Examples of recommendation systems ?

What can be recommended?

How recommender systems arrived on the internet

From the street shop to the on-line shop
From the material content to the digital content

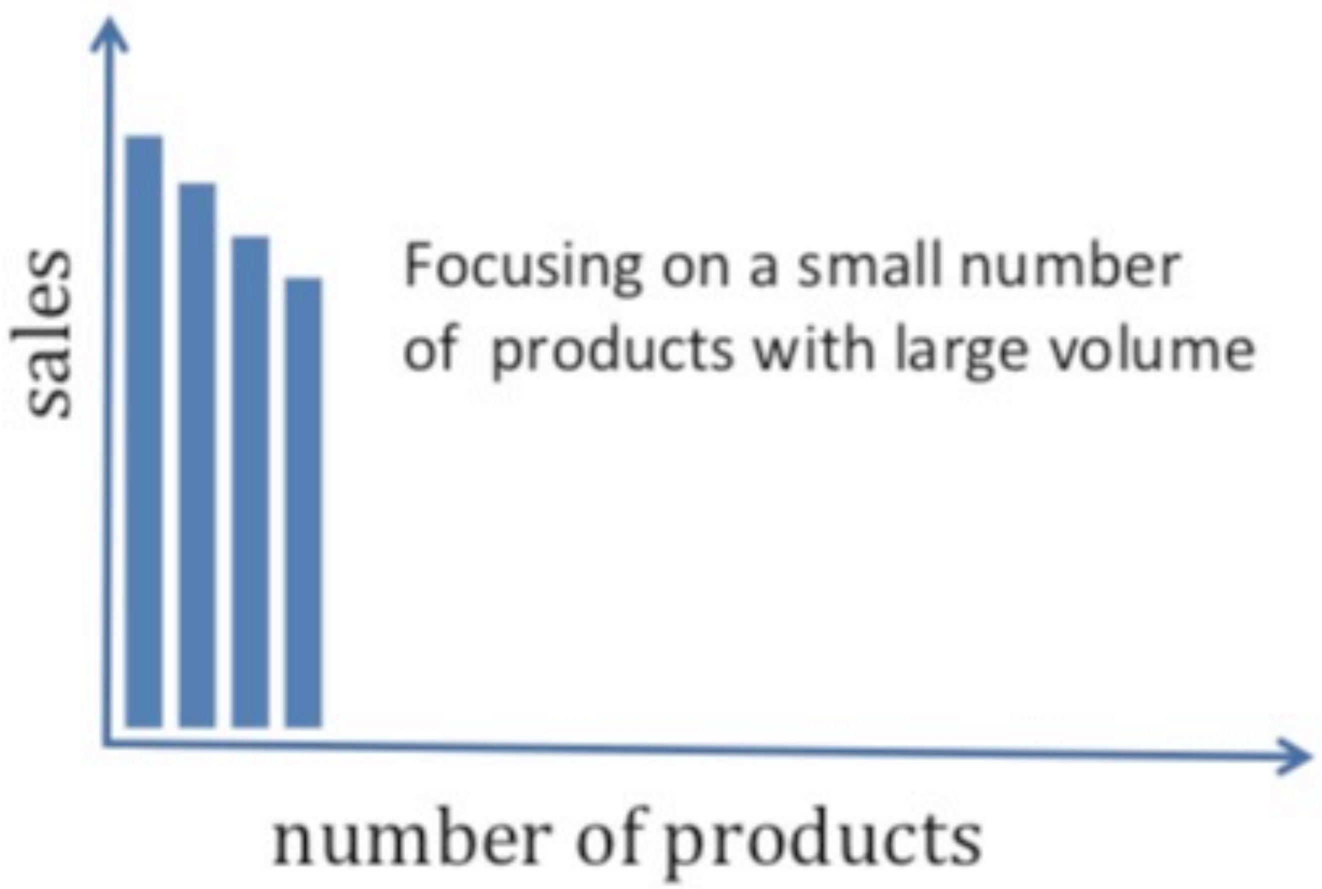
POLYTECH NANTES

Storage, shipping, duplication, costs (shops and items) have evolved, helping make large catalogues available

+

Properties of large catalogues

POLYTECH NANTES



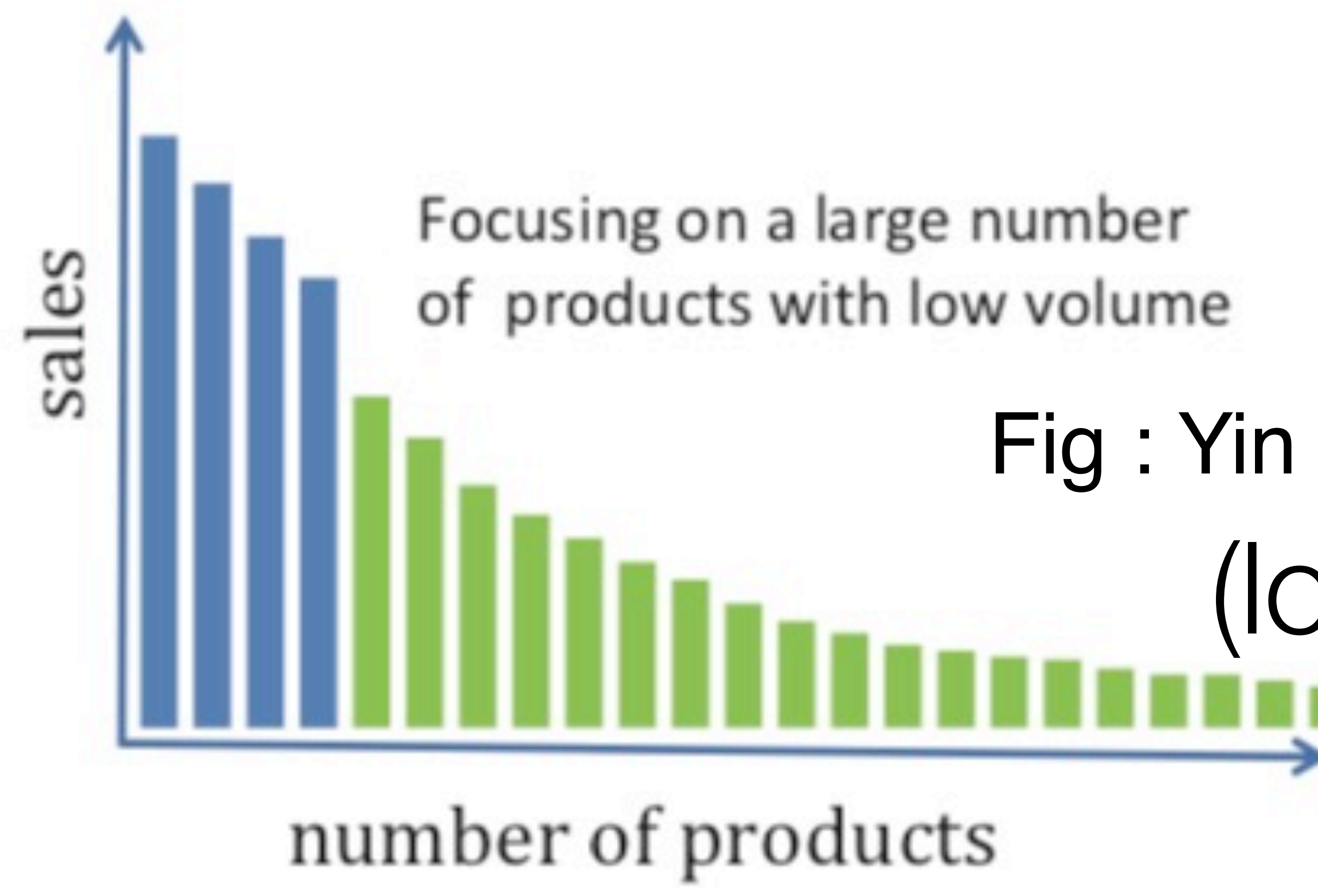


Fig : Yin et al. PVLDB 2012

(long trail)

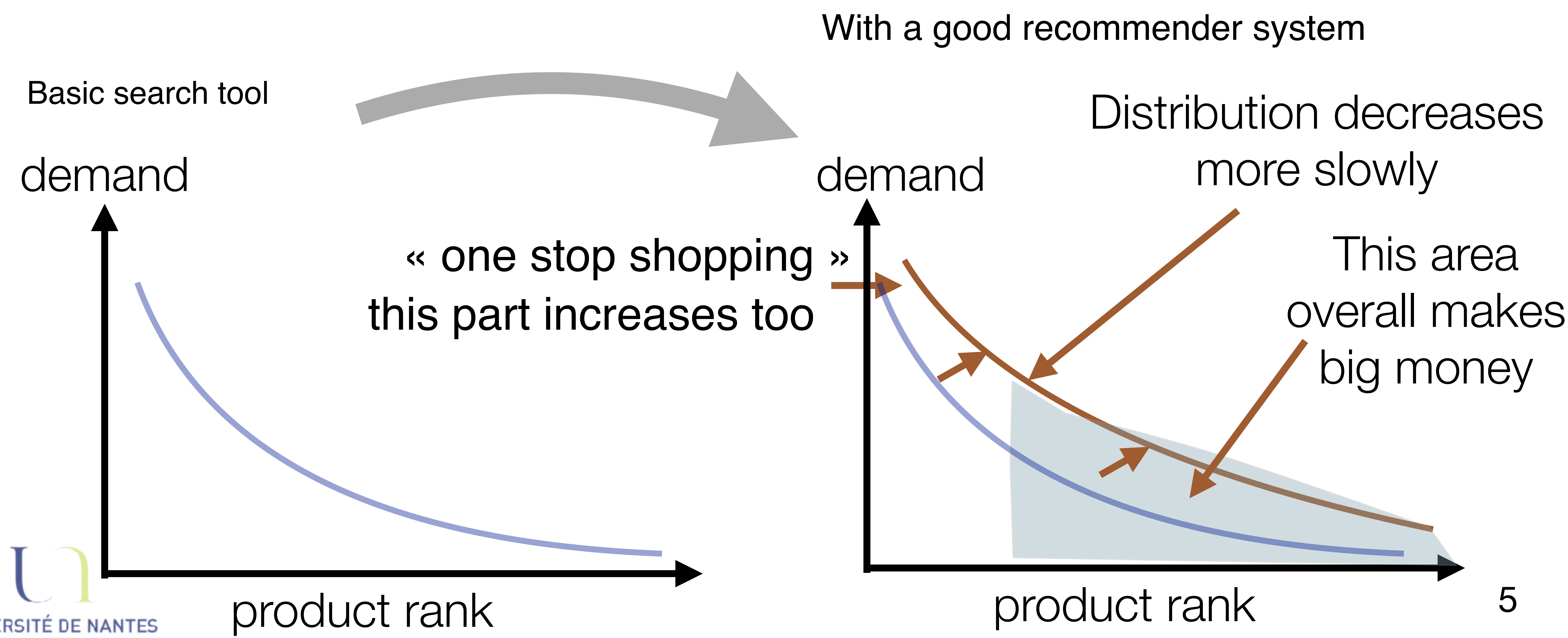
Margin is higher on rare items (less competition)
Bonus of one-stop shopping

Nice to have a large catalogue (sub-contracted ?)
But need to be able to exploit it
(otherwise the long train might not be beneficial)

UNIVERSITÉ DE NANTES

Large catalogue => hard to exploit to sales
because querying and browsing are not so easy

Experiments show you to can make the demand much higher
by adding recommendation to browsing and querying

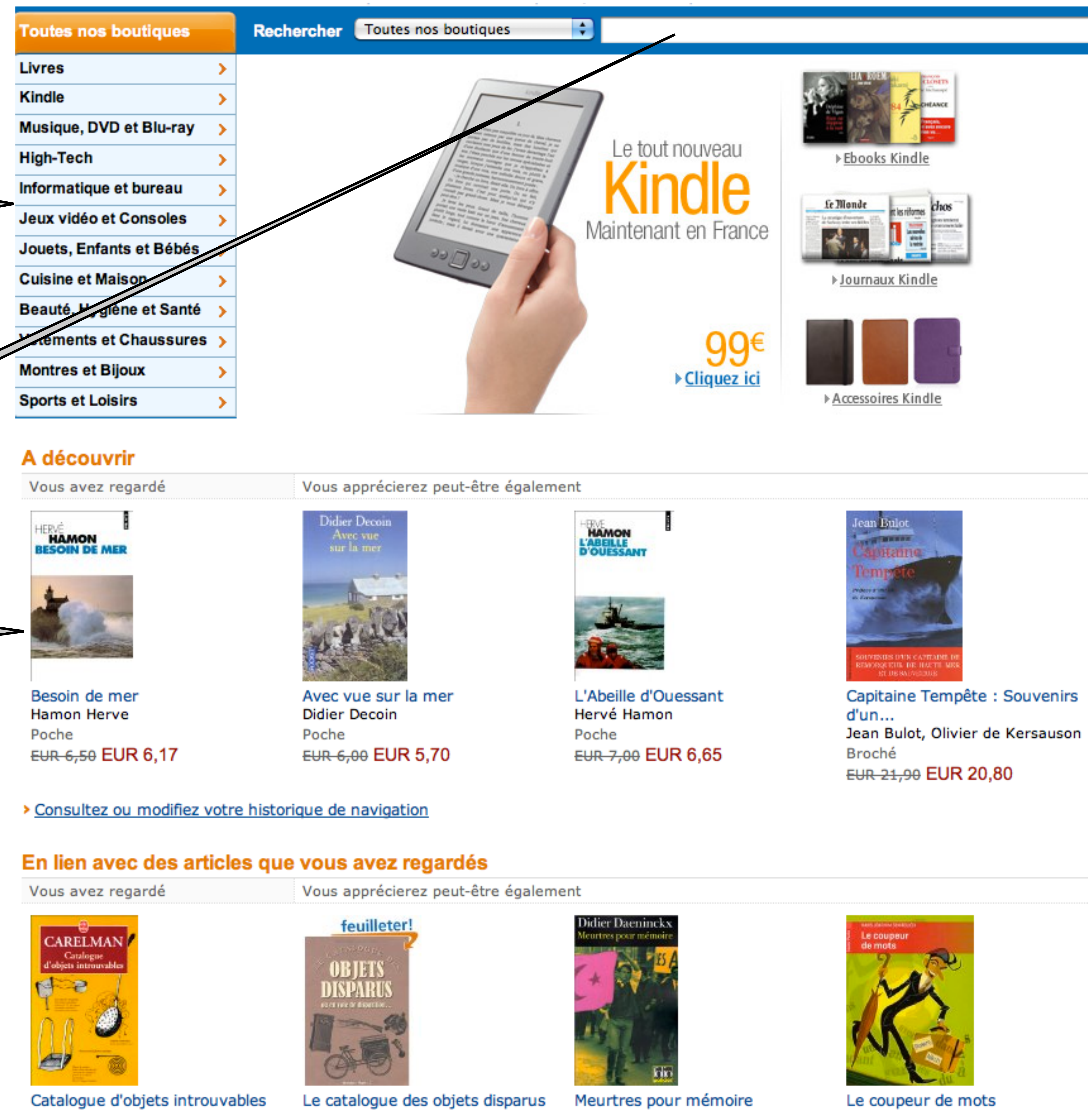


3 approaches to content selection example : Amazon

Browsing

Querying

Recommendation



But wasn't browsing/querying enough ?

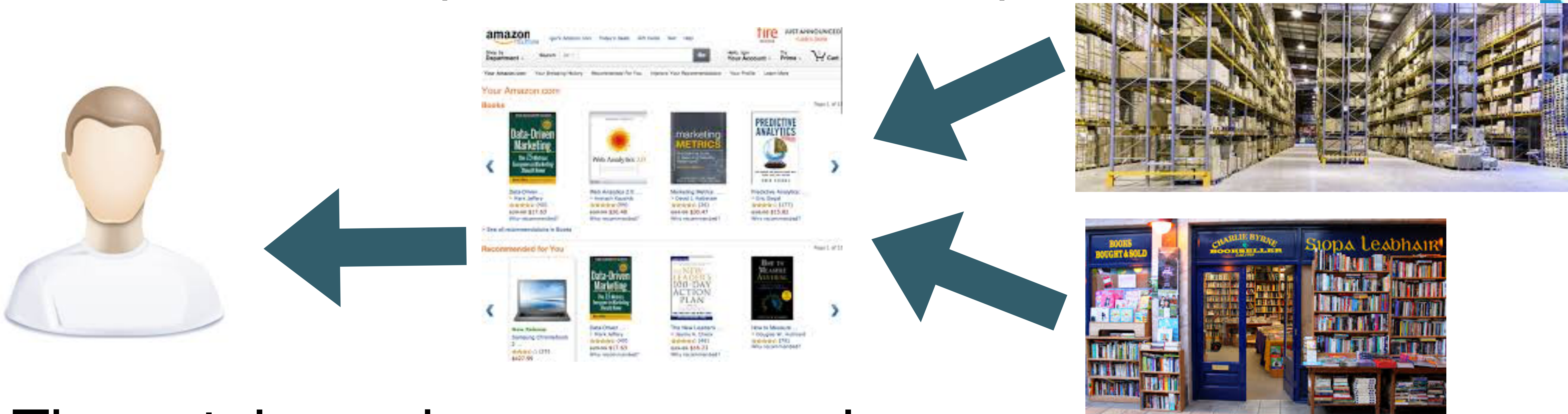
Catalogue = a **search space**

Querying, browsing (generalization) :
they enable exploring the space,
but most corners remain hard to reach
for most user knowledge

Serendipity : is a goal, a property (not a method) consisting in finding :
- interesting content
- that would not been found easily

Ability to jump to areas hard to reach spontaneously
by querying/browsing : this is one of the goals
of recommendation

From the on-line shop to the intermediation platform



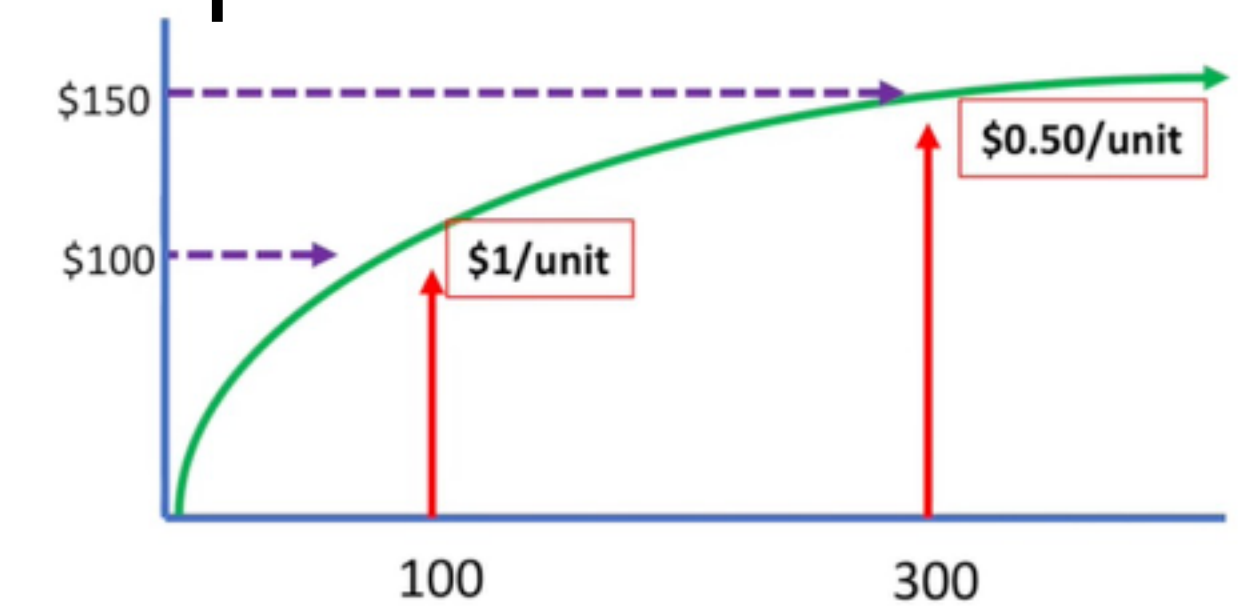
The catalogue becomes even larger
And plateforms can share providers :

The value of the portal is the catalogue AND the power of the search tool (including recommendation)

Network effects on platforms

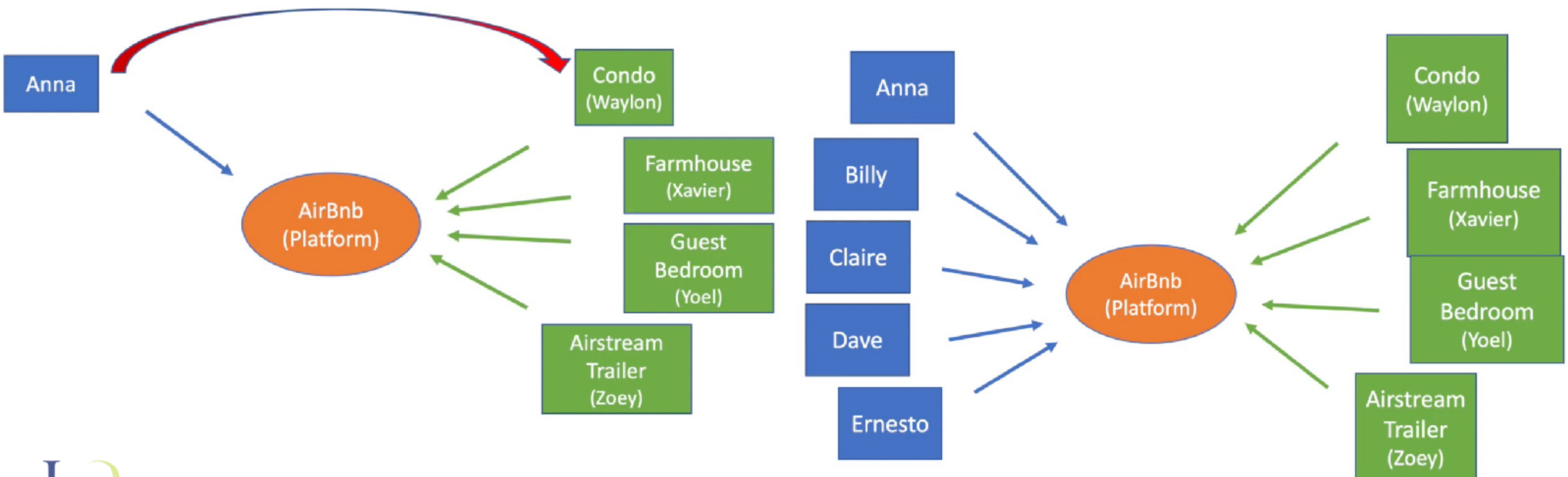
- Zoom
- MS Excel
- Instagram

cost for producer



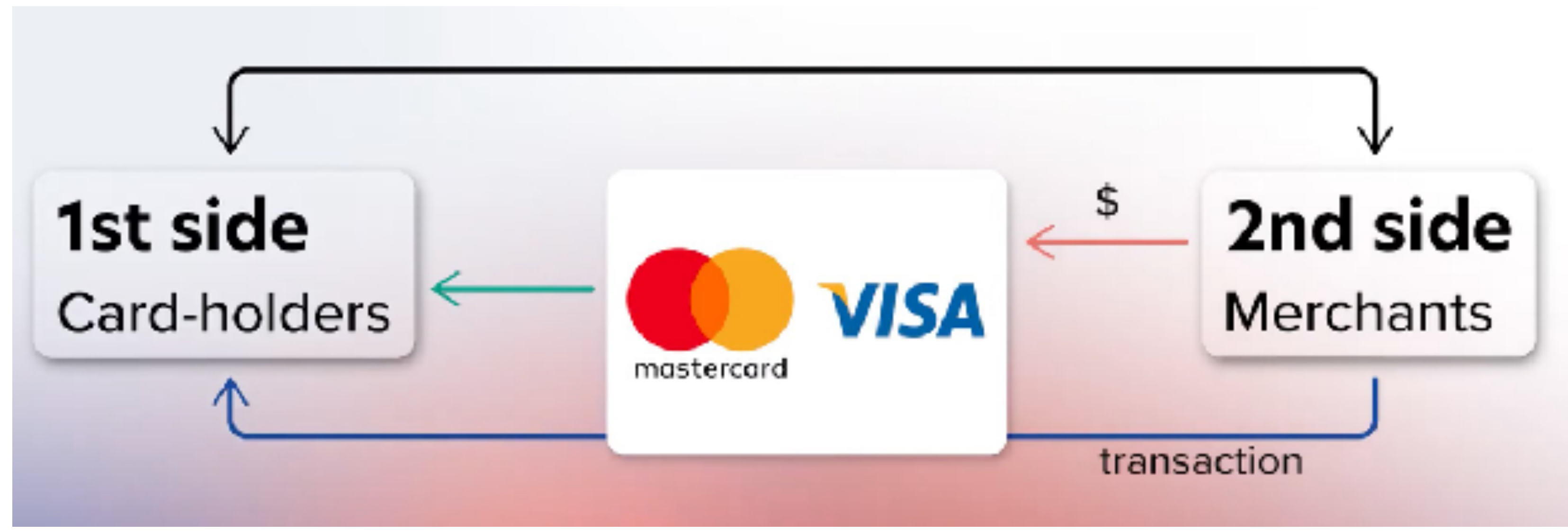
production

Economy of scale



Offer, prices, competition

The platform is a two-sided market (an economics concept)
(providers, consumers)
with cross-side network effects



ex. Uber, Airbnb, Youtube :
+ more consumers => better for providers
+ more providers => better for consumers
making links = their business model !

Difficult to start a business (chicken and egg start), hence huge existing systems.

Free of charge for one side helps create dynamics on this side.
Either one or the other side has to pay for the platform, sometimes both.

Recommender systems are often used for helping establish the link
between customer and content from an external content provider,
i.e. they operate on a plateform.

The recommender system is a core tool for building the quality of a platform

Stakeholders

Who does Spotify serve ?

- End-users ?
- Singers ?
- Spotify ?

Optimizing recommendation is optimization a function involving all 3

Crowdsourcing, description, evaluation

A society
of data crowdsourcing and « digital labor »

- recommender systems rely massively on content provided by the crowd of users, who « collaborate » to create a corpus that benefits to « all » fourniture de contenus à recommander
- Users provide:
  - content (stuff for others to read, view, buy,…)
  - meta-data over their content and other people's content (eg film topics)
  - Evaluations feed computation of reputation of everything on the plateform
    - of content
    - of people (peers)

A society
of data crowdsourcing and « digital labor »

- Plateforms present themselves as a community of peers+content that cooperate, to hide a centralized+commercial control and goals and favour free work by users
- They provide incentives (non financial) :
  - offer social visibility (better than IRL) to people, contents, opinion
  - social recognition of expertise and taste
  - provide everyone with numerical indicators of social position
  - privileged access to rights, content, jobs,….
- Sometimes digital labor is a real job, poorly paid as micro-tasks controlled by automata.

Listen to :
(france culture)

UNIVERSITÉ DE NANTES

Crowdsourcing

Is also a scientific topic about modelling, algorithms, maths, economics.

How to recruit and retain contributors ?

How to design and compute incentives ?

- evaluate quality of contributors and contributions
  - evaluation by peers, combined with
  - automated statistical analysis
- how to combine contributions ?
  - need to aggregate evaluations and to take contributor reputation into account in this process
  - this is a joint problem that needs to be solved iteratively

Tagging

Tagging is categorization
but categories may overlap (multiple tags)

Simple Man-Machine Interaction :
- values but no attributes (even implicit)
- values from open vocabulary (often)
- requires little user competence
- puts little constraints on user

-> incentive for tagging
-> suitable for crowdsourced tagging

- Tagging your own items
- Tagging existing items

-> incentive = creators are motivated to tag their content well so that it gets retrieved by other people

Tags

Ecomarathon • Eurospeedway Lausitz • Germany • Europe • 2010 • Polytech' Nantes • Polyjoule • prototype

Collaborative tagging :

Tags

Ecomarathon • Eurospeedway Lausitz • Germany • Europe • 2010 • Polytech' Nantes • Polyjoule • prototype

Tags given by user U1

Tags given by user U2

No constraint vocabulary :
incremental collaborative construction :
not only are tags crowdsourced,
but also vocabulary construction and maintenance.

Camion mecede 809
Enjoliveur mercedez
Interieur cuir noir mescedes 190
MERCEDEE E 320 avangarde
Mercedesse

Such a vocabulary (:=folksonomy) can be unstable

- bad for matching .
- but good for including new concepts with no central control
- overall, people tend to tag using popular tags (« rich get richer »)

Co-occurrence statistical analysis from tagged items

Goal : find pseudo-semantic relations between tags,
by computing statistical co-occurrences.

Ex. : Polytech and Polyjoule frequently present together on Flickr

Ex. : When tag *scipy* on Stackoverflow, *python* also present 90% of time
but when *python* is present, *scipy* present only 12% of time
=> *scipy* is a subpart of *python* (*scipy* is a famous *python* package)

These statistics are more reliable
if there are few tags used many times rather than many tags used once

=> Use these findings for query expansion

further possibility : consider <u>who</u> tags in the analysis

Analyze the (temporal) dynamics of tags

Goal : identify short events (vs. more stable entities)

ex. GreenCodingChallenge in Nantes with #GCLChallenge

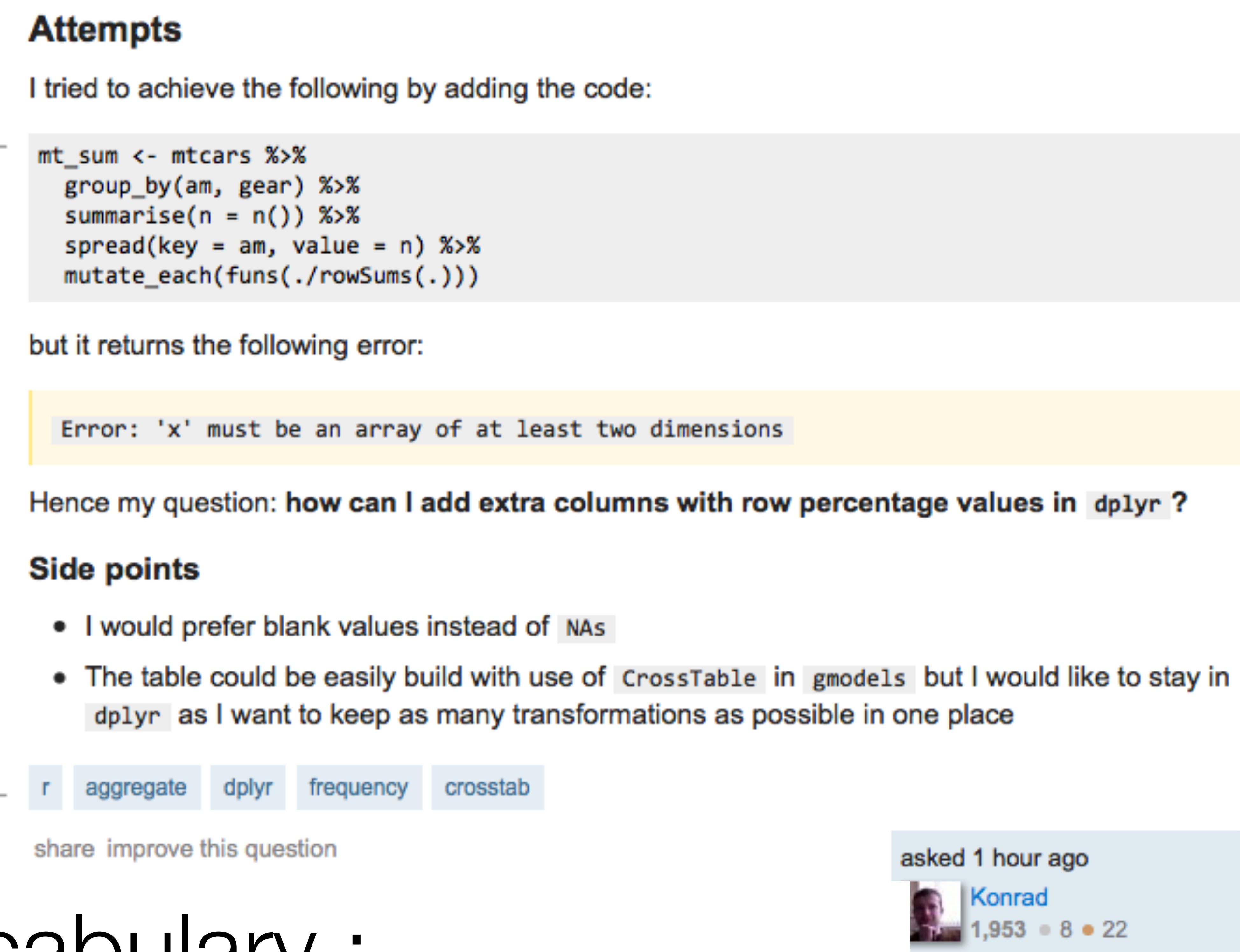
highly active during 3 days

can also find relation
to (GPS-like) geolocation, if available.

IUTNantes-INFO @IUTNantes_INFO 25m
#GCLChallenge @IUTNantes avec un peu de @CentraleNantes

IUTNantes-INFO @IUTNantes_INFO 27m
#GCLChallenge @IUTNantes

Re-introducing some control over the vocabulary
To further stabilize the vocabulary
systems may suggest tags (ex. Amazon, Stackoverflow)



Avoid divergence of vocabulary :

- You must use existing tag (stackoverflow when beginner)
- your question is more likely to be answered if you tag according to common practice

Describing objects :
automatically using external (web) resources
for processing annotations (correcting + extending). Ex. : wikipedia

Evaluating objects

- Ratings (1,2,3,4,5), good/not good ;   ordering (« better than »)
- Text (algorithms for «sentiment analysis» on blogs/facebook posts/tweets/ amazon «book damaged vs. uninteresting»)
- Unary value (bought/read/like),
  the alternative does not necessarily mean «bad» but maybe «no information»
  (various possible interpretations : not seen, already bought elsewhere, ...)
- I agree, I don't agree (evaluating an evaluation)
- Agregate multiple contributions (tags, evaluations) considering reputation