**Mining of Massive Dataset**

**Step 3:** Locality – Sensitive Hashing – Tập trung vào các chữ ký từ các tài liệu tương tự

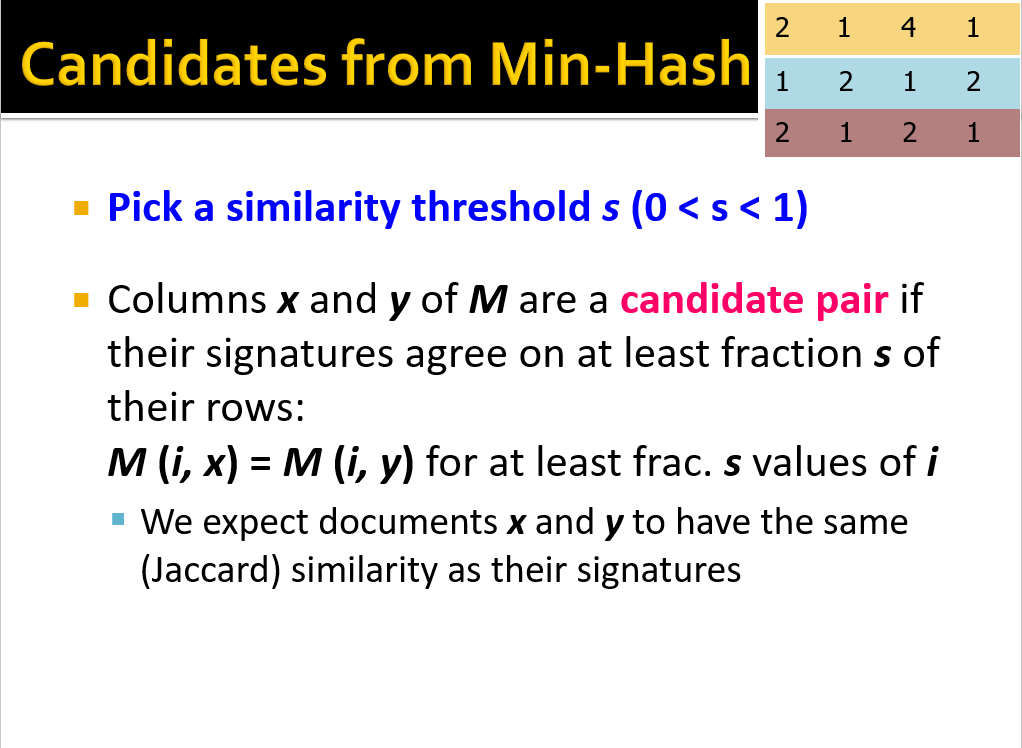
Mục tiêu: Tìm tài liệu có độ tương tự ít nhất s(đối với một số ngưỡng tương tự, ví dụ: s=0.8)

LSH - Ý tưởng chung: Sử dụng một hàm f (x, y) cho biết x và y có phải là một cặp ứng cử viên hay không: một cặp phần tử có độ giống nhau phải được đánh giá

Đối với ma trận Min-Hash:

Băm các cột của ma trận chữ ký M thành nhiều nhóm

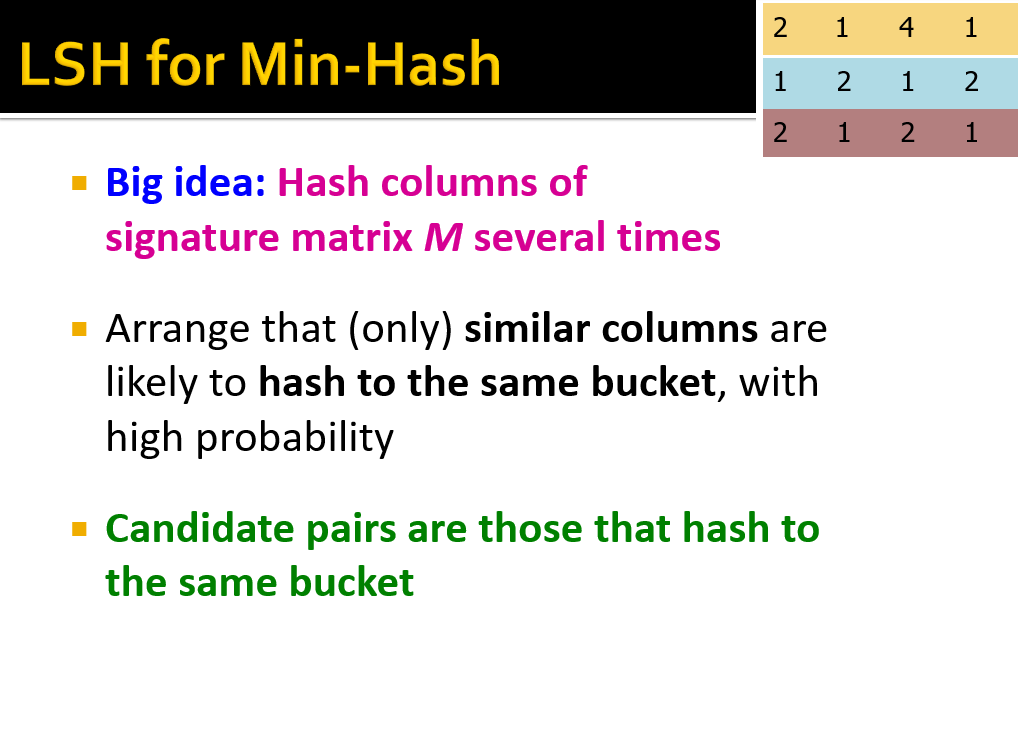
Mỗi cặp tài liệu được băm vào cùng một nhóm là một cặp ứng viên



Chọn ngưỡng tương tự s (0<s<1)

Các cột x và y của M là một cặp ứng cử viên nếu chữ ký của chúng đồng ý trên ít nhất phần s trong các hàng của chúng: M (i, x) = M (i, y) cho ít nhất là phân số. Giá trị s của i

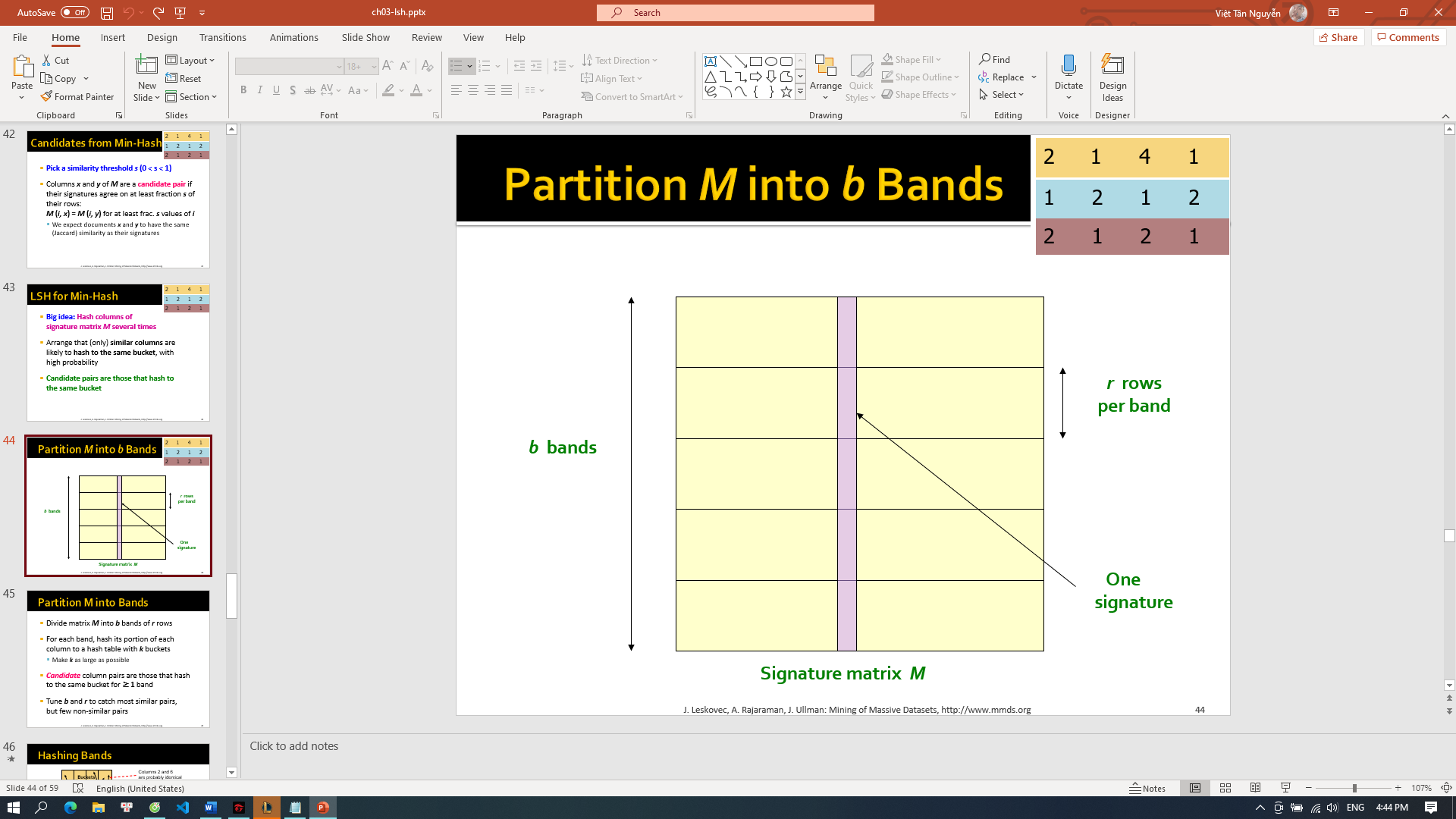
Mong đợi các tài liệu x và y có cùng (Jaccard) giống nhau như chữ ký của chúng



Ý tưởng: Băm các cột của ma trận chữ ký M nhiều lần

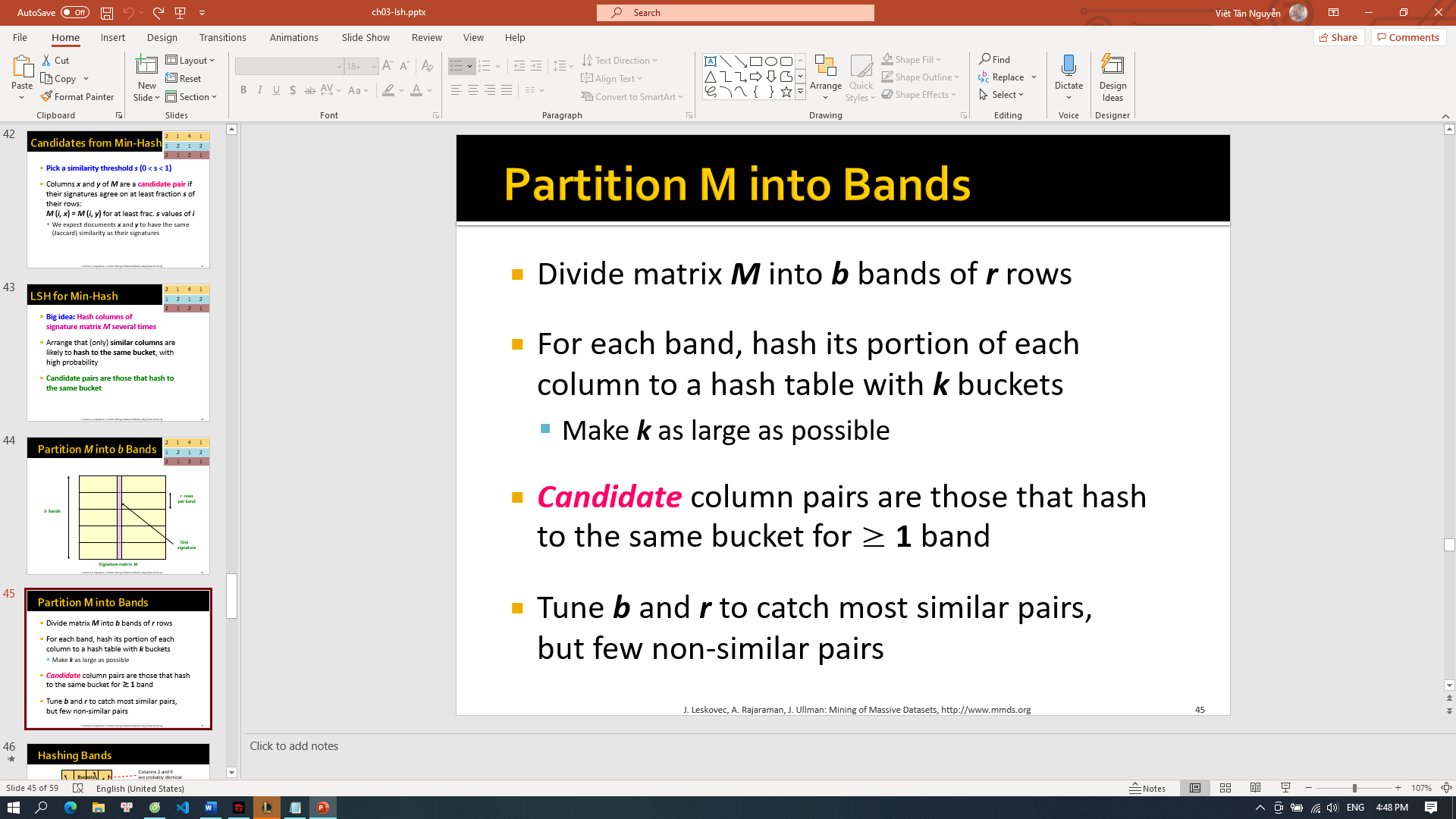
Sắp xếp để các cột tương tự có khả năng băm vào cùng một nhóm, với xác suất cao

Các cặp ứng cử viên là những cặp được băm vào cùng một nhóm



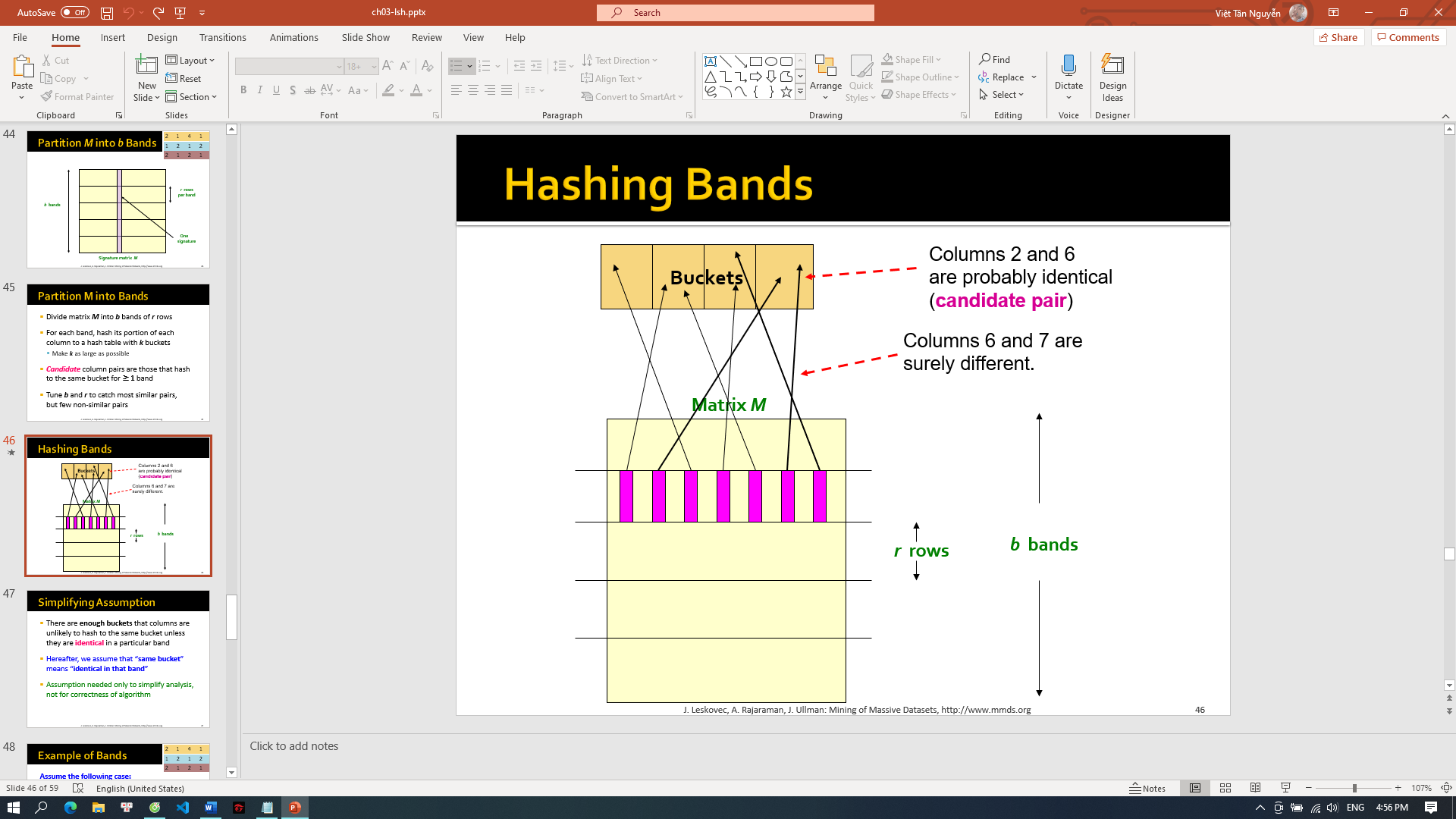
Vùng màu vàng là ma trận chữ ký M. Mỗi cột tương ứng với một chữ ký và mỗi hàng là một trong những thành phần của tất cả các chữ ký

Chia các hàng thành các dải b cho một số b. Kết quả là có r hàng trên mỗi dải, trong đó b nhân với r là tổng chiều dài của các chữ ký. Đó là, số lượng các hàm băm chính mà chúng ta sử dụng để tạo chữ ký. Chúng ta sẽ tạo một hàm băm từ mỗi băng tần.



Chúng ta đã chia ma trận chữ ký M thành b dải r hàng mỗi dải.

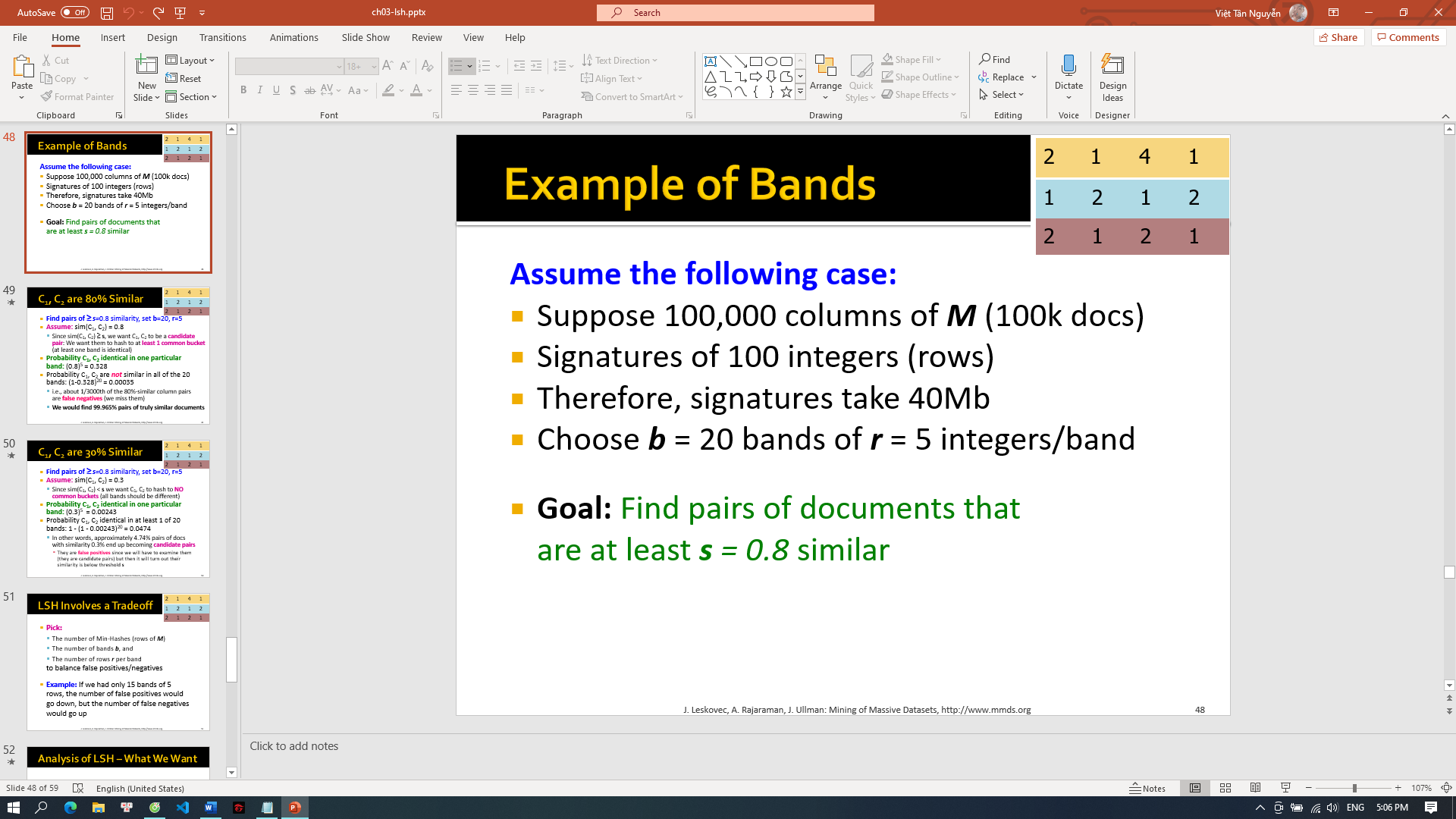
Từ mỗi dải, chúng ta tạo một hàm băm. Hàm băm này chỉ băm các giá trị mà một cột nhất định có trong dải đó. Chúng ta sẽ tạo một nhóm cho mỗi vectơ có thể có giá trị b mà một cột có thể có trong dải đó. Đó là, chúng ta muốn có quá nhiều nhóm để hàm băm thực sự là hàm nhận dạng, nhưng đó có lẽ là quá nhiều nhóm



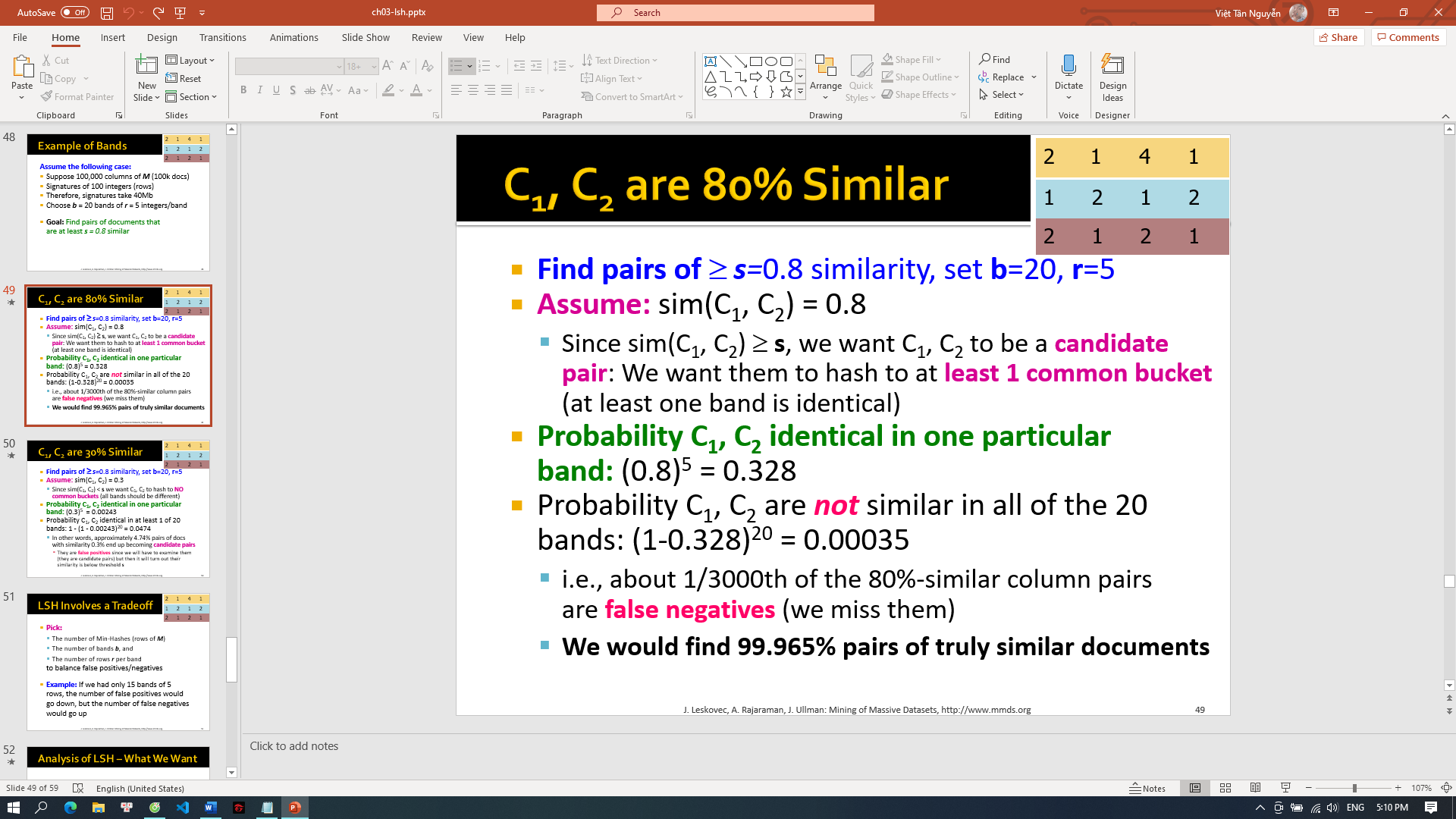
Đây là hình ảnh về những gì mà hàm băm cho LSH trên ma trận chữ ký. Chúng ta thấy một trong những dải b, tất nhiên là dải bao gồm r hàng. Nó cũng hiển thị ma trận bao gồm một số, trong số bảy cột hoặc chữ ký.

Và mỗi đại diện màu hồng, hình chữ nhật đại diện cho phần cột trong 1 dải mà chúng ta tập trung vào. Cột 6 và cột 7 băm thành các nhóm khác nhau. Do đó chúng chắc chắn khác nhau. Cặp 6 và 7 không được tạo thành 1 cặp ứng cử viên bởi hàm băm LSH này.

Mặt khác, cột 2 và 6 thực hiện việc băm cho cùng một nhóm. Vì vậy, cột 2 và 6 là một cặp ứng cử viên bất kể điều gì xảy ra trong các nhóm khác



Giả sử có 100.000 cột. Chúng ta đang tìm kiếm các tài liệu tương tự trong số 100.000 tài liệu. Chữ ký có độ dài là 100, chúng ta sử dụng các hàm minhash 100 để tạo chữ ký, do đó ma trận chữ ký M là 100 hàng x 100.000 cột. Chữ ký yêu cầu lấy 40Mb

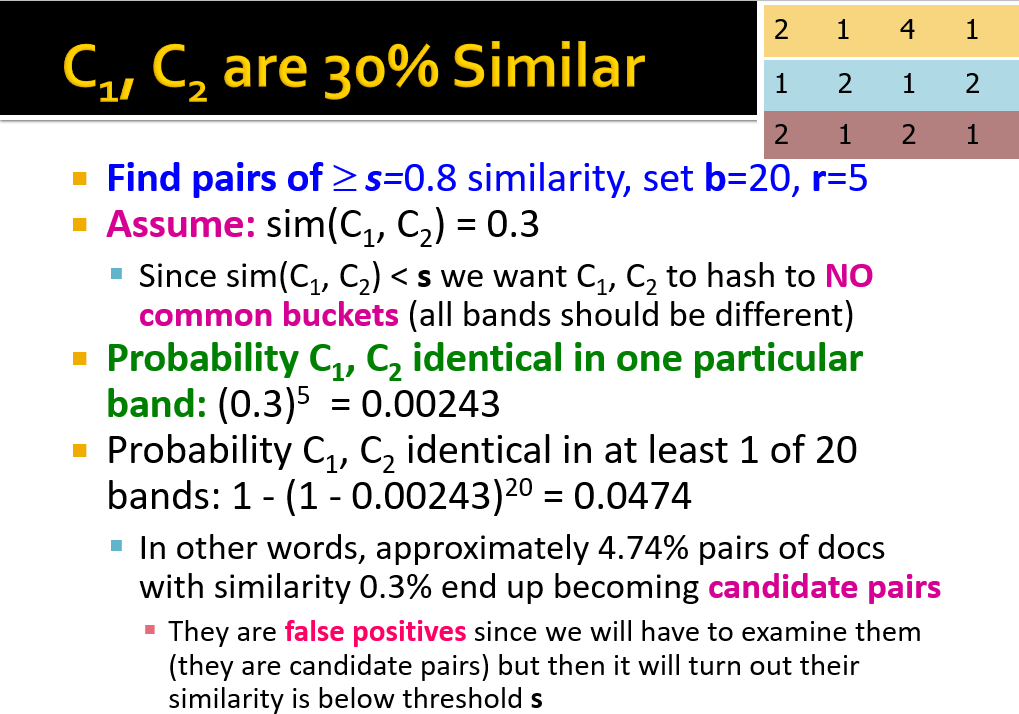


Đầu tiên xem xét 2 cột C1 và C2, đại diện cho các bộ có độ tương tự Jaccard 0.8. Do tính ngẫu nhiên liên quan đến băm nhỏ, các cột C1 và C2 có thể đồng ý trong nhiều hơn hoặc ít hơn 80 hàng của họ, nhưng rất có thể chúng sẽ có khoảng 80 hàng bằng nhau

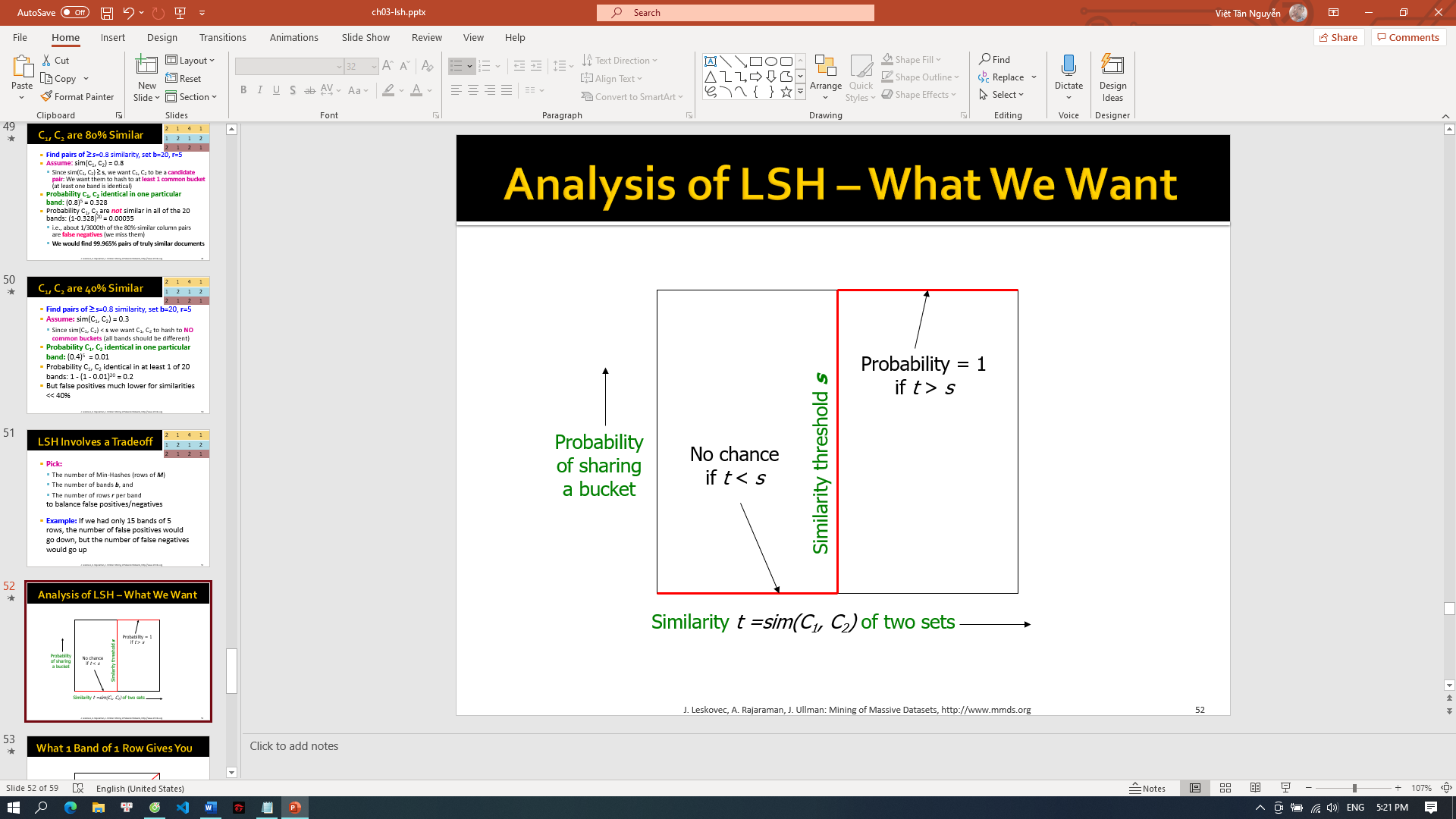
Xác suất để chúng đồng ý trong một hàng bất kỳ là 0.8, xác suất để hai cột đồng ý ở tất cả năm hàng của một dải được nâng lên 0,8 lên lũy thừa thứ năm, hoặc xấp xỉ 0,328.

Chúng ta có 20 cơ hội để biến cặp cột trở thành cặp ứng cử viên. Xác suất để chúng không băm vào cùng một nhóm trong một dải là 1 - 0,328 hoặc 0,672

Xác suất các cột không thể băm thành cùng 1 nhóm cho bất kỳ dải nào trong số 20 dải là giá trị 0.672 được nâng lên thành lũy thừa 20, đó là một con số nhỏ, là 0.00035. Cơ hội mà cặp C1 và C2 sẽ là 1 cặp ứng cử viên là 1-0.00035, là 0.99965. Xác suất âm tính giả, một cặp bộ có độ giống Jaccard 80%, nhưng những chữ ký không trở thành một cặp ứng cử viên là 0,00035, hoặc khoảng 1/3.000.



Bây giờ , đến với 1 cặp có độ giống nhau của Jaccard là 0.3. Xác suất để chữ ký của chúng băm vào cùng một nhóm trong ít nhất 1 trong số 20 dải chắc chắn không nhiều hơn 20 lần.



Một cách để xem xét vấn đề thiết kế một lược đồ LSH từ ma trận minhash là cái này. Chúng ta muốn xác suất của hai cột chia sẻ một nhóm là một hàm bước với ngưỡng t bằng giá trị mà tại đó chúng ta coi các tập cơ bản tương tự nhau. Tức là, nếu độ tương tự Jaccard của các tập cơ bản nhỏ hơn t, chúng ta muốn không có cơ hội các chữ ký sẽ chia sẻ một nhóm cho một trong các băm và do đó trở thành một cặp ứng cử viên. Tuy nhiên, nếu mức tương tự Jaccard cơ bản vượt quá 2, chúng ta muốn cặp chữ ký chắc chắn trở thành một cặp ứng cử viên.

