Якість даних: причини помилок

- Зібрані неправильні дані
- Форми написані почерком, який важко розібрати
- Немає кодів або зазначені старі коди
- Неправильне використання скорочень
- Помилки під час введення даних
- Дані подані кілька разів

Якість даних: причини помилок

- Недостатньо даних (дата народження, початку захворювання, вакцинації)
- Округлення чисел
- Систематичні помилки, викликані відсутністю репрезентативності
  - Більш важкі випадки
  - Міські > сільські
  - Джерело не представлено (приватний сектор, лікарі загальної практики)

Що таке очистка даних і чому вона важлива?

- Виявляє та усуває помилки та невідповідність даних з метою покращення якості даних
- Важлива для будь-яких даних, не лише для комп'ютерних баз даних
- Вказує на проблеми з якістю даних, що існують у базі даних
- Точність даних є важливою для надання користувачам можливості прийняти рішення

Основні процеси очищення даних

- Перевірка даних і виявлення помилок
- Виправлення помилок
- Підтвердження правильності даних (валідація)
- Попередження виникнення помилок

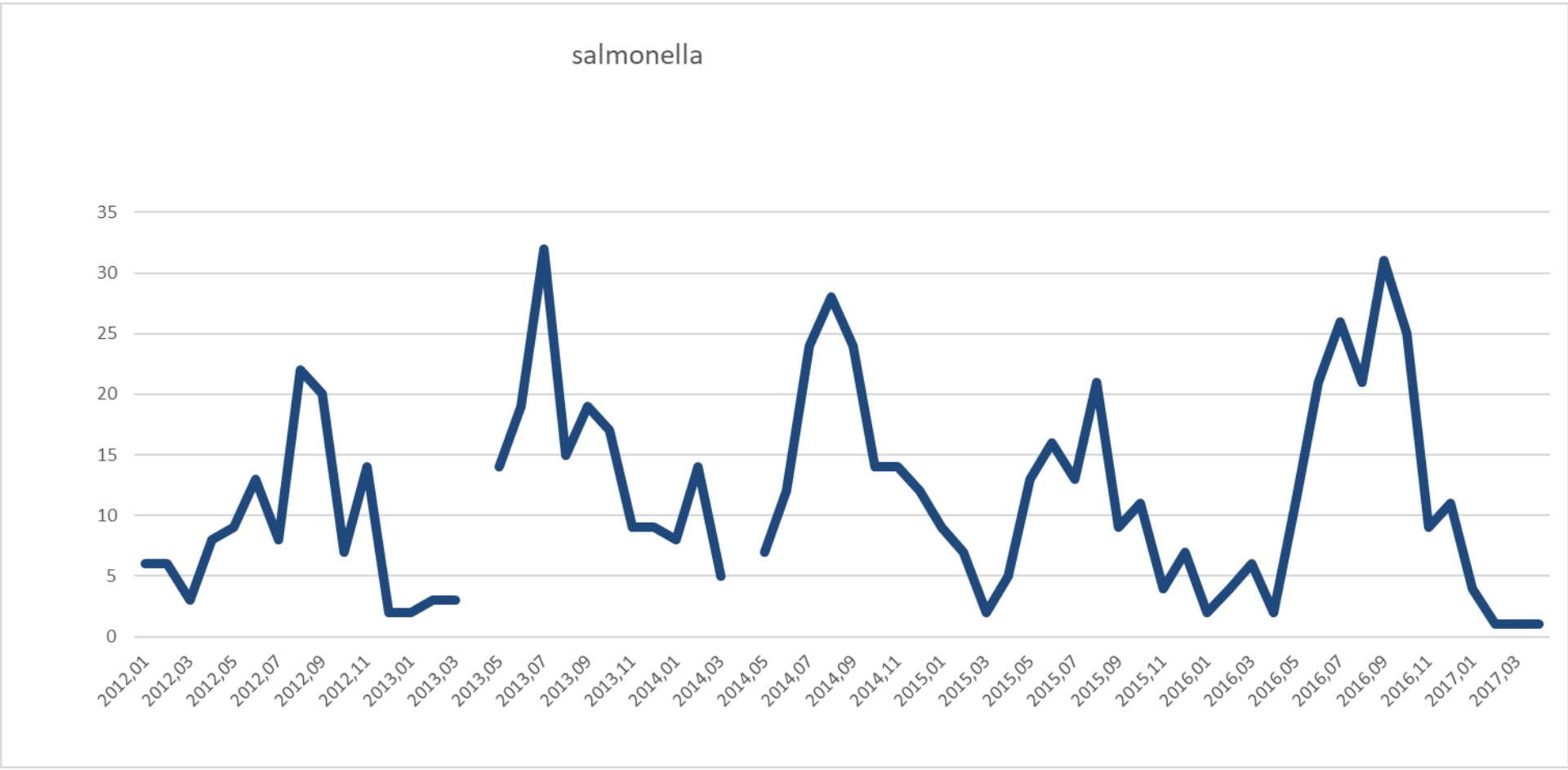
Методи очищення даних

1) Методи, що ґрунтуються на базі знань
   - Використовують знання про дані для очищення бази даних (ідентифікують те, що може бути помилкою)
   - Можуть бути використані для виявлення помилок при вводі даних, помилок у написанні слів і скорочень
   - Можна створити алгоритм на комп'ютері, який би знаходив неможливі значення (наприклад, вік: 167 років)

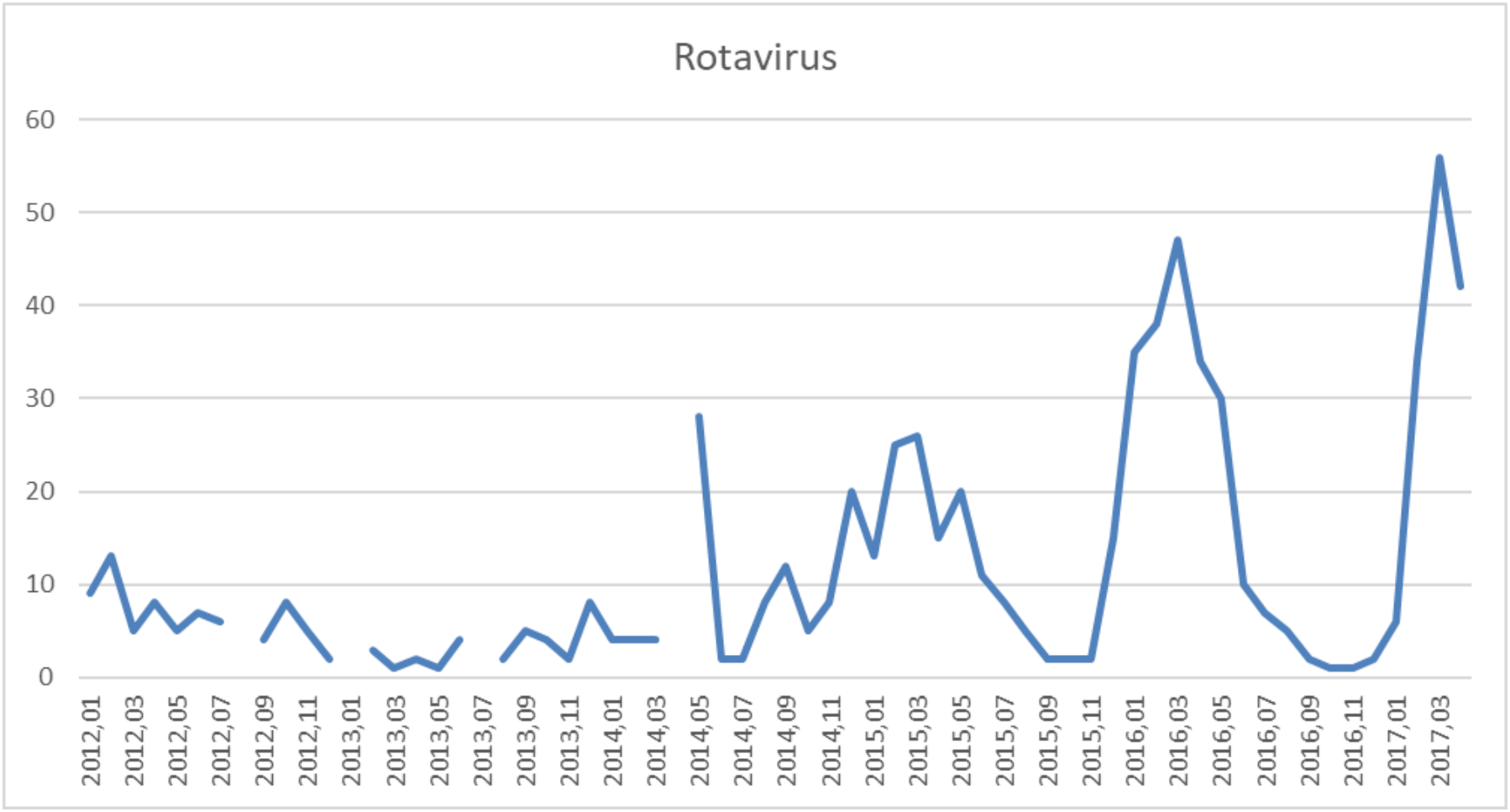
2) Об'єднання загальних баз даних
   - Дає можливість виявлення очевидних помилок

Проста процедура перевірки на присутність помилок

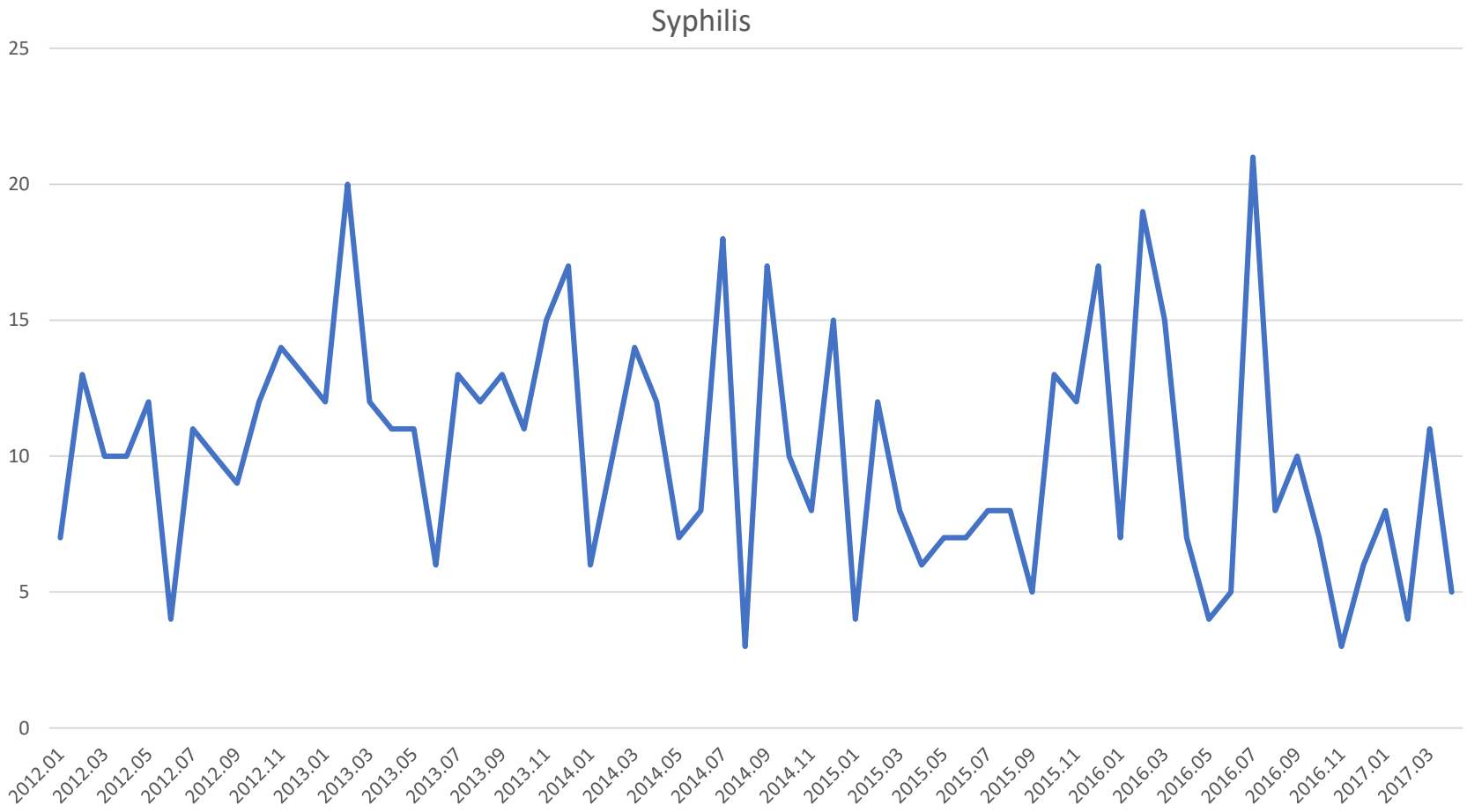
- Заблокуйте сукупність дат для пошуку незрозумілих результатів (таблиці частот повторюваності)
- Якщо дані зібрані в базі, відсортуйте дані за категоріями та перевірте, чи є повтори або надзвичайно рідкі результати
- Шукайте невідповідні дані, порівнюючи 2 чи більше змінних (наприклад, вік і робота)
- При перевірці якості введення даних, перевірте вибірку (5%) форм



Що сталося в квітні 2013 і 2014 року!



Чи були досліджені піки 15 травня, 16 та 17 березня?



Це достовірна тенденція сифілісу? Чи графік відображає надходження зібраних сповіщень?

Як виправити дані

- Записувати зміни (копія файлу)
- Робити зміни лише тоді, коли є помилка, а не тоді, коли результат здається неправдоподібним
- Якщо помилок буде багато (>5%), розгляньте можливість змінити процедуру збору та введення даних

Як попередити

- Зменшити кількість кроків від збору до зберігання даних
- Часто перевіряти дані
- Використовувати комп'ютер і базу даних (для швидкого управління даними)
- Створити стандартні алгоритми для систематичної перевірки якості та очищення даних
- Вчитися на помилках: після виявлення помилки спробувати представити, як уникнути подібної помилки в майбутньому

Де виправляти дані

- Дані слід виправляти на самому початку: місцеві органи охорони здоров'я повинні один за одним перевіряти дані, отримані від лікаря чи пацієнта
- Комп'ютерна система може включати в себе кілька інструментів для корекції (також за допомогою простого Xcell!), проте
- Комп'ютерна система на обласному рівні не матиме можливості виправляти якість введених першоджерелом даних

GIGOLO?

No!: G.I.G.O. LAW
Garbage In … Garbage Out