Research Data Services

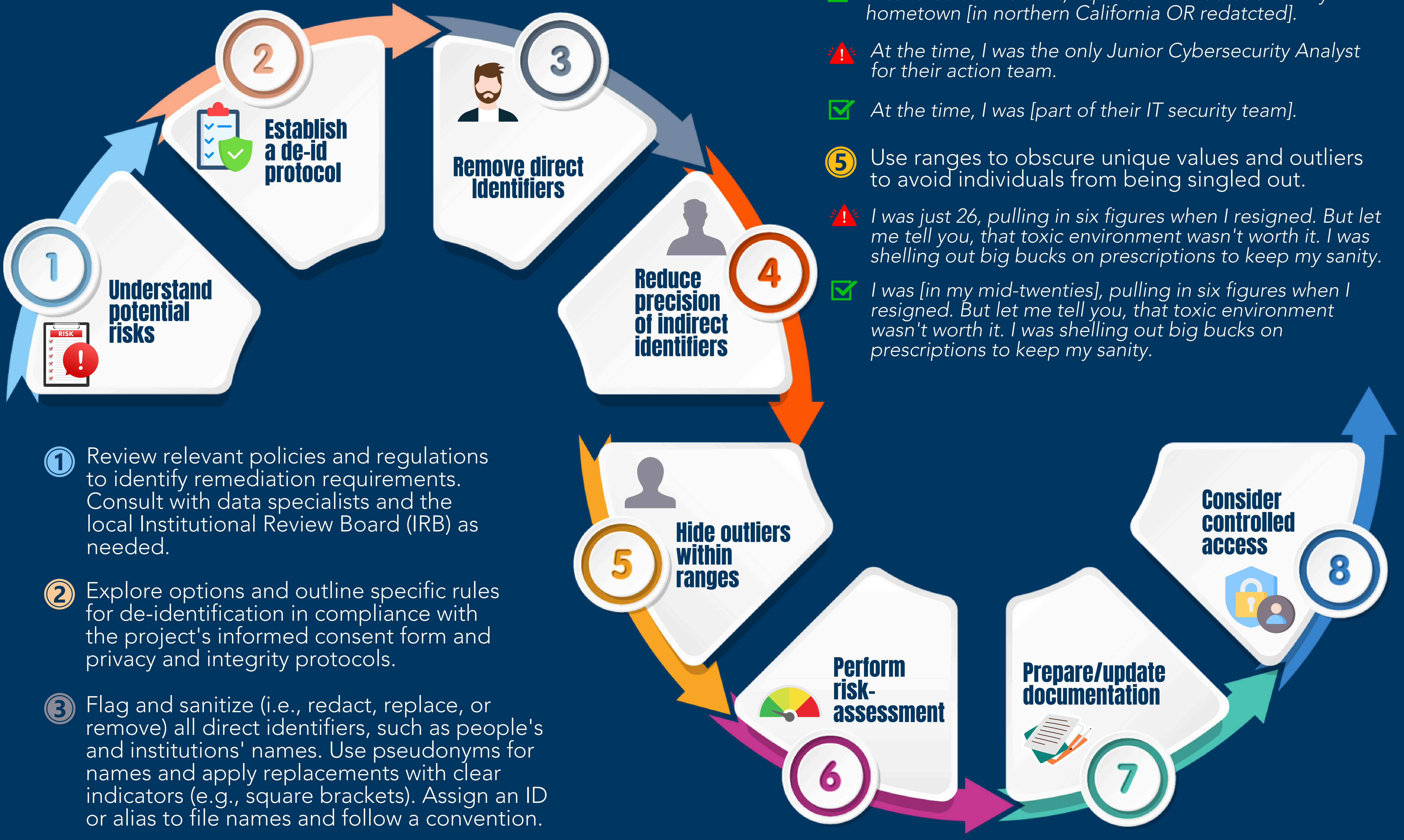
Data Literacy Series, n.05/2024

Sanitizing Sensitive Information from Textual Data

De-identifying text with human participant data is challenging due to its unstructured and sensitive nature, requiring careful measures for safe and ethical sharing. Here are eight essential steps to minimize re-identification risks while preserving reusability.

In the workflow below, we exemplify actions to safeguard the privacy and confidentiality of a research subject interviewed for a project examining workplace harassment and its impact on the mental health of tech company employees from Silicon Valley.

Name: Luke M. Hills
Age: 30
Job title: Junior Cybersecurity Analyst (2018-2020)
Company: ACME Tech
Boss: Dr. John Rules
Medical condition: burnout and anxiety


1 Understand potential risks
RISK
2 Establish a de-id protocol
3 Remove direct identifiers
4 Reduce precision of indirect identifiers
5 Hide outliers within ranges
6 Perform risk-assessment
7 Prepare/update documentation
8 Consider controlled access

① Review relevant policies and regulations to identify remediation requirements. Consult with data specialists and the local Institutional Review Board (IRB) as needed.

② Explore options and outline specific rules for de-identification in compliance with the project's informed consent form and privacy and integrity protocols.

③ Flag and sanitize (i.e., redact, replace, or remove) all direct identifiers, such as people's and institutions' names. Use pseudonyms for names and apply replacements with clear indicators (e.g., square brackets). Assign an ID or alias to file names and follow a convention.

⚠ *At ACME Tech, I had this real doozy with my boss, Dr. Rules.*

☑ *At [Company A], I had this real doozy with my boss, [name redacted].*

④ Examine data for context-specific details that could lead to re-identification. Redact or generalize these details.

⚠ *After that last incident, I quit and moved back to my hometown, up north, in Paynes Creek, California.*

☑ *After that last incident, I quit and moved back to my hometown [in northern California OR redatcted].*

⚠ *At the time, I was the only Junior Cybersecurity Analyst for their action team.*

☑ *At the time, I was [part of their IT security team].*

⑤ Use ranges to obscure unique values and outliers to avoid individuals from being singled out.

⚠ *I was just 26, pulling in six figures when I resigned. But let me tell you, that toxic environment wasn't worth it. I was shelling out big bucks on prescriptions to keep my sanity.*

☑ *I was [in my mid-twenties], pulling in six figures when I resigned. But let me tell you, that toxic environment wasn't worth it. I was shelling out big bucks on prescriptions to keep my sanity.*

⑥ Review the de-identified data for oversights and risk of re-identification. Check for compliance with steps 1 and 2.

⑦ Describe the de-identification processes performed. Include that description in the project documentation.

⑧ Determine whether a Data Use Agreement (DUA) or access controls are still necessary to preserve confidentiality and privacy requirements while maintaining data utility.

Want to learn more?
rds@library.ucsb.edu

UC **SANTA BARBARA**
Library

www.library.ucsb.edu