Creating High-Quality Financial Datasets from Unstructured Public Data

Doug Burdick
IBM Research

Public Data Financial Analytics

- Growing volume of public data around companies, their financial performance and key events
  - Regulatory filings, newsfeeds, government data, research articles
  - Primarily in unstructured format, authored by heterogeneous set of entities
- Creating a structured representation of entities and relationships enables modeling and analyzing institutions and industries
- Developing community of academics, industry, and regulators
  - Data Science for Macro-Modeling (DSMM) workshops
  - Financial Entity Identification and Information Integration (FEIII) Challenge

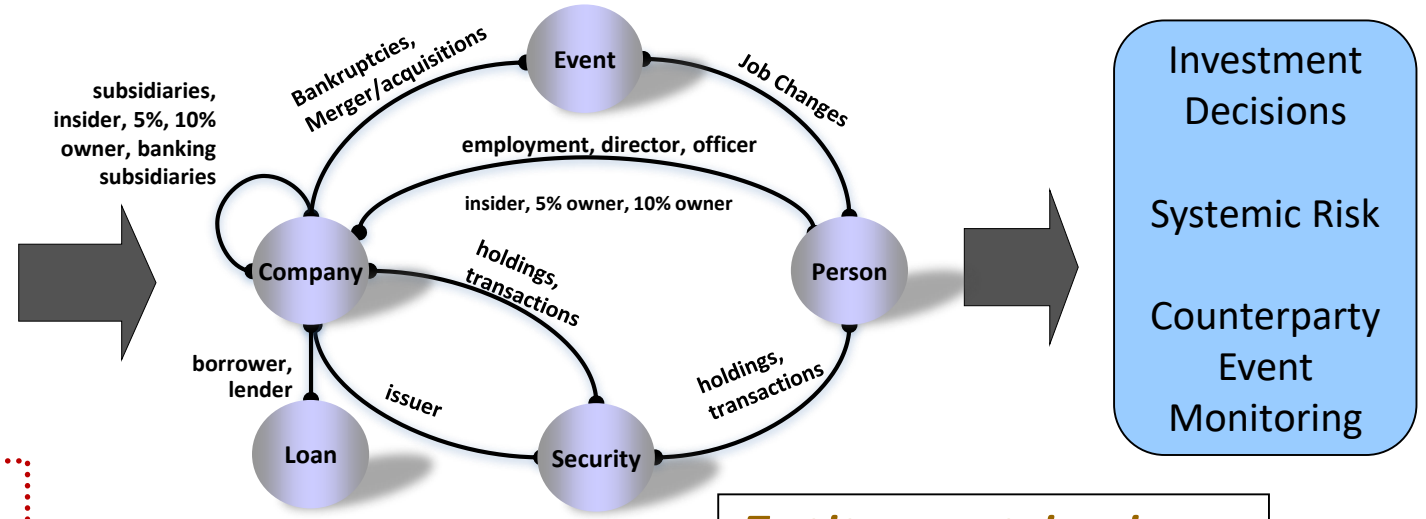
Case Study : Counterparty Relationships from Regulatory Filings

Filing timeline

2005 — 2017

SEC/FDIC Filings of Financial Companies
(Forms 10-K,8-k, 10-Q, DEF 14A, 3/4/5, 13F, SC 13D SC 13 G FDIC Call Reports)

Extract | Integrate

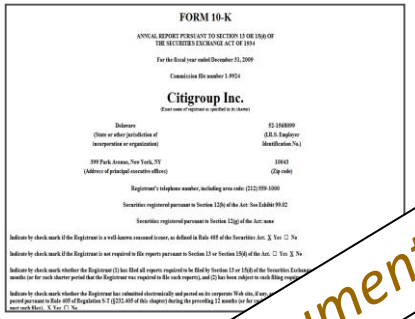


Investment Decisions

Systemic Risk

Counterparty Event Monitoring

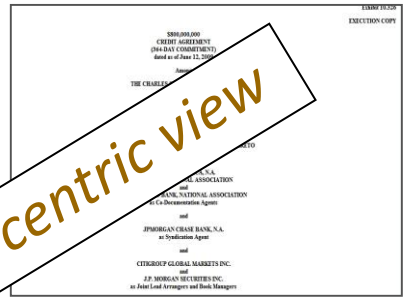
Annual Report



FORM 10-K

Citigroup Inc.

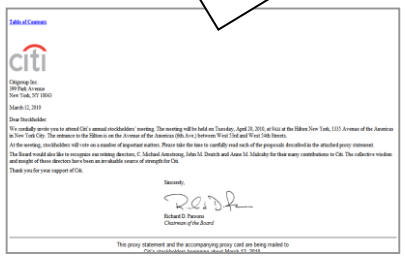
Loan Agreement



EXECUTION COPY

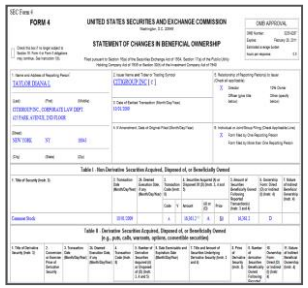
Document-centric view

Proxy St…nt



citi

Insider Transaction

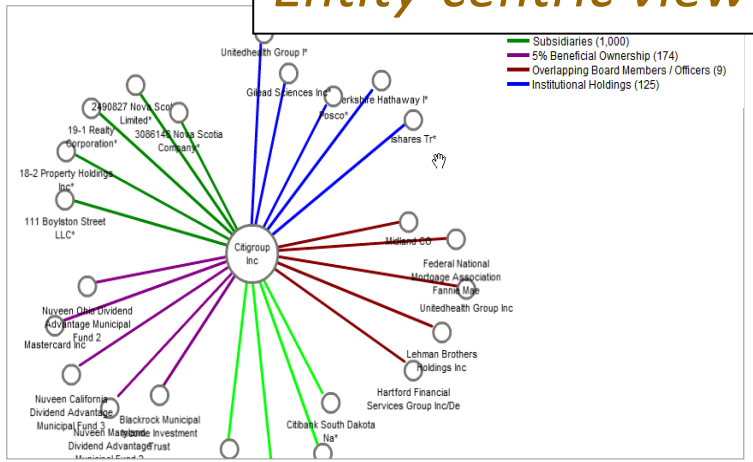


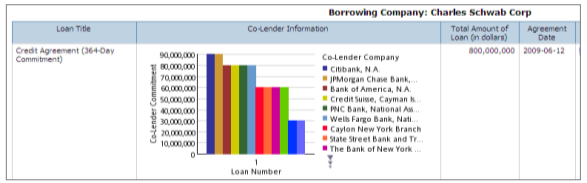
FORM 4

UNITED STATES SECURITIES AND EXCHANGE COMMISSION

STATEMENT OF CHANGES IN BENEFICIAL OWNERSHIP

Entity-centric view





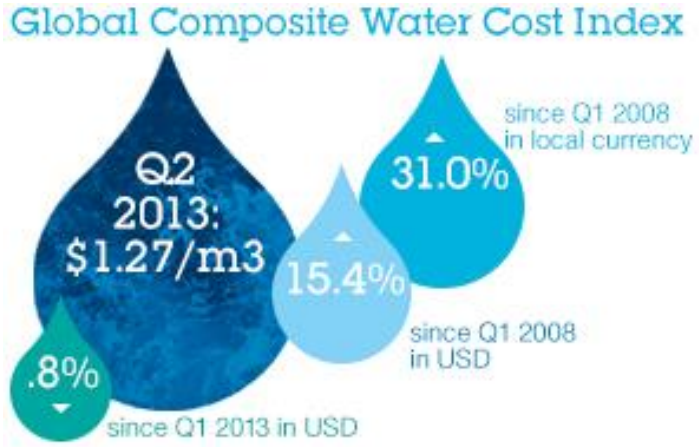
Case Study : Developing Water Cost Index through Big Data Analytics

Problem : What is the cost of water in different regions?

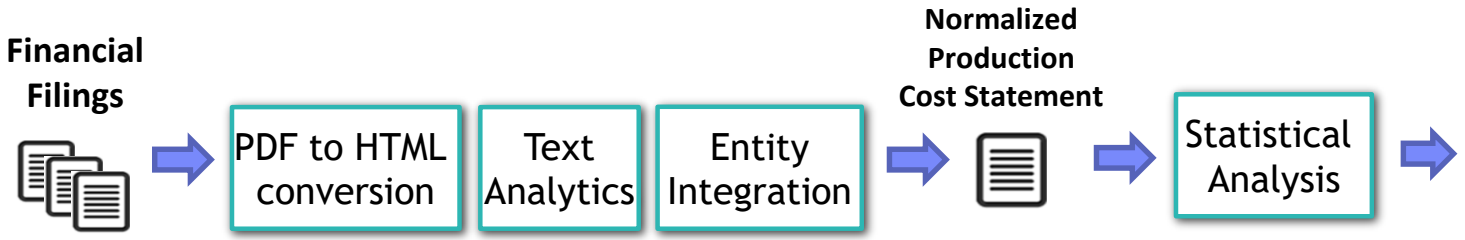
- Provides market benchmark
- Index enables financial products for water

Challenges

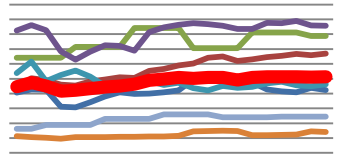
- Unstructured Public Data
- Isolate cost variables
- Identify (in) direct cost subsidies
- Account for filing discrepancies
- Address missing values



Architecture



WCI Published starting Oct 2013



Water cost index computed and updated periodically for each region

Case Study : Enhanced Credit Exposure Analysis using Public Data

- Large Investment Management company wants to analyze public data for
  - *Credit Exposure Analysis* : Construct comprehensive view of municipal bonds and related entities from unstructured data sources like regulatory filings and news articles
    - Related Entities such as Issuer (e.g., water municipality agency) and guarantors (e.g., city, state)
    - Financial Information such as Rating covenants (e.g., underlying revenue stream pledged), Issuer financial health (e.g., changes in number of customers and accounts)

City of Chicago Water Bonds EMMA filing July 2006

Water System Accounts

Year Ended December 31	Non-metered	Metered	Total
1996	333,202	158,573	491,775
1997	332,279	159,309	491,588
1998	331,554	160,520	492,074
1999	329,756	159,353	489,109
2000	328,327	160,895	489,222
2001	327,276	163,051	490,327
2002	326,778	164,067	490,845
2003	325,789	165,440	491,229
2004	324,689	167,545	492,234
2005	323,740	169,664	493,404

Customer Accounts

Entity	Year	Total Customers
…	…	…
Chicago Water System	2004	492,234
Chicago Water System	2005	493,404

  - Event monitoring to help identify adverse events around individual securities and related entities from news articles and web sources
    - E.g., credit rating changes, budget deficits, revenue and spending outlook changes

Bloomberg news article, Apr 2011

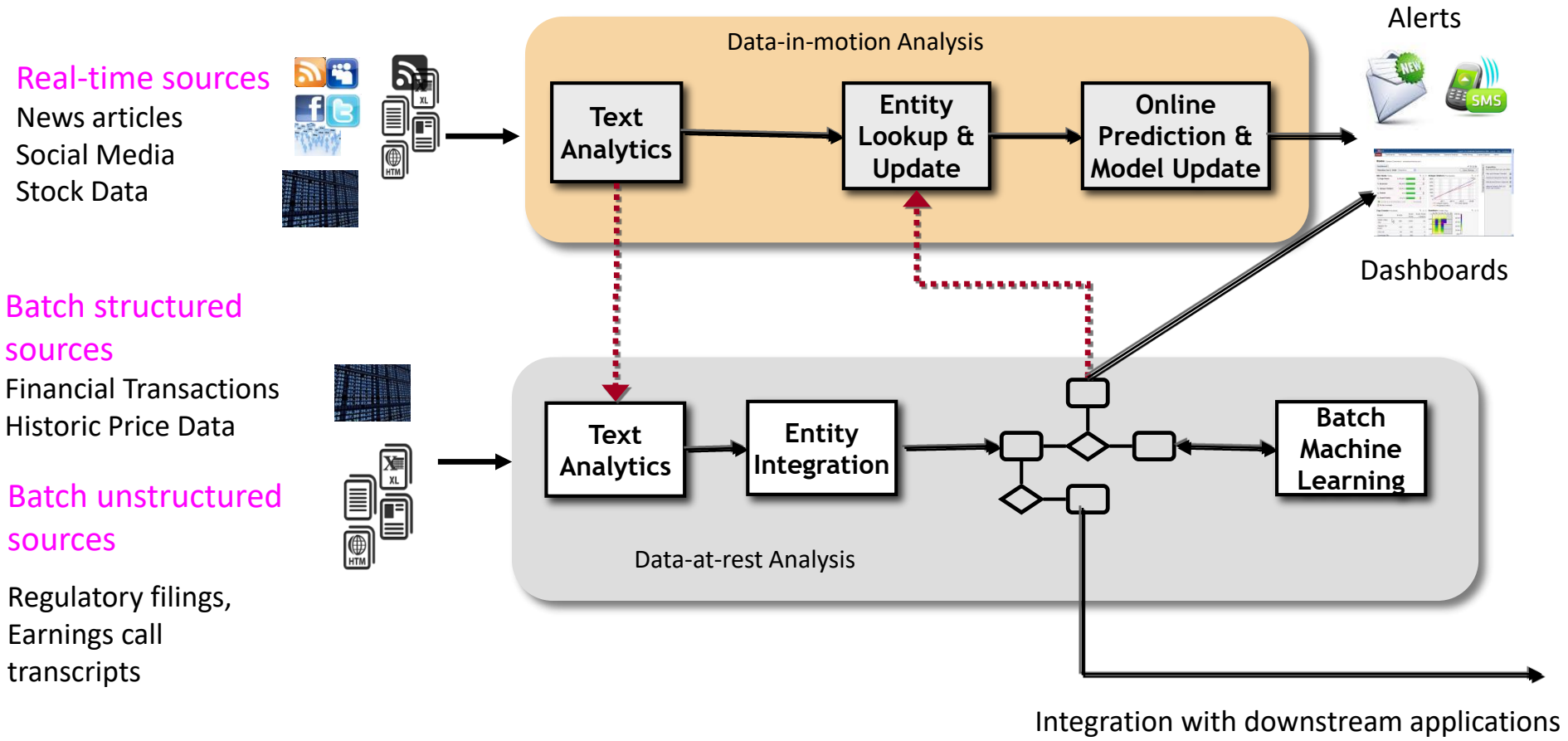
Repeated use of such reserve funds by the current mayor to balance budgets led Standard & Poor's to cut Chicago's credit rating on Nov. 5 by one level to A+

Rating Change

Entity	Agency	Rating	Change	ChangeMagnitude
Chicago	Standard & Poor's	A+	downgrade	-1

- Applications:
  - Enabling credit analysts and investors to understand financial health of issuer and guarantor
  - Identify data variables that help build enhanced valuation and credit risk models
    - E.g., Revenue / spending history and outlook as data variable for predicting bond yield spread

Financial Analytics Architecture



Community Building

- Joint work: University of Maryland, Office of Financial Research (OFR), NIST, IBM
- Build a community of academic researchers, industry practitioners, and financial regulators intersecting finance and big data
- Enable development:
  - Toolkits to create high-quality datasets from public data
  - Collections of high-quality structured datasets
  - Novel modeling techniques to fully leverage these datasets
- Developing community of academics, industry, and regulators
  - Data Science for Macro-Modeling (DSMM) workshops
  - Financial Entity Identification and Information Integration (FEIII) Challenge (sponsored by NIST and OFR)

FEIII Data Challenge

- Sponsored by NIST and US OFR, with multiple participating teams from academia, government, and industry.

- **Goal**: Build a reference financial-entity knowledge base by linking records across multiple financial datasets, including **FFIEC, SEC, LEI.**
  - **FFIEC:** dataset containing **6,652** records corresponding to the banks in the US
  - **SEC:** dataset containing **129,311** records from SEC corresponding to US publicly traded companies
  - **LEI:** dataset containing **53,958** records corresponding to “legal entities” (US and globally)

- Four concrete record linkage tasks:
  - Task1: FFIEC to LEI
  - Task2: FFIEC to SEC
  - Task3: FFIEC ids that match both LEI and SEC
  - Task4: LEI to SEC

- For the first 2 Tasks, ground truth data was provided by the organizers.

Why is the Problem Challenging

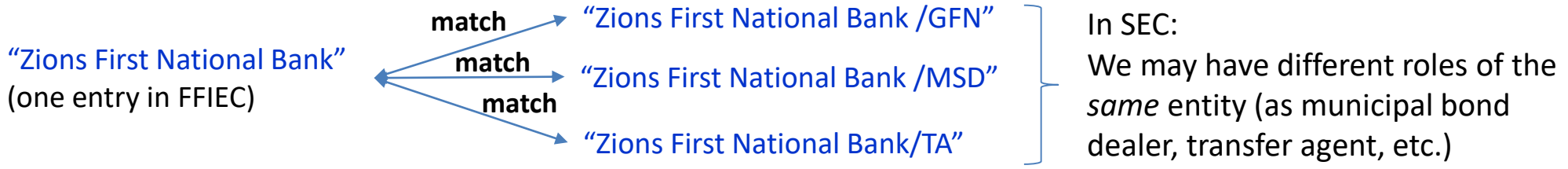
- *Sparsity of the data*
  - Only two attributes were made available for matching: **name** and **location**
    - (e.g., "PNC Bank", in "Wilmington, DE")
  - Even more, enough records with **missing location information**.
  - The **structure of the entity name** is especially important:
    - company suffixes ("Corp"), bank suffixes ("National Association"), acronyms ("N.A."), special characters ("&")

"Isabella Bank" (a bank) **not a match** "Isabella Bank Corp" (parent holding company)

- *Ambiguity*

"Bank of America N.A." **match** "Bank of America National Association"

  - In many situations, the matches are 1-to-1 (i.e., only one true match on each side)
  - However, in other situations, the matches are not 1-to-1:



Ingredients of Our Approach

- **HIL: High-Level Entity Integration Language** *[Hernandez et al, EDBT'13]*
  - Abstraction for the fundamental operations needed to **integrate, link and fuse entities** over structured and unstructured data.
  - Part of IBM BigData analytics suite
  - *Features:*
    - *Ability to define and combine multiple entity resolution algorithms, each exploiting some combination of available attributes*
    - *Constructs or policies to disambiguate among multiple candidate matches*
    - Also, includes **ETL-like operations** (data transformation, mapping and fusion)

Leads to high precision and recall, even over sparse data

- **STG (Smart Term Generation)**
  - Customizable technology to parse structure of entity names
  - We use it as a ***normalization function*** for company names:
    - normalize ("3rd Federal Savings & Loan Bank") = "Third Federal Savings and Loan Bank"
    - normalize ("Bank of America, N.A.") = "Bank of America National Association"

FEIII Data Challenge: Concrete Implementation in HIL

```
create link FFIEC_SEC_Links_1 as
select [ffiec_id: f.IDRSSD, sec_id: s.CIK]
from SEC_REF s, FFIEC_REF f
match using
    rule1:
        normalize  (s.CONFORMED_NAME) = normalize (f.Financial_Institution_Name_Cleaned)
        and toUpper (s.B_CITY) = toUpper (f.Financial_Institution_City)
        and toUpper (s.B_STPR) = toUpper (f.Financial_Institution_State)
    rule2:
        // …  similar rule for legal address
```

A first ER algorithm based on ***normalized name*** and ***location***

- **Logic for FFIEC → SEC task: ~20 lines of HIL**
- **Similar HIL logic (modulo renaming of attributes) was reused for the other tasks**

```
create link FFIEC_SEC_Links_ByName as
select [ffiec_id: f.IDRSSD, sec_id: s.CIK]
from SEC_REF s, FFIEC_REF f
match using
    rule1:
        normalize (s.CONFORMED_NAME) = normalize (f.Financial_Institution_Name_Cleaned)
cardinality ffiec_id 1:1 sec_id;
```

A second ER algorithm to match entities when ***location is missing*** or ***not the same***
-- Uses ***1:1 cardinality constraint*** to strengthen precision

"PNC Bank", in "Wilmington, DE" (in FFIEC) **match** "PNC Bank", in "Pittsburgh, PA" (in SEC)

Successful because there is only one PNC Bank on each side

FEIII Challenge Year 1: Results and Conclusion

- Our submitted results:
  - FFIEC → SEC: **261** matches, precision: 92.82%, recall: 84.35%, F1 score: 88.38% -- highest submitted
  - FFIEC → LEI: **480** matches, precision: 99.14%, recall: 92.54%, F1 score: 95.72% (best: 97.44%)
  - We also submitted full results for LEI → SEC (larger set of matches: **4,325**).

  - We also submitted true negatives (optional for Tasks 1 and 2). To find FFIEC ids that are not in LEI:
    - Used a relaxed HIL algorithm to find an "envelope" containing "all" the plausible matches from FFIEC to LEI, by matching just on name.
    - Applied complementation to obtain a set of FFIEC ids that we are confident do not appear in LEI.

- Conclusion
  - By using HIL and STG, we could rapidly express highly accurate matching algorithms, even in a sparse context.
  - The high-level language also facilitated the rapid modification of the code, in order to transfer the logic from one task to another.

Ongoing work

- Data Science for Macro-Modeling (DSMM) Workshops (http://dsmmworkshop.org)
  - Co-located with ACM SIGMOD
  - 3rd iteration @ SIGMOD 2017

- FEIII Data Challenge (https://ir.nist.gov/dsfin/): Provide interesting datasets to researchers at the intersection of finance and big data

- FEIII Data Challenges: Specific evaluation tasks with provided "raw" datasets
  - Year 1: Identifier alignment
  - Year 2: Identifying and understanding financial entities relationships