ЮУРГУ
ВШ ЭКН
Кафедра «Защита информации»

Идентифицирующие атрибуты при определении вероятности идентификации физического лица в базе персональных данных

Мищенко Е.Ю., аспирант
научный руководитель Соколов А.Н., к.т.н.

Вероятность идентификации

Персональные данные

Фамилия	Имя	Отчество	Улица	Дом	Квартира	Прочее
Иванов	Антон	А	Л	1	5	xxxxxx
Иванов	Борис	Б	С	2	6	уууууу
Иванов	Влад	В	Т	3	7	вввввв
Петров	Антон	А	Б	4	8	рррррр

Записей с атрибутом «Фамилия» и значением «Иванов» в базе данных – 3
Вероятность идентификации человека с Фамилией «Иванов»: ВИ = 1/3

Записей с атрибутом «Имя» и значением «Антон» в базе данных – 2
Вероятность идентификации человека с Именем «Антон»: ВИ = 1/2

Идентифицирующие атрибуты. Теория

- **Набор:**
  - Фамилия, имя, отчество
  - дата рождения (год, месяц, день)
  - место рождения (страна, регион, населенный пункт)
  - адрес проживания (населенный пункт, улица, дом, квартира)
  - биометрия (отпечатки пальцев, фото, ДНК)
- **Синтаксис:**
  - ФИО, место рождения, населенный пункт и улица проживания - ТЕКСТ
  - дата рождения – ДАТА
  - дом, квартира проживания - ЦИФРА
  - биометрия - ГРАФИКА

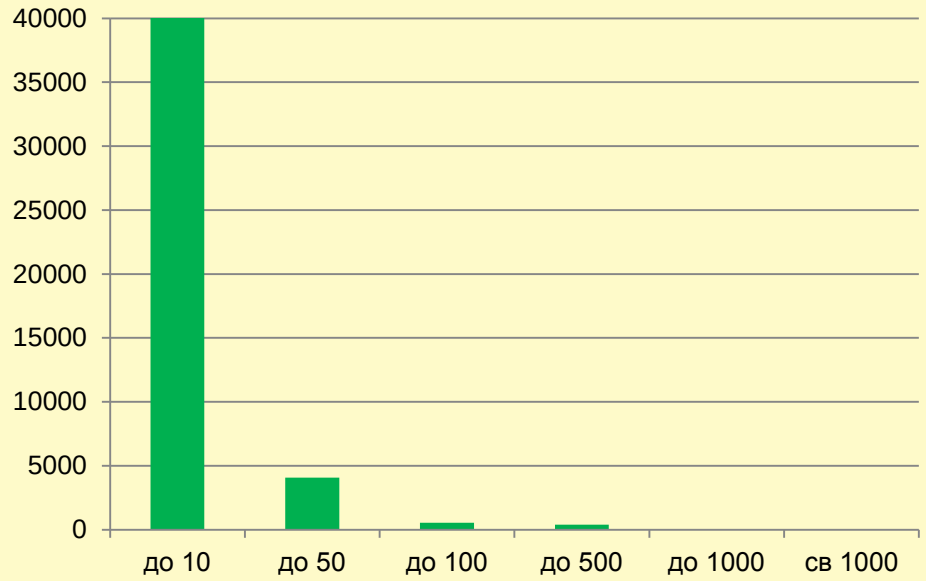
Идентифицирующие атрибуты. Эксперимент

- **Объем обработки ПД – 310 000 физических лиц**

- **Набор идентификаторов:**
  - **Фамилия – 45 100 значений**
  - **Имя – 690 значений**
  - **Отчество – 380 значений**
  - **Улица проживания – 890 значений**
  - **Дом – 320 номеров**
  - **Квартира – 630 номеров**

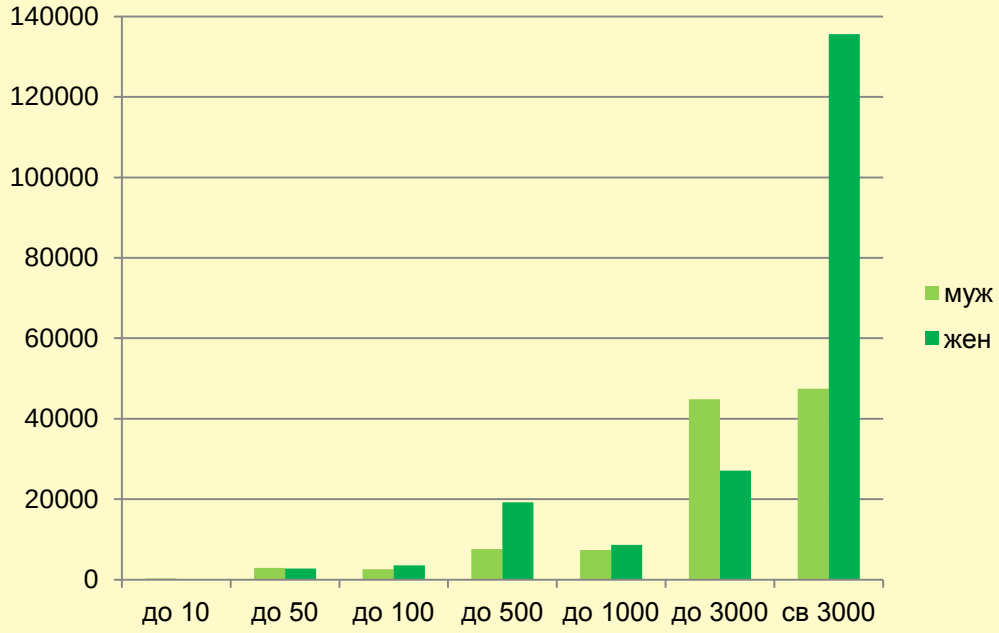
Фамилия. Вероятность

Ср. количество однофамильцев - 7
Эксперимент – 40000 фамилий имеют от 1 до 10 человек
Для 88% людей ВИ = 0,1 – 1,0
Вывод – обезличивание фамилии обязательно



Имя. Вероятность

Женщин – 197000 человек, 370 имен
Мужчин – 113000 человек, 320 имен
Для 0,27% мужчин и 0,07% женщин ВИ = 0,1 – 1,0
Вывод – обезличивание имени не нужно



Отчество

Женщин – 197000 человек, 370 отчеств
Мужчин – 113000 человек, 310 отчеств

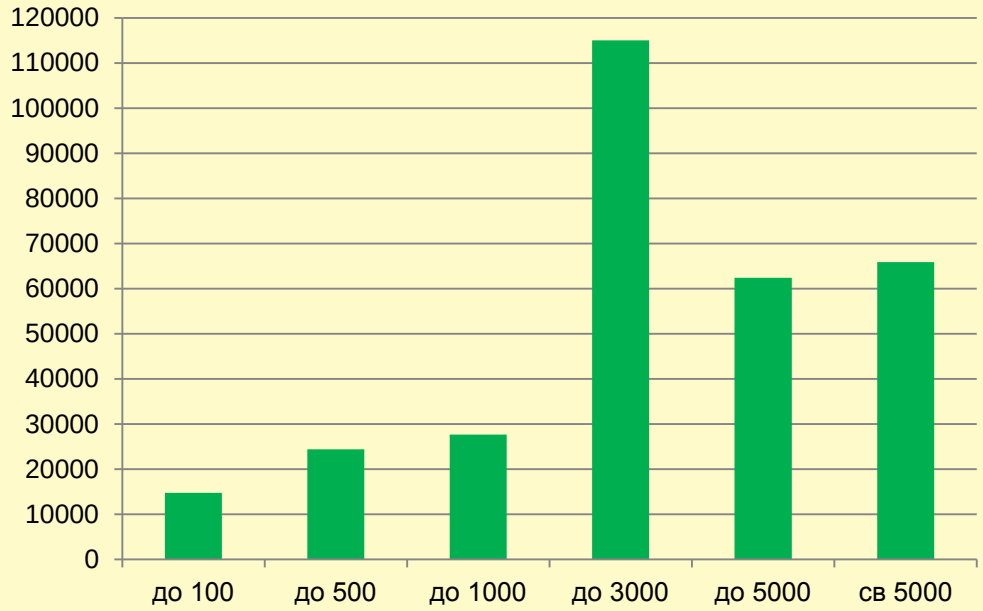
Отчество коррелирует с мужскими именами, есть отличие в распределении по буквам, ВИ в целом совпадает
Вывод – обезличивание отчества не нужно

Улица

Теория – отсутствует
Эксперимент – всего 890 улиц, в среднем на 1 улице живет 350 человек

Улица. Вероятность

Эксперимент – на 220 улицах живет более 95% людей
На 670 улицах живет 14750 человек, среднее – 22 человека
Для 5% людей ВИ = 0,01 – 0,5
Вывод – обезличивание улицы зависит от объема базы

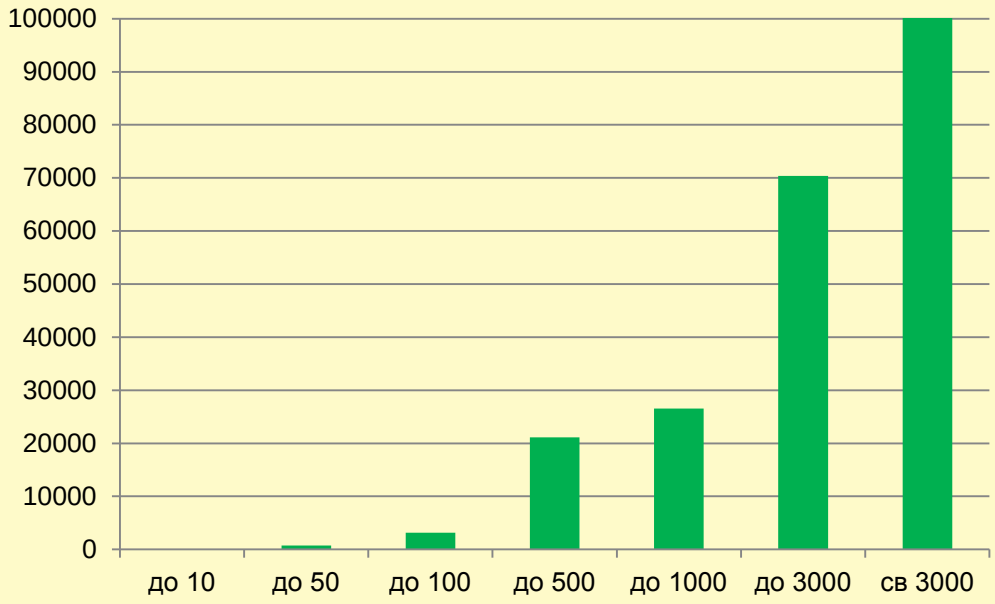


Номер дома

Теория – отсутствует
Эксперимент – номера домов с 1 по 10 есть на 98% улиц

Номер дома. Вероятность

Эксперимент – в 1% домов живет менее, чем по 10 человек
Для 0,03% людей ВИ = 0,1 – 0,5
Вывод – обезличивание номера дома не нужно

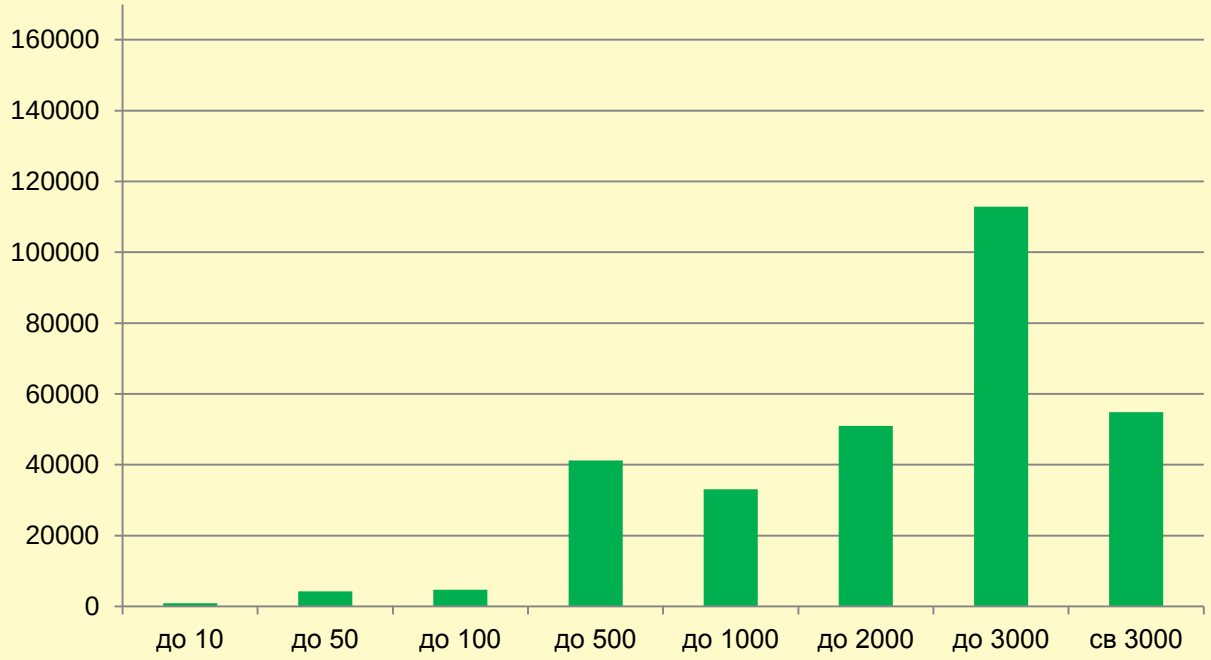


Номер квартиры

Теория – отсутствует
Эксперимент – номера квартир с 1 по 8 есть во всех домах

Номер квартиры. Вероятность

Эксперимент – в 1% квартир живет менее, чем по 10 человек
Для 0,3% людей ВИ = 0,1 – 1
Вывод – обезличивание номера квартиры не нужно



Планируемые исследования

1) Вероятность идентификации по группам: номер дома + номер квартиры, имя + номер дома - мала, требуется уточнение

2) Вероятность идентификации по группе имя + отчество, улица + номер дома, улица + имя – зависит от объема базы, требуется уточнение

3) Вероятность идентификации по группе первых букв ФИО - зависит от объема базы, требуется уточнение

Выводы

1) Вероятность идентификации прямо от размера базы не зависит
2) Есть косвенная зависимость. Чем меньше объем базы – тем больше становится абсолютное количество малых процентных долей – ВИ будет увеличиваться
3) Для объема базы менее 10000 придется обезличивать все идентификаторы

Спасибо за внимание

Вопросы?