Introduction to Data Management

MIT Center for
Transportation & Logistics

ctl.mit.edu

Information Growth

1 Exabyte =
1 billion Gigabytes

- The amount of digital information being produced is increasing . . .
  - A total of 1.5 Exabytes of unique information was produced in 1999. This grew to 5 Exabytes by 2003.(1)
  - IDC calculated that from 2006 to 2012 the amount of data generated increased from 161 to 2837 Exabytes! (2)

- This information is coming from a wide variety of sources . . .

Facebook's 1.2 billion active daily users post 300 million photos and 4.5 billion "likes" a day!

Twitter averages 6,000 tweets posted every second!

Google processes over 40,000 search queries every second on average!

1. Lyman, P. and Varian H., (2003) "How Much Information"
2. Gantz, J. and Reinsel, D., (2012) "The Digital Universe in 2020"

To include supply chains . . . .

Amazon sells more than 480 million unique items to 244 million customers!

Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.

Every day, on average, UPS delivers over 20 million packages and documents to more than 8.4 million delivery points using over 100,000 delivery vehicles.

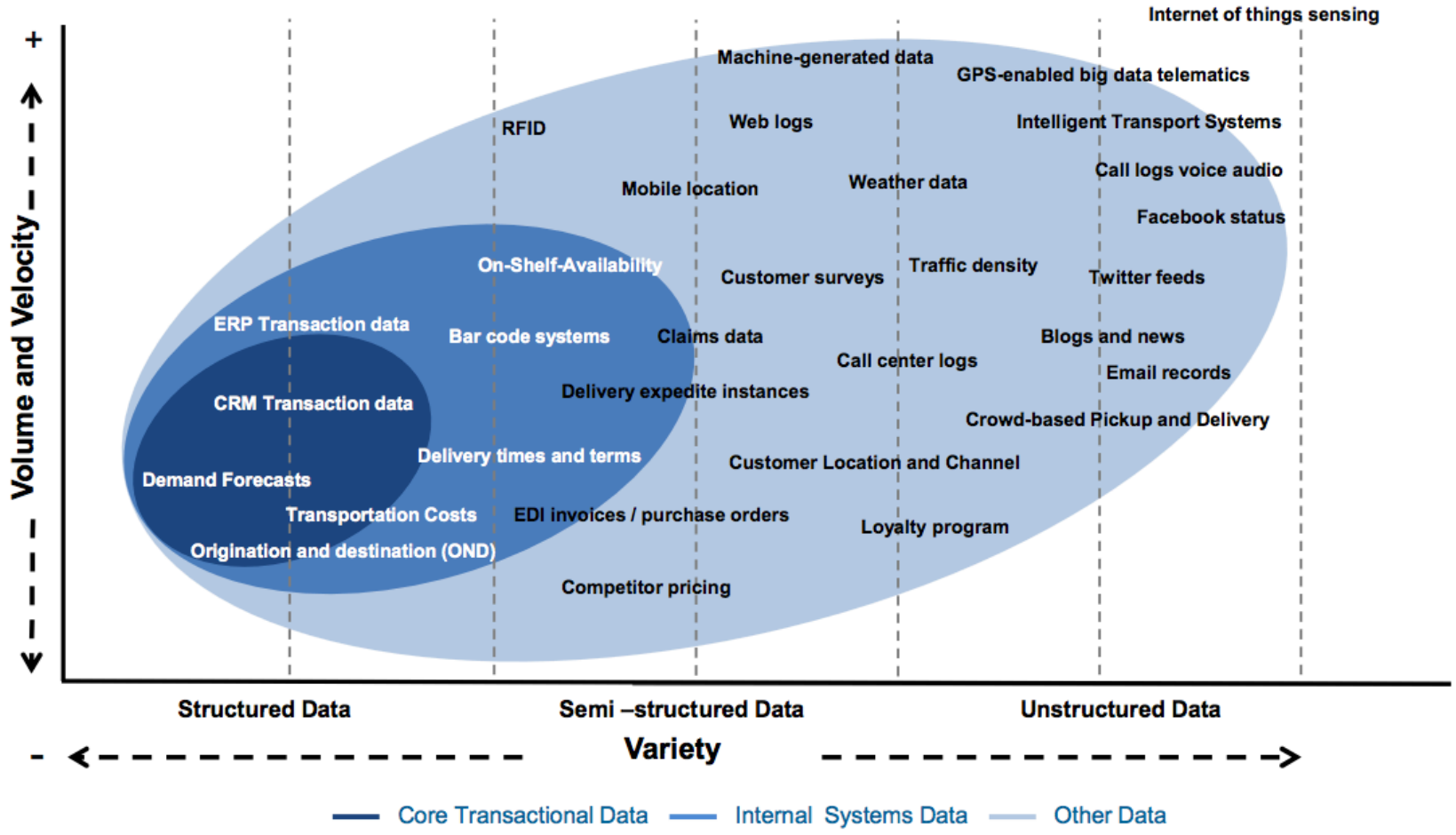
Macy's adjusts pricing in near-real time for 73 million items, based on demand and inventory, every day.

Other contributing factors . . .

- Sensors – generating ever more data
  - Automobiles have 60-100 sensors today and will increase to over 200
  - Smart packaging for shipment of perishable items captures temperature, humidity, etc. while in transit
  - Automated diagnostic tests on computers and servers
- Internet of Things – machines communicating with machines
  - There are ~6.4 billion connected things in use worldwide in 2016.
  - Gartner forecasts that number will grow to ~21 billion by 2020.
  - Estimates of 170 million wearable devices online by 2019.
- Regulations – increasing the information required to store
  - 2013 Drug Supply Chain Security Act (DSCSA) requires "serialized numeric identifier" (SNI)
    - Example (Basta): Transaction records for a single batch of 10,000 units
      - Pre-Serialization: 4 transactions (Create batch, Pick pallet, Ship pallet, Send ASN) ~ 2KB data
      - Post-Serialization: 60k transactions (10k each for Provisioning, Commissioning, Aggregating, Picking, Shipping, and Sending SNI data) ~ 11 MB data

Source: Basta, N. (2016) "Pharma Traceability: Year in Review"

It is not just the size, though . . .



Complexity is driven by Volume, Velocity, Variety, and Veracity!

MIT Center for Transportation & Logistics

Source: Rozados, I. and B. Tjahjono, (2014), "Big Data Analytics in Supply Chain Management"

Example: TinyCo

TinyCo

- You have recently been hired by TinyCo within their Supply Chain Management group.
- As your first assignment, you have been asked by your boss, Edith, VP Supply Chain, to help understand the demand and forecasting processes. As part of your learning, she has told you to look into a specific store, meet with the store manager, and report back to her. Other members of your team will be exploring other stores.
- You have asked the store managers for data on their sales for the last 3 years. The person in charge of the data informs you that all of the stores are closed on Sundays. He also has sent you a spreadsheet with the following email message:

> *Welcome to ToyCo! I have attached the data that you asked for in the spreadsheet. It has six tabs: sku masterand separate tabs for each of the following stores: 312, 323, 415, 521, and 632. The store tabs contain all of the daily sales information for the last two years (that is all I could find) for that respective store. It lists the database or transaction ID, date, the SKU, quantity sold that day, and total revenue from those sales. The sku master tab contains some information on each SKU, such as its name, weight, cube, unit cost, etc. This was really hard to get – so I hope that this is all you need.*

TinyCo Data

SKU Master Table

Department	Class	Style	Color	SKU	Cost	Price	Vendor
800	05	20	02	8000520021	7.50	9.99	MA Excellent Products
800	04	51	11	8000451112	9.00	12.99	MA Excellent Products
731	24	55	52	7312455520	25.00	31.99	MA Excellent Products
731	24	55	53	7312455530	14.50	22.99	GA General Wholesales
5001	201	12	4	50012011240	2.50	7.99	China Imports
5001	201	12	5	50012011250	7.50	9.99	China Imports
5001	201	13	4	50012011341	8.00	6.99	China Imports
5001	300	01	1	50013000110	6.50	12.99	China Imports

Transaction File for Store 312

DB_ID	SKU	Store	Date	Unit Sales	Dollar Sales
79444	50012011250	312	8/3/14	3	29.97
79445	50012011250	312	8/4/14	2	19.98
79446	50012011250	312	8/5/14	5	49.95
79447	50012011250	312	8/6/14	3	29.97
79448	50012011250	312	8/7/14	7	69.93
79449	50012011250	312	8/8/14	4	39.96
79450	50012011250	312	8/10/14	4	39.96
79451	50012011250	312	8/11/14	4	39.96
79452	50012011250	312	8/12/14	1	9.99
79453	50012011250	312	8/13/14	3	29.97
79454	50012011250	312	8/14/14	4	39.96
79455	50012011250	312	8/15/14	1	9.99
79456	50012011250	312	8/17/14	2	19.98
. . .	. . .	. . .	. . .	. . .	. . .

Edith, initially, wants you to characterize the total sales in both units and dollar value.

Recall How to Characterize a Distribution

- Central Tendency
  - Mode – value that appears most frequently
  - Median – value in the "middle" of a distribution, separating the lower from the higher half
  - Mean (μ) – sum of values multiplied by their probability (expected value
- Spread
  - Range – maximum value minus minimum value
  - Inner Quartiles – 75th percentile value minus the 25th percentile value
  - Variance (σ^2) - expectation of the squared deviation around the mean
  - Standard Deviation (σ) - Square root of the variance
  - Coefficient of Variation (CV) – Standard deviation over the mean = σ/μ

$$E[X] = \bar{x} = \mu = \sum_{i=1}^{n} p_i x_i \qquad Var[X] = \sigma^2 = \sum_{i=1}^{n} p_i \left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} p_i \left(x_i - \mu\right)^2$$

TinyCo Continued

TinyCo – Store 312 Sales Summary

I am guessing many of you had results something like this . . .

	Unit	Dollar
Minimum	-6	-59.94
25th Pctl	2	15.98
Mode	2	29.97
Median	3	31.99
Average	108.08	1,097.78
75th Pctl	5	68.97
Maximum	60000	599400
Range	60006	599460
InnerQuartile	3	52.99
Variance	4,165,470	415,683,414
Std. Dev.	2,041	20,388
CV	18.88	18.57

So . . . does anything look strange here?

- How can I have negative sales?
- Why is my average so much larger than the median?
- Why is the maximum so large?
- Whoa! That is a huge coefficient of variation!

Big warning flags and sirens should be going off in your head!

- All "mathematical methods" always assume **clean** data!
- As you describe data always keep in mind the following questions:
  - Is the data clean?
  - Is the data complete?
  - What assumptions are you making about the data?
  - Are the results making sense? How can I check?

Cleaning Data is Never an Afterthought!!!

- Always plan enough time for basic data checks
  - Garbage in, garbage out
- Typical checks
  - Invalid values - negative, text, too small, too big, missing
  - Mismatches between related data sets - # of rows, # of cols
  - Duplication – unique identifiers
  - Human error – wrong dates, invalid assumptions
  - Always explore the outliers – they are the most interesting!
- Be organized
  - Versioning
  - Keep track of data changes

TinyCo – Store 312 Sales Summary – a 2nd Look

Looking a little closer . . .

Issue	Action
21 records with negative sales	Probably returns (Check on TinyCo Policy) keep in for now
34 records with zero value	Removed
50 blank sales records (null)	Removed
30 records with non-numeric data	Removed
11 records with very large values	Appear to be multiplied by 10,000 – probably fat fingered – flag and correct to right order of magnitude
3 duplicated entries	Removed duplicates (2001601, 2001453, 2001331)
321 records w/ non-matching SKUs	Flag and keep for now (8 @ 50*12011250 and 313 @50012011340) – Check for SKU changes

- Some Tips
  - Always log issues and actions – keep an audit mindset
  - Never completely delete data – keep a complete data archive
  - Create a "clean data set" to use in analysis
  - Always validate decisions with process owners
  - Use graphs (Scatter Plots, Histograms, etc.)
  - Use pivot tables (if you must stay in spreadsheets)

TinyCo – Store 312 Sales Summary – a 2nd Look

After some initial cleaning, my summary is as follows . . .

	Unit	Dollar
Minimum	-6	-59.94
25th Pctl	2	15.98
Mode	2	29.97
Median	3	31.99
Average	5.83	76.44
75th Pctl	5	68.97
Maximum	42	699.93
Range	48	759.87
InnerQuartile	3	52.99
Variance	82	14,219
Std. Dev.	9	119
CV	1.55	1.56

- But this is barely the tip of the iceberg on what we want to do with this data!
  - How do the different SKUs behave?
  - How do sales differ by month? by week? by day of week?
  - Are sales prices per SKU consistent?
  - Are there trends in sales over time?
  - Do the returns (negative sales) correlate to other sales?
  - How are sales of the different SKUs related to each other?

Querying the Data

TinyCo II

- You are still working for TinyCo. Edith would now like to understand on how the the products sourced from MA Excellent Products behave over time.
- Specifically, she would like to know the following for those selected SKUs:
  - Which month has the highest average sales in dollars for store 312?
  - Which specific week had the highest sales in dollars for store 312?
  - Are the sales (in units) between the different SKUs from MA Excellent Products correlated for store 312?
  - How does the profit margin for sales in store 312 change over the time period?
  - TinyCo is thinking of running a one day promotion each week – which day of week makes the most sense for store 312?
  - How do these SKUs behave differently in the other stores?

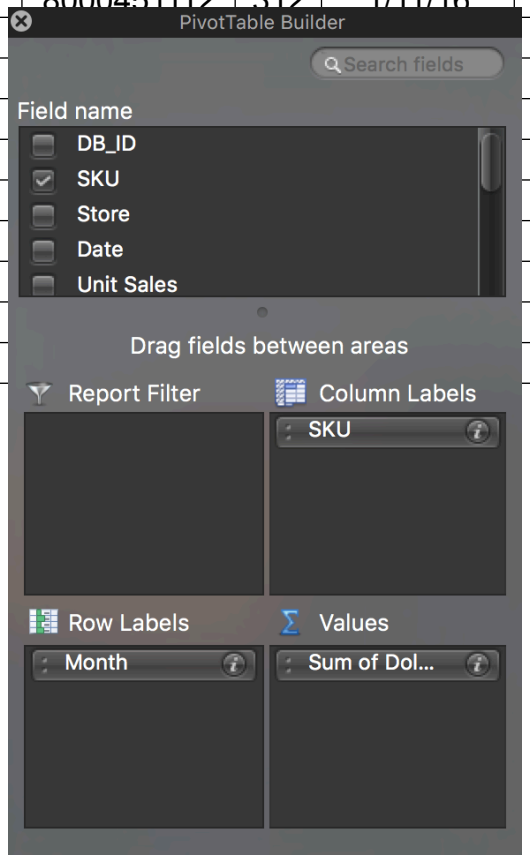
How would you answer these queries in a spreadsheet?

Querying the Data - Spreadsheets

- You *can* do this analysis in a spreadsheet, but it takes a little bit of work.
- Pivot Tables can help
  - Data summarization tool found in LibreOffice, Google Sheets, and Excel
  - They automatically sort, count, total or average the data stored in one table or spreadsheet, displaying the results in a second table showing the summarized data.
  - Very useful in tabulating and cross-tabulating data

Pivot Tables

DB_ID	SKU	Store	Date	Unit Sales	Dollar Sales	DOW	WOY	Month	Year	Week Year	Month Year
79568	50012011250	312	12/30/14	7	699.93	3	53	12	2014	201453	201412
85081	50013000110	312	5/26/16	42	545.58	5	22	5	2016	201622	201605
89834	8000451112	312	8/8/15	42	545.58	7	32	8	2015	201532	201508
89899	8000451112	312	11/14/15	42	545.58	7	46	11	2015	201546	201511
89946	8000451112	312	1/11/16	42	545.58	2	3	1	2016	201603	201601
89969				42							
89982				42							
90010				42							
90028				42							
90071				42							
90093				42							
90105				42							
90115				42							
90158				42							



Sum of Dollar Sales	Column Labels			
Row Labels	7312455520	8000451112	8000520021	Grand Total
1	$1,477	$17,387	$2,018	$20,882
2	$1,683	$12,573	$2,308	$16,564
3	$2,239	$14,211	$2,208	$18,658
4	$1,693	$15,387	$2,458	$19,537
5	$1,801	$13,401	$2,887	$18,090
6	$1,553	$11,397	$2,757	$15,708
7	$2,058	$16,767	$2,677	$21,503
8	$1,487	$15,882	$2,537	$19,907
9	$1,551	$7,800	$1,499	$10,850
10	$2,327	$7,797	$1,838	$11,962
11	$1,204	$9,432	$2,138	$12,773
12	$1,697	$10,285	$2,068	$14,050
Grand Total	**$20,774**	**$152,318**	**$27,393**	**$200,484**

The Big Data Challenge

The Opportunity . . .

- The Digital Disruption is here . . .
  - World's largest taxi company owns no taxis (Uber)
  - Largest accommodation provider owns no real estate (Airbnb)
  - Largest phone company owns no telco infrastructure (Skype, WeChat)
  - World's most valuable retailer has no inventory (Alibaba)
  - Most popular media owner creates no content (Facebook)
  - Fastest growing banks have no actual money (SocietyOne)
  - World's largest movie house owns no cinemas (Netflix)
  - Largest software vendors don't write apps (Apple/Google)

- Data Analytics have become pervasive . .
  - Retail and CPG - Demand and sales forecasting, Promotions, On-line pricing and product recommendations
  - Financial Services – Underwriting, Credit risk, Fraud analysis, Litigation mitigation
  - Telecommunications - TV networks, Product subscriptions
  - Marketing - Direct marketing, email, mobile ads, SEO, CRM, Loyalty programs, Customer acquisition and retention,
  - Distribution network and supply chain – Real-time track & trace, Micro-segmentation, Predictive cost estimation, Asset monitoring (digital twins), Serialization

Source: [www.ibmforentrepreneurs.com] and Cassa, C. (2016) Course notes SCM.264

The Challenge . . .

"It's an absolute myth that you can send an algorithm
over raw data and have insights pop up."
Jeffrey Heer, co-founder of Trifacta

"Facts are simple and facts are straight
Facts are lazy and facts are late
Facts all come with points of view
Facts don't do what I want them to
Facts just twist the truth around
Facts are living turned inside out"
The Talking Heads

There is no magical data science machine turning data into insights!

- Data is messy – always requires lots of cleaning and usually some programming
- Data is usually siloed – need to combine data from multiple disparate sources
- Data is big and getting bigger – standard techniques and tools (i.e., spreadsheets) are no longer sufficient

Source: Cassa, C. (2016) Course notes SCM.264

Why are we weaning you off of spreadsheets?

- Essentially it is a choice of Structured vs. Unstructured data
- A database is a structured way of storing data
  - Impose rules, constraints, relationships
  - Abstraction: Separates data use from how and where the data is stored. This allows systems to grow and makes them easier to develop and maintain through modularity.
  - Performance: Database may be tuned for high performance for the task that needs to be done (many reads, many writes, concurrency)
- Spreadsheets are essentially unstructured data
  - You may be provided with one giant spreadsheet and you need to do lots of different things with it
  - Great for informal, causal, and one-off analysis and prototyping
  - Not suited for repeatable, auditable, or high performance production

Source: Cassa, C. (2016) Course notes SCM.264

Other problems with unstructured data storage

- **Redundancy**:
  - the same data could be represented multiple times in a file
- **Clarity**:
  - hard to know relationships between elements in a dataset and business process rules
- **Consistency**:
  - hard to ensure that dataset has the same value in all locations where it appears in files / spreadsheets
- **Security**:
  - Cannot control multi-user access to a file
- **Scalability**:
  - Imagine the trouble if there are hundreds or thousands of users, vendors, brokers, products, stores, etc.

Source: Cassa, C. (2016) Course notes SCM.264

Questions, Comments, Suggestions?
Use the Discussion Forum!

"Cleo and Sadie wanting to go from data to outside, not insight"

MIT Center for
Transportation & Logistics

caplice@mit.edu
ctl.mit.edu