Big Plateaus of Big Data on the Big Island

Todd Walter

Chief Technologist

Teradata Corporation

ca.linkedin.com/in/ToddAWalter

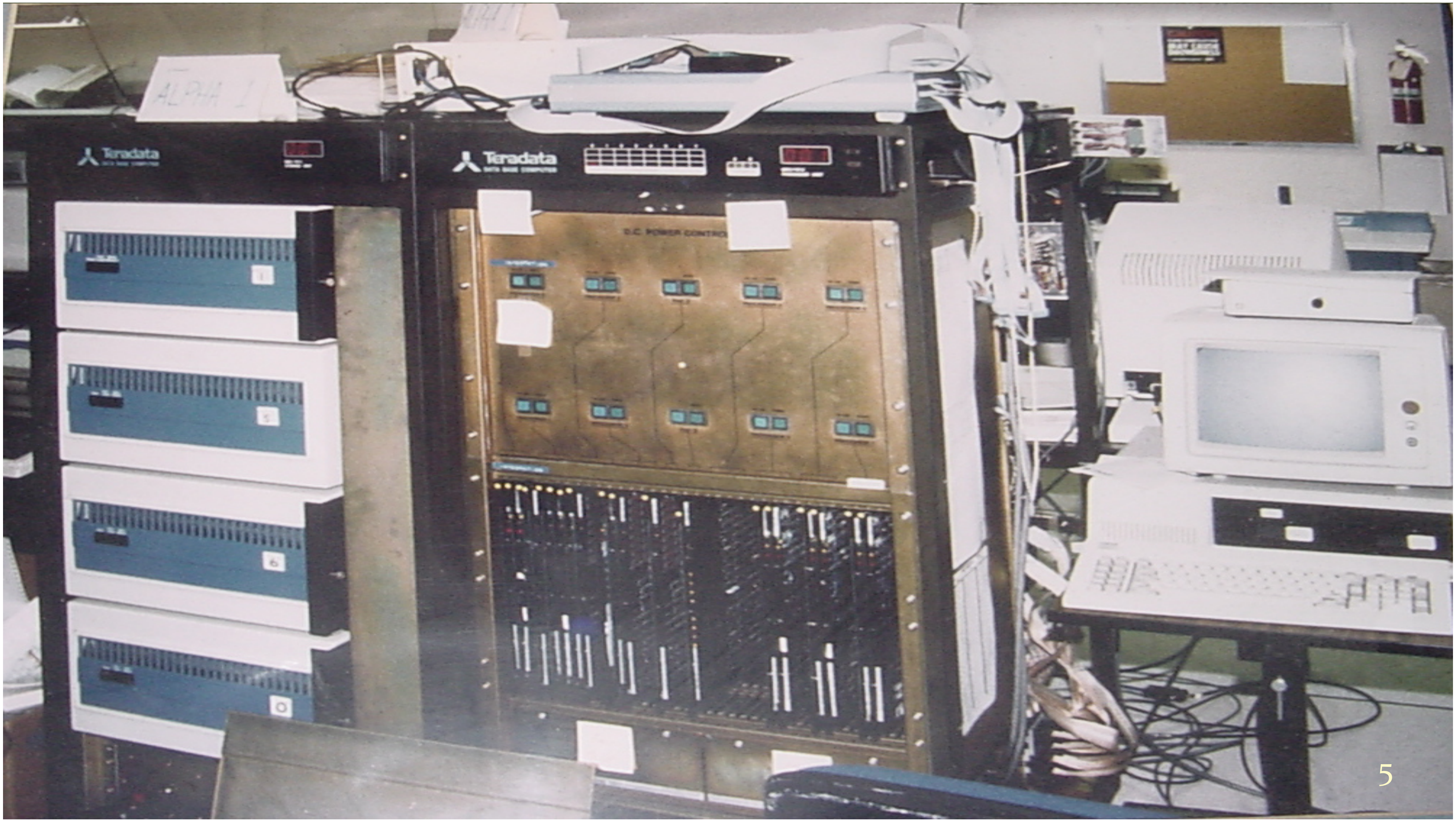
9/2/2015

Disclaimer

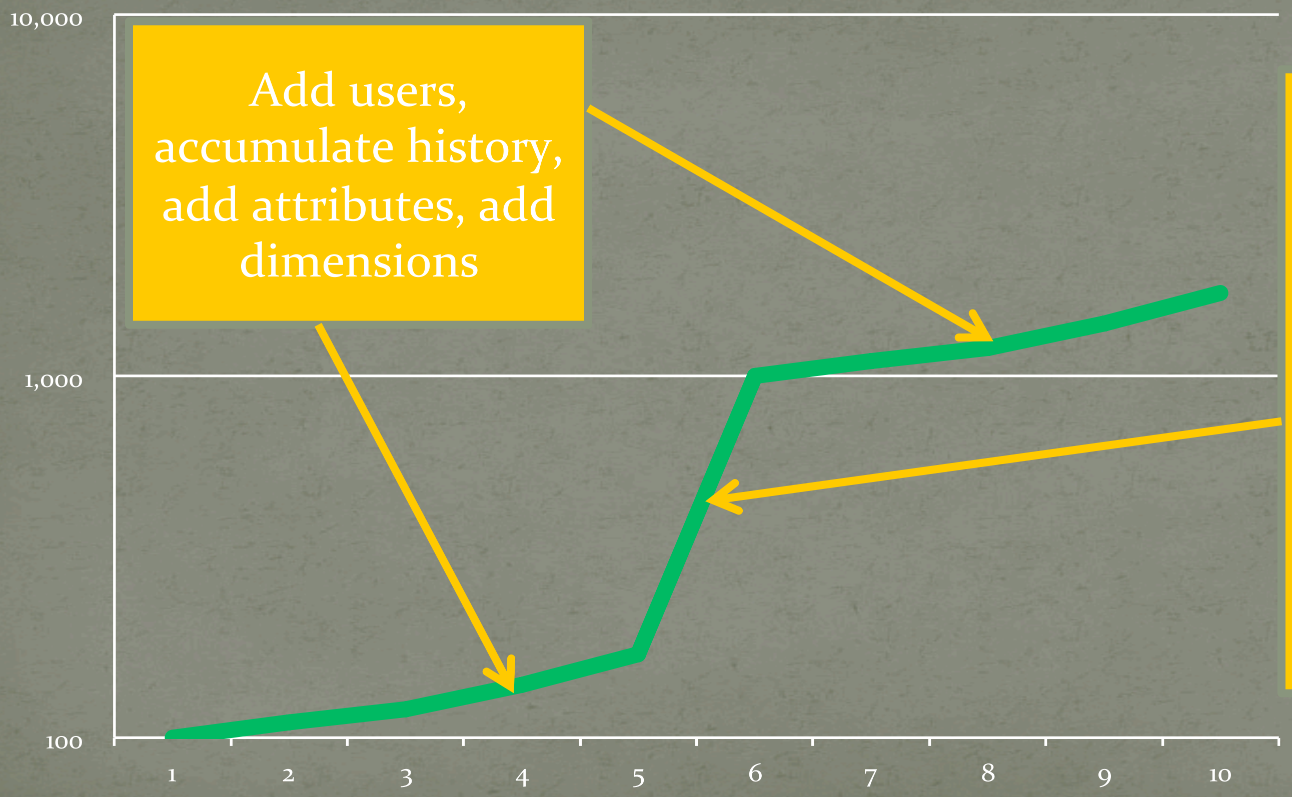
- The following solely represents the opinions of Todd Walter not the opinions of Teradata Corporation
- Nothing in this document may be construed to be a promise of future functionality, capability or product from Teradata Corporation at any point in the future.

What is Big?

- 1979 - Teradata was founded to solve enterprise Terabyte problems utilizing commodity microprocessors and commodity disks
  - Microprocessor – 8086
  - Disk drive - 200MB weighs 30lbs and requires a washing machine sized cabinet
- “Citibank is installing a Teradata Corp database system … With modular disks added, the DBC/1012 system can handle up to a trillion bytes of information, more than any organization is believed capable of using” --- American Banker, Sept 1984
- “Citibank is installing a fourth computer from Teradata … A 168 processor configuration … will achieve more than 170M instructions per second and include 70GB of disk storage for very large relational databases. … will contain 138 disk drives of 515MB each” --- American Banker, March 1986


ALPHA 1
Teradata
DATA BASE COMPUTER
Teradata
DATA BASE COMPUTER
D.C. POWER CONTROL

Customer Data Plateaus

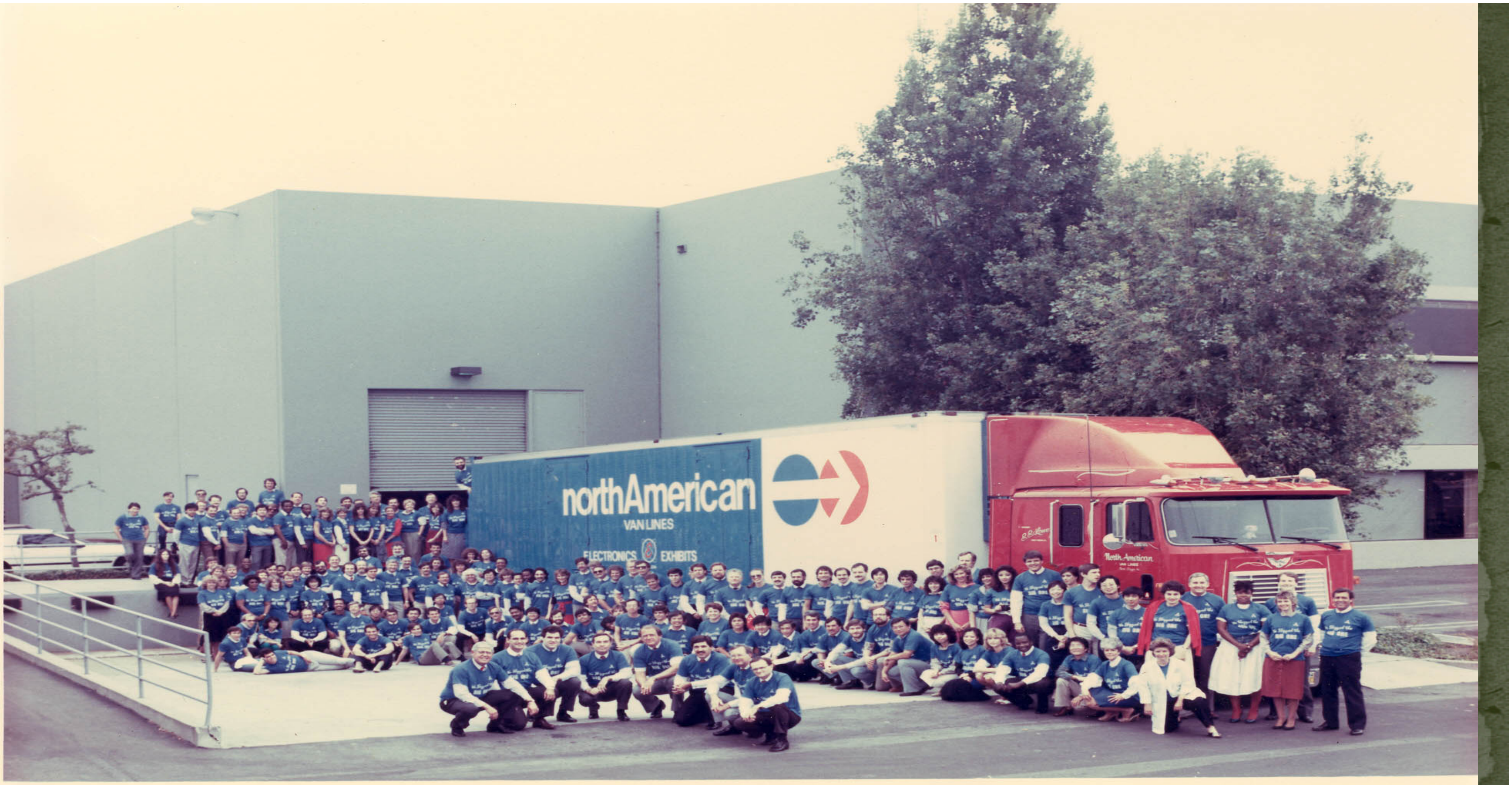


Add users, accumulate history, add attributes, add dimensions

Entirely new data set enabled by new technology at new price point – value has to exceed cost (ROI). Earliest adopters have vision ahead of ROI

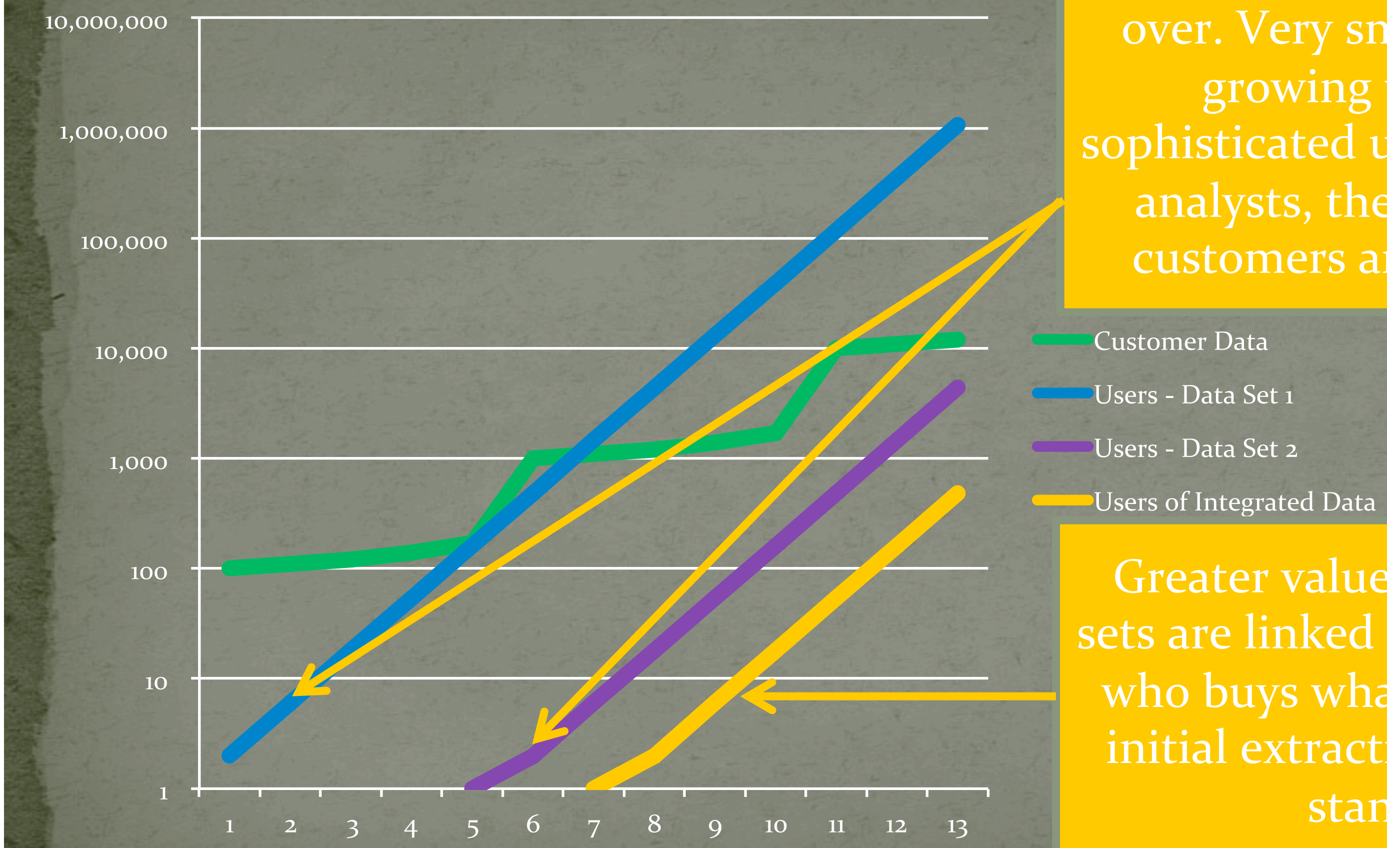
Market Basket

- Retailers were tracking summary sales data
  - <store, item, day>
  - Aggregate locally in the store, send once per day
- Visionaries believed that value could be derived from market basket detail
  - Send every line item
  - 250,000 records per store per day -> 10M records per store per day
  - Batch first, then send continuously
- New analytics needed
  - Affinity analysis rather than simple counts/sums



We Shipped the BIG ONE - June 2, 1986

User Adoption of New Data



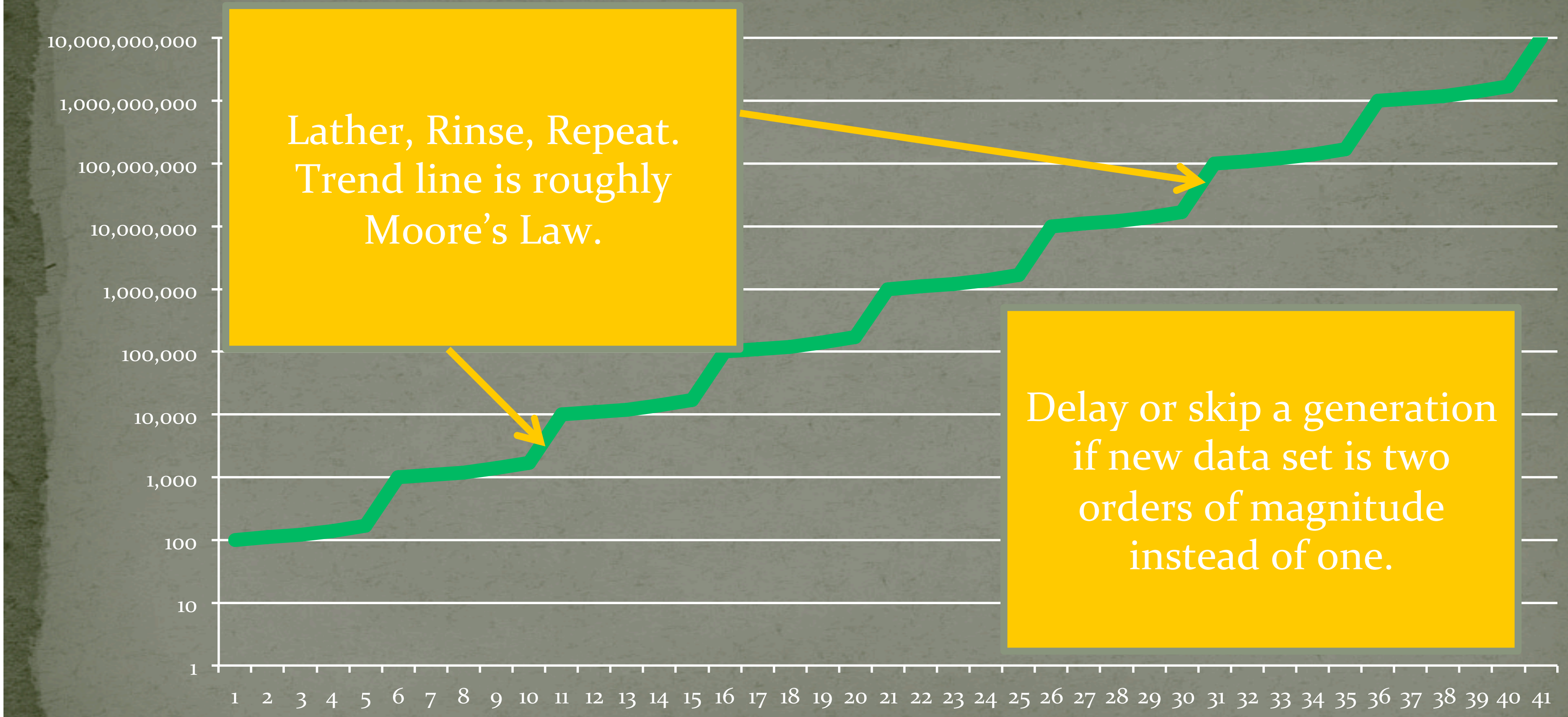
User adoption of new data sets starts over. Very small number of experts growing to wider audience, sophisticated users moving to business analysts, then business users, B2B customers and even to consumers

Greater value is derived when data sets are linked – see bigger picture (eg who buys what, when). Comes after initial extraction of easy value from standalone data

Market Basket

- A couple very sophisticated users doing very complex hand coded SQL to derive basket attributes/types, product affinity
- Add HR - product sales trends by time for stocking and lane staffing, then employee efficiency (scans per minute)
- Add marketing – basket contents linked to shopper and shopper behavior (coupons, buy only on sale, items bought together by segment/person, location of item in store)
- Add customer service – returns matched against receipt, update receipt to record return
- Add B2B – vendors can track movement of items through store/supply chain
- Add consumer conversation – warrantee/recall, consumption rate, when did I last buy
- Now market basket data is an integrated, shared resource that informs many applications and serves millions of users

Data Size Plateaus over Time



Lather, Rinse, Repeat. Trend line is roughly Moore's Law.

Delay or skip a generation if new data set is two orders of magnitude instead of one.

Retail Plateaus

- <Store, Item, Week>
- <Store, Item, Day>
  - Simple aggregations

2B records total for 9 quarters

- Market Basket
  - Affinity
  - Link to person, demographics, HR

2B records per day, keep 9 quarters

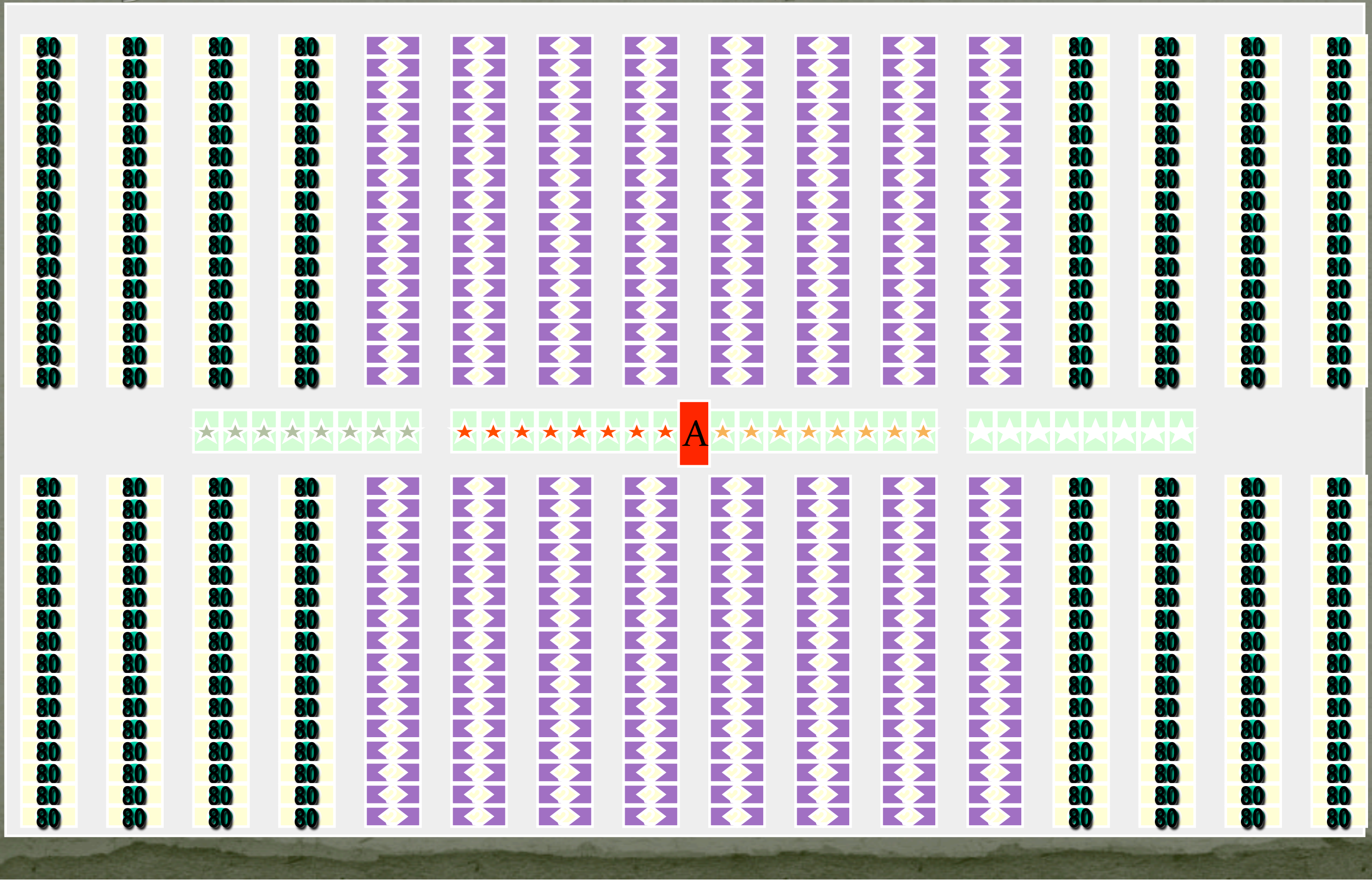
- Inventory by SKU by store
  - Temporal, time series, forecasting
  - Link to product, marketing, market basket

Retail Plateaus

30B records per day

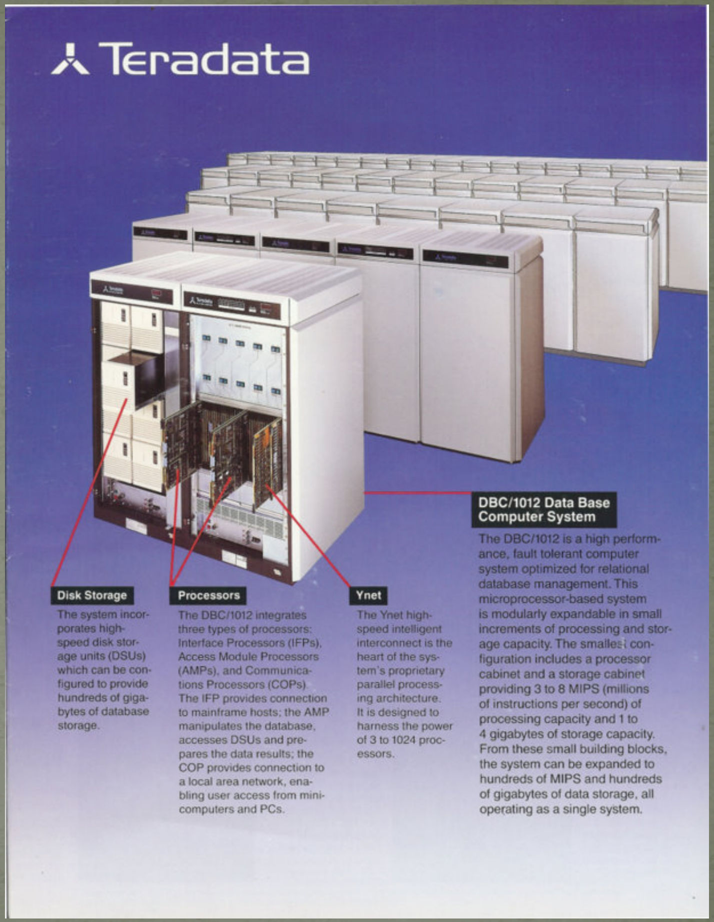
- Web Logs and traffic
  - Behavioral patterns – eg path linked to person, offers, other channels
  - Operations of the web site
- Supply chain sensors – sampled at major event
  - Activity Based Costing
  - Link to customer, product, HR, planning
- Social Media
  - Text analysis, Filtering, languages
  - Link to customer, sales, other channel interactions
- Supply chain sensors – sampled at minutes or seconds
  - Telematics
  - Real time, Event detection, trending, static and dynamic rules
  - Link to HR, thresholds, forecasts, routing, planning

2000 - Theoretical 1 Petabyte System
512 Nodes, 20K Disks (72GB)



Production Systems

- 1986: 500GB user data space, 1TB spinning disk
- 1992: 1TB user data space
- 1996: 11TB
- 1997: 24TB
- 1999: 130TB
- 2005: 570TB
- 2009: 1.2PB
- 2013: 46PB



Teradata

Disk Storage

The system incorporates high-speed disk storage units (DSUs) which can be configured to provide hundreds of gigabytes of database storage.

Processors

The DBC/1012 integrates three types of processors: Interface Processors (IFPs), Access Module Processors (AMPs), and Communications Processors (COPs). The IFP provides connection to mainframe hosts; the AMP manipulates the database, accesses DSUs and prepares the data results; the COP provides connection to a local area network, enabling user access from minicomputers and PCs.

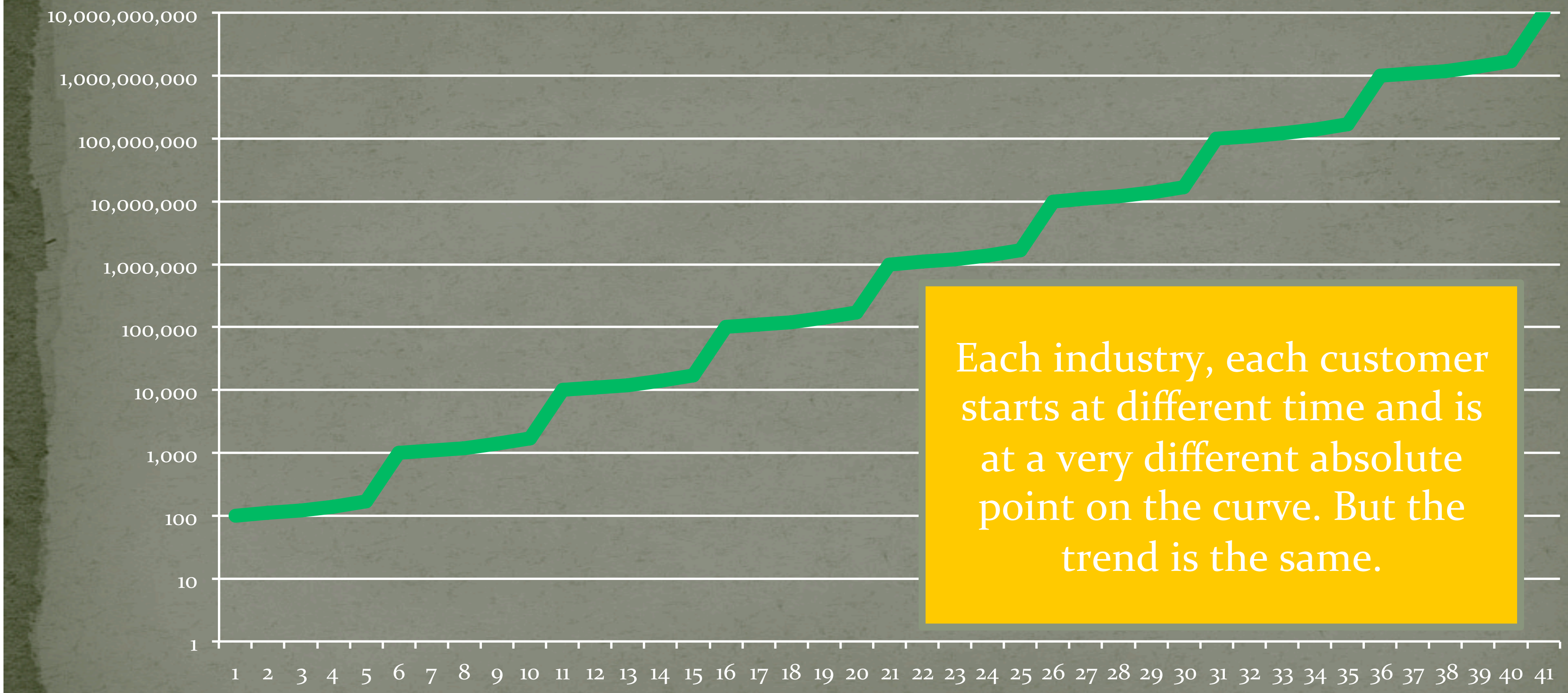
Ynet

The Ynet high-speed intelligent interconnect is the heart of the system's proprietary parallel processing architecture. It is designed to harness the power of 3 to 1024 processors.

DBC/1012 Data Base Computer System

The DBC/1012 is a high performance, fault tolerant computer system optimized for relational database management. This microprocessor-based system is modularly expandable in small increments of processing and storage capacity. The smallest configuration includes a processor cabinet and a storage cabinet providing 3 to 8 MIPS (millions of instructions per second) of processing capacity and 1 to 4 gigabytes of storage capacity. From these small building blocks, the system can be expanded to hundreds of MIPS and hundreds of gigabytes of data storage, all operating as a single system.

Data Size Plateaus over Time



Each industry, each customer starts at different time and is at a very different absolute point on the curve. But the trend is the same.

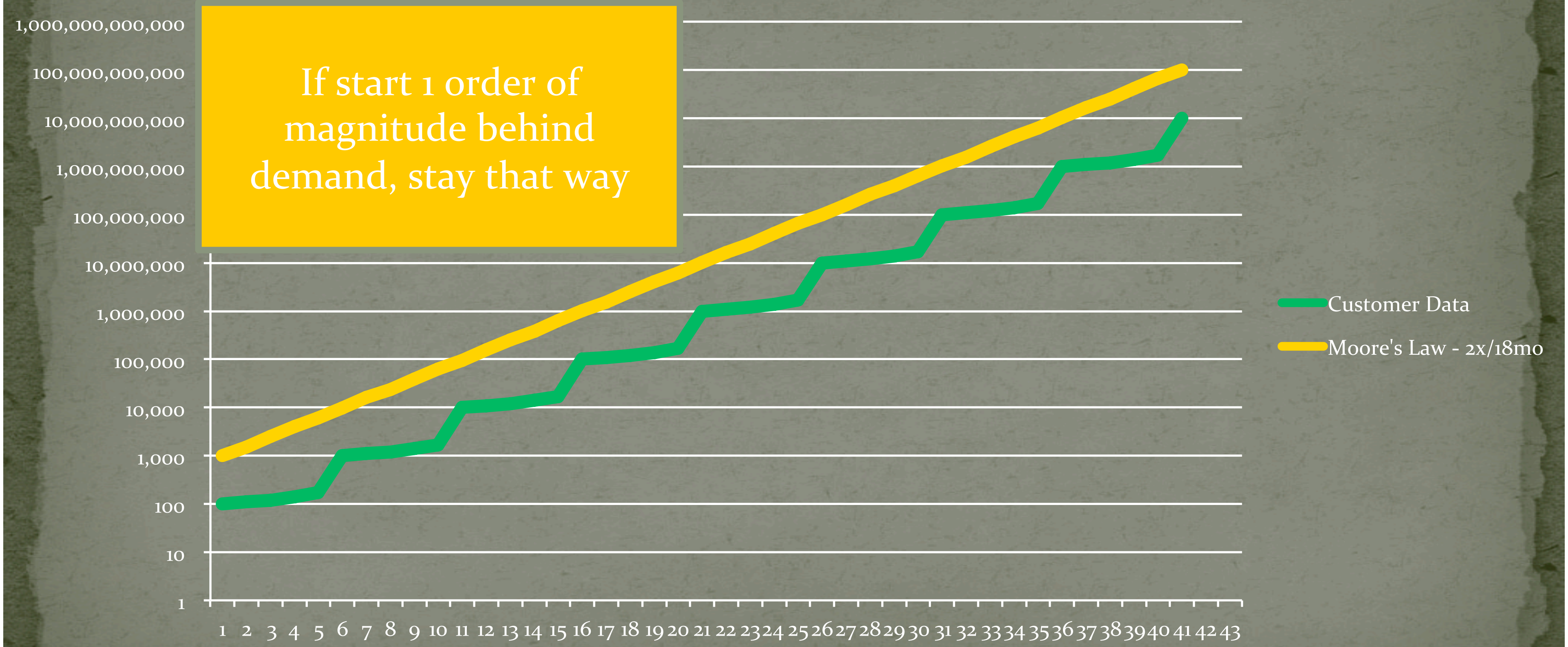
Package Tracking

- Every scan captured
- Begin with optimization questions
  - Reduce scans per package, grouping for routes, packing planes
- Add Pricing
  - How many packages of what attributes from where to where
  - Did customer meet commitments for discounts or violate terms?
  - Links to customer, individual package, location, deliver activity based costing
- Delivery fidelity
  - If we hold this plane/truck for late arriving truck, how many more packages arrive on time – cost benefit
  - Links to customer, finance, cost, real time vehicle information
- Eventually to customer
  - Where is my package?
  - 40M requests per day
- Package scan data now a critical shared resource of the company

Utilities

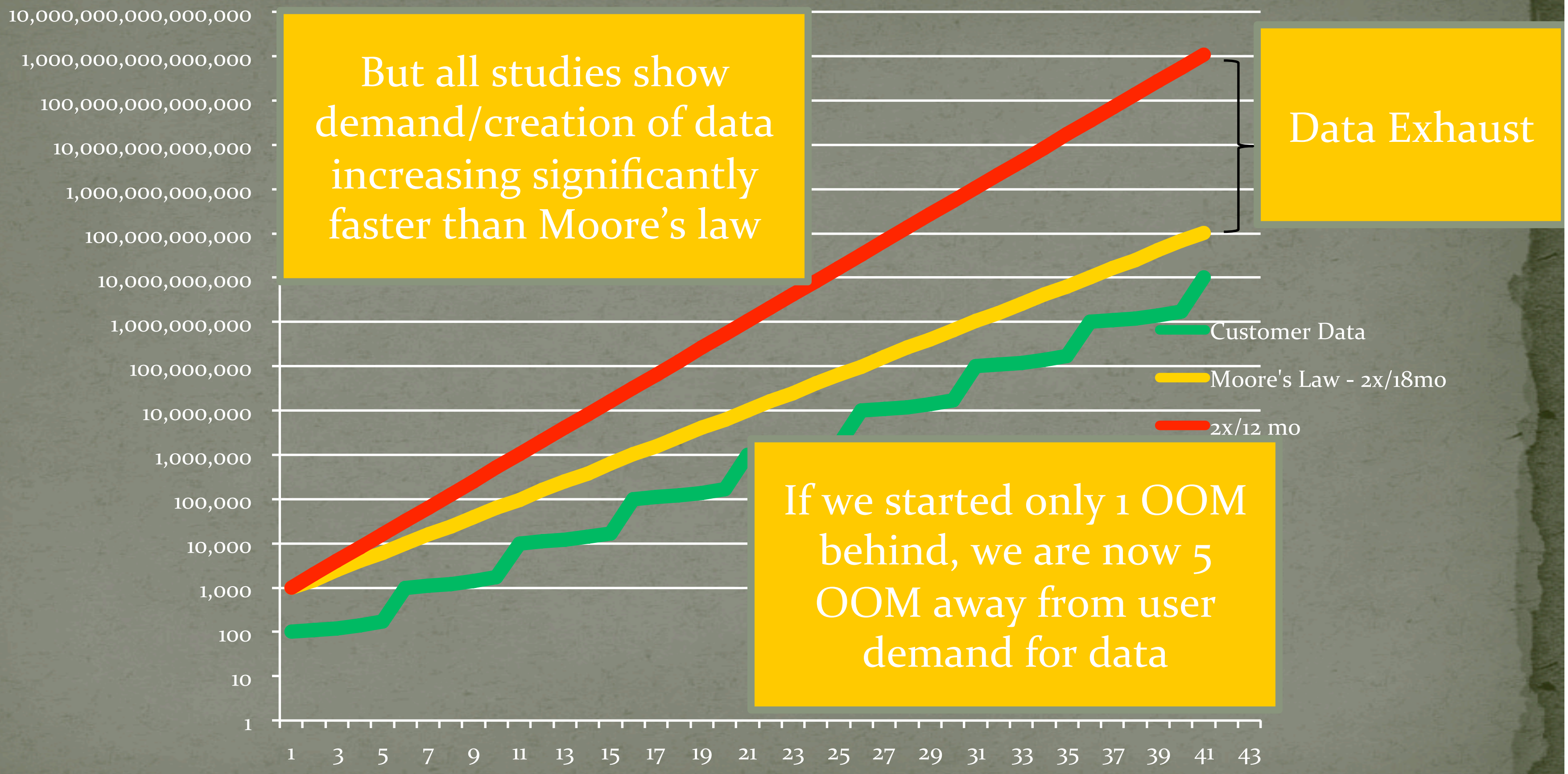
- Spent their whole life with one record per customer per month
- Smart meters arrive sending one record per customer per 15 minutes
  - 96 records per day, 2880 per month, 34560 per year
- 20M customers
  - More than 690B records per year
- Next is smart appliances – x 10 again
- Plus the smart grid with every transformer and connection point reporting
- Billing by time, network and supply management, forecasting by time of day, special offers to cause customers to move/smooth usage, identify waste, fraud and crime
- Organizational chaos for folks that have been dealing with 20M records per month using 30 year old programs

Customer Data Size vs. Moore's Law



If start 1 order of magnitude behind demand, stay that way

Customer Data Size vs. 2x per 12mo

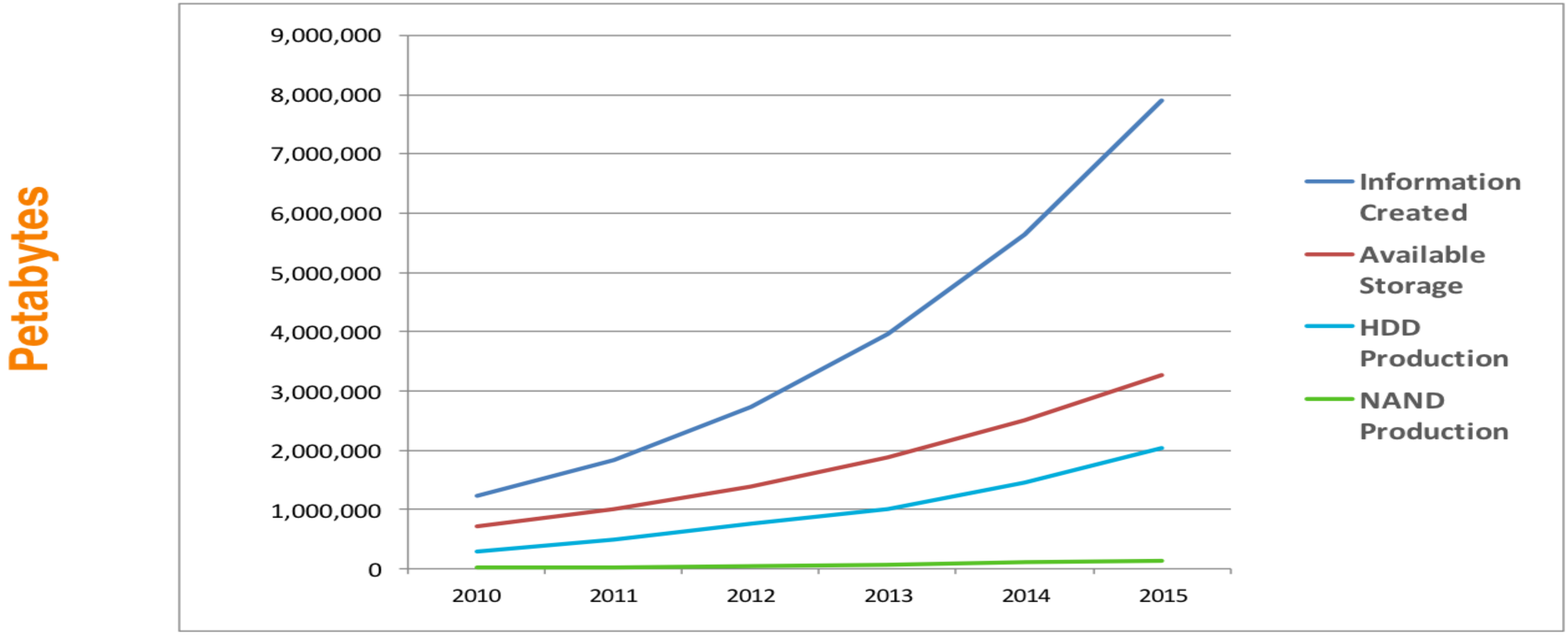


But all studies show demand/creation of data increasing significantly faster than Moore's law

Data Exhaust

If we started only 1 OOM behind, we are now 5 OOM away from user demand for data

Data Generation Outstrips Storage Capacity



Source: IDC IVIEW, "Extracting Value from Chaos," June 2011, and
Gartner, Market Trends: "Enlarging the Library of Forms in Which Storage is a Necessary Element," June 30, 2011

Today

- Next OOM or two are challenging organizations and technologies
  - 200B incoming network records per day, 500B/day in 2 years
- But the pattern is the same
  - Start with basic ingest of the data set and a couple scientists hacking at it
  - Develop new analytic algorithms to deal with the new questions that the new data enables
  - Grow usage across the organization – users and applications
  - Integrate with other organizational data

eBay

1 Trillion records, 7PB in one table – but linked to thousands of tables to derive the bigger picture

- Clickstream first
  - Identify user patterns, site behavior, A/B testing of site changes
- Add every event inside the site
  - 20-200 events per click
- Usage
  - Start with a few ops folks who want to understand how the site behaves, capacity planning, server utilization
  - Add site developers – everyone who develops for the site now can see continuously how their page is used and behaves
  - Add marketing to see effect of different presentation, ad placement, search results optimization
- Thousands of users across the company running millions of queries per day
  - Shared resource across the whole organization

Crystal Ball

- Not hard to predict that the pattern will hold
- Today’s “Big Data” will become tomorrow’s organization wide resource
- Today’s Data Science projects will be tomorrow’s mainline applications
- Today’s developer class tools will have to give way to self service applications for vastly larger numbers of increasingly less specialized analysts and users
- Today’s coding of each question has to give way to pretty pictures on executive desks and big “easy buttons”
- Just like every plateau in the past

What is Different?

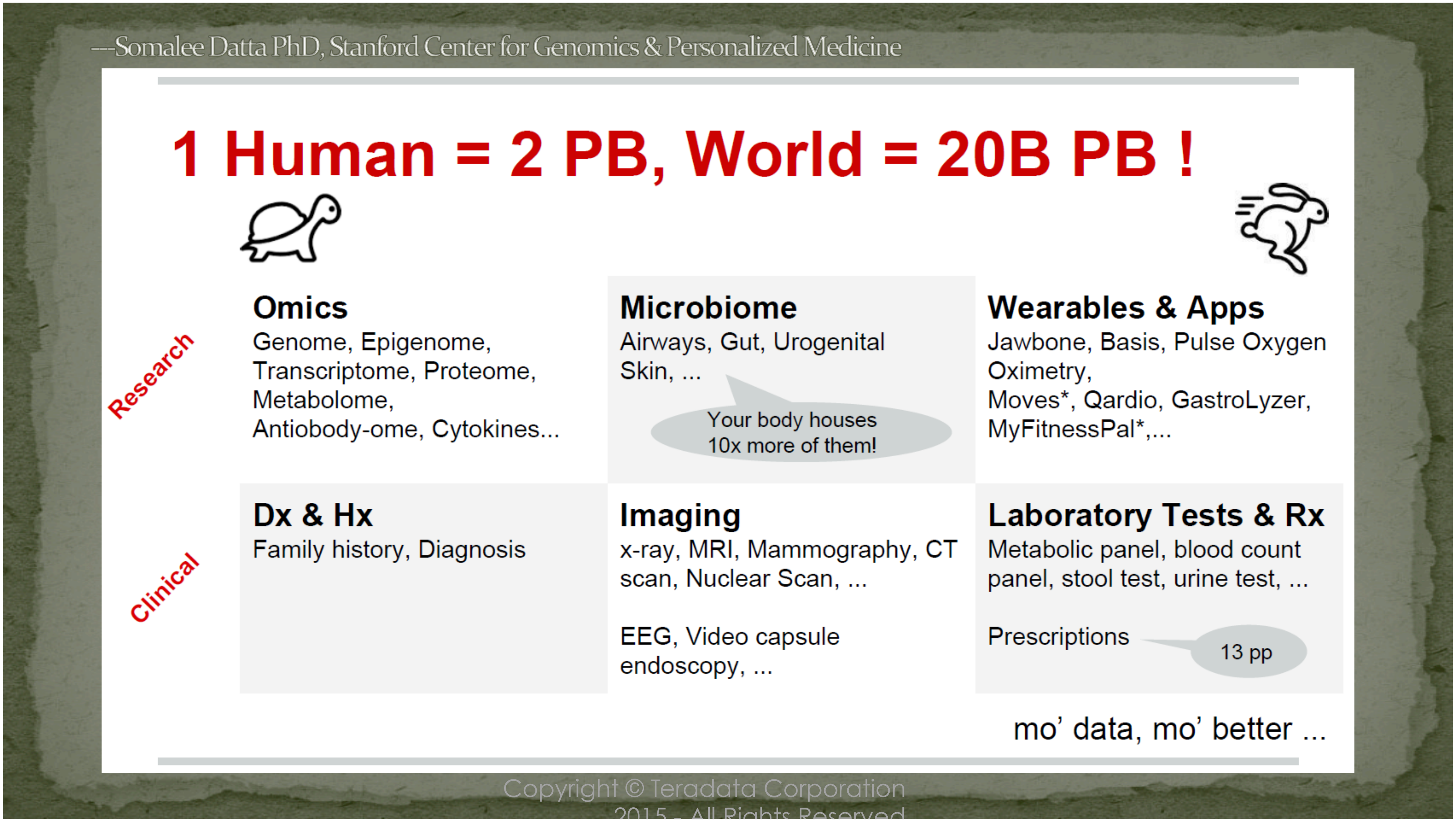
- Character of data is changing
  - Transactions -> Interactions -> Machinations
- Structure of data non-uniform and drastically more dynamic
  - Web Logs, machine data have widely varying forms from one record to the next
    - Different pages report different attributes
    - Different (versions of) sensors report different data
    - Many sensors in one machine each with own format
    - Each unique attribute is sparsely represented
  - No way to create or keep up with a traditional data model

Vehicle Sensors

- No standards, every manufacturer defines their own data reporting for each sensor
- Bitwise representation – back to the stone age of 4 bits mean temperature variation from 75F (or was that 30C??)
- Hundreds of sensors per vehicle each reporting what it cares about
- Hundreds of humans tracking metadata by hand to make sense of the stream – aargh!
- Today a moderate number of vehicles fully instrumented and reporting at high sampling rates
  - Many more reporting aggregates at periodic intervals
  - First 1M vehicles at detailed reporting == 4PB/mo
- Today a few scientists who can interpret the data looking for the big patterns
- Tomorrow
  - You get an alert on your dash screen that your water pump is about to fail, an appointment has been made for you at the dealer three exits ahead and it has been verified that the parts are in stock
  - Safety issues identified early, saving lives and $$
  - Fully integrated with all enterprise data and conversing with the consumer

What is Different?

- Form of data is expanding
  - Transactions were well known data types understood from earliest days of computing
  - Now –finally- becoming practical to deal with content and analysis on text, audio, images, medical scans, video, … at scale
  - Specialized CPU intensive algorithms
- “New” analytic patterns
  - Path, graph,
  - Machine learning
  - Specialized per domain – genomics, well drilling,…



---Somalee Datta PhD, Stanford Center for Genomics & Personalized Medicine

1 Human = 2 PB, World = 20B PB !

Research	**Omics** Genome, Epigenome, Transcriptome, Proteome, Metabolome, Antiobody-ome, Cytokines...	**Microbiome** Airways, Gut, Urogenital Skin, ... Your body houses 10x more of them!	**Wearables & Apps** Jawbone, Basis, Pulse Oxygen Oximetry, Moves*, Qardio, GastroLyzer, MyFitnessPal*,...
Clinical	**Dx & Hx** Family history, Diagnosis	**Imaging** x-ray, MRI, Mammography, CT scan, Nuclear Scan, ... EEG, Video capsule endoscopy, ...	**Laboratory Tests & Rx** Metabolic panel, blood count panel, stool test, urine test, ... Prescriptions 13 pp

mo' data, mo' better ...

What Do Users Need?

- Representation
  - New usable representations required to manage the diversity
  - NVP, JSON, XML, … are a start but not very user friendly
  - New end user language/tools needed to provide access
  - Today’s new tools are a step backwards, developers building for developers
- Curation
  - Schema on read is powerful but…
  - Kicks the can down the road to be the users’ responsibility
  - At the same time that the data is messier than ever before
  - Need new language and tools that make it easy to iteratively curate data
  - And automation of understanding new data

What Do Users Need?

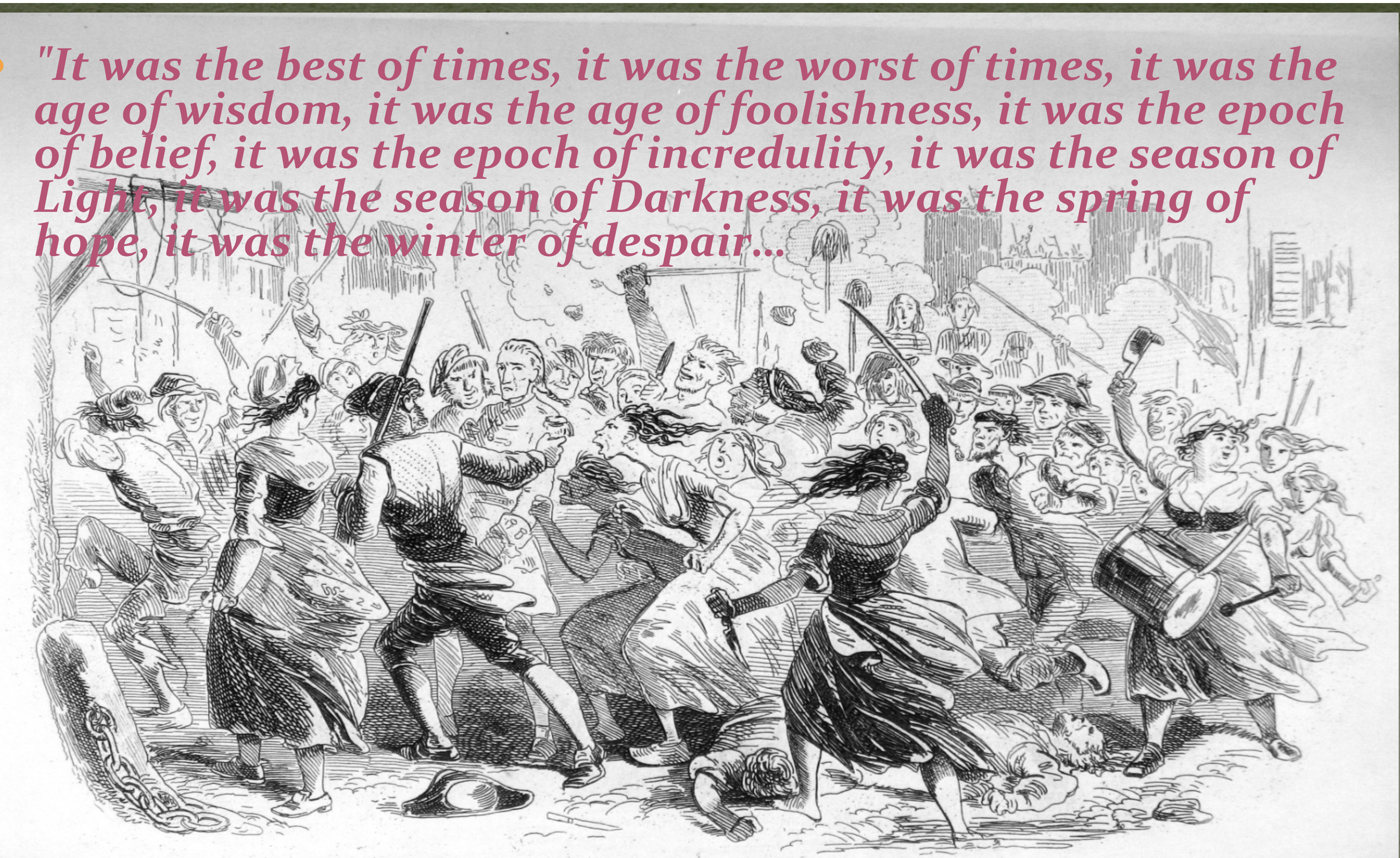
- Integration
  - Integration with all of the data in the enterprise is needed to realize the value from the new data
    - Todd's opinion – 10/90. Only 10% of the value realized in isolation
  - Customer, product, marketing, HR, supply chain, finance, pricing, transactions, … - traditional data of the organization still carries much of the value
- Access
  - Most end users today are using Excel, Access, R, …
  - Many analytic algorithms being developed today (see R library) but few are scalable
  - Non-scalable tools forcing users to use tiny samples, extreme subsets, single instances
  - Must be able to scale algorithms and access with ease
- Workload
  - Easy scaling of widely ranging workloads utilizing shared, very large data recsources

ROAD
CLOSED

And Extra Credit

- The Anti-Database
- Signal to Noise ratio really bad in much of the Big Data world
- How to throw away the RIGHT data?
  - In an intelligent way
  - Learning
  - Adjusting
  - Related, connected streams
  - Easy rules specification
- How many readings do we need to tell us the temperature did not change over the last hour?
- CERN spends more compute power throwing away data than storing and analyzing it

- *"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair…*



- ***--- Charles Dickens – "A Tale of Two Cities"***