A COMPONENT-BASED DATA MASKING SOLUTION

PROCESS

Mask Personally Identifiable Information:

1. DISCOVER SENSITIVE DATA
2. FIGURE OUT THE TYPES OF ATTACK
3. DE_IDENTIFY: COMMON METHODS + REMOVE THE OUTLIERS
4. PROVE THE STATISTICS

FIRST, DEFINE THE MODEL

PII DEFINITION: WHAT DEFINES PERSON IDENTITY?

The term “PII,” as defined in OMB Memorandum M-07-1616 refers to information that can be used to **distinguish or trace an individual’s identity**, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual.
US General Services Administration

Personally Identifiable Information is a sensitive and critical organizational resource.


Social Security Numbers


Names


Credit Card Numbers


DOBs

IS THERE OTHER DATA?

YES!

TONS OF IT:

- Medications
- Results
- Vitals
- Problems

WHO WOULD KNOW THE DATA?

Researchers, grad students domestic and international

If your professor is on seven medications – can you find him?

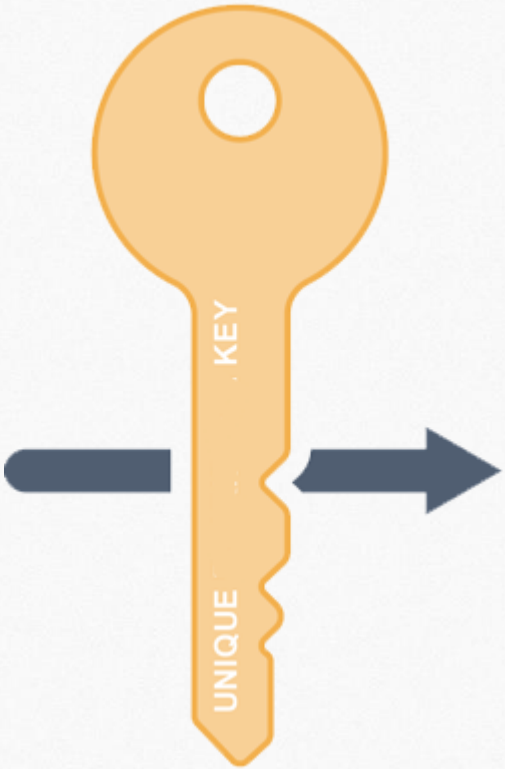
Will you trust Kim Jong-un with your meds?

PROBLEM 1: UNIQUE IDENTIFYING ELEMENTS

CIPHER IS NOT GOOD ENOUGH

UNIQUE DATA

- Social security number
- **(123-45-6789)**
- Passport number
- **(C00001234)**
- Credit card
- **(4234-5678-9123-4567)**
- Driver's license
- **(123-456-789)**
- Account Numbers, phones, faxes, VINs

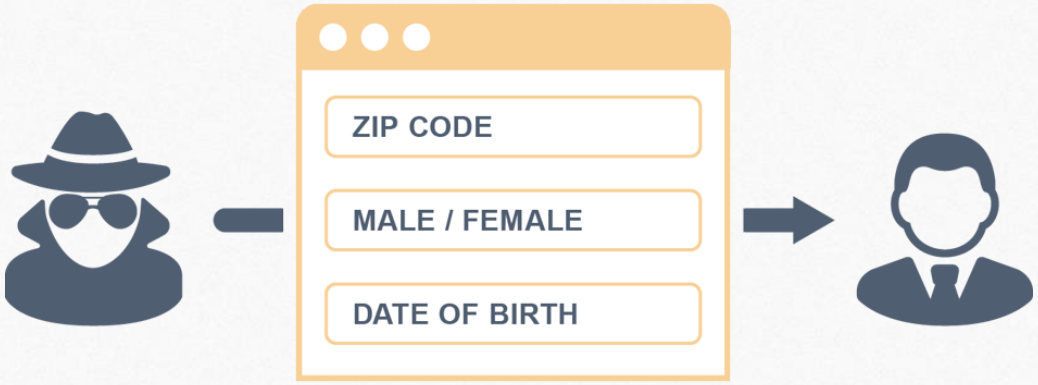


SDM: MASKED DATA

- **987-65-4321**
- **A00009876**
- **4276-5432-1987-6543**
- **654-987-321**

PROBLEM 2: STATISTICS AS AN ENEMY

PUBLIC DATA SETS: K-ANONYMITY, LATANIYA SWEENY



Statistics:

- **Zip code 27615: 45,090 people (Raleigh, NC)**
- **Over 65 → 5,190**
- **Female over 65 → 2,595**
- **Female over 65 with birthday on April 3 → 7**
- **DOB: April 3, 1946 → 1**
- **87% of people in the US can be re-identified with Gender, Zip & DOB**

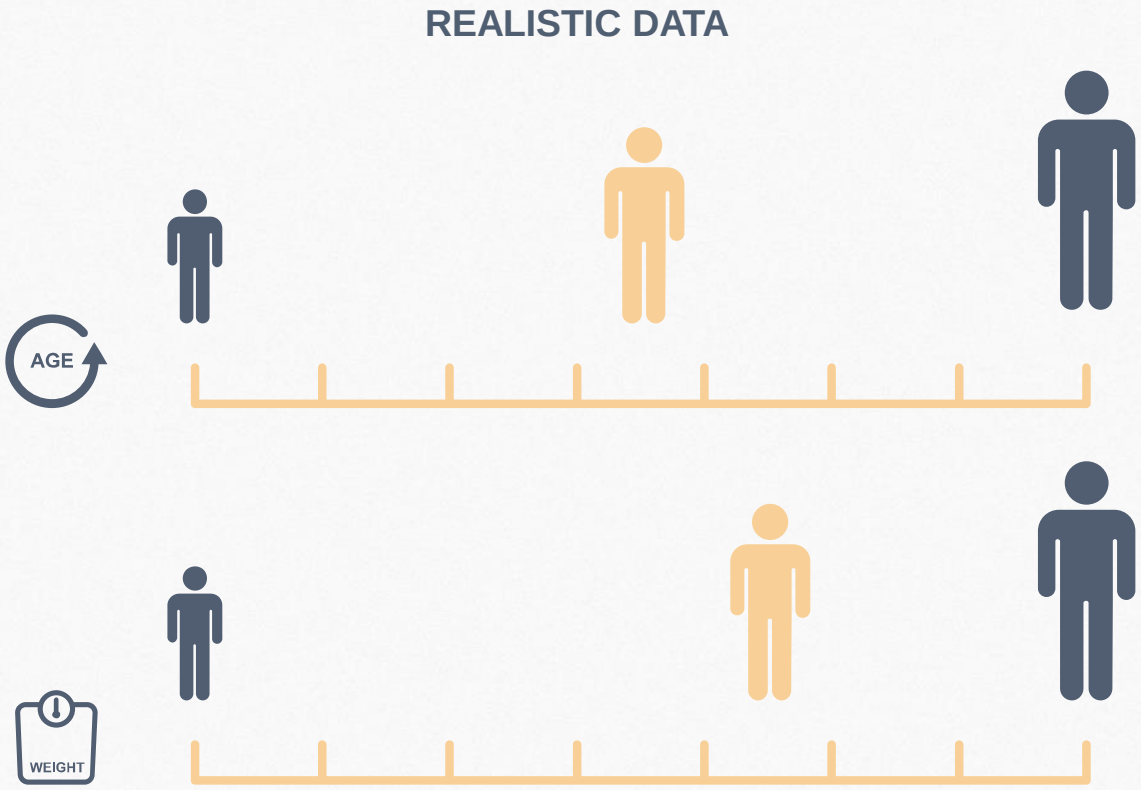
PROBLEM 3: YOUR NEIGHBOUR KNOWS THINGS ABOUT YOU

L-DIVERSITY and more

RELATIONSHIPS WITHIN THE ELEMENTS:

VACCINATION DATE > DATE OF BIRTH, HEIGH / WEIGHT > BMI

PROBLEM 3: YOUR NEIGHBOUR KNOWS THINGS ABOUT YOU



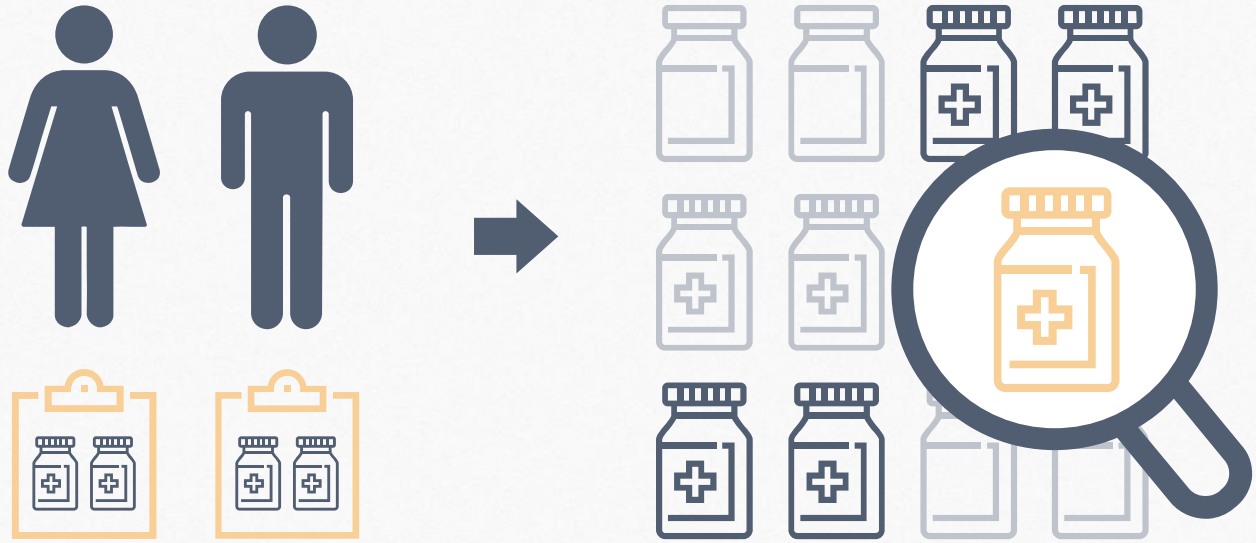
Vitals & Results?

HARD TO CHANGE – HAVE TO BE WITHIN RANGE TO BE REALISTIC

CAN'T CHANGE NORMAL > ABNORMAL

PROBLEM 3: YOUR NEIGHBOUR KNOWS THINGS ABOUT YOU

YOUR ATTRIBUTES IDENTIFY YOU:



Medications

- **WHO ELSE HAS EXACTLY THE SAME MEDICATION LIST?**

Problems

- **WHO ELSE HAS EXACTLY THE SAME PROBLEM LIST?**
- **WHO ELSE HAS EXACTLY THE SAME FAMILY HISTORY, SURGICAL HISTORY?**

PATENT PENDING ALGORITHMS TO ADDRESS THESE ISSUES

HUSHHUSH SOLUTION: PROPER DATA MASKING

VARIETY OF ALGORITHMS

Unique Elements:

- Stay Unique (SSN, SIN, PHONE, IP, etc.)

Non Unique Elements :
Statistical Challenge – Patent Pending Algorithms (ZIP, Names, Addresses, URLs)

- Automate the job of the experts: **statistical distortion**
- Variance algorithms (date, number)

Shuffle
Generic Alpha Numeric
Medicine-specific

Industry Grade Metrics: **k-anonymity, L-diversity**

Maintaining Integrity:

- Newly masked values always map consistently.

Support New Test Data Creation and As an Alternative : Same Ol' Random:

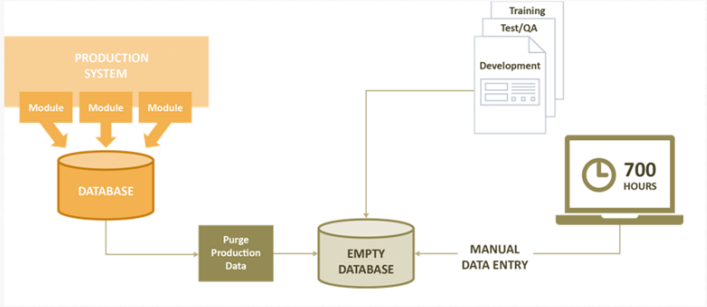
- Randomly created **values**

WHEN WE USE MASKING

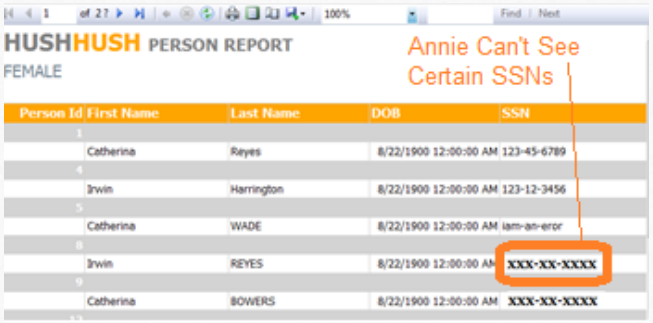
1. In development
    - Within an organization
    - At client site
2. In Sales Demos
    - More realistic data
3. In co-development with third parties
4. Privacy in Design
    - Statistical & medical research at client site

ON-PREMISES, IN THE CLOUD, SERVICES

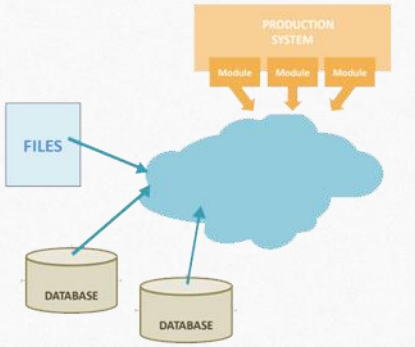
SDLC



Privacy in Design



Moving to The Cloud



How do we integrate with the data in the cloud?

Research

HUSHHUSH DE-IDENTIFICATION SOLUTIONS

NOW:

- **Client does no masking (no revenue, no compliance, no research)**
- **Client does inappropriate masking (monetary and reputational dangers)**
- **Client does masking without tool (expensive, monetary and reputational dangers)**

PROPOSITION:

We can figure out current risks of re-identification

We can improve your de-identification:

- **Remove the fear, enable clients to do research**
- **State of the art obfuscation tool – high confidence in removing risks**
- **Change the Risk / Reward Ratio in favor of being able to sell the data**

Identify the risk of re-identification before and after