Carnegie Mellon University

Improving Security in Software Acquisition with Data Retention Specifications

Daniel Smullen,

Travis Breaux

Navy Postgraduate School Acquisition Research Symposium

May 4, 2016

isr institute for SOFTWARE RESEARCH

Motivation

- DoD and the Defense Industrial Base (DIB) leverage 3rd party service compositions to outsource infrastructure and services.
  - What sort of data do they want?
  - What will they do with my data once they have it?
  - What am I willing to give them?

- In prior work: we studied actions for collection, use and sharing [RE'15]
- What about data retention?

[RE'15] T. Breaux, D. Smullen, H. Hibshi, "Detecting Repurposing and Over-collection in Multi-Party Privacy Requirements Specifications." *IEEE 23rd International Requirements Engineering Conference*, Ottawa, Canada, pp. 166-175, Sep. 2015.

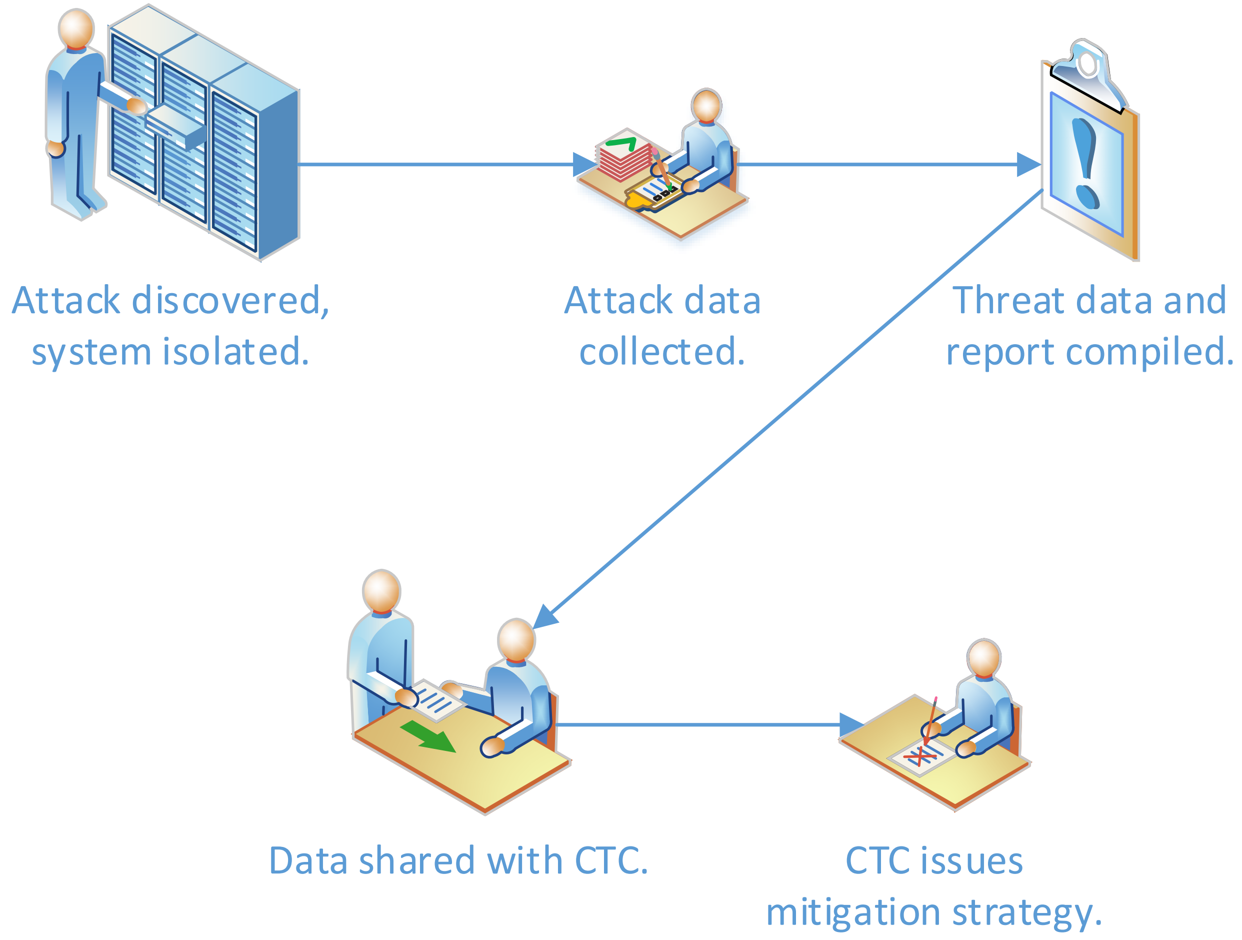
Data Retention Actions

- Redaction
  - Directly remove some type of data from a collection.
- Perturbation
  - Don’t directly remove a specific type of data, just reduce the quality of the data, or remove specific data points.
- Data Append
  - Combine disjoint data to infer something, create a new type of data.

Running Example

- Cyber threat information sharing portal.
- A **cyber threat clearinghouse** and **DoD contractor** have recently crafted a ***data sharing agreement*** that enables them to collaborate to share **cyber threat information** via this portal.

Running Example



Attack discovered, system isolated.

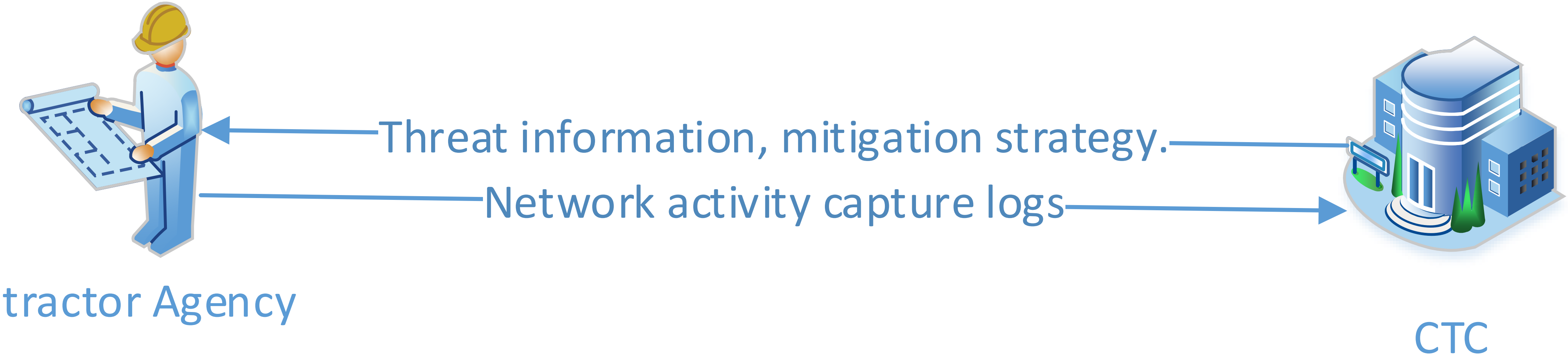
Attack data collected.

Threat data and report compiled.

Data shared with CTC.

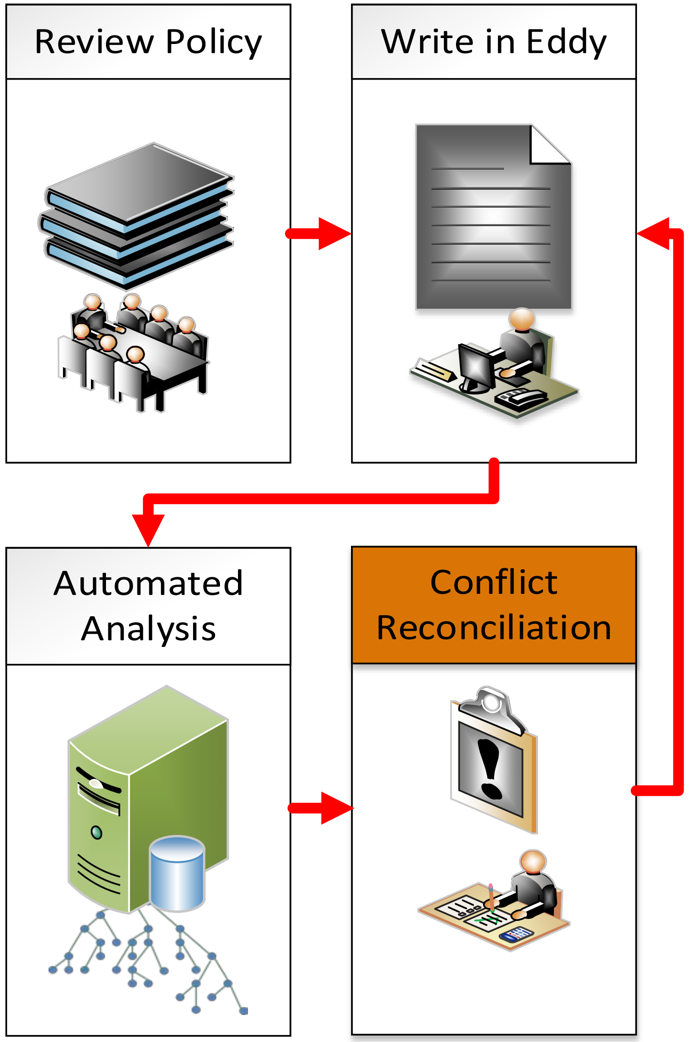
CTC issues mitigation strategy.

Running Example


Threat information, mitigation strategy.
Network activity capture logs
Contractor Agency
CTC

Technical Approach

- Eddy; formal language with semantics based on Description Logic (DL).
- Specify and analyze data flows.
- Use this as a tool to measure how specified actions affect *unanticipated disclosure*.



Eddy Language Structure

- Two sections; header section, policy section.

- Define data, actors, purposes in header.
  - All concepts can have sub-concepts described through DL *subsumption*.
  - Can have equivalent concepts described through DL *equivalence*.

- Define actions (with modality – permission, prohibition, obligation) in policy section.

Using Eddy

1. Analyze policy text to extract requirements, code into Eddy:

Modal obligation. Purpose specified. [illegible]

"All O-2 or higher service members must collect the name and rank from network analysts in order to complete a network analysis authorization form. This must be done to assure compliance with standing orders."

```
SPEC HEADER

D authorization_form > name, rank

SPEC POLICY

O COLLECT authorization_form FROM network_analysts FOR assure_compliance
```

Using Eddy

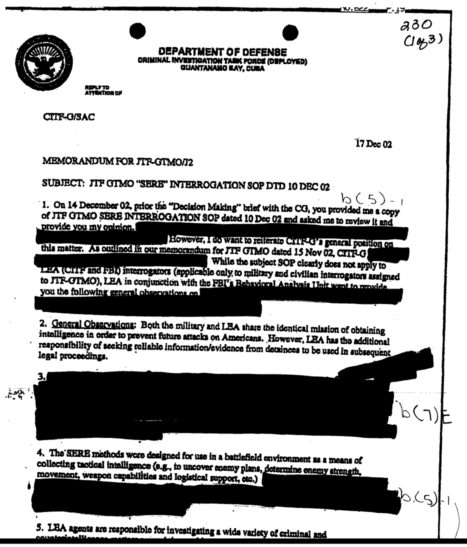
2. Tool compiles Eddy into Description Logic:

- $\text{name} \sqsubseteq \text{authorization\_form}$
- $\text{rank} \sqsubseteq \text{authorization\_form}$
- $p_1 \equiv \text{COLLECT} \sqcap \exists\text{hasObject.authorization\_form} \sqcap \exists\text{hasSource.network\_analysts} \sqcap \exists\text{hasPurpose.assure\_compliance}$
- $p_1 \sqsubseteq \text{Obligation}$

```
SPEC HEADER
       ATTR NAMESPACE "http://gaius.isri.cmu.edu/example.owl"
       ATTR DESC "This is an example policy for medical data, written in Eddy."
       P treatment > diagnosis, prescription, blood-tests
       D patient-labs > bloodwork
       A medical-professional > phlebotomist, doctor, nurse
       A laboratory > phlebotomist
SPEC POLICY
       P COLLECT bloodwork FROM phlebotomist FOR treatment
       P COLLECT bloodwork FROM laboratory FOR treatment
       P USE bloodwork FROM phlebotomist FOR marketing
       P USE patient-labs FROM phlebotomist FOR anything
       P USE bloodwork FROM medical-professional FOR diagnosis
       R USE bloodwork FROM medical-professional FOR anything
```

Further Extending Eddy for Data Retention

- ***Redaction*** means to remove data elements from a dataset.
- Redaction is useful as a data minimization strategy when data cannot be shared, or when it can be pared down to the minimum necessary.



230
(1 of 3)

DEPARTMENT OF DEFENSE
CRIMINAL INVESTIGATION TASK FORCE (DEPLOYED)
GUANTANAMO BAY, CUBA

REPLY TO
ATTENTION OF

CITF-G/SAC

17 Dec 02

MEMORANDUM FOR JTF-GTMO/J2

SUBJECT: JTF GTMO "SERE" INTERROGATION SOP DTD 10 DEC 02

b(5)-1

1. On 14 December 02, prior the "Decision Making" brief with the CG, you provided me a copy of JTF GTMO SERE INTERROGATION SOP dated 10 Dec 02 and asked me to review it and provide you my opinion. [redacted] However, I do want to reiterate CITF-G's general position on this matter. As outlined in our memorandum for JTF GTMO dated 15 Nov 02, CITF-G [redacted] While the subject SOP clearly does not apply to LEA (CITF and FBI) interrogators (applicable only to military and civilian interrogators assigned to JTF-GTMO), LEA in conjunction with the FBI's Behavioral Analysis Unit want to provide you the following general observations on [redacted]

2. General Observations: Both the military and LEA share the identical mission of obtaining intelligence in order to prevent future attacks on Americans. However, LEA has the additional responsibility of seeking reliable information/evidence from detainees to be used in subsequent legal proceedings.

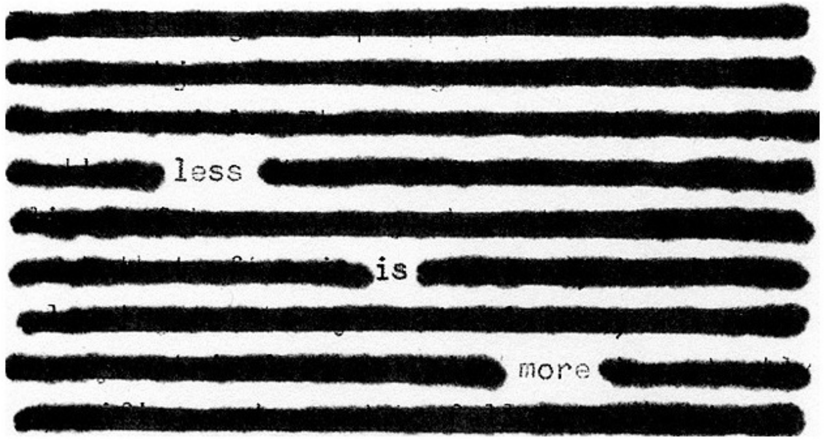
3. [redacted]

b(7)E

4. The SERE methods were designed for use in a battlefield environment as a means of collecting tactical intelligence (e.g., to uncover enemy plans, determine enemy strength, movement, weapon capabilities and logistical support, etc.) [redacted]

b(5)-1

5. LEA agents are responsible for investigating a wide variety of criminal and



[redacted] less [redacted] is [redacted] more [redacted]

These are paper redactions, but we can redact data from digital sharing, too.

Further Extending Eddy for Data Retention

- ***Data append*** refers to a general class of methods that link two or more data elements together.
- By *prohibiting* data append, downstream parties are bound to limit the use of a redacted dataset; cannot combine to recreate original.
- Fixed requirement for the data prior to sharing, assuring disjointness from other datasets post-transfer to a third party.

Further Extending Eddy for Data Retention

- ***Perturbation*** refers to a general class of methods that introduces statistical inaccuracies into data; conforms to statistical profile of original data;
  - Changing data values.
  - Removing values.
  - Adding new values.

- Eddy language does not assume that data perturbation is implemented by any particular method.

Experimental Design

- *How do data append, redaction and perturbation systemically affect data subject unanticipated disclosure?*

- Microsimulation [Lov16]; a technique for analyzing real-world situations based on synthetic data.
- Combine with data sharing agreement profile (Eddy specification).
- Perform analysis.

[Lov16] R. Lovlace, “Spatial Microsimulation in R”, CRC Press, 2016.

Why Generate Synthetic Datasets?

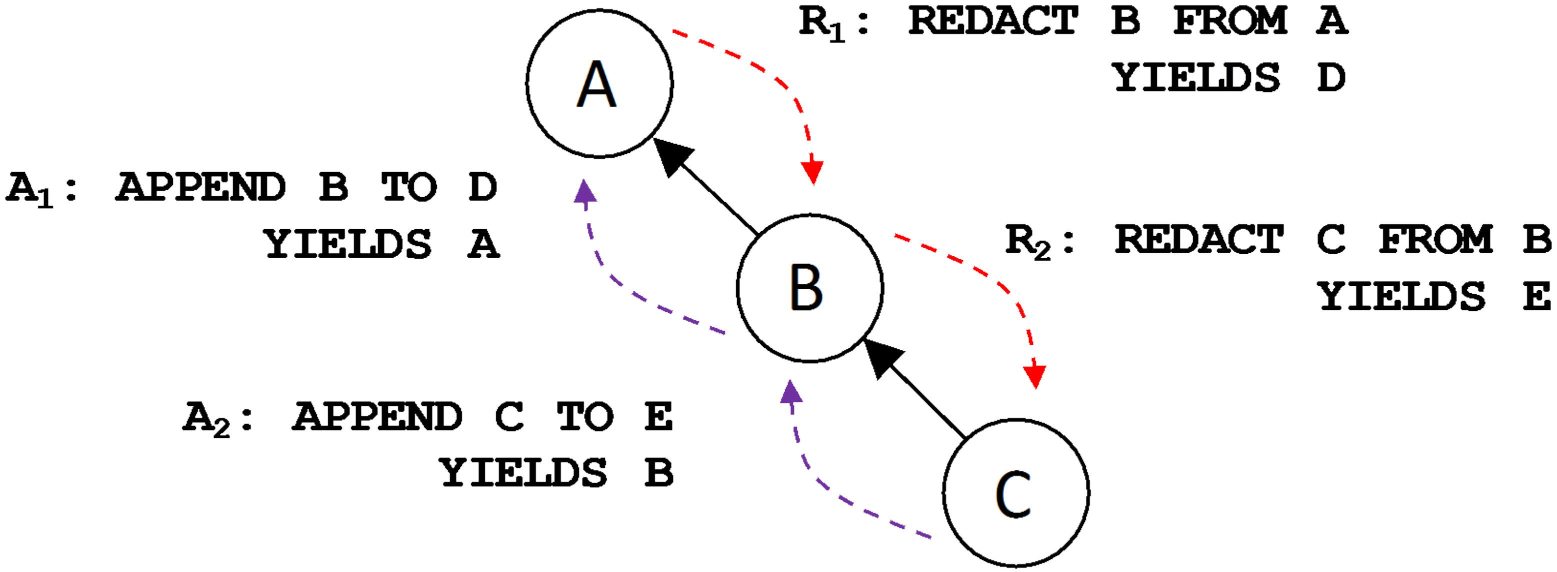
- Third-party suppliers want to see the technology works before sharing their data.
- De-anonymized datasets for public release lack the sensitive data we want to protect in our analysis.

Our Approach:

- Empirically motivated, micro-simulation based on Monte Carlo to achieve technical realism.

How to Sample Data Retention Specifications?

- Based on randomly sampling from the population of data retention agreements expressed in the extended Eddy.



R_1: REDACT B FROM A
YIELDS D

A_1: APPEND B TO D
YIELDS A

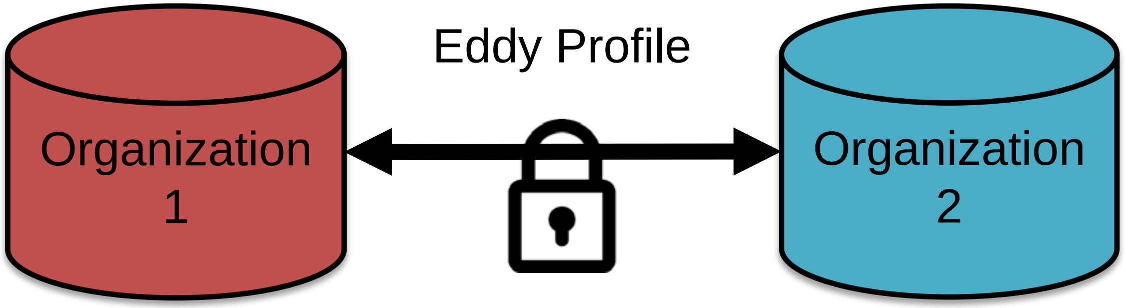
R_2: REDACT C FROM B
YIELDS E

A_2: APPEND C TO E
YIELDS B

Sampling Data Retention Specifications

- Requires expert analysis to perform data segmentation; 3 steps.
  1. Examine captured data.
  2. Determine what data is associated with a threat.
  3. Determine what data is extraneous.

- Experts use this analysis to refine the data attributes described in Eddy.



Proposed Evaluation Metrics

- Categorical re-identification approach proposed by [Swe02].
- Number of possible results (threats) proportional to uniqueness of threat with respect to data attributes.

- Threat employed by one organization easier to identify compared to threat used by many organizations.
- Threat associated with one individual easier to identify compared to association with many individuals.

[Swe02] L. Sweeney, “k-Anonymity: a model for protecting privacy”. *International Journal for Uncertainty in Fuzzy Knowledge Based Systems*, 557-567, 2002.

Proposed Evaluation Metrics

- Dataset is queried to match threat information with likely threat.

- The results of this query will be *possible* threats.

- Probability of correct identification, given threat is identified;
  - $P(correct\ identification | threat\ identified) = \frac{1}{count(ident.\ query\ results)}$

Conclusions, Lessons Learned

- Synthetic data generation process requires aggregate data, deep knowledge of data characteristics.
- Designing evaluation functions to determine data utility is bound to queries on data that are being executed.
- Tension between *business value* derived from data and *risk of unanticipated disclosure* of confidential information.

Conclusions, Lessons Learned

- Impossible to measure or predict impact of data retention actions without knowing how data will be used.

- System integrators/acquisition specialists must recognize that both data **and** queries on data have confidentiality risks built in.

- *We have proposed a method to calculate and engineer confidentiality impact; analysts must have intuition that system has inherent confidentiality risk.*

Future Work

- Use machine learning to augment and/or replace expert judgement in data segmentation process.

- Reduces the need for personnel to analyze the data, instead it can be broken up into categories automatically.

Future Work

- Specification enforcement mechanisms;
  - Prevent data from being used in unspecified way, rather than check conformance.
- Seed data with predictable data points that show evidence of unauthorized uses.
- Feedback mechanisms;
  - Collect and report/deny/redact data at runtime.

Future Work

- Specification enforcement mechanisms;

  Would allow integrators/analysts to:

  - Take control of how 3rd party services use downstream data.
  - Evaluate whether to use a 3rd party service based on whether it is truthful about specified data practices.

- Allows enforcement of new requirements for 3rd party based on analysis;
  - Can force redaction of certain sensitive data,
  - Prevent data mixture with respect to classification levels, or data types,
  - Require data to always be combined.