Big Data Analytics in
Large Scale Socio-economic Development

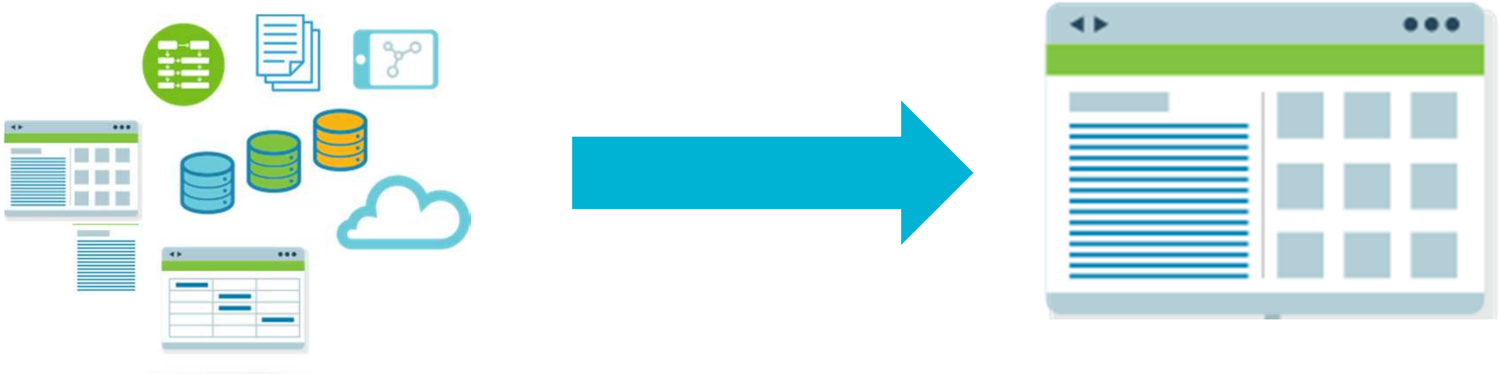
Ihab F. Ilyas

University of Waterloo
@ihabilyas

UNIVERSITY OF
WATERLOO

Data Science: Many Definitions and One Goal

- **Extract Value from Data**



Statistics + Machine Learning + Data Management + Systems ..

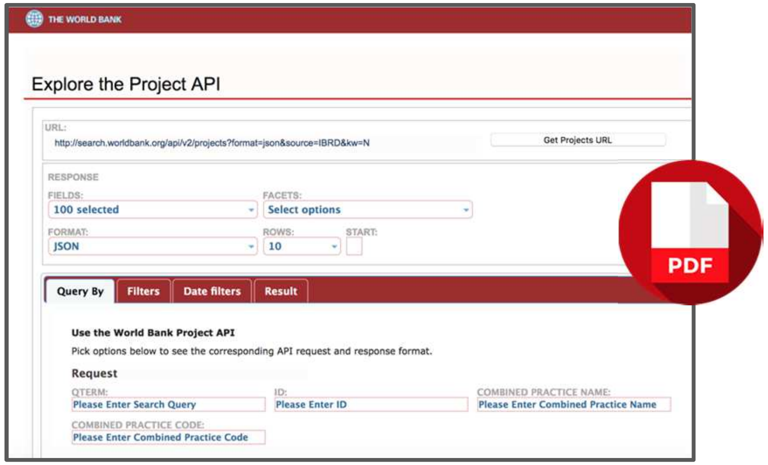
Case 1
The Effectiveness of the World Bank Funded Projects

The World Bank

- Increase the transparency and accountability of international development projects
- Visualize the location of Bank-financed projects to better monitor development impact
- Integrate location data with procurement and spending on local projects
- Measure spending against local economic and development indicators

Extract, Integrate, Map

Project Reports



THE WORLD BANK

Explore the Project API

URL: http://search.worldbank.org/api/v2/projects?format=json&source=IBRD&kw=N

Get Projects URL

RESPONSE

FIELDS: 100 selected

FACETS: Select options

FORMAT: JSON

ROWS: 10

START:

Use the World Bank Project API

Pick options below to see the corresponding API request and response format.

Request

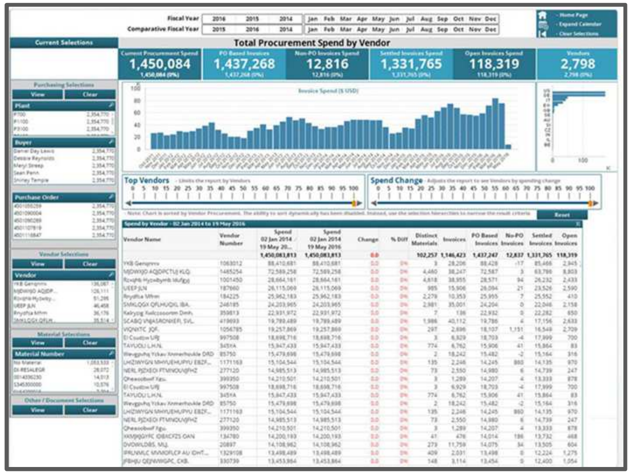
QTERM: Please Enter Search Query

ID: Please Enter ID

COMBINED PRACTICE NAME: Please Enter Combined Practice Name

COMBINED PRACTICE CODE: Please Enter Combined Practice Code

Procurement Data



Fiscal Year: 2016 2015 2014 | Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Comparative Fiscal Year: 2015 2016 2014 | Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Total Procurement Spend by Vendor

Current Procurement Spend	PO Based Invoices	Non-PO Invoices Spend	Settled Invoices Spend	Open Invoices Spend	Vendors
1,450,084	1,437,268	12,816	1,331,765	118,319	2,798
1,450,084 (0%)	1,437,268 (0%)	12,816 (0%)	1,331,765 (0%)	118,319 (0%)	2,798 (0%)

- Note: Chart is sorted by Vendor Procurement. The ability to sort dynamically has been disabled. Instead, use the selection hierarchies to narrow the result criteria

Spend by Vendor - 02 Jan 2014 to 19 May 2016

Vendor Name	Vendor Number	Spend 02 Jan 2014 - 19 May 20...	Spend 02 Jan 2014 19 May 2016	Change	% Diff	Distinct Materials	Invoices	PO Based Invoices	No-PO Invoices	Settled Invoices	Open Invoices
		1,450,083,813	1,450,083,813	0.0		102,257	1,146,423	1,437,247	12,837	1,331,765	118,319
YKB Genqmmv	1063012	88,410,681	88,410,681	0.0	0%	3	28,206	88,428	-17	85,466	2,945
MJDWIJO AQJDPCTUJ KLQ	1465254	72,589,258	72,589,258	0.0	0%	4,460	38,247	72,587	3	63,786	8,803
Rzxqhb Hyzwbymb Mufjxq	1001450	28,664,161	28,664,161	0.0	0%	4,618	38,955	28,571	94	26,232	2,433
UEEP JLN	187660	26,115,069	26,115,069	0.0	0%	985	15,906	26,094	21	23,526	2,590
Rnydfsa Mfnn	184225	25,962,183	25,962,183	0.0	0%	2,279	10,353	25,955	7	25,552	410
SMKLQGX QFLHUQXL IBA.	246185	24,203,965	24,203,965	0.0	0%	2,981	35,001	24,204	0	22,046	2,158
Kalryzg Xwlcsosorm Dmh.	359813	22,931,972	22,931,972	0.0	0%	7	136	22,932	0	22,282	650
SCABQ VNJASRONXEFL SVL.	419693	19,789,489	19,789,489	0.0	0%	1,986	40,112	19,786	4	17,156	2,633
VIQNXTC JQF.	1056785	19,257,869	19,257,869	0.0	0%	297	2,896	18,107	1,151	16,549	2,709
El Csudtow URJ	997508	18,698,716	18,698,716	0.0	0%	3	6,929	18,703	-4	17,999	700
TAYUOU LH.N.	345YA	15,947,433	15,947,433	0.0	0%	774	6,762	15,906	41	15,864	83
Wevgpvhq Yckav Xnmerhoukle DRD	85750	15,479,698	15,479,698	0.0	0%	2	18,242	15,482	-2	15,164	316
LHZIWYGN MHYUEHUPYU EBZF...	1171163	15,104,544	15,104,544	0.0	0%	135	2,246	14,245	860	14,135	970
NERL PJZXEOI FTMNOUVJFHZ	277120	14,985,513	14,985,513	0.0	0%	73	2,550	14,980	6	14,739	247
Qheasobwf Fgu.	399350	14,210,501	14,210,501	0.0	0%	3	1,289	14,207	4	13,333	878
El Csudtow URJ	997508	18,698,716	18,698,716	0.0	0%	3	6,929	18,703	-4	17,999	700
TAYUOU LH.N.	345YA	15,947,433	15,947,433	0.0	0%	774	6,762	15,906	41	15,864	83
Wevgpvhq Yckav Xnmerhoukle DRD	85750	15,479,698	15,479,698	0.0	0%	2	18,242	15,482	-2	15,164	316
LHZIWYGN MHYUEHUPYU EBZF...	1171163	15,104,544	15,104,544	0.0	0%	135	2,246	14,245	860	14,135	970
NERL PJZXEOI FTMNOUVJFHZ	277120	14,985,513	14,985,513	0.0	0%	73	2,550	14,980	6	14,739	247
Qheasobwf Fgu.	399350	14,210,501	14,210,501	0.0	0%	3	1,289	14,207	4	13,333	878
HXMJXJGYFC IDBACFZS OAN	134780	14,200,193	14,200,193	0.0	0%	41	476	14,014	186	13,732	468
DVDWLDBS, MLJ.	20897	14,108,962	14,108,962	0.0	0%	279	11,759	14,075	34	13,505	604
IPRUNMLC MVMOFLCP AU IDHT...	1329108	13,498,489	13,498,489	0.0	0%	409	2,031	13,498	0	12,224	1,275
JFBHJU QEJNWGPC, CKB.	330739	13,453,864	13,453,864	0.0	0%	148	3,114	13,454	0	12,400	1,054

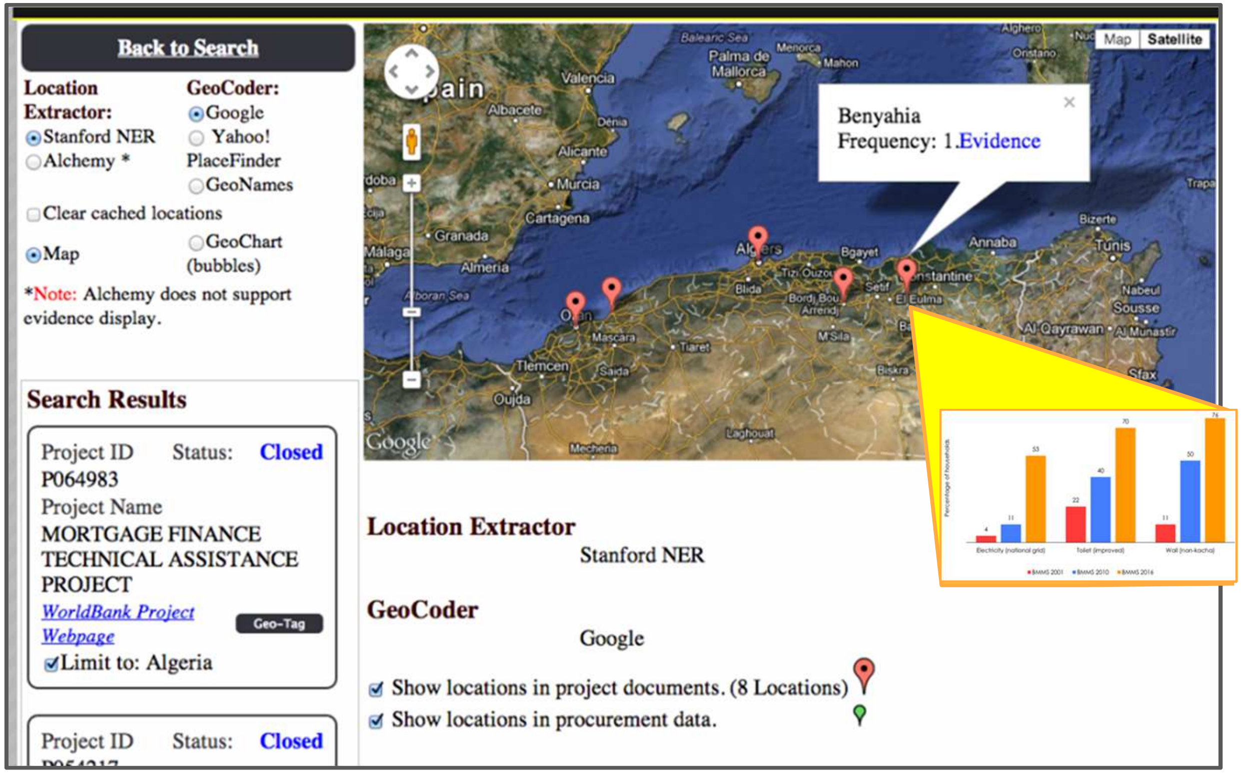
Stanford University Natural Language Processing NLP

GeoNames

AlchemyAPI

Read and Process Project Reports

- Geo-tagging
- Link location-specific procurement data
- Link to local Socio-economic indicators



Back to Search

Location Extractor:
◉ Stanford NER
○ Alchemy *

GeoCoder:
◉ Google
○ Yahoo! PlaceFinder
○ GeoNames

□ Clear cached locations

◉ Map
○ GeoChart (bubbles)

*Note: Alchemy does not support evidence display.

Search Results

Project ID P064983
Status: Closed
Project Name
MORTGAGE FINANCE TECHNICAL ASSISTANCE PROJECT
WorldBank Project Webpage
Geo-Tag
☑ Limit to: Algeria

Project ID
Status: Closed

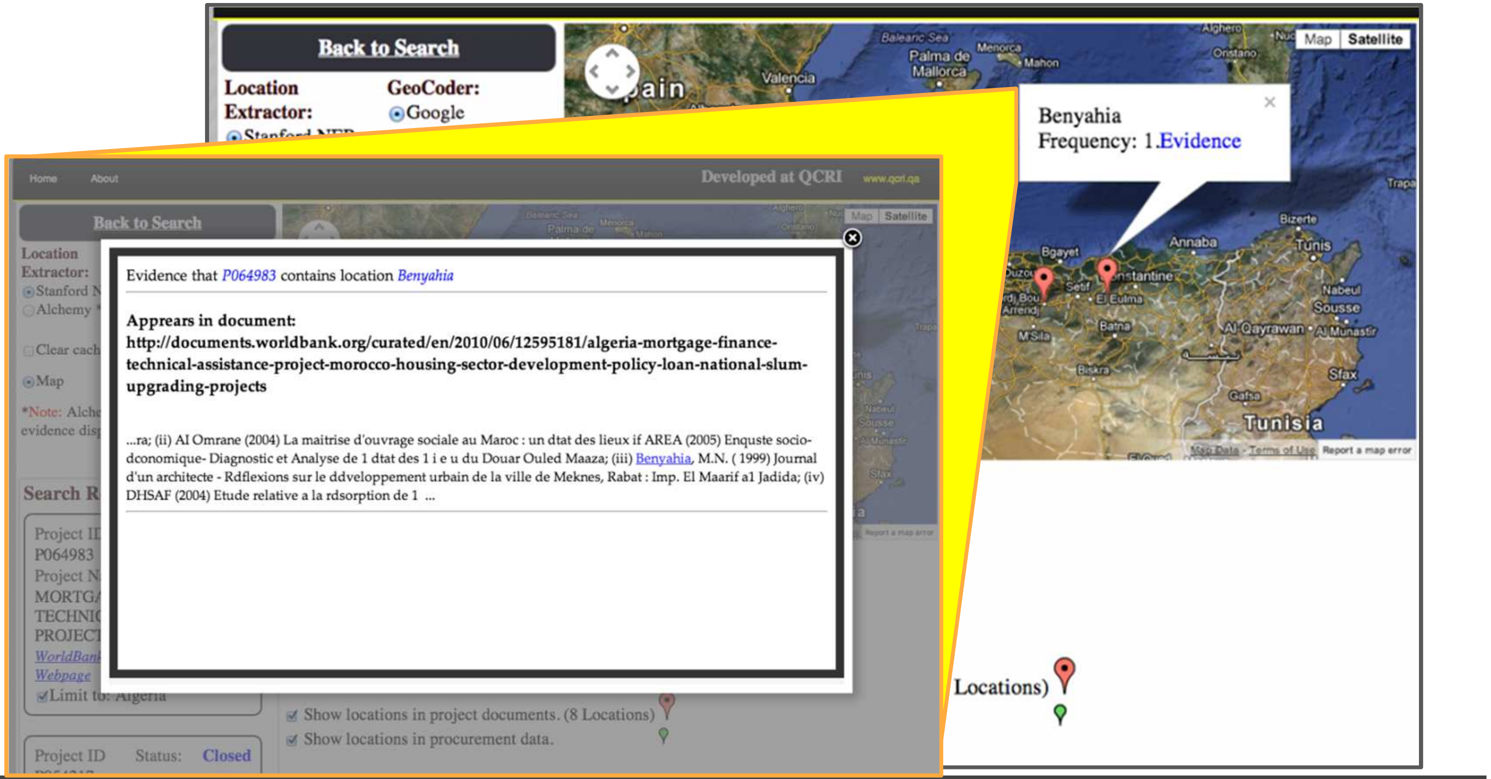
Benyahia
Frequency: 1.Evidence

Location Extractor
Stanford NER

GeoCoder
Google

☑ Show locations in project documents. (8 Locations)
☑ Show locations in procurement data.

Trace back Origins of Fact



Evidence that *P064983* contains location *Benyahia*

Apprears in document: http://documents.worldbank.org/curated/en/2010/06/12595181/algeria-mortgage-finance-technical-assistance-project-morocco-housing-sector-development-policy-loan-national-slum-upgrading-projects

...ra; (ii) AI Omrane (2004) La maitrise d'ouvrage sociale au Maroc : un dtat des lieux if AREA (2005) Enquste socio-dconomique- Diagnostic et Analyse de 1 dtat des 1 i e u du Douar Ouled Maaza; (iii) Benyahia, M.N. (1999) Journal d'un architecte - Rdflexions sur le ddveloppement urbain de la ville de Meknes, Rabat : Imp. El Maarif a1 Jadida; (iv) DHSAF (2004) Etude relative a la rdsorption de 1 ...

Case 2

Media Promotion Analytics

Internews - Information changes lives

- An international NGO
- Ensure access to trusted, quality information that empowers people to have a voice in their future

Linking Internews Projects and Aid Data

USASPENDING.gov BETA

GRANTS.GOV℠

USAID FROM THE AMERICAN PEOPLE

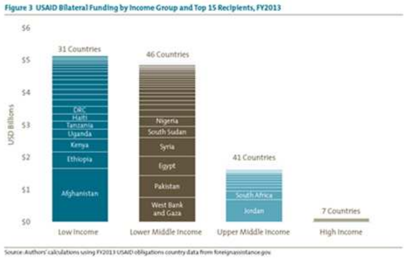
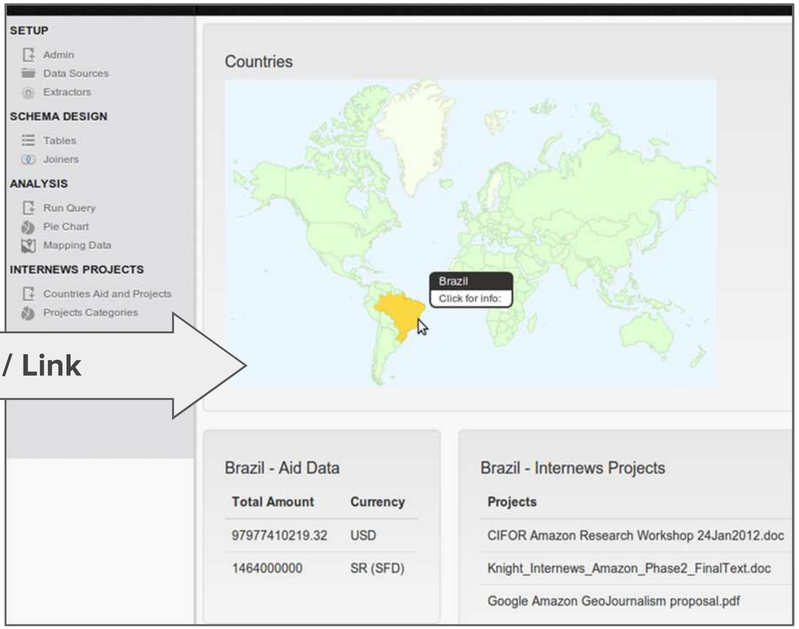


Figure 3 USAID Bilateral Funding by Income Group and Top 15 Recipients, FY2013

Source: Authors' calculations using FY2013 USAID obligations country data from foreignassistance.gov

Classify / Extract / Link



Brazil - Aid Data

Total Amount	Currency
97977410219.32	USD
1464000000	SR (SFD)

Brazil - Internews Projects

Projects
CIFOR Amazon Research Workshop 24Jan2012.doc
Knight_Internews_Amazon_Phase2_FinalText.doc
Google Amazon GeoJournalism proposal.pdf

Case 3

Behavioral Insights and Policy

How *Data Science* affects Policy

- How people react to different messages
- Behavioural insights combines statistics, economics, and psychology
- Use data science to answer the question:

“What works for whom?”

- Tailor policy based on demographics/populations

Predictive Analytics - Randomized Controlled Trials

1. Send different messages to different populations

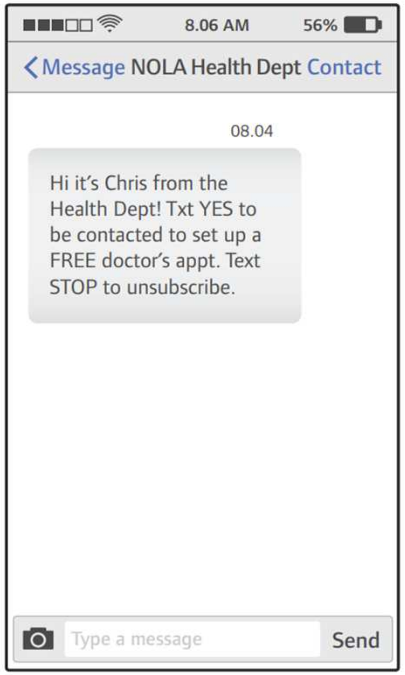
2. Analyze how they respond

3. Build predictive models

Behavioural Insights - Healthcare Appointments

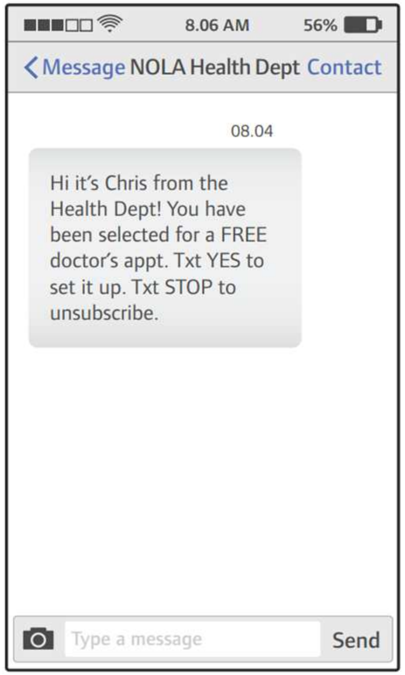
People respond differently to different messages

- Simple and generic
- Feeling special
- Promoting social life



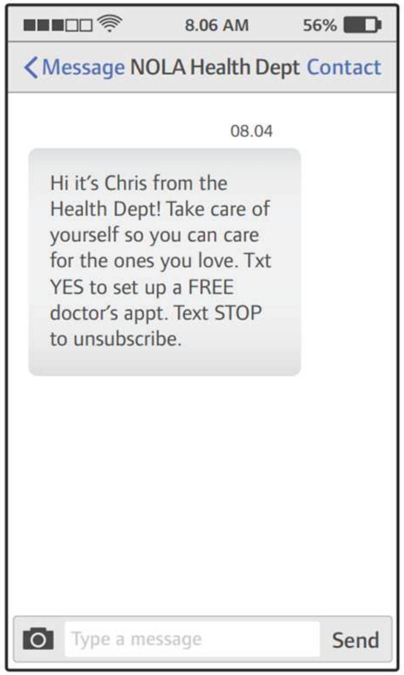
Hi it's Chris from the Health Dept! Txt YES to be contacted to set up a FREE doctor's appt. Text STOP to unsubscribe.

Simple



Hi it's Chris from the Health Dept! You have been selected for a FREE doctor's appt. Txt YES to set it up. Txt STOP to unsubscribe.

Unique



Hi it's Chris from the Health Dept! Take care of yourself so you can care for the ones you love. Txt YES to set up a FREE doctor's appt. Text STOP to unsubscribe.

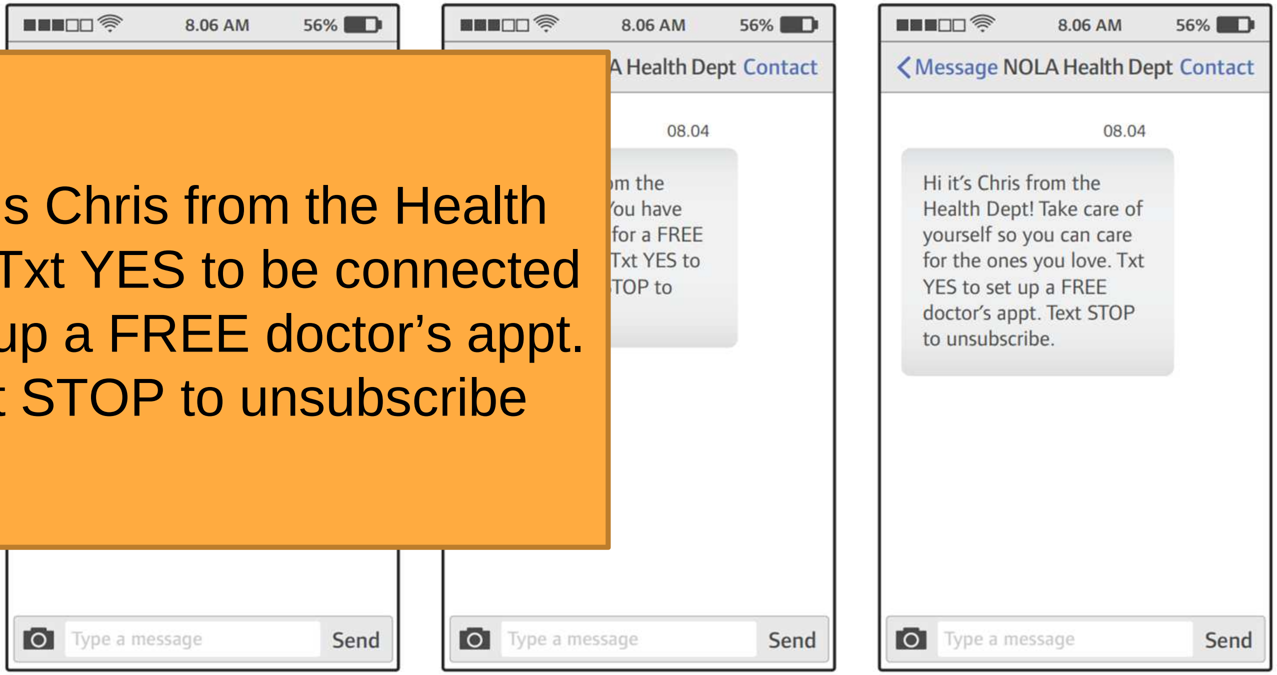
Prosocial

Behavioural Insights - Healthcare Appointments

People respond differently to diffe messages

- Simple and g
- Feeling spec
- Promoting social life

Hi, It’s Chris from the Health Dept! Txt YES to be connected to set up a FREE doctor’s appt. Text STOP to unsubscribe



om the
You have
for a FREE
Txt YES to
TOP to

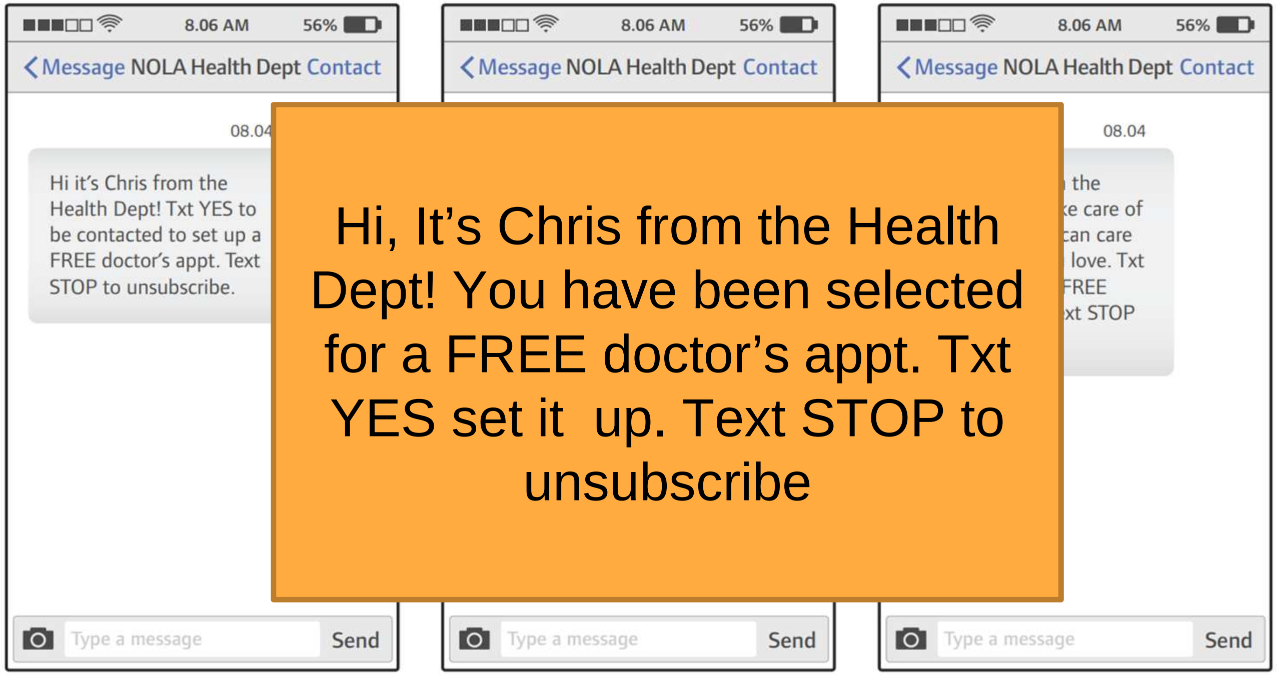
Hi it’s Chris from the Health Dept! Take care of yourself so you can care for the ones you love. Txt YES to set up a FREE doctor’s appt. Text STOP to unsubscribe.

Simple Unique Prosocial

Behavioural Insights - Healthcare Appointments

People respond differently to different messages

- Simple and generic
- Feeling special
- Promoting social life



Hi it's Chris from the Health Dept! Txt YES to be contacted to set up a FREE doctor's appt. Text STOP to unsubscribe.

the
e care of
can care
love. Txt
FREE
xt STOP

Hi, It's Chris from the Health Dept! You have been selected for a FREE doctor's appt. Txt YES set it up. Text STOP to unsubscribe

Simple

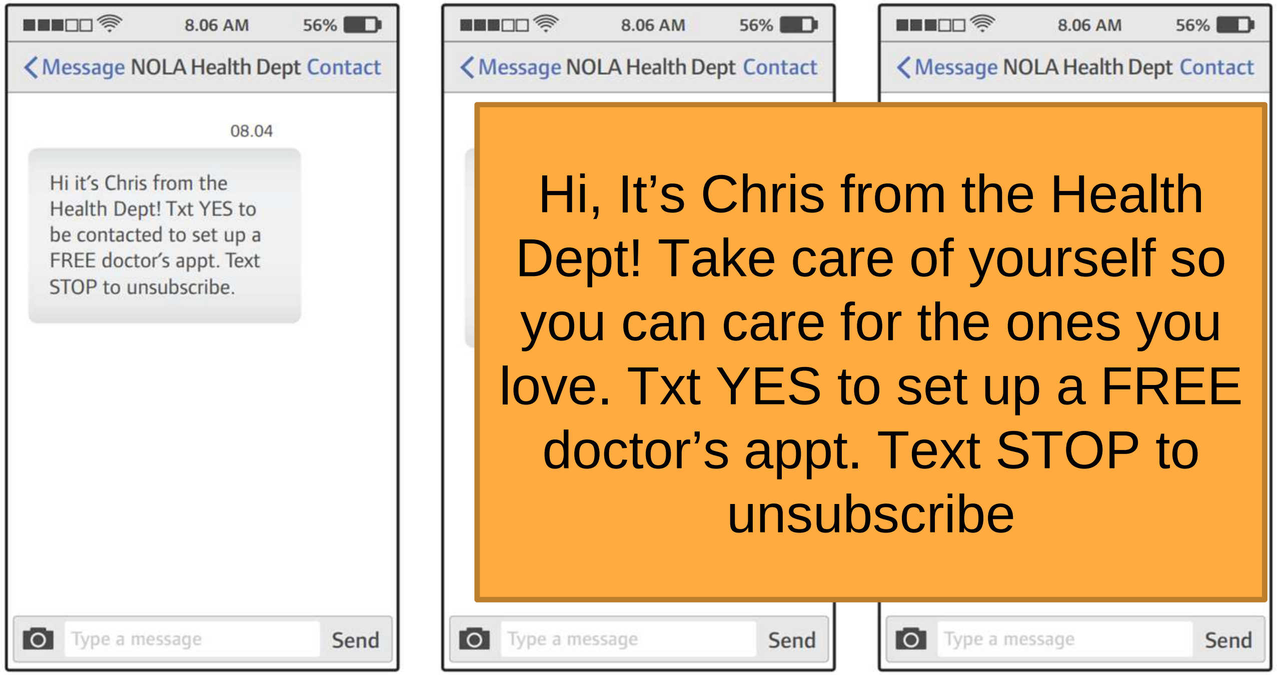
Unique

Prosocial

Behavioural Insights - Healthcare Appointments

People respond differently to different messages

- Simple and generic
- Feeling special
- Promoting social life



Hi it's Chris from the Health Dept! Txt YES to be contacted to set up a FREE doctor's appt. Text STOP to unsubscribe.

Hi, It's Chris from the Health Dept! Take care of yourself so you can care for the ones you love. Txt YES to set up a FREE doctor's appt. Text STOP to unsubscribe

Simple

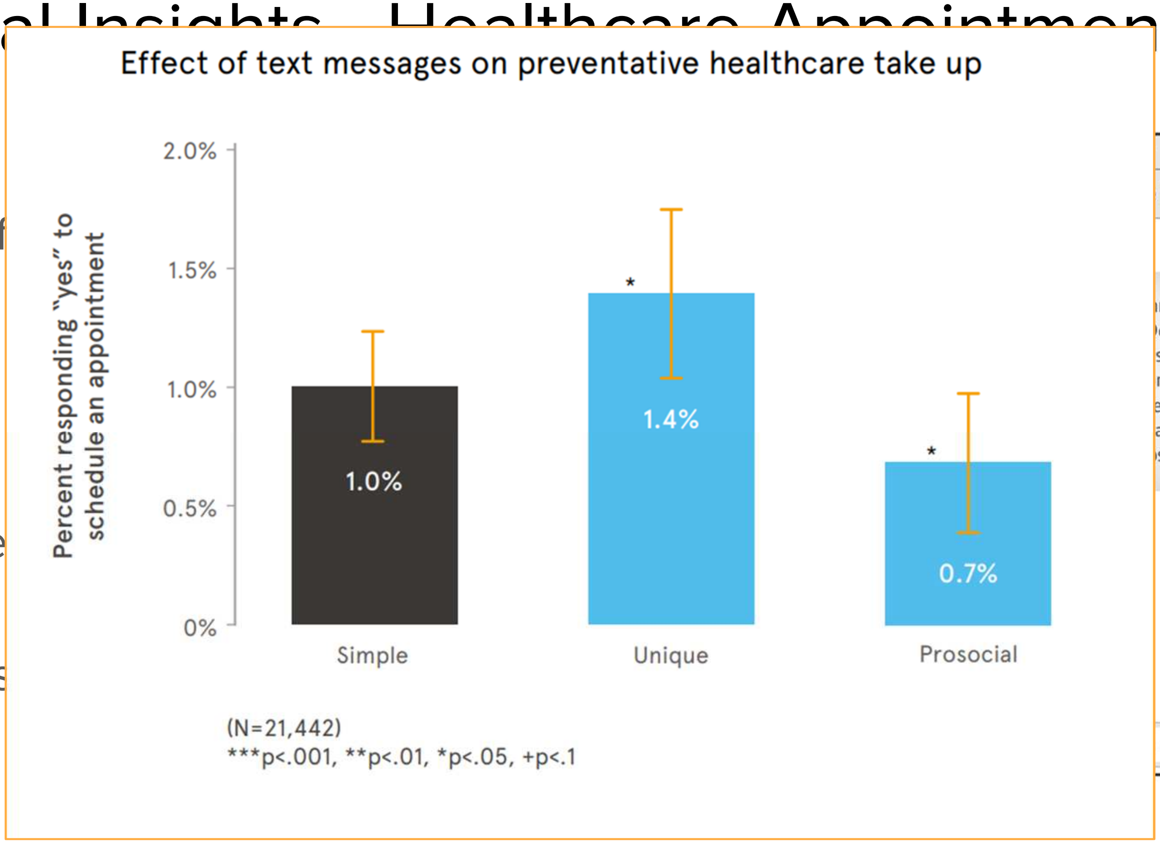
Unique

Prosocial

Behavioural Insights – Healthcare Appointments

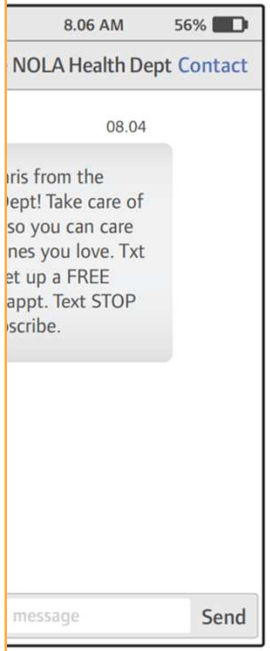
People respond differently to diff messages

- Simple and
- Feeling spe
- Promoting s



Effect of text messages on preventative healthcare take up

(N=21,442)
***p<.001, **p<.01, *p<.05, +p<.1



ıris from the
ept! Take care of
so you can care
nes you love. Txt
et up a FREE
appt. Text STOP
ıscribe.

Prosocial

Data Analytics to Fight Crime

- Predict Criminal Profiles
- Detect Human Trafficking

Pre-trial Criminal Justice

- Pre-trial has been estimated at over $9 billion per year in the U.S
- Data science can improve how decisions are made during the earliest part of the criminal justice process

Pre-trial Criminal Justice - Public Safety

- The District Attorney of New Jersey used data analysis
  - Camden, New Jersey, reduced murder by 41% and crime in the city general by 26%
- Use rigorous statistical analysis to classify defendants into **low**, **moderate**, and **high**-risk
  - Avoid releasing dangerous people
  - Avoid the cost of keeping low-risk offenders in jail waiting for trials

Predictive Analytics for Criminal Justice

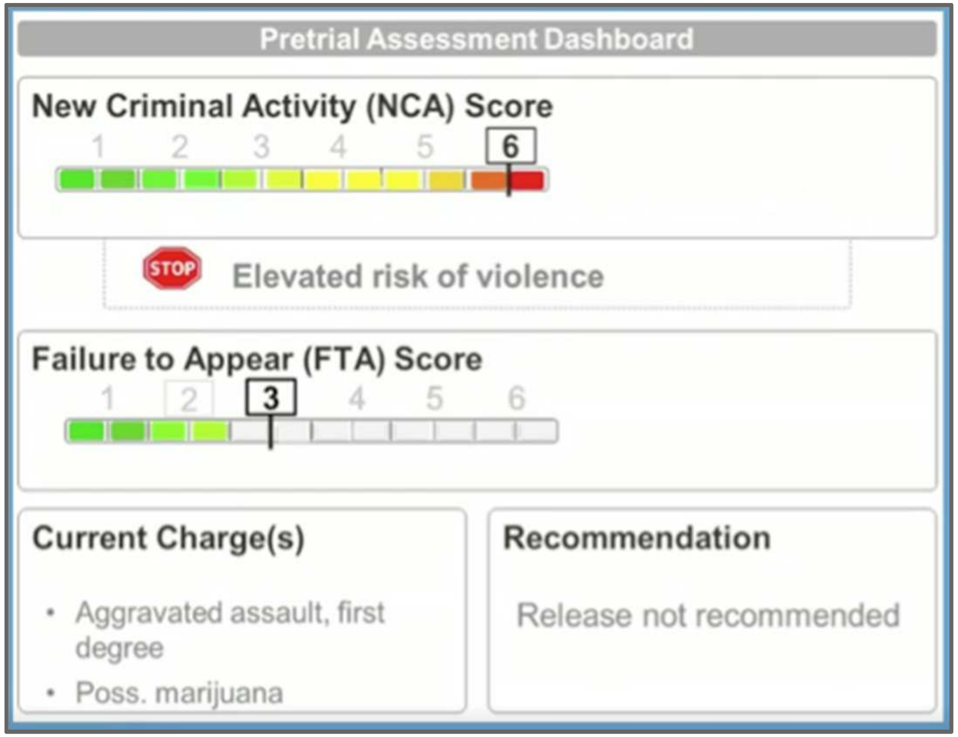
A team of Data Scientist examined

1.5 Million cases

Build a **risk-assessment tool**

Predict whether or not,
if someone is released, they will:

1. Commit a new crime.
2. Commit an act of violence.
3. Come back to court



Pretrial Assessment Dashboard

New Criminal Activity (NCA) Score

Elevated risk of violence

Failure to Appear (FTA) Score

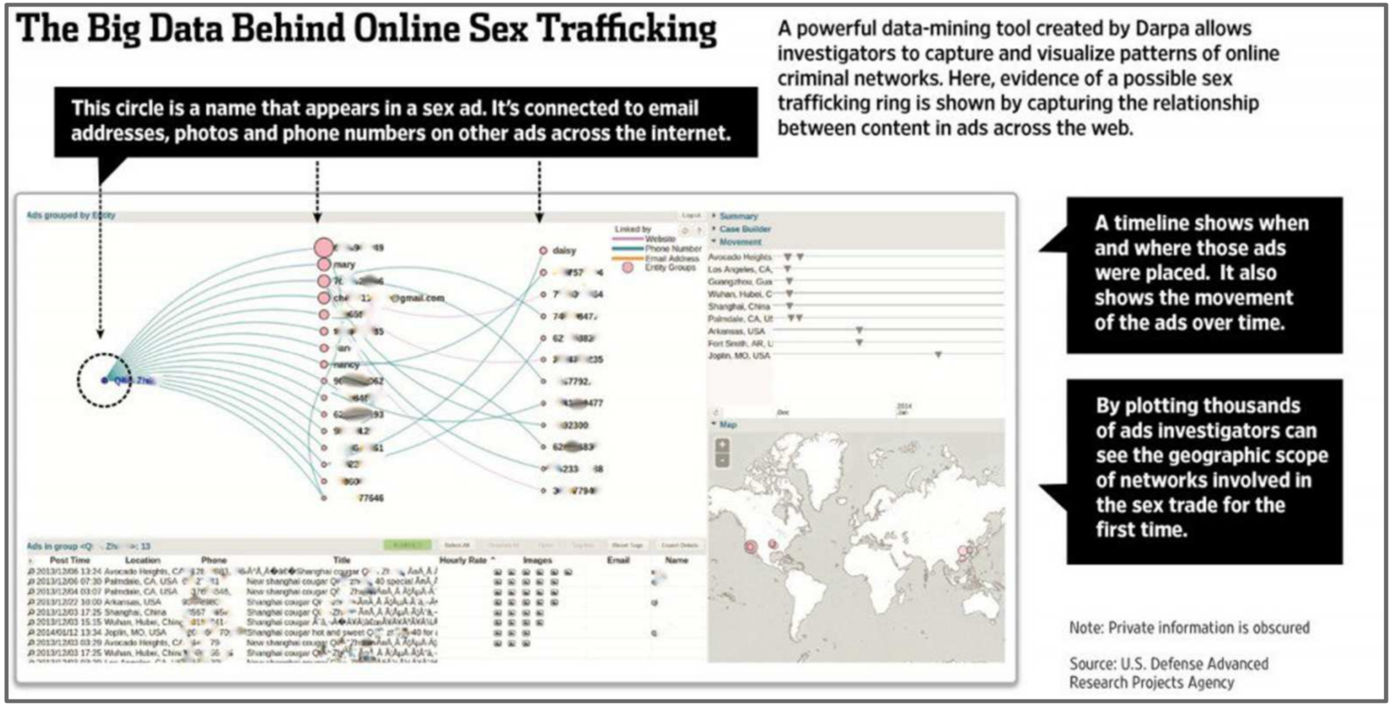
Current Charge(s)

- Aggravated assault, first degree
- Poss. marijuana

Recommendation

Release not recommended

Human Trafficking in the Deep Web



The Big Data Behind Online Sex Trafficking

A powerful data-mining tool created by Darpa allows investigators to capture and visualize patterns of online criminal networks. Here, evidence of a possible sex trafficking ring is shown by capturing the relationship between content in ads across the web.

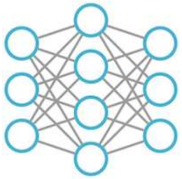
This circle is a name that appears in a sex ad. It's connected to email addresses, photos and phone numbers on other ads across the internet.

A timeline shows when and where those ads were placed. It also shows the movement of the ads over time.

By plotting thousands of ads investigators can see the geographic scope of networks involved in the sex trade for the first time.

Note: Private information is obscured

Source: U.S. Defense Advanced Research Projects Agency

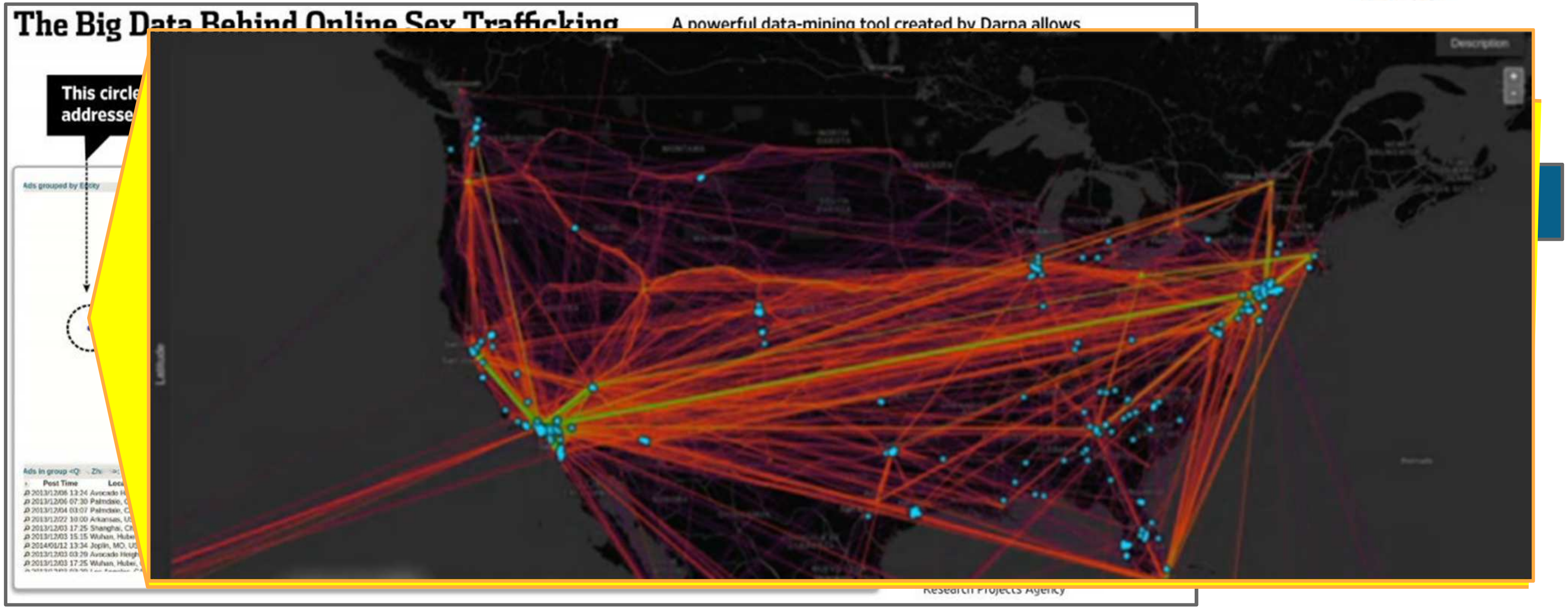


Information Extraction

Machine Learning

Statistical Inference

Human Trafficking - MEMEX



The Big Data Behind Online Sex Trafficking

A powerful data-mining tool created by Darpa allows

This circle

addresse

Ads grouped by Entity

Ads in group <Q Zh >

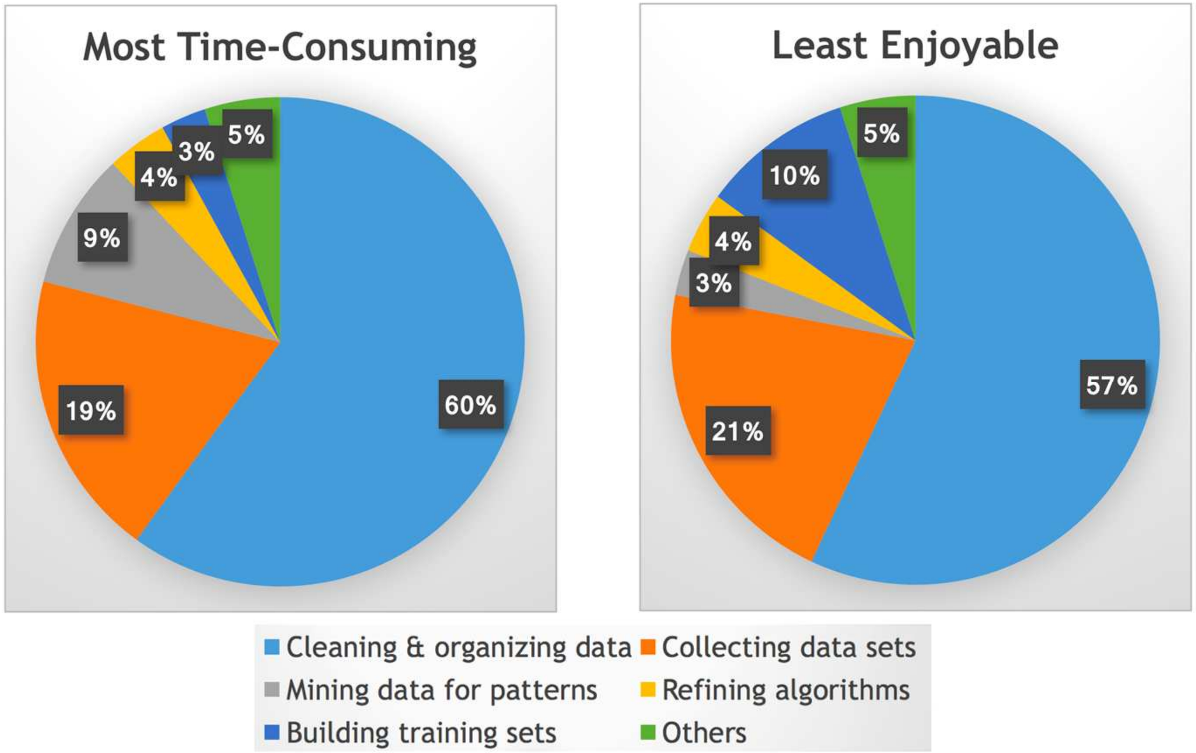
	Post Time	Loca
⌕	2013/12/06 13:24	Avocado H
⌕	2013/12/06 07:30	Palmdale, C
⌕	2013/12/04 03:07	Palmdale, C
⌕	2013/12/22 10:00	Arkansas, US
⌕	2013/12/03 17:25	Shanghai, Ch
⌕	2013/12/03 15:15	Wuhan, Hube
⌕	2014/01/12 13:34	Joplin, MO, US
⌕	2013/12/03 03:29	Avocado Heigh
⌕	2013/12/03 17:25	Wuhan, Hubei,

Research Projects Agency

Major Challenges

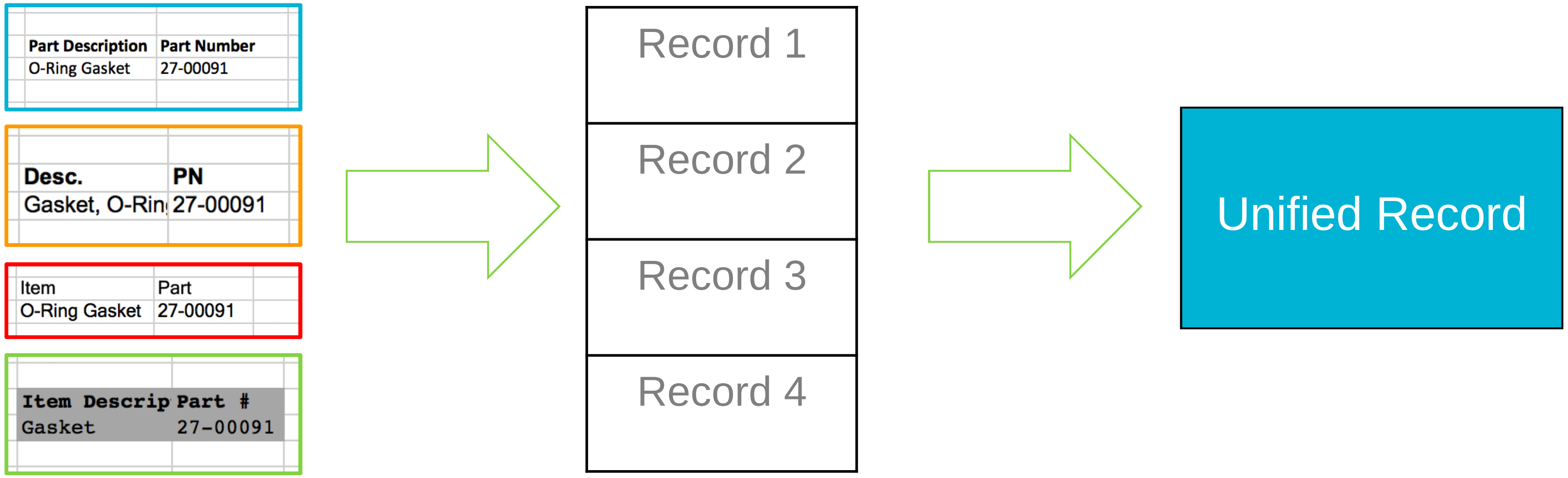
Data Management and Quality

Data Curation: Most Time-Consuming, Least Enjoyable



Unification: Dedup

Record Linkage and Deduplication



Part Description	Part Number
O-Ring Gasket	27-00091

Desc.	PN
Gasket, O-Rin	27-00091

Item	Part
O-Ring Gasket	27-00091

Item Descrip	Part #
Gasket	27-00091

Record 1

Record 2

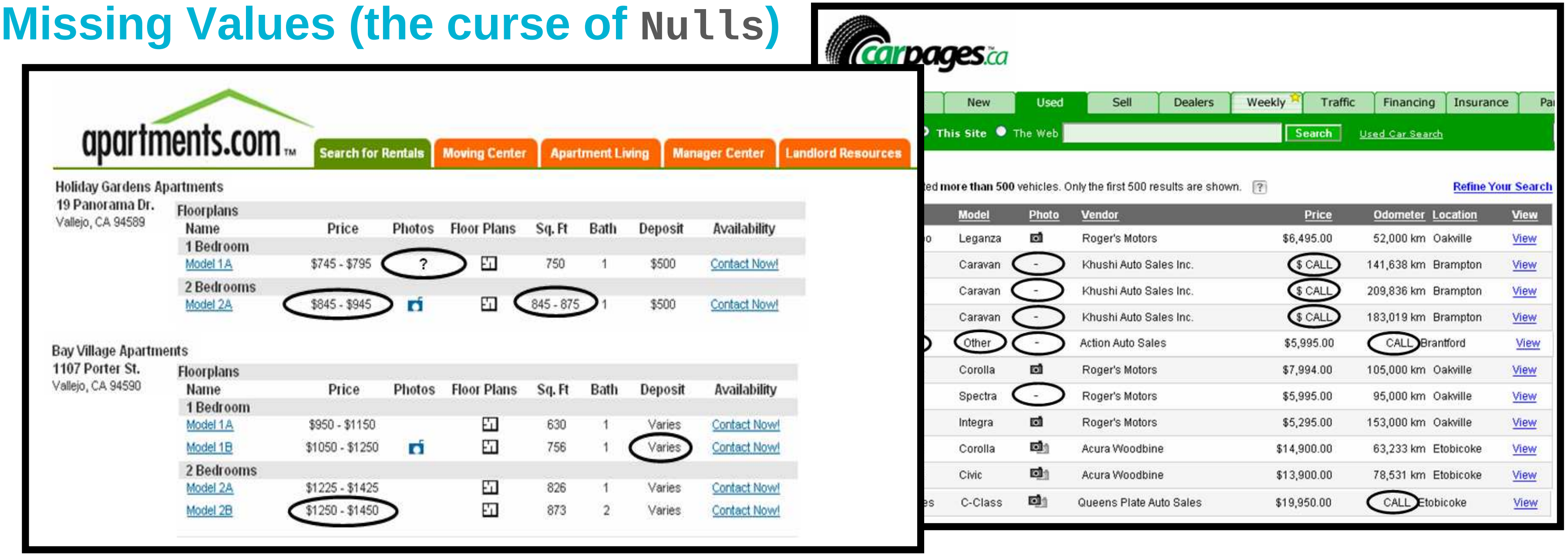
Record 3

Record 4

Unified Record

Cleaning: Missing Values

- **Missing Values (the curse of `Nulls`)**



Holiday Gardens Apartments
19 Panorama Dr.
Vallejo, CA 94589

Floorplans							
Name	Price	Photos	Floor Plans	Sq. Ft	Bath	Deposit	Availability
1 Bedroom							
Model 1A	$745 - $795	?	▦	750	1	$500	Contact Now!
2 Bedrooms							
Model 2A	$845 - $945	📷	▦	845 - 875	1	$500	Contact Now!

Bay Village Apartments
1107 Porter St.
Vallejo, CA 94590

Floorplans							
Name	Price	Photos	Floor Plans	Sq. Ft	Bath	Deposit	Availability
1 Bedroom							
Model 1A	$950 - $1150		▦	630	1	Varies	Contact Now!
Model 1B	$1050 - $1250	📷	▦	756	1	Varies	Contact Now!
2 Bedrooms							
Model 2A	$1225 - $1425		▦	826	1	Varies	Contact Now!
Model 2B	$1250 - $1450		▦	873	2	Varies	Contact Now!

ted **more than 500** vehicles. Only the first 500 results are shown. [?]

	Model	Photo	Vendor	Price	Odometer	Location	View
o	Leganza	📷	Roger's Motors	$6,495.00	52,000 km	Oakville	View
	Caravan	-	Khushi Auto Sales Inc.	$ CALL	141,638 km	Brampton	View
	Caravan	-	Khushi Auto Sales Inc.	$ CALL	209,836 km	Brampton	View
	Caravan	-	Khushi Auto Sales Inc.	$ CALL	183,019 km	Brampton	View
	Other	-	Action Auto Sales	$5,995.00	CALL	Brantford	View
	Corolla	📷	Roger's Motors	$7,994.00	105,000 km	Oakville	View
	Spectra	-	Roger's Motors	$5,995.00	95,000 km	Oakville	View
	Integra	📷	Roger's Motors	$5,295.00	153,000 km	Oakville	View
	Corolla	📷	Acura Woodbine	$14,900.00	63,233 km	Etobicoke	View
	Civic	📷	Acura Woodbine	$13,900.00	78,531 km	Etobicoke	View
es	C-Class	📷	Queens Plate Auto Sales	$19,950.00	CALL	Etobicoke	View

Real data is full of of N/A or nulls, special values (99999) etc.

Cleaning: Rule Violations

Integrity Constraints

ID	FN	LN	ROLE	CITY	ST	SAL
105	Anne	Nash	M	NYC	NY	110
211	Mark	White	E	SJ	CA	80
386	Mark	Lee	E	NYC	AZ	75
235	John	Smith	M	NYC	NY	1200

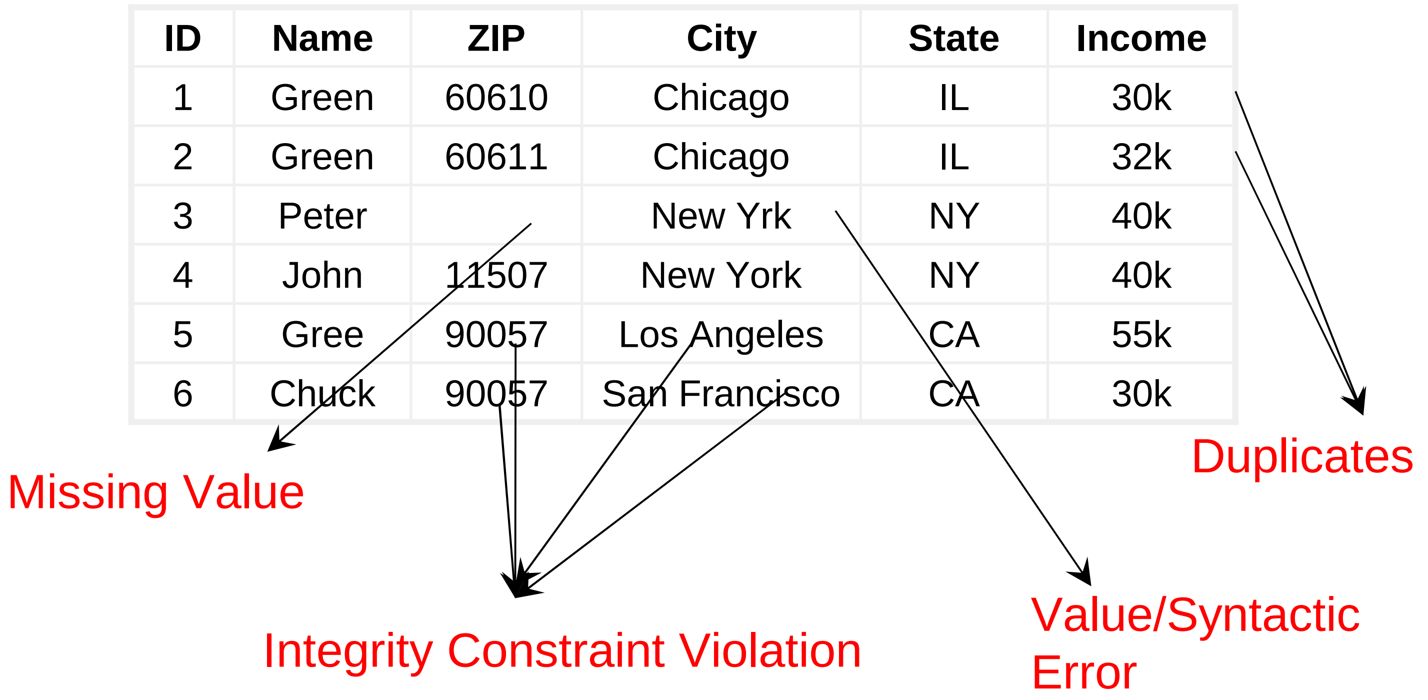
$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ROLE = t_\beta.ROLE \wedge t_\alpha.CITY = \text{“}NYC\text{”} \\ \wedge t_\beta.CITY \neq \text{“}NYC\text{”} \wedge t_\alpha.SAL < t_\beta.SAL)$$

Business Rule

Two employees of the same role, the one who lives in NYC cannot earn less than the one who does not live in NYC

Rarely expressed in practice. Most curation tools are rule-based implemented in imperative language

And They Don’t Come Piece-meal



ID	Name	ZIP	City	State	Income
1	Green	60610	Chicago	IL	30k
2	Green	60611	Chicago	IL	32k
3	Peter		New Yrk	NY	40k
4	John	11507	New York	NY	40k
5	Gree	90057	Los Angeles	CA	55k
6	Chuck	90057	San Francisco	CA	30k

Duplicates

Missing Value

Value/Syntactic Error

Integrity Constraint Violation

Multiple Efforts for Automation

8443322821
vs 844-332-2821

Pattern enforcement

TRIFACTA

Program synthesis

“Jeff Bezos”
vs “J. Bezos”

Entity Resolution
Schema Integration

tamr

ML+ Expert sourcing

VLDB17, Chicago
Vs VLDB17, Munich

Error repairs & imputation

HOLOCLEAN

Scalable Statistical Inference

Machine Learning is at the Heart of Data Science

- **Engineering and plumbing is ~ 80% of the exercise**
  - In-situ data preparation and signals computation: **Feature engineering**
  - Expert registration and engagement: **Training data management**
  - Blocking and pruning the candidate space: **Scale**
  - Provenance and lineage maintenance: **Explain and rollback**
  - Continuous model monitoring and validation: **Model management**

Data Repair is a statistical leaning and inference problem

Thank You!

@ihabilyas