Évaluation

Types d’expérimentations

- comparatif ou descriptif
- expérimental / quasi-expérimental / in situ
- longitudinal / ponctuel

Éléments à mettre en commun

- éthique : formulaire de consentement, autorisation d'utilisation et de conservation des données

- recrutement des sujets

- questionnaires standard pour mesurer l'UX, l'utilisabilité (SUS), la motivation, etc.

- analyse des résultats (tests statistiques) et visu des réponses

- structuration du protocole et des résultats ; éléments de design (description des itérations)

Tendances à approfondir

- Ne plus utiliser les p-value

http://www.aviz.fr/badstats (la page contient un chapitre de livre super intéressant écrit par Pierre et des examples d'articles de mises en oeuvre)

- La reproductibilité, partager de quoi reproduire l'expérience

https://www.researchgate.net/publication/241623180_RepliCHI_SIG_From_a_panel_to_a_new_submission_venue_for_replication

Usability Evaluation Considered Harmful (Some of the Time)

Saul Greenberg
Department of Computer Science
University of Calgary
Calgary, Alberta, T2N 1N4, Canada
saul.greenberg@ucalgary.ca

Bill Buxton
Principle Researcher
Microsoft Research
Redmond, WA, USA
bibuxton@microsoft.com

ABSTRACT
Current practice in Human Computer Interaction as encouraged by educational institutes, academic review processes, and institutions with usability groups advocate usability evaluation as a critical part of every design process. This is for good reason: usability evaluation has a significant role to play when conditions warrant it. Yet evaluation can be ineffective and even harmful if naively done 'by rule' rather than 'by thought'. If done during early stage design, it can mute creative ideas that do not conform to current interface norms. If done to test radical innovations, the many interface issues that would likely arise from an immature technology can quash what could have been an inspired vision. If done to validate an

INTRODUCTION
Usability evaluation is one of the major cornerstones of user interface design. This is for good reason. As Dix et al., remind us, such evaluation helps us "assess our designs and test our systems to ensure that they actually behave as we expect and meet the requirements of the user" [7]. This is typically done by using an evaluation method to measure or predict how effective, efficient and/or satisfied people would be when using the interface to perform one or more tasks. As commonly practiced, these usability evaluation methods range from laboratory-based user observations, controlled user studies, and/or inspection techniques [7,22,1]. The scope of this paper concerns these methods.

Welcome to the BELIV Workshop 2016

Beyond Time And Errors: Novel Evaluation Methods For Visualization

In conjunction with VIS 2016

News

Panelists announced

We are happy to announce that *Daniel Weisskopf, Laura McNamara, Mark Whiting, Niklas Elmqvist, and Tamara Munzner* have agreed to participate in our panel "On the Future of Evaluation and BELIV". We are looking forward to some fantastic discussions with this very strong and diverse set of panelists. The panel will be in the last session from 4:15pm - 5:30pm. Don't miss it!

Important dates

~~June 19, 2016: Paper submission due~~
~~July 10, 2016: First notification~~
~~July 20, 2016: Revisions due~~
~~Aug 5, 2016: Final notification~~
~~Aug 31, 2016: Camera-ready due~~
~~Sep 5, 2016: Revised CR deadline~~
Oct 24, 2016: BELIV workshop (1-day)

Plan

Évaluation et tests :

- Introduction
- Approches d’évaluation
- Méthodes analytiques
- Méthodes empiriques
- Évaluation 2.0 : passer à l’échelle
- Design expérimental

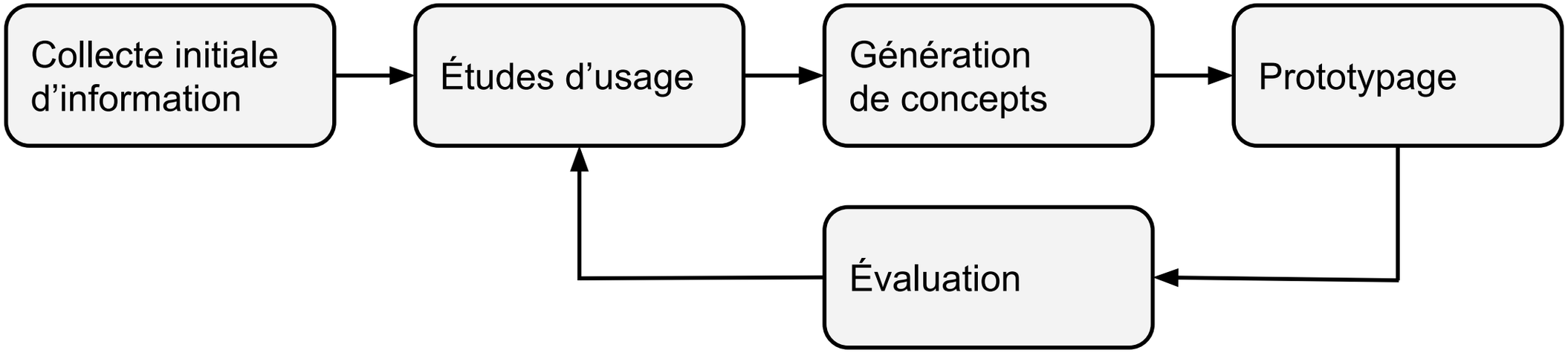
Évaluation et tests

- **Introduction**
- Approches d’évaluation
- Méthodes analytiques
- Méthodes empiriques
- Évaluation 2.0 : passer à l’échelle
- Design expérimental

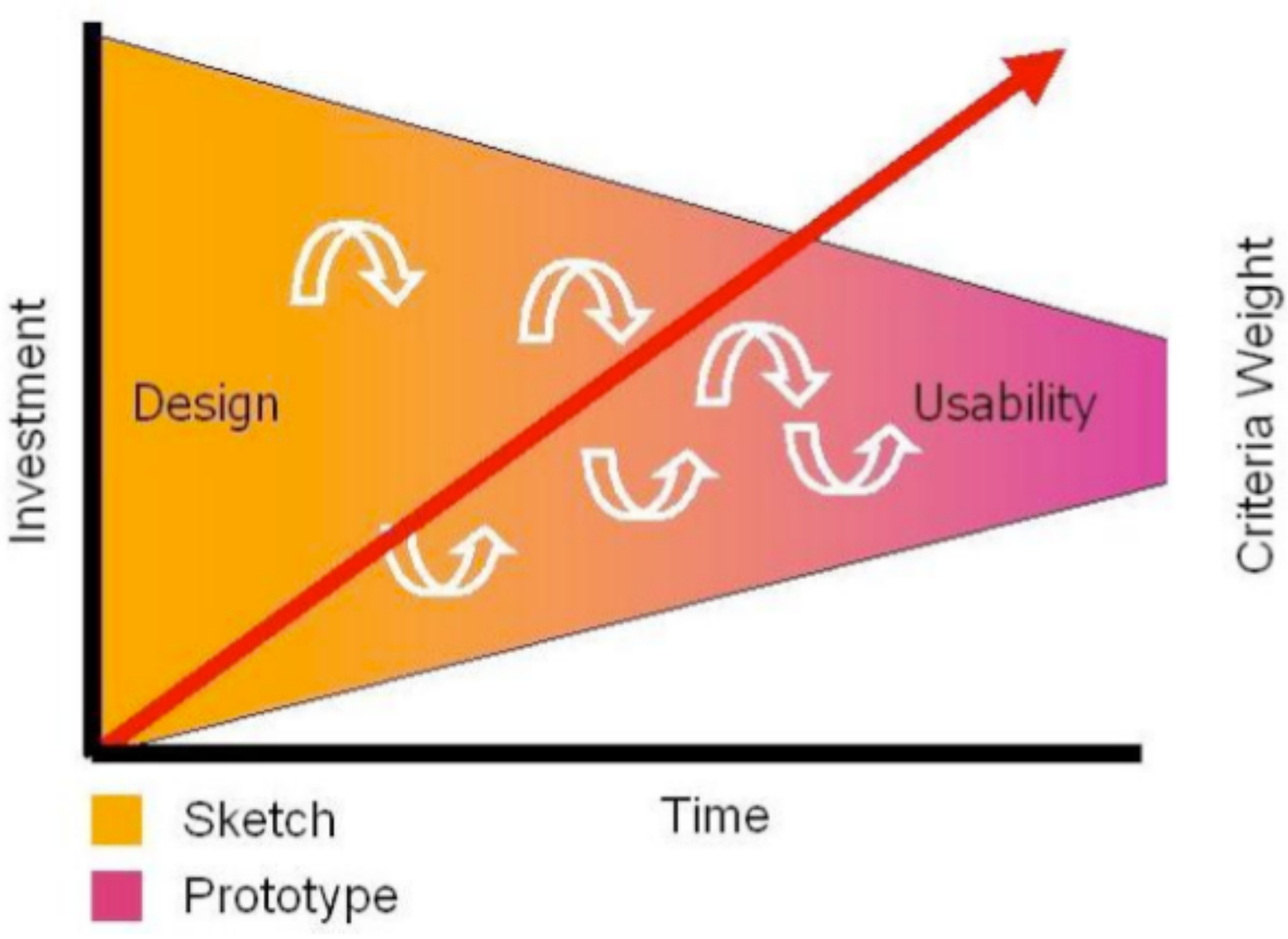
Le rôle de l’évaluation dans le processus UX

- Une partie du cycle itératif : *design-build-evaluate*
- Une comparaison entre ce qui est “construit” et ce qui était prévu
- Un endroit pour réfléchir sur cette différence et la prochaine phase de conception

Quand évaluer



Le processus de conception



Être agile

Rater rapidement pour réussir plus tôt :

- Itérations sur des prototypes basse fidélité
- Design en parallèle : construire et tester plusieurs prototypes
- Explorer des alternatives

Augmenter progressivement la fidélité

Affronter la réalité, concevoir pour des cas d’utilisation pas des spécifications

Modified from Evolution of Evaluation in HCI by J. Kaye

A bit of history

Evaluation by Engineers / Computer Scientists

Evaluation by Experimental Psychologists
& Cognitive Scientists

Evaluation by HCI Professionals

Evaluation in Context

Evaluation by Engineers/Computer

50’s to 70’s

Users are engineers & mathematicians:

.:Evaluators are engineers

.:The limiting factor is reliability

Users are programmers:

.:Evaluators are programmers

.:The speed of the machine is the limiting factor

Evaluation by Experimental Psychologists & Cognitive Scientists

Modified from Evolution of Evaluation in HCI by J. Kaye

end of 70’s, 80’s

Users are users:

.:the computer is a tool, not an end result

Evaluators are cognitive scientists and experimental psychologists:

.:they’re used to measuring things through experiments

The limiting factor is what the human can do

Case Study of Evaluation: Text Editors

Roberts & Moran, 1982, 1983.

Their methodology for evaluating text editors had 3 criteria:

.:Objectivity

.:“implies that the methodology not be biased in favor of any particular editor’s conceptual structure”

.:Thoroughness

.:“implies that multiple aspects of editor use be considered”

.:Ease-of-use (of the method, not the editor itself)

.:“the methodology should be usable by editor designers, managers of word processing centers, or other non-psychologists who need this kind of

Evaluation by HCI Professionals

80’s to now

Usability

Expertise

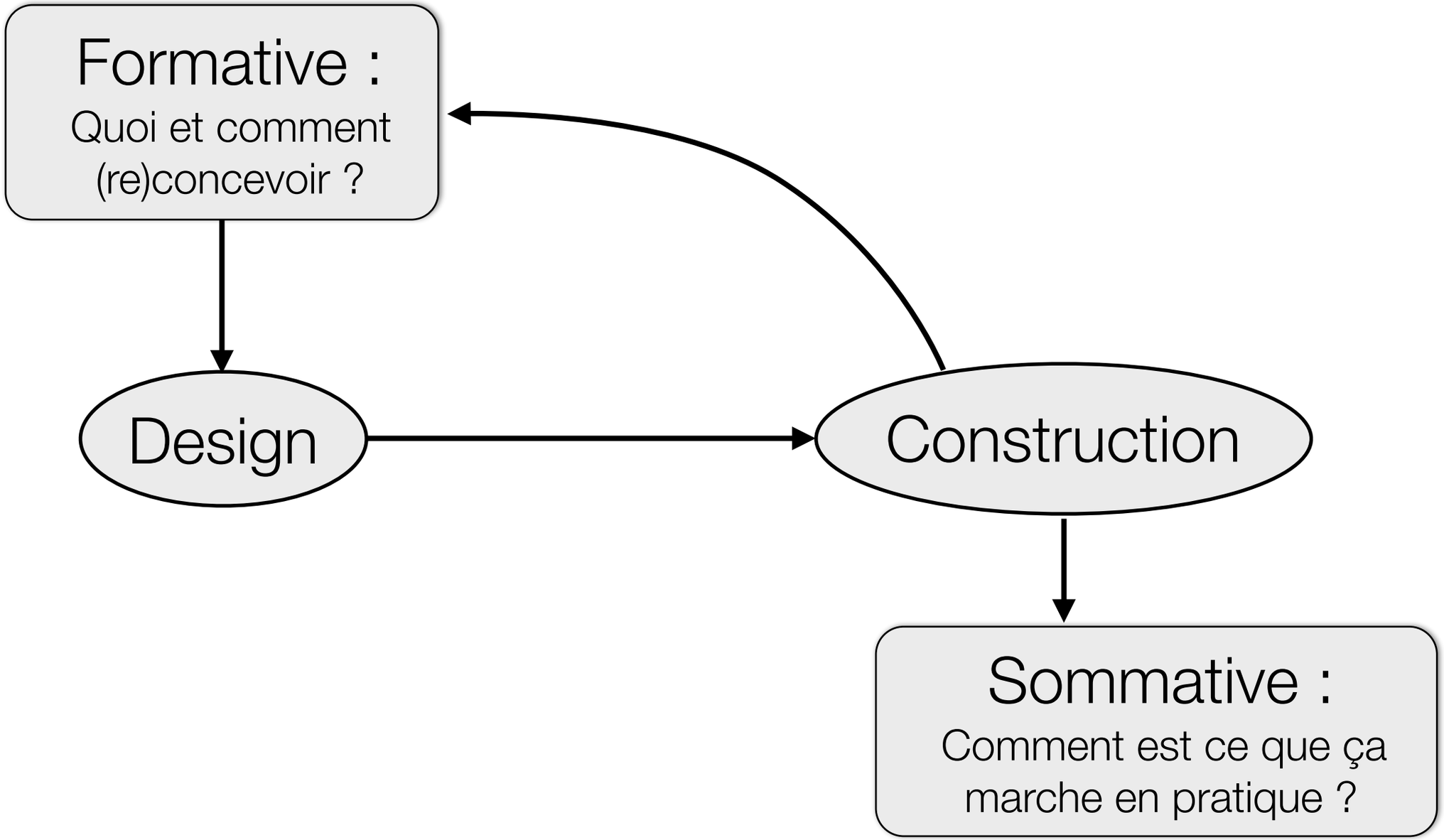
Focus on better results, regardless of whether they were experimentally provable or not.

Evaluation in Context

Évaluation et tests

- Introduction
- **Approches d’évaluation**
- Méthodes analytiques
- Méthodes empiriques
- Évaluation 2.0 : passer à l’échelle
- Design expérimental

Évaluation Formative ou Sommative ?



M. Scriven: The methodology of evaluation, 1967

Évaluation Analytique vs. Empirique

"If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you may just study the kind and speed of the cuts it makes in the hands of a good axeman."

[Scriven, 1967]

Des méthodes complémentaires

L'évaluation empirique permet de comprendre les implications des propriétés de l'objet

- La hache va elle trancher tel buche ?

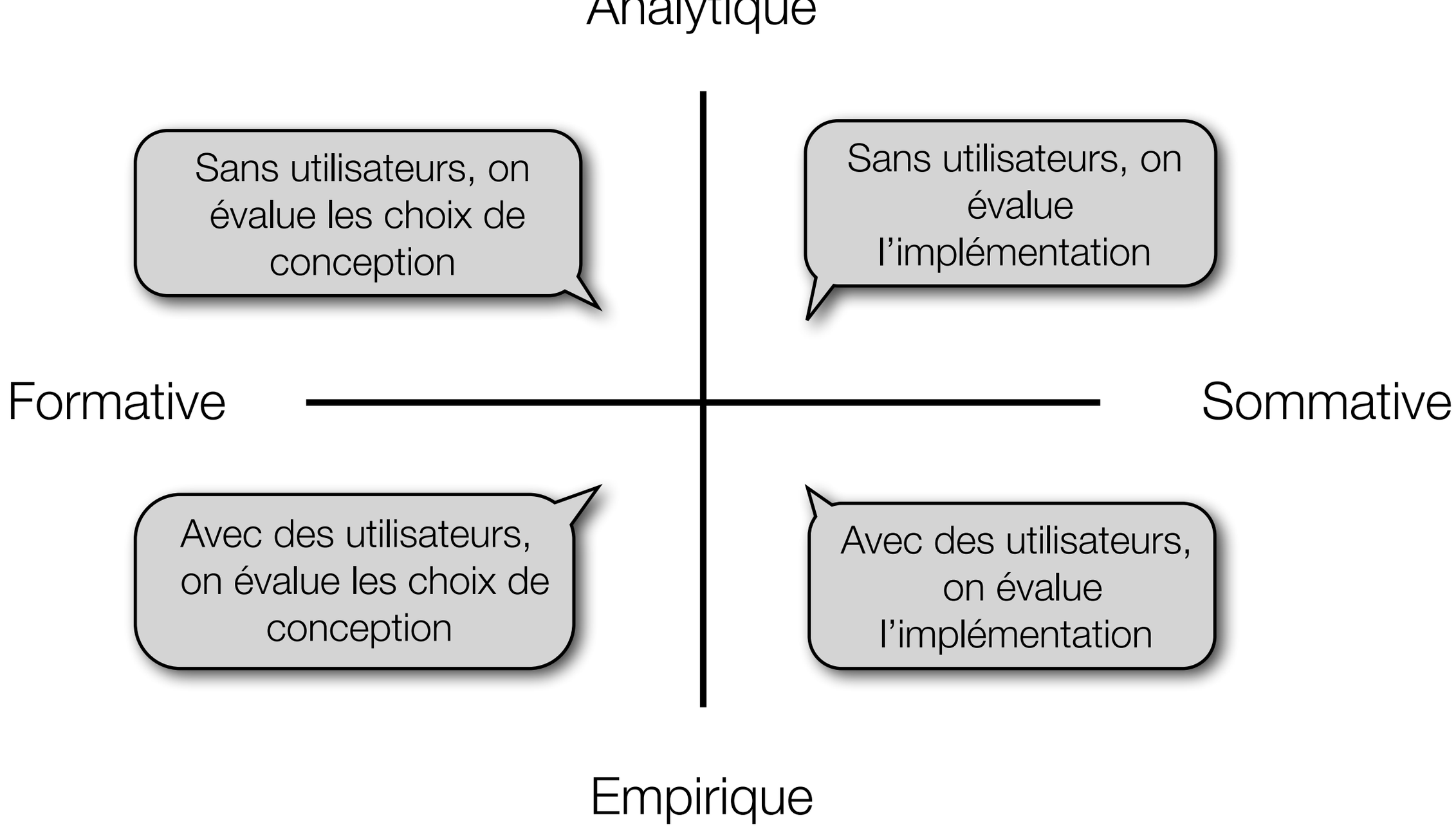
L'évaluation analytique identifie offre une grille critique sur les propriétés importantes

- Le manche de la hache est il compatible avec les gauchers ?

Dans les deux cas :

- Productions de faits qui doivent être interprétés

Des approches orthogonales



Une évaluation sans critère ne sert à rien !

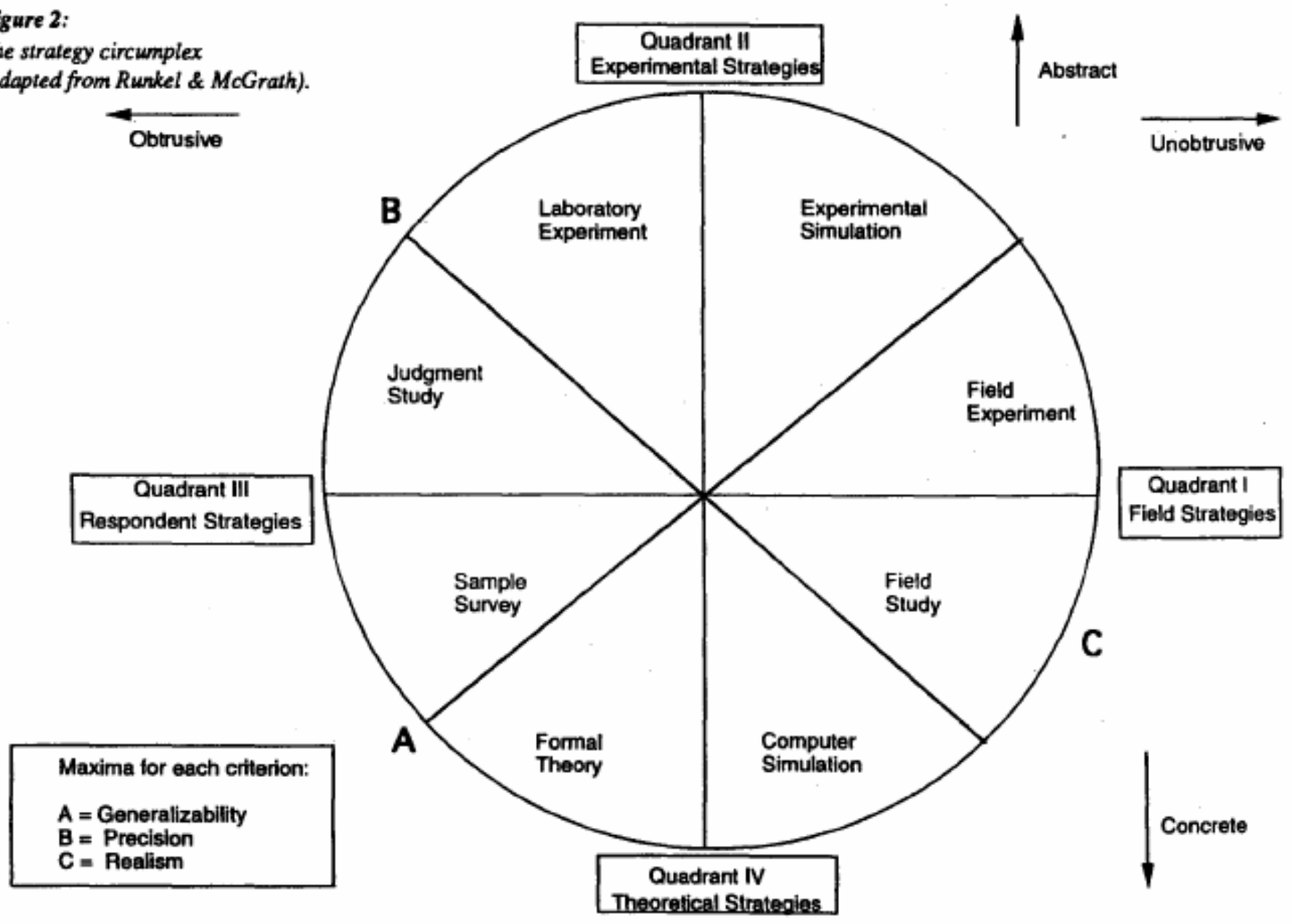
If faut définir ce qu'on veut savoir avant d'évaluer !

Critères possible (parmi bien d'autres) :

- Test informel d'une idée contre une autre
- Analyse statistique de la performance moyenne
- Acceptation par un groupe d'utilisateur réaliste
- Vérification de critères heuristiques / ergonomiques
- Mesure de métriques d'utilisabilité lié au design

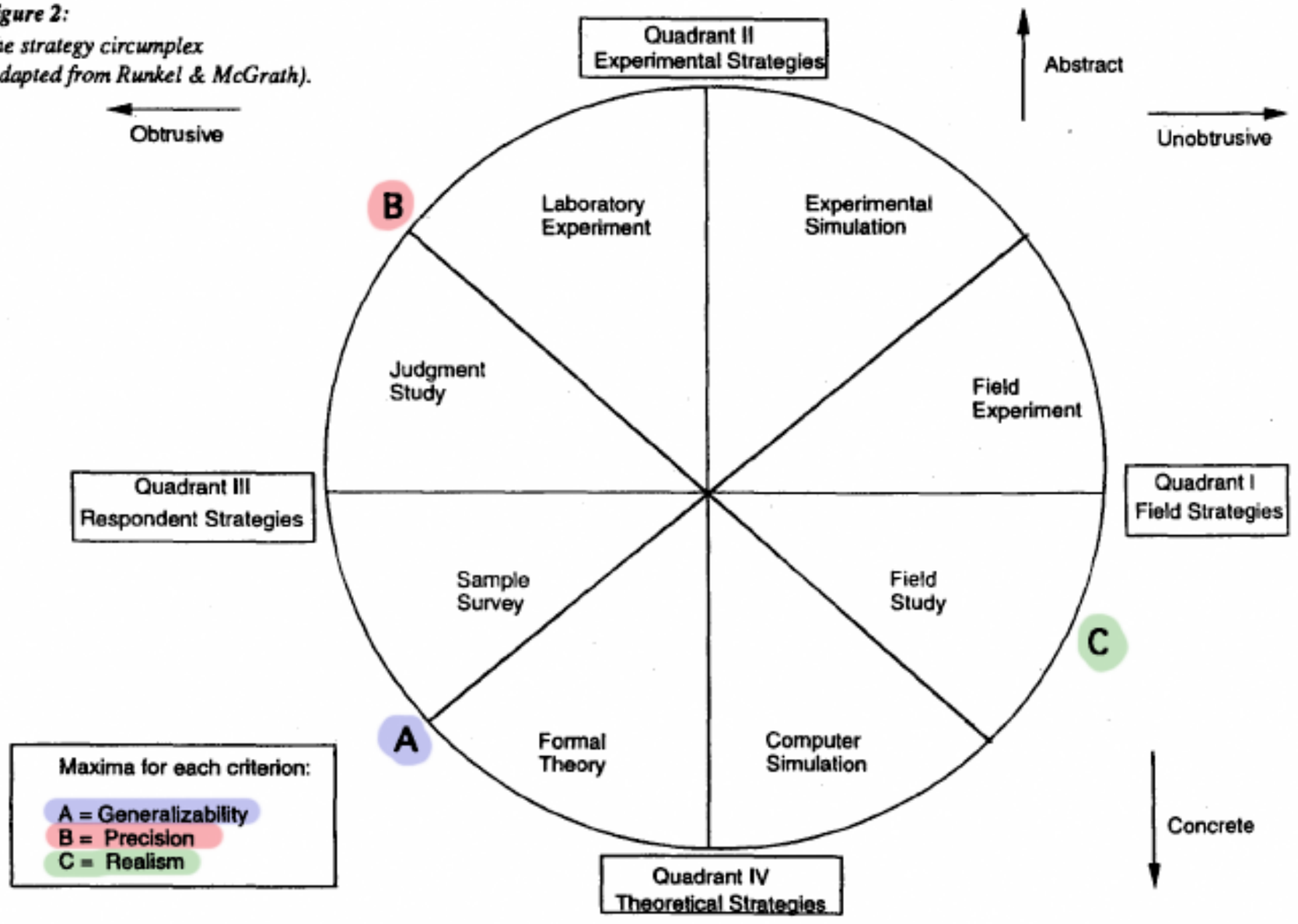
Taxonomy of Methods [McGrath et al. 1994]

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).



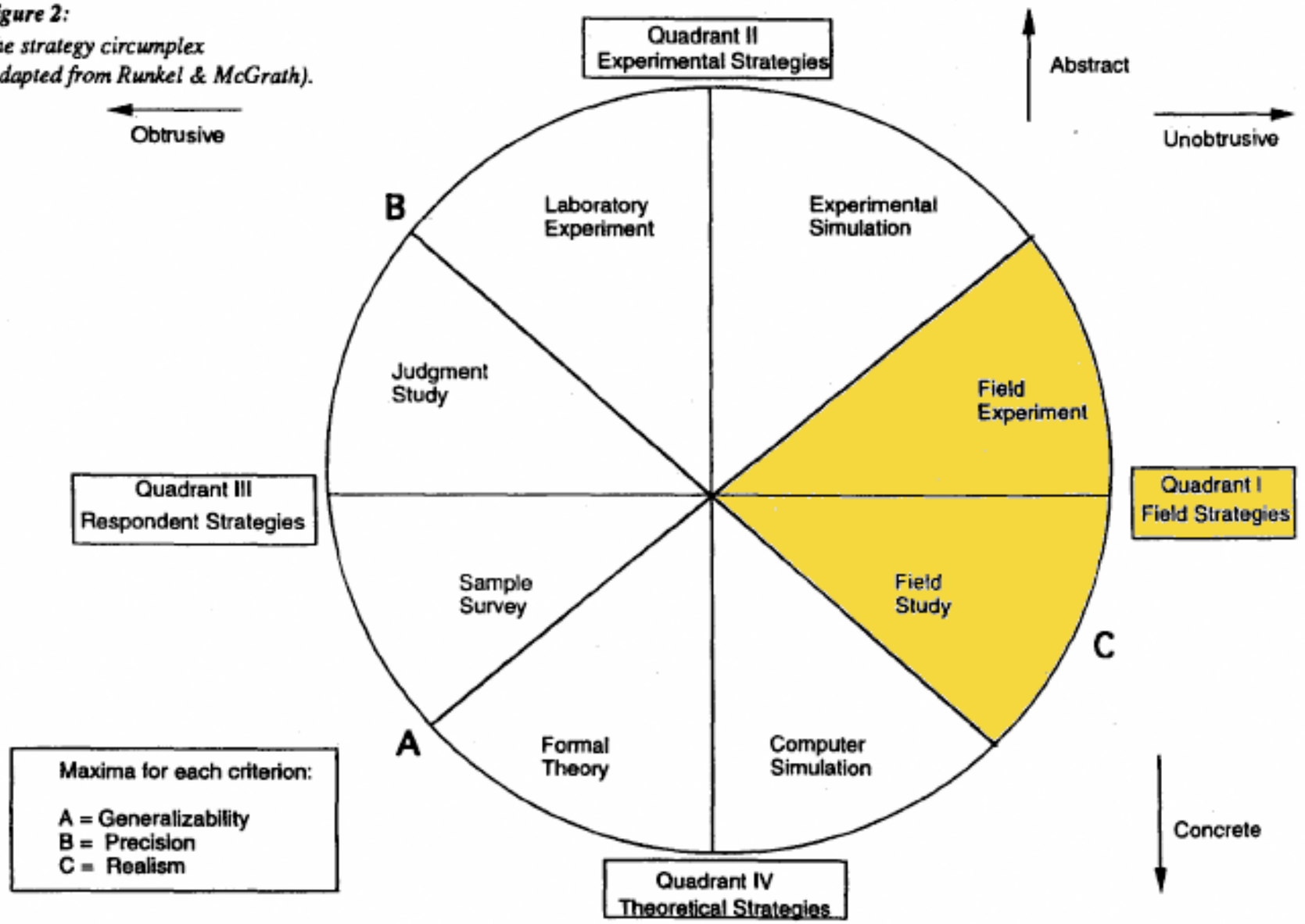
Taxonomy of Methods [McGrath et al. 1994]

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).



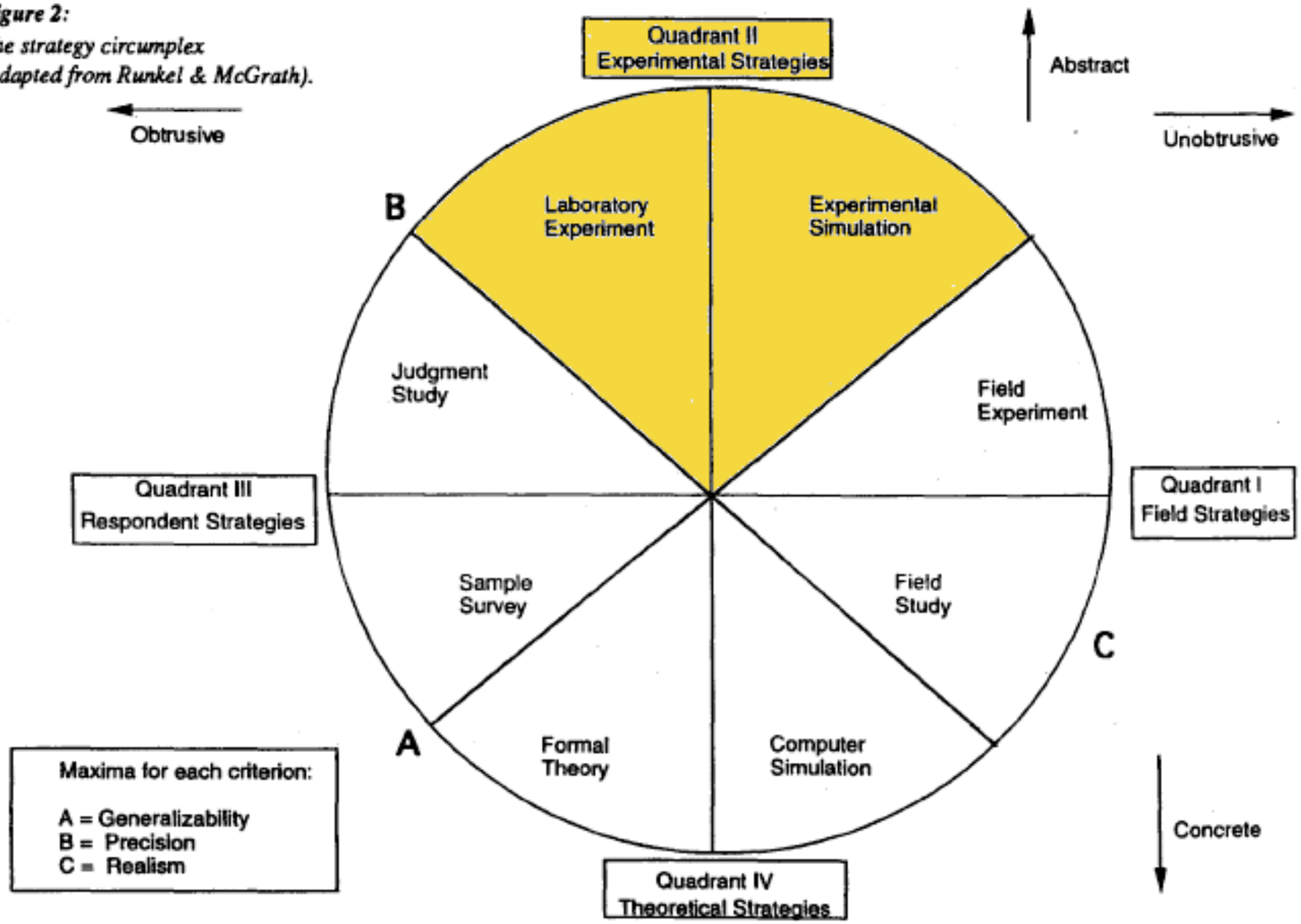
Taxonomy of Methods [McGrath et al. 1994]

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).



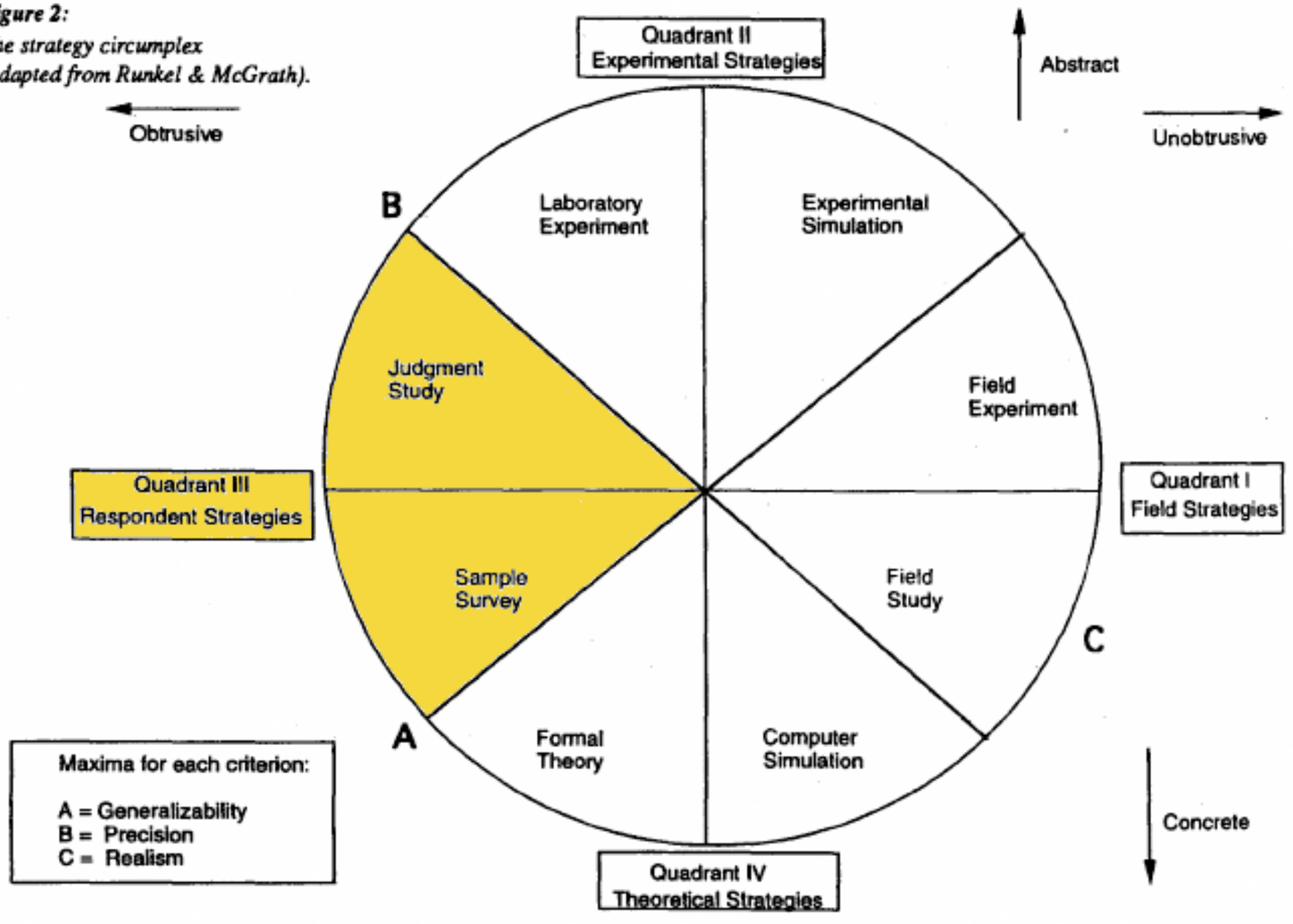
Taxonomy of Methods [McGrath et al. 1994]

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).



Taxonomy of Methods [McGrath et al. 1994]

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).



Taxonomy of Methods [McGrath et al. 1994]

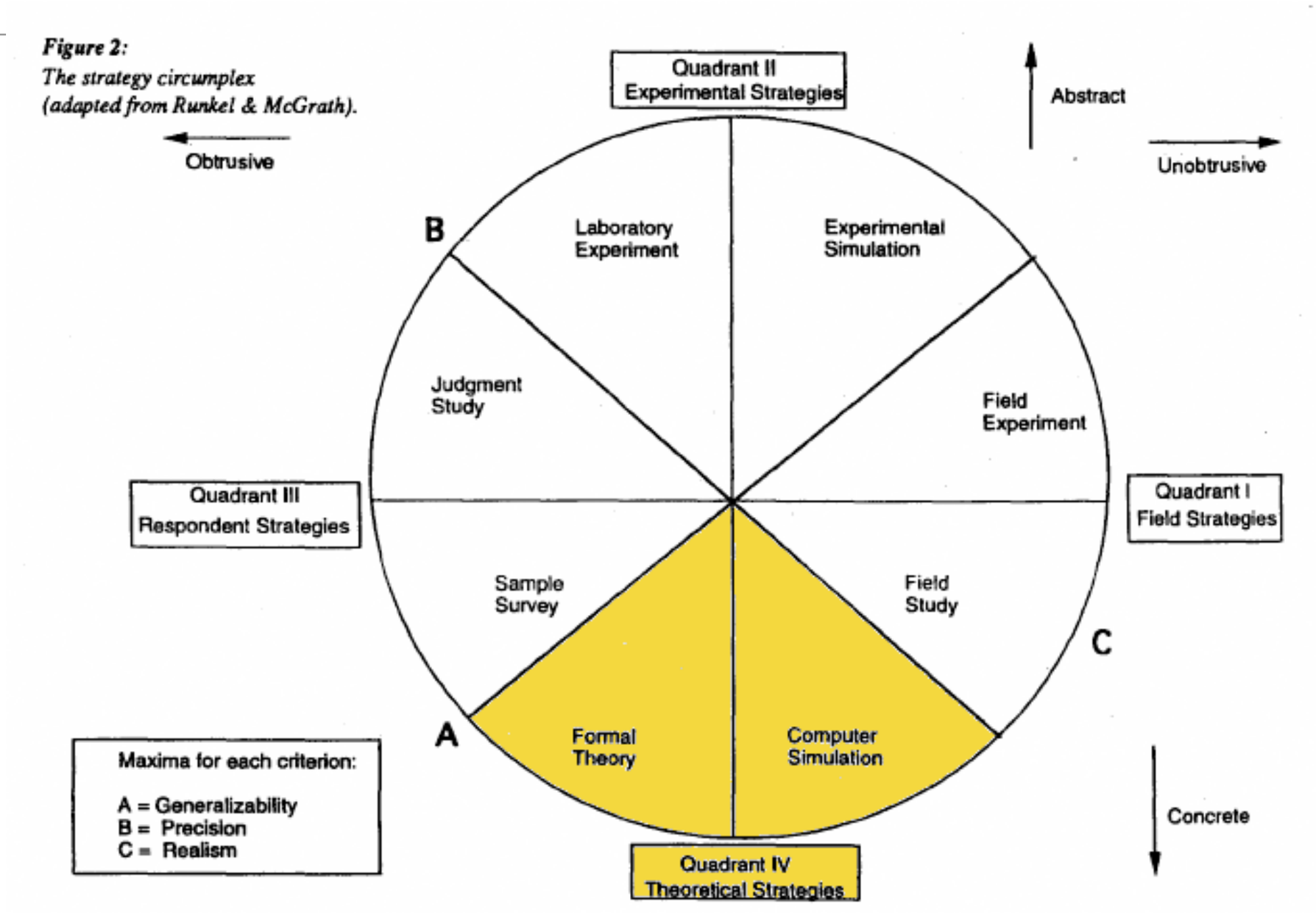


Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).

Évaluation et tests

- Introduction
- Approches d'évaluation
- **Méthodes analytiques**
- Méthodes empiriques
- Évaluation 2.0 : passer à l'échelle
- Design expérimental

Les grands types d’évaluation analytique

Basée sur des modèles

- Évaluation selon des modèles d’interaction

Basée sur l’inspection

- Review d’experts / critique du design
- Cognitive walkthrough
- Évaluation heuristique

Évaluation basée sur des modèles

GOMS (Goals, Operators, Methods, and Selection rules)

- Les objectifs (Goals) correspondent à ce que l'utilisateur essaie de faire
- Les opérateurs sont les actions faites pour atteindre cet objectif.
- Les méthodes sont les séquences d'opérateur qui permettent d'accomplir un objectif. Il peut y avoir plus d'une méthode pour un même objectif, dans ce cas :
- Les règles de sélection (Selection rules) sont utilisées pour décrire quand un utilisateur voudrait choisir tel méthode plutôt qu'une autre. On ignore souvent ces règles dans une analyse GOMS simple.

KLM

- On analyse une action et on la découpe en étapes atomiques
- On cherche la durée de chaque étape dans une table
- On prédit la durée de l'action complète
- *Permet la prédire avant d'implémenter !*

From http://www.cs.umd.edu/class/fall2002/cmsc838s/tichi/printer/goms.html

[John & Kieras, 1996]

GOMS analysis

```
GOAL: EDIT-MANUSCRIPT
.     GOAL: EDIT-UNIT-TASK ... repeat until no more unit tasks
.     .     GOAL: ACQUIRE UNIT-TASK
.     .      .      GOAL: GET-NEXT-PAGE ... if at end of manuscript page
.     .      .      GOAL: GET-FROM-MANUSCRIPT
.     .      GOAL: EXECUTE-UNIT-TASK ... if a unit task was found
.     .      .      GOAL: MODIFY-TEXT
.     .      .      .   [select: GOAL: MOVE-TEXT* ...if text is to be moved
.     .      .      .       GOAL: DELETE-PHRASE ...if a phrase is to be deleted
.     .      .      .       GOAL: INSERT-WORD] ... if a word is to be inserted
.     .      .      .     VERIFY-EDIT

*Expansion of MOVE-TEXT goal
GOAL: MOVE-TEXT
.     GOAL: CUT-TEXT
.     .     GOAL: HIGHLIGHT-TEXT
.     .     .     [select**: GOAL: HIGHLIGHT-WORD
.     .     .     .   MOVE-CURSOR-TO-WORD
.     .     .     .   DOUBLE-CLICK-MOUSE-BUTTON
.     .     .     .   VERIFY-HIGHLIGHT
.     .     .           GOAL: HIGHLIGHT-ARBITRARY-TEXT
.     .     .     .      MOVE-CURSOR-TO-BEGINNING    1.10
.     .     .     .      CLICK-MOUSE-BUTTON          0.20
.     .     .     .      MOVE-CURSOR-TO-END          1.10
.     .     .     .      SHIFT-CLICK-MOUSE-BUTTON    0.48
.     .     .     .      VERIFY-HIGHLIGHT]           1.35
.     .     GOAL: ISSUE-CUT-COMMAND
.     .     .     MOVE-CURSOR-TO-EDIT-MENU          1.10
.     .     .     PRESS-MOUSE-BUTTON                0.10
.     .     .     MOVE-CURSOR-TO-CUT-ITEM           1.10
.     .     .     VERIFY-HIGHLIGHT                  1.35
.     .     .     RELEASE-MOUSE-BUTTON              0.10
```

• • •

From http://www.cs.umd.edu/class/fall2

[John & Kieras, 1996]

GOMS analysis

```
*Expansion of MOVE-TEXT goal
GOAL: MOVE-TEXT
.     GOAL: CUT-TEXT
.     .     GOAL: HIGHLIGHT-TEXT

...

.     .    GOAL: ISSUE-CUT-COMMAND
.     .    .     MOVE-CURSOR-TO-EDIT-MENU          1.10
.     .    .     PRESS-MOUSE-BUTTON                0.10
.     .    .     MOVE-CURSOR-TO-CUT-ITEM           1.10
.     .    .     VERIFY-HIGHLIGHT                  1.35
.     .    .     RELEASE-MOUSE-BUTTON              0.10
.     GOAL: PASTE-TEXT
.     .     GOAL: POSITION-CURSOR-AT-INSERTION-POINT
.     .    MOVE-CURSOR-TO-INSERTION-POIONT         1.10
.     .    CLICK-MOUSE-BUTTON                      0.20
.     .    VERIFY-POSITION                         1.35
.     .    GOAL: ISSUE-PASTE-COMMAND
.     .    .     MOVE-CURSOR-TO-EDIT-MENU          1.10
.     .    .     PRESS-MOUSE-BUTTON                0.10
.     .    .     MOVE-MOUSE-TO-PASTE-ITEM          1.10
.     .    .     VERIFY-HIGHLIGHT                  1.35
.     .    .     RELEASE-MOUSE-BUTTON              0.10
TOTAL TIME PREDICTED (SEC)                         14.38

Based on the above GOMS analysis, it should take 14.38 seconds to move text.
```

From http://www.cs.umd.edu/class/fall2002/cmsc838s/tichi/printer/goms.html

KLM

Description	Operation	Time (sec)
Reach for mouse	H[mouse]	0.40
Move pointer to "Replace" button	P[menu item]	1.10
Click on "Replace" command	K[mouse]	0.20
Home on keyboard	H[keyboard]	0.40
Specify word to be replaced	M4K[word]	2.15
Reach for mouse	H[mouse]	0.40
Point to correct field	P[field]	1.10
Click on field	K[mouse]	0.20
Home on keyboard	H[keyboard]	0.40
Type new word	M4K[word]	2.15
Reach for mouse	H[mouse]	0.40
Move pointer on Replace-all	P[replace-all]	1.10
Click on field	K[mouse]	0.20
Total		10.2

Limites

- Des prédictions valides pour un utilisateur expert qui ne fait pas d'erreur
  - les experts font aussi des erreurs !
  - pas de prise en compte des utilisateurs novices ou intermédiaires qui font des erreurs occasionnelles.
  - il existe des extensions qui essaient de modéliser l'apprentissage
- Toutes les tâches ont un objectif clair
  - Beaucoup de tâches ne sont pas si dirigées spécialement en design UX.
- Ne prend pas en compte les différences individuelles entre utilisateurs
  - Basé sur des moyennes statistiques
- Ne prend pas en compte les aspect sociaux et organisationnels du produit
- Pas d'information sur la qualité d'utilisation et le plaisir provoqué par le produit.
- Pas représentatif des théories actuelles sur la cognition humaine
  - Supposition d'un model linéaire de la cognition avec une activité faite à la fois.

Inspections et critiques d’experts

- Tout au long du processus de développement
- Conduite par des développeurs et des experts (internes ou externes)
- Outil pour identifier des problèmes
- Peut aller d’une heure à une semaine de travail
- Préférer une approche structurée
  - Les reviewers doivent pouvoir communiquer sur tous les problèmes (sans fâcher l’équipe)
  - Les critiques ne doivent pas être aggressive envers les développeurs / designers
  - L’objectif principal est d’identifier les problèmes (pas leur source)
- Des solutions peuvent être suggérées a l’équipe

Méthodes d'inspection

Inspection des "Guidelines"

- Vérifier que l'interface respecte bien un ensemble de règles.

Inspection centrée cohérence

- Vérifier que l'interface est cohérente / consistante
  avec elle-même, avec les applications liées, avec l'OS
- Une vue générale peut aider, ex : impression des pages clés du site collées au mur.
- On peut forcer la consistance par les outils, ex : via les CSS pour les sites Web

Procédure d’inspection

- Trouver des experts/reviewers
- Définir un plan avec des limites temporelles
- Préparer le materiel pour les reviewers, y compris les critères d’intérêt
- Sur place ou sur un autre site
- Rédaction d’un rapport et définition des conséquences

+/- des évaluations par experts

- Results of informal reviews and inspections are often directly used to change the product
  - ... still state of the art in many companies!
  - The personal view of the CEO, or his partner ...

- Really helpful evaluation
  - Is explicit
  - Has clearly documented findings
  - Can increase the quality significantly

- Expert reviews and inspections are a starting point for change

Les critères ergonomiques (usability guidelines)

- Don Norman’s principles:
  - visibility, affordances, natural mapping, and feedback
- Ben Shneiderman’s 8 Golden Rules of UI design
- Bruce Tognazzini’s 16 principles:
  - http://www.asktog.com/basics/firstPrinciples.html
- Christian Bastien’s Ergonomic Criteria
- Jakob Nielsen’s Heuristics

L’évaluation heuristique

Conçue comme une méthode d’évaluation “discount” basée sur l’inspection :

- Évaluation rapide, pas cher et facile d’interfaces
- http://www.useit.com/papers/heuristic/

Principes :

- Il y a une liste de propriétés désirable dans une interface : les “heuristiques”
- Ces heuristiques peuvent être vérifiées par des experts avec un résultat clair et précis

Évaluation et tests

Séance passée :

- Introduction
- Approches d’évaluation
- Méthodes analytiques

Cette séance :

- **Méthodes empiriques**
- **Évaluation 2.0 : passer à l’échelle**

Prochaine séance :

- Design expérimental

Pourquoi utiliser des méthodes empiriques ?

Il est difficile de connaître la qualité d'une expérience avant que des gens n'aient essayé votre produit.

Les méthodes "expertes" :

- les experts en savent trop
- les experts n'en savent pas assez (sur les tâches par ex.)
- difficile de prédire ce que vont faire de "vrais" utilisateurs

Pourquoi utiliser des méthodes empiriques ?

Identifier des problèmes d'utilisabilité du produit

Rassembler des données sur les performances du produit

Connaître la satisfaction des participants face au produit

Évaluation empiriques

- Focus groups
- Études de terrain
- Étude d’utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Évaluation empiriques

- Focus groups
- Études de terrain
  - Trouver les problèmes en contexte
  - Permet d’évaluer un prototype fonctionnel sur de “vraies” tâches
  - Plutôt des observations qualitatives
- Étude d’utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Évaluation empiriques

- Focus groups
- Études de terrain
- Étude d'utilisabilité
  - Trouver des problèmes pour la prochaine itération
  - Évaluer des prototypes en laboratoire sur des tâches prédéfinies
  - Observations qualitatives (problèmes d'utilisabilité / d'ergonomie)
- Mesures physiologiques
- Expérience contrôlée

Évaluation empiriques

- Focus groups
- Études de terrain
- Étude d’utilisabilité
- Mesures physiologiques
  - ex : eye tracking
- Expérience contrôlée

Évaluation empiriques

- Focus groups
- Études de terrain
- Étude d'utilisabilité
- Mesures physiologiques
- Expérience contrôlée
  - Tester une hypothèse (ex : l'interface X est plus rapide que l'interface Y)
  - Évaluer un prototype fonctionnel ou une application stable, dans un environnement contrôlé sur des tâches définies.
  - Observations quantitatives (temps, taux d'erreur, satisfaction)

Question

Quel type d'évaluation pour :

décider d'introduire un nouveau type de clavier sur iOS ?

Parmi :

- Focus groups
- Études de terrain
- Étude d'utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Évaluation empiriques

- **Focus groups**
- Études de terrain
- Étude d’utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Focus Groups

- Discussion informelle et qualitative en groupe
  - Recueillir des informations sur ce que pense et ressentent les gens
  - Recueillir des opinions, attitudes, sentiments, besoins idées
  - Comprendre pourquoi les gens agissent et se comporte de telle manière
- Tôt dans la conception, avant le design d'interface
- Complémentaire d'études quantitative plus poussées
- Setup :
  - Groupes de 6 à 8 participants
  - Mené par un modérateur
  - Durée de 1h30 à 2h
- Analyse de la discussion, enregistrement vidéo => rapport simple avec des citations

Focus groups + et –

Avantages

- Rapide, facile, peu cher
- Information sur les opinions, motivations et objectifs des gens
- Flexibilité et exploration de différents thèmes

Inconvénients

- Pas représentatif, dur de généraliser
- Ce que les utilisateurs pensent vs. ce qu'ils font vraiment
- L'analyse peut être lourde/laborieuse
- Peut être biaisé par le modérateur ou des participants avec des opinions bien trempées

Évaluation empiriques

- Focus groups
- **Études de terrain**
- Étude d’utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Études de terrain

Activités étudiées en situation

Avantages :

- Permet de mieux comprendre l'acceptation
- Permet des études longues, pour comprendre les dynamiques d'apprentissage, de collaboration ou d'adaptation

Problèmes :

- Cher
- Il faut un produit (ou prototype) fiable
- Collecte d'observations lourde

Principes communs à toutes ces études

- Focus groups
- Études de terrain
- Étude d’utilisabilité
- Mesures physiologiques
- Expérience contrôlée

Recruter des participants

Des utilisateurs représentatifs

- connaissances du domaine
- maitrise des tâches

On peut s’en rapprocher

- système pour les médecins
  > recruter des étudiants de médecine

Utiliser des “carottes” pour recruter des participants

L’éthique !

Il a une pression (même involontaire) sur les participants :

- Angoisse de la performance
- Ressenti d’évaluation d’intelligence du participant
- Comparaison entre participants
- Paraître bête en face des observateurs
- Compétition avec d’autres participants with other subjects

Respect et contrôle

Temps

- Ne pas le gâcher

Confort

- Rendre la session agréable et mettre le participant à l'aise

Consentement éclairé

- Informer l'utilisateur autant que possible

Vie privée

- Les données collectées sont anonymes et restent privées

Contrôle

- Les participants peuvent arrêter n'importe quand

Avant le test

Temps

- Avoir conduit des pilotes pour tester les tâches et le matériel/contenu

Confort

- On teste le système, pas vous
- Tout problème rencontré est “la faute du système”. On a besoin de votre aide pour identifier ces problèmes.

Vie privée

- Les résultats du test sont confidentiels et protégés.

Information

- Présentation rapide de la raison de l’étude
- Expliquer comment l’information est capturée, et qu’on peut l’arrêter à tout moment

Pendant le test

- Temps
  - Se concentrer sur les tâches nécessaires
- Confort
  - Atmosphère calme
  - Prise de pauses lorsque les sessions durent
  - Ne jamais paraître déçu
  - Donner une tâche à la fois
  - La première tâche doit être plutôt facile pour donner une expérience positive
- Vie privée
  - Pas de chef autour
  - répondre aux questions (en évitant d'introduire des biais)
- Controle
  - Les utilisateurs peuvent arrêter à tout moment

Après le test

Confort

- Expliquer ce à quoi ils ont contribué

Information

- Répondre à toutes les questions auquel vous n’avez pas pu répondre pendant l’expérience

Vie privée

- Ne pas publier qui permettent d’identifier un participant
- Ne pas montrer de vidéo ou d’audio sans la permission explicite des participants

Évaluation empiriques

- Focus groups
- Études de terrain
- **Étude d’utilisabilité**
- Mesures physiologiques
- Expérience contrôlée

https://fr.wikipedia.org/wiki/Test_utilisateur

Qu’est qu’un test d’utilisabilité

Un test d’utilisabilité est un moyen de mesurer comment un artefact (comme une page web, un interface, un document, ou un dispositif) répond à ce pourquoi il a été conçu.

Métriques

Facilité d'apprentissage

- temps d'apprentissage, …

Facilité d'utilisation

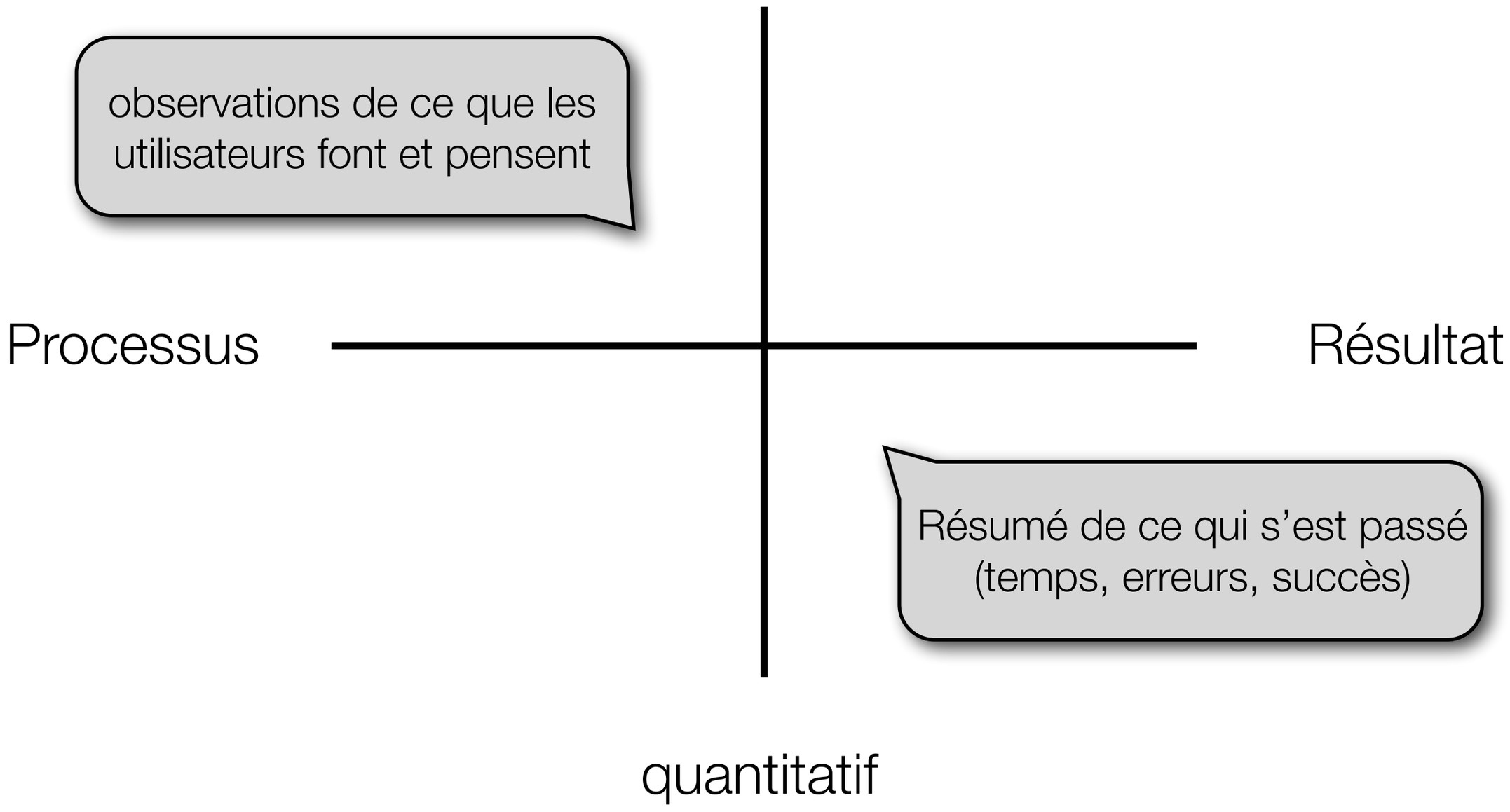
- temps de réalisation, taux d'erreur…

Satisfaction des utilisateurs

- questionnaires…

Pas "intuitif"! **Par "naturel"!**

Quel type de données capturer ?



Quoi capturer (quantitatif)

- Taux de succès
- Taux et types d'erreurs : Combien d'erreurs faites par les participants ? Étaient elles fatales ou corrigible avec la bonne information ?
- Temps de réalisation : Combien de temps prennent les participants pour réaliser une tâche de base ? (Par exemple, faire un achat, créer un compte, commander un produit.)
- Pages visitées, le nombre d'étapes pour réaliser une tâche…
- Souvenir : Combien les utilisateurs se souviennent après une période de non utilisation ?
- Réponse émotionnelle : Notes de questionnaire, comment la personne pense avoir réalisé la tâche (Confiance, stress, désir de recommender le système…)

Quoi capturer (qualitatif)

- Comment les participants ont réagit au système
- Ce que participants l'ont compris
- Quels chemin les participants ont pris
- Quels problèmes les participants ont rencontrés
- Ce que les gens ont dit (/pensé) pendant l'activité
- Les réponses des participants à des questions ouvertes

Il faut un plan !

Un bon plan contient :

- un objectif / une tâche, càd quoi faire, ou une questions à laquelle il faut trouver une réponse
- des données qu'auraient normalement un utilisateur réalisant la tâche

Cela peut être une simple phrase expliquant le but :

- acheter un billet d'avion pour l'Espagne en juillet.
- un scénario plus détaillé pour clarifier la motivation : partir en vacance avec des amis cet été au chaud.

Les participants

Au plus proches des utilisateurs du produit

Il faut filtrer les participants (ne pas prendre le 1e venu)

Prévoir que le recrutement de participants a un cout

- en temps, ou
- en argent…

Tester !

Vérifier :

- Le prototype à tester
- La configuration technique (ordinateur, écran, résolution, connexion)
- Outils de prise de note papier ou sur ordinateur
- Formulaires de consentement (avec stylo)
- Questionnaires, au besoin
- Une copie du scénario
- Des caméras, micros et autres outils d’enregistrement

Faire des essais !

Avant de commencer

Il faut savoir et avoir mis au clair :

- l'objectif
- une description du système testé
- l'environnement et le matériel
- les participants
- la méthodologie
- les tâches
- les mesures

Cela aidera à concevoir une bonne étude

Cela aidera à l'analyse des données

Les laboratoires de test

- Pièces spécialement conçues
  - Avec des outils d'enregistrement
  - e.g. micros, caméras
- Pièce d'observation séparée
  - Souvent connectée à la salle d'étude
  - avec un mirror sans tain et de l'audio
- Les participants réalisent les scénarios
  - Technique de "*Think aloud*"
  - Décider quand interrompre ou non
  - Minimiser les variations entre tests

From C|Net "How Google tested Google Instant"
http://news.cnet.com/8301-30684_3-20019652-265.html

Think aloud : penser tout haut

Besoin de savoir ce que les gens pensent pas seulement ce qu'ils font

Demander aux participants de parler pour dire

- ce qu'ils pensent
- ce qu'ils essaient de faire
- les questions qui émergent dans leur tête
- ce qu'ils voient / lisent

Relancer **très** régulièrement

- "*Dites moi à quoi vous pensez*"
- Bien enregistrer tous les moments ou vous donnez de l'aide

Démarche

Faire un enregistrement (et/ou des notes)

- confirmer que vous voyez ce qu'il se passe
- utiliser une montre / un chronomètre
- prendre des notes, un enregistrement audio & vidéo et des logs
- bien enregistrer tous les moments ou vous donnez de l'aide

Analyse

- Résumer les données
  - faire une liste des incidents critiques (positifs et négatifs)
  - inclure des données pour “montrer”
  - essayer d’estimer pourquoi certains ont rencontré telle ou telle difficulté
- Que disent les données ?
  - Est ce que l’interface marche comme elle devrait
  - Est ce que les utilisateurs se comportent de manière attendue
  - Est ce qu’il manque quelque chose?
- Mettre à jour l’analyse des tâches et repenser le design
  - Noter la sévérité et la facilité / complexité à réparer les problèmes

Mesures d'utilisabilité

Situations pour lesquelles les chiffres sont utiles :

- temps pour réaliser une tâche
- tâches réussies
- comparer deux designs en termes de vitesse ou de nb d'erreurs

Mesures

- le temps est facile à enregistrer
- les erreurs et les succès plus difficiles,
  définir en avance à quoi cela correspond

Ne pas combiner mesures d'efficacité et thinking-aloud.

- parler / s'expliquer va affecter les performances

Laboratoire d'utilisabilité basique

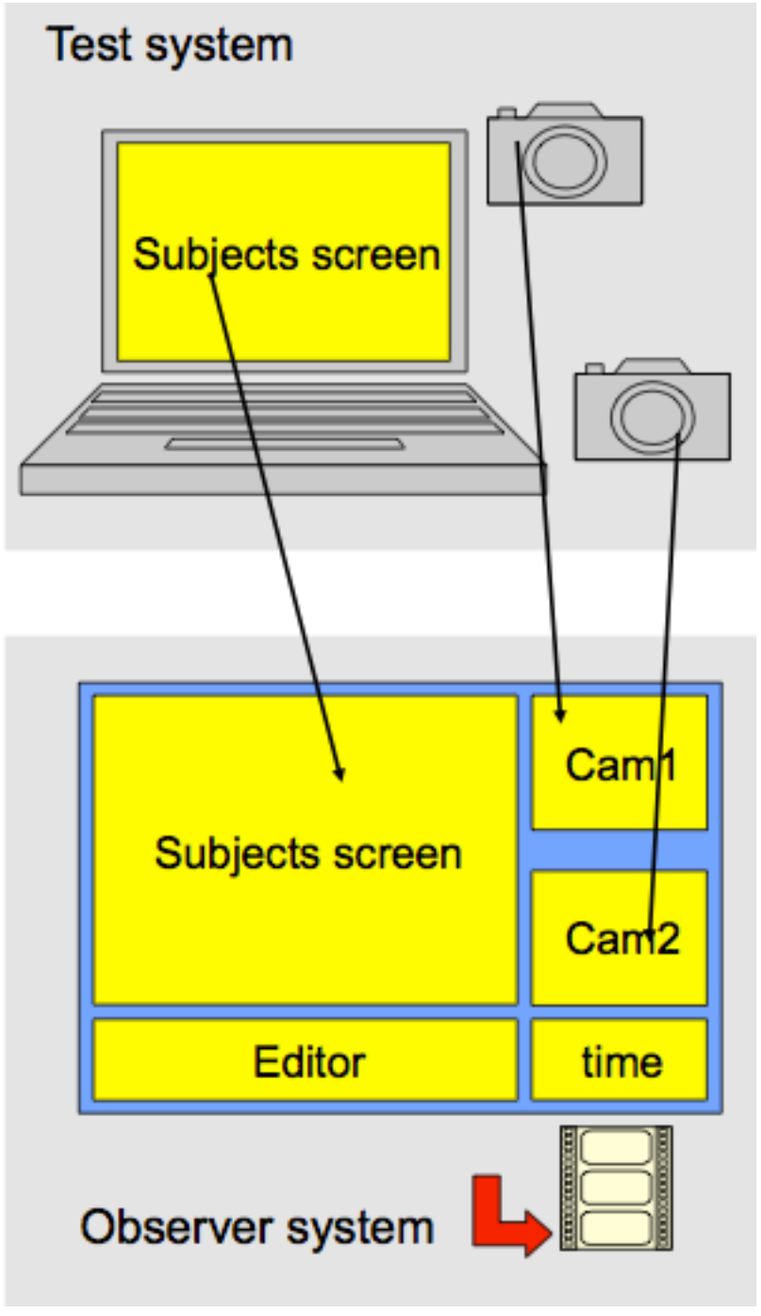
Objectif : avoir plusieurs vues

- Capturer l'écran (avec le pointeur)
- Voir la personne interagir
- Voir l'environnement

Mise en place :

- Un ordinateur pour le participant
- Un ordinateur pour l'observateur

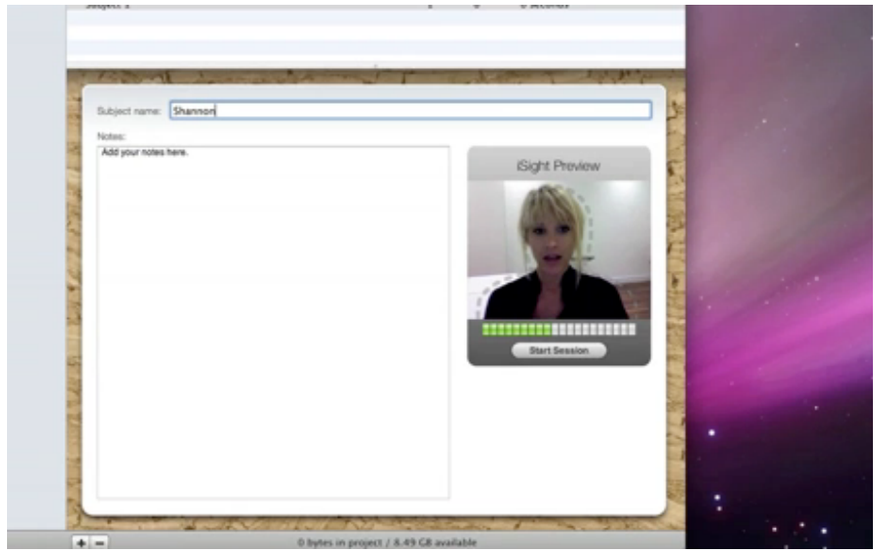
Un debrief à la fin



Outils existants

Ovo studio

- gratuit pour les étudiants
- http://www.ovostudios.com



Silverback

- http://silverbackapp.com/

Morae

- http://www.techsmith.com/morae.html

Vraiment rapide et peu cher

Évaluation empiriques

- Focus groups
- Études de terrain
- Étude d’utilisabilité
- **Mesures physiologiques**

Mesure physiologiques

Eye tracking

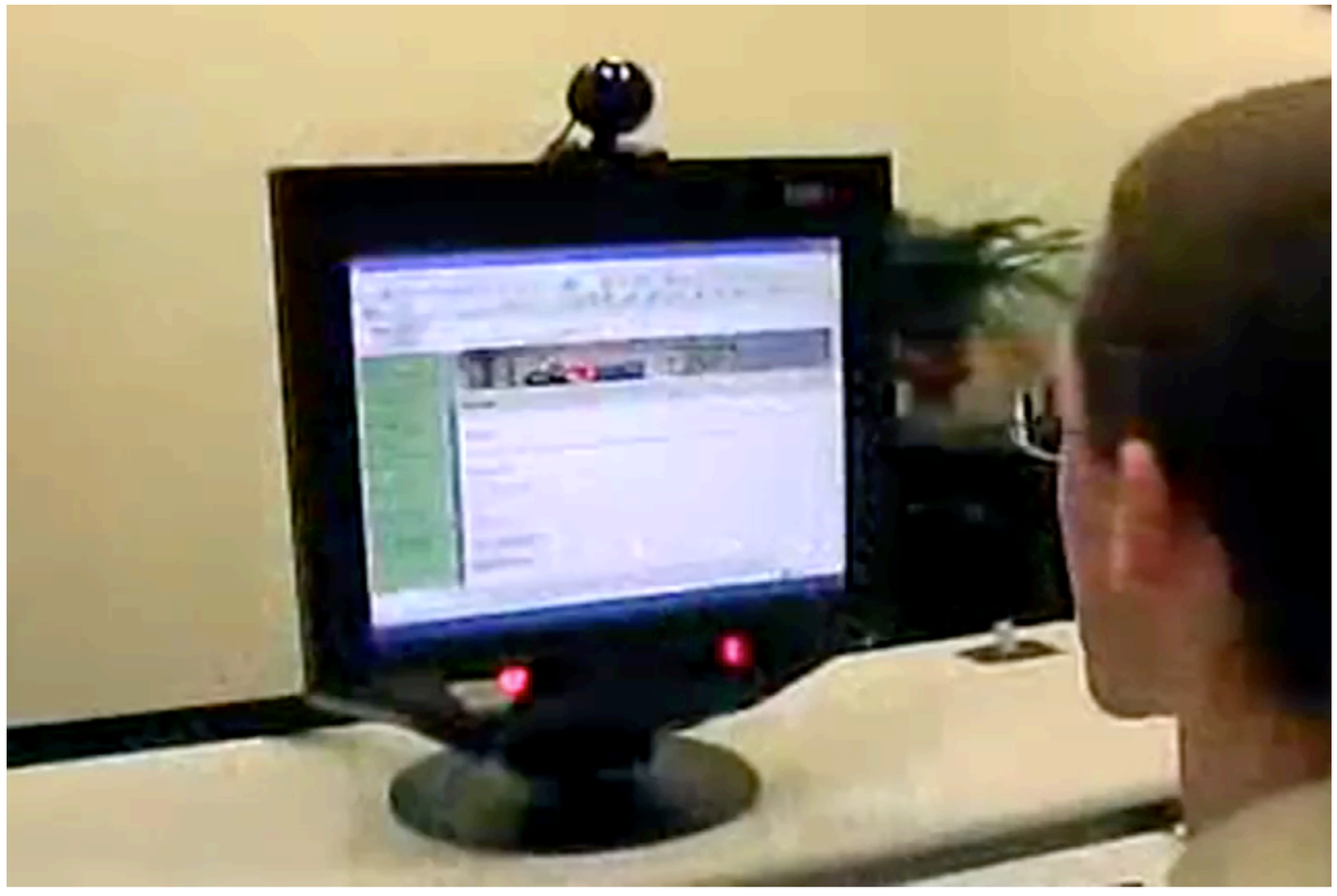
- Bien développé
- Robuste
- Nouveaux outils peu cher

Stress

- Conductivité de la peau

Activité cérébrale

- expérimental



Eye-tracker - © Kent State University (US)

Évaluation et tests

- Introduction
- Approches d'évaluation
- Méthodes analytiques
- Méthodes empiriques
- **Évaluation 2.0 : passer à l'échelle**
- Design expérimental

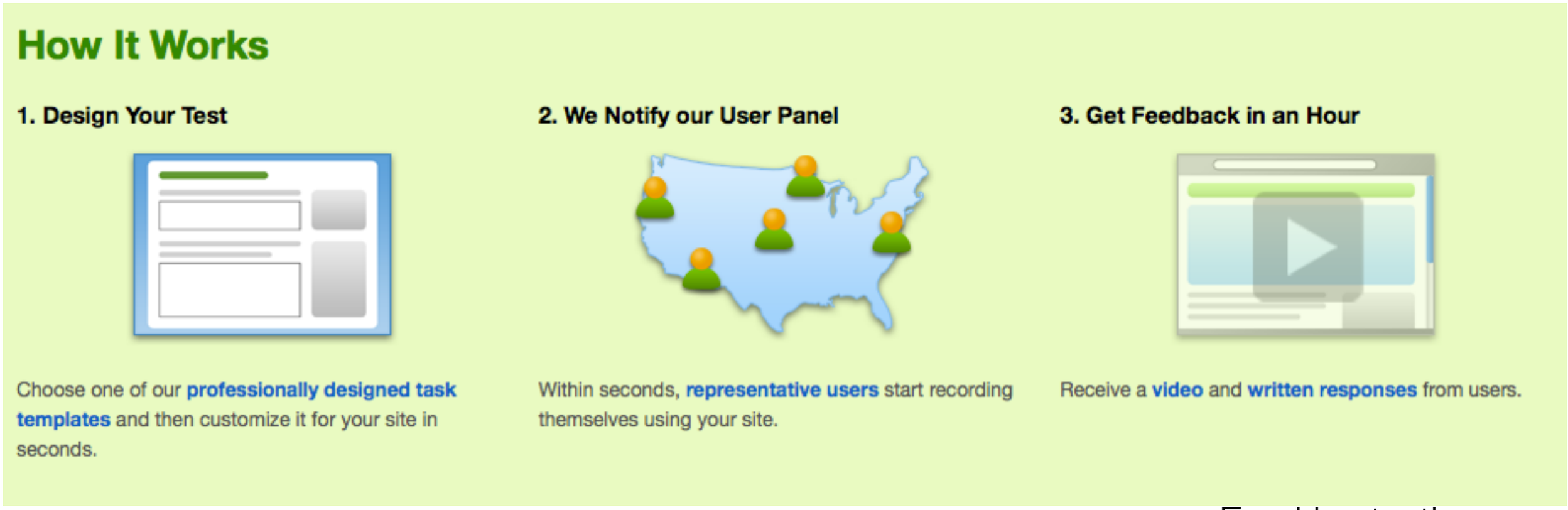
Passer les tests d’utilisabilité à l’échelle

Grandes audiences sur le Web

Grandes audiences sur les plateformes mobiles

Distribution facile et mise à jour rapides

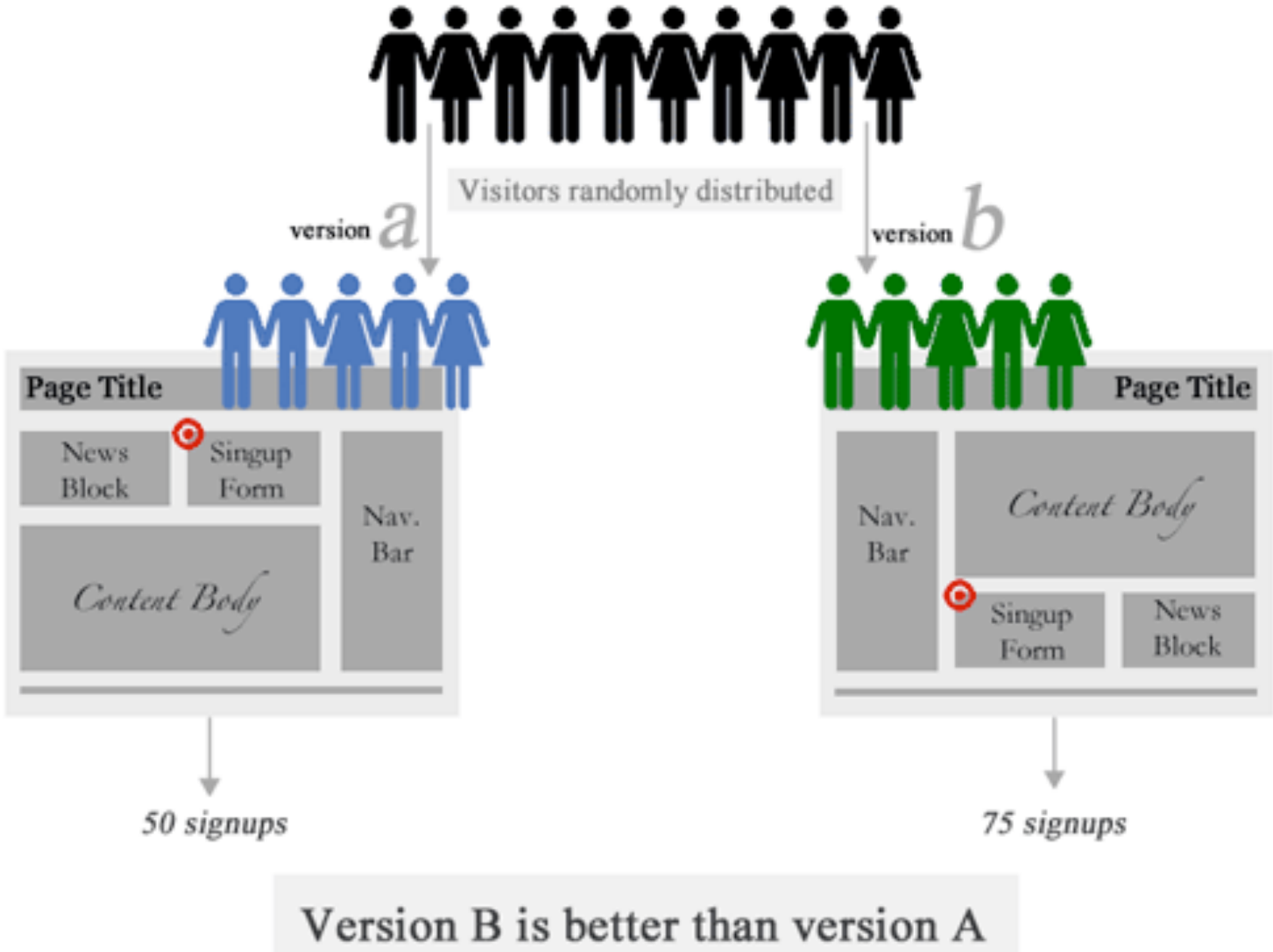
Etudes d’utilisabilité à distance



E.g. Usertesting.com

From http://www.smashingmagazine.com/2010/06/24/the-ultimate-guide-to-a-b-testing/

Tests A / B



Tests A/B

ex : http://optimizely.com/

Tester une meilleure page d'accueil

- un design de formulaire plus efficace
- un meilleur taux de conversion

Limites :

- Ne remplace pas les études utilisateurs !
- Ne fournit pas d'explication.
- Les changements arbitraires peuvent être dérangeant
- Compliqué lors qu'il y a des interactions sociales (ex : Facebook)
- Souvent utilisés pour comparer des changements incrémentaux, devient compliqué dans le cas de re-designs complets

Distribution contrôlées de versions Beta

Évaluation et tests

- Introduction
- Approches d'évaluation
- Méthodes analytiques
- Méthodes empiriques
- Évaluation 2.0 : passer à l'échelle
- **Design expérimental**

Design expérimental

- Introduction et exemples
- Les éléments de base d'une expérience
- Définition d'une expérience
- Conduite de l'expérience
- Récupération et nettoyage des données
- Analyse des données
  - Exploratoire
  - Statistique

Design expérimental

- **Introduction et exemples**
- Les éléments de base d'une expérience
- Définition d'une expérience
- Conduite de l'expérience
- Récupération et nettoyage des données
- Analyse des données
  - Exploratoire
  - Statistique

Expériences contrôlées

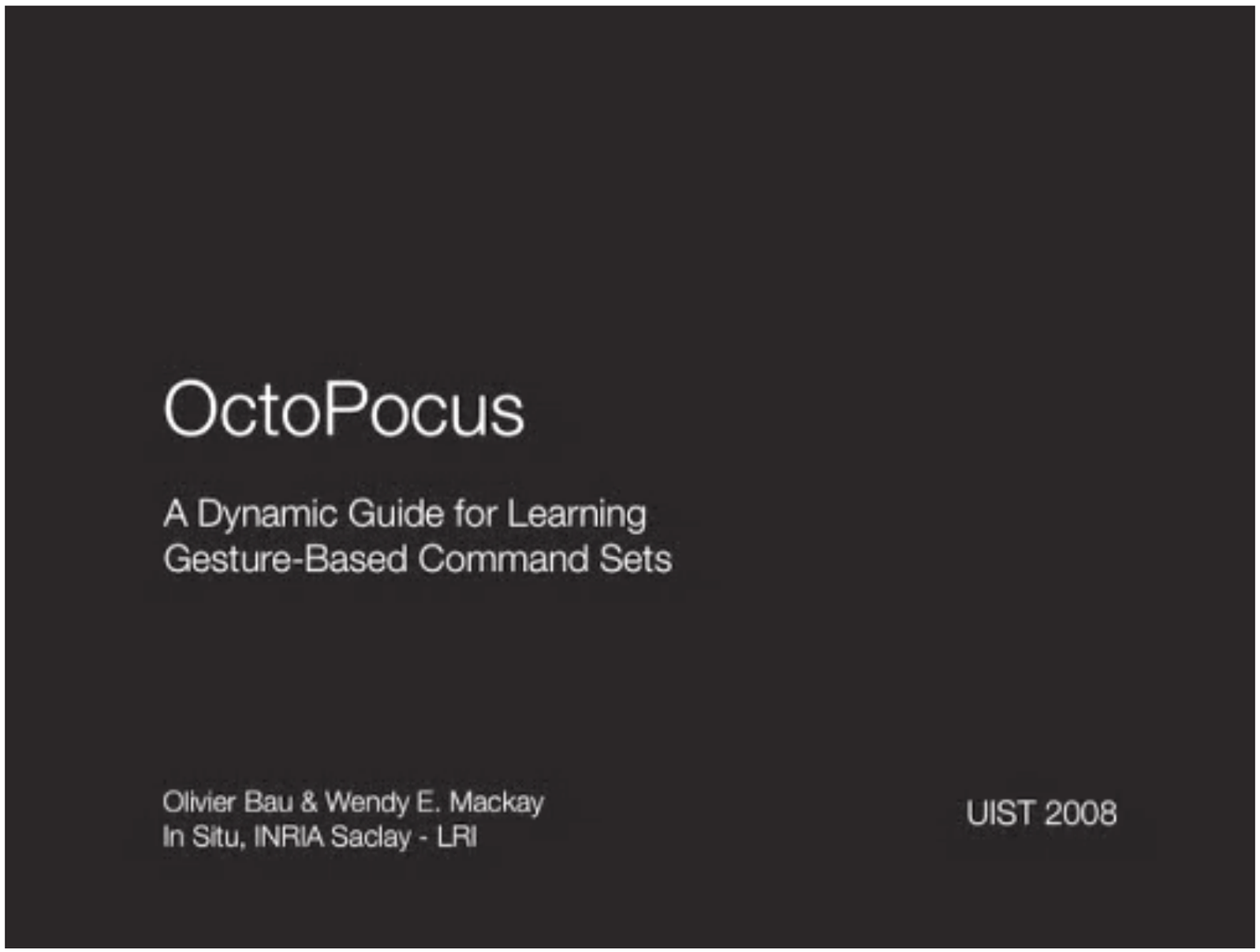
Une approche scientifique

- Réponses spécifiques à des questions
  - Performance
  - Apprentissage
  - Satisfaction
- Connaissance généralisables dans plusieurs contextes
- Montrer un lien de causalité
  - corrélation : montrer qu'un changement dans A entraine un changement dans B
  - ordre : montre que A arrive avant B
  - pas de cause cachée : montrer qu'il n'y a pas de C avec C -> A et C -> B

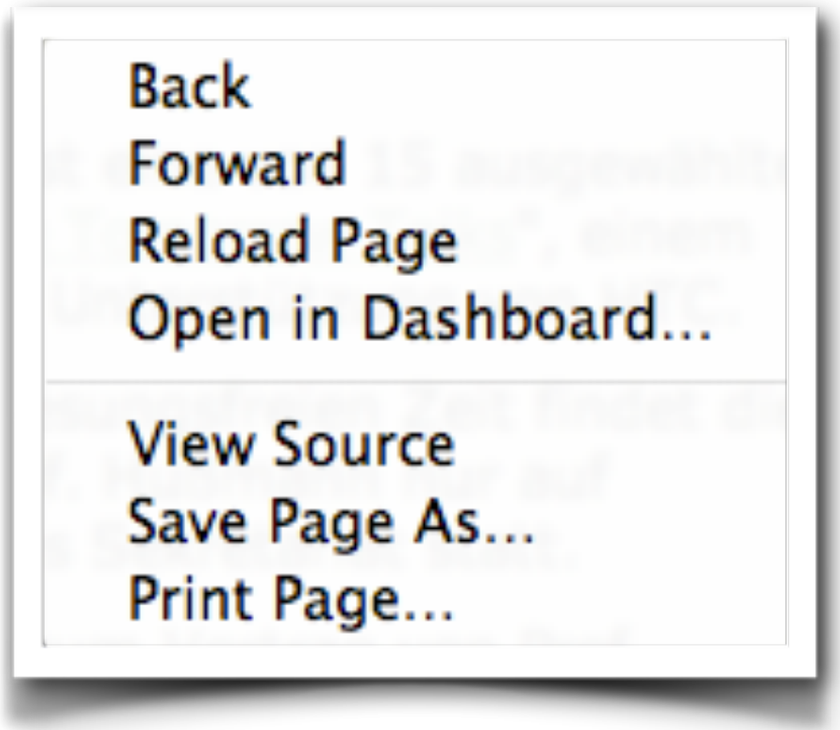
Exemple : Dispositifs de contrôle en entrée

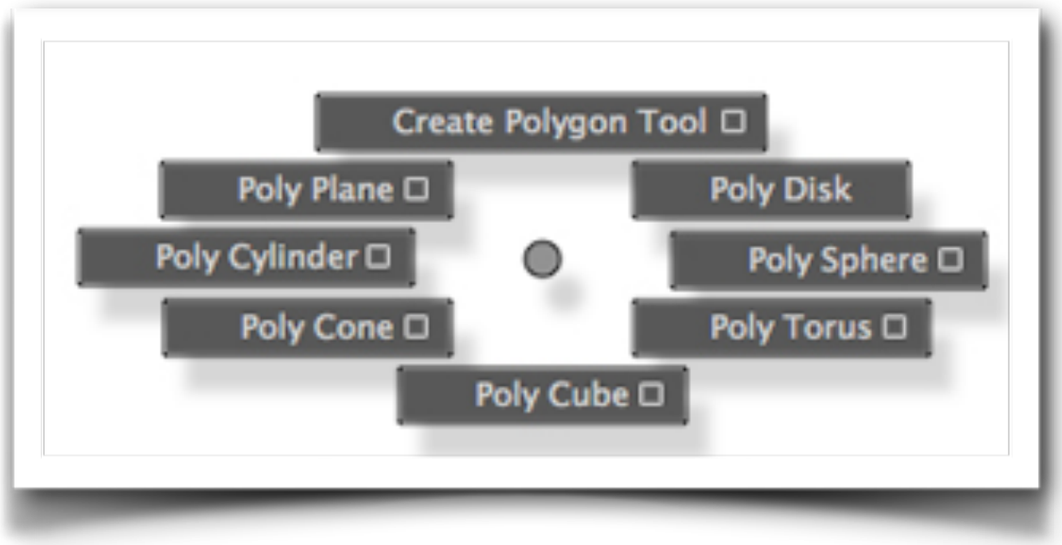
Dispositif	Étude	IP (bits/s)
Main	Fitts (1954)	10,6
Souris	Card, English, & Burr (1978)	10,4
Joystick	Card, English, & Burr (1978)	5
Trackball	Epps (1986)	2,9
Touchpad	Epps (1986)	1,6
Eyetracker	Ware & Mikaelian (1987)	13,7

Exemple : Apprentissage de gestes



Exemple : comparer deux designs de menu





Design expérimental

- Introduction et exemples
- **Les éléments de base d'une expérience**
- Définition d'une expérience
- Conduite de l'expérience
- Récupération et nettoyage des données
- Analyse des données
  - Exploratoire
  - Statistique

Méthode 1.

1. Définir ce qu'on cherche : formuler une hypothèse
   - Le menu circulaire réduit le temps de recherche des éléments
2. Concevoir l'expérience, choisir variables et paramètres fixes
   - Définir la structure des menus
3. Conduire une expérience pilote pour tester l'expérience
   - Améliorer le design des menus

Méthode 2.

4. Sélectionner des participants
   - Des étudiants passant plus de 2h / j sur leur ordinateur
5. Conduire l’expérience et collecter les données
6. Analyser les données pour accepter ou rejeter l’hypothèse et mesurer l’effet de la variation
   - Temps de recherche moyen : 2.26 (Circulaire), 2.64 (Classique)
   - La différence est significative : $p<.05$

Les éléments d'une expérience

- Facteurs (ou **variables indépendantes**)
  - Les variables qu'on va changer pour chaque condition
  - La quantité d'éléments dans un menu, ou le nombre de sous-menus
- Niveaux (les valeurs possibles des variables indépendantes)
  - Un menu avec 8 éléments ou un menu avec 12
- Mesures (ou réponses, ou **variables dépendantes**)
  - Les résultats mesurés de l'expérience.
  - Le temps de sélection d'un élément
- Réplication
  - le nombre de participants pour chaque niveau

Les variables indépendantes (facteurs)

- Les conditions d'une expérience sont définies par ces variables
  - The number of items in a list, text size, font, color
- Le nombre de variables différentes constitue le niveau
  - E.g., font can be times or arial (2 levels), background can be blue, green, or white (3 levels). This results in 6 experimental conditions (times on blue, times, on green, ..., arial on white)

Les variables dépendantes

Les variables dépendantes sont les variables qui vont être mesurées :

- Mesures objectives : e.g. temps pour réaliser une tâche, nombre d'erreurs, etc.
- Mesures subjectives : plaisir, frustration, etc.
- Elles ne devraient varier qu'en fonction de changements dans les variables indépendantes (= être fixes sinon).

Exercice

Identifier les variables indépendantes (facteurs) et les variables dépendantes (mesures) dans chaque scénario. Donner des niveaux possibles des variables indépendantes.

- Une étude pour voir si les gens ayant suivi une formation en sécurité utilisent des mots de passe plus sécurisés.
- Une étude qui cherche à voir qui d'un joystick ou d'une souris est plus efficace pour sélectionner des cibles statiques ou des cibles qui bougent
- Une étude pour savoir si les équipes qui utilisent du video Hangout/Skype sont plus productives que celles qui utilisent seulement des chats textes.

Les participants / sujets

Intra-sujets (Within-subjects design) :

- Participants exposés à toutes les conditions

+Besoin de peu d'utilisateurs (10 - 20)

- Effet d'apprentissage

Inter-sujets (Between-subjects design) :

- Les participants sont séparés en groupes et chaque groupe exposé à une condition (contrôle et traitement)

+Pas d'effet d'apprentissage

- demande plus d'utilisateurs

Gérer les effets d'ordre

L'ordre de présentation des traitements peut avoir un effet sur les mesures :

- apprentissage
- fatigue
- contraste (le premier traitement se reporte sur les réponses du 2e traitement)

Solution

- repos entre les traitements
- Équilibrer (counterbalancing), mais ça peut devenir compliqué
- Carré Latin

Carré Latin

via http://hci.rwth-aachen.de/~chat/StatLecture/prerequisite.pdf

Chaque condition apparait dans chaque position possible

Chaque condition précède chaque autre condition une fois

Exemple pour 6 traitements :

1	A	B	F	C	E	D
2	B	C	A	D	F	E
3	C	D	B	E	A	F
4	D	E	C	F	B	A
5	E	F	D	A	C	B
6	F	A	E	B	D	C

Exercice

Quel type de design expérimental ?

- Une étude pour voir si les gens ayant suivi une formation en sécurité utilisent des mots de passe plus sécurisés.
- Une étude qui cherche à voir qui d'un joystick ou d'une souris est plus efficace pour sélectionner des cibles statiques ou des cibles qui bougent
- Une étude pour savoir si les équipes qui utilisent du video Hangout/Skype sont plus productives que celles qui utilisent seulement des chats textes.

Hypothèses

- Prédire le résultat d'une expérience
- Dire comment un changement dans les variables indépendantes va impacter les variables dépendantes

Approche classique

- Formuler une hypothèse de travail H_1
- Formuler une hypothèse nulle H_0
  - intuition (naive) : si H_0 est faux alors H_1 doit être vrai
- Conduire une expérience et l'analyse statistique qui va avec pour falsifier cette hypothèse
- Si l'analyse statistique montre une différence significative il est probable que le résultat de soit pas dû au hasard

Exercice

Quel H_0 et H_1 ?

- Une étude pour voir si les gens ayant suivi une formation en sécurité utilisent des mots de passe plus sécurisés.
- Une étude qui cherche à voir qui d'un joystick ou d'une souris est plus efficace pour sélectionner des cibles statiques ou des cibles qui bougent
- Une étude pour savoir si les équipes qui utilisent du video Hangout/Skype sont plus productives que celles qui utilisent seulement des chats textes.

Validité

Validité interne

- Validité des résultats en cas de réplication

Validité externe

- Confiance de généralisation des résultats au monde réel

Cas pratique

Comparaison de la vitesse d'écriture entre un clavier d'ordinateur et un clavier mobile.

- Quelles variables indépendantes (facteurs) ?
- Quelles variables dépendantes (mesures) ?
- inter- ou intra- participants ?
- Quelles hypothèses ?

Design expérimental

- Introduction et exemples
- Les éléments de base d'une expérience
- Définition d'une expérience
- **Conduite de l'expérience**
- Récupération et nettoyage des données
- Analyse des données
  - Exploratoire
  - Statistique

Récupération des données

Observations de ce qu'on fait les utilisateurs

Données de logs

- Données des variables dépendantes (temps, erreurs)

Structure en tableau :

userid	group	condition	executiontime	error

Cas pratique

- Quel protocole ?
- Quel format des données collectées

Allez voir

- http://10fastfingers.com/ pour vous donner une idée.

Design expérimental

- Introduction et exemples
- Les éléments de base d'une expérience
- Définition d'une expérience
- Conduite de l'expérience
- **Récupération et nettoyage des données**
- Analyse des données
  - Exploratoire
  - Statistique

Récupération des données

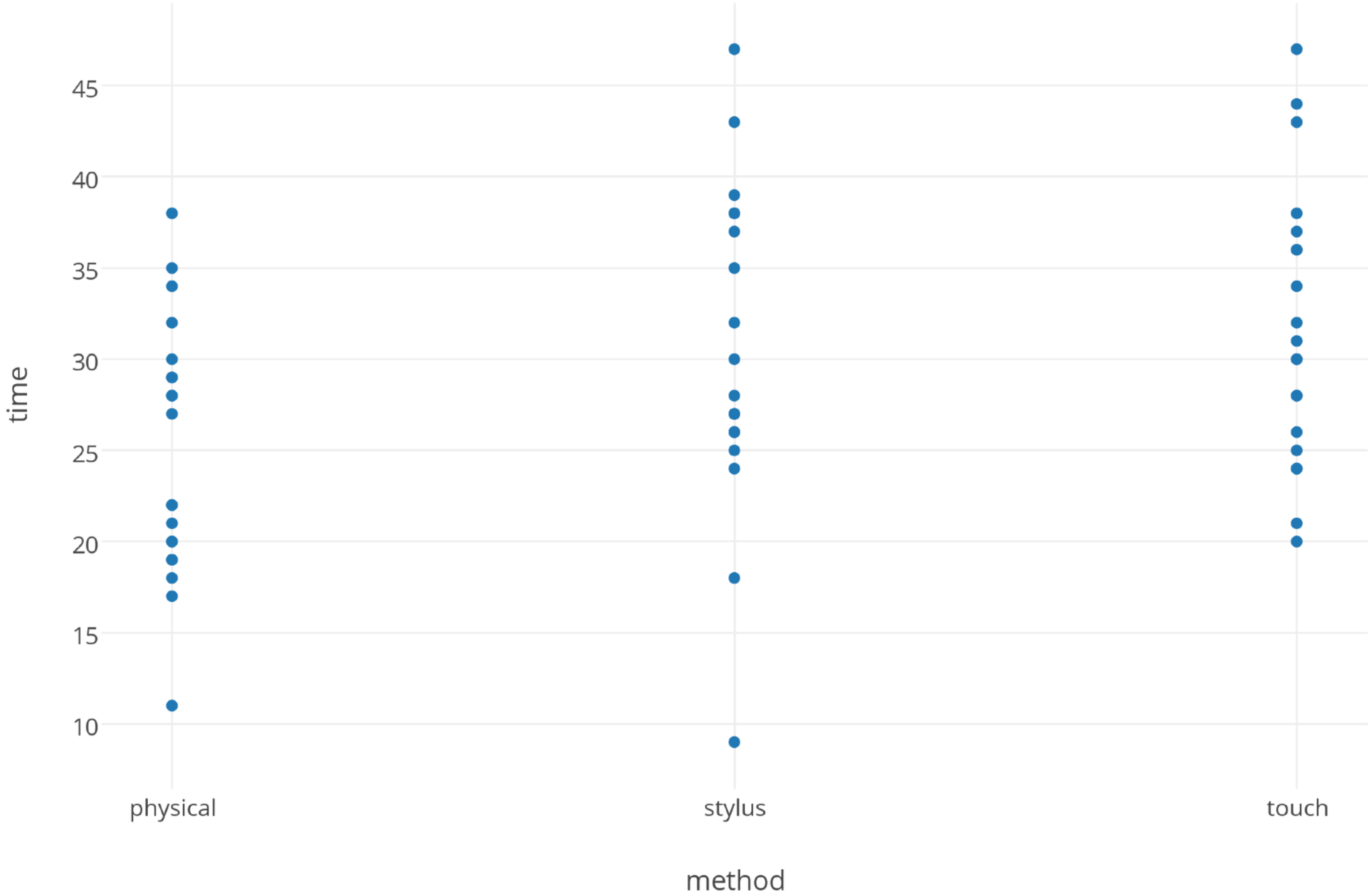
Agréger les données de tous les groupes dans un pad avec un format csv (comma separated value).

Les copier sur votre comptes plot.ly

Design expérimental

- Introduction et exemples
- Les éléments de base d'une expérience
- Définition d'une expérience
- Conduite de l'expérience
- Récupération et nettoyage des données
- **Analyse des données**
  - Exploratoire
  - Statistique

Qui est le plus rapide ?



Qui est le plus rapide

Ça dépend de savoir

- la différence entre les médianes
- la distribution des données (la déviation standard)
- la taille de l'échantillon
- si les moyennes sont significativement différentes

> Premier regard sur les données avec quelques graphes et des statistiques de base pour se faire une idée.

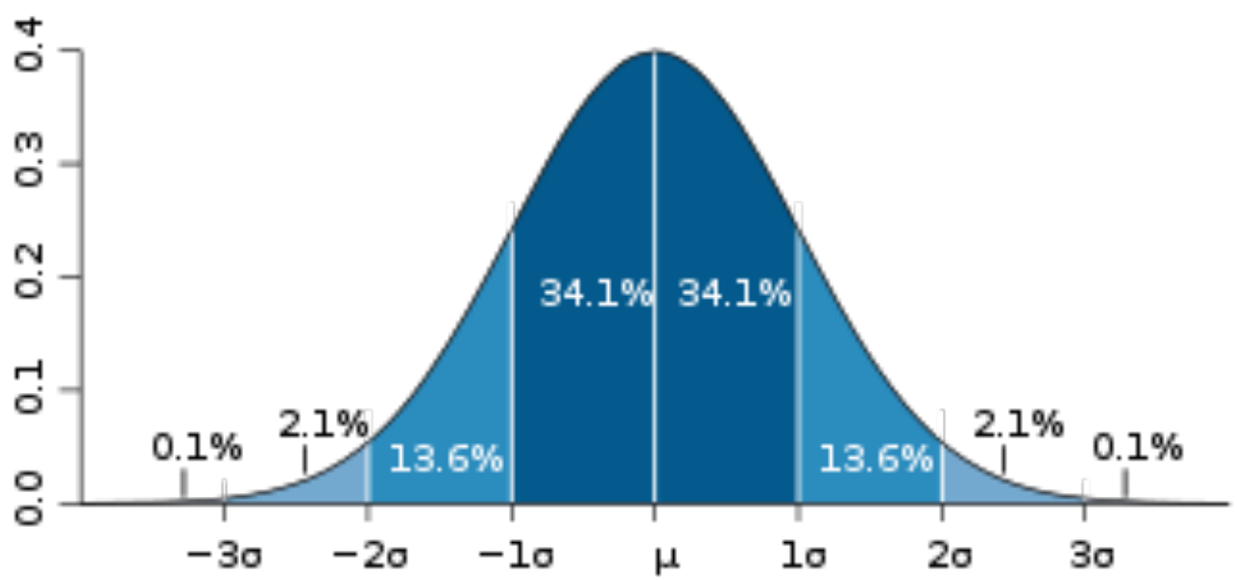
- partie exploratoire

(Student's) t-test

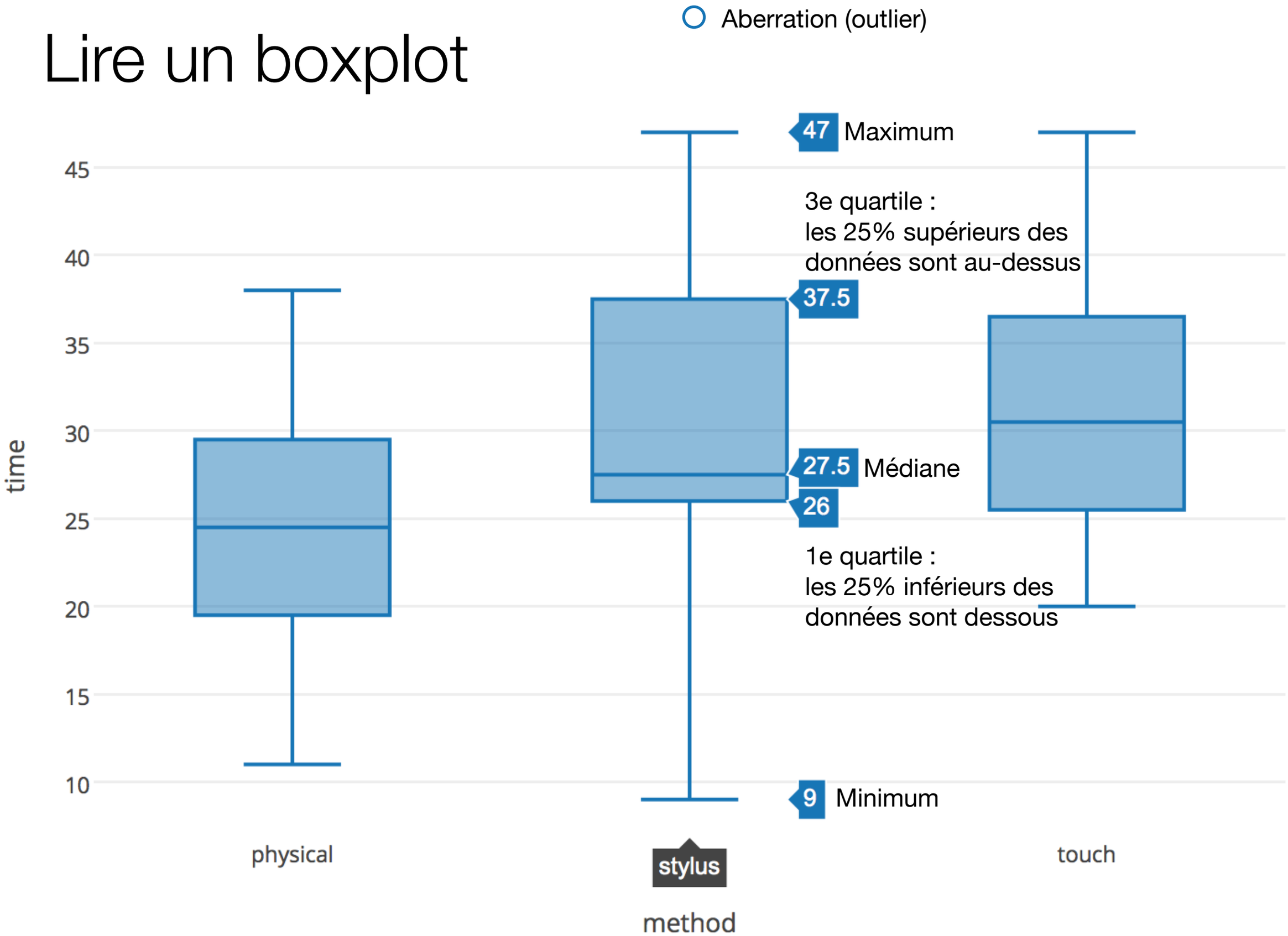
On regarde la relation entre deux jeux de données

Conçu pour :

- un petit échantillon (= peu de mesures)
- une déviation standard (et une moyenne) indéfinie
- mais une **distribution normale**



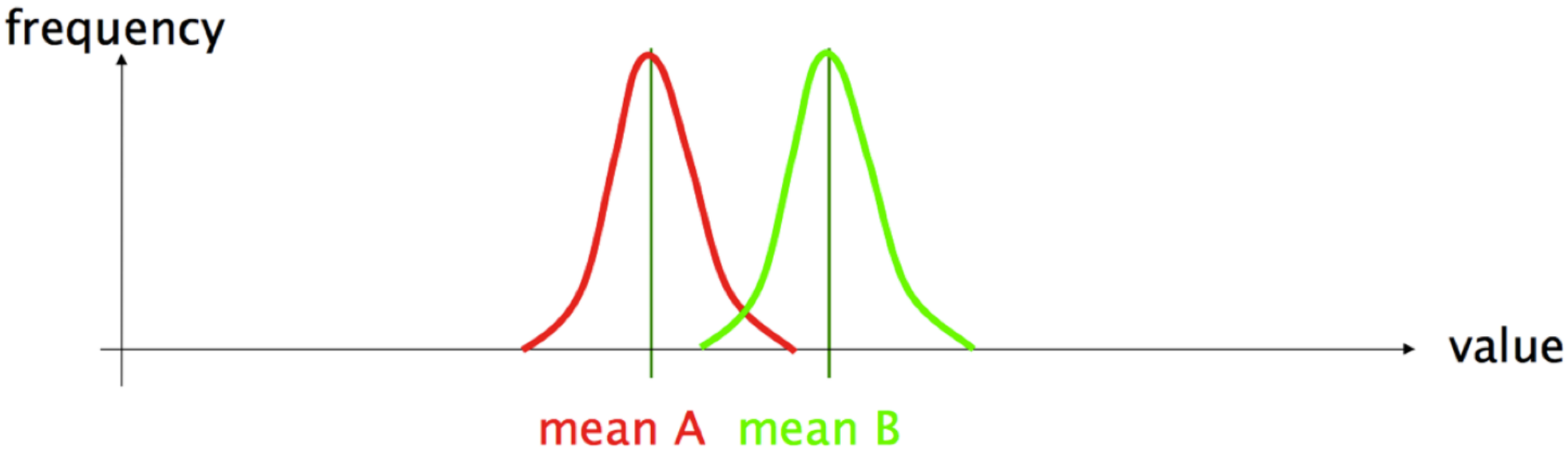
Lire un boxplot

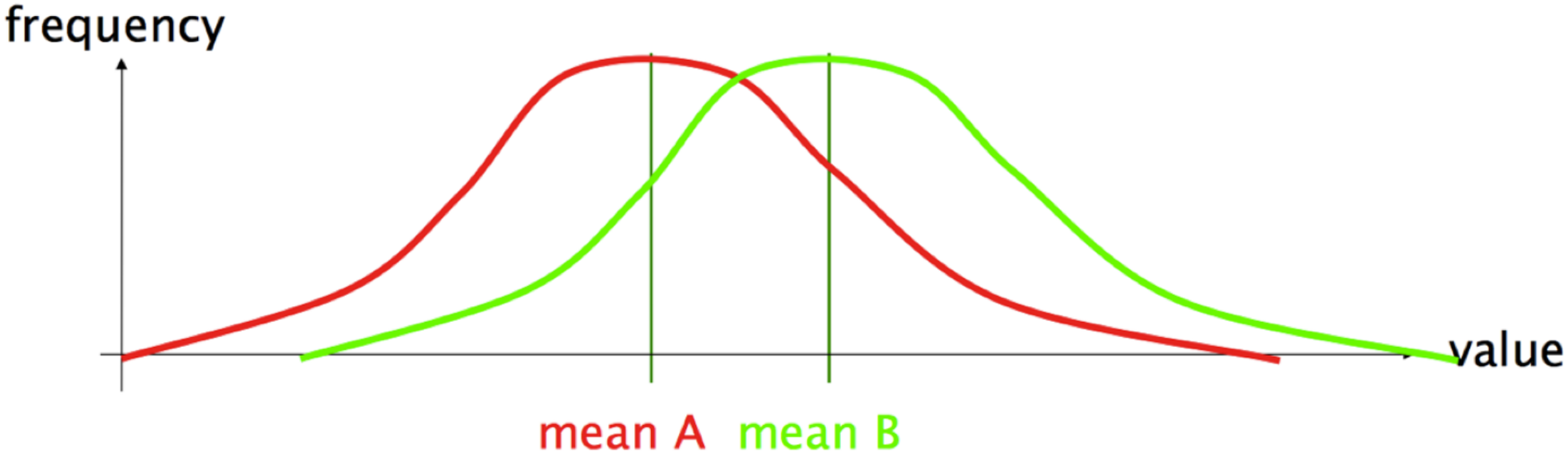


Comparer des valeurs

via http://www.medien.ifi.lmu.de/lehre/ws1213/mmi2/uebung/slides10.pdf

Y a t'il une différence significative entre deux mesures ?





t-test

Donne p :

- **la probabilité que les deux populations aient la même moyenne**
- Et non la probabilité que le résultat soit par hasard…

En UX :

- $p < 0.05$ (= 5% de probabilité) est la convention (ou 0.01)
- un p plus faible (ex : 0.00001) ne veut pas dire que le résultat est “plus” significatif.
- un résultat significatif est différent d’un résultat est important

NE PAS

Si p>0.05 dire :

- “*notre test a montré qu’il n’y avait pas de différence*”.

- différence significative -> il s’est passé quelque chose
- pas de différence significative -> rien

On ne peut pas montrer qu’il n’y a pas de différence
(avec les outils statistiques dont j’ai parlé)

t-statistique