

## **Yahoo! Data Quality: Embedded DQ Strategy, Statistical Influences, and Challenges in Yahoo!'s Massive Data Environment**

### **ABSTRACT**

---

While data is vital to Yahoo!, quality data is a key to its success. Consistent, accurate, and overall high-quality data are needed to give reliable insights on reach, engagement and monetization. This presentation describes the Data Quality team's approach to ensure the highest data quality by applying recognized best practices with a customized approach in the very large data-intensive organization. Recently, to encourage ownership of data quality throughout the organization, Yahoo! instituted an embedded, company-wide data quality program. In addition, the DQ team began investing in improvements to data forecasting and monitoring to aid the new program. These improvements reduced errors and encouraged buy-in from across the organization. The presentation also focuses on the technical challenges associated with protecting data accuracy from malicious traffic in Yahoo!'s massive data environment.

### **BIOGRAPHY**

---

#### **Jeff Kibler**

Data Quality Lead, Audience Management & Analytics  
Yahoo!

Jeff Kibler earned his Bachelors in Computer Science and his Masters of Business Administration from the University of Illinois at Urbana-Champaign. At Motorola, he designed and developed test policies and procedures using six sigma methodologies within the cellular phone business. He currently acts as the dq lead for audience measurement and analytics within the Yahoo! data quality team.



#### **Oleksiy Chayka**

Yahoo! / University of Trento, Italy

Oleksiy Chayka received his joint Masters in Computer Science from the University of Trento ( Italy ) and RWTH Aachen ( Germany ) in 2007. In his current PhD coursework at the University of Trento , he researches data quality management within the European project OKKAM and the MASTER project. Since March 2010, he works as a DQ intern at Yahoo!. Oleksiy also works in system analysis and application of data mining techniques to data quality assessment. He is a member of W3C Incubator group on data provenance and Database Group of Trento DBTrento).



## Yahoo! Data Quality

Embedded DQ Strategy, Statistical influences, and Challenges in  
Yahoo!'s Massive Data Environment



1

MIT IQIS '10  
Jeff Kibler  
Oleksiy Chayka

**YAHOO!**

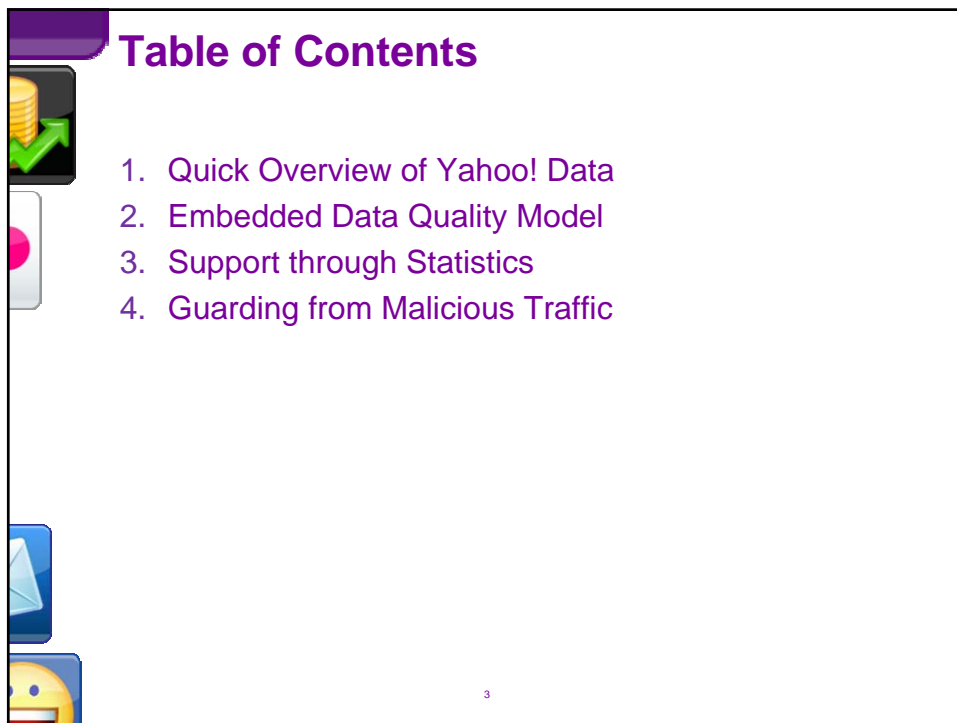
6/29/2010

### Abstract



While data is vital to Yahoo!, quality data is a key to its success. Consistent, accurate, and overall high-quality data are needed to give reliable insights on reach, engagement and monetization. This presentation describes the Data Quality team's approach to ensure the highest data quality by applying recognized best practices with a customized approach in the very large data-intensive organization. Recently, to encourage ownership of data quality throughout the organization, Yahoo! instituted an embedded, company-wide data quality program. In addition, the DQ team began investing in improvements to data forecasting and monitoring to aid the new program. These improvements reduced errors and encouraged buy-in from across the organization. The presentation also focuses on the technical challenges associated with protecting data accuracy from malicious traffic in Yahoo!'s massive data environment.

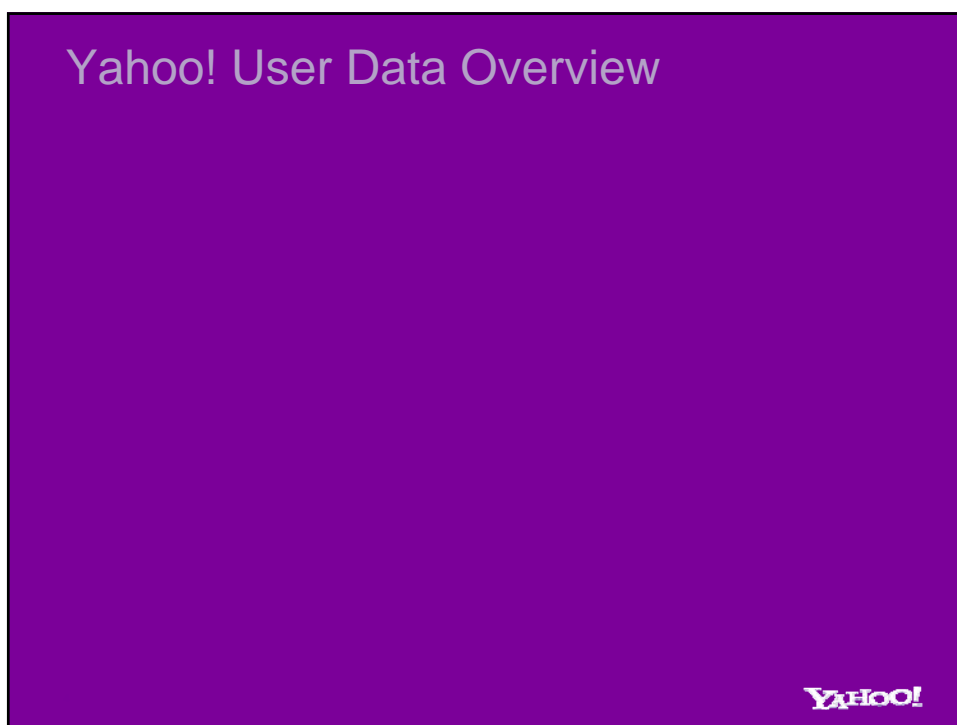
2

The slide features a purple header bar with the title 'Table of Contents' in white. To the left of the list are four icons: a stack of gold coins with a green arrow pointing up, a pink circle, a blue envelope, and a yellow smiley face.

## Table of Contents

1. Quick Overview of Yahoo! Data
2. Embedded Data Quality Model
3. Support through Statistics
4. Guarding from Malicious Traffic

3



## Yahoo! User Data Volume and Definition

Collects over a dozen terabytes of data per day [1]

- U.S. Library of Congress equivalents *every day*

Leading Internet Portal and Software Supplier [2]

- Serves 150+ MM US users – 75%+ of US internet users
- Top ranked in 11 sites (Mail, Messenger, FP, Finance, etc.)



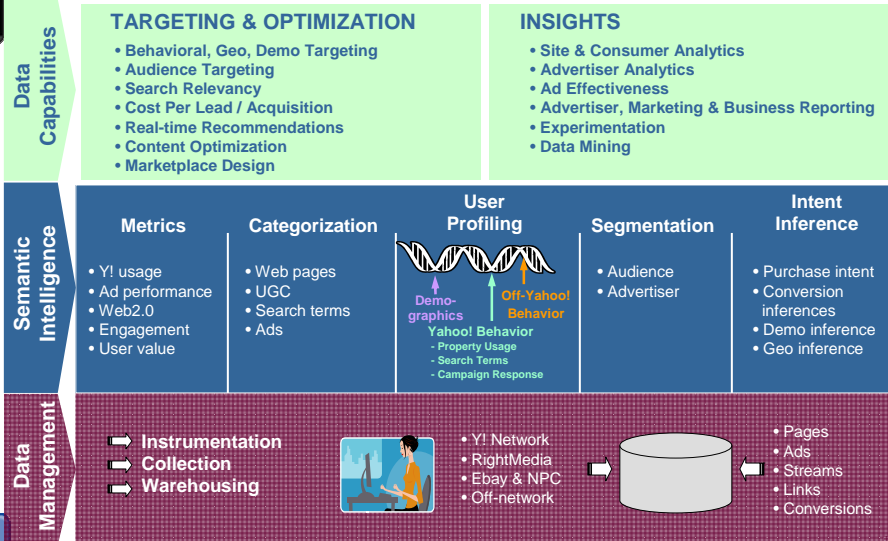
User Data & Analytics – Organization Dedicated to Data R&D

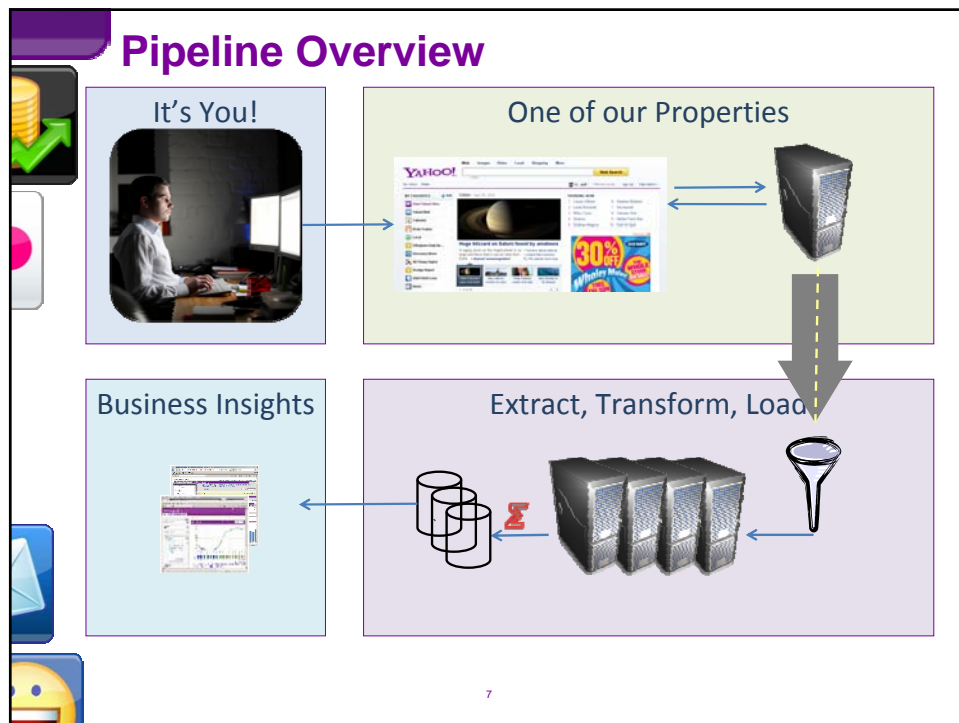
- Use of data for business insights
  - Improving the product through “advanced usability testing”
  - Leader in advanced user profiling & behavioral targeting
  - Ads become “content” and user experience is enhanced
- Data is Key to Measuring Company Growth
  - Quarterly reporting of key metrics to Wall Street, the press, etc.
  - Engagement: unique users, time spent and return users
  - Effective metrics program is critical to guiding the business

[1] Baeza-Yates, Ricardo and Raghu Ramakrishnan. Data Challenges at Yahoo!. EDBT 2008. March 2008.

[2] August 2008 US Yahoo! Audience Measurement Report. comScore, June 2009.

## Yahoo! User Data Management and Application



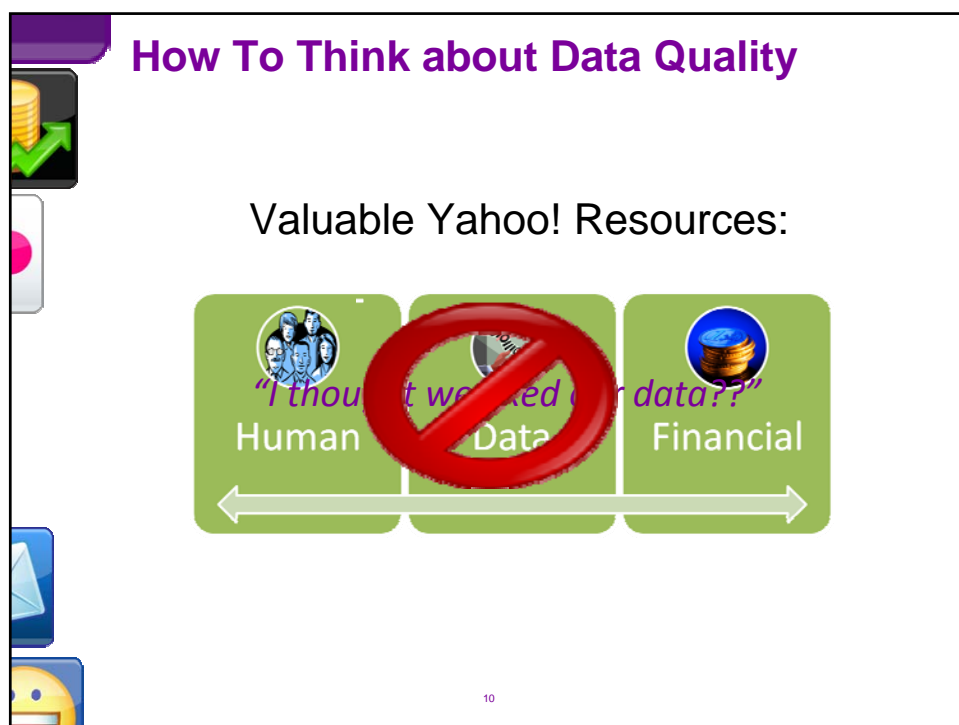
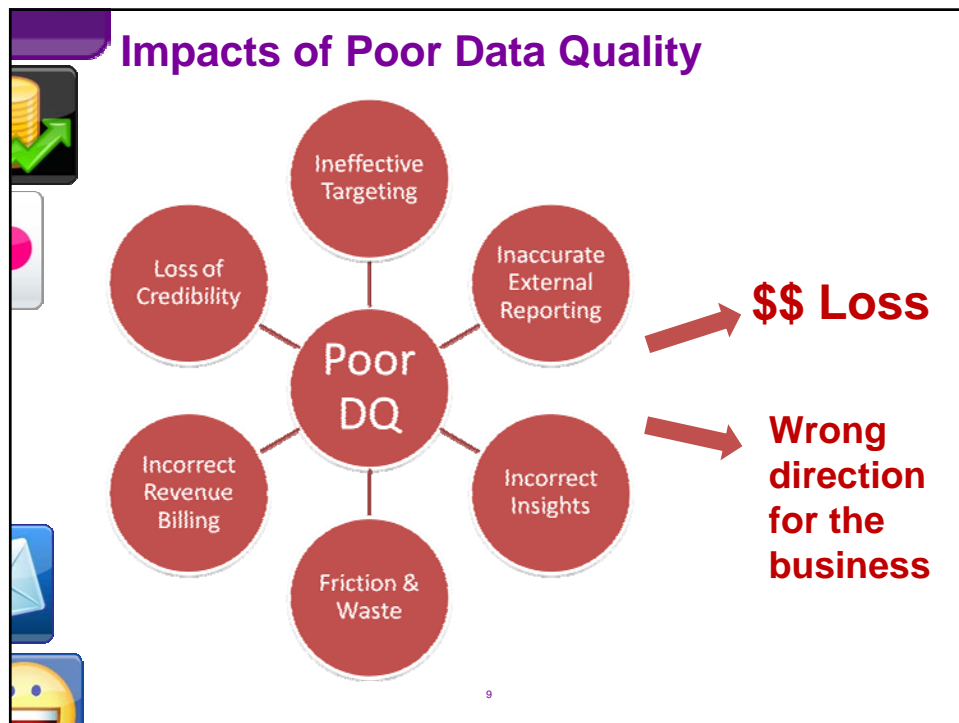


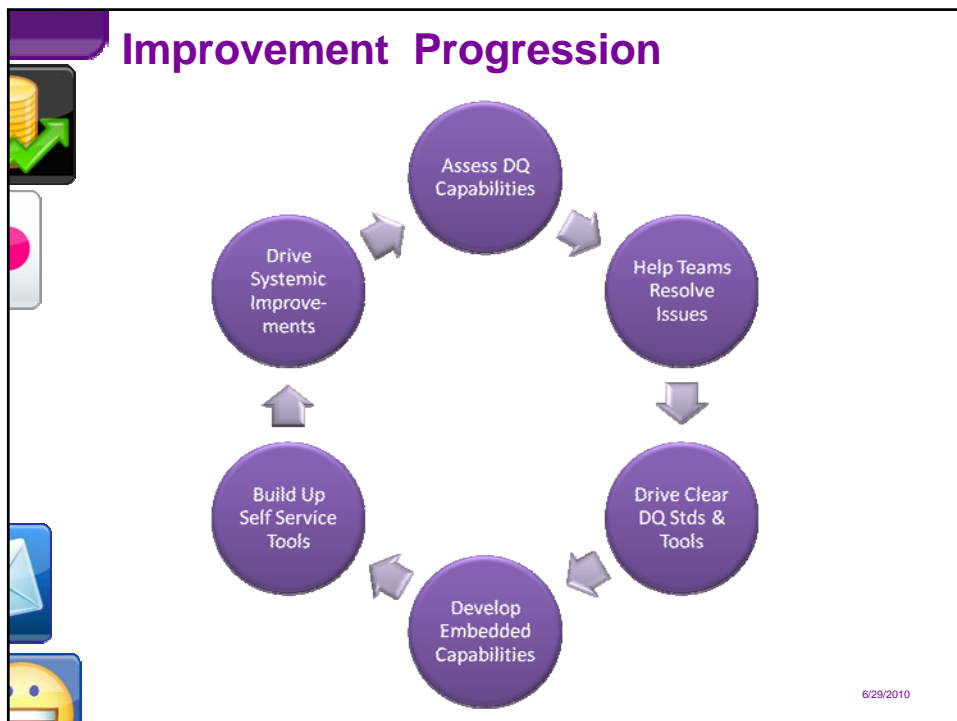
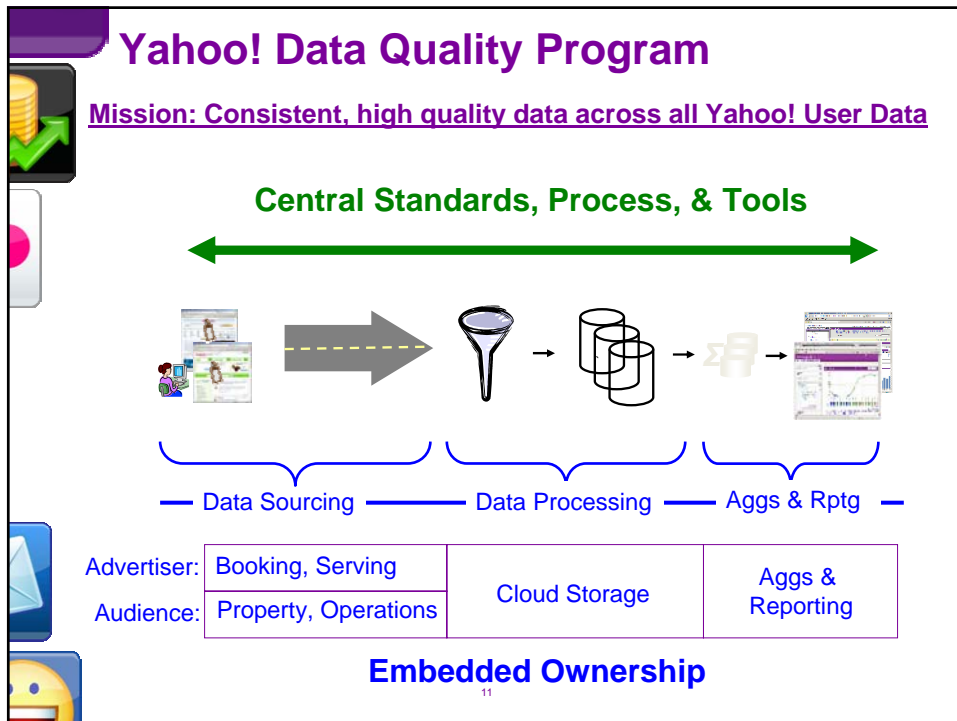
## Embedded Data Quality Model

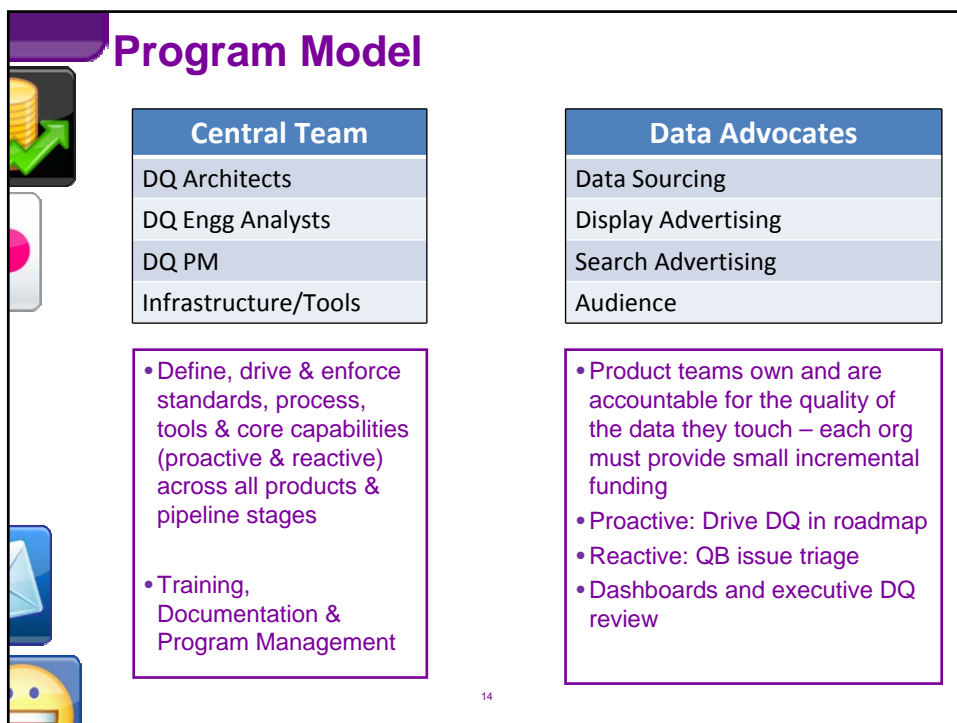
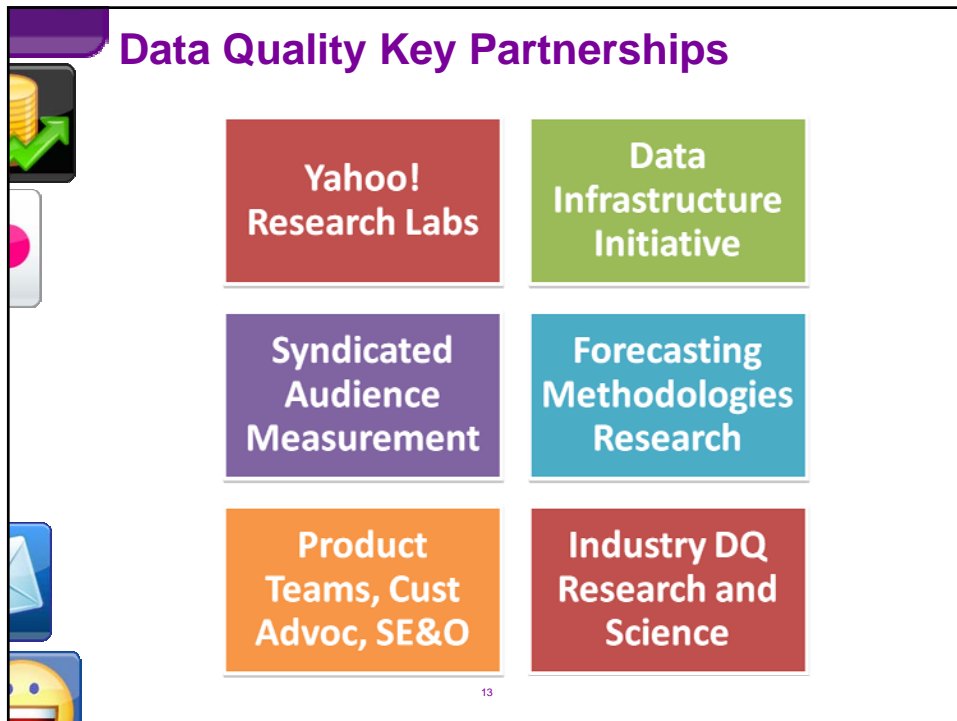
### Yahoo!'s Cross-Organizational Ownership





Jeff Kibler

YAHOO!









## Organizational Success Factors

Engage key DQ leverage points

- Organic alliances with people who feel DQ Pain
- “Data Quality is Everyone’s Job”
- Clear charter and communication in teams’ language

Evolutionary approach

- Start with inventory, then simple monitoring and add on
- Learn as you go and become recognized experts

Education, communication and teaming

- Why DQ is important and current status against goals
- How to use DQ tools and processes to improve
- What is needed from each team for DQ improvement?

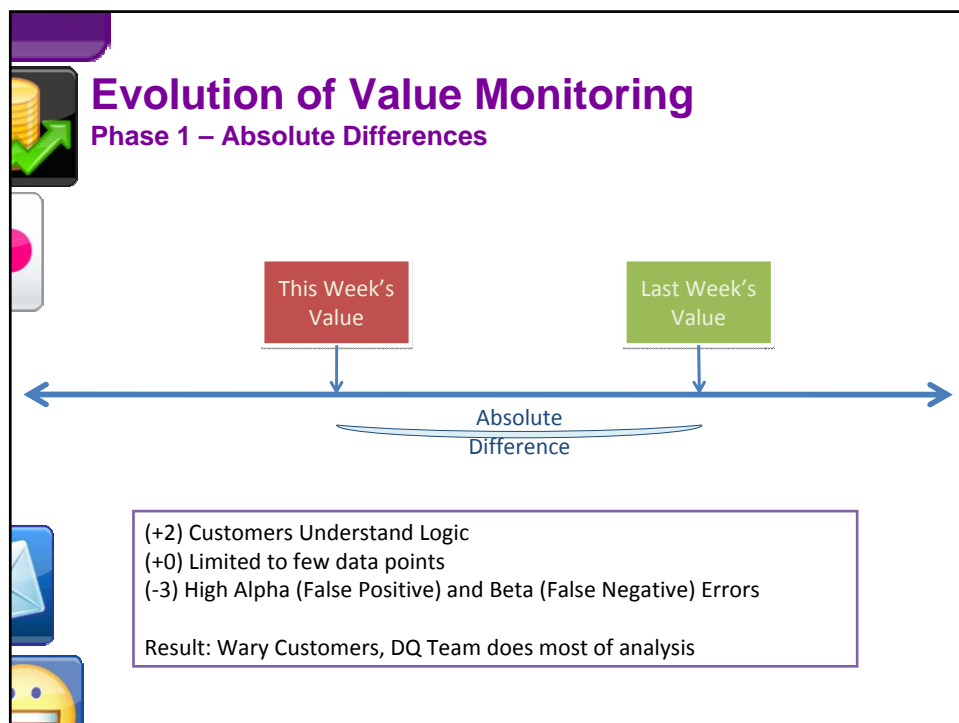
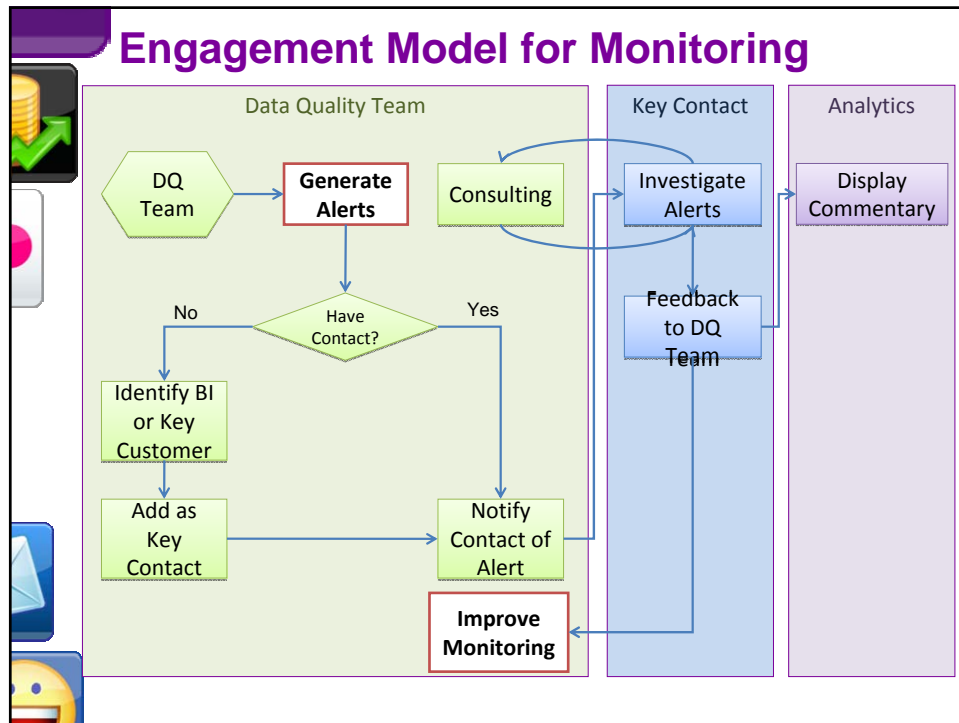
15

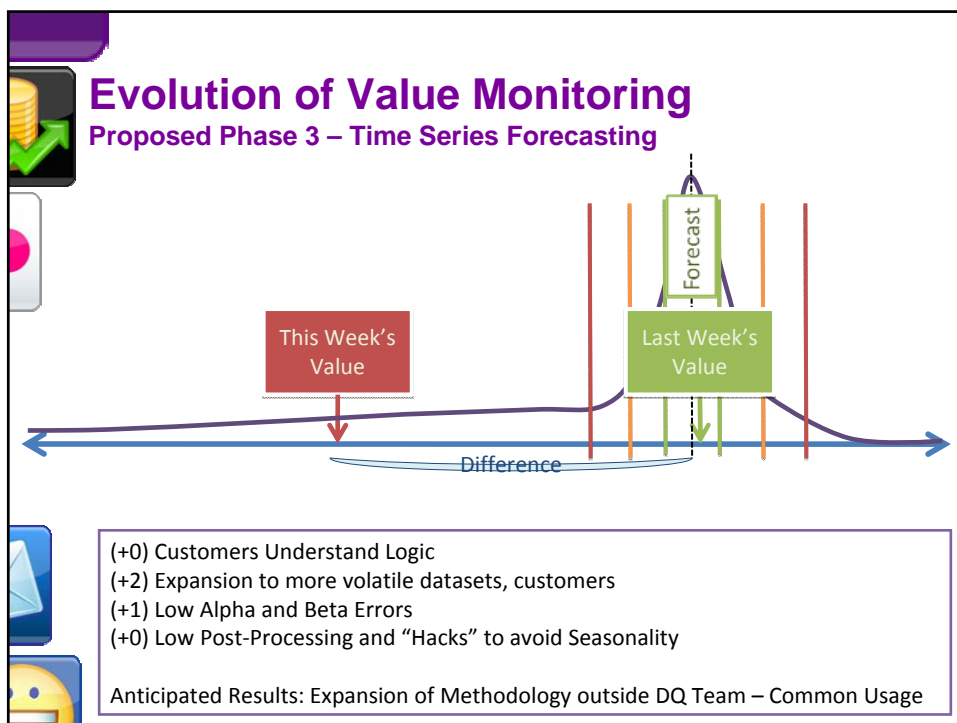
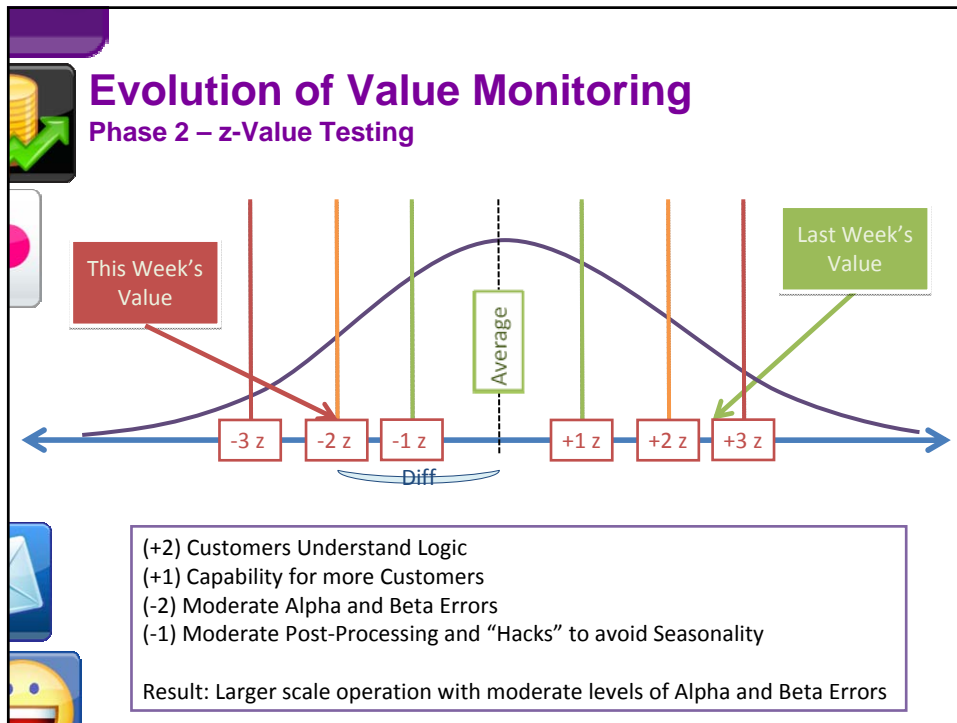
## Support through Statistics

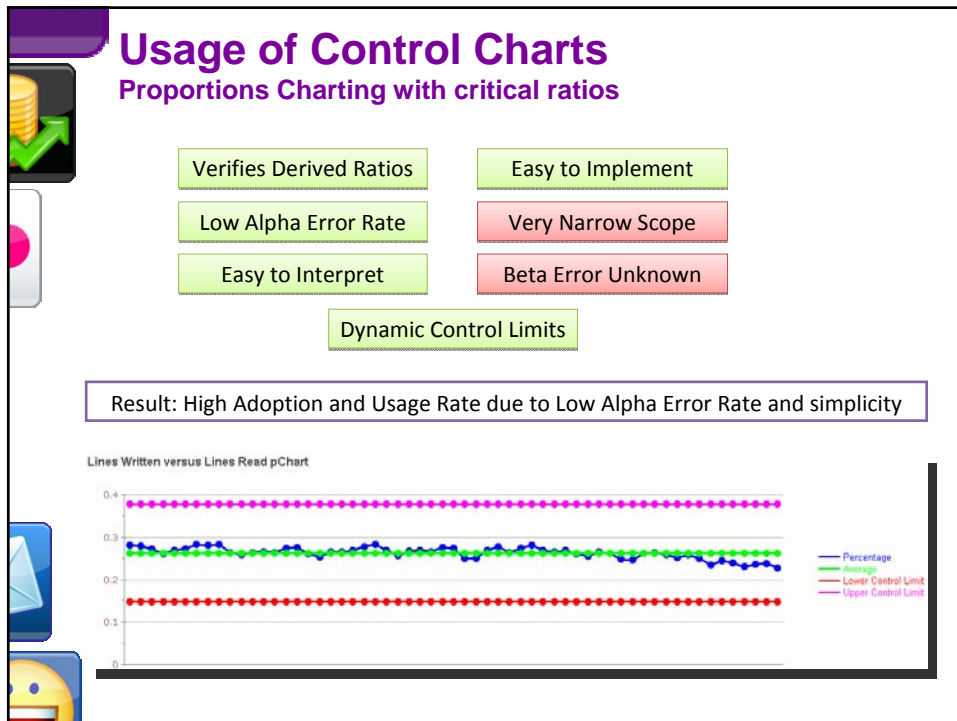
### Selling DQ through Data Quality Monitoring

Jeff Kibler









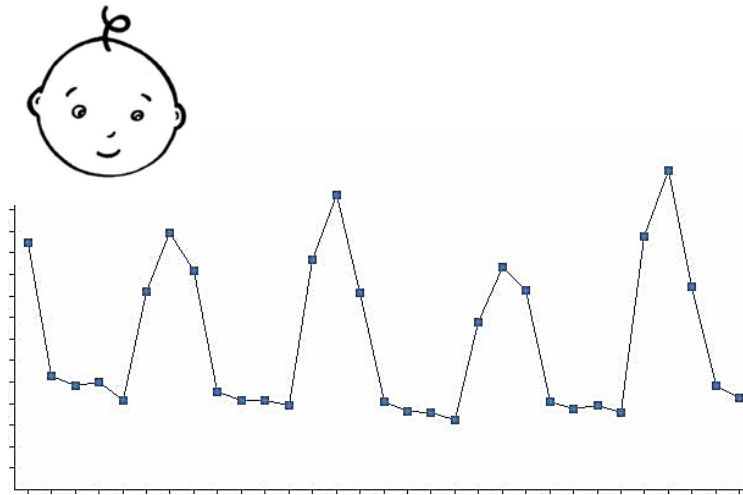
## Guarding from Malicious Traffic

### Traffic Protection in Yahoo!

Aleksey Chayka

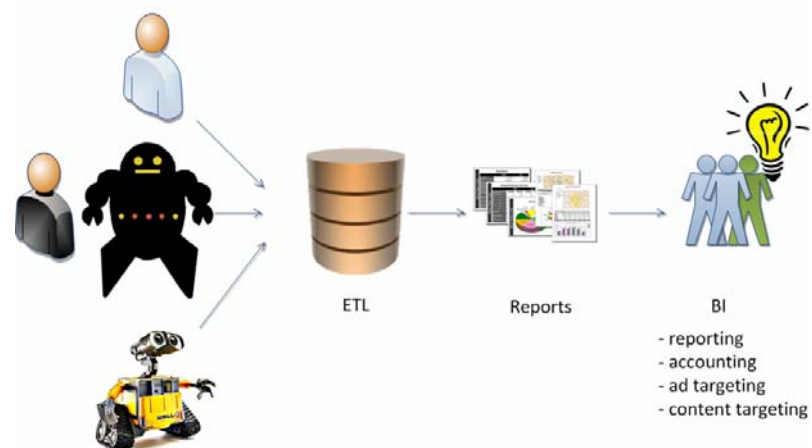


## Understanding people behind traffic



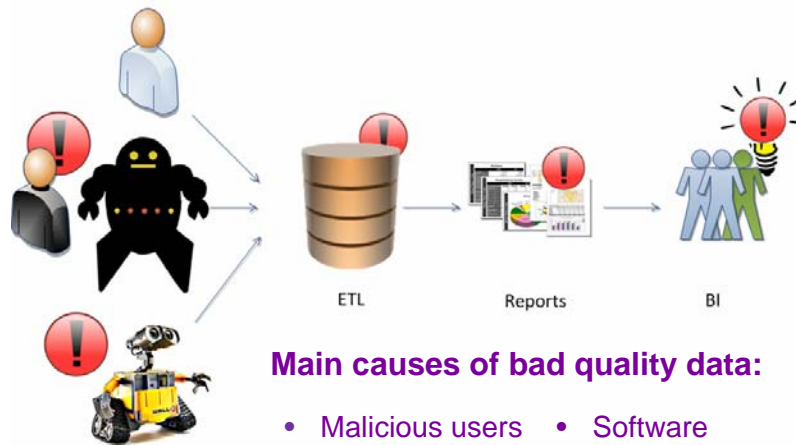
23

## Traffic Protection: Motivation



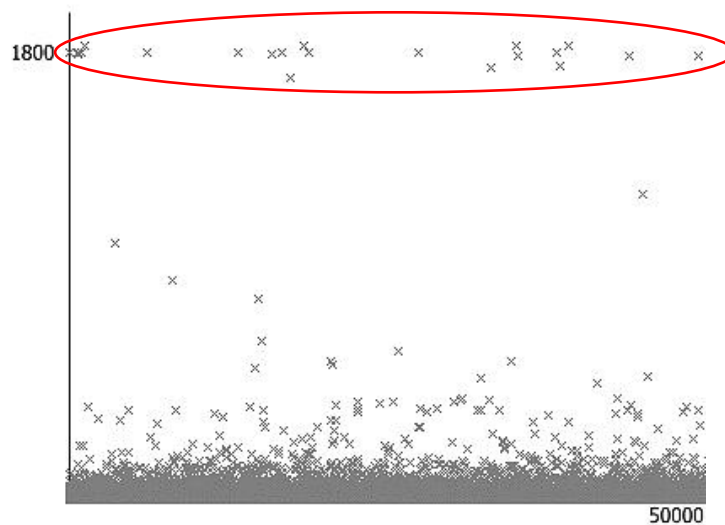
24


## Traffic Protection: Motivation



25

## Data analysis: Filtering spikes and detectable issues





## Traffic Protection: attributes for analysis

### Yahoo! ID

- ✓ User identification (?)
- ✓ Consistent view on user activities in Yahoo! network

- Many people uses Yahoo! without registration (e.g., search)
- Each user may have many yuid
- Each yuid may be used by many users/robots

### IP address

- ✓ User's PC identification (?)
- ✓ Geo location


- Public PC
- Dynamic IP address

### Browser Cookie


- ✓ Bind to one PC/user (?)
- ✓ Reach statistics
- ✓ Most traffic has browser cookies

- Public PC
- Cookie churn


27



## Abuser identification

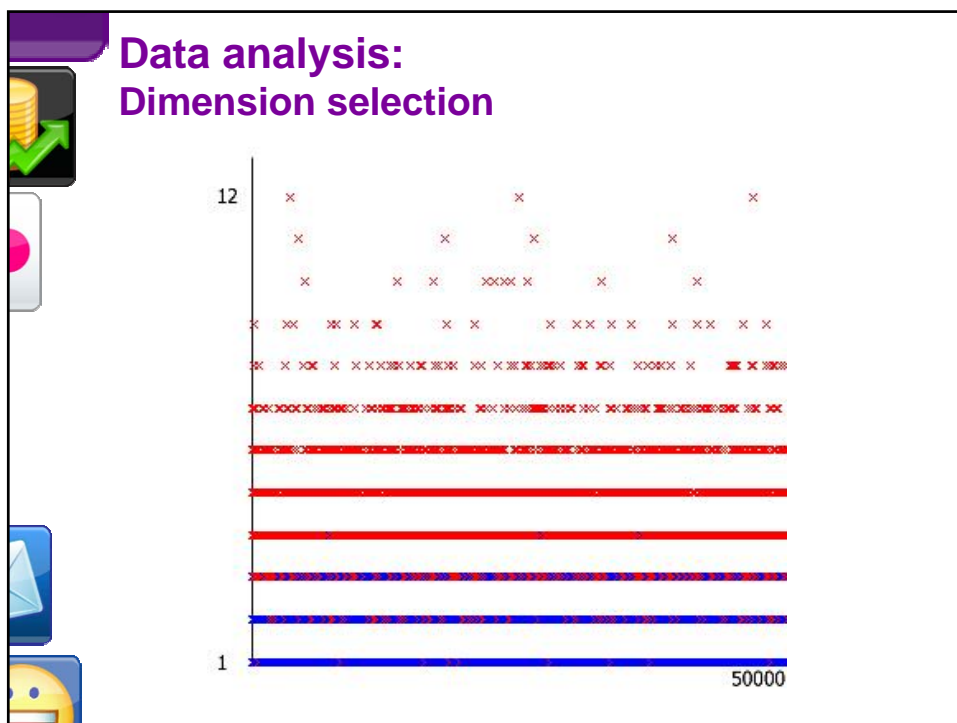
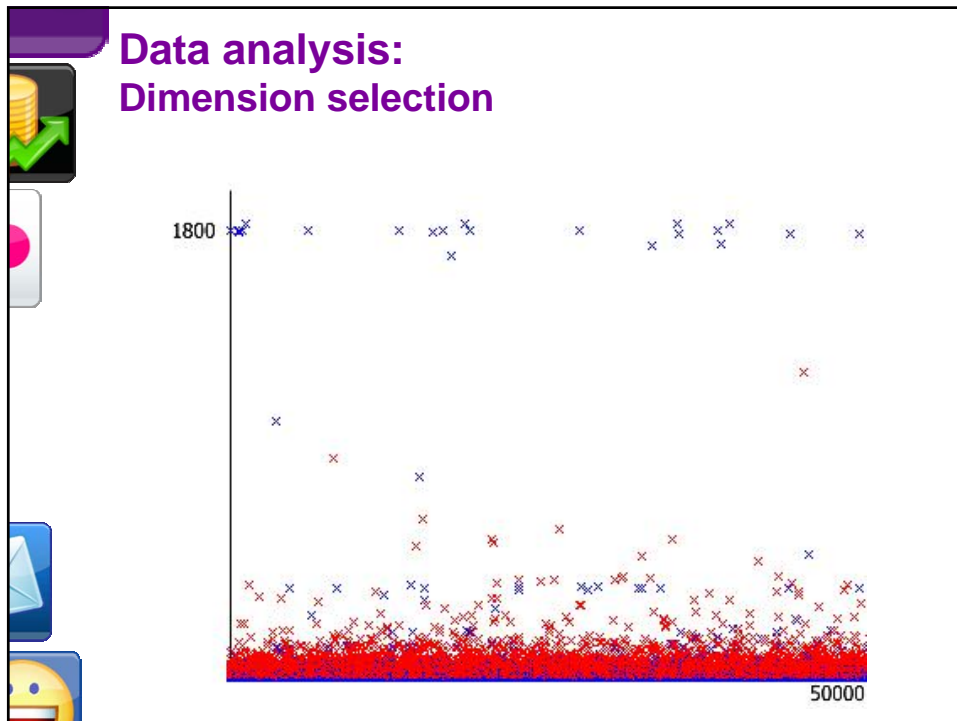


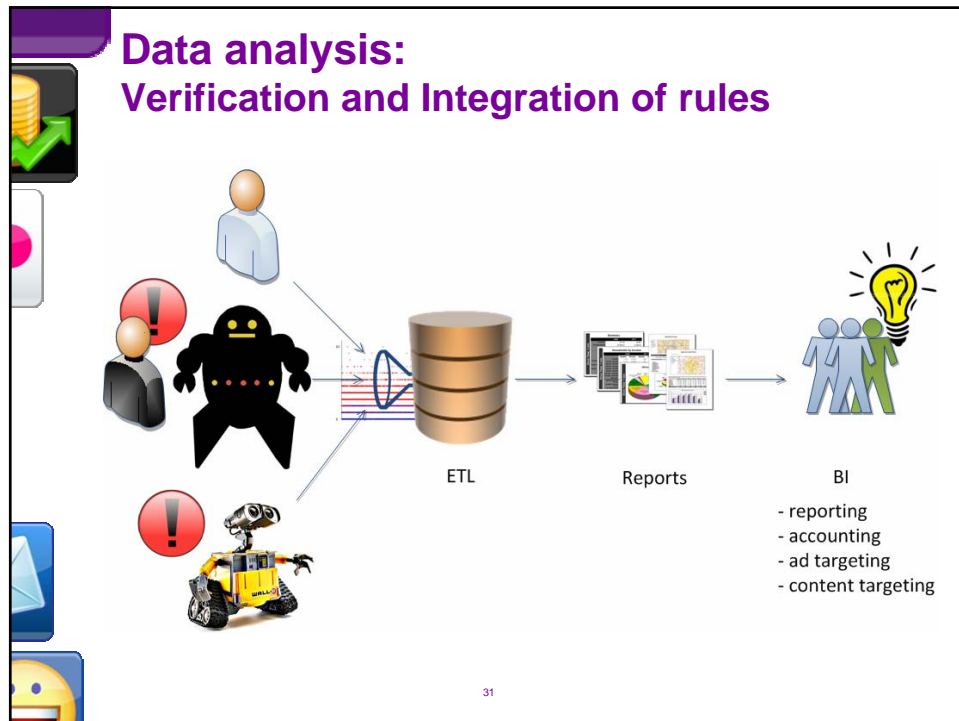
- No. of clicks
- No. of page views
- No. of properties visited
- No. of yuid used
- No. of IP used
- Average time spent on page or property, etc.



## Machine learning

28





## Questions

Jeff Kibler / Oleksiy Chayka

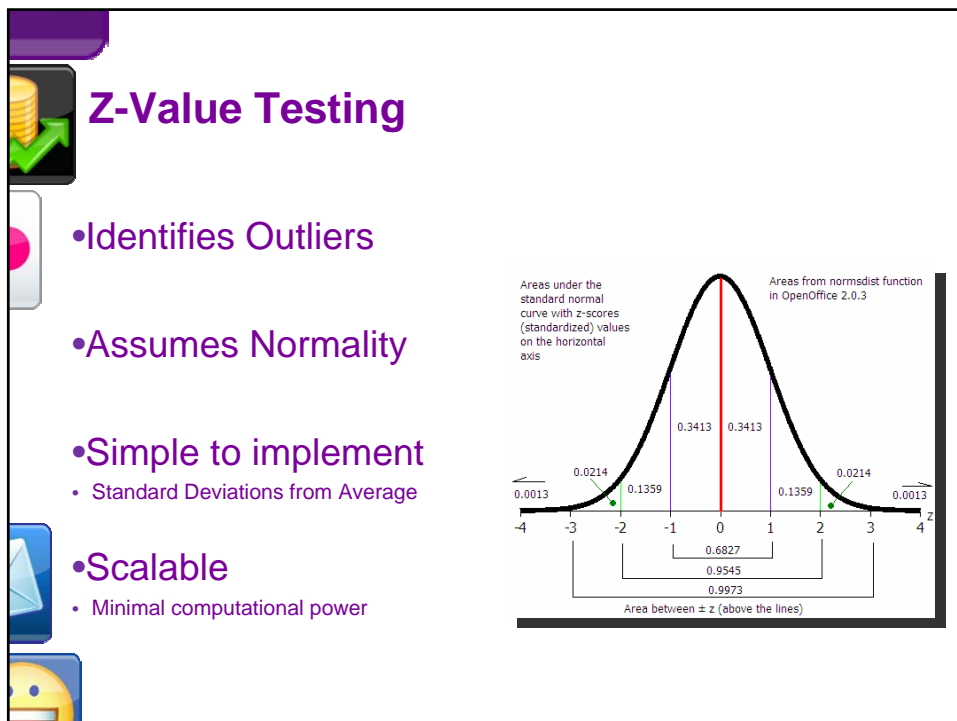
[jkibler@yahoo-inc.com](mailto:jkibler@yahoo-inc.com)  
[aleksey@yahoo-inc.com](mailto:aleksey@yahoo-inc.com)

2021 S. First Street  
Champaign, IL 61822

**YAHOO!**

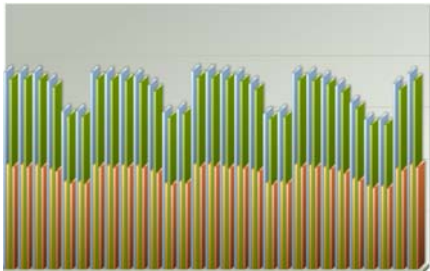
## Backup

YAHOO!



## Yahoo!'s Usage of Z-Value Testing

- Oracle-Driven
  - Stored Procedure (AIR)
  - Scalar Processing time
- Large-Scale Pipeline Monitoring
  - Adequate for consistent behavior
- Alert Generation



## Limitations of Z-Value Testing in Yahoo!

- Post-Processing
  - Alpha/Beta Errors
- Outlier Removal
- Skew / Bimodal
- External
  - Seasonality
  - World Events
  - Weekend/Weekday
- Normalcy/Granularity

## pChart Control Charting

### Benefits

1. Quick Adoption of DQ Principles by Stakeholders
2. Simple to Read
3. Fairly easy to Implement
4. Few Beta/Alpha Errors
5. Knowledge Transfer

### Drawbacks

1. No Root Cause Analysis
2. No Accuracy Verification

37

## Data analysis: Dimension selection

