What is Big Data?

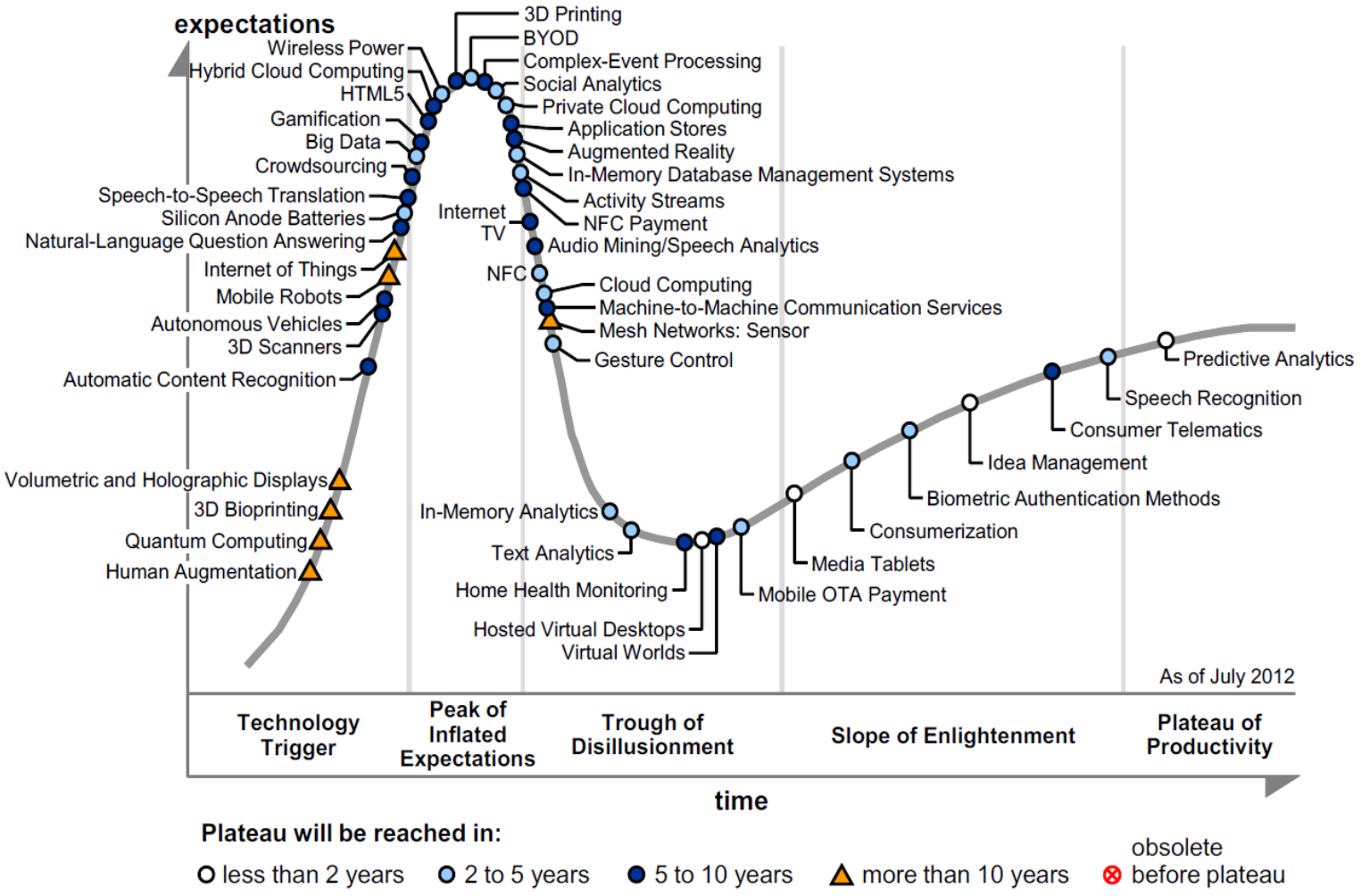
Bryan Smith

bryan.smith@microsoft.com

Trend


Google Trends, Mar 03 2013
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Hype



Source: Gartner (July 2012)

Definition

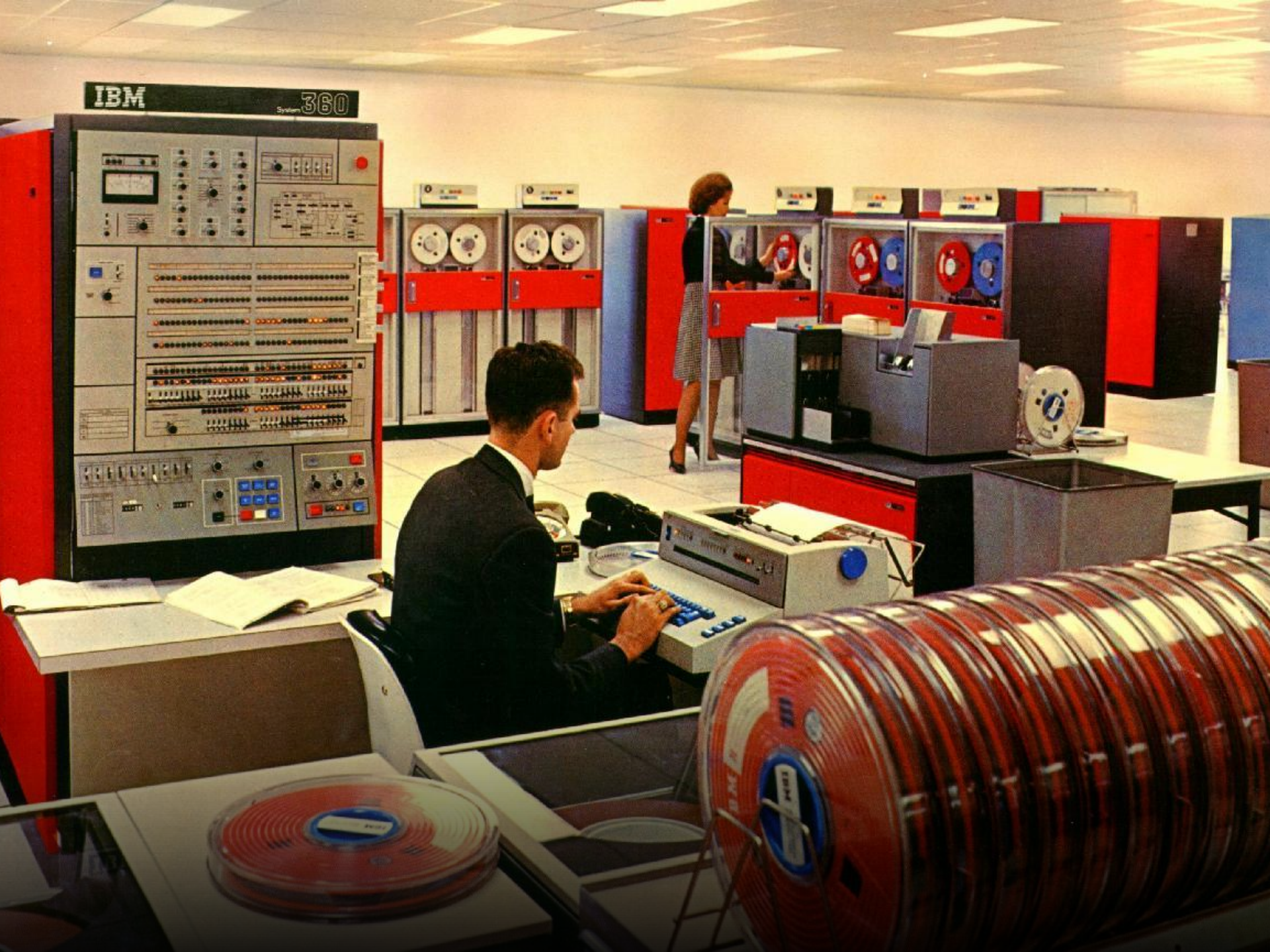
The phrase “big data” is now beyond completely meaningless. For those of us who have been in the industry long enough, the mere mention of the phrase is enough to induce a big data headache — please pass the big data Advil. (Editor’s note: We couldn’t agree more!)

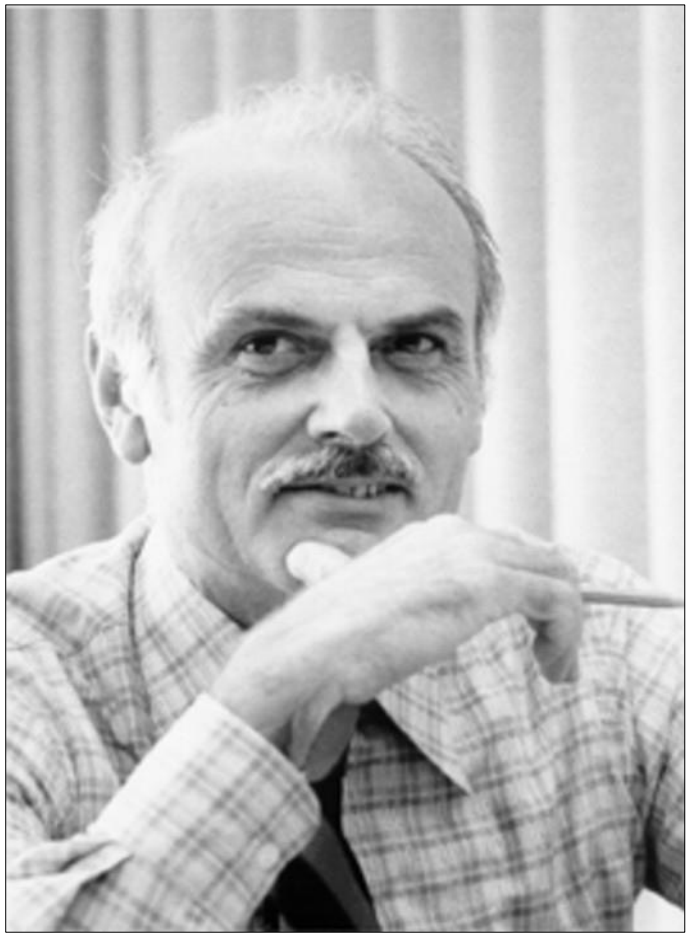
VB

‘Big Data’ is Dead! What’s next?
John De Goes, Precog CEO & Guest Columnist
VentureBeat, Feb 22 2013

Alternatives

- Smart Data
- Data Science
- Predictive Analytics
- New SQL


IBM
System 360



A Relational Model of Data for Large Shared Data Banks

E. F. CODD
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity
CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for noninferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

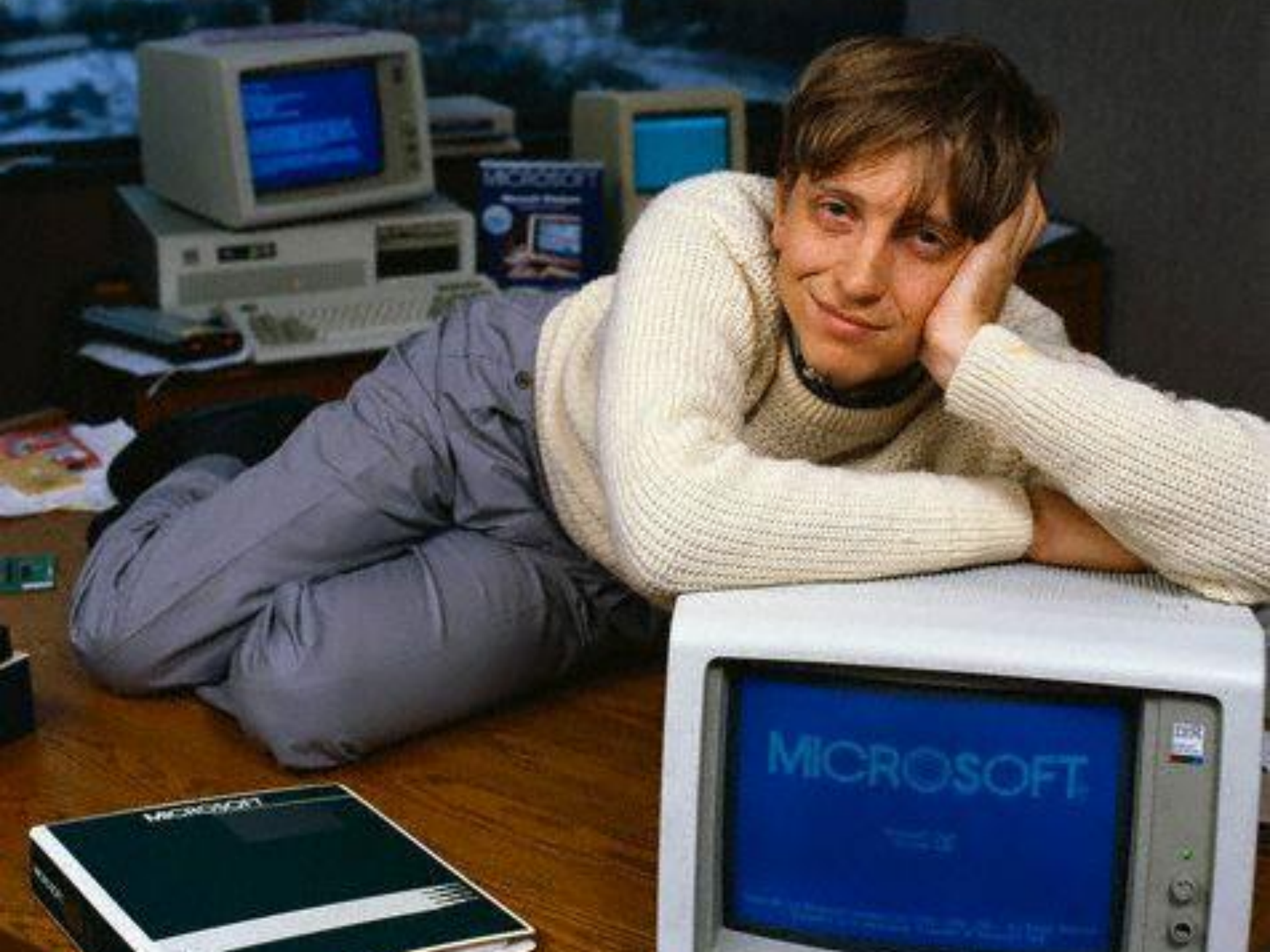
A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

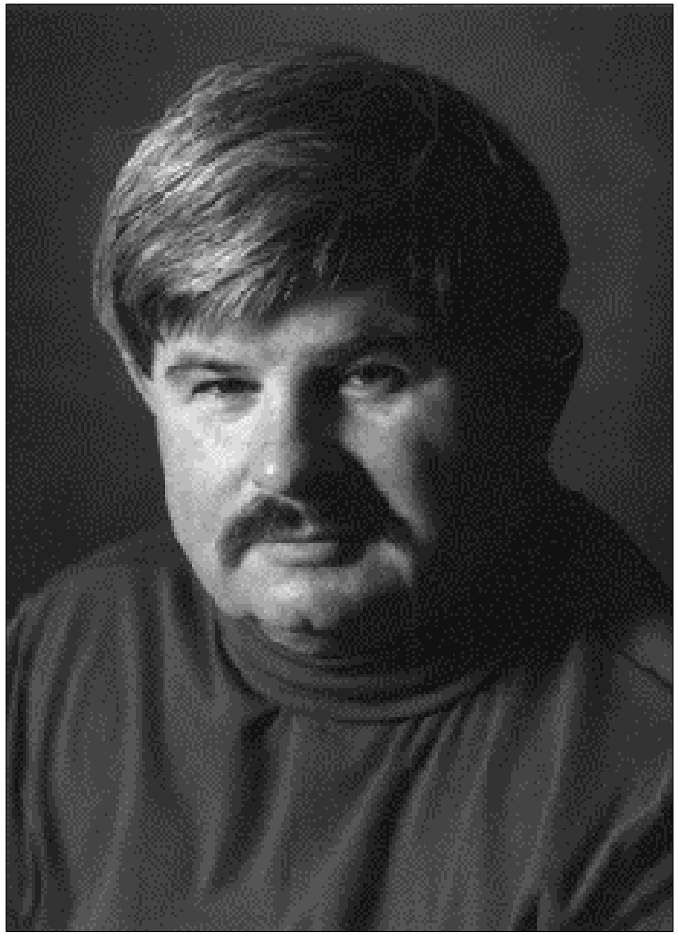
Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. Data Dependencies in Present Systems

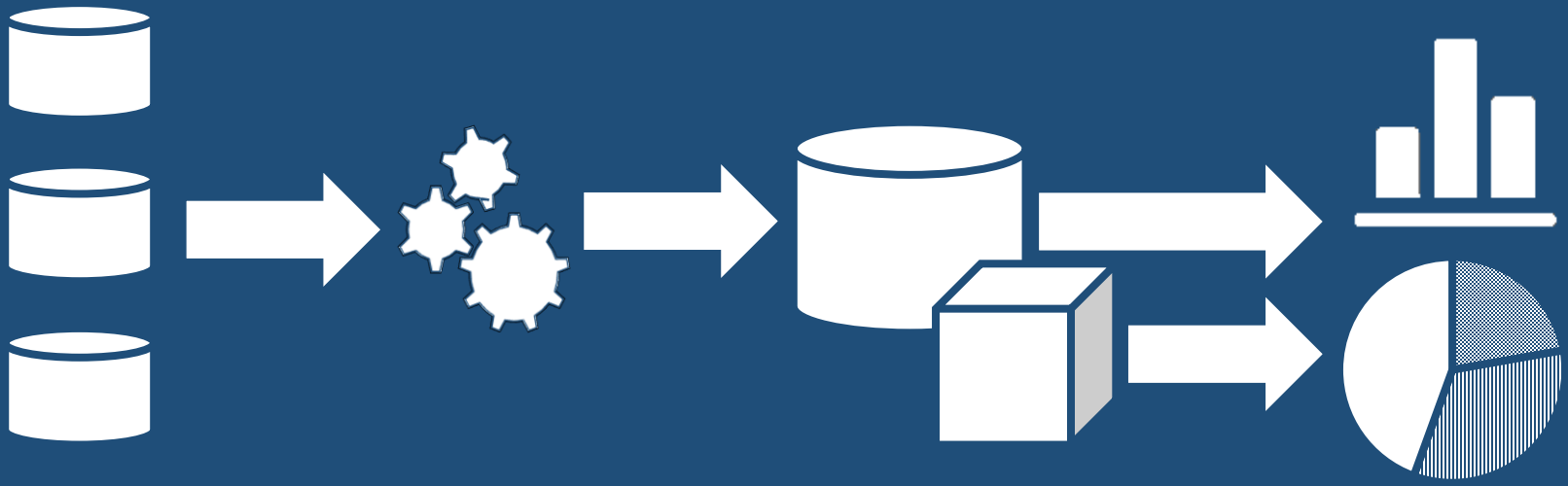
The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed *without logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. *Ordering Dependence.* Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

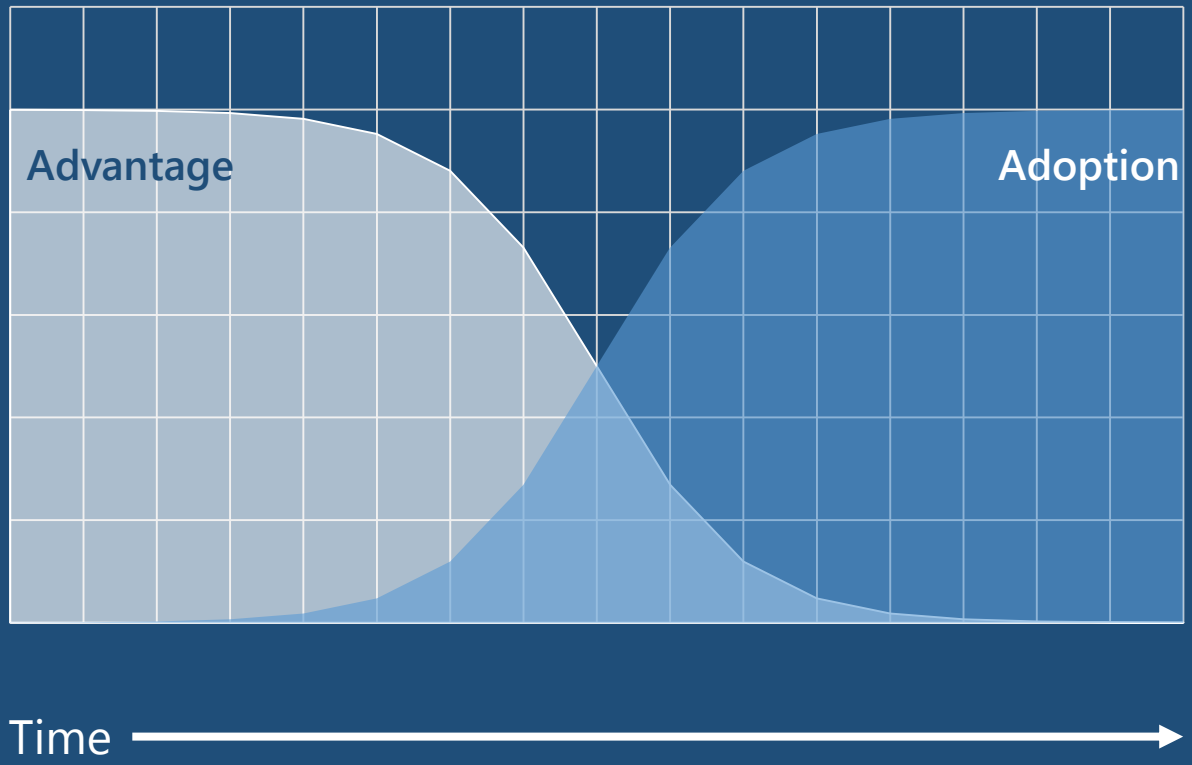

MICROSOFT
MICROSOFT



Business Intelligence



Advantage



Innovation

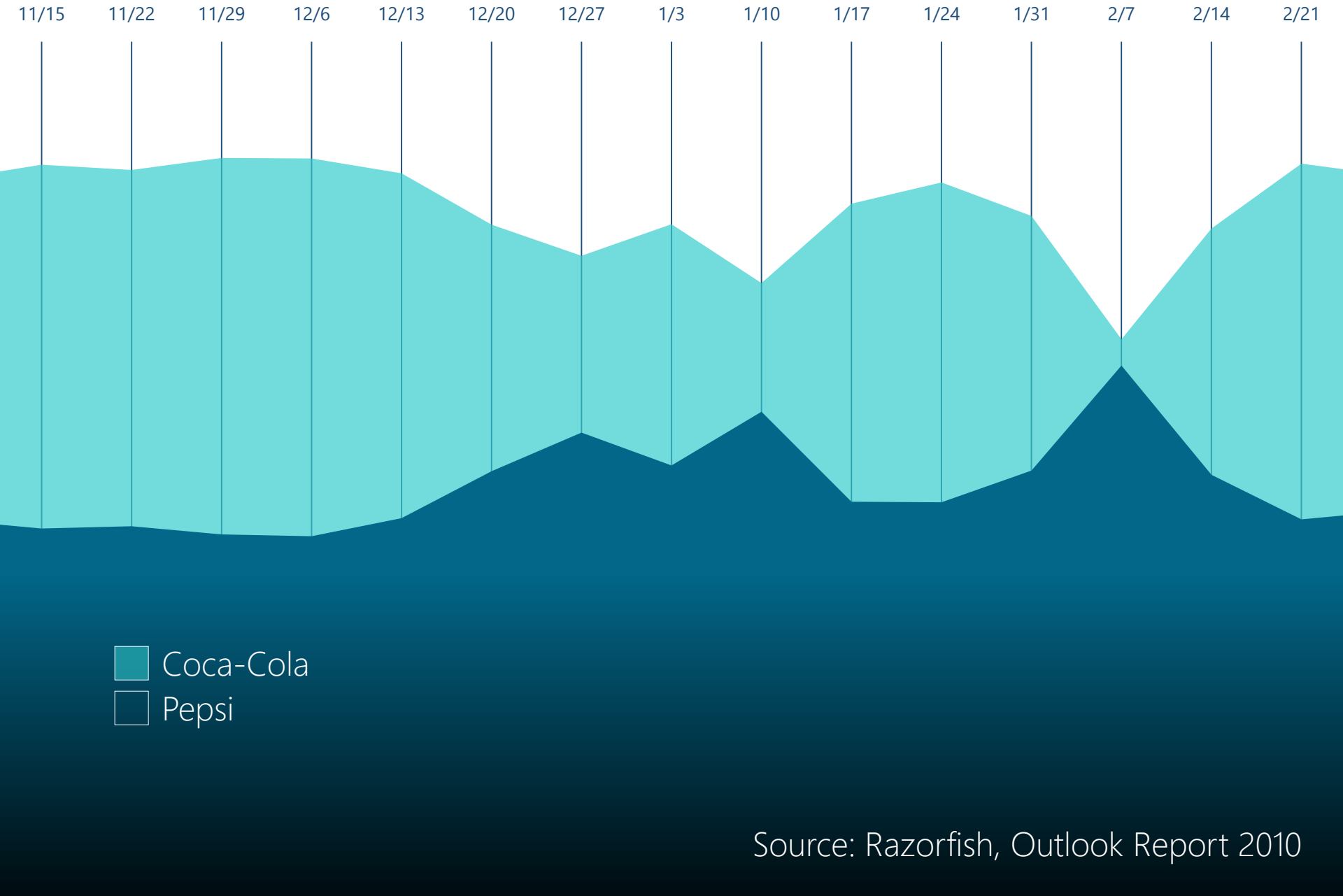
- More advanced techniques
- More sophisticated visualizations
- New sources of information

Connected Consumer

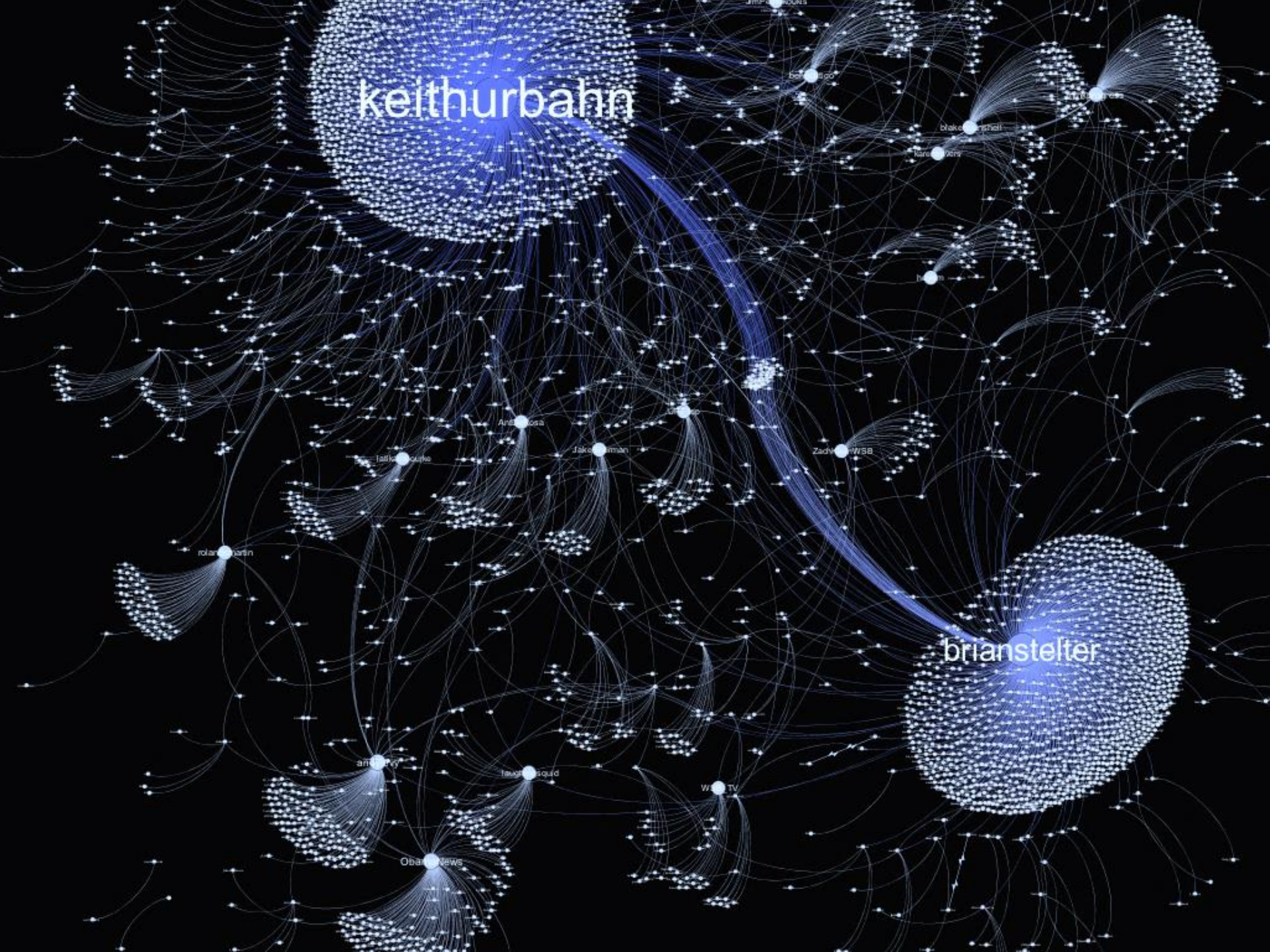
The Internet of Things

Unreported Side Effects of Drugs Are Found Using Internet Search Data, Study Finds

New York Times, Mar 06 2013



Source: Razorfish, Outlook Report 2010


keithurbahn
brianstelter


HALO 4

We can track where a car associated with a murder suspect is currently located and where it's been over the past several days, weeks or months.

Ray Kelly, NYPD Commissioner

Use Cases

Customer Insight & Engagement

Sentiment Analysis Churn Analysis Segmentation Targeting
Market Basket Analysis Recommendations Personalization

Marketing

Demand Forecasting Market Intelligence Product Usage Tracking
Competitor Sentiment Experimentation Data as a Product

Operations

Telemetry Analysis Monitoring & Alerting Automated Maintenance
Automated Problem Resolution Automated Resource Provisioning

Fraud Detection & Risk Management

Fraud Detection Crime Prevention Cyber-Attack Detection & Deflection
Insider Threat Detection Compliance Enforcement Surveillance

Big Data is New Data

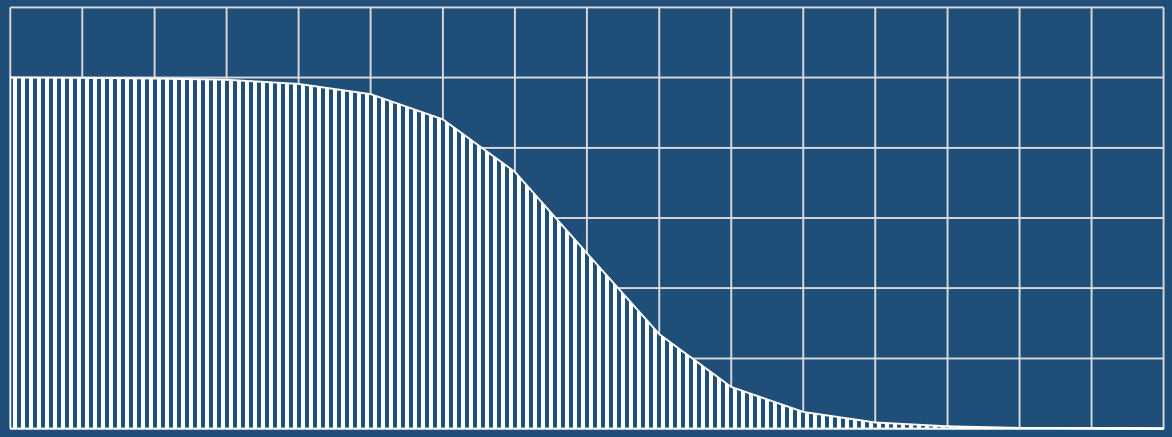
New Data Demands New Technology

Volume
Velocity
Variety

Volume

2002-08-05 00:51:37 192.168.1.104 - 192.168.1.103 80 GET /exchweb/img/form-prev.gif - 200 Mozilla/4.76+[en]+(X11;+U;+Linux+2.4.9-ac7+i686;+
2002-08-05 00:51:40 192.168.1.104 administrator 192.168.1.103 80 GET /exchange/Administrator/Inbox/ Cmd=contents&Page=1 200 Mozilla/4.76+
2002-08-05 03:29:27 192.168.1.100 - 192.168.1.103 80 GET /index.htm - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:29:27 192.168.1.100 - 192.168.1.103 80 GET /index_files/main4.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:29:27 192.168.1.100 - 192.168.1.103 80 GET /index_files/space.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:29:27 192.168.1.100 - 192.168.1.103 80 GET /index_files/105x60_acoda-direct_2.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Win
2002-08-05 03:29:27 192.168.1.100 - 192.168.1.103 80 GET /index_files/w513_105_pet100_blue.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Wi
2002-08-05 03:29:29 192.168.1.100 - 192.168.1.103 80 GET /index_files/main4.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:29:29 192.168.1.100 - 192.168.1.103 80 GET /index_files/105x60_acoda-direct_2.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Win
2002-08-05 03:30:56 192.168.1.100 - 192.168.1.103 80 GET /index_files/space.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:30:56 192.168.1.100 - 192.168.1.103 80 GET /index_files/w513_105_pet100_blue.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Wi
2002-08-05 03:30:56 192.168.1.100 - 192.168.1.103 80 GET /index_files/odd772.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 03:30:56 192.168.1.100 - 192.168.1.103 80 GET /index_files/lrec_ron_063002.js - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+N
2002-08-05 03:30:56 192.168.1.100 - 192.168.1.103 80 GET /index_files/yho-105X60-25fr-c12.gif - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windo
2002-08-05 04:15:31 192.168.1.100 - 192.168.1.103 80 GET /news/ - 302 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:15:31 192.168.1.100 - 192.168.1.103 80 GET /news/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:15:36 192.168.1.100 - 192.168.1.103 80 GET /news/Dynamic+disks,+part+1.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windo
2002-08-05 04:15:38 192.168.1.100 - 192.168.1.103 80 GET /news/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:15:40 192.168.1.100 - 192.168.1.103 80 GET /news/The+ABCs+of+IP+address+classes.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0
2002-08-05 04:15:40 192.168.1.100 - 192.168.1.103 80 GET /news/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:15:42 192.168.1.100 - 192.168.1.103 80 GET /news/NEWS%3A++Proxim+Enhances+Wireless+Network+Security+and+Management+
2002-08-05 04:15:42 192.168.1.100 - 192.168.1.103 80 GET /news/NEWS:++Proxim+Enhances+Wireless+Network+Security+and+Management+for+
2002-08-05 04:15:49 192.168.1.100 - 192.168.1.103 80 GET /news/RE%3A+Dude.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1
2002-08-05 04:15:49 192.168.1.100 - 192.168.1.103 80 GET /news/RE:+Dude.EML - 404 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:15:55 192.168.1.100 - 192.168.1.103 80 GET /news/Network+Computing+Newsletter+#181.EML - 200 Mozilla/4.0+(compatible;+MS
2002-08-05 04:15:57 192.168.1.100 - 192.168.1.103 80 GET /news/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:16:14 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:16:14 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/ - 500 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:16:18 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/Accelerated+VNC.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+V
2002-08-05 04:16:18 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/Accelerated+VNC.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+V
2002-08-05 04:16:27 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:16:31 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/Accelerated+VNC.EML - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+V
2002-08-05 04:16:45 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/ - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:17:31 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/Re%3A+VNC+Viewer+on+PocketPC+(Casio+E-125)%3F.EML - 200 M
2002-08-05 04:17:31 192.168.1.100 - 192.168.1.103 80 GET /news/VNC+Group/Re:+VNC+Viewer+on+PocketPC+(Casio+E-125) .EML 404 Mozilla/4.0
2002-08-05 04:19:30 192.168.1.100 - 192.168.1.103 80 GET /webstuffs/index.htm - 404 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:19:32 192.168.1.100 - 192.168.1.103 80 GET /webstuffs/index.htm - 404 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
2002-08-05 04:19:32 192.168.1.100 - 192.168.1.103 80 GET /webstuffs/index.htm - 404 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)

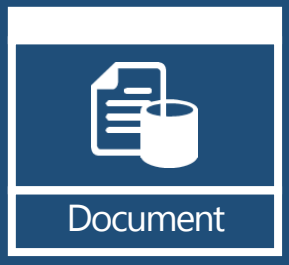
Velocity

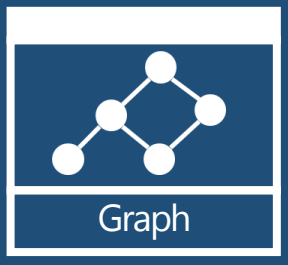


Variety

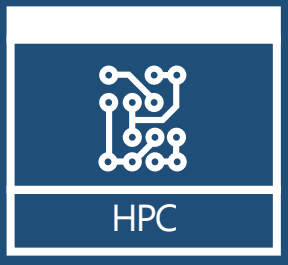

```
{ "created_at":"Thu, 14 Feb 2013 19:28:02 +0000",
  "entities":{
    "hashtags":[ {"text":"WindowsPhone","indices":[30,43]},
                 {"text":"SQLServer","indices":[52,62]}
                 ],
    "urls":[{ "url":"http:\/\/t.co\/bqNyAMkc",
              "expanded_url":"http:\/\/spr.ly\/6014nzJu",
              "display_url":"spr.ly\/6014nzJu",
              "indices":[79,99]
              }],
    "user_mentions":[ {"screen_name":"SQLServer","name":"Microsoft SQL Server", … }]
    },
  …"from_user_id":22303580, … "iso_language_code":"en", …
  "source":"&lt;a href=&quot;http:\/\/twitter.com\/&quot;&gt;web&lt;\/a&gt;",

  "text":
        "RT @SQLServer: Just released! #WindowsPhone App for
        #SQLServer Support Content http:\/\/t.co\/bqNyAMkc",

  "to_user":null,"to_user_id":0,"to_user_id_str":"0","to_user_name":null }
```

Platforms





NoSQL



Other

Summary

- Big Data is a shift from operationally-derived BI & a wider embrace of data analytics
- Non-operational data presents new technological challenges that drive adoption of new platforms

From Data to Insights

Big data. Small data. All data.

Microsoft

Register at https://vts.inxpo.com/Launch/QReg.htm?ShowKey=13140

Date: March 26, 2013
Time: 9:30am - 10:30am PST
Duration: 60 minutes

Big data brings big challenges, but also opportunities--for people and companies to gain insight, spot trends and make decisions in ways never before possible before. In the battle of the buzzwords, "big data" is poised to render "guestimation" obsolete.

To better understand the impact of big data on the future of global business, Microsoft is hosting an exclusive webcast briefing, "From data to insights", produced in association with the Economist.

In the webcast, you'll hear from **Tom Standage, digital editor of the Economist**, on the social and economic benefits of mining data, followed by a moderated discussion featuring two Microsoft data experts, **VP/Technical Fellow Dave Campbell** and **Technical Fellow Raghu Ramakrishnan**, for an insider's view of the trends and technologies driving the business of big data, as well as Microsoft's big data strategy.