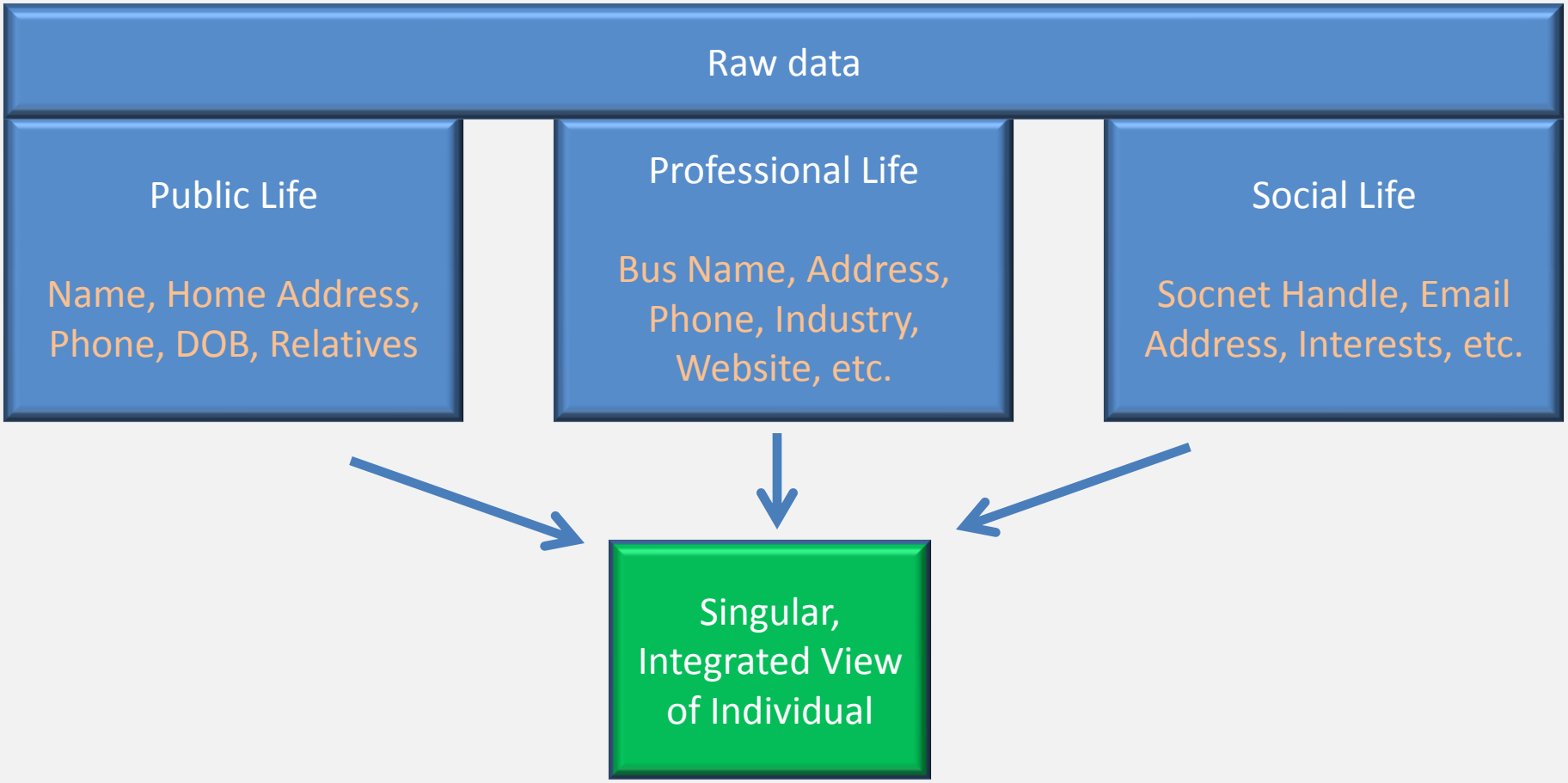
Navigation of Cascading Data Errors

Kevin Marcus, CTO, Intelius

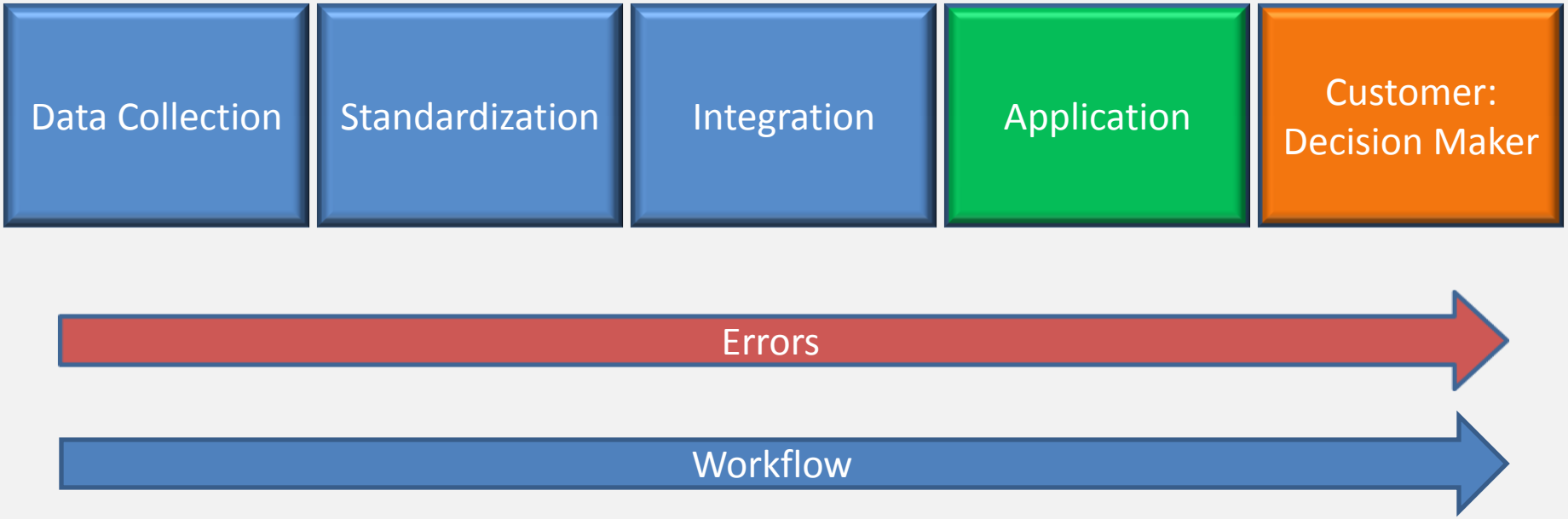
The Problem



- Apps want a single id: “People, Business, Assets”
- No universal globally available identifiers
- Data scattered in places, formats, attributes

Workflow

- Output of each step is input to next.
- Impact of errors are magnified as they progress.
- Any manipulation of original data has potential to introduce error



Data Collection

- Easy to do, *Hard to do well*.
- Entry Errors
- Digitization Errors
- False Information
- Precision
- Verification can be difficult.
- *Businesses may not be incented to collect high quality data!*

FirstName	MiddleName	Las
ROBERT	M	BE
BOB	S	BE
ROBERT	MARTIN	BE
ROBERT	M	BE
BOB	SHELTON	BE
ROBERT	M	BE
ROBERT	S	BE
ROBERT	M	BE
ROBERT	M	BE

```
MAR     JANE     BE
MARY    JEAN     ZE
```

```
54    6        19410615
54    4        19830309
54    2        19410600
```

```
OLD WOODINVILLE DUVALL RD NE
OLD WOODINVILLE DUVALL R
OLD WOODINVILLE DUVA RD
NE OLD WOODINVILLE DUVAL
```

Standardization

- Formatting ambiguities
- Some industries more advanced than others (Finance, HR)
- No “global data model”
  - Best friend: Multisampling convergence over time.
- Address standardization is hard and still *unsolved*.

Standardize Address: “11148 vANDERunG waY, round rock, tx”

http://zip4.usps.com/zip4/zcl_0_results.jsp

We're sorry! We were unable to process your request.

The address was not found. Please check the address below.
You may want to utilize the Yellow Pages and/or White Pages below.

* Required Fields

* Address 1: 11148 VANDERUNG WAY
Address 2: Apt, floor, suite, etc.
* City: ROUND ROCK
* State: TX Find state abbreviation

http://semaphorecorp.com/cgi/zp4.acgi$find

Search Result

Your Input	Search Result
11148 vANDERunG waY round rock TX	11148 vANDERunG waY round rock TX

Search warnings
4.1 Address not found
6.1 Multiple streets match

bing 11148 vANDERunG waY, Round Rock, TX

Businesses | People | Collections | Locations

location result Modify search

There was more than one match for your location.

Location suggestions:
15148
16148

Google maps 11148 vANDERunG waY, Round Rock, TX

Find businesses, addresses and places of interest. Learn

Get Directions | My Maps

Did you mean:

Wandering Vine Trail, Round Rock, Williamson, TX 78665

YAHOO! Web Images Video Local Shopping More

11148 vANDERunG waY, Round Rock, TX

Search Pad

SearchScan - On

0 results for
11148 vANDERunG waY,...

We did not find results for: **11148 vANDERunG waY, Round Rock, TX**.
Try the suggestions below or type a new query above.

Suggestions:

- Best Guess? 20 miles away – 11148 Wandering Way

Address: 11148 vANDERunG waY

City: round rock State: tx Zip:

qk: 20

Submit Query new dup (DBG) (TMR) (Cache) (Con

Version: v20081129a-

Remy's
SplitAddrM 11148 vANDERunG waY
Input: qa=**11148 vANDERunG waY**

Delay:32.799959182739 msec
Number=**11148**
Predirection=
Street=**WANDERING**
StreetType=**WAY**
Direction=
Unit=
City=**AUSTIN**
State=**TX**
Zip=**78753**
BuildingName=
Quality Tag=**Q:A-R2dist14-Z2-A-A-Klev2002^11**
APIParserVersion=**v20081129a-V1.23-1.33-3**

"Low quality match"

Record Integration is Hard

- Conflation
  - Are two records the same?
  - Statistical models are not perfect.
- De-dupe: Apps want a single ID!
- Orphans and Black Holes
  - Orphan: record unlinkable to others.
  - Black Hole: record matches too many others.
- Data Symmetry
  - Data interface restrictions. (e.g. CLID)
  - Offset with better conflation, caching

```
{  'addresses': [  {  'address_street1': '0 Street',
                      'city': 'SomeCity',
                      'postal_code': '12345',
                      'state': 'AA'}],
   'all_tags': [  'ABC county',
                  'member of ABC County Association Of Realtors',
                  'works at ABC Inc',
                  'realtor',
                  'AA',
                  'real estate agent',
                  'ABC'],
   'emails': [{'value': 'email@email.com'}],
   'names': [  {  'display_name': 'L B',
                  'first_name': 'L',
                  'last_name': 'B',
   'phone_numbers': [{'number_type': 'fax', 'sources': [0], 'value': '1234567890'}],
   'tags': [  {'value': 'ABC county'},
              {'value': 'member of ABC County Association Of Realtors'},
              {'value': 'works at ABC Inc'},
              {'value': 'realtor'},
              {'value': 'real estate agent'}]}
```

```
{  'addresses': [  {  'address_street1': '123 Street',
                      'city': 'SomeCity',
                      'postal_code': '12345',
                      'state': 'AA'}],
   'all_tags': ['Physician Assistant', 'General practitioner', 'ABC', 'physician Assistant', 'AA'],
   'jobs': [{'job_title': 'Physician Assistant'}],
   'names': [{'first_name': 'L', 'last_name': 'B', 'middle_name': 'I'}],
   'tags': [  {'value': 'physician Assistant'},
              {'value': 'General practitioner'}]}
```


PartyID	FirstName	MiddleName	LastName	HouseNum	StreetName	City	State	Zip
22	L	I	B	456	Street	SomeCity	AA	12345
22	L	I	B	123	Street	SomeCity	AA	12345
96	L		B	0	Street	SomeCity	AA	12345

Applications and Decisions

- Historical: Data + App often originated at same organization.
  - No longer true!
  - Data is bought, sold, processed and sliced from various sources!
- Apps: Guilty until proven innocent
- Computers should *help* a human decide
  - Should not be the final judgment!
  - AI not quite good enough yet ☺

M W M II: Parental Control!

- Unique name: Only 3 people in US; 2 in WA!
- 43 Criminal records
- 10 Civil cases
- DOB available on many – not all!
- "Use suffix when writing signature"
- Obvious issues: "Given the filing date, would it make sense for II?"

Record 14 of 43

IDENTIFICATION

[redacted]

M

DOB: [redacted] 1963

OFFENSE DETAILS

CASE NUMBER: [redacted] - Confirm Case at the Courthouse

COURT NAME: SKAGIT COUNTY DISTRICT COURT

CASE TYPE: NON-TRAFFIC

STATE CODE: WA

OFFENSE ID: [redacted]

OFFENSE DATE: 09/03/2004

OFFENSE CODE: 26.50.110

OFFENSE: VIOLATION OF PROTECTION ORDER

RECORD 3

CIVIL RECORD VERIFICATION: [redacted]

DEFENDANT: [redacted]

CASE NUMBER: [redacted]

FILING TYPE: CIVIL JUDGMENT

ENTITY TYPE: INDIVIDUAL RECORD

FILING DATE: 19970100

AMOUNT LIABILITY: $5,400.00

ADDRESS: [redacted]

CITY: LYNNWOOD

STATE: WA

ZIP: 98036

PLAINTIFF: BANK OF AMERICA

COURT CODE: WA099

COURT NAME: SNOHOMISH COUNTY, SOUTH/LYNNWOOD DISTRICT COURT

UNLAWFUL DETAINER: N

RECORD 4

CIVIL RECORD VERIFICATION: [redacted]

DEFENDANT: [redacted]

CASE NUMBER: [redacted]

FILING TYPE: CHAPTER 13 FILING (NEW)

ENTITY TYPE: INDIVIDUAL RECORD

ADDRESS: [redacted]

CITY:

STATE: WA

ZIP: [redacted]

ATTORNEY: PRO SE

ASSETS AVAILABLE: Y

COURT CODE: WAKINF1

COURT NAME: WASHINGTON FED COURT-SEATTLE

JUDGE INITIALS: TTG

ACTION TYPE: BANKRUPTCY

Future Needs

- More data collection transparency.
- Better error detection and correction tools.
- Improved record conflation mechanisms.
- International influences
  - Privacy and collection issues vary
  - Schemas may need to change.
- Easier global “fix” for individuals affected.
  - Centralized repository?
  - Better authentication and validation systems?