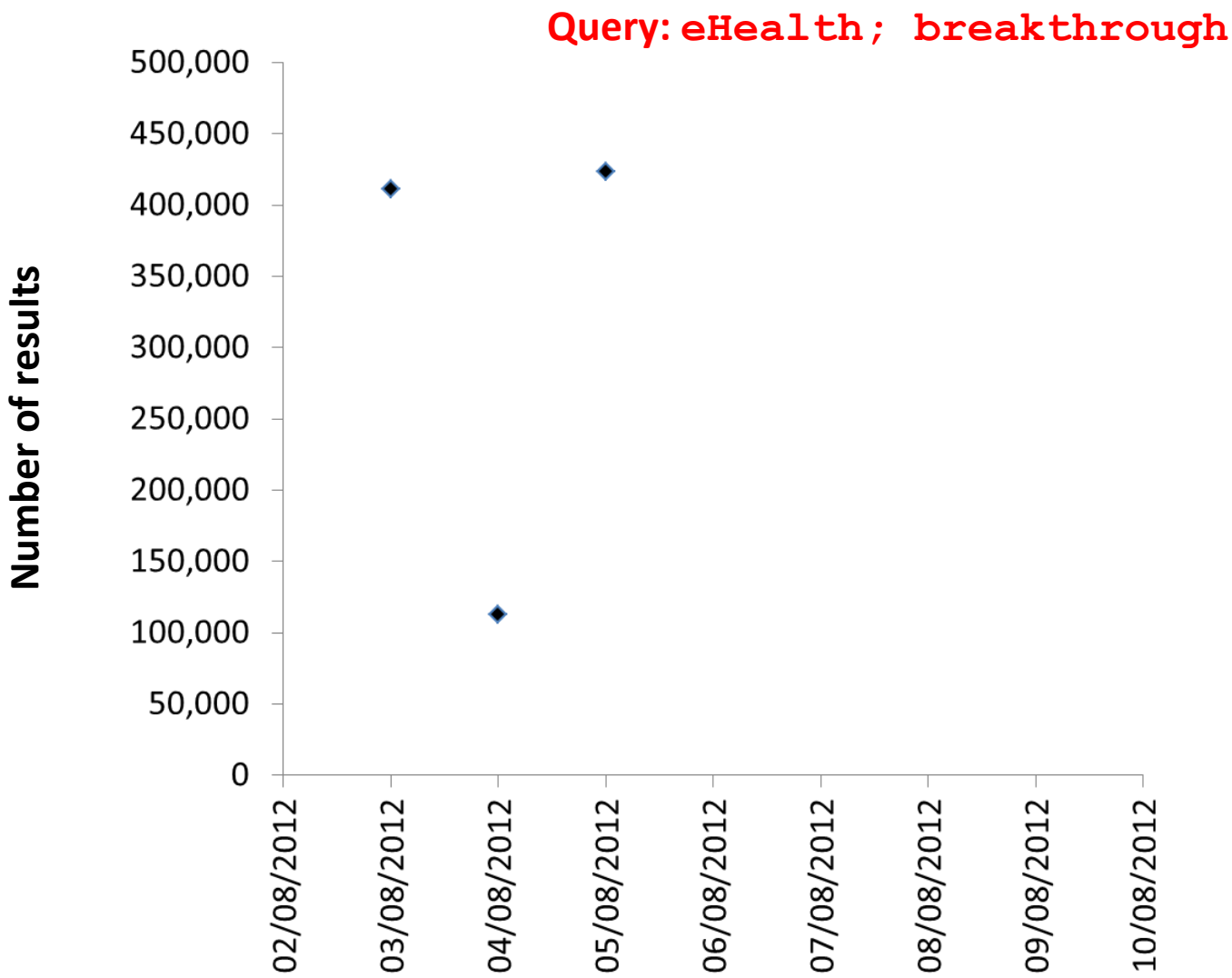
# Google results



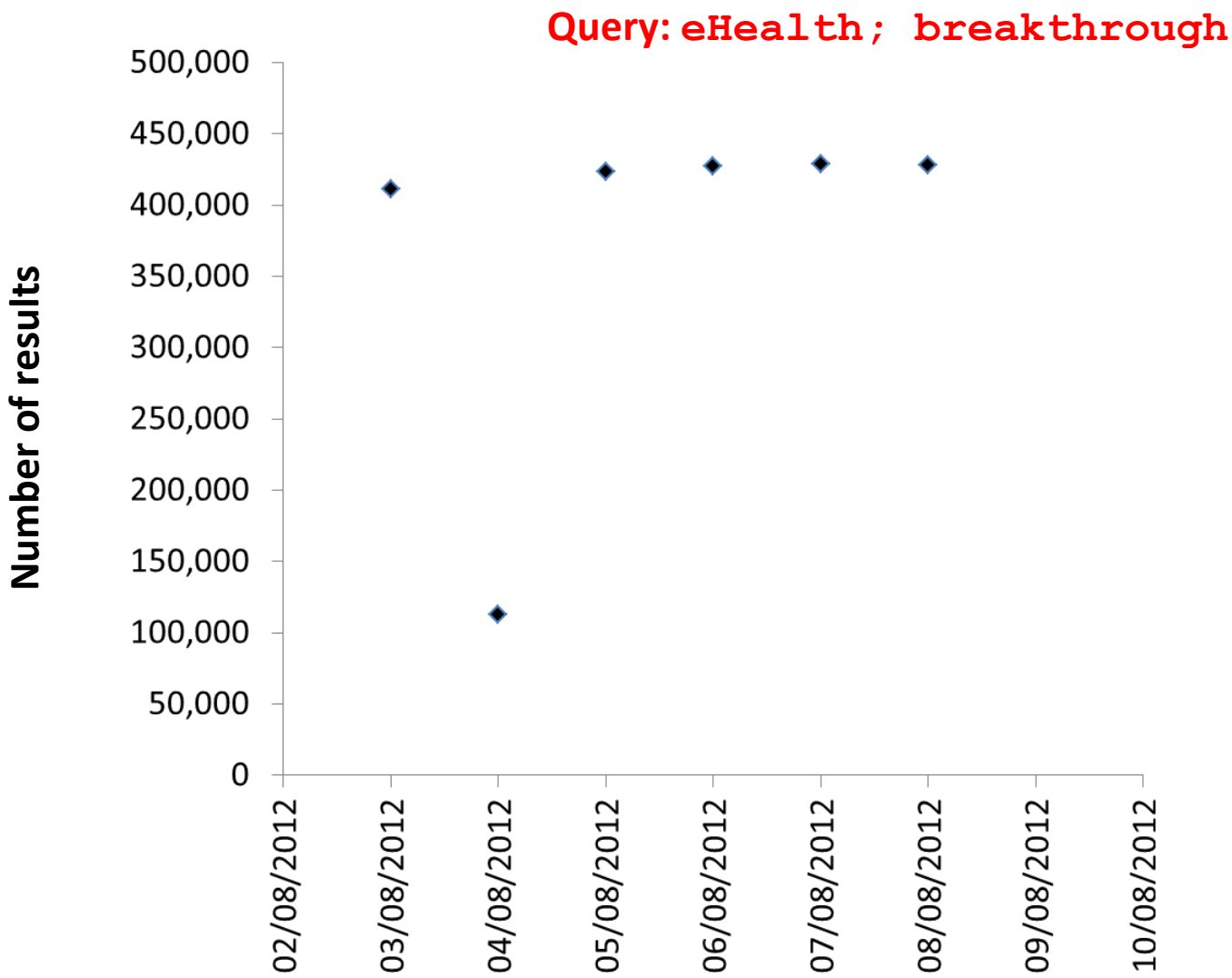
# Google results



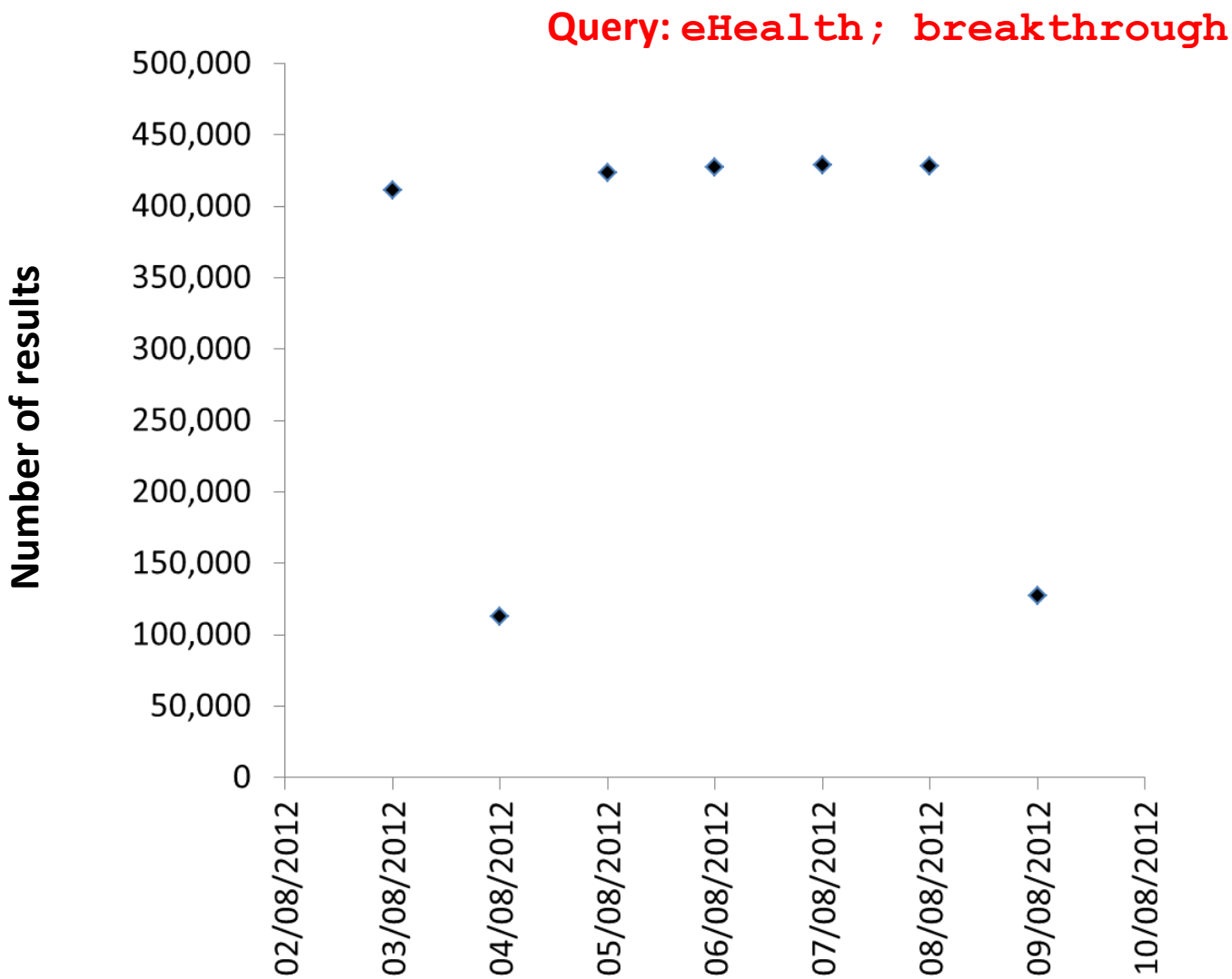
# Google results



# Google results

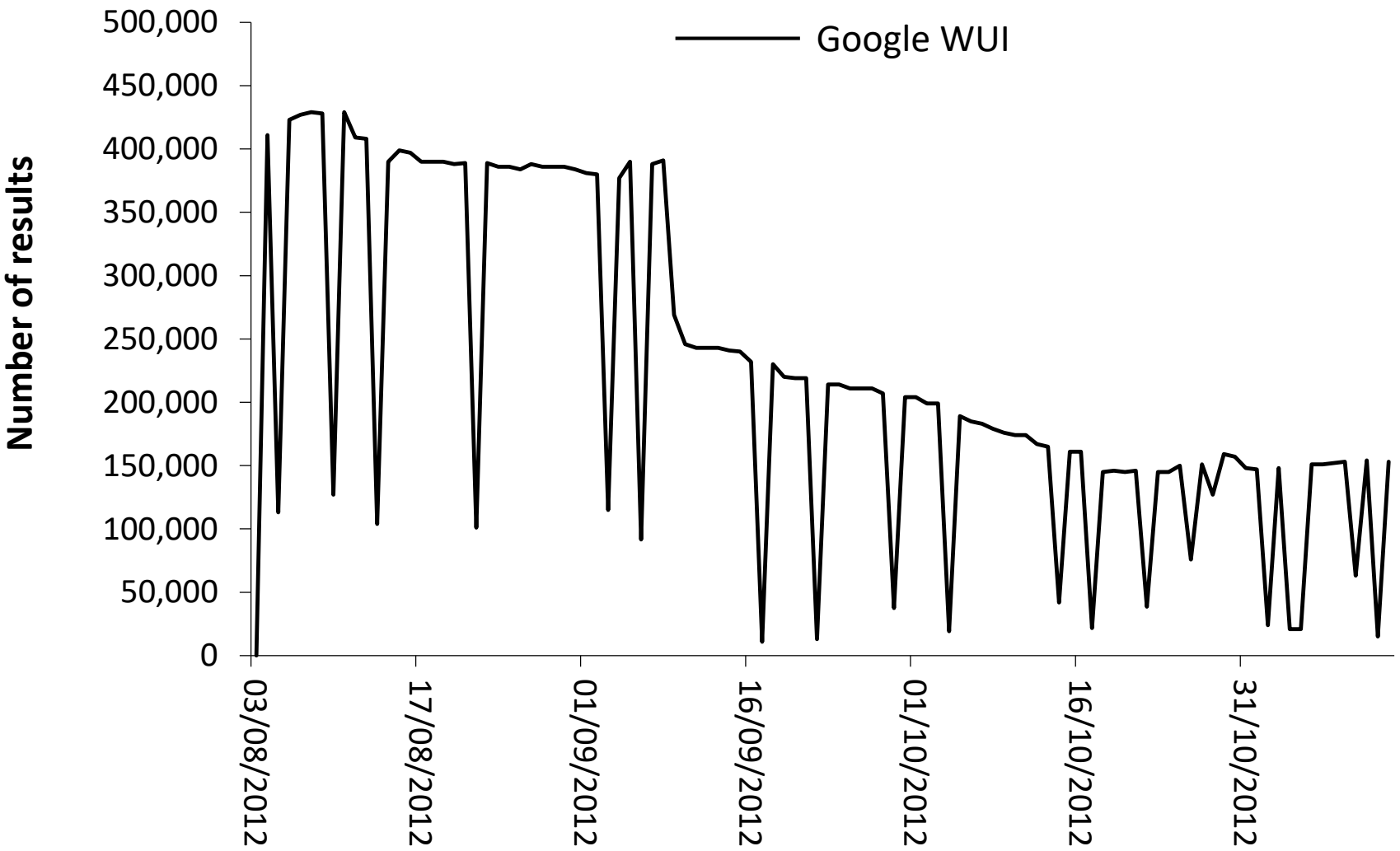


# Google results



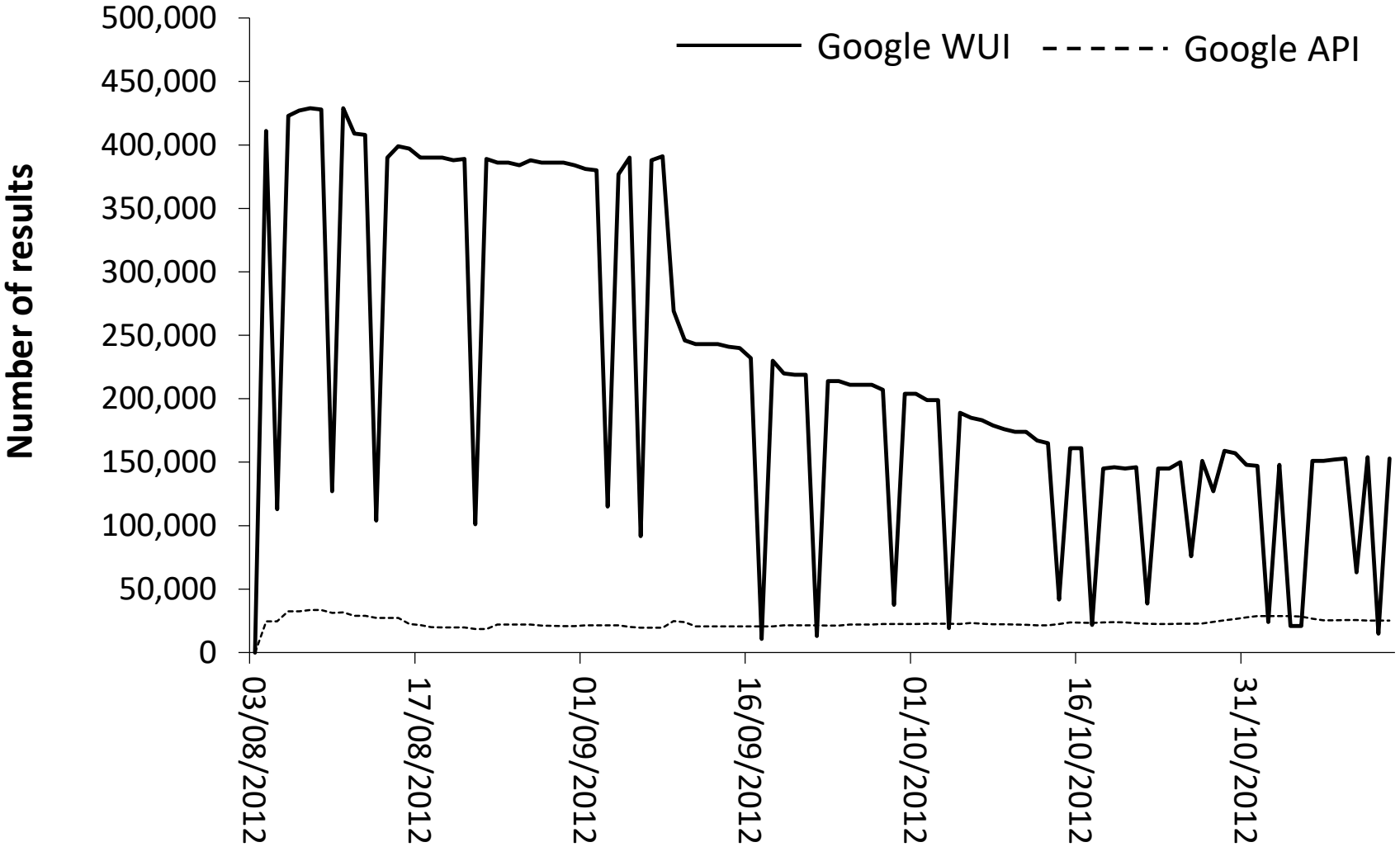
# Google results

**Query: `eHealth; breakthrough`**



# Google results

**Query: eHealth; breakthrough**



# Same day instability

| Query | Google's WUI 05/12/2012 09:45 | Google's WUI 05/12/2012 17:45 | Google's API 05/12/2012 09:40 | Google's API 05/12/2012 17:40 |
|---|---|---|---|---|
| ehealth; breakthrough | 167,000 | 167,000 | 21,500 | 21,400 |
| telemed; breakthrough | **377** | **197** | 669 | 665 |
| ehealth; "closer to reality" | **3,340** | **3,470** | 470 | 469 |
| telemed; "closer to reality" | 12,300 | 12,300 | 457 | 457 |
| ehealth; "paves the way" | **9,350** | **8,830** | 1,460 | 1,460 |
| telemed; "paves the way" | 22,800 | 22,800 | 1,140 | 1,170 |
| ehealth; "previously impossible" | 82 | 82 | 103 | 103 |
| telemed; "previously impossible" | **61** | **603** | 501 | 502 |
| ehealth; revolutionary | 21,200 | 21,200 | 14,200 | 14,400 |
| telemed; revolutionary | **364** | **594** | 1,740 | 1,750 |
| ehealth; unprecedented | **12,100** | **54,200** | 9,650 | 9,650 |
| telemed; unprecedented | **589** | **585** | 621 | 623 |

# Observations – Google study

- M. Palomino, et al., *"Instability in Search Engine Results: Lessons Learned in the Context of Horizon Scanning Applications"*, in Proceedings of the 24th International Conference on Database and Expert Systems Applications (DEXA), Prague, Czech Republic, August 2013, pp. 53 - 57.

- Google's API offers less documents than Google's WUI, but the documents are **more recent**.

- The same behaviour shown by Google's interfaces appears to be shown by **Bing's interfaces**.