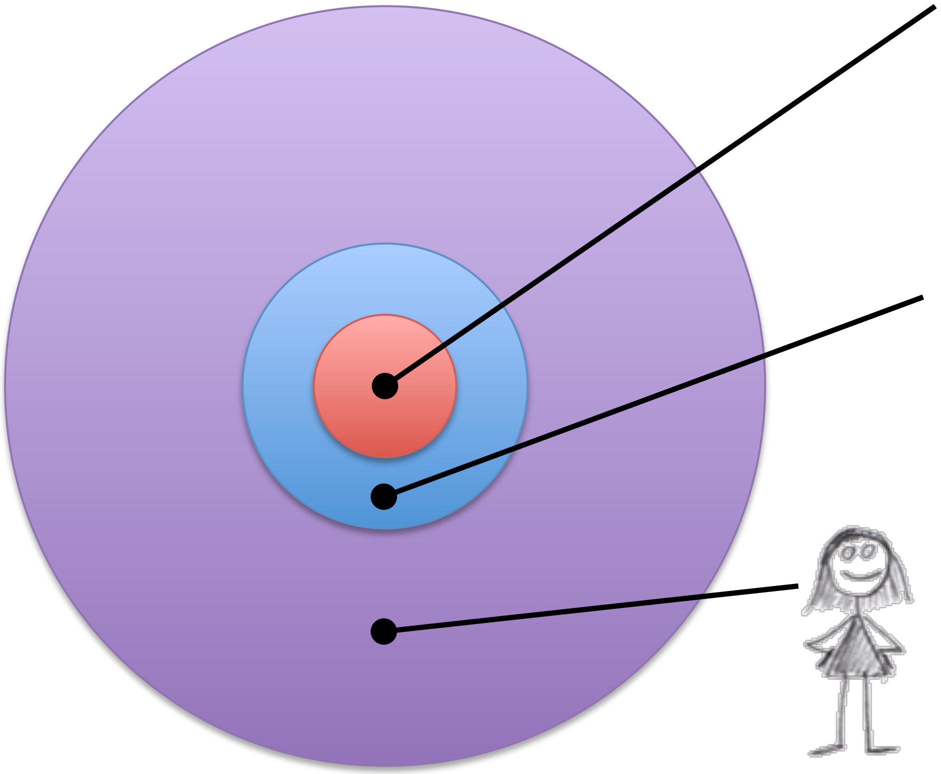
Low-Impact Data Mining

Alex Selkirk
Principal, Shan Gao Ma LLC
Board President, Common Data Project
alex.selkirk@commondataproject.org
(646) 217-0918

Evolving Data Use



Reporting

Business metrics, customer information

We made $3 MM this year

We have 150 MM customers

Business Intelligence

QOS, KPIs, Trends

Data changes the way you run your business. E.g. Wal-mart

Data-mining

Behavior analysis, correlations between data

Data-driven products and services

Cust/Personalization

e.g. Amazon's
recommendation engine

Today: High Impact Data Mining

Explosion in data sources.

Explosion in data uses, re-uses.

New markets for buying and selling data.

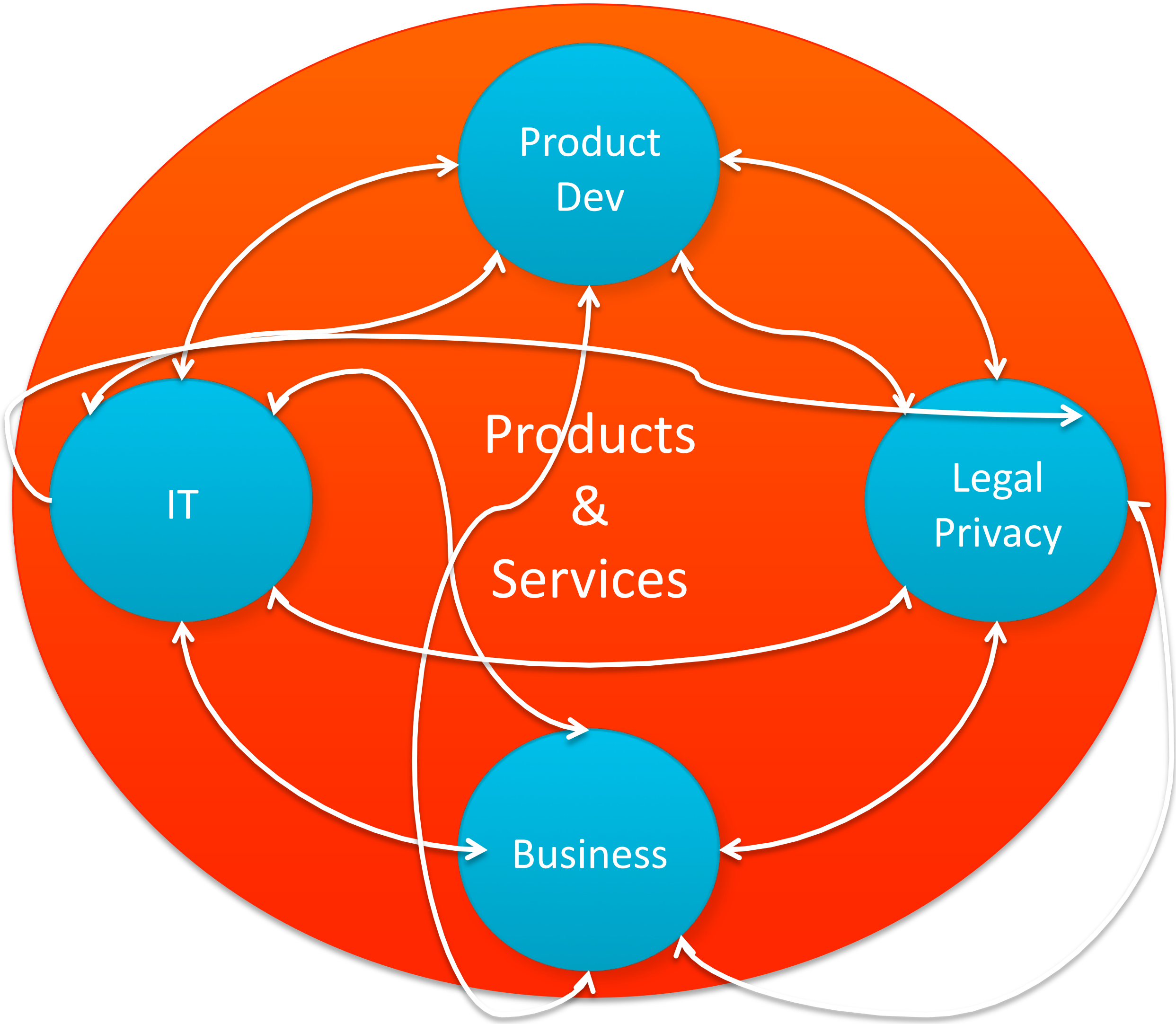
= Data-mining is no longer a periodic operation.

= Privacy policies quickly get out of sync with data-mining reality.

"Anonymous" promises usually mean broken promises.

1. De-identification doesn't work.
2. Re-identification is easy. Combining Zip code, gender, birth date can reliably identify 87% of the population. http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable

Today: High-Impact Privacy



- Knowledge of privacy decisions uneven across organization.
- Privacy is expensive.
- Enforcing policy is hard. Changes are hard to manage.
- Privacy often an afterthought.

How Do We Lower Impact?

INFRASTRUCTURE

Distributed & Collaborative Documentation System for Data-Mining==Useful by-product to loop-in legal and IT for privacy reviews and audits.

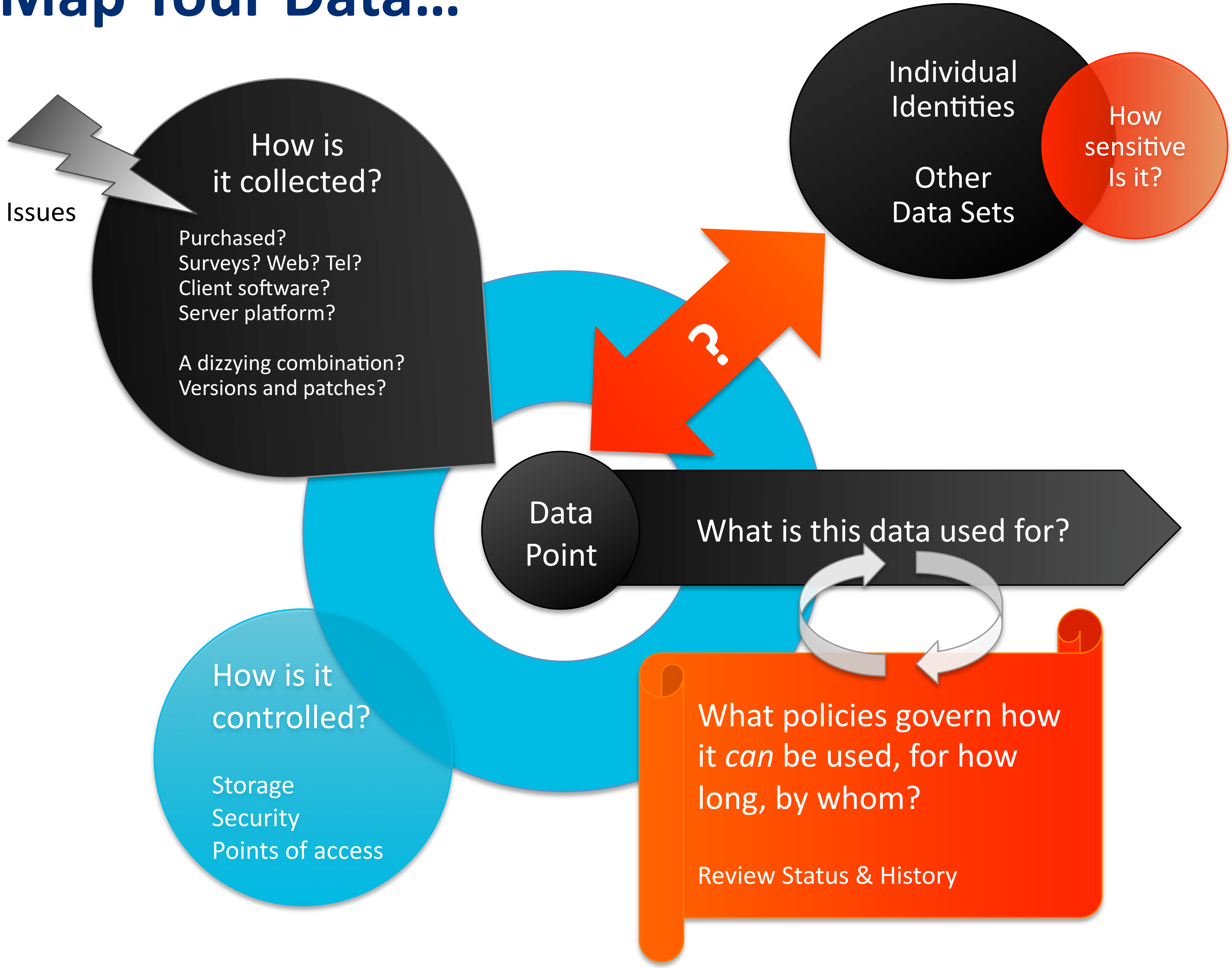
PROCESS

This enables a more iterative   review process.

It’s 2010, Do you know where your data is?

You're **collecting** how many? terabytes of what? data per day collected from which? **customers** for what? **purposes**? You are **storing** the data how? And where? for how long? Who? has **access** to it using what? protocols. What data? is directly **linked to individual identities**. What data? can’t be **combined** with what data?. What? is re-**sold**. What? is **shared with third-parties**. What data? is covered by which? **policies**. What? has **changed** recently and with what? consequences? There was a privacy **breach** where?, that **affected** what data? and which? customers?

Map Your Data…



Issues

How is it collected?

Purchased?
Surveys? Web? Tel?
Client software?
Server platform?

A dizzying combination?
Versions and patches?

Individual Identities

Other Data Sets

How sensitive Is it?

?

Data Point

What is this data used for?

How is it controlled?

Storage
Security
Points of access

What policies govern how it *can* be used, for how long, by whom?

Review Status & History

What does all this documentation enable?

INTERNALLY

Design and enforce “low-impact” privacy practices.

EXTERNALLY

Rationalize what you promise with what you can deliver.

Mitigate the impact of privacy breaches.

Resources

What is data science? (O'Reilly Radar)
http://radar.oreilly.com/2010/06/what-is-data-science.html

De-identification doesn't work. http://paulohm.com/

Re-identification with Zip code+Gender+Birthdate.
http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable

http://shangaoma.com

http://commondataproject.org

THANK YOU

Alex Selkirk

SHAN GAO MA?

The Common Data Project