Proveniência de Dados

Troy Costa Kohwalter

troy@ic.uff.br

Instituto de Computação UFF

Dados de pesquisa

Por que compartilhar?

Como compartilhar?

COMPARTILHAMENTO DE DADOS DE PESQUISA

Dados de Pesquisa

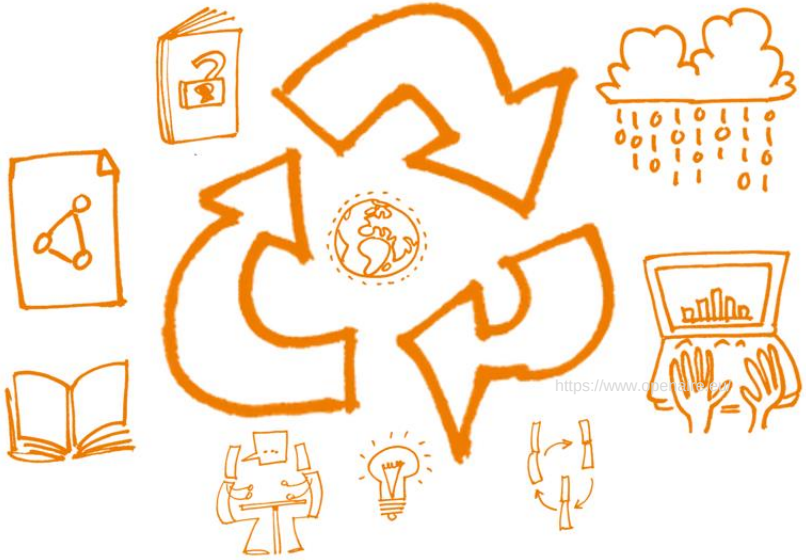
- Definição:

> *"Dados são fatos, observações ou experiências sobre as quais um argumento ou teoria é construído ou testado. Os dados podem ser numéricos, descritivos, auditivos ou visuais. Os dados podem ser brutos, abstraídos ou analisados, experimentais ou observacionais."*
>
> - Política de Dados de Pesquisa da UCL (*University College London*)

https://blog.opinionbox.com/

Dados de Pesquisa

- São as fontes ou materiais originais que você criou ou reuniu para conduzir seu projeto de pesquisa
- Podem ser digitais ou não digitais
- Resposta à sua *Pergunta de Pesquisa* é baseada na análise desses dados de pesquisa



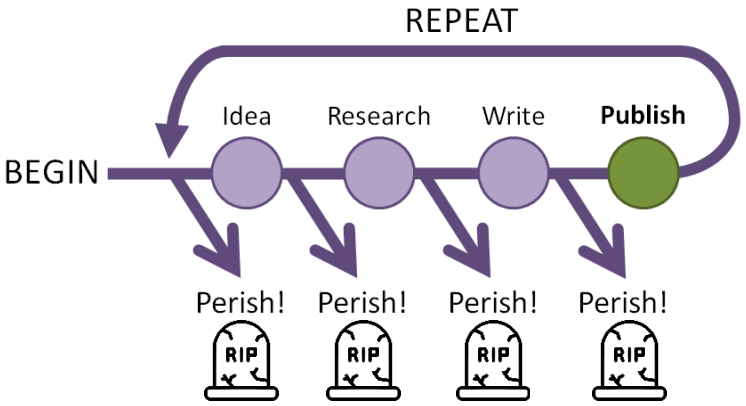
Dados de Pesquisa

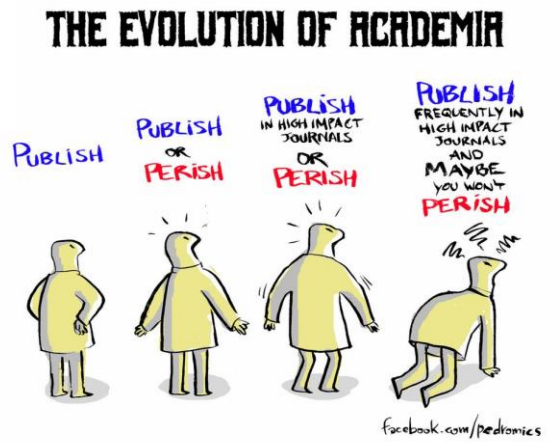
- Os dados de pesquisa são recursos valiosos!
- Geralmente exige muito tempo e dinheiro para serem produzidos
  - Pode levar diversos anos!
- Muitos dados têm um valor significativo além do uso para a pesquisa original
  - Colaborações
  - Surgimento de novas linhas de pesquisa

https://www.openaire.eu/

Por que compartilhar?

"Publish or Perish!"





Por que compartilhar?

- Muitas instituições desejam compartilhar dados de pesquisa
  - *Aumentar o impacto*
  - *Visibilidade de suas pesquisas*
  - *Disseminar o conhecimento*
  - *Estabelecer novas parcerias*

"If we're all in a ship together, and the ship has some holes in it, and we're sort of bailing water out of it, and we have a great design for a bucket, then even if we're bailing out way better than everyone else, we should probably still share the bucket design."

-Elon Musk, 2014

Por que compartilhar?

- Compartilhando os dados de pesquisa:

Incentiva a investigação científica e o debate

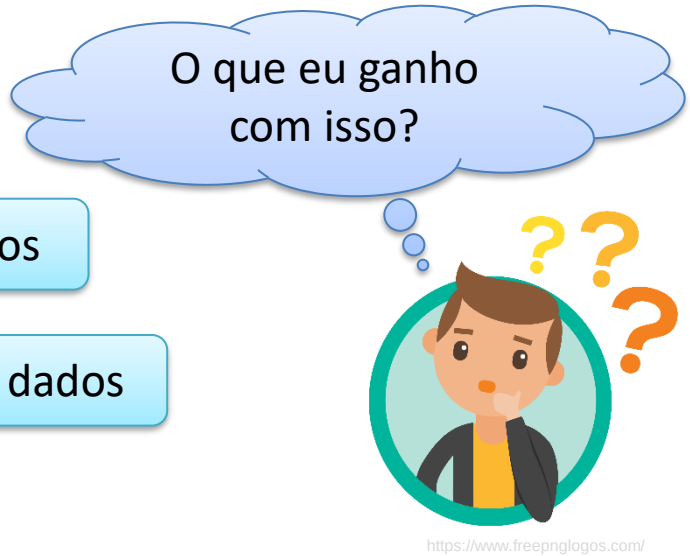
Promove inovação e novos usos potenciais dos dados

Leva a novas colaborações entre usuários e criadores de dados

Maximiza a transparência e responsabilidade

Incentiva a melhoria e validação de métodos de pesquisa

Permite o apuração dos resultados da pesquisa



Por que compartilhar?

- Compartilhando os dados de pesquisa:

Reduz o custo de duplicação da coleta de dados

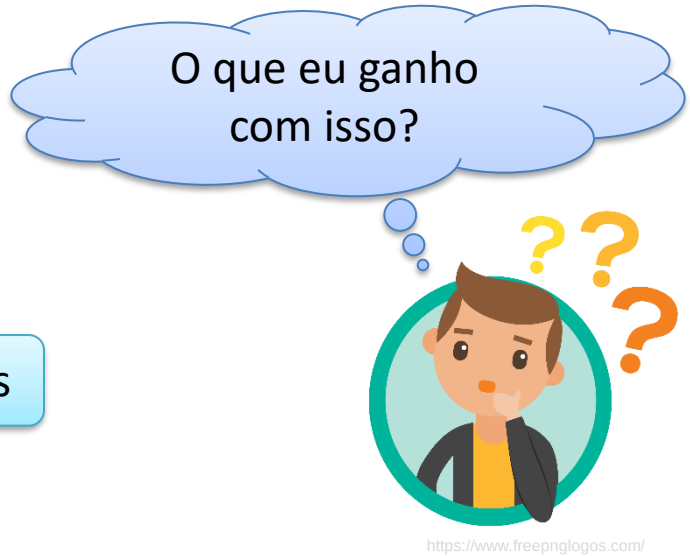
Aumenta o impacto e a visibilidade da pesquisa

Promove a pesquisa que criou os dados e seus resultados

Pode fornecer crédito ao pesquisador

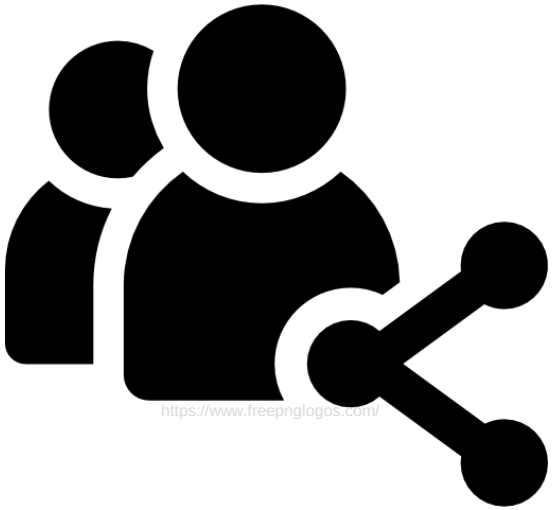
Fornece recursos importantes para educação e treinamento

Aumenta o perfil acadêmico dos pesquisadores



Como compartilhar?

- Existem várias maneiras de compartilhar dados de pesquisa:
  - Colocar em um data center especializado
  - Colocar em um repositório institucional
  - Disponibilizar online por meio de um projeto ou site institucional
  - Disponibilizar informalmente entre pesquisadores
  - ***Escrever artigos científicos***!
    - Submetê-los para uma revista ou conferencia



Conferências e Periódicos

- Exigem cada vez mais que os dados usados nas publicações sejam compartilhados ou disponibilizados em um banco de dados ou repositório acessível
  - Ajuda a tornar seus dados **citáveis**, **rastreáveis** e **localizáveis**!

- Dados de pesquisa e as publicações baseadas nesses dados fazem parte da **produção científica**
  - *Publicar é bastante importante para os pesquisadores!*

Dados Pessoais
Dados Confidenciais
Dados Pessoais Sensíveis
Declaração de Consentimento
Anonimização dos Dados

https://sebraeinteligenciasetorial.com.br/

ÉTICA E CONSENTIMENTO

Ética e Consentimento

- Dois fatores importantes para se compartilhar Dados de Pesquisa:
  - **Consentimento**
  - **Confidencialidade**

https://rockcontent.com/

Ética e Consentimento

- Estratégias para lidar com a confidencialidade dependem da natureza da pesquisa
  - São essencialmente informadas pelas obrigações éticas e legais do pesquisador!

- O dever de confidencialidade para com os informantes pode ser explícito através de um termo assinado pelo participante

Ética e Consentimento

- Tipos de dados:

Dados Pessoais

Dados Confidenciais

Dados Pessoais Sensíveis

https://towardsdatascience.com/

https://www.security.org/

Ética e Consentimento

- **Dados Pessoais**
  - São dados que se referem a um indivíduo
  - Pode ser usado para identificar o indivíduo
    - *Nome, sexo, endereço, telefone, CPF, etc.*
  - Também inclui qualquer expressão de opinião sobre o indivíduo

https://towardsdatascience.com/

Ética e Consentimento

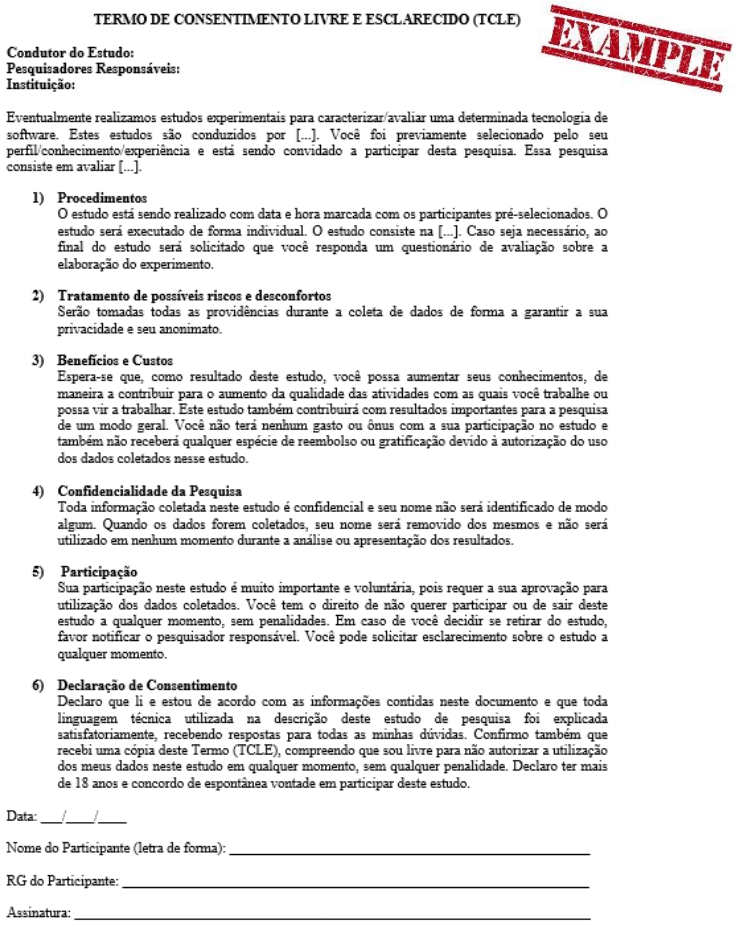
- **Dados Confidenciais**
  - São dados fornecidos em sigilo ou dados concordados em serem mantidos em sigilo
  - Não são de domínio público

Ética e Consentimento

- **Dados Pessoais Confidenciais**
  - Raça
  - Origem étnica
  - Opinião política
  - Crenças religiosas ou semelhantes
  - Filiação a sindicatos
  - Saúde ou condição física ou mental
  - Etc.

Como conseguir o consentimento?

- Uma possibilidade é pelo TCLE
  - ***T**ermo de **C**onsentimento **L**ivre e **E**sclarecido*
- É um documento que explica os procedimentos do experimento que deve ser assinado pelo participante
- Espera-se que os pesquisadores obtenham consentimento para que as pessoas participem da pesquisa e que possa usar as informações coletadas
- Sempre que possível, o consentimento também deve levar em consideração:
  - Qualquer uso futuro de dados
  - Compartilhamento
  - Preservação
  - Uso de longo prazo de dados de pesquisa



TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

EXAMPLE

Condutor do Estudo:
Pesquisadores Responsáveis:
Instituição:

Eventualmente realizamos estudos experimentais para caracterizar/avaliar uma determinada tecnologia de software. Estes estudos são conduzidos por [...]. Você foi previamente selecionado pelo seu perfil/conhecimento/experiência e está sendo convidado a participar desta pesquisa. Essa pesquisa consiste em avaliar [...].

1) **Procedimentos**
O estudo está sendo realizado com data e hora marcada com os participantes pré-selecionados. O estudo será executado de forma individual. O estudo consiste na [...]. Caso seja necessário, ao final do estudo será solicitado que você responda um questionário de avaliação sobre a elaboração do experimento.

2) **Tratamento de possíveis riscos e desconfortos**
Serão tomadas todas as providências durante a coleta de dados de forma a garantir a sua privacidade e seu anonimato.

3) **Benefícios e Custos**
Espera-se que, como resultado deste estudo, você possa aumentar seus conhecimentos, de maneira a contribuir para o aumento da qualidade das atividades com as quais você trabalhe ou possa vir a trabalhar. Este estudo também contribuirá com resultados importantes para a pesquisa de um modo geral. Você não terá nenhum gasto ou ônus com a sua participação no estudo e também não receberá qualquer espécie de reembolso ou gratificação devido à autorização do uso dos dados coletados nesse estudo.

4) **Confidencialidade da Pesquisa**
Toda informação coletada neste estudo é confidencial e seu nome não será identificado de modo algum. Quando os dados forem coletados, seu nome será removido dos mesmos e não será utilizado em nenhum momento durante a análise ou apresentação dos resultados.

5) **Participação**
Sua participação neste estudo é muito importante e voluntária, pois requer a sua aprovação para utilização dos dados coletados. Você tem o direito de não querer participar ou de sair deste estudo a qualquer momento, sem penalidades. Em caso de você decidir se retirar do estudo, favor notificar o pesquisador responsável. Você pode solicitar esclarecimento sobre o estudo a qualquer momento.

6) **Declaração de Consentimento**
Declaro que li e estou de acordo com as informações contidas neste documento e que toda linguagem técnica utilizada na descrição deste estudo de pesquisa foi explicada satisfatoriamente, recebendo respostas para todas as minhas dúvidas. Confirmo também que recebi uma cópia deste Termo (TCLE), compreendo que sou livre para não autorizar a utilização dos meus dados neste estudo em qualquer momento, sem qualquer penalidade. Declaro ter mais de 18 anos e concordo de espontânea vontade em participar deste estudo.

Data: __/___/____

Nome do Participante (letra de forma): _____

RG do Participante: _____

Assinatura: _____

Ética e Consentimento

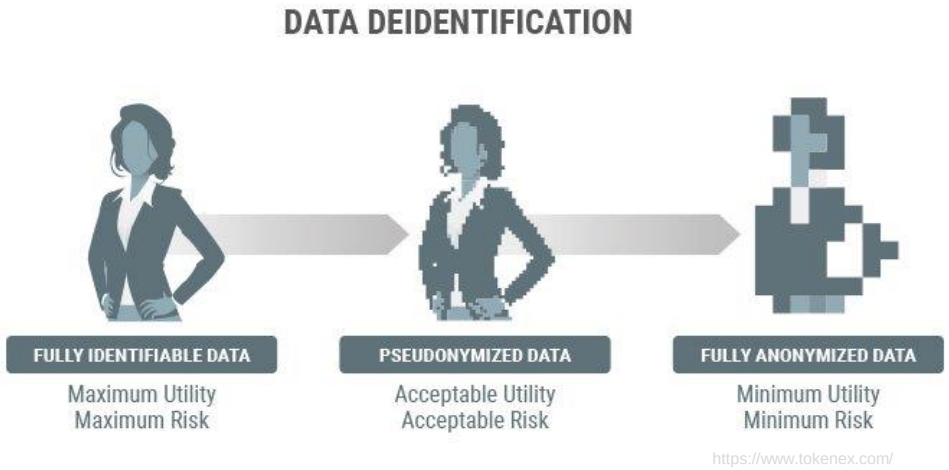
- Os pesquisadores **devem**:
  - Informar aos participantes como os dados da pesquisa serão armazenados, preservados e usados a longo prazo
    - *Questionários, vídeo, áudio, etc.*
  - Informar aos participantes como a confidencialidade será mantida
    - *Anonimizando dados*
  - Obter consentimento informado, por escrito ou verbal, para compartilhamento de dados
    - *Termo de consentimento*
    - *Explicito na gravação, caso seja verbal*

Anonimização dos Dados

- Antes de usar os dados obtidos em pesquisa...
  - **Verificar se precisa torná-los anônimo!**

- Caso afirmativo:
  - Preservar a identidade dos participantes, organizações, empresas, etc.
  - Não ser possível identificar a pessoa a partir dos dados de pesquisa

*De acordo com **London's Global University**, anonimato é o processo de **remoção de identificadores pessoais, diretos** e **indiretos**, que podem levar à identificação de um indivíduo.*



Anonimização dos Dados

- Anonimato pode ser necessário por razões éticas para proteger a identidade das pessoas
  - Também pode ser por razões legais para não divulgar dados pessoais

- Dados pessoais não devem ser divulgados a partir de informações de pesquisa!!!
  - Exceto quando o entrevistado tenha dado consentimento explicito e específico para fazê-lo!

https://medium.com/@PiwikPro/

GO FAIR

Findable Accessible Interoperable Reusable

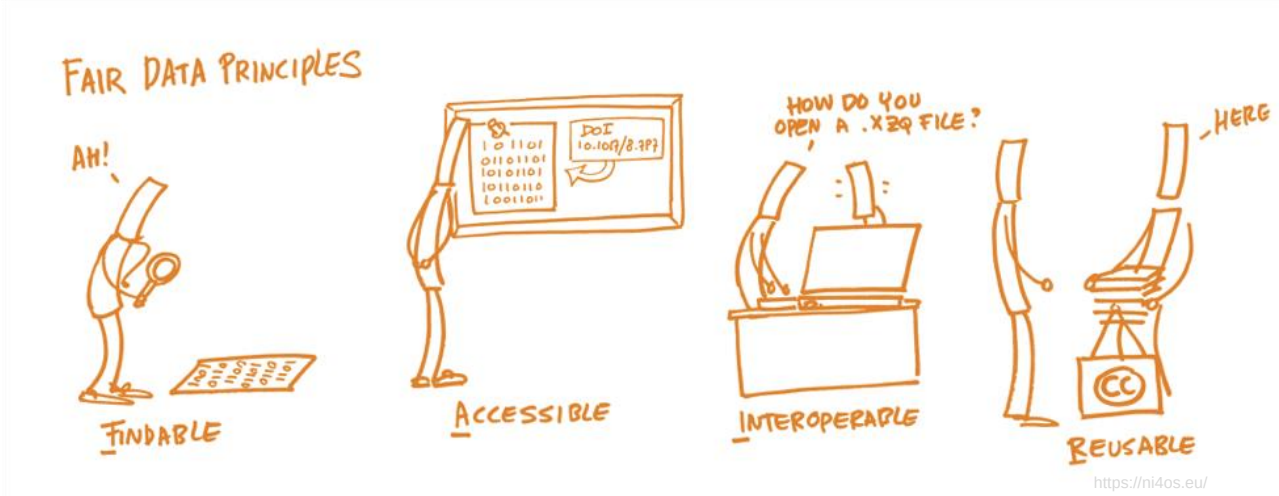
Findability

Accessibility

Interoperability

Reuse

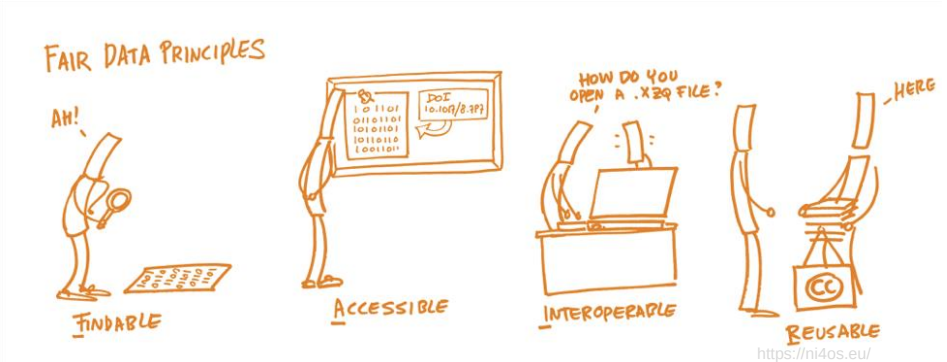
F.A.I.R.



F.A.I.R.

- **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse
  - Criado em 2016

- Fornece diretrizes para melhorar:
  - **Encontrabilidade**
  - **Acessibilidade**
  - **Interoperabilidade**
  - **Reutilização** de ativos digitais



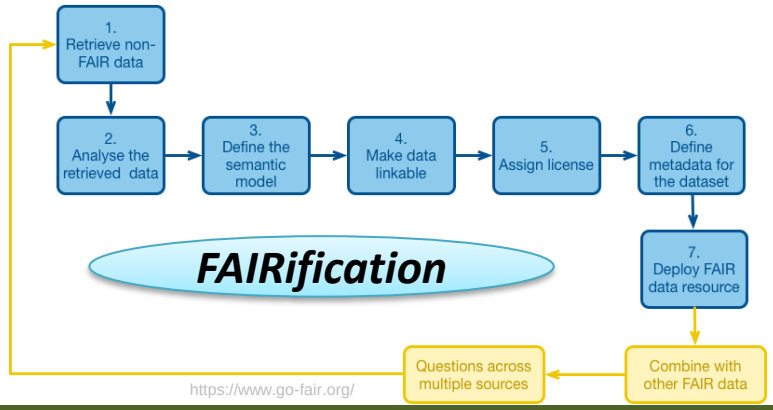
- Os princípios enfatizam a ação da máquina
  - A capacidade dos sistemas computacionais de encontrar, acessar, interoperar e reutilizar dados com nenhuma ou mínima intervenção humana

- Pois nós, humanos, dependemos cada vez mais do suporte computacional para lidar com os dados
  - Alto volume de dados existentes!
  - Complexidade crescente

Encontrabilidade

- O primeiro passo para (re)usar dados é encontrá-los!
- Metadados e dados devem ser fáceis de encontrar para humanos e computadores
- Metadados legíveis por máquina são essenciais para a descoberta automática
- Componente essencial do processo de ***FAIRification***

https://www.interaction-design.org/



https://www.go-fair.org/

Encontrabilidade

- **F1:** (Meta) dados são atribuídos a um identificador globalmente único e persistente
  - *Ex.: Orcid, DOI*
- **F2:** Os dados são descritos com metadados ricos (*definidos no R1*)
  - *Informações descritivas sobre o contexto, qualidade e condição ou características dos dados*
- **F3:** Metadados incluem claramente e explicitamente o identificador dos dados que descrevem
  - *Metadados e o conjunto de dados que eles descrevem geralmente são arquivos separados...*
- **F4:** (Meta) dados são registrados ou indexados em um recurso pesquisável
  - *Ex.: Google conseguir indexar as páginas para aparecer nas buscas*

Acessibilidade

- Não basta apenas encontrar os dados
  - Precisa saber como eles podem ser acessados!

- Possivelmente inclui também a autenticação e autorização

https://www.concordcarlisle.org/

Acessibilidade

- **A1:** (Meta) dados são recuperáveis por seu identificador usando um protocolo de comunicação padronizado
  - *A recuperação de dados deve ser mediada sem ferramentas especializadas ou proprietárias ou métodos de comunicação*
  - *A maioria dos produtores de dados usa http(s) ou ftp*
- **A1.1:** O protocolo é aberto, gratuito e universalmente implementável
  - *Este critério impacta na escolha do repositório onde irá compartilhar os dados*
- **A1.2:** O protocolo permite um procedimento de autenticação e autorização, quando necessário
  - *Muitas vezes faz sentido solicitar aos usuários que criem uma conta de usuário para um repositório para definir direitos específicos do usuário*
- **A2:** Os metadados são acessíveis, mesmo quando os dados não estão mais disponíveis
  - *Datasets tendem a degradar ou desaparecer com o tempo*
  - *Relacionado aos problema de registro e indexação descritos em F4*
  - *Útil para rastrear pessoas, instituições ou publicações associadas à pesquisa original*

Interoperabilidade

- Dados geralmente precisam ser integrados a outros dados

- Dados as vezes também precisam interoperar com aplicativos ou fluxos de trabalho para análise, armazenamento e processamento

Interoperabilidade

- **I1:** Os (meta) dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento
  - *Humanos devem ser capazes de trocar e interpretar os dados uns dos outros*
  - *Dados também devem ser legíveis para máquinas, sem a necessidade de algoritmos, tradutores ou mapeamentos especializados ou ad hoc*
  - *Ex.: JSON, RDF*

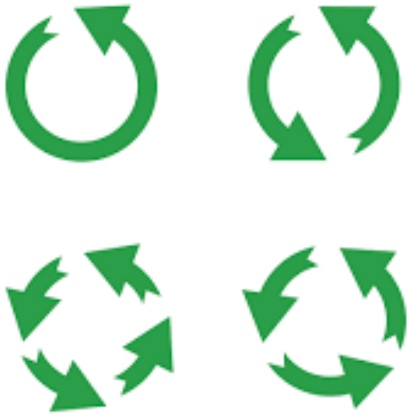
- **I2:** (Meta) dados usam vocabulários que seguem os princípios FAIR
  - *Vocabulário usado precisa ser documentado e resolvido usando identificadores globais únicos e persistentes*
  - *Documentação precisa ser facilmente encontrada e acessível por qualquer pessoa que use o conjunto de dados*
  - *Ex.: FAIR Data Point*

- **I3:** (Meta) dados incluem referências qualificadas a outros (meta) dados
  - Especificar se um conjunto de dados se baseia em outro conjunto de dados
  - Se conjuntos de dados adicionais são necessários para completar os dados
  - Se informações complementares são armazenadas em um conjunto de dados diferente

Reuso

- O objetivo final do FAIR é **otimizar a reutilização de dados**

- Metadados e dados devem ser bem descritos!
  - Assim eles podem ser replicados e/ou combinados em diferentes configurações



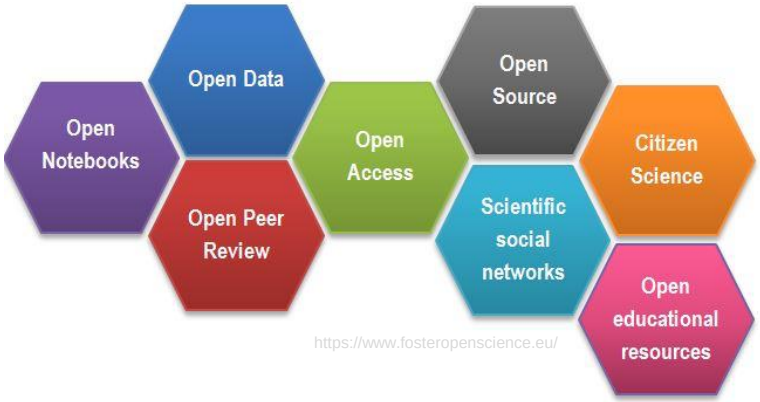
Reuso

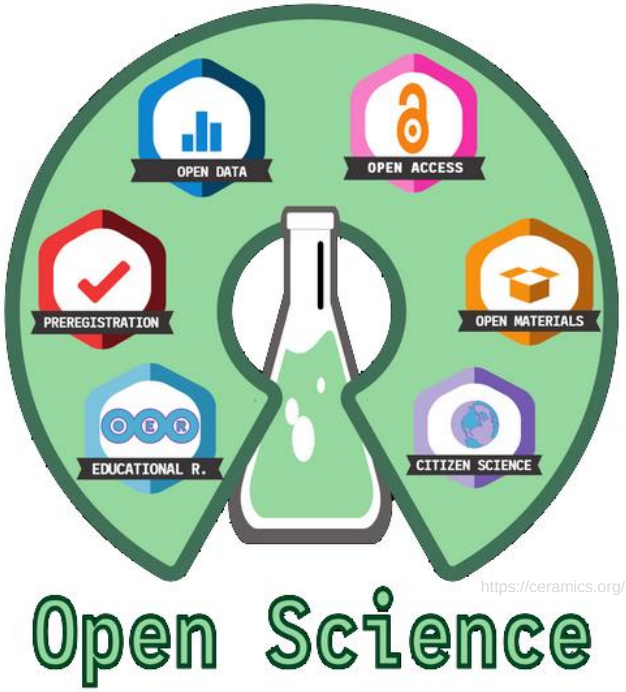
- **R1:** Meta (dados) são ricamente descritos com uma pluralidade de atributos precisos e relevantes
  - *Está relacionado ao princípio F2, mas se concentra na capacidade de um usuário decidir se os dados são realmente úteis em um determinado contexto*
  - *O autor dos metadados deve ser o mais generoso possível, incluindo informações que podem parecer irrelevantes*
  - *Ex.: Descrever o escopo de seus dados, mencionar quaisquer particularidades ou limitações, especificar a data de geração/coleta dos dados*

- **R1.1:** (Meta) dados são liberados com uma licença de uso de dados clara e acessível
  - *Quais direitos de uso você atribui aos seus dados?*
  - *As condições sob as quais os dados podem ser usados devem ser claras para máquinas e humanos*
  - *Ex.: Licença MIT*

- **R1.2 :** (Meta) dados são associados à proveniência detalhada
  - *Quem o gerou ou coletou? Como foi processado? Já foi publicado antes? Ele contém dados de outra pessoa que você pode ter transformado ou completado?*
  - *Incluir uma descrição do fluxo de trabalho que levou aos seus dados*

- **R1.3 :** (Meta) dados atendem aos padrões da comunidade relevantes ao domínio
  - *Seguir os padrões estabelecidos pela comunidade para arquivamento e compartilhamento*
  - *Se divergir, justificar no metadado*

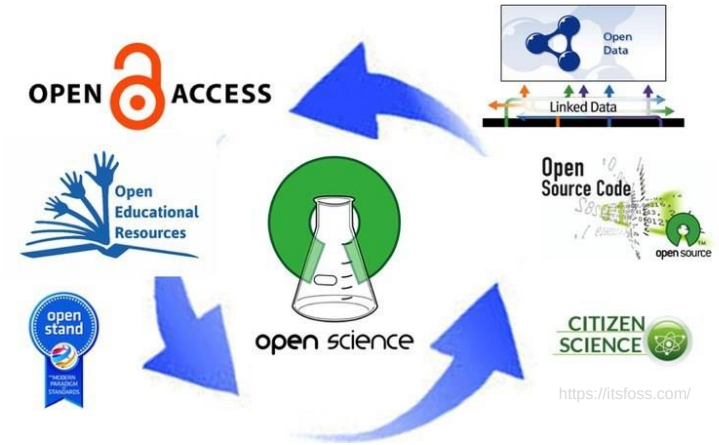




Ciência aberta x Ciência Fechada

Como seguir na direção de ciência aberta?

CIÊNCIA ABERTA



Qual caminho seguir?

Ciência aberta

Ciência fechada

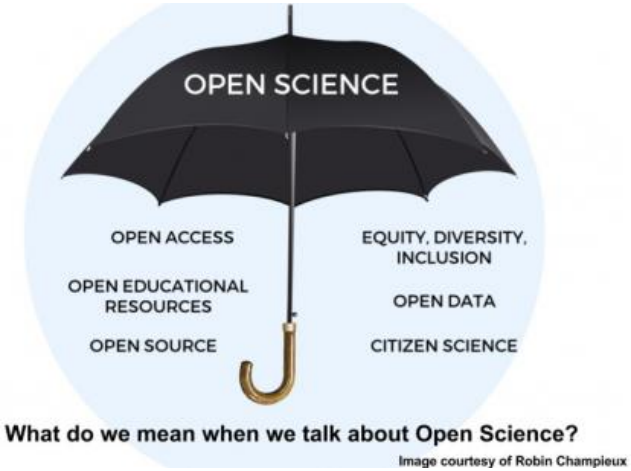
Qual caminho seguir?

Ciência Aberta

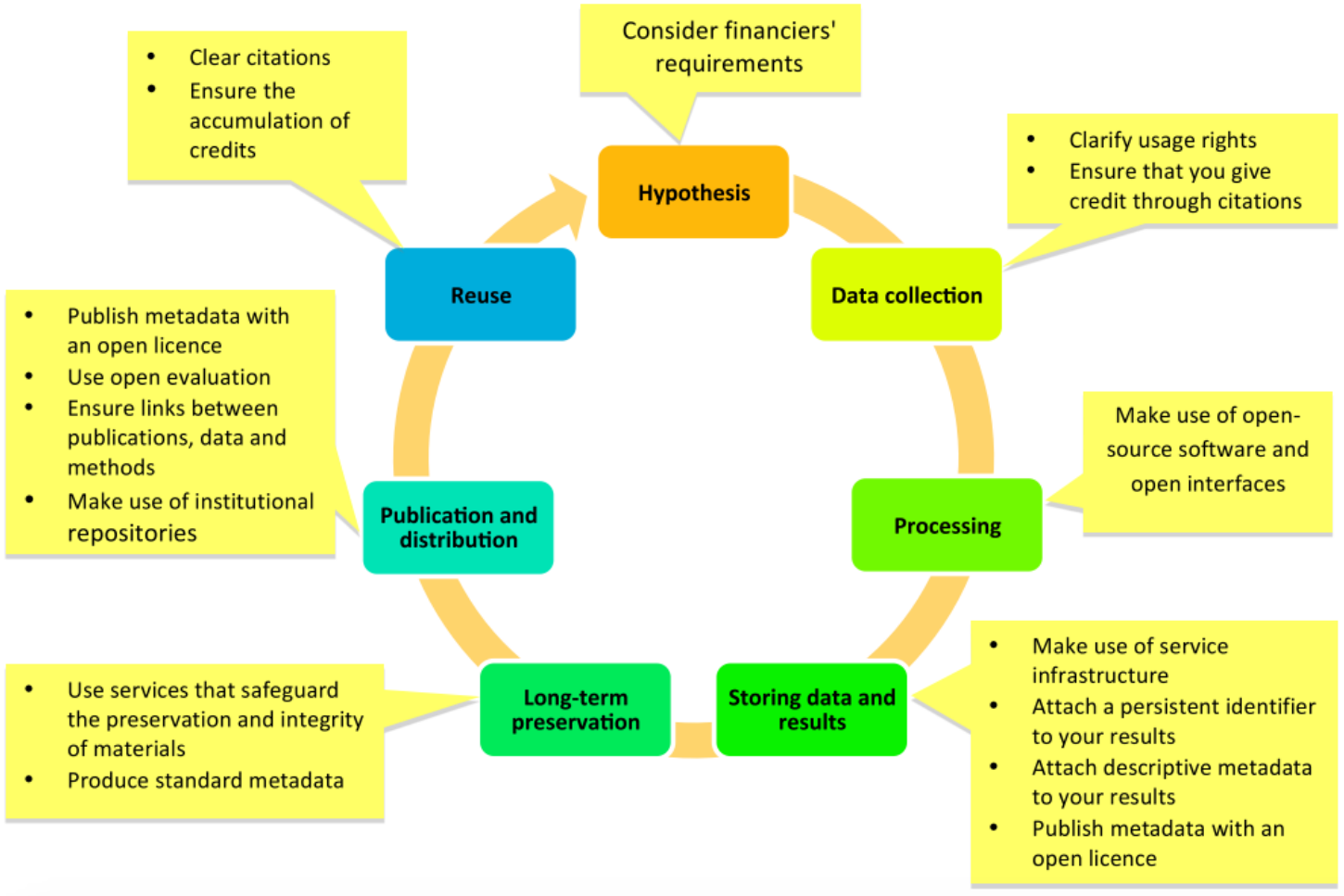
- Disponibilizar conhecimento
- Universidades
- Artigo cientifico
- Mestrado e doutorado

Ciência Fechada

- Pesquisa fechada
- Empresa privada
  - Investimento privado
- Patentes

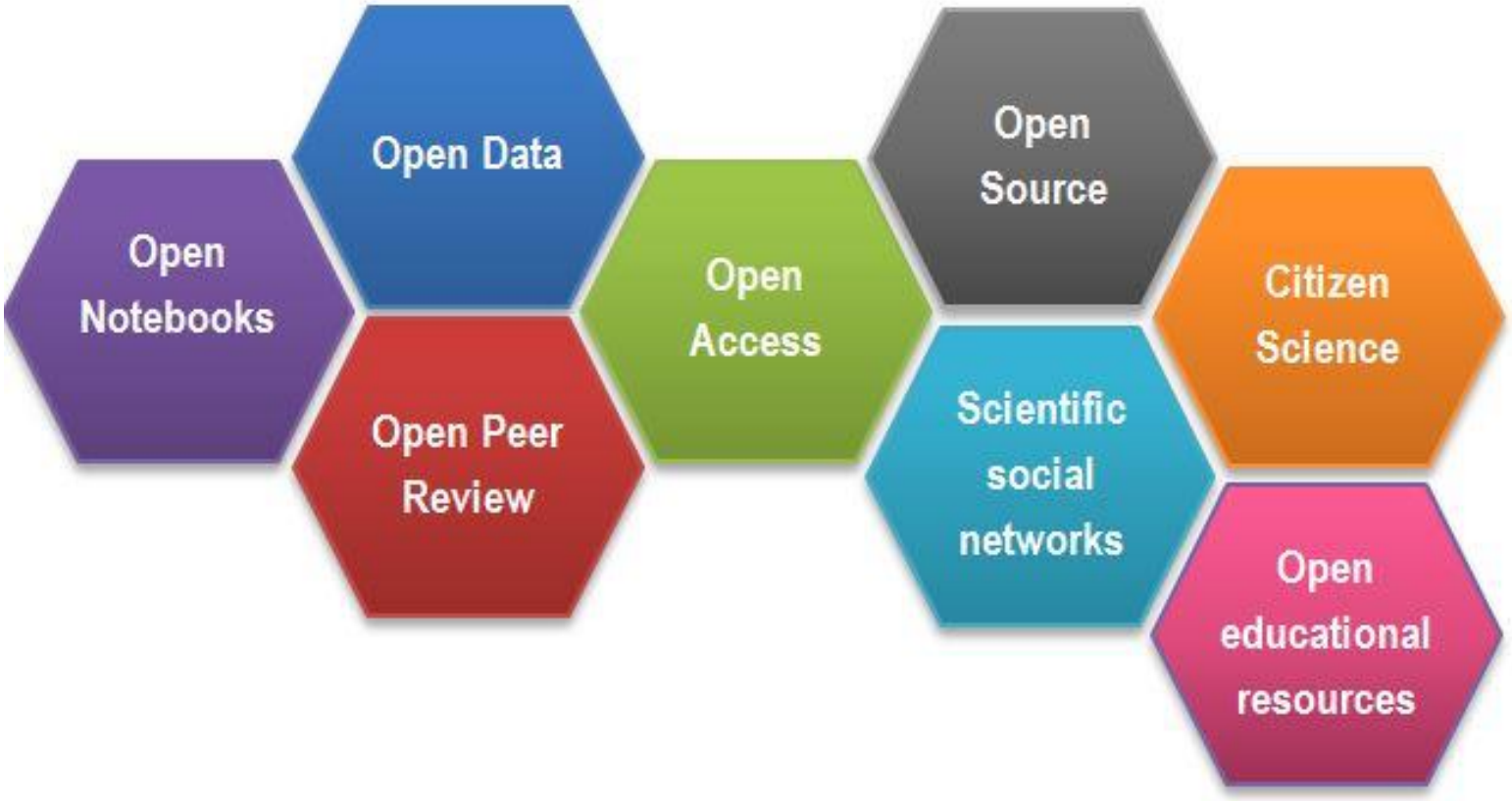


Ciência aberta...



https://www.fosteropenscience.eu/content/what-open-science-introduction

Ciência aberta...



https://www.fosteropenscience.eu/content/what-open-science-introduction

Como seguir na direção de ciência aberta?

- Adotar um serviço de hospedagem de código
  - Ex. GitHub
- Escolher uma licença apropriada
  - Ex. MIT
- Disponibilizar todos os materiais
  - Ex. código, dados, protocolos, exemplos, etc.
- Documentar adequadamente o projeto
- Facilitar o processo de uso/instalação
- Acolher potenciais usuários
- Publicar!

Licença MIT

Copyright <YEAR> <COPYRIGHT HOLDER>
Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Como seguir na direção de ciência aberta?

- Adotar um serviço de hospedagem de código
  - Ex. GitHub
- Escolher uma licença apropriada
  - Ex. MIT
- Disponibilizar todos os materiais
  - Ex. código, dados, protocolos, exemplos, etc.
- Documentar adequadamente o projeto
- Facilitar o processo de uso/instalação
- Acolher potenciais usuários
- Publicar!

Por exemplo...

Grupo de Evolução e Manutenção de Software (GEMS)

Universidade Federal Fluminense (UFF) http://gems.ic.uff.br/

Repositories 36 | Packages | People 1 | Projects

Find a repository... | Type: All | Language: All

merge-nature

Companion website for the paper "On the Nature of Software Merge Conflicts"

Java 0 0 0 0 Updated 17 days ago

noworkflow

Supporting infrastructure to run scientific experiments without a scientific workflow management system.

Jupyter Notebook MIT 19 87 48 1 Updated 19 days ago

Top languages

Java Python Jupyter Notebook JavaScript Ruby

People 1 >

PROVENIÊNCIA

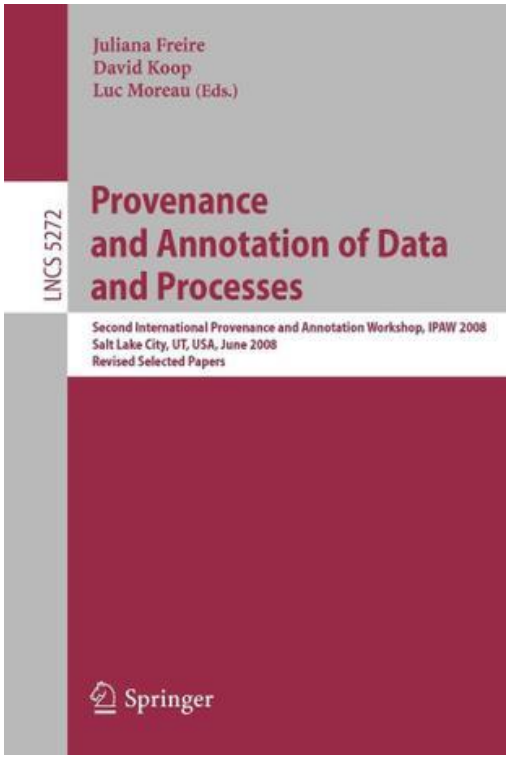
Instituto de Computação

Prov·e·nance

noun

The place of origin or earliest known history of something.
A record of ownership of a work of art or an antique, used as a guide to authenticity or quality.

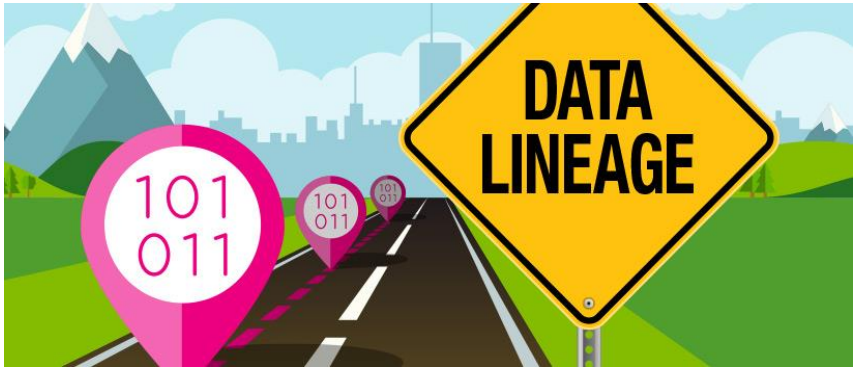
Latin pro- (forth) + Latin venire (come) → Latin provenire (come) → French provenir (come or stem from) → provenance



PROVENIÊNCIA DE DADOS

Proveniência de Dados

- Também conhecida como "*linhagem de dados*"
- São metadados emparelhados com registros que detalham a origem, alterações e detalhes que suportam a confiança ou validade dos dados
- É importante para rastrear erros nos dados e atribuí-los às fontes
- Útil em relatórios e auditoria para processos de negócios e de pesquisa



Proveniência de Dados

- Simplificando...
- Ajuda a responder a perguntas como:
  - “por que os dados foram produzidos?”
  - “como os dados foram produzidos?”
  - “onde os dados foram produzidos?”
  - “quando os dados foram produzidos?”
  - “por quem os dados foram produzidos?”

Mas como surgiu?

- Proveniência é a cronologia da propriedade, custódia ou localização de um objeto histórico
  - Origem é do francês, **provenir**, 'vir de / para trás'
- Surgiu em 1780

provenir

[pʀɔv(ə)niʀ]

Full verb table **INTRANSITIVE VERB**

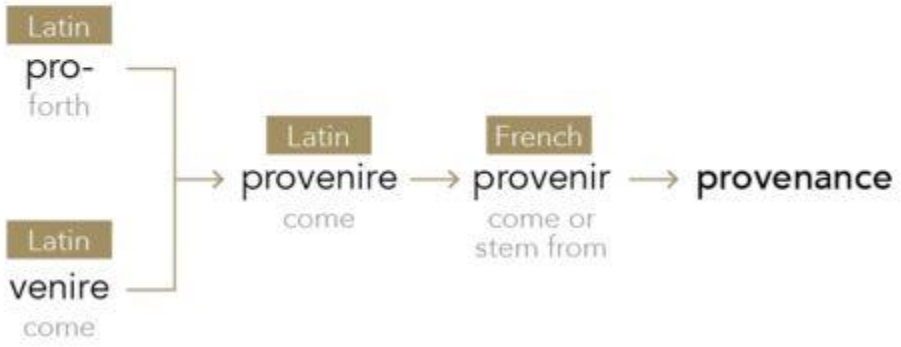
Word Frequency

1. provenir de
2. (= *être originaire de*) to come from
   Ces tomates proviennent d'Espagne. These tomatoes come from Spain.
3. (= *résulter de*) to be due to ♦ to be the result of
   Cela provient d'un manque d'organisation. This is due to a lack of organization. ♦ This is the result of a lack of organization.

Prov·e·nance

noun

The place of origin or earliest known history of something. A record of ownership of a work of art or an antique, used as a guide to authenticity or quality.



Mas como surgiu?

- Termo originalmente usado em obras de arte
  - Hoje é usado em sentidos semelhantes em uma ampla gama de campos
  - Arqueologia, paleontologia, arquivos, manuscritos, livros impressos, ciência, computação...

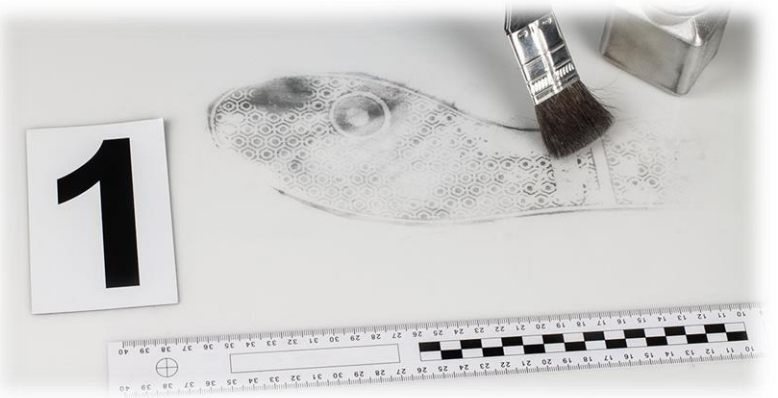
Propósito

- Fornecer evidências contextuais e circunstanciais para a produção ou descoberta original do objeto
  - Estabelecer, na medida do possível, sua história posterior
  - Sequências de sua propriedade formal, custódia e locais de armazenamento

Propósito

- A prática tem um valor particular para ajudar a autenticar objetos!!!
  - Também podem ser utilizadas técnicas comparativas, opiniões de especialistas e resultados de testes científicos para esses fins...
  - Mas o estabelecimento da proveniência é essencialmente uma questão de documentação
- A proveniência é conceitualmente comparável ao termo legal **cadeia de custódia**

Art. 158*-A* do Código de Processo Penal: *Considera-se cadeia de custódia o conjunto de todos os procedimentos utilizados para manter e documentar a história cronológica do vestígio coletado em locais ou em vítimas de crimes, para rastrear sua posse e manuseio a partir de seu reconhecimento até o descarte.*



Propósito

- Museus e o comércio de arte
- Estabelecer a autoria e a autenticidade de um objeto
- Estabelecer a validade moral e jurídica de uma cadeia de custódia
  - Quantidade crescente de arte saqueada!

Propósito

- Essas questões se tornaram uma grande preocupação em relação às obras que mudaram de mãos em áreas controladas durante as guerras e invasões
- Muitos museus começaram a compilar registros proativos dessas obras e de sua história!

E na Informática?

- Dentro da informática, o termo "proveniência" também significa a **linhagem dos dados**
- É recente... 2006!
  - Workshop Internacional de Anotação de Proveniência (IPAW)
  - Teoria e Prática de Proveniência (TaPP)

- Modelo conceitual de causalidade
  - Relação dos *processos* que atuam sobre *dados* e *agentes* que são responsáveis por esses processos

Desafios de Proveniência (*Provenance Challenges*)

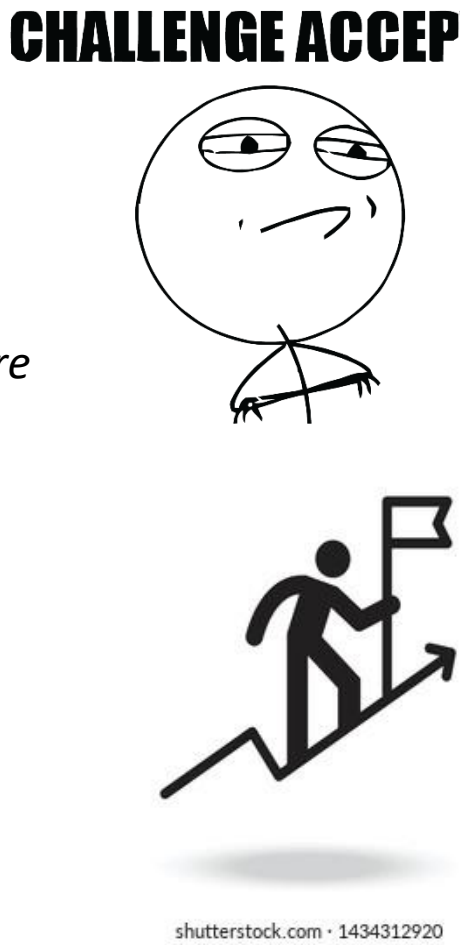
- Proveniência é um conceito crítico em workflows científicos, permite cientistas entender:
  - Origem de seus resultados
  - Repetir seus experimentos
  - Validar os processos que foram usados para derivar os dados

- Discussão sobre padronização de proveniência no IPAW'06
  - Comunidade decidiu que precisa entender as diferentes representações usadas para proveniência, seus aspectos comuns e as razões para suas diferenças

- "*Desafios de Proveniência*" foram criados para comparar e compreender as abordagens existentes

Desafios de Proveniência (*Provenance Challenges*)

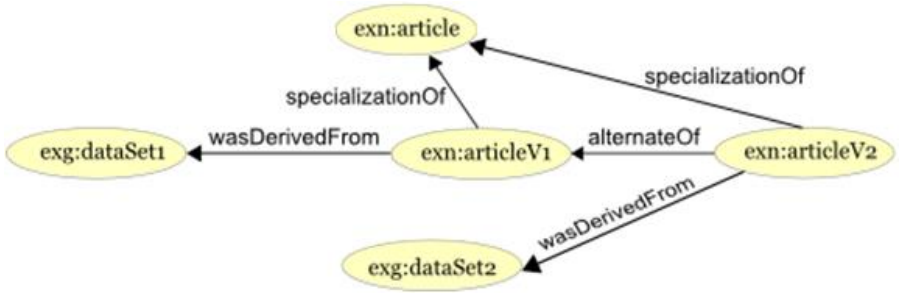
- ***2006 – 2010***
  - ***1º Desafio***: *entender as capacidades dos diferentes sistemas de proveniência e expressar suas representações de proveniência (2006)*
  - ***2º Desafio***: *estabelecimento de interoperabilidade entre os sistemas, por meio de troca de informações de proveniência (2007)*
  - ***3º Desafio***: *avaliar o modelo OPM em um ambiente prático (2009)*
  - ***4º Desafio***: *interrompido pelo lançamento de um novo modelo de proveniência (2010)*



Proveniência

"Refere-se à documentação histórica de um objeto, ou sua trajetória de vida"

- Fornece um fundamento essencial para avaliar a autenticidade de dados
  - Permite confiabilidade e reprodutibilidade
- Responde às perguntas "*por que*" e "*como os dados foram produzidos*", "*onde*", "*quando*" e "*por quem*"
- Modelos:
  - OPM (2007)
  - PROV (2010)
- Grafo de Proveniência
  - Grafo de Causalidade



Proveniência de Dados

- A definição deixa claro que toda a ideia de proveniência diz respeito à **confiança**, **credibilidade** e **reprodutibilidade**

- Em pesquisas intensivas de dados, os usuários de dados provavelmente não são os produtores de dados

Proveniência de Dados

- Os produtores de dados podem:
  - Configurar um instrumento ou simulação de uma determinada maneira para coletar dados primários
  - Aplicar certas metodologias e processos para extrair, transformar e analisar dados de entrada para produzir um produto de dados de saída

- Os metadados de proveniência é importante para determinar a **qualidade**, o **grau de confiança** que se pode depositar nos resultados, a **reprodutibilidade** dos resultados e a **reutilização** dos dados

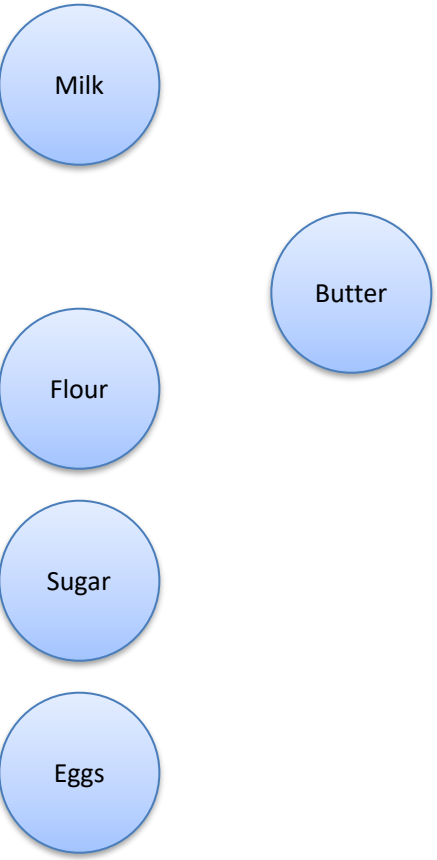
Proveniência de Dados

- Para os usuários de dados:
  - Base científica de sua análise e a responsabilidade de sua pesquisa dependem da **credibilidade** e **confiabilidade** de seus dados
  - Verificar a **qualidade** dos dados junto com o nível esperado de imprecisão

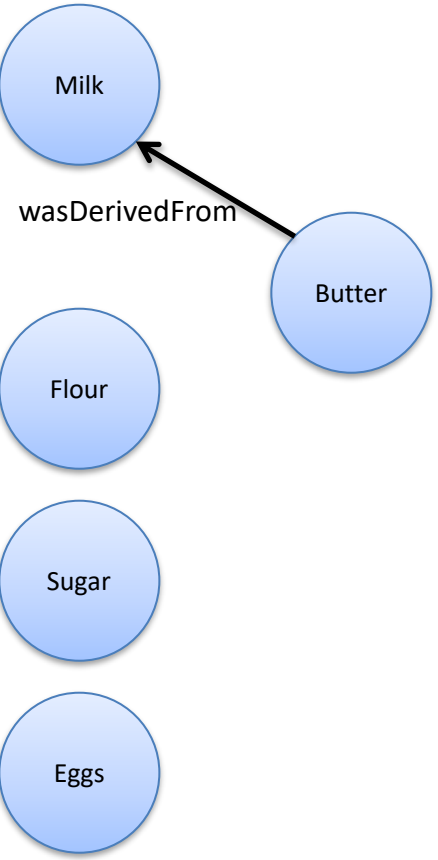
Tá, mas vamos para um exemplo simples!

- Proveniência de dados...
  - Para fazer um bolo com cobertura!
- O que temos?
  - Ingredientes
  - Utensílios de cozinha
  - Fogão
  - O bolo!
- O que precisamos fazer?
  - As operações...
  - Mexer
  - Assar
  - Decorar com calda
- Quem faz?
  - O cozinheiro

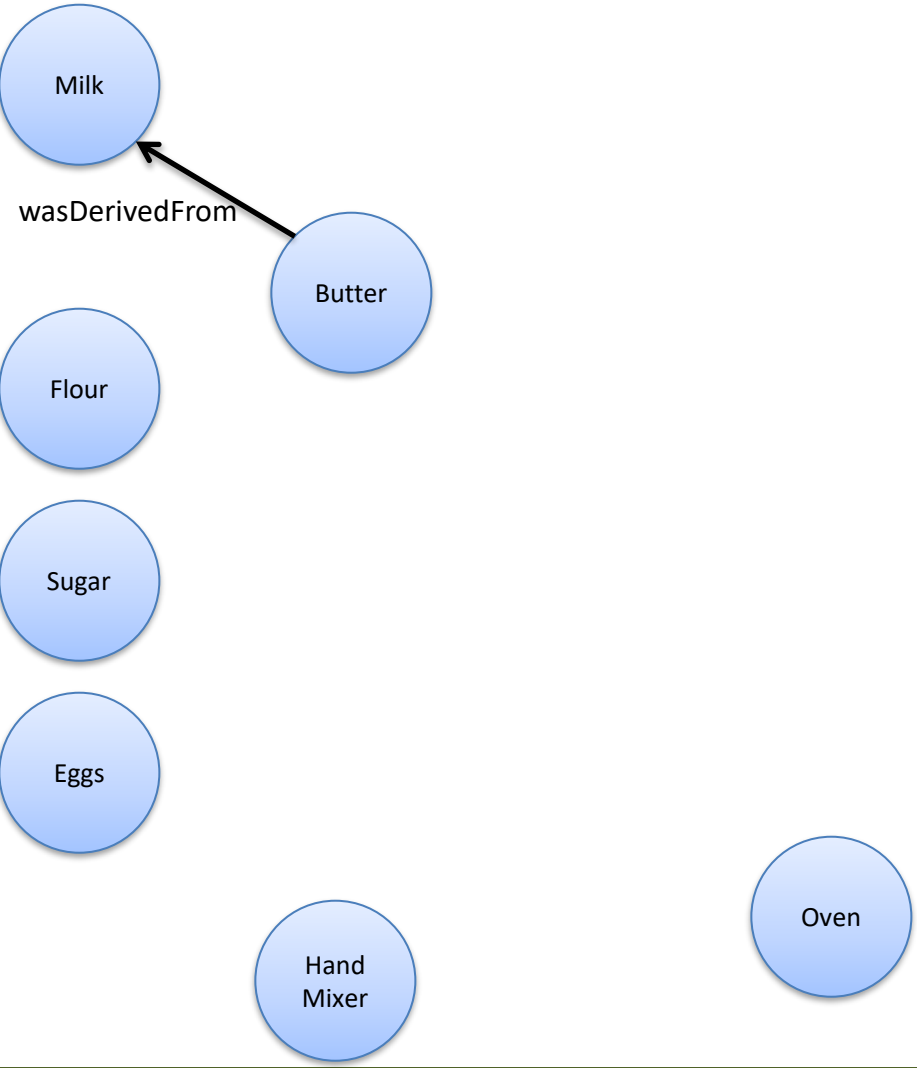
Proveniência



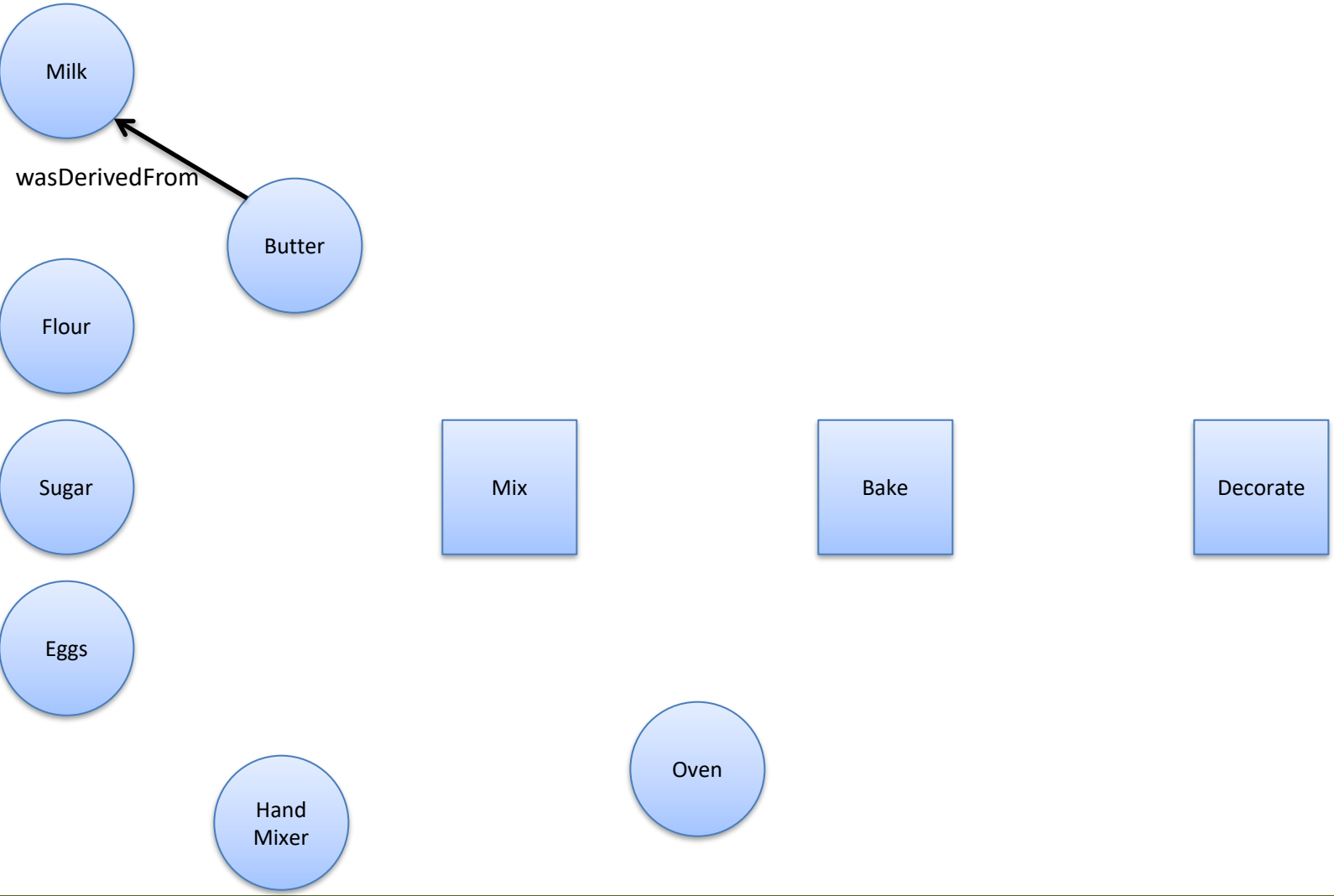
Proveniência



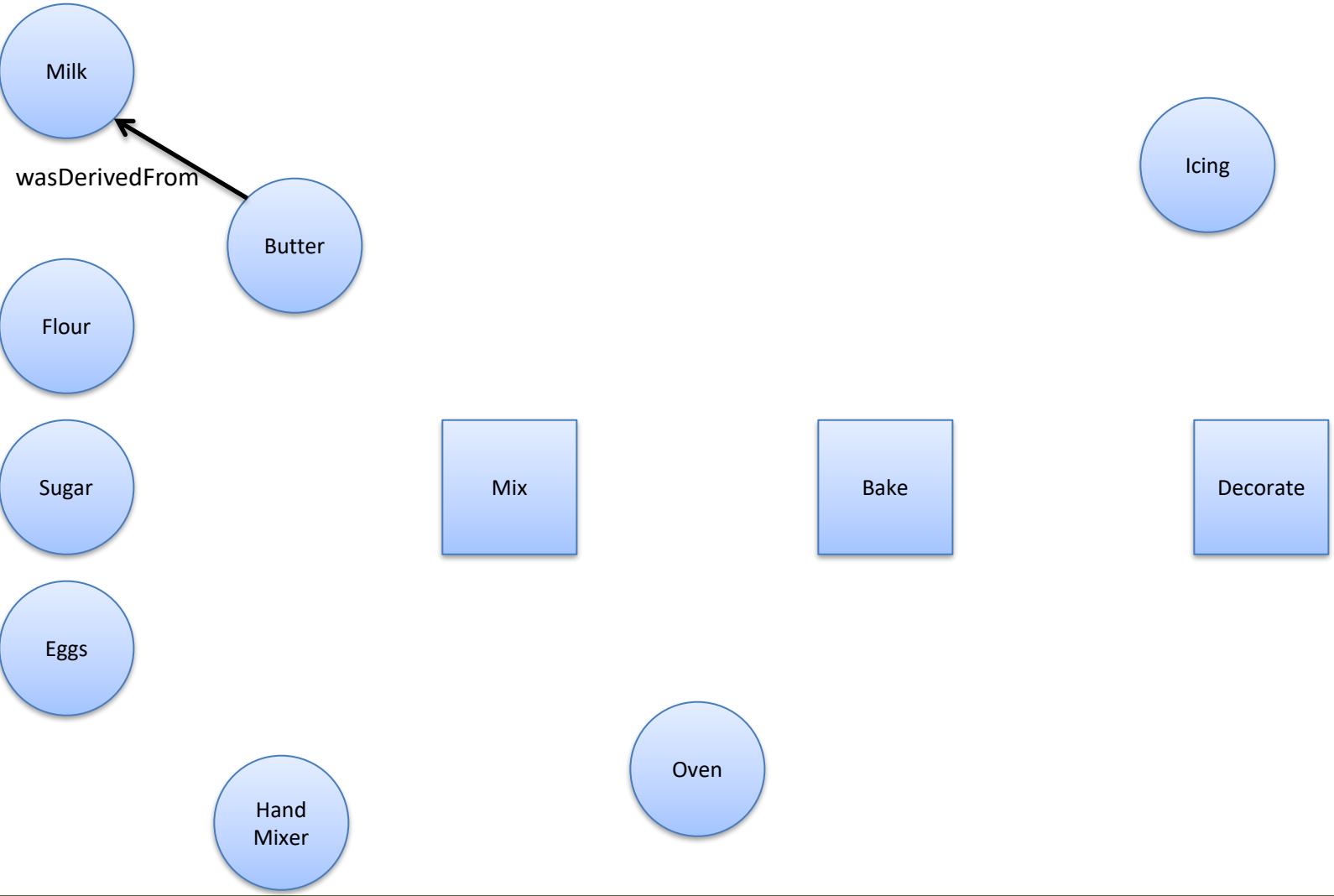
Proveniência



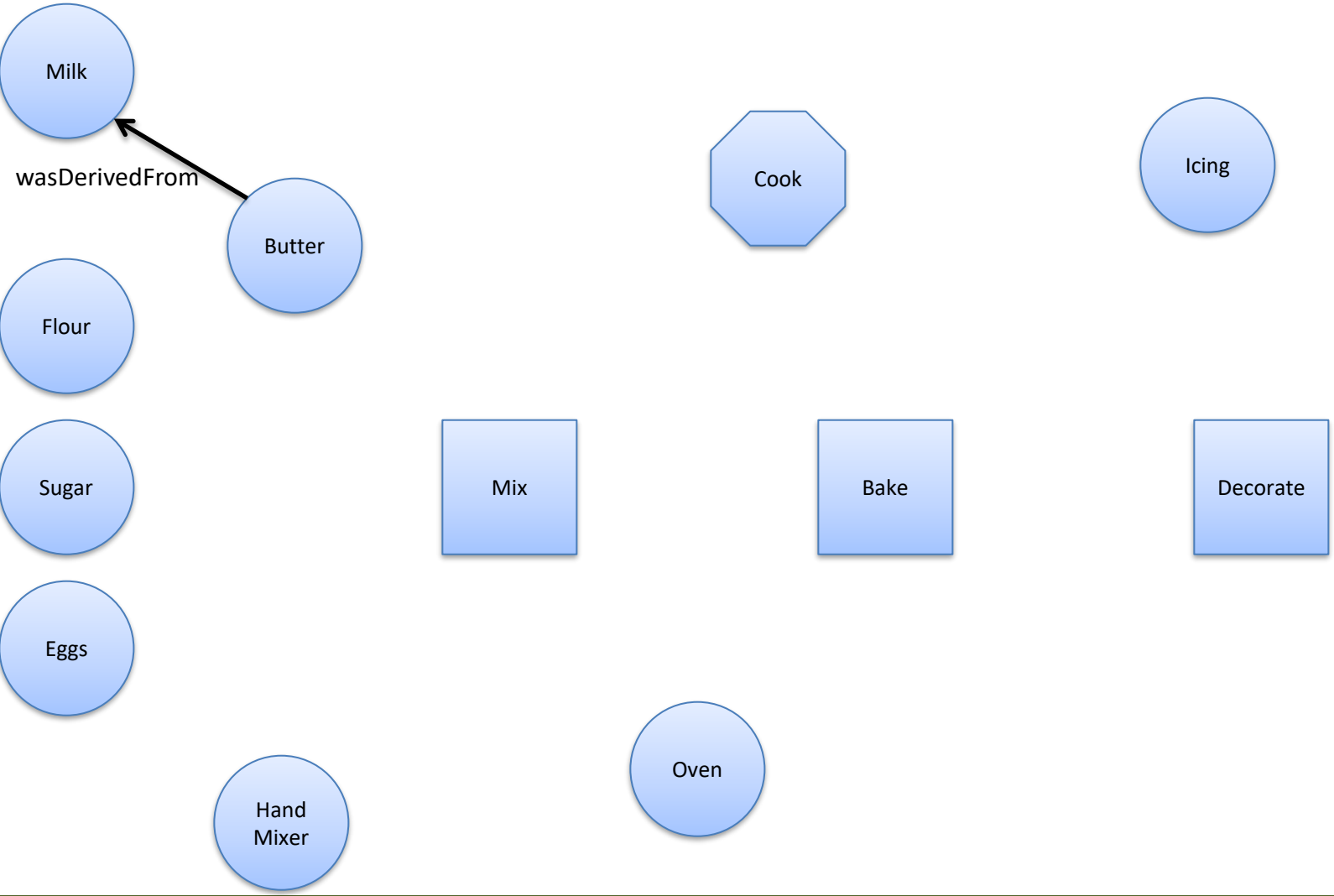
Proveniência



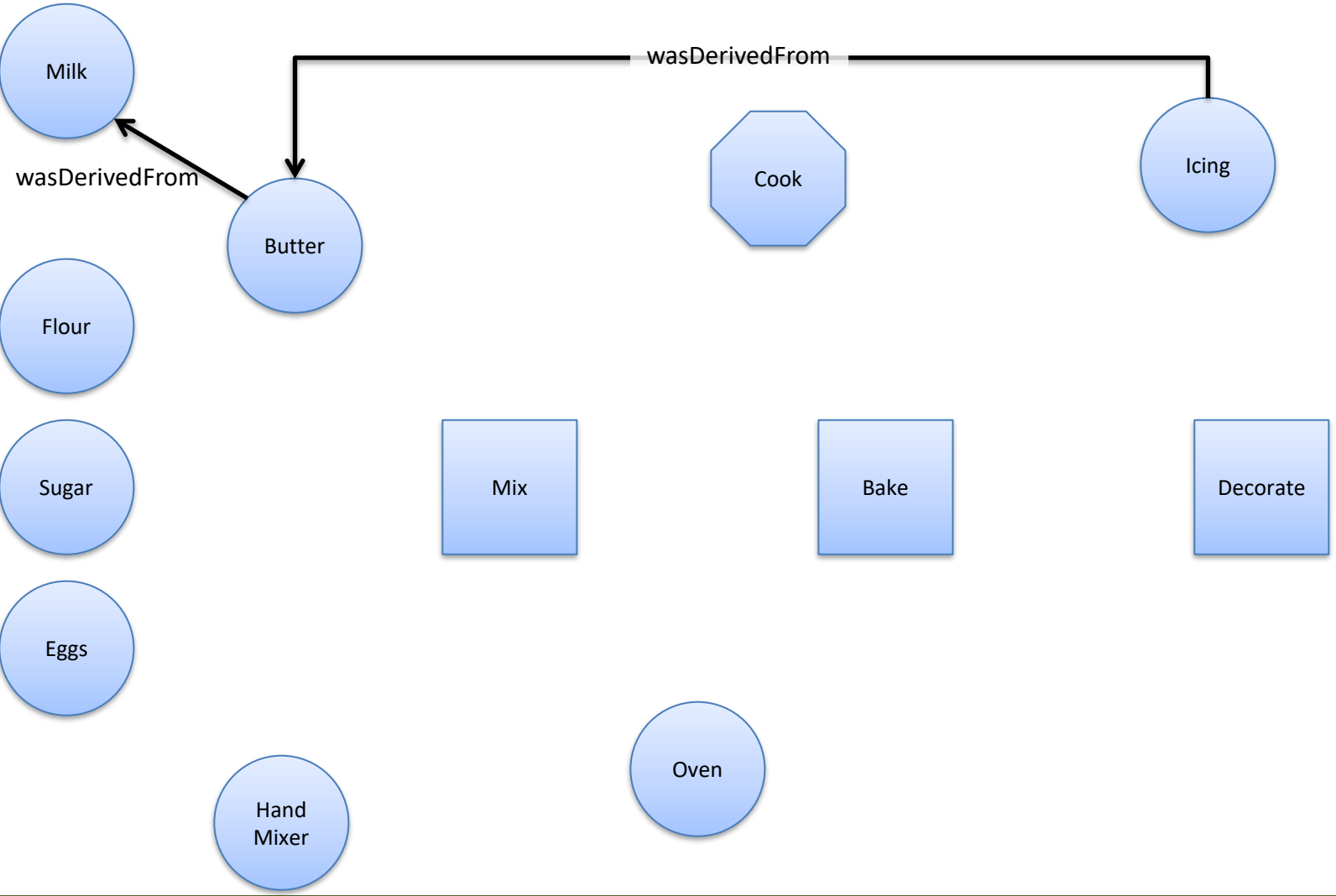
Proveniência



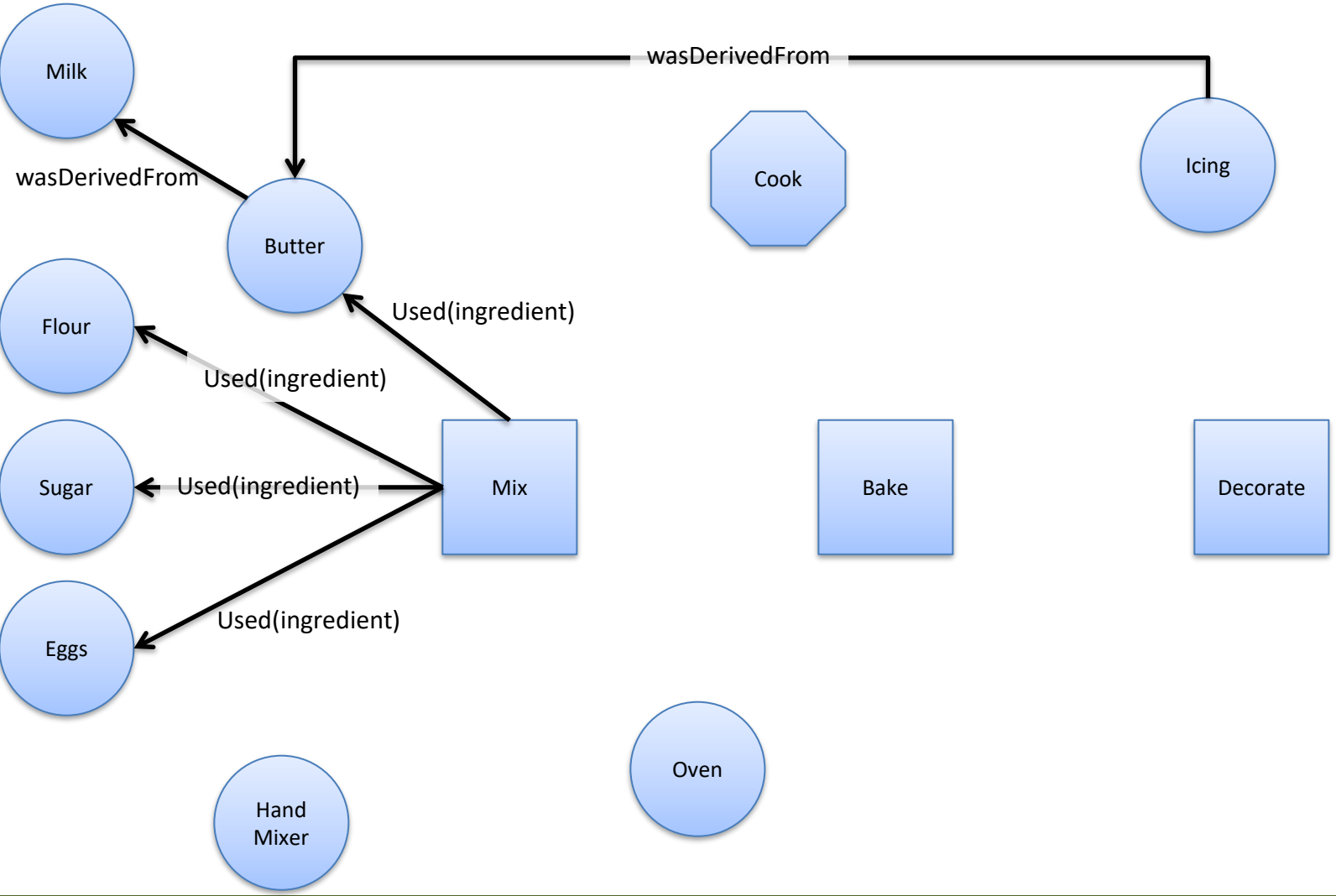
Proveniência



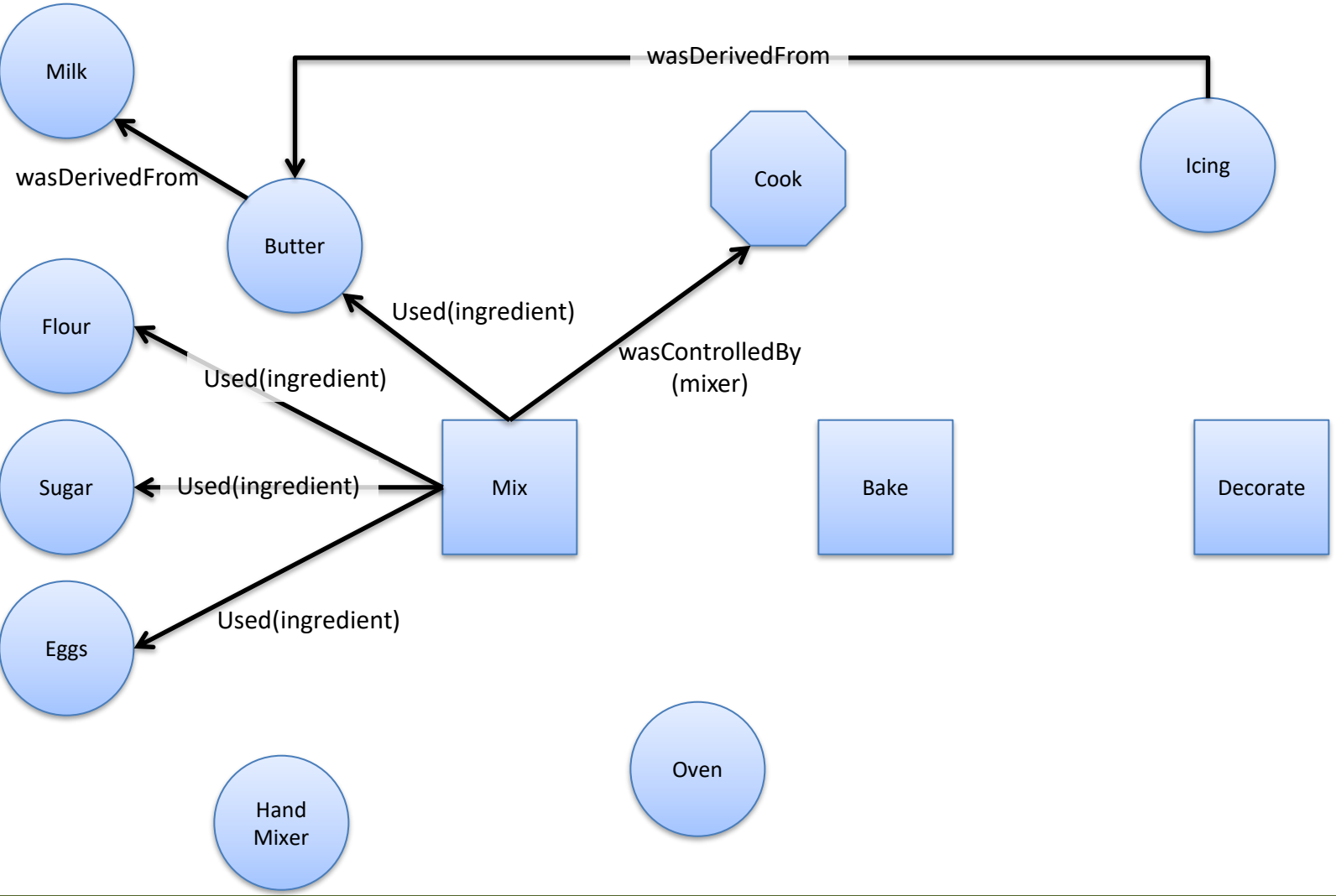
Proveniência



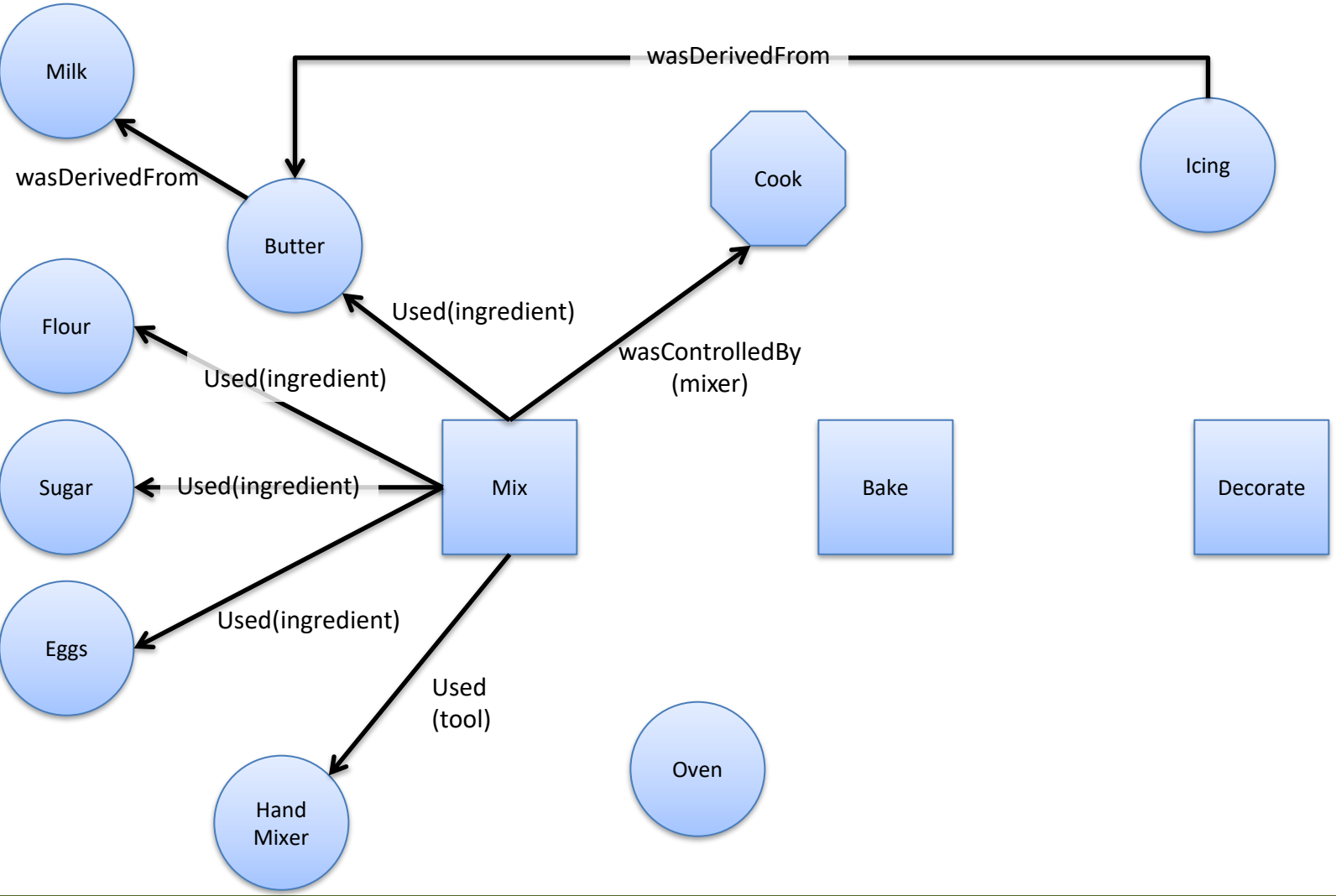
Proveniência



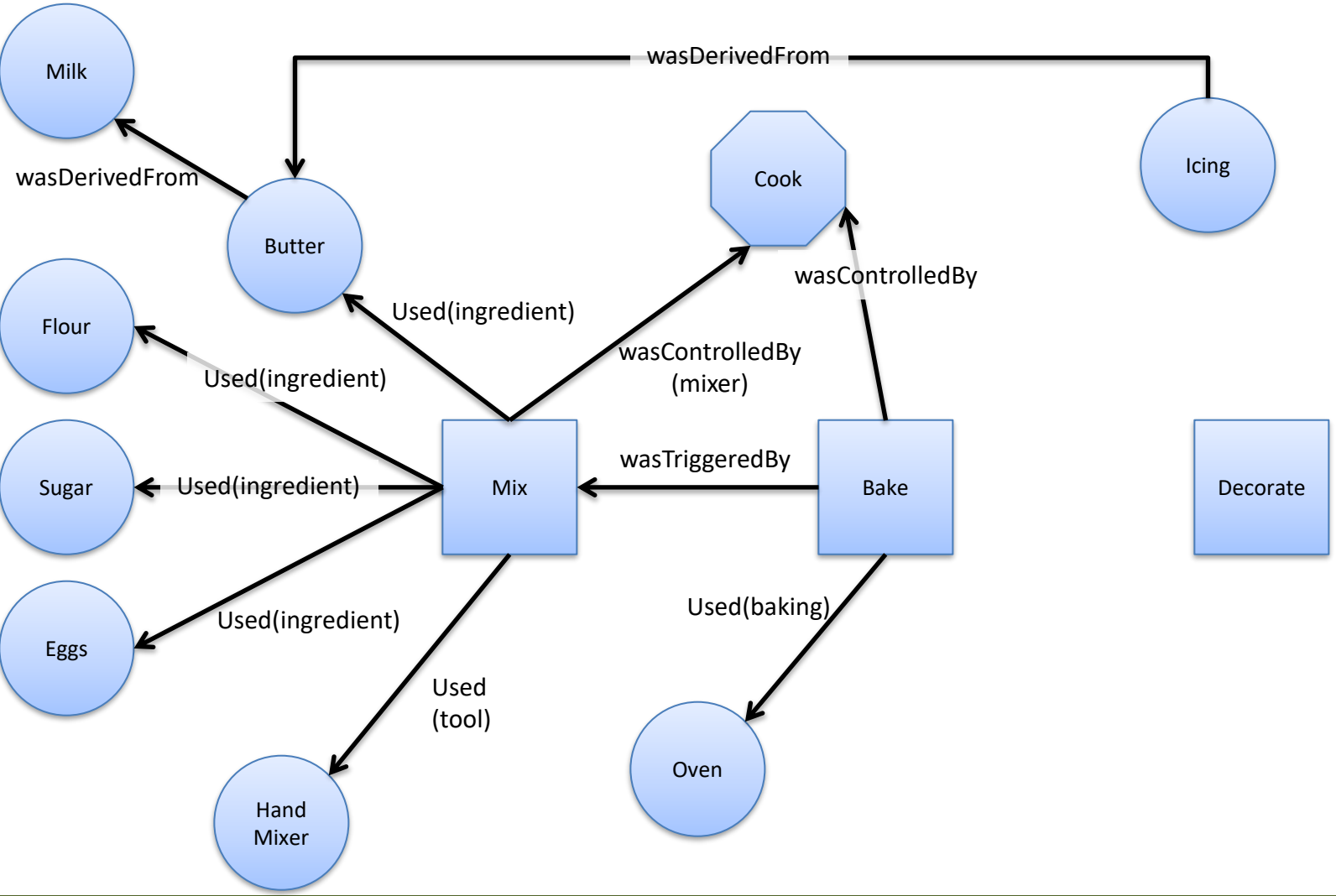
Proveniência



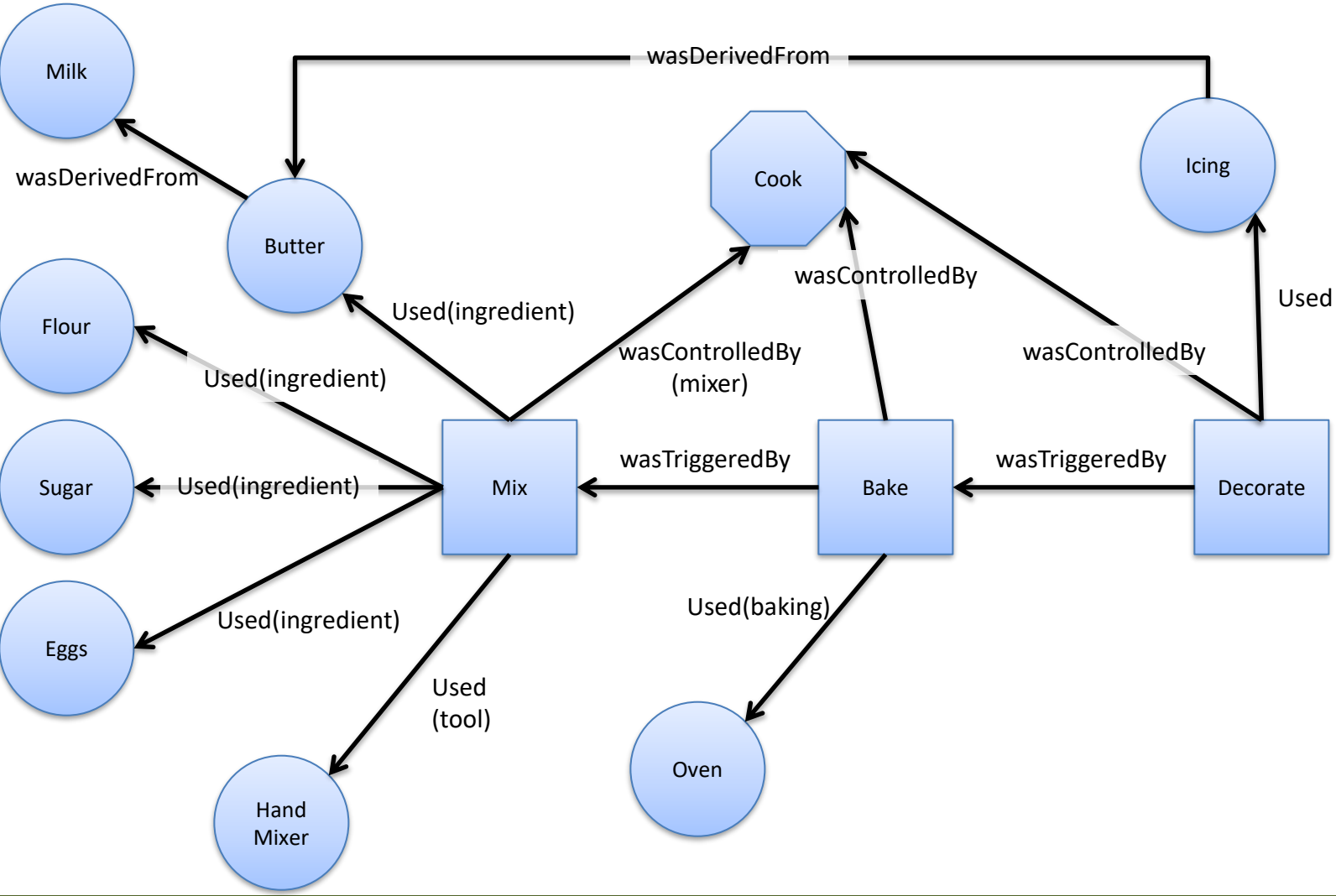
Proveniência



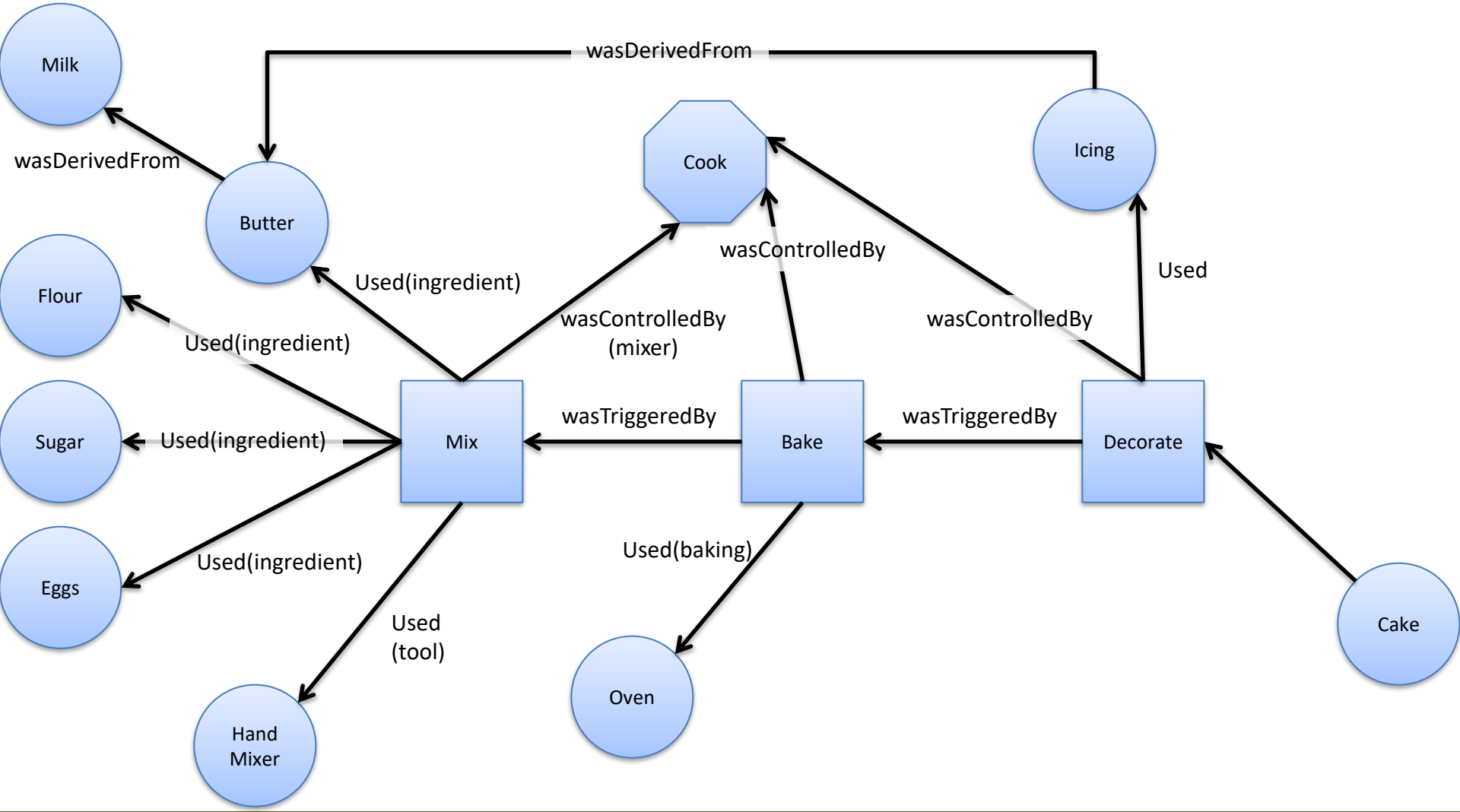
Proveniência



Proveniência



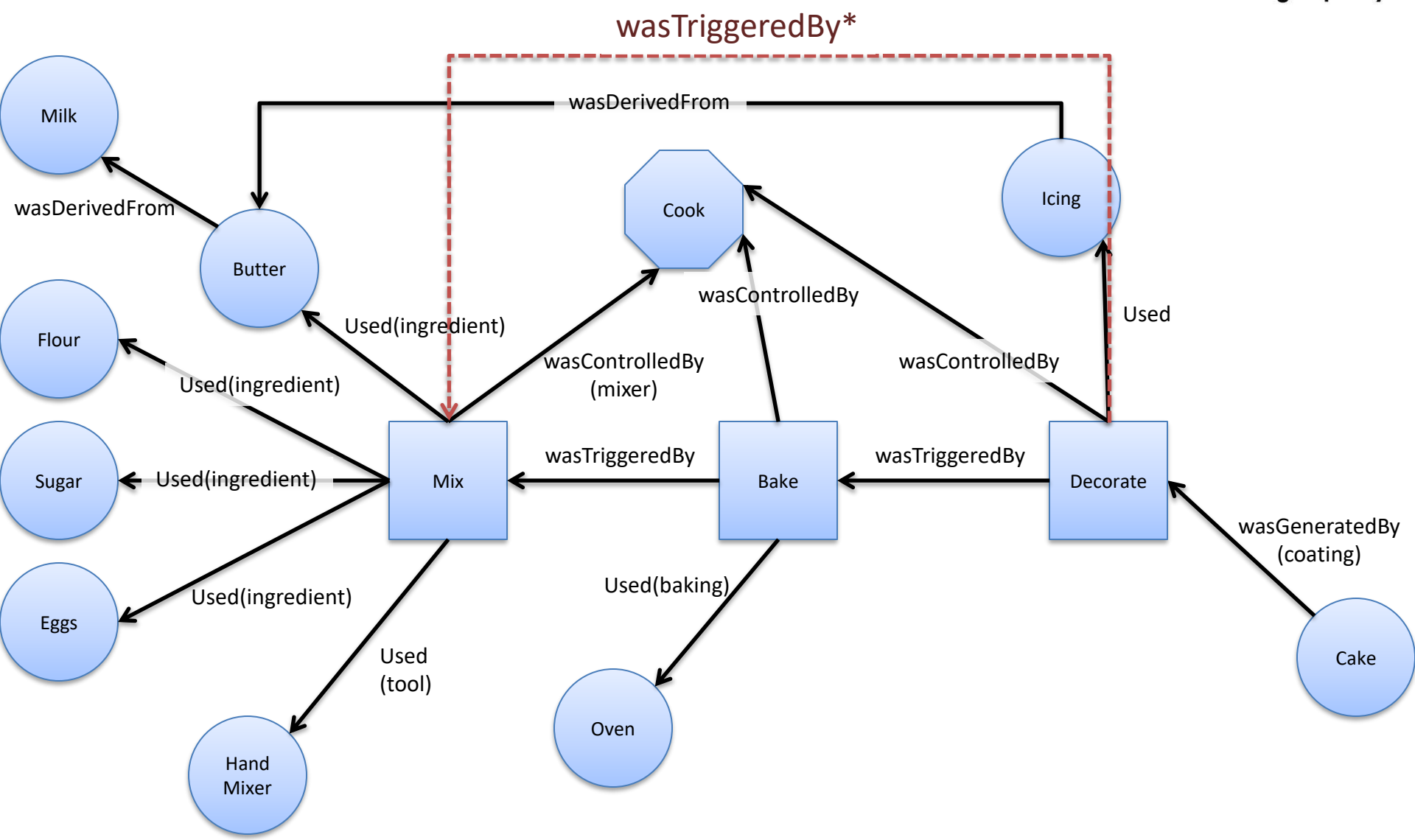
Proveniência



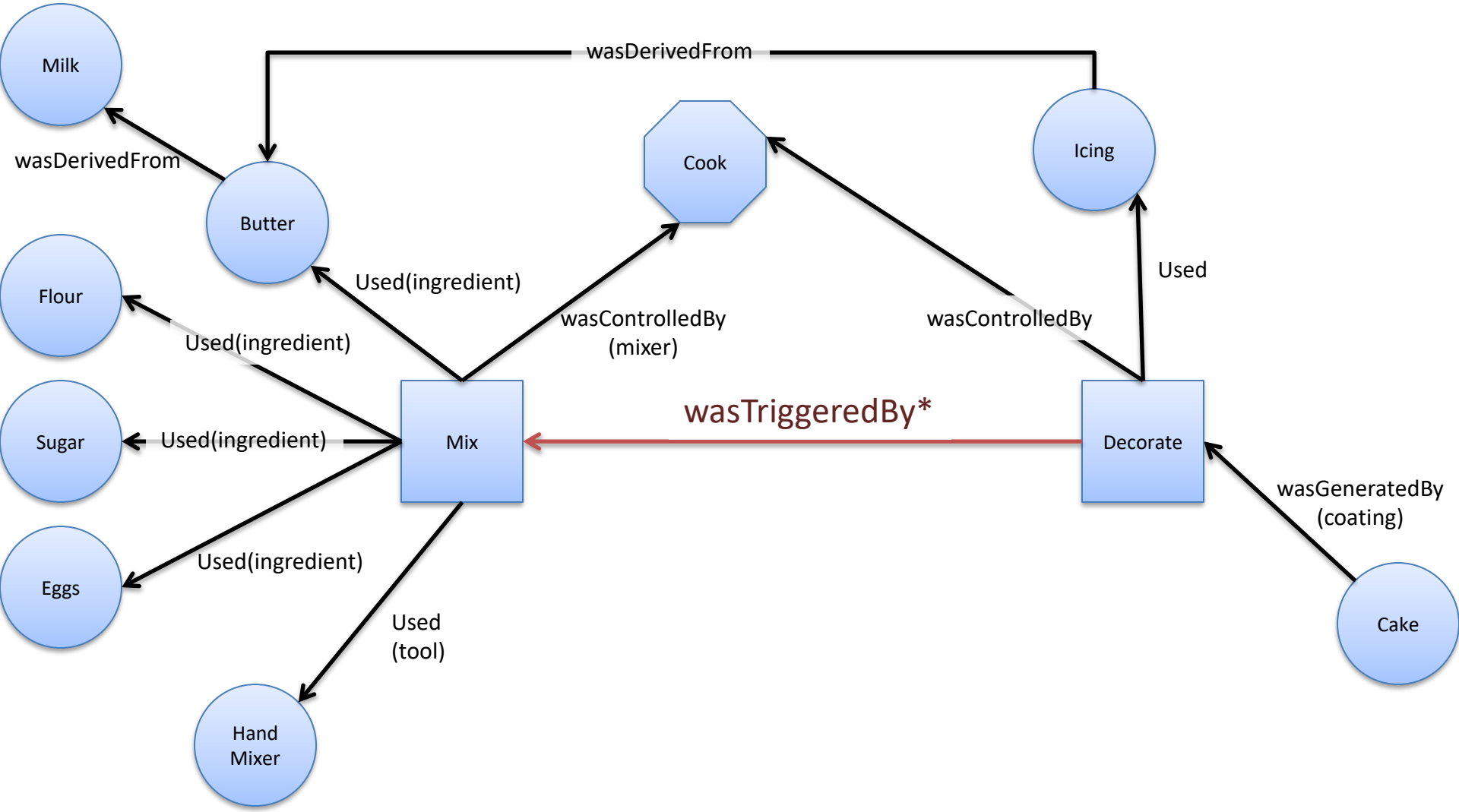
Proveniência

- Com isso temos um grafo de proveniência
- Esse grafo nos permite fazer algumas operações interessantes de navegação
- Por exemplo, inferências

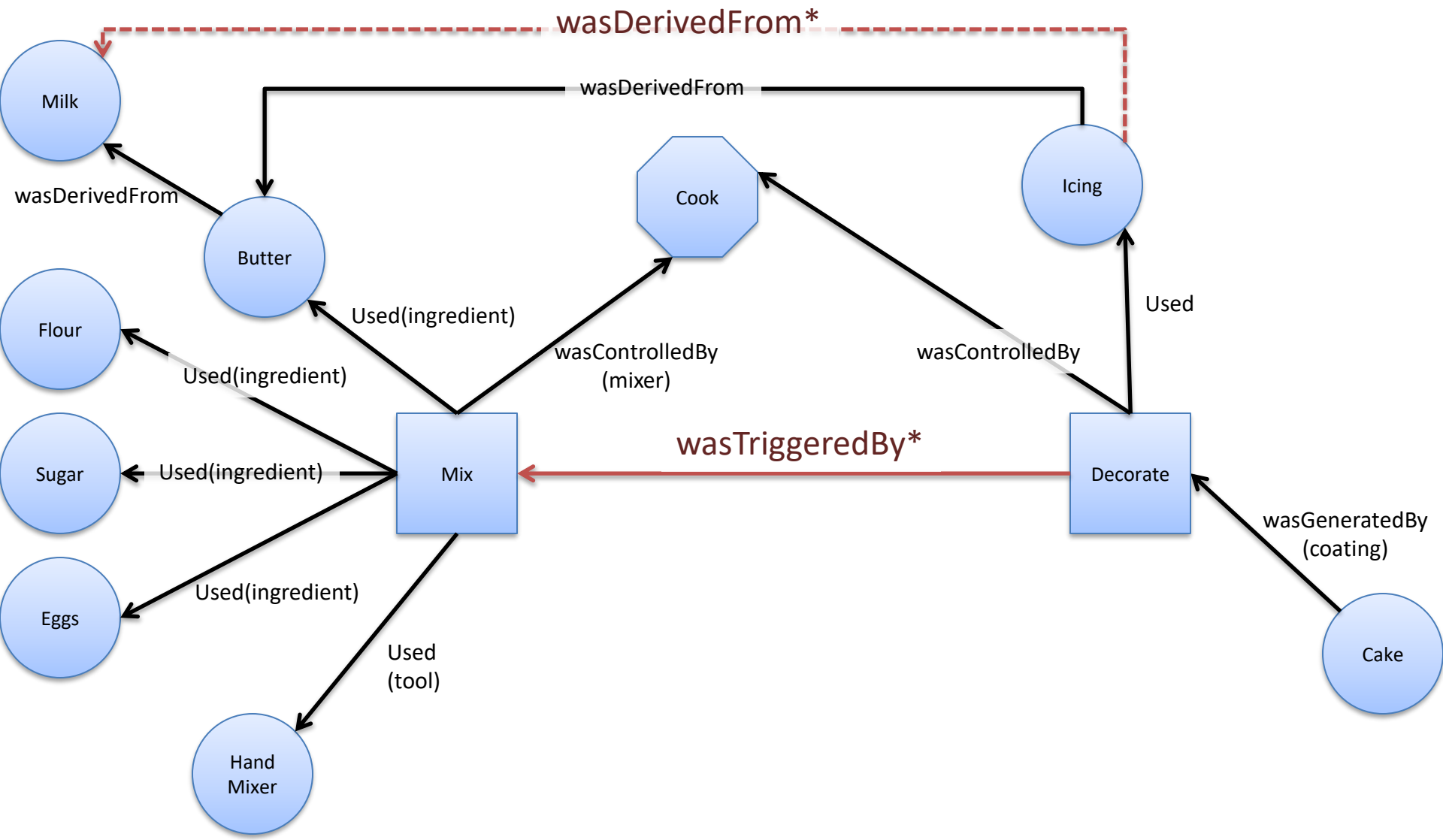
Inferência


wasTriggeredBy*
Milk
wasDerivedFrom
wasDerivedFrom
Butter
Cook
Icing
Flour
Used(ingredient)
wasControlledBy
(mixer)
wasControlledBy
wasControlledBy
Used
Sugar
Used(ingredient)
Mix
wasTriggeredBy
Bake
wasTriggeredBy
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used(baking)
Used
(tool)
Cake
Oven
Hand
Mixer

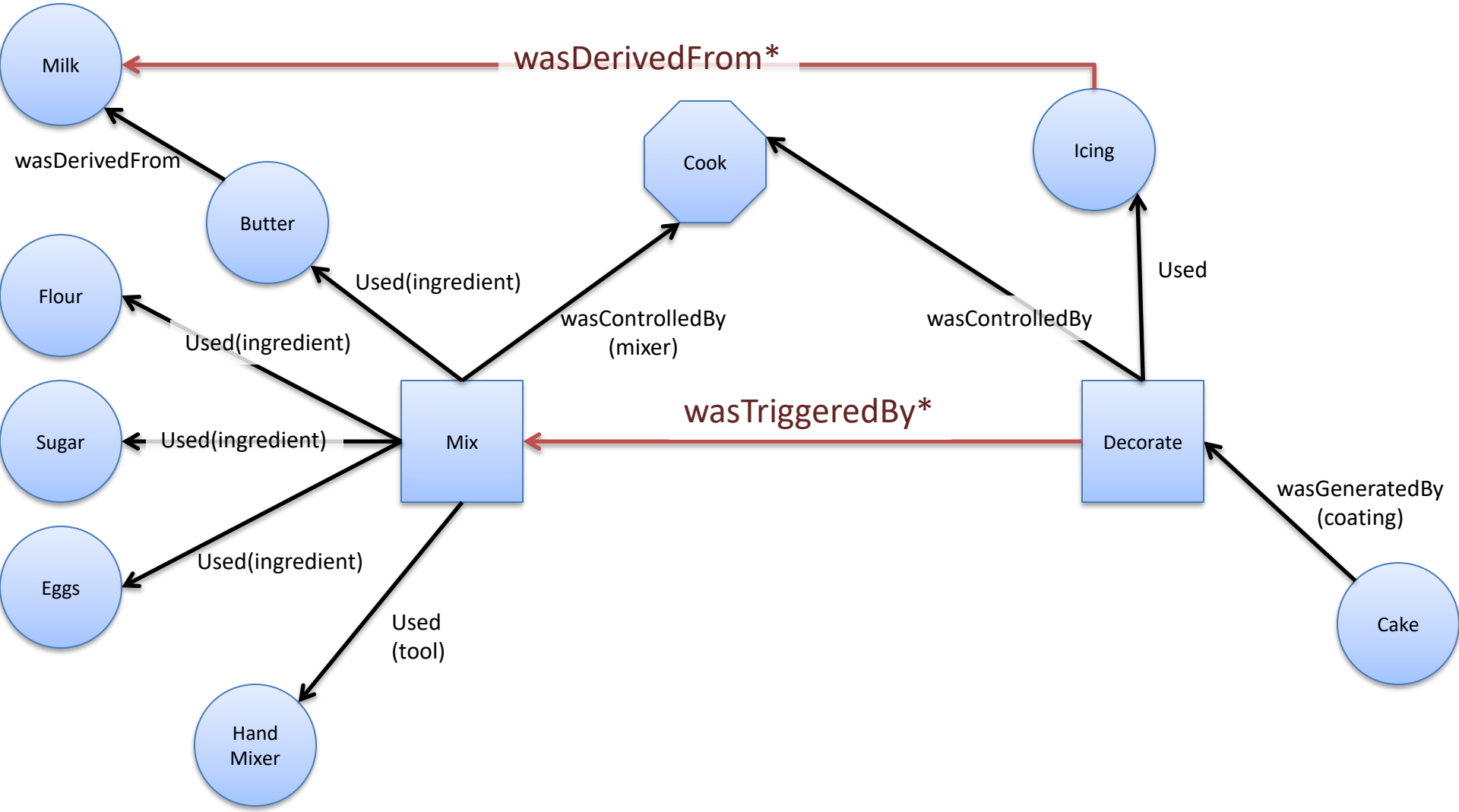
Inferência


Milk
wasDerivedFrom
wasDerivedFrom
Butter
Cook
Icing
Flour
Used(ingredient)
wasControlledBy
(mixer)
wasControlledBy
Used
Used(ingredient)
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used
(tool)
Cake
Hand
Mixer

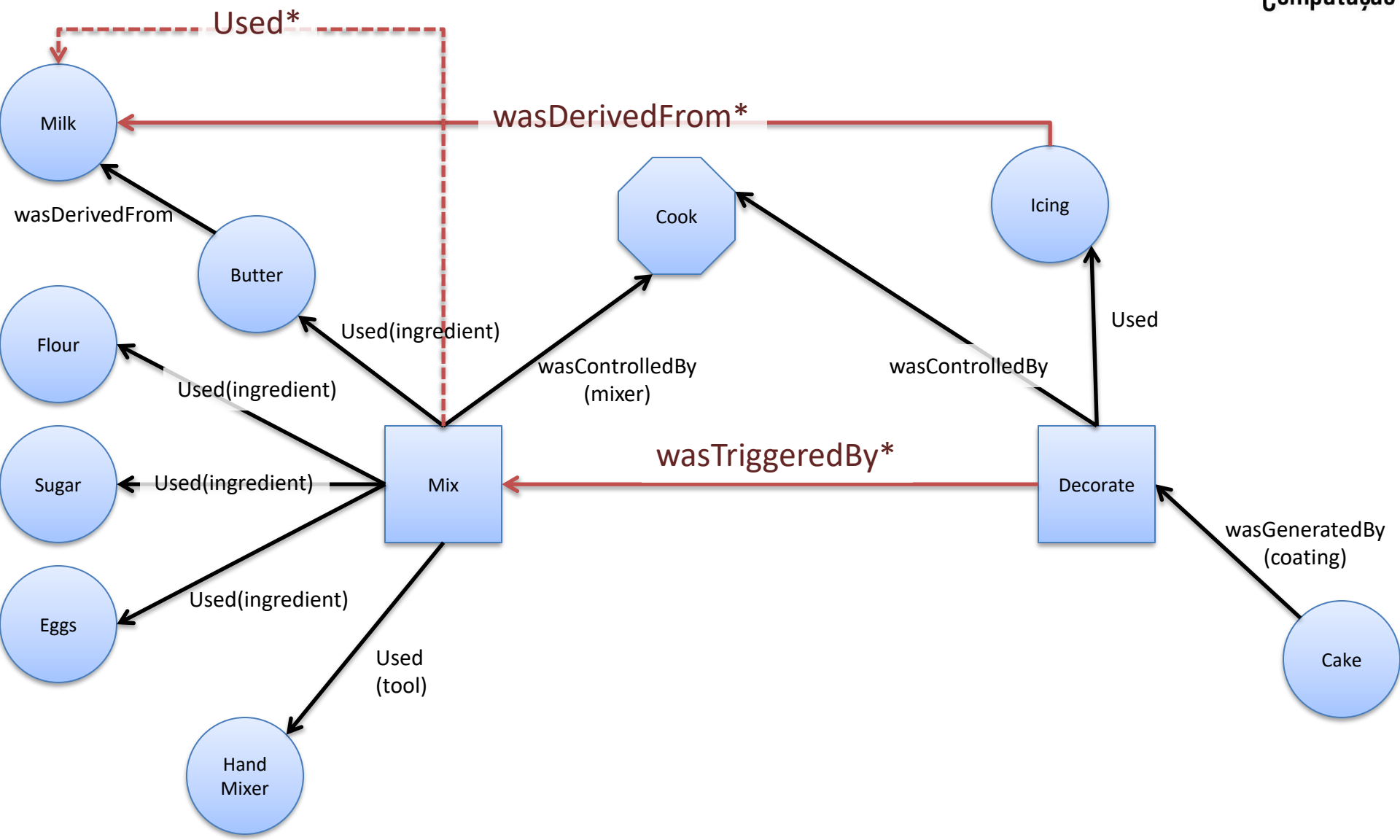
Inferência


wasDerivedFrom*
Milk
wasDerivedFrom
wasDerivedFrom
Cook
Icing
Butter
Used(ingredient)
Flour
wasControlledBy
(mixer)
wasControlledBy
Used
Used(ingredient)
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used
(tool)
Cake
Hand
Mixer

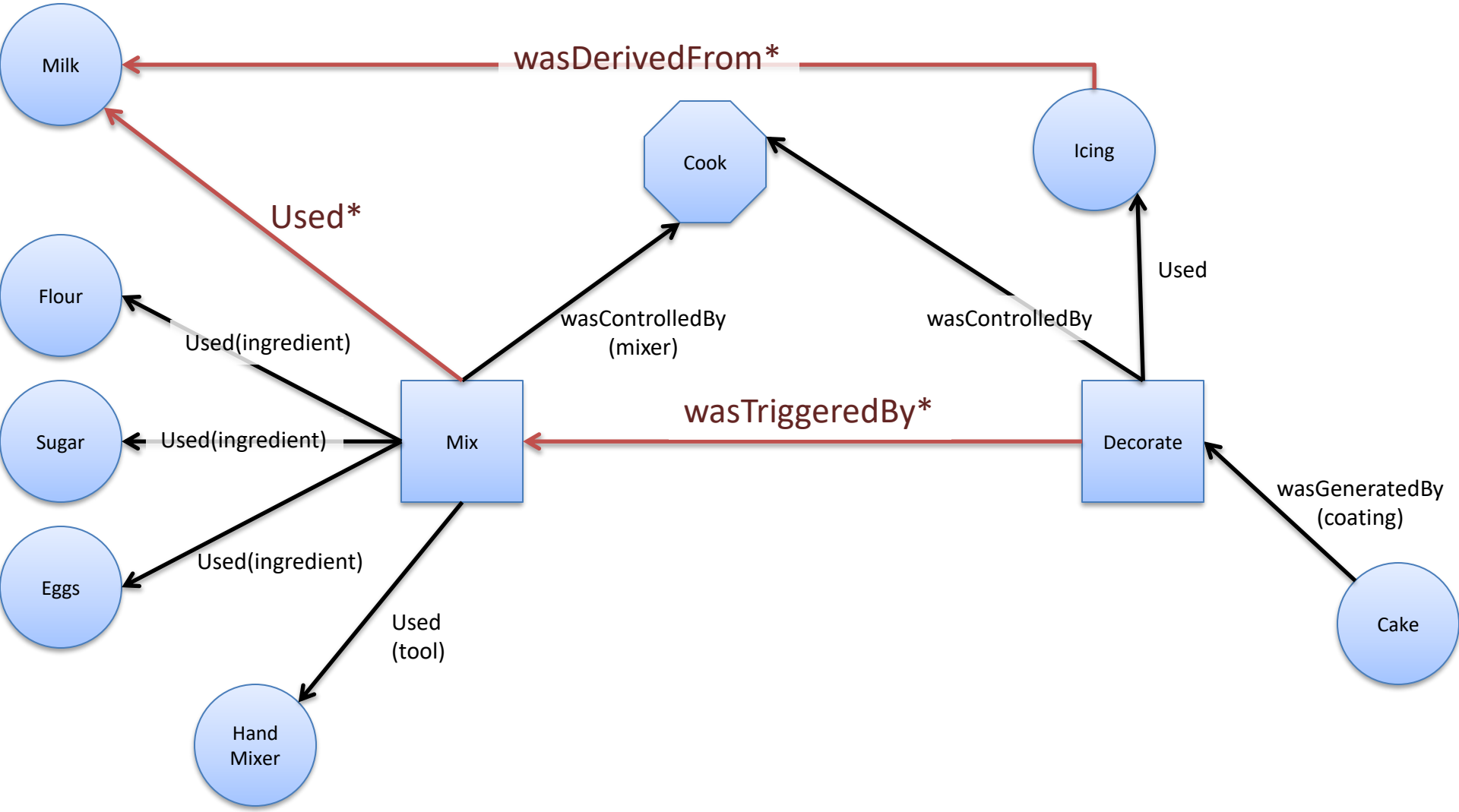
Inferência


Milk
wasDerivedFrom*
wasDerivedFrom
Butter
Cook
Icing
Flour
Used(ingredient)
wasControlledBy
(mixer)
wasControlledBy
Used
Used(ingredient)
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used
(tool)
Cake
Hand
Mixer

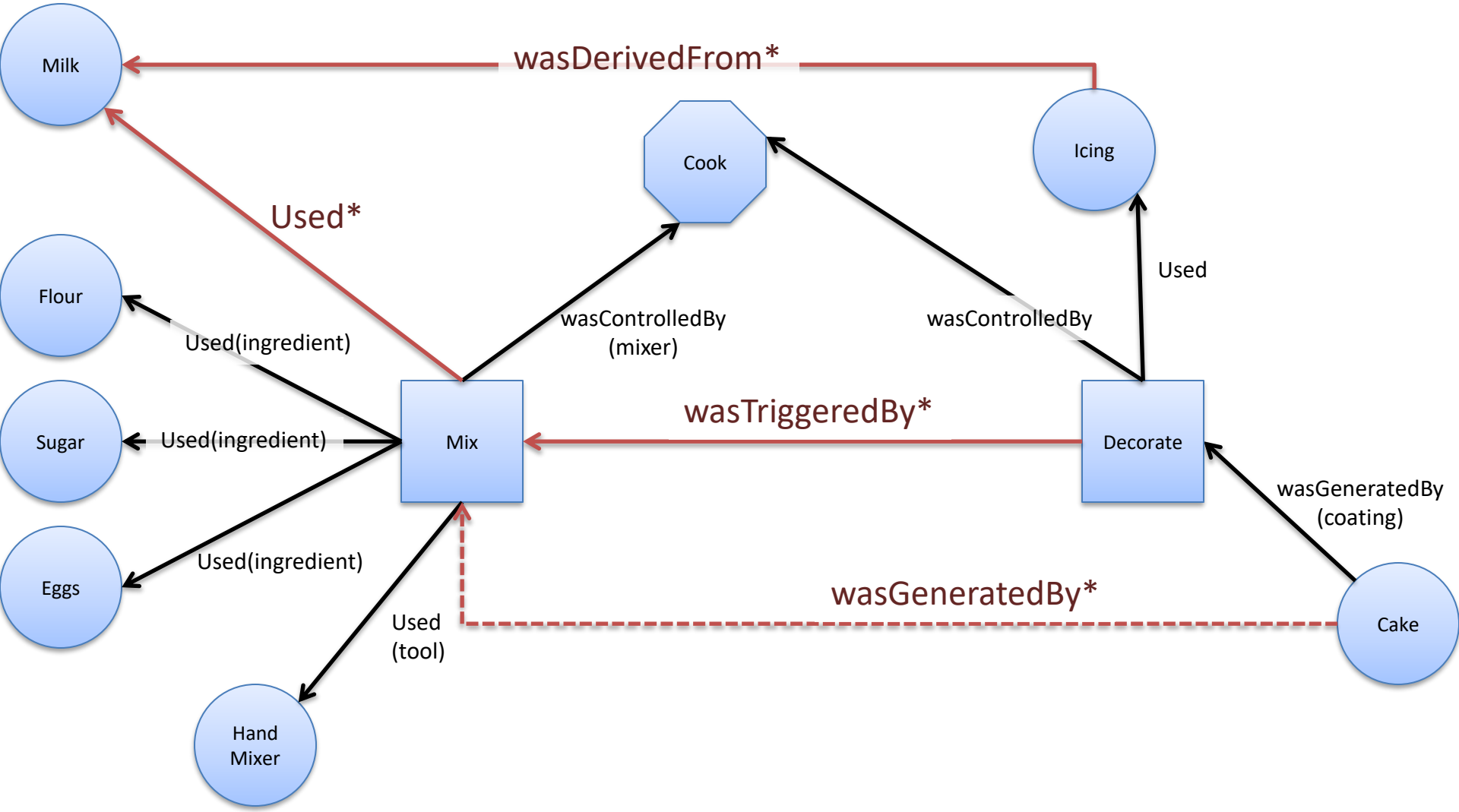
Inferência


Used*
Milk
wasDerivedFrom*
wasDerivedFrom
Butter
Cook
Icing
Used(ingredient)
Flour
wasControlledBy
(mixer)
wasControlledBy
Used
Used(ingredient)
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used
(tool)
Cake
Hand
Mixer

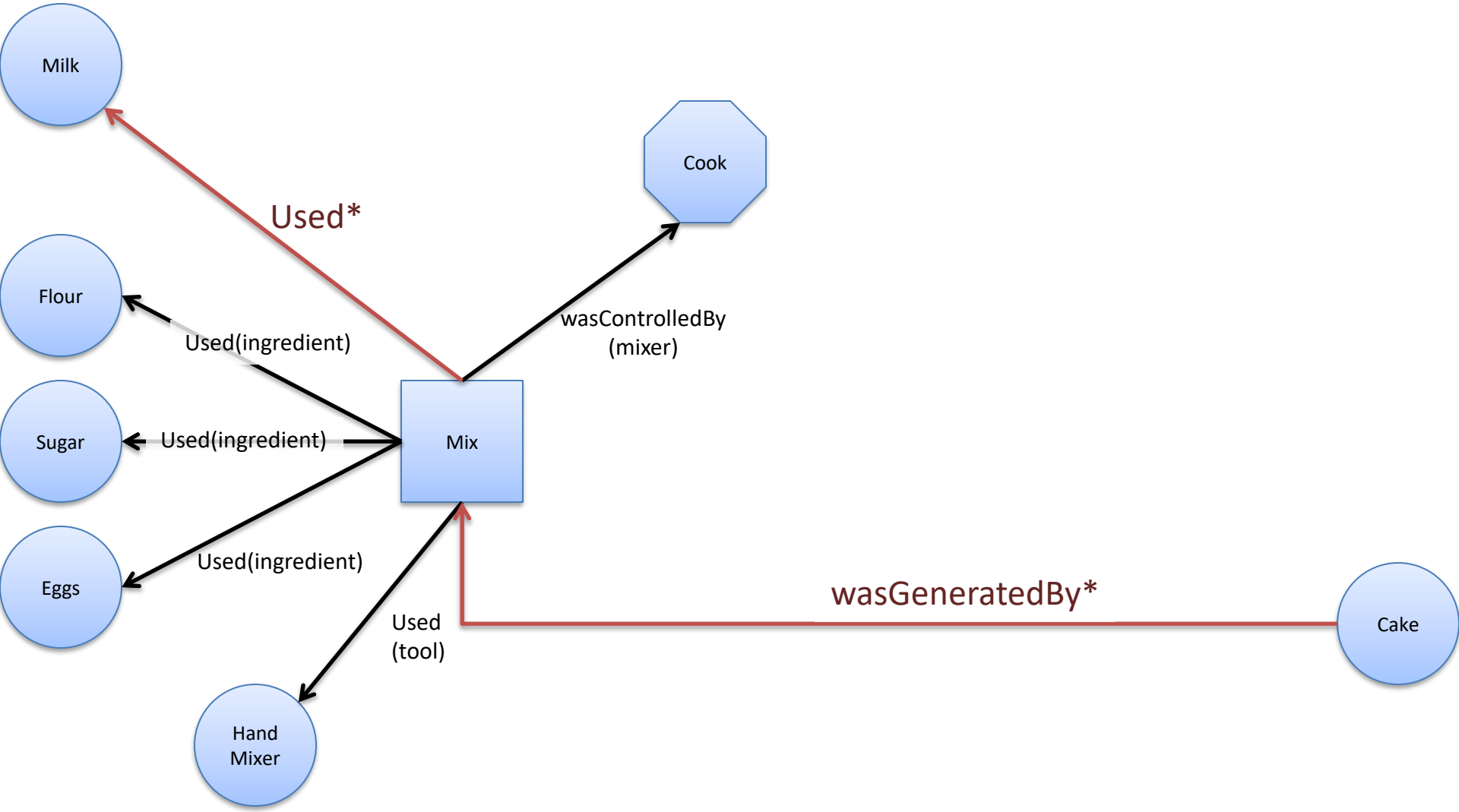
Inferência


Milk
wasDerivedFrom*
Cook
Icing
Used*
Flour
Used(ingredient)
wasControlledBy
(mixer)
wasControlledBy
Used
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
Used
(tool)
Cake
Hand
Mixer

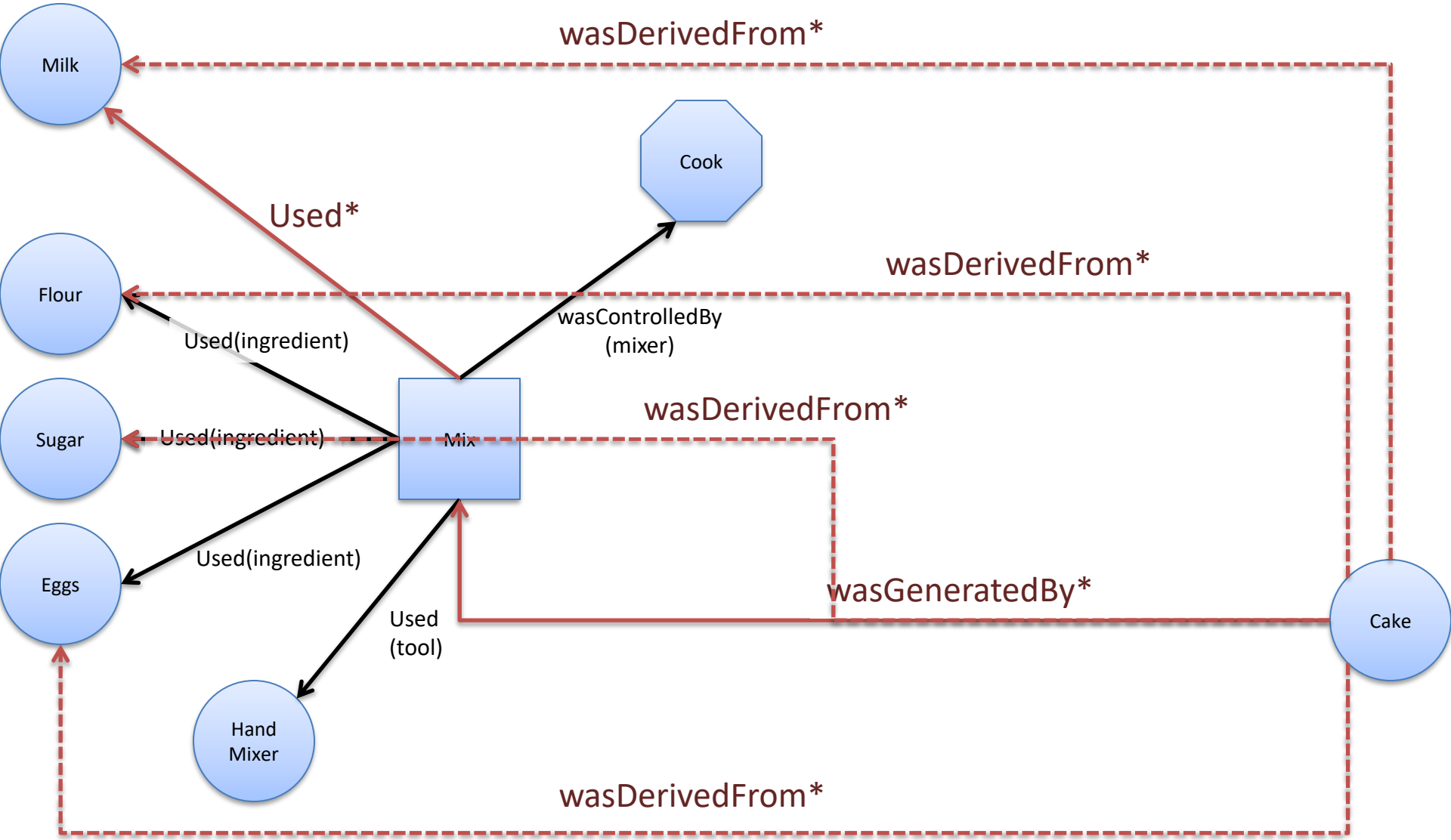
Inferência


Milk
wasDerivedFrom*
Icing
Cook
Used*
Flour
Used(ingredient)
wasControlledBy
(mixer)
wasControlledBy
Used
Sugar
Used(ingredient)
Mix
wasTriggeredBy*
Decorate
wasGeneratedBy
(coating)
Eggs
Used(ingredient)
wasGeneratedBy*
Used
(tool)
Cake
Hand
Mixer

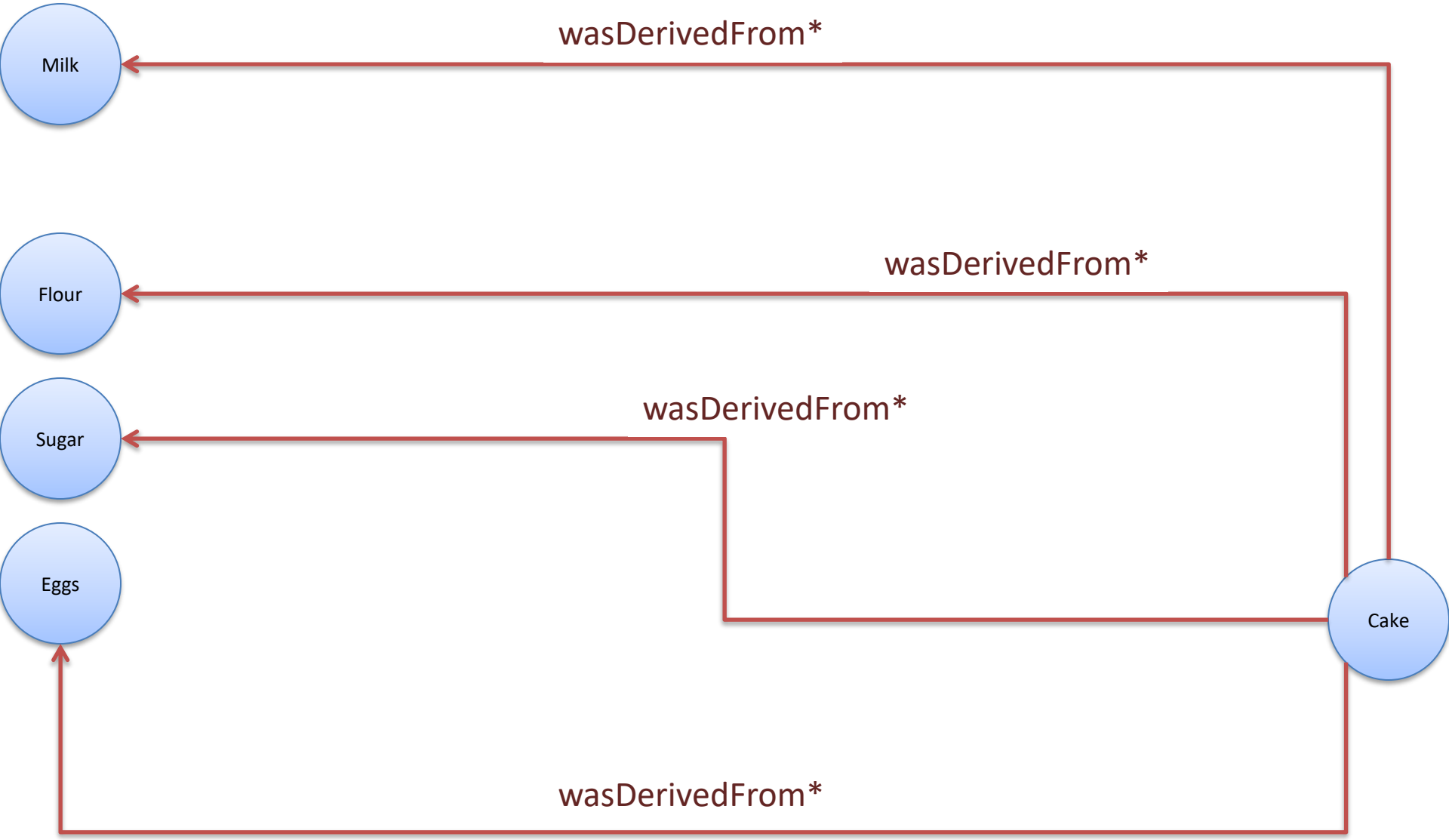
Inferência


Milk
Cook
Used*
Flour
Used(ingredient)
wasControlledBy
(mixer)
Sugar
Used(ingredient)
Mix
Eggs
Used(ingredient)
wasGeneratedBy*
Cake
Used
(tool)
Hand
Mixer

Inferência


wasDerivedFrom*
Milk
Cook
Used*
Flour
wasDerivedFrom*
Used(ingredient)
wasControlledBy
(mixer)
Sugar
Used(ingredient)
Mix
wasDerivedFrom*
Eggs
Used(ingredient)
wasGeneratedBy*
Cake
Used
(tool)
Hand
Mixer
wasDerivedFrom*

Inferência


Milk
wasDerivedFrom*
Flour
wasDerivedFrom*
Sugar
wasDerivedFrom*
Eggs
Cake
wasDerivedFrom*

REGISTRO E GERENCIAMENTO

Como registrar e gerenciar

- A proveniência é registrada como um tipo de metadados sobre o produto de dados
- Muitos campos de metadados coletados rotineiramente se enquadram na categoria de informações de proveniência
  - Data de criação
  - Criador
  - Instrumento ou software usado
  - Métodos de processamento de dados
  - etc.
- Assim, um bom gerenciamento de dados constitui a base para o registro preciso da proveniência!

Como registrar e gerenciar

- As abordagens para capturar e representar a proveniência podem ser descritas em várias dimensões:
  - Gravado em uma string de texto
  - Usando esquema genérico ou específico
  - Um modelo de dados de proveniência
  - Capturado internamente em uma ferramenta de software ou programa ou em um sistema externo
  - Representado em forma legível por máquina ou por humanos

Como registrar e gerenciar

- Em sua forma mais simples, a proveniência pode ser registrada em um único arquivo de texto README
  - Descreve a coleta de dados e os métodos de processamento usados
- A proveniência também pode ser registrada de uma forma mais estruturada
  - Usar elementos específicos, padrões de metadados
- Uma alternativa é usar diretamente o Modelo de Dados de Proveniência do W3C (PROV-DM)!

Como registrar e gerenciar

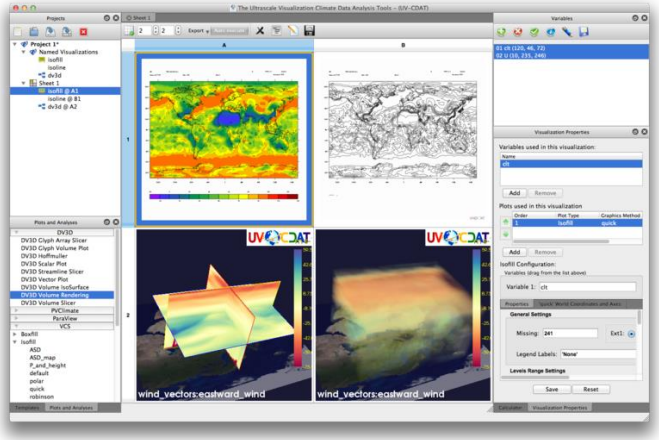
- Proveniência pode ser capturada internamente por ferramentas de software durante sua atividade de processamento
  - **Kepler**
  - **Galaxy**
  - **Taverna**
  - **noWorkflow**

- Proveniência esta normalmente disponível apenas para usuários do mesmo sistema
  - As vezes podem ser exportadas para uso externo!

- Os sistemas que adotam a abordagem interna tendem a capturar a proveniência de maneiras proprietárias
- Os sistemas que adotam uma abordagem externa geralmente usam um padrão como o W3C PROV
  - Precisam interagir com muitos tipos diferentes de sistemas

Como registrar e gerenciar

- Proveniência pode ser capturada de uma forma que permite que usuários leiam mais facilmente
  - Isso pode ser apenas uma descrição textual
  - Também pode envolver uma visualização da representação legível
    - Grafos, imagens, diagramas, etc.



Metadados de Proveniência

- Os principais pontos de metadados relacionados à proveniência dos dados incluem:
  - Origem de dados
  - Como os dados são transformados
  - Onde os dados se encontram com o passar do tempo

- Os principais usos da proveniência dos dados incluem:
  - Validação de dados
  - Depuração de dados defeituosos
  - Regeneração de dados perdidos ou corrompidos
  - Análise de Dependência de Dados
  - Auditoria e conformidade

PROV-N

- É uma sintaxe projetada para escrever instâncias do modelo de dados PROV
- As expressões PROV-N adotam uma notação funcional que consiste em um nome e uma lista de argumentos entre parênteses
- A interpretação dos argumentos PROV-N é definida de acordo com sua posição na lista de argumentos
- Esta convenção permite uma notação compacta

PROV-N

- O modelo de dados PROV define identificadores
  - Em PROV-N, são expressos como um nome local opcionalmente precedido de um prefixo e dois pontos
- Argumentos opcionais de PROV-N não precisam ser especificados
  - Regra geral para argumentos opcionais é que, se nenhum deles for usado na expressão, eles serão simplesmente omitidos, resultando em uma expressão mais simples
- No entanto, pode ser que apenas alguns dos argumentos opcionais precisem ser especificados
- Como a posição dos argumentos na expressão é importante, nesse caso, um marcador adicional deve ser usado para indicar que um determinado termo não está disponível
  - O marcador sintático '-' é usado para esse propósito

PROV-N

- A maioria das expressões inclui um identificador e um conjunto de pares de valor de atributo
  - Ambos são opcionais, a menos que especificado de outra forma
  - Por convenção, o identificador ocorre na primeira posição e o conjunto de pares atributo-valor na última posição

- Consistente com a convenção sobre argumentos, o marcador '-' pode ser usado quando o identificador não está disponível ou pode ser omitido completamente sem surgimento de ambiguidade
  - Para eliminar a ambiguidade de expressões que contêm um identificador opcional, o identificador ou marcador opcional deve ser seguido por ';'

Entidades
Atividades
Geração
Uso
Comunicação
Derivação
Agentes
Atribuição
Associação
Delegação

ESTRUTURAS BÁSICAS DE PROV

Entidade

*"Uma **entidade** é uma coisa física, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias"*

- `entity(id, [attr1=val1, ...])`
  - ***Id**:* identificador para a entidade
  - ***attributes***: um conjunto opcional de pares de atributo-valor ((attr1, val1), ...) representando informações adicionais sobre os aspectos fixos desta entidade

Atividade

*"Uma **atividade** é algo que ocorre durante um período de tempo e atua sobre ou com entidades; pode incluir consumir, processar, transformar, modificar, realocar, usar ou gerar entidades."*

- Atividades abrangem uma ampla gama de noções
  - Atividades de processamento de informações podem incluir mover, copiar ou duplicar entidades digitais
  - Atividades físicas podem incluir dirigir um carro entre dois locais ou imprimir algo

- Exemplos de atividades:
  - A publicação de um documento na Web
  - O envio de uma mensagem no Twitter
  - A extração de metadados embutidos em um arquivo
  - A condução de um carro de Boston a Cambridge
  - A montagem de um conjunto de dados baseado em um conjunto de medições
  - A realização de uma análise estatística sobre um conjunto de dados
  - Executando uma consulta SPARQL em um armazenamento triplo
  - Editando um arquivo

Atividade

- `activity(id, st, et, [attr1=val1, ...])`
  - ***id***: identificador para a atividade
  - ***startTime***: um tempo opcional (st) para o início da atividade
  - ***endTime***: um tempo opcional (et) para o fim da atividade
  - ***attributes***: um conjunto opcional de pares de atributo-valor ((attr1, val1), ...) representando informações adicionais sobre os aspectos fixos desta atividade

Atividade

- As atividades e entidades estão associadas entre si de duas maneiras diferentes:
  - As *atividades* ***utilizam*** *entidades*
  - As *atividades* ***produzem*** *entidades*

- O ato de ***utilizar*** ou ***produzir*** uma entidade pode ter uma duração
  - O termo '***geração***' refere-se à conclusão do ato de produzir
  - O termo '***uso***' se refere ao início do ato de utilizar entidades

Geração

*"A **geração** é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração."*

- Exemplos de geração são:
  - A criação completa de um arquivo por um programa
  - A criação completa de um conjunto de dados
  - A publicação completa de uma nova versão de um documento

- Dado que uma geração é a conclusão da produção de uma entidade, ela é instantânea!

Geração

- `wasGeneratedBy(id; e, a, t, attrs)`
  - ***id***: identificador opcional para a geração
  - ***entity***: identificador (e) para a entidade criada
  - ***activity***: identificador opcional (a) para a atividade que criou a entidade
  - ***time***: um "tempo de geração" opcional (t), o tempo em que a entidade foi completamente criada
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre esta geração

- *Obs: Embora **id**, **activity**, **time** e **attributes** sejam opcionais, pelo menos um deles deve estar presente*

Uso

*"**Uso** é o início da utilização de uma entidade por uma atividade. Antes do uso, a atividade não havia começado a utilizar esta entidade e não poderia ter sido afetada pela entidade."*

- Os exemplos de uso incluem:
  - Um procedimento que começa a consumir um argumento
  - Um serviço que começa a ler um valor
  - Um programa que começa a ler um arquivo de configuração
  - O ponto em que um ingrediente, como ovos, está sendo adicionado em uma atividade
- A mesma entidade pode ser usada várias vezes, possivelmente por diferentes atividades
  - Por exemplo, um arquivo em um sistema de arquivos pode ser lido indefinidamente
- Mas também pode consumir inteiramente uma entidade
  - Por exemplo, os ovos não estão mais disponíveis após serem adicionados à mistura

Uso

- `used(id; a, e, t, attrs)`
  - ***id***: identificador opcional para o uso;
  - ***activity***: identificador (a) para a atividade que usou a entidade;
  - ***entity***: identificador opcional (e) para a entidade usada
  - ***time***: um "tempo de uso" opcional (t), o tempo em que a entidade começou a ser usada
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre o uso

- *Obs1: Embora **id**, **entity**, **time** e **attributes** sejam opcionais, pelo menos um deles deve estar presente*

- *Obs2: Uma referência a uma determinada entidade pode aparecer em vários usos que compartilham um determinado identificador de atividade*

Comunicação

*"**Comunicação** é a troca de alguma entidade não especificada por duas atividades, uma atividade usando alguma entidade gerada pela outra."*

- A geração de uma entidade por uma atividade e seu uso subsequente por outra atividade é denominada comunicação
  - A atividade de escrever um artigo de celebridade foi informada por (uma instância de comunicação) a atividade de interceptar mensagens de correio de voz

- Uma comunicação implica que a atividade *a2* é dependente de outra *a1*, por meio de alguma entidade não especificada que é gerada por *a1* e usada por *a2*

Comunicação

- `wasInformedBy(id; a2, a1, attrs)`
  - ***id***: identificador opcional para identificar a relação;
  - ***informed***: identificador (a2) da atividade informada;
  - ***informant***: idenfiticador (a1) da atividade informante;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre essa comunicação.

Derivação

*"Uma **derivação** é a transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente."*

- As atividades utilizam entidades e produzem entidades
- Em alguns casos, a utilização de uma entidade influencia a criação de outra de alguma forma.
  - Essa noção de 'influência' é capturada por derivações

- Exemplos de derivação incluem:
  - A transformação de uma tabela relacional em um conjunto de dados vinculado
  - A transformação de uma tela em uma pintura
  - O transporte de uma obra de arte de Londres para Nova York
  - Uma transformação física, como o derretimento de gelo em água

Derivação

- O foco da derivação é conectar uma entidade gerada a uma entidade usada.
- O conceito de derivação pode ser bastante sutil
  - Implícita está a noção de que a entidade gerada foi afetada de alguma forma pela entidade usada
  - Se um artefato foi usado por uma atividade que também gerou um novo artefato, nem sempre significa que o segundo artefato foi derivado do primeiro
- Na atividade de criação de uma pintura, um artista pode ter misturado alguma tinta que nunca foi realmente aplicada à tela
  - A pintura normalmente não seria considerada uma derivação da tinta não utilizada
- PROV não tenta especificar as condições sob as quais existem derivações
  - A derivação é considerada como tendo sido determinada por meios não especificados
  - Embora uma cadeia de uso e geração seja necessária para que uma derivação seja mantida entre entidades, ela não é suficiente
  - Alguma forma de influência que ocorre durante as atividades envolvidas também é necessária

Derivação

- `wasDerivedFrom(id; e2, e1, -, -, -, attrs)`
  - ***id***: identificador opcional para uma derivação;
  - ***generatedEntity***: identificador (e2) da entidade gerada pela derivação;
  - ***usedEntity***: identificador (e1) da entidade usada pela derivação;
  - ***attributes***: conjunto opcional (attrs) de pares atributo-valor que representam informações adicionais sobre esta derivação.

Agente

*"Um **agente** é algo que tem alguma forma de responsabilidade pela realização de uma atividade, pela existência de uma entidade ou pela atividade de outro agente."*

- Para muitos propósitos, uma consideração chave para decidir se algo é confiável ou confiável é saber *quem ou o que foi responsável* por sua produção
  - Os dados publicados por uma organização famosa podem ser mais confiáveis do que os de uma organização desconhecida
  - Uma afirmação de um cientista conhecido com um histórico estabelecido pode ser mais acreditada do que uma afirmação de um novo aluno
  - Um cálculo realizado por uma biblioteca de software estabelecida pode ser mais confiável do que por um programa único

Agente

- Um agente pode ser um tipo específico de *entidade* ou *atividade*
  - Isso significa que o modelo pode ser usado para expressar a proveniência dos próprios agentes
- Exemplos de agentes:
  - O software para verificar o uso da gramática em um documento pode ser definido como um agente de uma atividade de preparação de documento
  - Um site que vende livros na Web
  - Os serviços envolvidos no processamento de pedidos
  - As empresas que os hospedam
- Os agentes podem estar relacionados a entidades, atividades e outros agentes

Agente

- `agent(id, [attr1=val1, ...])`
  - ***id***: identificador para um agente;
  - ***attributes***: conjunto de pares de valor de atributo ((attr1, val1), ...) representando informações adicionais sobre este agente.

Atribuição

*"**Atribuição** é a atribuição de uma entidade a um agente."*

- Uma postagem de blog pode ser atribuída a um autor
- Um telefone celular ao fabricante

- `wasAttributedTo(id; e, ag, attrs)`
  - ***id***: identificador opcional para a relação;
  - ***entity***: identificador de entidade (e);
  - ***agent***: o identificador (ag) do agente ao qual a entidade é atribuída e, portanto, tem alguma responsabilidade por sua existência;
  - ***attributes***: um conjunto opcional (attrs) de pares atributo-valor que representam informações adicionais sobre esta atribuição.

Associação

*"Uma **associação** de atividade é uma atribuição de responsabilidade a um agente por uma atividade, indicando que o agente tinha uma função na atividade."*

- Os agentes são definidos como tendo algum tipo de responsabilidade pelas atividades

- Exemplos de associação entre uma atividade e um agente são:
  - Criação de página web sob orientação de designer
  - Várias formas de participação em um painel de discussão, incluindo membro da audiência, membro do painel ou presidente do painel
  - Um evento público, patrocinado por uma empresa e hospedado por um museu

Associação

- `wasAssociatedWith(id; a, ag, -, attrs)`
  - ***id***: um identificador opcional para a associação entre uma atividade e um agente;
  - ***activity***: um identificador (a) para a atividade;
  - ***agent***: um identificador opcional (ag) para o agente associado à atividade;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo representando informações adicionais sobre esta associação desta atividade com este agente.

- *Obs: Embora **id**, **agent** e **atributes** sejam opcionais, pelo menos um deles deve estar presente*

Delegação

*"**Delegação** é a atribuição de autoridade e responsabilidade a um agente (por si mesmo ou por outro agente) para realizar uma atividade específica como delegado ou representante, enquanto o agente por quem atua retém alguma responsabilidade pelo resultado do trabalho delegado."*

- A natureza desta relação é ampla, incluindo a relação contratual, mas também a iniciativa altruísta do agente representativo
  - Um aluno publicando uma página da web que descreve um departamento acadêmico pode resultar no aluno e no departamento como agentes associados à atividade
  - Pode não importar qual aluno real publicou uma página da web
  - Pode ser significativamente importante que o departamento tenha dito ao aluno para publicar a página da web

Delegação

- `actedOnBehalfOf(id; ag2, ag1, -, attrs)`
  - ***id***: um identificador opcional para o link de delegação entre o delegado e o responsável;
  - ***delegate***: um identificador (ag2) para o agente associado a uma atividade, agindo em nome do agente responsável;
  - ***responsible***: um identificador (ag1) para o agente, em nome do qual o agente delegado agiu;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre este link de delegação.

Subtipagem

Relações expandidas

Novas relações

Atributos específicos do PROV

ESTRUTURAS ESTENDIDAS DE PROV

Estruturas estendidas de PROV

- Estruturas estendidas são definidas por uma variedade de mecanismos:
  - **Subtipagem**
  - **Relações expandidas**
  - **Novas relações**
  - **Atributos específicos do PROV**

Subtipagem

- A subtipagem pode ser aplicada nos tipos básicos
  - Um “**agente de software**” é um tipo especial de agente
  - Uma “**organização**” é uma instituição social ou legal, como uma empresa, sociedade, etc.
  - Agentes “**pessoa**” são pessoas
- A subtipagem também pode ser aplicada às relações básicas
  - Uma **revisão** é um tipo especial de derivação
  - Uma **citação** é um tipo especial de derivação

Revisão

*"Uma **revisão** é uma derivação para a qual a entidade resultante é uma versão revisada de algum original."*

- A implicação é que a entidade resultante contém conteúdo substancial do original
- Uma relação de revisão é um tipo de relação de derivação de uma entidade revisada para uma entidade anterior
- O tipo de uma relação de revisão é denotado por:
  - **prov:Revision**.
- PROV não define atributos específicos de revisão.

Citação

*"Uma **citação** é a repetição de (parte ou toda) uma entidade, como texto ou imagem, por alguém que pode ou não ser seu autor original*

- Uma relação de cotação é um tipo de relação de derivação, para a qual uma entidade foi derivada de uma entidade precedente copiando, ou "citando", parte dela ou toda ela

- O tipo de relação de citação é denotado por:
  - **prov:Quotation**.

- PROV não define atributos específicos de citação

Relações Expandidas

- Mostrei os sete conceitos de PROV que são mapeados para relações binárias

- Porém...
  - Alguns usos avançados desses conceitos não podem ser capturados por uma relação binária!
  - Requerem que as relações sejam expandidas para relações n-árias (!!!)

- As relações binárias são atalhos que podem ser "abertos" por aplicativos e preenchidos com mais detalhes

Relações Expandidas

- Por exemplo:
  - A **derivação** é um relacionamento de nível muito alto entre duas entidades
  - Um aplicativo pode decidir 'abrir' esse relacionamento em uma relação expandida que descreve como uma entidade foi derivada de outra em virtude da listagem da **geração**, **uso** e **atividade envolvida** na relação de derivação

- Os usuários/aplicativos são livres para decidir o nível de granularidade que desejam descrever e o PROV fornece a maneira de fazer isso!

Relações Expandidas

- Para ilustrar as relações expandidas, revisitamos o conceito de **associação**
- Os agentes podem contar com planos
  - Conjuntos de ações ou etapas, para atingir seus objetivos no contexto de uma atividade
- Consequentemente, uma forma expandida de relação de associação permite que um **plano** seja especificado
- O **plano** é definido por **subtipagem** e a **associação** por uma relação expandida

Plano e Associação

*"Um **plano** é uma entidade que representa um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos."*

- O tipo de entidade Plano é denotado por:
  - **prov: Plan**
- PROV não define atributos específicos do plano

*"Uma **associação** de atividade é uma atribuição de responsabilidade a um agente por uma atividade, indicando que o agente tinha uma função na atividade."*

- Pode ter um **plano** vinculado a associação!

Associação

- `wasAssociatedWith(id; a, ag, pl, attrs)`
  - ***id***: um identificador opcional para a associação entre uma atividade e um agente;
  - ***activity***: um identificador (a) para a atividade;
  - ***agent***: um identificador opcional (ag) para o agente associado à atividade;
  - ***plan***: um identificador opcional (pl) para o plano em que o agente confiava no contexto desta atividade;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo representando informações adicionais sobre esta associação desta atividade com este agente.

- *Obs: Embora **id**, **agent**, **plan** e **atributes** sejam opcionais, pelo menos um deles deve estar presente*

Relações Expandidas

- Não existem requisitos prescritivos sobre a natureza dos planos, sua representação, as ações ou etapas em que consistem, ou seus objetivos pretendidos

- Uma vez que os planos podem evoluir com o tempo, pode ser necessário rastrear sua proveniência, então os próprios planos são entidades

- Representar o plano explicitamente na proveniência pode ser útil para várias tarefas:
  - Validar a execução conforme representada no registro de proveniência
  - Gerenciar falhas de expectativa
  - Fornecer explicações

Relações Expandidas

- Exemplo de **associação** entre uma **atividade** e um **agente** envolvendo um **plano** é:
  - Uma transformação XSLT (uma **atividade**) lançada por um usuário (um **agente**) com base em uma folha de estilo XSL (um **plano**)

Derivação

- `wasDerivedFrom(id; e2, e1, a, g2, u1, attrs)`
  - ***id***: identificador opcional para uma derivação;
  - ***generatedEntity***: identificador (e2) da entidade gerada pela derivação;
  - ***usedEntity***: identificador (e1) da entidade usada pela derivação;
  - ***activity***: identificador opcional (a) para a atividade usando e gerando as entidades acima;
  - ***generation***: identificador opcional (g2) para a geração envolvendo a entidade gerada (e2) e a atividade (a);
  - ***usage***: identificador opcional (u1) para o uso envolvendo a entidade usada (e1) e atividade (a);
  - ***attributes***: conjunto opcional (attrs) de pares atributo-valor que representam informações adicionais sobre esta derivação.

Delegação

- `actedOnBehalfOf(id; ag2, ag1, a, attrs)`
  - ***id***: um identificador opcional para o link de delegação entre o delegado e o responsável;
  - ***delegate***: um identificador (ag2) para o agente associado a uma atividade, agindo em nome do agente responsável;
  - ***responsible***: um identificador (ag1) para o agente, em nome do qual o agente delegado agiu;
  - ***activity***: um identificador opcional (a) de uma atividade para a qual o link de delegação é válido;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre este link de delegação.

Inicio

*"**Início** é quando uma atividade é considerada iniciada por uma entidade, conhecida como **gatilho**. A atividade não existia antes de seu início. Qualquer uso, geração ou invalidação envolvendo uma atividade segue o início da atividade. Um início pode se referir a uma entidade de gatilho que desencadeou a atividade, ou a uma atividade, conhecida como iniciador, que gerou o gatilho."*

- Visto que um início é quando uma atividade é considerada iniciada, ela é instantânea
- `wasStartedBy(id; a2, e, a1, t, attrs)`
  - ***id***: identificador opcional para identificar o inicio da atividade;
  - ***activity***: identificador (a2) para a atividade iniciada;
  - ***trigger***: identificador opcional (e) para a entidade que esta acionando a atividade;
  - ***starter***: identificador opcional (a1) para a atividade que gerou a entidade (e);
  - ***time***: momento opcional (t) em que a atividade foi iniciada;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre o inicio desta atividade.

- *Obs: Embora **id**, **trigger**, **starter**, **time** e **atributes** sejam opcionais, pelo menos um deles deve estar presente*

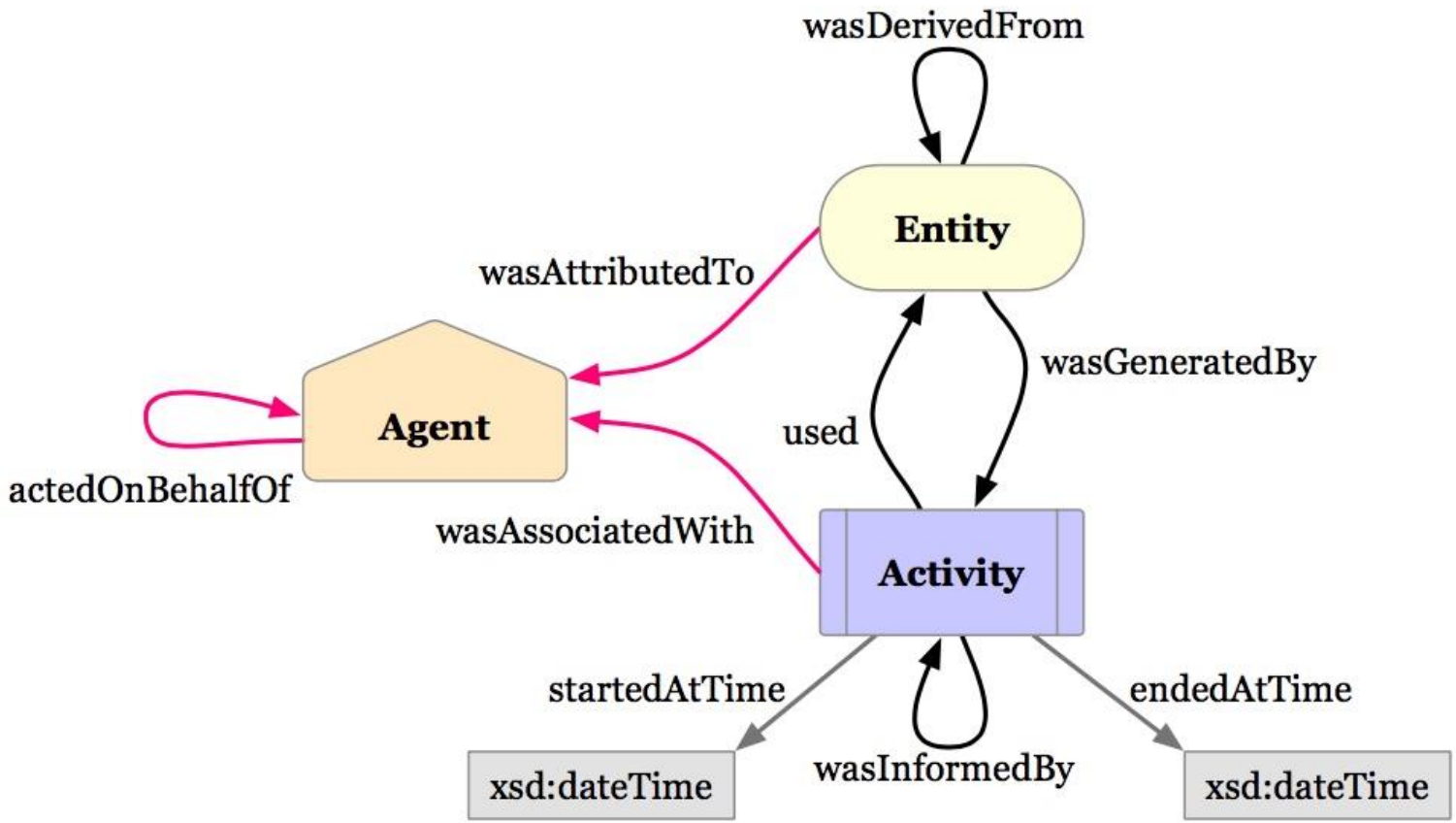
Fim

*"**Fim** é quando uma atividade é considerada encerrada por uma entidade, conhecida como **gatilho**. A atividade não existe mais após seu término. Qualquer uso, geração ou invalidação envolvendo uma atividade precede o fim da atividade. Um fim pode se referir a uma entidade de gatilho que encerrou a atividade, ou a uma atividade, conhecida como **ender** que gerou o gatilho."*

- Dado que um fim é quando uma atividade é considerada encerrada, é instantâneo
- `wasEndedBy(id; a2, e, a1, t, attrs)`
  - ***id***: identificador opcional para o fim da atividade;
  - ***activity***: identificador (a2) para a atividade finalizada;
  - ***trigger***: identificador opcional (e) para a entidade que dispara o término da atividade;
  - ***ender***: identificador opcional (a1) para a atividade que gerou a entidade (e);
  - ***time***: momento opcional (t) em que a atividade foi finalizada;
  - ***attributes***: um conjunto opcional (attrs) de pares de valor-atributo que representam informações adicionais sobre o fim desta atividade.

- *Obs: Embora **id**, **trigger**, **ender**, **time** e **atributes** sejam opcionais, pelo menos um deles deve estar presente*

Resumindo...


wasDerivedFrom
Entity
wasAttributedTo
wasGeneratedBy
Agent
used
actedOnBehalfOf
wasAssociatedWith
Activity
startedAtTime
endedAtTime
wasInformedBy
xsd:dateTime
xsd:dateTime

Novas relações

- PROV oferece suporte a relações adicionais que não são *subtipos* ou *versões expandidas* de relações existentes!

- Exemplos:
  - Influência
  - Especialização
  - Alternativa
  - Invalidação

Influência

*"**Influência** é a capacidade de uma entidade, atividade ou agente de ter um efeito sobre o caráter, desenvolvimento ou comportamento de outra por meio de uso, início, fim, geração, invalidação, comunicação, derivação, atribuição, associação ou delegação."*

- Uma relação de influência entre dois objetos *o2* e *o1* é uma dependência genérica de *o2* em *o1* que significa alguma forma de influência de *o1* em *o2*
  - As relações de **uso**, **início**, **fim**, **geração**, **invalidação**, **comunicação**, **derivação**, **atribuição**, **associação** e **delegação** também são uma influência

- Recomenda-se adotar essas relações mais específicas ao escrever descrições de proveniência
- Antecipa-se que a relação de Influência pode ser útil para expressar dúvidas sobre informações de proveniência

Influência

- `wasInfluencedBy(id; o2, o1, attrs)`
  - ***id***: um identificador opcional que identifica a relação;
  - ***influencee***: um identificador (o2) para uma entidade, atividade ou agente;
  - ***influencer***: um identificador (o1) para uma entidade, atividade ou agente ancestral do qual o primeiro depende;
  - ***attributes***: um conjunto opcional (attrs) de pares atributo-valor que representam informações adicionais sobre esta relação.

Especialização

*"Uma entidade que é uma **especialização** de outra compartilha todos os aspectos desta e, adicionalmente, apresenta aspectos mais específicos da mesma coisa que a última. Em particular, o tempo de vida da entidade sendo especializada contém o de qualquer especialização."*

- Exemplos de aspectos incluem:
  - Um período de tempo
  - Uma abstração
  - Um contexto associado à entidade

Especialização

- `specializationOf(infra, supra)`
  - ***specificEntity***: um identificador (infra) da entidade que é uma especialização da entidade geral (supra);
  - ***generalEntity***: um identificador (supra) da entidade que está sendo especializada.

- *Obs: Uma especialização não é uma influência e, portanto, não tem um **id** e **atributos***

Alternativa

*"Duas entidades **alternativas** apresentam aspectos da mesma coisa. Esses aspectos podem ser iguais ou diferentes, e as entidades alternativas podem ou não se sobrepor no tempo."*

- `alternateOf(e1, e2)`
  - ***alternate1***: um identificador (e1) da primeira das duas entidades;
  - ***alternate2***: um identificador (e2) da segunda das duas entidades.

- *Obs: Uma alternativa não é uma influência e, portanto, não tem um id e atributos*

- Observe que alternateOf é uma relação muito geral que apenas afirma que as duas entidades alternativas fixam alguns aspectos de alguma coisa comum e há alguma conexão relevante entre a proveniência das alternativas

Invalidação

*"A **invalidação** é o início da destruição, cessação ou expiração de uma entidade existente por uma atividade. A entidade não está mais disponível para uso (ou invalidação posterior) após a invalidação. Qualquer geração ou uso de uma entidade precede sua invalidação."*

- Visto que uma invalidação é o início da destruição, cessação ou expiração, ela é instantânea

- As entidades têm uma duração
- A geração marca o início de uma entidade, enquanto a invalidação marca o seu fim

Invalidação

- A vida útil de uma entidade pode terminar por diversos motivos:
  - Uma entidade foi destruída: por exemplo uma pintura foi destruída pelo fogo
  - Uma entidade foi consumida: por ex. Bob comeu toda a sopa
  - Uma entidade expira: por exemplo a oferta "compre uma cerveja, ganhe outra grátis" é válida durante o happy hour (19h-20h)
  - Uma entidade é limitada no tempo: por exemplo, o site de notícias da BBC em 3 de abril de 2012
  - Um atributo de entidade está mudando: por exemplo, o semáforo mudou de verde para vermelho

- Nos dois primeiros casos, a entidade desapareceu fisicamente após seu término: não há mais sopa, nem pintura
- No terceiro caso, pode haver um "voucher de oferta" que ainda exista, mas não é mais válido
- Da mesma forma, em 4 de abril, o site de notícias da BBC ainda existe, mas não é a mesma entidade que o site de notícias da BBC em 3 de abril
- O semáforo verde (uma entidade com semáforo verde de aspecto fixo) tornou-se o semáforo vermelho (outra entidade com semáforo vermelho de aspecto fixo)

Invalidação

- `wasInvalidatedBy(id; e, a, t, attrs)`
  - ***id***: identificador opcional para a invalidação;
  - ***entity***: identificador para a entidade invalidada;
  - ***activity***: identificador opcional para a atividade que invalidou a entidade;
  - ***time***: uma "hora de invalidação" opcional, a hora em que a entidade começou a ser invalidada;
  - ***attributes***: um conjunto opcional de pares de valor de atributo representando informações adicionais sobre esta invalidação.

- *Obs: Embora **id**, **activity**,**time** e **atributes** sejam opcionais, pelo menos um deles deve estar presente*

Atributos específicos de PROV

- ***prov:label***
  - O atributo **prov:label** fornece uma representação legível por humanos de uma instância de um tipo ou relação PROV-DM
  - O *valor* associado ao atributo **prov:label** deve ser uma string
- ***prov:location***
  - Um local pode ser um local geográfico identificável (ISO 19112), mas também pode ser um local não geográfico, como um diretório, linha ou coluna
  - Existem várias maneiras pelas quais a localização pode ser expressa, como por uma coordenada, endereço, ponto de referência, etc.
  - O atributo **prov:location** é um atributo opcional de **Entidade**, **Atividade**, **Agente**, **Uso**, **Geração**, **Invalidação**, **Início** e **Fim**.
- ***prov:role***
  - Um papel é a função de uma **entidade** ou **agente** com respeito a uma **atividade**, no contexto de um **uso**, **geração**, **invalidação**, **associação**, **início** e **fim**
  - O atributo **prov:role** pode ocorrer várias vezes em uma lista de pares de atributo-valor

Atributos específicos de PROV

- ***prov:type***
  - O atributo **prov:type** fornece mais informações de digitação para qualquer construção com um conjunto opcional de pares de atributo-valor
- ***prov:value***
  - O atributo **prov:value** fornece um *valor* que é uma representação direta de uma **entidade**
  - O atributo **prov:value** é um atributo opcional da **entidade**

- *Valor*
  - Um valor é uma constante, como string, número, hora, nome, dados binários codificados, etc.
  - A interpretação está fora do escopo de PROV

EXEMPLOS NO PROV-N

Exemplos no PROV-N

- Uma atividade com o identificador **a1** e um tipo de atributo com o valor **createFile**

```
activity(a1, [ prov:type="createFile" ])
```

- Duas entidades com identificadores **e1** e **e2**

```
entity(e1)
entity(e2)
```

- A atividade **a1** usou **e1** e **e2** foi gerada por **a1**

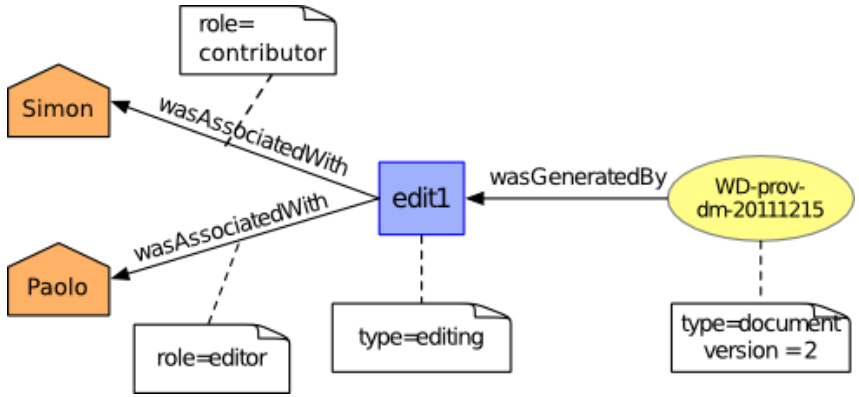
```
used(a1, e1)
wasGeneratedBy(e2, a1)
```

- As mesmas descrições, mas com um identificador explícito **u1** para o uso, e o marcador sintático '**-**' para marcar a ausência do identificador na geração. Ambos são seguidos por '**;**'

```
used(u1; a1, e1)
wasGeneratedBy(-; e2, a1)
```

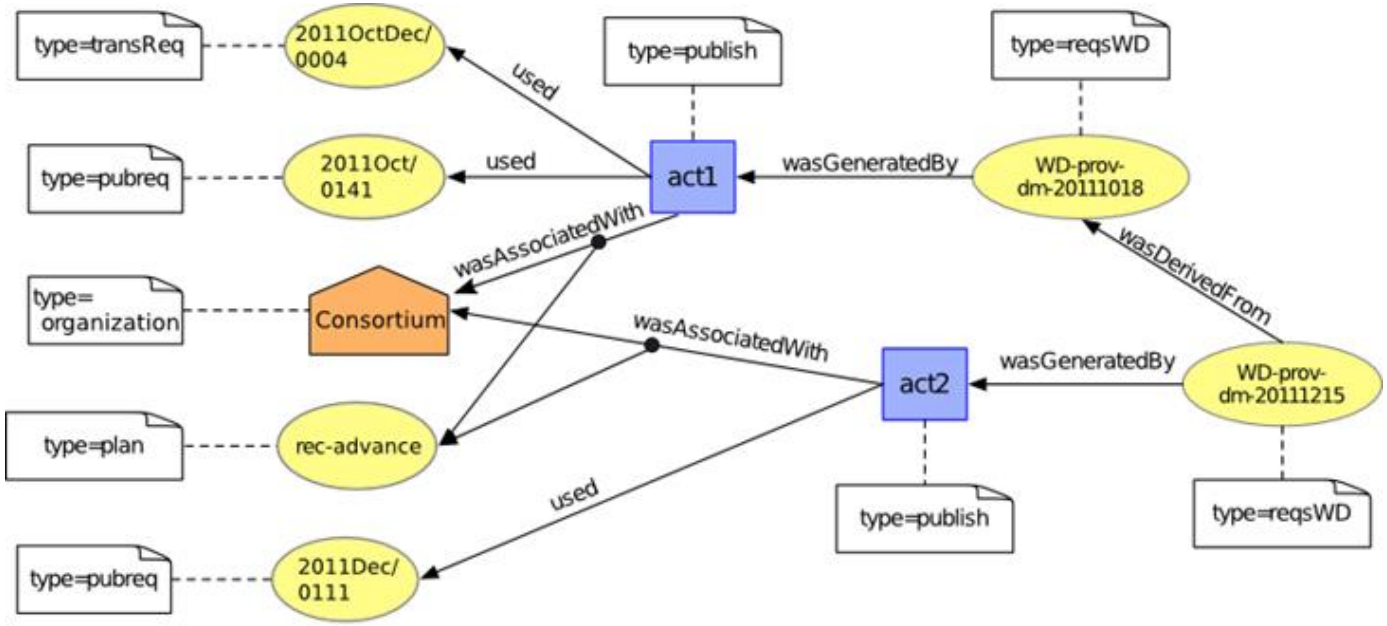
Exemplos no PROV-N

entity(tr:WD-prov-dm-20111215, [prov:type="document", ex:version="2"])
activity(ex:edit1, [prov:type="editing"])
wasGeneratedBy(tr:WD-prov-dm-20111215, ex:edit1, -)
agent(ex:Paolo, [prov:type='prov:Person'])
agent(ex:Simon, [prov:type='prov:Person'])
wasAssociatedWith(ex:edit1, ex:Paolo, -, [prov:role="editor"])
wasAssociatedWith(ex:edit1, ex:Simon, -, [prov:role="contributor"])



Exemplos no PROV-N

entity(tr:WD-prov-dm-20111215, [prov:type='rec54:WD'])
activity(ex:act2, [prov:type="publish"])
wasGeneratedBy(tr:WD-prov-dm-20111215, ex:act2, -)
wasDerivedFrom(tr:WD-prov-dm-20111215, tr:WD-prov-dm-20111018)
used(ex:act2, email:2011Dec/0111, -)
wasAssociatedWith(ex:act2, w3:Consortium, process:rec-advance)
[...]



Pacote

Coleção

TÓPICOS AVANÇADOS DE PROV

Pacote

*"Um **pacote** é um conjunto nomeado de descrições de proveniência e é, ele próprio, uma entidade, permitindo assim que a proveniência da proveniência seja expressa."*

- `bundle id description_1 ... description_n endBundle`
  - ***id***: identificador do pacote;
  - ***descriptions***: um conjunto de descrições de proveniência `description_1`, ..., `description_n`.

- Exemplo:

```
bundle bob:bundle1
  entity(ex:report1, [ prov:type="report", ex:version=1 ])
  wasGeneratedBy(ex:report1, -, 2012-05-24T10:00:01)
endBundle

bundle alice:bundle2
  entity(ex:report1)
  entity(ex:report2, [ prov:type="report", ex:version=2 ])
  wasGeneratedBy(ex:report2, -, 2012-05-25T11:00:01)
  wasDerivedFrom(ex:report2, ex:report1)
endBundle
```

Coleção

*"Uma **coleção** é uma entidade que fornece uma estrutura para alguns constituintes que devem ser entidades. Esses constituintes são considerados membros das coleções. Uma coleção vazia é uma coleção sem membros."*

- Existem muitos tipos diferentes de coleções, como conjuntos, dicionários ou listas
- Usando Coleções, pode-se expressar a proveniência da coleção em si, além da dos membros
- Um exemplo de coleção é um arquivo de documentos
  - Cada documento tem sua própria proveniência
  - O arquivo em si também tem alguma proveniência: quem o manteve, quais documentos ele continha em que momento, como foi montado, etc.

Filiação

*"A **filiação** informa que uma entidade pertence a uma coleção."*

- Uma relação de filiação/associação é definida para indicar os membros de uma coleção

- `hadMember(c, e)`
  - ***coleção***: um identificador (c) para a coleção cujo membro é afirmado;
  - ***entidade***: o identificador e de uma entidade que faz parte da coleção.

- *Obs: A filiação não é uma influência e, portanto, não tem um id e atributos.*

Coleção e Filiação

- Exemplo:

```
entity(e0)
entity(e1)
entity(e2)

entity(c, [prov:type='prov:Collection'  ])      // c is a
collection, with unknown content

hadMember(c, e0)
hadMember(c, e1)
hadMember(c, e2)
```

Informações Finais sobre Proveniência

- O W3C Provenance Working Group recomenda:
  - PROV Primer,
  - PROV Ontology (PROV-O),
  - PROV Data Model (PROV-DM),
  - PROV Notation (PROV-N),
  - PROV Constraints,
  - PROV Access and query.

- O Workshop Internacional de Proveniência e Anotação (IPAW) é um workshop semestral que se preocupa com questões de proveniência de dados, derivação de dados e anotação de dados
  - Reúne cientistas da computação de diferentes áreas e usuários de proveniência para discutir problemas em aberto relacionados à procedência de artefatos computacionais e não computacionais

noWorkflow

visTrails

PinG + Prov Viewer

FERRAMENTAS

noWorkflow

- O projeto noWorkflow visa permitir que os cientistas se beneficiem da análise de dados de proveniência, mesmo quando não usam um sistema de workflow
- Objetivo é permitir que eles evitem o uso de convenções de nomenclatura para armazenar arquivos originados em execuções anteriores

https://github.com/gems-uff/noworkflow

noWorkflow

- Ao invés de rodar
  - $ python experiment.py
- Instalar o noWorkflow (apenas uma vez)
  - $ pip install noworkflow[all]
- E executar
  - $ now run experiment.py

https://github.com/gems-uff/noworkflow

noWorkflow

```
$ now run -v experiment.py
[now] removing noWorkflow boilerplate
[now] setting up local provenance store
[now] collecting definition provenance
[now] registering user-defined functions
[now] collecting deployment provenance
[now] registering environment attributes
[now] searching for module dependencies
[now] registering provenance from 703 modules
[now] collecting execution provenance
[now] executing the script
[now] the execution of trial 1 finished successfully
```

Código do script

Ambiente (módulos, versões, SO, etc.)

Início, fim, Valores de variáveis, etc.

noWorkflow

- **Coleta proveniência** de scripts **Python** de forma **transparente**
- Considerera multiplas execuções (**ensaios**)
- Fornece **visualizações** para análise de proveniência
- Permite **consultar** proveniência em diferentes linguagens (SQL e Prolog)

https://github.com/gems-uff/noworkflow

noWorkflow

visTrails

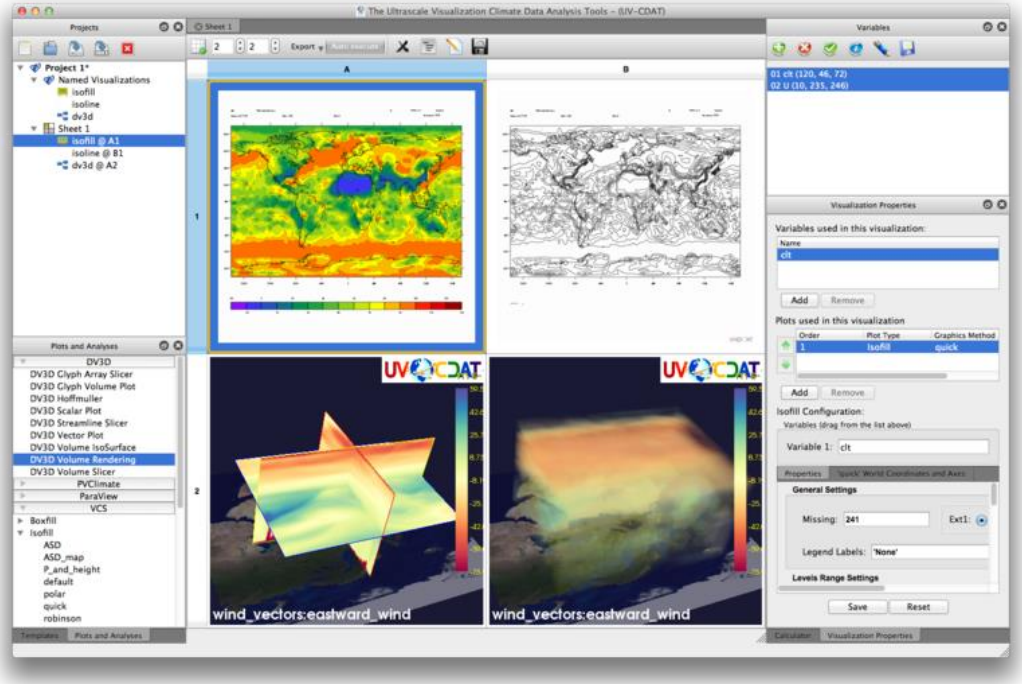
PinG + Prov Viewer

VISTRAILS

https://www.vistrails.org/index.php/Main_Page

visTrails

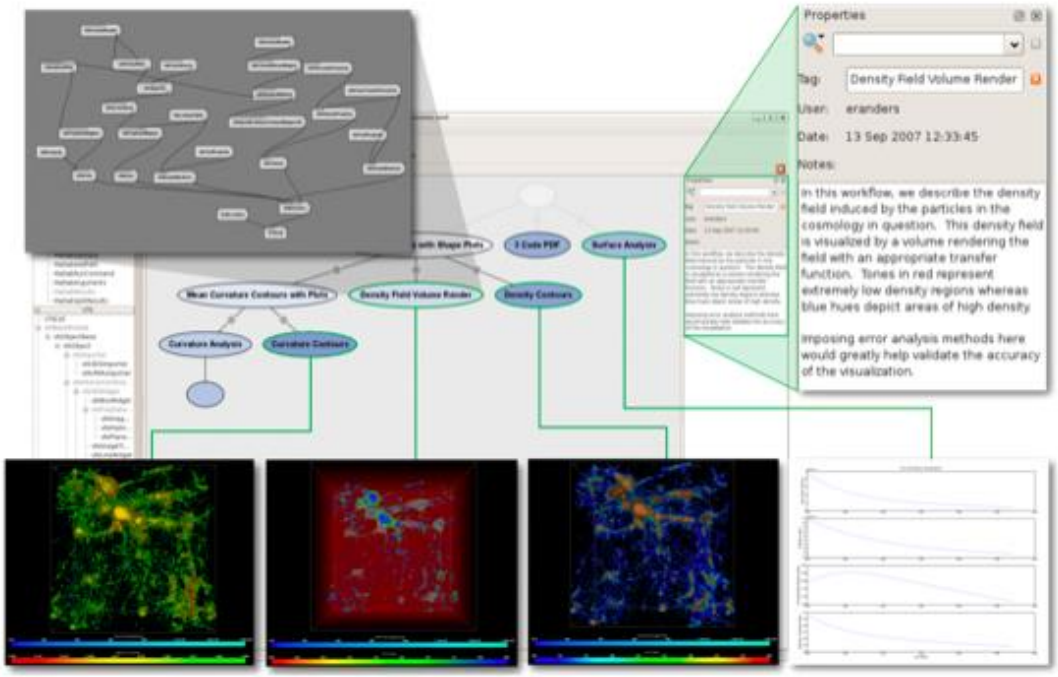
- VisTrails é um workflow científico de código aberto e sistema de gerenciamento de proveniência
- Fornece suporte para simulações, exploração de dados e visualização
- VisTrails foi projetado para gerenciar essas evoluções rápidas dos workflows



- Workflows têm sido tradicionalmente usados para automatizar tarefas repetitivas, para aplicativos de natureza exploratória, como simulações, análise de dados e visualização
- Conforme um engenheiro ou cientista gera e avalia hipóteses sobre os dados em estudo, uma série de workflows diferentes são criados enquanto um workflow é ajustado em um processo interativo

visTrails

- Mantém a proveniência de produtos de dados, dos workflows que derivam esses produtos e suas execuções
- Essas informações são persistidas como arquivos XML ou em um banco de dados relacional
- Os usuários podem:
  1. Navegar nas versões do workflow de forma intuitiva
  2. Desfazer as alterações (sem perder nenhum resultado)
  3. Comparar visualmente diferentes workflows e seus resultados
  4. Examinar as ações que levou a um resultado



noWorkflow

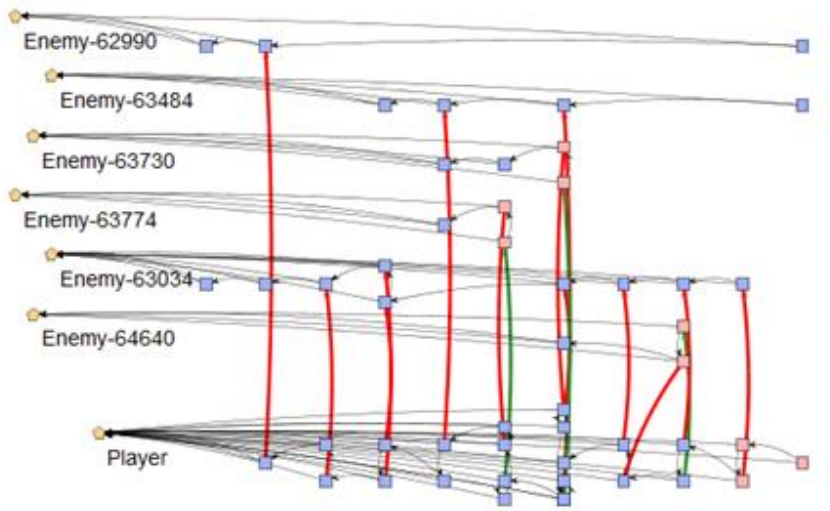
visTrails

PinG + Prov Viewer

PING + PROV VIEWER

Provenance in Games (PinG)

- Mapeamento de domínio
  - Proveniência para jogos
  - **Primeiro trabalho que trouxe proveniência para o contexto de jogos**
- Coleta de dados
  - Proveniência
  - Relações causais



PinG

(Kohwalter et al., 2012)

- Entidades
  - Objetos
- Atividades
  - Ações
  - Eventos
- Agentes
  - NPCs
  - Player

PinG

(Kohwalter et al., 2012)

- Entidades
  - Objetos
- Atividades
  - Ações
  - Eventos
- Agentes
  - NPCs
  - Player

PinG

(Kohwalter et al., 2012)

- Entidades
  - Objetos
- Atividades
  - Ações
  - Eventos
- Agentes
  - NPCs
  - Player

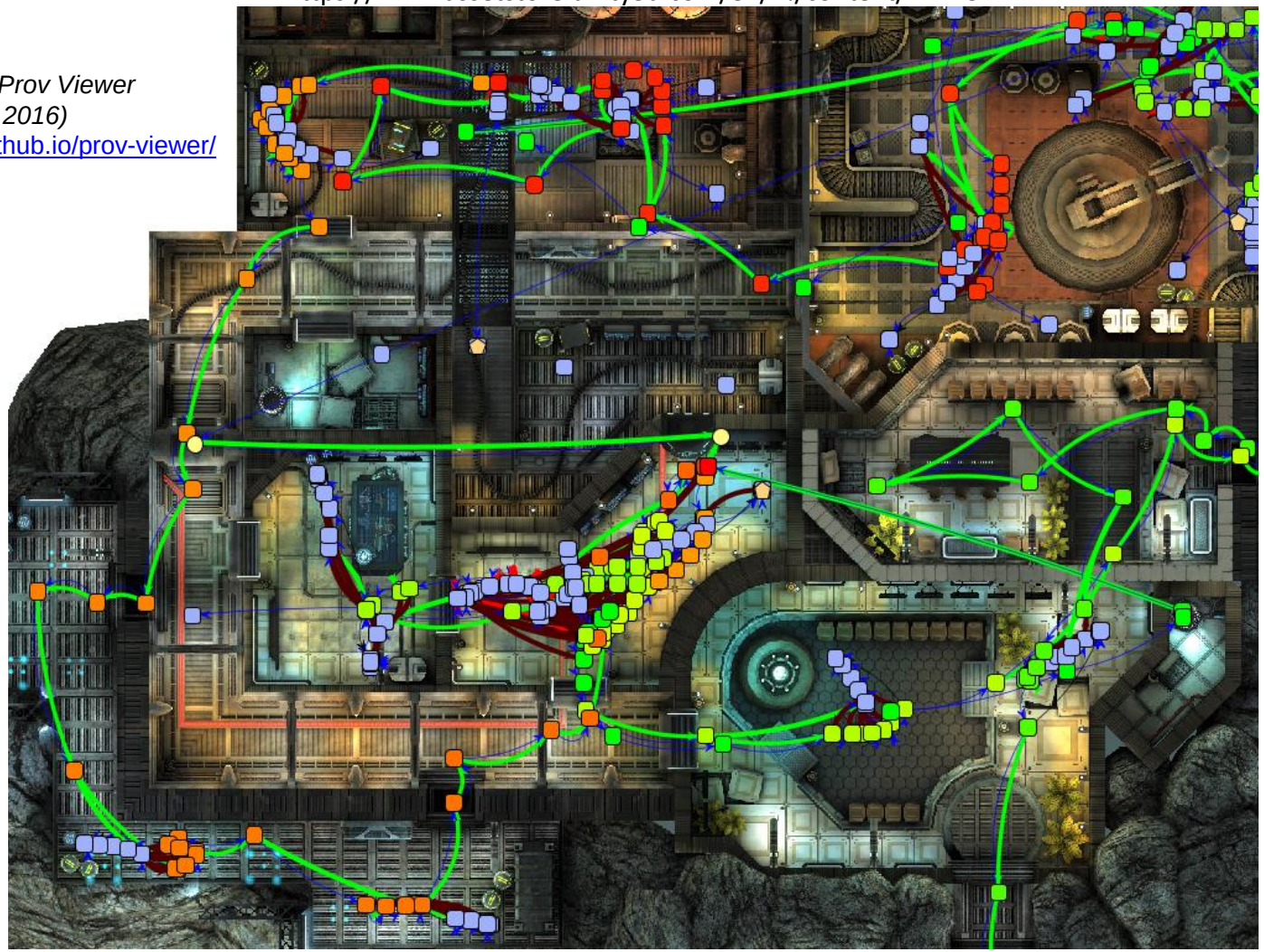





Angry Bots

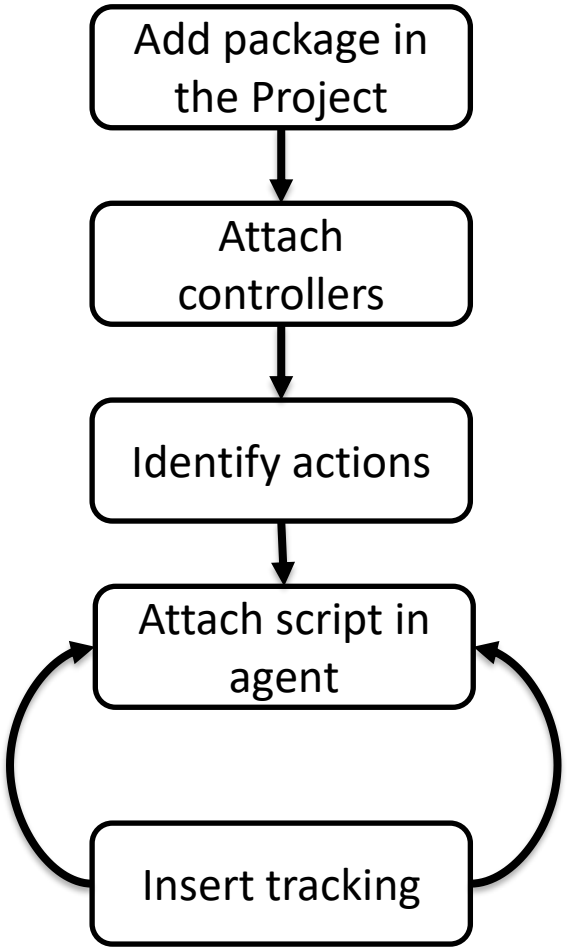
https://www.assetstore.unity3d.com/en/#!/content/12175

Grafo gerado no *Prov Viewer (Kohwalter et al., 2016)*
http://gems-uff.github.io/prov-viewer/



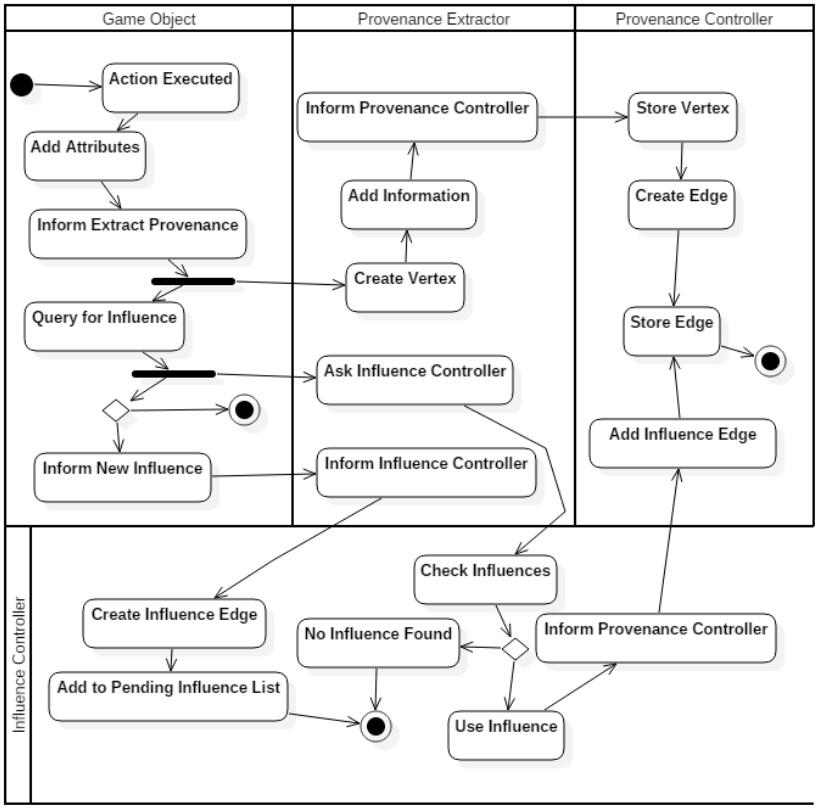
Mas como?

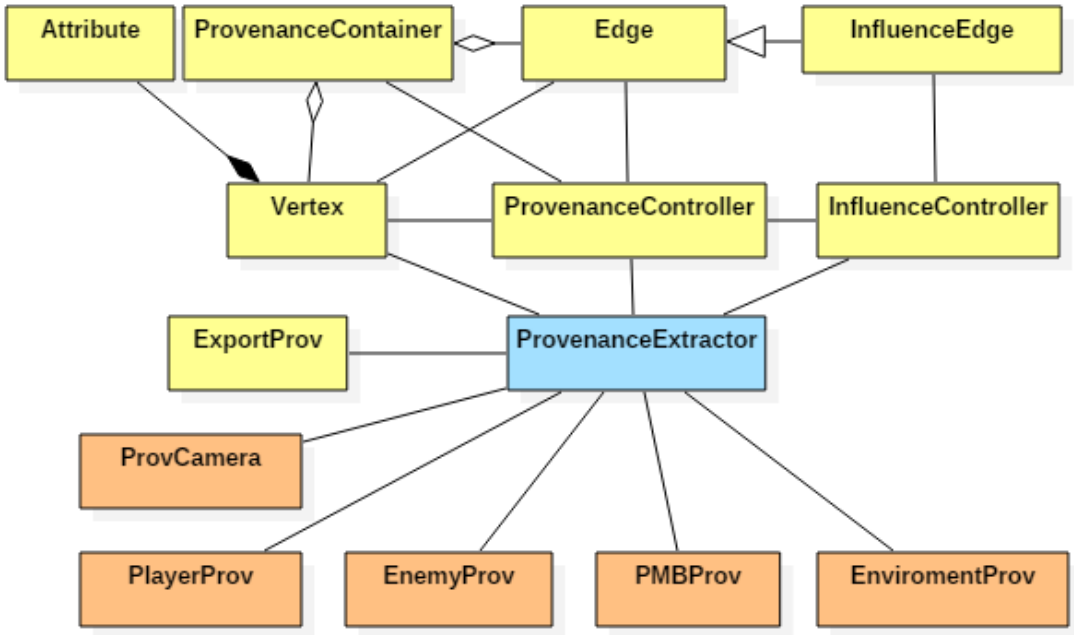
1. Import PinGU
2. Provenance controllers
   1. *Provenance Controller*
   2. *Influence Controller*
3. Game Design
   1. Identify actions
   2. Identify scripts
4. Provenance extractor
   1. *Extract Provenance*
5. Provenance tracking functions
   1. Domain specific
   2. Insert in existing scripts



PinG & PinGU

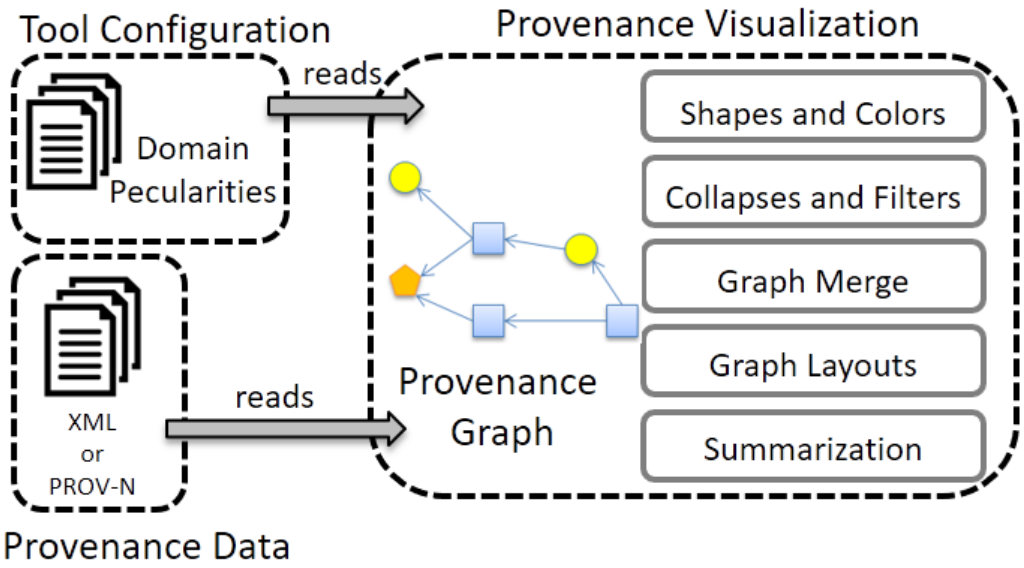
- Captura de Proveniência





Prov Viewer

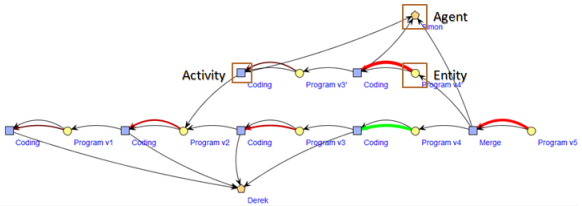
- Ferramenta interativa para grafos de proveniência



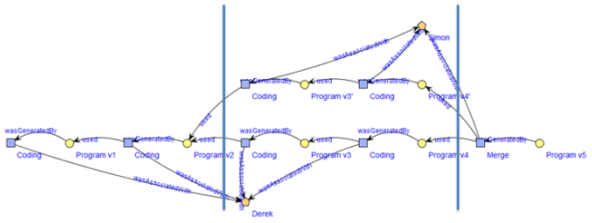
Prov Viewer

Shapes and Colors

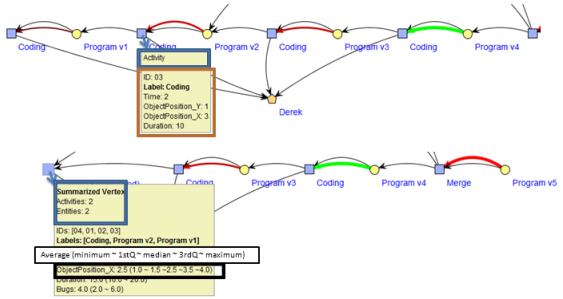
- PROV shapes
  - Pentagon: Agent
  - Circle: Entity
  - Square: Activity
- Default PROV colors
- Color Schema
  - Traffic light color based on metadata information



Temporal Filter

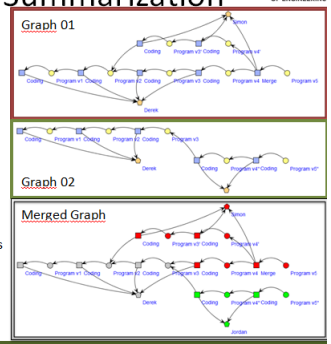


Vertex Metadata

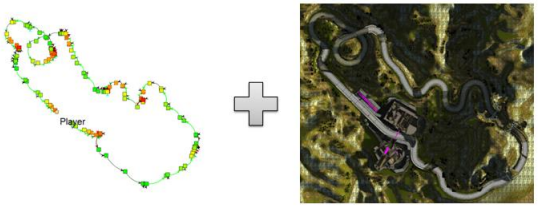


Automatic Summarization

- Combine similar vertices
  - Compare vertices
  - Detect similarity
  - No information loss
- Similarity
  - Vertex type
  - Attributes
  - Values
- Between graphs
  - Graph merge
  - Analyze multiple trials or sessions
- Within same graph
  - Sequential vertices
  - Deduplication



Spatial referencing data

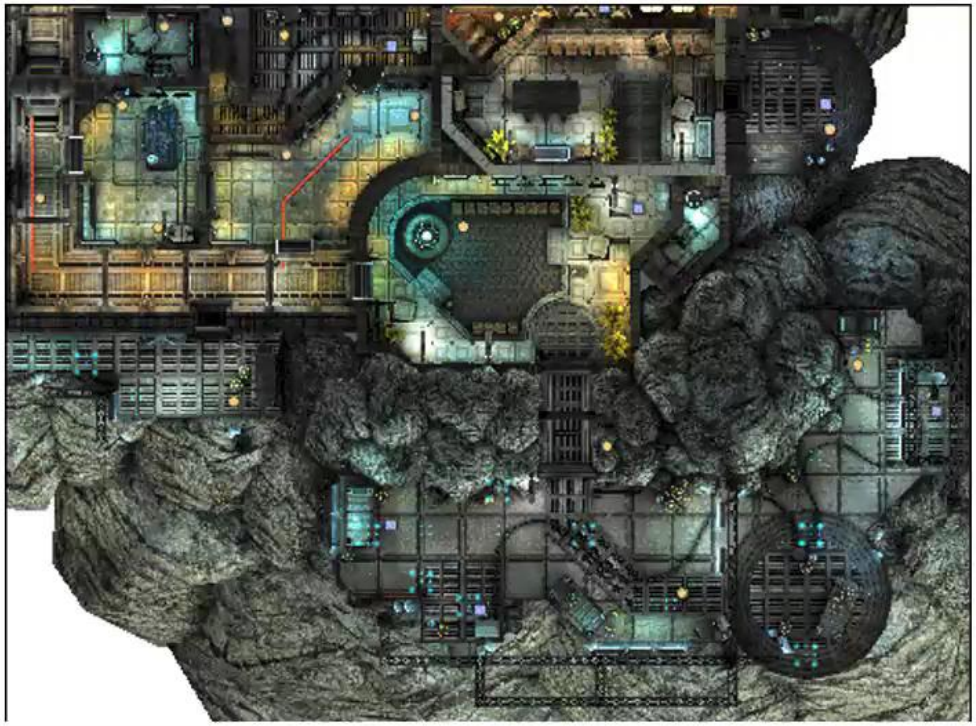


Zoom

Exemplo

- Angry Bots
  https://www.assetstore.unity3d.com/en/#!/content/12175

Grafo gerado no *Prov Viewer (Kohwalter et al., 2016)*
http://gems-uff.github.io/prov-viewer/

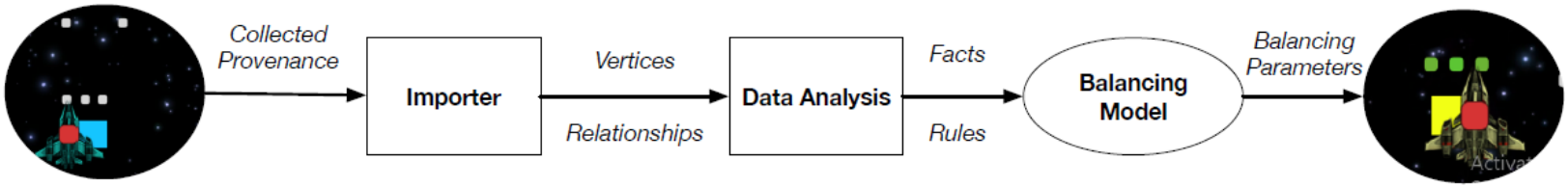


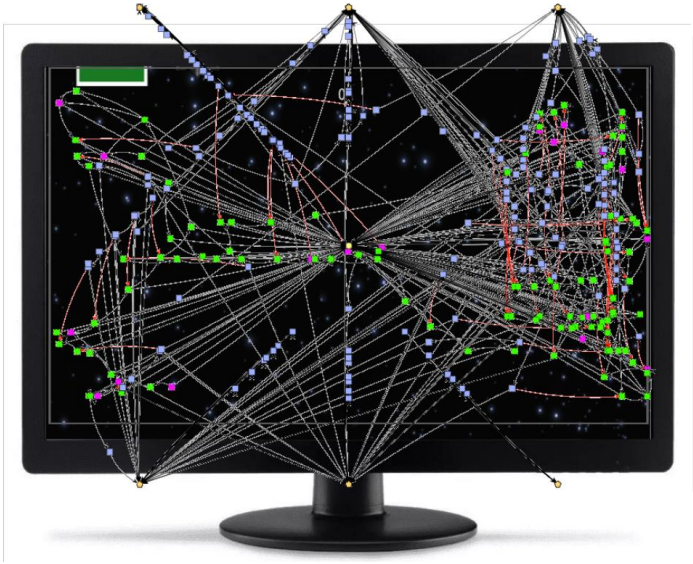
Exemplo

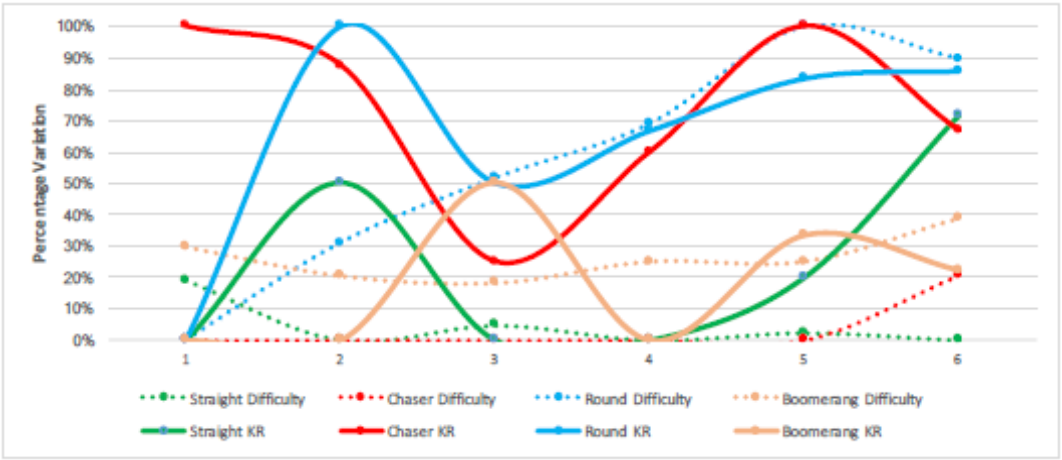
- Car Tutorial
  https://www.assetstore.unity3d.com/en/#!/content/10

Grafo gerado no *Prov Viewer (Kohwalter et al., 2016)*
http://gems-uff.github.io/prov-viewer/

Balanceamento Dinâmico



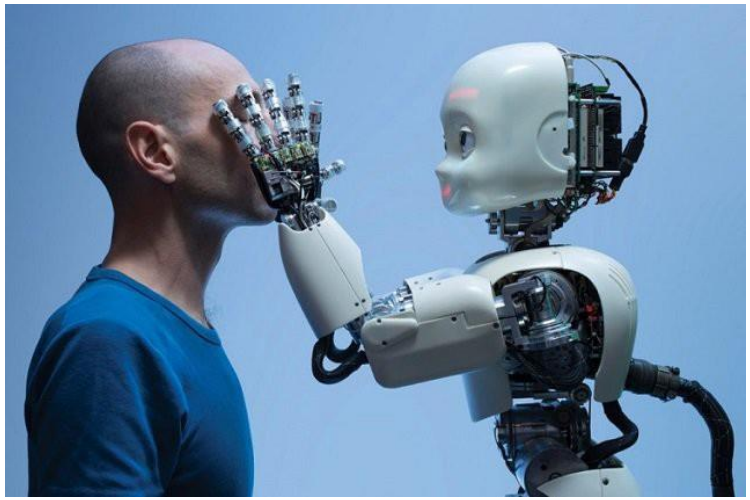




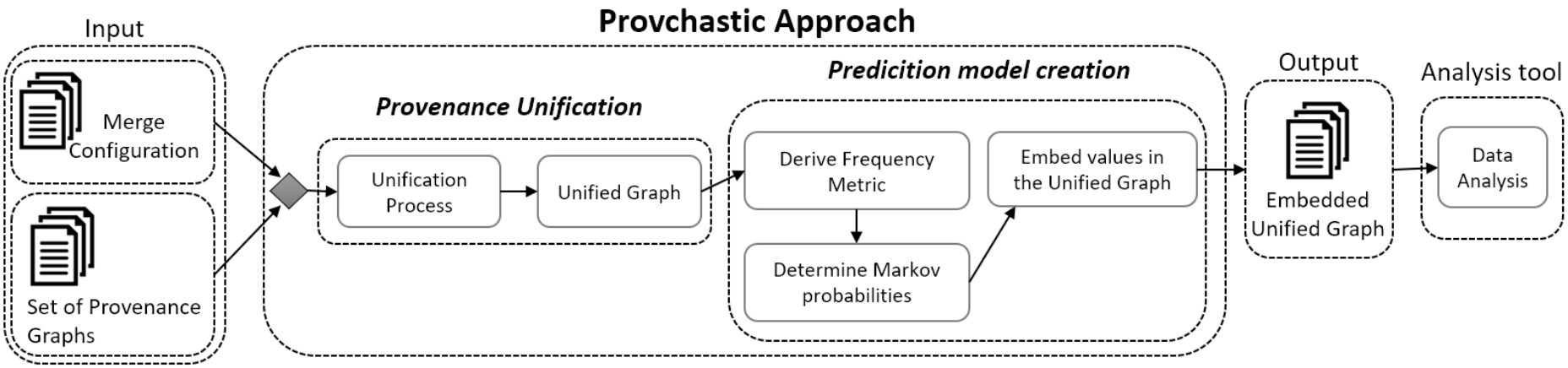
Imitation Learning

"Refere-se à aquisição por parte do agente de habilidades ou comportamentos, observando um professor que demonstra uma determinada tarefa"

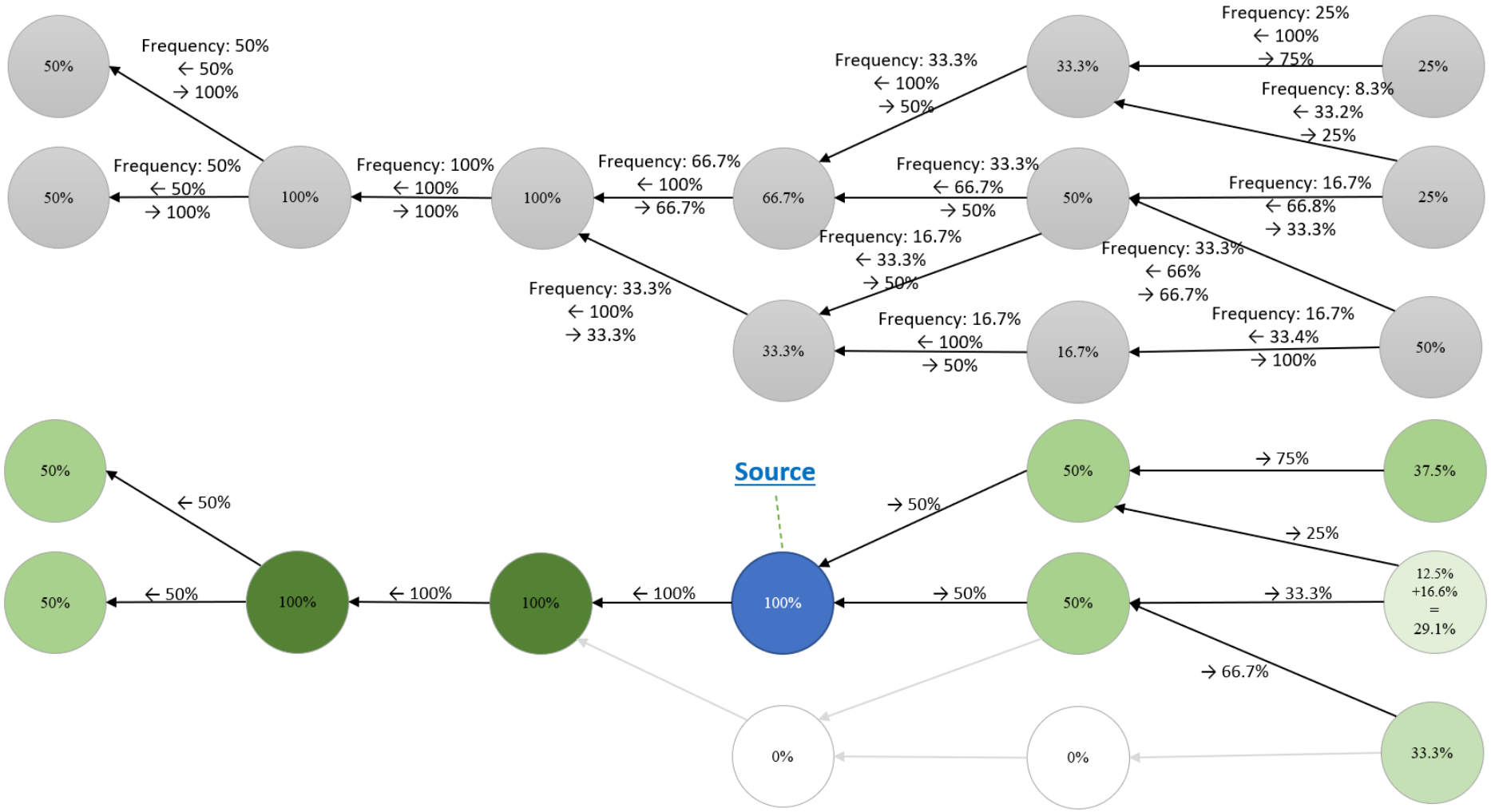
- Aprender como os jogadores jogam
  - Desenvolver agentes autônomos



Análises Preditivas

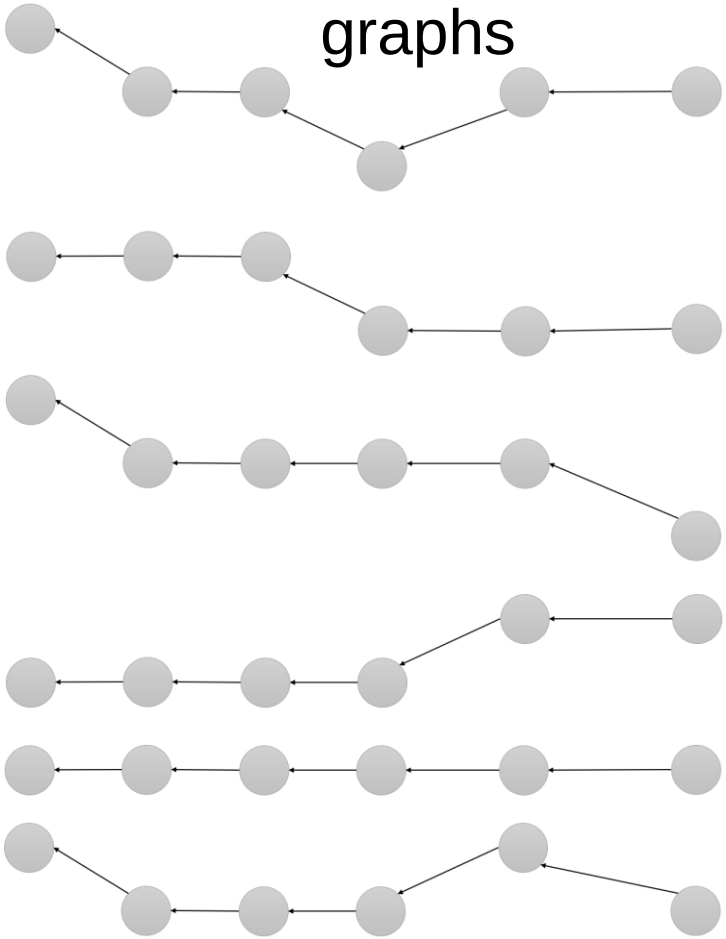


Predições curtas e distantes

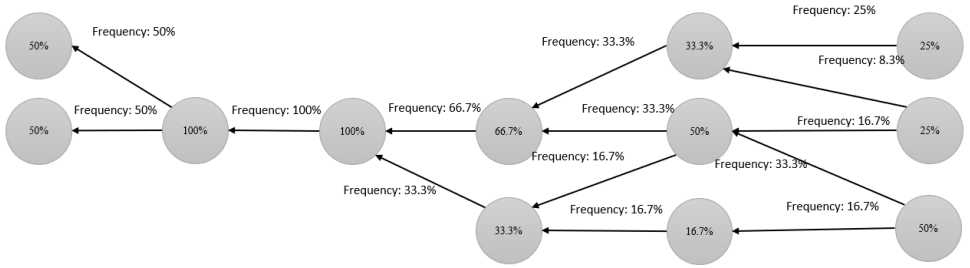


Como? Prov Unification

N provenance graphs



Unified provenance graph



noWorkflow

visTrails

PinG + Prov Viewer

DEMAIS FERRAMENTAS

Outras Ferramentas...

- Kepler
  - https://code.kepler-project.org/code/kepler/trunk/modules/provenance/docs/provenance.pdf
- Taverna
  - http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
- Galaxy
  - https://docs.galaxyproject.org/en/master/api/api.html#module-galaxy.webapps.galaxy.api.provenance

Principais referências bibliográficas

- Eynden V., et al., **Managing and Sharing Data**, 3ª edição, 2011
- **London's Global University**: https://www.ucl.ac.uk/
- **Go FAIR**: https://www.go-fair.org/
- **Prov-DM**: https://www.w3.org/TR/prov-dm/

Proveniência de Dados

Troy Costa Kohwalter

troy@ic.uff.br

Instituto de Computação UFF