Software Engineering: Science or Art

Ingeniería del Software: Ciencia o Arte

Cortis Cooke
Oregon State University
cooke@eecs.orst.edu

Greggie Rothermeli
Oregon State University
grotheri@eecs.orst.edu

(Artículo de **REFLEXIÓN**. Recibido el 12-07-2010. Aprobado el 20-10-2010)

Abstract *— If we could count the times we've done, or we have made, the question of whether Software Engineering is science or art, it would take several sheets of paper to take notes and answer. The answer always varies depending on the person and the environment in which to develop the repeated scenes. In this study we conclude that, no matter who asks or who respond or where this discussion, in reality the answer is "both".*

Keywords*: software engineering, evidence, evidential force.*

Resumen — Si pudiéramos contabilizar las veces que nos hemos hecho, o nos han hecho, la pregunta de si la Ingeniería del Software es ciencia o es arte, serían necesarias muchas hojas de papel para anotarlas y responderlas. La respuesta siempre varía dependiendo de la persona y el entorno en el cual se desarrollan las repetidas escenas. En este trabajo llegamos a la conclusión de que, sin importar quien pregunte o quien responda o dónde se presente la discusión, en realidad la respuesta es: "ambas".

Palabras clave: Ingeniería del Software, evidencia, fuerza evidencial.

1. Introducción

Lo que distingue a la ciencia del arte es la forma en que nosotros, como gerentes y profesionales, tomamos decisiones: mediante la formulación de argumentos racionales a partir de las evidencias que tenemos —evidencias que provienen tanto de nuestra experiencia como de la investigación relacionada. Es decir, pasamos de una parte a un todo examinando el cuerpo de la evidencia, para determinar lo que sabemos acerca de la mejor forma de desarrollar un buen producto software. Este punto de vista no es particular de la ingeniería del software e incluso de las ciencias matemáticas, es lo que caracteriza en general a la buena ciencia (Wagner, 19912).

En este artículo se examinan las formas en las que la comprensión cuidadosa de la argumentación y de las evidencias puede conducir a una ingeniería del software más efectiva —y en última instancia a una mejor toma de decisiones, y a procesos y productos de software de mayor calidad.

2. El arte de la ciencia

Como profesionales, muchos tratamos de ampliar nuestros horizontes y experiencias con base en la lectura de revistas y artículos científicos acerca de las tecnologías del software: paradigmas, procesos, técnicas y herramientas, que utilizamos para especificar, diseñar, construir, probar, usar y evaluar nuestros productos software. Los estudios presentados y reportados en revistas y actas de congresos representan un rico y creciente cuerpo de evidencias sobre muchos aspectos del desarrollo y evolución del software; pero estas evidencias no son suficientes; hay que utilizarlas más eficientemente para construir argumentos y preparar casos acerca del qué hacer y qué evitar.

Por lo general, una acción de este tipo consta de tres partes:

1. Una o más solicitudes que un conjunto de propiedades debe cumplir
2. Un cuerpo de evidencias —de fuentes diversas— que soportan las solicitudes
3. Un conjunto de argumentos que vinculan las solicitudes con las evidencias.

Los investigadores de la ingeniería del software a menudo se centran demasiado en la producción de evidencias y muy poco en la construcción de los argumentos asociados. Así como a los profesionales, nos es difícil saber qué leer, qué creer y cómo unir las piezas. Además, a veces suponemos que una pieza de evidencia es muy similar a otra. De hecho, como veremos, hay formas de evaluar cada pieza de la evidencia y de usar características evidenciales para construir argumentos más sólidos y convincentes.

Para ello, debemos reconocer que podemos utilizar las evidencias de dos maneras distintas: para generar hipótesis y para probarlas. La necesidad de establecer y probar hipótesis plantea cuestiones clave que debemos responder para poder comprender el papel de las evidencias:

- ¿Qué queremos decir cuando hablamos de una tecnología "funcional"? Antes de utilizar los resultados de un estudio para evaluar si una tecnología es efectiva, debemos conocer cuál es la efectividad de tales medios. En particular, debe haber alguna manera medible o demostrable para probar que se utilizó apropiadamente la tecnología.

- ¿Qué tipos de evidencias —y cuántas— tenemos que demostrar que funcionan? Algunas tecnologías muestran sus efectos incluso antes de que el software se entregue, pero algunas —como las que afectan la fiabilidad— pueden ser evaluadas sólo después de que el sistema está en uso. Tenemos que reunir evidencias para demostrar los efectos, tanto antes como después. Estas evidencias también deben permitir la comparación entre el uso y no uso de la tecnología (Pfleeger & Kitchenham, 1994-1995)

- ¿Quién proporciona y quién revisa la evidencia? Muchos proveedores aportan evidencias de la efectividad de sus productos, y muchos investigadores están ansiosos por demostrar que sus nuevas ideas o herramientas pueden marcar una diferencia positiva para los desarrolladores o los usuarios. Pero este afán pueden sesgar los resultados, aun cuando los vendedores y los investigadores hacen grandes esfuerzos por evitarlo. Casi siempre es preferible una evaluación independiente.

- Si una tecnología funciona en un dominio, ¿para qué nos hablan de otros dominios? Las evidencias recogidas en un contexto o entorno puede que no se apliquen en otros.

- ¿Cómo puede informarnos una evidencia acerca de las ventajas y desventajas de utilizar una tecnología imperfecta en lo social, lo económico o lo político? Ninguna tecnología es perfecta, y tenemos que ser capaces de utilizar evidencias para tomar decisiones acerca de los riesgos involucrados en la adopción de imperfecciones. En particular, la evidencia debe ser compatible con las decisiones políticas y las estrategias de mitigación de riesgos.

3. El papel del riesgo y la incertidumbre

Nos gusta pensar de la ciencia con certidumbre, donde los resultados son claros cuando se entienden y aceptan las reglas. Pero la ciencia está llena de incertidumbres, y debemos reconocer su papel y los riesgos derivados que tomamos, tanto al generar evidencias como cuando la usamos para construir argumentos. Los abogados reconocen la incertidumbre asociada con diversos tipos de evidencias, por lo que buscan las piezas de las evidencias que en conjunto tienen más "valor probatorio" que cuando se usan por separado. Por ejemplo, David Schum (1994) identifica cuatro categorías distintas de evidencias, algunas con más incertidumbre que otras:

1. *Evidencia tangible*, que puede ser examinada directamente para ver qué revela. Algunos ejemplos incluyen objetos —como el código—, documentos —como las especificación de requisitos—, imágenes —como el registro de sensores—, medidas —como las líneas de código—, y tablas —como histogramas o pulsaciones por período de tiempo.

2. *Evidencia testimonial*, que es entregada por una persona que reporta acerca de lo ocurrido. Existen dos tipos de evidencias testimoniales:

   - *Evidencia testimonial inequívoca*, que consiste en la observación directa o en un rumor: "Vi que el usuario introdujo esta secuencia de pulsaciones de teclas, con lo cual el sistema falló"; o "Ana me dijo que los informes se imprimieron incorrectamente cuando se utilizaron los datos de la semana pasada".

   - *Evidencia testimonial dudosa*, que es probabilística. "Juan cree que el fallo del sistema puede estar relacionado con la inusual carga que ha tenido esta semana"; o "Gabriel no está seguro de si elimina los archivos antiguos antes de crear los nuevos".

3. *Evidencias ausentes*, pueden ser tangibles o testimoniales, y pueden ser útiles para soportar un argumento. Por ejemplo, el no descubrir registro de defectos durante las pruebas unitarias puede ser evidencia de pruebas unitarias inadecuadas. Del mismo modo, la ausencia de un informe de revisión de requisitos podría significar que nadie realizó la revisión de requisitos.

4. *Hechos aceptados*, que son registros autorizados, tales como incidencias o resultados de pruebas de aceptación. Estos hechos suelen tener credibilidad sustancial, ya que son el resultado de procedimientos seguidos y verificados cuidadosamente.

En lugar de estar consternados por la naturaleza de la incertidumbre y la variabilidad evidencial, podemos utilizar la incertidumbre a nuestro favor. La fuerza evidencial es el grado en que cada pieza de evidencia contribuye a o disminuye al argumento que utiliza. Por ejemplo, dos piezas de evidencia débil juntas podrían ser más fuertes que cualquiera de ellas por separado. O una pieza de evidencia podría anular otra, si tuviera resultados opuestos obtenidos en las mismas circunstancias. Schum (1994) ofrece algunas directrices acerca de cómo determinar la naturaleza y el grado de fuerza evidencial de una determinada evidencia. Por ejemplo, para un determinado resultado podemos hacer varias preguntas, con base en las cinco características evidenciales representadas en la Fi. 1.

En primer lugar, debemos determinar el tipo de evidencia: ¿Está documentada o sólo es un rumor?

¿Existen detalles suficientes como para que el estudio se pueda replicar? ¿Están claramente definidas y medidas las variables importantes? Y ¿Fue bien diseñado el estudio? Estas directrices pueden ayudar a determinar el grado en que cada pieza de evidencia está sólidamente diseñada e implementada (Kitchenham *et al.*, 2002).

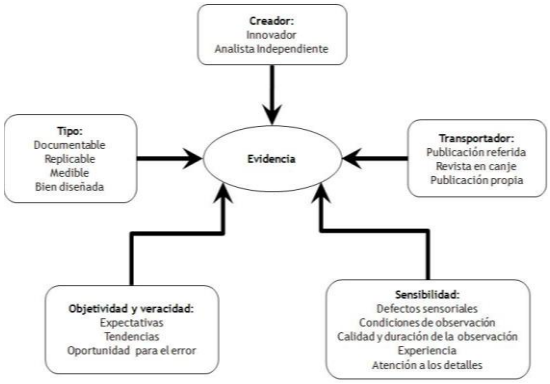

Fig. 1. Características para evaluar la contribución a la fuerza evidencial (Adaptado de Schum (1994))

A continuación, nos fijamos en quién creó la evidencia. Si fue diseñada y producida por el innovador cuya tecnología se está evaluando, la evidencia tiene menos fuerza evidencial que si fuera producto de un análisis independiente. Del mismo modo, tomamos nota del portador de la evidencia. Una publicación referenciada tiene más credibilidad que una revista profesional —que puede tener el creador de la tecnología como un anunciante o colaborador—, que a su vez tiene más credibilidad que un artículo auto-publicado —al igual que muchas publicaciones en la Web.

Como profesionales, podemos determinar la credibilidad evidencial, en parte, mediante un examen profundo al mismo estudio para ver qué tan sensible es a un error. Por ejemplo, podría ser objeto de defectos sensoriales, como cuando dos observadores toma notas diferentes del mismo resultado. Las condiciones de observación pueden ser diferentes, como la calidad y la duración de la observación, la experiencia de los observadores, y su atención al detalle. Por ejemplo, escribir reportes de fallas cuando un producto es nuevo, a menudo es más detallado que los que se escriben meses más tarde, cuando otros productos capturan cada vez más la atención y el equipo del proyecto se está asignando a otras tareas (Pfleeger & Hatton, 1997).

Por último, debemos entender el grado de objetividad y veracidad de los investigadores. A veces, los investigadores están ansiosos por mostrar resultados positivos, incluso cuando tratan de diseñar un estudio objetivo. Esta situación es particularmente cierta para los candidatos a doctorado y para los creadores de nuevas tecnologías, que prefieren resultados positivos en lugar de negativos. Una vez más, a mayor objetividad mayor fuerza evidencial.

4. Formando un todo

La conclusión es que debemos ver un cuerpo de evidencias en el contexto de nuestra experiencia y nuestra necesidad para soportar la toma de decisiones. A pesar de que una pieza determinada de evidencia tiene su propia fuerza evidencial, en última instancia debe considerarse en el contexto de cómo contribuye a la discusión general que se está realizando. Lo ideal sería que cada pieza de evidencia se sumara a la fuerza global del argumento, pero en la práctica ahí hay problemas. A veces un estudio tiene dudosa credibilidad, especialmente cuando no es fácilmente replicable. En otras ocasiones, la evidencia no existe, a menudo cuando la información del propietario está oculta. La evidencia puede ser ambigua, particularmente cuando hay variables confusas que enturbian el poder explicatorio de las variables independientes. Y la evidencia puede estar en conflicto, como en el debate acerca de si los equipos de inspección son necesarios.

Este tipo de conflicto no se limita a los estudios de ingeniería del software. El campo de la medicina, generalmente considerado el estándar dorado de la ciencia para producir y evaluar evidencias, proporciona muchos ejemplos de estudios conflictivos. Por ejemplo, Gina Kolata (2003) informó de un grave conflicto: hace algún tiempo un estudio en salud bien diseñado indicó que la terapia de reemplazo hormonal —HRT— ayudaba a proteger a las mujeres menopáusicas contra las enfermedades del corazón; pero un estudio más reciente realizado por the Women's Health Initiative en the US National Institutes of Health sugiere lo contrario: que la HRT incrementa el riesgo. El último estudio fue tan convincente que las mujeres abandonaron su medicación HRT en el verano de 2002.

La idea de la fuerza de un argumento no es nueva. En 1827, Jeremy Bentham propuso la utilización de una escala numérica para evaluar los argumentos jurídicos, "fuerza inferencial". Él valoraba cada pieza de evidencia desde -10 hasta +10. Un número positivo significa que la evidencia favorecía la hipótesis planteada, y un número negativo pesaba en su contra —cero significa que no hubo influencia. Para ayudar a asignar la valoración, Bentham hacía cuatro preguntas sobre cada pieza de evidencia:

1. ¿Qué tan seguro está el testigo respecto de la verdad del evento que afirma?
2. ¿Cómo se conforma el evento respecto a la experiencia de la audiencia general? Es decir, ¿qué tan raro es que acurra?
3. ¿El testigo es digno de confianza?
4. ¿El testimonio es apoyado o refutado por otras evidencias?

Schum (1994) amplía las ideas de Bentham con un análisis estadístico para sugerir maneras de evaluar la fuerza de un argumento. En primer lugar, describe la *clasificación de evidencias*: formas de elicitar información para minimizar la tendencia, corroborar hechos, y mejorar la credibilidad de cada pieza de evidencia siempre que sea posible. Añade el *análisis Bayesiano*, que permite expresar una pieza de evidencia utilizando el grado de incertidumbre con que se asocia. Schum conecta las piezas de las evidencias utilizando una representación formal de las *cadenas de razonamiento*. Estas representaciones gráficas de los argumentos permiten ver cómo cada pieza de evidencia se relaciona con las afirmaciones hechas. Por último, la cadena de evidencias puede servir de base para *medidas de probabilidad*: la probabilidad de que la hipótesis sea verdadera, dada la evidencia.

5. Argumentos de múltiples fuentes

Bloomfield y Littlewood (2003) describen la fuerza evidencial en términos de la naturaleza de cada pieza de evidencia. Señalan que los argumentos con diversidad de evidencias son más fuertes que los argumentos con múltiples repeticiones de la misma clase de evidencia. Su trabajo fue motivado por la necesidad de una evidencia fuerte cuando se toman decisiones acerca de los sistemas de seguridad crítica. Por ejemplo, es posible reforzar evidencias basadas en procesos sobre la fiabilidad probable de un sistema —como una revisión de las prácticas— mediante la adición de evidencias basadas en el producto —como el análisis de código estático.

Del mismo modo, el UK Defence Standard 00-55 describe un argumento con dos partes: una demanda de inferencia lógica basada en pruebas y una demanda probabilística basada en el análisis estadístico. Este "argumento multi-fuentes", donde cada fuente maneja un tipo diferente de evidencias, puede ser más fácil de analizar que un argumento completo. Por otra parte, las fuentes no tienen que ser independientes. Las fuentes extras suelen proporcionar más confianza que una fuente aislada, pero el costo adicional de la confianza extra se debe justificar. Lars Bratthall y Jørgensen (2002) toman exactamente este enfoque al investigar los plazos y factores clave de éxito para mantener en alta disponibilidad un sistema de comercio electrónico; mostraron que un estudio de caso con múltiples fuentes de datos tiene mayor validez que cualquiera de las evidencias de una fuente de datos aislada.

Cuando se trabaja en proyectos de software con plazos estrictos, podríamos tener la tentación de buscar una sola pieza o un tipo de evidencia para informar nuestras decisiones. Pero hay dos razones de peso para buscar diversidad evidencial: 1) la primera radica en la naturaleza de los modelos que construimos para soportar nuestros argumentos. Cada modelo tiene algunas presunciones subyacentes, y estos supuestos pueden, de alguna manera, ser débiles. Por ejemplo, podemos utilizar métodos formales para demostrar que una especificación es correcta; pero la prueba será una pieza de evidencia fuerte sólo si la especificación es una representación exacta de los requisitos de más alto nivel. 2) La evidencia misma podría tener debilidades innatas. Por ejemplo, a menudo utilizamos resultados de pruebas para respaldar los argumentos acerca de la fiabilidad, pero la prueba completa rara vez es posible, por lo que la evidencia sólo será tan fuerte como el rigor y la exactitud de la prueba.

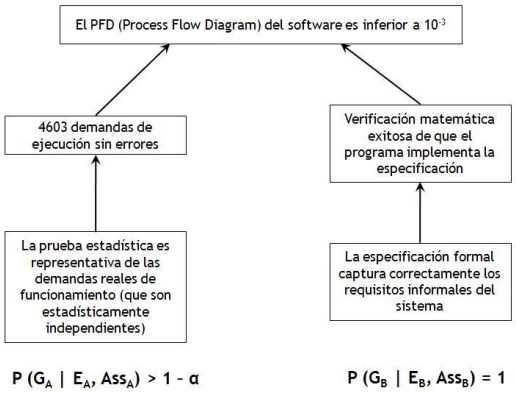

Fig. 2. Un ejemplo de la argumentación acerca de un objetivo de seguridad (Adaptado de Bloomfield & Littlewood (2002))

Cuando se construyen cuidadosamente piezas fuertes y diversas, el argumento general se hace más fuerte, para que podamos tener más confianza en nuestras decisiones. Bloomfield y Littlewood (2003) ilustran esta idea con el ejemplo que se muestra en la Fig. 2. El cuadro en la parte superior es una aserción acerca de un sistema de software. Cada uno de los dos cuadros debajo de él representa evidencias de que la aserción es correcta. Pero los cuadros en la parte inferior contienen suposiciones sobre la evidencia que deben ser ciertas para que la evidencia sea útil. La evidencia y los argumentos forman un árbol, y podemos asociar una probabilidad a cada rama con base en su naturaleza. La rama izquierda en la Fig. 2 —rama A— muestra la probabilidad de que la aserción (G_A) esté soportada, dada la evidencia (E_A) y la correctitud de la suposición acerca de la evidencia (Ass_A). Del mismo modo, la rama B tiene una aserción con evidencia (E_B) y la suposición (Ass_B). De esta manera, un argumento multi-fuente puede ser más fuerte que cualquiera de sus fuentes solas.

Podemos poner estos enfoques juntos, lo que permite evaluar la fuerza de un argumento. Para cada argumento, debemos considerar cinco cosas:

1. La extensión de las evidencias que soportan la aserción

2. Nuestra confianza en los supuestos acerca de cada pieza de evidencia
3. La dificultad en la asignación de valores numéricos a las evidencias medibles y las evaluaciones de probabilidad
4. La necesidad de simplificar los supuestos en los modelos subyacentes
5. La contribución de cada pieza de evidencia al conjunto completo.

Ahora, examinemos dos ejemplos para ver lo que está involucrado en la aplicación de estas técnicas a evidencias imperfectas.

5.1 Un ejemplo en medicina

El trabajo de la Corporación RAND —Research ANd Development—, de interpretar evidencias para aportar a la discusión acerca de las decisiones políticas, ofrece muchas lecciones evidenciales de las que los ingenieros de software pueden aprender. Por ejemplo, durante el tiempo que la efedra fue popular como suplemento dietético, se presentaron por lo menos 18000 reportes acerca de sus efectos adversos, incluyendo muerte y enfermedades. RAND examinó las evidencias para determinar si soportaban o refutaban los reportes (Shekelle *et al.*, 2003). La US Food and Drug Administration no tiene a los suplementos alimenticios sujetos a los mismos estándares rigurosos de los medicamentos. En su lugar, sólo busca evidencias de que no exista "riesgo significativo o irrazonable de lesión o enfermedad". Así, los fabricantes no tienen que demostrar evidencias de la seguridad de la efedra antes de que se lleve al mercado. Por lo tanto, no existe un cuerpo de evidencias para demostrar con certeza científica que la efedra es segura. Entonces, ¿cuál era el estado de la evidencia? Existían reportes en varias formas y desde varias fuentes:

- 52 reportes publicados y no publicados de los ensayos acerca de la pérdida de peso o disminución del rendimiento deportivo
- 1820 reportes de quejas de los consumidores a la US Food and Drug Administration
- 71 reportes en la literatura médica
- 15951 reportes proporcionados por Metabolife, un fabricante de suplementos que contienen efedra.

Algunos de los 52 reportes incluían sólo un pequeño número de personas; otros evaluaban el uso de la sustancia durante un corto período de tiempo; y otros tenían limitaciones como el uso de una muestra no representativa. Los diversos reportes reflejaban estudios que los investigadores realizaron en diferentes lugares, por diferentes razones. Tenían:

- *Objetivos diferentes*: pérdida de peso, rendimiento deportivo, y efectos adversos para la salud
- *Compuestos químicos diferentes*: la efedrina química, la hierba efedra, y la efedrina o efedra combinada con otras sustancias químicas que podrían haber afectado el resultado
- *Diseños de estudio diferentes*
- *Tratamientos diferentes*
- *Medidas de resultado diferentes*, como ejercicios diferentes.

RAND ejecutó tres pasos para hacer frente a la evidencia como un solo cuerpo de la información. En primer lugar, los investigadores de RAND establecieron *criterios de confianza* en la evidencia. Para hacer frente a la pérdida de peso, seleccionaron estudios que evaluaran la efedra, efedrina o efedrina plus, además de otros compuestos. De estos, eligieron sólo los estudios con un período de al menos ocho semanas, y descartaron los estudios con otras serias limitaciones, lo que redujo el número final de los estudios a 20. Del mismo modo, restringieron los estudios de rendimiento deportivo a los que evaluaron la efedra, efedrina o efedrina plus además de otros compuestos, como la cafeína, que se utilizan para mejorar el rendimiento atlético. Por último, limitaron los estudios de seguridad a los que tenían documentación de un evento adverso, incluyendo la documentación sin otras causas posibles. Esto dejó 284 estudios posibles de estudiar.

El segundo paso de la evaluación incluyó la *categorización de tratamientos*. Compararon todos los estudios de pérdida de peso en seis categorías —Fig. 3—, mientras evaluaban los estudios de rendimiento deportivo por separado, ya que cada prueba incluía un tipo diferente de ejercicio.

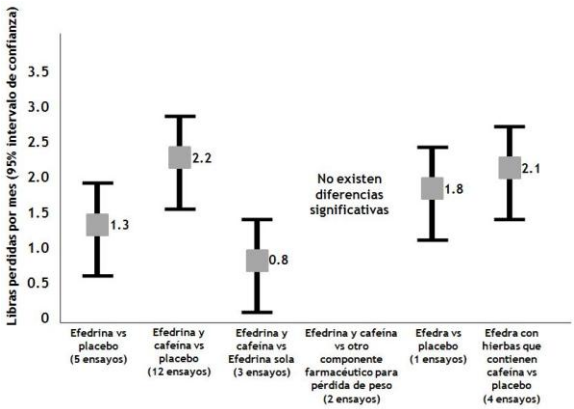

Fig. 3. Resultados del análisis de pérdida de peso (Cortesía RAND)

El paso preparatorio final en la evaluación incluyó *establecer medidas de resultado*. Los estudios de pérdida de peso indicaban pérdida de peso como una reducción en libras o kilogramos. Los estudios de rendimiento deportivo midieron algún indicador de rendimiento en el ejercicio, tales

como el consumo de oxígeno, tiempo hasta el agotamiento, la producción de dióxido de carbono, la resistencia muscular, o el tiempo de reacción. Los estudios de seguridad agruparon los síntomas en categorías de efectos clínicamente similares.

La Fig. 3 muestra que, a pesar de los diferentes tipos de estudios y resultados, utilizar efedra o efedrina, en efecto da lugar a pérdida de peso. Un análisis similar indica que pueden aparecer reacciones adversas. Posteriormente, a finales de 2003, la Federal Drug Administration prohibió el uso de la efedra y la efedrina en los EE.UU. (Kaufman, 2003).

5.2 Un ejemplo en desarrollo de software

Podemos aplicar estos pasos —*fijar criterios de confianza, categorizar tratamientos,* y *establecer medidas de resultados*— a los importantes pero imperfectos componentes de evidencias del desarrollo de software, lo que lleva a una mejor toma de decisiones. Para ver cómo, considere la abundante literatura acerca de las revisiones e inspecciones. Existe una multiplicidad de estudios incluso para un simple tipo de revisión, como la revisión de requisitos. Por ejemplo, Regnell, Runeson y Thelin (2000) analizaron algunos estudios de caso publicados y basados en escenarios. En estos estudios, los diferentes encuestados utilizan diferentes enfoques para leer documentos de requisitos, y para buscar defectos que eventualmente podrían conducir a errores de diseño o de código. Al igual que con los estudios de la efedra, el número de sujetos de cada estudio fue pequeño —el más grande tenía 66 participantes—, y muchos aspectos de estudio variados:

- *Objetivos diferentes*. Algunos miraban como técnica la efectividad, otros miraban la eficiencia.

- *Poblaciones diferentes*. Algunos utilizaron a estudiantes, otros utilizaron a profesionales. En cada caso, el número de años y tipo de experiencia variaron.

- *Diseños de estudio diferentes*. Algunos estudios incluyeron varios equipos diferentes, otros simularon artificialmente efectos de equipo construyendo diferentes combinaciones de resultados. Algunos estudios incluyeron una reunión del equipo, mientras que otros no lo hicieron.

- *Tratamientos diferentes*. Algunos estudios incluyeron análisis basados en defectos, en las que cada revisor buscó un tipo particular de defecto. Otros utilizan un enfoque basado en la perspectiva, donde cada revisor actuó como un usuario, diseñador, o probador. Algunos compararon la técnica aplicada con listas de verificación o ad hoc, algunos con listas de control solamente, algunos sólo con técnicas ad hoc, y algunos compararon dos variaciones de análisis basados en la perspectiva.

- *Medidas de resultado diferentes*. Un conjunto de medidas incluyó el tiempo que cada revisor emplea en la preparación para la revisión, la realización de la revisión, o la participación en la reunión del equipo. Otro conjunto registró el número de defectos. La eficiencia se midió como el número de defectos por unidad de tiempo, y la eficacia como el número de defectos encontrados en un porcentaje de todos los defectos conocidos.

Sin embargo, a diferencia de los estudios de la efedra, generalmente los resultados no muestran que los estudios basados en escenarios sean efectivos. Will Hayes (1999) realizó un meta-análisis de algunos de los estudios y señaló que algunos de los resultados se contradecían entre sí. Incluso, las repeticiones de algunos de los mismos "paquetes de laboratorio" no produjo resultados consistentes; Regnell, Runeson, y Thelin (2000) replicaron parcialmente los experimentos previos basados en las perspectivas, y no encontraron diferencias significativas entre perspectivas en la tasa de detección de defectos, el número de defectos encontrados por hora, o la cobertura de defectos.

Debido a que muchos de los estudios replican otros —usando diferentes grupos de estudiantes pero con los mismos documentos y protocolos—, es necesario un argumento multi-fuente para demostrar el grado en que los trabajos se basan en escenarios. Es decir, podría ser el momento para buscar otro tipo de evidencias. Por ejemplo, los investigadores pueden realizar un análisis matemático formal de la especificación resultante, o comparar los resultados de la pruebas después de analizar dos piezas de software, una desarrollada usando una revisión con base en escenarios y otra que no. Cuando tenemos un cuerpo tan diverso de evidencias, podemos tomar mejores decisiones sobre qué tipo de revisiones hacer y cuándo hacerlas.

6. Conclusiones

Estos estudios ofrecen muchas lecciones para profesionales y gerentes. En primer lugar, incluso siendo perfecta, la replicación de los estudios no existe o no se puede realizar, pero no todo está perdido; existen técnicas para combinar los datos y los resultados de estudios imperfectos que conducen a decisiones válidas. Participando en estos estudios y usando sus resultados podemos mejorar nuestros procesos y productos. En segundo lugar, incluir incertidumbre en el proceso de toma de decisiones nos ayuda a evaluar los riesgos.

A pesar de la incertidumbre asociada a cada estudio o categoría de investigación, los peligros de la efedra son aún bastante claros. Por otro lado, los resultados reportados en los estudios basados en la perspectiva son incompatibles; la decisión en este caso debe esperar por más y variados estudios, tal vez con más practicantes, antes de seleccionar la técnica adecuada para un proyecto específico. En tercer lugar, la investigación empírica es un proceso, no un fin en sí mismo. A medida que se realicen más estudios, hay que revisar el cuerpo de evidencias para ver si nuestras conclusiones se mantienen todavía. Y en cuarto lugar, mediante el uso de estos enfoques podremos llegar a ser más sofisticados —y conocedores— en ingeniería del software.

Por otra parte, aun cuando los proyectos que participan en un estudio son grandes y realistas, replicar las condiciones en que se ejecutaron es muy difícil. Por estas razones, los estudios de la efedra nos ofrecen un modelo para combinar de forma alternativa —no replicar— estudios similares, y sin embargo socavar información acerca de las causas y defectos subyacentes a la ingeniería del software. Del mismo modo, la noción de un argumento multi-fuente nos permite combinar diferentes tipos de estudios para describir una mayor —y mejor— imagen que esté disponible para un solo estudio.

Podemos aplicar estas lecciones tomando algunos conceptos para mejorar la ingeniería del software:

- Se puede buscar y participar en familias de estudios en lugar de depender sólo de un estudio —las familias pueden abordar colectivamente lo que los estudios solos no pueden.

- Para cada familia de estudios, hay que establecer criterios para la confianza en la evidencia, clasificar tratamientos, y establecer medidas de resultado.

- Podemos tener un plan de imperfecciones en lugar de esperar a ver lo que va mal con un estudio en particular. Frecuentemente leemos, acerca de un estudio, que se quejan de sus limitaciones y lo desestiman. En su lugar, podemos utilizar las limitaciones para ayudarnos a buscar otros estudios que, de acuerdo con los estudios existentes, lleve a conclusiones más firmes. Es decir, cada pieza de evidencia podría tener fuerza evidencial pequeña, pero colectivamente pueden demostrar un punto fuerte.

- Debemos tratar adecuadamente con la incertidumbre. Podemos aplicar técnicas como el análisis Bayesiano y el razonamiento causal a argumentos evidenciales, para poder evaluar la probabilidad de que las evidencias justifican nuestra conclusión. De la misma manera, podemos incorporar la incertidumbre en nuestro análisis de riesgo; reduciendo la incertidumbre —incluso si no podemos eliminarla— reduciremos nuestro riesgo.

- Podemos revisar la credibilidad de cada pieza de evidencia, teniendo en cuenta no sólo cuándo fue publicada o presentada, sino si los investigadores, profesionales, o los vendedores tienen intereses creados en un resultado positivo.

- Podemos inspeccionar evidencias existentes para ver qué cuerpos de evidencia necesitan reforzarse con estudios complementarios. Como profesionales, podemos ser voluntarios para participar en estos importantes estudios.

- Antes de leer o participar en un estudio, se debe determinar si el estudio involucra generar o probar una hipótesis. Entonces podremos evaluar si las evidencias que vamos a ofrecer nos dan confianza en la hipótesis.

Aunque la ingeniería de software es más sofisticada ahora que cuando comenzó la medición del software, aún enfrentamos obstáculos para encontrar un buen equilibrio entre "ciencia" y "arte". En particular, tendemos a centrarnos —tanto en la lectura como en la práctica— en estudios individuales o pequeños aspectos de la tecnología, en lugar de familias de estudios y prácticas importantes. Nos hemos concentrado en las ciencias “duras” y en el diseño experimental basado en estadística para nuestros modelos, y hemos descuidado las áreas de las ciencias sociales y los meta-análisis que pueden ayudarnos a incrementar la confianza en la fuerza de nuestras evidencias.

Para empeorar las cosas, seguimos esperando a que la evidencia no esté en conflicto, en lugar de hacer planes para cuando esté. Teniendo una visión más amplia, leyendo mucho, y usando los resultados y los métodos que han tenido éxito en otras disciplinas, podremos alejar al desarrollo de software de su posición inestable sobre una colección dispar de resultados interesantes y colocarlo sobre una base más sólida: una fértil disciplina con resultados entretejidos y sinérgicos.

Referencias

- Bentham, J. (1827). Rationale of Judicial Evidence: Specially Applied to English Practice. New York: Hunt and Clarke. 662 p.

- Bloomfield, R. & Littlewood, B. (2003). Multi-legged Arguments: The Impact of Diversity upon Confidence in Dependability Arguments. Proceedings in 2003 International Conference on Dependable Systems and Networks, DSN 03. IEEE CS Press, pp. 25—34.
- Bratthall, L. & Jørgensen, M. (2002). Can You Trust a Single Data Source Exploratory Software Engineering Case Study? Empirical Software Engineering, Vol. 7, No. 1, pp. 9-26.
- Hayes, W. (1999). Research Synthesis in Software Engineering: A Case for Meta-Analysis. Proceedings in Sixth International Software Metrics Symposium, Boca Raton, FL, USA, pp. 143—151.
- Kaufman, M. (2003). U.S. to Stop Ephedra Sales. Washington Post, 31 Dec. 2003; www.washingtonpost.com/wp-dyn/articles/A43065-2003Dec30.html, May 2010.
- Kitchenham. B., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C. & Emam, K. (2002). Preliminary Guidelines for Empirical Research in Software Engineering. IEEE Transactions Software Engineering, Vol. 28, No. 8, pp. 721—734.
- Kolata, G. (2003). Hormone Studies: What Went Wrong? New York Times, 22 Apr. 2003; www.nytimes.com/2003/04/22/health/womenshealth/22HORM.html, May 2010.
- Pfleeger, S. L. & Hatton, L. (1997). Investigating the Influence of Formal Methods. Computer, Vol. 30, No. 2, pp. 33—43.
- Pfleeger, S. L. & Kitchenham, B. A. (1994-1995). Series on experimental design and analysis in software engineering. ACM SIGSOFT Software Engineering Notes.
- Regnell, B., Runeson, P. & Thelin, T. (2000). Are the Perspectives Really Different? Further Experimentation on Scenario-Based Reading of Requirements. Empirical Software Enginerring, Vol. 5, No. 4, pp. 331—356.
- Schum, D. A. (1994). Evidential Foundations of Probabilistic Reasoning. Wiley-Interscience. 568 p.
- Shekelle, P. G., Hardy, M. L., Morton, S. C., Maglione, M., Mojica, W. A., Suttorp, M. J., Rhodes, S. L., Jungvig, L. & Gagné, J. (2003). Efficacy and Safety of Ephedra and Ephedrine for Weight Loss and Athletic Performance: A Meta-Analysis. The Journal of American Medical Association, Vol. 289, No. 12, pp. 1537—1545.
- Shekelle, P. G., Maglione, M. & Morton, S. C. (2003). Preponderance of Evidence: Judging What to Do About Ephedra. RAND Review, Vol. 27, No. 1, pp. 16—21.
- Wagner, J. (1991). The Search for Signs of Intelligent Life in the Universe. London: Perennial Press. 240 p. **Ω**