7. Experimentos

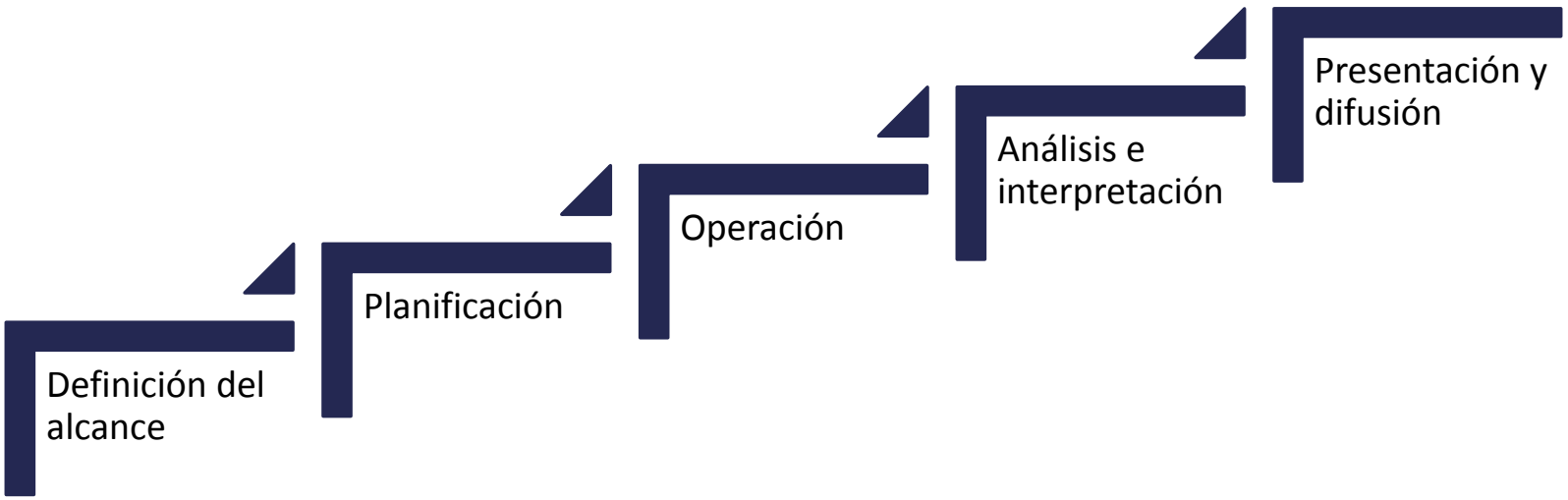
Introducción

Un experimento es un método de investigación empírica que manipula una o varias variables (denominadas variables independientes) midiendo el efecto que tienen sobre otra variable (denominada dependiente o "de respuesta")

Tipos de experimentos

- Experimentos controlados
  - Se asignan los tratamientos de forma aleatoria
- Cuasi experimentos
  - No es posible la asignación de tratamientos de forma aleatoria
- Orientados a las personas
  - Los sujetos aplican los tratamientos a los objetos de estudio
- Orientados a la tecnología
  - Se aplican herramientas o programas a los objetos de estudio

Proceso Experimental – Actividades (Wohlin 2012)



Definición del alcance

- Se definen los objetivos del experimento y las preguntas de investigación
- Resulta útil utilizar el método GQM en el cual se propone la siguiente plantilla:
  - **Analizar** *<Objeto(s) de estudio>* - ¿qué es lo que se estudia?
  - **con el propósito de** *<Propósito>* - ¿qué intención tiene el estudio?
  - **con respecto a** *<aspecto de calidad>* - ¿qué efecto se estudia?
  - **desde el punto de vista de** *<Perspectiva>* - ¿quién se ve afectado?
  - **en el contexto de** *<Contexto>* - ¿dónde, cómo, cuándo y por quién se lleva a cabo el estudio?

Planificación

1. Selección del contexto:
   - Offline vs. Online
   - Estudiantes vs. Profesionales
   - "De juguete" vs. Proyectos reales
   - Específico vs. General

Si se quisiera evaluar una tecnología emergente ¿qué contexto les parece más adecuado?

1. Formulación de hipótesis
   - Establecer hipótesis nulas (H_0) y alternativas (H_1)
   - Elegir el test estadístico más adecuado
     - Potencia del test
     - Requisitos de ese test

Planificación (cont.)

3. **Selección/identificación de variables**: dependientes, independientes, controladas, confusoras, aleatorias

3. **Selección de sujetos**: quienes van a aplicar los tratamientos (o niveles)
   - Es importante que la muestra sea grande y representativa

3. **Elección del diseño**:
   - Se deben aplicar principios de diseño de:
     - Aleatorización
     - Bloqueo
     - Balanceo o equilibrado

Aplicación de aleatorización, bloqueo y balanceo

- Tenemos que tener una variable confusora:
  - Por ejemplo, la experiencia de un programador, la cual no queremos que influya en la productividad.
- ¿Qué hacemos?
  - Aleatorización -> discusión
  - Bloqueo -> discusión
  - Balanceo/equilibrado -> discusión
- Las estrategias no son mutuo-excluyentes, se pueden aplicar varias a la vez.

Planificación – Diseño experimental

6. Elección del diseño experimental
   - Preguntas relevantes:
     - ¿Cuántas variables independientes?
       - Experimentos simples/ Experimentos factoriales
     - ¿Cuántos tratamientos por sujeto?
       - IES (Inter-sujetos) / IAS (Intra-sujetos)
     - ¿Cómo controlar los factores/variables confusoras?
       - Bloqueo/Aleatorización
     - ¿Cómo combinar los niveles de las variables ?
       - Diseño cruzado/Diseño anidado

Estas respuestas dependen de las amenazas a la validez que queramos controlar

Planificación – Instrumentación

7. Instrumentación: proporciona los medios para realizar y supervisar el experimento
   - **Objetos experimentales**: programas, diseños, especificaciones de requisitos
   - **Guías**: descripciones de procesos, templates, checklist, entrenamiento adicional
   - **Instrumentos de medición**: formularios, entrevistas, encuestas

Planificación - Evaluación de la validez

- Evaluación de la validez (factores que las influencian)
  - Validez interna: ¿El tratamiento causa "realmente" el efecto?
    - Cómo se seleccionan y agrupan los sujetos
    - Eventos inesperados
    - Materiales
  - Validez externa: ¿Pueden generalizarse los resultados obtenidos?
    - No contar con los sujetos adecuados
    - Entorno de ejecución equivocado
  - Validez del constructo: ¿Hasta qué punto las medidas seleccionadas miden las variables que aparecen en la hipótesis?
    - Que las variables elegidas no midan correctamente los conceptos
  - Validez de la conclusión: ¿Hasta qué punto las conclusiones son estadísticamente válidas?
    - Bajo poder estadístico
    - Violar suposiciones de tests estadísticos
    - Falta de fiabilidad de las medidas

Operación

1. Preparación
   - Conformar el grupo de sujetos: motivarlos e informarlos
   - Entrenamiento de sujetos
   - Se aconseja realizar un experimento de prueba (**piloto**)
2. Ejecución
   - Mismo lugar físico: facilita la recolección de datos y de que se supervise el experimento
   - Recolección de datos: desde totalmente manual hasta totalmente automatizada
   - Importante verificar que los sujetos entienden los formularios y los rellenan correctamente, están adecuadamente motivados y se aplican los tratamientos correctamente y en el orden adecuado
3. Validación de los datos
   - Validar que son “razonables” y que se han recolectado correctamente
   - Son importantes las notas que hayan surgido de la etapa de ejecución

Análisis e Interpretación

- Factores importantes al elegir las técnicas de análisis:
  - La naturaleza de los datos: cualitativa vs. cuantitativa y las escalas de medición
  - Tipo de diseño experimental
- Análisis cuantitativo
  - Estadísticos descriptivos
  - Reducción de datos
  - Contraste de hipótesis
    - Distribución normal: tests paramétricos
    - Distribución no normal: tests no paramétricos

Test estadísticos de acuerdo al diseño experimental

Tipo de diseño	Test paramétricos	Test no paramétricos
Un factor, un tratamiento		Test binomial Chi 2
Un factor, dos tratamientos	Test t Test F Test t emparejado	Mann-Whitney Chi 2 Wilcoxon
Un factor, más de dos tratamientos	ANOVA	Kruskal-Wallis Chi 2
Más de un factor	ANOVA	

Presentación y difusión

- Formas de comunicar los hallazgos:
  - Artículo en congreso o revista
  - Reporte técnico
  - Paquete para replicación
  - Material educativo
- De acuerdo al objetivo de la publicación, puede ser necesario extender con más detalle el proceso experimental
  - Martín Solari (Universidad ORT) trabajó en una propuesta de paquete de laboratorio para experimentos en Ingeniería de Software (2012) con el objetivo de brindar una estructura para la documentación de experimentos la cual favorezca la replicación y continuidad de la investigación.

Replicación de experimentos

Replicación: repetición o reproducción del experimento original

- Beneficios:
  - Permite dilucidar entre casualidad vs. causalidad
- Tipos de replicaciones
  - *Según el Sitio*:
    - Internas: experimentos repetidos por los mismos investigadores que llevaron a cabo el experimento original
    - Externas: experimentos que son realizados por otros investigadores, en otro sitio diferente
  - *Según el diseño experimental*: Idénticas, cercanas, diferenciadas o distintas
  - *Según la dependencia*: dependientes, semi-independientes, dependientes

Replicación de experimentos (cont.)

- Todos los tipos de replicaciones son válidas
  - Replicaciones diferenciadas ayudan a comprender mejor las relaciones entre las variables
- Para que las replicaciones sean efectivas se necesita además:
  - Colaboración efectiva entre los investigadores originales y los que realizarán la replicación para transmitir el “conocimiento tácito”

Familias de experimentos

Definición: Conjunto de estudios relacionados en un marco común de investigación

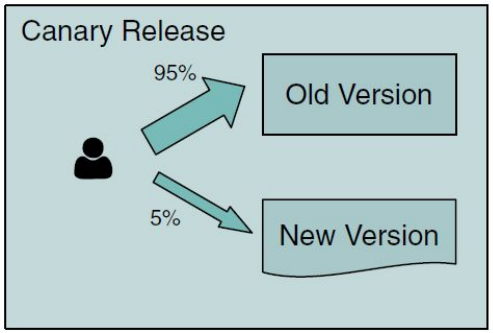
- Comparten un mismo objetivo
  - Los experimentos de la familia contribuyen a alcanzar dicho objetivo
  - Son visualizados de forma conjunta y no aislada
  - Las hipótesis de los experimentos se complementan y ayudan a generar un cuerpo de conocimiento más sólido
- Fomentan y facilitan la tarea de replicación

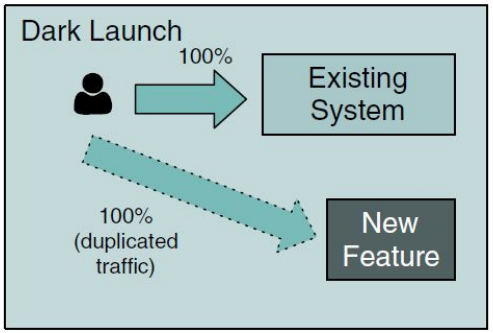
Aplicación de realización de experimentos en enfoques de desarrollo continuo

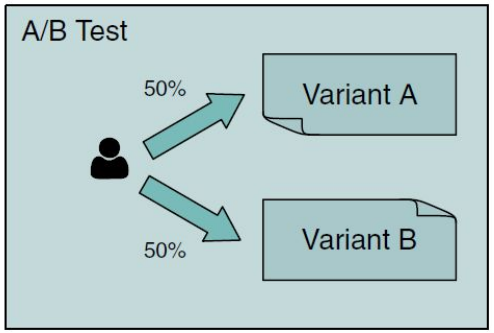
Experimentación continua

- En base al nivel de adopción puede considerarse como una "práctica" de desarrollo de software o como un "enfoque de desarrollo" propiamente dicho, en donde los datos obtenidos de la realización de experimentos guían de forma eficiente la toma de decisiones en el proceso de desarrollo de software.
- Puede verse caracterizado dentro de lo que se llama "Desarrollo dirigido por los datos" (Data Driven Development), ya que los datos que se obtienen a través de la realización de los experimentos son los que guían el desarrollo de software.

Prácticas o usos más comunes







Modelo de evolución de las organizaciones a la experimentación continua

	Category/ Phase	Crawl	Walk	Run	Fly
Technical Evolution	**Technical focus of product dev. Activities** 	**(1) Logging of signals** **(2) Work on data quality issues** **(3) Manual analysis of experiments** Transitioning from the debugging logs to a format that can be used for data-driven development.	**(1) Setting-up a reliable pipeline** **(2) Creation of simple metrics** Combining signals with analysis units. Four types of metrics are created: debug metrics (largest group), success metrics, guardrail metrics and data quality metrics.	**(1) Learning experiments** **(2) Comprehensive metrics** Creation of comprehensive set of metrics using the knowledge from the learning experiments.	**(1) Standardized process for metric design and evaluation, and OEC improvement**
	Experimentation platform complexity 	**No experimentation platform** An initial experiment can be coded manually (ad-hoc).	**Platform is required** 3rd party platform can be used or internally developed. The following two features are required: • **Power Analysis** • **Pre-Experiment A/A testing**	**New platform features** The experimentation platform should be extended with the following features: • **Alerting** • **Control of carry-over effect** • **Experiment iteration support**	**Advanced platform features** The following features are needed: • **Interaction control and detection** • **Near real-time detection and automatic shutdown of harmful experiments** • **Institutional memory**
	Experimentation pervasiveness 	**Generating management support** Experimenting with e.g. design options for which it's not a priori clear which one is better. To generate management support to move to the next stage.	**Experiment on individual feature level** Broadening the types of experiments run on a limited set of features (design to performance, from performance to infrastructure experiments)	**Expanding to (1) more features and (2) other products** Experiment on most new features and most products.	**Experiment with every minor change to portfolio** Experiment with any change on all products in the portfolio. Even to e.g. small bug fixes on feature level.
Organizational Evolution	**Engineering team self-sufficiency** 	**Limited understanding** External Data Scientist knowledge is needed in order to set-up, execute and analyse a controlled experiment.	**Creation and set-up of experiments** Creating the experiment (instrumentation, A/A testing, assigning traffic) is managed by the local Experiment Owners. Data scientists responsible for the platform supervise Experiment Owners and correct errors.	**Creation and execution of experiments** Includes monitoring for bad experiments, making ramp-up and shut-down decisions, designing and deploying experiment-specific metrics.	**Creation, execution and analyses of experiments** Scorecards showing the experiment results are intuitive for interpretation and conclusion making.
	Experimentation team organization 	**Standalone** Fully centralized data science team. In product teams, however, no or very little data science skills. The standalone team needs to train the local product teams on experimentation. We introduce the role of Experiment Owner (EO).	**Embedded** Data science team that implemented the platform supports different product teams and their Experiment Owners. Product teams do not have their own data scientists that would analyse experiments independently.	**Partnership** Product teams hire their own data scientists that create a strong unity with business. Learning between the teams is limited to their communication.	**Partnership** Small data science teams in each of the product teams. Learnings from experiments are shared automatically across organization via the institutional memory features.
Business Evolution	**Overall Evaluation Criteria (OEC)**	OEC is **defined** for the first set of experiments with a few key signals that will help ground expectations and evaluation of the experiment results.	**OEC evolves** from a few key signals to a structured set of metrics consisting of Success, Guardrail and Data Quality metrics. Debug metrics are not a part of OEC.	OEC is **tailored** with the findings from the learning experiments. Single metric as a weighted combination of others is desired.	OEC is **stable**, only periodic changes allowed (e.g. 1 per year). It is also used for setting the performance goals for teams within the organization.

Herramientas para apoyar a la definición de experimentos

- Kit de hipótesis para A/B testing :
  - http://experimentationhub.com/hypothesis-kit.html
- Provee un framework para la definición de experimentos del tipo A/B testing e hipotesis driven experimentation