A Company's Digital Twin

TUM Data Innovation Lab

Team:
Sagarika Kathuria, Pooreumoe Kim,
Frederik Wenkel, Jieyi Zhang

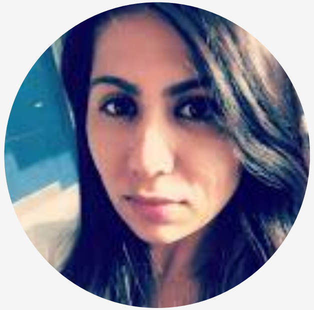
Mentors:
Simon Brand, Anton Kurz, Sebastian Rossner

Co-Mentor:
Laure Vuaille

Project Leader:
Dr. Ricardo Acevedo Cabra

Supervisor:
Prof. Dr. Massimo Fornasier

Our Team



Sagarika Kathuria
Mathematics in Data Science

Pooreumoe Kim
Data Engineering and Analytics

Frederik Wenkel
Mathematics

Jieyi Zhang
Data Engineering and Analytics

AGENDA

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

AGENDA

01 Motivation

02 Digital Twin Concept

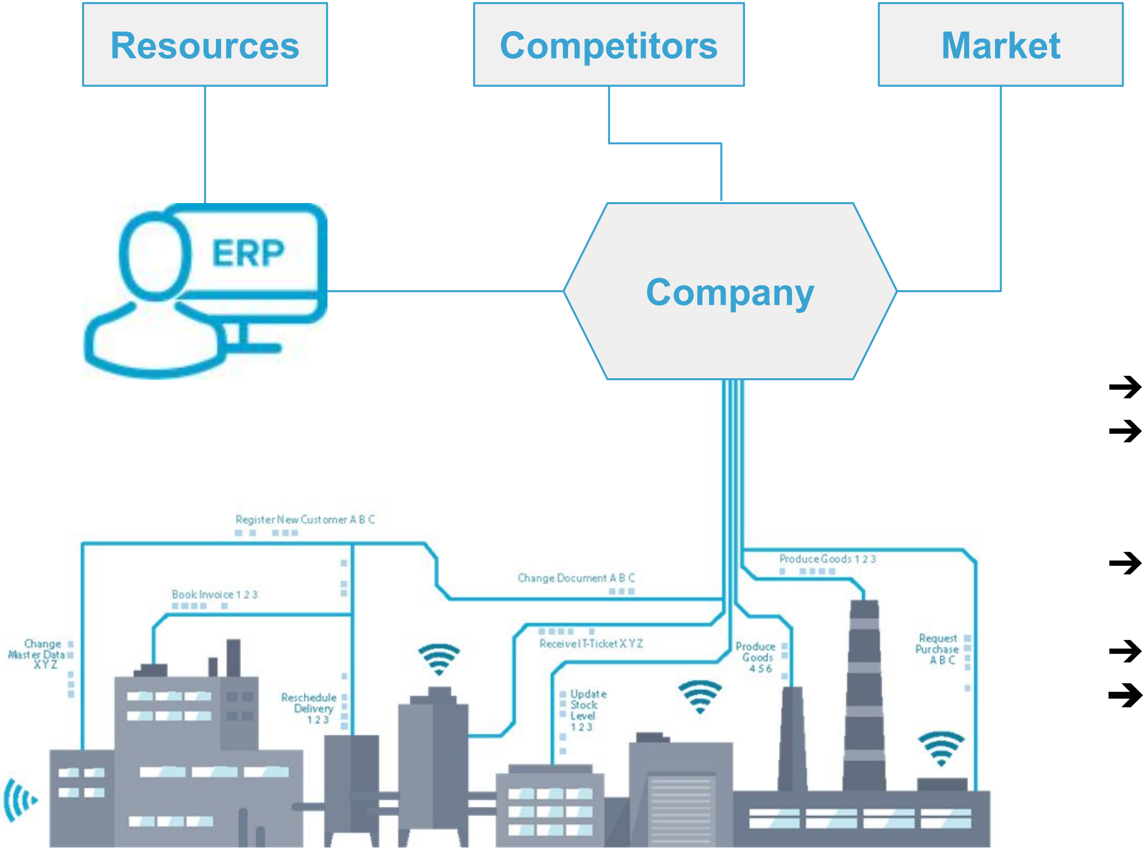
03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

Company as a Complex System



➔ Company's future performance?
➔ **Forecasting systems**

➔ Sales forecasting, predictions for volatile resources
➔ "Patchwork" of forecasts
➔ **No complete picture**

AGENDA

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

Digital Twin Concept

Digital Twin Concept

Major characteristics

- ➔ Similar behavior to original organization
- ➔ **Family resemblance**

- ➔ No exact copy
- ➔ Digital Twin can be **raised differently**

AGENDA

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

Process Mining at Celonis

celonis

- ➔ Every process leaves a digital footprint in the company's environment
- ➔ Celonis Process Mining enables an organization to structure its raw data of a process into a **data model**
- ➔ The main object : Activity Table
  - ➔ Consists of a number of activities represented by **cases**
  - ➔ Each case has its own **case key** to distinguish different cases
  - ➔ Each Activity has an **event time**,a **sorting** and a **category**

_CASE_KEY	ACTIVITY_EN	EVENTTIME	CATEGORY_NAME	_SORTING
201807010	Status: New	2018.7.2 3:08	Category 1	10
201807010	Status: Open	2018.7.2 3:14	Category 1	10
201807010	Change assignee	2018.7.2 3:14	Category 1	10
201807010	Status: On Hold	2018.7.2 3:49	Category 1	10
201807010	Status: Open	2018.7.2 6:00	Category 1	10
201807010	Status: Closed	2018.7.2 12:06	Category 1	10
201807011	Create Ticket	2018.7.1 18:21	Category 2	10
201807011	Change priority	2018.7.1 18:21	Category 2	10
201807011	Status: Closed	2018.7.1 21:39	Category 2	10

ACTIVITY TABLE

celonis

➔ **Analysis** in Celonis Data Mining:

- ➔ Helps visualize and understand the data using tools like Process Query Language(PQL) and Key Performance Indicators (KPI’s)
- ➔ Helps identify bottlenecks and inefficiencies within the process
- ➔ Consists of various objects like: Process Explorers, Variants Explorer, Standard KPI's
- ➔ Customizable with tables, charts, diagrams and custom KPI’s

ANALYSIS

AGENDA

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

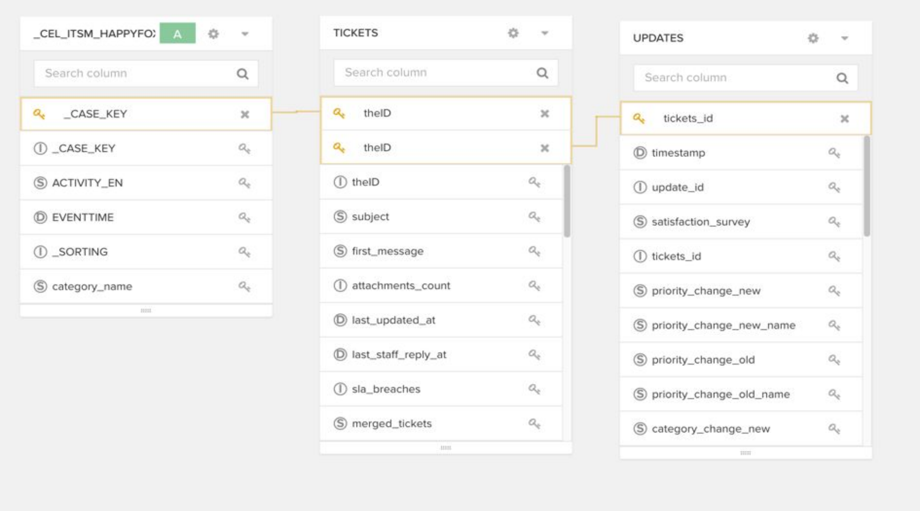
05 Specific Use Case at Celonis

06 Summary & Outlook

Explorative Data Analysis

celonis

- ➔ Input Dataset : Celonis Happyfox IT Service Management
- ➔ Happyfox - a Saas Platform, offers help desk ticketing system for approx. 12000 companies.
- ➔ Celonis Happyfox ITSM Data Model:
  - ➔ Three data tables which contains the information for each ticket - _CEL_ITSM_happyfox_ACTIVITIES, Tickets, Updates
  - ➔ Data stored in the form of string, boolean, number and date time values
- ➔ Data accessed using PQL (Process Query Language) in Celonis and Python API.



Explorative Data Analysis

➔ ITSM Ticket Processing

➔ Each ticket - has its own process - each process type called a Variant

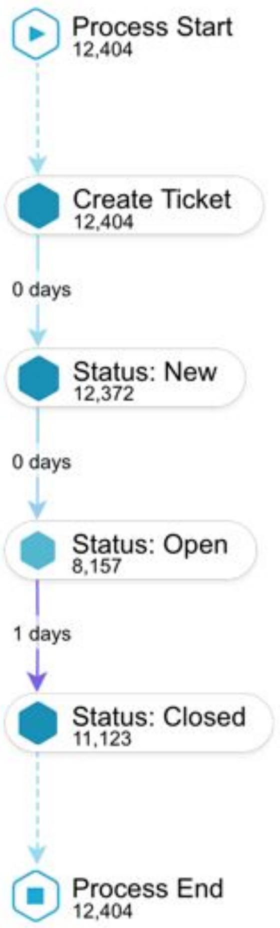
➔ Happy Path - variant that happens most frequently

➔ Throughput time -

$$T_{tp} = ts_{act} - ts_{act\_previous}$$

$$T_{total\_tp} = ts_{last\_act} - ts_{first\_act}$$

➔ Total Variants : 1850 , 20% cases follow Happy Path

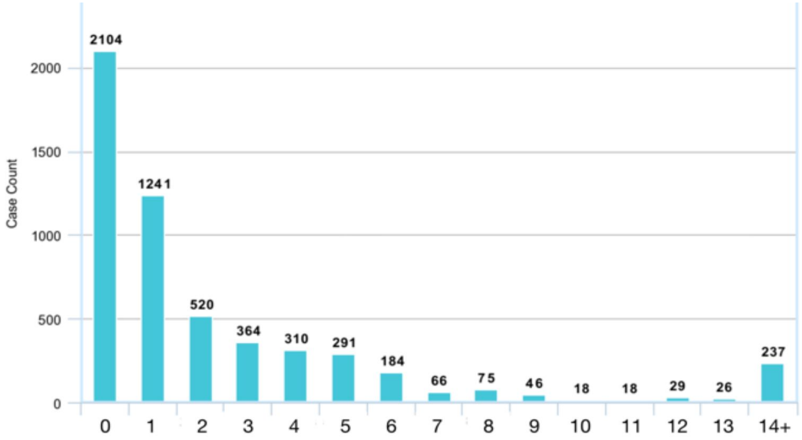


celonis

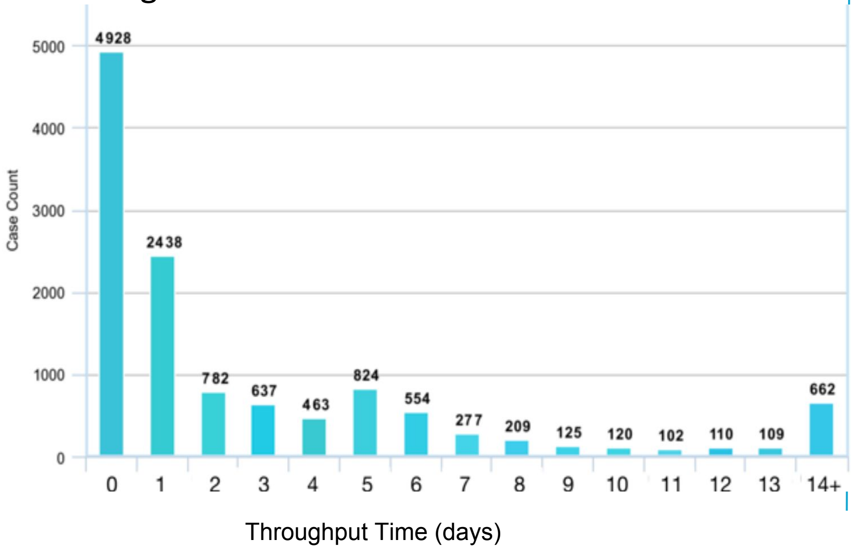
Explorative Data Analysis

- ➔ Throughput Time Analysis
  - ➔ 2 cases - tickets created in 2017 & 2018
  - ➔ Median - close to 0 days
    Max Throughput Time - 276 days
  - ➔ Digital Twin model considers outliers as well

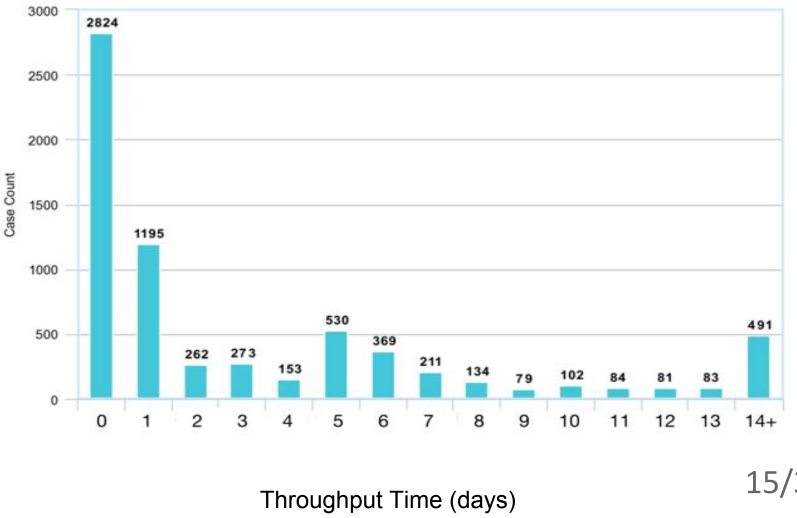
2017



17/18



2018



Explorative Data Analysis

celonis

Throughput Time Analysis

- ➔ Kolmogorov–Smirnov(KS) Test
  - ➔ Tests the distance between the empirical distribution function of the data and the cumulative distribution function (CDF) of the reference distribution.
  - ➔ H_0 = Two distributions follow the same distribution
  - ➔ H_0 holds if p-value > 0.05
- ➔ Need non-parametric methods to simulate the Data such as Bins Method, Kernel Density Estimation

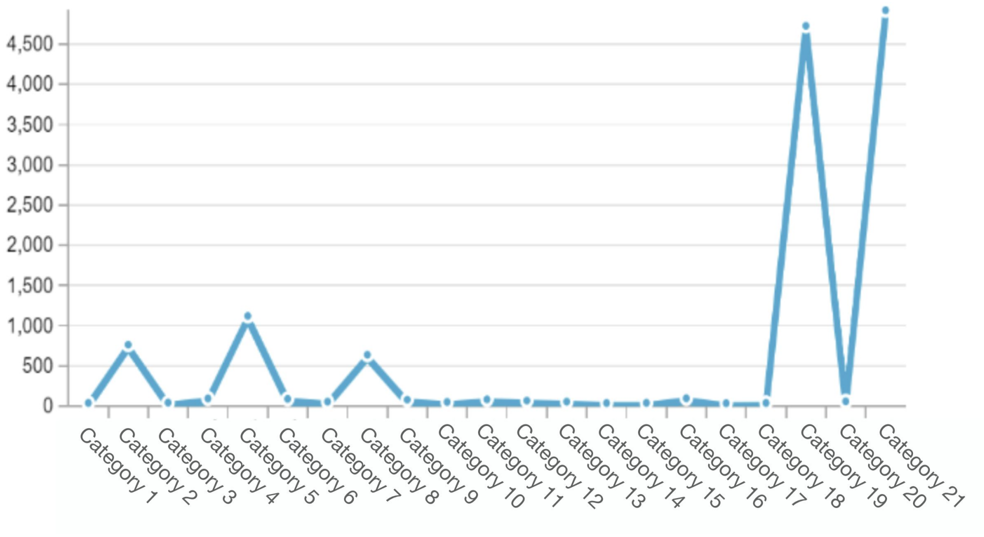
Distribution	p-value	Is passed
Poisson Distribution	4.675e-10	No
Exponential Distribution	0.0	No
Birnbaum-Saunders	0.0	No

KS Test Results

celonis

Explorative Data Analysis

- ➔ Category Based Analysis
  - ➔ Total Categories - 21
  - ➔ Categories of Importance - 5
  - ➔ Digital Twin simulates categories in cases based on their percentages

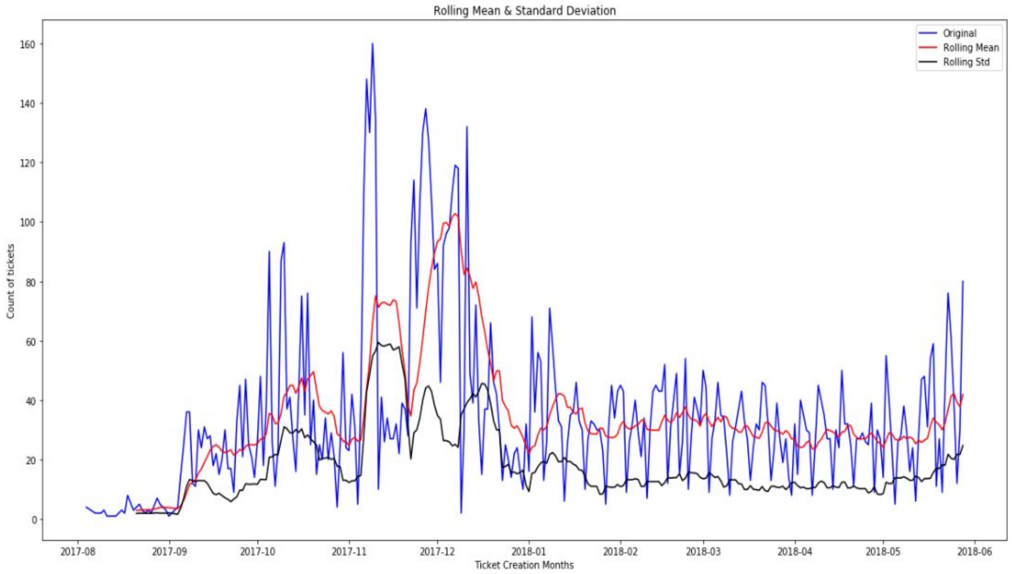


Count of Tickets based on Categories

Explorative Data Analysis

- ➔ Ticket Number Analysis
  - ➔ Total Tickets - 12000
  - ➔ Dickey Fuller Test - Statistical test for checking stationarity.

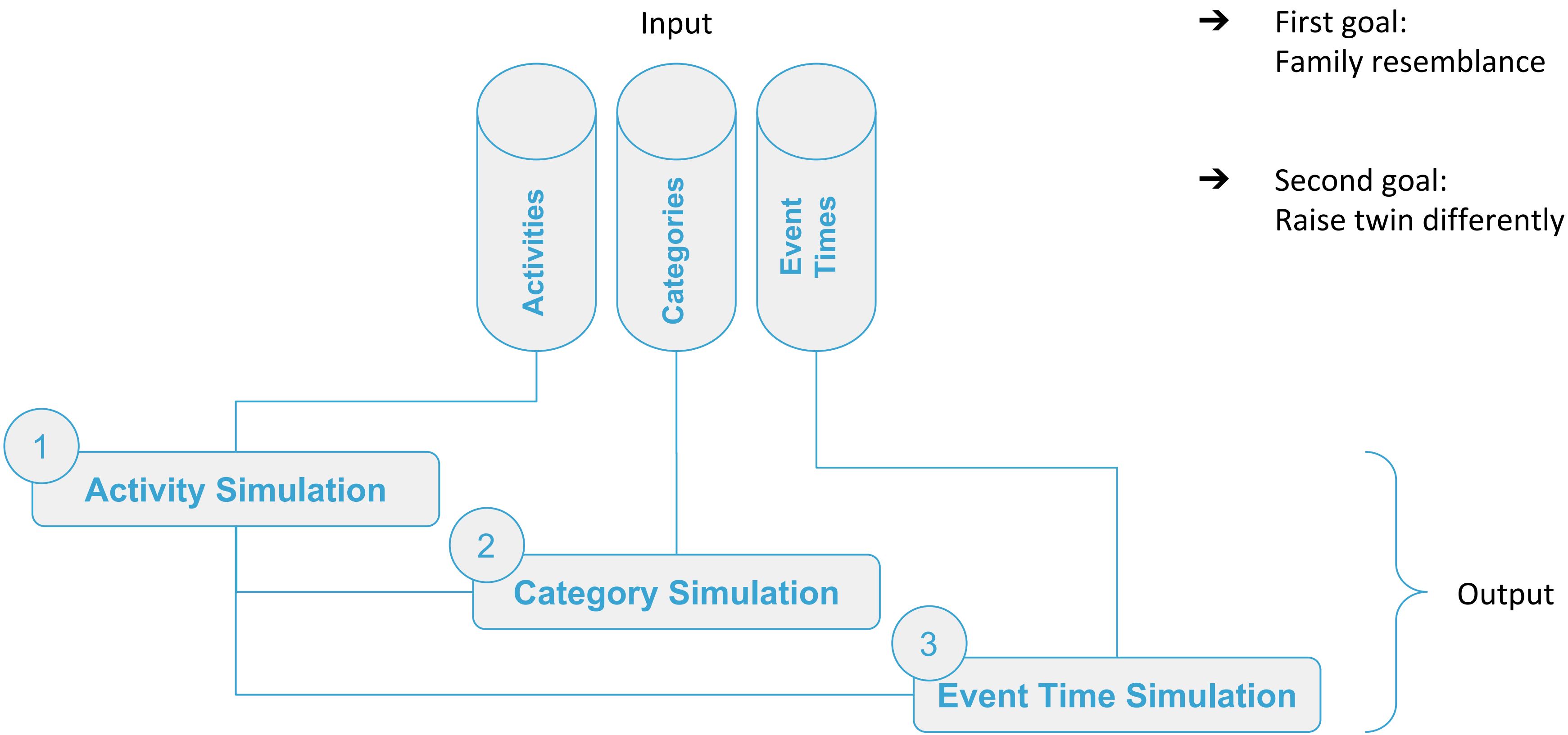
    H_0 - Series is non-stationary
    p-value : 0.091
    Cannot reject H_0 as p-value > 0.05
  - ➔ Use Transformations like log
  - ➔ Estimate trend and seasonality to predict number of tickets using Time Series Methods




```
Results of Dickey-Fuller Test:
Test Statistic                  -2.608286
p-value                          0.091297
#Lags Used                      13.000000
Number of Observations Used    274.000000
Critical Value (1%)             -3.454444
Critical Value (5%)             -2.872147
Critical Value (10%)            -2.572422
dtype: float64
```

Count of Tickets Based on Creation Date

General Simulation Approach



➔ First goal:
Family resemblance

➔ Second goal:
Raise twin differently

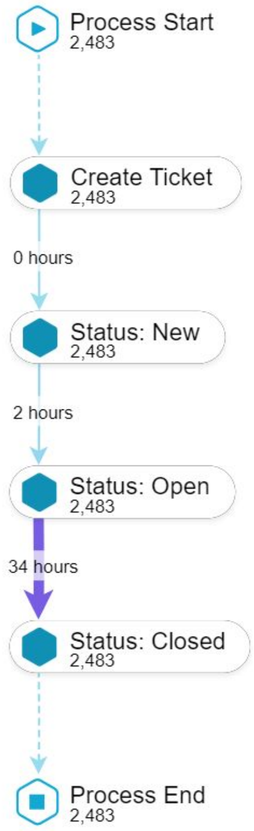
Activity Simulation

- ➔ First building block of simulation
- ➔ Importance for subsequent blocks

- ➔ Measure of precision: Occurrences of most frequent activity flows from input data in output

- ➔ According to empirical observations from input
- ➔ **Markov Process**

- ➔ Manageable matrix representation
- ➔ **No dependencies captured** on activity history



Activity Simulation

➔ **Preprocessing** yields improvement

➔ Treat activity flows with different starting activities separately

➔ **Linear Additive Markov Process (LAMP)** instead of ordinary Markov Process

$$P\left(x_{n+1} | x_0, \ldots, x_n\right) = \sum_{i=1}^{M} w_i \cdot P\left(x_{\max\{n-i+1,0\}}, x_{n+1}\right)$$

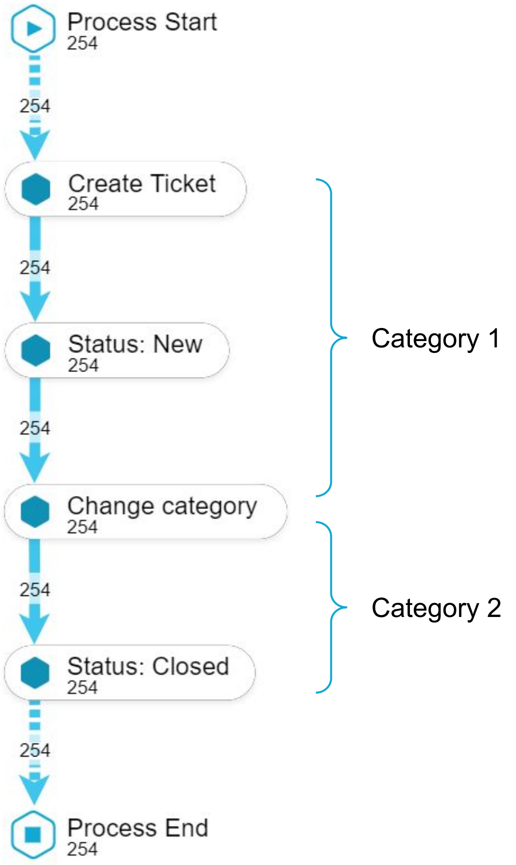
➔ Parameters w, P have to be learned minimizing the following negative log likelihood

$$L\left(w, P, c_1, \ldots, c_B\right) = -\sum_{k=1}^{B}\left[\sum_{l=1}^{L_k} \log\left(\sum_{i=1}^{M} w_i \cdot P\left(x^k_{\max\{l-i,0\}}, x^k_l\right)\right)\right]$$

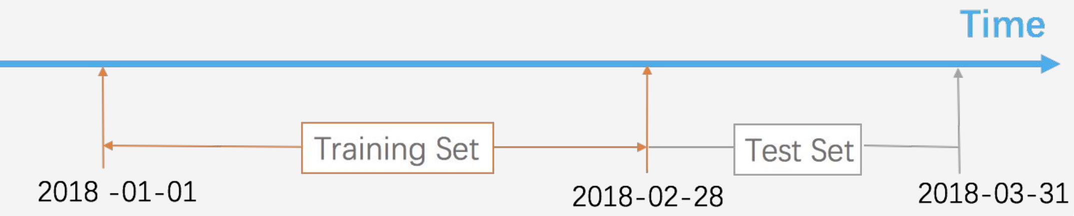
Category Simulation

➔ Category assignment to every activity flow from activity simulation

➔ **Starting category** assignment
➔ **Markov Process** for category changes

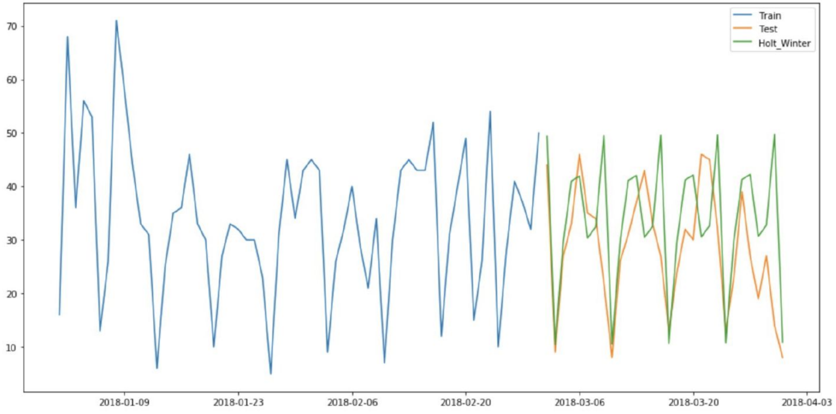


The number of Tickets Prediction



- Train time series models from training set
  - Data Separation
  - ARIMA, Holt Winter's ...

- Compare the prediction result to test set
  - Measure errors



Holt Winter - Creation Date Vs Count of Tickets

The number of Tickets Prediction

Model	RMSE Value
Naive Approach	34.51
Simple Average	19.18
Moving Average	20.02
Simple Exponential Smoothing	31.08
Double Exponential Smoothing	20.59
Holt Winter's / Triple Exponential Smoothing	12.15
ARIMA	22.56

➔ **Choose Holt Winter's Model**

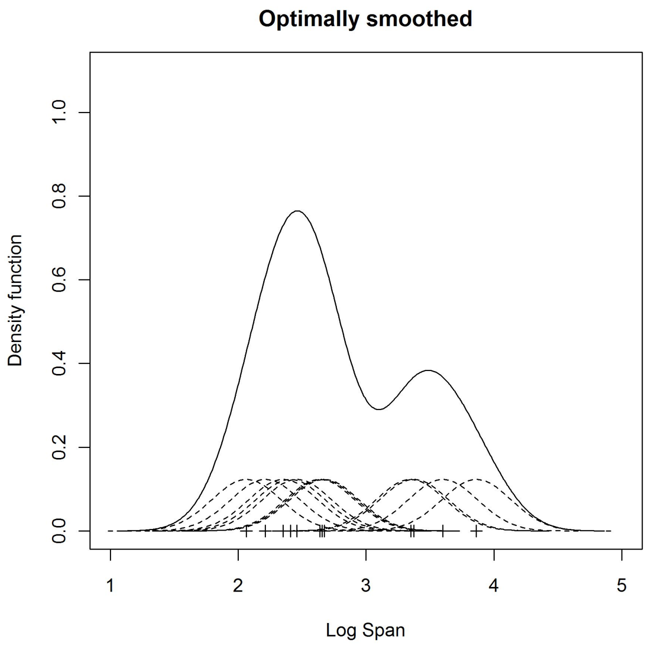
Throughput Time Simulation: KDE

- Kernel Density Estimation(KDE):
  Non-parametric way to estimate the probability density function(PDF) of a random variable.

- $$\hat{f}(x) = \frac{1}{nh} \Sigma_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

  h : Bandwidth
  K(.) : Kernel function

- Two important parameters



Throughput Time Simulation: KS-Test

- Kolmogorow-Smirnow-Test (KS-Test)

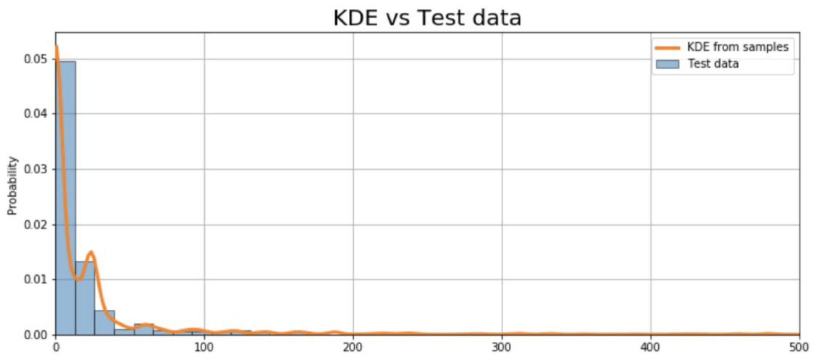
- $H_0: F_X(x) = F_Y(x)$

  $H_1: F_X(x) \neq F_Y(x)$

- p-value > 0.05:

  Two sets have same distribution

  The higher, the more identical



Throughput time
From <Status: New> to <Change assignee>

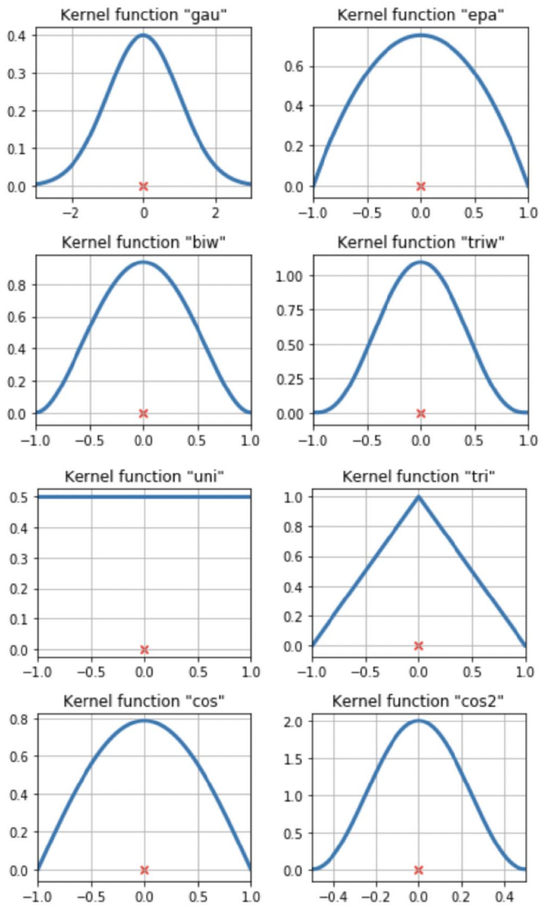
Throughput Time Simulation: Kernel Selection

- Simulating power(KS-Test)?
  - Identical
  - Table 7 (Doc)

- Negative value generation?
  - Except Gaussian, no negative if bandwidth <= 0.1
  - Table 8 (Doc)

- Internal sampling method?
  - Gaussian & Uniform

➔ **Choose Uniform Kernel**



Throughput Time Simulation: Bandwidth Selection

- Three methods for bandwidth:
  - Constant bandwidth (0.1)
  - Gridsearch
  - Gridsearch with Cross validation

- Analysis of variance(ANOVA)

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \mu_i \neq \mu_j$$

F-value	p−value
0.839	0.362

ANOVA result: cannot reject H_0

Bandwidth	Running Times(1,000 rounds)
0.1 (Constant)	4.0061
Gridsearch	12.5433
Cross-Validation	21.5929

Comparison of running times

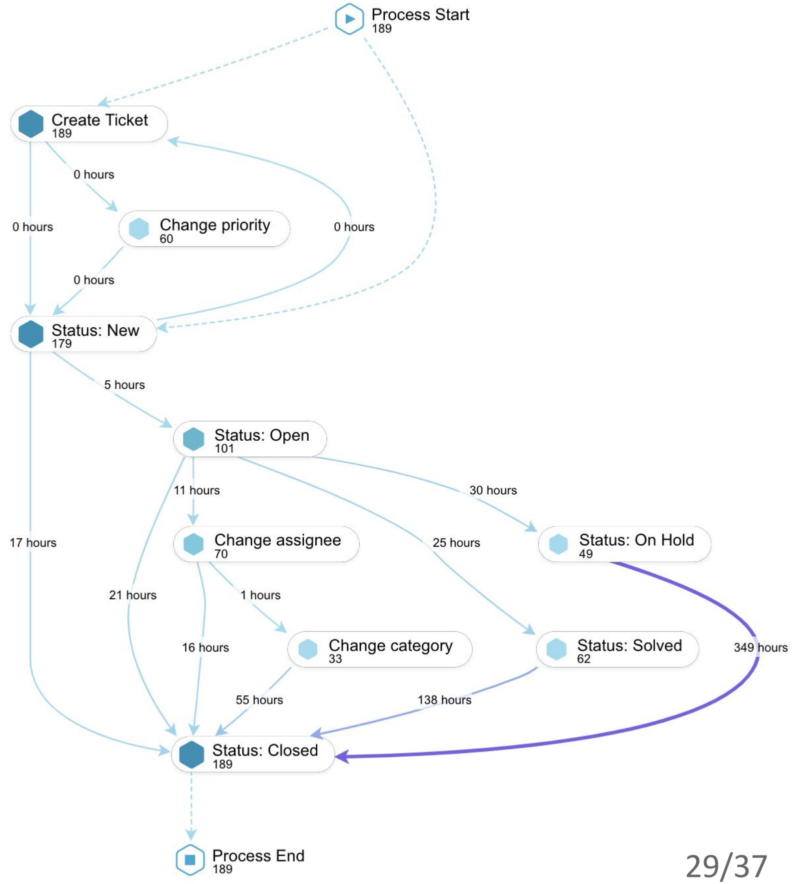
➔ **Choose Constant Bandwidth(0.1)**

Digital Twin: Integration of the Methodologies

- ❑ Activity Simulation
- ❑ Category Simulation
- ❑ Number of Cases
- ❑ Throughput Time

➔ **Generates Virtual Activity Table**

_CASE_KEY	ACTIVITY_EN	EVENTTIME	CATEGORY_NAME	_SORTING
201807010	Status: New	2018.7.2 3:08	Category 1	10
201807010	Status: Open	2018.7.2 3:14	Category 1	10
201807010	Change assignee	2018.7.2 3:14	Category 1	10
201807010	Status: On Hold	2018.7.2 3:49	Category 1	10
201807010	Status: Open	2018.7.2 6:00	Category 1	10
201807010	Status: Closed	2018.7.2 12:06	Category 1	10

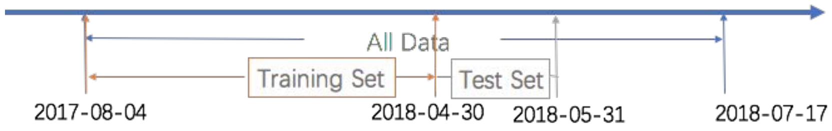


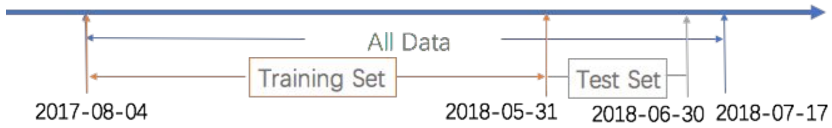
Model Validation

- Training Data

Model	Markov Process	KDE (Bandwidth=0.1)	HoltWinter (Period = 7 days)
M1	2 months	2 months	All
M2	2 months	2 months	4 months
M3	1 month	1 month	4 months

- Cross Validation





Model Validation

- Simulated Values Validation (Prediction for May 2018)

	Cases per day	Events per Day	Avg Total Throughput time	Trimmed Avg Total Throughput Time	Sample Size
rel. Error M1	31.25%	28.57%	15.52%	16.05%	**21.45%**
rel. Error M2	25.00%	32.38%	**11.20%**	**11.11%**	23.55%
rel. Error M3 ☀	**18.75%**	**20.95%**	18.10%	14.81%	25.45%

- Simulated Values Validation (Prediction for June 2018)

	Cases per day	Events per Day	Avg Total Throughput time	Trimmed Avg Total Throughput Time	Sample Size
rel. Error M1	45.16%	36.99%	43.53%	33.33%	**13.40%**
rel. Error M2	38.71%	30.64%	42.35%	31.82%	14.87%
rel. Error M3 ☀	**36.67%**	**27.75%**	**28.24%**	**18.18%**	14.87%

celonis

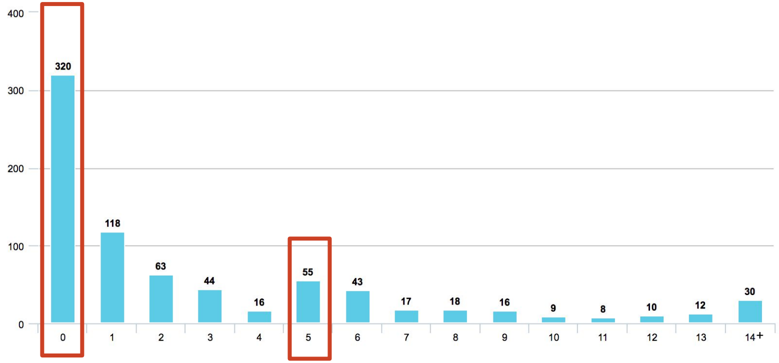
Model Validation

- Model 3

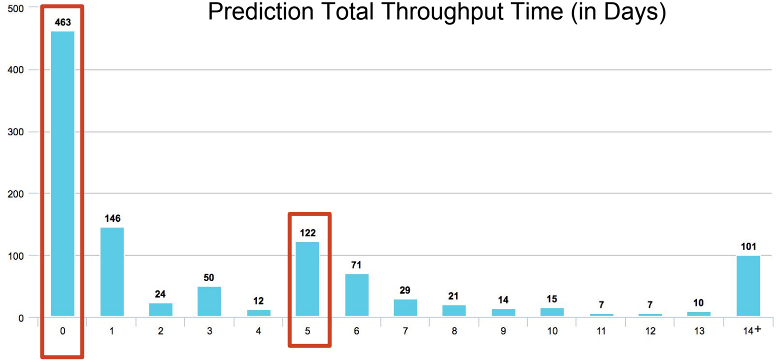
Activity Frequency Validation (2018-05)

Activity	Frequency Real Data	Frequency Prediction	Rel. Error In %
Create Ticket	**21.34%**	**21.36%**	0.09
Status: New	**16.46%**	**17.95%**	9.05
Change assignee	**12.20%**	**9.40%**	22.95
Status: Open	**11.59%**	**12.82%**	10.61
Status: Closed	**10.37%**	**9.40%**	9.35
Change Category	9,76%	**9.40%**	3.69
Status: Solved	7.93%	6.84%	13.75
Status: On Hold	5.49%	6.84%	24.59
Change Priority	4.89%	6.84%	39.87

Total Throughput Time Distribution (2018-05)



Prediction Total Throughput Time (in Days)



Real Data Total Throughput Time (in Days)

celonis

AGENDA

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

First Level Service Automation

What if I buy a chatbot and automate the first-level service?

How does this affect my throughput time?
How much people could I reallocate?
...

Reduce the throughput time of steps which belong to first level service to zero and simulate the cases.

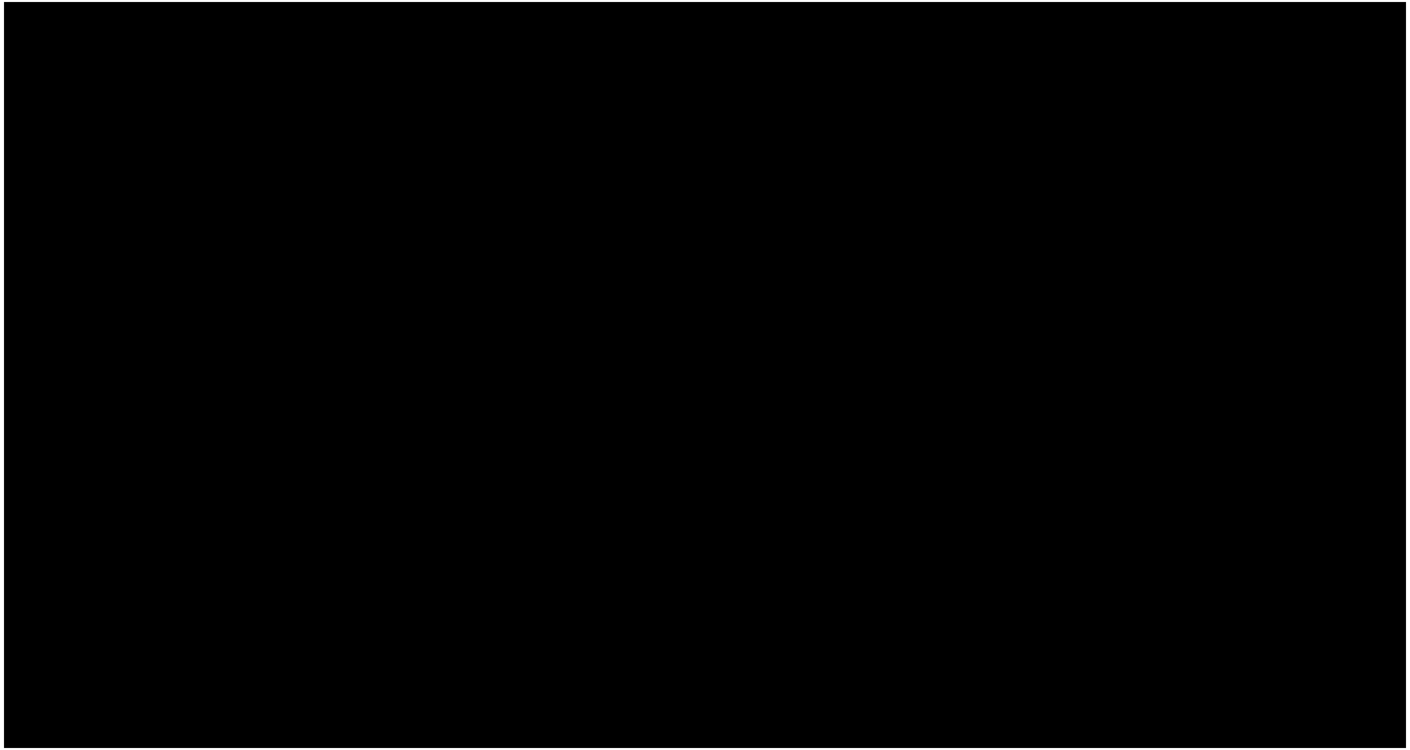
Simulated Value for May 2018

	Cases per day	Events per Day	Avg Total Throughput time	Trimmed Avg Total Throughput Time	Sample Size
Percentage of Change	-6.25%	-4.76%	-68.10%	-78.75%	-24.09%

First Level Service Automation

Real World **Automated**

- More tickets solved within a short time
- Average throughput time overall reduced
- Same Happy Paths



Tickets Created in May 2018 (Celonis Process Explorer)

AGENDA

celonis

01 Motivation

02 Digital Twin Concept

03 Process Mining at Celonis

04 Methodology

05 Specific Use Case at Celonis

06 Summary & Outlook

Summary and Outlook

Summary

- Process mining and Digital Twin Model
- Happyfox ITSM Data Model
- Data exploration
- Introduction of the techniques applied
- Model training and validation
- Application of Digital Twin

Future Works

- More sophisticated model
- More what-if questions
    - Number of tickets change
    - 24/7 customer support
- More input data
- ...

Thank you!