Valence Health

Evolution of vBond: Linking Data from Diverse Storage Systems to Support Health Care Analytics

NOVEMBER 12TH, 2012
BART PHILLIPS, MS
VALENCE HEALTH

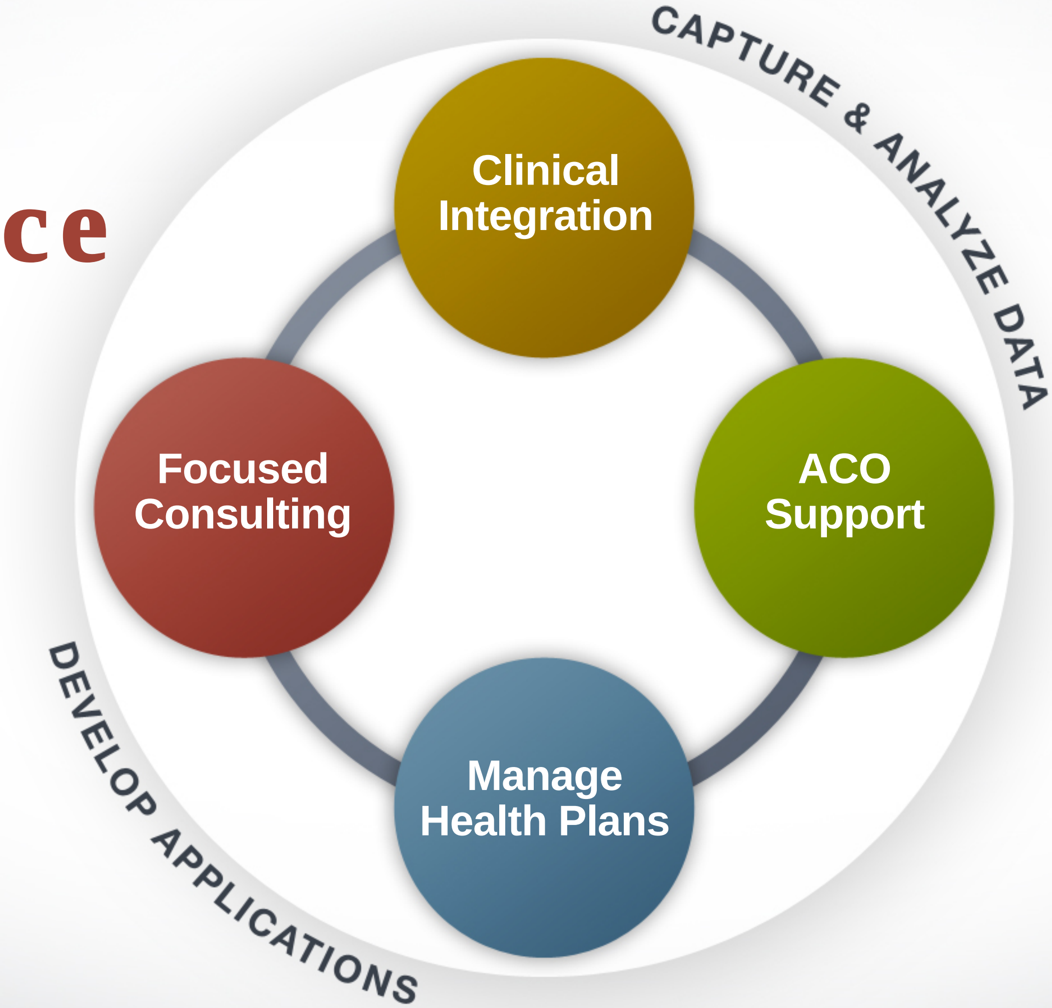
BETTER CARE IS ELEMENTAL.

Overview

- Background on Valence Health
- Linking data industry
- Common tricks
- vBond – the evolution of the Valence Health linking solution
- Where does vBond go from here?

Valence Health BETTER CARE IS ELEMENTAL.

What Valence Does



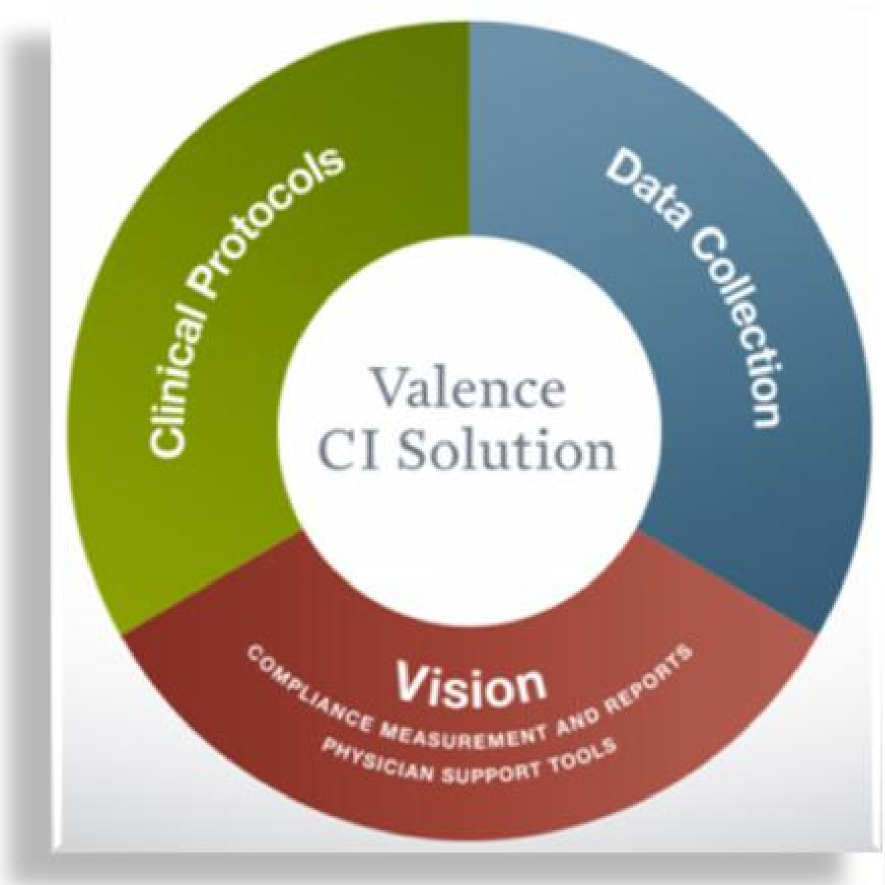
Valence delivers patient-centered, data-driven solutions so providers can achieve optimal reward for quality care.

Valence Health What Others Say

- **“V**alence can now provide alerts about patients before they visit a practice, so doctors have the information they need to ensure compliance with care guidelines.” – SAS
- “[Valence Health] arms providers with the ability to prove that new metrics are truly being met in order to achieve optimal reward.” – North Bridge
- “For 15 years Valence Health has been leading the way in enabling healthcare providers to optimize their systems to deliver quality care.”

Valence Health BETTER CARE IS ELEMENTAL.

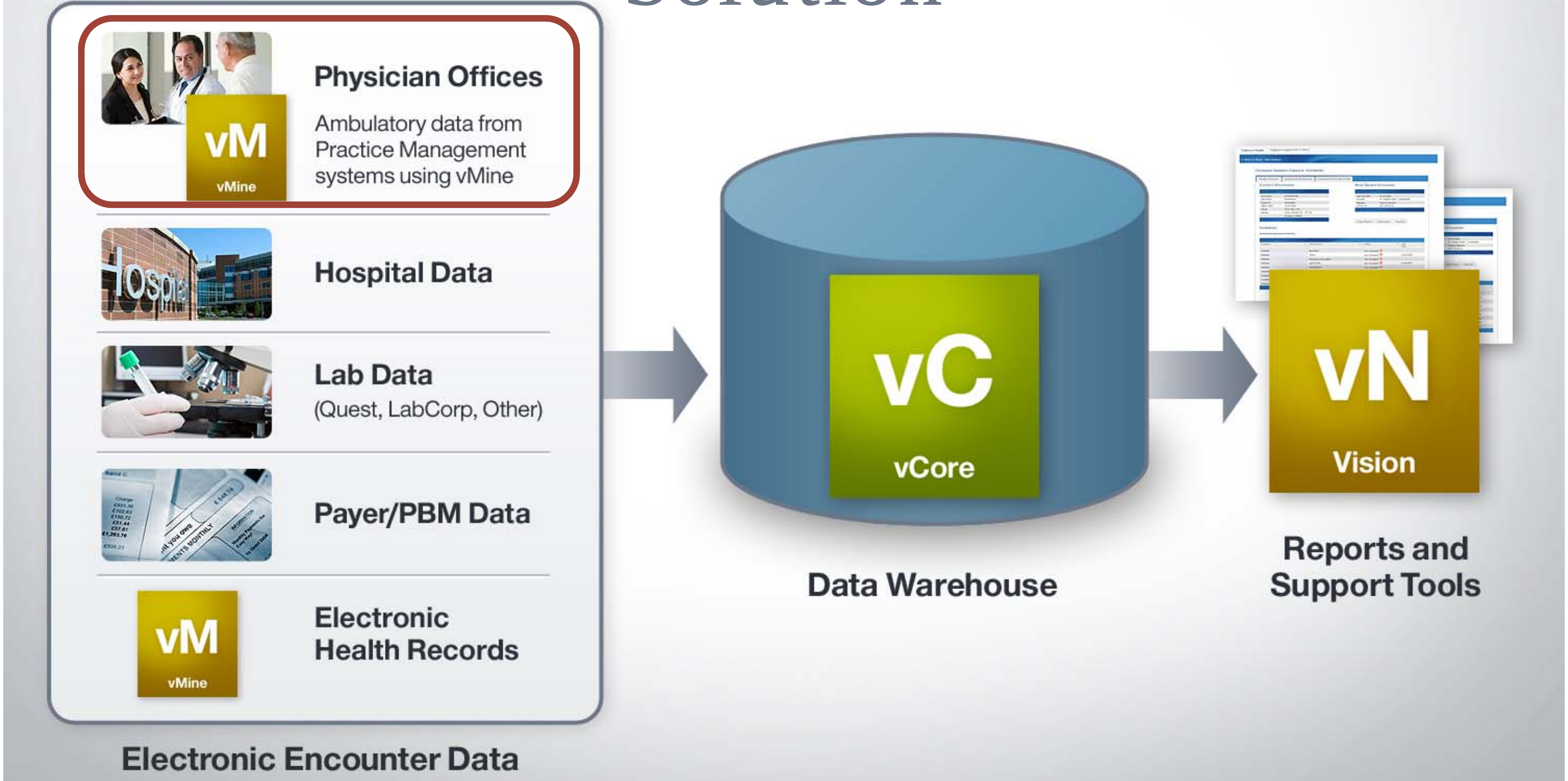
Our Clinical Integration Solution



Our Philosophy:

- **Patient-centric approach**
- **Provide an opportunity for all specialties to participate, > 90 protocols**
- **An option where clinical data is collected, and physicians DON'T do more administrative work**

Data Commonly Used in our Solution



vM
vMine

Physician Offices

Ambulatory data from Practice Management systems using vMine

Hospital Data

Lab Data

(Quest, LabCorp, Other)

Payer/PBM Data

vM
vMine

Electronic Health Records

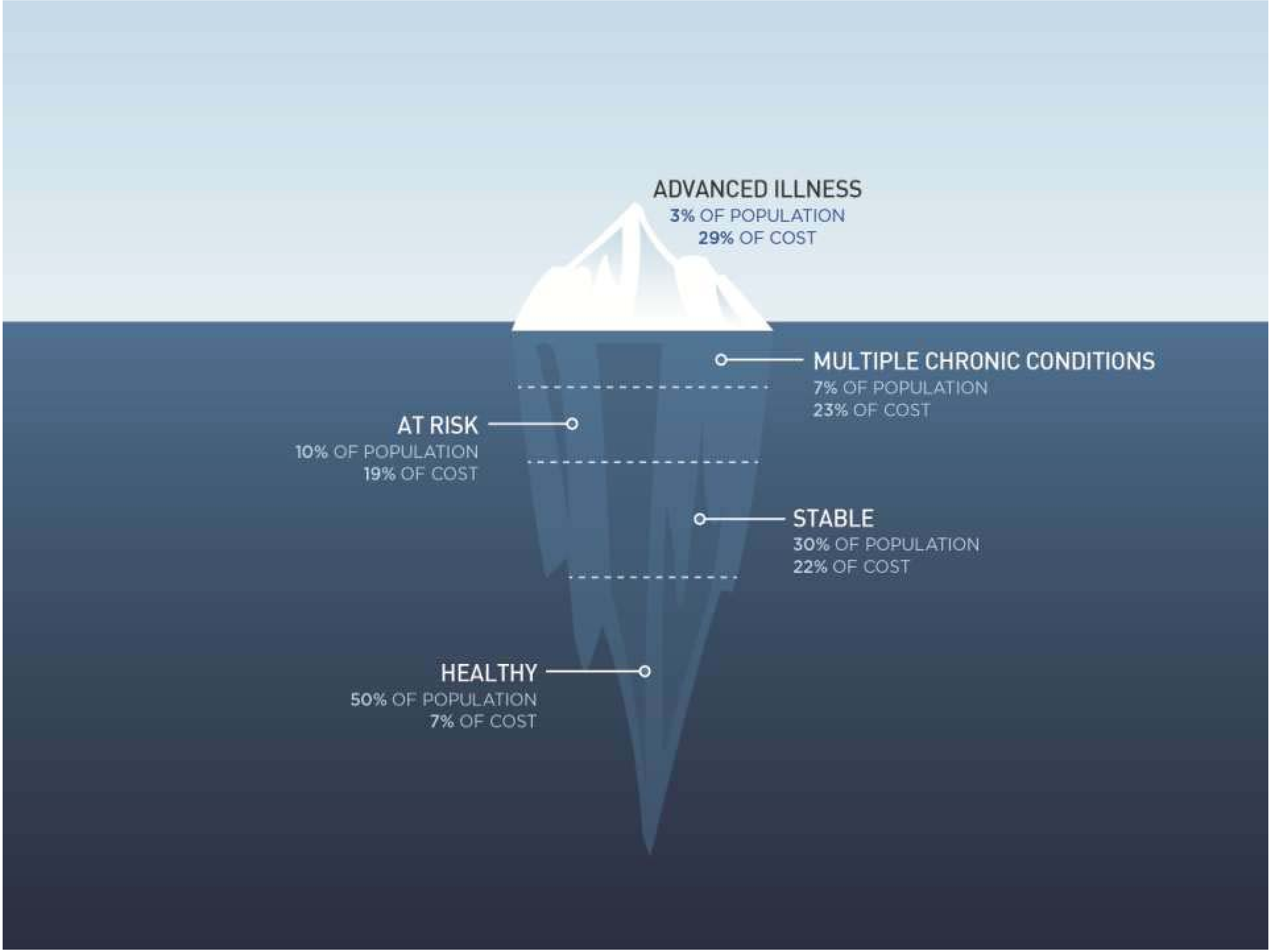
Electronic Encounter Data

vC
vCore

Data Warehouse

vN
Vision

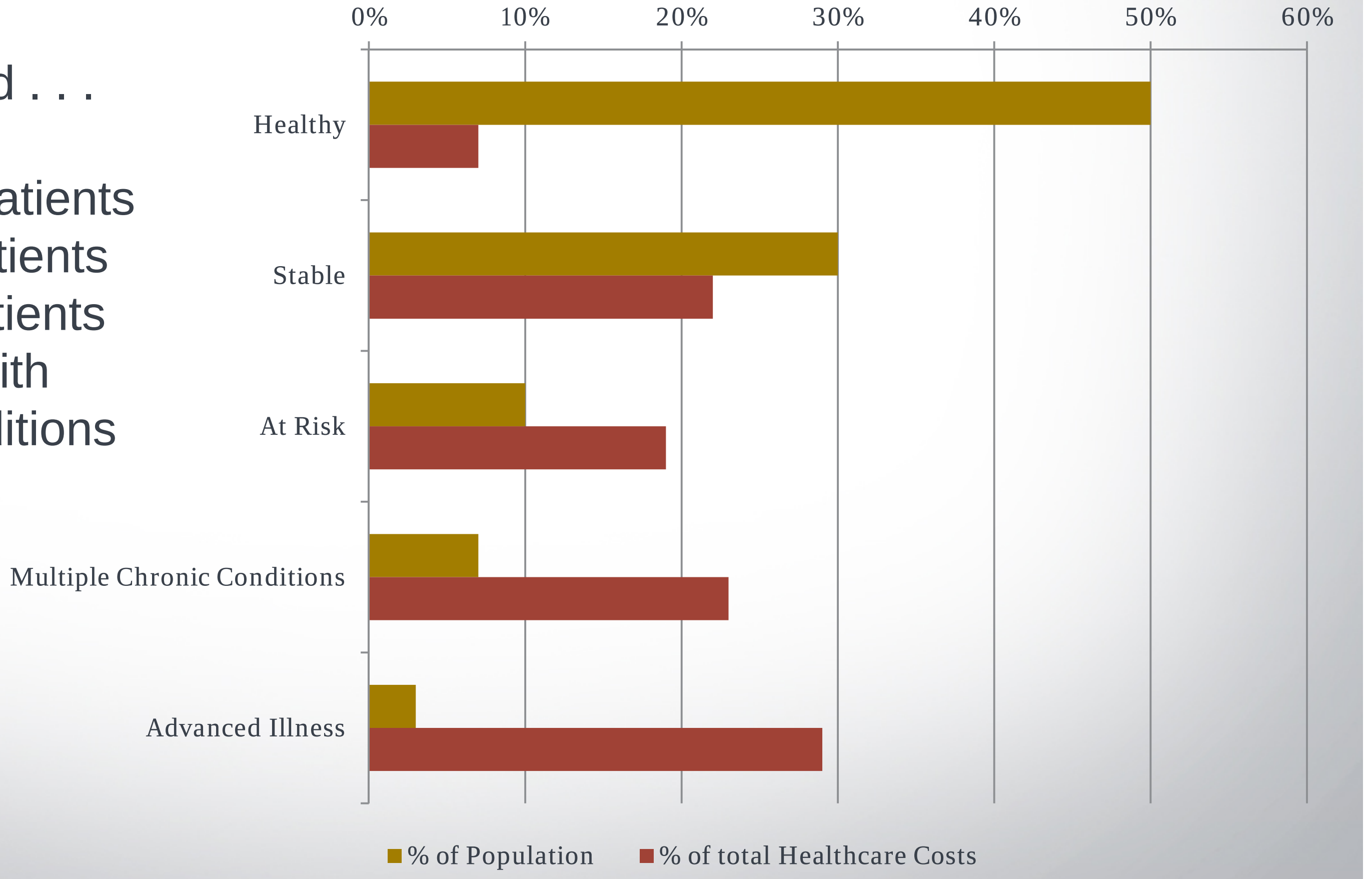
Reports and Support Tools


ADVANCED ILLNESS
3% OF POPULATION
29% OF COST
MULTIPLE CHRONIC CONDITIONS
7% OF POPULATION
23% OF COST
AT RISK
10% OF POPULATION
19% OF COST
STABLE
30% OF POPULATION
22% OF COST
HEALTHY
50% OF POPULATION
7% OF COST

The Sickest 10% Account for Half of Healthcare Costs

Per Capita: Patients with Advanced Illnesses spend . . .

70 x more than Healthy patients
13 x more than Stable patients
5 x more than at Risk patients
3 x more than Patients with multiple chronic conditions



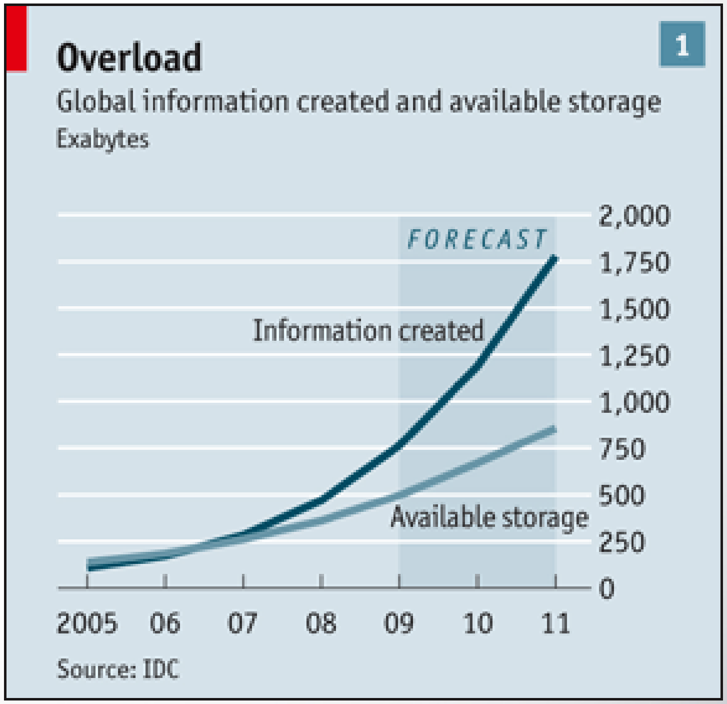
Data, Data and More Data

There is no shortage of data to review.

In 2010 enterprises stored 7 BILLION gigabytes of data.

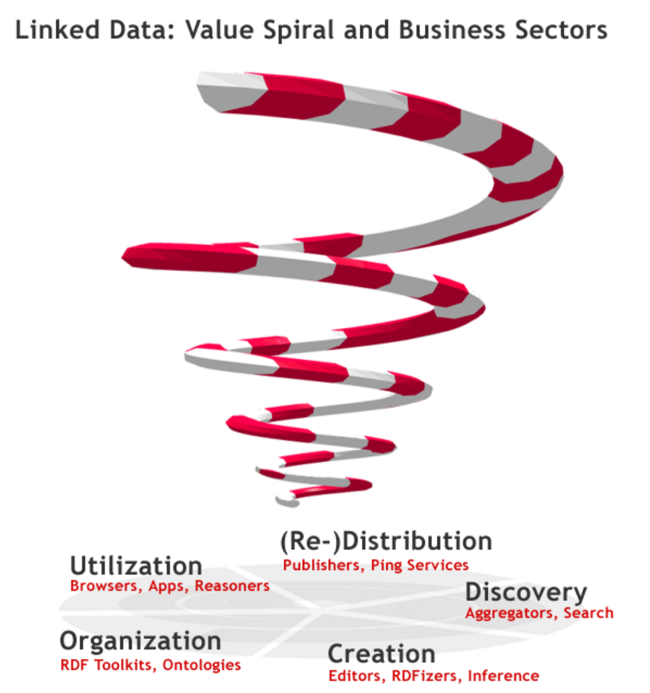
90% of the worlds data has been generated in the past 2 years[2]

In recent years Oracle, IBM, Microsoft and SAP between them have spent more than $15 billion on buying software firms specializing in data management and analytics



Linking Data – The Industry

- **Data mining was $100 Billion industry in 2010, with10% annual growth[1]**
- **Over 168 companies provide consulting on mining and/or analytics products[2]**
- **Data-driven Industries:**
  - Technology
  - Insurance
  - Sports teams
  - Financial
  - Marketing
  - Medicine


Linked Data: Value Spiral and Business Sectors

Vendors – Link Plus

- **Offer Registry Plus Software**
  - Developed by Center for Disease Control (CDC) for the National Program of Cancer Registries (NPCR).
    - **De-duplicates cancer registry data**
    - **Links cancer registry with an external file**
  - Cost effective
    - **Low low price of $0.00**
    - **Easy to use**
    - **Robust**
- **Input: Last name, First Name, SSN, DOB, Sex**
- **http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm**

Vendors – AutoMatch

- Uses probabilistic logic for matching
- Uses iterative, multiple pass executions
- Does better when greater sensitivity or overall Accuracy is desired.
- Input data: SSN, Last name, First name, DOB, Race, Phone#, Sex
- www.netrics.com

Vendors –Netrics

- **Netrics (Tibco) Matching Engine**
  - In-memory database search application that can be attached to virtually any data source including Oracle, Microsoft SQL Server, IBM DB2, MySQL, and many others.
  - The engine can provide sustained real-time, highly accurate search capabilities for small, medium, large and really humongous databases.
    - **Can handle any size database (billions of records) with sub-second latency.**
- **Input data: First name, last name, street, city, zip, state**
- **www.Netrics.com (http://www.tibco.com/products/automation/application-integration/pattern-matching/default.jsp)**

Vendors – SAS DataFlux

- **Uses a Customer Data Integration**
  - Combines a customer data repository, a tightly-integrated data quality solution, and a service-oriented architecture (SOA).
  - With these components, an organization can:
    - **Build a central reference file for customer data (the repository)**
    - **Create accurate and consistent information within the reference file (using data quality technology)**
    - **Build a way to share customer data throughout the organization (with the SOA)**
- **www.dataflux.com**

Linking Challenges in Healthcare

- **Sensitive data**
  - Privacy – not all info can or is willingly shared
  - SSN has a decreasing value
- **Limited data**
  - Different data elements are available from different data sources
- **Non unique demographic information and standardizing challenges**
  - How many John Smiths are there really?
  - Apt vs apartment or street vs st.
- **Data entry errors**
  - Fat fingers
    - 123555789 vs 123456789 means off by 2/9 = 22.2%.
    - 123567892 vs 123456789 means off by 1/9 = 11.1%
- **Population specific challenges**
  - Children
  - Illegal immigrants
  - Married women

Linking Challenges – SSN

- **9 digit numerical codes (ex: 876-54-3210)**
  - First three digit represent the state and states have ranges
    - **E.g. CA is 900s**
  - Next two are the office that dispensed the number
  - The last four are non-randomly assigned
- **National identifier since FDR administration – Mid 1930s**
  - Considered the most powerful piece of information about a person.

As a patient identifier

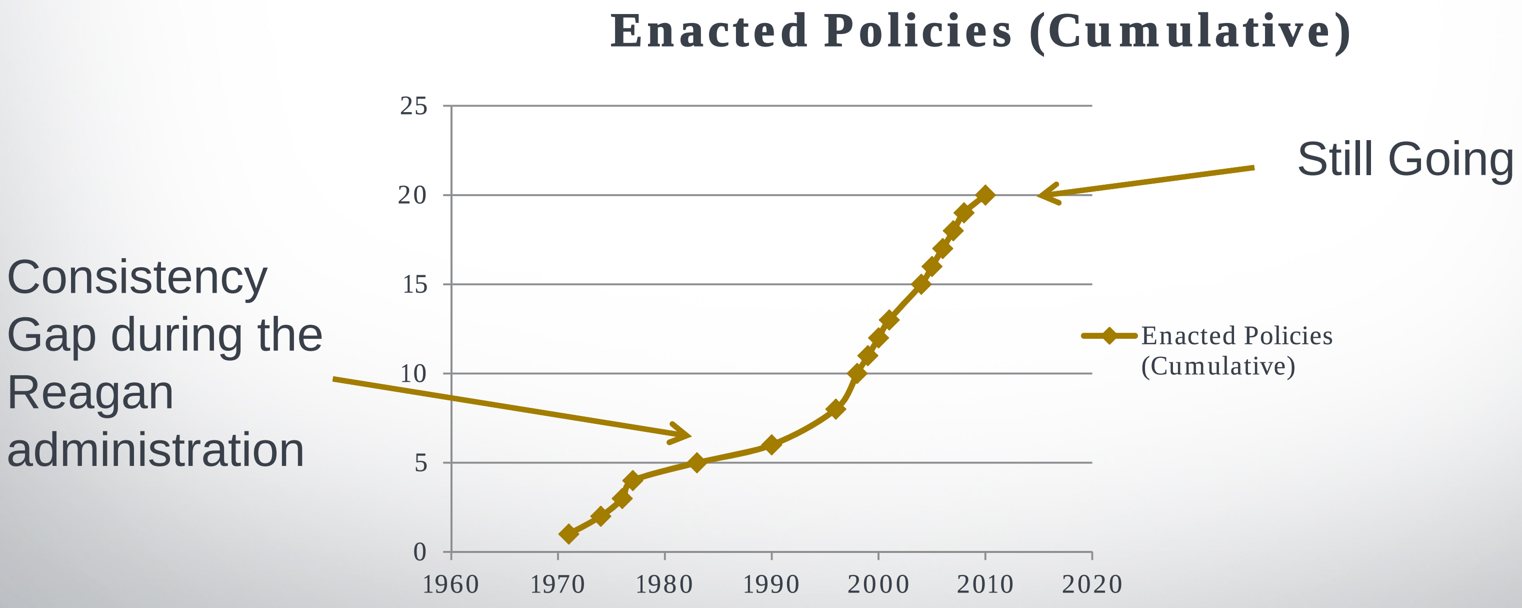
- Next to name, address, sex, and birth date, the Social Security number is probably the most frequently collected piece of information.

Linking Challenges – SSN

- **Pros**
  - All living U.S. citizens have a unique SSN
    - **Making it easy to organize and identify**
  - Commonly captured
  - Easily stored and indexed
  - People generally remember
- **Cons**
  - Leading cause of identity theft. (ex: If you forget the password to your bank account, some banks ask for your SSN as one of the ways to log back in)
    - **Sacrificing personal privacy because of the mistaken impression that nothing better is available.**

Linking Challenges – SSN

- **History of Congressional SSN Restriction**
  - Over time, Congress has (incrementally) restricted the usage of SSN.
  - Legislation passed overtime restricting SSN usage:



Linking Challenges – SSN

- 5 States either restrict the solicitation of SSNs or prohibit denying goods and services to an individual who declines to give an SSN
- 19 States restrict the printing of SSNs on ID cards required to access products or services
- 22 States restrict intentionally communicating SSNs to the public and/or intentional public posting and display
- 17 States restrict mailing of SSN’s within the mailing envelope

Conclusion on (f)Utility of SSN?

Phasing out use of SSN is like the setting sun: Interesting but you better prepare for the dark

Linking Challenges -- Limited Data

- There are myriad data sources which must be linked to provide a picture of a given patient's medical treatment. An incomplete list includes Payor, Pharmacy (Prescription Benefit ManAge or Retail), Laboratory, Hospital and Professional Services.
- Each data source has characteristics which make linking a challenge
- Several examples:
  - phone numbers are not provided from lab sources
  - Some practices don't collect address information

Linking Challenges – Non Unique Values and Standardization

- **It is not uncommon to see different people with the same name**
- **Bad SSNs can be commonly used**
  - 111111111, 222222222, 333333333, etc
  - Values need to be cleansed
- **Address Information**
  - &prefix.address1=tranwrd(&prefix.address1," ALLEY"," ALY");
  - &prefix.address1=tranwrd(&prefix.address1," ANNEX"," ANX");
  - &prefix.address1=tranwrd(&prefix.address1," ARCADE"," ARC");
  - &prefix.address1=tranwrd(&prefix.address1," AVENUE"," AVE");
  - &prefix.address1=tranwrd(&prefix.address1," BAYOU"," BYU");
  - &prefix.address1=tranwrd(&prefix.address1," BEACH"," BCH");
  - &prefix.address1=tranwrd(&prefix.address1," BEND"," BND");
- **USPS data source to drive consolidation**

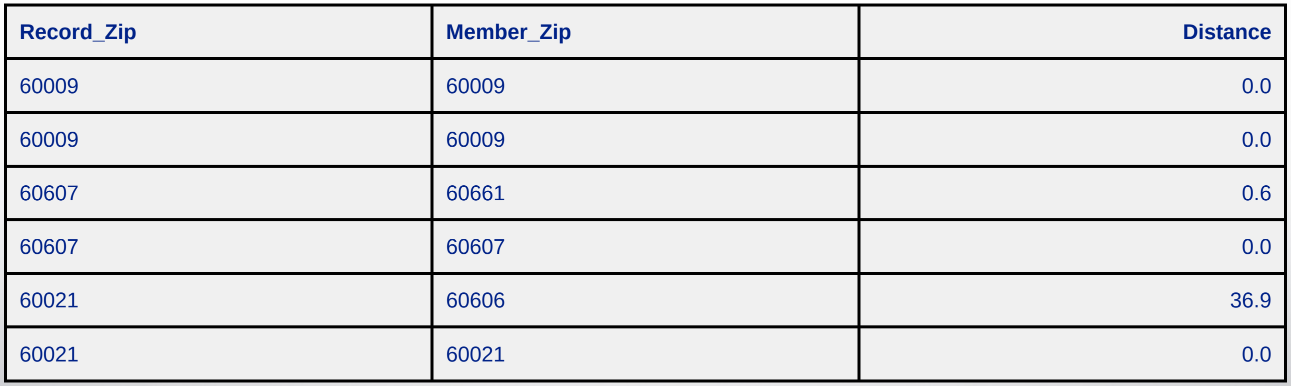
SAS CODE SNIPPET

SAS Code to Compare Distance

- Distance = zipcitydistance(Record_Zip,Member_Zip);

- Currently, we accept a proximity match for 0 <= Distance <=10

- Examples below



Record_Zip	Member_Zip	Distance
60009	60009	0.0
60009	60009	0.0
60607	60661	0.6
60607	60607	0.0
60021	60606	36.9
60021	60021	0.0

Linking Challenges – Non Unique Values

- **How many John Smiths are there?**
- **Common Names from a 13,288,308 person sample**

Name	Occurrences	% of Total
Jayne Doe	2058	0. 015%
James Smith	1602	0.012%
Robert Smith	1489	0.011%
Mary Smith	1144	0.009%
Smith, Johnson, Miller, Rodriguez, Garcia as surnames	1098	0.008%

Linking Challenges – Non Unique Values and Standardization

- Most Common Name for Valence Clients

Client	Name (LAST.FIRST)	Occurrences	% of total
A	GARCIA, MARIA	378	0.030%
B	HERNANDEZ, MARIA	152	0.039%
C	DOE, JAYNE	2057	0.243%
D	SMITH, JAMES	1241	0.014%
E	KIM, YOUNG	197	0.014%

Challenges – Sub Populations

- **Children**
  - Newborns and young children often use parent's SSN or don't have complete info at all
    - **80% of children 10 and under, 67% of children age 11-20**
- **Last Name Changes**
  - Marriage rate: 6.8 per 1,000 total population[1]
    - **Divorce Rate: 3.4 per 1,000 population[1]**
- **Biases for data completeness on sub populations**
  - Illegal immigrants are more likely to be undocumented/uncounted
  - Sicker/older populations are more likely to seek care
  - More affluent populations are more likely to have health insurance
- **Younger populations are more likely to change address[2]**
  - 18% of people age 16-24 move each year versus 11% age 25-64 and 3% over 65

[1] http://www.cdc.gov/nchs/fastats/divorce.htm

[2] http://www.census.gov/hhes/migration/data/cps/cps2011.html

SEGWAY

Moving

Right

Along

Common Tricks for Linking Patient Information

- **First , Last, Middle names**
  - “Sounds like” – SOUNDEX Function
  - Nicknames
  - Name reversal (last name flipped with first name)
  - Mother’s maiden names
- **Date of Birth**

  -Use month & Day

  -Use transposing digits

  -Use consistency in date style and order (Month/Day/Year)
- **Social Security Number**

  -Use transposing digits

SAS CODE SNIPPET

SOUNDEX Function

- The SOUNDEX function encodes a character string according to an algorithm that was originally developed by Margaret K. Odell and Robert C. Russel (US Patents 1261167 (1918) and 1435663 (1922)).
- The SOUNDEX algorithm is English-biased and is less useful for languages other than English.
- **Step 1**: Retain the first letter in the argument and discard the following letters:
  - A E H I O U W Y
- **Step 2**: Assign the following numbers to these classes of letters:
  - 1: B F P V
  - 2: C G J K Q S X Z
  - 3: D T
  - 4: L
  - 5: M N
  - 6: R
- **Step 3**: If two or more adjacent letters have the same classification from Step 2, then discard all but the first. (Adjacent refers to the position in the word before discarding letters.)

Common Tricks for Linking Patient Information

- **Address standardization is important and available thru the help of two sources**
  - **CASS (Coding Accuracy Support System)**
    - The customer address information with the USPS address database**.**
  - **NCOA (National Change of Address)**
    - compares the customer address information with the USPS Move Update Database.
      - If an exact match is found, then the customer's address information is updated with the new address

Advantages and Functions of CASS

- The input of:

1 MICROWSOFT
REDMUND WA

- Produces the output of:

1 MICROSOFT WAY
REDMOND WA 98052-8300

- Here the street and city name misspellings have been corrected
  - street suffix, ZIP code and ZIP+4 add-on have been added; and, in this case the address was determined to be the location of a business

- **CASS software can also return descriptive information about the address.**
  - If the address was successfully processed, or if not, why not
  - Information on how to deliver the mailing.

Other Common Elements to Use

- Personal e-mail addresses
- Internet user IDs and passwords
- Driver's license numbers
- Insurance Policy ID
- Relationship status
- Ordering Provider

SEGWAY

Moving

Right

Along

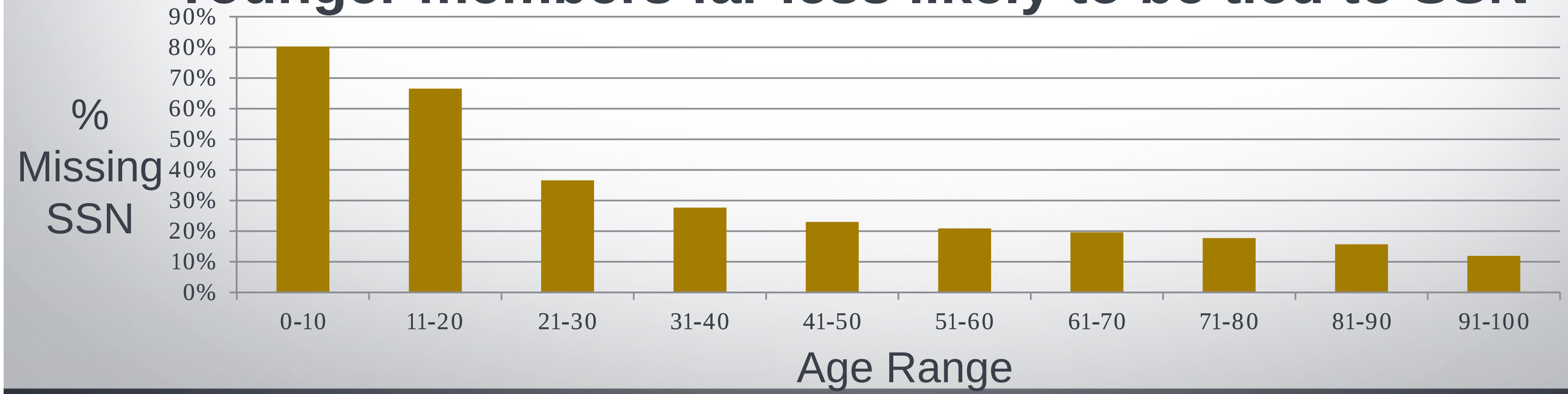
The Evolution of vBond

- **First, we used SSN as the primary key and deterministic linking**
  - Pros – easy to implement and reliable -- .4% false positive rate
  - Cons – decreasing population and other challenges already mentioned
- **Then we developed a probabilistic approach**
  - Based off of work done by National Center of Health Statistics (NCHS)*
- **Then we transitioned to a Bayesian Approach**
  - Leverages conditional probabilities for more reliable matching
- **Then we developed a deterministic 2nd pass to fix/find more matches made earlier in the process**

*National Center for Health Statistics. Office of Analysis and Epidemiology, The National Health Interview Survey (1986-2004) Linked Mortality Files, mortality follow-up through 2006: Matching Methodology, May 2009. Hyattsville, Maryland. (Available at the following address: http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf)

SSN as the Primary Key

- **We reviewed SSN data for Valence Client Membership**
  - SSN was not available for between 23% and 81% of client roster (missing for 32% for all members)
- **Of those 68% of members with non-null SSN . . .**
  - SSN was unique for between 78% and 95% of client rosters (average of 93% Unique SSNs for all members)
- **Younger members far less likely to be tied to SSN**



Developing a Probabilistic Approach

- **Probabilistic Matching - Process of using statistical methods to determine the overall likelihood that two records truly match.**
  - Preferred method for matching large data sets or when a large number of attributes are involved in the matching process.
  - Example: An uncommon name such as “Barack Obama” is less likely to appear in the database than a “John Smith” and thus has a higher weight when a match is found.

Probabilistic Approach – NCHS Approach

- **Step 1 – Sufficiency test – Confirm records have the following:**
  - SSN, Sex, DOB
  - Last name, first name, month of birth, year of birth
  - Last name, first initial, SSN
- **Step 2 – Creating the match**
  - Works through 7 different methods
- **Step 3 – Create the match score**
  - Probabilistic statistics
- **Step 4 – Determine if match is maintained**
  - Comparison of score to threshold

Probabilistic Approach – NCHS Approach

- **Step 3 – Create the Match Score[1]**

$$\mathit{Score} = \{\Sigma W_{SSN1} + \ldots + W_{SSN9}{}^{4}\} + W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthdate} + W_{birthmonth} + W_{birthyear} + W_{stateofbirth} + W_{stateofresidence}$$

NCHS developed weights known as binit weights, based upon the frequency of occurrence of the 12 data items in the NDI files for years 1979 to 2000, which represents about 49 million persons.

Weights = [Log^2 (1/p^i)] : the base 2 logarithm of the inverse of the probability of occurrence of the value of the identifying data item on the submission record

vBond – Probabilistic Approach

- **Step 1 – Data Cleaning**
  - Variable standardization
    - First name, last name, address, and city variables are screened for non-character values and those values are removed.
    - First name and last name variables are converted to sound functions to avoid spelling discrepancies in potential matching.
    - Phone numbers of any length other than 10 and false numbers are set to missing values.
    - Zip codes retain only the first 5 digits if less than length 5 or if unknown to the USPS registry, the fields are set to missing values.

vBond – Probabilistic Approach (cont'd)

- **Step 2 – Probability Creation**
  - Data variables counted and assigned percents based on probability of occurrence, which are merged to respective fields. These fields are used to calculate probabilistic matching likelihood if variable value successfully matches to another.
- **Step 3 – Data Matching**
  - Incomplete claim lines are matched against complete lines using various combinations of the aforementioned variables.

vBond – Probabilistic Approach (cont'd)

- **Step 4 – Threshold Comparison**
  - The matched claim lines are separated into classes based on probabilistic scoring. The user defines a minimum threshold allowance for acceptable matches and those matched claim lines exceeding the predetermined value are assigned the permanent member identification value.
- **Step 5 – Non-matched Member Identification**
  - Members containing complete identification data with no presence of SSN are then assigned unique identification values (Note: Recall, members finding eventual matches to SSN based members are assigned member ID of 1 + SSN) of 5 + unique 9 digit number.
  - Remaining claim lines without assigned member IDs are then submitted back through steps 3-4, where user can establish new minimum classification standard for matches.

Probabilistic Approach: Example

Score

28

Field	Input	Comparison	Comments
SSN		123456789	
First Name	BILL	WILLIAM	Nickname Match
Last Name		JENSEN	
Date of Birth	1/1/1931	1/1/1931	
Gender		M	
Address Line 1	123 MAINSTREET #123	123 MAIN ST #123	Fuzzy Match
City		CHICAGO	
State			
Zip	60607	60661	Proximity Match
Phone	123-456-7890	123-456-7890	

76 Year Old Male (at Date of Service) with 5 Input Cells
=> Required Scoring Threshold ≥ 23
=> Linked Record

Valence Health

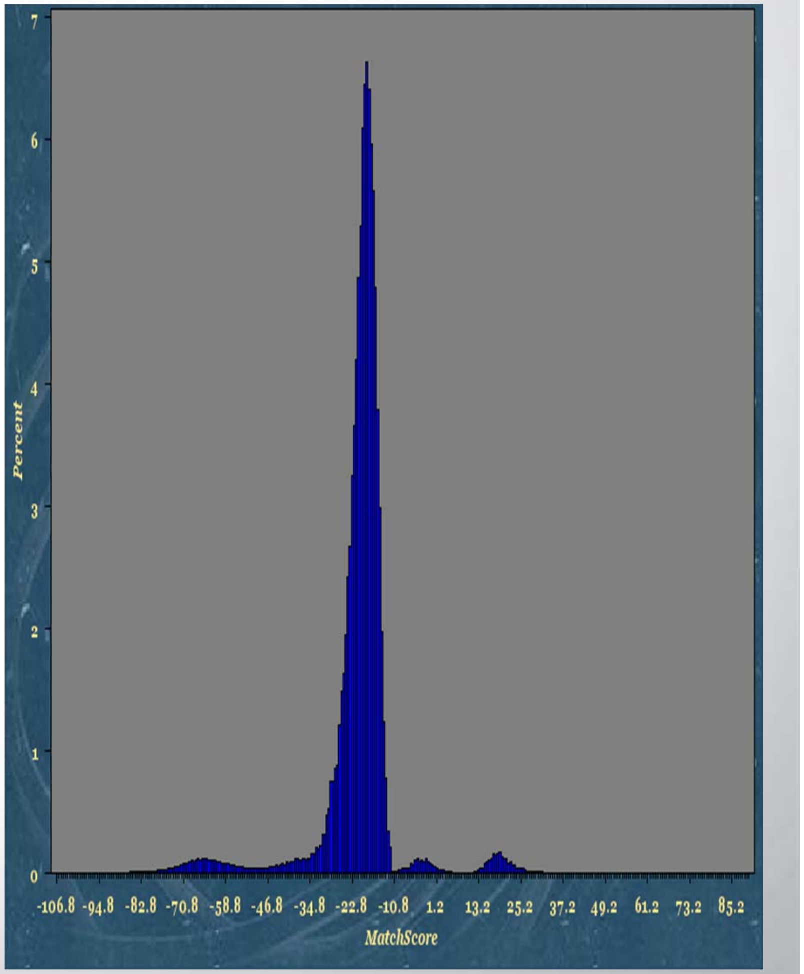
BETTER CARE IS ELEMENTAL.

Probabilistic Approach: A Starting Point

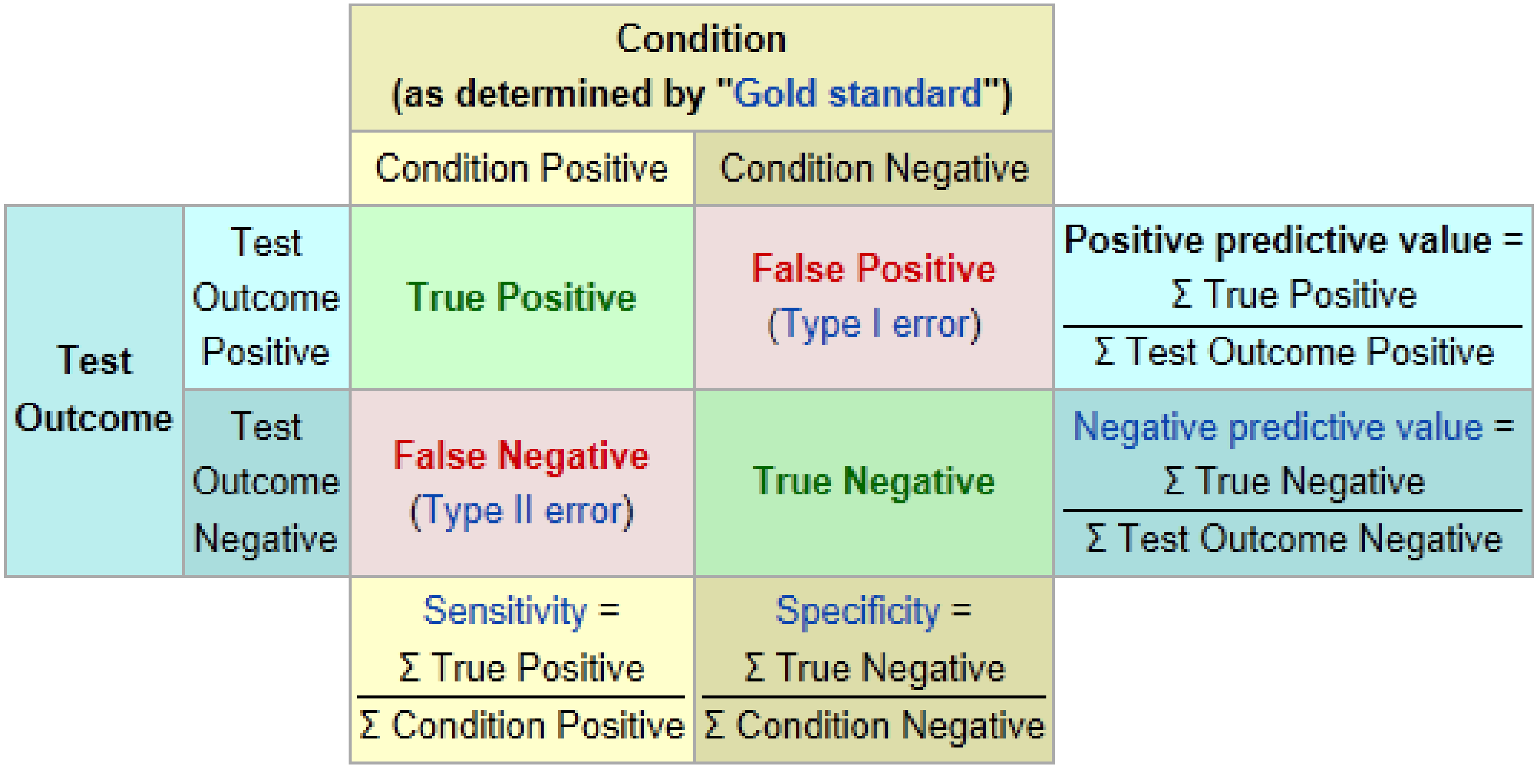
TPR	PPV	FPR	NPV
98.9%	99.9%	1.5%	86.5%

Definitions:

- **TPR:** true positive rate, sensitivity, or ability to identify potential matches
- **PPV:** positive predictive value or ability to correctly confirm a match
- **FPR:** false positive rate, 1-specificity, or rate of mismatch
- **NPV:** negative predictive value or ability to correctly confirm a non-matching combination



Sensitivity, Specificity, and Positive/Negative Predictive Values

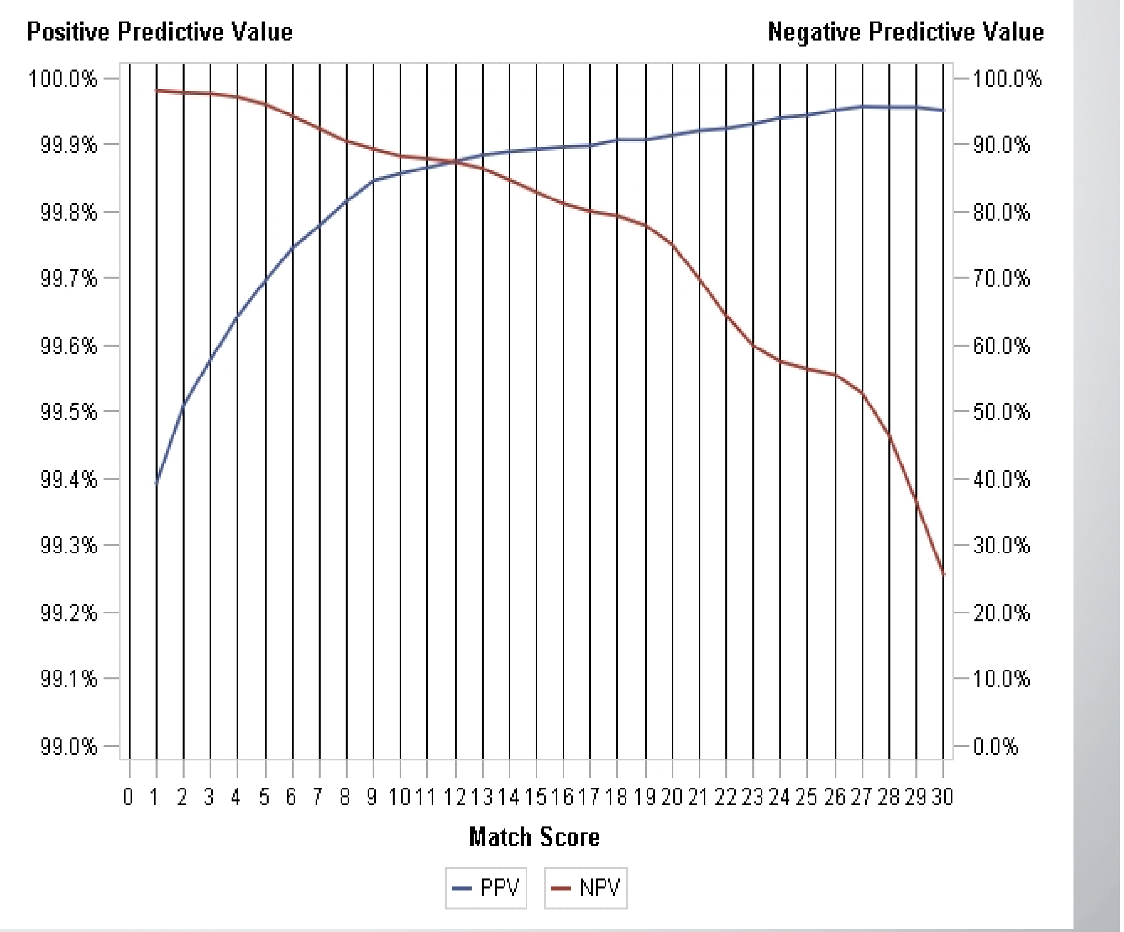


		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	**Positive predictive value** = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Positive and Negative PV by Match Score

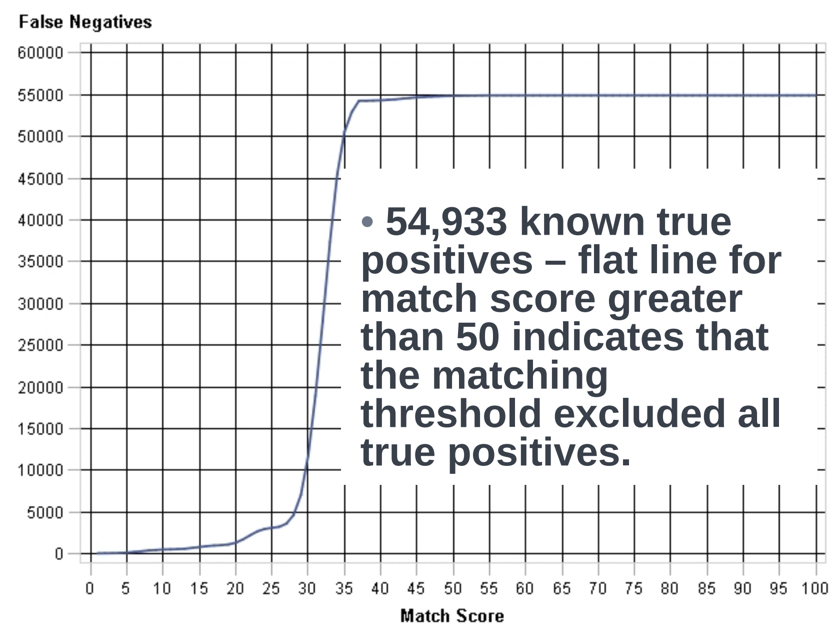
- **Positive predictive values increase and negative PV decrease with increasing match score**

Match Score	7	10	26
PPV	99.8%	99.9%	100%
NPV	92.4%	88.3%	55.6%

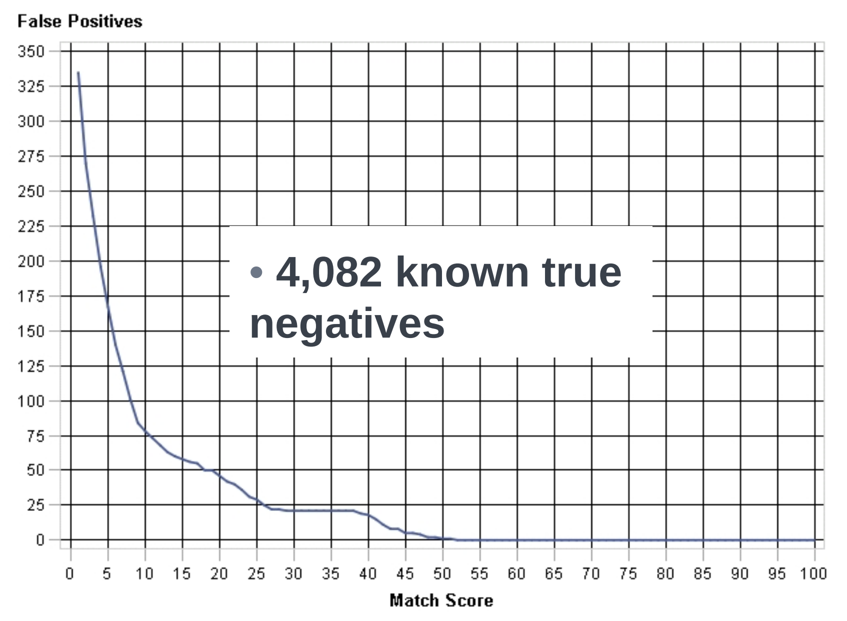


- Optimization for both predictive values is reached when match score equals 10.

Incorrectly Determined Results



- 54,933 known true positives – flat line for match score greater than 50 indicates that the matching threshold excluded all true positives.



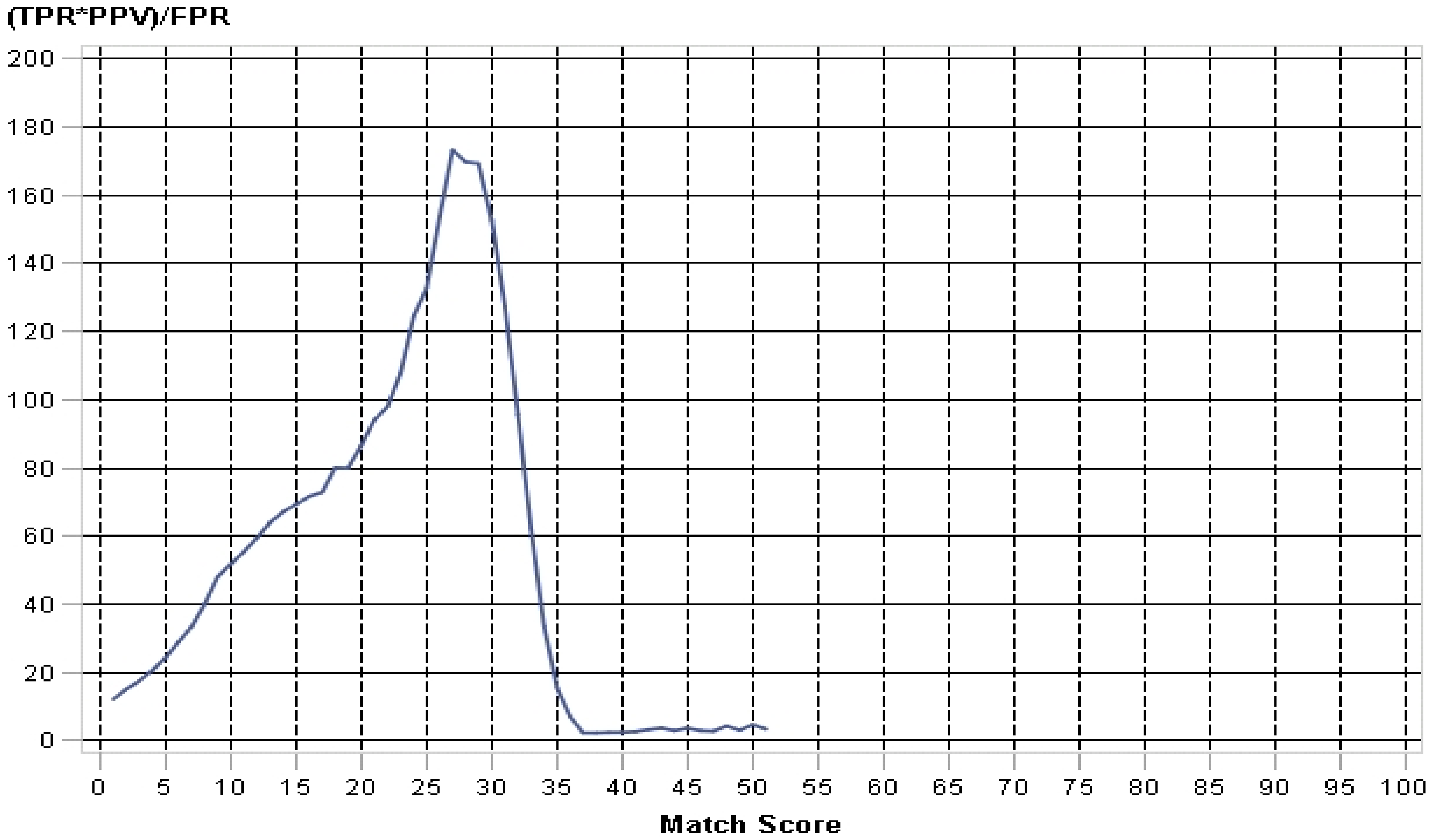
- 4,082 known true negatives

Notable false negative and false positive statistics by match score

Match Score	7	10	26
False Negatives	326	533	3,245
True Positive Rate	99.4%	99.0%	94.1%
False Positives	121	78	25
False Positive Rate	3.0%	1.9%	0.6%

Valence Health BETTER CARE IS ELEMENTAL.

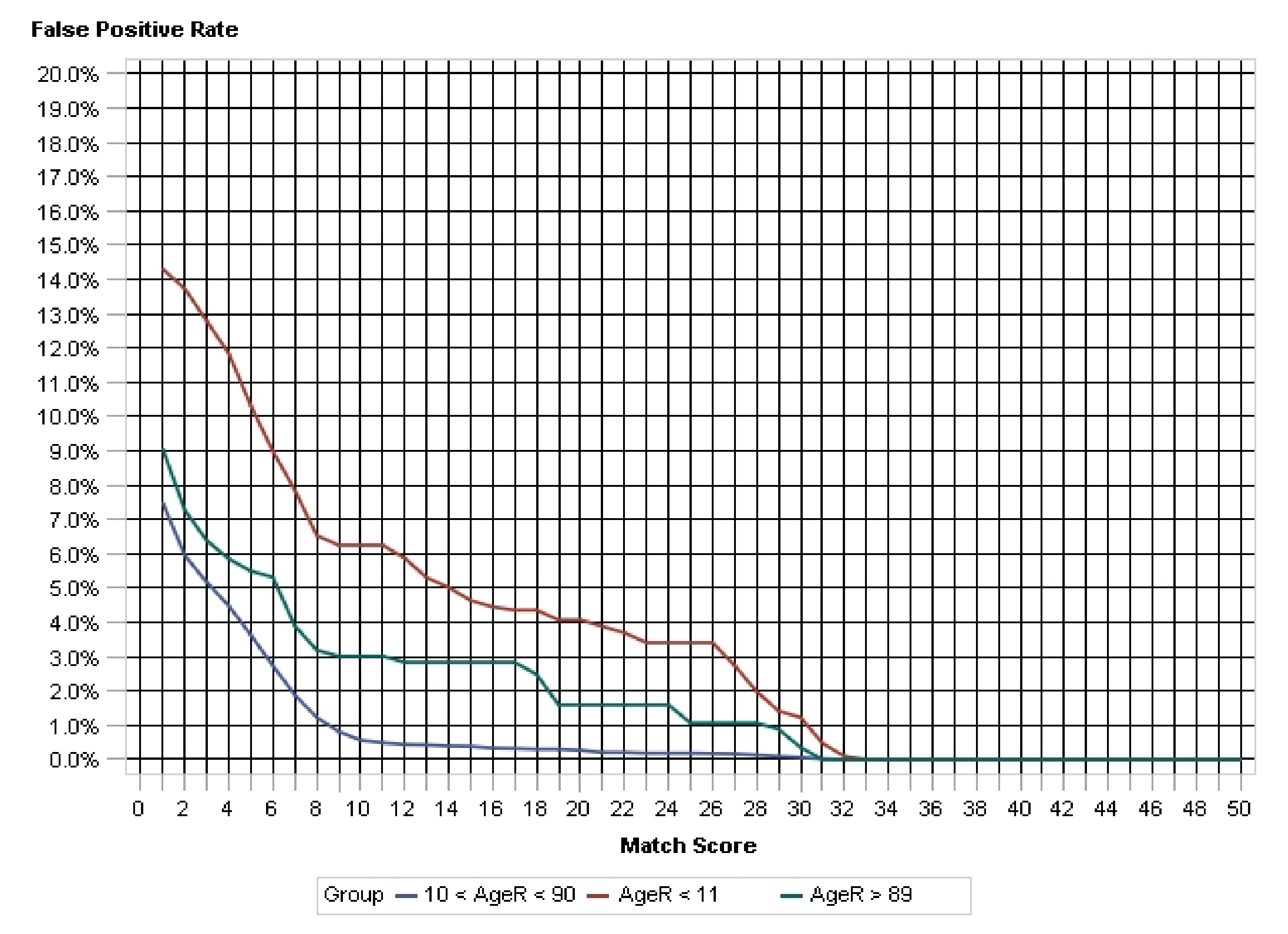
What Metric Will Guide Us?



Valence Health

BETTER CARE IS ELEMENTAL.

Validation Additions – Age Analysis



Valence Health BETTER CARE IS ELEMENTAL.

The Decision – Age Group Scoring Thresholds for Positive Match Status

Group	Local Max	TPR	PPV	FPR	NPV
Age < 11	29	73.52%	99.53%	1.42%	47.53%
Age < 11	32	24.89%	99.91%	0.09%	24.46%
10 < Age < 90	18	98.09%	99.93%	0.30%	92.07%
10 < Age < 90	21	96.76%	99.95%	0.21%	87.25%
10 < Age < 90	23	95.17%	99.96%	0.19%	82.14%
10 < Age < 90	31	59.56%	99.99%	0.02%	35.48%
Age > 89	19	96.58%	99.72%	1.60%	83.21%
Age > 89	25	86.84%	99.79%	1.06%	56.42%
Age > 89	30	55.47%	99.89%	0.35%	27.82%

Group	Percent of Data
Age < 11	2.20%
10 < Age < 90	96.78%
Age > 89	1.02%

Probabilistic Approach: Final Thresholds

Based on the rate of false positive observance under the traditional SSN-based linking approach, we identified the following match scores to be acceptable thresholds

Age Group	Match Score
0-10	32
11-89	23
90+	30

Note that the same calibration technique produced different thresholds when applied to a different client

Age Group	Match Score
0-10	18
11-89	15
90+	26

Bayesian Approach

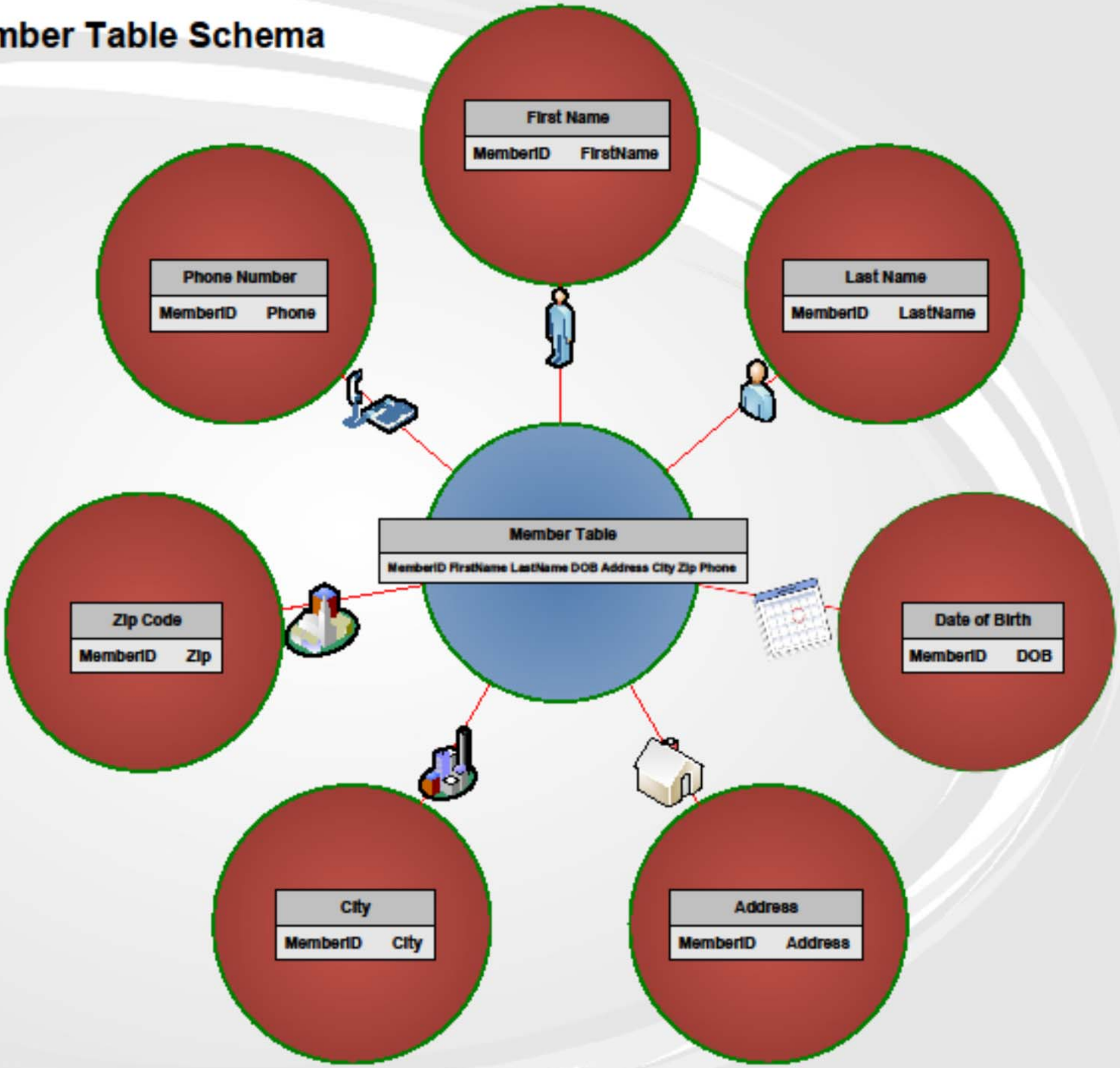
Thomas Bayes
(c. 1702 – April 17, 1761)

- Bayesian probability interprets the concept of probability as "an abstract concept, a quantity that we assign theoretically, for the purpose of representing a state of knowledge, or that we calculate from previously assigned probabilities," in contrast to interpreting it as a frequency or "propensity" of some phenomenon
- **Probability quantifies a "personal belief" that can evolve as new data becomes available**
- Bayes' theorem gives the relationship between the probabilities of ***A*** and ***B***, ***P(A)*** and ***P(B***), and the conditional probabilities of ***A*** given ***B*** and ***B*** given ***A***, ***P(A|B)*** and ***P(B|A)*** with the following formula.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Linking Algorithm Member Table Schema

LEGEND	
Member Table	The member table contains the single most common representation for every single field by member ID. It permits a single line of data for each unique member ID.
Alias Table	An alias table only contains values within the respective field which have ever been linked to the unique member ID. It permits an unlimited number of member ID and field combinations.



Member Table
MemberID FirstName LastName DOB Address City Zip Phone

First Name	
MemberID	FirstName

Last Name	
MemberID	LastName

Phone Number	
MemberID	Phone

Zip Code	
MemberID	Zip

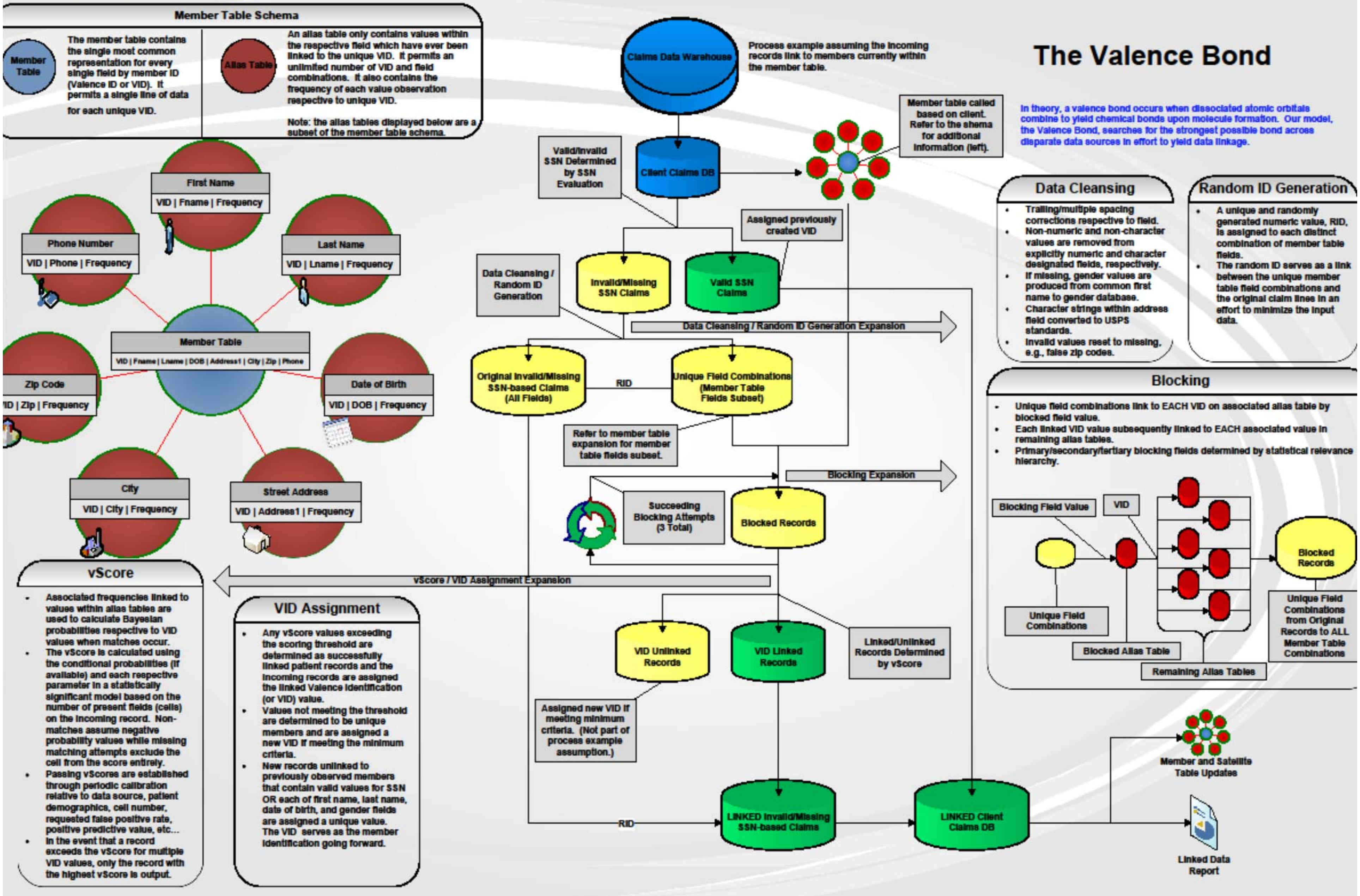
Date of Birth	
MemberID	DOB

City	
MemberID	City

Address	
MemberID	Address

The Valence Bond

In theory, a valence bond occurs when dissociated atomic orbitals combine to yield chemical bonds upon molecule formation. Our model, the Valence Bond, searches for the strongest possible bond across disparate data sources in effort to yield data linkage.



Member Table Schema

Member Table: The member table contains the single most common representation for every single field by member ID (Valence ID or VID). It permits a single line of data for each unique VID.

Alias Table: An alias table only contains values within the respective field which have ever been linked to the unique VID. It permits an unlimited number of VID and field combinations. It also contains the frequency of each value observation respective to unique VID.

Note: the alias tables displayed below are a subset of the member table schema.

Process example assuming the incoming records link to members currently within the member table.

Claims Data Warehouse
Client Claims DB
Valid/Invalid SSN Determined by SSN Evaluation
Member table called based on client. Refer to the shema for additional information (left).
Assigned previously created VID
Invalid/Missing SSN Claims
Valid SSN Claims
Data Cleansing / Random ID Generation
Data Cleansing / Random ID Generation Expansion
Original Invalid/Missing SSN-based Claims (All Fields)
RID
Unique Field Combinations (Member Table Fields Subset)
Refer to member table expansion for member table fields subset.
Blocking Expansion
Succeeding Blocking Attempts (3 Total)
Blocked Records
vScore / VID Assignment Expansion
VID Unlinked Records
VID Linked Records
Linked/Unlinked Records Determined by vScore
Assigned new VID if meeting minimum criteria. (Not part of process example assumption.)
RID
LINKED Invalid/Missing SSN-based Claims
LINKED Client Claims DB
Member and Satellite Table Updates
Linked Data Report

Data Cleansing

- Trailing/multiple spacing corrections respective to field.
- Non-numeric and non-character values are removed from explicitly numeric and character designated fields, respectively.
- If missing, gender values are produced from common first name to gender database.
- Character strings within address field converted to USPS standards.
- Invalid values reset to missing, e.g., false zip codes.

Random ID Generation

- A unique and randomly generated numeric value, RID, is assigned to each distinct combination of member table fields.
- The random ID serves as a link between the unique member table field combinations and the original claim lines in an effort to minimize the input data.

Blocking

- Unique field combinations link to EACH VID on associated alias table by blocked field value.
- Each linked VID value subsequently linked to EACH associated value in remaining alias tables.
- Primary/secondary/tertiary blocking fields determined by statistical relevance hierarchy.

Blocking Field Value
VID
Unique Field Combinations
Blocked Alias Table
Remaining Alias Tables
Blocked Records
Unique Field Combinations from Original Records to ALL Member Table Combinations

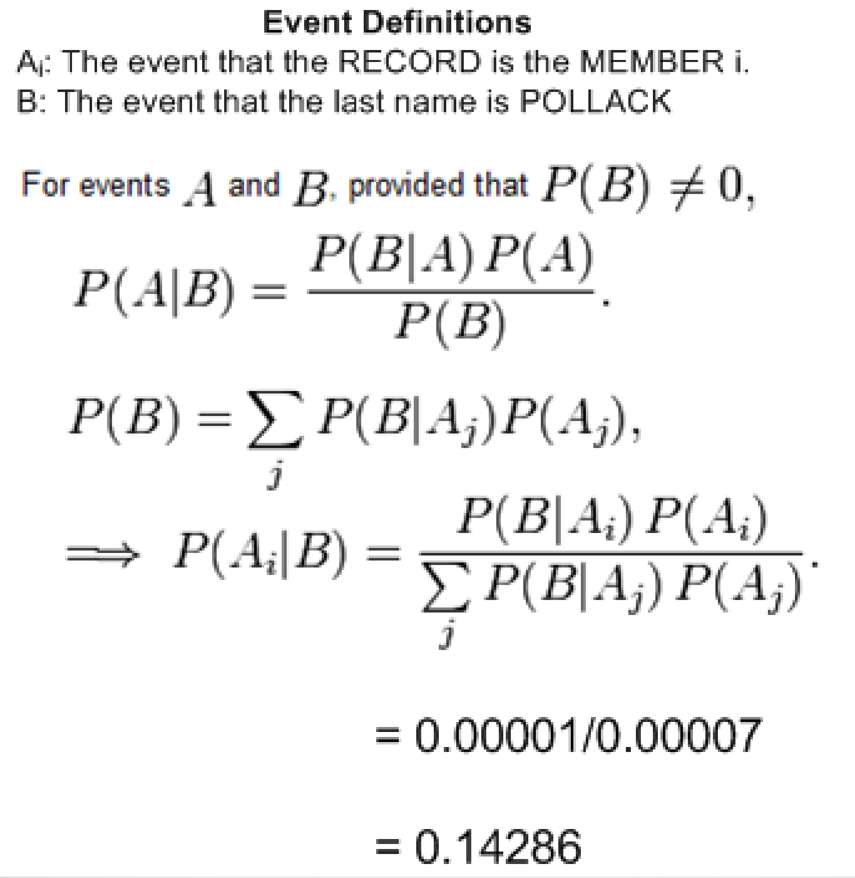
vScore

- Associated frequencies linked to values within alias tables are used to calculate Bayesian probabilities respective to VID values when matches occur.
- The vScore is calculated using the conditional probabilities (if available) and each respective parameter in a statistically significant model based on the number of present fields (cells) on the incoming record. Non-matches assume negative probability values while missing matching attempts exclude the cell from the score entirely.
- Passing vScores are established through periodic calibration relative to data source, patient demographics, cell number, requested false positive rate, positive predictive value, etc...
- In the event that a record exceeds the vScore for multiple VID values, only the record with the highest vScore is output.

VID Assignment

- Any vScore values exceeding the scoring threshold are determined as successfully linked patient records and the incoming records are assigned the linked Valence Identification (or VID) value.
- Values not meeting the threshold are determined to be unique members and are assigned a new VID if meeting the minimum criteria.
- New records unlinked to previously observed members that contain valid values for SSN OR each of first name, last name, date of birth, and gender fields are assigned a unique value. The VID serves as the member identification going forward.

Bayesian Analytics



Event Definitions

A_i: The event that the RECORD is the MEMBER i.

B: The event that the last name is POLLACK

For events A and B, provided that $P(B) \neq 0$,

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

$$P(B) = \sum_j P(B|A_j)P(A_j),$$

$$\implies P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_j P(B|A_j)\,P(A_j)}.$$

= 0.00001/0.00007

= 0.14286

Bayesian Linking Example

- Existing records in VH_EMPI
- Person key 1 for LEAHANNA MORALES, maps to patient key 101, with counter 10
- Person key 2 for LEAHANNA MORALE, maps to patient key 101, with counter 1
- Person key 3 for JOHN MORALES, maps to patient key 102, with counter 5

- Conditional probability of patient key 101 given fname=LEAHANNA is 11/11=1
- Conditional probability of patient key 101 given lname=MORALES is 10/15=.67
- Conditional probability of patient key 101 given lname=MORALE is 1/1=1
- Conditional probability of patient key 102 given fname=JOHN is 5/5=1
- Conditional probability of patient key 102 given lname=MORALES is 5/15=.33

- So, when LEANNA get matched to LEAHANNA based on SOUNDEX, the fname weight is the conditional probability that the patient key is 101 given the fname=LEAHANNA, which is 1

Bayesian Linking Example

weight1 = 0.002365 (lname=MORALES)

weight2 = 1 (fname=LEAHANNA)

weight3 = 0 (address does not match)

weight4 = -0.000183 (penalty for zip)

weight5 = 0 (Paula's record has no phone)

weight6 = -0.02222 (penalty for DOB)

weight7 = 0.000002146 (state=TX)

cells=5

mscore=0.7

matchscore=0.97996

matchscore > mscore AND matchscore > sum(weight1, weight4, weight7)

Deterministic 2nd Pass

- **Deterministic Matching** - a rules-based process to determine a match between two records.

- The process works best for simple, easily-defined matches.
  - Linking the same [Social Security Number, Phone Number, Name, Address, Driver's License/State ID Number, etc.] between two records.
- A considerable amount of data cleaning is performed BUT slightly different approaches are taken than prior to probabilistic methodology

Deterministic 2nd Pass

- **False Positives**
  - Fixing incorrect links by looking for different people within one patient key
- **False Negatives**
  - Fixing incorrect non-links by looking again across patient population but from a deterministic perspective

Deterministic 2nd Pass – False Positive

- Within a member key - compare each data element independently to find how many unique values
- Combine all data elements to perform a final comparison
- Combining multiple data elements to perform exact match comparisons allows false positives to be identified
- Algorithmic strategy needs to be aligned with database design

Deterministic 2nd Pass– False Negatives

- For each pair (or group) of member keys, keep the permutation that has the most number of variables.
- Then compare the values from one member key to other member keys
- Below grid is an example of what combination of variables, when they each match, would constitute match

sex_fname	lname	dob	phone	address1	zip	permutation
y	y	y	y			1
y	y	y		y		2
y	y	y			y	3
y	y		y	y		4
y	y		y		y	5
y	y			y	y	6
y		y	y	y		7
y		y	y		y	8
						9
	y	y	y	y		10
						11
y	y	y	y	y		12
						…

Deterministic 2nd Pass – False Negative

- If the first name, last name, and address of 2 records match, but the dob is different than could it be the son or daughter of the parent if the difference is greater than 16years or it could be a simple typo?

SAS CODE SNIPPET

```
%macro count_digit_diff(m_varprefix,m_var1,m_var2);
    if length(&m_var1.) le length(&m_var2.) then do; cdd1_&m_varprefix.=&m_var1.; cdd2_&m_varprefix.=&m_var2.; end;
    else do; cdd1_&m_varprefix.=&m_var2.; cdd2_&m_varprefix.=&m_var1.; end;
    &m_varprefix._digit_diff=length(cdd2_&m_varprefix.)-length(cdd1_&m_varprefix.);
    do m_i=1 to length(cdd1_&m_varprefix.);
        if substr(cdd1_&m_varprefix.,m_i,1) ne substr(cdd2_&m_varprefix.,m_i,1) then do;
            if m_i=length(cdd1_&m_varprefix.) then do;
                &m_varprefix._digit_diff=&m_varprefix._digit_diff+1;
                &m_varprefix._digit_1stdiff=min(&m_varprefix._digit_1stdiff,m_i);
            end;
            else if substr(cdd1_&m_varprefix.,m_i+1)=substr(cdd2_&m_varprefix.,m_i,length(substr(cdd1_&m_varprefix.,m_i+1))) or
                substr(cdd2_&m_varprefix.,m_i+1)=substr(cdd1_&m_varprefix.,m_i,length(substr(cdd2_&m_varprefix.,m_i+1))) then do;
                &m_varprefix._digit_diff=&m_varprefix._digit_diff+1;
                &m_varprefix._digit_1stdiff=min(&m_varprefix._digit_1stdiff,m_i);
                m_i=length(cdd1_&m_varprefix.)+1;
            end;
            else do;
                &m_varprefix._digit_diff=&m_varprefix._digit_diff+1;
                &m_varprefix._digit_1stdiff=min(&m_varprefix._digit_1stdiff,m_i);
            end;
        end;
    end;
    &m_varprefix._digit_diffpct=&m_varprefix._digit_diff/length(cdd2_&m_varprefix.);
    cdd1_&m_varprefix.=''; cdd2_&m_varprefix.='';
    drop m_i cdd1_&m_varprefix. cdd2_&m_varprefix.;
%mend count_digit_diff;
```

SEGWAY

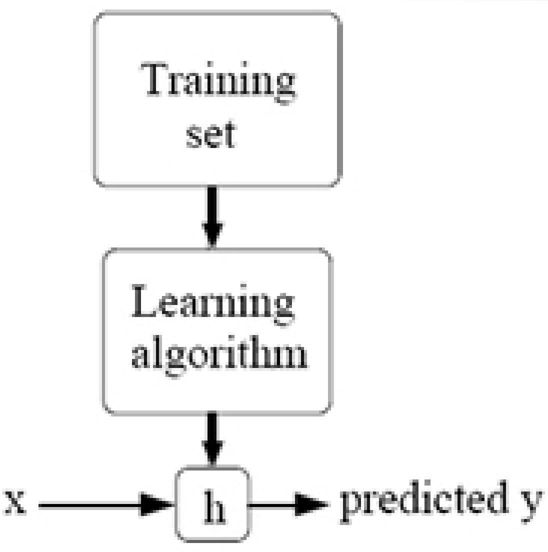
Moving

Right

Along

Machine Learning: The Future

- **Machine Learning** - A branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been fed into it.



- How can it be leveraged?
  - **Feeding it consistent information**

Machine Learning: The Future

- **Given:**
  - Machine learning solutions will only gain more intelligence with additional data and techniques.
  - In machine learning, the machine never stops learning.

Therefore:

- The potential and possibilities of machine learning are endless
- “In the future every business will be a data-driven enterprise” -Alexander Gray – CTO Skytree

Machine Learning: vBond Application

- Age and Cell count specific threshold set at 3
- A high percentage of these links are being member fixed based on deterministic phase
- Feedback of % of member fixes is feed back into program that sets age and cell count threshold
- Age and Cell count threshold is raised to 3.1
- % of member fixes is monitored and falls back to expected levels
- Age and Cell count threshold remains at 3.1

Machine Learning Universities Working with Companies

Universities	Businesses
University of Toronto	Google
University of Washington	Netflix
University of Michigan	Amazon
Carnegie Mellon University	Blizzard
University of Edinburgh	Valve
Ohio State University	Knewton
John Hopkins University	Symantec
Standford University	Sense Networks
Massachusetts Institute of Technology	Hunch.com

Acknowledgements

- **Omar ‘The Unstoppable Intern’ Hafeez**
- **Brandon ‘The Original’ Barber**
- **Tim ‘Street’ Dollear**
- **Brandon ‘Long Story Short’ Fletcher**
- **G ‘Mystery Man’ Liu**

Contact Information

Bart Phillips

bphillips@valencehealth.com