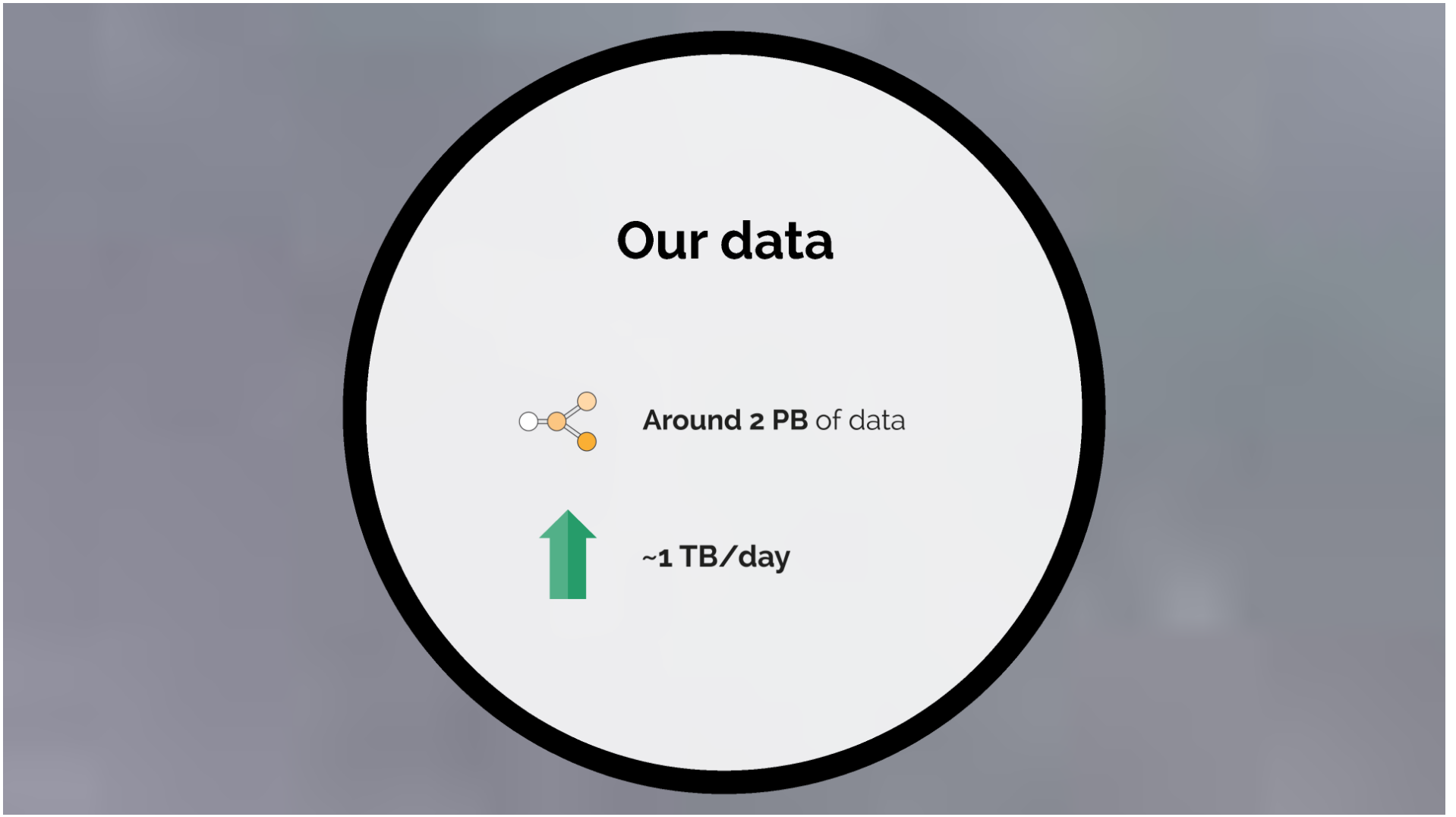
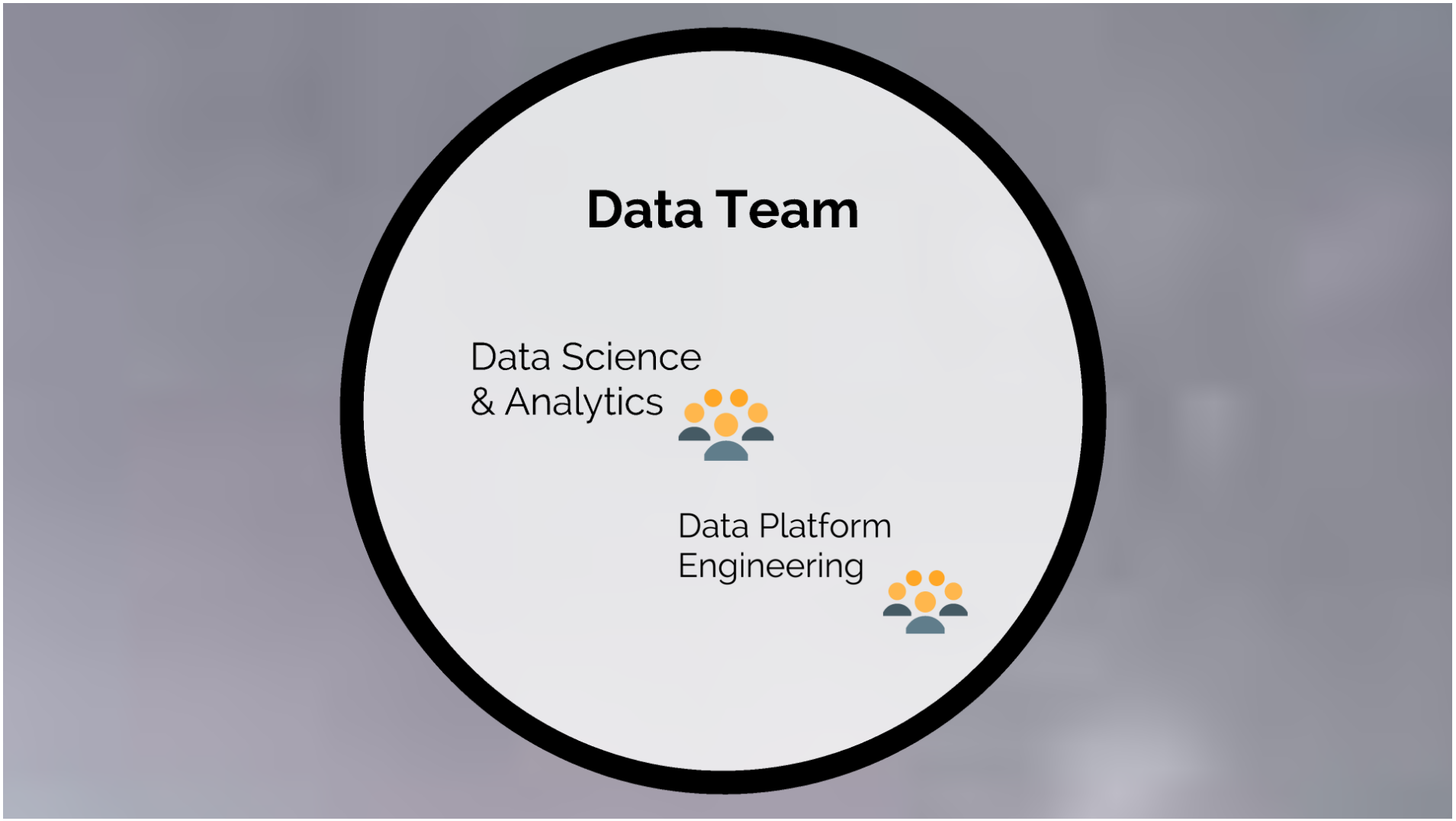
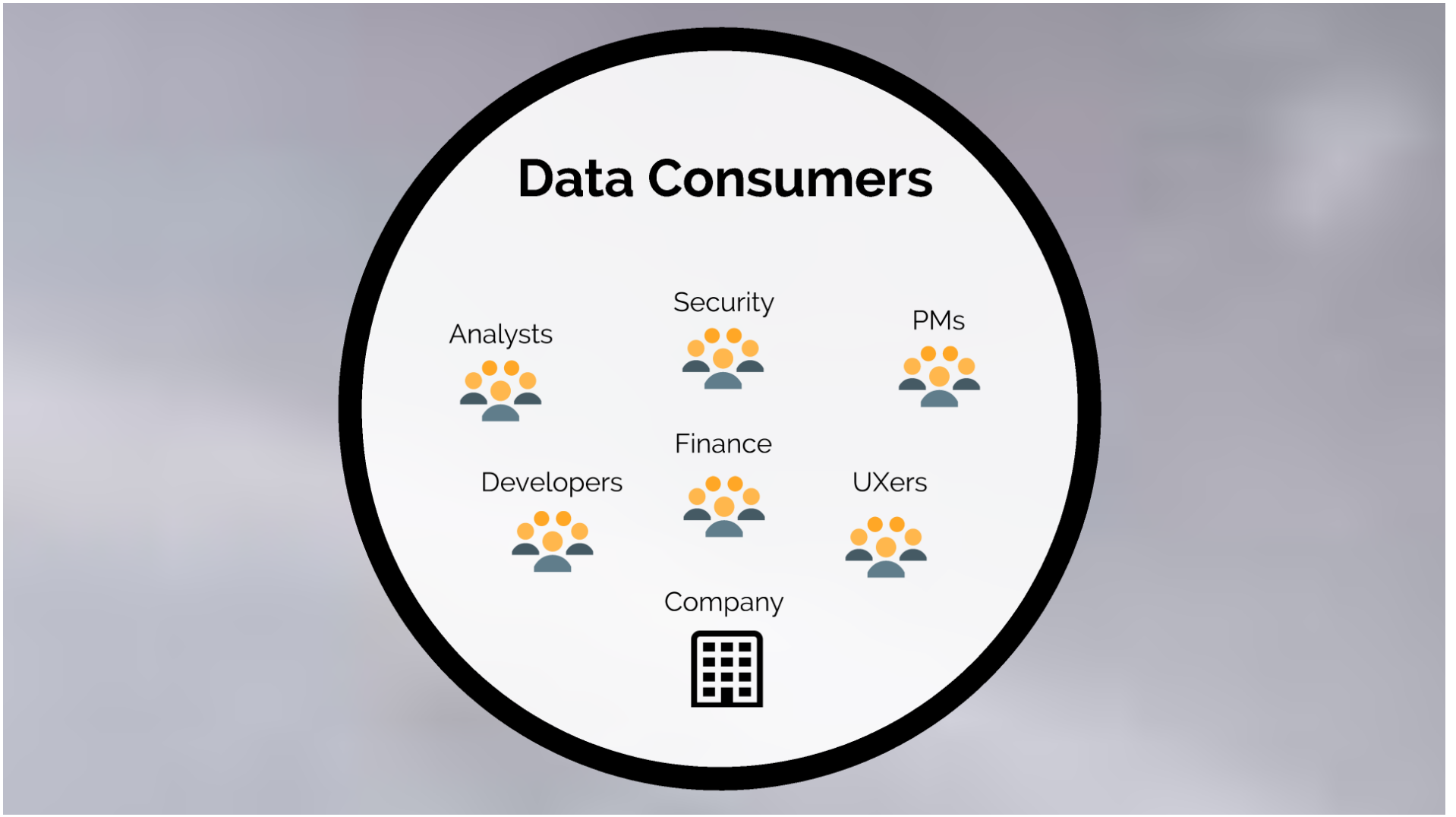

Evolution of Data Ingestion at Prezi
Tamas Nemeth
https://linkedin.com/in/nemetht/
@treff7es
Gergely Krasznai
https://www.linkedin.com/in/krasznai/
@gkrasznai


Evolution of Data Ingestion at Prezi


About Prezi
100+ million users
3.5+ billion prezi views
San Francisco
Budapest
Riga
infogr.am
Prezi
Data
Data Team
Data Consumers


Our data
Around 2 PB of data
~1 TB/day


Data Team
Data Science
& Analytics
Data Platform
Engineering


Data Consumers
Analysts
Security
PMs
Finance
Developers
UXers
Company


Evolution of Data Ingestion at Prezi


Evolution of Data Ingestion at Prezi

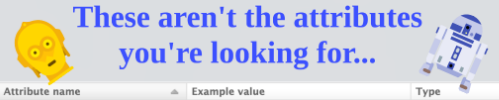

Evolution of Data Ingestion at Prezi
Challenges

Challenges

Log everything!	→	Hard to answer questions.
"Logcatalog" vs reality	→	Not in sync because not enforced.
Live logs on central machine	→	Bash power needed.
Lack of quality checks	→	ETL pipeline breaks or skews analysis if uncaught.
Importance of understanding and cleaning data	→	Time from question/idea to insight is long.

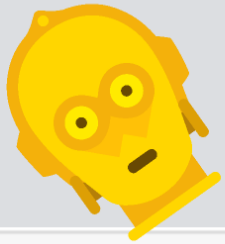
Challenges

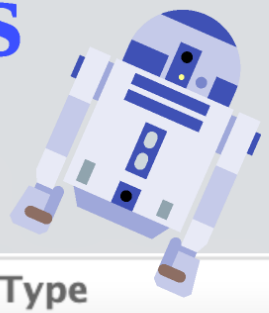
Log everything!	→	Hard to answer questions.
"Logcatalog" vs reality	→	Not in sync because not enforced.
Live logs on central machine	→	Bash power needed.
Lack of quality checks	→	ETL pipeline breaks or skews analysis if uncaught.
Importance of understanding and cleaning data	→	Time from question/idea to insight is long.



Attribute name	Example value	Type
error		string
error	no license found	string
errorcode		string
errorcode	integer	integer
errorcode	[reason]	string
errorDetails		string
errorMessage		string
errormessage	my dummy error	string
errorObj		string
errors	Authentication problem	string
error_case		string
error_code		string
error_domain	NSURLErrorDomain	string
error_message		string
error_message	user_not_found	string
error_msg	token expired	string
error_msg	fjjttj	string
error_phase		integer
error_reason	subprocess failed	string
error_string	offline OR backend_issue etc	string
error_type		string
error_type	not_compatible	string
level	ERROR	fixed list of values
media_error		integer

These aren't the attributes you're looking for...

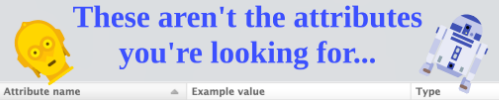




Attribute name	Example value	Type
error		string
error	no license found	string
errorcode		string
errorcode	integer	integer
errorcode	[reason]	string
errorDetails		string

Challenges

Log everything!	→	Hard to answer questions.
"Logcatalog" vs reality	→	Not in sync because not enforced.
Live logs on central machine	→	Bash power needed.
Lack of quality checks	→	ETL pipeline breaks or skews analysis if uncaught.
Importance of understanding and cleaning data	→	Time from question/idea to insight is long.



Attribute name	Example value	Type
error		string
error	no license found	string
errorcode		string
errorcode	integer	integer
errorcode	[reason]	string
errorDetails		string
errorMessage		string
errormessage	my dummy error	string
errorObj		string
errors	Authentication problem	string
error_case		string
error_code		string
error_domain	NSURLErrorDomain	string
error_message		string
error_message	user_not_found	string
error_msg	token expired	string
error_msg	fjjttj	string
error_phase		integer
error_reason	subprocess failed	string
error_string	offline OR backend_issue etc	string
error_type		string
error_type	not_compatible	string
level	ERROR	fixed list of values
media_error		integer

Challenges

Log everything!	→	Hard to answer questions.
"Logcatalog" vs reality	→	Not in sync because not enforced.
Live logs on central machine	→	Bash power needed.
Lack of quality checks	→	ETL pipeline breaks or skews analysis if uncaught.
Importance of understanding and cleaning data	→	Time from question/idea to insight is long.


Evolution of Data Ingestion at Prezi
Challenges

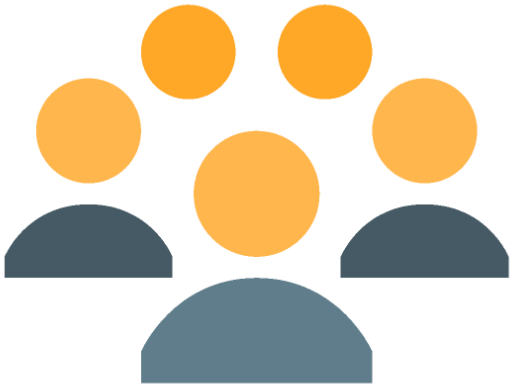

Evolution of Data Ingestion at Prezi
Challenges
XFTs

At the birth of data

"Glassbox" project with clear value proposition

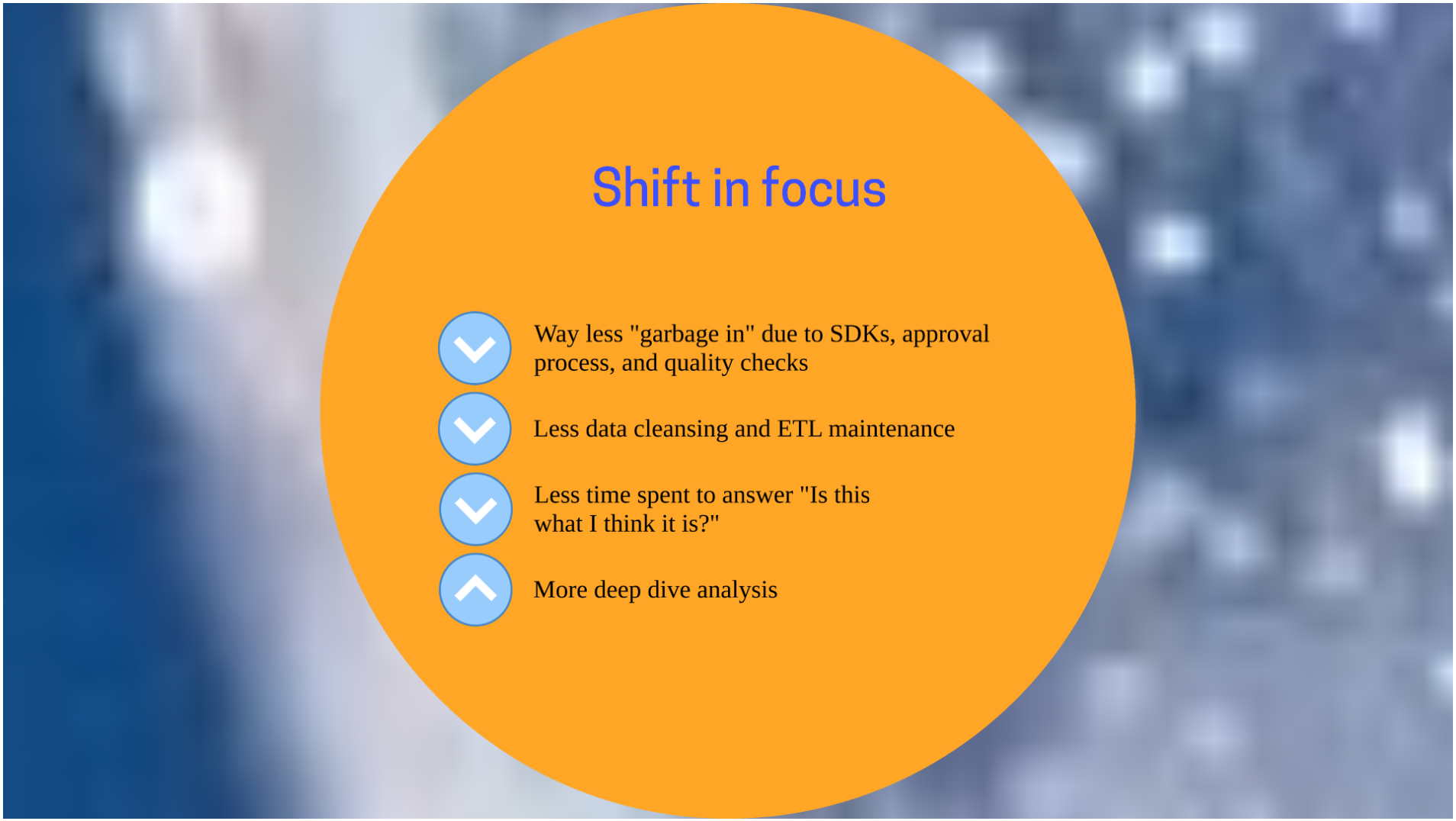
Support and effort from both bottom and top

Rewiring the whole product



Glassbox

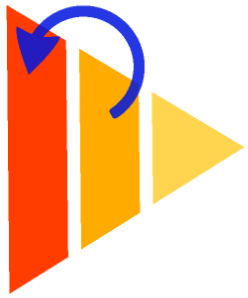
1. Reduce time from question/idea to insight
2. Improve quality and consistency of logs
3. Empower more parts of the organisation


Shift in focus
Way less "garbage in" due to SDKs, approval process, and quality checks
Less data cleansing and ETL maintenance
Less time spent to answer "Is this what I think it is?"
More deep dive analysis


Evolution of Data Ingestion at Prezi
Challenges
XFTs

Evolution of Data Ingestion at Prezi

Enforced review process

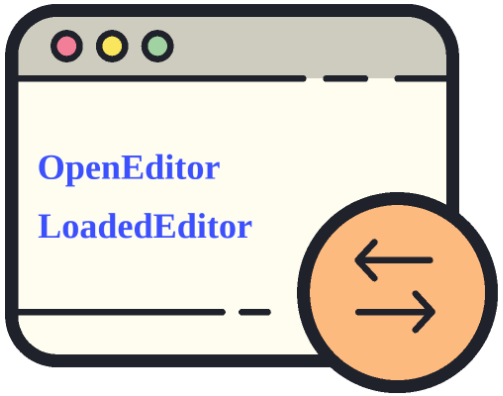


1. Transparency // job story loops, click flows

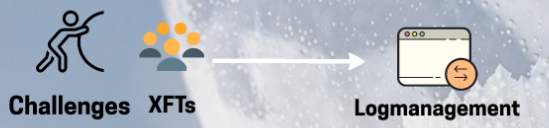
2. Naming convention // track user intent
   (and some technical events)

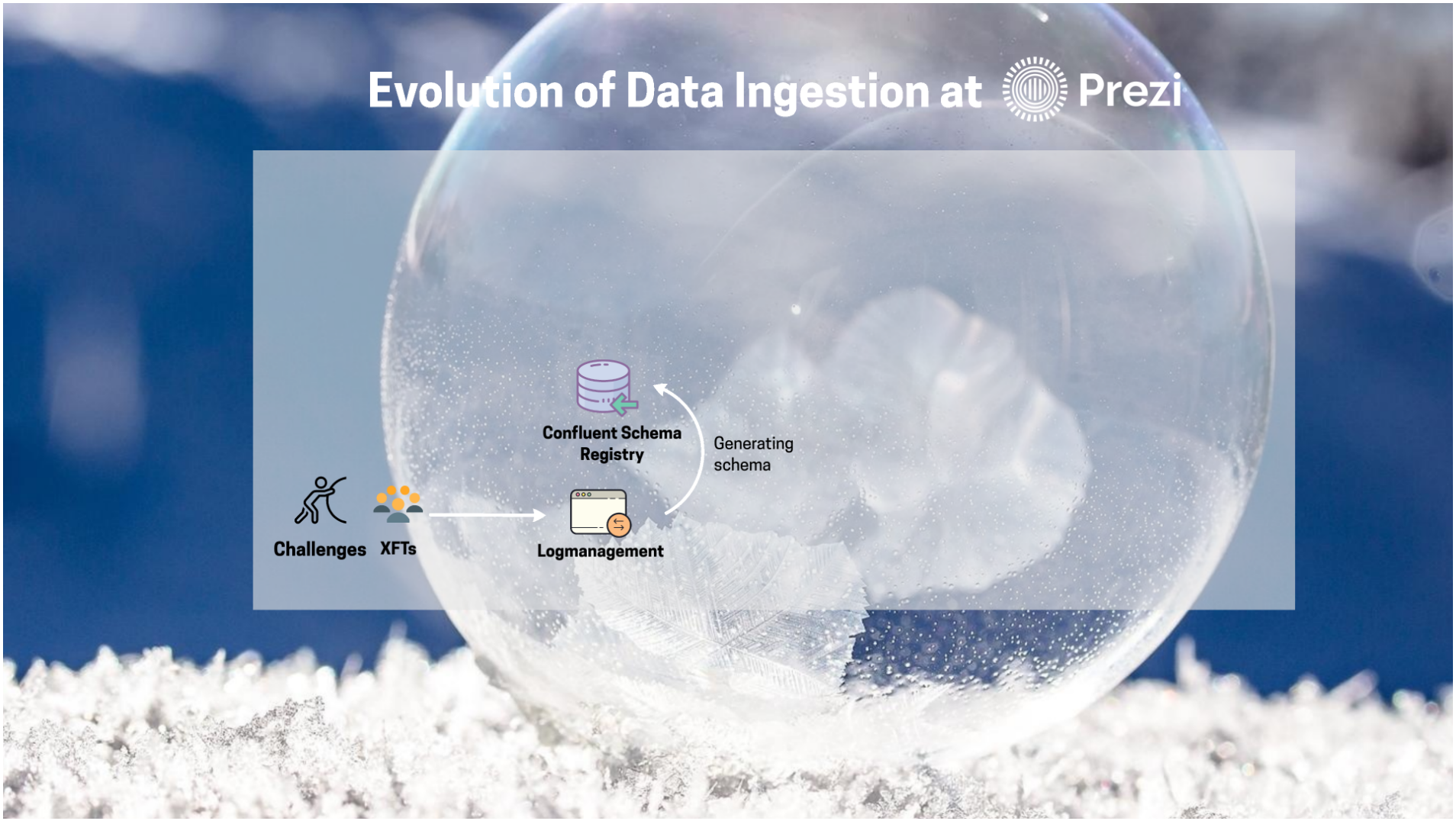
3. Events and info attached universal across platforms

Each event lives in one place and has to get **APPROVAL** from the admins before developers can touch the code



Evolution of Data Ingestion at Prezi




Evolution of Data Ingestion at Prezi
Confluent Schema Registry
Generating schema
Challenges
XFTs
Logmanagement

Generating schema for every platform

Platform specific strict schema

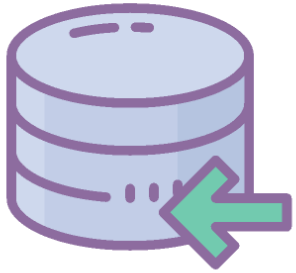
```
{
  "name": "platform",
  "type": [
    "type": "record",
    "name": "Platform",
    "fields": [{
      "name": "type",
      "type": [
        "type": "enum",
        "name": "PlatformType",
        "symbols": ["Desktop"]
      ],
      "doc": "Type of the platform"
    }, {
      "name": "device_id",
      "type": "string",
      "doc": "Device ID",
      "generated_by": "ClientSide"
    }]
  ]
}
```

No Compatibility

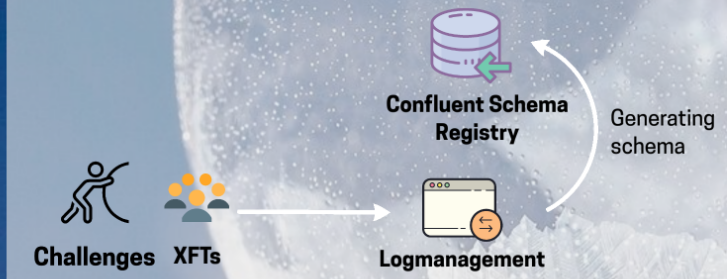
Generic loose schema

```
{
  "name": "platform",
  "type": [
    "type": "record",
    "name": "Platform",
    "fields": [{
      "name": "type",
      "type": "string",
      "doc": "Type of the platform",
      "valid_values": ["Desktop", "Web", "Mobile"]
    }, {
      "name": "device_id",
      "type": ["null", "string"],
      "doc": "Device ID",
      "default": null,
      "generated_by": "ClientSide"
    }]
  ]
}
```

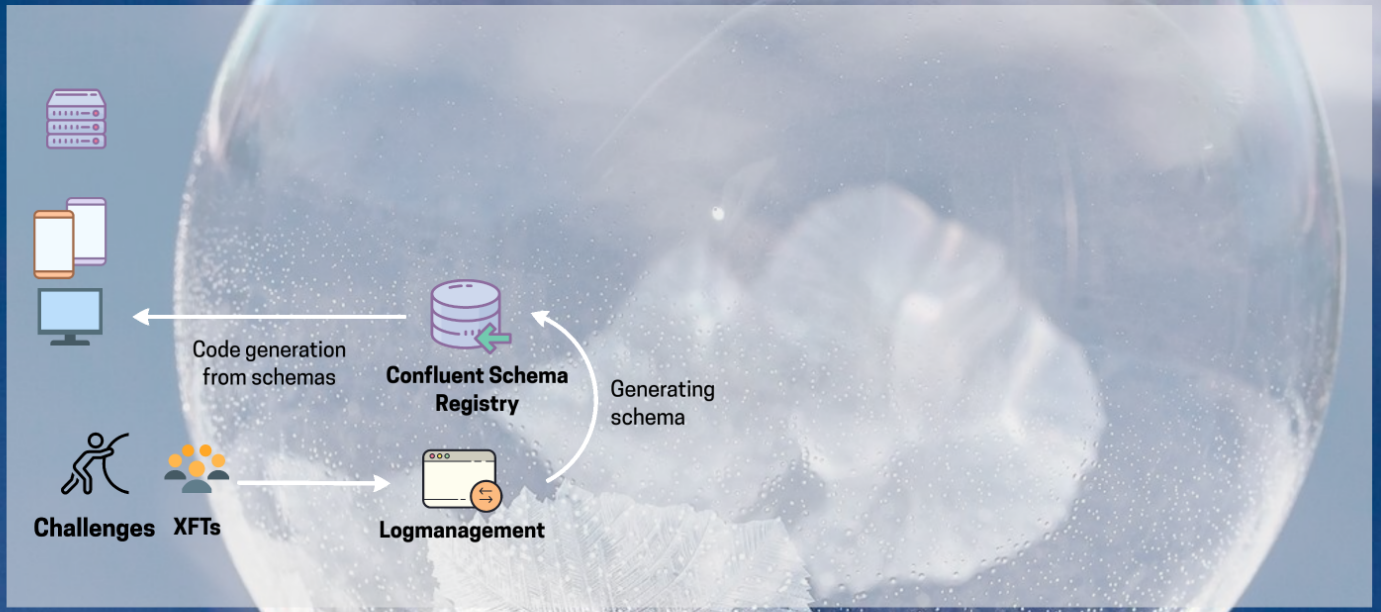
Full compatibility



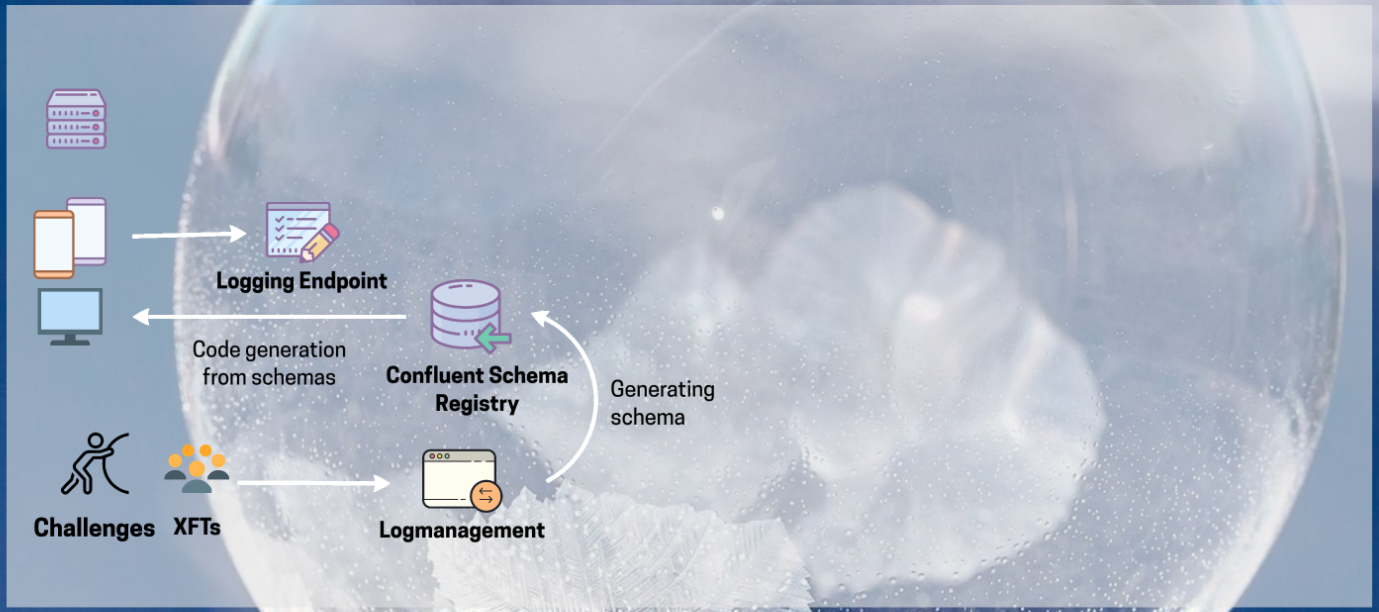
Evolution of Data Ingestion at Prezi



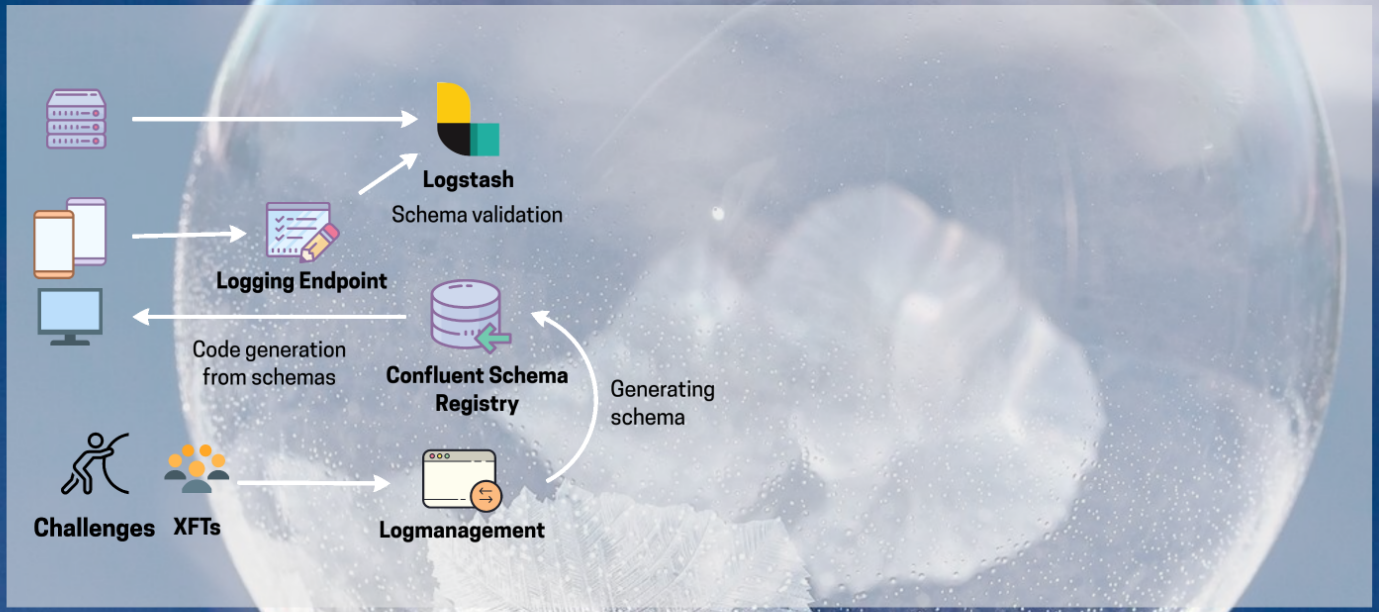
Evolution of Data Ingestion at Prezi



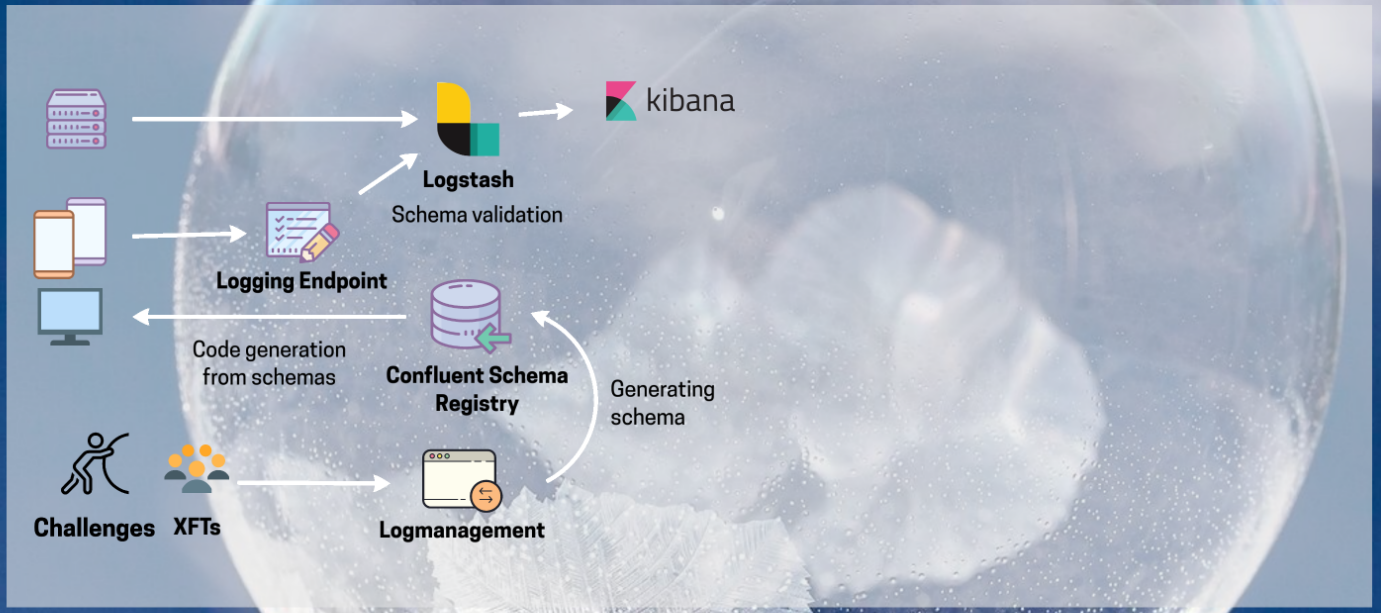
Evolution of Data Ingestion at Prezi



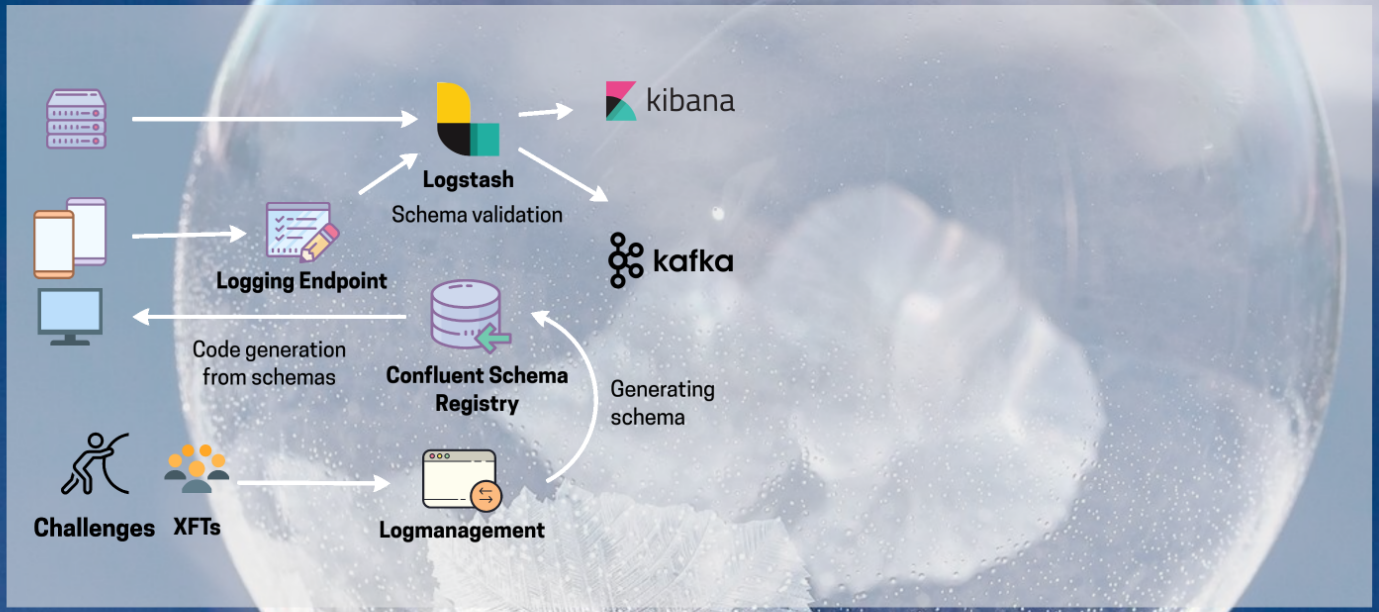
Evolution of Data Ingestion at Prezi



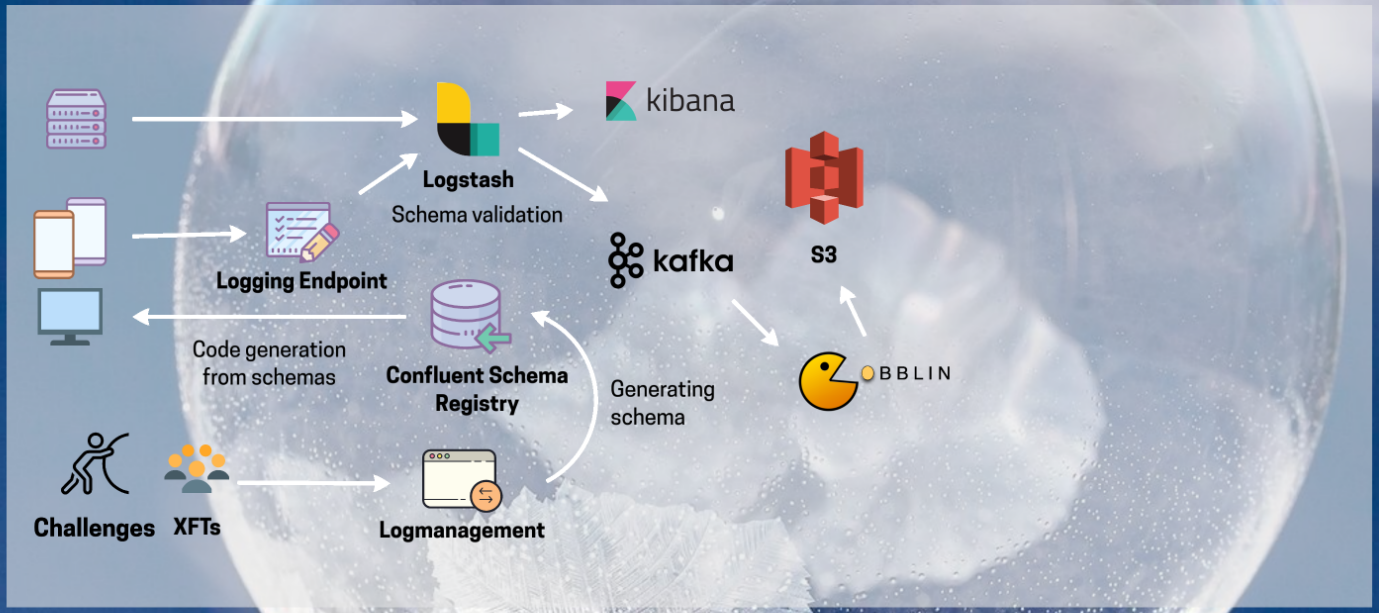
Evolution of Data Ingestion at Prezi



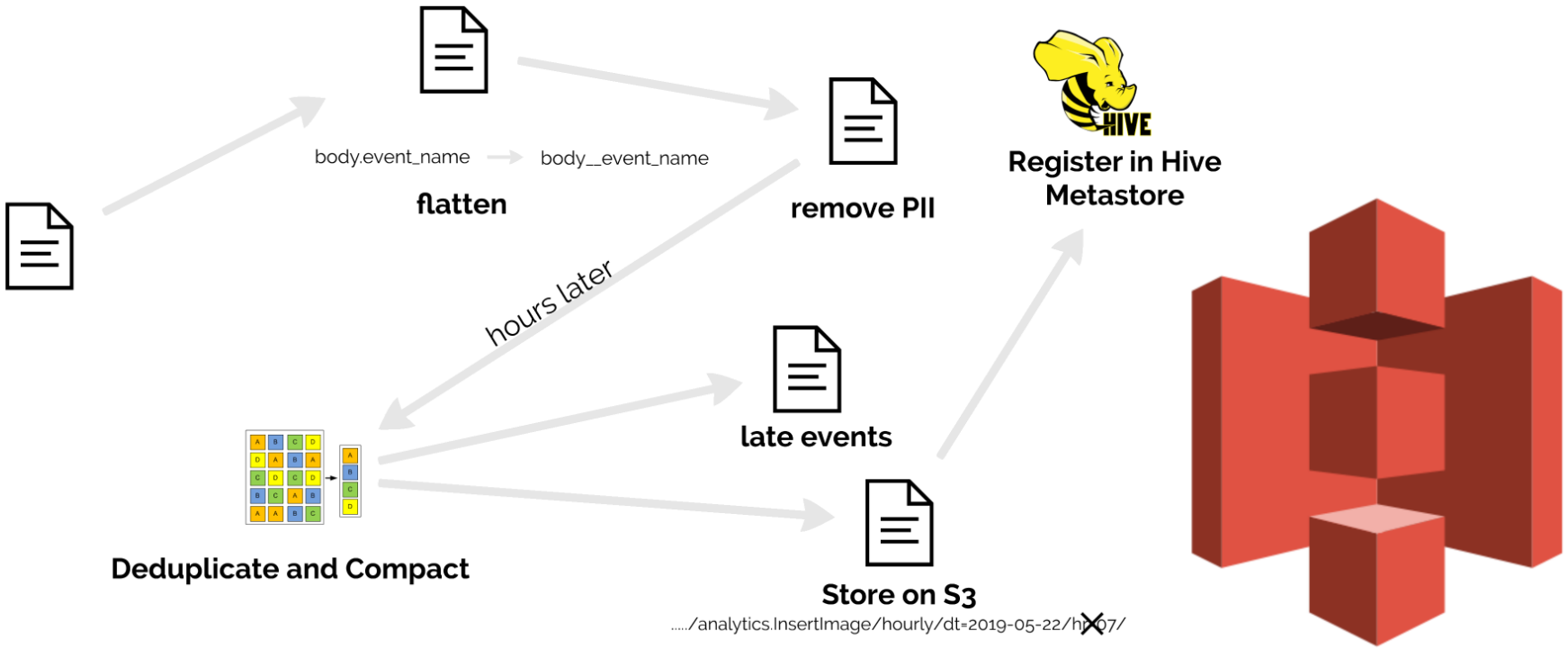
Evolution of Data Ingestion at Prezi



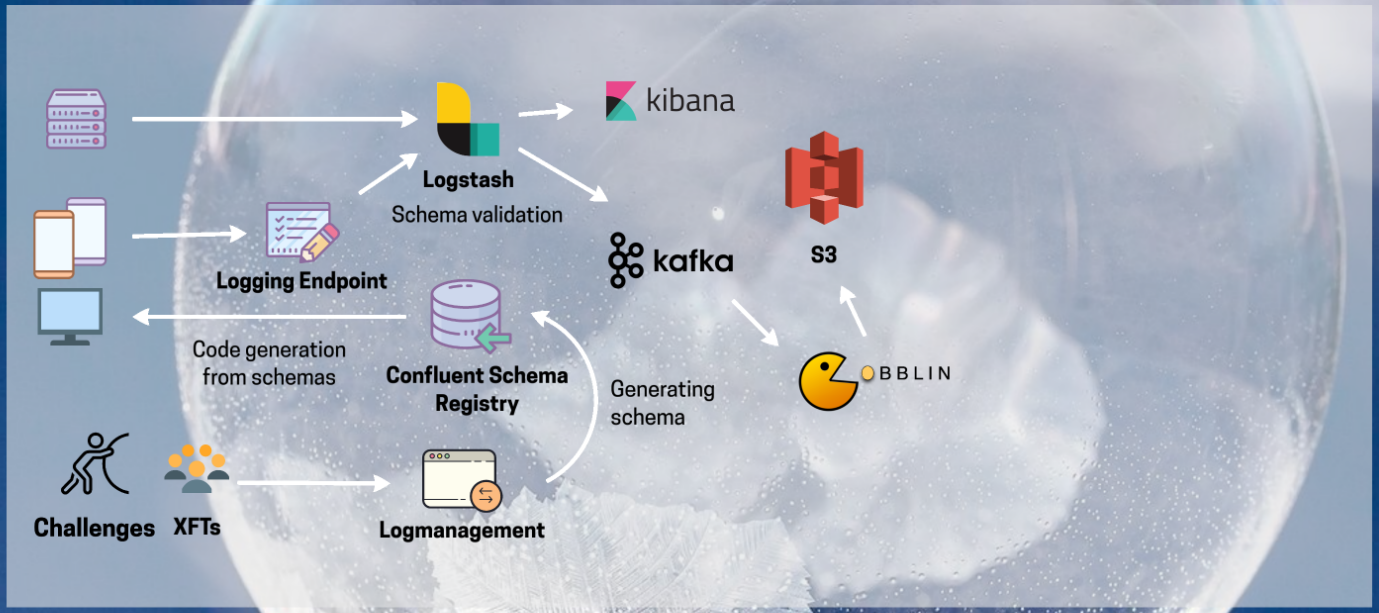
Evolution of Data Ingestion at Prezi

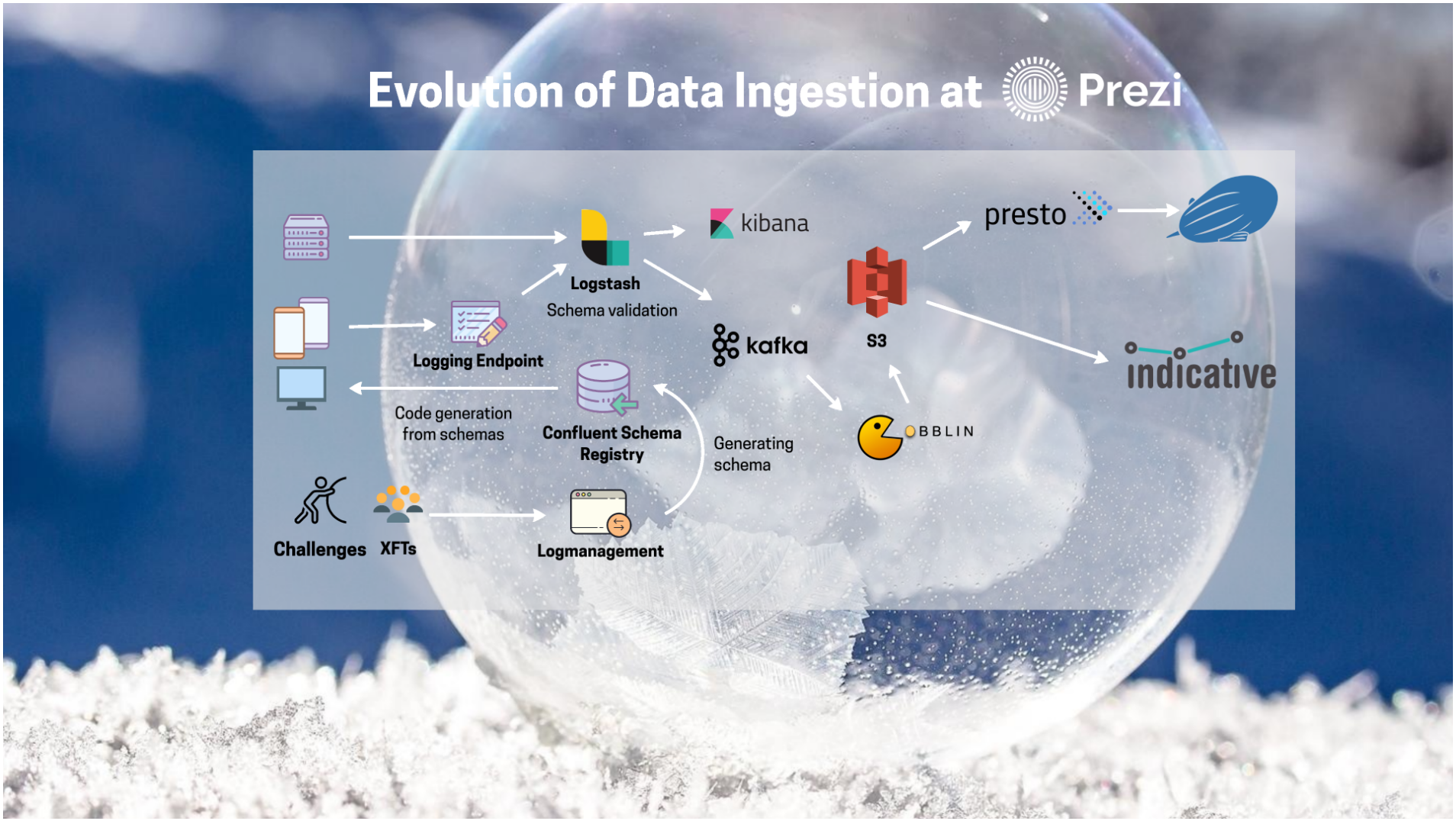


Storing events on S3

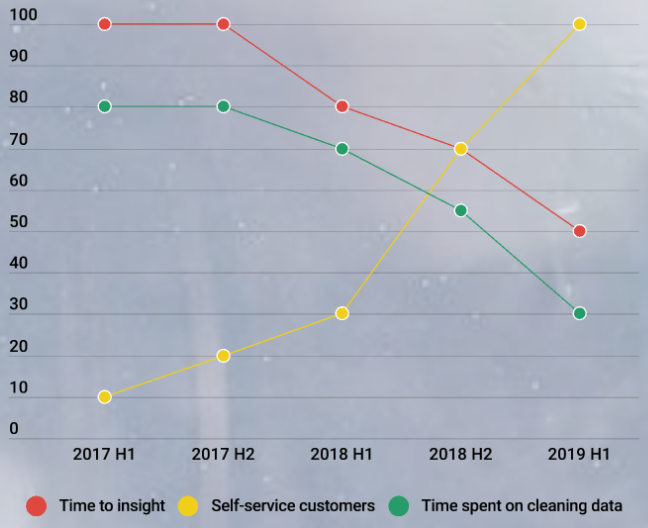


Evolution of Data Ingestion at Prezi




Evolution of Data Ingestion at Prezi
Logstash
Schema validation
kibana
kafka
S3
presto
indicative
Logging Endpoint
Code generation
from schemas
Confluent Schema
Registry
Generating
schema
BBLIN
Challenges
XFTs
Logmanagement

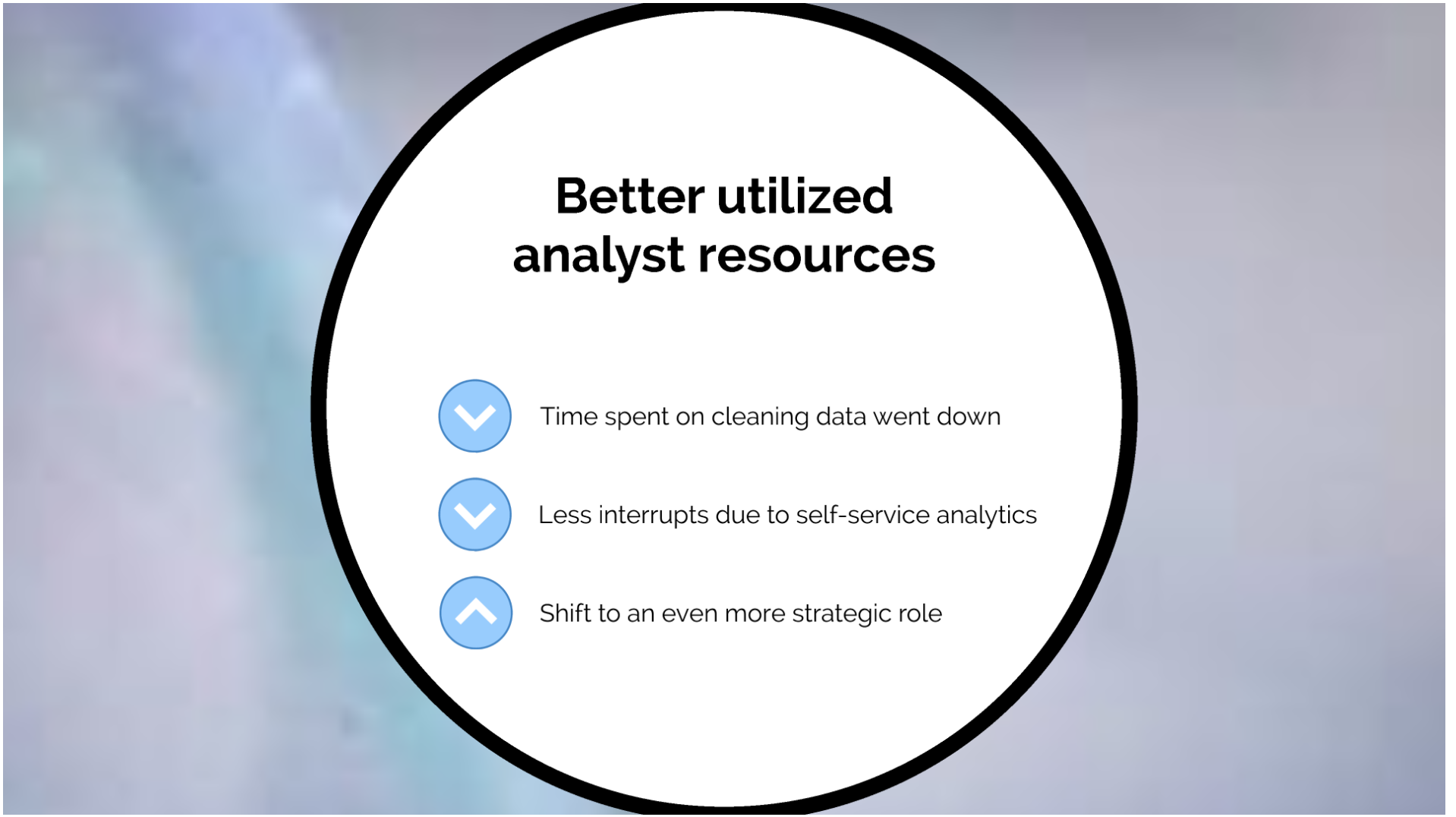
Results

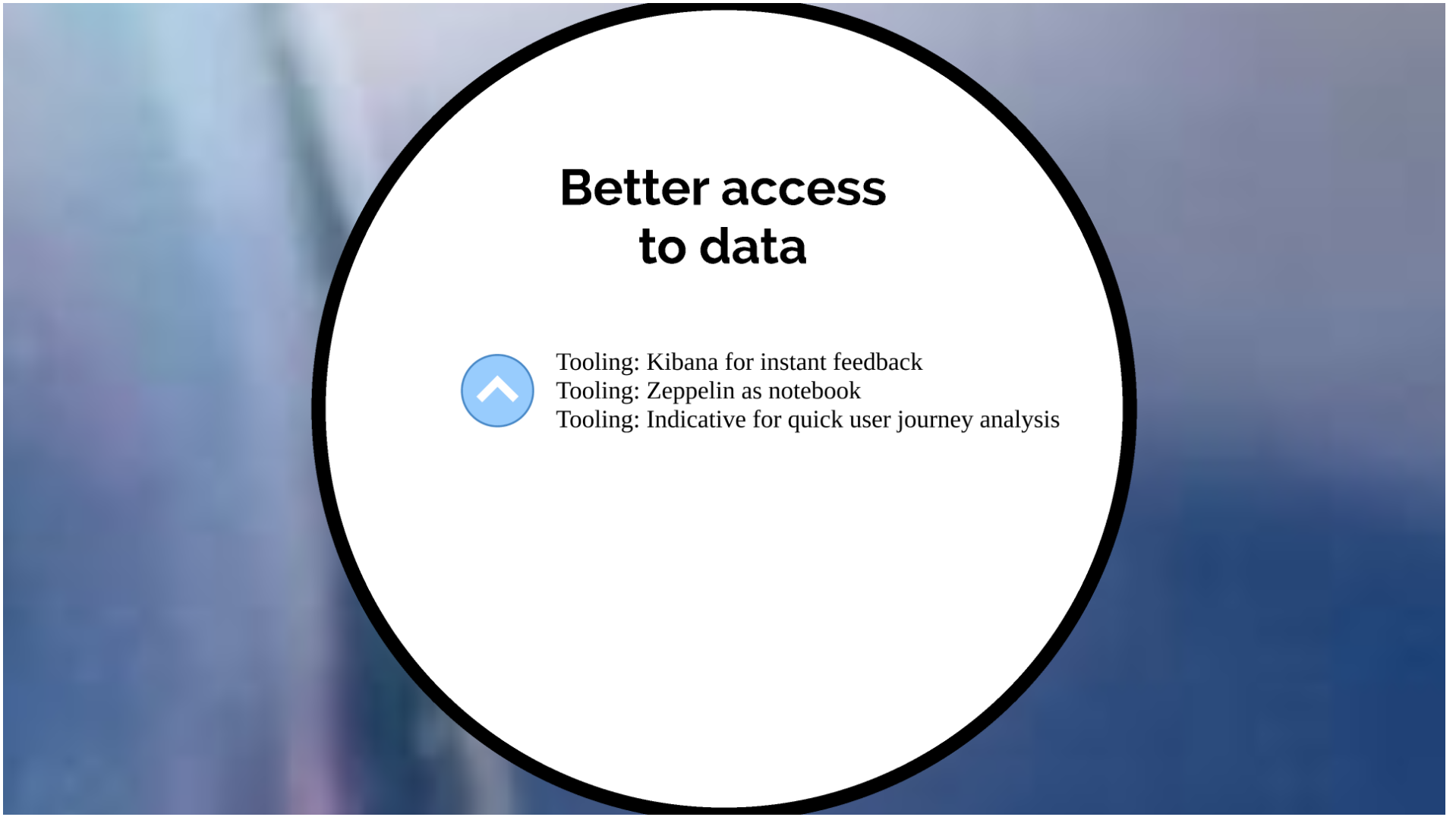


DS&A team

Company

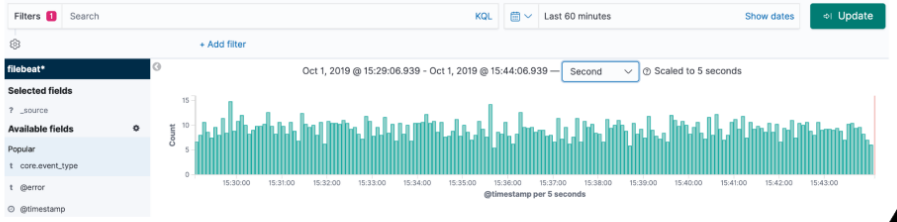
Next steps

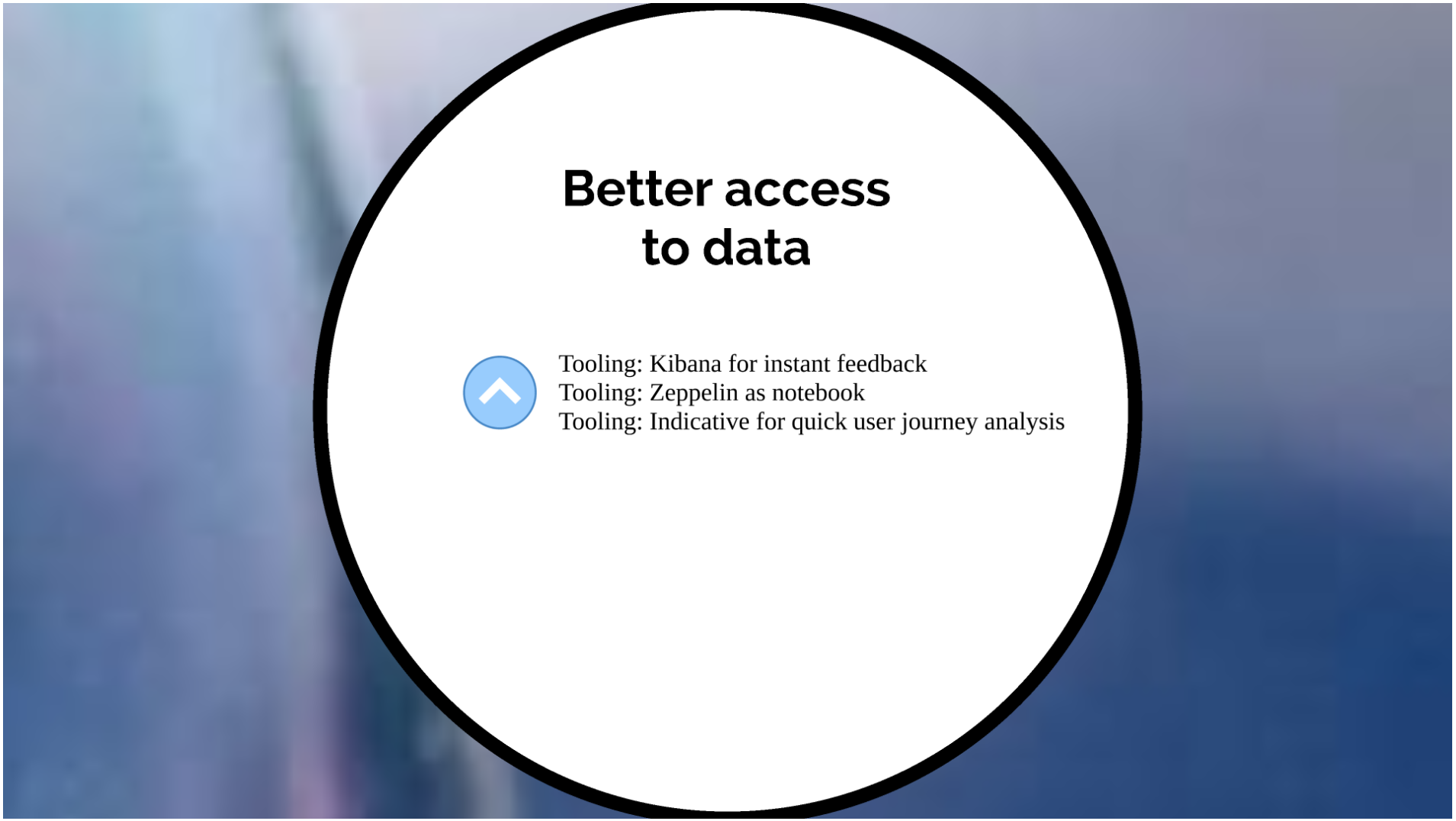

Better utilized analyst resources
Time spent on cleaning data went down
Less interrupts due to self-service analytics
Shift to an even more strategic role


Better access
to data
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis

Better access to data

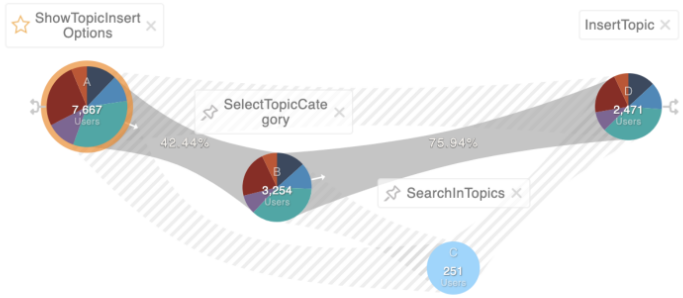
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis

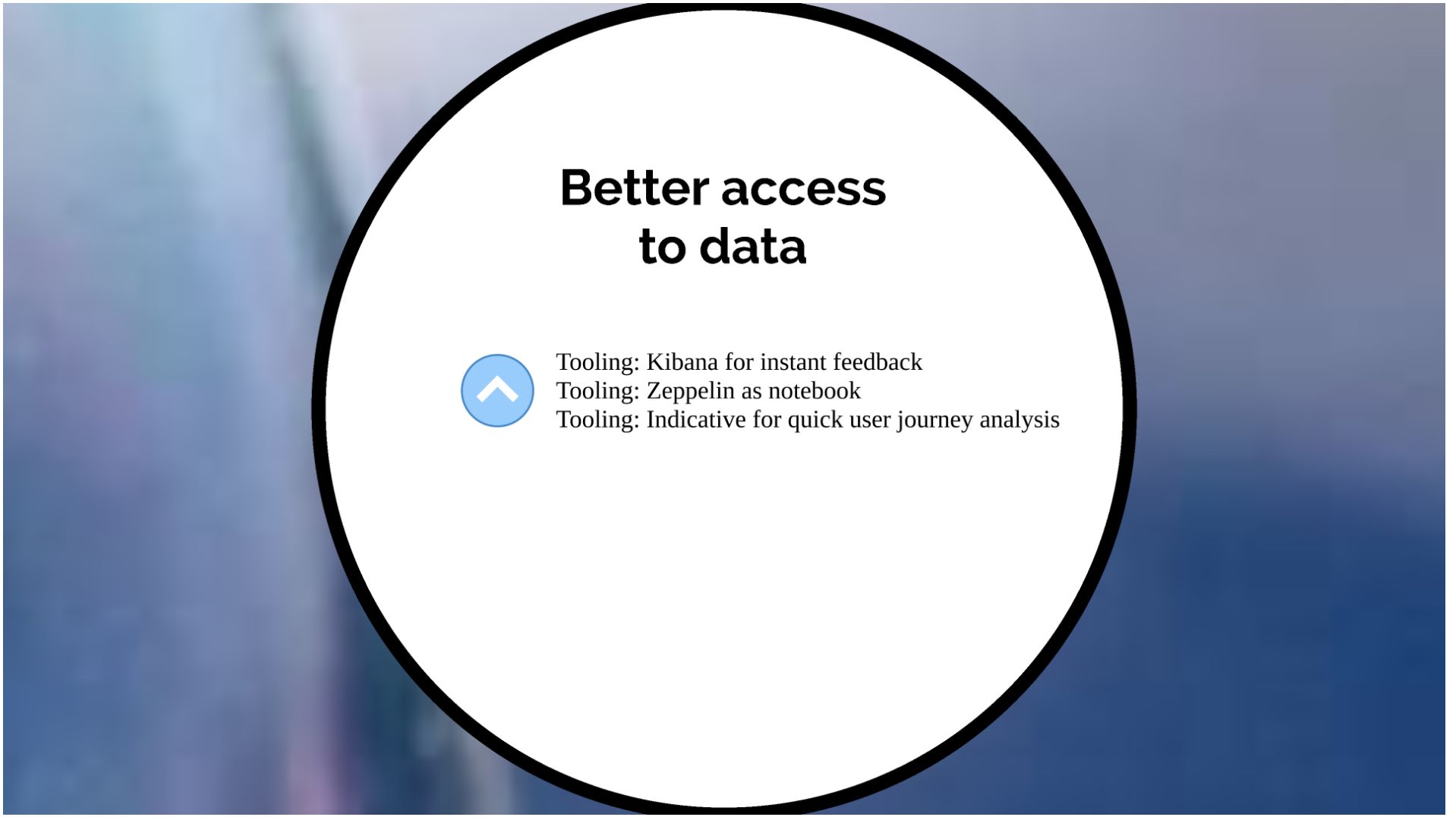


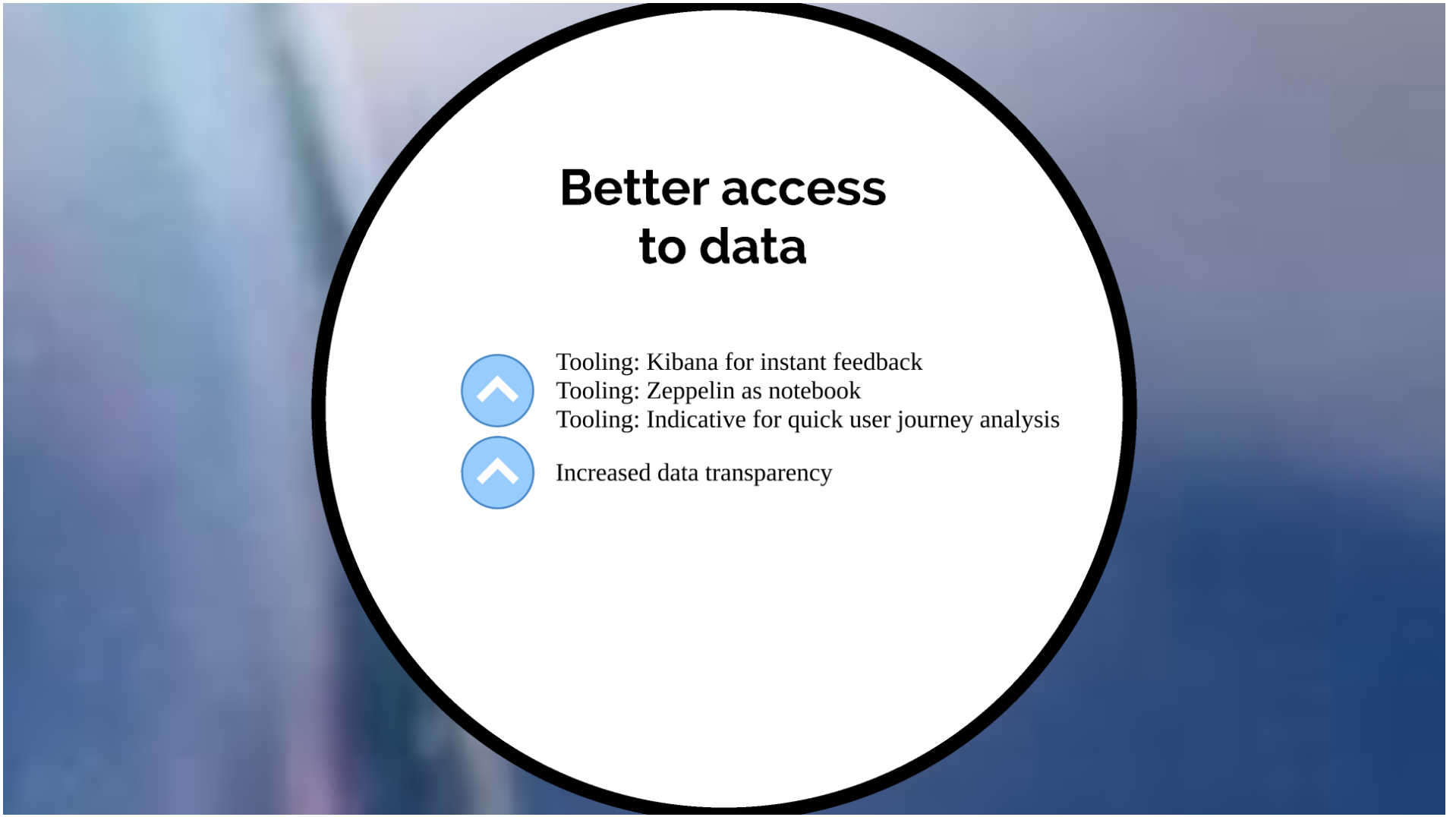

Better access
to data
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis

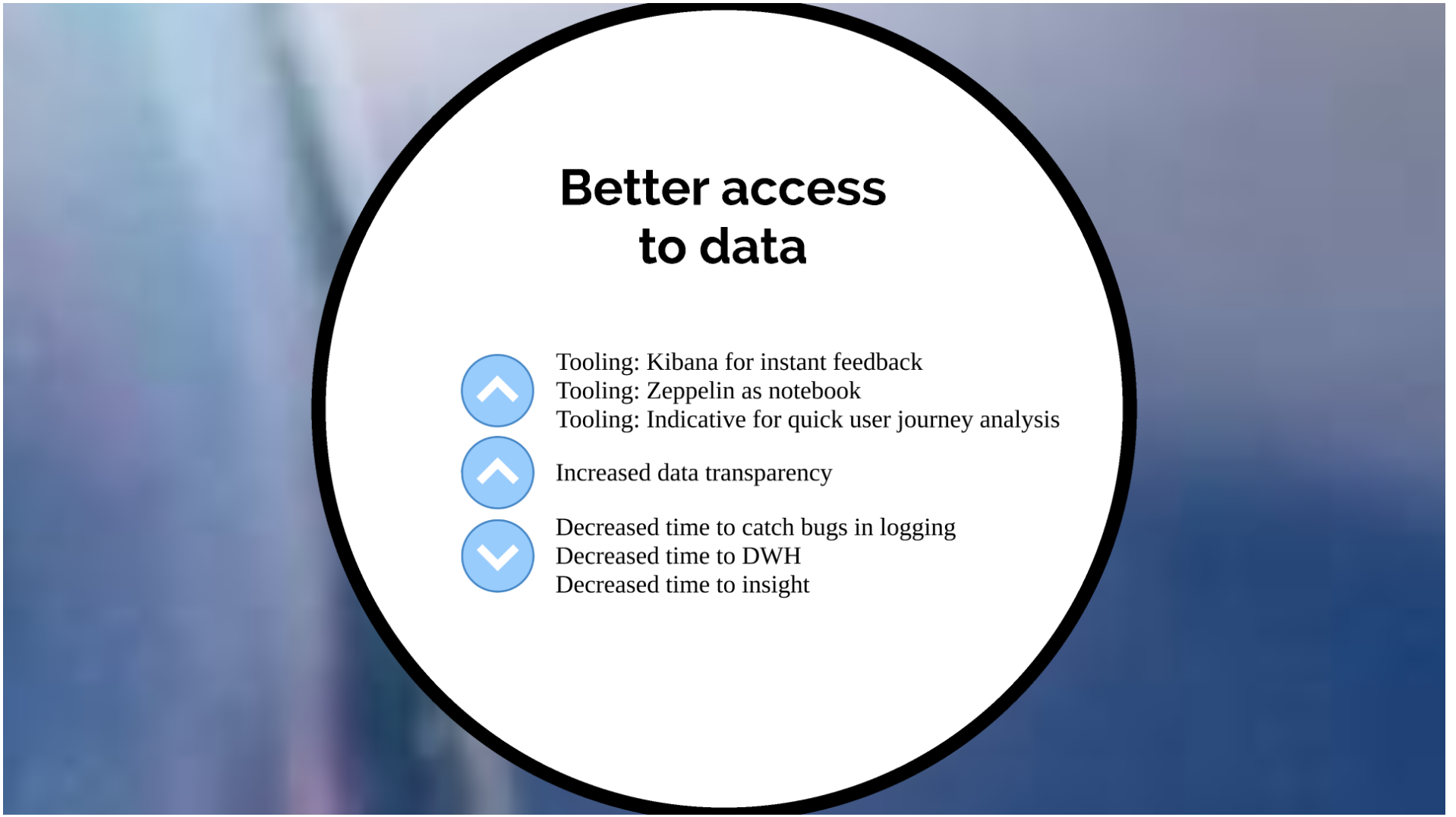
Better access to data

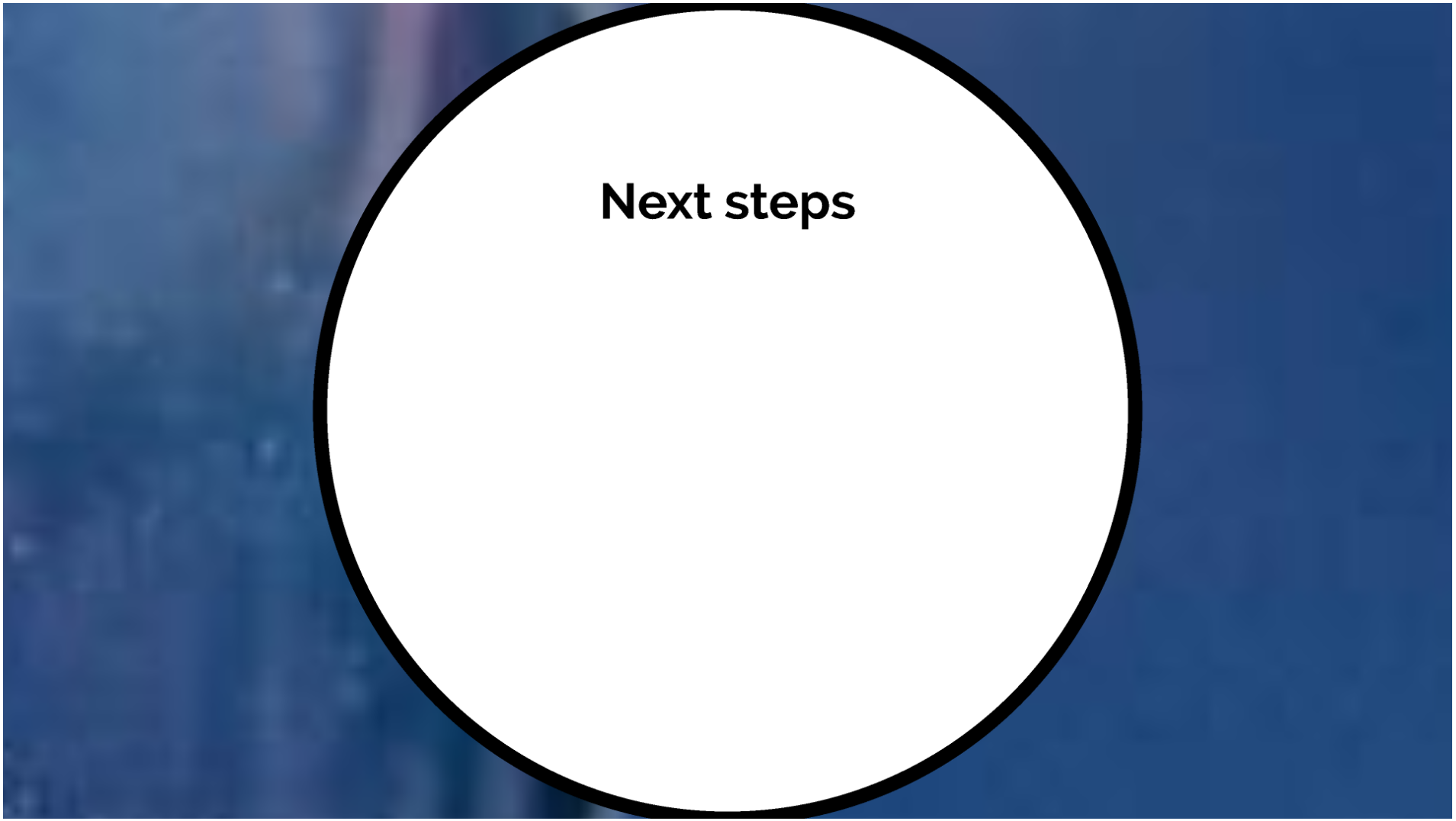
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis

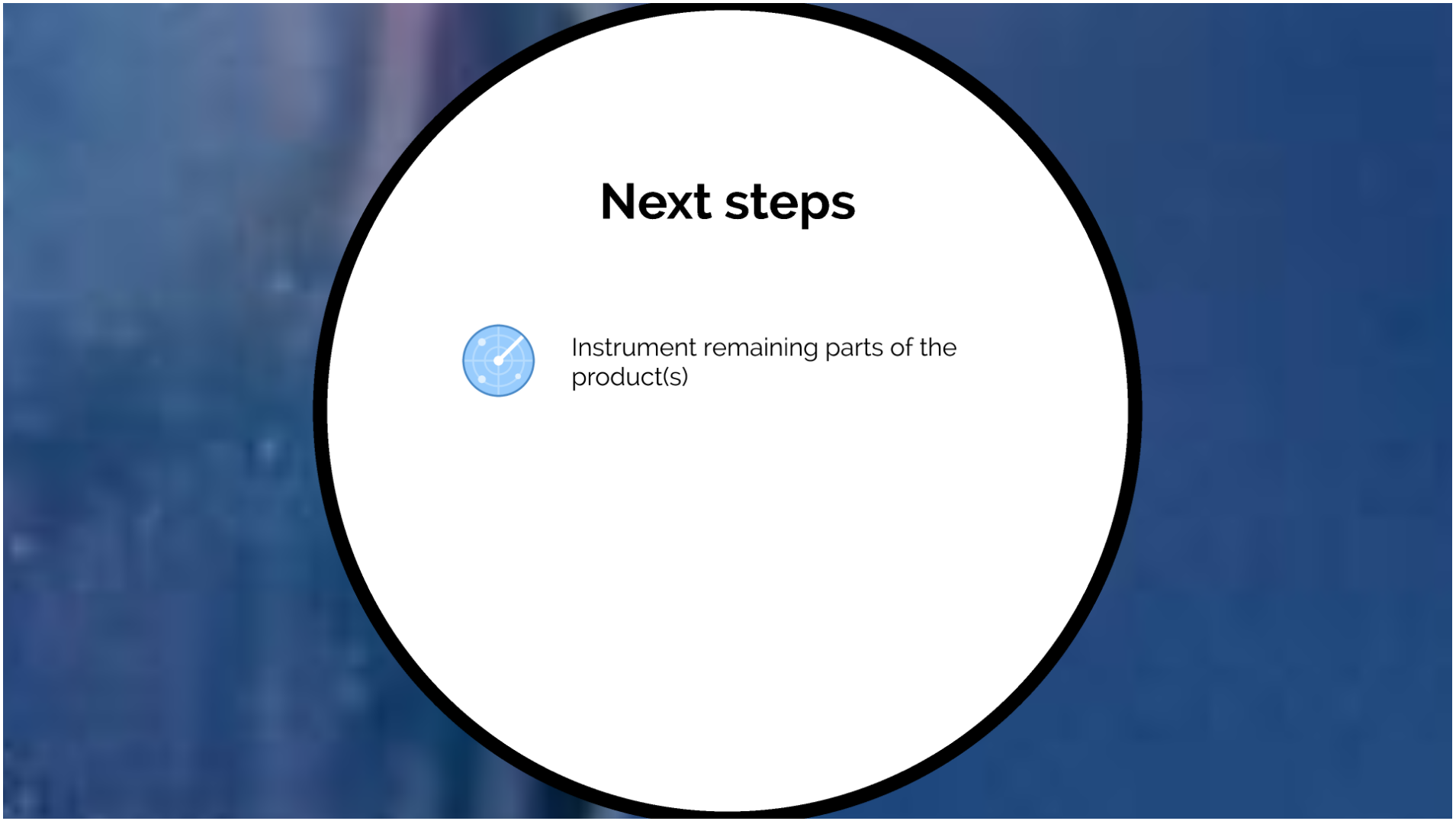



Better access
to data
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis


Better access
to data
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis
Increased data transparency


Better access
to data
Tooling: Kibana for instant feedback
Tooling: Zeppelin as notebook
Tooling: Indicative for quick user journey analysis
Increased data transparency
Decreased time to catch bugs in logging
Decreased time to DWH
Decreased time to insight


Next steps


Next steps
Instrument remaining parts of the product(s)

Next steps

Instrument remaining parts of the product(s)

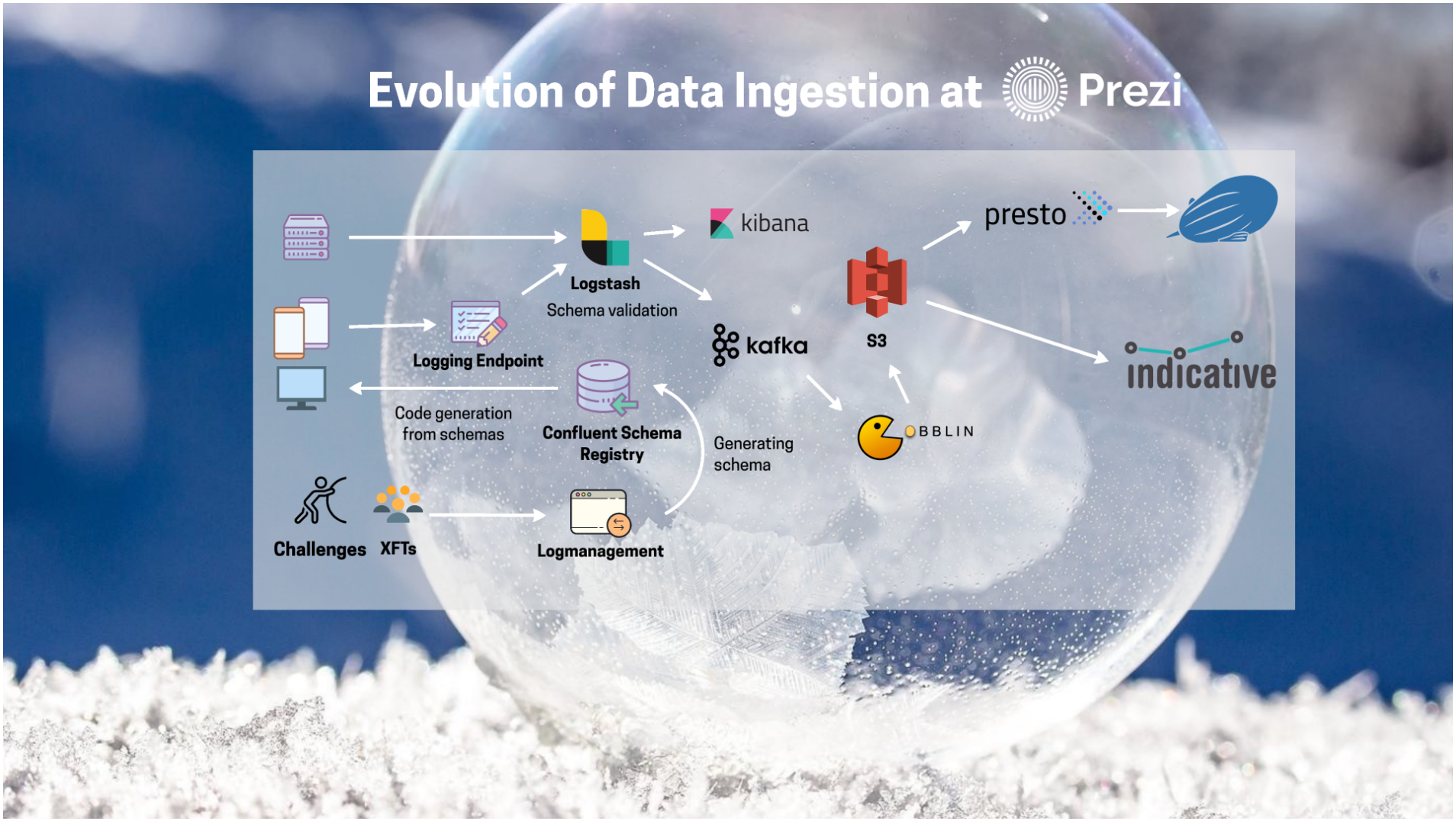
Onboarding non-analytical events

Next steps

Instrument remaining parts of the product(s)

Onboarding non-analytical events

More widespread use of real-time usage triggers


Evolution of Data Ingestion at Prezi
Logstash
Schema validation
kibana
presto
Logging Endpoint
kafka
S3
indicative
Code generation
from schemas
Confluent Schema
Registry
Generating
schema
BBLIN
Challenges
XFTs
Logmanagement


Evolution of Data Ingestion at Prezi


Evolution of Data Ingestion at Prezi
Tamas Nemeth
https://linkedin.com/in/nemetht/
@treff7es
Gergely Krasznai
https://www.linkedin.com/in/krasznai/
@gkrasznai