HPCC Systems Machine Learning Framework: Scalable and Extensible ML Platform for Big Data

Dr. Edin Muharemagic and Jo Prichard, Data Scientists, LexisNexis

ICMLA 2012, Boca Raton, FL Dec 13, 2012

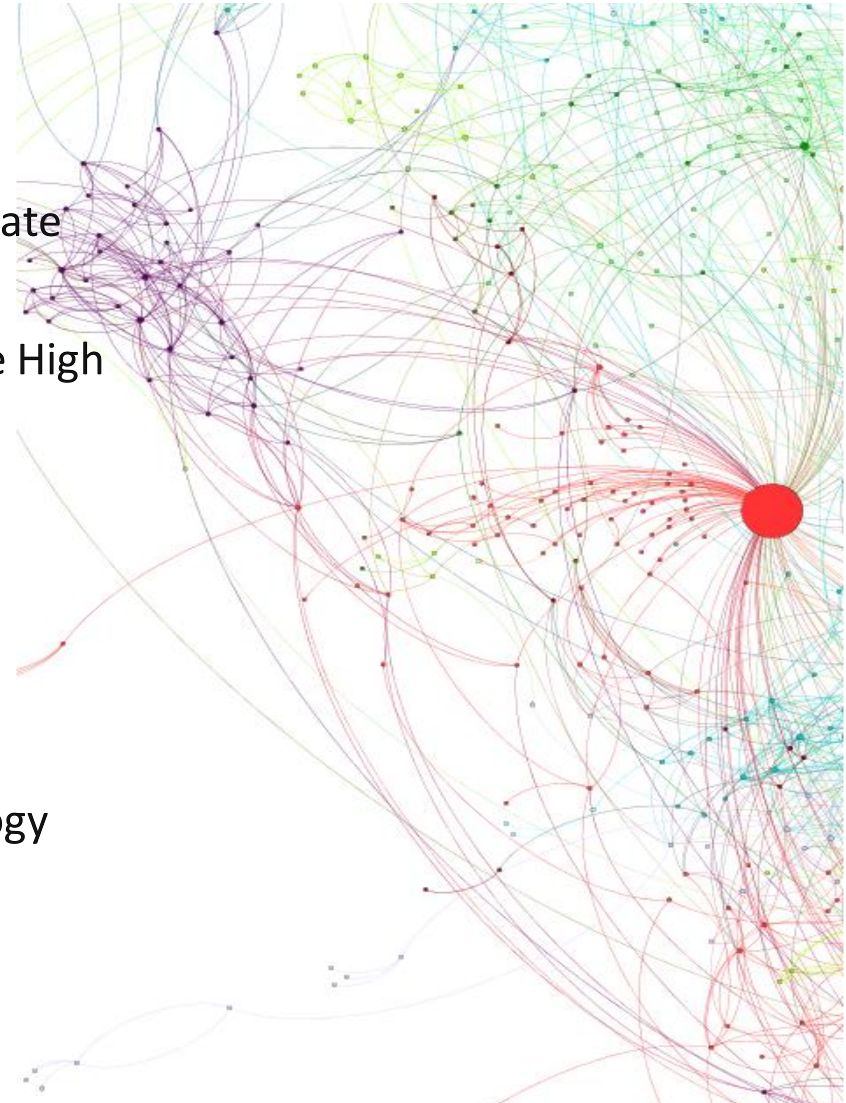
LexisNexis

HPCC Systems
http://hpccsystems.com

Risk Solutions

Overview

- LexisNexis Risk Solutions
- Public Data Social Graph
- Key ingredients for a successful fraud syndicate
- Mortgage Fraud Example Walkthrough
- HPCC Systems – Open Source Data Intensive High Performance Supercomputer
- Machine Learning on HPCC: ECL-ML
- LexisNexis – FAU Cooperative Research: ML Extensions and Applications
- Kettle: GUI to ECL-ML
- HPCC-PaperBoat ML
- SALT – Scalable Automated Linking Technology
- Kettle: GUI to SALT
- KEL – Knowledge Engineering Language
- Raffle and Giveaways



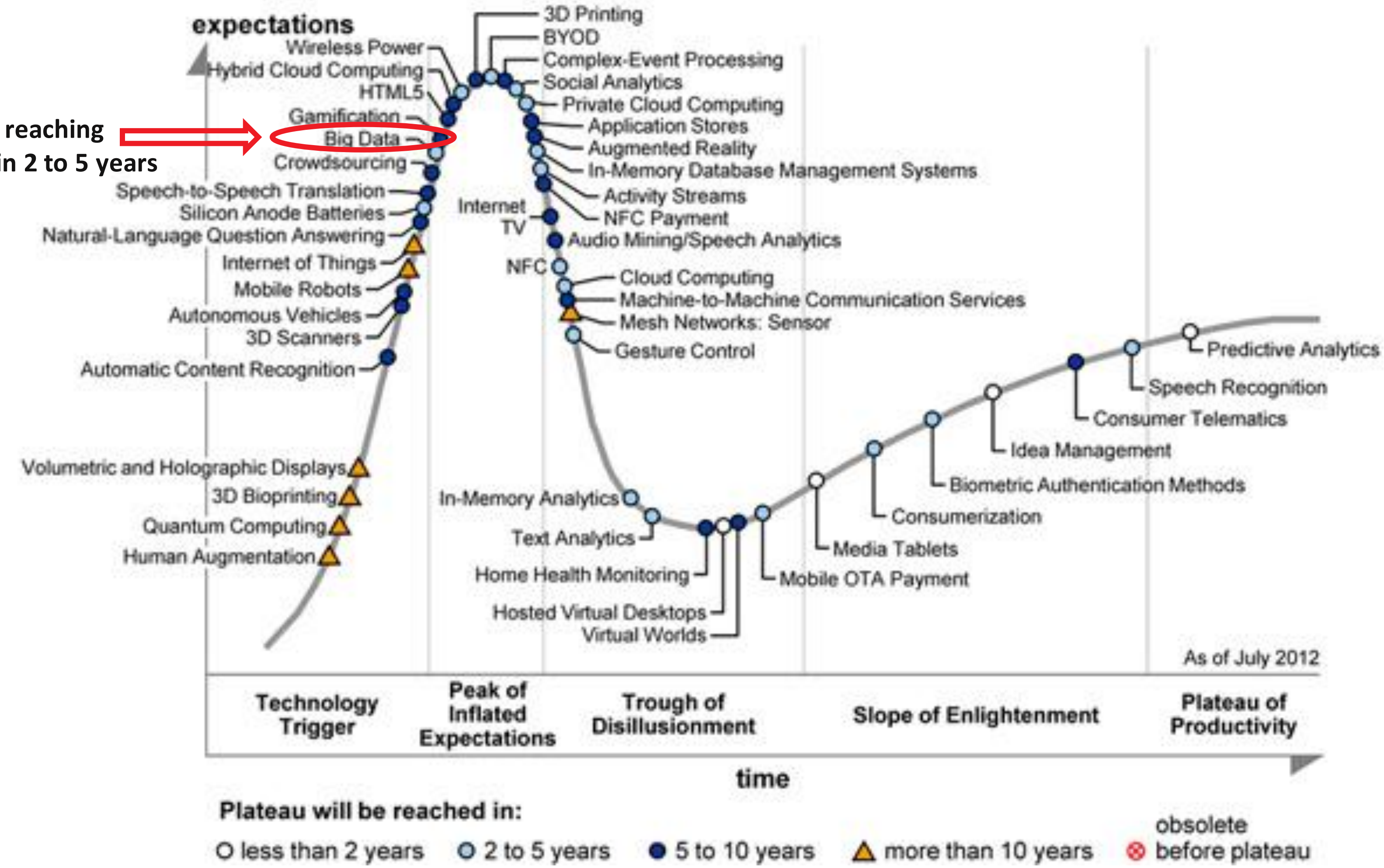
(Big) Data-Driven World

- Science Paradigms (Jim Gray, Microsoft)
  - Empirical – Describing natural phenomena
  - Theoretical – Using models, generalizations
  - Computational – Simulating complex phenomena
  - **Data Exploration – Data Intensive Scientific Discovery**
- NSF Big Data Solicitation: "The Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)"

Big Data in Gartner's Hype Cycle

HPCC Systems
http://hpccsystems.com

Big Data reaching
Plateau in 2 to 5 years



Source: Gartner

We leverage data about people, businesses, and assets to assess risk and opportunity associated with industry-specific problems

Identity Authentication and Risk Assessment

Contributory Data	Public Records	Credit Headers	Phone Listings

Who are you?

Where are you?

How much risk is associated with you?

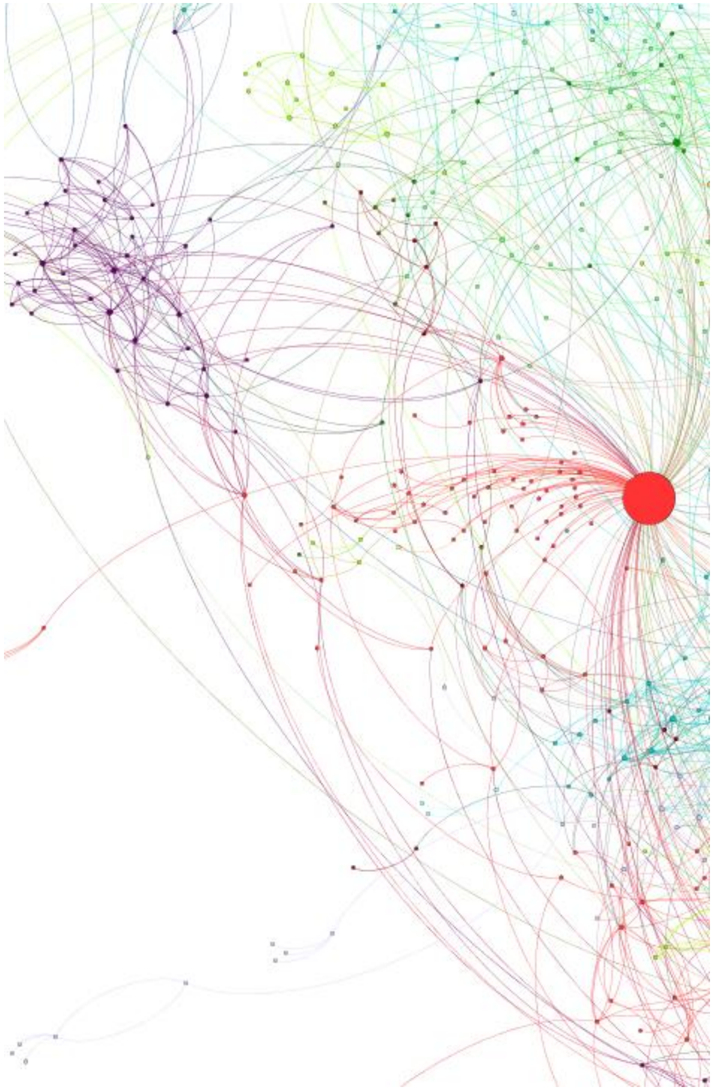
Insurance	Assess underwriting risk and verify applicant data; prevent/ investigate fraudulent claims; optimize policy administration.
Financial Services	Prevent/investigate money laundering and comply with laws. Prevent/investigate identity fraud .
Background Screening	Verify applicant identity and credentials. Avoid legal exposure by identifying criminal records and drug use.
Receivables Management	Assist collections by locating delinquent debtors. Assess collectability of debts and prioritize collection efforts.
Legal	Locate/vet witnesses, assess assets/associations of parties in legal actions; Perform diligence on prospective clients (KYC).
Government	Locate missing children/suspects; research/document cases; reduce entitlement fraud; accelerate revenue collection.
Health Care	Verify patient identity, eligibility, and ability to pay. Validate provider credentials; prevent/investigate fraud.

LexisNexis works with Fortune 1000 and mid-market clients across all industries and both federal and state governments

Our Customers

- **100%** of U.S. P&C insurance carriers
- **100%** of the top 50 U.S. banks
- **90%** of the Fortune 500 companies
- **All 50** U.S. states 70% of local governments, 80% of U.S. Federal agencies
- **97** of AM Law 100 firms

LexisNexis Public Data Social Graph



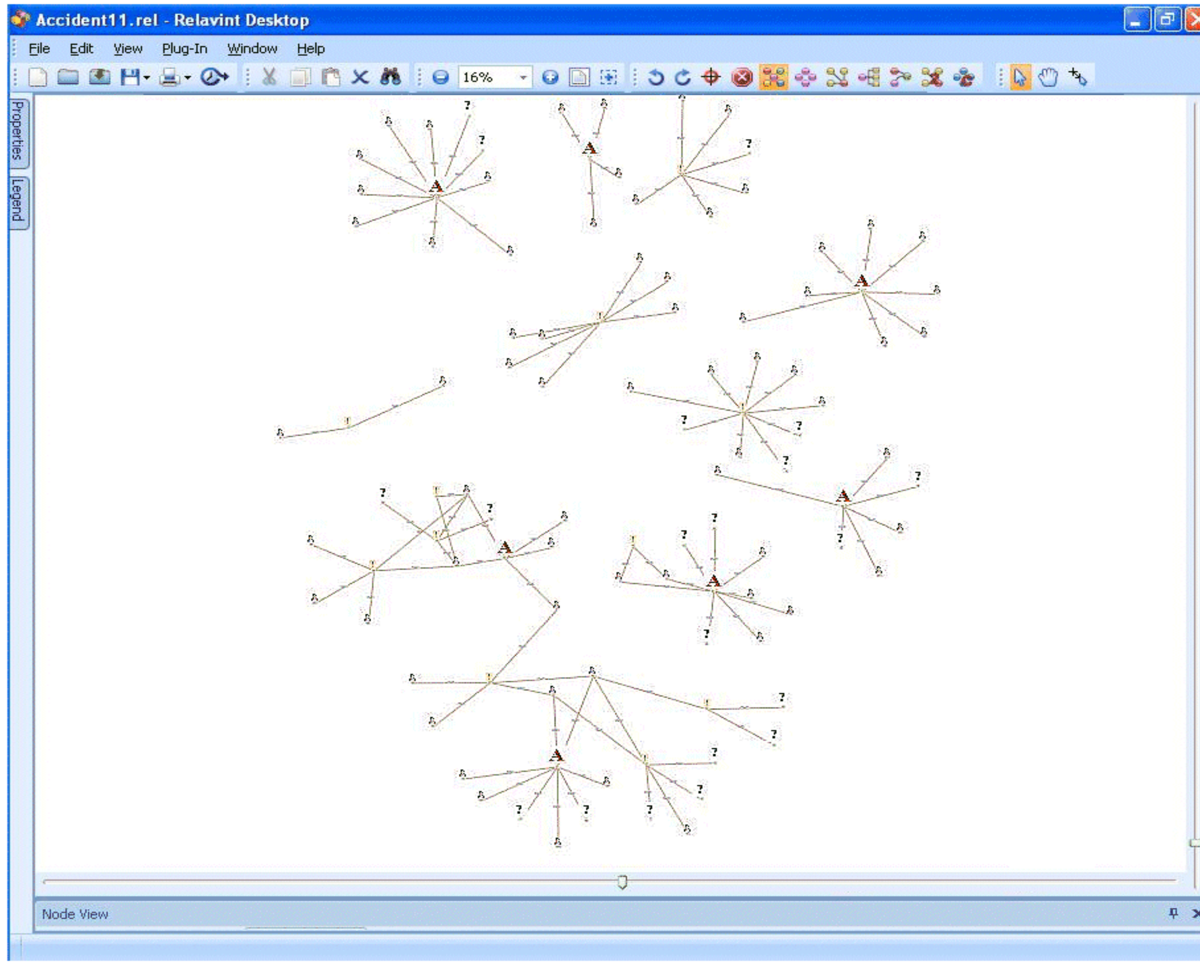
- **LexisNexis Public Data Social Graph (PDSG)**
  - Public Data relationships inferred from 50 TB of public data.
  - People connected to people, assets, businesses,…
  - High Value relationships for Mapping trusted networks.
- **Large Scale Data Fabrication and Analytics**.
  - Thousands of data sources to ingest, clean, aggregate and link.
  - 270 million people, 4 billion relationships, 700 million deeds.
  - 140 billion intermediate data points when running collusion analysis on property transfers.
- **HPCC Systems from LexisNexis Risk Solutions**
  - Open Source Data Intensive high performance supercomputer. (http://hpccsystems.com)
- **Innovative Examples leveraging the LexisNexis PDSG**
  - Healthcare.
    - Medicaid\Medicare Fraud.
    - Drug Seeking Behavior (excessive access to controlled or expensive drugs)
  - Financial Services.
    - Mortgage Fraud.
    - Money Laundering.
    - “Bust out” Fraud (open CC get $$ and vanish)
  - Insurance.
    - Staged Accident Fraud.

Real Life Graph Analytics

Scenario

This view of carrier data shows seven known fraud claims and an additional linked claim.

The Insurance company data **only finds a connection between two of the seven claims**, and only identified one other claim as being weakly connected.



Real Life Graph Analytics

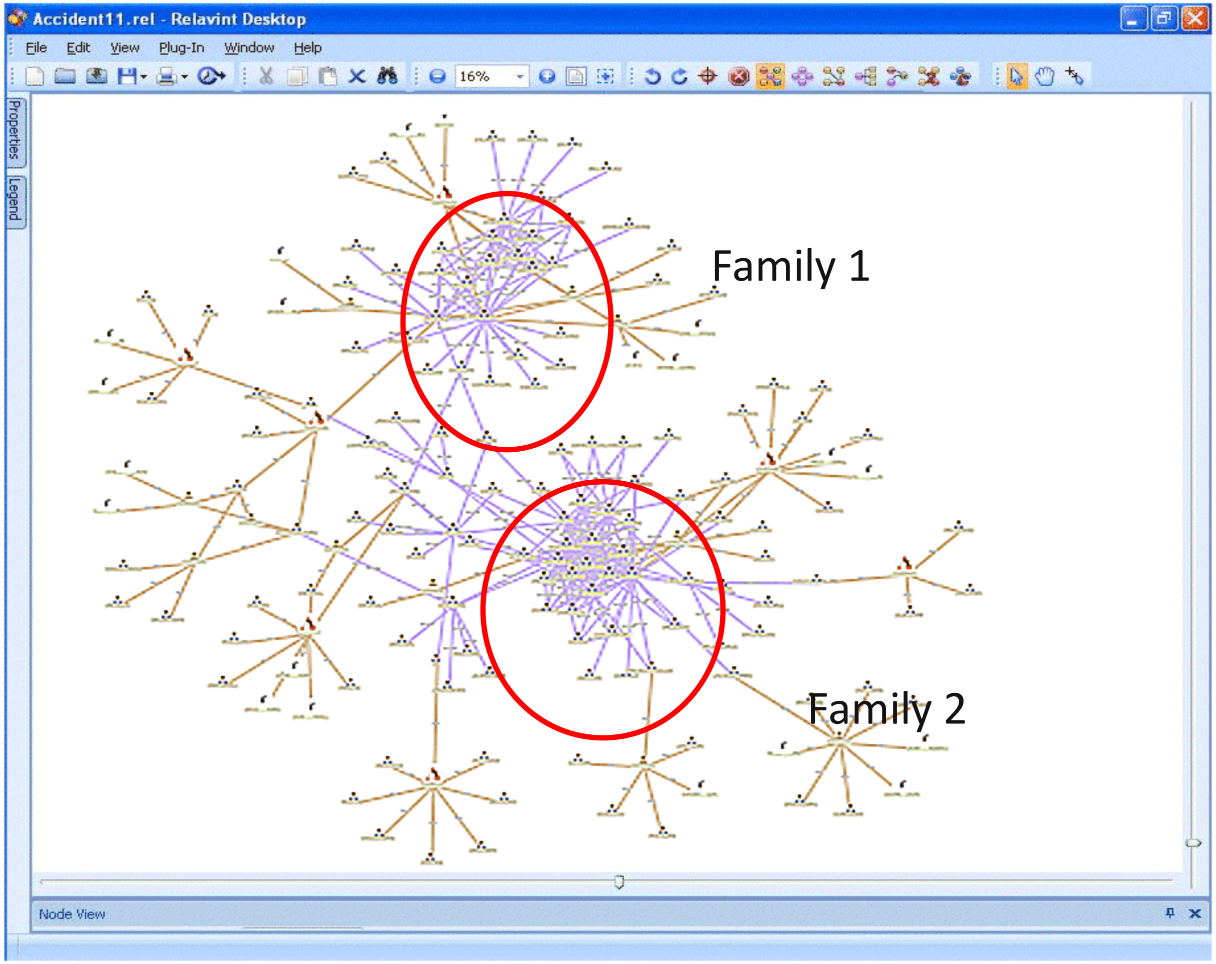
Task

After adding the LexID to the carrier Data, LexisNexis HPCC technology then explored 2 additional degrees of relative separation

Result

The results showed **two family groups interconnected on all of these seven claims**.

The links were much stronger than the carrier data previously supported.



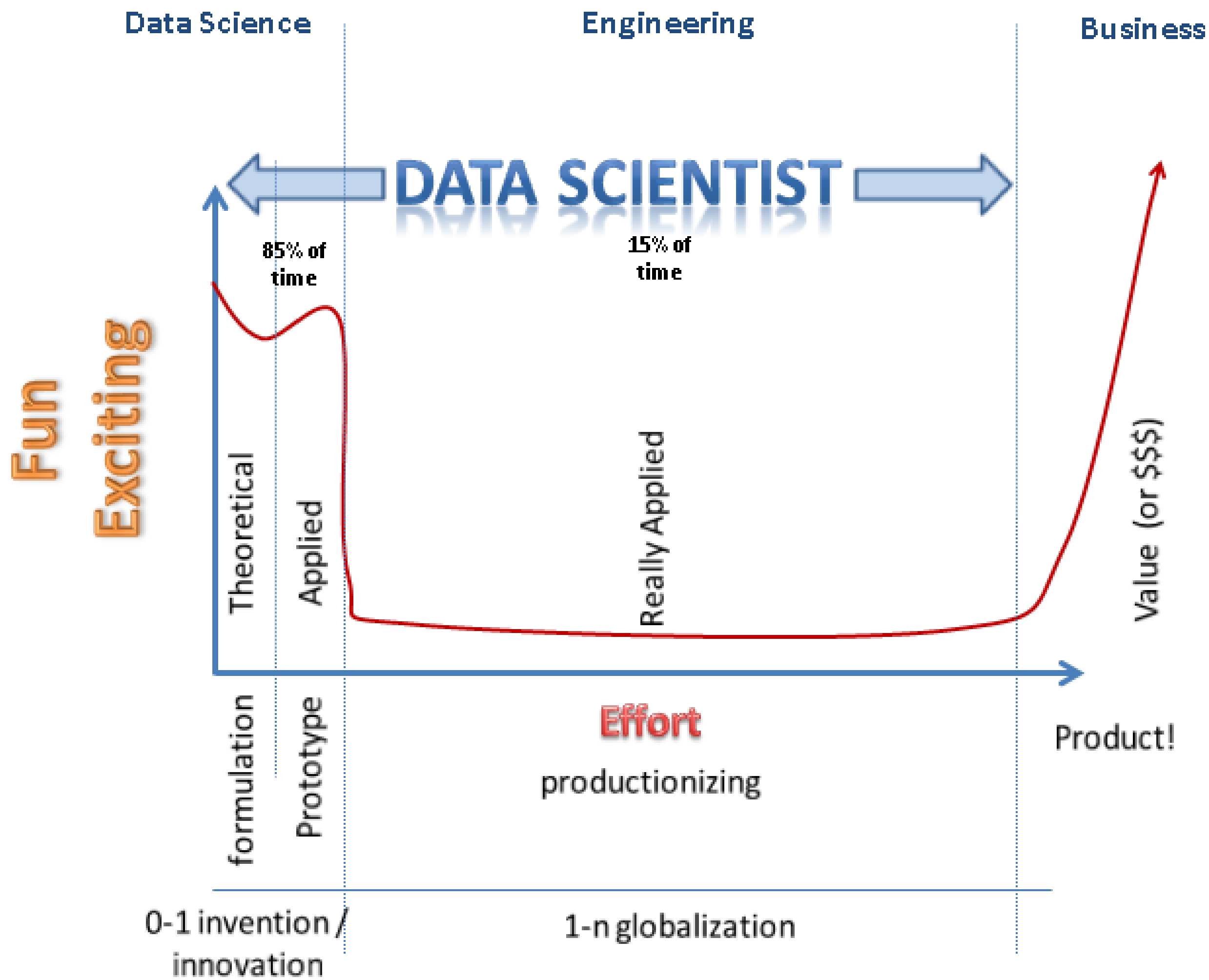
The important thing is to not stop questioning…

"The mere formulation of a problem is far more essential than its solution, which may be merely a matter of mathematical or experimental skills. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science."

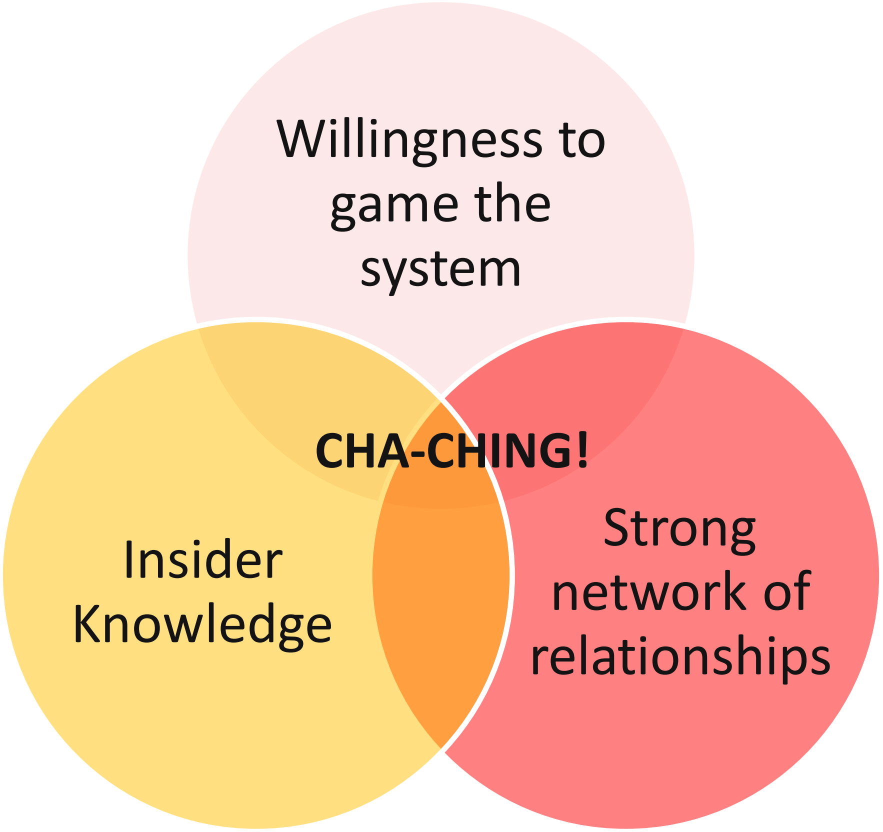
"The important thing is to not stop questioning. Curiosity has its own reason for existing.“

- Albert Einstein

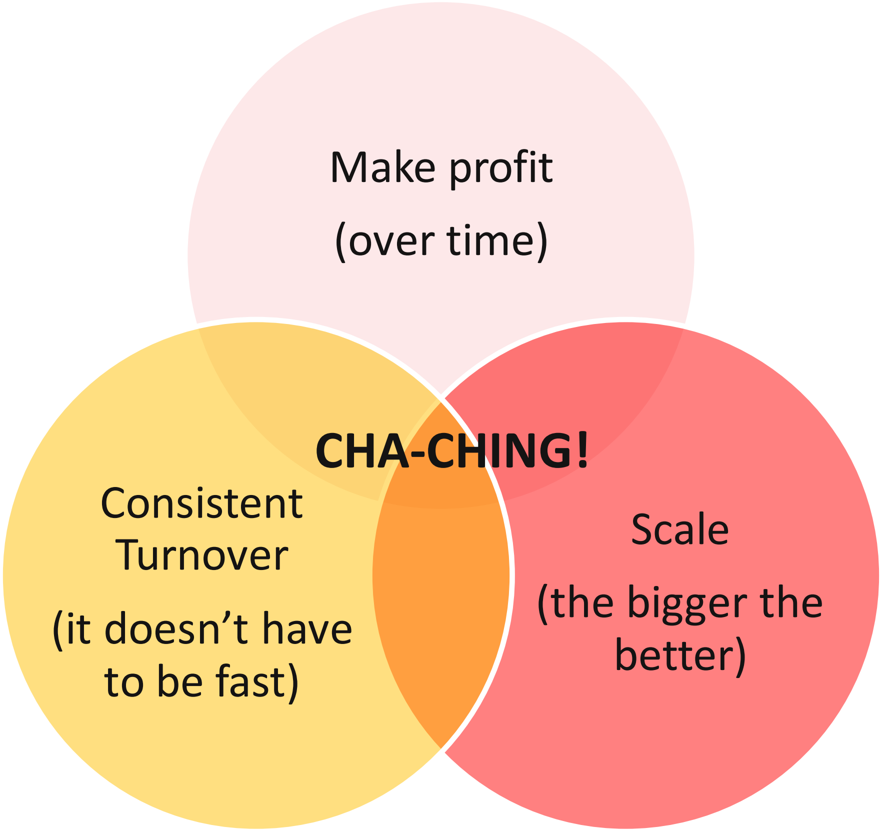
A Note about Data Science



Key ingredients for a successful large scale fraud syndicate



Key ingredients for a successful large scale fraud syndicate
Real Estate Strategy



Perfect borrower… John Smith's profile

HPCC Systems
http://hpccsystems.com

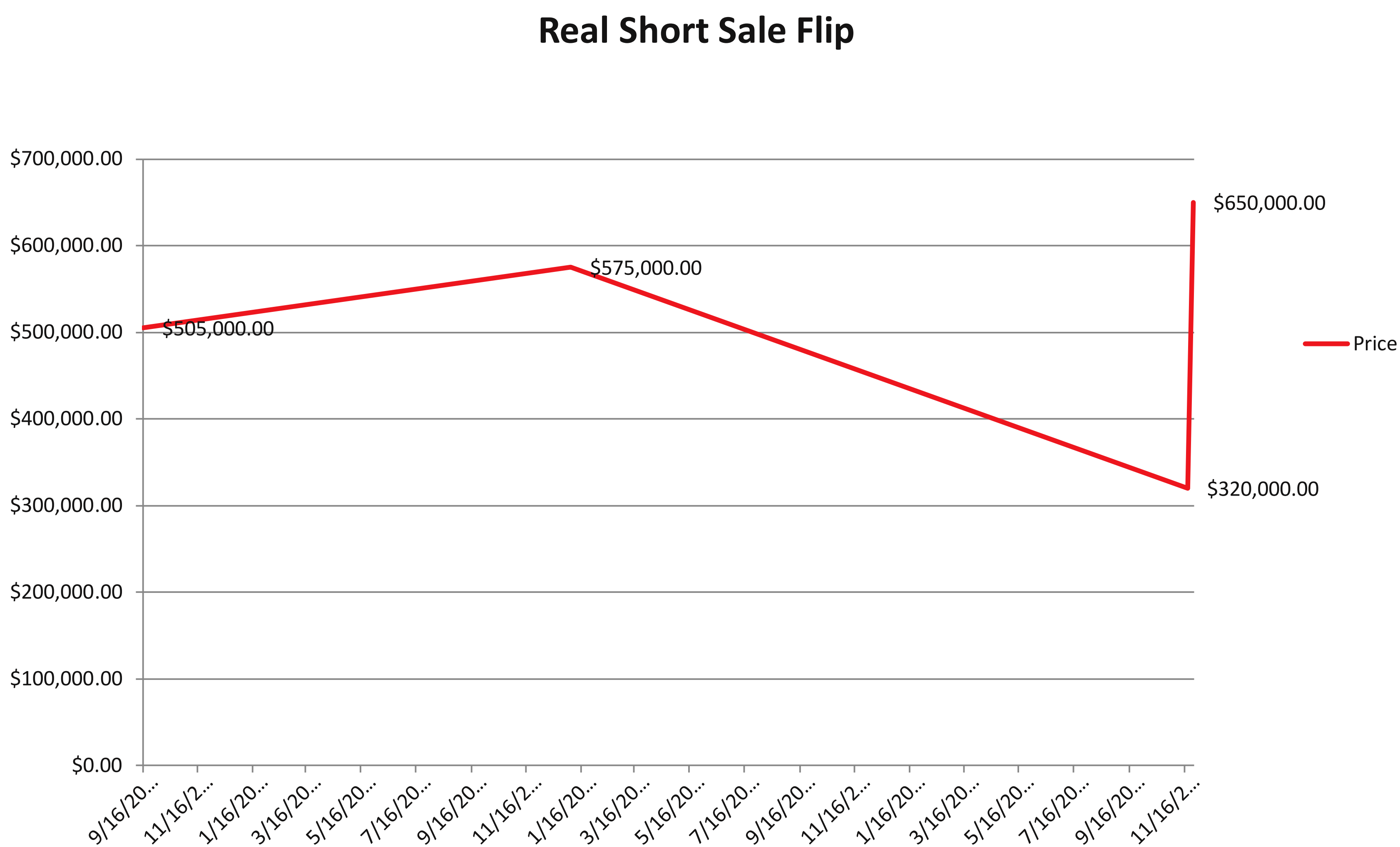
Credit Score	=	785
Credit History	=	No late payments
Homes Owned	=	10 Homes
Employment	=	10 years with ABC
Income	=	$150,000 / year
Cash / Assets	=	$1,000,000
LTV	=	avg. 75% or less
Doc Type	=	Full Doc
DTI	=	avg. 18%
Debt	=	$400 a month

A typical example of mortgage fraud

HPCC Systems
http://hpccsystems.com

Ms. Gordon		Mr. Michael		Mr. Smith		Mr. James
Bought Sep 2009	90 days	Bought Jan 2010	330 days	Bought Nov 2010	7 days	Bought Nov 2010
Buys: $505K	Flip	Buys: $575K	Short Sale	Buys: $320K	Flop	Buys: $650K
	Gain: +$70K		Loss: ($255K)		Gain: +$330K	

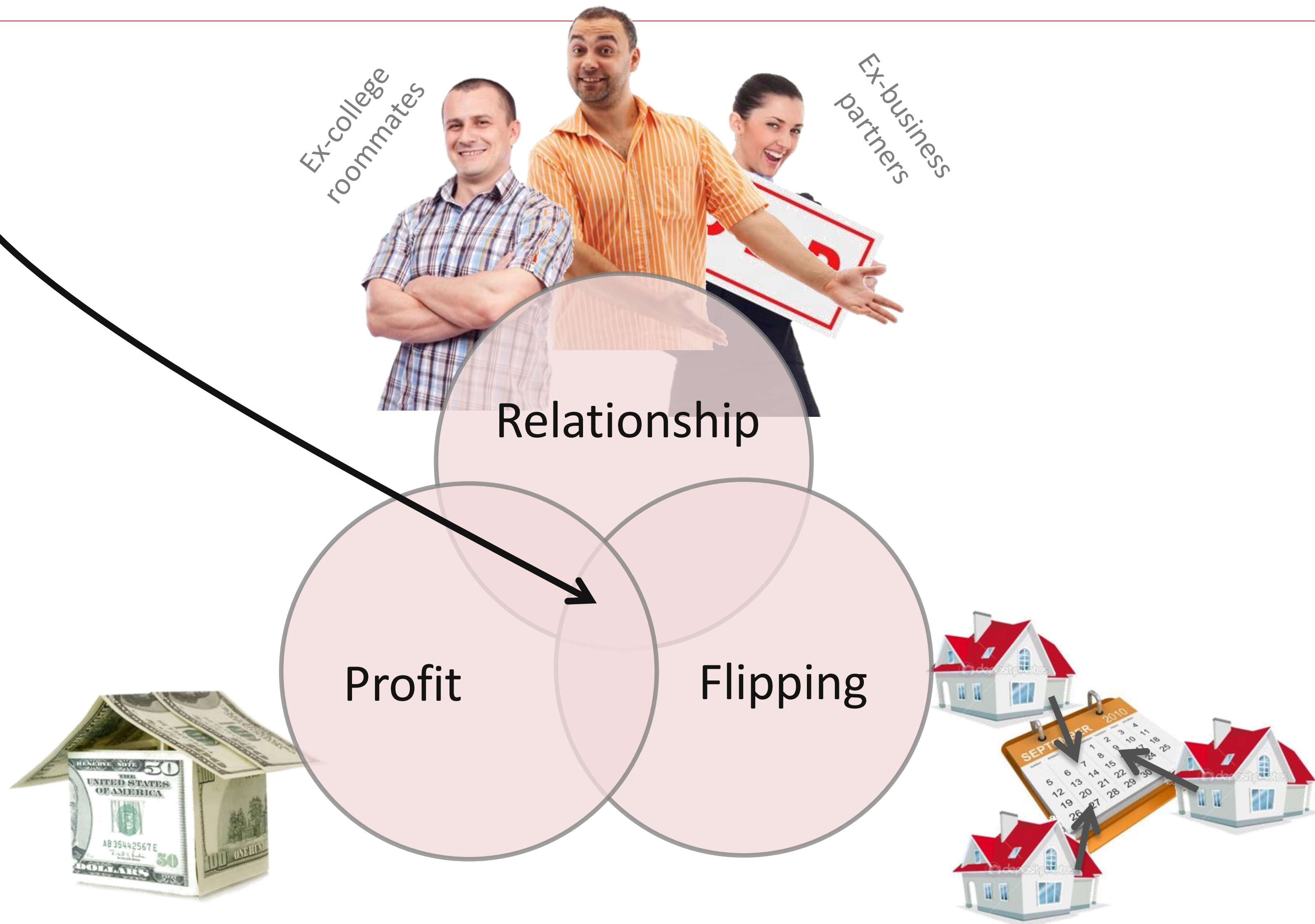
A typical example of mortgage fraud



Get rich quick

	Ring Leader	Business Partner	College Roommate
$ Split ($330K)	$150,000	$30,000	$150,000
Commission	0	$58,200	0
**Rental **Currently renting the home.	$15,000	0	$15,000
Total	**$165,000**	**$88,200**	**$165,000**

Hidden

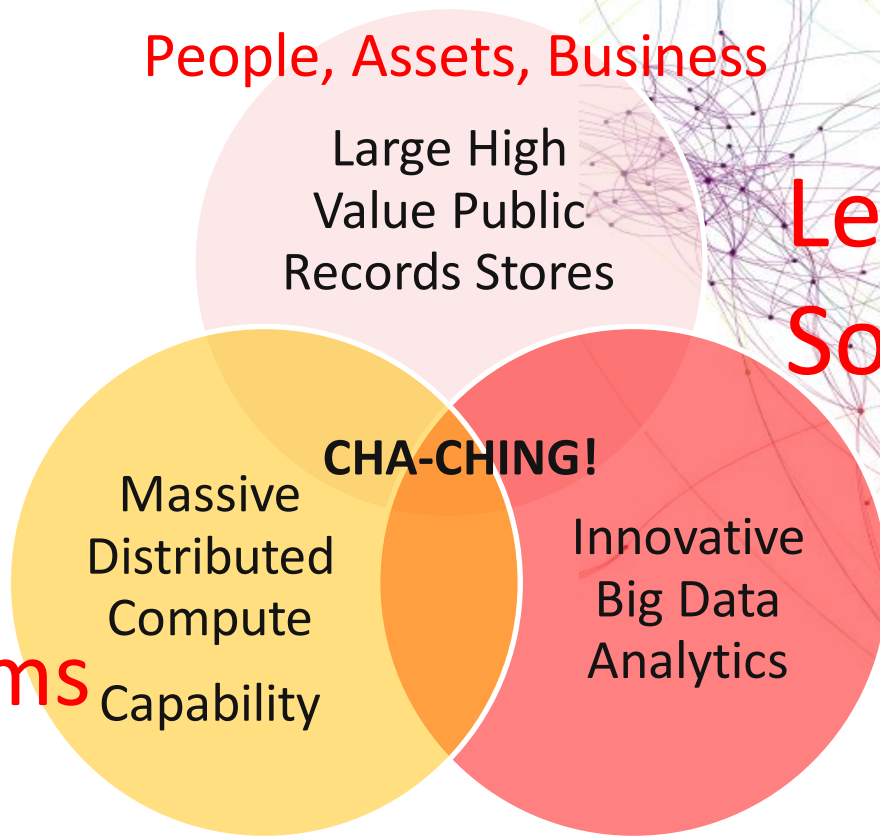

Ex-college roommates
Ex-business partners
Relationship
Profit
Flipping

Invisible Sharks and the advantages of networking fraud

Fraudsters know all the thresholds and game the system.

- Rules based detection creates misdirection.
  - Designed to detect the big fish.
- Organized social networks fly under the radar.
  - Suspicious data points are sparse and financial institutions only have small slices of activity.
  - Great granularity but no big picture.
- Key to solving the puzzle is BIG DATA
  - BIG DATA to stitch a more complete picture together.
  - Measure the social context of seemingly low risk events within social groups.

HPCC Systems
http://hpccsystems.com



LexId
Social Graph

HPCC Systems
Hpccsystems.com

Visualization of a suspicious seller network

HPCC Systems
http://hpccsystems.com

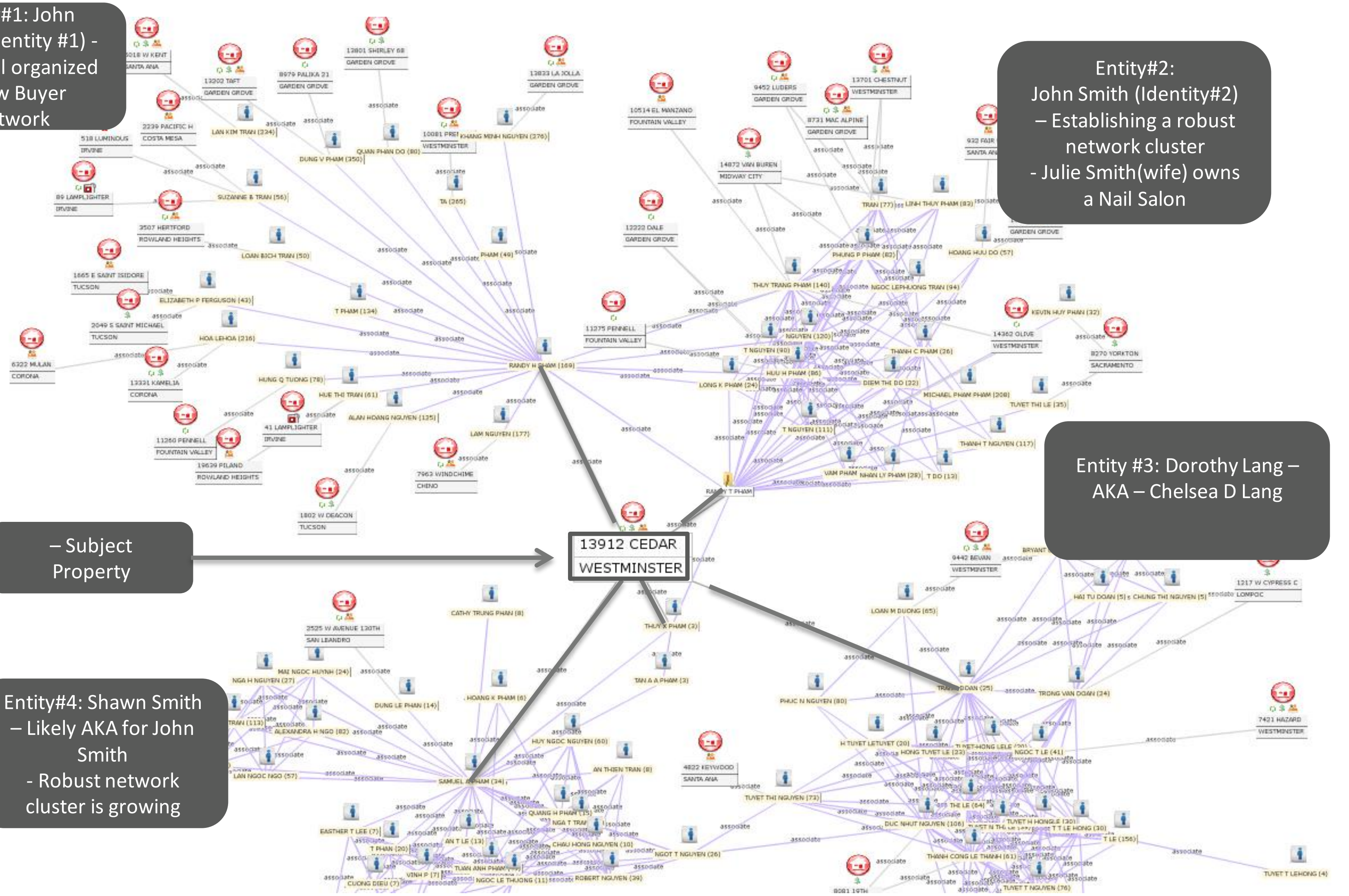
Entity#1: John Smith(Identity #1) - Potential organized Straw Buyer network

Entity#2:
John Smith (Identity#2) – Establishing a robust network cluster
- Julie Smith(wife) owns a Nail Salon

Entity #3: Dorothy Lang – AKA – Chelsea D Lang

– Subject Property

Entity#4: Shawn Smith – Likely AKA for John Smith
- Robust network cluster is growing



Variables to describe context temporally and socially.

HPCC Systems
http://hpccsystems.com

- Network and cluster high risk behavior and rank score clusters (most egregious at the top)

Cluster with the highest number of sales between known associates.

Flips within 120 days between known associates.

Shows the network is well connected to each other

Shows that the flipping is spread evenly over the actors.

##	cluster id	sales count	flip count	in network count	in network flip c...	in network high profit	prop network cohesivity	prop 1st degrees	prop 2nd degrees	sales count stdd	flip count actors	flip_count_stdd
231	9835802	189	15	45	5	7	1.537634	21	30	3.730517	12	0.690353
3...	1044786435	178	13	24	4	2	1.727193	10	38	3.951771	11	0.647936
28	670840426	116	12	18	3	1	1.814815	3	23	5.058777	9	0.831479
1	161256572	93	10	10	2	0	1.892857	1	12	5.899239	7	0.914732
10	32629685	79	10	10	2	0	1.916667	1	11	6.277716	7	0.912871
1...	17342174	82	10	8	2	0	1.850000	1	8	5.899152	7	0.871780
1...	1033860591	79	10	10	2	0	1.500000	2	2	6.277716	7	0.912871
1...	948662502	149	15	19	2	1	1.833333	4	33	4.345076	11	0.744032
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

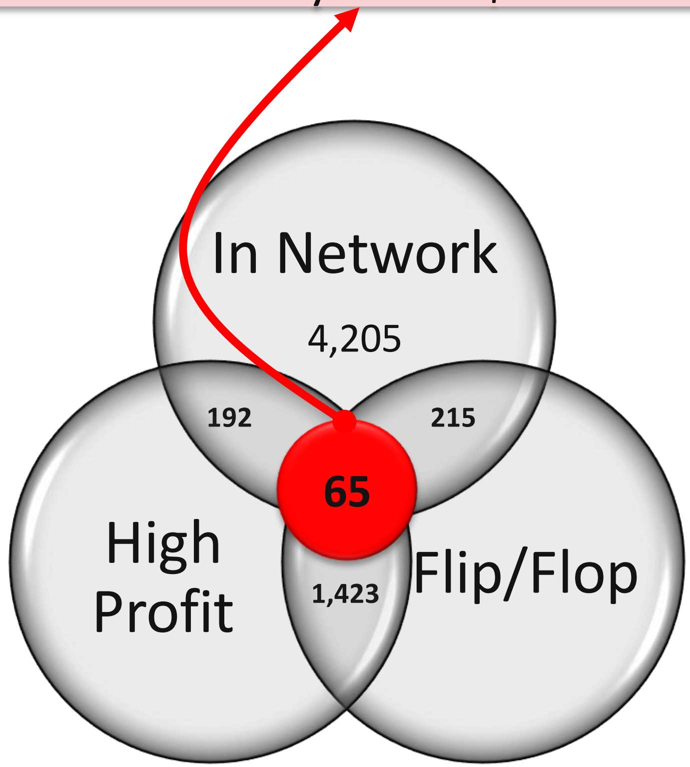
Person at the center of the cluster.

9835802

- Highest risk cluster does NOT include any of the high velocity flippers.
- Risk behavior is evenly spread across the actors which inhibits detection (traditionally).
- Person at the center of the cluster is best connected yet personally not listed on any of the deeds of flipped properties.

We found the needle in the haystack

LexisNexis analyzed 150,000 loans for a large mortgage originator and found the 65 most risky files. >$17 million at risk!



High Velocity Streets - Local San Francisco Interest

HPCC Systems
http://hpccsystems.com

prim_name	suffix	street_flip_count	street_in_network_flip_count	street_high_profit_flip_count	street_high_profit_in_network_flip
14TH	ST	83	5	4	2
22ND	ST	204	2	8	1
CHICAGO	WAY	50	1	3	1
HARRIET	ST	36	1	2	1
HOLLOWAY	AVE	71	5	1	1
LYON	ST	155	9	6	1
04TH ST 404		0	0	0	0
04TH ST 909		0	0	0	0
07TH ST 351		0	0	0	0
08 10&12 DIAMOND	ST	0	0	0	0
08TH AVE 3		0	0	0	0

Medicaid Fraud

HPCC Systems
http://hpccsystems.com

Scenario

Proof of concept for Office of the Medicaid Inspector Generation (OMIG) of large Northeastern state. Social groups game the Medicaid system which results in fraud and improper payments.

Task

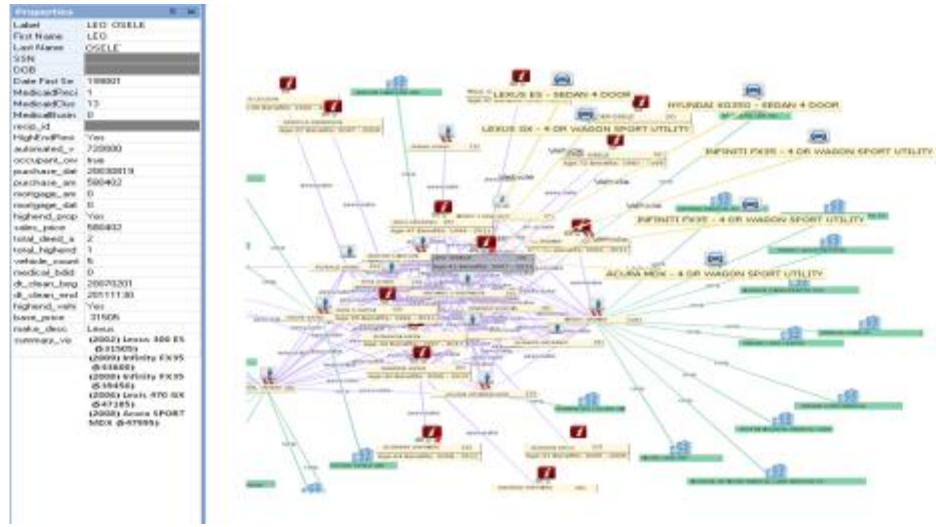
Given a large list of names and addresses, identify social clusters of Medicaid recipients living in expensive houses, driving expensive cars.

Result

Interesting recipients were identified using asset variables, revealing hundreds of high-end automobiles and properties.

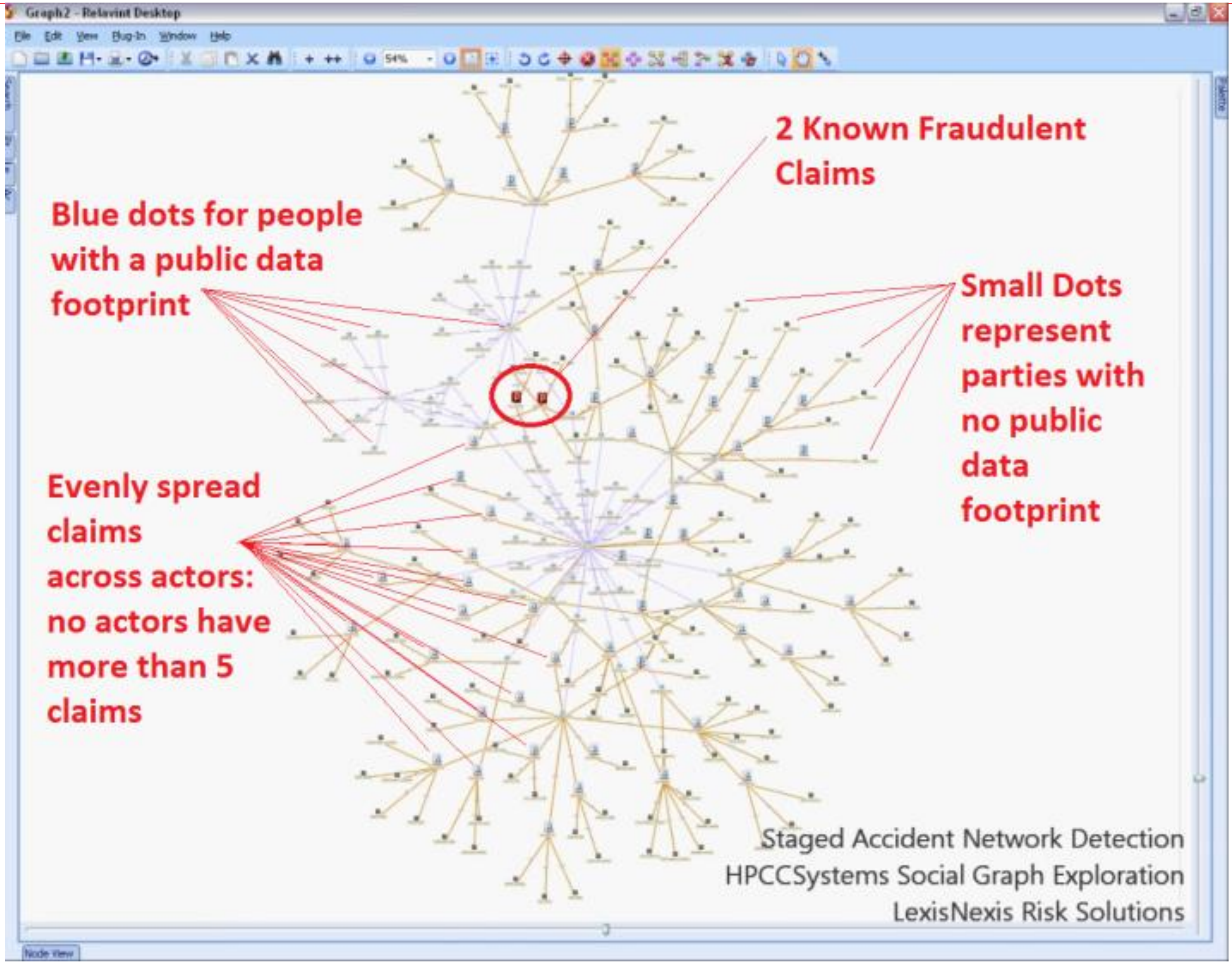
Leveraging the Public Data Social Graph, large social groups of interesting recipients were identified along with links to provider networks.

The analysis identified key individuals not in the data supplied along with connections to suspicious volumes of “property flipping” potentially indicative of mortgage fraud and money laundering



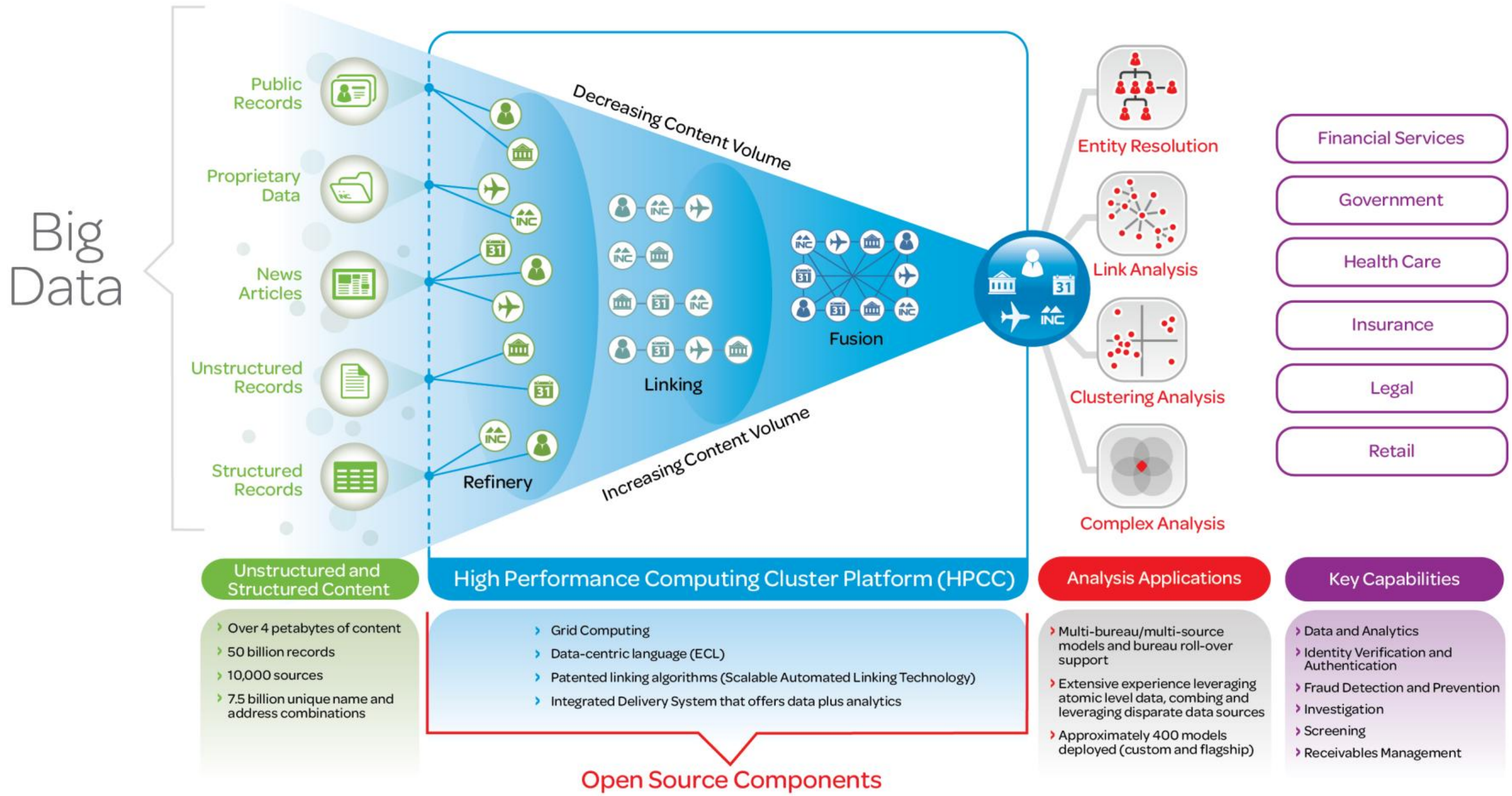
Make Description	#	Make Description	#
Mercedes-Benz	46	Chevrolet	2
Lexus	41	Hummer	2
BMW	27	Jeep	2
Infiniti	13	Nissan	2
Acura	9	Toyota	2
Lincoln	8	Aston Martin	1
Audi	7	Bentley	1
Land Rover	7	Cadillac	1
Porsche	6	GMC	1
Jaguar	5	Honda	1
Mercedes Benz	3	Volkswagen	1
Saab	3	Volvo	1

Staged Accident Network Detection



We **connect the dots** between millions of public records and transactions, resulting in actionable information our customers use to advance their goals

HPCC Systems
http://hpccsystems.com



- **High Performance Computing Cluster Platform (HPCC)** enables data integration on a scale not previously available and real-time answers to millions of users. Built for big data and proven for 10 years with enterprise customers.
- **Offers a single architecture**, two data platforms (query and refinery) and a consistent data-intensive programming language (ECL)
- **ECL Parallel Programming Language** optimized for business differentiating data intensive applications

The secret sauce in our portfolio is LexID℠

LexID℠

The fastest linking technology platform available with results that help you make intelligent information connections.

LexID℠ is the ingredient behind our products that turns disparate information into meaningful insights. This technology enables customers using our products to identify, link and organize information quickly with a high degree of accuracy.

Get a more complete picture

Make intelligent information connections beyond the obvious by drawing insights from both traditional and new sources of data.

Better results, faster

Use the fastest technology for processing large amounts of data to help you solve cases more quickly and confidently.

Protect private information

Keep customer SSNs and FEINs secure and enjoy peace of mind knowing you are taking steps to observe the highest levels of privacy and compliance.

Questions / Contact Info

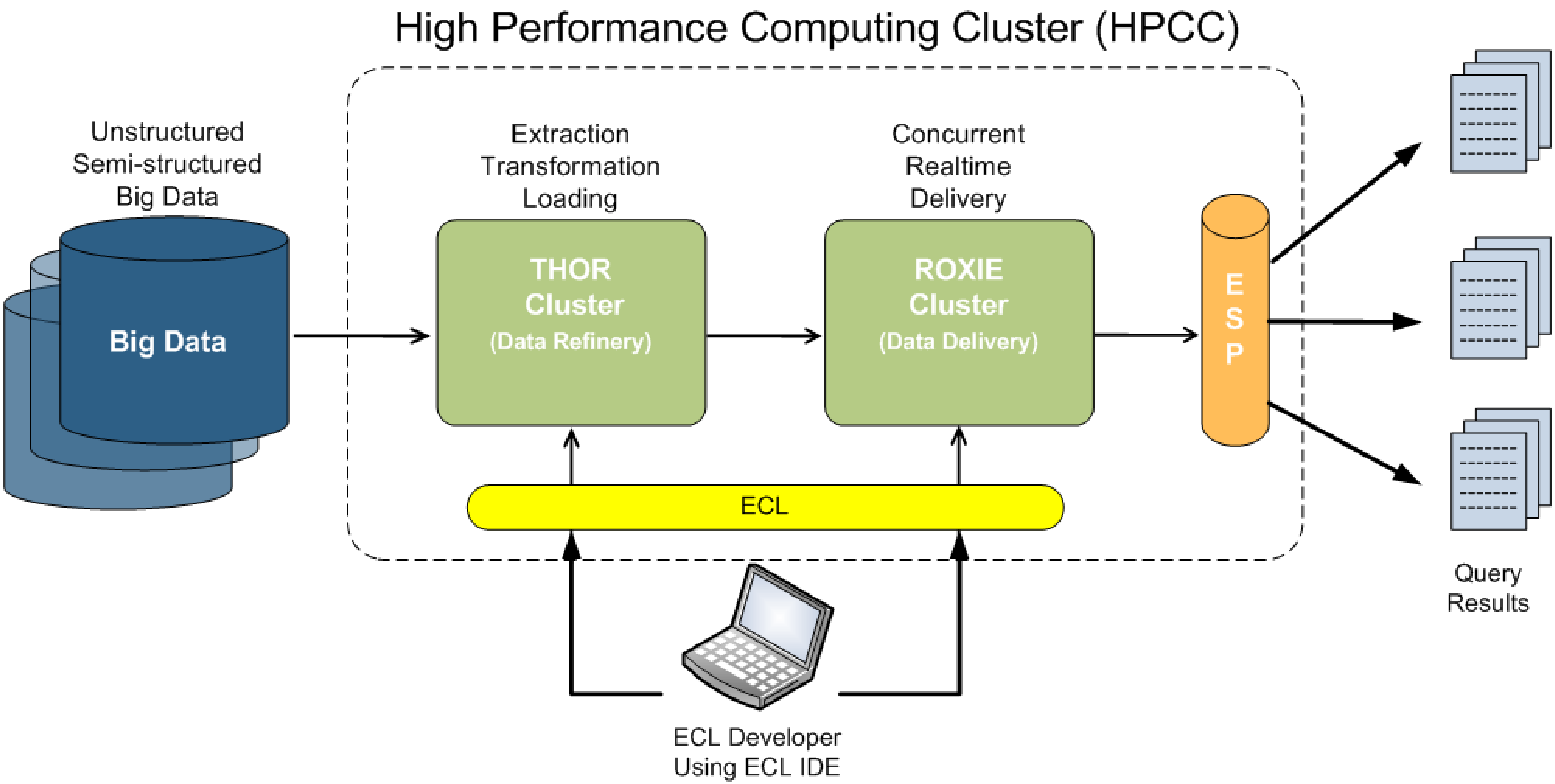
Any Questions?

Contact Jo Prichard:
Twitter: @joprichard
Email: Jo.prichard@lexisnexis.com

http://hpccsystems.com

The HPCC Systems platform

HPCC Systems
http://hpccsystems.com



Enterprise Control Language (ECL)

HPCC Systems
http://hpccsystems.com

Declarative programming language: Describe **what** needs to be done and **not how** to do it

Powerful: Unlike Java, high level primitives as JOIN, TRANSFORM, PROJECT, SORT, DISTRIBUTE, MAP, etc. are available. Higher level code means fewer programmers & shortens time to delivery

Extensible: As new attributes are defined, they become primitives that other programmers can use

Implicitly parallel: Parallelism is built into the underlying platform. The programmer needs not be concerned with it

Maintainable: A high level programming language, no side effects and attribute encapsulation provide for more succinct, reliable and easier to troubleshoot code

Complete: ECL provides for a complete data programming paradigm

Homogeneous: One language to express data algorithms across the entire HPCC platform, including data ETL and high speed data delivery

```
// Initialize output log
log_out_init := project(log_init,
                        transform(layout_logo
                        self := left,
                   self := []));

// Create error log
outerrorfile := join(log_seq,
                     log_out_init,
            left.linenum = right.linenum,
            transform(recordof(log_seq),
                     self := left),
            left only,
            hash);

// Denormalize key value pairs
outlogfile := sort(denormalize(distribute(log_
                         sort(distribute(key
              left.linenum = right.linenum,
              transform(layout_logout,
                        self.keyvals := left
                                row({ri
                   self := left),
```

Benefits

Speed

- Scales to extreme workloads quickly and easily
- Increase speed of development leads to faster production/delivery
- Improved developer productivity

Capacity

- Enables massive joins, merges, sorts and data transformations
- State of the art Big Data workflow management
- Increases business responsiveness
- Accelerates creation of new services via rapid prototyping capabilities
- Offers a platform for collaboration and innovation leading to better results

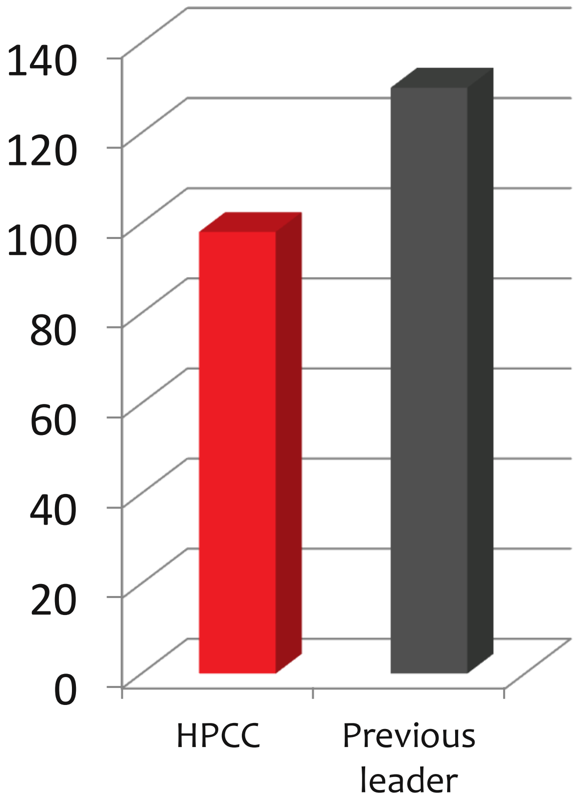
Cost Savings

- Commodity hardware and fewer people can do much more in less time
- Uses IT resources efficiently via sharing and higher system utilization

Terasort Benchmark results

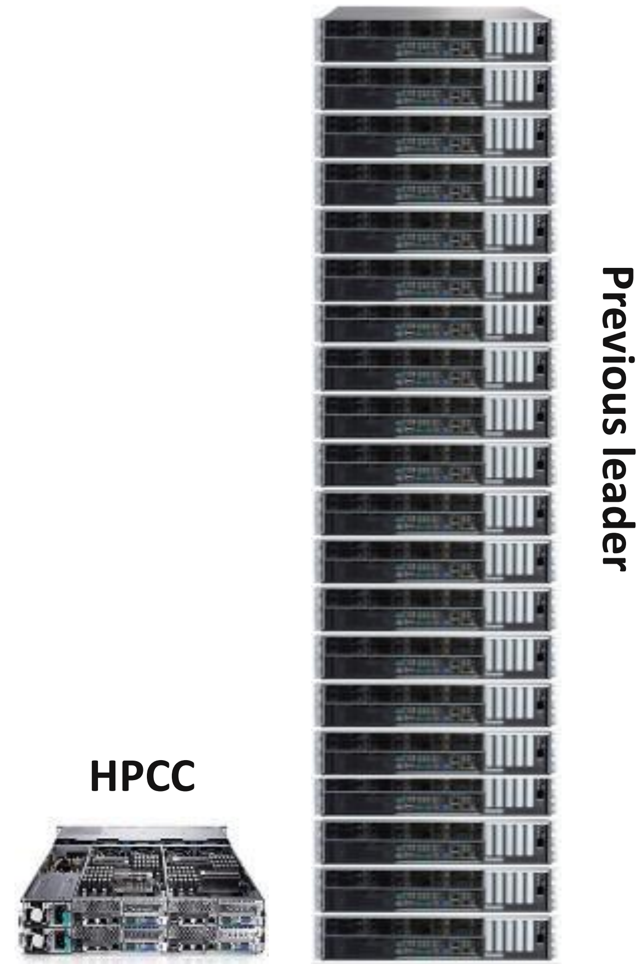
HPCC Systems
http://hpccsystems.com

Execution Time (seconds)



Productivity

Space/Cost



ECL-ML: Machine Learning on HPCC

General aspects

- Based on a distributed ECL linear algebra framework
- New algorithms can be quickly developed and implemented
- Common interface to classification (pluggable classifiers)

ML algorithms

- Linear regression
- Several Classifiers
- Multiple clustering methods
- Association analysis

Document manipulation and Natural Language Processing

- Tokenization
- CoLocation

Statistics

- General statistical methods
- Correlation
- Cardinality
- Ranking

ECL-ML: Extensible ML on HPCC (ii)

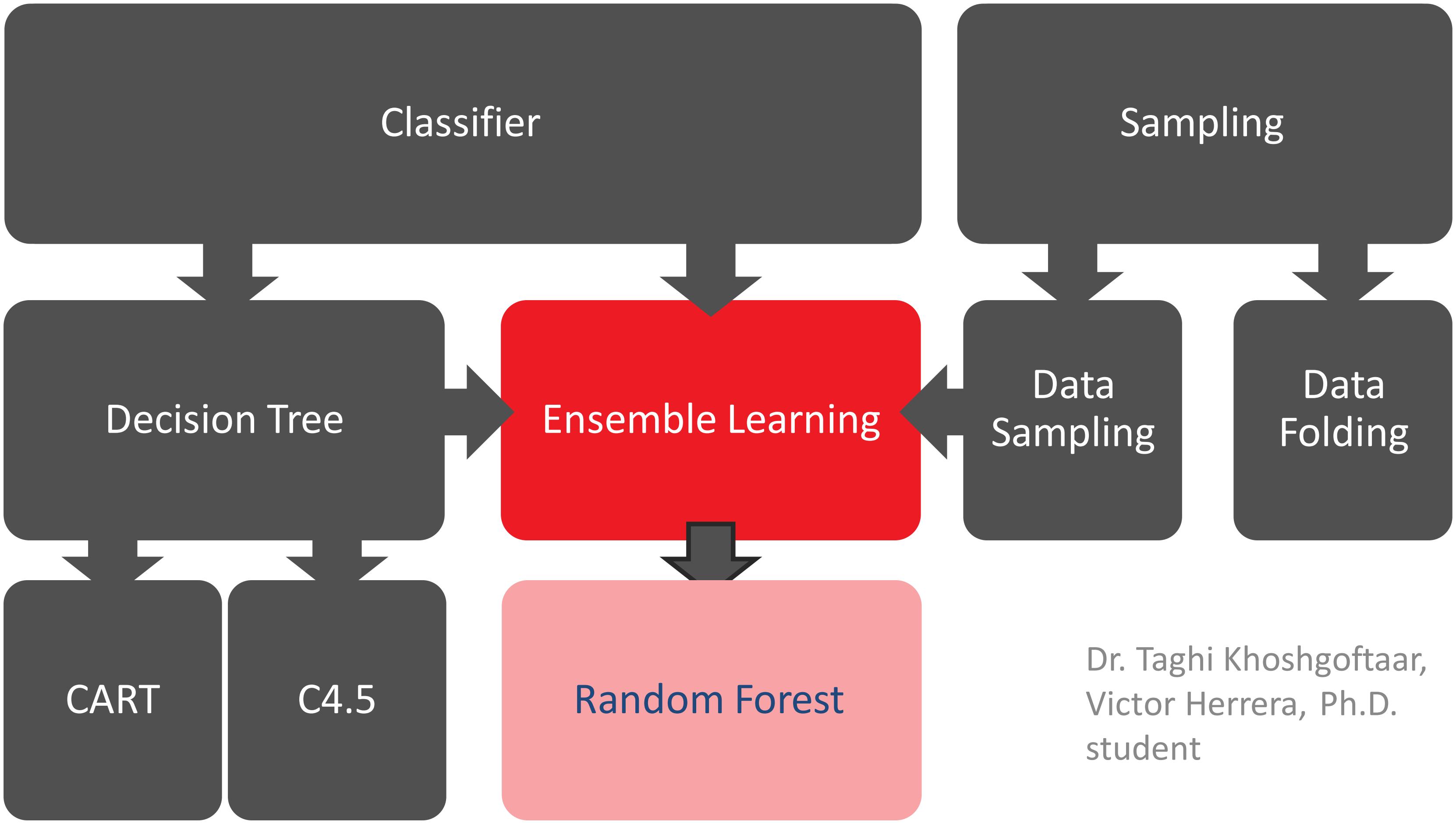
Linear Algebra library

- Support for sparse matrices
- Standard underlying matrix/vector data structures
- Basic operations (addition, products, transpositions)
- Determinant/inversions
- Factorization/SVD/Cholesky/Rank/LU
- PCA
- Eigenvectors/Eigenvalues
- Interpolation (Lanczos)
- Identity
- Covariance
- KD Trees

Open Source and available at: : http://hpccsystems.com/ml

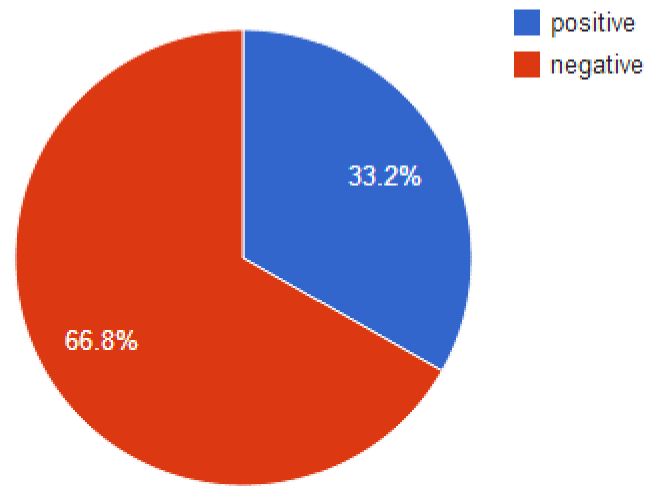
LN – FAU Cooperative Research
ECL-ML Extension

HPCC Systems
http://hpccsystems.com



Dr. Taghi Khoshgoftaar, Victor Herrera, Ph.D. student

LN – FAU Cooperative Research (ii)
Twitter Sentiment Analysis



Dataset: Result 1

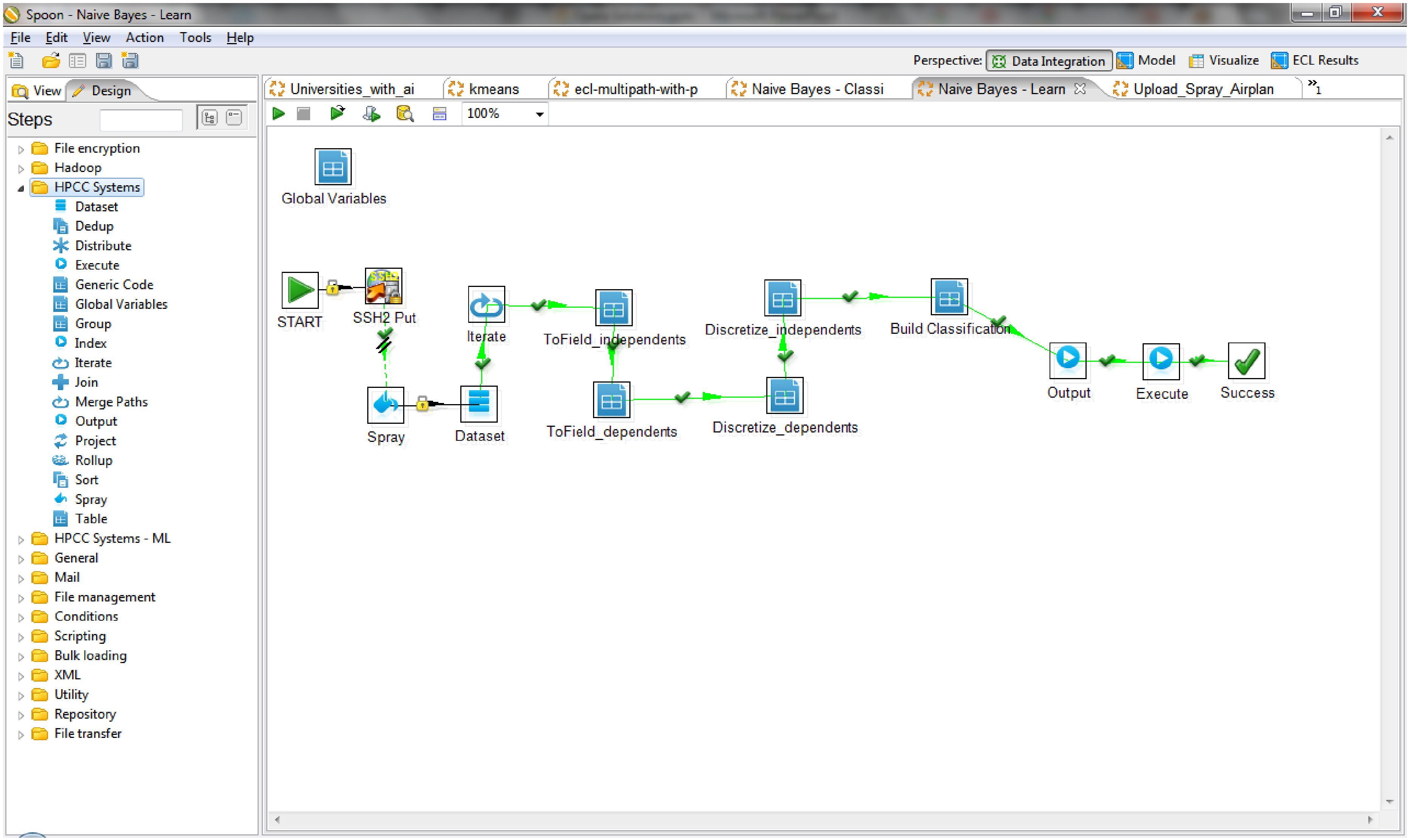
	id	date	tweet	sentiment
1	1	Mon Dec 10 23:47:51 +0000 2012	RT @ComedyPosts: ALL I WANT FOR CHRISTMAS IS YOOOOOUUUUU.... to fall off a cliff.	-1
2	2	Mon Dec 10 23:47:50 +0000 2012	If I hear one more thing about the \"fiscal cliff\", I'm going to lose it. Refuse to negotiate and blame it on the other guys. #whatsnew?	1
3	3	Mon Dec 10 23:47:50 +0000 2012	cliff is fired	-1
4	4	Mon Dec 10 23:47:49 +0000 2012	@ShawGahh ahaha wanna see the cliff we're studying.. This piece of shizz http://t.co/jvY2IvX0	-1
5	5	Mon Dec 10 23:47:48 +0000 2012	RT @ComedyPosts: ALL I WANT FOR CHRISTMAS IS YOOOOOUUUUU.... to fall off a cliff.	-1

Charlene Gilbert, MS Student

Twitter Sentiment Analysis

- Building Sentiment Analysis Classifier
  - Build Training Data
    - Harvest Tweets
    - Label tweets based on Emoticons
    - Spray onto Cluster
  - Train Bayes Classifier using data
    - Create Bayes Model: ML.Classify.NaiveBayes.LearnD(dfIndep,dfDep)
    - Create Vocabulary: ML.Docs.Tokenize.Split(ML.Docs.Tokenize.Clean(Tweets))
- Build Roxie Twitter Sentiment Analysis
  - Get 1500 recent tweets:
    - sPipeQuery := '/usr/local/bin/twitsearch "' + searchquery_value + '"';
    - PIPE(sPipeQuery,rPipe,CSV(SEPARATOR(''),TERMINATOR('$end\n')));
  - Classify Tweets: ML.Classify.NaiveBayes.ClassifyD(dfIndep,TrainModel)
  - Visualize Results:
    - dPieFormatted:=VL.FormatCartesianData(dPieData,sentiment);
    - VL.Google.Cartesian('PieChart','SentimentBreakdown',dPieFormatted);

Kettle: a GUI to ECL-ML



HPCC-PaperBoat Integration

- Completed in about 2 weeks by one person without prior ECL experience
- Represents about 100,000 lines of existing C++ libraries
- Uses embedded C++ wrappers, ECL macros and ECL functions
- Seamless from an ECL programming standpoint

Nikolaos Vasiloglou, Ph.D. GA Tech

Integration code of PaperBoat with ECL

The goal of this section is to give more details about the integration mechanism of PaperBoat. As mentioned in *C++ code overview* the key point is to override some methods of the `WorkSpace` class.

All the files for connecting PaperBoat with ECL can be found inside the directory

```
ecl-pb/ecl-pb-glue
```

Let's start with the `workspace.h` file.

We define a new `WorkSpace` class that inherits from the `fl::ws::WorkSpace`. We need to provide new methods that translate an HPCC Dataset to a PaperBoat Table and loads it in the workspace. We also need to provide a method that does the opposite, exports a PaperBoat Table into an HPCC Dataset.

```
namespace fl { namespace hpcc {

class WorkSpace : public fl::ws::WorkSpace {
  public:

    // We have some type definitions that we need to ommit for
    // the shake of simplification
    ......

    /**
     * @brief This method reads an HPCC Dataset, converts it to
     *        a PaperBoat Table and stores it inside the workspace
     *        The function is templatized over the precision T of the Dataset.
     */
    template<typename T>
    void LoadDenseHPCCDataSet(const std::string &name,
        index_t n_attributes,
        index_t n_entries,
        const void *in_data);

    /**
     * @brief This method exports a PaperBoat Table residing in the workspace
     *        to an HPCC Dataset. The function is templatized over the
     *        precision T of the HPCC Dataset.
     */
    template<typename T>
```

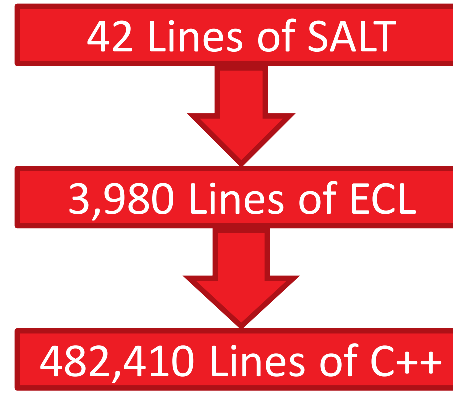
HPCC-PaperBoat ML

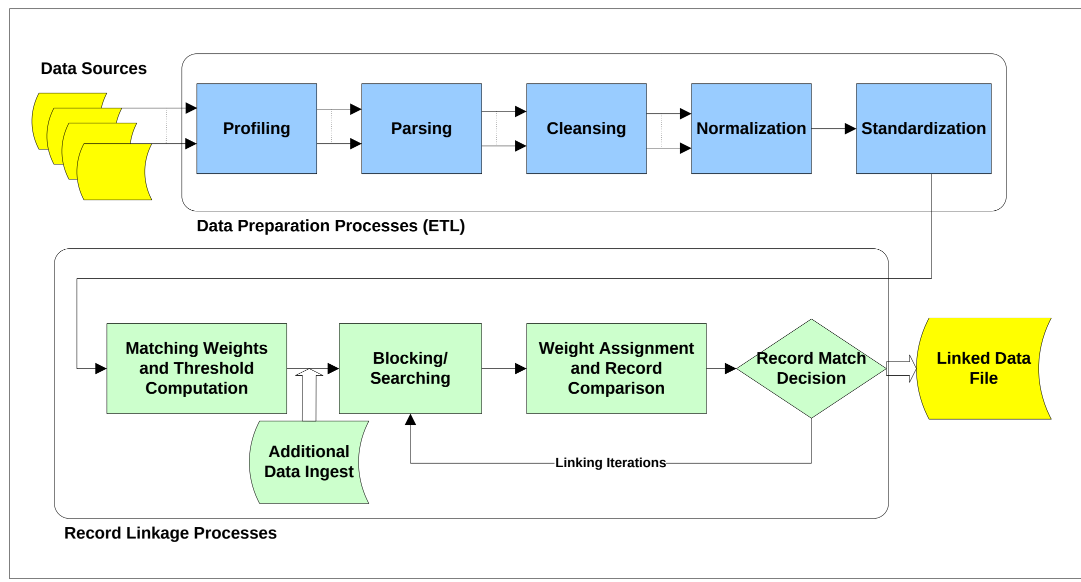
- **Available now**
  - All nearest neighbors
  - Kernel Density Estimation and non-parametric Bayes Classifier
  - Linear Regression
  - LASSO
  - Support Vector Machine
  - Non-Negative Matrix Factorization
  - Singular Value Decomposition
- **Coming soon**
  - Non-parametric regression
  - Decision trees
  - Principal Component Analysis
  - Orthogonal Range Search
  - Mean Shift
  - Ensemble Singular Value Decomposition
  - Multi-Time Series Prediction
  - Maximum Variance Unfolding
  - 3D Tensor Factorization
  - Graph Formation/Diffusion

Beyond ECL: SALT

HPCC Systems
http://hpccsystems.com

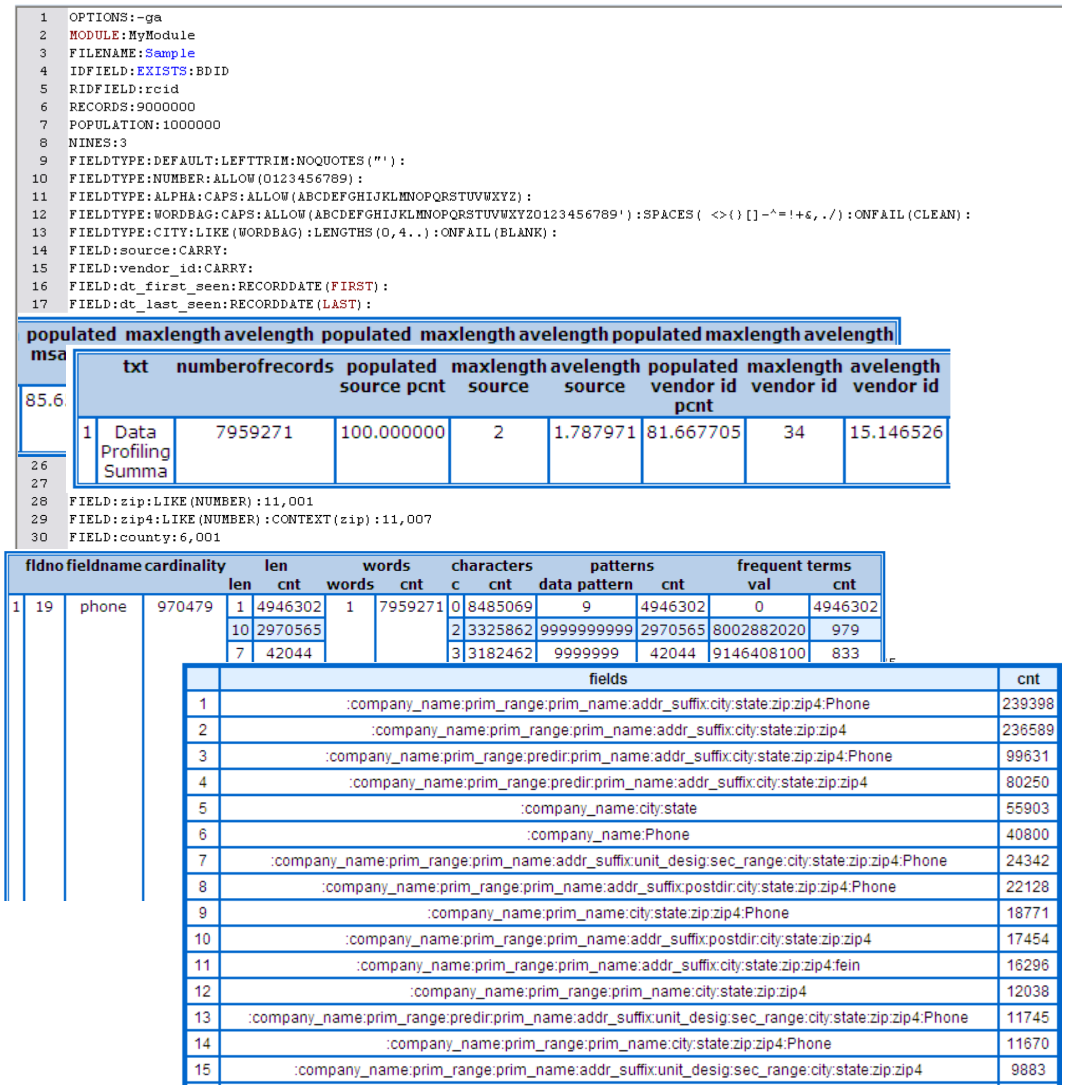
- The acronym stands for “Scalable Automated Linking Technology”
- Templates based ECL code generator
- Provides for automated data profiling, parsing, cleansing, normalization and standardization
- Sophisticated specificity and relatives based linking and clustering





Beyond ECL: SALT (ii)

- Calculates record matching field weights based on term specificity and matching weights
- What is the chance that two records for “John Smith” refer to the same person? How about “Edin Muharemagic”?
- It also takes into account transitive relationships
- What if these two records for “John Smith” were already linked to “Edin Muharemagic”? How many “John Smiths” does “Edin Muharemagic” know? Are these two “John Smith” records referring to the same person now?



```
OPTIONS:-ga
MODULE:MyModule
FILENAME:Sample
IDFIELD:EXISTS:BDID
RIDFIELD:rcid
RECORDS:9000000
POPULATION:1000000
NINES:3
FIELDTYPE:DEFAULT:LEFTTRIM:NOQUOTES("'):
FIELDTYPE:NUMBER:ALLOW(0123456789):
FIELDTYPE:ALPHA:CAPS:ALLOW(ABCDEFGHIJKLMNOPQRSTUVWXYZ):
FIELDTYPE:WORDBAG:CAPS:ALLOW(ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789'):SPACES( <>{}[]-^=!+&,./):ONFAIL(CLEAN):
FIELDTYPE:CITY:LIKE(WORDBAG):LENGTHS(0,4..):ONFAIL(BLANK):
FIELD:source:CARRY:
FIELD:vendor_id:CARRY:
FIELD:dt_first_seen:RECORDDATE(FIRST):
FIELD:dt_last_seen:RECORDDATE(LAST):
...
FIELD:zip:LIKE(NUMBER):11,001
FIELD:zip4:LIKE(NUMBER):CONTEXT(zip):11,007
FIELD:county:6,001
```

populated maxlength avelength populated maxlength avelength populated maxlength avelength
msa
85.6.

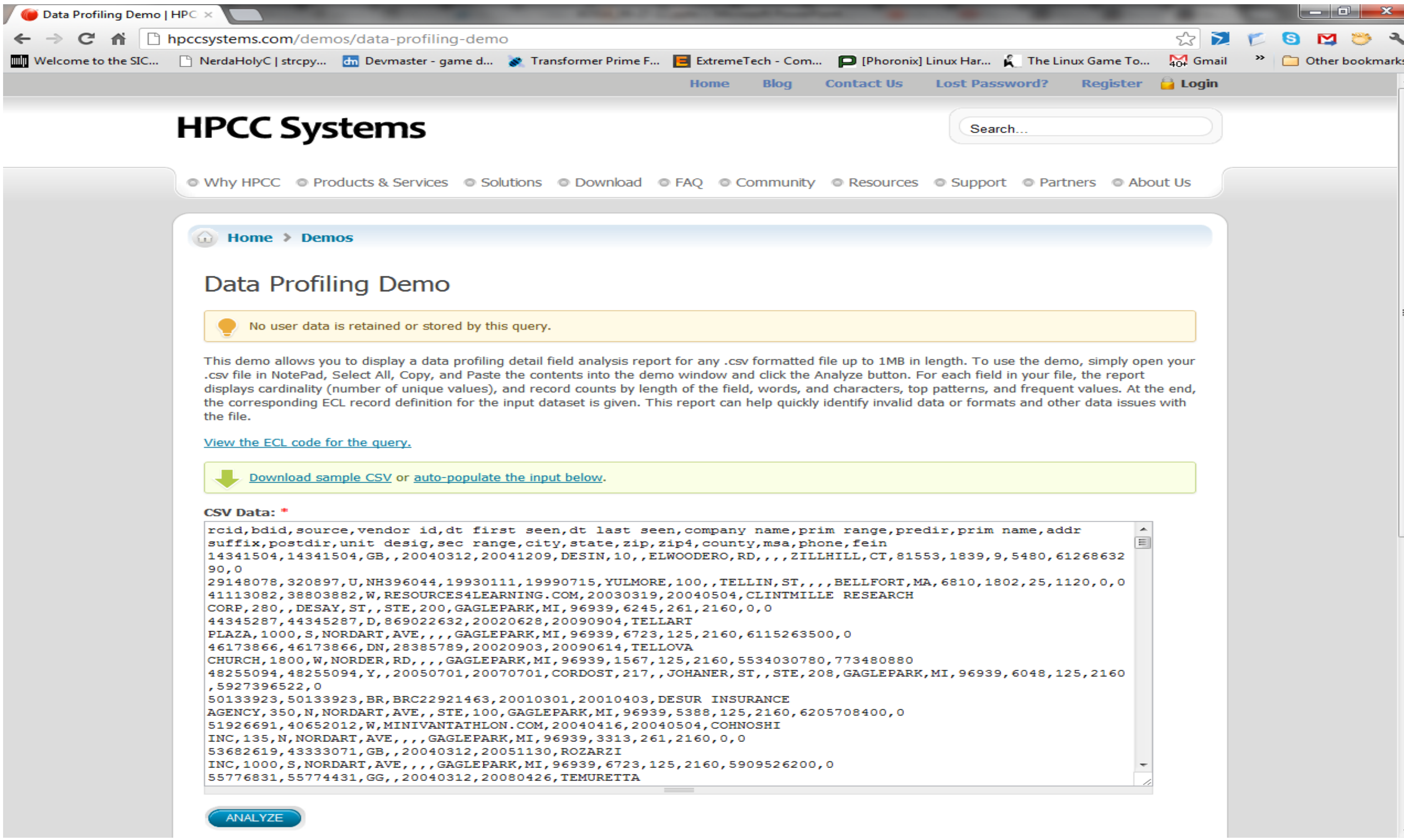
	txt	numberofrecords	populated source pcnt	maxlength source	avelength source	populated vendor id pcnt	maxlength vendor id	avelength vendor id
1	Data Profiling Summa	7959271	100.000000	2	1.787971	81.667705	34	15.146526

	fldno	fieldname	cardinality	len: len	len: cnt	words: words	words: cnt	characters: c	characters: cnt	patterns: data pattern	patterns: cnt	frequent terms: val	frequent terms: cnt
1	19	phone	970479	1	4946302	1	7959271	0	8485069	9	4946302	0	4946302
				10	2970565			2	3325862	9999999999	2970565	8002882020	979
				7	42044			3	3182462	9999999	42044	9146408100	833

	fields	cnt
1	:company_name:prim_range:prim_name:addr_suffix:city:state:zip:zip4:Phone	239398
2	:company_name:prim_range:prim_name:addr_suffix:city:state:zip:zip4	236589
3	:company_name:prim_range:predir:prim_name:addr_suffix:city:state:zip:zip4:Phone	99631
4	:company_name:prim_range:predir:prim_name:addr_suffix:city:state:zip:zip4	80250
5	:company_name:city:state	55903
6	:company_name:Phone	40800
7	:company_name:prim_range:prim_name:addr_suffix:unit_desig:sec_range:city:state:zip:zip4:Phone	24342
8	:company_name:prim_range:prim_name:addr_suffix:postdir:city:state:zip:zip4:Phone	22128
9	:company_name:prim_name:city:state:zip:zip4:Phone	18771
10	:company_name:prim_range:prim_name:addr_suffix:postdir:city:state:zip:zip4	17454
11	:company_name:prim_range:prim_name:addr_suffix:city:state:zip:zip4:fein	16296
12	:company_name:prim_range:prim_name:city:state:zip:zip4	12038
13	:company_name:prim_range:predir:prim_name:addr_suffix:unit_desig:sec_range:city:state:zip:zip4:Phone	11745
14	:company_name:prim_range:prim_name:city:state:zip:zip4:Phone	11670
15	:company_name:prim_range:prim_name:addr_suffix:unit_desig:sec_range:city:state:zip:zip4	9883

SALT Demo

HPCC Systems
http://hpccsystems.com



Data Profiling Demo | HPC
hpccsystems.com/demos/data-profiling-demo

Home Blog Contact Us Lost Password? Register Login

HPCC Systems

Why HPCC · Products & Services · Solutions · Download · FAQ · Community · Resources · Support · Partners · About Us

Home > Demos

Data Profiling Demo

No user data is retained or stored by this query.

This demo allows you to display a data profiling detail field analysis report for any .csv formatted file up to 1MB in length. To use the demo, simply open your .csv file in NotePad, Select All, Copy, and Paste the contents into the demo window and click the Analyze button. For each field in your file, the report displays cardinality (number of unique values), and record counts by length of the field, words, and characters, top patterns, and frequent values. At the end, the corresponding ECL record definition for the input dataset is given. This report can help quickly identify invalid data or formats and other data issues with the file.

View the ECL code for the query.

Download sample CSV or auto-populate the input below.

CSV Data: *

```
rcid,bdid,source,vendor id,dt first seen,dt last seen,company name,prim range,predir,prim name,addr
suffix,postdir,unit desig,sec range,city,state,zip,zip4,county,msa,phone,fein
14341504,14341504,GB,,20040312,20041209,DESIN,10,,ELWOODERO,RD,,,,ZILLHILL,CT,81553,1839,9,5480,61268632
90,0
29148078,320897,U,NH396044,19930111,19990715,YULMORE,100,,TELLIN,ST,,,,BELLFORT,MA,6810,1802,25,1120,0,0
41113082,38803882,W,RESOURCES4LEARNING.COM,20030319,20040504,CLINTMILLE RESEARCH
CORP,280,,DESAY,ST,,STE,200,GAGLEPARK,MI,96939,6245,261,2160,0,0
44345287,44345287,D,869022632,20020628,20090904,TELLART
PLAZA,1000,S,NORDART,AVE,,,,GAGLEPARK,MI,96939,6723,125,2160,6115263500,0
46173866,46173866,DN,28385789,20020903,20090614,TELLOVA
CHURCH,1800,W,NORDER,RD,,,,GAGLEPARK,MI,96939,1567,125,2160,5534030780,773480880
48255094,48255094,Y,,20050701,20070701,CORDOST,217,,JOHANER,ST,,STE,208,GAGLEPARK,MI,96939,6048,125,2160
,5927396522,0
50133923,50133923,BR,BRC22921463,20010301,20010403,DESUR INSURANCE
AGENCY,350,N,NORDART,AVE,,STE,100,GAGLEPARK,MI,96939,5388,125,2160,6205708400,0
51926691,40652012,W,MINIVANTATHLON.COM,20040416,20040504,COHNOSHI
INC,135,N,NORDART,AVE,,,,GAGLEPARK,MI,96939,3313,261,2160,0,0
53682619,43333071,GB,,20040312,20051130,ROZARZI
INC,1000,S,NORDART,AVE,,,,GAGLEPARK,MI,96939,6723,125,2160,5909526200,0
55776831,55774431,GG,,20040312,20080426,TEMURETTA
```

ANALYZE

SALT Demo

HPCC Systems
http://hpccsystems.com

fldno	fieldname	cardinality	len		words		characters		patterns		frequent terms	
			len	cnt	words	cnt	c	cnt	data pattern	cnt	val	cnt
1	rcid	1000	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
2	bdid	259	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
3	source	45	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
4	vendor id	954	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
5	dt first seen	243	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
6	dt last seen	243	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
7	company name	291	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
8	prim range	153	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
9	predir	7	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
10	prim name	77	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
11	addr suffix	14	[-] Hide		[-] Hide		[-] Hide		[-] Hide		[-] Hide	
			2	822	1	1000	T	710	AA	822	ST	709
			3	127			S	709	AAA	127	AVE	122
			4	44			V	165	AAAA	44	RD	88
			0	6			D	154		6	BLVD	43
			6	1			A	123	AAAAAA	1	DR	22
							E	123				6
							R	114			HWY	2
							L	46			LN	2
							B	43			CIR	1
							W	5			CT	1
							N	3			HOLW	1
							Y	3			NORDER	1
							H	3			ROW	1
							O	3			WAY	1
							C	2				
							I	1				
12	postdir	3	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
13	unit desig	10	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
14	sec range	76	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
15	city	36	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
16	state	18	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
17	zip	40	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
18	zip4	197	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
19	county	25	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
20	msa	15	[+] Show		[+] Show		[+] Show		[+] Show		[+] Show	
21	phone	33	[-] Hide		[-] Hide		[-] Hide		[-] Hide		[-] Hide	
			1	130	1	170	0	196	9	130	0	130
			10	37			6	59	9999999999	37	1844656000	6
			9	2			5	53	999999999	2	5836375160	3
			4	1			4	40	9999	1	3539847492	2
							1	36			1560083360	1
							8	35			1560162872	1

SALT Demo

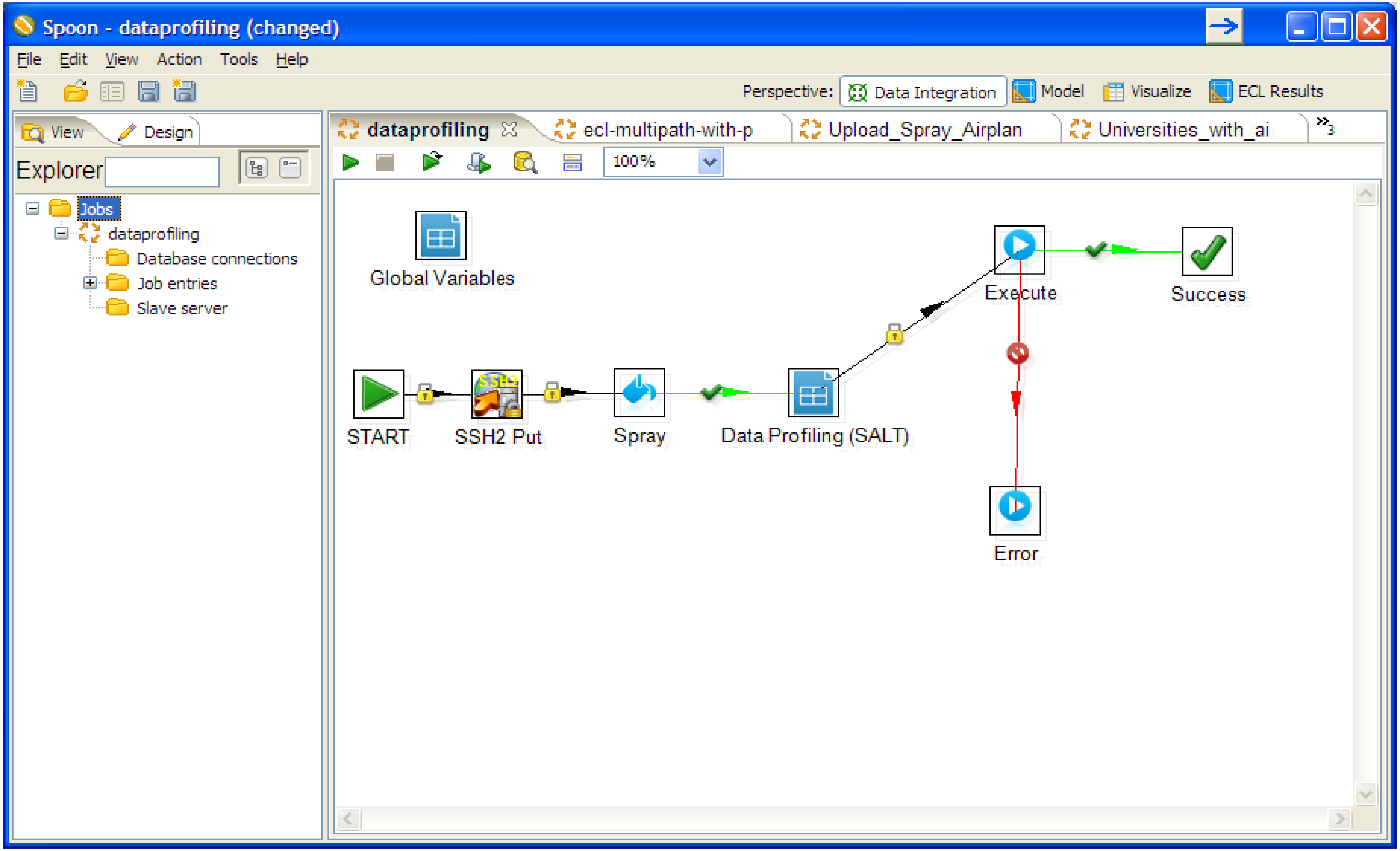
14	sec range	76	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
15	city	36	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
16	state	18	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
17	zip	40	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
18	zip4	197	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
19	county	25	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
20	msa	15	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
21	phone	33	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show
22	fein	4	[+] Show	[+] Show	[+] Show	[+] Show	[+] Show

Record Layout

```
Record_Layout := RECORD
UNSIGNED5 rcid;
UNSIGNED5 bdid;
STRING2 source;
STRING34 vendor id;
STRING8 dt first seen;
UNSIGNED4 dt last seen;
STRING37 company name;
STRING24 prim range;
STRING4 predir;
STRING11 prim name;
STRING6 addr suffix;
STRING2 postdir;
STRING4 unit desig;
STRING9 sec range;
STRING12 city;
STRING9 state;
STRING5 zip;
UNSIGNED3 zip4;
UNSIGNED2 county;
UNSIGNED2 msa;
UNSIGNED5 phone;
UNSIGNED5 fein;
END;
```

LexisNexis | Risk Solutions
HPCC Systems
1000 Alderman Dr.
Alpharetta, GA 30005
info@hpccsystems.com
US: 1.877.316.9669
Intl: 1.678.694.2200

SALT/Kettle integration



The Future: Knowledge Engineering

At the end of the day

- Do I really care about the format of the data?
- Do I even care about the placement of the data?
- I do care (a lot!) about what can be inferred from the data
- The context is important as long as it affects my inference process
- I want to leverage existing algorithms

ECL	KEL
Generates C++ (1->100)	Generates ECL (1->12)
Files and Records	Entities and associations
DECLARE how you want the data manipulated without worrying about the underlying process that makes that happen	DECLARE how you want the information to be derived; without worrying about the underlying data manipulations to make that happen
Can write graph and statistical algorithms	Major algorithms built in
Thor/Roxie split by human design	Thor/Roxie split by system design
Solid, reliable and mature	R&D

KEL by Example

- Actor := ENTITY(FLAT(UID(ActorName), Actor=ActorName))
- Movie := ENTITY(FLAT(UID(MovieName), Title=MovieName))
- Appearance := ASSOCIATION(FLAT(Actor Who, Movie What))

- USE IMDB.File_Actors(FLAT, Actor ,Movie ,Appearance)

- CoStar := ASSOCIATION(FLAT(Actor Who, Actor WhoElse))

- GLOBAL: Appearance(#1,#2) Appearance(#3,#2) => CoStar(#1,#3)

- QUERY:FindActors(_Actor) <= Actor(_Actor)
- QUERY:FindMovies(_Actor) <= Movie(UID IN Appearance(Who IN Actor(_Actor){UID}){What})
- QUERY:FindCostars(_Actor) <= Actor(UID IN CoStar(Who IN Actor(_Actor){UID}){WhoElse})
- QUERY:FindAll(_Actor) <= FindActors(_Actor), FindMovies(_Actor), FindCostars(_Actor)

What's next?

- Version 3.8.6 of the HPCC Systems platform is out!
- Ongoing R&D (3.10, 4.0 and beyond):
  - Level 3 BLAS support
  - More Machine Learning related algorithms
  - Heterogeneous/hybrid computing (FPGA, memory computing, GPU)
  - Knowledge Engineering Language driving ML
  - General usability enhancements (GUI to SALT data profiling and linking, etc.)
  - Integration with other third party systems (R, for example)

Useful links

- LexisNexis Open Source HPCC Systems Platform: http://hpccsystems.com
- Machine Learning portal: http://hpccsystems.com/ml
- PaperBoat integration: http://ismion.com/documentation/ecl-pb/index.html
- The HPCC Systems blog: http://hpccsystems.com/blog
- Our GitHub portal: https://github.com/hpcc-systems
- Community Forums: http://hpccsystems.com/bb

Call For Papers: Big Data Symposium

HPCC Systems
http://hpccsystems.com

Call for papers for the Big Data Symposium at the 2013 International Conference on Collaboration Technologies and Systems (CTS 2013).

The purpose of the Big Data Symposium is to address, explore and exchange information on the state-of-the-art and practice in the broad multidisciplinary field of Big Data. Participation is extended to researchers, designers, educators and interested parties in all disciplines and specialties (such as computer science, linguistics, psychology, statistics, sociology, multimedia and semantic web technologies, and more).

The deadline for the call for papers is January 21, 2013 and the link with details is here:
http://hpccsystems.com/community/training-events/events/cta_bddac2013

The Symposium track is being organized by a Program Committee led by HPCC Systems from LexisNexis. HPCC Systems is the open source, enterprise-proven Big Data Analytics Platform from LexisNexis.

Have questions or need more information?

Contact 1.877.316.9669 or info@hpccsystems.com

Trivia!

HPCC Systems
http://hpccsystems.com

Answer the question and win a prize!

- What is the name of the data refinery engine that provides batch oriented data manipulation?
- What is the name of the data delivery engine that provides real-time analytics?
- What is the query processing language supporting the HPCC Systems platform?
- What can be used to generate ECL code for automating data profiling, parsing and cleansing?
- Name three machine learning algorithms supported in ECL-ML.
- What does KEL stand for?

Questions???

HPCC Systems
http://hpccsystems.com