Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

Knowledge Discovery and Data Mining 1

(Data Mining Algorithms 1)

Wintersemester 2019/20

Agenda

1. Introduction
2. Basics
3. Supervised Methods
4. Unsupervised Methods
5. Process Mining
   - 5.1 Introduction
   - 5.2 Process Models – An Overview
   - 5.3 Process Discovery
   - 5.4 Conformance Checking
   - 5.5 Additional Mining Tasks
   - 5.6 Streams

Agenda

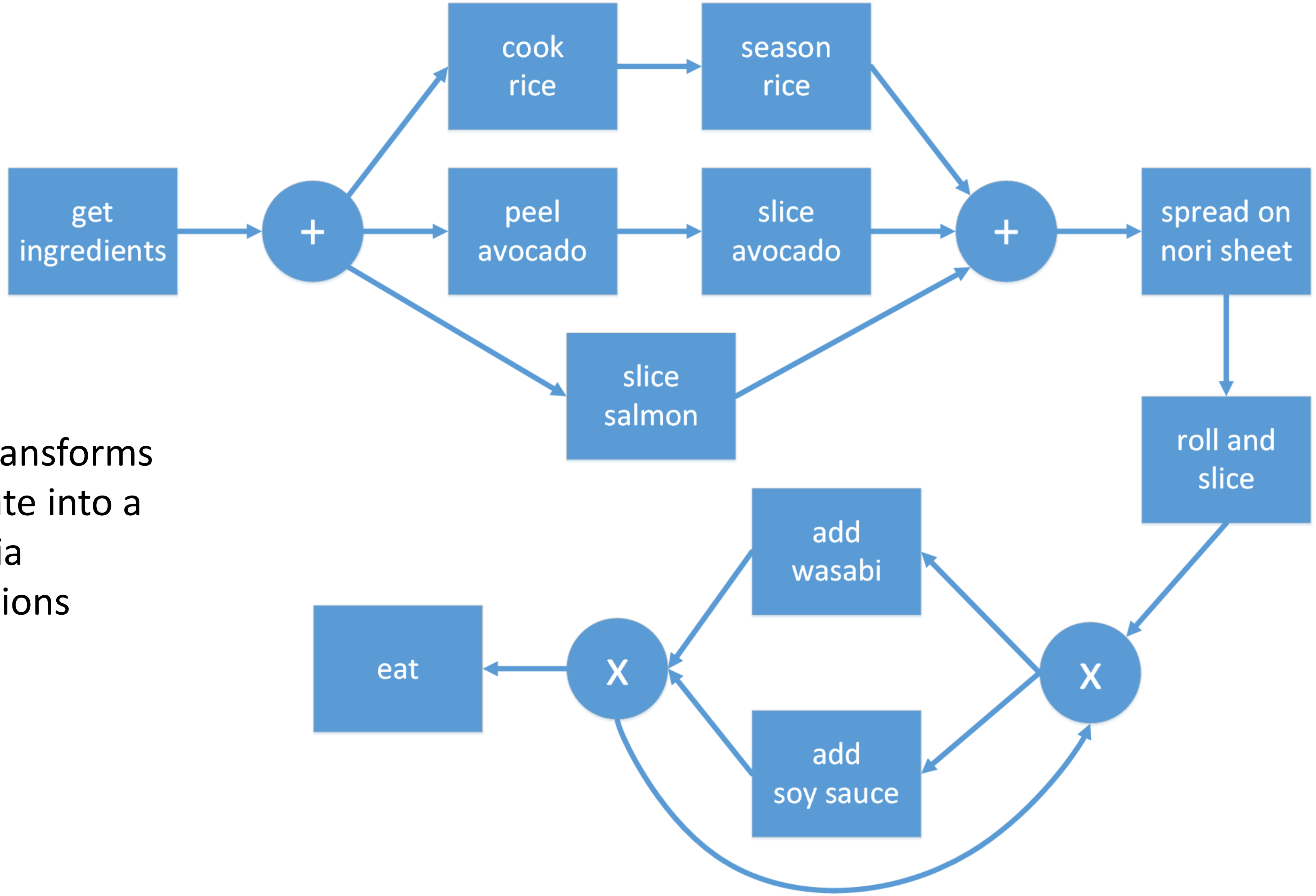
1. Introduction
2. Basics
3. Supervised Methods
4. Unsupervised Methods
5. Process Mining

5.1 Introduction
- Motivation
- Getting the Data

5.2 Process Models – An Overview
5.3 Process Discovery
5.4 Conformance Checking
5.5 Additional Mining Tasks
5.6 Streams

Processes in Applications

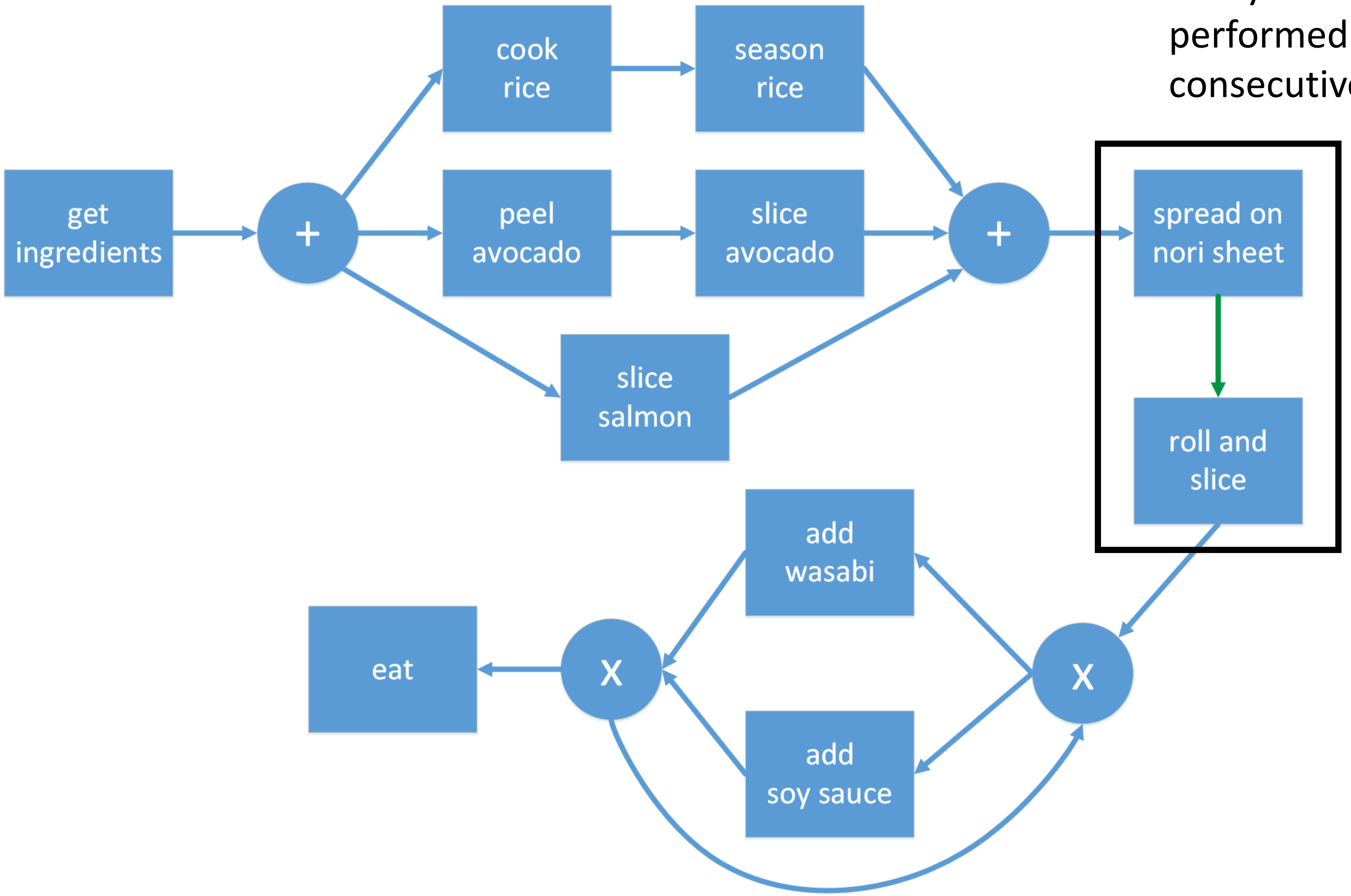
Example: The Sushi Process



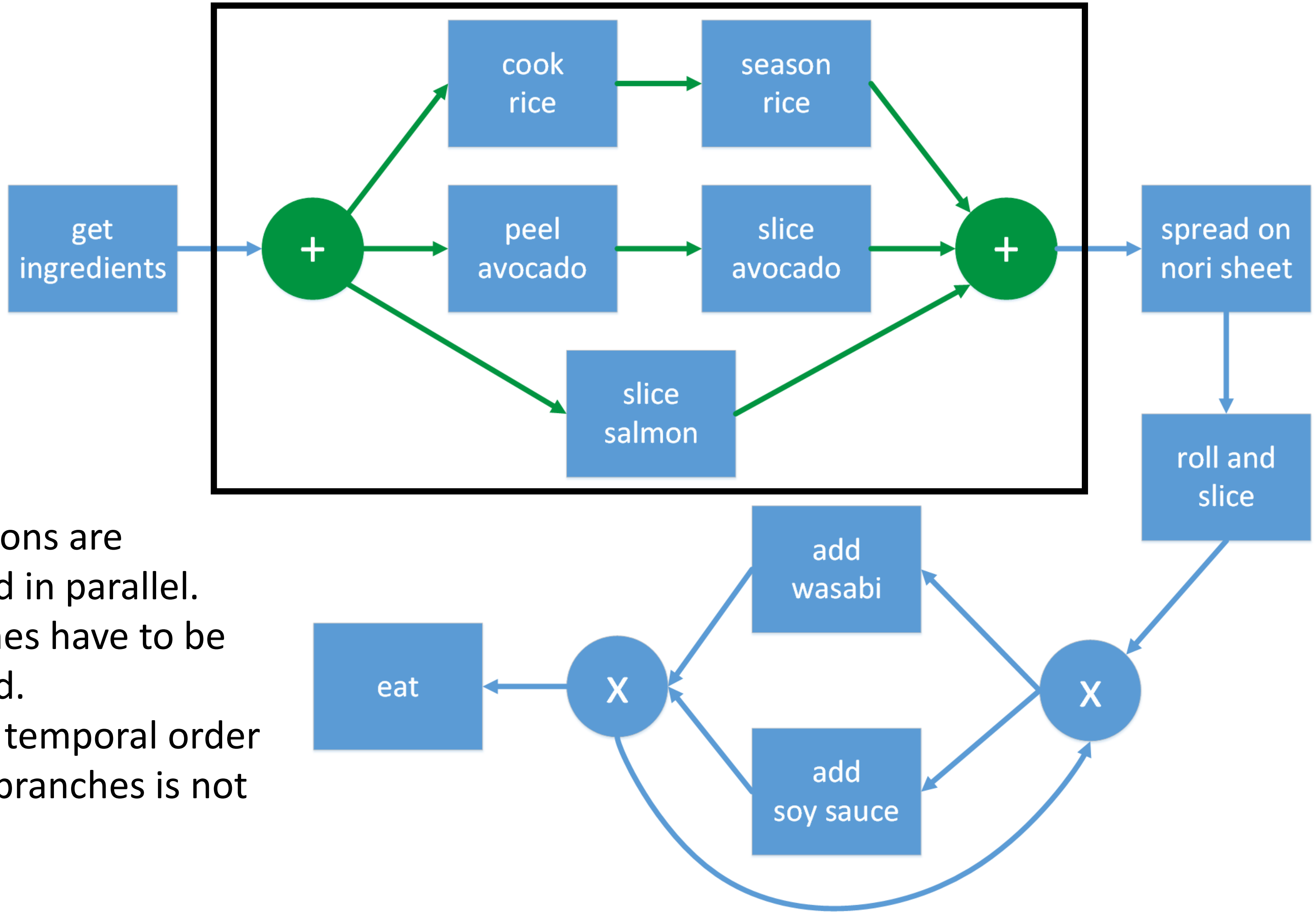
A process transforms an initial state into a final state via multiple actions

Process Properties: Sequence

- Many actions are performed in consecutive order

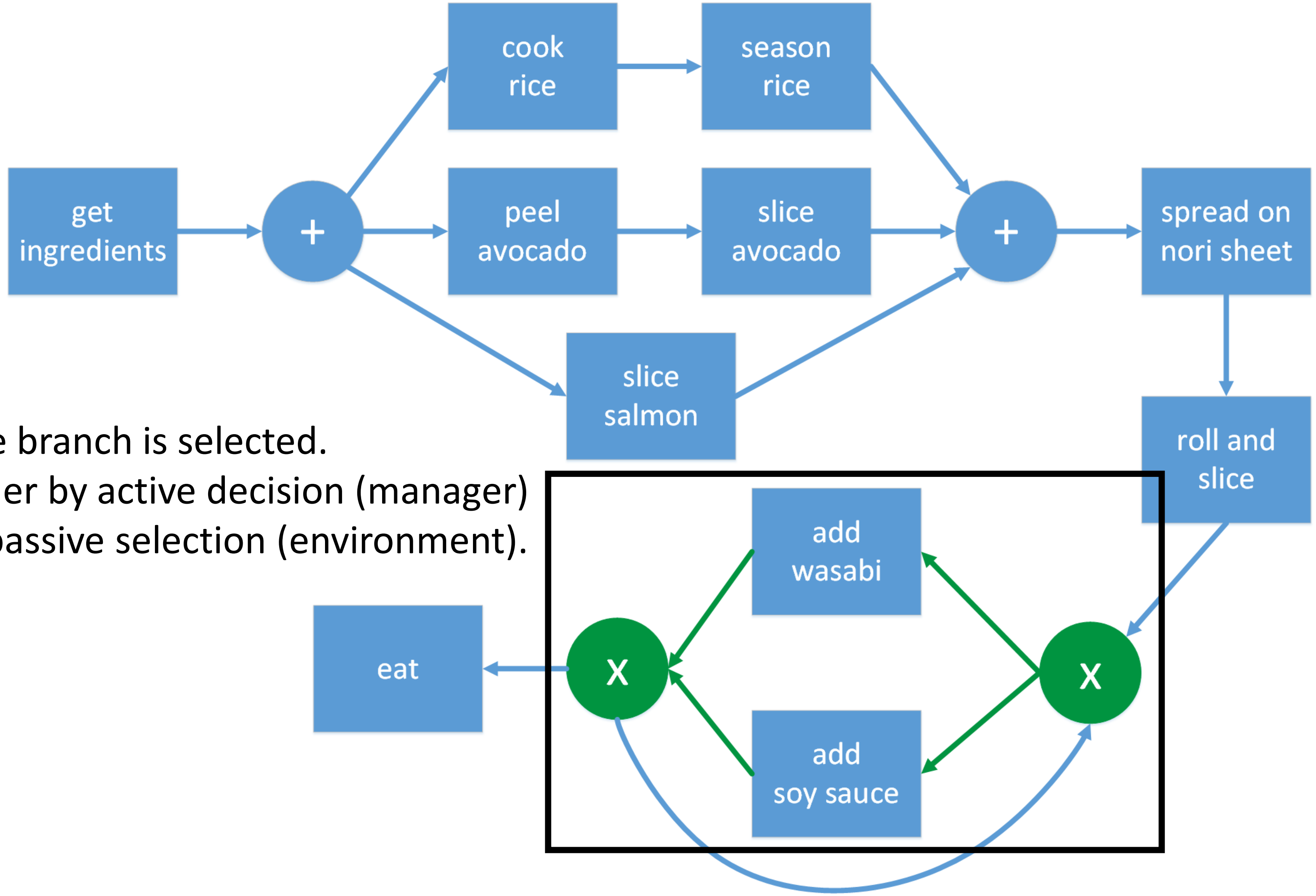


Process Properties: Concurrency



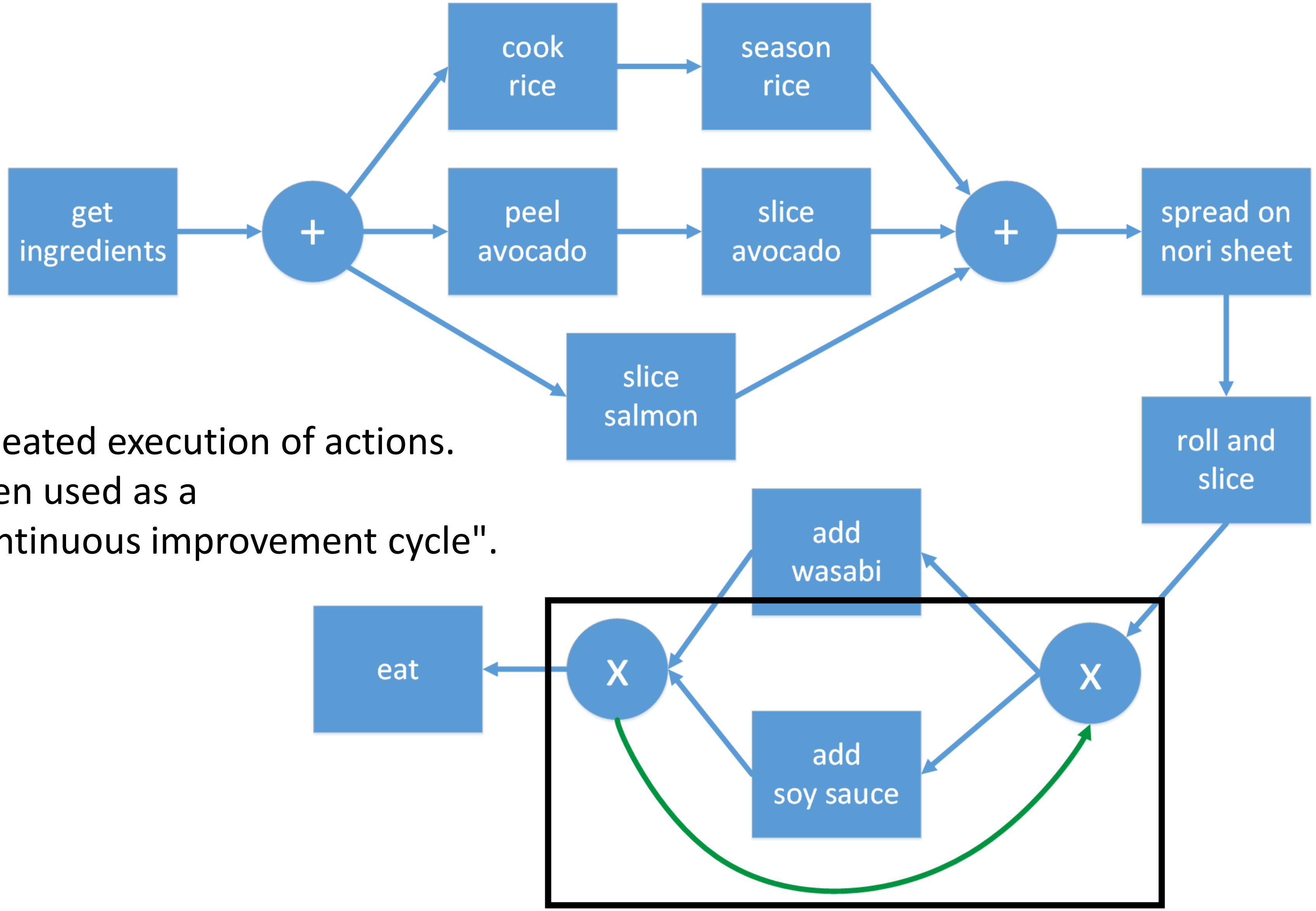
- Some actions are performed in parallel.
- All branches have to be performed.
- The exact temporal order between branches is not strict.

Process Properties: Choice



- One branch is selected.
- Either by active decision (manager) or passive selection (environment).

Process Properties: Loop

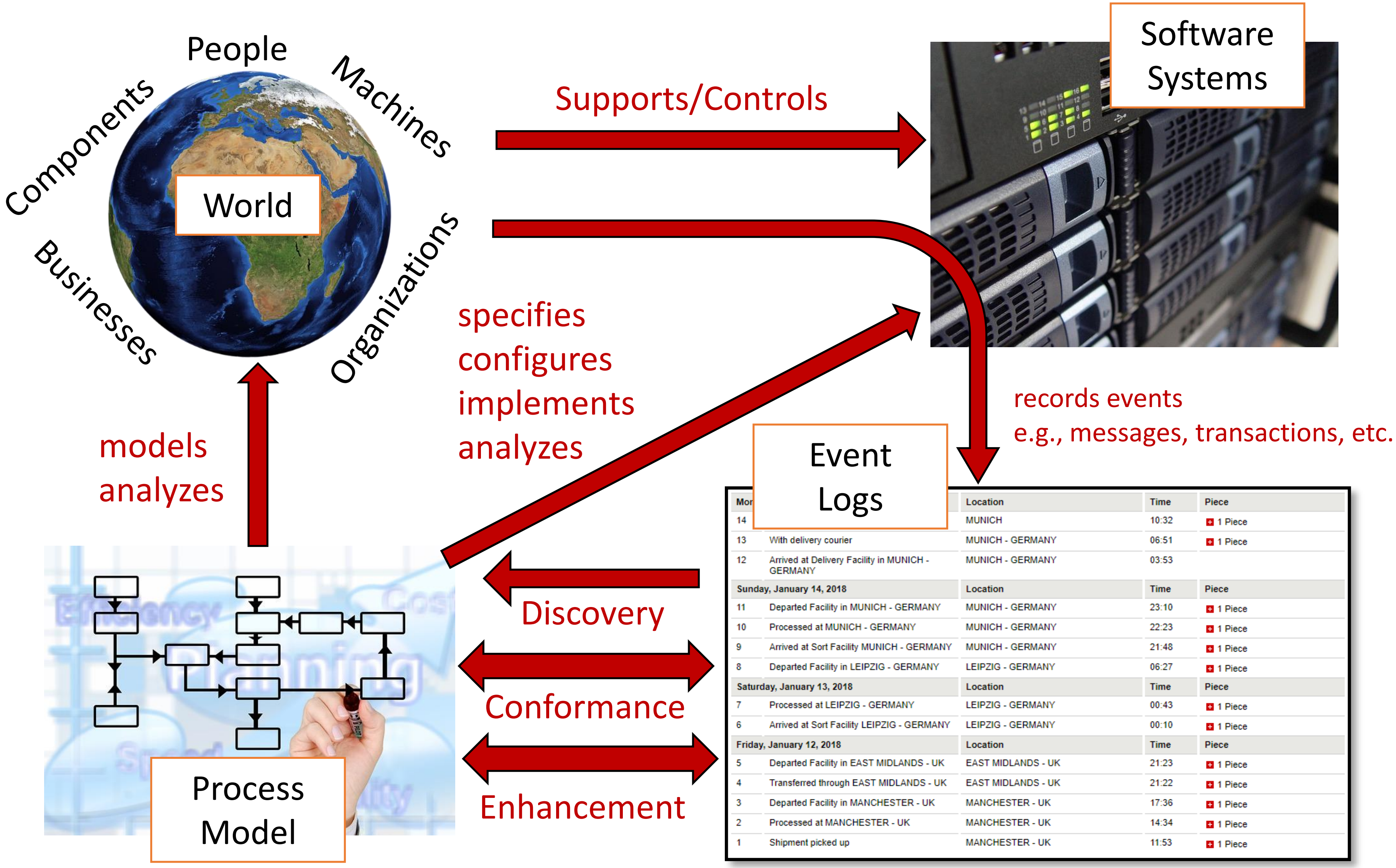


- Repeated execution of actions.
- Often used as a "continuous improvement cycle".

Benefits of Process Models

- Insights by changing perspectives and highlights.
- Specification / Documentation for certifications or legal contract purposes.
- Verification of executions to reveal problems.
- Performance analysis to identify issues like bottlenecks.
- Simulation (digital twin) to experiment virtually with changed settings.

Information Flow of Event Data



Mor		Location	Time	Piece
14		MUNICH	10:32	1 Piece
13	With delivery courier	MUNICH - GERMANY	06:51	1 Piece
12	Arrived at Delivery Facility in MUNICH - GERMANY	MUNICH - GERMANY	03:53	
Sunday, January 14, 2018		Location	Time	Piece
11	Departed Facility in MUNICH - GERMANY	MUNICH - GERMANY	23:10	1 Piece
10	Processed at MUNICH - GERMANY	MUNICH - GERMANY	22:23	1 Piece
9	Arrived at Sort Facility MUNICH - GERMANY	MUNICH - GERMANY	21:48	1 Piece
8	Departed Facility in LEIPZIG - GERMANY	LEIPZIG - GERMANY	06:27	1 Piece
Saturday, January 13, 2018		Location	Time	Piece
7	Processed at LEIPZIG - GERMANY	LEIPZIG - GERMANY	00:43	1 Piece
6	Arrived at Sort Facility LEIPZIG - GERMANY	LEIPZIG - GERMANY	00:10	1 Piece
Friday, January 12, 2018		Location	Time	Piece
5	Departed Facility in EAST MIDLANDS - UK	EAST MIDLANDS - UK	21:23	1 Piece
4	Transferred through EAST MIDLANDS - UK	EAST MIDLANDS - UK	21:22	1 Piece
3	Departed Facility in MANCHESTER - UK	MANCHESTER - UK	17:36	1 Piece
2	Processed at MANCHESTER - UK	MANCHESTER - UK	14:34	1 Piece
1	Shipment picked up	MANCHESTER - UK	11:53	1 Piece

Event Logs as Starting Point

case id	activity	timestamp	resource 1	resource 2	execution quality
		…			
Sushi 113	get ingredients	09:31	Andreas	bag	good
Sushi 239	slice salmon	09:35	Bianca	knife 1	medium
Sushi 239	spread on nori sheet	09:42	Bianca		very good
Sushi 248	eat	09:43	Charlie		-
Sushi 249	get ingredients	09:47	Andreas	bag	good
Sushi 113	cook rice	09:51	Bianca	rice cooker 3	poor
Sushi 239	roll and slice	09:51	Charlie	knife 1	good
Sushi 113	peel avocado	09:53	Andreas	knife 2	poor
Sushi 239	add soy sauce	09:54	Bianca		good
Sushi 239	add soy sauce	09:55	Bianca		poor
Sushi 239	eat	09:57	Andreas		-
		…			

Event Logs Technically

- Data collection mostly fully automated.
- Process-Aware Information Systems (PAIS)
  - ERP (Enterprise-Resource Planning) [SAP, Oracle]
  - BPM (Business Process Management) [IBM BPM]
  - CRM (Customer Relationship Management)
- Popular data format: XES
  - XML-based
  - easy to understand


```xml
<?xml version="1.0" encoding="UTF-8" ?>
<log xes.version="2.0" xes.features="arbitrary-depth" xmlns="http://www.xes-standard.org
    /">
    <extension name="Concept" prefix="concept" uri="http://www.xes-standard.org/concept.
        xesext"/>
    <extension name="Time" prefix="time" uri="http://www.xes-standard.org/time.xesext"/>
    <global scope="trace">
        <string key="concept:name" value=""/>
    </global>
    <global scope="event">
        <string key="concept:name" value=""/>
        <date key="time:timestamp" value="1970-01-01T00:00:00.000+00:00"/>
        <string key="system" value=""/>
    </global>
    <classifier name="Activity" keys="concept:name"/>
    <classifier name="Another" keys="concept:name system"/>
    <float key="log attribute" value="2335.23"/>
    <trace>
        <string key="concept:name" value="Trace number one"/>
        <event>
            <string key="concept:name" value="Register client"/>
            <string key="system" value="alpha"/>
            <date key="time:timestamp" value="2009-11-25T14:12:45:000+02:00"/>
            <int key="attempt" value="23">
                <boolean key="tried hard" value="false"/>
            </int>
        </event>
        <event>
            <string key="concept:name" value="Mail rejection"/>
            <string key="system" value="beta"/>
            <date key="time:timestamp" value="2009-11-28T11:18:45:000+02:00"/>
        </event>
    </trace>
</log>
```

Event Logs Formally

An **event** $\boldsymbol{e}$ is a tuple $e = (c, a, t, \dots)$ containing
a case identifier c,
an activity label a and
a timestamp t.

An event can contain additional attributes.

case id	activity	timestamp	resource 1	resource 2	execution quality
		…			
Sushi 113	get ingredients	09:31	Andreas	bag	good
Sushi 239	slice salmon	09:35	Bianca	knife 1	medium
Sushi 239	spread on nori sheet	09:42	Bianca		very good
Sushi 248	eat	09:43	Charlie		-
Sushi 249	get ingredients	09:47	Andreas	bag	good
Sushi 113	cook rice	09:51	Bianca	rice cooker 3	poor
Sushi 239	roll and slice	09:51	Charlie	knife 1	good
Sushi 113	peel avocado	09:53	Andreas	knife 2	poor
Sushi 239	add soy sauce	09:54	Bianca		good
Sushi 239	add soy sauce	09:55	Bianca		poor
Sushi 239	eat	09:57	Andreas		-
		…			

For an event $e = (c, a, t)$, we define the projections
$\#_{case}(e) = c$, $\#_{activity}(e) = a$, and $\#_{time}(e) = t$.

An **event log** $\boldsymbol{L}$ is a multiset of events.

Event Logs Formally

case id	activity	timestamp	resource 1	resource 2	execution quality
		…			
Sushi 113	get ingredients	09:31	Andreas	bag	good
Sushi 239	slice salmon	09:35	Bianca	knife 1	medium
Sushi 239	spread on nori sheet	09:42	Bianca		very good
Sushi 248	eat	09:43	Charlie		-
Sushi 249	get ingredients	09:47	Andreas	bag	good
Sushi 113	cook rice	09:51	Bianca	rice cooker 3	poor
Sushi 239	roll and slice	09:51	Charlie	knife 1	good
Sushi 113	peel avocado	09:53	Andreas	knife 2	poor
Sushi 239	add soy sauce	09:54	Bianca		good
Sushi 239	add soy sauce	09:55	Bianca		poor
Sushi 239	eat	09:57	Andreas		-
		…			

A **case** $\boldsymbol{C}$, identified by c in the log, is the set of events

$$C = \{e \in L \mid \#_{case}(e) = c\}$$

A **trace** $\boldsymbol{\sigma_c}$ is the sequence of activities for a case $C = \{e_1, \dots, e_n\}$ with

$$\sigma_c = \#_{activity}(e_{\pi(1)}), \dots, \#_{activity}(e_{\pi(n)})$$

such that $\#_{timestamp}(e_{\pi(i)}) < \#_{timestamp}(e_{\pi(j)})$ for $\pi(i) < \pi(j)$.

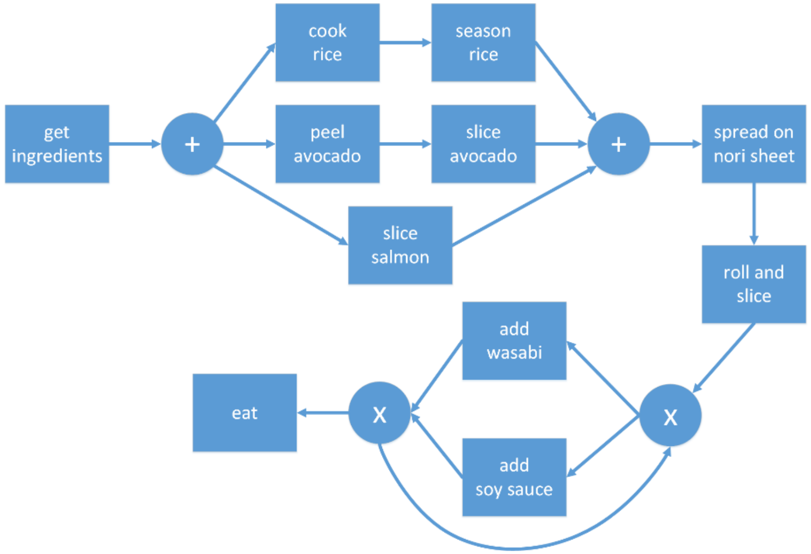
Integration into the Data Mining World

Itemsets
(e.g. frequent itemset mining)

Processes

Sequences
(e.g. sequential pattern mining)

$\{rice, avocado, salmon\}$



$get\ ingredients$
$\rightarrow prepare\ ingredients$
$\rightarrow spread\ on\ nori\ sheet$
$\rightarrow roll\ and\ slice$
$\rightarrow season\ with\ wasabi$
$\rightarrow season\ with\ soy\ sauce$
$\rightarrow eat$

no order ⟷ total order

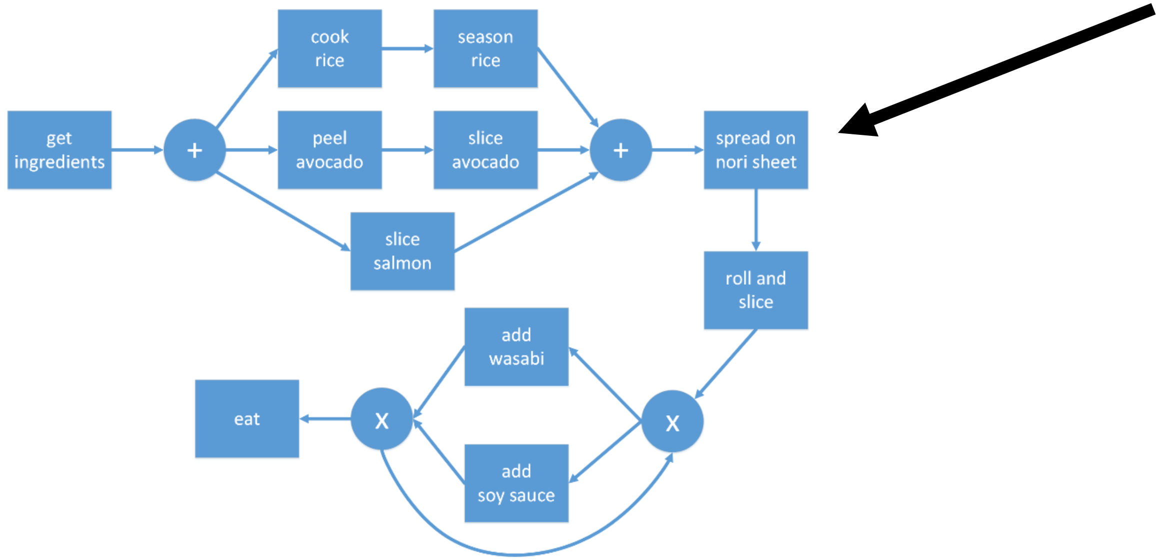
- unordered
- set-based

- partially ordered
- sequences can occur, models are directed graphs
- branches break order (concurrency)

- strictly totally ordered
- sequence-based

Process Mining Task: Discovery

- Given an event log, find a process model which
  - must be able to replay the log ⇒ *Fitness*
  - simplifies as far as possible ⇒ *Simplicity*
  - does not overfit the log ⇒ *Generalization*
  - does not underfit the log ⇒ *Precision*

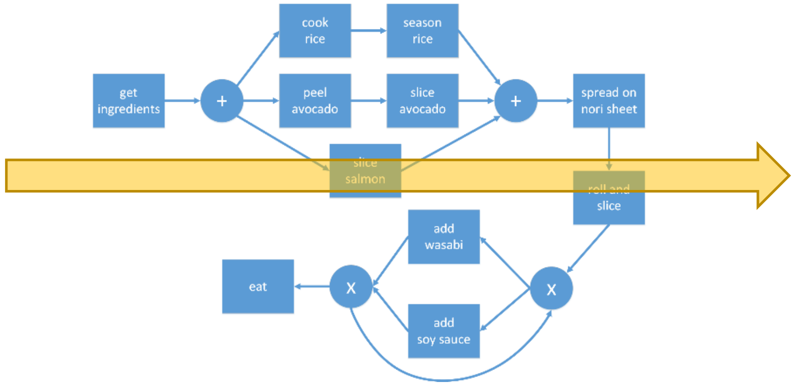


case id	activity	timestamp
…		
Sushi 113	get ingredients	09:31
Sushi 239	slice salmon	09:35
Sushi 239	spread on nori sheet	09:42
Sushi 248	eat	09:43
Sushi 249	get ingredients	09:47
Sushi 113	cook rice	09:51
Sushi 239	roll and slice	09:51
Sushi 113	peel avocado	09:53
Sushi 239	add soy sauce	09:54
Sushi 239	add soy sauce	09:55
Sushi 239	eat	09:57
…		

Process Mining Task: Conformance Checking

- Given an event log and a process model, decide for each case whether it conforms to the model or not. If not, give the issues.

cook rice, add wasabi, roll and slice, eat

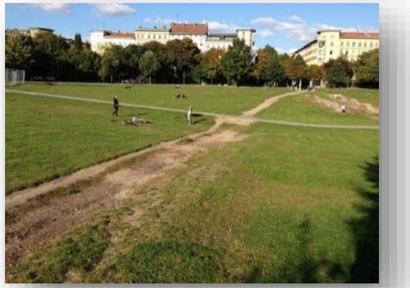


conform

non-conform

- A case instance can perform better than others. Then reveal the beneficial deviations to improve the general workflow.
- If the case performs worse, identify the root cause to avoid misbehavior.

Housebreaking

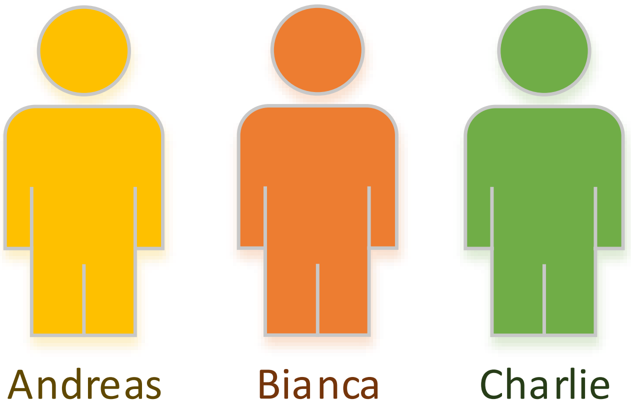


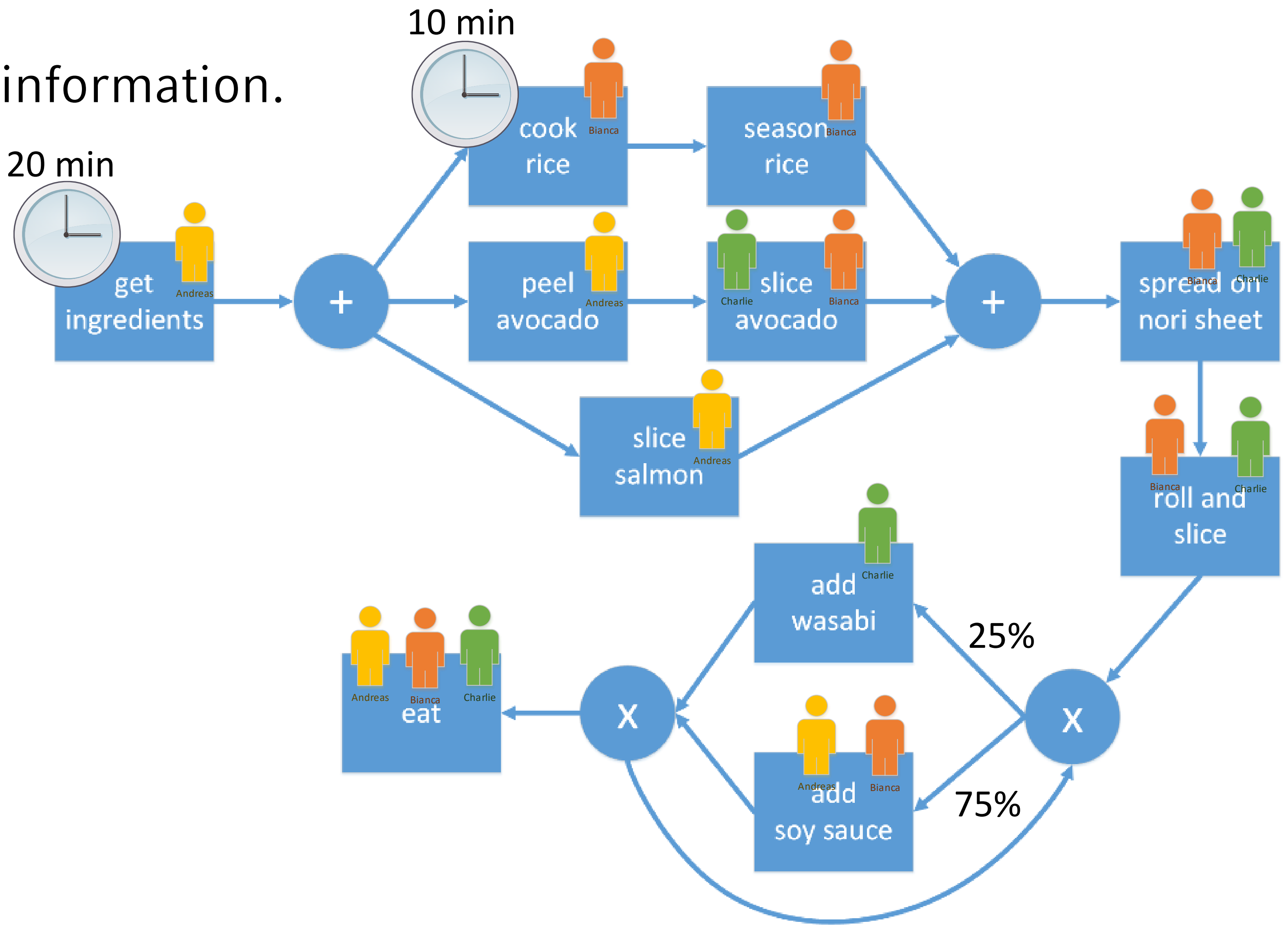
Trails

Tool choice

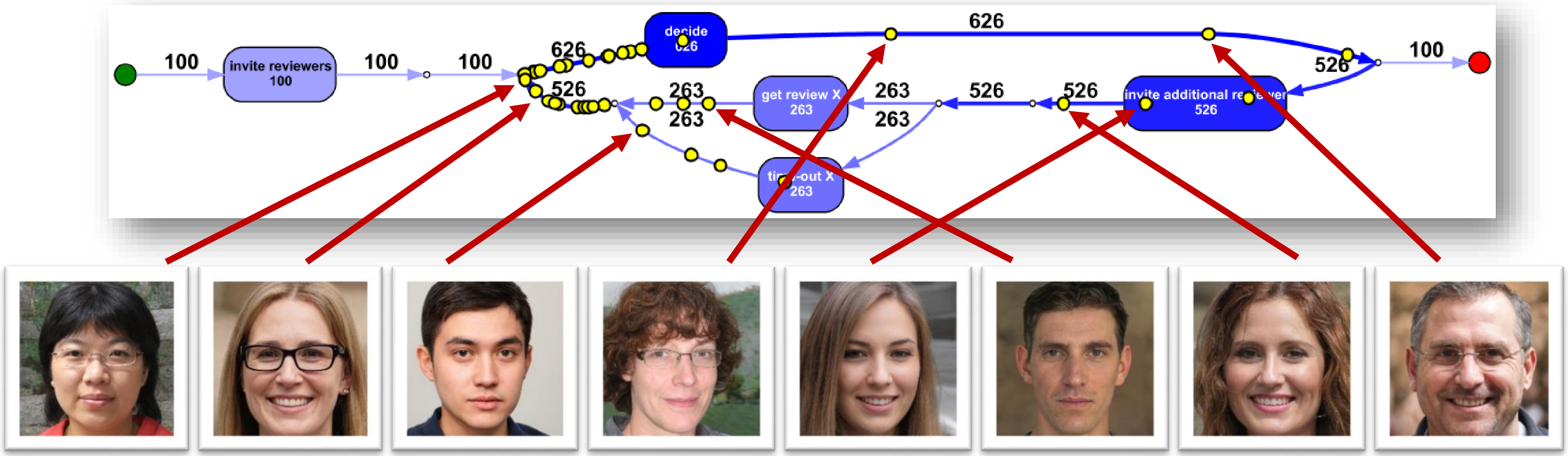
Process Mining Task: Enhancement

- Given a process model,
  augment with additional information.
  - Temporal information
  - Social networks
  - Organisational roles
  - Decision rules





Process Mining Risks and Green Data Science

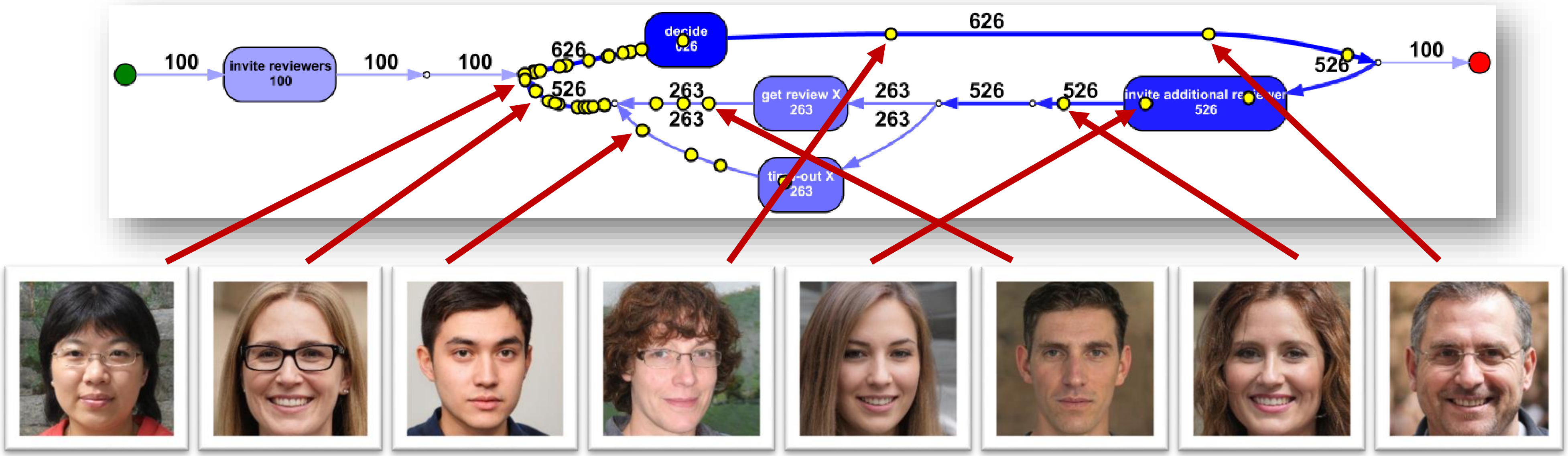


Made with StyleGAN
arXiv:1912.04958

- Mostly: Cases related to people. But what is in the data?
  - Students *Who asks the most questions?*
  - Employees *Who is associated with long execution terms?*
  - Tenants *Who needs maintenance often?*
  - Clients *Who calls most for service?*

neutral,
objective,
data-oriented

Process Mining Risks and Green Data Science

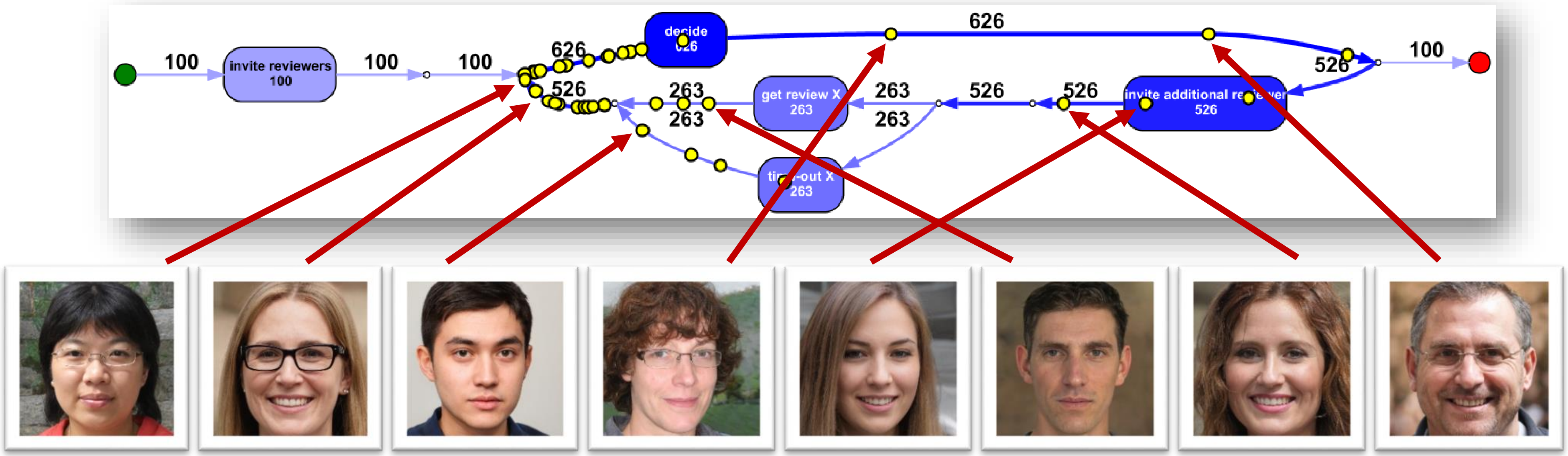


Made with StyleGAN
arXiv:1912.04958

- Same results, but with intentional mindset:
  - Students *Who is the least intelligent student?*
  - Employees *Who is the slowest worker?*
  - Tenants *Who caused the most repairs?*
  - Clients *Who complains the most?*

bad intention, negative-subjective, pessimistic

Process Mining Risks and Green Data Science

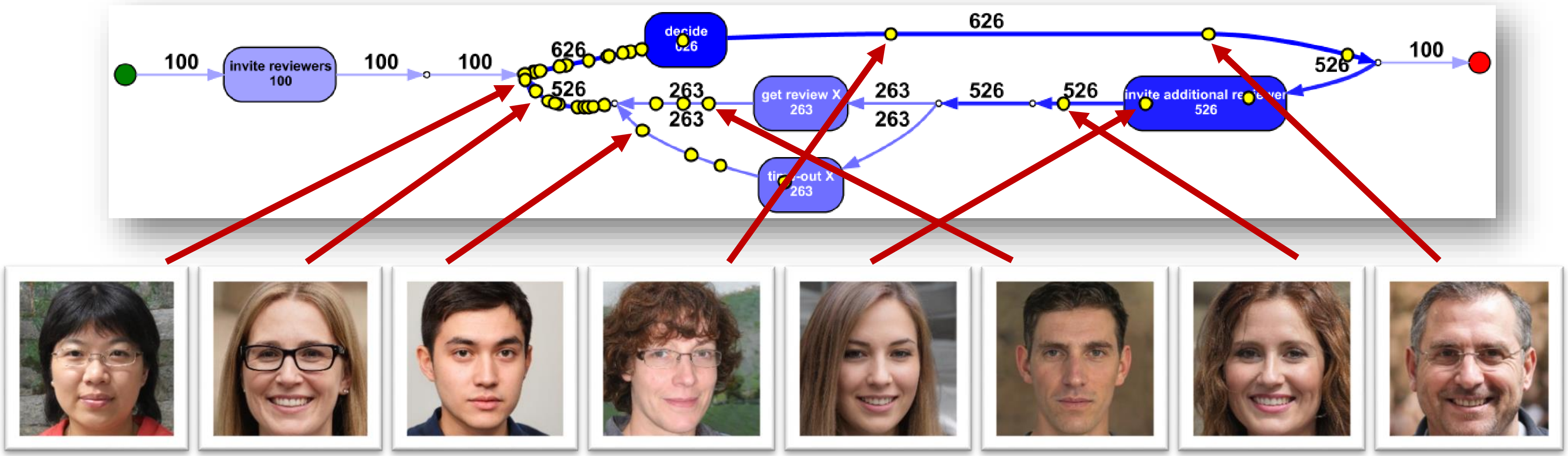


Made with StyleGAN
arXiv:1912.04958

- And the other extreme, changed mindset:
  - Students *Who is the most interested student?*
  - Employees *Who handles the most difficult tasks?*
  - Tenants *Who takes care of the rental property?*
  - Clients *Who gives a lot of constructive feedback?*

good intention,
positive-
subjective,
optimistic

Process Mining Risks and Green Data Science



Made with StyleGAN
arXiv:1912.04958

- Be careful with interpretations.
- Even if you are objective, can your results be interpreted otherwise?
- Can you obscure the results so they stay meaningful, but protect individuals?
  e.g. cluster individuals, top-k-rankings, k-anonymity, hashing, noise addition,…

Scientific Process Mining Tools

- PROM:
  - First version in 2010.
  - Java-based.
  - Provides many algorithms in a GUI.

- pm4py:
  - First version in 2019
  - Python-based
  - Documentation: https://pm4py.fit.fraunhofer.de/
  - Several algorithms available

Agenda

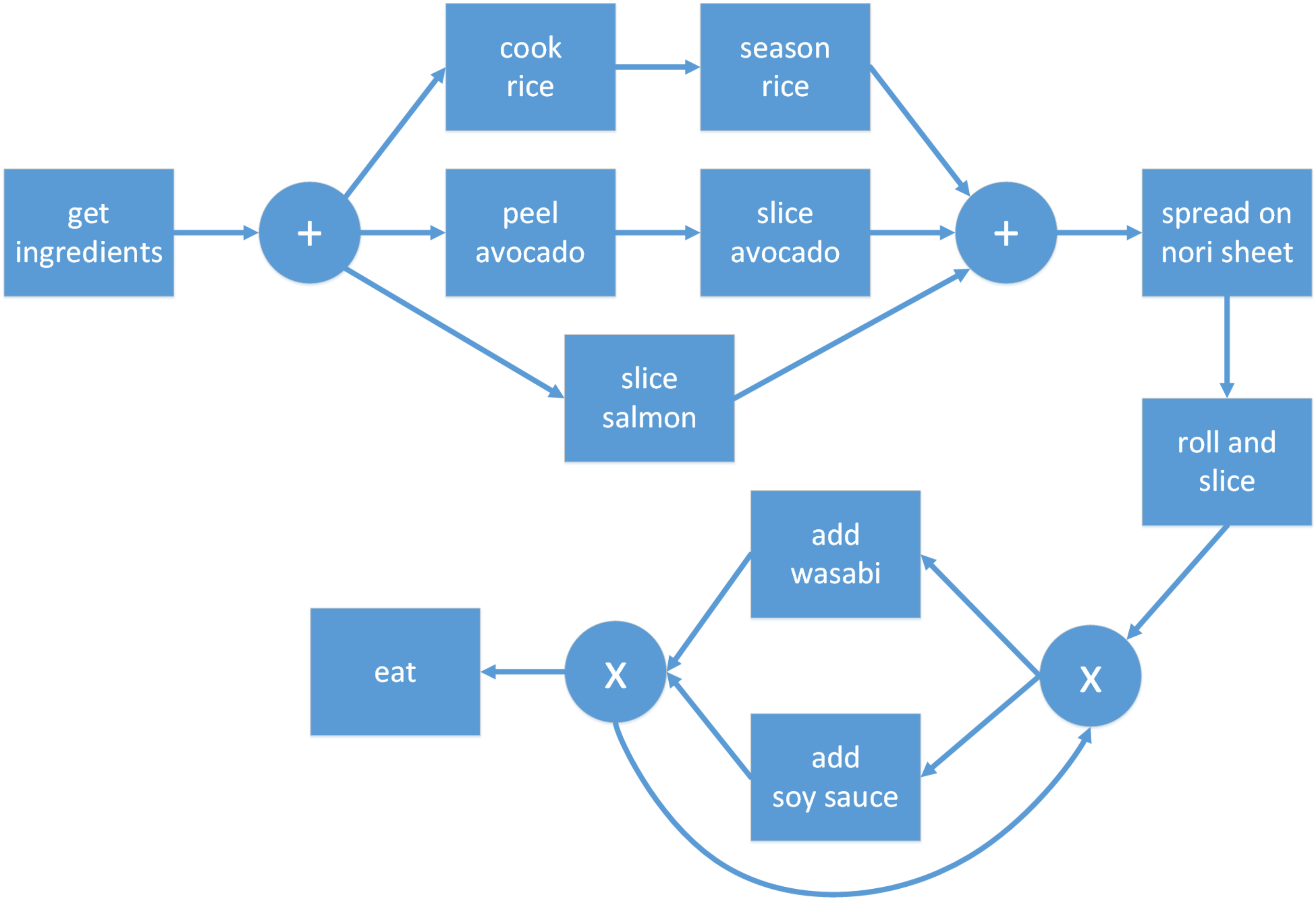
Motivation

Why do we need Process Models?

- Predetermine operational processes in the form of guidelines
  - Descriptive vs. Normative model
- Visualization of processes
- Process reasoning
- Analysis of given processes
  - Starting point for initial implementation and re-design
  - Distribution of responsibilities
  - Planning and controlling
  - Compliance checking
  - Performance prediction via simulation
  - …

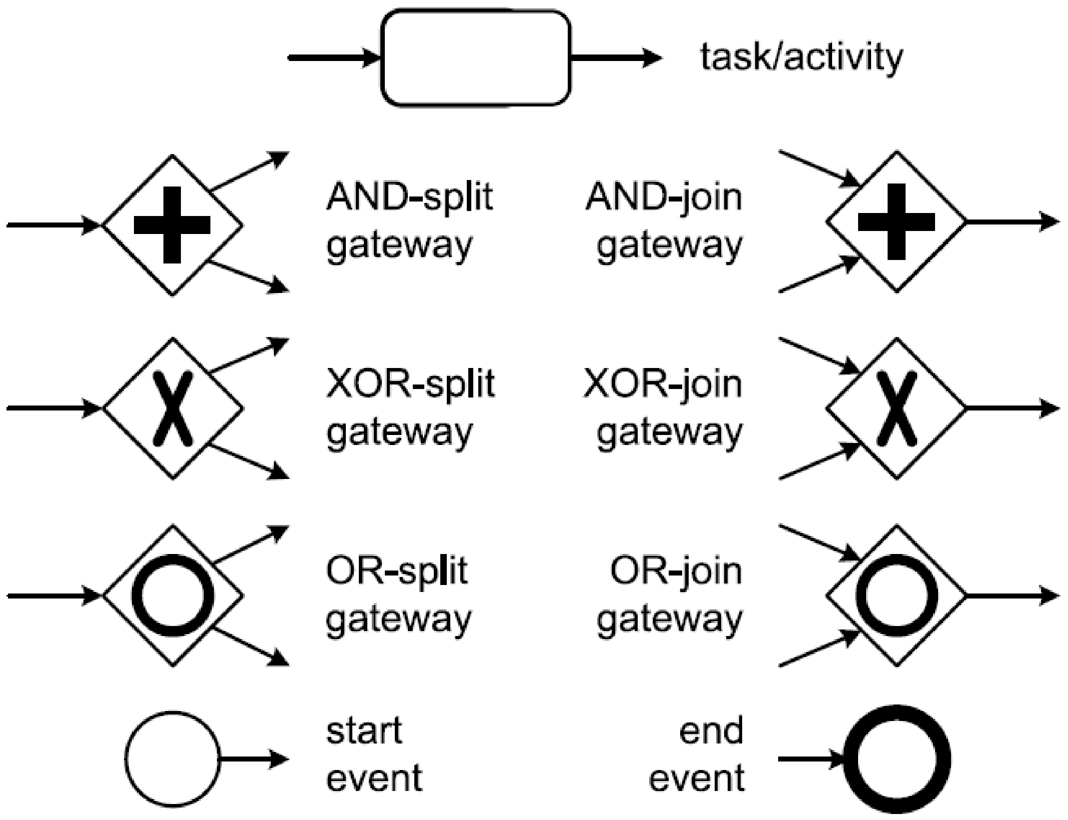
Process Model – BPMN (Business Process Modeling Notation)

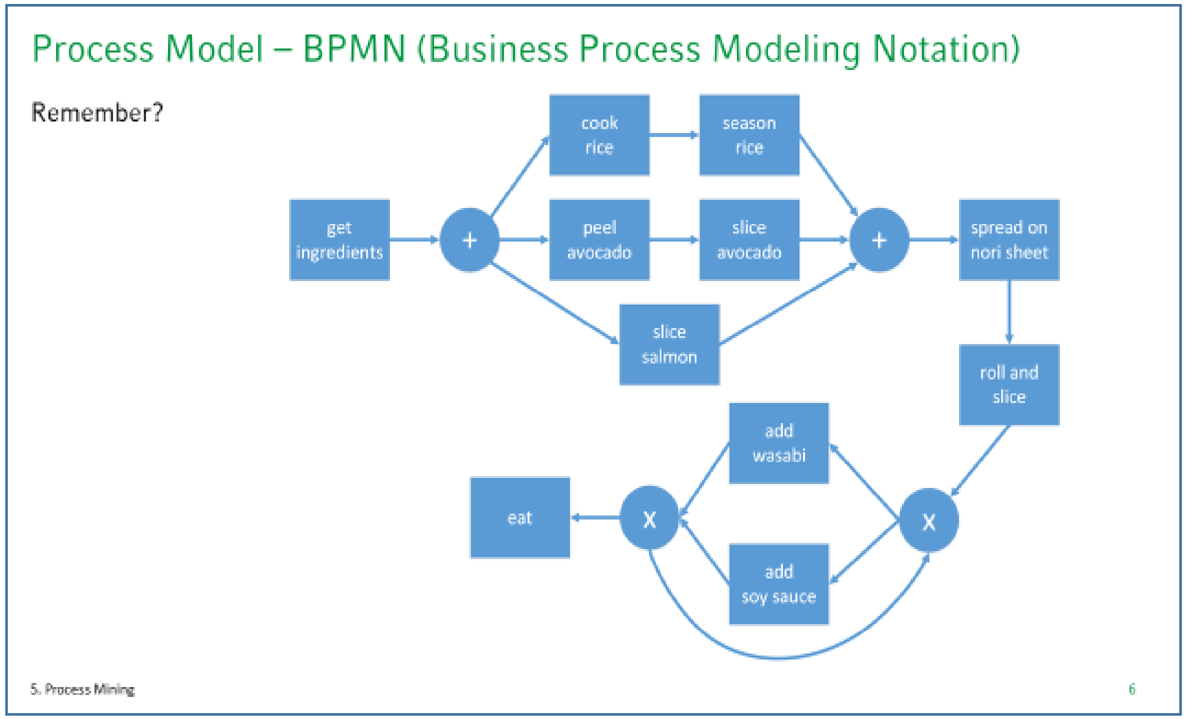
Remember?



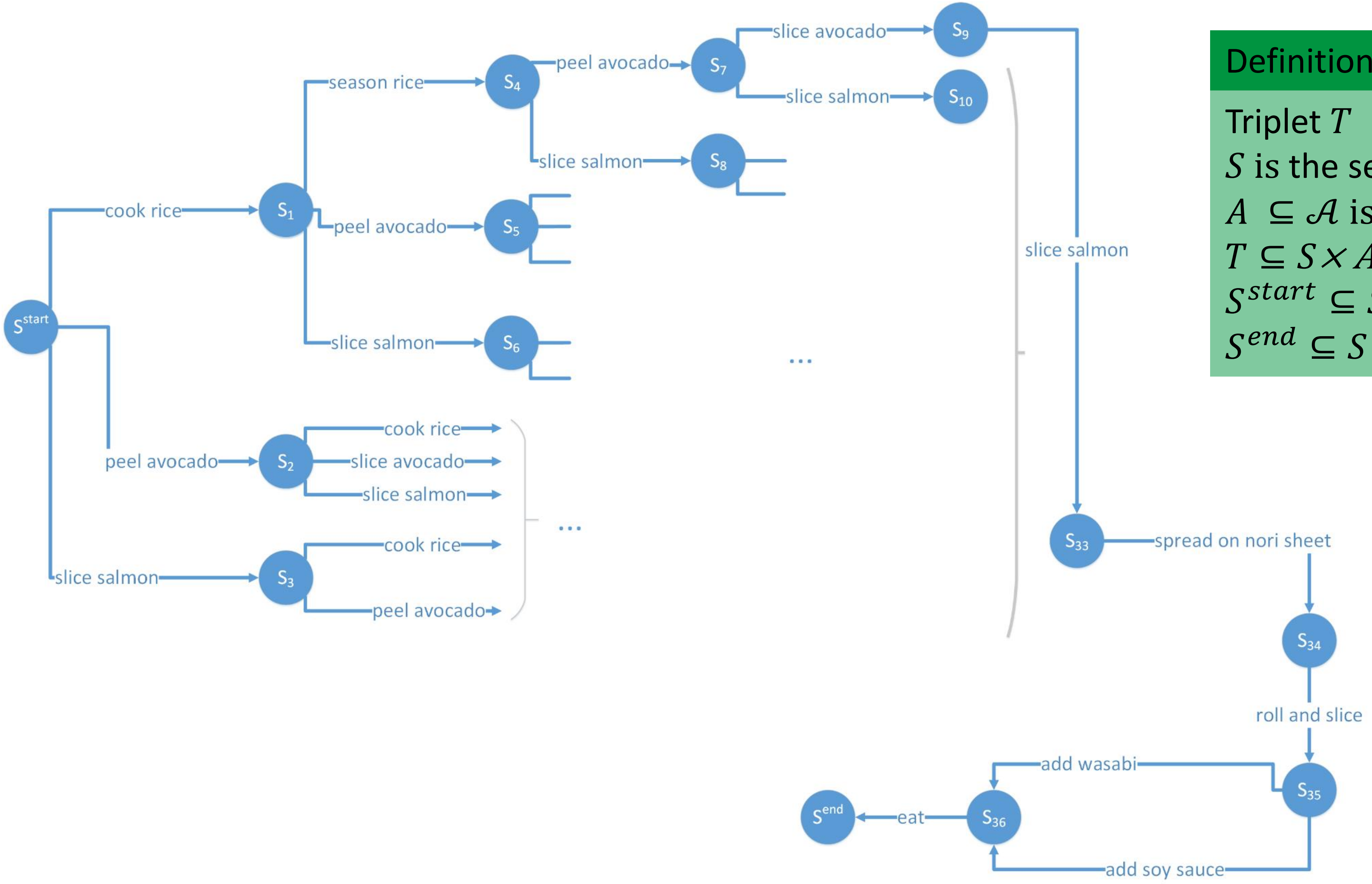
Process Model – BPMN (Business Process Modeling Notation)

Exemplary subset of elements contained in BPMN





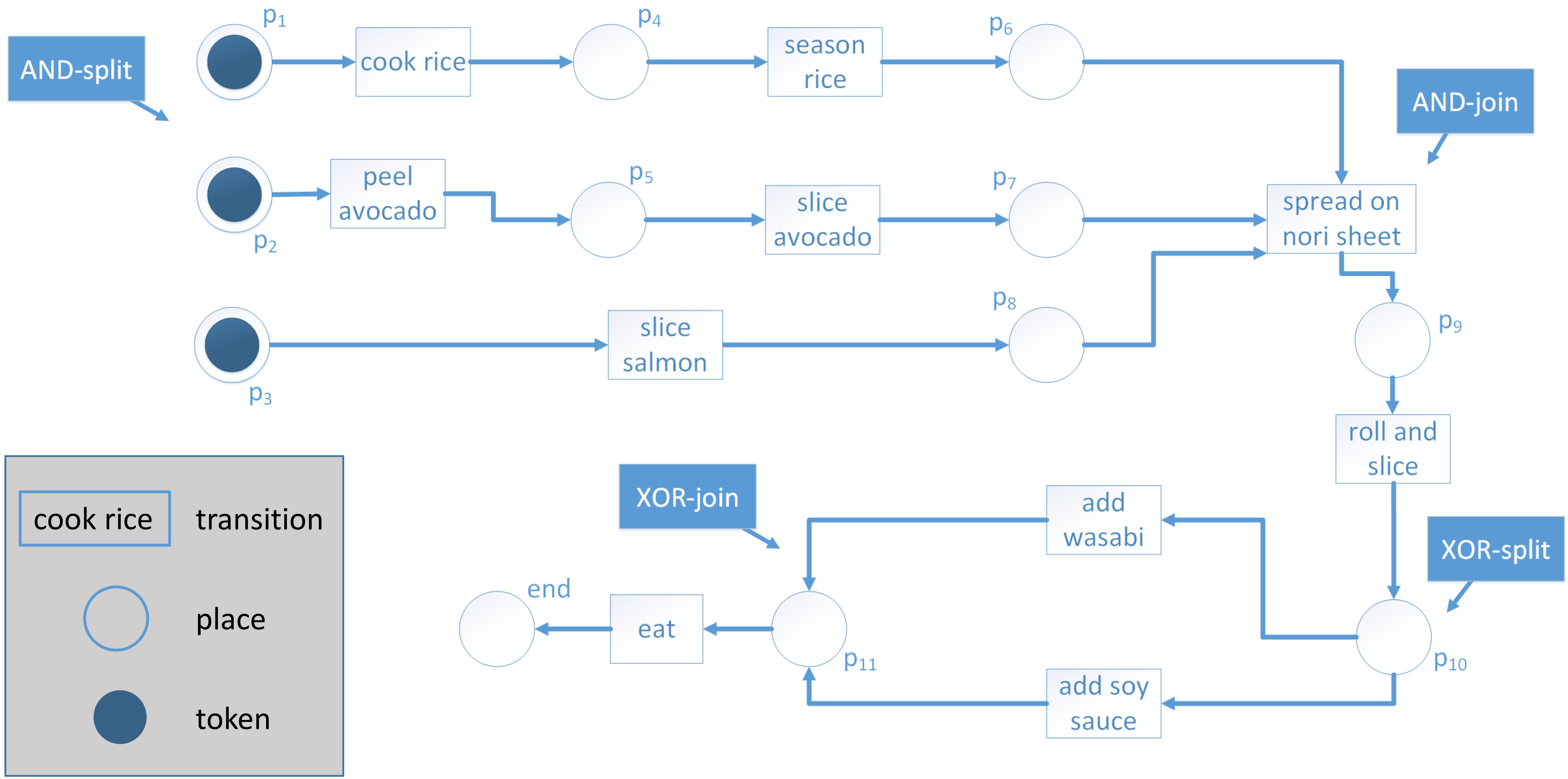
Process Model – Transition System



Definition (Transition system)

Triplet $T = (S, A, T)$, where
S is the set of *states*
$A \subseteq \mathcal{A}$ is the set of *activities*
$T \subseteq S \times A \times S$ is the set of *transitions*
$S^{start} \subseteq S$ is the set of *inital states*
$S^{end} \subseteq S$ is the set of *final states*

Process Model – Petri Nets

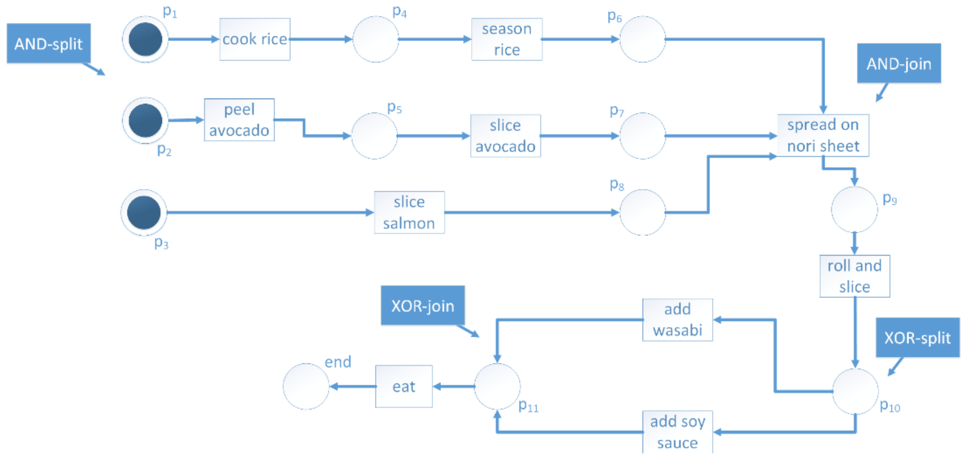


Process Model – Petri Nets

As already seen the Petri net is a bipartite graph.

Definition (Petri Net)

Triplet N $= (P, T, F)$, where
P is a finite set of *places*
T is a finite set of *transitions*, $P \cap T = \emptyset$
$F \subseteq (T \times P) \cup (P \times T)$ is a set of *directed arcs* (called *flow relation*)



Exemplary formalization of given Petri Net:

P = {p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, end}

T = {cook rice, season rice, peel avocado, slice avocado, slice salmon,
spread on nori sheet, roll and slice, add wasabi, add soy sauce, eat}

F = {(p_1, cook rice), (p_2, peel avocado), (p3, slice salmon), (cook rice, p4), (peel avocado, p5), …}

Process Models – Workflow-Nets (WF-Nets)

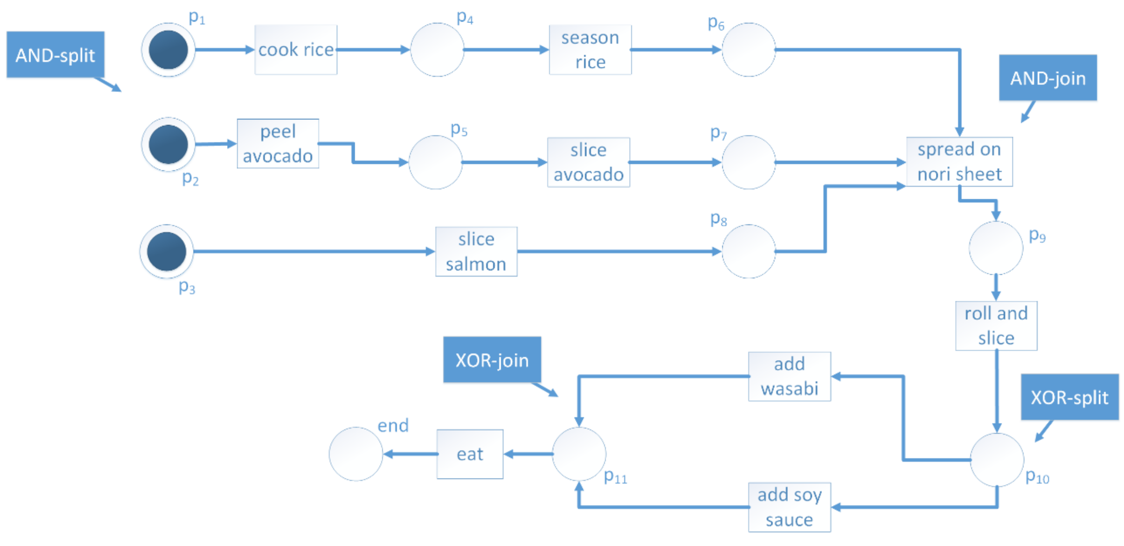
Subclass of Petri Nets

Definition (Workflow Net)

$N = (P, T, F)$, where
(P, T, F) is a Petri net as already defined
N is a *workflow net* iff.
a) P contains a source place i s. t. $\bullet\, i = \emptyset$
b) P contains a sink place o s. t. $o \bullet = \emptyset$
c) If we add a transition t^* to N which connects o with i
i. e. $\bullet t^* = \{o\}$ and $t^* \bullet = \{i\}$, then
the resulting Petri net is strongly connected.

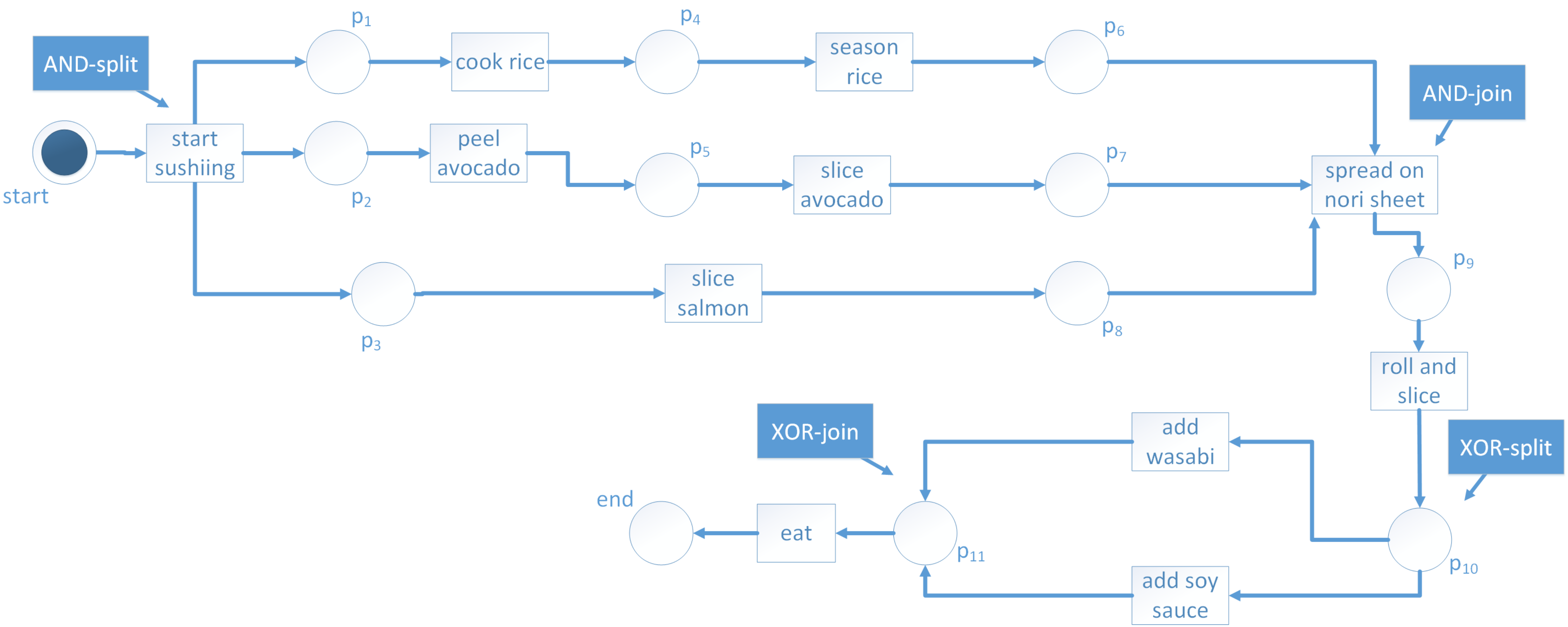
Definition (Strongly connected)

A Petri net is strongly connected iff for every pair of nodes (i.e. places and transitions) *x* and *y*, there is a path leading from *x* to *y*



Can the Petri Net shown be considered a Workflow Net?

Process Models – Workflow-Nets (WF-Nets)



Process Models – Additional Criterion (Soundness)

A WF-net does not necessarily represent a correct process

→ Deadlocks, livelocks, not activatable activities etc. are possible

Definition (Soundness)

Let $N = (P, T, F)$ be a $workflow\ net$ with i and o as input and output places.
N is *sound* iff.

- *(safeness)* Places do not hold multiple tokens at the same time
- *(proper completion)* The moment the procedure terminates there is a token in place *o* and all the other places are empty
- *(option to complete)* For any case the procedure will terminate eventually
- *(absence of dead parts)* For any $t \in T$ there is a firing sequence enabling t

Process Models – Methods (Verification)

Verification is a method to analyze process models against specific properties (*Model checking*).

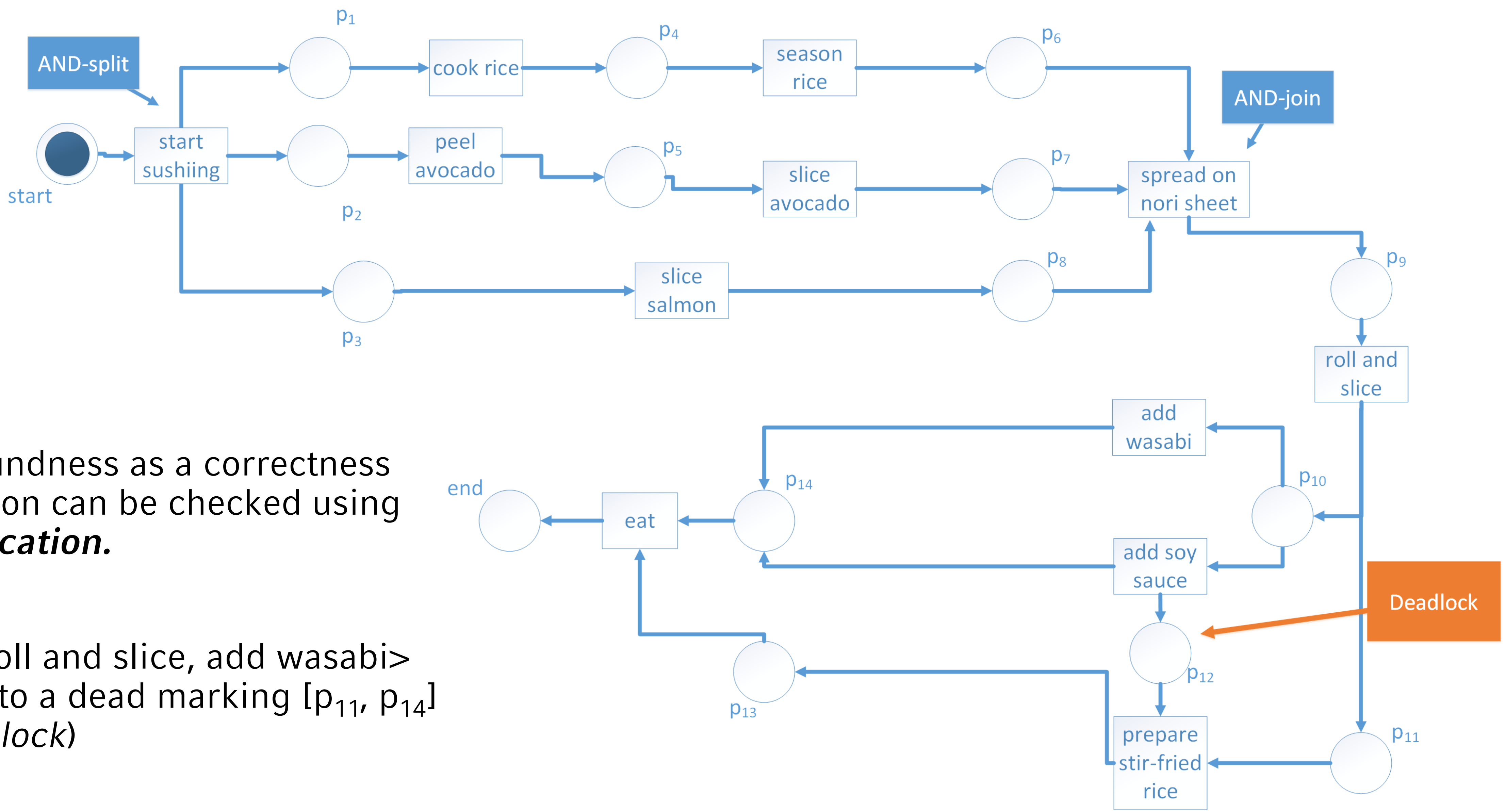
- Those properties can be expressed in temporal logic.
- Specifically in LTL (Linear Temporal Logic) which is an significant example in relation to process models.

<u>Two further exemplary verification tasks in the following:</u>

1. Two process models can be checked against each other using ***Verification***.

E.g. Trying to match a descriptive and a normative model to see where reality differs from guidelines

Process Models – Methods (Verification)



2. Soundness as a correctness criterion can be checked using ***Verification.***

<…, roll and slice, add wasabi> leads to a dead marking [p_{11}, p_{14}] (*Deadlock*)

Process Models – Roundup

Known process model types so far:

- Transitions systems
- BPMN
- Petri Nets
- Workflow Nets

There are still others like

- Reachability graphs
- Causal nets
- …

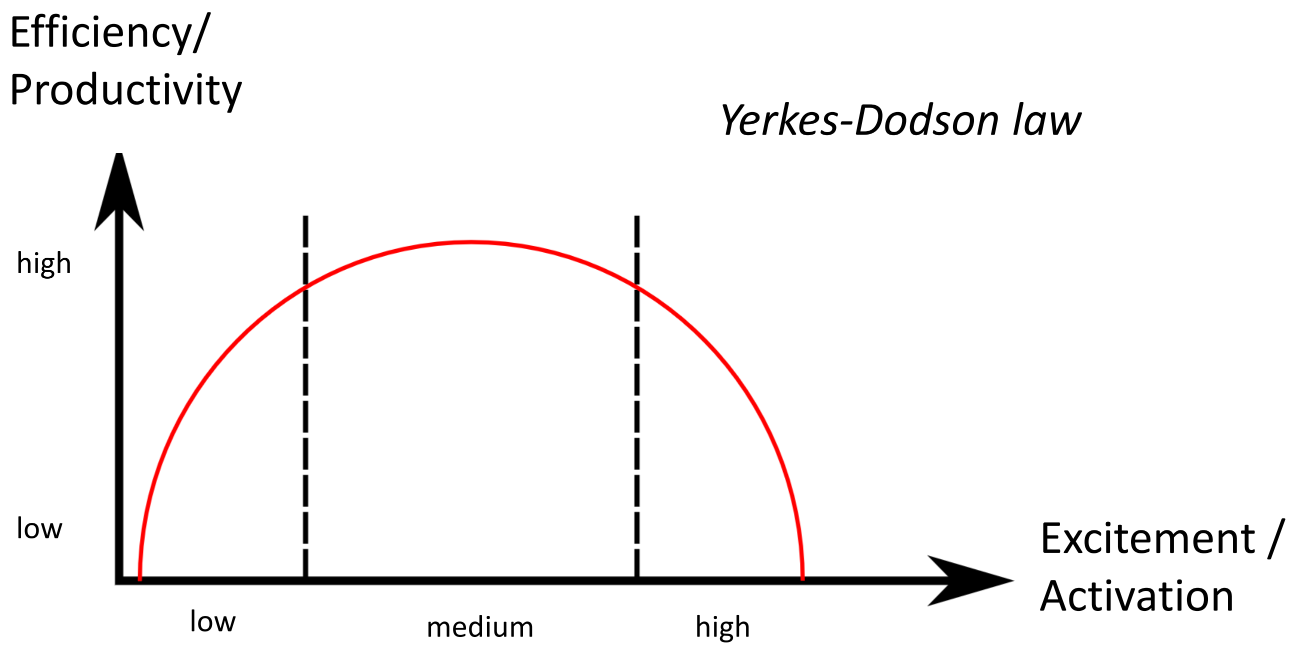
Benefit:

- Process analysis gets simplified
- Predict performance via simulation
- Predetermine guidelines
- Purpose determines outcome
- …

Process Models – Discussion

Creating a model is not an easy task

- **Capturing human behavior**
  - Human covers multiple processes with different priorities → dependencies evolve
    → Difficult to model one process in isolation
  - Productivity of a human is varying over time.
    It also depends on other factors e.g. Yerkes-Dodson law



Process Models – Discussion (cont.)

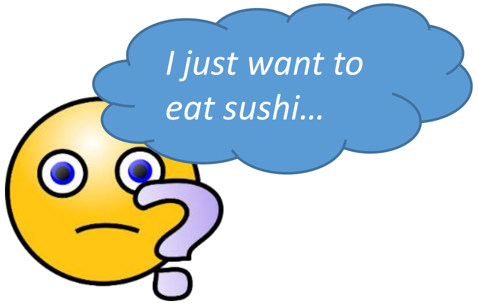
- **Idealization of reality**
  - Hand-made models tend to be
    - subjective
    - oversimplified
  - The choice of a representative sample of cases is crucial
    → Biased focus on *normal / desirable* behavior

Actually it only covers 80% of cases but is seen as a representative

Remaining 20% could possibly cover high amount of problems

Process Models – Discussion (cont.)

- **Granularity**
  E.g. there are many types of sushi: Nigiri, Sashimi, Maki, Uramaki…



E.g. **discrete** vs. **continuous**

cook rice vs. get pot → get rice → fill rice into pot → add water → …

⇒ A suitable granularity for the process model depends on
  - **the input data**
  - **the model‘s purpose**

References

Yerkes, R.M., & Dodson, J.D. (1908). The Relation of Strength of Stimulus to Rapidity of Habit Formation. *Journal of Comparative Neurology & Psychology, 18,* 459–482. https://doi.org/10.1002/cne.920180503

Wil van der Aalst. 2016. *Process Mining: Data Science in Action* (2nd. ed.). Springer Publishing Company, Incorporated

Wil van der Aalst. *(1998). "The application of Petri nets to workflow management" (PDF). Journal of Circuits, Systems and Computers.* ***8*** *(1): 21–66.*