Demystifying the Data Scientist

A case study From BankID Norway

Thomas Hillesøy and Jørn Kristian Stensby, *Sens Consulting AS*
Erik Larsen, *BankID Norway AS*

October 2018

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

splunk> .conf18

Speakers

Jørn Kristian Stensby

Senior Solution Architect

Thomas Hillesøy

Senior Solution Architect

Erik Larsen

Head of Operation

Agenda

Demystifying the data scientist

▸ BankID Norway and their needs

▸ Data collection

▸ Data engineering

▸ Data analysis

▸ Q&A

*"A **data scientist** is a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist **extracts insights from data sets** for product development, and evaluates and identifies strategic opportunities."*

splunk> .conf18

BankID

WE ENSURE SAFE AND SECURE ONLINE IDENTIFICATION AND SIGNING

▸ BankID's goal: Secure digitalization of the society

▸ 84% of Norway's eligible population uses BankID
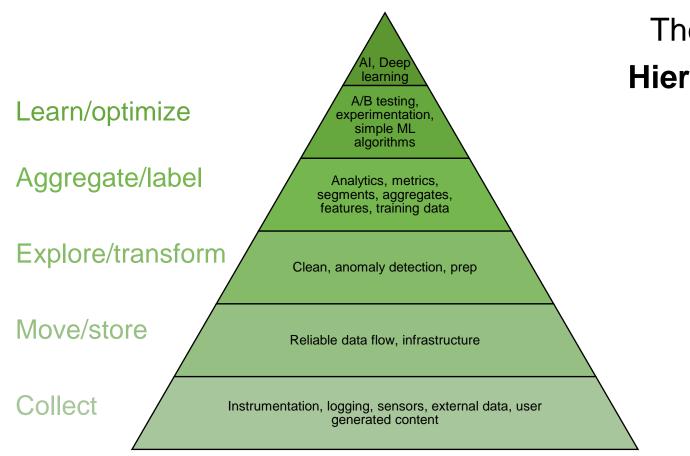
▸ About 1.7 million transactions daily

▸ 2018: Released new services

▸ Needed a way to aggregate billing transactions and have insights into the service

splunk> .conf18

Our goals

- Aggregate customer transaction statistics and ensure accuracy and reliability

- Give operational– and business insights about the new services

How to get to data science?

Building data trust

The data science

Hierarchy of needs

Learn/optimize

AI, Deep learning

A/B testing, experimentation, simple ML algorithms

Aggregate/label

Analytics, metrics, segments, aggregates, features, training data

Explore/transform

Clean, anomaly detection, prep

Move/store

Reliable data flow, infrastructure

Collect

Instrumentation, logging, sensors, external data, user generated content

Best practices for data analytics projects

Facilitate interdisciplinary collaboration

Create short learning loops

Enable others

Data analysis

Explore together

Make data producers become consumers

Get exposed to prod data early

Data engineering
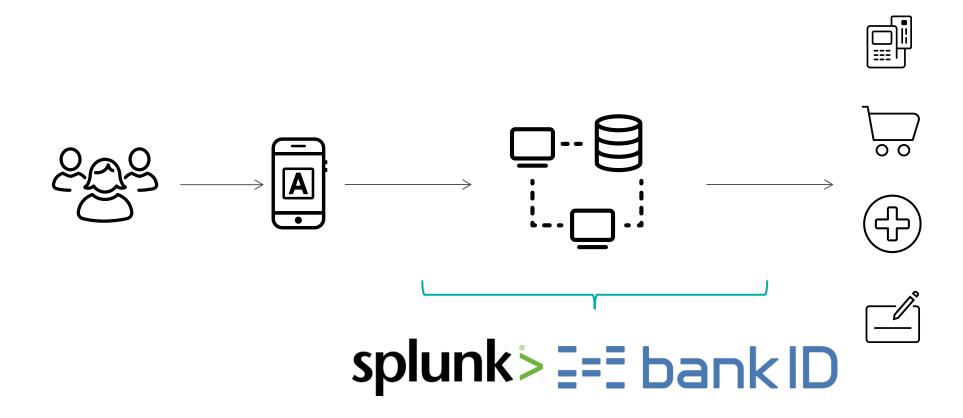
Be use-case driven

Get new data right

Data collection

Data Collection

Grab what you need, all the time

Where is our data?

Data collection – Data engineering – Data analysis

Be use-case driven

- Have a set goal for what you want to achieve
- Start with only the necessary data sources
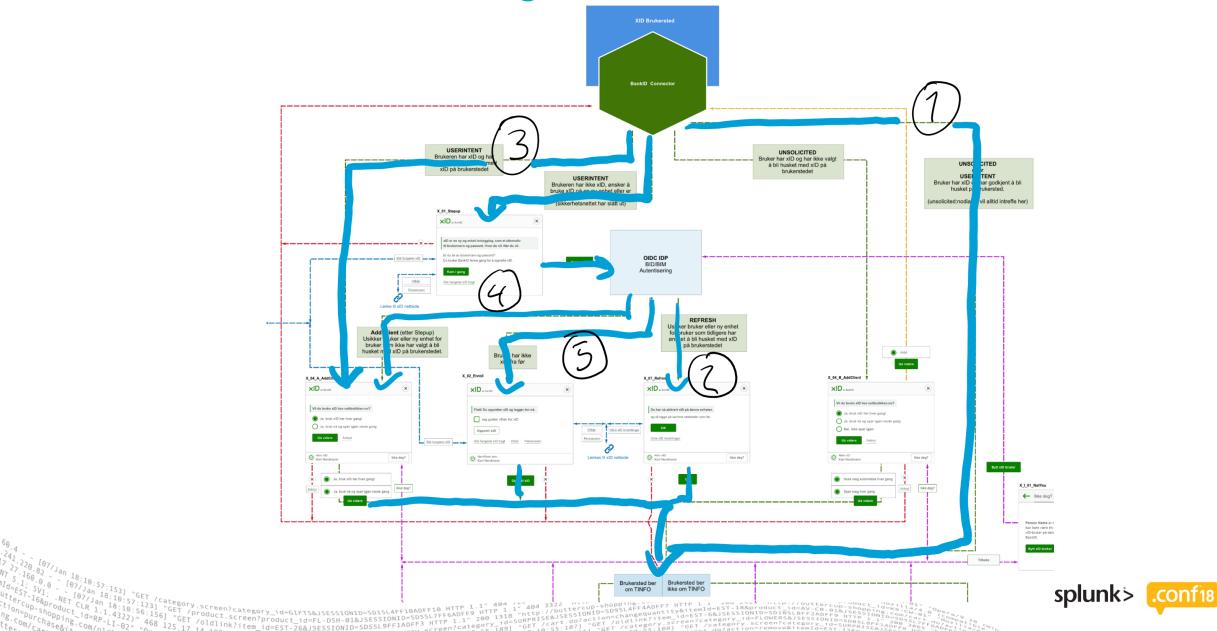
Getting new data right

- Guide developers (data producers)
- Requirements for creating logs
- Build data trust
- Normalize to your data model

Data Engineering

Creating data trust

splunk> .conf18

Starting with the basics

Building a transaction

1528721737672 app=XID-CLIENT sid=47cf08fd-fcd5-4c57-9ca1-75264e68eb85 cid=e208317d-e8d3-4e08-9114-6dd6313fb223 merchant=
xID_av_BankID_Norge-Nordea level=1 type=ACTION code=16432 value=null message="stepup continue"

1528721775287 app=XID-CLIENT sid=47cf08fd-fcd5-4c57-9ca1-75264e68eb85 cid=c06b5053-a474-47c3-a3d1-c81c2a966fca merchant=
xID_av_BankID_Norge-Nordea level=1 type=ACTION code=16416 value=null message="refresh onContinue"
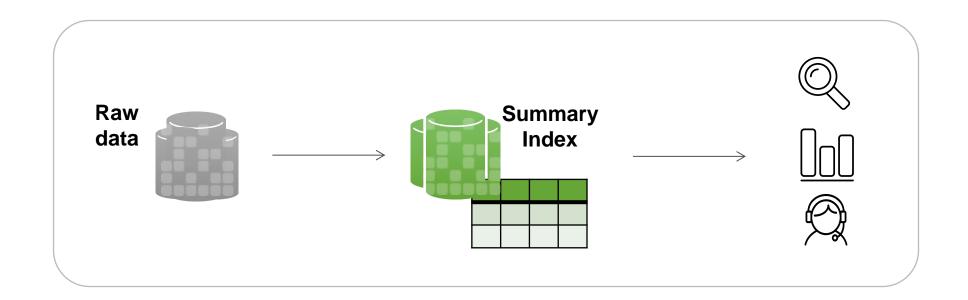
Categorizing customer journeys

i	Time	Event		_time ⇕	Duration ⇕	App ⇕	MerchantID ⇕	Context ⇕	ResponseStatus ⇕
⌄	27/08/2018 15:21:35.981	1535376095981 app=XID-CLIENT sid=da6c09ca-c0ac-4d46-a436-92d67a4ec2d9 cid=eeb41fd2-ceb2-43e8-baf5-f73d2ca271eb merchant=example_with_security_rights level=1 type=PERFORMANCE code=2036 value=66 message=REQUEST_LOAD_TIME_PERFORMANCE sid = da6c09ca-c0ac-4d46-a436-92d67a4ec2d9		2018-08-26 12:51:41.818	11.839	XID-CLIENT	xIDLPage	authenticate initSession logData setEndUserChoice	200 201 204 400
⌄	27/08/2018 15:21:35.981	1535376095981 app=XID-CLIENT sid=da6c09ca-c0ac-4d46-a436-92d67a4ec2d9 cid=eeb41fd2-ceb2-43e8-baf5-f73d2ca271eb merchant=example_with_security_rights level=1 type=PERFORMANCE code=2037 value=743 message=PAGE_LOAD_TIME_PERFORMANCE sid = da6c09ca-c0ac-4d46-a436-92d67a4ec2d9		2018-08-26 09:53:46.933	33.833	XID-CLIENT	xIDLPage	enrollEndUser enrollOrUpdateBrowser findEndUser	200 201 204
⌄	27/08/2018 15:21:35.980	1535376095980 app=XID-CLIENT sid=da6c09ca-c0ac-4d46-a436-92d67a4ec2d9 cid=eeb41fd2-ceb2-43e8-baf5-f73d2ca271eb merchant=example_with_security_rights level=1 type=PERFORMANCE code=2035 value=421 message=PAGE_RENDER_TIME_PERFORMANCE sid = da6c09ca-c0ac-4d46-a436-92d67a4ec2d9						initSession logData setEndUserChoice	400
⌄	27/08/2018 15...	1535376095931 app=XID-CLIENT sid=da6c09ca-c0ac-4d46-a436-92d67a4ec2d9 cid=eeb41fd2-ceb2-43e8-baf5-f73d2ca271eb merchant... level=1 type=PERFORMANCE code=2038 value=733 message=TOTAL_LOAD_TIME_PERFORMANCE ...67a4ec2d9		2018-08-26 18:06:29.102	1.792	XID-CLIENT	storebrand-xid-prod	authenticate initSession logData	200 201 400
⌄	2... 15...	...da6c09ca-c0ac-4d46-a436-92d67a4ec2d9 cid=eeb41fd2-ceb2-43e8-baf5-f73d2ca271eb mercha... level=1 type=LOG message="[Startup][useragent=Mozilla/5.0 (Macintosh; Intel Mac OS X ...ML, like Gecko) Chrome/68.0.3440.106 Safari/537.36][cid=eeb41fd2-ceb2-43e8-baf5-f73d2 ca271eb][sid=da6c09ca-c0ac-4d46-a436-92d67a4ec2d9][locale=nb][mode=authenticate][coiUrl=https://xid-preprod.bankidapi		2018-0...		...IENT	xIDLPage	enrollEndUser enrollOrUpdateBrowser findEndUser initSession logData	200 201 204 400

Group events with same ID

Collect fields specifying user actions

_time ⇕	Duration ⇕	App ⇕	MerchantID ⇕	CustomerJourneyID ⇕		Duration ⇕	MerchantID ⇕	CustomerJourneyID ⇕	Invoice ⇕	ReadTerms ⇕	Summary ⇕
2018-08-26 12:51:41.818	11.839	XID-CLIENT	xIDLPage	465489b36f365948bff132d962956595	2:51:41.818	11.839	xIDLPage	465489b36f365948bff132d962956595	Yes	No	User logs in
2018-08-26 09:53:46.933	33.833	XID-CLIENT	xIDLPage	35918232956d6b9d2f1da7e8a6805119	9:53:46.933	33.833	xIDLPage	35918232956d6b9d2f1da7e8a6805119	Yes	Yes	User enrolls
2018-08-26 18:06:29.102	1.792	XID-CLIENT	storebrand-xid-prod	1349c38816e653476155a1951550fda3	8:06:29.102	1.792	storebrand-xid-prod	1349c38816e653476155a1951550fda3	Yes	No	User auto-logins
2018-08-26 12:53:09.617	69.571	XID-CLIENT	xIDLPage	7509d2969d594c6c3c97808e7399a182	2:53:09.617	69.571	xIDLPage	7509d2969d594c6c3c97808e7399a182	No	No	User experiences error
2018-08-26 14:34:19.957	12.617	XID-CLIENT	storebrand-xid-prod	35918232956d6b9d2f1da7e8a6805119	4:34:19.957	12.617	storebrand-xid-prod	35918232956d6b9d2f1da7e8a6805119	Yes	Yes	User enrolls
2018-0...		D-CLIENT	storebrand-xid-prod	1349c38816e653476155a1951550fda3	9:11:28...			...49c38816e653476155a1951550fda3	Yes	No	User auto-logins
2018-0...		D-CLIENT	storebrand-xid-prod	465489b36f365948bff132d962956595	5:40:36...			...5489b36f365948bff132d962956595	Yes	No	User logs in
2018-0...		D-CLIENT	storebrand-xid-prod	465489b36f365948bff132d962956595	4:33:17...			...5489b36f365948bff132d962956595	Yes	No	User logs in
2018-08-26 11:09:47.540	7.040	XID-CLIENT	storebrand-xid-prod	1349c38816e653476155a1951550fda3	1:09:47.540	7.040	storebrand-xid-prod	1349c38816e653476155a1951550fda3	Yes	No	User auto-logins

Create customer journey hash ID

Use a lookup to add information

Collecting customer journeys

▸ Each night a scheduled search runs which gathers user transactions, categorizes them and collects them into a summary index

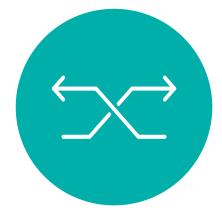▸ Information extraction using the SI is faster and more uniform

Raw data → **Summary Index**

splunk> .conf18

Data collection – Data engineering – Data analysis

Explore together

▪ Use commands such as correlate, transaction and cluster

▪ Balance reverse engineering and business domain knowledge

Make data producers become consumers

▪ Have developers actively use visualizations/alerts

▪ Be inviting and prioritize time to guide others

▪ Avoid getting dirty data

Get exposed to prod data early

▪ Mind unexpected scenarios

▪ Don't overcomplicate searches

▪ Make sure the developers knows the necessity of a concise log format

splunk> .conf18
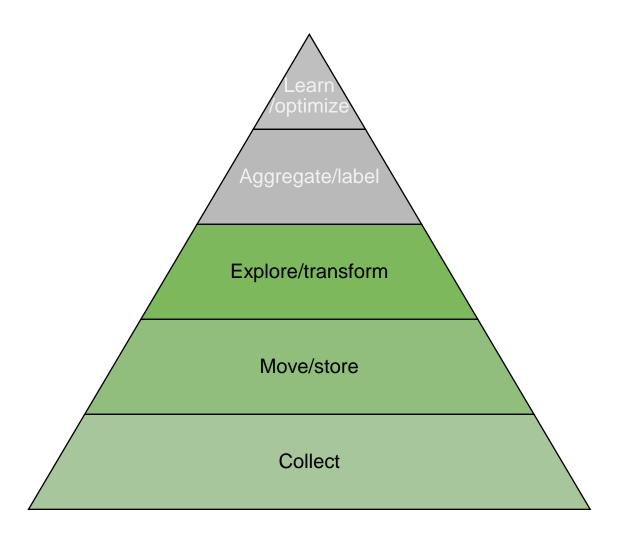
Data Analysis

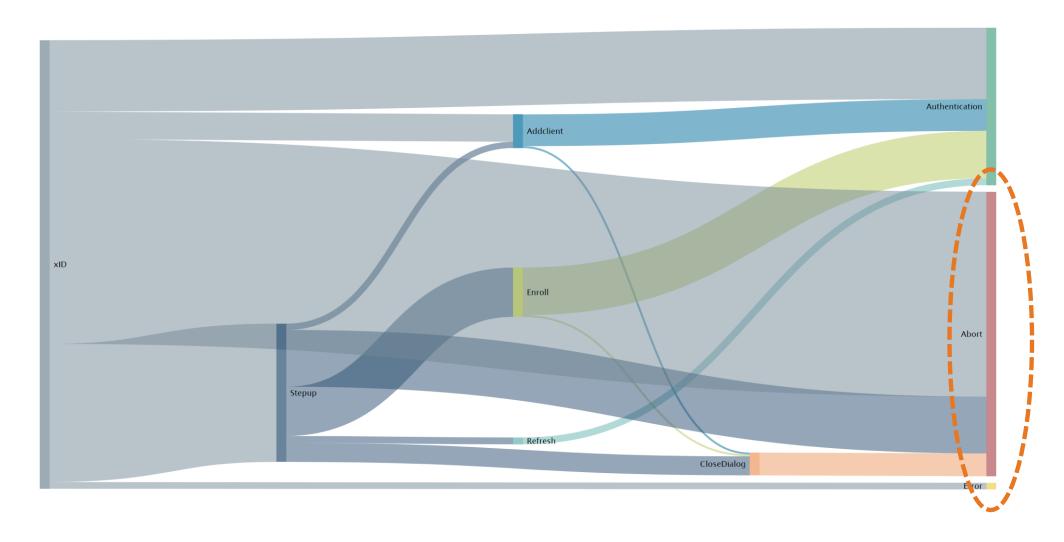**Collaboration makes better**

splunk> .conf18

Extracting insights

▶ Data analysis is about reaching the top of the pyramid

▶ Using the scientific approach

- Launching hypotheses

- Qualitative study with interviews and user observations
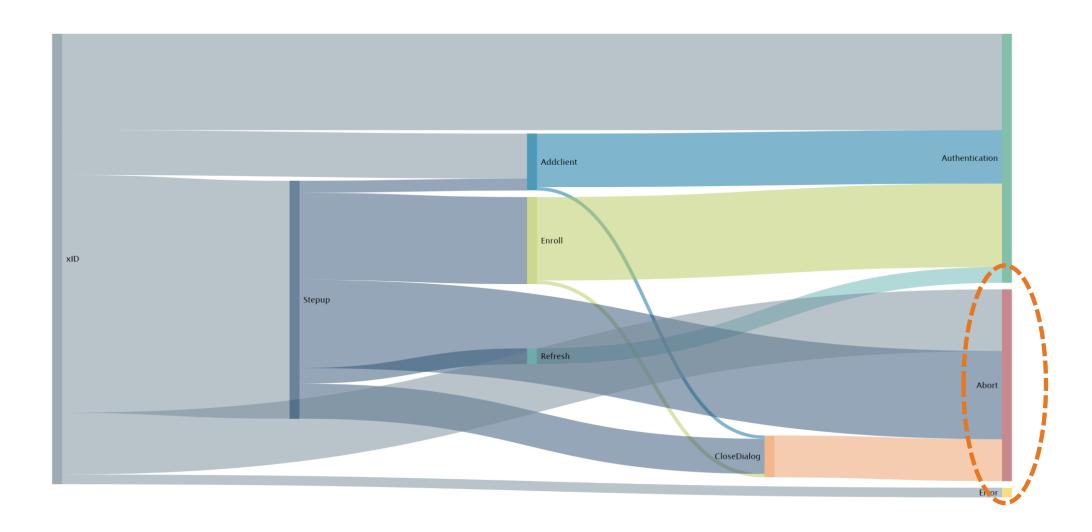
- Quantitative research using the data set

Learn /optimize

Aggregate/label

Explore/transform

Move/store

Collect

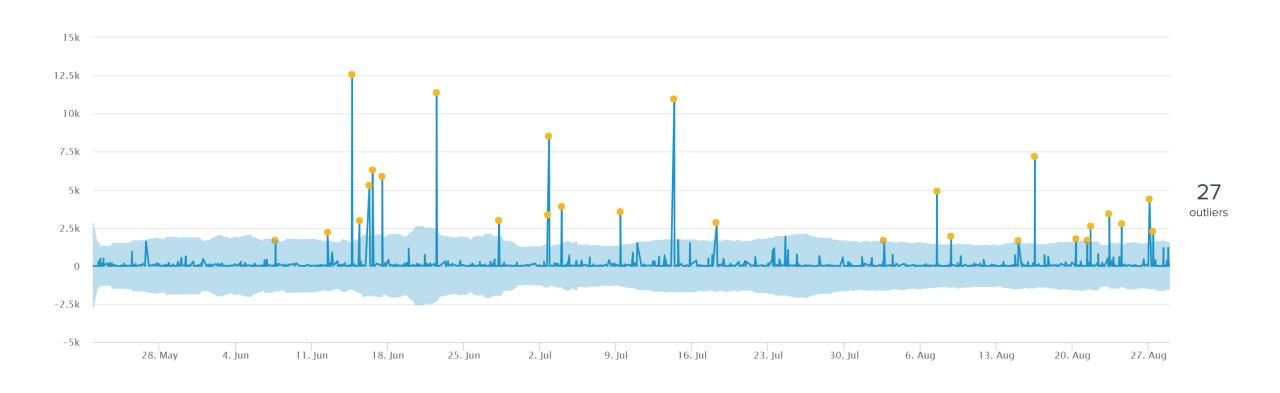# Customer journey analysis

Customer journey analysis

Customer journey analysis

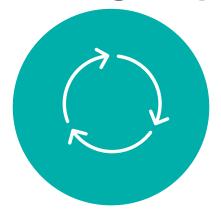## Finding outliers in transaction time usage

Data collection – Data engineering – Data analysis

Facilitate interdisciplinary collaboration

- Workshops with developers
- Present dashboards for HR and Sales

Create short learning loops

- Minimum viable product approach
- Include feedback as a topic in weekly meetings
- Information in the office space
- Scientific approach

Enable others

- Make people from core business capable of doing analysis on their own data
- Splunk Fundamentals training

splunk> .conf18

Closing Thoughts

1. Data science is not magic

2. For everyone and every organization with data

3. Keep the logic simple

4. A simple and flexible tool still needs to be managed properly

splunk> .conf18

Q&A

Thomas Hillesøy | Solution Archtect

Jørn Kristian Stensby | Solution Archtect

Erik Larsen | Head of Operations

splunk> .conf18

Thank You

**Don't forget to rate this session
in the .conf18 mobile app**

Supporting material

1. Articles

- https://hbr.org/2013/12/datas-credibility-problem
- https://hbr.org/2013/01/why-it-fumbles-analytics
- https://hbr.org/2018/03/what-happens-when-data-scientists-and-designers-work-together
- https://hbr.org/2014/03/what-makes-big-data-projects-succeed

2. SPLs

- https://github.com/thilles/splunkconf2018

splunk> .conf18