REUTERS/Yves Herman

BUILDING DATA SCIENCE

Harvard Data Science

December 2015

@monavernon @bulicny

mona.vernon@tr.com brian.ulicny@tr.com

THOMSON REUTERS

Who is Thomson Reuters?

FINANCIAL & RISK

Critical news, information & analytics, enables transactions, and connects trading, investing, financial and corporate professionals.

LEGAL

Critical information, decision support tools, software & services to legal, investigation, business and government professionals.

TAX & ACCOUNTING

Integrated tax compliance and accounting information, software & services for professionals in accounting firms, corporations, law firms and government.

INTELLECTUAL PROPERTY & SCIENCE

Comprehensive IP & scientific information, decision support tools & services to enable governments, academia, publishers, corporations & law firms.

REUTERS NEWS

Powered by more than 2,800 journalists reporting in 20 languages from bureaus around the world, **Reuters** is the world's largest international news organization

Data Overview, Single Company: Boehringer Ingelheim

Financial & Risk

Legal

Tax & Accounting

IP & Science

48269	16268	180	86753 docs
News	Case Law	Editorial Analyses	Scientific Articles
Broker Research	Admin Decisions		Patents
Bonds	Public Records		Trademarks
Fundamentals	Dockets		Domain Names
Press Releases	Arbitration		Clinical Trials
			Drugs

Three Vs at TR:

Velocity from fractions of seconds to quarterly filings.

Volume: all the data needed by target professionals

Variety: multiple disparate content, formats, languages.

THOMSON REUTERS

Big data is not new to Thomson Reuters

- On an average day we distribute **10 billion bytes of real-time pricing data alone**, around every corner of the world. At peak we run up to 8 million per second.
- We process and collect more data in a day than we did in a month just five years ago.
- We use our own **predictive analytics** to improve how we find, extract and tag data – enabling customers to use data in ways not possible before.
- We use **semantic analysis** and learning machines to generate sentiment on news and social media.
- We screen **100 million websites a day** to help our customers identify hidden risks to help them protect their business.
- We run **machine learning algorithms** to spot suspicious trading patterns and potential fraud, and detect problems

GETTING VALUE OUT OF DATA

INTRODUCE THE DATA INNOVATION LAB

DATA MONETIZATION BUSINESS MODELS

DEMONSTRATION OF RECENT DATA SCIENCE PROJECTS

ABOUT THE DATA INNOVATION LAB

We are located in Boston's Innovation District and closely connected to the MIT and Cambridge Startup Ecosystem

Currently, 10 data scientists and visualization experts

Mission

- **Deliver insights** for customers with cutting-edge data science proof-of-concepts
- **Create value** with all the content from across Thomson Reuters enterprise, and from external partners, including open data

DATA INNOVATION LAB: HOW WE WORK

Project Team

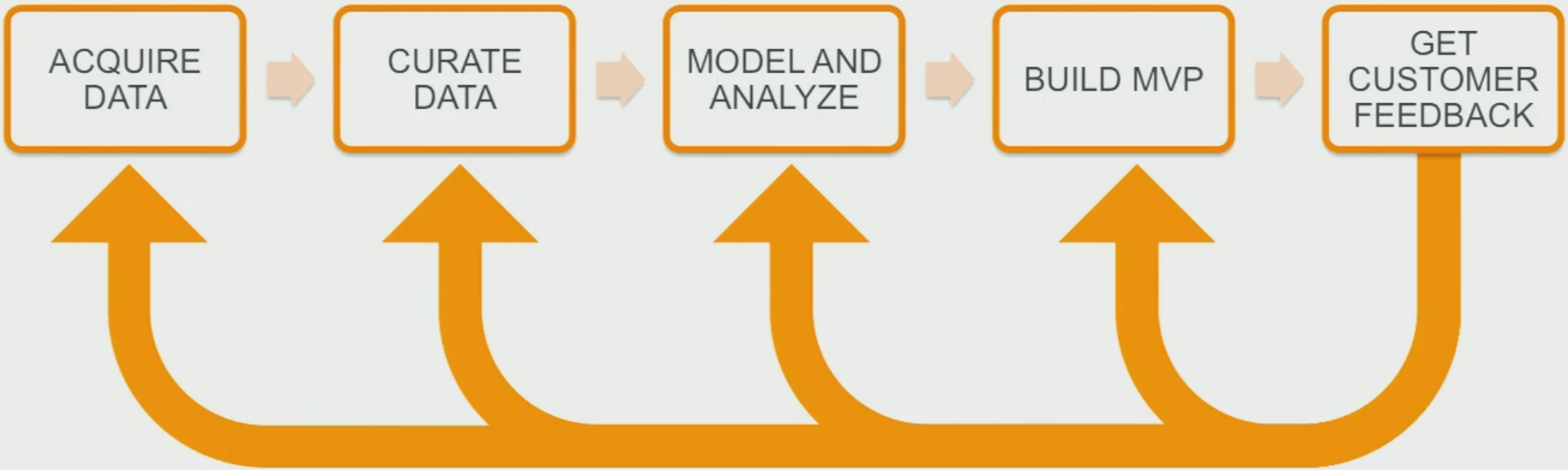
- Data Scientists
  (Data Innovation Lab)
- Business Lead & Content Experts
  (Thomson Reuters)
- External partner team member(s)
  (customers, vendors)

Deliverables

- Proof-of-concepts to gain customer feedback
- Analytical models
- Interactive demos and visualizations

LEAN DATA SCIENCE SPRINTS

Agile Data Science Projects, iterative, 2-week sprints
to build MVPs (Minimum Viable Products (UI, API, Visualization)



DATA INNOVATION LABS: PROJECTS

Sample Projects:

- Integration of data across Thomson Reuters
  Patents, News, Legal, Tax & Accounting, etc. via PermID
- Mining new data sources for Finance
  Industry partnerships
  Open data
- Knowledge Graphs
  Graph databases
  Semantic web
  Visualization
- Startup Ecosystem
  Collaborating and co-developing solutions for customers

DATA INNOVATION LAB: CAPABILITIES

Data Science

machine learning, classification, regression, anomaly detection

Text Mining and NLP

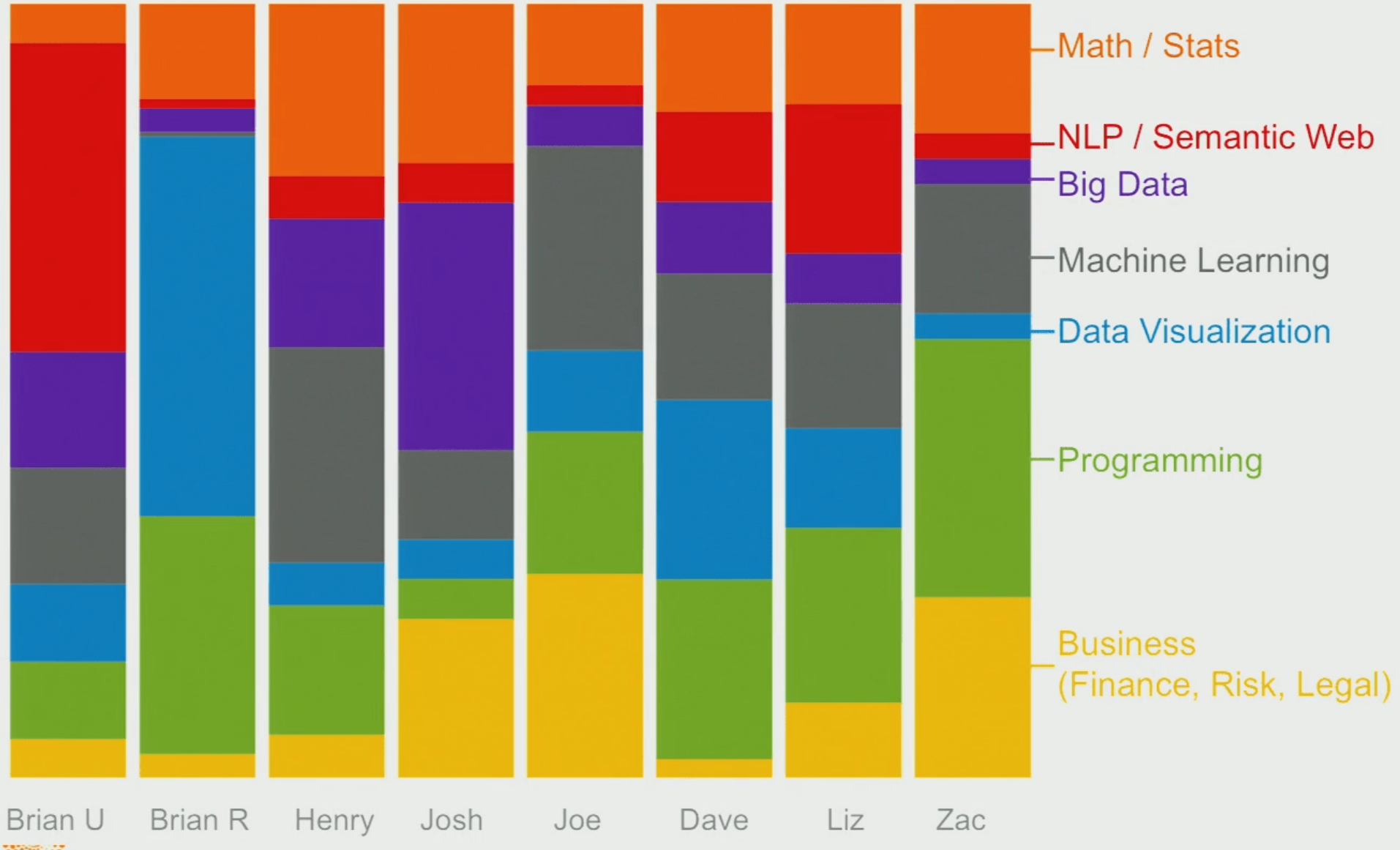
Big Data

Hadoop, Spark, Hive, Graph DB, etc.

Quantitative Modeling & Financial Research

Rapid Prototyping & Data Visualization

Data Science Skills In the Lab



GETTING VALUE OUT OF DATA

INTRODUCE THE DATA INNOVATION LAB

DATA MONETIZATION BUSINESS MODELS

DEMONSTRATION OF RECENT DATA SCIENCE PROJECTS

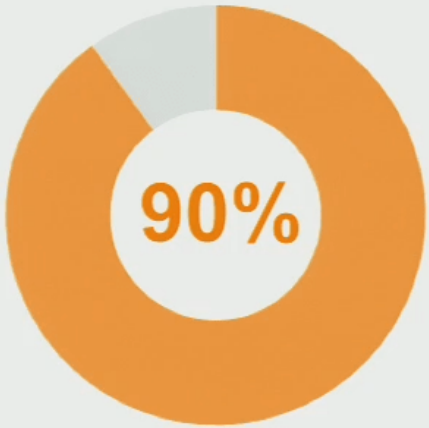
We define the following terms as…

- Raw Data: Data as it is collected from the source – not manipulated or processed (e.g. sensor data)
- Time Series: Measurements of prices, economic data through time
- Reference Data: Terms and conditions, financials statements
- Analytics: Discovery and communication of meaningful patterns in data
- Predictive Analytics: Extracting information from data to determine patterns and predict future outcomes and trends
- Data Exhaust: Data generated as information byproducts resulting from digital or online activities
- Entity Analytics: The analysis of the connections/linkages between different entities or information

Data is generated at an unprecedented rate

2.5 quintillion bytes
(2.3 trillion gigabytes) of data created each day

At least **100 terabytes** of data stored by most companies in the US



of data in the world created in the last 2 years

40 zettabytes (43 trillion gigabytes) of data will be created by 2020 – **300 times increase** from 2005

1 terabyte of trade information captured by the New York Stock Exchange in each trading session

Source: IBM

Big data and analytics market is expected to reach $125 billion in 2015

of large organizations purchase external data today

70%

IoT analytics is expected to grow at a 5-year CAGR of 30%

Applications with advanced and predictive analytics will grow 65% faster than applications without predictive functionality

100%

of large organizations will purchase external data by 2019

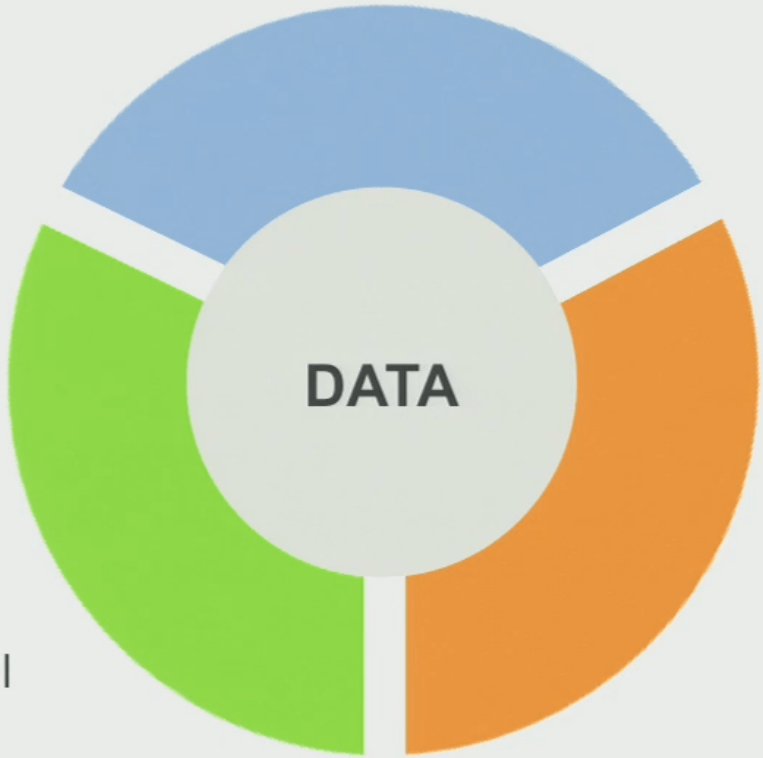
More organizations will begin to monetize their data

Source: IDC Worldwide Big Data and Analytics Predictions for 2015

Data has value outside traditional business operations

CORE BUSINESS OPERATIONS

Data generated and used as part of traditional business operations



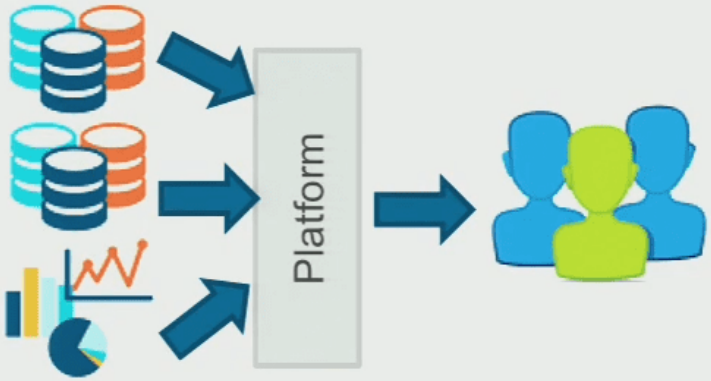
INTERNAL ANALYTICS

Analytics to drive strategy, business intelligence, and marketing analytics. Enable deeper customer engagement, optimize operations, unlock cross-sell and up-sell opportunities.

DATA MONETIZATION MODELS

Data or data-driven insights (external-facing analytics) provided to customers. Enter adjacent and new markets.

There are 3 main data monetization models



Sell Raw Data	Provide Data Analytics	Develop Data Platform
Provide raw data directly to customers or through distributors	Develop new analytics from internal data alone or combined with other data sources. Raw data may or may not be provided in the solution.	Provide a marketplace and platform for multiple data sources and analytics applications.
Easy to accomplish and short time-to-market	Differentiation via data-driven decision making internally or as a service	Multi-sided network effects
Value is a function of exclusivity or differentiated access	Need data scientists to develop the capabilities in-house	Need data science, software development and business expertise in platform business strategy

Selling raw data is most valuable when the data is unique to the providing firm

Value of raw data increases when the data is inimitable.

Firms may also provide raw data free or in a freemium model to increase customer engagement.

- Target provides downloads of its point of sale (POS) data to suppliers who use this information to monitor inventory levels and customer demand in real-time.
- Walmart provides all of the sell though data to their suppliers by SKU, hour and store with their Retail Link system.
- Nielsen sells customer shopper behavior in 250,000 households in 25 countries to consumer goods suppliers and retailers, which in turn use this data to increase marketing and sales effectiveness.
- INRIX provides real-time traffic information to Ford Motor Company to be integrated in its in-car navigation system.

Providing data analytics

Powerful analytics applications connect and link data from different sources.

- ARPA-E's Project TERRA collects performance data from plants in the field, incorporates genomic data and builds algorithms to correlate a gene to a specific trait or plant performance.
- Interactions Marketing combines retail POS data and regional weather data to provide insights into customer behavior to retailers and manufacturers.
- Weather Underground provides weather data to businesses to look for patterns in their sales with respect to weather changes.
- Boston-based start-up CargoMetrics combines location of ships, reported through tracking systems, with the contents of ships to sell commodities movement information to hedge funds.

Example: Analytics adds significant value to commodities trading

Maritime movement of commodities data

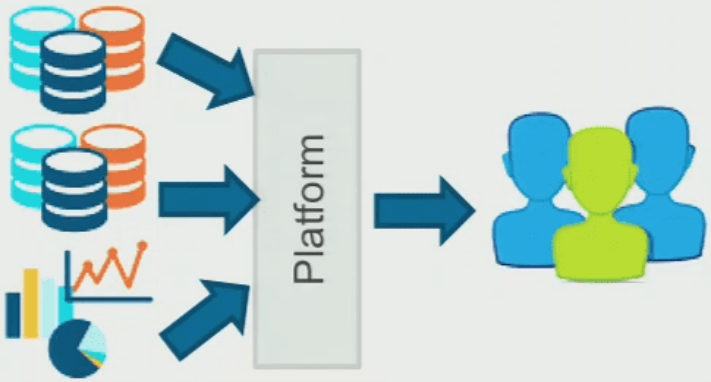
Maritime trade accounts for more than 80% of the global commerce

Proprietary metadata
Advanced cargo-modeling algorithms and analytics

Novel trading strategies utilized in internal hedge fund

Boston-based analytics startup CargoMetrics raised $2.14M in 2014

Data platforms can harness powerful network effects



Algorithms that match customers with the data they need is a key component for creating value in data platforms.

- Airbnb, Uber, Salesforce provide platforms for data and service providers to connect with users.
- **Thomson Reuters App Studio** connects app builders to the Eikon platform customers
- Riskpulse combines analytics with its platform where third parties can supply hazard data to help businesses identify risks related to weather conditions, natural disasters (earthquakes, volcano eruptions, etc.) and port strikes.

Common Issues and Challenges

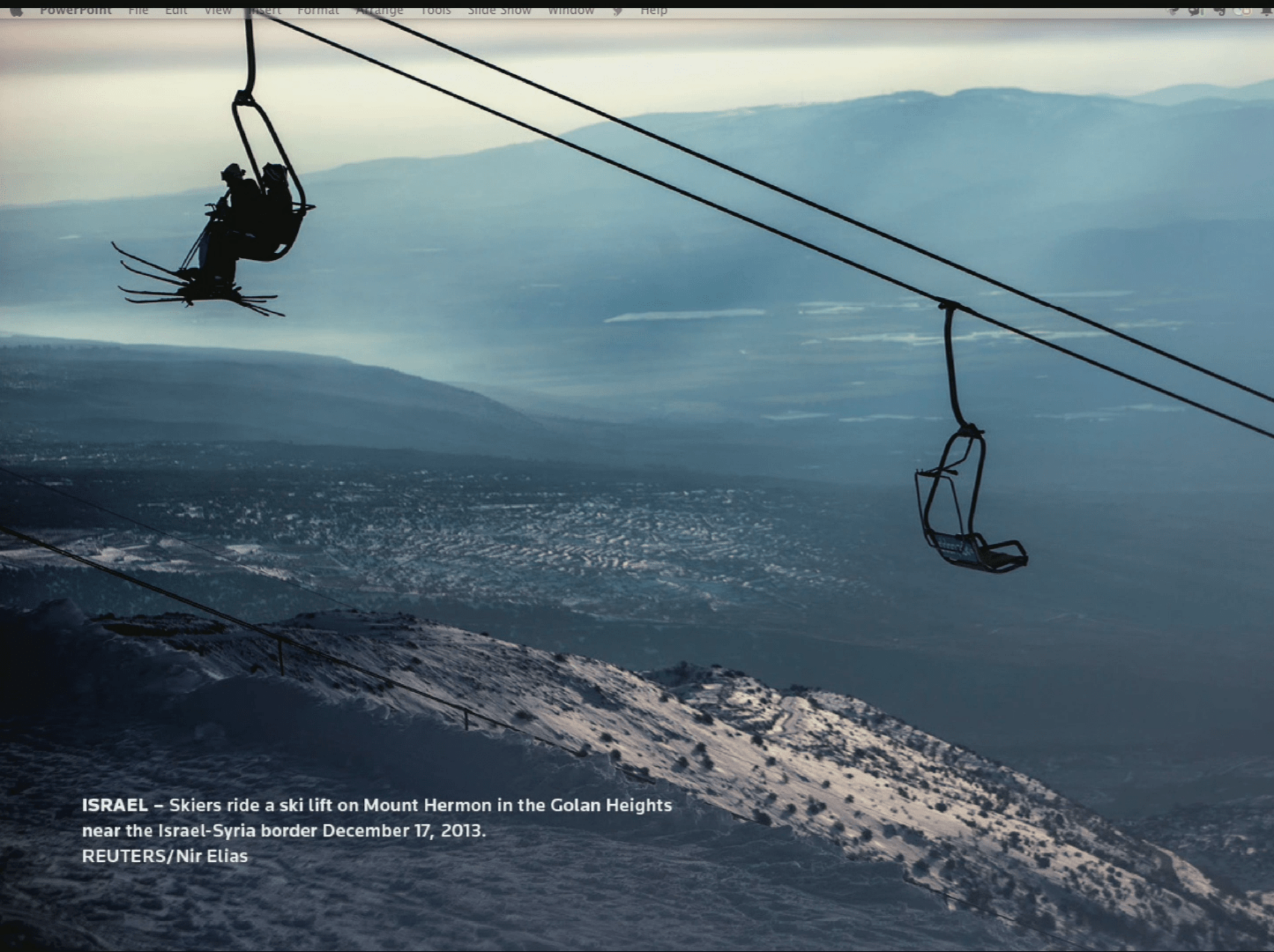
- Finding the initial hypothesis
  - eg crop yields affects commodity prices, affects farmer profitability, affects equipment purchases, affects John Deere Sales, affects John Deere EPS, effects JD share price performance.
- Finding a reliable, repeatable method to collect the data with the required frequency
- Establishing history where possible
- Establishing rights to the data /
- Establishing Internal Agreement to Externalize the data
- Striking balance between source protection and maintaining value
- Confirming the data is not trumped by another better, faster, cheaper source
- Concording the data
- Linking to a tradable security

GETTING VALUE OUT OF DATA

INTRODUCE THE DATA INNOVATION LAB

DATA MONETIZATION BUSINESS MODELS

DEMONSTRATION OF RECENT DATA SCIENCE/DATA MONETIZATION PROJECTS



ISRAEL – Skiers ride a ski lift on Mount Hermon in the Golan Heights near the Israel-Syria border December 17, 2013.
REUTERS/Nir Elias

Invitation

Please plan to join us Saturday, Jan 23 to Sunday, Jan 24

Data Science against Slavery Hackathon

District Hall (Seaport, at Courthouse Station on Silver Line)

Sponsored by Thomson Reuters

In affiliation with Demand Abolition and CEASE Network.

All levels of experience welcome.

Questions: brian.ulicny@thomsonreuters.com